Nonparametric and Semiparametric Econometrics: Theory and Applications[1]

ZONGWU CAI

Department of Economics, University of Kansas
E-mail: caiz@ku.edu

December 10, 2025

---

[1]This manuscript may be printed and reproduced for individual or instructional use, but may not be printed for commercial purposes.

Preface

This is the advanced level of nonparametric econometrics with theory and applications. Here, the focus is on both the theory and methodology with the strong skills to analyze real data using nonparametric econometric techniques and statistical softwares such as R or Python. This is along the line with the spirit "STRONG THEORETICAL FOUNDATION with EXCELLENT SKILLS". In other words, this course covers the advanced topics in analyzing economic and financial data using nonparametric techniques, particularly in nonlinear time series models and some models related to economic and financial applications. The topics covered start from classical approaches to modern modeling techniques even up to the research frontiers. The difference between this course and others is that you will learn not only the theory but also step by step how to build a model based on data (or so-called *let data speak themselves*) through real data examples using statistical softwares or how to explore the real data using what you have learned. Therefore, there is no a single book serviced as a textbook for this course so that materials from some books and articles will be provided. However, some necessary handouts, including computer codes like R codes, will be provided.

Several projects (two or three), including the heavy computer coding works, are assigned throughout the term. The purpose of doing projects is to train students to understand the theoretical concepts and to know how to apply the methodologies learned in class to real problems. The group discussion is allowed to do the projects, particularly writing the computer codes. But, writing the final report to each project must be in your own language. If you use the R language, you can download it from the public web site at http://www.r-project.org/ and install it into your own computer. You are STRONGLY encouraged to use (but not limited to) the package R or Python since it is a very convenient programming language for doing statistical analysis and Monte Carol simulations as well as various applications in quantitative economics and finance.

Why do we need to study nonparametric and semiparametric econometrics? Here is a motivated example. For example, let us go back to the classical sample selection (Heckman) model in the books by Cameron and Trivedi (2005) and Dvidson and MacKinnon (2004). That is, the model setting is given by

$$y_t^0 = g_1(\mathbf{X}_t) + u_t, \quad \text{and} \quad z_t^0 = g_2(\mathbf{W}_t) + v_t.$$

We only observe data $\{y_t, \mathbf{X}_t, \mathbf{W}_t, z_t\}_{t=1}^n$, where $y_t = y_t^0$ if $z_t = 1$ with $z_t = I(z_t^0 > 0)$. Then, without normality assumption, we have the following model

$$\begin{aligned} \mathbb{E}(y_t \,|\mathbf{X}_t, \mathbf{W}_t) &= \mathbb{E}(y_t^0 |\, \mathbf{X}_t, \mathbf{W}_t, z_t = 1) = g_1(X_t) + \mathbb{E}(u_t |\, \mathbf{X}_t, \mathbf{W}_t, v_t > -g_2(\mathbf{W}_t)) \\ &= g_1(\mathbf{X}_t) + \mathbb{E}(u_t |\, \mathbf{W}_t, v_t > -g_2(\mathbf{W}_t) \equiv g_1(\mathbf{X}_t) + g_3(\mathbf{W}_t), \end{aligned} \tag{1}$$

which is an additive model, where $g_3(\mathbf{W}_t)$ denotes the second term on the right hand side in the above equation. Therefore, this is a generalization of the Heckman model (sample selection). To estimate $g_1(\mathbf{x})$, we need to learn nonparametric methods. If $g_1(\mathbf{X}_t) = \boldsymbol{\beta}^\top \mathbf{X}_t$, then, the above model becomes a semiparametric model. To estimate $\boldsymbol{\beta}$, one needs to learn the econometric tools for modeling semiparametric models.

Contents

1 Density, Distribution & Quantile Estimations **1**

1.1 Time Series Structure . . . 1
1.1.1 Mixing Conditions . . . 1
1.1.2 Martingale and Mixingale . . . 3
1.2 Kolmogorov-Smirnov Tests . . . 4
1.3 Kernel Density Estimation . . . 8
1.3.1 Estimation Procedure . . . 8
1.3.2 Optimality . . . 17
1.3.3 Data-Driven Bandwidth Selection Methods . . . 19
1.3.4 Boundary Problems . . . 23
1.3.5 Curse of Dimensionality . . . 26
1.4 Semiparametric Estimation of Density Function . . . 28
1.5 Theoretical Applications . . . 29
1.5.1 Distribution Estimation . . . 29
1.5.2 Quantile Estimation . . . 32
1.5.3 Value-at-Risk and Expected Shortfall . . . 32
1.5.4 Smoothed Quantile Estimation . . . 33
1.6 Computer Code . . . 35

2 Regression Models **44**

2.1 Instrocution . . . 44
2.2 Nadaraya-Watson Estimation . . . 47
2.2.1 Asymptotic Properties . . . 48
2.2.2 Boundary Behavior . . . 50
2.3 Local Polynomial Estimate . . . 51
2.3.1 Formulation . . . 51
2.3.2 Implementation in **R** and A Real Example . . . 52
2.3.3 Properties of Local Polynomial Estimator . . . 54
2.3.4 Complexity of Local Polynomial Estimator . . . 59
2.3.5 Bandwidth Selection . . . 61
2.4 Weighted Nadaraya-Watson Estimation . . . 65
2.5 Functional Coefficient Model . . . 67
2.5.1 Model and Its Properties . . . 67
2.5.2 Local Linear Estimation . . . 69
2.5.3 Smoothing Variable Selection and Bandwidth Selection . . . 70

2.5.4 Goodness-of-Fit Test . . . 70
2.5.5 Asymptotic Results . . . 73
2.5.6 Assumptions and Theoretical Proofs . . . 75
2.5.7 Applications . . . 82
2.6 Additive Model . . . 85
2.6.1 Model Framework . . . 85
2.6.2 Backfitting Algorithm . . . 87
2.6.3 Projection Method . . . 89
2.6.4 Two-Stage Procedure . . . 90
2.6.5 Analysis of the Boston House Price Data via Additive Model . . . 92
2.7 Semiparametric Models . . . 94
2.7.1 Partially Linear Models . . . 94
2.7.2 Single Index Models . . . 96
2.7.3 Functional Coefficient Index Models . . . 103
2.7.4 Distributional Index Models . . . 103
2.8 Time-Varying Coefficient Models . . . 104
2.9 Computer Codes . . . 106
2.9.1 Codes for Example 2.1 . . . 106
2.9.2 Codes for Additive Modeling Analysis of Boston Data . . . 111

3 Quantile Regression Models 113
3.1 Introduction . . . 113
3.2 Parametric Quantile Models . . . 119
3.3 Nonparametric Modeling Procedures . . . 122
3.3.1 Local Linear Quantile Estimate . . . 122
3.3.2 Asymptotic Results . . . 123
3.3.3 Bandwidth Selection . . . 128
3.3.4 Covariance Estimate . . . 130
3.3.5 Additive Quantile Regression Model . . . 132
3.4 Semiparametric Models . . . 132
3.5 Empirical Examples . . . 134
3.5.1 A Simulated Example . . . 134
3.5.2 Real Data Examples . . . 137
3.6 Mathematical Derivations . . . 146
3.6.1 Proofs of Main Results . . . 146
3.6.2 Proofs of Lemmas . . . 150
3.7 Composite Quantile Regression . . . 153

4 Nonparametric Measures of Risk 156
4.1 Introduction . . . 156
4.2 Nonparametric Model Setup . . . 159
4.3 Nonparametric Estimating Procedures . . . 161
4.3.1 Estimation of Conditional PDF and CDF . . . 161
4.3.2 Estimation of Conditional VaR and ES . . . 165
4.4 Asymptotic Theories . . . 165

4.4.1 Assumptions . . . 165
4.4.2 Asymptotic Properties for Conditional PDF and CDF . . . 167
4.4.3 Asymptotic Theory for CVaR and CES . . . 169
4.4.4 Theoretical Proofs . . . 173
4.5 Empirical Examples . . . 182
4.5.1 Bandwidth Selection . . . 182
4.5.2 Simulated Examples . . . 183
4.5.3 Real Examples . . . 188
4.6 Semiparametric Expectile Regressions . . . 191
4.6.1 Instruction . . . 191
4.6.2 Relationship Between Expectile and ES . . . 193
4.6.3 Estimation Procedures . . . 194

5 Nonparametric Models with Nonstationary Covariates 198
5.1 Introduction . . . 198
5.2 Statistical Properties . . . 200
5.2.1 Local Linear Estimation . . . 200
5.2.2 Notations and Assumptions . . . 201
5.2.3 Asymptotic Results . . . 203
5.2.4 Nadaraya-Watson Estimation . . . 205
5.3 Functional Coefficient Models . . . 206
5.4 Specification Tests . . . 207
5.4.1 Nonparametric Tests of Mean Function . . . 207
5.4.2 Nonparametric Test of Heteroskedasticity . . . 209
5.5 Empirical Applications . . . 211
5.6 Discussions . . . 215
5.7 Theoretical Proofs . . . 215

6 Nonparametric Estimation Equations 220
6.1 Introduction . . . 220
6.2 Estimating Equations . . . 221
6.2.1 Nonparametric Estimating Equations . . . 221
6.2.2 Examples . . . 224
6.3 Asymptotic Theory . . . 225
6.4 Theoretical Proofs . . . 227

7 Nonparametric and Semiparametric Models for Casual Inferences 234
7.1 Introduction . . . 234
7.2 Estimation Methods of ATE . . . 235
7.2.1 Propensity Score Based Approaches . . . 236
7.2.2 Covariate Balancing Methods . . . 238
7.2.3 CB Approach Based on Machine Learning Method . . . 239
7.2.4 Synthetic Control Methods . . . 254
7.2.5 Quasi-SC Method for Nonlinear Models . . . 256
7.2.6 Panel Data Approaches and Modified SC Methods . . . 260

7.3 Estimation of QTE 263
7.3.1 Unconditional QTE 263
7.3.2 Partially Conditional QTE 264
7.3.3 Nonparametric Tests of Heterogeneity 269
7.4 QTE for Panel Data 271
7.4.1 Model Setup 271
7.4.2 Inference Procedures 272
7.4.3 An Empirical Application 275
7.5 Estimating Counterfactual Distribution Functions 281
7.5.1 Estimation Approach 281
7.5.2 Implementation of Finding Optimal Weights 285
7.5.3 Extension to High-Dimensional Case 287
7.5.4 Testing Stochastic Dominance 288

List of Tables

1.1 The median and standard deviation in parentheses of 10,000 values of $f_n(x)$ for $\delta = 0$ (random walk case). 14
1.2 The median and standard deviation in parentheses of 10,000 values of $f_n(x)$ for $\delta = 1$ (nearly integrated case). 17
1.3 Sample sizes required for p-dimensional nonparametric estimate to have comparable performance with that of 1-dimensional nonparametric estimate using size $n_1 = 100$. 27

3.1 The Median and Standard Deviation of 500 MADE Values 136
3.2 The Post-Sample Predictive Intervals For Exchange Rate Data 147

7.1 The median (in the top panel) of biases and root mean square error (in the bottom panel) of various estimators under M1. 249
7.2 The median (in the top panel) of biases and root mean square error (in the bottom panel) of various estimators under M2. 250
7.3 The median (in the top panel) of biases and root mean square error (in the bottom panel) of various estimators under M3. 251
7.4 The median (in the top panel) of biases and root mean square error (in the bottom panel) of various estimators under M4. 252
7.5 A summary of the chosen covariates. 253
7.6 The number of rejections (Rej.) and acceptances (Accept) among 200 trials. 253
7.7 The performance of various estimators on the Twins dataset. 254
7.8 Descriptive statistics and symmetry testing results. 267
7.9 Test results for testing if PCQTE function changes over mother's age. 271
7.10 Descriptive Statistics of Monthly Return 277
7.11 Descriptive Statistics of Monthly Volatility 277

List of Figures

1.1 Left panel: Together with the true density (**black line**), bandwidth is taken to be 0.25 (red line), 0.5 (green line), 1.0 (blue line) and the optimal one (cyan line line, see later) with the Epanechnikov kernel. Right panel: the kernel density estimates for two different kernel functions: Gaussian (**lack line**) and Epanechnikov (red line). 12

1.2 The ACF and PACF plots for the original data (top panel), denoted by X_t, and the first difference (middle panel), denoted by $r_t = X_t - X_{t-1}$, which can be regarded as the simple return. The bottom left panel is for $f_n(x)$ (black solid line) by using the built-in function **density()** and the bottom right panel is for the own code, respectively, together with the density curve of the standard normal (red dashed line). 13

1.3 The kernel density estimator for the random walk case ($\delta = 0$). Top panel: Boxplots for $f_n(x)$ with bandwidth, $d = 0.5$; Middle panel: Boxplots for $f_n(x)$ with bandwidth, $d = 1$; and Bottom panel: Boxplots for $f_n(x)$ with bandwidth, $d = 2$. In all panels, n is taken to be 200, 1000 and 5000, and x is taken to be -5 (magenta), -2.5 (red), 0 (orange), 2.5 (blue) and 5 (green). 15

1.4 The kernel density estimator for the random walk case ($\delta = 0$). From the left to the right panel, $n = 200$, $n = 1000$, and $n = 5000$. In all panels, bandwidth $h = 0.5\, n^{-1/10}$ and x is taken to be -5 (magenta), -2.5 (red), 0 (orange), 2.5 (blue) and 5 (green). 16

1.5 The kernel density estimator for the random walk case ($\delta = 0$). From the left and the right panel, $n = 200$, $n = 1000$, and $n = 5000$. In all panels, bandwidth $h = 0.5\, n^{-1/10}$ and $x = a\sqrt{n}$ with a taken to be -0.5 (magenta), -0.25 (red), 0 (orange), 0.25 (blue) and 0.5 (green). 17

1.6 The Epanechnikov and Gaussian kernels. 19

2.1 The plot of three loss functions: quadratic loss (black solid line), Huber loss (red dashed line) with $M = 6$, the check function ($\tau = 0.05$, green dotted line), and the check function ($\tau = 0.90$, blue dashed-dotted line). 46

2.2 Scatterplots of ΔX_t, $|\Delta X_t|$, and $(\Delta X_t)^2$ versus $x(t) = X_t$ with the smoothed curves computed using **scatter.smooth()** and the local constant estimation. 54

2.3 Scatterplots of ΔX_t, $|\Delta X_t|$, and $(\Delta X_t)^2$ versus $x(t)$ with the smoothed curves computed using **scatter.smooth()** and the local linear estimation. 55

2.4 The results from model (2.55). 93

2.5 (a) Residual plot for model (2.55). (b) Plot of $g_1(x_6)$ versus x_6. (c) Residual plot for model (2.56). (d) Density estimate of Y. 93

3.1 For $\tau = 0.15$ and 0.90 and $\zeta = 0.10$, the plot of the check function and the smoothed check function in (a) and the plot of the difference of the smoothed check function and the check function in (b). 121

3.2 *Simulated Example*: The plots of the estimated coefficient functions for three quantiles $\tau = 0.05$ (dashed line), $\tau = 0.50$ (dotted line), and $\tau = 0.95$ (dot-dashed line) with their true functions (solid line): $\sigma(u)$ versus u in (a), $a_1(u)$ versus u in (b), and $a_2(u)$ versus u in (c), together with the 95% point-wise confidence interval (thick line) with the bias ignored for the $\tau = 0.5$ quantile estimate. 137

3.3 *Boston Housing Price Data*: Displayed in (a)-(d) are the scatter plots of the house price versus the covariates U, X_1, X_2 and $\log(X_2)$, respectively. 139

3.4 *Boston Housing Price Data*: The plots of the estimated coefficient functions for three quantiles $\tau = 0.05$ (solid line), $\tau = 0.50$ (dashed line), and $\tau = 0.95$ (dotted line), and the mean regression (dot-dashed line): $\widehat{a}_{0,\tau}(u)$ and $\widehat{a}_0(u)$ versus u in (e), $\widehat{a}_{1,\tau}(u)$ and $\widehat{a}_1(u)$ versus u in (f), and $\widehat{a}_{2,\tau}(u)$ and $\widehat{a}_2(u)$ versus u in (g). The thick dashed lines indicate the 95% point-wise confidence interval for the median estimate with the bias ignored. 140

3.5 *Exchange Rate Series:* (a) Japanese-dollar exchange rate return series $\{Y_t\}$; (b) autocorrelation function of $\{Y_t\}$; (c) moving average trading technique rule. 143

3.6 *Exchange Rate Series:* The plots of the estimated coefficient functions for three quantiles $\tau = 0.05$ (solid line), $\tau = 0.50$ (dashed line), and $\tau = 0.95$ (dotted line), and the mean regression (dot-dashed line): $\widehat{a}_{0,0.50}(u)$ and $\widehat{a}_0(u)$ versus u in (d), $\widehat{a}_{0,0.05}(u)$ and $\widehat{a}_{0,0.95}(u)$ versus u in (e), $\widehat{a}_{1,\tau}(u)$ and $\widehat{a}_1(u)$ versus u in (f), and $\widehat{a}_{2,\tau}(u)$ and $\widehat{a}_2(u)$ versus u in (g). The thick dashed lines indicate the 95% point-wise confidence interval for the median estimate with the bias ignored. 145

4.1 Simulation results for Example 4.1 when $p = 0.05$. Displayed in (a) - (c) are the true CVaR functions (solid lines), the estimated WDKLL CVaR functions (dashed lines), and the estimated NW CVaR functions (dotted lines) for $n = 250, 500$ and 1000, respectively. Box-plots of the 500 MADE values for both the WDKLL and NW estimations of CVaR are plotted in (d). 184

4.2 Simulation results for Example 4.1 when $p = 0.05$. Displayed in (a) - (c) are the true CES functions (solid lines), the estimated WDKLL CES functions (dashed lines), and the estimated NW CES functions (dotted lines) for $n = 250, 500$ and 1000, respectively. Box-plots of the 500 MADE values for both the WDKLL and NW estimations of CES are plotted in (d). 185

4.3 Simulation results for Example 4.1 when $p = 0.01$. Displayed in (a) - (c) are the true CVaR functions (solid lines), the estimated WDKLL CVaR functions (dashed lines), and the estimated NW CVaR functions (dotted lines) for $n = 250, 500$ and 1000 , respectively. Box-plots of the 500 MADE values for both WDKLL and NW estimation of the conditional VaR are plotted in (d). . . . 186

4.4 Simulation results for 4.1 when $p = 0.01$. Displayed in (a) - (c) are the true CES functions (solid lines), the estimated WDKLL CES functions (dashed lines), and the estimated NW CES functions (dotted lines) for $n = 250, 500$ and 1000, respectively. Box-plots of the 500 MADE values for both the WD-KLL and NW estimations of CVaR are plotted in (d). 187
4.5 Simulation results for Example 4.2 when $p = 0.05$. (a) Box-plots of MADEs for both the WDKLL and NW estimates for CVaR. (b) Box-plots of MADEs for Both the WDKLL and NW estimates for CES. 188
4.6 (a) 5% CVaR estimate for DJI index. (b) 5% CES estimate for DJI index. . 189
4.7 (a) 5% CVaR estimates for IBM stock returns. (b) 5% CES estimates for IBM stock returns index. (c) 5% CVaR estimates for three different values of lagged negative IBM returns $(-0.275, -0.025, 0.325)$. (d) 5% CVaR estimates for three different values of lagged negative DJI returns $(-0.225, 0.025, 0.425)$. (e) 5% CES estimates for three different values of lagged negative IBM returns $(-0.275, -0.025, 0.325)$. (f) 5% CES estimates for three different values of lagged negative DJI returns $(-0.225, 0.025, 0.425)$. 190

5.1 (a) Time series plot of Z_t; (b) ACF plot of Z_t; (c) ACF plot of $Z_t - Z_{t-1}$; (d) Estimated curve of $\beta(z)$. 213
5.2 Plots for $\widehat{\beta}_1(z)$ in the back solid line and $-\widehat{\beta}_2(z)$ in the red dashed line. The left panel is for the US-Canada case and the right panel is for the US-China case. 214

7.1 The kernel density estimation of infant birth weight for white. The solid line is for $Y(0)$ and the dotted line for $Y(1)$. The left panel is for whites and the right panel is for blacks. 267
7.2 Parametric estimation results for the partially conditional quantile treatment effects 268
7.3 Estimated PCQTEs for whites (the left panel) and blacks (th right panel) for three quantile levels $\tau = 0.10$, $\tau = 0.25$ and $\tau = 0.50$, respectively, together with the estimated unconditional 0.5-QTEs and their 95% confidence intervals. 269
7.4 The plot of the estimated density for the pre-treatment, post-treatment and whole sample VIX of CSI 300 index. 275
7.5 The plot of the estimated QTE is in the red line, $\widehat{\Delta}_\tau$ versus τ, together with its 95% CI (the shaded area) based on the blockwise Bootstrap proposed in Cai et al. (2026). The horizontal (blue) line is $\widehat{\Delta}_1$, the ATE calculated by the HCW's approach. 279
7.6 The plot of the estimated QTE is in the red line, $\widehat{\Delta}_\tau$ versus τ, together with its 95% CI (the red shaded area) based on the blockwise Bootstrap proposed in Cai et al. (2026). The horizontal (blue) line is $\widehat{\Delta}_1$, the ATE calculated by the HCW's approach. 280

Chapter 1

Density, Distribution & Quantile Estimations

1.1 Time Series Structure

Since most of economic and financial data are time series, we discuss our methodologies and theory under the framework of time series. For linear models, the time series structure can be often assumed to have some well known forms such as an autoregressive moving average (ARMA) model or nonlinear parametric time series models such as threshold autoregressive (TAR) model. However, under nonparametric setting, this assumption might not be valid. Therefore, we can assume a more general time series dependence, which is commonly used in the literature, described as follows.

1.1.1 Mixing Conditions

Mixing dependence is commonly used to characterize the dependent structure and it is often referred to as short range dependence or weak dependence, which means that the distance between two observations goes farther and farther, the dependence becomes weaker and weaker very faster. It is well known that α-mixing (strong mixing) includes many time series models as a special case. In fact, under very mild assumptions, linear processes, including linear autoregressive models and more generally bilinear time series models are α-mixing with mixing coefficients decaying exponentially. Many nonlinear time series models, such as functional coefficient autoregressive processes with/without exogenous variables, nonlinear additive autoregressive models with/without exogenous variables, ARCH and GARCH type processes, stochastic volatility models, and many continuous time diffusion models (including the Black-Scholes type models) are strong mixing under some mild conditions. See

Genon-Catalot et al. (2000), Cai and Masry (2000), Cai (2002a), Carrasco and Chen (2002), and Chen and Tang (2005) for more details.

To simplify notation, we only introduce mixing conditions for (strictly) stationary processes (in spite of the fact that a mixing process is not necessarily stationary), denoted by I(0). The idea is to define mixing coefficients to measure the strength (in different ways) of dependence for the two segments of a time series which are apart from each other in time. Let $\{X_t\}_{t=-\infty}^{\infty}$ be a strictly stationary time series. For $n \geq 1$, define

$$\alpha(n) = \sup_{A \in \mathcal{F}_{-\infty}^{0}; B \in \mathcal{F}_{n}^{\infty}} |\mathbb{P}(A)\mathbb{P}(B) - \mathbb{P}(AB)|,$$

where $\mathcal{F}_i^j$ denotes the σ-algebra generated by $\{X_t; i \leq t \leq j\}$. Note that $\mathcal{F}_n^{\infty} \downarrow$. If $\alpha(n) \to 0$ as $n \to \infty, \{X_t\}$ is called α-mixing or strong mixing. There are several other mixing conditions such as ρ-mixing, β-mixing, ϕ-mixing, and ψ-mixing; see the books by Hall and Heyde (1980) and Fan and Yao (2003) for details. Indeed,

$$\begin{aligned}
\beta(n) &= \mathbb{E}\left\{ \sup_{A \in \mathcal{F}_{n}^{\infty}} | \mathbb{P}(A) - \mathbb{P}\left(A \mid X_t, t \leq 0\right) \right\}, \\
\rho(n) &= \sup_{X \in \mathcal{F}_{-\infty}^{0}; Y \in \mathcal{F}_{n}^{\infty}} \mathrm{Corr}(X, Y)|, \\
\phi(n) &= \sup_{A \in \mathcal{F}_{-\infty}^{0}; B \in \mathcal{F}_{n}^{\infty}, \mathbb{P}(A) > 0} |\mathbb{P}(B) - \mathbb{P}(B \mid A)|,
\end{aligned}$$

and

$$\psi(n) = \sup_{A \in \mathcal{F}_{-\infty}^{0}; B \in \mathcal{F}_{n}^{\infty}, \mathbb{P}(A)\mathbb{P}(B) > 0} |1 - \mathbb{P}(B \mid A)/\mathbb{P}(B)|.$$

It is well known that the relationships among the mixing conditions are

$$\alpha(n) \leq \frac{1}{4}\rho(n) \leq \frac{1}{2}\phi(n),$$

so that ψ-mixing $\Longrightarrow$ ϕ-mixing $\Longrightarrow$ ρ-mixing $\Longrightarrow$ α-mixing as well as β-mixing $\Longrightarrow$ α-mixing. Note that all our theoretical results are derived under mixing conditions. Therefore, the following inequalities are very useful in our theoretical derivations, which can be found in the book by Hall and Heyde (1980).

Lemma 1.1: (Davydov's inequality) (i) If $E\left|X_t\right|^p + E\left|X_s\right|^q < \infty$ for some $p \geq 1$ and $q \geq 1$ and $1/p + 1/q < 1$, it holds that

$$|\mathrm{Cov}\left(X_i, X_s\right)| \leq 8\alpha^{1/r}(|s-t|)||X_t\left\|_{|}\right|X_s\|_q,$$

where $r = (1 - 1/p - 1/q)^{-1}$.
(ii) If $\mathbb{P}(|X_t| \leq C_1) = 1$ and $\mathbb{P}(|X_s| \leq C_2) = 1$ for some constants C_1 and C_2, it holds that

$$|\mathrm{Cov}(X_t, X_s)| \leq 4\alpha(|s-t|)C_1C_2$$

Note that if we allow X_t and X_s to be complex-valued random variables, (ii) still holds with the coefficient "4" on the right hand side of the inequality replaced by "16".
(iii) If $\mathbb{P}(|X_t| \leq C_1) = 1$ and $E|X_s|^p < \infty$ for some constants C_1 and $p > 1$, then,

$$|\mathrm{Cov}(X_t, X_s)| \leq 6C_1 \|X_j\|_p \alpha^{1-p^{-1}}(|s-t|).$$

Lemma 1.2: If $E|X_i|^p + E|X_j|^q < \infty$ for some $p \geq 1$ and $q \geq 1$ and $1/p + 1/q = 1$, it holds that

$$|\mathrm{Cov}(X_t, X_s)| \leq 2\phi^{1/p}(|s-t|)||X_t\|_p ||X_s\|_q .$$

1.1.2 Martingale and Mixingale

Martingale is very useful in applications. Here is the definition. Let $\{X_n, n \in \mathcal{N}\}$ be a sequence of random variables on a probability space $(\Omega, \mathcal{F}, P)$, and let $\{\mathcal{F}_n, n \in \mathcal{N}\}$ be an increasing sequence of sub-σ-fields of $\mathcal{F}$. Suppose that the sequence $\{X_n, n \in \mathcal{N}\}$ satisfies
(i) X_n is measurable with respect to $\mathcal{F}_n$,
(ii) $E|X_n| < \infty$,
(iii) $\mathbb{E}[X_n \mid \mathcal{F}_m] = X_m$ for all $m < n, n \in \mathcal{N}$.
Then, the sequence $\{X_n, n \in \mathcal{N}\}$ is said to be a martingale with respect to $\{\mathcal{F}_n, n \in \mathcal{N}\}$. We write that $\{X_n, \mathcal{F}_n, n \in \mathcal{N}\}$ is a martingale. If (i) and (ii) are retained and (iii) is replaced by the inequality $\mathbb{E}[X_n \mid \mathcal{F}_m] \geq X_m$ ($\mathbb{E}[X_n \mid \mathcal{F}_m] \leq X_m$), then $\{X_n, \mathcal{F}_n, n \in \mathcal{N}\}$ is called a sub-martingale (super-martingale). Define $Y_n = X_n - X_{n-1}$. Then $\{Y_n, \mathcal{F}_n, n \in \mathcal{N}\}$ is called a martingale difference (MD) if $\{X_n, \mathcal{F}_n, n \in \mathcal{N}\}$ is called a martingale. Clearly, $\mathbb{E}[Y_n \mid \mathcal{F}_{n-1}] = 0$, which means that a MD is not predicable based on the past information. In a finance language, a stock market is *efficient*. Equivalently, it is a MD.

Another type of dependent structure is called mixingale, which is the so-called asymptotic martingale. The concept of mixingale, introduced by McLeish (1975), is defined as follows. Let $\{X_n, n \geq 1\}$ be a sequence of square-integrable random variables on a probability space $(\Omega, \mathcal{F}, P)$, and let $\{\mathcal{F}_n, -\infty < n < \infty\}$ be an increasing sequence of sub-σ-fields of $\mathcal{F}$. Then,

$\{X_n, \mathcal{F}_n\}$ is called a L_r-mixingale (difference) sequence for $r \geq 1$ if, for some sequences of nonnegative constants c_n and ψ_m, where $\psi_m \to 0$ as $m \to \infty$, we have

$$(i) \quad \|\mathbb{E}(X_n \mid \mathcal{F}_{n-m})\|_r \leq \psi_m c_n, \quad \text{and} \quad (ii) \quad \|X_n - \mathbb{E}(X_n \mid \mathcal{F}_{n-m})\|_r \leq \psi_{m+1} c_n,$$

for all $n \geq 1$ and $m \geq 0$. The idea of mixingale is to try to build a bridge between martingale and mixing. The following examples give the idea of the scope of L_2-mixingale.

Examples:

1. A square-integrable martingale is a mixingale with $c_n = \|X_n\|$ and $\psi_0 = 1$ and $\psi_m = 0$ for $m \geq 1$.

2. A linear process is given by $X_n = \sum_{i=-\infty}^{\infty} \alpha_{i-n} \xi_i$ with $\{\xi_i\}$ independently and identically distributed (iid), mean zero, and variance σ^2 and $\sum_{i=-\infty}^{\infty} \alpha_i^2 < \infty$. Then, $\{X_n, \mathcal{F}_n\}$ is a mixingale with all $c_n = \sigma$ and $\psi_m^2 = \sum_{|i| \geq m} \alpha_i^2$.

3. If $\{X_n\}$ is a square-integrable sequence of ϕ-mixing, then it is a mixingale with $c_n = 2\|X_n\|_2$ and $\psi_m = \phi^{1/2}(m)$, where $\phi(m)$ is the ϕ-mixing coefficient.

4. If $\{X_n\}$ is a sequence of α-mixing with $\|X_n\|_p < \infty$ for some $p > 2$, then it is a mixingale with $c_n = 2(\sqrt{2}+1)\|X_n\|_2$ and $\psi_m = \alpha^{1/2-1/p}(m)$, where $\alpha(m)$ is the α-mixing coefficient.

Note that Examples 3 and 4 can be derived from the following inequality, due to McLeish (1975).

Lemma 1.3: (McLeish's inequality) Suppose that X is a random variable measurable with respect to $\mathcal{A}$, and $\|X\|_r < \infty$ for some $1 \leq p \leq r \leq \infty$. Then

$$\|\mathbb{E}(X \mid \mathcal{F}) - \mathbb{E}(X)\|_p \leq \begin{cases} 2[\phi(\mathcal{F}, \mathcal{A})]^{1-1/r} \|X\|_r, & \text{for } \phi\text{-mixing}, \\ 2\left(2^{1/p}+1\right)[\alpha(\mathcal{F}, \mathcal{A})]^{1/p-1/r} \|X\|_r, & \text{for } \alpha\text{-mixing}. \end{cases}$$

1.2 Kolmogorov-Smirnov Tests

The Kolmogorov-Smirnov (KS) test is one of the classical nonparametric statistical tests used to determine if a sample comes from a specific distribution (one-sample test) or if two samples come from the same distribution (two-sample test). It is based on the cumulative distribution function (CDF) and is calculated by finding the maximum vertical distance between two CDFs. The test is used to assess the "goodness-of-fit" between a sample and a hypothesized distribution or between two samples. It is named after Andrey Kolmogorov and Nikolai Smirnov, who developed it in the 1930s.

Let $\{X_t\}_{t=1}^n$ be a random sample with a (unknown) marginal distribution $F(\cdot)$ (CDF) and its probability density function (PDF) $f(\cdot)$. The question is how to estimate $f(\cdot)$ and $F(\cdot)$ nonparametrically. Since

$$F(x) = \mathbb{P}\left(X_t \leq x\right) = \mathbb{E}\left[I\left(X_t \leq x\right)\right] = \int_{-\infty}^{x} f(u)\, du,$$

by the method of moment estimation (MME), where I_A is the indicator function of any set A, $F(x)$ can be estimated by

$$F_n(x) = \frac{1}{n}\sum_{t=1}^{n} I\left(X_t \leq x\right), \tag{1.1}$$

which is called the empirical cumulative distribution function (ecdf). Next, we need to examine the asymptotic properties of $F_n(x)$. If $\{X_t\}$ is stationary, then, $\mathbb{E}\left[F_n(x)\right] = F(x)$ and

$$\begin{aligned} n\text{Var}\left(F_n(x)\right) =& \text{Var}\left(I\left(X_t \leq x\right)\right) + 2\sum_{t=2}^{n}\left(1 - \frac{t-1}{n}\right)\text{Cov}\left(I\left(X_1 \leq x\right), I\left(X_t \leq x\right)\right) \\ =& \underbrace{F(x)[1-F(x)] + 2\sum_{t=2}^{n}\text{Cov}\left(I\left(X_1 \leq x\right), I\left(X_t \leq x\right)\right)}_{\to \sigma_F^2(x) \quad \text{by assuming that } \sigma_F^2(x) < \infty} \\ & \underbrace{-2\sum_{t=2}^{n}\frac{t-1}{n}\text{Cov}\left(I\left(X_1 \leq x\right), I\left(X_t \leq x\right)\right)}_{\to 0 \text{ by Kronecker Lemma}} \\ \to\ & \sigma_F^2(x) \equiv F(x)[1-F(x)] + \underbrace{2\sum_{t=2}^{\infty}\text{Cov}\left(I\left(X_1 \leq x\right), I\left(X_t \leq x\right)\right)}_{\text{This term is called } A_d(x)}. \end{aligned}$$

Therefore,

$$n\text{Var}\left(F_n(x)\right) \ \to\ \sigma_F^2(x). \tag{1.2}$$

One can show based on the mixing theory that

$$\sqrt{n}\left[F_n(x) - F(x)\right] \ \xrightarrow{d}\ N\left(0, \sigma_F^2(x),\right), \tag{1.3}$$

where "$\xrightarrow{d}$" denotes the convergence in distribution, which can be derived in the same way as in the proof of Theorem 2.2 in Section 2.5; see Section 2.5.6 for details. It is clear that $A_d(x) = 0$ if $\{X_t\}$ are independent and identically distributed, so that $\sigma_F^2(x) = F(x)[1 -$

$F(x)]$. If $A_d(x) \neq 0$, the question is how to estimate it. For each given x, one can use the heteroskedasticity consistent (HC) estimator by White (1980) or the heteroskedasticity and autocorrelation consistent (HAC) estimator by Newey and West (1987) or the kernel version method of HAC by Andrews (1991).

The results in (1.3) can be used to construct a nonparametric test statistic to test the null hypothesis

$$H_0 : F(x) = F_0(x) \quad \text{versus} \quad H_a : F(x) \neq (>)(<)F_0(x), \tag{1.4}$$

where $F_0(\cdot)$ is a known distribution. A nonparametric test statistic is the well-known Kolmogorov-Smirnov test statistic, defined as

$$D_n = \sqrt{n} \sup_{-\infty<x<\infty} |F_n(x) - F_0(x)|$$

for the two-sided test. One can show, see, for example, Serfling (1980) or Billingsley (1999), that under some regularity conditions,

$$\mathbb{P}\left(D_n \leq d\right) \quad \to \quad 1 - 2\sum_{j=1}^{\infty}(-1)^{j+1}\exp\left(-2j^2d^2\right) \tag{1.5}$$

and

$$\mathbb{P}\left(D_n^+ \leq d\right) = \mathbb{P}\left(\sqrt{n}D_n^- \geq -d\right) \quad \to \quad 1 - \exp\left(-2d^2\right), \tag{1.6}$$

where $D_n^+ = \sqrt{n} \sup_{-\infty<x<\infty} [F_n(x) - F_0(x)]$ and $D_n^- = \sqrt{n} \sup_{-\infty<x<\infty} [F_0(x) - F_n(x)]$ for one-sided tests. In **R**, there is a built-in command for the Kolmogorov-Smirnov test, which is **ks.test()**.

Remark 1.1: *Note the most important assumptions on the classical Kolmogorov-Smirnov test formulated in* (1.4)*,* (1.5) *and* (1.6) *are as follows. First,* $F_0(\cdot)$ *in* (1.4) *must be known. Second, the results in* (1.5) *and* (1.6) *hold only for the iid data. If any one of the above assumptions is not satisfied, the classical Kolmogorov-Smirnov test results as in* (1.5) *and* (1.6) *do not hold.*

To overcome the above problems, let us consider the following hypothesis

$$H_0 : F(x) = F_0(x, \theta) \quad \text{versus} \quad H_a : F(x) \neq (>)(<)F_0(x, \theta),$$

where $F_0(\cdot)$ is a known distribution with unknown parameter θ. A nonparametric test statistic is defined as

$$D_{n,\theta} = \sqrt{n} \sup_{-\infty<x<\infty} \left| F_n(x) - F_0(x, \widehat{\theta}_n) \right|$$

for the two-sided test, where $\widehat{\theta}_n$ is the maximum likelihood estimate (MLE) of θ. Clearly, $D_{n,\theta}$ can be expressed

$$D_{n,\theta} = \sup_{-\infty<x<\infty} \left| \sqrt{n}(F_n(x) - F_0(x,\theta)) - \sqrt{n}(F_0(x, \widehat{\theta}_n) - F_0(x,\theta)) \right|.$$

For this case that the data are iid, Babu and Rao (2004) proposed a Bootstrap approach, termed as nonparametric Bootstrap Kolmogorov-Smirnov (NPBKS) test, to compute the p-value. For this end, using the Bootstrap (asymptotic) theory, we can approximate the distribution of $\sqrt{n}(F_n(x) - F_0(x,\theta))$ and $\sqrt{n}(F_0(x, \widehat{\theta}_n) - F_0(x,\theta))$ by that of $\sqrt{n}(F_n^{(b)}(x) - F_n(x))$ and $\sqrt{n}(F_0(x, \widehat{\theta}_n^{(b)}) - F_0(x, \widehat{\theta}_n))$, where $F_n^{(b)}(x)$ is the ECDF based on the b-th Bootstrap sample and $\widehat{\theta}_n^{(b)}$ is the MLE from the b-th Bootstrap sample for $b \in \{1, \cdots, B\}$ for some large B. Therefore, $D_{n,\theta}$ can be approximated by

$$D_{n,\theta}^{(b)} = \sup_{-\infty<x<\infty} \left| \sqrt{n} \left(F_n^{(b)}(x) - F_0(x, \widehat{\theta}_n^{(b)}) \right) - C_n(x) \right|, \tag{1.7}$$

where $C_n(x) = \sqrt{n}(F_n(x) - F_0(x, \widehat{\theta}_n))$ is regarded as the estimated bias term. The NPBKS procedure is to repeat the following steps for $b \in \{1, \cdots, B\}$.

1. Generate $X_1^{(b)}, \cdots, X_n^{(b)}$ by sampling $X_1, \cdots, X_n$ with replacement. Note that this step is the classical nonparametric Bootstrap procedure.
2. Obtain parametrically fitted parameters $\widehat{\theta}_n^{(b)}$ of θ from $X_1^{(b)}, \cdots, X_n^{(b)}$.
3. Obtain the empirical distribution function $F_n^{(b)}(x)$ of $X_1^{(b)}, \cdots, X_n^{(b)}$.
4. Calculate Bootstrap KS statistic $D_{n,\theta}^{(b)}$.

Then, the p-value of the basic Bootstrap KS test can be approximated as $p = \sum_{b=1}^{B} 1\{D_{n,\theta}^{(b)} > D_{n,\theta}\}/B$. Babu and Rao (2004) proved that for the iid data, $D_{n,\theta}^{(b)}$ and $D_{n,\theta}$ have the same limiting distribution for almost all samples $X_1, \cdots, X_n$.

Furthermore, Chandy et al. (2025) generalized the NPBKS test as in Babu and Rao (2004) to the time series case by using a nonparametric block Bootstrap procedure, termed NPBBKS, summarized as follows.

1. Generate $X_1^{(b)}, \cdots, X_n^{(b)}$ by applying circular block Bootstrap on the original sample as defined in Politis and Romano (1992).
2. Obtain parametrically fitted parameters $\widehat{\theta}_n^{(b)}$ of θ from $X_1^{(b)}, \cdots, X_n^{(b)}$.
3. Obtain the empirical distribution function $F_n^{(b)}(x)$ of $X_1^{(b)}, \cdots, X_n^{(b)}$.
4. Calculate Bootstrap KS statistic

$$B_{n,\theta}^{(b)} = \sup_{-\infty<x<\infty} \left| \sqrt{n} \left(F_n^{(b)}(x) - F_0(x, \widehat{\theta}_n^{(b)}) \right) - K_n(x) \right|,$$

where $K_n(x) = \sqrt{n} \left(\mathbb{E}^*[F_n^{(b)}(x)] - F_0(x, \theta^*) \right)$ is the estimated bias term, and the expected values $\mathbb{E}^*[F_n^{(b)}(x)]$ and $\theta^* = \mathbb{E}[\widehat{\theta}_n^{(b)}]$ can be approximated by, respectively, $\mathbb{E}_B^*[F_n^{(b)}(x)] = \sum_{b=1}^B F_n^{(b)}(x)/B$ and $\mathbb{E}_B[\widehat{\theta}_n^{(b)}] = \sum_{b=1}^B \widehat{\theta}_n^{(b)}/B$.

The, the p-value of the block Bootstrap KS test can be approximated as $p = \sum_{b=1}^B 1\{B_{n,\theta}^{(b)} > D_{n,\theta}\}/B$. Note that in the above circular block Bootstrap procedure, one can apply the plug-in approach for choosing the circular Bootstrap block size introduced in Politis and White (2004). Finally, by assuming that the time series $\{X_t\}$ is stationary and strongly mixing and using arguments similar to those in Künsch (1989), under certain conditions, Chandy et al. (2025) justified the validity of the NPBB procedure of the KS test. However, the choice of the block size might not be easy in practice.

1.3 Kernel Density Estimation

1.3.1 Estimation Procedure

Since $f(x) = F'(x)$, alternatively, $f(x)$ can be re-expressed as

$$f(x) = \lim_{h \downarrow 0} \frac{F(x+h) - F(x-h)}{2h} \approx \frac{F(x+h) - F(x-h)}{2h}$$

if h is very small, so that $f(x)$ can be estimated by using $F_n(x)$ in (1.1) as

$$f_n(x) = \frac{F_n(x+h) - F_n(x-h)}{2h} = \frac{1}{n} \sum_{t=1}^n K_h \left(X_t - x \right), \tag{1.8}$$

where $K(u) = I(|u| \leq 1)/2$ and $K_h(u) = K(u/h)/h$. Indeed, the kernel function $K(u)$ can be taken to be any **symmetric** density function. Here, h is called the bandwidth. Initially, $f_n(x)$ was proposed by Rosenblatt (1956) and Parzen (1962) explored its properties in detail. Therefore, it is called the Rosenblatt-Parzen kernel density estimate.

Remark 1.2: *Let* $R(h) = f(x) - [F(x+h) - F(x-h)/2h$ *so that* $f(x) = [F(x+h) - F(x-h)/2h + R(h)$. *Then,* $R(h) = O(h^2)$ *is the second order approximation of* $f(x)$ *if* h *is small and the second derivative of* $f(x)$ *is continuous. Therefore,* $f_n(x)$ *is not the unbiased estimate due to the approximation error.*

Exercise: Please show that $F_n(x)$ is an unbiased estimate of $F(x)$ but $f_n(x)$ is a biased estimate of $f(x)$. **Think about intuitively**

(1) why $f_n(x)$ is biased

(2) where the bias comes from

(3) why $K(\cdot)$ should be symmetric.

Next, we derive the asymptotic variance for $f_n(x)$. First, define $Z_t = K_h(X_t - x)$. Then,

$$\begin{aligned}\mathbb{E}[Z_1 Z_t] &= \iint K_h(u-x)K_h(v-x)f_{1,t}(u,v)dudv \\ &= \iint K(u)K(v)f_{1,t}(x+uh, x+vh)dudv \quad \to \quad f_{1,t}(x,x),\end{aligned}$$

where $f_{1,t}(u,v)$ is the joint density of (X_1, X_t), so that

$$\mathrm{Cov}(Z_1, Z_t) \quad \to \quad f_{1,t}(x,x) - f^2(x).$$

It is easy to show that

$$h\mathrm{Var}(Z_1) \quad \to \quad \nu_0(K)f(x),$$

where $\nu_j(K) = \int u^j K^2(u)du$ for $j \geq 0$. Therefore,

$$n\,h\,\mathrm{Var}(f_n(x)) = h\mathrm{Var}(Z_1) + \underbrace{2h\sum_{t=2}^{n}\left(1 - \frac{t-1}{n}\right)\mathrm{Cov}(Z_1, Z_t)}_{\equiv A_f \to 0 \quad \text{under some assumptions}} \quad \to \quad \nu_0(K)f(x)$$

To show that $A_f \to 0$, let $d_n \to \infty$ and $d_n h \to 0$. Then,

$$|A_f| \leq h\sum_{t=2}^{d_n}|\mathrm{Cov}(Z_1, Z_t)| + h\sum_{t=d_n+1}^{n}|\mathrm{Cov}(Z_1, Z_t)|.$$

For the first term, if $f_{1,i}(u,v) \leq M_1$, then, it is bounded by $h\,d_n = o(1)$. For the second term, we apply the Davydov's inequality (see Lemma 1.1) to obtain

$$h\sum_{t=d_n+1}^{n}|\mathrm{Cov}(Z_1, Z_t)| \leq M_2\sum_{t=d_n+1}^{n}\alpha(t)/h = O\left(d_n^{-\beta+1}h^{-1}\right)$$

if $\alpha(n) = O\left(n^{-\beta}\right)$ for some $\beta > 2$. If $d_n = O\left(h^{-2/\beta}\right)$, then, the second term is dominated by $O\left(h^{1-2/\beta}\right)$ which goes to 0 as $n \to \infty$. Hence,

$$n\,h\,\mathrm{Var}\left(f_n(x)\right) \;\;\to\;\; \nu_0(K)f(x) \equiv \sigma_f^2(x). \tag{1.9}$$

By a comparison of (1.2) and (1.9), one can see clearly that there is an infinity term involved $(A_d(x))$ in $\sigma_F^2(x)$ due to the dependence but the asymptotic variance in (1.9) is the same as that for the iid case (without the infinity term). We can establish the following asymptotic normality for $f_n(x)$ but the proof can be derived in the same way as in the proof of Theorem 2.2 in Section 2.5; see Section 2.5.6 for details.

Theorem 1.1: *Under some regularity conditions, we have*

$$\sqrt{n\,h}\left[f_n(x) - f(x) - \frac{h^2}{2}\mu_2(K)f''(x) + o_p\left(h^2\right)\right] \;\xrightarrow{d}\; N\left(0, \sigma_f^2(x)\right),$$

where the term $\frac{h^2}{2}\mu_2(K)f''(x)$ *is called the asymptotic bias with* $\mu_2(K) = \int u^2K(u)du$ *and* $\sigma_f^2(x)$ *is defined in* (2.10).

Remark 1.3: *Note that Theorem 1.1 can be proved by using the Linderburg-Feller or Lyapunov central limit theorem (CLT)*[1] *for triangular arrays, if* $\{X_t\}$ *are independent. But, for time series cases, the proof is different and is similar to that for Theorem 2.2 (see Section 2.5 later) so that you can follow the idea in Section 2.5 to establish Theorem 1.1 for time series cases.*

According to Theorem 1.1, $f_n(x) \;\to\; f(x)$ for each x as $n \to \infty$ so that $f_n(x) = O_p(1)$, when $\{X_t\}$ is stationary. However, when $\{X_t\}$ is a nonstationary process like a random walk (integrated process), denoted by I(1), defined in (1.11) later, then, one can show that $f_n(x) = O_p(1/\sqrt{n})$; see, for example, the papers by Phillips and Park (1998), Cai et al. (2009) and Cai (2011) for details. Thus, the order of magnitude of the density estimate $f_n(x)$ in the integrated case is smaller than in the stationary case when $n \to \infty$. This is explained by the fact that an integrated process like X_t eventually (as $t \to \infty$) has a bigger probability of being away from a given point x than a stationary process and the kernel function $K(\cdot)$ assigns smaller values to the more distant points. This has important implications for kernel

[1]The Lyapunov CLT says that if triangular arrays $\{Z_{nt}\}_{t=1}^n$ are independent, and the Lyapunov condition $\sum_{t=1}^n \mathbb{E}\left(|Z_{nt}|^{2+\delta}\right)/s_n^{2+\delta} \to 0$ for some $\delta > 0$ holds, where $s_n^2 = \sum_{t=1}^n \mathrm{Var}\left(X_{nt}\right), \sum_{t=1}^n \left[Z_{nt} - \mathbb{E}\left(Z_{nt}\right)\right]/s_n$ converges to the standard normal. See, for example, Serfling (1980) for details.

regression with nonstationary time series. In effect, this reduces the rate of convergence of the kernel estimate of the density function. Indeed, the asymptotic distribution of $f_n(x)$ for nonstationary $\{X_t\}$ is totally different from that in Theorem 1.1 for stationary case, which is given by

$$\sqrt{n}\, f_n(x) \xrightarrow{d} \xi$$

in distribution, where ξ is a non-normal random variable (a local time of a Brownian motion), and the rate of convergence of the kernel estimate of the density function is much slower than $\sqrt{n\,h}$ for the stationary case. See Theorem 3.1 in Phillips and Park (1998) and Lemma B.1 in Cai et al. (2009) for details. Indeed, from Lemma B.1 in Cai et al. (2009), one can see that

$$\xi = \begin{cases} L(1,0)/\sigma_u, & \text{if } \ x \ \text{ is fixed,} \\ L(1,a)/\sigma_u, & \text{if } \ x = a\sqrt{n} \ \text{ for any fixed } a, \end{cases}$$

where σ_u^2 is the variance of $u_t = X_t - X_{t-1}$, and $L(t,x)$ is the local time t of the standard Brownian motion at x, given by

$$L(t,x) = \lim_{\epsilon \downarrow 0} \frac{1}{2\,\epsilon} \int_0^t I(|W_u(s) - x| \le \epsilon) ds \tag{1.10}$$

with $W_u(t)$ being the standard Brownian motion generated by $\{u_t\}$. For the definition and its properties, see, for example, the book by Marcus and Rosen (2006) for details.

Exercise: First, by comparing (1.2) with (1.9), what can you observe? Second, if $\{X_t\}$ is a sequence of nonstationary (say, unit root) random variable, what does $f_n(x)$ estimate? Please think about this problem. You will be asked to do a simulation to see what you can observe in your next homework assignment.

Example 1.1: Let us examine how importance the choice of bandwidth is. The data $\{X_i\}_{i=1}^n$ are generated from $N(0,1)$ iid and $n = 300$. The grid points are taken to be $[-4, 4]$ with an increment $\Delta = 0.1$. Bandwidth is taken to be 0.25 (red line), 0.5 (green line) and 1.0 (blue), and h_{opt} (cyan line), given in (1.12) later, respectively, and the kernel can be the Epanechnikov kernel $K(u) = 0.75\,(1-u^2)\,I(|u| \le 1)$ or the Gaussian kernel $K(u) = \frac{1}{\sqrt{2\pi}} \exp\,(-u^2/2)$. Comparisons are given in Figure 1.1 (the left panel) for different choices of h. Note that the comparison between two kernels: Gaussian (black line) and Epanechnikov (red line) is displaced in the right panel of Figure 1.1. This simulation shows that the choice of bandwidth h is critical but the choice of $K(u)$ is not so sensitive.

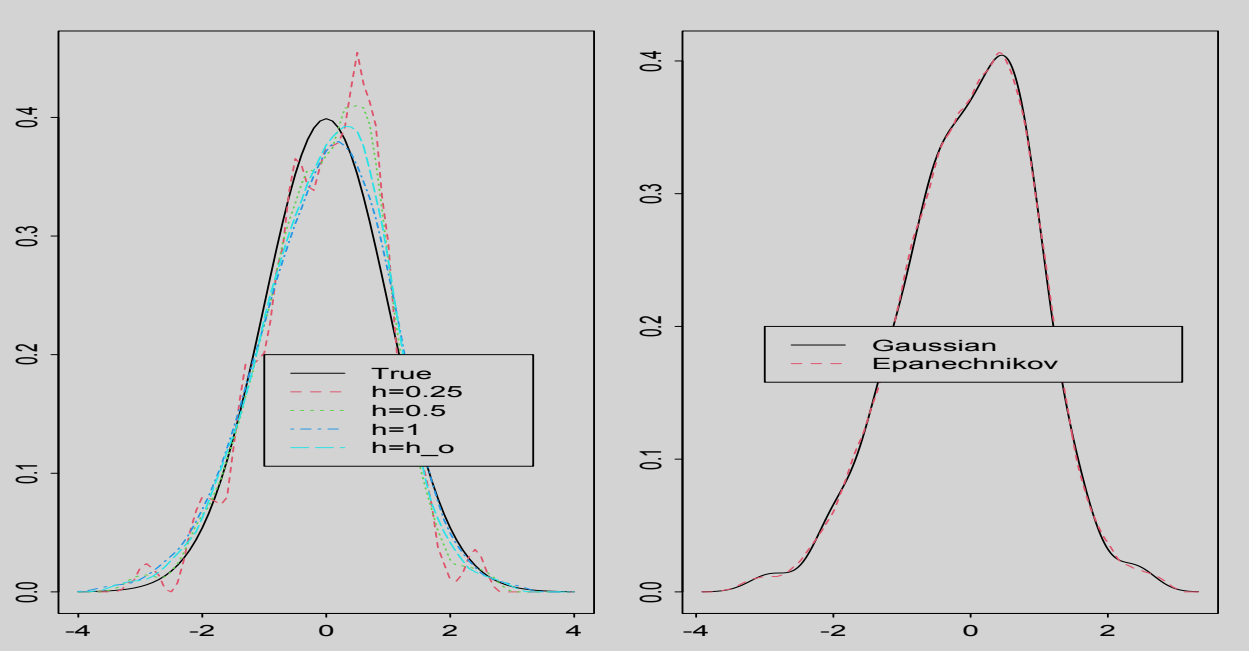

Figure 1.1: Left panel: Together with the true density (**black line**), bandwidth is taken to be 0.25 (red line), 0.5 (green line), 1.0 (blue line) and the optimal one (cyan line line, see later) with the Epanechnikov kernel. Right panel: the kernel density estimates for two different kernel functions: Gaussian (**lack line**) and Epanechnikov (red line).

Example 1.2: Next, we apply the kernel density estimation to estimate the density of the weekly 3-month Treasury bill (Secondary Market Rate, Discount Basis) from January 8, 1954 to September 23, 2022.[2] Figure 1.2 displays the ACF and PACF plots for the original data (top panel), denoted by X_t, and the first difference (middle panel), denoted by $r_t = X_t - X_{t-1}$, and the estimated density of the differencing series r_t together with the true standard normal density: the bottom left panel is for $f_n(x)$ (**black** solid line) by using the built-in function **density()** and the bottom right panel is for the own code, respectively, together with the with the density curve of the standard normal (red dashed line). From the top penal in Figure 1.2 first, one can conclude clearly that X_t is nonstationary (possible unit root) so that the differencing of X_t is is needed. Then, define $r_t = X_t - X_{t-1}$ and the ACF and PACF plots of $\{r_t\}$ are given in the middle panel of Figure 1.2 from which, one can see that r_t is autocorrelated. Finally, the bottom panel concludes that the distribution of r_t is not normal although its distribution looks symmetric and uni-mod. But, at 0, there is a high peak and two tails are heavy, which support the stylized factors about the distribution of the return.

[2]The dataset can be updated to today and can be downloaded from Federal Bank of St. Louis at https://fred.stlouisfed.org/series/DTB3.

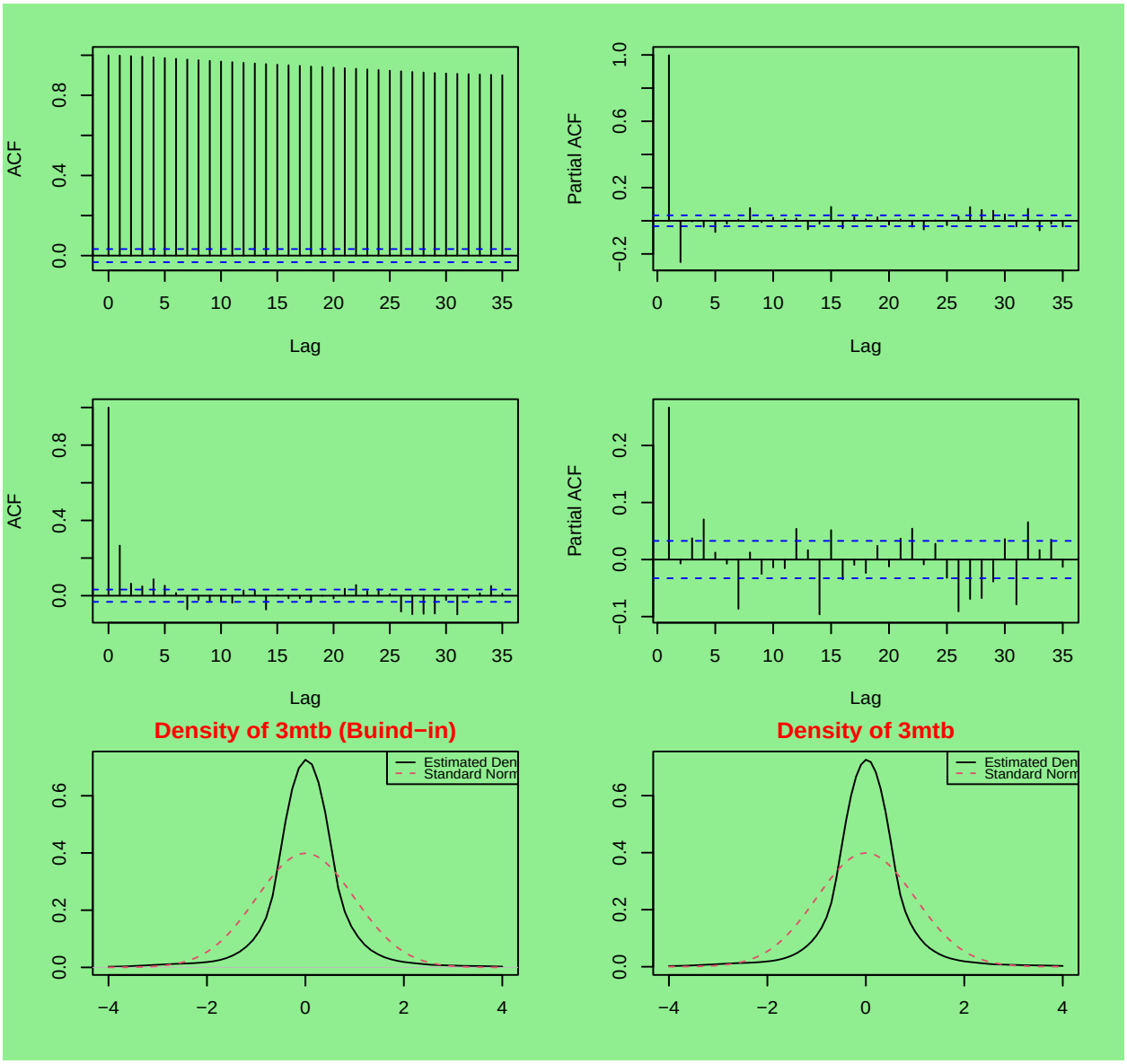

Figure 1.2: The ACF and PACF plots for the original data (top panel), denoted by X_t, and the first difference (middle panel), denoted by $r_t = X_t - X_{t-1}$, which can be regarded as the simple return. The bottom left panel is for $f_n(x)$ (black solid line) by using the built-in function **density()** and the bottom right panel is for the own code, respectively, together with the density curve of the standard normal (red dashed line).

Also, one can see that there is no difference between computings based on **density()** and the own code.

Example 1.3: In this example, we consider the case that $\{X_t\}_{t=1}^n$ is nonstationary and investigate the asymptotic properties of both $f_n(x)$ and $\sqrt{n} f_n(x)$. The data generating

process is

$$X_t = (1 - \delta/n)X_{t-1} + u_t \tag{1.11}$$

with $X_0 = 0$ for some $\delta \geq 0$. Here, we consider two cases $\delta = 0$ (random walk) and $\delta = 1$ (nearly integrated process or nearly random walk or nearly unit root), where $t = 1, \cdots, n$, and $u_t \overset{i.i.d}{\sim} N(0, 1)$. Bandwidth $h = d\, n^{-1/10}$ (see, for example, Phillips and Park (1998) and Cai et al. (2009) for details on why h is chosen as this form[3]) with d taken to be 0.5, 1 and 2, respectively. The sample size is taken to be 200, 1000 and 5000, respectively. For each setting, the simulation is repeated 10,000 times, and $f_n(x)$ calculated for x being fixed with -5 (magenta), -2.5 (red), 0 (orange), 2.5 (blue) and 5 (green), respectively. The simulation results are given in Figure 1.3 for boxplots and Table 1.1 for reporting the median and standard devision of the 10,000 values of $f_n(x)$ for each x, each sample size, and each d value, respectively. It is clear from both Figure 1.3 and Table 1.1, for each setting, $f_n(x)$ is closer to 0 as the sample size gets larger, which verifies the theory in Remark 1.3 that $f_n(x)$ converges to 0 as n goes to infinity. Note that as for how to choose empirically the optimal bandwidth h_{opt} for the nonstationary case, the reader is referred to the paper by Sun and Li (2011) for details.

Table 1.1: The median and standard deviation in parentheses of 10,000 values of $f_n(x)$ for $\delta = 0$ (random walk case).

The kernel density estimator for a random walk

	$d = 0.5$					$d = 1$					$d = 2$				
n	$x=-5$	$x=-2.5$	$x=0$	$x=2.5$	$x=5$	$x=-5$	$x=-2.5$	$x=0$	$x=2.5$	$x=5$	$x=-5$	$x=-2.5$	$x=0$	$x=2.5$	$x=5$
200	0.017	0.034	0.042	0.034	0.025	0.021	0.034	0.047	0.034	0.025	0.025	0.038	0.045	0.038	0.025
	(0.040)	(0.043)	(0.044)	(0.042)	(0.041)	(0.038)	(0.040)	(0.041)	(0.040)	(0.038)	(0.036)	(0.038)	(0.038)	(0.038)	(0.036)
1000	0.016	0.018	0.020	0.018	0.016	0.016	0.018	0.021	0.018	0.016	0.016	0.019	0.020	0.018	0.016
	(0.019)	(0.020)	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)
5000	0.008	0.009	0.009	0.009	0.008	0.009	0.009	0.010	0.009	0.009	0.009	0.009	0.009	0.009	0.009
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.008)	(0.008)	(0.009)	(0.009)	(0.008)

Next, for $h = 0.5\, n^{-1/10}$, $\sqrt{n}\, f_n(x)$ is calculated for x being fixed with -5 (magenta), -2.5 (red), 0 (orange), 2.5 (blue) and 5 (green) respectively. For each setting, the simulation is repeated 10,000 times. The estimated density curves are plotted in Figure 1.4, from which, one can observe that the estimated curves get closer to each other as the sample size gets

[3]Indeed, this is the optional bandwidth when X_t has the form as in (1.11).

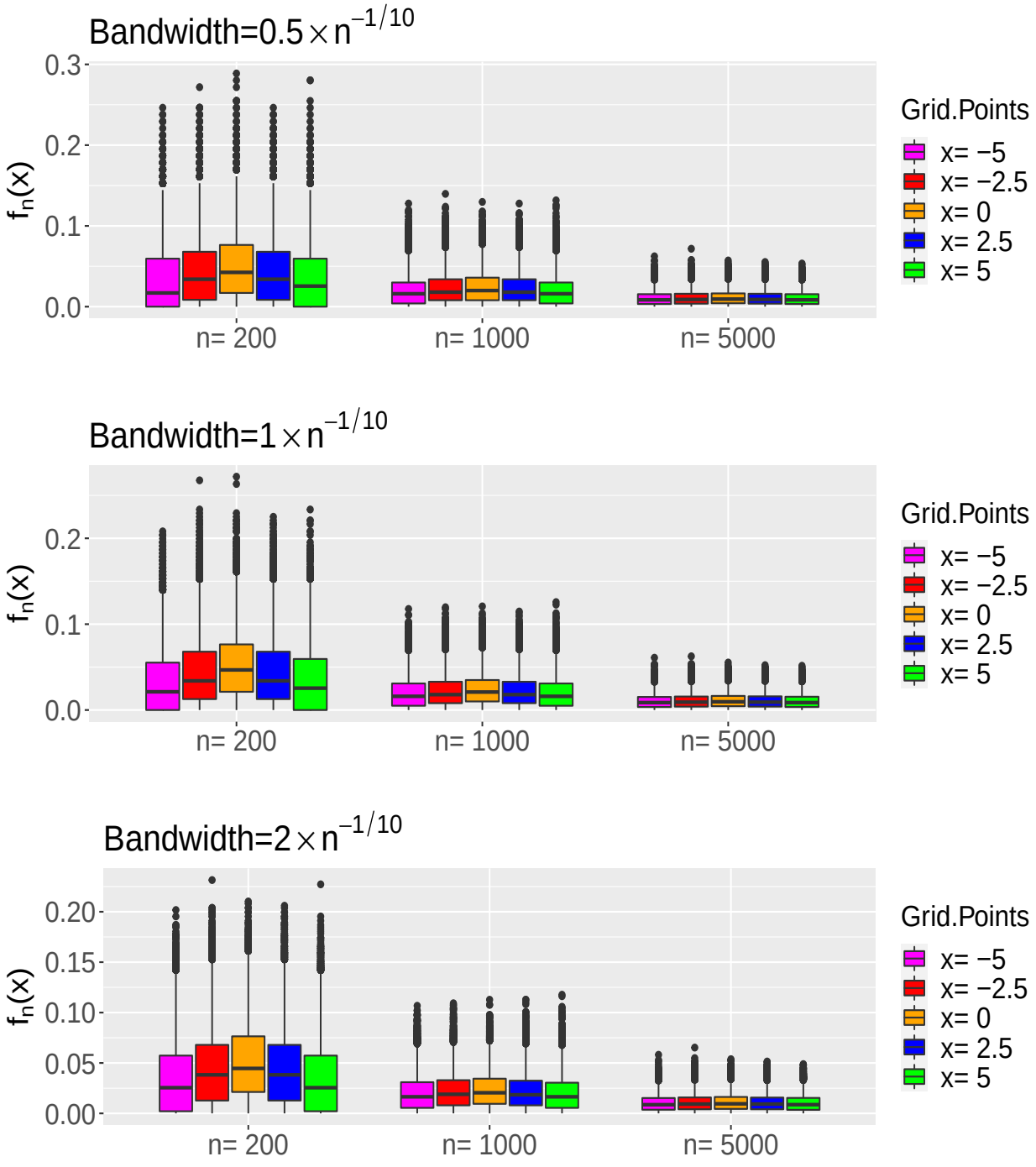

Figure 1.3: The kernel density estimator for the random walk case ($\delta = 0$). Top panel: Boxplots for $f_n(x)$ with bandwidth, $d = 0.5$; Middle panel: Boxplots for $f_n(x)$ with bandwidth, $d = 1$; and Bottom panel: Boxplots for $f_n(x)$ with bandwidth, $d = 2$. In all panels, n is taken to be 200, 1000 and 5000, and x is taken to be -5 (magenta), -2.5 (red), 0 (orange), 2.5 (blue) and 5 (green).

larger. Therefore, $\sqrt{n}\, f_n(x)$ can be used to approximate the distribution of the local time $L(1,0)$ of a Brownian motion.

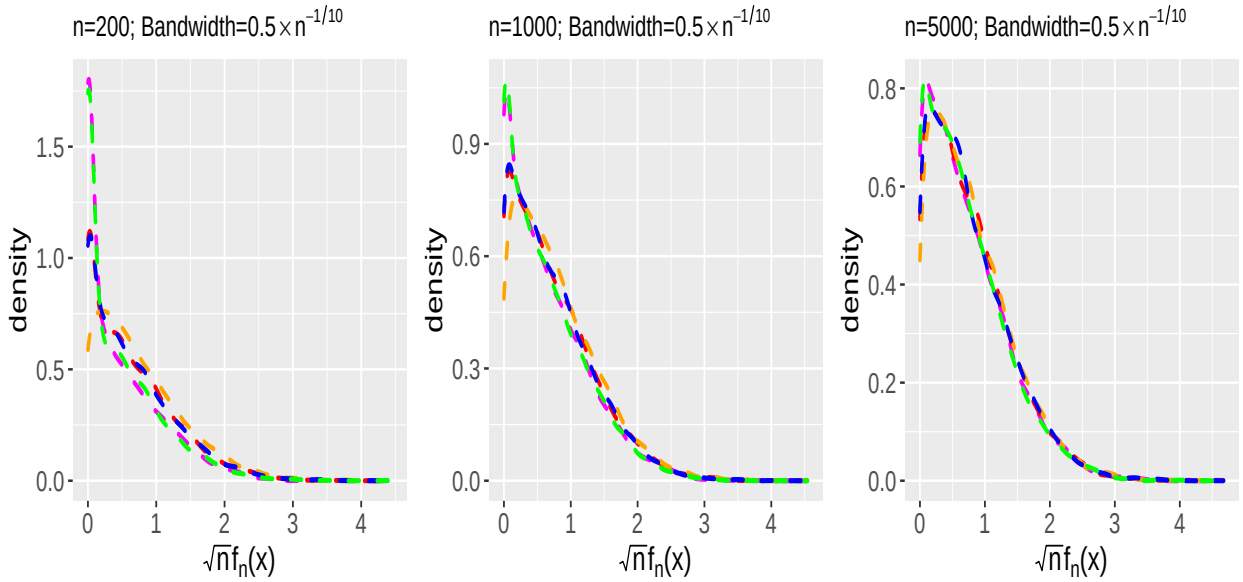

Figure 1.4: The kernel density estimator for the random walk case ($\delta = 0$). From the left to the right panel, $n = 200$, $n = 1000$, and $n = 5000$. In all panels, bandwidth $h = 0.5\, n^{-1/10}$ and x is taken to be -5 (magenta), -2.5 (red), 0 (orange), 2.5 (blue) and 5 (green).

Finally, repeat the above procedures with $x = a\sqrt{n}$ with a taken to be -0.5 (magenta), -0.25 (red), 0 (orange), 0.25 (blue) and 0.5 (green), respectively. The estimated density curves are displayed in Figure 1.5, from which, one can see that the estimated curves approximate different distributions as the sample size increases, which is dependent on the value of a. Therefore, $\sqrt{n}\, f_n(a\sqrt{n})$ can be used to approximate the distribution of the local time $L(1,a)$ for any $a \neq 0$.

Note that we also consider the case that $\delta = 1$ (nearly integrated case). The simulation results are presented in Table 1.2 for each setting and the same conclusions to those for a random walk scenario can be made. The figures similar to Figures 1.3 - 1.5 can re-produced in the same way, but, to save the space, they are not depicted here.

Note that the computer code in **R** for the above two examples can be found in Section 1.6. **R** has a built-in function **density()** for computing the nonparametric density estimation. Also, you can use the command **plot(density())** to plot the estimated density. Further, **R** has a built-in function **ecdf()** for computing the empirical cumulative distribution function estimation and **plot(ecdf())** for plotting the step function.

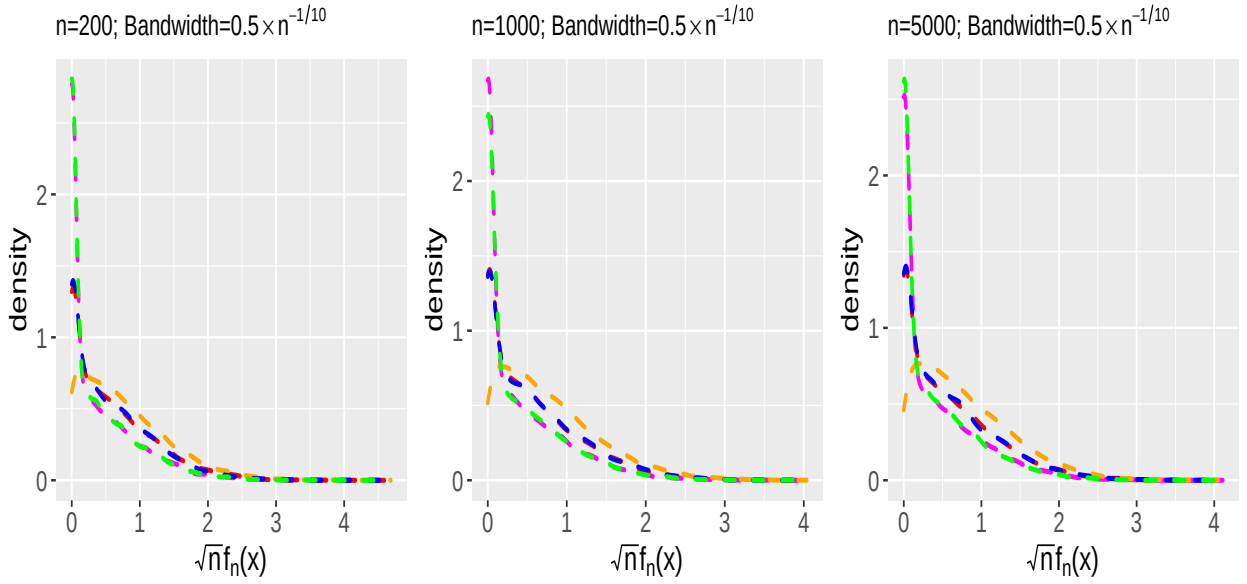

Figure 1.5: The kernel density estimator for the random walk case ($\delta = 0$). From the left and the right panel, $n = 200$, $n = 1000$, and $n = 5000$. In all panels, bandwidth $h = 0.5\, n^{-1/10}$ and $x = a\sqrt{n}$ with a taken to be -0.5 (magenta), -0.25 (red), 0 (orange), 0.25 (blue) and 0.5 (green).

Table 1.2: The median and standard deviation in parentheses of 10,000 values of $f_n(x)$ for $\delta = 1$ (nearly integrated case).

The kernel density estimator for a nearly random walk

	$d = 0.5$					$d = 1$					$d = 2$				
n	$x=-5$	$x=-2.5$	$x=0$	$x=2.5$	$x=5$	$x=-5$	$x=-2.5$	$x=0$	$x=2.5$	$x=5$	$x=-5$	$x=-2.5$	$x=0$	$x=2.5$	$x=5$
200	0.025	0.042	0.059	0.042	0.025	0.030	0.047	0.055	0.047	0.030	0.032	0.049	0.057	0.049	0.034
	(0.042)	(0.045)	(0.046)	(0.045)	(0.043)	(0.040)	(0.042)	(0.042)	(0.042)	(0.040)	(0.038)	(0.039)	(0.039)	(0.039)	(0.038)
1000	0.020	0.024	0.026	0.024	0.020	0.021	0.024	0.026	0.024	0.021	0.021	0.024	0.026	0.024	0.021
	(0.020)	(0.020)	(0.020)	(0.020)	(0.020)	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)
5000	0.011	0.012	0.012	0.011	0.011	0.011	0.011	0.012	0.011	0.011	0.011	0.012	0.012	0.011	0.011
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)

1.3.2 Optimality

As we already have shown that

$$\mathbb{E}\left(f_n(x)\right) = f(x) + \frac{h^2}{2}\mu_2(K) f''(x) + o\left(h^2\right), \quad \text{and} \quad \text{Var}\left(f_n(x)\right) = \frac{\nu_0(K) f(x)}{nh} + o\left((nh)^{-1}\right),$$

so that the asymptotic mean integrated squares error (AMISE) is

$$\text{AMISE} = \frac{h^4}{4}\mu_2^2(K) \int \left[f''(x)\right]^2 + \frac{\nu_0(K)}{nh}.$$

Minimizing the AMISE gives the

$$h_{\text{opt}} = C_1(K) \|f''\|_2^{-2/5} n^{-1/5} = d\, n^{-1/5}, \tag{1.12}$$

where $C_1(K) = [\nu_0(K)/\mu_2^2(K)]^{1/5}$ and $d = C_1(K) \|f''\|_2^{-2/5}$ depending on $K(\cdot)$ and $f(\cdot)$. With this asymptotically optimal bandwidth, the optimal AMISE is given by

$$\text{AMISE}_{\text{opt}} = \frac{5}{4} C_2(K) \|f''\|_2^{2/5} n^{-4/5},$$

where $C_2(K) = [\nu_0^2(K)\mu_2(K)]^{2/5}$. To choose the best kernel, it suffices to choose one to minimize $C_2(K)$.

Proposition 1: *The nonnegative probability density function K minimizing $C_2(K)$ is a re-scaling of the Epanechnikov kernel:*

$$K_{opt}(u) = \frac{3}{4a}\left(1 - u^2/a^2\right)_+$$

for any $a > 0$.

Proof: First of all, we note that $C_2(K_h) = C_2(K)$ for any $h > 0$. Let K_0 be the Epanechnikov kernel. For any other nonnegative K, by re-scaling if necessary, we assume that $\mu_2(K) = \mu_2(K_0)$. Thus, we need only to show that $\nu_0(K_0) \le \nu_0(K)$. Let $G = K - K_0$. Then,

$$\int G(u)\, du = 0 \text{ and } \int u^2 G(u)\, du = 0,$$

which implies that

$$\int \left(1 - u^2\right) G(u)\, du = 0.$$

Using this and the fact that $K_0(\cdot)$ has the support $[-1, 1]$, we have

$$\begin{aligned}\int G(u)K_0(u)\, du =& \frac{3}{4}\int_{|u|\le 1} G(u)\left(1 - u^2\right)\, du \\ =& -\frac{3}{4}\int_{|u|>1} G(u)\left(1 - u^2\right)\, du = \frac{3}{4}\int_{|u|>1} K(u)\left(u^2 - 1\right)\, du.\end{aligned}$$

Since $K(\cdot)$ is nonnegative, so is the last term. Therefore,

$$\int K^2(u)\, du = \int K_0^2(u)\, du + 2\int K_0(u)G(u)\, du + \int G^2(u)\, du \ge \int K_0^2(u)\, du,$$

which proves that $K_0(\cdot)$ is the optimal kernel. □

Remark 1.4: *This proposition implies that the Epanechnikov kernel with $a = 1$ should be used in practice. Clearly, the Epanechnikov kernel is symmetric and has a finite support as well as is differentiable within its support. The difference between the Epanechnikov and Gaussian kernels can be evidenced from Figure 1.6. As seen in Figure 1.1b, the difference of using two kernels to estimate $f(x)$ is not distinguishable.*

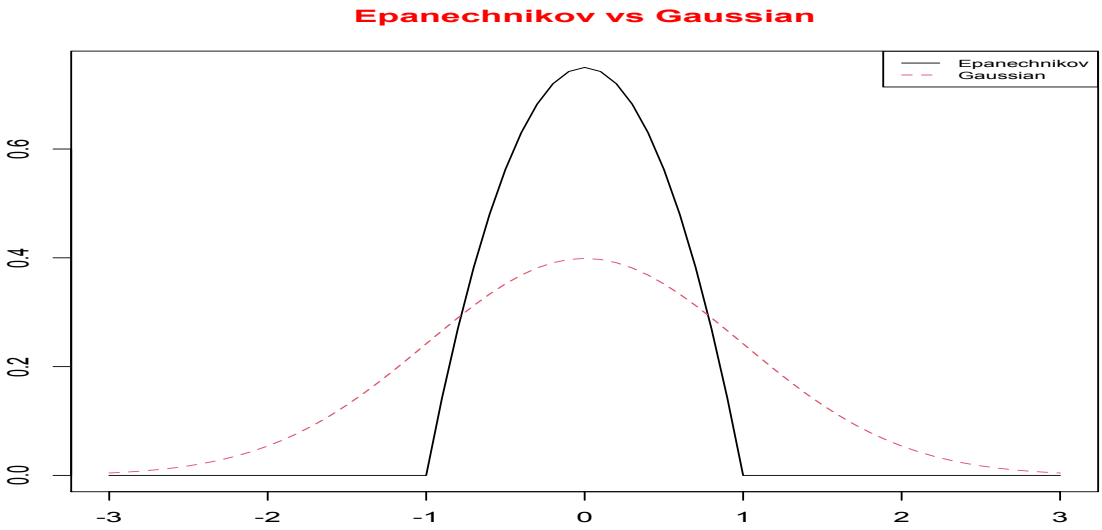

Figure 1.6: The Epanechnikov and Gaussian kernels.

1.3.3 Data-Driven Bandwidth Selection Methods

A. Simple Bandwidth Selectors

I. Normal Reference

The optimal bandwidth (1.12) is not directly usable since it depends on the unknown parameter $\|f''\|_2$. When $f(x)$ is a Gaussian density with standard deviation σ, it is easy to see from (1.12) that $\|f''\|_2^2 = 3/[8\sqrt{\pi}\sigma^5]$ so that

$$h_{opt} = (8\sqrt{\pi}/3)^{1/5} C_1(K)\sigma n^{-1/5},$$

which is called the normal reference bandwidth selector in literature, obtained by replacing the unknown parameter σ in the above equation by its sample standard deviation s. In particular, after calculating the constant $C_1(K)$ numerically, we have the following normal

reference bandwidth selector

$$\widehat{h}_{opt,n} = \begin{cases} 1.06\, s\, n^{-1/5} & \text{for the Gaussian kernel} \\ 2.34\, s\, n^{-1/5} & \text{for the Epanechnikov kernel.} \end{cases}$$

Clearly, if the true density of X_t is close to normal, then, the normal reference bandwidth selector should work well and it is often used in practice due to its simplicity. Of course, the true density of X_t is not close to normal, say, having multiple modes, then, the normal reference bandwidth selector should not be used.

If $f(x)$ has a unique mode, one might use Laplace (saddle-point) approximation to $f(x)$ as

$$f(x) \approx f(x_m)\sqrt{2\,\pi}\sigma_m\, \phi((x - x_m)/\sigma_m),$$

where x_m is the mode of $f(x)$, $\phi(x)$ is the density of the standard normal, and $\sigma^2 = -1/h'(x_m)$ with $h(x) = f'(x)/f(x)$. Then, one can use the normal reference bandwidth selector multiply a constance, which might need an estimate.

II. Edgeworth Expansion

Hjort and Jones (1996a) proposed an improved rule obtained by using an Edgeworth expansion for $f(x)$ around the Gaussian density. Such a rule is given by

$$\widehat{h}_{opt,e} = \widehat{h}_{opt,n}\left(1 + \frac{35}{48}\widehat{\gamma}_4 + \frac{35}{32}\widehat{\gamma}_3^2 + \frac{385}{1024}\widehat{\gamma}_4^2\right)^{-1/5},$$

where $\widehat{\gamma}_3$ and $\widehat{\gamma}_4$ are respectively the sample skewness and kurtosis. For details about the Edgeworth expansion, please see the book by Hall (1992).

III. Plug-in Method

Note that the normal reference bandwidth selector is only a simple rule of thumb. It is a good selector when the data are nearly Gaussian distributed, and is often reasonable in many applications. However, it can lead to over-smooth when the underlying distribution is asymmetric or multi-modal. In that case, one can either subjectively tune the bandwidth, or select the bandwidth by more sophisticated bandwidth selectors. One can also transform data first to make their distribution closer to normal, then estimate the density using the normal reference bandwidth selector and apply the inverse transform to obtain an estimated density for the original data. Such a method is called the transformation method. There

are quite a few important techniques for selecting the bandwidth such as cross-validation (CV) and plug-in bandwidth selectors. A conceptually simple technique, with theoretical justification and good empirical performance, is the plug-in technique. This technique relies on finding an estimate of the functional $\|f''\|_2$, which can be obtained by using a pilot bandwidth. An implementation of this approach is proposed by Sheather and Jones (1991) and an overview on the progress of bandwidth selection can be found in Jones et al. (1996).

Function **dpik()** in the package **KernSmooth** in **R** selects a bandwidth for estimating the kernel density estimation using the plug-in method.

IV. Cross-Validation Method

The integrated squared error (ISE) of $f_n(x)$ is defined by

$$\mathrm{ISE}(h) = \int [f_n(x) - f(x)]^2 \, dx.$$

A commonly used measure of discrepancy between $f_n(x)$ and $f(x)$ is the mean integrated squared error (MISE) $\mathrm{MSE}(h) = \mathbb{E}[\mathrm{ISE}(h)]$. It can be shown easily, see, e.g., Chiu (1991), that $\mathrm{MISE}(h) \approx \mathrm{AMISE}(h)$. The optimal bandwidth minimizing the AMISE is given in (1.12). The least squares cross-validation (LSCV) method proposed by Rudemo (1982) and Bowman (1984) is a popular method to estimate the optimal bandwidth h_{opt} . Cross-validation is very useful for assessing the performance of an estimator via estimating its prediction error. The basic idea is to set one of the data point aside for validation of a model and use the remaining data to build the model. The main idea is to choose h to minimize $\mathrm{ISE}(h)$. Since

$$\mathrm{ISE}(h) = \int f_n^2(x)dx - 2\int f(x)f_n(x)dx + \int f^2(x)dx,$$

the question is how to estimate the second term on the right hand side. Well, let us consider the simplest case when $\{X_t\}$ are iid. Re-express $f_n(x)$ as

$$f_n(x) = \frac{n-1}{n} f_n^{(-s)}(x) + \frac{1}{n} K_h\left(X_s - x\right)$$

for any $1 \le s \le n$, where

$$f_n^{(-s)}(x) = \frac{1}{n-1} \sum_{t \neq s}^{n} K_h\left(X_t - x\right),$$

which is the kernel density estimate without the sth observation, commonly called the **jackknife** estimate or leave-one-out estimate. It is easy to see that for any $1 \leq s \leq n$,

$$f_n(x) \approx f_n^{(-s)}(x).$$

Let $\mathcal{D}_{-s} = \{X_1, \cdots, X_{s-1}, X_{s+1}, \cdots, X_n\}$. Then,

$$\mathbb{E}\left[f_n^{(-s)}(X_s) \mid \mathcal{D}_{-s}\right] \equiv^4 \int f_n^{(-s)}(x) f(x) dx \approx \int f_n(x) f(x) dx, \tag{1.13}$$

if $\{X_t\}_{t=1}^n$ are iid, which, by using the method of moment, can be estimated by $\frac{1}{n}\sum_{s=1}^n f_n^{(-s)}(X_s)$. Therefore, the cross-validation is

$$\mathrm{CV}(h) = \int f_n^2(x) dx - \frac{2}{n}\sum_{s=1}^n f_n^{(-s)}(X_s) = \frac{1}{n^2}\sum_{s,t} K_h^*(X_s - X_t) - \frac{2}{n(n-1)}\sum_{t\neq s}^n K_h(X_s - X_t),$$

where $K^*(\cdot)$ is the convolution of $K(\cdot)$ and $K(\cdot)$ as

$$K^*(u) = \int K(v) K(u - v) dv. \tag{1.14}$$

Let $\widehat{h}_{cv}$ be the minimizer of $\mathrm{CV}(h)$. Then, it is called the optimal bandwidth based on the cross-validation. Stone (1984) showed that if $\{X_t\}_{t=1}^n$ are iid, $\widehat{h}_{cv}$ is a consistent estimate of the optimal bandwidth h_{opt} in the sense that $\widehat{h}_{cv}/h_{\text{opt}}$ converges to 1 in probability.

Function **lscv()** in the package **locfit** in **R** selects a bandwidth for estimating the kernel density estimation using the least squares cross-validation method.

Remark 1.5: *Note that the above cross-validation method does not work well for time seres cases since (1.13) holds only for the iid data. Indeed, the leave-one-out cross-validation method was challenged by Shao (1993), which claimed that the popular leave-one-out cross-validation method, which is asymptotically equivalent to many other model selection methods such as the Akaike Information Criterion (AIC), the C_p, and the Bootstrap, is asymptotically inconsistent in the sense that the probability of selecting the model with the best predictive ability does not converge to 1 as the total number of observations $n \to \infty$ and also, Shao (1993) showed that the inconsistency of the leave-one-out cross-validation can be rectified by using a leave-n_ν-out cross-validation with n_ν (block-wise cross-validation), the number of observations reserved for validation, satisfying $n_\nu/n \to 0$ and $n_\nu \to \infty$ as $n \to \infty$. The reader is referred to the paper by Shao (1993) for details. However, the choice of the block size n_ν is not an easy task in practice.*

[4] This equality holds only for the iid data but not for dependent data.

Finally, to pay attention to the structure of stationary time series data, one can use other data-driven methods to choose $\widehat{h}$ such as the nonparametric AIC[5] approach proposed in Cai and Tiwari (2000); see details in Section 2.3.5 or the modified multi-fold cross-validation criterion as in Cai et al. (2000); see Section 2.3.5 for details.

1.3.4 Boundary Problems

In many applications, the density $f(\cdot)$ has a bounded support. For example, the interest rate cannot be less than zero and the income is always nonnegative. It is reasonable to assume that the interest rate has support $[0, 1]$. However, because a kernel density estimator spreads smoothly point masses around the observed data points, some of those near the boundary of the support are distributed outside the support of the density. Therefore, the kernel density estimator under estimates the density in the boundary regions. The problem is more severe for large bandwidth and for the left boundary where the density is high. Therefore, some adjustments are needed. To gain some further insights, let us assume without loss of generality that the density function $f(\cdot)$ has a bounded support $[0, 1]$ and we deal with the density estimate at the left boundary. For simplicity, suppose that $K(\cdot)$ has a support $[-1, 1]$. For the left boundary point $x = ch (0 \leq c < 1)$, it can easily be seen that as $h \to 0$

$$\mathbb{E}\left(f_n(ch)\right) = \int_{-c}^{1/h-c} f(ch + hu)K(u)\, du = f(0+)\mu_{0,c}(K) + hf'(0+)\left[c\mu_{0,c}(K) + \mu_{1,c}(K)\right] + o(h), \tag{1.15}$$

where $f(0+) = \lim_{x \downarrow 0} f(x)$,

$$\mu_{j,c} = \int_{-c}^{\infty} u^j K(u)du, \quad \text{and} \quad \nu_{j,c}(K) = \int_{-c}^{\infty} u^j K^2(u)du.$$

Also, we can show that $\mathrm{Var}\left(f_n(ch)\right) = O(1/nh)$. Therefore,

$$f_n(ch) = f(0+)\mu_{0,c}(K) + hf'(0+)\left[c\mu_{0,c}(K) + \mu_{1,c}(K)\right] + o_p(h). \tag{1.16}$$

Particularly, if $c = 0$ and $K(\cdot)$ is symmetric, then $\mathbb{E}\left(f_n(0)\right) = f(0)/2 + o(1)$.

There are several methods to deal with the density estimation at boundary points. Possible approaches include the boundary kernel as in Gasser and Müller (979), reflection as Schuster (1985) and Hall and Wehrly (1991), transformation as in Wand et al. (1991) and

[5]For the detailed information, please see my lecture notes on **"A Summary of Classical and Modern Model Selection Methods"**.

Marron and Ruppert (1994), and local polynomial fitting as in Hjort and Jones (1996b) and Loader (1996), and others.

A. Boundary Kernel

One way of choosing a boundary kernel is to use the following boundary kernel

$$K_{(c)}(u) = \frac{12}{(1+c)^4}(1+u)\left\{(1-2c)u + \frac{3c^2-2c+1}{2}\right\} I_{[-1,c]}.$$

Note $K_{(1)}(t) = K(t)$, the Epanechnikov kernel as defined above. Moreover, Zhang and Karunamuni (1998) showed that this kernel is optimal in the sense of minimizing the MSE in the class of all kernels in order $(0, 2)$ with exactly one change of sign in their support. The downside to the boundary kernel is that it is not necessarily non-negative, as will be seen on densities where $f(0) = 0$. Actually, this boundary kernel is commonly used in applied research.

B. Reflection

The reflection method is to construct the kernel density estimate based on the synthetic data $\{\pm X_t; 1 \le t \le n\}$ where "reflected" data are $\{-X_t; 1 \le t \le n\}$ and the original data a re $\{X_t; 1 \le t \le n\}$. This results in the estimate

$$f_n(x) = \frac{1}{n}\left\{\sum_{t=1}^{n} K_h\left(X_t - x\right) + \sum_{t=1}^{n} K_h\left(-X_t - x\right)\right\}, \quad \text{for } x \ge 0.$$

Note that when x is away from the boundary, the second term in the above is practically negligible. Hence, it only corrects the estimate in the boundary region. This estimator is twice the kernel density estimate based on the synthetic data $\{\pm X_t; 1 \le t \le n\}$. See Schuster (1985) and Hall and Wehrly (1991).

C. Transformation

The transformation method is to first transform the data by $Y_t = g\left(X_t\right)$, where $g(\cdot)$ is a given monotone increasing function, ranging from $-\infty$ to ∞. Now apply the kernel density estimator to this transformed data set to obtain the estimate $f_n(y)$ for Y and apply the inverse transform to obtain the density of X. Therefore,

$$f_n(x) = |g^{\top}(x)|\frac{1}{n}\sum_{t=1}^{n} K_h\left(g\left(X_t\right) - g(x)\right).$$

The density at $x = 0$ corresponds to the tail density of the transformed data since $\log(0) = -\infty$, which can not usually be estimated well due to lack of the data at tails. Except at this point, the transformation method does a fairly good job. If $g(\cdot)$ is unknown in applications, similar to the Box-Cox transformation, Karunamuni and Alberts (2005) suggested a parametric form and then estimated the parameter by the profile likelihood estimation. Also, Karunamuni and Alberts (2005) considered other types of transformations.

D. Local Likelihood Fitting

The main idea is to consider the approximation $\log\left(f\left(X_t\right)\right) \approx \mathbb{P}\left(X_t - x\right)$, where $P(u-x) = \sum_{j=0}^{q} a_j(u-x)^j$ with the localized version of log-likelihood

$$\sum_{t=1}^{n} \log\left(f\left(X_t\right)\right) K_h\left(X_t - x\right) - n \int K_h(u-x) f(u) du.$$

With this approximation, the local likelihood proposed in Tibshirani and Hastie (1987) is employed here to estimate $f(x)$, as

$$\mathcal{L}\left(a_0, \cdots, d_q\right) = \sum_{t=1}^{n} \mathbb{P}\left(X_t - x\right) K_h\left(X_t - x\right) - n \int K_h(u-x) \exp(P(u-x)) du.$$

Let $\{\widehat{a}_j\}$ be the maximizer of the above local likelihood $\mathcal{L}\left(a_0, \cdots, d_q\right)$. Then, the local likelihood density estimate is

$$f_n(x) = \exp\left(\widehat{a}_0\right).$$

If the maximizer does not exist, then, $f_n(x) = 0$. See Loader (1996) and Hjort and Jones (1996b) for more details. Note that indeed, Loader (1996) and Hjort and Jones (1996b) showed that the local likelihood density estimate does not have the so-called boundary problem as described above. Finally, if **R** is used for the local fit for density estimation, please use the function **density.lf()** in the package **locfit**.

Exercise: Please conduct a Monte Carol simulation to see what the boundary effects are and how the correction methods work. For example, you can consider some distribution densities with a finite support such as beta-distribution.

Remark 1.6: *From Cai (2011), it is very interesting to know that the boundary problem does not exist for the Rosenblatt-Parzen estimator when X_t is nonstationary since X_t has a*

higher probability of taking very large values. Indeed, as argued by Cai (2011), for any fixed a, one has

$$\mathbb{P}(|X_t| \geq a\sqrt{n}) = \mathbb{P}(|X_t|/\sigma_u\sqrt{n} \geq a/\sigma_u) \approx \mathbb{P}(|W_u(r)| \geq a/\sigma_u) = 2[1 - \Phi(a/\sqrt{r}\sigma_u)] > 0,$$

where $t = rn$ for $0 < r < 1$, $W_u(\cdot)$ is the standard Brownian motion generated by $\{u_t\}$, and $\Phi(\cdot)$ is the CDF for the standard normal.

1.3.5 Curse of Dimensionality

As we discussed earlier, the kernel density or distribution estimation is basically one-dimensional. For the multivariate case $\mathbf{X}_t = (X_{1t}, \cdots, X_{pt}) \in \mathbb{R}^p$, the kernel density estimate in (1.8) is extended to the following; see, e.g., Robinson (1983),

$$f_n(\mathbf{x}) = \frac{1}{n}\sum_{t=1}^{n} K_H\left(\mathbf{X}_t - \mathbf{x}\right), \tag{1.17}$$

where $K_h(\mathbf{u}) = K\left(\mathbf{H}^{-1}\mathbf{u}\right)/\det(\mathbf{H}), K(\mathbf{u})$ is a multivariate kernel function, and $\mathbf{H}$ is the bandwidth matrix such as for all $1 \leq i, j \leq p, n\, h_{ij} \to \infty$ and $h_{ij} \to 0$ where h_{ij} is the (i, j) th element of $\mathbf{H}$. The bandwidth matrix is introduced to capture the dependent structure in the independent variables. Particularly, if $\mathbf{H}$ is a diagonal matrix and $K(\mathbf{u}) = \prod_{j=1}^{p} K_j\left(u_j\right)$ where $K_j(\cdot)$ is a univariate kernel function, then, $f_n(\mathbf{x})$ becomes

$$f_n(\mathbf{x}) = \frac{1}{n}\sum_{t=1}^{n}\prod_{j=1}^{p} K_{h_j}\left(X_{jt} - x_j\right),$$

which is called the product kernel density estimation. This case is commonly used in practice. Similar to the univariate case, it is easy to derive the theoretical results for the multivariate case, which is left as an exercise. See Wand and Jones (1994) for details.

For the product kernel estimate with $h_j = h$, we can show easily that

$$\mathbb{E}\left(f_n(\mathbf{x})\right) = f(\mathbf{x}) + \frac{h^2}{2}\operatorname{tr}\left(\mu_2(K)f''(\mathbf{x})\right) + o\left(h^2\right),$$

where $\mu_2(K) = \int \mathbf{u}\,\mathbf{u}^\top K(\mathbf{u})d\mathbf{u}$, $\operatorname{tr}(\mathbf{A})$ denoted the trace of matrix $\mathbf{A}$, and

$$\operatorname{Var}\left(f_n(\mathbf{x})\right) = \frac{\nu_0(K)f(\mathbf{x})}{nh^p} + o\left((nh^p)^{-1}\right),$$

so that the AMSE is given by

$$\text{AMSE} = \frac{\nu_0(K)f(\mathbf{x})}{nh^p} + \frac{h^4}{4}B(\mathbf{x}),$$

where $B(\mathbf{x}) = \left(\text{tr}\left(\mu_2(K)f''(\mathbf{x})\right)\right)^2$. By minimizing the AMSE, we obtain the optimal pointwise bandwidth

$$h_{\text{opt}} = \left(\frac{p\nu_0(K)f(x)}{B(\mathbf{x})}\right)^{1/(p+4)} n^{-1/(p+4)},$$

depending on $\mathbf{x}$, which leads to the optimal rate of convergence for MSE which is $O\left(n^{-4/(4+p)}\right)$ by trading off the rates between the bias and variance. When p is large, the so called *curse of dimensionality* exists. To understand this problem quantitatively, let us look at the rate of convergence. To have a comparable performance with one-dimensional nonparametric regression with n_1 data points, for p-dimensional nonparametric regression, we need the number of data points n_p,

$$O\left(n_p^{-4/(4+p)}\right) = O\left(n_1^{-4/5}\right),$$

or $n_] = O\left(n_1^{(p+4)/5}\right)$. Note that here we only emphasize on the rate of convergence for MSE by ignoring the constant part. Table 1.3 shows the result with $n_1 = 100$. The increase of required sample sizes for higher dimension is in a polynomial rate.

Table 1.3: Sample sizes required for p-dimensional nonparametric estimate to have comparable performance with that of 1-dimensional nonparametric estimate using size $n_1 = 100$.

dimension (p)	2	3	4	5	6	7	8	9	10
sample size (n_p)	252	631	1,585	3,982	10,000	25,119	63,096	158,490	398,108

Exercise: Please derive the asymptotic results given in (1.17) for the general multivariate case.

In **R**, the built-in function **density()** is only for univariate case. For multivariate situations, there are threepackages **ash**, **ks** and **KernSmooth**. Function **kde()** in **ks** can compute the multivariate density estimate for 2 to 6 dimensional data and Function **bkde2D()** in **KernSmooth** computes the 2D kernel density estimate. Also, **ks** provides some functions for some bandwidth matrix selection such as **Hbcv()** and **Hscv()** for 2D case and **Hlscv()** and **Hpi()**. As recommended by Deng and Wickham (2011), **ASH** and **KernSmooth** should be used due to thier excellence: they are both fast, accurate, and well-maintained.

Exercise: Please read the papers by Aït-Sahalia and Lo (1998), Pritsker (1998), Aït-Sahalia and Lo (2000), and Hong and Li (2005) on how to apply the kernel density estimation to the nonparametric estimation of the state-price densities (SPD) or risk neutral densities and

nonparametric risk estimation based on the state-price density. Please download the data from http://finance.yahoo.com/ (say, S&P500 index) to estimate the SPD.

1.4 Semiparametric Estimation of Density Function

To overcome the so-called *curse of dimensionality* as described in Section 1.3.5, Gallant and Nychka (1987) proposed the so-called semiparametric estimation to density function, termed as SNP estimation. Actually, the SNP method can be regarded a sieve or series method. The SNP estimator is based on the class of densities (for univariate case)

$$\mathcal{F}_K = \left\{ f_K : f_K(x, \boldsymbol{\theta}) = \left[\sum_{j=0}^{K} \theta_j x^j \right]^2 \exp(-x^2/2) + \epsilon_0 \phi(x), \boldsymbol{\theta} \in \Theta_K \right\},$$

where $\Theta_K = \left\{ \boldsymbol{\theta} : \boldsymbol{\theta} = (\theta_0, \theta_1, \cdots, \theta_K), \int f_K(x, \boldsymbol{\theta}) dx = 1 \right\}$, ϕ denotes the standard normal density, ϵ_0 is a small positive number, and $K = 0, 1, \cdots$. Estimation is by quasi-maximum likelihood based on the data $\{X_t\}_{t=1}^n$

$$\widehat{f}_K = \text{argmax}_{f \in \mathcal{F}_K} \sum_{t=1}^{n} \log \left[\frac{1}{\sigma} f \left(\frac{X_t - \mu}{\sigma} \right) \right],$$

where $K = K_n \to \infty$ in some way. It is convenient to rewrite the SNP density in terms of normalized Hermite polynomials

$$H_{e_j}(x) = (\sqrt{2\pi} j!)^{-1/2} \sum_{m=0}^{\lfloor j/2 \rfloor} (-1)^m \frac{j!}{m! 2^m (j-2m)!} x^{j-2m},$$

which is the so-called Hermite series expansion estimator of density function. Now, the above $\mathcal{F}_K$ becomes the following

$$\mathcal{F}_n = \left\{ f_n : f_n(x, \boldsymbol{\theta}) = \left[\sum_{j=0}^{K_n} \theta_j H_{e_j}(x) \right]^2 \exp(-x^2/2) + \epsilon_0 \phi(x), \boldsymbol{\theta} \in \Theta_n \right\}$$

with $\Theta_n = \left\{ \boldsymbol{\theta} : \boldsymbol{\theta} = (\theta_0, \theta_1, \cdots, \theta_{K_n}), \sum_{j=0}^{K_n} \theta_j^2 + \epsilon_0 = 1 \right\} \subset \mathbb{R}^{K_n+1}$. Clearly, the choice of K_n is very crucial, similar to that for h in Section 1.3.3. In practice, one can use the cross-validation described in Section 1.3.3 to choose K_n empirically, as suggested by Coppejansa and Gallant (2002). For details, please see the papers by Gallant and Nychka (1987) and Coppejansa and Gallant (2002) for the asymptotic theory and practical issues. Note that it is not difficult to generalize the above SNP method to the multivariate case $\mathbf{X}_t$. For details, please read the paper by Del Brio et al. (2010).

1.5 Theoretical Applications

In this section, for simplification, our focus is only on the univariate case; that is $p = 1$.

1.5.1 Distribution Estimation

A. Smoothed Distribution Estimation

The question is how to obtain a smoothed estimate of CDF $F(x)$. Well, one way of doing so is to integrate the estimated PDF $f_n(x)$, given by

$$\widehat{F}_n(x) = \int_{-\infty}^{x} f_n(u)du = \frac{1}{n}\sum_{i=1}^{n} \mathcal{K}\left(\frac{x - X_i}{h}\right), \tag{1.18}$$

where $\mathcal{K}(x) = \int_{-\infty}^{x} K(u)du$; the distribution of $K(\cdot)$. Why do we need this smoothed estimate of CDF? To answer this question, we need to consider the mean squares error.

First, we derive the asymptotic bias. By the integration by parts, we have

$$\begin{aligned}\mathbb{E}\left[\widehat{F}_n(x)\right] &= \mathbb{E}\left[\mathcal{K}\left(\frac{x - X_i}{h}\right)\right] = \int F(x - hu)K(u)du \\ &= F(x) + \frac{h^2}{2}\mu_2(K)f'(x) + o\left(h^2\right)\end{aligned}$$

Next, we derive the asymptotic variance.

$$\mathbb{E}\left[\mathcal{K}^2\left(\frac{x - X_i}{h}\right)\right] = \int F(x - hu)b(u)du = F(x) - hf(x)\theta + o(h),$$

where $b(u) = 2K(u)\mathcal{K}(u)$ and $\theta = \int ub(u)du$. Then,

$$\text{Var}\left[\mathcal{K}\left(\frac{x - X_i}{h}\right)\right] = F(x)[1 - F(x)] - hf(x)\theta + o(h).$$

Define $I_j(x) = \text{Cov}\left(I\left(X_1 \leq x\right), I\left(X_{j+1} \leq t\right)\right) = F_j(x, x) - F^2(x)$ and

$$I_{nj}(x) = \text{Cov}\left(\mathcal{K}\left(\frac{x - X_1}{h}\right), \mathcal{K}\left(\frac{x - X_{j+1}}{h}\right)\right).$$

By means of Lemma 2 in Lehmann (1966), the covariance $I_{nj}(x)$ may be written as follows

$$\begin{aligned}I_{nj}(t) = \int \Bigg\{ & P\left[\mathcal{K}\left(\frac{x - X_1}{h}\right) > u, \mathcal{K}\left(\frac{x - X_{j+1}}{h}\right) > v\right] \\ & - P\left[\mathcal{K}\left(\frac{x - X_1}{h}\right) > u\right] P\left[\mathcal{K}\left(\frac{x - X_{j+1}}{h}\right) > v\right]\Bigg\} dudv.\end{aligned}$$

Inverting the CDF, $\mathcal{K}(\cdot)$ and making two changes of variables, the above relation becomes

$$I_{nj}(x) = \int \left[F_j(x-hu, x-hv) - F(x-hu)F(x-hv)\right] K(u)K(v)dudv.$$

Expanding the right-hand side of the above equation according to Taylor's formula, we obtain

$$|I_{nj}(x) - I_j(x)| \leq C\, h^2.$$

By the Davydov's inequality (see Lemma1.1), we have

$$|I_{nj}(x) - I_j(x)| \leq C\, \alpha(j),$$

so that for any $1/2 < \tau < 1$,

$$|I_{nj}(x) - I_j(x)| \leq C\, h^{2\tau}\, \alpha^{1-\tau}(j).$$

Therefore,

$$\frac{1}{n}\sum_{j=1}^{n-1}(n-j)\,|I_{nj}(x) - I_j(x)| \leq \sum_{j=1}^{n-1}|I_{nj}(x) - I_j(x)| \leq C\, h^{2\tau}\sum_{j=1}^{\infty}\alpha^{1-\tau}(j) = O\left(h^{2\tau}\right)$$

provided that $\sum_{j=1}^{\infty}\alpha^{1-\tau}(j) < \infty$ for some $1/2 < \tau < 1$. Indeed, this assumption is satisfied if $\alpha(n) = O\left(n^{-\beta}\right)$ for some $\beta > 2$. By the stationarity, it is clear that

$$n\text{Var}\left(\widehat{F}_n(x)\right) = \text{Var}\left(\mathcal{K}\left(\frac{x - X_1}{h}\right)\right) + \frac{2}{n}\sum_{j=1}^{n-1}(n-j)I_{nj}(x).$$

Therefore,

$$\begin{aligned} n\text{Var}\left(\widehat{F}_n(x)\right) &= F(x)[1-F(x)] - hf(x)\theta + o(h) + 2\sum_{j=1}^{\infty} I_j(x) + O\left(h^{2\tau}\right) \\ &= \sigma_F^2(x) - hf(x)\theta + o(h). \end{aligned}$$

We can establish the following asymptotic normality for $\widehat{F}_n(x)$ but the proof will be discussed later.

Theorem 1.2: *Under some regularity conditions, we have*

$$\sqrt{n}\left[\widehat{F}_n(x) - F(x) - \frac{h^2}{2}\mu_2(K)f'(x) + o_p\left(h^2\right)\right] \xrightarrow{d} N\left(0, \sigma_F^2(x)\right).$$

Similarly, we have

$$n\, AMSE\left(\widehat{F}_n(x)\right) = \frac{nh^4}{4}\mu_2^2(K)\left[f'(x)\right]^2 + \sigma_F^2(x) - hf(x)\theta.$$

If $\theta > 0$, minimizing the AMSE gives the

$$h_{opt} = \left(\frac{\theta f(x)}{\mu_2^2(K) \left[f'(x) \right]^2} \right)^{1/3} n^{-1/3},$$

and with this asymptotically optimal bandwidth, the optimal AMSE is given by

$$n\, AMSE_{opt} \left(\widehat{F}_n(x) \right) = \sigma_F^2(x) - \frac{3}{4} \left(\frac{\theta^2 f^2(x)}{\mu_2(K) f'(x)} \right)^{2/3} n^{-1/3}.$$

Remark 1.7: *From the aforementioned equation, we can see that if $\theta > 0$, the AMSE of $\widehat{F}_n(x)$ can be smaller than that for $F_n(x)$ in the second order. Also, it is easy to that if $K(\cdot)$ is the Epanechnikov kernel, $\theta > 0$.*

B. Relative Efficiency and Deficiency

To measure the relative efficiency and deficiency of $\widehat{F}_n(x)$ over $F_n(x)$, we define

$$i(n) = \min \left\{ k \in \{1, 2, \cdots\}; \mathrm{MSE}\left(F_k(x)\right) \leq \mathrm{MSE}\left(\widehat{F}_n(x) \right) \right\}$$

We have the following results without the detailed proofs which can be found in Cai and Roussas (1998).

Proposition 2: *(i) Under some regularity conditions,*

$$\frac{i(n)}{n} \;\to\; 1, \quad \textit{if and only if} \quad nh_n^4 \;\to\; 0.$$

(ii) Under some regularity conditions,

$$\frac{i(n) - n}{n\, h} \;\to\; \theta(x), \quad \textit{if and only if} \quad h_n^3 \;\to\; 0,$$

where $\theta(x) = f(x)\theta / \sigma_F^2(x)$.

Remark 1.8: *It is clear that the quantity $\theta(x)$ may be looked upon as a way of measuring the performance of the estimate $\widehat{F}_n(x)$. Suppose that the kernel $K(\cdot)$ is chosen, so that $\theta > 0$, which is equivalent to $\theta(x) > 0$. Then, for sufficiently large n, $i(n) > n + nh_n(\theta(x) - \varepsilon)$. Thus, $i(n)$ is substantially larger than n, and, indeed, $i(n) - n$ tends to ∞. Actually, Reiss (1981) and Falk (1983) posed the question of determining the exact value of the superiority of θ over a certain class of kernels. More specifically, let $\mathcal{K}_m$ be the class of kernels $\mathcal{K} : [-1, 1] \to \Re$ which are absolutely continuous and satisfy the requirements: $\mathcal{K}(-1) = 0, \mathcal{K}(1) = 1$, and $\int_{-1}^{1} u^\mu K(u) du = 0, \mu = 1, \cdots, m$, for some $m = 0, 1, \cdots$ (where the moment condition is*

vacuous for $m = 0$). Set $\Psi_m = \sup\{\theta; \mathcal{K} \in \mathcal{K}_m\}$. Then, Mammitzsch (1984) answered the question posed by showing in an elegant manner. See Cai and Roussas (1998) for more details and simulation results.

Exercise: Please conduct a Monte Carol simulation to see what the differences are for smoothed and non-smoothed distribution estimations.

1.5.2 Quantile Estimation

It is well known in the literature that hat the quantile for a multivariate case is not uniquely defined; see, e.g., the papers by Cai (2010a) and Galvao and Montes-Rojas (2025) for details, so that in section, we consider only the univariate case. To this end, let $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ denote the order statistics of $\{X_t\}_{t=1}^n$. Define the inverse of $F(x)$ as $F^{-1}(\tau) = \inf\{x \in \mathbb{R}; F(x) \geq \tau\}$. The traditional estimate of $F(x)$ has been the empirical distribution function $F_n(x)$ based on $X_1, \cdots, X_n$, as defined in (1.1) in Section 1.2, while the estimate of the τ-th quantile $\xi_\tau = F^{-1}(\tau)$, $0 < \tau < 1$, is the sample quantile function $\xi_{\tau,n} = F_n^{-1}(\tau) = X_{([n\tau])}$, where $[x]$ denotes the integer part of x. It is a consistent estimator of ξ_τ for α-mixing data; see, e.g., Yoshihara (1995). However, as stated in Falk (1983), $F_n(x)$ does not take into account the smoothness of $F(x)$; i.e., the existence of a probability density function $f(x)$. In order to incorporate this characteristic, investigators proposed several smoothed quantile estimates, one of which is based on $\widehat{F}_n(x)$ obtained as a convolution between $F_n(x)$ and a properly scaled kernel function; see the previous section. Finally, note that **R** has a command **quantile()** which can be used for computing $\xi_{\tau,n}$, the nonparametric estimate of quantile.

1.5.3 Value-at-Risk and Expected Shortfall

Value at Risk (VaR) is a popular measure of market risk associated with an asset or a portfolio of assets. It has been chosen by the Basel Committee on Banking Supervision as a benchmark risk measure and has been used by financial institutions for asset management and minimization of risk. Let $\{X_t\}_{t=1}^n$ be the market value of an asset over n periods of $t = 1$ a time unit, and let $Y_t = -\log(X_t/X_{t-1})$ be the negative log-returns (loss). Suppose $\{Y_t\}_{j=1}^n$ is a strictly stationary dependent process with marginal distribution function $F(y)$. Given a positive value τ close to zero, the $1-\tau$ level VaR is

$$\nu_\tau = \inf\{u : F(u) \geq 1 - \tau\} = F^{-1}(1-\tau), \tag{1.19}$$

which specifies the smallest amount of loss such that the probability of the loss in market value being larger than ν_τ is less than τ. Comprehensive discussions on VaR are available in Duffie and Pan (1997) and Jorion (2001), and references therein. Therefore, VaR can be regarded as a special case of quantile. **R** has a built-in package called **cvar** for a set of methods for calculation of VaR, particularly, for some parametric models such as the General Pareto Distribution (GPD). But, it is well known in the literature that the restrict parametric specifications might be misspecified.

A more general form for the generalized Pareto distribution with shape parameter $k \neq 0$, scale parameter σ, and threshold parameter θ, is

$$f(x) = \frac{1}{\sigma}\left(1 + k\frac{x-\theta}{\sigma}\right)^{-1/k-1}, \quad \text{and} \quad F(x) = 1 - \left(1 + k\frac{x-\theta}{\sigma}\right)^{-1/k}$$

for $\theta < x$, when $k > 0$. In the limit for $k = 0$, the density is $f(x) = \frac{1}{\sigma}\exp(-(x-\theta)/\sigma)$ for $\theta < x$. If $k = 0$ and $\theta = 0$, the generalized Pareto distribution is equivalent to the exponential distribution. If $k > 0$ and $\theta = \sigma$, the generalized Pareto distribution is equivalent to the Pareto distribution.

Another popular risk measure is the expected shortfall (ES) which is the expected loss, given that the loss is at least as large as some given quantile of the loss distribution (e.g., VaR), defined as

$$\mu_\tau = \mathbb{E}\left(Y_t \mid Y_t > \nu_\tau\right) = \int_{\nu_\tau}^{\infty} y f(y) dy/\tau. \tag{1.20}$$

It is well known from Artzner et al. (1999) that ES is a coherent risk measure such as it satisfies the four axioms: homogeneity (increasing the size of a portfolio by a factor should scale its risk measure by the same factor), monotonicity (a portfolio must have greater risk if it has systematically lower values than another), risk-free condition or translation invariance (adding some amount of cash to a portfolio should reduce its risk by the same amount), and subadditivity (the risk of a portfolio must be less than the sum of separate risks or merging portfolios cannot increase risk). VaR satisfies homogeneity, monotonicity, and risk-free condition but is not sub-additive. See Artzner et al. (1999) for details.

1.5.4 Smoothed Quantile Estimation

The smoothed sample quantile estimate of $\xi_\tau, \widehat{\xi_\tau}$, based on $\widehat{F}_n(x)$, is defined by:

$$\widehat{\xi_\tau} = \widehat{F}_n^{-1}(\tau) = \inf\left\{x \in \mathbb{R}; \widehat{F}_n(x) \geq \tau\right\}.$$

$\widehat{\xi}_\tau$ is referred to in literature as the perturbed (smoothed) sample quantile. Asymptotic properties of $\widehat{\xi}_\tau$, both under independence as well as under certain modes of dependence, have been investigated extensively in literature; see Cai and Roussas (1997) and Chen and Tang (2005).

By the differentiability of $\widehat{F}_n(x)$, we use the Taylor expansion and ignore the higher terms to obtain

$$\widehat{F}_n\left(\widehat{\xi}_\tau\right) = \tau \approx \widehat{F}_n\left(\xi_\tau\right) - f_n\left(\xi_\tau\right)\left(\widehat{\xi}_\tau - \xi_\tau\right),$$

then,

$$\widehat{\xi}_\tau - \xi_\tau \approx \left[\widehat{F}_n\left(\xi_\tau\right) - \tau)\right] / f_n\left(\xi_\tau\right) \approx \left[\widehat{F}_n\left(\xi_\tau\right) - \tau)\right] / f\left(\xi_\tau\right),$$

since $f_n(x)$ is a consistent estimator of $f(x)$. As an application of Theorem 1.2, we can establish the following theorem for the asymptotic normality of $\widehat{\xi}_\tau$ but the proof is omitted since it is similar to that for Theorem 1.2.

Theorem 1.3: *Under some regularity conditions, we have*

$$\sqrt{n}\left[\widehat{\xi}_\tau - \xi_\tau - \frac{h^2}{2}\mu_2(K) f'\left(\xi_\tau\right) / f\left(\xi_\tau\right) + o_p(\left(h^2\right)\right] \xrightarrow{d} N\left(0, \sigma_F^2\left(\xi_\tau\right) / f^2\left(\xi_\tau\right)\right).$$

Next, let us examine the AMSE. To this effect, from Theorem 1.3, it is easy to derive the asymptotic bias and variance, which are $h^2\mu_2(K)f'(\xi_\tau)/[2\,f(\xi_\tau)]$ and $\sigma_F^2\left(\xi_\tau\right)/f^2(\xi_\tau) - h\theta/f(\xi_\tau)$, respectively, so that the AMSE is given by

$$n\text{AMSE}\left(\widehat{\xi}_\tau\right) = \frac{nh^4}{4}\mu_2^2(K)\left[f'\left(\xi_\tau\right)/f\left(\xi_\tau\right)\right]^2 + \sigma_F^2\left(\xi_\tau\right)/f^2\left(\xi_\tau\right) - h\theta/f\left(\xi_\tau\right).$$

If $\theta > 0$, minimizing the AMSE gives the

$$h_{opt} = \left(\frac{\theta f\left(\xi_\tau\right)}{\mu_2^2(K)\left[f'\left(\xi_\tau\right)\right]^2}\right)^{1/3} n^{-1/3},$$

and with this asymptotically optimal bandwidth, the optimal AMSE is given by

$$n\text{AMSE}_{\text{opt}}\left(\widehat{\xi}_\tau\right) = \sigma_F^2\left(\xi_\tau\right)/f^2\left(\xi_\tau\right) - \frac{3}{4}\left(\frac{\theta^2}{\mu_2(K) f'\left(\xi_\tau\right) f\left(\xi_\tau\right)}\right)^{2/3} n^{-1/3},$$

which indicates a reduction to the AMSE of the second order.

Now, by virtue of the above discussions, we can use $\widehat{F}_n(x)$ in (1.18) to estimate the VaR, ν_τ given in (1.19), as the smoothed estimate of ν_τ, denoted by $\widehat{\nu}_\tau$. Similarly, we can estimate

μ_τ in (1.19), denoted by $\widehat{\mu}_\tau$. Indeed, Chen and Tang (2005) conducted an intensive study on simulations to demonstrate the advantages of nonparametric estimation $\widehat{\nu}_\tau$ over the sample quantile version. We refer to the paper by Chen and Tang (2005) for simulation results and empirical examples.

Exercise: Please use the above procedures to estimate nonparametrically the ES and discuss its properties as well as conduct simulation studies and empirical applications.

1.6 Computer Code

```
###############
# Example 1.1
#############

################################################
 # Define the Epanechnikov kernel function
  kernel<-function(x){0.75*(1-x^2)*(abs(x)<=1)}
    # Define the kernel density estimator
  kernden=function(x,z,h,ker){
    # parameters: x=variable; h=bandwidth; z=grid point; ker=kernel
    nz<-length(z)
    nx<-length(x)
    x0=rep(1,nx*nz)
    dim(x0)=c(nx,nz)
    x1=t(x0)
    x0=x*x0
    x1=z*x1
    x0=x0-t(x1)
    if(ker==1){x1=kernel(x0/h)}          # Epanechnikov kernel
    if(ker==0){x1=dnorm(x0/h)}           # normal kernel
    f1=apply(x1,2,mean)/h
    return(f1)
 }
```

```
# Simulation for different bandwidths and different kernels
 n=300                                        # n=300

 ker=1                                         # ker=1 => Epan; ker=0 => Gaussian
 h0=c(0.25,0.5,1)                              # set initial bandwidths
 z=seq(-4,4,by=0.1)                            # grid points
 nz=length(z)                                  # number of grid points
 x=rnorm(n)                                    # simulate x ~ N(0, 1)
 if(ker==1){h_o=2.34*n^{-0.2}}  # bandwidth for Epanechnikov kernel
 if(ker==0){h_o=1.06*n^{-0.2}}  # bandwidth for normal kernel
 f1=kernden(x,z,h0[1],ker)
 f2=kernden(x,z,h0[2],ker)
 f3=kernden(x,z,h0[3],ker)
 f4=kernden(x,z,h_o,ker)
 text1=c("True","h=0.25","h=0.5","h=1","h=h_o")
 data=cbind(dnorm(z),f1,f2,f3,f4)        # combine them as a matrix

 quartz()
 matplot(z,data,type="l",lty=1:5,col=1:5,xlab="",ylab="")
 legend(-1,0.2,text1,lty=1:5,col=1:5)

 f5=density(x, kernel=c("gaussian"))$y
 z1=density(x, kernel=c("gaussian"))$x
 f6=density(x, kernel=c("epanechnikov"))$y
 data1=cbind(f5,f6)
 text2=c("Gaussian","Epanechnikov")

 quartz()
 matplot(z1,data1,type="l",lty=1:2,col=1:2,xlab="",ylab="")
 legend(-1,0.2,text2,lty=1:2,col=1:2)
```

```
  quartz()
  par(mfrow=c(1,2),mex=0.4,bg="light grey")
  matplot(z,data,type="l",lty=1:5,col=1:5,xlab="",ylab="")
  legend(-1,0.2,text1,lty=1:5,col=1:5)
  text1=c("Gauassian","Epanechnikov")
  matplot(z1,data1,type="l",lty=1:2,col=1:2,xlab="",ylab="")
  legend(-3,0.2,text2,lty=1:2,col=1:2)
#####################################################

 ##################
 # Example 1.2
 ##################

#####################################################
  z1=read.table(file="/NP_lecture_note/data/ex3-2.txt", header=F)
   # dada: weekly 3-month Treasury bill from 1954 to 2022
  x=z1[,4]/100                    # decimal
  n=length(x)
  y=diff(x)                       # Delta x_t=x_t-x_{t-1}=change rate
  x=x[1:(n-1)]
  n=n-1
  x_star=(x-mean(x))/sqrt(var(x))  # standardized
  den_3mtb=density(x_star,bw=0.30,kernel=c("epanechnikov"),
  from=-3,to=3,n=61)
  den_est=den_3mtb$y              # estimated density values
  z_star=seq(-3,3,by=0.1)
  text1=c("Estimated Density","Standard Norm")

  win.graph()  # for Windows
  # quartz()   # for macOS
  par(bg="light green")
  plot(den_3mtb,main="Density of 3mtb (Buind-in)",ylab="",xlab="",
```

```
  col.main="red")
  points(z_star,dnorm(z_star),type="l",lty=2,col=2,ylab="",xlab="")
  legend(0,0.45,text1,lty=c(1,2),col=c(1,2),cex=0.7)

  h_den=0.5
  f_hat=kernden(x_star,z_star,h_den,1)
  ff=cbind(f_hat,dnorm(z_star))

  win.graph()
  par(bg="light blue")
  matplot(z_star,ff,type="l",lty=c(1,2),col=c(1,2),ylab="",xlab="")
  title(main="Density of 3mtb",col.main="red")
  legend(0,0.55,text1,lty=c(1,2),col=c(1,2),cex=0.7)
#####################################################


##############################
 # Example 1.3 (delta=0)
##############################

#####################################################################
 # Load needed packages
 library(ggplot2)
 library(tidyverse)
 library(ggpubr)
 set.seed(1)                          # to create reproducible results
 cols <- c("magenta", "red", "orange","blue","green")
#####################################################################

#####################################################################
 # Define the Rosenblatt-Parzen density estimator
 RP_dens_est<-function(x,h,z){
```

```
    # parameters: x=observed variable; h=bandwidth; z=grid point;
    nz<-length(z)
    nx<-length(x)
    x0=rep(1,nx*nz)
    dim(x0)=c(nx,nz)
    x1=t(x0)
    x0=x*x0
    x1=z*x1
    x0=x0-t(x1)
    x1=0.5*(abs(x0/h)<=1)                    # the uniform kernel
    f1=apply(x1,2,mean)/h
    return(f1)                               # return fn(z)
 }
#######################################################################

#############################################################################
 #  The Kernel Density Estimator for a Random Walk
 # Simulation for different bandiwidths, sample sizes and values of fixed x.
 rm(list = c())                             # clean the previous variables
 x<-seq(-5,5,length.out=5)                  # take 5 values of fixed x
 nrep=1e4                                  # repeat the simulation nrep times
 ns= c(200,1000,5000)                       # sample size
 delta=1
 ds=c(0.5,1,2)
 quest1<-list(NULL,NULL,NULL)               # fn(x)
 for (n in ns) {                            # sample size
   for (i in 1:nrep) {
       Xt<-cumsum(rnorm(n))                # generate data from a random walk
       for (h in 1:length(ds)) {
              d<-ds[h]                        # bandwidth= d*n^(-1/10)
              quest1[[h]]<-c(quest1[[h]],     #compute fn(x)
                             RP_dens_est(Xt,h=d*n^(-1/10),x))
```

```
    }
   }
 }
tabmed<-list()                                  # save median  of fn(x)
tabsd<-list()                                    # save sd of fn(x)
fig1<-list()                                      # save box-plots
for (h in 1:length(ds)) {
  d<-ds[h]
  Quest1<-data.frame(quest=quest1[[h]],       # rearrange simulated data
                     Grid.Points=factor(rep(paste("x=",x),nrep*length(ns)),
                                        levels = paste("x=",x)),
                     n=factor(rep(paste("n=",ns),each=nrep*length(x)),
                              levels = paste("n=",ns)))
  tabmed[[h]]<-(with(Quest1,
                     tapply(quest, list( n=n,Grid.Points=Grid.Points),median))
                %>%as.data.frame())
  tabsd[[h]]<-(with(Quest1,
                     tapply(quest, list( n=n,Grid.Points=Grid.Points),sd))
                %>%as.data.frame())
  fig1[[h]]<-Quest1%>%
    ggplot(aes(y=quest,x=n,fill=Grid.Points))+
    geom_boxplot()+
    scale_fill_manual(values = cols)+
    xlab("")+
    ylab(expression(paste(f[n],'(x)')))+
    labs(title=bquote(paste('Bandwidth=',.(d[1])%*%n^{-1/10})))+
    theme(axis.title = element_text(size=19),
          plot.title = element_text(size=21),
          axis.text= element_text(size=17),
          legend.text= element_text(size=17),
          legend.title= element_text(size=17))
 }
```

```
    write.csv(tabmed,"median_of_densityest_rw.csv")
    write.csv(tabsd,"sd_of_densityest_rw.csv")
    ggsave("rwbarplots.pdf", plot = do.call(ggarrange, c(fig1,ncol=1,nrow=3)),
           width = 24, height = 25, dpi =1500, bg = "white",units = "cm")
#####################################################################

 ########################
 # Example 1.3 (delta=1)
 ########################
 ###########################################################################
 #  The Kernel Density Estimator for a Nearly Random Walk
 # Simulation for different bandiwidths, sample sizes and values of fixed x.
 rm(list = c())
 x<-seq(-5,5,length.out=5)               # take 5 values of fixed x
 nrep=1e4                               # repeat the simulation nrep times
 ns= c(200,1000,5000)                    # sample size
 delta=1
 ds=c(0.5,1,2)
 quest1<-list(NULL,NULL,NULL)            # fn(x)
 for (n in ns) {                         # sample size
    phi<-1-delta/n                         # coefficient for AR(1)
    Phi<-diag(1,ncol=n,nrow=n)
    for (j in 1:n) {
      Phi[j,-(1:j)]<-phi^(1:(n-j))
   }
  for (i in 1:nrep) {
    u<-matrix(rnorm(n),ncol=1)             # error term
    Xt<-as.numeric(Phi%*%u)              # generate data from a nearly random walk
    for (h in 1:length(ds)) {
      d<-ds[h]                           # bandwidth= d*n^(-1/10)
      quest1[[h]]<-c(quest1[[h]],           #compute fn(x)
                     RP_dens_est(Xt,h=d*n^(-1/10),x))
```

```
  }
 }
}
tabmed<-list()                              # save median  of fn(x)
tabsd<-list()                                 # save sd of fn(x)
fig1<-list()                               # save box-plots
for (h in 1:length(ds)) {
  d<-ds[h]
  Quest1<-data.frame(quest=quest1[[h]],      # rearrange simulated data
                     Grid.Points=factor(rep(paste("x=",x),nrep*length(ns)),
                                        levels = paste("x=",x)),
   n=factor(rep(paste("n=",ns),each=nrep*length(x)),
            levels = paste("n=",ns)))
  tabmed[[h]]<-(with(Quest1,
                     tapply(quest, list( n=n,Grid.Points=Grid.Points),median))
                %>%as.data.frame())
  tabsd[[h]]<-(with(Quest1,
                     tapply(quest, list( n=n,Grid.Points=Grid.Points),sd))
                %>%as.data.frame())
  fig1[[h]]<-Quest1%>%
    ggplot(aes(y=quest,x=n,fill=Grid.Points))+
    geom_boxplot()+
    scale_fill_manual(values = cols)+
    xlab("")+
    ylab(expression(paste(f[n],'(x)')))+
    labs(title=bquote(paste('Bandwidth=',.(d[1])%*%n^{-1/10})))+
    theme(axis.title = element_text(size=19),
          plot.title = element_text(size=21),
          axis.text= element_text(size=17),
          legend.text= element_text(size=17),
          legend.title= element_text(size=17))
}
```

```
 write.csv(tabmed,"median_of_densityest_nearrw.csv")
 write.csv(tabsd,"sd_of_densityest_nearrw.csv")
 ggsave("nearrwbarplots.pdf", plot = do.call(ggarrange, c(fig1,ncol=1,nrow=3)),
        width = 24, height = 25, dpi =1500, bg = "white",units = "cm")
#####################################################################
```

Chapter 2

Regression Models

2.1 Instrocution

Suppose that we want to forecast the future value, say Y_{t+h}, h-step ahead with $h \geq 0$, given the information set Ω_t at time t. There are several forecasting criteria available in the literature. The general form is

$$m\left(\Omega_t\right) = \min_a \mathbb{E}\left[\rho\left(Y_{t+h} - a\right) \mid \Omega_t\right],$$

where $\rho(\cdot)$ is an objective (loss) function, which might not be differentiable. Here are several major scenarios.

(1) If $\rho(z) = z^2$ is the quadratic function, then, $m\left(\Omega_t\right) = \mathbb{E}\left(Y_{t+h} \mid \Omega_t\right)$ is called the mean regression function for $h = 0$ and the prediction function for $h \geq 1$. Implicitly, it requires that the distribution of Y_t should be symmetric. If the distribution of Y_t is skewed, then, this loss function is not a good criterion. Chapter 2 is for discussing the nonparametric or semiparametric forms in detail.

(2) If $\rho(y) = \rho_\tau(y) = y\left(\tau - I_{\{y<0\}}\right)$, the so-called *check function* in the quantile literature, which is clearly not differentiable at $y = 0$ although it is a continuous function, where $\tau \in (0, 1)$, then, $m\left(\Omega_t\right)$ satisfies the following equation

$$\int_{-\infty}^{m(\Omega_t)} f\left(y \mid \Omega_t\right) du = F\left(m\left(\Omega_t\right) \mid \Omega_t\right) = \tau,$$

where $f(y \mid \Omega_t)$ and $F(m(\Omega_t) \mid \Omega_t)$ are the conditional PDF and CDF of Y_{t+h} given Ω_t, respectively. This $m\left(\Omega_t\right)$ becomes the conditional quantile or quantile regression, dented by $q_\tau\left(\Omega_t\right)$, proposed by Koenker and Bassett (1978, 1982). Particularly, if

$\tau = 1/2$ and $h = 0$, then, $m(\Omega_t)$ is the well known least absolute deviation (LAD) regression, which is robust against outliers. If $q_\tau(\Omega_t)$ is a linear function of regressors like $\boldsymbol{\beta}_\tau^\top \mathbf{X}_t$ as in Koenker and Bassett (1978, 1982), Koenker (2005) developed the **R** module **quantreg** to make statistical inferences on the linear quantile regression model. Chapter 3 is devoted to studying the nonparametric and semiparametric quantile regression models in detail.

(3) If $\rho(x) = \frac{1}{2}x^2 I_{|x|\le M} + M(|x| - M/2)I_{|x|>M}$, the so called Huber function in literature, then, it is the so-called Huber robust regression, which has a very important property that it is robust against outliers. We will not discuss this topic. If you have an interest, please read the paper by Huber (1964) and the book by Rousseeuw and Leroy (1987). In **R**, the library **MASS** has the function **rlm()** for robust linear model. Also, the library **lqs** contains functions for bounded-influence regression.

(4) If $\rho(\mathbf{W}_t) = -U(\mathbf{W}_t)$, where $U(\cdot)$ is a utility function such as constant absolute risk aversion (CARA) or constant relative risk aversion (CRRA) or Epstein-Zin recursive utility function as in Epstein and Zin (1989), $\mathbf{W}_t = \boldsymbol{\alpha}^\top \mathbf{R}_{t+1}$ is a portfolio, $\mathbf{R}_{t+1}$ is a vector of risky asset returns, and $\boldsymbol{\alpha}$ is a vector of allocations, then, it becomes a portfolio management problem. For details, please read the paper by Aït-Sahalia and Brant (2001) or the book by Jondeau et al. (2007). Chapter 6 delivers a detailed study to this case.

(5) If $\rho(v) = \ell_\eta(v) = \eta\, v^2\, I(v > 0) + (1-\eta)\, v^2\, I(v \le 0)$ for $0 \le \eta \le 1$, which is an asymmetric squared loss (ALS) function, then, $m(\Omega_t)$, denoted by $e_\eta(\Omega_t)$, is the conditional expectile or expectile regression; see, for example, Newey and Powell (1987), Efron (1991), Section 6.3 in Fan and Gijbels (1996), Yao and Tong (1996), Cai et al. (2018), and references therein for details. It is easy to see that the loss function combines both the quadratic loss and the check function and the expectile reduces to the mean regression function if $\eta = 1/2$. Section 4.6 in Chapter 4 considers some parametric, semiparametric and nonparametric models for modeling the expectile regression.

To see differences among above four cases for the loss function $\rho(\cdot)$, please look at the plot of loss functions given in Figure 2.1.

Remark 2.1: *Note that for the second and third cases, the regression functions usually do not have a close form of expression. Since the information set Ω_t contains too many variables*

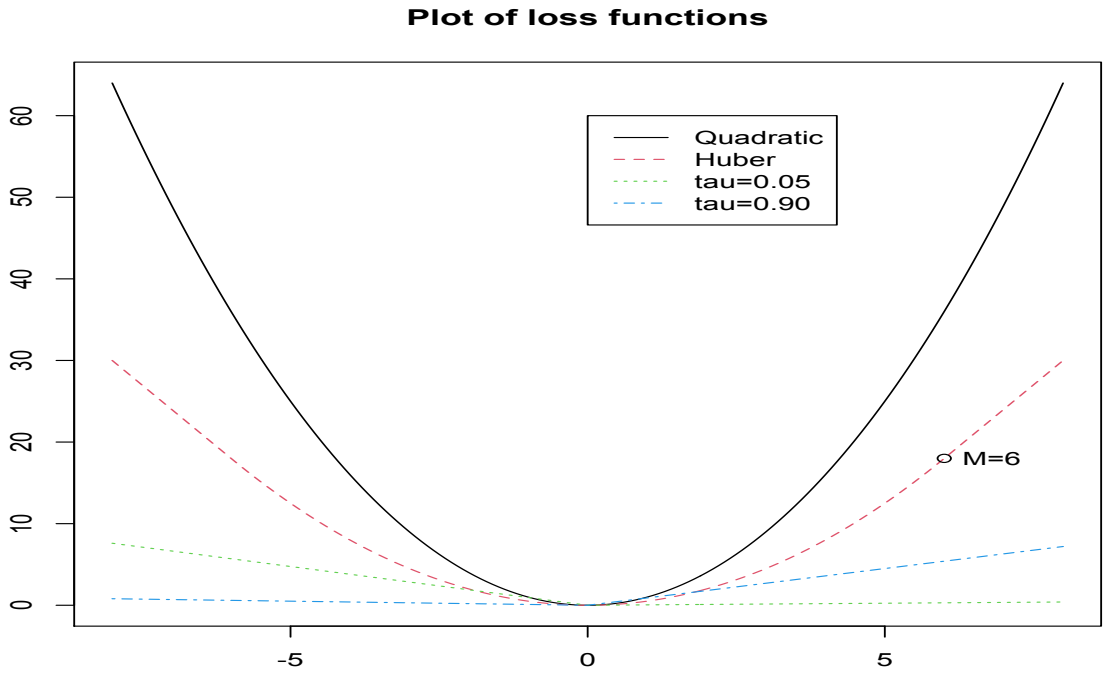

Figure 2.1: The plot of three loss functions: quadratic loss (black solid line), Huber loss (red dashed line) with $M = 6$, the check function ($\tau = 0.05$, green dotted line), and the check function ($\tau = 0.90$, blue dashed-dotted line).

(high dimension), it is often to approximate Ω_t *by some finite numbers of variables, say* $\mathbf{X}_t = (X_{t1}, \cdots, X_{tp})$ *with a fixed* $p \geq 1$*, including the lagged variables and exogenous variables. First, our focus is on the mean regression* $m(\mathbf{X}_t)$*. Of course, by the same token, we can consider the nonparametric estimation of the conditional variance* $\sigma^2(\mathbf{x}) = Var(Y_t \mid \mathbf{X}_t = \mathbf{x})$*. Why do we need to consider nonlinear (nonparametric) models in economic practice? To find the answer, please read the motivated example in* (1) *for the general sample selection mode, the paper by Engle et al. (1986) for the application to build a nonlinear relationship between electricity sales and temperature, and some examples in economics and finance in the book by Granger and Teräsvirta (1993).*

Remark 2.2: *Note that throughout this chapter, it is assumed that all regressors are continuous. For the case with discrete and/or partially discrete regressors, the reader is referred to the papers by Li and Racine (2003) and Hall et al. (2004) or the book by Li and Racine (2007) for details.*

2.2 Nadaraya-Watson Estimation

How to estimate $m(x) = \mathbb{E}(Y_t \,|\, \mathbf{X}_t = \mathbf{x})$ nonparametrically? For simplicity of notation, we focus on the univariate case that $p = 1$, so that $\mathbf{X}_t$ and $\mathbf{x}$ are re-written as X_t and x, respectively. Let us look at the Nadaraya-Watson estimate of the mean regression $m(x)$. The main idea is as follows:

$$m(x) = \int y f(y \mid x) dy = \frac{\int y f(x, y) dy}{\int f(x, y) dy},$$

where $f(x, y)$ is the joint PDF of X_t and Y_t. To estimate $m(x)$, we can apply the plug-in method based on the observations $\{(X_t, Y_t)\}_{t=1}^n$. That is, plug the nonparametric kernel density estimate $f_n(x, y)$ (product kernel method) into the right hand side of the above equation to obtain

$$\widehat{m}_{nw}(x) = \frac{\int y f_n(x, y) dy}{\int f_n(x, y) dy}$$

which can be easily simplified as

$$\widehat{m}_{nw}(x) = \frac{1}{n} \sum_{t=1}^{n} Y_t K_h \left(X_t - x\right) / f_n(x) = \sum_{t=1}^{n} W_t Y_t,$$

where $f_n(x)$ is the kernel density estimation of $f(x)$, defined in (1.8) in Chapter 1 , and

$$W_t = K_h \left(X_t - x\right) / \sum_{t=1}^{n} K_h \left(X_t - x\right).$$

$\widehat{m}_{nw}(x)$ is the well known Nadaraya-Watson (NW) estimator, proposed by Nadaraya (1964) and Watson (1964), respectively. Note that the weights $\{W_t\}_{t=1}^n$ do not depend on $\{Y_t\}$. Therefore, $\widehat{m}_{nw}(x)$ is called a linear estimator, similar to the ordinary least squares (OLS) estimate.

Let us look at the NW estimator from a different angle. $\widehat{m}_{nw}(x)$ can be re-expressed as the minimizer of the locally weighted least squares; that is,

$$\widehat{m}_{nw}(x) = \text{argmin}_a \sum_{t=1}^{n} \left(Y_t - a\right)^2 K_h \left(X_t - x\right). \tag{2.1}$$

A nice interpretation of the weight $K_h(X_t - x)$ in (2.1) is as follows. Unlike OLS estimate, it assigns a different weight to each observation and the weight is larger when X_t is closer to x.

This means that when X_t is in a neighborhood of $x, m(X_t)$ is approximated by a constant a (local approximation). Indeed, we consider the following working model

$$Y_t = m(X_t) + \varepsilon_t \approx a + \varepsilon_t$$

with the weights $\{K_h(X_t - x)\}_{t=1}^n$, where $\varepsilon_t = Y_t - \mathbb{E}(Y_t \mid X_t)$. Therefore, the Nadaraya-Watson estimator is also called the local constant estimator.

In the implementation, for each grid point x, we can fit the following transformed linear model

$$Y_t^* = \beta_1 X_t^* + \varepsilon_t,$$

where $Y_t^* = \sqrt{K_h(X_t - x)}Y_t$ and $X_t^* = \sqrt{K_h(X_t - x)}$. In **R**, we can use functions **lm()** or **glm()** with weights $\{K_h(X_t - x)\}_{t=1}^n$ to fit a weighted least squares or generalized linear model. Or, you can use the weighted least squares theory (matrix multiplication); see Section 2.9.

2.2.1 Asymptotic Properties

We derive the asymptotic properties of the nonparametric estimator for the time series situations. Note that the mathematical derivations are different for the iid case and time series situations since the key equality $\mathbb{E}[Y_t \mid X_1, \cdots, X_n] = \mathbb{E}[Y_t \mid X_t] = m(X_t)$ holds only for the iid case. To ease notation, we consider only the simple case when $p = 1$. A simple algebra leads to

$$\widehat{m}_{nw}(x) f_n(x) = \underbrace{\frac{1}{n}\sum_{t=1}^n m(X_t) K_h(X_t - x)}_{I_1} + \underbrace{\frac{1}{n}\sum_{t=1}^n K_h(X_t - x)\varepsilon_t}_{I_2},$$

where $f_n(x) = \sum_{t=1}^n K_h(X_t - x)/n$. We will show that I_1 contributes only the asymptotic bias and I_2 gives the asymptotic normality. First, we derive the asymptotic bias for the interior boundary points. By the Taylor's expansion, when X_t is in $(x - h, x + h)$, we have

$$m(X_t) = m(x) + m'(x)(X_t - x) + \frac{1}{2}m''(x_t)(X_t - x)^2,$$

where $m'(x)$ is the first derivative of $m(x)$, $m''(x)$ is the second derivative of $m(x)$, and $x_t = x + \theta\,(X_t - x)$ with $-1 < \theta < 1$. Then,

$$I_1 \equiv \frac{1}{n}\sum_{t=1}^{n} m\,(X_t)\,K_h\,(X_t - x) = m(x)f_n(x) + m'(x)\underbrace{\frac{1}{n}\sum_{t=1}^{n}(X_t - x)\,K_h\,(X_t - x)}_{J_1(x)}$$

$$+ \frac{1}{2}\underbrace{\frac{1}{n}\sum_{t=1}^{n} m''\,(x_t)\,(X_t - x)^2\,K_h\,(X_t - x)}_{J_2(x)}.$$

Then,

$$\mathbb{E}\,[J_1(x)] = \mathbb{E}\,[(X_t - x)\,K_h\,(X_t - x)] = \int (u - x)K_h(u - x)f(u)du$$
$$= h\int uK(u)f(x + hu)du = h^2 f'(x)\mu_2(K) + o\left(h^2\right).$$

Similar to the derivation of the variance of $f_n(x)$ in (1.9), we can show that under some conditions,

$$nh\mathrm{Var}\,(J_1(x)) = O(1).$$

Therefore, $J_1(x) = h^2 f'(x)\mu_2(K) + o_p\,(h^2)$. By the same token, we have

$$\mathbb{E}\,[J_2(x)] = \mathbb{E}\left[m''\,(x_t)\,(X_t - x)^2\,K_h\,(X_t - x)\right]$$
$$= h^2\int m''(x + \theta hu)u^2K(u)f(x + hu)du = h^2 m''(x)\mu_2(K)f(x) + o\left(h^2\right),$$

and $\mathrm{Var}\,(J_2(x)) = O(1/nh)$. Therefore, $J_2(x) = h^2 m''(x)\mu_2(K)f(x) + o_p\,(h^2)$. Hence,

$$I_1 = m(x)f(x) + m'(x)J_1(x) + \frac{1}{2}J_2(x)$$
$$= m(x)f(x) + \frac{h^2}{2}\mu_2(K)\left[m''(x) + 2m'(x)f'(x)/f(x)\right]f(x) + o_p\left(h^2\right)$$

by the fact that $f_n(x) = f(x) + o_p(1)$. The term $I_1 \approx f(x)\,[m(x) + B_{nw}(x)]$, where

$$B_{nw}(x) = \frac{h^2}{2}\mu_2(K)\left[m''(x) + \boxed{2m'(x)f'(x)/f(x)}\right] \tag{2.2}$$

is regarded as the asymptotic bias. The bias term involves not only curvatures of $m(x)$ ($m''(x)$) but also the unknown density function $f(x)$ and its derivative $f'(x)$ so that the design can not be adaptive; see, e.g., Fan and Gijbels (1996) for details.

Under some regularity conditions, similar to (1.9), we can show that for the given grid point x, an interior grid point,

$$nh\text{Var}(I_2) \quad \to \quad \nu_0(K)\sigma_\varepsilon^2(x)f(x) \equiv \sigma_m^2(x)f^2(x),$$

where $\sigma_\varepsilon^2(x) = \text{Var}\left(\varepsilon_t \mid X_t = x\right)$ and $\sigma_m^2(x) = \nu_0(K)\sigma_\varepsilon^2(x)/f(x)$. Further, by the fact that $f_n(x) = f(x) + o_p(1)$ and the Slutsky theorem, we can establish the asymptotic normality (the proof is provided later)

$$\sqrt{nh}\left[\widehat{m}_{nw}(x) - m(x) - B_{nw}(x) + o_p\left(h^2\right)\right] \quad \xrightarrow{d} \quad N\left(0, \sigma_m^2(x)\right), \tag{2.3}$$

where $B_{nw}(x)$ is given in (2.2).

2.2.2 Boundary Behavior

For expositional purpose, in what follows, we only consider the case when $p = 1$. As for the boundary behavior for the NW estimator, we can follow Fan and Gijbels (1996). Without loss of generality, we consider the left boundary point $x = ch, 0 < c < 1$. From Fan and Gijbels (1996), we take $K(\cdot)$ to have support $[-1, 1]$ and $m(\cdot)$ to have support $[0, 1]$. Similar to (1.15), it is easy to see that if $x = ch$,

$$\begin{aligned}
\mathbb{E}\left[J_1(ch)\right] &= \mathbb{E}\left[\left(X_t - ch\right)K_h\left(X_t - ch\right)\right] = \int_0^1 (u - ch)K_h(u - ch)f(u)du \\
&= h\int_{-c}^{1/h-c} uK(u)f(h(u + c))du \\
&= hf(0+)\mu_{1,c}(K) + h^2 f'(0+)\left[\mu_{2,c}(K) + c\mu_{1,c}(K)\right] + o\left(h^2\right),
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}\left[J_2(ch)\right] &= \mathbb{E}\left[m''\left(x_t\right)\left(X_t - ch\right)^2 K_h\left(X_t - ch\right)\right] \\
&= h^2\int_{-c}^{1/h-c} m''(h(c + \theta u))u^2K(u)f(h(u + c))du \\
&= h^2 m''(0+)\mu_{2,c}(K)f(0+) + o\left(h^2\right).
\end{aligned}$$

Also, we can see that

$$\text{Var}\left(J_1(ch)\right) = O(1/nh) \quad \text{and} \quad \text{Var}\left(J_2(ch)\right) = O(1/nh),$$

which imply that

$$J_1(ch) = hf(0+)\mu_{1,c}(K) + o_p(h) \quad \text{and} \quad J_2(ch) = h^2 m''(0+)\mu_{2,c}(K)f(0+) + o\left(h^2\right).$$

This, in conjunction with (1.16), gives

$$I_1 - m(ch) = m'(ch)J_1(ch)/f_n(ch) + \frac{1}{2}J_2(ch)/f_n(ch) = a(c,K)h + b(c,K)h^2 + o_p\left(h^2\right),$$

where

$$a(c,K) = \frac{m'(0+)\mu_{1,c}(K)}{\mu_{0,c}(K)},$$

and

$$b(c,K) = \frac{\mu_{2,c}(K)m''(0+)}{2\mu_{0,c}(K)} + \frac{f'(0+)m'(0+)\left[\mu_{2,c}(K)\mu_{0,c}(K) - \mu_{1,c}^2(K)\right]}{f(0+)\mu_{0,c}^2(K)}.$$

Here, $a(c,K)h + b(c,K)h^2$ serves as the asymptotic bias term, which has the order $O(h)$. Also, we can show that at the boundary point, the asymptotic variance has the following form

$$nh\mathrm{Var}\left(\widehat{m}_{nw}(x)\right) \quad \rightarrow \quad \nu_{0,c}(K)\sigma_m^2(0+)/\left[\mu_{0,c}(K)f(0+)\right],$$

which has the same order as that for the interior point although the scaling constant is different.

2.3 Local Polynomial Estimate

To overcome the above shortcomings of the local constant estimate, we can use the local polynomial fitting scheme; see Fan and Gijbels (1996) for details. The main idea is described as follows.

2.3.1 Formulation

Assume that the regression function $m(x)$ has $(q+1)$ th order continuous derivative. For ease notation, assume that $p = 1$. When $X_t \in (x-h, x+h)$, then,

$$m\left(X_t\right) \approx \sum_{j=0}^{q} \frac{m^{(j)}(x)}{j!}\left(X_t - x\right)^j = \sum_{j=0}^{q} \beta_j \left(X_t - x\right)^j,$$

where $\beta_j = m^{(j)}(x)/j\,!$. Therefore, when $X_t \in (x-h, x+h)$, the model becomes

$$Y_t \approx \sum_{j=0}^{q} \beta_j \left(X_t - x\right)^j + \varepsilon_t.$$

Hence, we can apply the weighted least squares method. Similar to (2.1), the locally weighted least squares becomes

$$\widehat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \sum_{t=1}^{n} \left(Y_t - \sum_{j=0}^{q} \beta_j \left(X_t - x \right)^j \right)^2 K_h \left(X_t - x \right), \tag{2.4}$$

where $\boldsymbol{\beta} = (\beta_0, \cdots, \beta_q)$, which leads to obtaining the local polynomial estimate $\widehat{\boldsymbol{\beta}}$;

$$\widehat{\boldsymbol{\beta}} = \left(\mathbb{X}^\top \mathbf{W} \mathbb{X} \right)^{-1} \mathbb{X}^\top \mathbf{W} \mathbf{Y}, \tag{2.5}$$

where $\mathbf{W} = \text{diag}\left\{ K_h \left(X_1 - x \right), \cdots, K_h \left(X_n - x \right) \right\}$, the design matrix

$$\mathbb{X} = \begin{pmatrix} 1 & (X_1 - x) & \cdots & (X_1 - x)^q \\ 1 & (X_2 - x) & \cdots & (X_2 - x)^q \\ \vdots & \vdots & \ddots & \vdots \\ 1 & (X_n - x) & \cdots & (X_n - x)^q \end{pmatrix}, \quad \text{and} \quad \mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$

Therefore, for $1 \leq j \leq q$,

$$\widehat{m}^{(j)}(x) = j! \widehat{\beta}_j.$$

This means that the local polynomial method estimates not only the regression function itself but also derivatives of regression.

2.3.2 Implementation in R and A Real Example

There are several ways of implementing the local polynomial estimator. One way you can do so is to write your own code by using matrix multiplication as in 2.5 or employing function **lm()** or **glm()** with weights $\{K_h \left(X_t - x \right)\}$. Importantly, in **R**, there are some build-in packages for implementing the local polynomial estimate. For example, the package **KernSmooth** contains several functions. Function **bkde()** computes the kernel density estimate and Function **bkde2D()** computes the 2D kernel density estimate as well as Function **bkfe()** computes the kernel functional (derivative) density estimate. Function **dpik()** selects a bandwidth for estimating the kernel density estimation using the plug-in method and Function **dpill()** chooses a bandwidth for the local linear ($q = 1$) regression estimation using the plug-in approach. Finally, Function **locpoly()** is for the local polynomial fitting including a local polynomial estimate of the density of X (or its derivative) if the dependent variable is omitted. Finally, the package **np** provides a variety of nonparametric and semiparametric kernel methods that seamlessly handle a mix of continuous, unordered, and

ordered factor data types (unordered and ordered factors are often referred to as "nominal" and "ordinal" categorical variables respectively). A vignette containing many of the examples found in the help files accompanying the **np** package that is intended to serve as a gentle introduction to this package can be accessed via *vignette("np", package="np")* in **R**.

Example 2.1: We apply the kernel regression estimation and local polynomial fitting methods to estimate the drift and diffusion of the weekly 3-month Treasury bill from January 2, 1970 to December 26, 1997[1]. Let x_t denote the weekly 3-month Treasury bill. It is often to model X_t by assuming that it satisfies the continuous-time stochastic differential equation (Black-Scholes model)

$$dX_t = \mu(X_t)\,dt + \sigma(X_t)\,dW_t,$$

where W_t is a Wiener process, $\mu(X_t)$ is called the drift function and $\sigma(X_t)$ is called the diffusion function. Our interest is to identify $\mu(X_t)$ and $\sigma(X_t)$. Assume a time series sequence $\{X_{t\Delta}, 1 \le t \le n\}$ is observed at **equally spaced** time points. Using the **infinitesimal generator**, see, for example, Øksendal (1985), the first-order approximations of moments of X_t, a discretized version of the Ito's process, are given by Stanton (1997) as follows

$$y(t) = \Delta X_t = \mu(X_t)\,\Delta + \sigma(X_t)\,\varepsilon\sqrt{\Delta},$$

where $\Delta X_t = X_{t+\Delta} - X_t, \varepsilon \sim N(0,1)$, and x_t and ε_t are independent. Therefore,

$$\mu(X_t) = \lim_{\Delta\to 0} \mathbb{E}\left[\Delta X_t \mid X_t\right]/\Delta \quad \text{and} \quad \sigma^2(X_t) = \lim_{\Delta\to 0} \mathbb{E}\left[(\Delta X_t)^2 \mid X_t\right]/\Delta.$$

For the higher order approximations, see, for example, Fan and Zhang (2003). Hence, estimating $\mu(x)$ and $\sigma^2(x)$ becomes a nonparametric regression problem. We can use both local constant and local polynomial method to estimate $\mu(x)$ and $\sigma^2(x)$. As a result, the local constant estimators (red line) together with the **lowess()** smoothers (black line) and the scatterplots of Δx_t in (a), $|\Delta x_t|$ in (b), and $(\Delta x_t)^2$ in (c) versus x_t are presented in Figure 2.2 and the local linear estimators (red line) together with the **lowess()** smoothers (black line) and the scatterplots of ΔX_t in (a), $|\Delta X_t|$ in (b), and $(\Delta X_t)^2$ in (c) versus X_t are displaced in Figure 2.3. An alternative approach can be found in Aït-Sahalia (1996) to estimate $\mu(x)$ due to the domination of $\sigma(X_t)\,\varepsilon\sqrt{\Delta}$ over $\mu(X_t)\,\Delta$; see, for example, the paper by Cai and Hong (2009) for more details on this regard. From the above observations, can you conclude that the short rate drift is actually nonlinear or linear? For the detailed arguments, please see

[1]Similar to Example 1.2, the data set can be updated to today and it covers a longer period.

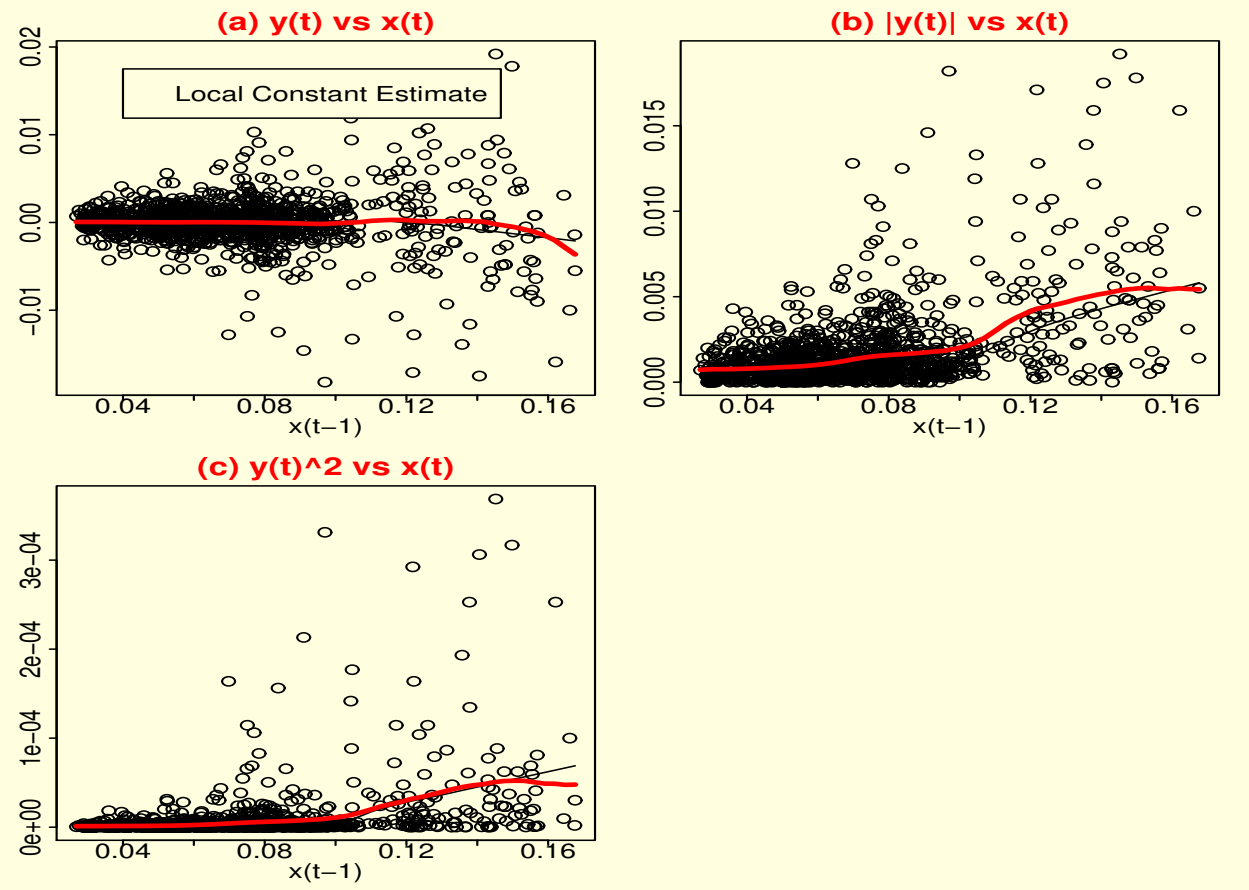

Figure 2.2: Scatterplots of $\Delta X_t, |\Delta X_t|$, and $(\Delta X_t)^2$ versus $x(t) = X_t$ with the smoothed curves computed using **scatter.smooth()** and the local constant estimation.

the papers by Chapman et al. (1999) and Chapman and Pearson (2000) from both financial and econometric point views.

2.3.3 Properties of Local Polynomial Estimator

Define, for $0 \leq j \leq q$,

$$s_{n,j}(x) = \sum_{t=1}^{n} (X_t - x)^j K_h (X_t - x),$$

and $\mathbf{S}_n(x) = \mathbb{X}^\top \mathbf{W} \mathbb{X}$. Then, the $(i+1, j+1)$ th element of $\mathbf{S}_n(x)$ is $s_{n,i+j}(x)$. Similar to the evaluation of I_{11}, we can show easily that

$$s_{n,j}(x) = nh^j \mu_j(K) f(x) \{1 + o_p(1)\}.$$

Define, $\mathbf{H} = \text{diag}\{1, h, \cdots, h^q\}$ and $\mathbf{S} = (\mu_{i+j}(K))_{0 \leq i,j \leq q}$. Then, it is not difficult to show that $\mathbf{S}_n(x) = nf(x)\mathbf{H}S\mathbf{H}\{1 + o_p(1)\}$.

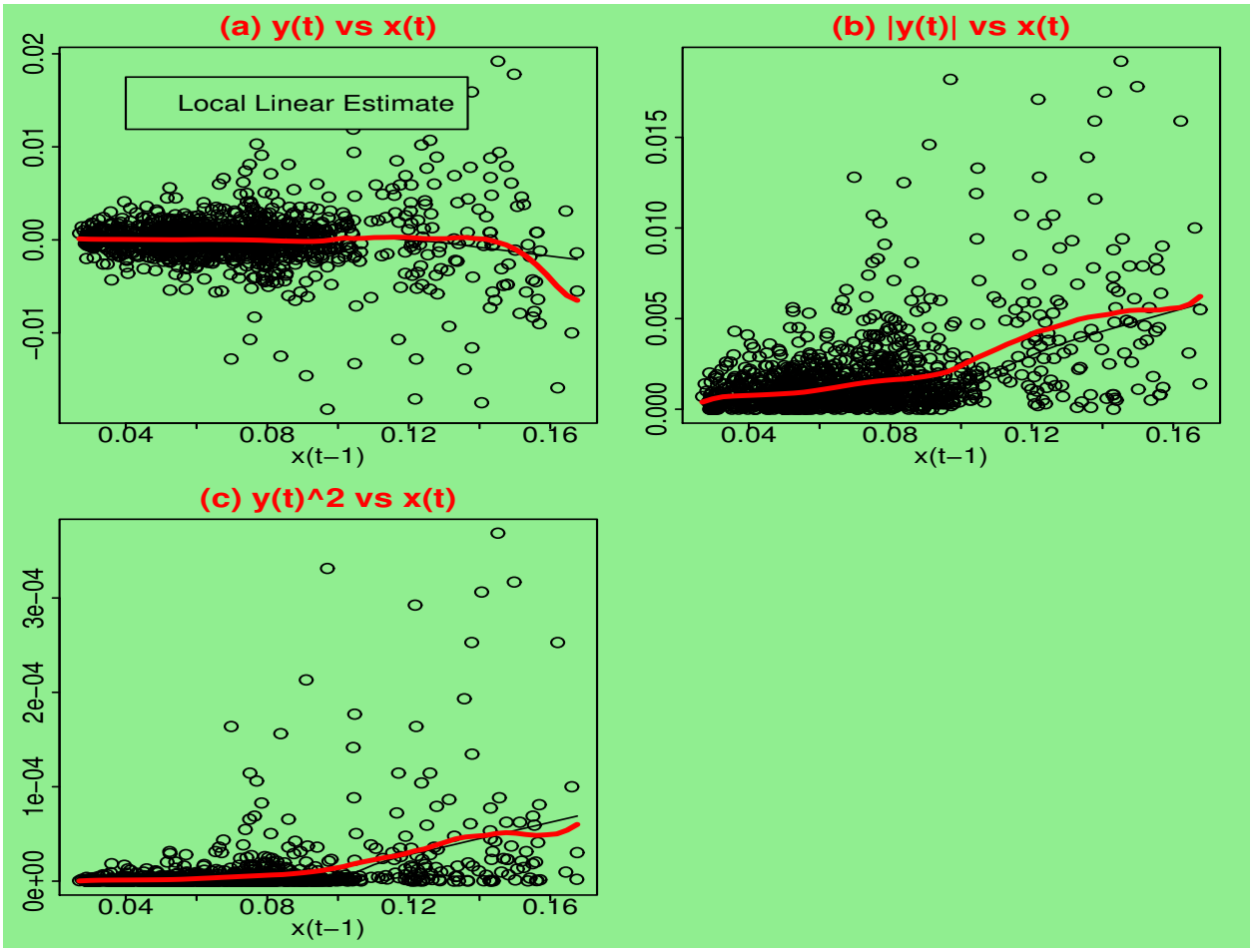

Figure 2.3: Scatterplots of $\Delta X_t, |\Delta X_t|$, and $(\Delta X_t)^2$ versus $x(t)$ with the smoothed curves computed using **scatter.smooth()** and the local linear estimation.

First of all, for $0 \leq j \leq q$, let e_j be a $(q+1) \times 1$ vector with $(j+1)$ th element being one and zero otherwise. Then, $\widehat{\beta}_j$ can be re-expressed as

$$\widehat{\beta}_j = \mathbf{e}_j^\top \widehat{\boldsymbol{\beta}} = \sum_{t=1}^{n} W_{j,n,h}\left(X_t - x\right) Y_t,$$

where $\mathbf{e}_j$ is a $(q+1) \times 1$ vector with its jth element being 1 and other elements being zero and $W_{j,n,h}\left(X_t - x\right)$ is called the effective kernel in Fan and Gijbels (1996) and Fan and Yao (2003), given by

$$W_{j,n,h}\left(X_t - x\right) = \left(1, \left(X_t - x\right), \cdots, \left(X_t - x\right)^q\right) \mathbf{S}_n(x)^{-1} \mathbf{e}_j K_h\left(X_t - x\right).$$

It is not difficult to show (based on the least square theory) that $W_{j,n,h}\left(X_t - x\right)$ satisfies the

following the so-called discrete moment conditions

$$\sum_{t=1}^{n} (X_t - x)^l W_{j,n,h}(X_t - x) = \begin{cases} 1 & \text{if } l = j \\ 0 & \text{otherwise} \end{cases} \tag{2.6}$$

Note that the local constant estimator does not have this property; see $J_1(x)$ in Section 2.2.1. This property implies that the local polynomial estimator is unbiased for estimating β_j, when the true regression function $m(x)$ is a polynomial of order q.

To gain more insights about the local polynomial estimator, define the equivalent kernel as in Fan and Gijbels (1996))

$$W_j(u) = (1, u, \cdots, u^q)\, \mathbf{S}^{-1} \mathbf{e}_j K(u).$$

Then, it can be shown, see, for example, Fan and Gijbels (1996), that

$$W_{j,n,h}(X_t - x) = \frac{1}{nh^{j+1} f(x)} W_j\left((X_t - x)/h\right)\{1 + o_p(1)\}$$

and

$$\int u^l W_j(u) du = \begin{cases} 1 & \text{if } l = j \\ 0 & \text{otherwise.} \end{cases}$$

The implications of these results are summarized as follows.

As pointed out by Fan and Yao (2003), the local polynomial estimator works like a kernel regression estimation with a known design density $f(x)$. This explains why the local polynomial fit adapts to various design densities. In contrast, the kernel regression estimator has large bias at the region where the derivative of $f(x)$ is large, namely it cannot adapt to highly-skewed designs. To see that, imagine the true regression function has large slope in this region. Since the derivative of design density is large, for a given x, there are more points on one side of x than the other. When the local average is taken, the Nadaraya-Watson estimate is biased towards the side with more local data points because the local data are asymmetrically distributed. This issue is more pronounced at the boundary regions, since the local data are even more asymmetric. On the other hand, the local polynomial fit creates asymmetric weights, if needed, to compensate for this kind of design bias. Hence, it is adaptive to various design densities and to the boundary regions.

We next derive the asymptotic bias and variance expression for local polynomial estimators. For independent data, we can obtain the bias and variance expression via conditioning

on the design matrix $\mathbb{X}$. However, for time series data, conditioning on $\mathbf{X}$ would mean conditioning on nearly the entire series. Hence, we derive the asymptotic bias and variance using the asymptotic normality rather than conditional expectation. As explained in Chapter 1 localizing in the state domain weakens the dependent structure for the local data. Hence, one would expect that the result for the independent data continues to hold for the stationary process with certain mixing conditions. The mixing condition and the bandwidth should be related, which can be seen later.

Set $\mathbf{B}_n(x) = (b_1(x), \cdots, b_n(x))^\top$, where, for $0 \leq j \leq q$,

$$b_{j+1}(x) = \sum_{t=1}^{n} \left[m(X_t) - \sum_{j=0}^{q} \frac{m^{(j)}(x)}{j!} (X_t - x)^j \right] (X_t - x)^j K_h(X_t - x).$$

Then,

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \left(\mathbf{X}^\top \mathbf{W} \mathbf{X}\right)^{-1} B_n(x) + \left(\mathbf{X}^\top \mathbf{W} \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{W} \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_n)^\top$. It is easy to show that if q is odd,

$$\mathbf{B}_n(x) = nh^{q+1} \mathbf{H} f(x) \frac{m^{(q+1)}(x)}{(q+1)!} \mathbf{c}_{1,q} \{1 + o_p(1)\},$$

where, for $1 \leq k \leq 3$, $\mathbf{c}_{k,q} = (\mu_{q+k}(K), \cdots, \mu_{2q+k}(K))^\top$. If q is even,

$$\mathbf{B}_n(x) = nh^{q+2} \mathbf{H} f(x) \left[\mathbf{c}_{2,q} \frac{m^{(q+1)}(x) f'(x)}{f(x)(q+1)!} + \mathbf{c}_{3,q} \frac{m^{(q+2)}(x)}{(q+2)!} \right] \{1 + o_p(1)\}.$$

Note that $f'(x)/f(x)$ does not appear in the right hand side of $\mathbf{B}_n(x)$ when q is odd. In either case, we can show that

$$nh\mathrm{Var}[H(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})] \quad \to \quad \sigma^2(x) \mathbf{S}^{-1} \mathbf{S}^* \mathbf{S}^{-1} / f(x) = \Sigma(x),$$

where $\mathbf{S}^*$ is a $(q+1) \times (q+1)$ matrix with the (i,j) th element being $\nu_{i+j-2}(K)$.

This shows that the leading conditional bias term depends on whether q is odd or even. By a Taylor series expansion argument, we know that when considering $|X_t - x| < h$, the remainder term from a q th order polynomial expansion should be of order $O(h^{q+1})$, so the result for odd q is quite easy to understand. When q is even, $(q+1)$ is odd hence the term h^{q+1} is associated with $\mu_l(K)$ for l odd, and this term is zero because $K(u)$ is a even function. Therefore, the h^{q+1} term disappears, while the remainder term becomes $O(h^{q+2})$. Since q is either odd or even, then we see that the bias term is an even power of h. This

is similar to the case where one uses higher order kernel functions based upon a symmetric kernel function (an even function), where the bias is always an even power of h.

Finally, we can show that when q is odd,

$$\sqrt{nh}[\mathbf{H}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})-\mathbf{B}(x)+o_p(h^{q+1})] \xrightarrow{d} N(0,\Sigma(x)),$$

the asymptotic bias term for the local polynomial estimator is

$$\mathbf{B}(x)=\frac{h^{q+1}}{(q+1)!}m^{(q+1)}(x)\mathbf{S}^{-1}\mathbf{c}_{1,q}\left\{1+o_p(1)\right\}.$$

Or

$$\sqrt{nh^{2j+1}}\left[\widehat{m}^{(j)}(x)-m^{(j)}(x)-B_j(x)+o_p(h^{q+1-j})\right] \xrightarrow{d} N\left(0,\sigma_{jj}(x)\right),$$

where the asymptotic bias and variance for the local polynomial estimator of $m^{(j)}(x)$ are

$$B_j(x)=\frac{j!h^{q+1-j}}{(q+1)!}m^{(q+1)}(x)\mu_{q+1}(W_j)\left\{1+o_p(1)\right\} \quad \text{and} \quad \sigma_{jj}(x)=\frac{(j!)^2\sigma^2(x)}{f(x)}\nu_0(W_j),$$

respectively. Especially, when $q=1$ and $j=0$, $\widehat{m}^{(j)}(x)$ is the local linear estimate of $m(x)$, denoted by $\widehat{m}_{\text{ll}}(x)$. Thus, we have the following asymptotic normality as

$$\sqrt{nh}\left[\widehat{m}_{\text{ll}}(x)-m(x)-B_{\text{ll}}(x)+o_p(h^2)\right] \xrightarrow{d} N\left(0,\sigma_m^2(x)\right), \tag{2.7}$$

where $B_{\text{ll}}(x)=\frac{h^2}{2}\mu_2(K)m''(x)$ and $\sigma_m^2(x)$ is given in (2.3). Clearly, by a comparison of (2.7) with (2.3), one can see that $B_{\text{ll}}(x)$ is totally different from $B_{\text{nw}}(x)$ in (2.2), although both have the same the asymptotic variance as $\sigma_m^2(x)$.

Similarly, we can derive the asymptotic bias and variance at boundary points if the regression function has a finite support. For details, see the books by Fan and Gijbels (1996), Fan and Yao (2003), and Ruppert and Wand (1994) for details. Indeed, define $\mathbf{S}_c$, $\mathbf{S}_c^*$, and $\mathbf{c}_{k,q,c}$ similarly to $\mathbf{S}$, $\mathbf{S}^*$ and $\mathbf{c}_{k,q}$ with $\mu_j(K)$ and $\nu_j(K)$ replaced by $\mu_{j,c}(K)$ and $\nu_{j,c}(K)$ respectively. We can show that

$$\sqrt{nh}\left[\mathbf{H}(\widehat{\boldsymbol{\beta}}(ch)-\boldsymbol{\beta}(ch))-\mathbf{B}_c(0)+o_p(h^{q+1})\right] \xrightarrow{d} N\left(0,\Sigma_c(0)\right), \tag{2.8}$$

where the asymptotic bias term for the local polynomial estimator at the left boundary point is

$$\mathbf{B}_c(0)=\frac{h^{q+1}}{(q+1)!}m^{(q+1)}(0)\mathbf{S}_c^{-1}\mathbf{c}_{1,q,c}\left\{1+o_p(1)\right\},$$

and the asymptotic variance is $\Sigma_c(0) = \sigma^2(0)\mathbf{S}_c^{-1}\mathbf{S}_c^*\mathbf{S}_c^{-1}/f(0)$. Alternatively,

$$\sqrt{nh^{2j+1}}\left[\widehat{m}^{(j)}(ch) - m^{(j)}(ch) - B_{j,c}(0) + o_p(h^{q+1-j})\right] \xrightarrow{d} N\left(0, \sigma_{jj,c}(0)\right), \tag{2.9}$$

where with $W_{j,c}(u) = (1, u, \cdots, u^q)\,\mathbf{S}_c^{-1}\mathbf{e}_j K(u)$,

$$B_{j,c}(0) = \frac{j!h^{q+1-j}}{(q+1)!}m^{(q+1)}(0)\int_{-c}^{\infty} u^{q+1}W_{j,c}(u)du\,\{1 + o_p(1)\}$$

and

$$\sigma_{jj,c}(0) = \frac{(j!)^2\sigma^2(0)}{f(0)}\int_{-c}^{\infty} W_{j,c}^2(u)du.$$

The asymptotic result in (2.9) implies that the asymptotic behavior at boundary is the same as that for the interior point.

Exercise: Please derive the asymptotic properties for the local polynomial estimator; that is to prove (2.8) or (2.9).

The above conclusions show that when $q - j$ is odd, the bias at the boundary is of the same order as that for points on the interior. Hence, the local polynomial fit does not create excessive boundary bias when $q - j$ is odd. Thus, the appealing boundary behavior for local polynomial mean estimation extends to derivative estimation. However, when $q - j$ is even, the bias at the boundary is larger than in the interior, and the bias can also be large at points where $f(x)$ is discontinuous. This is referred to as boundary effect. For these reasons (and the minimax efficiency arguments), it is recommended that one strictly set $q - j$ to be odd when estimating $m^{(j)}(x)$. It is indeed an odd world!

2.3.4 Complexity of Local Polynomial Estimator

To implement the local polynomial estimator, we have to choose the order of the polynomial q, the bandwidth h and the kernel function $K(\cdot)$. These parameters are of course confounded each other. Clearly, when $h = \infty$, the local polynomial fitting becomes a global polynomial fitting and the order q determines the model complexity. Unlike in the parametric models, the complexity of local polynomial fitting is primarily controlled by the bandwidth, as shown in Fan and Gijbels (1996) and Fan and Yao (2003). Hence q is usually small and the issue of choosing q becomes less critical. We discuss those issues in detail as follows.

(1) If the objective is to estimate $m^{(j)}(\cdot)(j \geq 0)$, the local polynomial fitting corrects automatically the boundary bias when $q-j$ is is odd. Further, when $q-j$ is odd, comparing with the order $q-1$ fit (so that $q-j-1$ is even), the order q fit contains one extra parameter without increasing the variance for estimating $m^{(j)}(\cdot)$. But this extra parameter creates opportunities for bias reduction, particularly in the boundary regions; see the next section and the books by Fan and Gijbels (1996) and Ruppert and Wand (1994). For these reasons, the odd order fits (the order q is chosen so that $q-j$ is odd) outperforms the even order fits [the order $(q-1)$ fit so that q is even]. Based on theoretical and practical considerations, the order $q = j+1$ is recommended in Fan and Gijbels (1996). If the primary objective is to estimate the regression function, one uses local linear fit and if the target function is the first order derivative, one uses the local quadratic fit and so on.

(2) It is well known that the choice of the bandwidth h plays an important role in any kernel smoothing, including the local polynomial fitting. A too large bandwidth causes over-smoothing (reducing variance), creating excessive modeling bias, while a too small bandwidth results in under-smoothing (reducing bias but increasing variance), obtaining wiggly estimates. The bandwidth can be subjectively chosen by users via visually inspecting resulting estimates, or automatically chosen by data via minimizing an estimated theoretical risk (discussed later). Since the choice of bandwidth is not easy task, it is often attacked by people who do not know well nonparametric techniques.

(3) Since the estimate is based on the local regression (2.4), it is reasonable to require a non-negative weight function $K(\cdot)$. It can be shown (see Fan and Gijbels (1996)) that for all choices of q and j, the optimal weight function is $K(z) = 3/4\,(1-z^2)_+$, the Epanechnikov kernel, based on minimizing the asymptotic variance of the local polynomial estimator. Thus, it is a universal weighting scheme and provides a useful benchmark for other kernels to compare with. As shown in Fan and Gijbels (1996) and Fan and Yao (2003), other kernels have nearly the same efficiency for practical use of q and j. Hence, the choice of the kernel function is not critical.

The local polynomial estimator compares favorably with other estimators, including the Nadaraya-Watson (local constant) estimator and other linear estimators such as the Gasser and Müller estimator as in Gasser and Müller (979) and the Priestley and Chao estimator as in Priestley and Chao (1972). Indeed, it was shown by Fan (1993) that the local linear fitting is asymptotically minimax based on the quadratic loss function among all linear estimators

and is nearly minimax among all possible linear estimators. This minimax property is extended by Fan et al. (1997) to more general local polynomial fitting. For the detailed comparisons of the above four estimators, see Fan and Gijbels (1996).

Remark 2.3: *Note that the Gasser and Müller estimator and the Priestley and Chao estimator are particularly useful for the fixed design, including the case that $X_t = t$, characterizing the time trend. For more details, the reader is referred to the paper by Cai (2007) for the trending time series setting or* (2.73) *in Section 2.8. Let $s_t = (2t+1)/2$ $(t = 1, \cdots, n-1)$ with $s_0 = -\infty$ and $s_n = \infty$. The Gasser and Müller estimator is*

$$\widehat{m}_{\mathrm{gm}}(t_0) = \sum_{t=1}^{n} \int_{s_{t-1}}^{s_t} K_h(u - t_0)\, du Y_t.$$

Unlike the local constant estimator, no denominator is needed since the total weight

$$\sum_{t=1}^{n} \int_{s_{t-1}}^{s_t} K_h(u - t_0)\, du = 1.$$

Indeed, the Gasser and Müller estimator is an improved version of the Priestley and Chao estimator, which is defined as

$$\widehat{m}_{\mathrm{pc}}(t_0) = \sum_{t=1}^{n} K_h(t - t_0)\, Y_t.$$

Note that the Priestley and Chao estimator is only applicable for the equi-space setting.

Finally, note that when X_t is nonstationary such as unit root or nearly unit root, the theory on the NW and local linear estimators will be totally different from that for the stationary case, which will be investigated in Chapter 5 in detail.

2.3.5 Bandwidth Selection

As seen in previous sections, for stationary sequences of data under certain mixing conditions, the local polynomial estimator performs very much like that for independent data, because windowing reduces dependency among local data. Partially because of this, there are not many studies on bandwidth selection for these problems. However, it is reasonable to expect the bandwidth selectors for independent data continue to work for dependent data with certain mixing conditions. Below, we summarize a few of useful approaches. When data do not have strong enough mixing, the general strategy is to increase bandwidth in order to reduce the variance.

A. Cross-Validation Type Approaches

As what we had already seen for the nonparametric density estimation, the cross-validation method is a very useful tool for assessing the performance of an estimator via estimating its prediction error. The basic idea is to set one of the data point aside for validation of a model and use the remaining data to build the model. It is defined as

$$\mathrm{CV}(h) = \sum_{s=1}^{n} \left[Y_s - \widehat{m}_{-s}\left(X_s\right)\right]^2,$$

where $\widehat{m}_{-s}\left(X_s\right)$ is the local polynomial estimator with $j = 0$ and bandwidth h, but without using the sth observation. The above summand is indeed a squared-prediction error of the sth data point using the training set $\{(X_t, Y_t) : t \neq s\}$. This idea of the cross-validation method is simple but is computationally intensive. An improved version, in terms of computation, is the generalized cross-validation (GCV), proposed by Craven and Wahba (1979). This criterion can be described as follows. The fitted values $\widehat{Y} = \left(\widehat{m}\left(X_1\right), \cdots, \widehat{m}\left(X_n\right)\right)^\top$ can be expressed as $\widehat{Y} = H(h)Y$, where $H(h)$ is an $n \times n$ hat matrix, depending on the $\mathbf{X}$-variate and bandwidth h, and it is also called a smoothing matrix. Then the GCV approach selects the bandwidth h that minimizes

$$\mathrm{GCV}(h) = \left[n^{-1}\mathrm{tr}(I - H(h))\right]^{-2} \mathrm{MASE}(h),$$

where $\mathrm{MASE}(h) = \sum_{t=1}^{n} \left(Y_t - \widehat{m}\left(X_t\right)\right)^2 / n$ is the average of squared residuals.

A drawback of the cross-validation type method is its inherited variability, see, for example, Hall and Johnstone (1992). Further, it cannot be directly applied to select bandwidths for estimating derivative curves. As pointed out by Fan et al. (1995), the cross-validation type method performs poorly due to its large sample variation, even worse for dependent data; see, for example, Shao (1993). Plug-in methods avoid these problems. The basic idea is to find a bandwidth h minimizing estimated mean integrated square error (MISE). See Ruppert et al. (1995) and Fan and Gijbels (1996) for details.

B. Nonparametric AIC Selector

Inspired by the nonparametric version of the Akaike final prediction error criterion proposed by Tjøstheim and Auestad (1994b) for the lag selection in nonparametric setting, Cai and Tiwari (2000) proposed a simple and quick method to select bandwidth for the foregoing estimation procedures, which can be regarded as a nonparametric version of the

AIC to be attentive to the structure of time series data and the overfitting or under-fitting tendency. Note that the idea is also motivated by its analogue of Cai and Tiwari (2000). The basic idea is described as follows.

By recalling the classical AIC for linear models under the likelihood setting

$$-2(\text{ maximized log likelihood}) + 2 \text{ (number of estimated parameters)},$$

Cai and Tiwari (2000) proposed the following nonparametric AIC to select h minimizing

$$\text{AIC}(h) = \log\{\text{MASE}\} + \psi(\text{tr}(H(h)), n), \tag{2.10}$$

where $\psi(\text{tr}(H(h)), n)$ is chosen particularly to be the form of the bias-corrected version of the AIC, due to Hurvich and Tsai (1989),

$$\psi(\text{tr}(H(h)), n) = 2\{\text{tr}(H(h)) + 1\}/[n - \{\text{tr}(H(h)) + 2\}], \tag{2.11}$$

and $\text{tr}(H(h))$ is the trace of the smoothing matrix $H(h)$, regarded as the nonparametric version of degrees of freedom, called the effective number of parameters, denoted by df. See the book by Hastie and Tibshirani (1990) for the detailed discussions on this aspect for nonparametric models.[2] Note that actually, (2.10) is a generalization of the AIC for the parametric regression and autoregressive time series contexts, in which $\text{tr}(H(h))$ is the number of regression (autoregressive) parameters in the fitting model. In view of (2.11), when $\psi(\text{tr}(H(h)), n) = -2\log(1 - \text{tr}(H(h))/n)$, then, (2.10) becomes the generalized cross-validation (GCV) criterion, commonly used to select the bandwidth in the time series literature even in the iid setting, when $\psi(\text{tr}(H(h)), n) = 2\text{tr}(H(h))/n$, then, (2.10) is the classical AIC discussed in Engle et al. (1986) for time series data, and when $\psi(\text{tr}(H(h)), n) = -\log(1 - 2\text{tr}(H(h))/n)$, (2.10) is the T-criterion, proposed and studied by Rice (1984) for the iid samples. It is clear that when $\text{tr}(H(h))/n \to 0$, then the nonparametric AIC, the GCV and the T-criterion are asymptotically equivalent. However, the T-criterion requires $\text{tr}(H(h))/n < 1/2$, and, when $\text{tr}(H(h))/n$ is large, the GCV has relatively weak penalty. This is especially true for the nonparametric setting. Therefore, the criterion proposed here counteracts the over-fitting tendency of the GCV. Note that Hurvich et al. (1998) gave the detailed derivation of the nonparametric AIC for the nonparametric

[2]Indeed, the df can be fined as either $\text{df} = \text{tr}\left(H(h)H(h)^\top\right)$ or $\text{tr}\left(2H(h) - H(h)H(h)^\top\right)$ or the average of the aforementioned two since $\text{tr}(H(h)) \le \text{tr}\left(H(h)H(h)^\top\right) \le \text{tr}\left(2H(h) - H(h)H(h)^\top\right)$. See Hastie and Tibshirani (1990) for details.

regression problems under the iid Gaussian error setting and they argued that the nonparametric AIC performs reasonably well and better than some existing methods in the literature.

C. Multi-Fold Cross-Validation Criterion

Various existing bandwidth selection techniques for nonparametric regression can be adapted for the foregoing estimation; see, e.g., the nonparametric AIC as discussed above. Also, Fan and Gijbels (1996) and Ruppert et al. (1995) developed data-driven bandwidth selection schemes based on asymptotic formulas for the optimal bandwidths, which are less variable and more effective than the conventional data-driven bandwidth selectors such as the cross-validation bandwidth rule.

Indeed, Cai et al. (2000) proposed a simple and quick method for selecting bandwidth h. It can be regarded as a modified multi-fold cross-validation criterion that is attentive to the structure of stationary time series data. Let m and Q be two given positive integers and $n > mQ$. The basic idea is first to use Q sub-series of lengths $n - qm(q = 1, , \cdots, Q)$ to estimate the unknown coefficient functions and then compute the one-step forecasting errors of the next section of the time series of length m based on the estimated models. More precisely, we choose h that minimizes the average mean squared (AMS) error

$$\mathrm{AMS}(h) = \sum_{q=1}^{Q} \mathrm{AMS}_q(h), \tag{2.12}$$

where for $q = 1, \cdots, Q$,

$$\mathrm{AMS}_q(h) = \frac{1}{m} \sum_{t=n-qm+1}^{n-qm+m} \left\{ Y_t - \widehat{m}(X_t) \right\}^2,$$

and $\widehat{m}(\cdot)$ is computed from the sample $\{(X_t, Y_t), 1 \le t \le n - qm\}$ with bandwidth equal $h[n/(n - qm)]^{1/5}$. Note that we re-scale bandwidth h for different sample sizes according to its optimal rate, i.e. $h \propto n^{-1/5}$. In practical implementations, we may use $m = [0.1n]$ and $Q = 4$. The selected bandwidth does not depend critically on the choice of m and Q, as long as mQ is reasonably large so that the evaluation of prediction errors is stable. A weighted version of AMS(h) can be used, if one wishes to down-weight the prediction errors at an earlier time. We believe that this bandwidth should be good for modeling and forecasting for time series.

2.4 Weighted Nadaraya-Watson Estimation

Assume that the process $\{(X_t, Y_t)\}_{t=-\infty}^{\infty}$ is stationary with $X_t \in \mathbb{R}^1$. Of interest is to estimate nonparametrically the regression function $m(x) = \mathbb{E}(\phi(Y_t) \,|\, \mathbf{X}_t = \mathbf{x})$, where $\phi(\cdot)$ is an arbitrary measurable function on the real line and it is assumed that $E|\phi(Y_t)| < \infty$. The introduction of $\phi(\cdot)$ allows us to estimate not only regression function but also conditional distribution ($\phi(Y_t) = I(Y_t \leq y)$ for any fixed y) as in Hall et al. (1999), Cai (2002a), and Cai and Wang (2008), and conditional moment ($\phi(Y_t) = Y_t^r$, $r > 0$). Again, for simplicity, it is assumed that $p = 1$. If $m(u)$ is assumed to has $(q+1)$th continuous derivative, then it can be approximated by a polynomial function as $m(u) \approx \sum_{j=0}^{q} \beta_j \, (u - x)^j$ at x. For the given data $\{(X_t, Y_t)\}_{t=1}^{n}$, at the grid point x, the (locally) weighted least-squared approach is used to find the estimate of $m(x)$. More precisely,

$$\sum_{t=1}^{n} \left[\phi(Y_t) - \sum_{j=0}^{q} \beta_j \, (X_t - x)^j \right]^2 c_{n,t}(x) \, K_h(X_t - x) \tag{2.13}$$

is minimized with respect to $\beta_0, \cdots, \beta_q$, where $c_{n,t}(x)$ is a weight function, dependent on the data $X_1, \cdots, X_n$. The estimator of $m(x)$ is $\widehat{\beta}_0$.

Clearly, when $q = 0$ and $c_{n,t}(x) = 1$, the estimator of $m(x)$ becomes the well-known Nadaraya-Watson estimate, defined by

$$\widehat{m}_{\mathrm{nw}}(x) = \frac{\sum_{t=1}^{n} K_h(X_t - x) \, \phi(Y_t)}{\sum_{t=1}^{n} K_h(X_t - x)},$$

and when $q = 1$ and $c_{n,t}(x) = 1$, it is the local linear estimate

$$\widehat{m}_{\mathrm{ll}}(x) = \sum_{t=1}^{n} w_t \, \phi(Y_t),$$

where $w_t = K_h(X_t - x) \, \{S_{n,2} - (X_t - x) \, S_{n,1}\} / (S_{n,0} \, S_{n,2} - S_{n,1}^2)$ with $S_{n,j} = \sum_{t=1}^{n} K_h(X_t - x) \, (X_t - x)^j$. It is easy to show from Fan and Gijbels (1996) that the weights $\{w_t\}_{t=1}^{n}$ satisfy the following discrete moment conditions as in (2.6)

$$\sum_{t=1}^{n} (X_t - x)^j \, w_t = 0 \tag{2.14}$$

for $j = 0$ and 1. A direct consequence of this relation is that the finite sample bias is zero, but not for the Nadaraya-Watson estimator.

To take both advantages from the Nadaraya-Watson method and local polynomial fitting, we define the weighted Nadaraya-Watson estimate as follows. Let $\{w_t(x)\}_{t=1}^n$ denote the weight functions of the data $X_1, \cdots, X_n$ and the grid point x with the property that $\{w_t(x)\}_{t=1}^n$ satisfy

$$w_t(x) \geq 0, \qquad \sum_{t=1}^{n} w_t(x) = 1, \qquad \text{and} \qquad \sum_{t=1}^{n} (X_t - x)\, w_t(x)\, K_h(X_t - x) = 0. \tag{2.15}$$

Note that the constraint (2.15) is motivated by the property of local linear estimator such as (2.14) and it can be regarded as a generalization of the discrete moment conditions defined in (2.14). Of course, $\{w_t(x)\}_{t=1}^n$ satisfying these conditions are not uniquely defined, and they are specified by maximizing $\prod_{t=1}^n w_t(x)$ subject to the constraint (2.15). The weighted version of Nadaraya-Watson (WNW) estimator of conditional mean $m(x)$ of $\phi(Y_t)$ given $X_t = x$ is defined by

$$\widehat{m}_{\text{wnw}}(x) = \frac{\sum_{t=1}^n w_t(x)\, K_h(X_t - x)\, \phi(Y_t)}{\sum_{t=1}^n w_t(x)\, K_h(X_t - x)}$$

minimizing (2.13) with $c_{n,t}(x) = w_t(x)$ and $q = 0$. Clearly, unlike the local linear estimator, $\min_t\{\phi(Y_t)\} \leq \widehat{m}_{\text{wnw}}(x) \leq \max_t\{\phi(Y_t)\}$ for any x, which is particularly appealing in the estimation of conditional distribution and quantiles; see, e.g., Hall et al. (1999) and Cai (2002a). Cai (2001) showed that $\widehat{m}_{\text{wnw}}(x)$ is first-order equivalent to a local linear estimator and $\widehat{m}_{\text{wnw}}(x)$ has automatic good behavior at boundaries. Especially, note that when $\phi(Y_t) = I(Y_t \leq y)$, the weighted Nadaraya-Watson estimator $\widehat{m}_{\text{wnw}}(x)$ becomes the estimation of the conditional distribution of Y_t given $X_t = x$, which was studied by Hall et al. (1999) and Cai (2002a) for time series data. Also, it is easy to see that $0 \leq \widehat{m}_{\text{wnw}}(x) \leq 1$ and it is monotonically increasing in y.

The natural question arises is how to choose the weights. Because the weight functions $\{w_t(x)\}_{t=1}^n$ satisfy $w_t(x) \geq 0$ and $\sum_{t=1}^n w_t(x) = 1$ for each fixed point x, we can regard $\{w_t(x)\}_{t=1}^n$ as probabilities, so that $\prod_{t=1}^n w_t(x)$ can be viewed as the empirical likelihood function. Therefore, to find the best weight functions $\{w_t(x)\}_{t=1}^n$, we use the empirical likelihood method as in Owen (1988), proceeded as follows. For fixed point x, the logarithm of the empirical likelihood of $\{w_t(x)\}_{t=1}^n$ is $\sum_{t=1}^n \log\{w_t(x)\}$. We maximize the log-empirical likelihood subject to the constraint in (2.15) through the Lagrange multiplier; that is, we maximize

$$\sum_{t=1}^{n} \log\{w_t(x)\} + \lambda_1 \sum_{t=1}^{n} (X_t - x)\, w_t(x)\, K_h(X_t - x) + \lambda_2 \left[\sum_{t=1}^{n} w_t(x) - 1\right]. \tag{2.16}$$

By some algebraic work, the $\{w_t(x)\}$ are simplified to

$$w_t(x) = n^{-1} \left\{1 + \lambda \left(X_t - x\right) K_h(X_t - x)\right\}^{-1},$$

where λ, a function of data and x, is uniquely defined by (2.15), which ensues that $\sum_{t=1}^{n} w_t(x) = 1$. By substituting the above expression of $w_t(x)$ into the log-empirical likelihood, it is easy to see that the maximization of (2.16) is equivalent to choosing λ maximizing

$$L_n(\lambda) = \frac{1}{n\,h} \sum_{t=1}^{n} \log \left\{1 + \lambda \left(X_t - x\right) K_h(X_t - x)\right\},$$

or to finding the root of the equation $L_n'(\lambda) = 0$. In practice, we recommend to use the New-Raphson iteration scheme to find the root of equation $L_n'(\lambda) = 0$; see, for instance, Cai (2002a) for more discussions.

Finally, note that Cai (2001) established the asymptotic normality and weak consistency of the resulting estimator for α-mixing time series at both boundary and interior points, and Cai (2001) showed that the weighted Nadaraya-Watson estimator not only preserve the bias, variance, and more importantly, automatic good boundary behavior properties of local linear estimator, but also makes computation fast. Furthermore, the asymptotic minimax efficiency is discussed, which is almost close to that for the local linear estimator; see the discussions after Theorem 1 in Cai (2001). For details, please see the paper by Cai (2001). In particular, the WNW can be very appealing to estimating the conditional distribution as discussed in Chapter 4.

Remark 2.4: *It is worth pointing out that when X_t is nonstationary such as unit root, the local constant estimator shares exactly same asymptotic bias as that for the local linear estimator. See Reamark 5.3 in Section 5.2.4 for details.*

2.5 Functional Coefficient Model

2.5.1 Model and Its Properties

As mentioned earlier, when p is large, there exists the so called curse of dimensionality. To overcome this shortcoming, one way to do so is to consider some dimension reduction approaches, such as the functional coefficient model as studied in Cai et al. (2000), elaborated in detail next, the additive model discussed in Section 2.6, and some semiparametric models

as in Section 2.7. First, we study the functional coefficient model. To use the notation from Cai et al. (2000), we might change the notation from the previous sections.

Let $\{\mathbf{U}_t, \mathbf{X}_t, Y_t\}_{t=-\infty}^{\infty}$ be jointly strictly stationary processes with $\mathbf{U}_i$ taking values in $\mathbb{R}^k$ and $\mathbf{X}_i$ taking values in $\mathbb{R}^p$. Typically, k is small. Let $\mathbb{E}(Y_1^2) < \infty$. We define the multivariate regression function

$$m(\mathbf{u}, \mathbf{x}) = \mathbb{E}(Y \mid \mathbf{U} = \mathbf{u}, \mathbf{X} = \mathbf{x}), \tag{2.17}$$

where $(\mathbf{U}, \mathbf{X}, Y)$ has the same distribution as $(\mathbf{U}_i, \mathbf{X}_i, Y_i)$. In a pure time series context, both $\mathbf{U}_i$ and $\mathbf{X}_i$ consist of some lagged values of Y_i. The functional-coefficient regression model has the form

$$m(\mathbf{u}, \mathbf{x}) = \sum_{j=1}^{p} a_j(\mathbf{u}) x_j, \tag{2.18}$$

where the functions $\{a_j(\cdot)\}$ are measurable from $\mathbb{R}^k$ to $\mathbb{R}^1$ and $\mathbf{x} = (x_1, \cdots, x_p)^\top$. This model has been studied extensively in the literature; see Cai et al. (2000) for the detailed discussions.

For simplicity, in what follows, we consider only the case $k = 1$ in (2.18). Extension to the case $k > 1$ involves no fundamentally new ideas. Note that models with large k are often not practically useful due to the "curse of dimensionality". If k is large, to overcome the problem, one way to do so is to consider a functional coefficient index model proposed by Fan et al. (2003), given by

$$m(\mathbf{u}, \mathbf{x}) = \sum_{j=1}^{p} a_j\left(\boldsymbol{\beta}^\top \mathbf{u}\right) x_j, \tag{2.19}$$

where $\beta_1 = 1$, and Fan et al. (2003) studied the estimation procedures, bandwidth selection and applications. Furthermore, Cai et al. (2015) considered the model in (2.19) on how to select $\boldsymbol{\beta}$ and $\{a_j(\mathbf{u})\}$ by using the least absolute shrinkage and selection operator (LASSO) type method proposed by Tibshirani (1996).

As elaborated by Cai et al. (2006) and Cai (2010b), functional coefficient models are appropriate and flexible enough for many applications, in particular when additive separability of covariates is unsuitable for the problem at hand. For ease of notation, we assume here that $p = 1$ and $k = 1$. Indeed, by assuming that $m(x, u)$ has a higher order partial derivative with respect to x and applying Taylor expansion to $m(x, u)$, one obtains

$$m(u, x) = \sum_{j=1}^{\infty} \frac{\partial^j m(0, u)}{\partial x^j} \frac{x^j}{j!} \approx \sum_{j=0}^{p} a_j(u) x_j \tag{2.20}$$

for some p (large), where $a_j(u) = (j!)^{-1}\partial^j m(0,u)/\partial x^j$ and $x_j = x^j$. Equation (2.20) implies that a functional coefficient model in (2.18) might be a good approximation to a general nonparametric model in (2.17).

More importantly, as argued in Cai (2010b), the functional coefficient model in (2.18) has an ability to capture heteroscedasticity. To get insights about this, it is easy to see that

$$\mathrm{Var}\left(Y_t \mid \mathbf{U}_t\right) = \mathbf{a}\left(\mathbf{U}_t\right)^\top \mathrm{Var}\left(\mathbf{X}_t \mid \mathbf{U}_t\right)\mathbf{a}\left(\mathbf{U}_t\right) + \sigma_\varepsilon^2\left(\mathbf{U}_t\right),$$

where $\sigma_\varepsilon^2\left(\mathbf{U}_t\right) = \mathrm{Var}\left(\varepsilon_t \mid \mathbf{U}_t\right)$. Therefore, the first term in the above expression behaves as an ARCH type model. Furthermore, the functional coefficient approach allows appreciable flexibility on the structure of fitted models without suffering from the *curse of dimensionality* since the nonparametric estimation is conducted in $\mathbb{R}^k$ instead of $\mathbb{R}^{p+k}$.

Finally, functional coefficient model can be used as a tool to study covariate adjusted regression for situations where both predictors and response in a regression model are not directly observable, but are contaminated with a multiplicative factor that is determined by the value of an unknown function of an observable covariate (confounding variable); see Şentürk and Müller (2006) and Cai and Xu (2008) for more details. For more advantages for the model in (2.18), the reader is referred to the paper by Cai (2010b), in particular, about applying functional coefficient model to analyze economic and financial data. Actually, Hong and Lee (2003) considered the applications of model (2.19) to the exchange rates, Juhl (2005) studied the unit root behavior of nonlinear time series models, Li et al. (2002) modeled the production frontier using China's manufactural industry data, Şentürk and Müller (2006) modeled the nonparametric correlation between two variables using a functional coefficient model as in (2.19), and Cai et al. (2006) considered the nonparametric two-stage instrumental variable estimators for returns to education.

2.5.2 Local Linear Estimation

As recommended by Fan and Gijbels (1996), we estimate the coefficient functions $\{a_j(\cdot)\}$ using the local linear regression method from observations $\{U_t, \mathbf{X}_t, Y_t\}_{t=1}^n$, where $\mathbf{X}_t = (X_{1t}, \cdots, X_{pt})$. We assume throughout that $a_j(\cdot)$ has a continuous second derivative. Note that we may approximate $a_j(\cdot)$ locally at u_0 by a linear function $a_j(u) \approx a_j + b_j\,(u - u_0)$. The local linear estimator is defined as $\widehat{a}_j\,(u_0) = \widehat{a}_j$, where $\left\{\left(\widehat{a}_j, \widehat{b}_j\right)\right\}$ minimize the sum of

weighted squares

$$\sum_{t=1}^{n}\left[Y_t-\sum_{j=1}^{p}\left\{a_j+b_j\left(U_t-u_0\right)\right\}X_{tj}\right]^2 K_h\left(U_t-u_0\right). \tag{2.21}$$

Then, it follows from the least squares theory that

$$\widehat{a}_j\left(u_0\right)=\sum_{k=1}^{n}K_{n,j}\left(U_k-u_0,\mathbf{X}_k\right)Y_k,$$

where

$$K_{n,j}(u,\mathbf{x})=\mathbf{e}_{j,2p}^{\top}\left(\widetilde{\mathbb{X}}^{\top}\mathbf{W}\widetilde{\mathbb{X}}\right)^{-1}\begin{pmatrix}\mathbf{x}\\ u\mathbf{x}\end{pmatrix}K_h(u),$$

$\mathbf{e}_{j,2p}$ is the $2p\times 1$ vector with 1 at the j th position, $\widetilde{\mathbb{X}}$ denotes an $n\times 2p$ matrix with $\left(\mathbf{X}_i^{\top},\mathbf{X}_i^{\top}\left(U_i-u_0\right)\right)$ as its i th row, and $\mathbf{W}=\operatorname{diag}\left\{K_h\left(U_1-u_0\right),\cdots,K_h\left(U_n-u_0\right)\right\}$.

2.5.3 Smoothing Variable Selection and Bandwidth Selection

Of importance is to choose an appropriate smoothing variable U in applying functional coefficient regression models if U is a lagged variable. Knowledge on physical background of the data may be very helpful, as Cai et al. (2000) discussed in modeling the lynx data. Without any prior information, it is pertinent to choose U in terms of some data-driven methods such as the Akaike information criterion and its variants, cross-validation, and other criteria. Ideally, we would choose U as a linear function of given explanatory variables according to some optimal criterion, which can be fully explored in the work by Fan et al. (2003). Nevertheless, we propose here a simple and practical approach: let U be one of the given explanatory variables such that AMS defined in (2.12) obtains its minimum value. Obviously, this idea can be also extended to select p (number of lags) as well. Finally, note that the bandwidth selectors described in Section 2.3.5, particularly, the AMS defined in (2.12) can used to select h.

2.5.4 Goodness-of-Fit Test

To test whether model (2.18) holds with a specified parametric form which is popular in economic and financial applications, such as the threshold autoregressive (TAR) models

$$a_j(u)=\begin{cases}a_{j1}, & \text{if } u\le\eta\\ a_{j2}, & \text{if } u>\eta,\end{cases}$$

or generalized exponential autoregressive (EXPAR) models[3]

$$a_j(u) = \alpha_j + (\beta_j + \gamma_j u) \exp\left(-\theta_j u^2\right),$$

or smooth transition autoregressive (STAR) models

$$\begin{aligned} a_j(u) &= [1 - \exp(-\theta_j u)]^{-1} \quad (\text{ logistic}), \\ &\text{or} \\ a_j(u) &= 1 - \exp\left(-\theta_j u^2\right) \quad (\text{exponential}), \\ &\text{or} \\ a_j(u) &= [1 - \exp(-\theta_j |u|)]^{-1} \quad (\text{absolute}), \end{aligned}$$

this accounts to considering the following general null hypothesis

$$H_0 : a_j(u) = \alpha_j(u, \boldsymbol{\theta}), \quad 1 \le j \le p, \tag{2.22}$$

where $\alpha_j(\cdot, \boldsymbol{\theta})$ is a given family of functions indexed by unknown parameter vector $\boldsymbol{\theta}$. To test H_0 formulated in (2.22), which is about testing nonparametric versus nonlinear parametric, as addressed in Cai et al. (2000), the classical test procedures such as the Wald test, the likelihood ratio test, and the score test (Lagrange multiplier, LM test) for testing parameters in a constrained parameter space like $H_0 : \boldsymbol{\theta} \in \Theta_0$, can not be applied to testing H_0 in (2.22).

To test H_0 in (2.22), we propose a goodness-of-fit test based on the comparison of the residual sum of squares (RSS) from both parametric and nonparametric fittings. This method is closely related to the sieve likelihood method proposed by Fan et al. (2001), which demonstrated the optimality of this kind of procedures for independent samples. To this end, we define the RSS under the null hypothesis as

$$\mathrm{RSS}_0 = n^{-1} \sum_{i=1}^{n} \left\{ Y_i - \alpha_1\left(U_i, \widehat{\boldsymbol{\theta}}\right) X_{i1} - \cdots - \alpha_p\left(U_i, \widehat{\boldsymbol{\theta}}\right) X_{ip} \right\}^2,$$

where $\widehat{\boldsymbol{\theta}}$ is the estimator of $\boldsymbol{\theta}$ under H_0. Analogously, the RSS corresponding to model (2.18) is

$$\mathrm{RSS}_1 = n^{-1} \sum_{i=1}^{n} \left\{ Y_i - \widehat{a}_1(U_i) X_{i1} - \cdots - \widehat{a}_p(U_i) X_{ip} \right\}^2.$$

[3]For more discussions on those models, please see the survey paper by van Dijk et al. (2002).

The test statistic is defined as

$$T_n = \left(\mathrm{RSS}_0 - \mathrm{RSS}_1\right)/\mathrm{RSS}_1 = \mathrm{RSS}_0/\mathrm{RSS}_1 - 1, \tag{2.23}$$

and we reject the null hypothesis (2.22) for large value of T_n. Clearly, T_n can be re-expressed as

$$n\left(T_n + 1\right) \approx n \ln\left(\mathrm{RSS}_0/\mathrm{RSS}_1\right) = -2\log \text{ likelihood ratio },$$

if $\varepsilon_i \sim N\left(0, \sigma^2\right)$. Therefore, T_n is termed as a generalized likelihood ratio (GLR) test in Cai et al. (2000) and the generalized F-test in Cai and Tiwari (2000), which can be used to do testing when regressors are even persistent; see the paper by Zhu et al. (2023).

Since there is no asymptotic theory for the proposed test statistic T_n, we suggest using the following nonparametric Bootstrap approach to evaluate the p value of the test:

1. Generate the Bootstrap residuals $\{\varepsilon_i^*\}_{i=1}^n$ from the empirical distribution of the centered residuals $\{\widehat{\varepsilon}_i - \bar{\varepsilon}\}_{i=1}^n$, where

$$\widehat{\varepsilon}_i = Y_i - \widehat{a}_1\left(U_i\right) X_{i1} - \cdots - \widehat{a}_p\left(U_i\right) X_{ip}, \quad \bar{\widehat{\varepsilon}} = \frac{1}{n}\sum_{i=1}^n \widehat{\varepsilon}_i,$$

   and define

$$Y_i^* = \alpha_1\left(U_i, \widehat{\boldsymbol{\theta}}\right) X_{i1} + \cdots + \alpha_p\left(U_i, \widehat{\boldsymbol{\theta}}\right) X_{ip} + \varepsilon_i^*$$

2. Calculate the Bootstrap test statistic T_n^* based on the sample $\{U_i, \mathbf{X}_i, Y_i^*\}_{i=1}^n$.

3. Reject the null hypothesis H_0 when T_n is greater than the upper-α point of the conditional distribution of T_n^* given $\{(U_i, \mathbf{X}_i, Y_i)\}_{i=1}^n$.

The p-value of the test is simply the relative frequency of the event $\{T_n^* \geq T_n\}$ in the replications of the Bootstrap sampling. For the sake of simplicity, we use the same bandwidth in calculating T_n^* as that in T_n. Note that we Bootstrap the centralized residuals from the nonparametric fit instead of the parametric fit, because the nonparametric estimate of residuals is always consistent, no matter whether the null or the alternative hypothesis is correct. The method should provide a consistent estimator of the null distribution even when the null hypothesis does not hold. Actually, Kreiss et al. (1998) considered nonparametric Bootstrap

tests in a general nonparametric regression setting. They proved that, asymptotically, the conditional distribution of the Bootstrap test statistic is indeed the distribution of the test statistic under the null hypothesis. It may be proven that the similar result holds here as long as $\widehat{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}$ at the rate $n^{-1/2}$. Note that the above nonparametric Bootstrap does not work when the heterogeneity exists. If so, Cai (2007) suggested using the wild Bootstrap instead of the aforementioned nonparametric Bootstrap, see the paper by Cai (2007) for details.

Finally, note that it is a great challenge to derive the asymptotic property of the testing statistic T_n under time series context and some necessary assumptions. That is to show that

$$b_n \left[T_n - \lambda_n \right] \xrightarrow{d} N\left(0, \sigma^2\right) \tag{2.24}$$

for some normalization constants b_n and λ_n, which is a great project for future research. Note that Fan et al. (2001) derived the above result in (2.24) for the iid sample. But, the question that (2.24) is true or not for time series is still open.

2.5.5 Asymptotic Results

We first present a result on mean squared convergence that serves as a building block for our main result and is also of independent interest. We now introduce some notation. Let

$$\mathbf{S}_n = \mathbf{S}_n\left(u_0\right) = \begin{pmatrix} \mathbf{S}_{n,0} & \mathbf{S}_{n,1} \\ \mathbf{S}_{n,1} & \mathbf{S}_{n,2} \end{pmatrix}$$

and

$$\mathbf{T}_n = \mathbf{T}_n\left(u_0\right) = \begin{pmatrix} \mathbf{T}_{n,0}\left(u_0\right) \\ \mathbf{T}_{n,1}\left(u_0\right) \end{pmatrix}$$

with

$$\mathbf{S}_{n,j} = \mathbf{S}_{n,j}\left(u_0\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^{\top} \left(\frac{U_i - u_0}{h}\right)^j K_h\left(U_i - u_0\right)$$

and

$$\mathbf{T}_{n,j}\left(u_0\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i \left(\frac{U_i - u_0}{h}\right)^j K_h\left(U_i - u_0\right) Y_i. \tag{2.25}$$

Then, the solution to (2.21) can be expressed as

$$\widehat{\boldsymbol{\beta}} = \mathbf{H}^{-1}\mathbf{S}_n^{-1}\mathbf{T}_n, \tag{2.26}$$

where $\mathbf{H} = \text{diag}(1, \cdots, 1, h, \cdots, h)$ with p-diagonal elements 1's and p diagonal elements h 's. To facilitate the notation, we denote

$$\mathbf{\Omega} = (\omega_{l,m})_{p\times p} = \mathbb{E}\left(\mathbf{X}\mathbf{X}^\top \mid U = u_0\right)$$

Also, let $f(u, \mathbf{x})$ denote the joint density of $(U, \mathbf{X})$ and $f_u(u)$ be the marginal density of U. We use the following convention: if $U = X_{j_0}$ for some $1 \le j_0 \le p$, then $f(u, \mathbf{x})$ becomes $f(\mathbf{x})$ the joint density of $\mathbf{X}$.

Theorem 2.1: *Let Conditions B1 - B4 hold and $f(u, \mathbf{x})$ be continuous at the point u_0. Let $h_n \to 0$ and $nh_n \to \infty$, as $n \to \infty$. Then it holds that*

$$\mathbb{E}\left(\mathbf{S}_{n,j}\left(u_0\right)\right) \;\to\; f_u\left(u_0\right)\mathbf{\Omega}\left(u_0\right)\mu_j(K),$$

and

$$nh_n\, Var\left(\mathbf{S}_{n,j}\left(u_0\right)_{l,m}\right) \;\to\; f_u\left(u_0\right)\nu_{2j}(K)\omega_{l,m}$$

for each $0 \le j \le 3$ and $1 \le l, m \le p$.

As a consequence of Theorem 2.1, we have

$$\mathbf{S}_n \xrightarrow{\mathcal{P}} f_u\left(u_0\right)\mathbf{S}, \quad \text{and} \quad \mathbf{S}_{n,3} \xrightarrow{\mathcal{P}} \mu_3(K) f_u\left(u_0\right)\mathbf{\Omega}$$

in the sense that each element converges in probability, where

$$\mathbf{S} = \begin{pmatrix} \mathbf{\Omega} & \mu_1(K)\mathbf{\Omega} \\ \mu_1(K)\mathbf{\Omega} & \mu_2(K)\mathbf{\Omega} \end{pmatrix}$$

Put

$$\sigma^2(u, \mathbf{x}) = \text{Var}(Y \mid U = u, \mathbf{X} = \mathbf{x})$$

and

$$\mathbf{\Omega}^*\left(u_0\right) = \mathbb{E}\left[\mathbf{X}\mathbf{X}^\top \sigma^2(U, \mathbf{X}) \mid U = u_0\right].$$

Let $c_0 = \mu_2(K)/\left(\mu_2(K) - \mu_1^2(K)\right)$,$c_1 = -\mu_1(K)/\left(\mu_2(K) - \mu_1^2(K)\right)$, and $c_2 = c_0^2\nu_0(K) + 2c_0c_1\nu_1(K) + c_1^2\nu_2(K)$. When $\mu_1(K) = 0$, $c_0 = 1$, $c_1 = 0$, and $c_2 = \nu_0(K)$.

Theorem 2.2: *Let $\sigma^2(u, \mathbf{x})$ and $f(u, \mathbf{x})$ be continuous at the point u_0. Then under Conditions B1 - B8,*

$$\sqrt{nh_n}\left[\widehat{\mathbf{a}}\left(u_0\right) - \mathbf{a}\left(u_0\right) - \frac{h^2}{2}\frac{\mu_2^2(K) - \mu_1(K)\mu_3(K)}{\mu_2(K) - \mu_1^2(K)}\mathbf{a}''\left(u_0\right) + o_p(h^2)\right] \xrightarrow{d} N\left(0, \Theta^2\left(u_0\right)\right),$$

provided that $f_u\left(u_0\right) \neq 0$, where

$$\Theta^2\left(u_0\right) = \frac{c_2}{f_u\left(u_0\right)}\mathbf{\Omega}^{-1}\left(u_0\right)\mathbf{\Omega}^*\left(u_0\right)\mathbf{\Omega}^{-1}\left(u_0\right).$$

Theorem 2.2 indicates that the asymptotic bias of $\widehat{a}_j(u_0)$ is

$$\frac{h^2}{2}\frac{\mu_2^2(K)-\mu_1(K)\mu_3(K)}{\mu_2(K)-\mu_1^2(K)}a_j''(u_0)$$

and the asymptotic variance is $(nh_n)^{-1}\theta_j^2(u_0)$, where

$$\theta_j^2(u_0)=\frac{c_2}{f_u(u_0)}\mathbf{e}_{j,p}^\top\Omega^{-1}(u_0)\,\mathbf{\Omega}^*(u_0)\,\mathbf{\Omega}^{-1}(u_0)\,\mathbf{e}_{j,p}.$$

When $\mu_1(K)=0$, the bias and variance expressions can be simplified as $h^2\mu_2(K)a_j''(u_0)/2$ and

$$\theta_j^2(u_0)=\frac{\nu_0(K)}{f_u(u_0)}\mathbf{e}_{j,p}^\top\mathbf{\Omega}^{-1}(u_0)\,\mathbf{\Omega}^*(u_0)\,\mathbf{\Omega}^{-1}(u_0)\,\mathbf{e}_{j,p},$$

respectively. The optimal bandwidth for estimating $a_j(\cdot)$ can be defined to be the one that minimizes the squared bias plus variance. Then, when $\mu_1(K)=0$, The optimal bandwidth is given by

$$h_{j,\text{ opt}}=\left[\frac{\nu_0(K)\mathbf{e}_{j,p}^\top\mathbf{\Omega}^{-1}(u_0)\,\mathbf{\Omega}^*(u_0)\,\mathbf{\Omega}^{-1}(u_0)\,\mathbf{e}_{j,p}}{\mu_2^2(K)f_u(u_0)\left\{a_j''(u_0)\right\}^2}\right]^{1/5}\quad n^{-1/5}.$$

2.5.6 Assumptions and Theoretical Proofs

We first impose some conditions on the regression model but they might not be the weakest possible.

Assumptions:

(B1) The kernel function $K(\cdot)$ is a bounded density with a bounded support $[-1,1]$.

(B2) $|f(u,v\mid\mathbf{x}_0,\mathbf{x}_1;l)|\le M<\infty$, for all $l\ge 1$, where $f(u,v,\mid\mathbf{x}_0,\mathbf{x}_1;l)$ is the conditional density of $(U_0,U_l))$ given $(\mathbf{X}_0,\mathbf{X}_l)$, and $f(u\mid\mathbf{x})\le M<\infty$, where $f(u\mid\mathbf{x})$ is the conditional density of U given $\mathbf{X}=\mathbf{x}$.

(B3) The process $\{U_i,\mathbf{X}_i,Y_i\}$ is α-mixing with $\sum k^c[\alpha(k)]^{1-2/\delta}<\infty$ for some $\delta>2$ and $c>1-2/\delta$.

(B4) $E|\mathbf{X}|^{2\delta}<\infty$, where δ is given in Condition B3.

(B5) Assume that

$$\mathbb{E}\left\{Y_0^2 + Y_l^2 \mid U_0 = u, \mathbf{X}_0 = \mathbf{x}_0; U_l = v, \mathbf{X}_l = \mathbf{x}_1\right\} \leq M < \infty \tag{2.27}$$

for all $l \geq 1, \mathbf{x}_0, \mathbf{x}_1 \in \mathbb{R}^p, u$, and v in a neighborhood of u_0.

(B6) Assume that $h_n \to$ and $n\,h_n \to \infty$. Further, assume that there exists a sequence of positive integers s_n such that $s_n \to \infty$, $\quad s_n = o\left((n\,h_n)^{1/2}\right)$, and $(n/h_n)^{1/2}\,\alpha\,(s_n) \to 0$, as $n \to \infty$

(B7) There exists $\delta^* > \delta$, where δ is given in Condition B3, such that

$$\mathbb{E}\left\{|Y|^{\delta^*} \mid U = u, \mathbf{X} = \mathbf{x}\right\} \leq M_4 < \infty$$

for all $\mathbf{x} \in \mathbb{R}^p$ and u in a neighborhood of u_0, and

$$\alpha(n) = O\left(n^{-\theta^*}\right),$$

where $\theta^* \geq \delta\,\delta^*/\left\{2\,(\delta^* - \delta)\right\}$

(B8) $E|\mathbf{X}|^{2\delta^*} < \infty$, and $n^{1/2-\delta/4}h^{\delta/\delta^*-1/2-\delta/4} = O(1)$

Remark 2.5: *We provide a sufficient condition for the mixing coefficient $\alpha(n)$ to satisfy Conditions B3 and B6. Suppose that $h_n = An^{-\rho}(0 < \rho < 1, A > 0)$, $s_n = (nh_n/\log n)^{1/2}$ and $\alpha(n) = O\left(n^{-d}\right)$ for some $d > 0$. Then, Condition B3 is satisfied for $d > 2(1-1/\delta)/(1-2/\delta)$ and Condition B6 is satisfied if $d > (1+\rho)/(1-\rho)$. Hence both conditions are satisfied if*

$$\alpha(n) = O\left(n^{-d}\right), \qquad d > \max\left\{\frac{1+\rho}{1-\rho}, \frac{2(1-1/\delta)}{1-2/\delta}\right\}.$$

Note that this is a trade-off between the order δ of the moment of Y and the rate of decay of the mixing coefficient; the larger the order δ, the weaker the decay rate of $\alpha(n)$.

To study the joint asymptotic normality of $\widehat{\mathbf{a}}\,(u_0)$, we need to center the vector $\mathbf{T}_n\,(u_0)$ by replacing Y_i with $Y_i - m\,(U_i, \mathbf{X}_i)$ in the expression (2.25) of $\mathbf{T}_{n,j}\,(u_0)$. Let

$$\mathbf{T}^*_{n,j}\,(u_0) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\left(\frac{U_i - u_0}{h}\right)^j K_h\,(U_i - u_0)\,[Y_i - m\,(U_i, \mathbf{X}_i)]\,,$$

and

$$\mathbf{T}^*_n = \begin{pmatrix} \mathbf{T}^*_{n,0} \\ \mathbf{T}^*_{n,1} \end{pmatrix}.$$

Because the coefficient functions $a_j(u)$ are conducted in the neighborhood of $|U_i - u_0| < h$, by Taylor's expansion,

$$m\left(U_i, \mathbf{X}_i\right) = \mathbf{X}_i^{\top}\mathbf{a}\left(u_0\right) + \left(U_i - u_0\right)\mathbf{X}_i^{\top}\mathbf{a}^{\top}\left(u_0\right) + \frac{h^2}{2}\left(\frac{U_i - u_0}{h}\right)^2 \mathbf{X}_i^{\top}\mathbf{a}''\left(u_0\right) + o_p\left(h^2\right),$$

where $\mathbf{a}^{\top}\left(u_0\right)$ and $\mathbf{a}''\left(u_0\right)$ are the vectors consisting of the first and second derivatives of the functions $a_j(\cdot)$. Then,

$$\mathbf{T}_{n,0} - \mathbf{T}_{n,0}^{*} = \mathbf{S}_{n,0}\mathbf{a}\left(u_0\right) + h\mathbf{S}_{n,1}\mathbf{a}^{\top}\left(u_0\right) + \frac{h^2}{2}\mathbf{S}_{n,2}\mathbf{a}''\left(u_0\right) + o_p\left(h^2\right)$$

and

$$\mathbf{T}_{n,1} - \mathbf{T}_{n,1}^{*} = \mathbf{S}_{n,1}\mathbf{a}\left(u_0\right) + h\mathbf{S}_{n,2}\mathbf{a}^{\top}\left(u_0\right) + \frac{h^2}{2}\mathbf{S}_{n,3}\mathbf{a}''\left(u_0\right) + o_p\left(h^2\right)$$

so that

$$\mathbf{T}_n - \mathbf{T}_n^{*} = \mathbf{S}_n\mathbf{H}\boldsymbol{\beta} + \frac{h^2}{2}\begin{pmatrix} \mathbf{S}_{n,2} \\ \mathbf{S}_{n,3} \end{pmatrix}\mathbf{a}''\left(u_0\right) + o_p\left(h^2\right) \tag{2.28}$$

where $\boldsymbol{\beta} = \left(\mathbf{a}\left(u_0\right)^{\top}, \mathbf{a}^{\top}\left(u_0\right)^{\top}\right)^{\top}$. Thus it follows from (2.26), (2.28), and Theorem 2.1 that

$$\mathbf{H}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = f_u^{-1}\left(u_0\right)\mathbf{S}^{-1}\mathbf{T}_n^{*} + \frac{h^2}{2}\mathbf{S}^{-1}\begin{pmatrix} \mu_2(K)\boldsymbol{\Omega} \\ \mu_3(K)\boldsymbol{\Omega} \end{pmatrix}\mathbf{a}''\left(u_0\right) + o_p\left(h^2\right)$$

from which the bias term of $\widehat{\boldsymbol{\beta}}\left(u_0\right)$ is evident. Clearly,

$$\begin{aligned} \widehat{\mathbf{a}}\left(u_0\right) - \mathbf{a}\left(u_0\right) &= \frac{\boldsymbol{\Omega}^{-1}}{f_u\left(u_0\right)\left(\mu_2(K) - \mu_1^2(K)\right)}\left[\mu_2(K)\mathbf{T}_{n,0}^{*} - \mu_1(K)\mathbf{T}_{n,1}^{*}\right] \\ &\quad + \frac{h^2}{2}\frac{\mu_2^2(K) - \mu_1(K)\mu_3(K)}{\mu_2(K) - \mu_1^2(K)}\mathbf{a}''\left(u_0\right) + o_p\left(h^2\right). \end{aligned} \tag{2.29}$$

Thus, (2.29) indicates that the asymptotic bias of $\widehat{\mathbf{a}}\left(u_0\right)$ is

$$\frac{h^2}{2}\frac{\mu_2^2(K) - \mu_1(K)\mu_3(K)}{\mu_2(K) - \mu_1^2(K)}\mathbf{a}''\left(u_0\right)$$

Let

$$\mathbf{Q}_n = \frac{1}{n}\sum_{i=1}^{n}\mathbf{Z}_i, \tag{2.30}$$

where

$$\mathbf{Z}_i = \mathbf{X}_i\left[c_0 + c_1\left(\frac{U_i - u_0}{h}\right)\right]K_h\left(U_i - u_0\right)\left[Y_i - m\left(U_i, \mathbf{X}_i\right)\right].$$

Then, it follows from (2.29) and (2.30) that

$$\sqrt{nh_n}\left[\widehat{\mathbf{a}}\left(u_0\right) - \mathbf{a}\left(u_0\right) - \frac{h^2}{2}\frac{\mu_2^2(K) - \mu_1(K)\mu_3(K)}{\mu_2(K) - \mu_1^2(K)}\mathbf{a}''\left(u_0\right)\right] = \frac{\Omega^{-1}}{f_u\left(u_0\right)}\sqrt{nh_n}\mathbf{Q}_n + o_p(1).$$

To finish the proof, one needs the following lemma, whose proof is more involved than that for Theorem 2.1. Therefore, we prove only this lemma. Throughout this section, we let C denote a generic constant, which may take different values at different places.

Lemma 2.1: Under Conditions B1 - B8 and the assumption that $h_n \to 0$ and $nh_n \to \infty$, as $n \to \infty$, if $\sigma^2(u, \mathbf{x})$ and $f(u, \mathbf{x})$ are continuous at the point u_0, then we have

(a) $h_n \mathrm{Var}\left(\mathbf{Z}_1\right) \;\to\; c_2\, f_u\left(u_0\right) \Omega^*\left(u_0\right)$,

(b) $h_n \sum_{l=1}^{n-1|} \left|\mathrm{Cov}\left(\mathbf{Z}_1, \mathbf{Z}_{l+1}\right)\right| = o(1)$, and

(c) $nh_n \mathrm{Var}\left(\mathbf{Q}_n\right) \;\to\; c_2\, f_u\left(u_0\right) \mathbf{\Omega}^*\left(u_0\right)$.

Proof: First, by conditioning on $(U_1, \mathbf{X}_1)$ and using Theorem 1 of Sun (1984), we have

$$\begin{aligned}\mathrm{Var}\left(\mathbf{Z}_1\right) &= \mathbb{E}\left[\mathbf{X}_1\mathbf{X}_1^\top \sigma^2\left(U_1, \mathbf{X}_1\right)\left\{c_0 + c_1\left(\frac{U_1 - u_0}{h}\right)\right\}^2 K_h^2\left(U_1 - u_0\right)\right] \\ &= \frac{1}{h}\left[c_2\, f_u\left(u_0\right) \mathbf{\Omega}^*\left(u_0\right) + o(1)\right]\end{aligned}$$

The result (c) follows in an obvious manner from (a) and (b) along with

$$\mathrm{Var}\left(\mathbf{Q}_n\right) = \frac{1}{n}\mathrm{Var}\left(\mathbf{Z}_1\right) + \frac{2}{n}\sum_{l=1}^{n-1}\left(1 - \frac{l}{n}\right)\mathrm{Cov}\left(\mathbf{Z}_1, \mathbf{Z}_{l+1}\right).$$

It thus remains to prove part (b). To this end, let $d_n \to \infty$ be a sequence of positive integers such that $d_n\, h_n \to 0$. Define

$$J_1 = \sum_{l=1}^{d_n - 1}\left|\mathrm{Cov}\left(\mathbf{Z}_1, \mathbf{Z}_{l+1}\right)\right| \quad \text{and} \quad J_2 = \sum_{l=d_n}^{n-1}\left|\mathrm{Cov}\left(\mathbf{Z}_1, \mathbf{Z}_{l+1}\right)\right|$$

It remains to show that $J_1 = o\left(h^{-1}\right)$ and $J_2 = o\left(h^{-1}\right)$.

We remark that because $K(\cdot)$ has a bounded support $[-1, \quad 1]$, $a_j(u)$ is bounded in the neighborhood of $u \in [u_0 - h, \quad u_0 + h]$. Let $B = \max_{1\le j\le p} \sup_{|u-u_0|<h} |a_j(u)|$ and $g(\mathbf{x}) = \sum_{j=1}^p |x_j|$. Then $\sup_{|u-u_0|<h} |m(u, \mathbf{x})| \le B\, g(\mathbf{x})$. By conditioning on $(U_1, \mathbf{X}_1)$ and $(U_{l+1}, \mathbf{X}_{l+1})$, and using (2.27) and Condition B2, we have, for all $l \ge 1$,

$$\begin{aligned}&\left|\mathrm{Cov}\left(\mathbf{Z}_1, \mathbf{Z}_{l+1}\right)\right| \\ &\le C\mathbb{E}\left[\left|\mathbf{X}_1\mathbf{X}_{l+1}^\top\right|\left\{|Y_1| + Bg\left(\mathbf{X}_1\right)\right\}\left\{|Y_{l+1}| + Bg\left(\mathbf{X}_{l+1}\right)\right\} K_h\left(U_1 - u_0\right) K_h\left(U_{l+1} - u_0\right)\right] \\ &\le C\mathbb{E}\left[\left|\mathbf{X}_1\mathbf{X}_{l+1}^\top\right|\left\{M_2 + B^2 g^2\left(\mathbf{X}_1\right)\right\}^{1/2}\left\{M_2 + B^2 g^2\left(\mathbf{X}_{l+1}\right)\right\}^{1/2} K_h\left(U_1 - u_0\right) K_h\left(U_{l+1} - u_0\right)\right] \\ &\le C\mathbb{E}\left[\left|\mathbf{X}_1\mathbf{X}_{l+1}^\top\right|\left\{1 + g\left(\mathbf{X}_1\right)\right\}\left\{1 + g\left(\mathbf{X}_{l+1}\right)\right\}\right] \le C.\end{aligned}$$

It follows that

$$J_1 \leq Cd_n = o\left(h^{-1}\right)$$

by the choice of d_n. We next consider the upper bound of J_2. To this end, using the Davydov's inequality (see Lemma 1.1), we obtain, for all $1 \leq j, m \leq p$ and $l \geq 1$,

$$|\mathrm{Cov}\left(Z_{1j}, Z_{l+1,m}\right)| \leq C[\alpha(l)]^{1-2/\delta}\left[E\left|Z_j\right|^{\delta}\right]^{1/\delta}\left[E\left|Z_m\right|^{\delta}\right]^{1/\delta}. \tag{2.31}$$

By conditioning on $(U, \mathbf{X})$ and using Conditions B2 and B7, one has

$$\begin{aligned}
\mathbb{E}\left[|Z_j|^{\delta}\right] &\leq C\mathbb{E}\left[|X_j|^{\delta} K_h^{\delta}\left(U-u_0\right)\left\{|Y|^{\delta}+B^{\delta}g^{\delta}(\mathbf{X})\right\}\right] \\
&\leq C\mathbb{E}\left[|X_j|^{\delta} K_h^{\delta}\left(U-u_0\right)\left\{M_3+B^{\delta}g^{\delta}(\mathbf{X})\right\}\right] \\
&\leq Ch^{1-\delta}\mathbb{E}\left[|X_j|^{\delta}\left\{M_3+B^{\delta}g^{\delta}(\mathbf{X})\right\}\right] \leq Ch^{1-\delta}.
\end{aligned} \tag{2.32}$$

A combination of (2.31) and (2.32) leads to

$$J_2 \leq C\, h^{2/\delta-2}\sum_{l=d_n}^{\infty}[\alpha(l)]^{1-2/\delta} \leq C\, h^{2/\delta-2}d_n^{-c}\sum_{l=d_n}^{\infty} l^c[\alpha(l)]^{1-2/\delta} = o\left(h^{-1}\right)$$

by choosing d_n such that $h^{1-2/\delta}d_n^c = C$, so the requirement that $d_n\, h_n \to 0$ is satisfied. □

Proof of Theorem 2.2

We use the small-block and large-block technique-namely, partition $\{1, \cdots, n\}$ into $2q_n+1$ subsets with large block of size $r = r_n$ and small block of size $s = s_n$. Set

$$q = q_n = \left\lfloor \frac{n}{r_n + s_n} \right\rfloor. \tag{2.33}$$

We now use the CramÃ©r-Wold device to derive the asymptotic normality of $\mathbf{Q}_n$. For any unit vector $\mathbf{d} \in \mathbb{R}^p$, let $Z_{n,i} = \sqrt{h}\mathbf{d}^{\top}\mathbf{Z}_{i+1}, i = 0, \cdots, n-1$. Then,

$$\sqrt{nh}\mathbf{d}^{\top}\mathbf{Q}_n = \frac{1}{\sqrt{n}}\sum_{i=0}^{n-1} Z_{n,i},$$

and, by Lemma 2.1,

$$\mathrm{Var}\left(Z_{n,0}\right) \approx f_u\left(u_0\right)\mathbf{d}^{\top}\mathbf{\Omega}^*\left(u_0\right)\mathbf{d}\left[c_0^2\nu_0(K) + 2c_0c_1\nu_1 + c_1^2\nu_2\right] \equiv \theta^2\left(u_0\right)$$

and

$$\sum_{l=0}^{n-1}\left|\mathrm{Cov}\left(Z_{n,0}, Z_{n,l}\right)\right| = o(1).$$

Define the random variables, for $0 \leq j \leq q-1$,

$$\eta_j = \sum_{i=j(r+s)}^{j(r+s)+r-1} Z_{n,i}, \quad \xi_j = \sum_{i=j(r+s)+r}^{(j+1)(r+s)} Z_{n,i}, \quad \text{and} \quad \zeta_q = \sum_{i=q(r+s)}^{n-1} Z_{n,i}.$$

Then,

$$\sqrt{n\,h}\mathbf{d}^\top \mathbf{Q}_n = \frac{1}{\sqrt{n}} \left\{ \sum_{j=0}^{q-1} \eta_j + \sum_{j=0}^{q-1} \xi_j + \zeta_q \right\} \equiv \frac{1}{\sqrt{n}} \left\{ Q_{n,1} + Q_{n,2} + Q_{n,3} \right\}.$$

We show that as $n \to \infty$

$$\frac{1}{n}\mathbb{E}\left[Q_{n,2}\right]^2 \;\to\; 0, \quad \frac{1}{n}\mathbb{E}\left[Q_{n,3}\right]^2 \;\to\; 0, \tag{2.34}$$

$$\left| \mathbb{E}\left[\exp\left(itQ_{n,1}\right)\right] - \prod_{j=0}^{q-1} \mathbb{E}\left[\exp\left(it\eta_j\right)\right] \right| \;\to\; 0, \tag{2.35}$$

$$\frac{1}{n}\sum_{j=0}^{q-1} \mathbb{E}\left(\eta_j^2\right) \;\to\; \theta^2\left(u_0,\right), \tag{2.36}$$

and

$$\frac{1}{n}\sum_{j=0}^{q-1} \mathbb{E}\left[\eta_j^2 I\left\{|\eta_j| \geq \varepsilon\theta\left(u_0\right)\sqrt{n}\right\}\right] \;\to\; 0 \tag{2.37}$$

for every $\varepsilon > 0$. (2.33) implies that $Q_{n,2}$ and $Q_{n,3}$ are asymptotically negligible in probability, (2.35) shows that the summands η_j in $Q_{n,1}$ are asymptotically independent and (2.36) and (2.37) are the standard Lindeberg-Feller conditions for asymptotic normality of $Q_{n,1}$ for the independent setup.

First, we establish (2.34). For this purpose, we choose the large block size. Condition B6 implies that there is a sequence of positive constants $\gamma_n \to \infty$ such that $\gamma_n s_n = o\left(\sqrt{nh_n}\right)$ and

$$\gamma_n \left(n/h_n\right)^{1/2} \alpha\left(s_n\right) \;\to\; 0. \tag{2.38}$$

Define the large block size r_n by $r_n = \left\lfloor \left(n\,h_n\right)^{1/2}/\gamma_n \right\rfloor$ and the small block size s_n. Then it can easily be shown from (2.38) that as $n \to \infty$,

$$s_n/r_n \;\to\; 0, \quad r_n/n \;\to\; 0, \quad r_n\left(nh_n\right)^{-1/2} \;\to\; 0, \tag{2.39}$$

and

$$\left(n/r_n\right)\alpha\left(s_n\right) \;\to\; 0. \tag{2.40}$$

Observe that

$$\mathbb{E}\left[Q_{n,2}\right]^2 = \sum_{j=0}^{q-1} \mathrm{Var}\left(\xi_j\right) + 2 \sum_{0 \le i < j \le q-1} \mathrm{Cov}\left(\xi_i, \xi_j\right) \equiv I_1 + I_2. \tag{2.41}$$

It follows from stationarity and Lemma 2.1 that

$$I_1 = q_n \mathrm{Var}\left(\xi_1\right) = q_n \mathrm{Var}\left(\sum_{j=1}^{s_n} Z_{n,j}\right) = q_n s_n \left[\theta^2\left(u_0\right) + o(1)\right].$$

Next, consider the second term I_2 in the right side of (2.41). Let $r_j^* = j\left(r_n + s_n\right)$, then $r_j^* - r_i^* \ge r_n$ for all $j > i$. Thus, we have

$$\begin{aligned} |I_2| &\le 2 \sum_{0 \le i < j \le q-1} \sum_{j_1=1}^{s_n} \sum_{j_2=1}^{s_n} \left|\mathrm{Cov}\left(Z_{n, r_i^* + r_n + j_1}, Z_{n, r_j^* + r_n + j_2}\right)\right| \\ &\le 2 \sum_{j_1=1}^{n-r_n} \sum_{j_2=j_1+r_n}^{n} \left|\mathrm{Cov}\left(Z_{n,j_1}, Z_{n,j_2}\right)\right|. \end{aligned}$$

By stationarity and Lemma 2.1, one obtains

$$|I_2| \le 2n \sum_{j=r_n+1}^{n} \left|\mathrm{Cov}\left(Z_{n,1}, Z_{n,j}\right)\right| = o(n). \tag{2.42}$$

Hence, by (2.39) - (2.42), we have

$$\frac{1}{n}\mathbb{E}\left[Q_{n,2}\right]^2 = O\left(q_n s_n n^{-1}\right) + o(1) = o(1). \tag{2.43}$$

It follows from stationarity, (2.39), and Lemma 2.1 that

$$\mathrm{Var}\left[Q_{n,3}\right] = \mathrm{Var}\left(\sum_{j=1}^{n-q_n(r_n+s_n)} Z_{n,j}\right) = O\left(n - q_n\left(r_n + s_n\right)\right) = o(n). \tag{2.44}$$

Combining (2.39), (2.43), and (2.44), we establish (2.34). As for (2.36) by stationarity, (2.39), (2.40), and Lemma 2.1, it is easily seen that

$$\frac{1}{n}\sum_{j=0}^{q_n-1} \mathbb{E}\left(\eta_j^2\right) = \frac{q_n}{n}\mathbb{E}\left(\eta_1^2\right) = \frac{q_n r_n}{n} \cdot \frac{1}{r_n}\mathrm{Var}\left(\sum_{j=1}^{r_n} Z_{n,j}\right) \;\to\; \theta^2\left(u_0\right).$$

To establish (2.35), we use Lemma 1.1 of Volkonskii and Rozanov (1959); see, also Ibragimov and Linnik (1971), to obtain

$$\left|\mathbb{E}\left[\exp\left(itQ_{n,1}\right)\right] - \prod_{j=0}^{q_n-1} \mathbb{E}\left[\exp\left(it\eta_j\right)\right]\right| \le 16\left(n/r_n\right)\alpha\left(s_n\right)$$

tending to 0 by (2.40).

It remains to establish (2.37). For this purpose, we use Theorem 4.1 in Shao and Yu (1996) and Conditions B5 - B8 to obtain

$$\mathbb{E}\left[\eta_1^2 I\left\{|\eta_1| \geq \varepsilon\theta\left(u_0\right)\sqrt{n}\right\}\right] \leq Cn^{1-\delta/2}\mathbb{E}\left(|\eta_1|^{\delta}\right) \leq Cn^{1-\delta/2}r_n^{\delta/2}\left\{\mathbb{E}\left(|Z_{n,0}|^{\delta^*}\right)\right\}^{\delta/\delta^*} \quad (2.45)$$

As in (2.32),

$$\mathbb{E}\left(|Z_{n,0}|^{\delta^*}\right) \leq Ch^{1-\delta^*/2}. \quad (2.46)$$

Therefore, by (2.45) and (2.46),

$$\mathbb{E}\left[\eta_1^2 I\left\{|\eta_1| \geq \varepsilon\theta\left(u_0\right)\sqrt{n}\right\}\right] \leq Cn^{1-\delta/2}r_n^{\delta/2}h^{(2-\delta^*)\delta/(2\delta^*)}.$$

Thus, by (2.34) and the definition of r_n, and using Conditions B7 and B8, we obtain

$$\frac{1}{n}\sum_{j=0}^{q-1}\mathbb{E}\left[\eta_j^2 I\left\{|\eta_j| \geq \varepsilon\theta\left(u_0\right)\sqrt{n}\right\}\right] \leq C\gamma_n^{1-\delta/2}n^{1/2-\delta/4}h_n^{\delta/\delta^*-1/2-\delta/4} \quad \rightarrow \quad 0,$$

because $\gamma_n \rightarrow \infty$. This completes the proof of the theorem.

2.5.7 Applications

1. Analysis Of Boston Housing Data

A. Description of Data

The well known **Boston house price data** set[4] consists of 14 variables, collected on each of 506 different houses from a variety of locations. The Boston house-price data set was used originally by Harrison and Rubinfeld (1978) and it was re-analyzed in Belsley et al. (1980) by various transformations in the table on pages 244-261. Variables are, denoted by $X_1, \cdots, X_{13}$ and Y, in order: The dependent variable is Y, the median value of owner-occupied homes in $\$1,000$'s (house price). The major factors possibly affecting the house prices used in the literature are: X_{13} = proportion of population of lower educational status X_6 = the average number of rooms per house, X_1 = the per capita crime rate, X_{10} = the full property tax rate, and X_{11} = the pupil/teacher ratio. For the complete description of all 14 variables, see Harrison and Rubinfeld (1978) and Gilley et al. (1996) for corrections.

[4]This dataset can be downloaded from the web site at http://lib.stat.cmu.edu/datasets/boston.

CRIM	per capita crime rate by town
ZN	proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS	proportion of non-retail business acres per town
CHAS	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
NOX	nitric oxides concentration (parts per 10 million)
RM	average number of rooms per dwelling
AGE	proportion of owner-occupied units built prior to 1940
DIS	weighted distances to five Boston employment centers
RAD	index of accessibility to radial highways
TAX	full-value property-tax rate per 10,000USD
PTRATIO	pupil-teacher ratio by town
B	$1000(\text{Bk} - 0.63)\text{\^{}}2$ where Bk is the proportion of blacks by town
LSTAT	lower status of the population
MEDV	Median value of owner-occupied homes in $1000â s

B. Linear Models

Harrison and Rubinfeld (1978) was the first to analyze this data set using a standard regression model Y versus all 13 variables including some higher order terms or transformations on Y and X_j's. The purpose of this study is to see whether there are the effects of pollution on housing prices via hedonic pricing methodology. Belsley et al. (1980) used this data set to illustrate the effects of using robust regression and outlier detection strategies. From these results, we might conclude that the model might not be linear and there might exist outliers. Also, Pace and Gilley (1997) added a geo-referencing idea (spatial statistics) and used a spatial estimation method to consider this data set.

Exercise: Please use all possible methods to explore this dataset to see what is the best linear model you can obtain.

C. Fit a Varying-Coefficient Model

Şentürk and Müller (2006) studied the correlation between the house price Y and the crime rate X_1 adjusted by the confounding variable X_{13} through a varying coefficient model

and they concluded that the expected effect of increasing crime rate on declining house prices seems to be only observed for lower educational status neighborhoods in Boston. Finally, it is surprising that all the existing nonparametric models aforementioned above did not include the crime rate X_1, which may be an important factor affecting the housing price, and did not consider the interaction terms such as X_{13} and X_1. See the paper by Fan and Huang (2005) for fitting a varying coefficient partially linear model to the Boston housing data, which will be discussed in detail in Sections 2.6.5 and 2.7.1, respectively. To fit robustly this dataset, a quantile approach might be better due to the existence of some outliers; see Section 3.5.2 for the detailed analyses.

Exercise: Please fit a a varying coefficient model to the Boston housing data.

2. Functional Coefficient Capital Asset Pricing Model

The model in (2.17) was successfully applied by Cai et al. (2015) to study the conditional capital asset pricing model (CAPM) to argue that the $\boldsymbol{\beta}$ in the conventional CAPM changes over time.

$$r_t = \boldsymbol{\beta}(\mathbf{Z}_t)^\top \mathbf{X}_t + \varepsilon_t, \tag{2.47}$$

where r_t is the return for an asset, $\mathbf{X}_t$ is a vector of factors, say, factors in the three (four, five or six)-factors Fama-French type model, and $\mathbf{Z}_t$ is a variable that drives the $\boldsymbol{\beta}$ to change over time. If the dimension of $\mathbf{Z}_t$ is large, which is typically true for financial applications, because that $\mathbf{Z}_t$ might contain many macroeconomic variables and financial characteristics variables, Cai et al. (2015) extended the model in (2.47) into the following functional coefficient index model

$$r_t = \boldsymbol{\beta}(\gamma^\top \mathbf{Z}_t)^\top \mathbf{X}_t + \varepsilon_t, \tag{2.48}$$

which is a semiparametric model and will be discussed further in Section 2.7.3.

As argued by Cai et al. (2015), the functional coefficient representation relaxes the strict assumptions regarding the structure of betas and alpha by combining the predictors into an index. Appropriate index variables are selected by applying the smoothly clipped absolute deviation penalty. In such a way, estimation and variable selection can be done simultaneously. Based on the empirical studies, the proposed model performs better than the alternatives in explaining asset returns and they found no strong evidence to reject the conditional CAPM.

2.6 Additive Model

2.6.1 Model Framework

In this section, we use the notation from Cai (2002b). Let $\{(\mathbf{X}_t, Y_t, \mathbf{Z}_t)\}_{t=-\infty}^{\infty}$ be jointly stationary processes, where $\mathbf{X}_t$ and $\mathbf{Z}_t$ take values in $\mathbb{R}^p$ and $\mathbb{R}^q$ with $p, q \geq 0$, respectively. The regression surface is defined by

$$m(\mathbf{x}, \mathbf{z}) = \mathbb{E}\left\{Y_t \mid \mathbf{X}_t = \mathbf{x}, \mathbf{Z}_t = \mathbf{z}\right\}. \tag{2.49}$$

Here, it is assumed that $\mathbb{E}|Y_t| < \infty$. Note that the regression function $m(\cdot, \cdot)$ defined in (2.49) can identify only the sum

$$m(\mathbf{x}, \mathbf{z}) = \mu + g_1(\mathbf{x}) + g_2(\mathbf{z}), \tag{2.50}$$

which covers the motivative example in (1) as a special case. Such a decomposition in (2.50) holds, for example, for the following nonlinear additive autoregressive model with exogenous variables, termed as an ARX model in nonlinear time series literature,

$$Y_t = \mu + g_1\left(X_{t-j_1}, \cdots, X_{t-j_p}\right) + g_2\left(Y_{t-i_1}, \cdots, Y_{t-i_q}\right) + \eta_t,$$

and

$$X_{t-j_1} = g_3\left(X_{t-j_2}, \cdots, X_{t-j_p}\right) + \varepsilon_t.$$

For detailed discussions on the ARX model, the reader is referred to the papers by Masry and Tjøstheim (1997) and Cai and Masry (2000). For identifiability, it is assumed that $\mathbb{E}\left\{g_1\left(\mathbf{X}_t\right)\right\} = 0$ and $\mathbb{E}\left\{g_2\left(\mathbf{Z}_t\right)\right\} = 0$, where $\mathbf{Z}_t = \left(Y_{t-i_1}, \cdots, Y_{t-i_q}\right)^\top$. Then, $\mu = \mathbb{E}(Y_t)$ and the projection of $m(\mathbf{x}, \mathbf{y})$ on the $g_1(\mathbf{x})$-direction is defined by

$$\mathbb{E}\left\{m\left(\mathbf{x}, \mathbf{Z}_t\right)\right\} = \mu + g_1(\mathbf{x}) + \mathbb{E}\left\{g_2\left(\mathbf{Z}_t\right)\right\} = \mu + g_1(\mathbf{x}), \tag{2.51}$$

which implies that $g_1(\cdot)$ can be identified up to an additive constant and $g_2(\cdot)$ can be retrieved likewise.

A thorough discussion of additive time series models defined in (2.50) can be found in Chen and Tsay (1993). Additive components can be estimated with a one-dimensional nonparametric rate. In most papers, to estimate additive components, several methods have been proposed. For example, Chen and Tsay (1993) used the iterative backfitting procedures,

such as the ACE algorithm and the BRUTO approach; see, e.g., Hastie and Tibshirani (1990) for details. But, their asymptotic properties are not well understood due to the implicit definition of the resulting estimators. To attenuate the drawbacks of iterative procedures, Tjøstheim and Auestad (1994a) proposed a direct method based on an average regression surface idea, referred to as projection method in Tjøstheim and Auestad (1994a) for time series data. As pointed out by Cai and Fan (2000), a direct method has some advantages, such as it does not rely on iterations, it can make computation fast, and more importantly, it allows an asymptotic analysis. Finally, the projection method was extended to nonlinear ARX models by Masry and Tjøstheim (1997) using the kernel method and Cai and Masry (2000) coupled with the local polynomial approach. It should be remarked that the projection method, under the name of marginal integration, was proposed independently by Newey (1994) and Linton and Nielsen (1995) for the iid samples, and since then, some important progresses have been made by some authors. For example, by combining the marginal integration with one-step backfitting, Linton (1997, 2000) presented an efficient estimator, Mammen et al. (1999) established rigorously the asymptotic theory of the backfitting, Cai (2007) considered estimating each component using the weighted projection method coupled with the local linear fitting in an efficient way, and Sperlich et al. (2002) extended the efficient method to models with simple interactions.

The projection method has some disadvantages although it has the aforementioned merits. The projection method may not be efficient if covariates (endogenous or exogenous variables) are strongly correlated, which is particularly relevant for autoregressive models. The intuitive interpretation is that additive components are not orthogonal. To overcome this shortcoming, two efficient estimation methods have been proposed in the literature. The first one is called weight function procedure, proposed by Fan et al. (1998) for the iid samples and extended to time series situations by Cai and Fan (2000). With an appropriate choice of the weight function, additive components can be efficiently estimated in the sense that an additive component can be estimated with the same asymptotic bias and variance as if the rest of components were known. The second one is to combine the marginal integration with one-step backfitting, introduced by Linton (1997, 2000) for the iid samples and extended by Sperlich et al. (2002) to additive models with single interactions, but this method has not been advocated for time series situations. However, there has not been any attempt to discuss the bandwidth selection for the projection method and its variations in the litera-

ture due to their complexity. In practice, one bandwidth is usually used for all components although Cai (2007) argued that different bandwidths might be used theoretically to deal with the situation that additive components posses the different smoothness. Therefore, the projection method may not be optimal in practice in the sense that one bandwidth is used.

To estimate unknown additive components in (2.50) efficiently, following the spirit of the marginal integration with one-step backfitting proposed by Linton (1997) for the iid samples, a two-stage method is used, due to Linton (2000), coupled with the local linear (polynomial) method, which has some attractive properties, such as mathematical efficiency, bias reduction and adaptation of edge effect; see, e.g., Fan and Gijbels (1996). The basic idea of the two-stage approach is described as follows. In the first stage, one obtains the initial estimated values for all components. More precisely, the idea for estimating any additive component is first to estimate directly high-dimensional regression surface by the local linear method and then to average the regression surface over the rest of variables to stabilize variance. Such an initial estimate, in general, is under-smoothed so that the bias should be asymptotically negligible. In the second stage, the local linear (polynomial) technique is used again to estimate any additive component by using the initial estimated values of the rest of components. In such a way, it is shown that the estimate in the second stage is not only efficient in the sense of being equivalent to a procedure based on knowing other components, but also making the bandwidth selection much easier. Note that this technique is not novel to this chapter since the two-stage method is first used by Linton (1997, 2000) for the iid samples, but many details and insights are.

2.6.2 Backfitting Algorithm

The building block of the generalized additive model algorithm is the scatterplot smoother. We will first describe scatterplot smoothing in a simple setting, and then indicate how it is used in generalized additive modeling. Here y is a response or outcome variable, and x is a prognostic factor. We wish to fit a smooth curve $f(x)$ that summarizes the dependence of y on x. If we were to find the curve that simply minimizes $\sum_{i=1}^{n}\left[y_i - f\left(x_i\right)\right]^2$, the result would be an interpolating curve that would not be smooth at all. The cubic spline smoother imposes smoothness on $f(x)$. We seek the function $f(x)$ that minimizes

$$\sum_{i=1}^{n}\left[y_i - f\left(x_i\right)\right]^2 + \lambda \int \left[f''(x)\right]^2 dx. \tag{2.52}$$

Notice that $\int [f''(x)]^2 dx$ measures the "wiggliness" of the function $f(x)$: linear $f(x)$ s have $\int [f''(x)]^2 dx = 0$, while non-linear fs produce values bigger than zero. λ is a non-negative smoothing parameter that must be chosen by the data analyst. It governs the tradeoff between the goodness of fit to the data and (as measured by and wiggleness of the function. Larger values of λ force $f(x)$ to be smoother.

For any value of λ, the solution to (2.52) is a cubic spline, i.e., a piecewise cubic polynomial with pieces joined at the unique observed values of x in the dataset. Fast and stable numerical procedures are available for computation of the fitted curve. What value of did we use in practice? In fact it is not a convenient to express the desired smoothness of $f(x)$ in terms of λ, as the meaning of λ depends on the units of the prognostic factor x. Instead, it is possible to define an *effective number of parameters* or *degrees of freedom* of a cubic spline smoother, and then use a numerical search to determine the value of λ to yield this number. In practice, if we chose the effective number of parameters to be 5 , roughly speaking, this means that the complexity of the curve is about the same as a polynomial regression of degrees 4. However, the cubic spline smoother "spreads out" its parameters in a more even manner, and hence is much more flexible than a polynomial regression. Note that the degrees of freedom of a smoother need not be an integer.

The above discussion tells how to fit a curve to a single prognostic factor. With multiple prognostic factors, if x_{ij} denotes the value of the j th prognostic factor for the i th observation, we fit the additive model

$$y_i = \sum_{j=1}^{d} f_j\left(x_{ij}\right) + \varepsilon_i.$$

A criterion like (2.52) can be specified for this problem, and a simple iterative procedure exists for estimating the f_j s. We apply a cubic spline smoother to the outcome $y_i - \sum_{j \neq k}^{d} \widehat{f}_j\left(x_{ij}\right)$ as a function of x_{ik}, for each prognostic factor in turn. The process is continues until the estimates $\widehat{f}_j(x)$ stabilize. These procedure is known as "back-fitting" and the resulting fit is analogous to a multiple regression for linear models.

To fit an additive model or a partially additive model in **R**, the function is **gam()** in the package **gam**. For details, please look at the help command **help(gam)** after loading the package **gam** "**library(gam)**" . Note that the function **gam()** allows to fit a semiparametric

additive model as

$$Y = \boldsymbol{\beta}^\top \mathbf{X} + \sum_{j=1}^{p} g_j\left(Z_j\right) + \varepsilon,$$

which can be done by specifying some components without smooth.

2.6.3 Projection Method

This section is devoted to a brief review of the projection method and discusses its merits and disadvantages. It is assumed that all additive components have continuous second partial derivatives, so that $m(\mathbf{u}, \mathbf{v})$ can be locally approximated by a linear term in a neighborhood of $(\mathbf{x}, \mathbf{y})$, namely, $m(\mathbf{u}, \mathbf{v}) \approx \beta_0 + \boldsymbol{\beta}_1^\top(\mathbf{u} - \mathbf{x}) + \boldsymbol{\beta}_2^\top(\mathbf{v} - \mathbf{y})$ with $\left\{\boldsymbol{\beta}_j\right\}$ depending on $\mathbf{x}$ and $\mathbf{y}$, where $\boldsymbol{\beta}_1^\top$ denotes the transpose of $\boldsymbol{\beta}_1$.

Let $K(\cdot)$ and $L(\cdot)$ be symmetric kernel functions in $\mathbb{R}^p$ and $\mathbb{R}^q$, respectively, and $h_{11} = h_{11}(n) > 0$ and $h_{12} = h_{12}(n) > 0$ be bandwidths in the step of estimating the regression surface. Here, to handle various degrees of smoothness, Cai (2007) proposed using h_{11} and h_{12} differently although the implementation may not be easy in practice. The reader is referred to the paper by Cai (2007) for details. Given observations $\{\mathbf{X}_t, Y_t, \mathbf{Z}_t\}_{t=1}^n$, let $\widehat{\boldsymbol{\beta}}_j$ be the minimizer of the following locally weighted least squares

$$\sum_{t=1}^{n} \left\{Y_t - \beta_0 - \boldsymbol{\beta}_1^\top\left(\mathbf{X}_t - \mathbf{x}\right) - \boldsymbol{\beta}_2^\top\left(\mathbf{Z}_t - \mathbf{z}\right)\right\}^2 K_{h_{11}}\left(\mathbf{X}_t - \mathbf{x}\right) L_{h_{12}}\left(\mathbf{Z}_t - \mathbf{z}\right),$$

where $K_h(\cdot) = K(\cdot/h)/h^p$ and $L_h(\cdot) = L(\cdot/h)/h^q$. Then, the local linear estimator of the regression surface $m(\mathbf{x}, \mathbf{z})$ is $\widehat{m}(\mathbf{x}, \mathbf{z}) = \widehat{\beta}_0$. By computing the sample average of $\widehat{m}(\cdot, \cdot)$ based on (2.51), the projection estimators of $g_1(\cdot)$ and $g_2(\cdot)$ are defined as, respectively,

$$\widehat{g}_1(\mathbf{x}) = \frac{1}{n}\sum_{t=1}^{n} \widehat{m}\left(\mathbf{x}, \mathbf{Z}_t\right) - \widehat{\mu}, \quad \text{and} \quad \widehat{g}_2(\mathbf{z}) = \frac{1}{n}\sum_{t=1}^{n} \widehat{m}\left(\mathbf{X}_t, \mathbf{z}\right) - \widehat{\mu},$$

where $\widehat{\mu} = n^{-1}\sum_{t=1}^{n} Y_t$. Under some regularity conditions, by using the same arguments as those employed in the proof of Theorem 3 in Cai and Masry (2000), it can be shown (although not easy and tedious) that the asymptotic bias and asymptotic variance of $\widehat{g}_1(\mathbf{x})$ are, respectively, $h_{11}^2 \mathrm{tr}\left\{\mu_2(K) g_1''(\mathbf{x})\right\}/2$ and $v_1(\mathbf{x}) = \nu_0(K) A(\mathbf{x})$, where

$$A(\mathbf{x}) = \int p_2^2(\mathbf{y}) \sigma^2(\mathbf{x}, \mathbf{z}) p^{-1}(\mathbf{x}, \mathbf{z}) d\mathbf{z} \quad \text{and} \quad \sigma^2(\mathbf{x}, \mathbf{z}) = \mathrm{Var}\left(Z_t \mid \mathbf{X}_t = \mathbf{x}, \mathbf{Z}_t = \mathbf{z}\right).$$

Here, $p(\mathbf{x}, \mathbf{z})$ stands for the joint density of $\mathbf{X}_t$, $\mathbf{Z}_t$, $p_1(\mathbf{x})$ denotes the marginal density of $\mathbf{X}_t$, and $p_2(\mathbf{z})$ is the marginal density of $\mathbf{Z}_t$.

The foregoing method has some advantages, such as it is easy to understand, it can make computation fast, and it allows an asymptotic analysis. However, it can be quite inefficient in an asymptotic sense. To demonstrate this idea, let us consider the ideal situation that $g_2(\cdot)$ and μ are known. In such a case, one can estimate $g_1(\cdot)$ by directly regressing the partial error $\widetilde{Y}_t = Y_t - \mu - g_2(\mathbf{Z}_t)$ on $\mathbf{X}_t$ and such an ideal estimator is optimal in an asymptotic minimax sense; see, e.g., Fan and Gijbels (1996). The asymptotic bias for the ideal estimator is $h_{11}^2 \mathrm{tr}\{\mu_2(K) g_1''(\mathbf{x})\}/2$ and the asymptotic variance is

$$v_0(\mathbf{x}) = \nu_0(K) B(\mathbf{x}) \quad \text{with} \quad B(\mathbf{x}) = p_1^{-1}(\mathbf{x}) \mathbb{E}\left\{\sigma^2(\mathbf{X}_t, \mathbf{Z}_t) \mid \mathbf{X}_t = \mathbf{x}\right\};$$

see, e.g., Masry and Fan (1997). It is clear that $v_1(\mathbf{x}) = v_0(\mathbf{x})$ if $\mathbf{X}_t$ and $\mathbf{Z}_t$ are independent. If $\mathbf{X}_t$ and $\mathbf{Z}_t$ are correlated and when $\sigma^2(\mathbf{x}, \mathbf{z})$ is a constant, it follows from the CauchySchwarz inequality that

$$B(\mathbf{x}) = \frac{\sigma^2}{p_1(\mathbf{x})} \int p^{1/2}(\mathbf{z} \mid \mathbf{x}) \frac{p_2(\mathbf{z})}{p^{1/2}(\mathbf{z} \mid \mathbf{x})} d\mathbf{z} \leq \frac{\sigma^2}{p_1(\mathbf{x})} \int \frac{p_2^2(\mathbf{z})}{p(\mathbf{z} \mid \mathbf{x})} d\mathbf{z} = A(\mathbf{x})$$

which implies that the ideal estimator has always smaller asymptotic variance than the projection method although both have the same bias. This suggests that the projection method could lead to an inefficient estimation of $g_1(\cdot)$ and $g_2(\cdot)$ when $\mathbf{X}_t$ and $\mathbf{Z}_t$ are serially correlated, which is particularly relevant for autoregressive models. To alleviate this shortcoming, the following two-stage approach is proposed, described next.

2.6.4 Two-Stage Procedure

The two-stage method due to Linton (1997, 2000) is introduced. The basic idea is to get an initial estimate for $\widehat{g}_2(\cdot)$ using a small bandwidth h_{12}. The initial estimate can be obtained by the projection method and h_{12} can be chosen so small that the bias of estimating $\widehat{g}_2(\cdot)$ can be asymptotically negligible. Then, using the partial residuals $Y_t^* = Y_t - \widehat{\mu} - \widehat{g}_2(\mathbf{Z}_t)$, we apply the local linear regression technique to the pseudo regression model

$$Y_t^* = g_1(\mathbf{X}_t) + \varepsilon_t^*$$

to estimate $g_1(\cdot)$. This leads naturally to the weighted least-squares problem

$$\sum_{t=1}^{n} \left\{Y_t^* - \beta_1 - \boldsymbol{\beta}_2^\top (\mathbf{X}_t - \mathbf{x})\right\}^2 J_{h_2}(\mathbf{X}_t - \mathbf{x}), \tag{2.53}$$

where $J(\cdot)$ is the kernel function in $\mathbb{R}^p$ and $h_2 = h_2(n) > 0$ is the bandwidth in the second stage. The advantage of this is two-fold: the bandwidth h_2 can now be selected purposely for estimating $g_1(\cdot)$ only and any bandwidth selection technique for nonparametric regression can be applied here. Maximizing (2.53) with respect to β_1 and $\boldsymbol{\beta}_2$ gives the two-stage estimate of $g_1(\mathbf{x})$, denoted by $\widetilde{g}_1(\mathbf{x}) = \widehat{\beta}_1$, where $\widehat{\beta}_1$ and $\widehat{\boldsymbol{\beta}}_2$ are the minimizer of (2.53).

It is shown in Theorem 2.3 below that under some regularity conditions, the asymptotic bias and variance of the two-stage estimate $\widetilde{g}_1(\mathbf{x})$ are the same as those for the ideal estimator, provided that the initial bandwidth h_{12} satisfies $h_{12} = o(h_2)$. To establish the asymptotic normality of the two-stage estimator, it is assumed that the initial estimator satisfies a linear approximation; namely,

$$\widehat{g}_2(\mathbf{Z}_t) - g_2(\mathbf{Z}_t) \approx \frac{1}{n}\sum_{i=1}^{n} L_{h_{12}}(\mathbf{Z}_i - \mathbf{Z}_t)\,\Gamma(\mathbf{X}_i, \mathbf{Z}_t)\,\delta_i + \frac{1}{2}h_{12}^2 \mathrm{tr}\left\{\mu_2(L) g_2''(\mathbf{Z}_t)\right\}, \qquad (2.54)$$

where $\delta_t = Z_t - m(\mathbf{X}_t, \mathbf{Z}_t)$ and $\Gamma(\mathbf{x}, \mathbf{y}) = p_1(\mathbf{x})/p(\mathbf{x}, \mathbf{z})$. Note that under some regularity conditions, by following the same arguments as in Masry (1996a), one might show (although the proof is not easy, quite lengthy, and tedious) that (2.54) holds. Note that this assumption is also imposed in Linton (2000) for the iid samples to simplify the proof of the asymptotic results of the two-stage estimator. Now, the asymptotic normality for the two-stage estimator is stated here and its proof can be found in Cai (2002b).

Theorem 2.3: *Under* (2.54) *and Assumptions A1-A9 stated in Cai (2002b), if bandwidths* h_{12} *and* h_2 *are chosen such that* $h_{12} \to 0, nh_{12}^q \to \infty, h_2 \to 0$, *and* $nh_2^p \to \infty$ *as* $n \to \infty$, *then*

$$\sqrt{nh_2^p}\left[\widetilde{g}_1(\mathbf{x}) - g_1(\mathbf{x}) - bias(\mathbf{x}) + o_p\left(h_{12}^2 + h_2^2\right)\right] \quad \to \quad N\left\{0, v_0(\mathbf{x})\right\},$$

where the asymptotic bias is

$$bias(\mathbf{x}) = \frac{h_2^2}{2} tr\left\{\mu_2(J) g_1''(\mathbf{x})\right\} - \frac{h_{12}^2}{2} tr\left\{\mu_2(L)\mathbb{E}\left(g_2''(\mathbf{Z}_t) \mid \mathbf{X}_t = \mathbf{x}\right)\right\}$$

and the asymptotic variance is $v_0(\mathbf{x}) = \nu_0(J) B(\mathbf{x})$.

We remark that by Theorem 2.3, the asymptotic variance of the two-stage estimator is independent of the initial bandwidths. Thus, the initial bandwidths should be chosen as small as possible. This is another benefit of using the two-stage procedure: the bandwidth selection problem becomes relatively easy. In particular, when $h_{12} = o(h_2)$, the bias from the initial estimation can be asymptotically negligible. For the ideal situation that $g_2(\cdot)$

is known, Masry and Fan (1997) show that under some regularity conditions, the optimal estimate of $g_1(\mathbf{x})$, denoted by $\widehat{g}_1^*(\mathbf{x})$, by using (2.54) in which the partial residual Y_t^* is replaced by the partial error $\widetilde{Y}_t = Y_t - \mu - g_2(\mathbf{Z}_t)$, is asymptotically normally distributed,

$$\sqrt{nh_2^p}\left[\widehat{g}_1^*(\mathbf{x}) - g_1(\mathbf{x}) - \frac{h_2^2}{2}\mathrm{tr}\left\{\mu_2(J)g_1''(\mathbf{x})\right\} + o_p\left(h_2^2\right)\right] \quad \rightarrow \quad N\left\{0, v_0(\mathbf{x})\right\}.$$

This, in conjunction with Theorem 2.3, shows that the two-stage estimator and the ideal estimator share the same asymptotic bias and variance if $h_{12} = o(h_2)$.

Finally, note that the reader is referred to the paper by Cai (2002b) for the detailed Monte Carlo simulation results and applications. Also, one can see the paper by Mammen et al. (1999) for some more approaches on additive modeling.

2.6.5 Analysis of the Boston House Price Data via Additive Model

There have been several papers devoted to the analysis of this dataset using some nonparametric methods. For example, Breiman and Friedman (1985), Pace (1993), Chaudhuri et al. (1997), and Opsomer and Ruppert (1998) used four covariates: X_6, X_{10}, X_{11} and X_{13} or their transformations (including the transformation on Y) to fit the data through a mean additive regression model such as

$$\log(Y) = \mu + g_1(X_6) + g_2(X_{10}) + g_3(X_{11}) + g_4(X_{13}) + \varepsilon, \tag{2.55}$$

where the additive components $\{g_j(\cdot)\}$ are unspecified smooth functions. Pace (1993) and Chaudhuri et al. (1997) also considered the nonparametric estimation of the first derivative of each additive component which measures how much the response changes as one covariate is perturbed while the other covariates are held fixed; see Chaudhuri et al. (1997). Let us use model (2.55) to fit the Boston house price data. The results are summarized in Figure 2.4 (the **R** code can be found in Section 2.9.2). Also, we fit a semi-parametric additive model (partially linear model as in (2.57)) as

$$\log(Y) = \mu + g_1(X_6) + \beta_2 X_{10} + \beta_3 X_{11} + \beta_4 X_{13} + \varepsilon. \tag{2.56}$$

The results are summarized in Figure 2.5 (the **R** code can be found in Section 2.9.2).

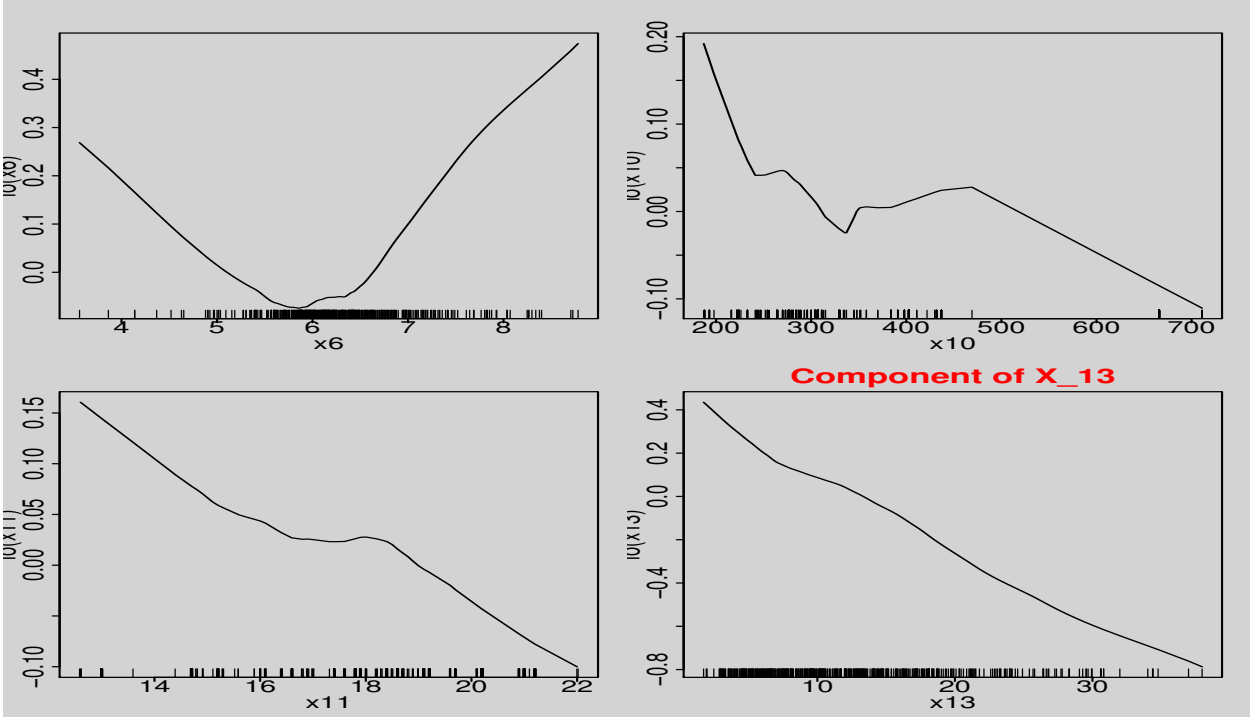

Figure 2.4: The results from model (2.55).

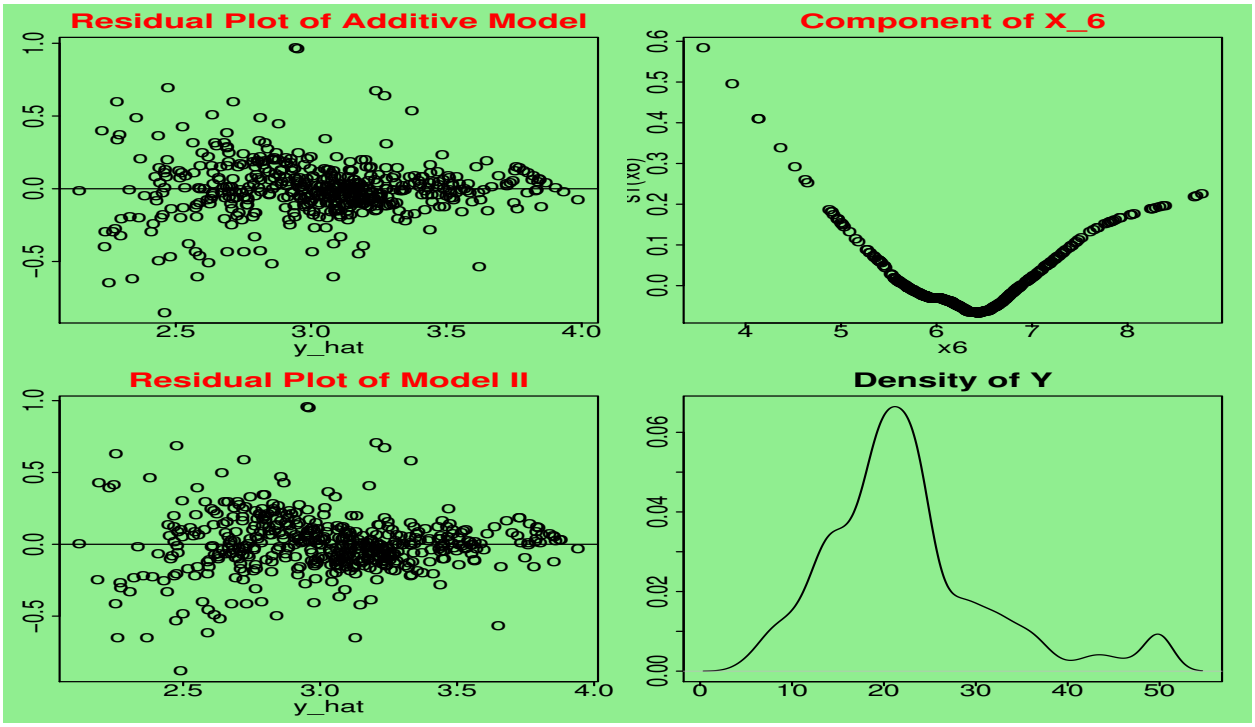

Figure 2.5: (a) Residual plot for model (2.55). (b) Plot of $g_1(x_6)$ versus x_6. (c) Residual plot for model (2.56). (d) Density estimate of Y.

2.7 Semiparametric Models

2.7.1 Partially Linear Models

Initiated by applications in economics as in Shiller (1984) for estimating the U-shaped cost curve from the electric utility industry and Engle et al. (1986) for a nonlinear relationship between electricity sales and temperature, we consider the following partially linear model

$$\mathbb{E}\left(Y_t \mid \mathbf{X}_t, \mathbf{Z}_t\right) = \boldsymbol{\beta}^\top \mathbf{X}_t + g(\mathbf{Z}_t), \tag{2.57}$$

where $g(\cdot)$ is an unknown link function. From (2.57), one can obtain

$$\mathbb{E}(Y_t \mid \mathbf{X}_t, \mathbf{Z}_t) - \mathbb{E}(Y_t|Z_t) = \boldsymbol{\beta}^\top [\mathbf{X}_t - \mathbb{E}(\mathbf{X}_t|\mathbf{Z}_t)],$$

which leads to the estimate of β by the method of moment estimation approach, proposed by Robinson (1988), given by

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{t=1}^{n} (\mathbf{X}_t - \widehat{m}_x(\mathbf{Z}_t))(\mathbf{X}_t - \widehat{m}_x(\mathbf{Z}_t))^\top\right)^{-1} \sum_{t=1}^{n} (\mathbf{X}_t - \widehat{m}_x(\mathbf{Z}_t))(Y_t - \widehat{m}_y(\mathbf{Z}_t)),$$

where $m_x(\mathbf{Z}_t) = \mathbb{E}(\mathbf{X}_t|Z_t)$ and $m_y(\mathbf{Z}_t) = \mathbb{E}(Y_t|\mathbf{Z}_t)$. Here, $\widehat{m}_x(\mathbf{z})$ is a nonparametric estimate of $m_x(\mathbf{z})$ and $\widehat{m}_y(\mathbf{z})$ is a nonparametric estimate of $m_y(\mathbf{z})$. Now, having estimated parameter vector β, it is possible to fit the nonlinear relation between $\mathbf{Z}_t$ and Y_t by simply estimating equation (2.58) presented below nonparametrically

$$\widehat{Y}_t = Y_t - \widehat{\boldsymbol{\beta}}^\top \mathbf{X}_t = g(\mathbf{Z}_t) + u_t. \tag{2.58}$$

Then, the local linear (or polynomial) technique can be applied here to estimate $g(\cdot)$. Robinson (1988) investigated the asymptotic properties of $\widehat{\boldsymbol{\beta}}$ (consistency and asymptotic normality). Especially, Robinson (1988) showed that $\widehat{\boldsymbol{\beta}}$ is efficient in the sense that its asymptotic variance of $\widehat{\boldsymbol{\beta}}$ is the smallest even $g(\cdot)$ is unknown. However, the method by Robinson (1988) needs to estimate both $m_x(\mathbf{Z}_t) = \mathbb{E}(\mathbf{X}_t|Z_t)$ and $m_y(\mathbf{Z}_t) = \mathbb{E}(Y_t|\mathbf{Z}_t)$ nonparametrically.

To avoid the above disadvantage, one can the so-called profile least squares method proposed by Speckman (1988), described as follows, which becomes profile likelihood estimation when the error is normally distributed; see, for example, Speckman (1988). Suppose that we have a random sample of size n, $\{(Y_t, \mathbf{X}_t, \mathbf{Z}_t)\}_{t=1}^n$ from model (2.58). For given β, (2.58) becomes

$$Y_t(\boldsymbol{\beta}) = Y_t - \boldsymbol{\beta}^\top \mathbf{X}_t = g(\mathbf{Z}_t) + u_t, \tag{2.59}$$

where $Y_t(\boldsymbol{\beta}) = Y_t - \boldsymbol{\beta}^\top \mathbf{X}_t$. This transforms the partially linear model in (2.58) into the conventional nonparametric regression model as in Section 2.3. Then, the local linear regression technique outlined in Section 2.3 is applied to estimating the function $g(\cdot)$ in (2.58). Thus, one can express $\widehat{g}(\mathbf{Z}_t)$ as

$$\begin{pmatrix} \widehat{g}(\mathbf{Z}_1) \\ \vdots \\ \widehat{g}(\mathbf{Z}_n) \end{pmatrix} = \mathbf{S}\,\mathbf{Y}(\boldsymbol{\beta}) = \mathbf{S}\,(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}),$$

where $\mathbf{S}$ is the smooth matrix, and substitute $\widehat{\mathbf{S}}\mathbf{Y}$ into (2.59) to obtain

$$(\mathbf{I} - \mathbf{S})\mathbf{Y} = (\mathbf{I} - \mathbf{S})\mathbf{X}\boldsymbol{\beta} + \mathbf{u}.$$

Applying the least squares method to obtain the ordinary least squares estimate of $\boldsymbol{\beta}$, denoted by $\widehat{\boldsymbol{\beta}}_{\text{pls}}$, given by

$$\widehat{\boldsymbol{\beta}}_{\text{pls}} = [\mathbf{X}^\top(\mathbf{I} - \mathbf{S})^\top(\mathbf{I} - \mathbf{S})\mathbf{X}]^{-1}\mathbf{X}^\top(\mathbf{I} - \mathbf{S})^\top(\mathbf{I} - \mathbf{S})\mathbf{Y}.$$

Moreover, $(\widehat{g}(\mathbf{Z}_1, \cdots, \widehat{g}(\mathbf{Z}_n))^\top = \mathbf{S}\,(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{\text{pls}})$. Speckman (1988) showed that $\widehat{\boldsymbol{\beta}}_{\text{pls}}$ is semi-parametrically efficient[5]. Clearly, the profile least squares method is better than that by Robinson (1988), since it needs only to estimate a nonparametric regression function of $Y_t(\boldsymbol{\beta})$ versus $\mathbf{Z}_t$.

Finally, the model in (2.58) was extended by Fan and Huang (2005) to the following varying-coefficient partially linear model

$$Y_t = \boldsymbol{\beta}^\top \mathbf{X}_t + a(\mathbf{Z}_t)^\top \mathbf{W}_t + v_t, \tag{2.60}$$

Then, Fan and Huang (2005) applied the profile least squares approach to estimate both $a(\cdot)$ and β and derived the asymptotic normality of the profile least-squares estimator. Also, they showed that the profile least squares estimator of β is semi-parametrically efficient and the model in (2.60) was employed by Fan and Huang (2005) to study the Boston housing data; see, for instance, Fan and Huang (2005) for details. Importantly, Fan and Huang (2005) also considered the test issues as in Section 2.5.4 and applied the GLR test proposed in Section 2.5.4 to test whether some of β in (2.60) are zero with deriving the asymptotic theory as in Fan et al. (2001), similar to the result in (2.24).

[5]Semiparametric Efficiency: The asymptotic variance of any semiparametric estimator is no smaller than the supremum of the Cramér-Rao lower bounds for all parametric sub-model.

2.7.2 Single Index Models

An object of interest such as the conditional density $f(y|\mathbf{x})$ or conditional distribution $F(y|\mathbf{x})$ or conditional mean $\mathbb{E}(Y_t \mid \mathbf{X}_t = \mathbf{x})$ is a single index model when it only depends on the vector $\mathbf{x}$ through a single linear combination of $\mathbf{x}$ as $\boldsymbol{\beta}^\top \mathbf{x}$. Indeed, most parametric models are single index, including, for example, normal regression, logit, probit, Tobit, and Poisson regression. In a semiparametric single index model, the object of interest depends on $\mathbf{x}$ through the function $\mathbb{E}(Y_t \mid \mathbf{X}_t = \mathbf{x}) = g\left(\boldsymbol{\beta}^\top \mathbf{x}\right)$, where $\beta \in \mathbb{R}^p$ and $g : \mathbb{R} \to \mathbb{R}$ are unknown. If $g(\cdot)$ is known, it is called a link function in generalized line model (GLM) literature. In single index models, there is only one nonparametric dimension. These methods fall in the class of dimension reduction techniques. The semiparametric single index regression model, proposed by Ichimuara (1993), is given by

$$\mathbb{E}(Y_t \mid \mathbf{X}_t = \mathbf{x}) = g\left(\boldsymbol{\beta}^\top \mathbf{x}\right), \tag{2.61}$$

where $g(\cdot)$ is an unknown function. The semiparametric single index binary choice model is

$$\mathbb{P}(Y_t = 1 \mid \mathbf{X}_t = \mathbf{x}) = \mathbb{E}(Y_t \mid \mathbf{X}_t = \mathbf{x}) = g\left(\boldsymbol{\beta}^\top \mathbf{x}\right), \tag{2.62}$$

where $g(\cdot)$ is an unknown distribution function.[6] We use $g(\cdot)$ (rather than, say, $F(\cdot)$) to emphasize the connection with the regression model.

In both contexts, the function $g(\cdot)$ includes any location and level shift, so the vector $\mathbf{X}_t$ cannot include an intercept. The level of $\boldsymbol{\beta}$ is not identified, so some normalization criterion for is needed. It is typically easier to impose this on β than on $g(\cdot)$. One approach is to set $\boldsymbol{\beta}^\top \boldsymbol{\beta} = 1$. A second approach is to set one component of β to equal one. (This second approach requires that this variable correctly has a non-zero coefficient.) The vector $\mathbf{X}_t$ must be dimension 2 or larger. If X_t is one-dimensional, then $\boldsymbol{\beta}$ is simply normalized to one, and the model is the one-dimensional nonparametric regression $\mathbb{E}(Y_t \mid \mathbf{X}_t = \mathbf{x}) = g(\mathbf{x})$ with no semiparametric component. Identification of β and $g(\cdot)$ also requires that $\mathbf{X}_t$ contains at least one continuously distributed variable, and that this variable has a non-zero coefficient. If not, $\boldsymbol{\beta}^\top \mathbf{x}$ only takes a discrete set of values, and it would be impossible to identify a continuous function $g(\cdot)$ on this discrete support. Therefore, in what follows, it is assumed that the model in (2.61) is identified without a further mention.

[6] If $g(\cdot)$ is a known link function, the model in (2.62) is either logist or probit model, popularly in many applications in economics and finance' see, the books by Hastie and Tibshirani (1990) and Cameron and Trivedi (2005) for details.

A. Ichumura's Estimator

The semiparametric single index regression model is

$$Y_t = g\left(\boldsymbol{\beta}^\top \mathbf{X}_t\right) + \epsilon_t$$

with $\mathbb{E}\left(\epsilon_t \mid \mathbf{X}_t\right) = 0$, and it generalizes the linear regression model (which sets $g(\cdot)$ to be linear), and is a restriction of the nonparametric regression model. The gain over full nonparametric setting is that there is only one nonparametric dimension, so the curse of dimensionality is avoided. Suppose that $g(\cdot)$ were known. Then, you could estimate β by a (nonlinear) least-squares (LS) with the LS criterion

$$S_n(\boldsymbol{\beta}, g) = \sum_{t=1}^{n} \left(Y_t - g\left(\boldsymbol{\beta}^\top \mathbf{X}_t\right)\right)^2. \tag{2.63}$$

You could think about to replacing $g(\cdot)$ by $\widehat{g}(\cdot)$. But, since $g(\cdot)$ is unknown conditional mean of Y_t given $\boldsymbol{\beta}^\top \mathbf{X}_t = z, g(\cdot)$ depends on $\boldsymbol{\beta}$, so that Ichimuara (1993) suggested a two-step (2LS) estimation procedure as follows. First, a leave-one out Nadaraya-Waston estimation of $g(\cdot)$ is used

$$\widehat{g}_{-t}\left(\boldsymbol{\beta}^\top \mathbf{X}_t\right) = \sum_{s \neq t}^{n} Y_s K\left(\boldsymbol{\beta}^\top \left(\mathbf{X}_s - \mathbf{X}_t\right)/h\right) / \sum_{s \neq t}^{n} K\left(\boldsymbol{\beta}^\top \left(\mathbf{X}_s - \mathbf{X}_t\right)/h\right), \tag{2.64}$$

and then, Ichimuara (1993) suggested replacing $g(\cdot)$ in $S_n(\boldsymbol{\beta}, g)$ by $\widehat{g}_{-t}\left(\boldsymbol{\beta}^\top \mathbf{X}_t\right)$,

$$l_n(\boldsymbol{\beta}) = \sum_{t=1}^{n} \left(Y_t - \widehat{g}_{-t}\left(\boldsymbol{\beta}^\top \mathbf{X}_t\right)\right)^2 I_t(b),$$

where $I_t(b)$ is a trimming function to make the computation easy. The Ichimura's estimator is $\widehat{\boldsymbol{\beta}}_{\text{2LS}} = \operatorname{argmin} l_n(\boldsymbol{\beta})$. However, Ichimuara (1993) did not discuss on how to choose $I_t(b)$ in $l_n(\boldsymbol{\beta})$. As pointed out by Härdle et al. (1993), the criterion in the Ichimura's estimator is somewhat similar to cross-validation so that the Ichimura's estimator may not be efficient (optimal). To obtain the efficient estimation of $\boldsymbol{\beta}$, Härdle et al. (1993) suggested picking $\boldsymbol{\beta}$ and the bandwidth h simultaneously by minimizing $l_n(\boldsymbol{\beta})$, denoted by $\widehat{\boldsymbol{\beta}}_{\text{hhi}}$.

Finally, for the asymptotic theory for $\widehat{\boldsymbol{\beta}}_{\text{2LS}}$ or $\widehat{\boldsymbol{\beta}}_{\text{hhi}}$, the reader is referred to the papers by Ichimuara (1993) and Härdle et al. (1993), respectively.

B. Klein and Spady's Binary Choice Estimator

Klein and Spady (1993) proposed an estimator of the semiparametric single index binary choice model which has strong similarities with Ichimura's estimator. The model is given in (2.62) and can be re-expressed as follows

$$Y_t = I\left(\boldsymbol{\beta}^\top \mathbf{X}_t > e_t\right),$$

where e_t is an error, which is a special case of (2.61). If e_t is independent of $\mathbf{X}_t$, and has distribution function $g(\cdot)$, then, the data satisfy the single-index regression as

$$\mathbb{E}\left(Y_t \mid \mathbf{X}_t\right) = g\left(\boldsymbol{\beta}^\top \mathbf{X}_t\right),$$

and it follows that Ichimura's estimator can be directly applied to this model. However, different from the Ichimura's estimator, Klein and Spady (1993) suggested a semiparametric likelihood approach. Given $g(\cdot)$, the log likelihood is

$$l_n(\boldsymbol{\beta}, g) = \sum_{t=1}^{n} \left[Y_t \ln\left(g\left(\boldsymbol{\beta}^\top \mathbf{X}_t\right)\right) + (1 - Y_t) \ln\left(1 - g\left(\boldsymbol{\beta}^\top \mathbf{X}_t\right)\right]\right.$$

Since $g(\cdot)$ is unknown, making this substitution of $g(\cdot)$ by $\widehat{g}_{-t}\left(\boldsymbol{\beta}^\top X_t\right)$ in (2.64), and adding trimming function, this leads to the feasible likelihood criterion

$$l_n(\boldsymbol{\beta}) = \sum_{t=1}^{n} \left[Y_t \ln\left(\widehat{g}_{-t}\left(\boldsymbol{\beta}^\top \mathbf{X}_t\right)\right) + (1 - Y_t) \ln\left(1 - \widehat{g}_{-t}\left(\boldsymbol{\beta}^\top \mathbf{X}_t\right)\right] I_t(b),\right.$$

where, as suggested by Klein and Spady (1993), the trimming indicator can be taken to be

$$I_t(b) = I\left(\widehat{f}_{\tilde{\boldsymbol{\beta}}^\top \mathbf{X}_t}\left(\tilde{\boldsymbol{\beta}}^\top \mathbf{X}_t\right) > b\right),$$

where $\tilde{\beta}$ is a preliminary estimator of $\boldsymbol{\beta}$ and $\widehat{f}(\cdot)$ is an estimation of the density function of $\tilde{\boldsymbol{\beta}}^\top \mathbf{X}_t$. It can be seen from Klein and Spady (1993) that trimming does not seem to matter in their simulations. Finally, the Klein and Spady estimator for $\boldsymbol{\beta}$ is the value $\widehat{\boldsymbol{\beta}}$ which maximizes $l_n(\boldsymbol{\beta})$ and in many respects, the Ichimura and Klein-Spady estimators are quite similar.

C. Average Derivative Estimator

Let $m\left(\mathbf{X}_t\right) = \mathbb{E}\left(Y_t \mid \mathbf{X}_t\right)$ and $m'(\mathbf{x})$ denote the first order derive of $m(\mathbf{x})$. Define the weighted derivative as follows

$$\delta = \mathbb{E}\left[m'(\mathbf{X}_t) w\left(\mathbf{X}_t\right)\right],$$

where $w(x)$ is a weight function, which is particularly convenient to set $w(x) = f_{\mathbf{X}}(\mathbf{x})$ with $f_{\mathbf{X}}(\mathbf{x})$ being the marginal density of $\mathbf{X}_t$, suggested by Powell et al. (1989). A simple algebra leads to the following expression

$$\boldsymbol{\delta} = \int m'(\mathbf{x}) f_{\mathbf{X}}^2(\mathbf{x}) = -2\mathbb{E}\left[Y_t f'_{\mathbf{X}}(\mathbf{X}_t)\right],$$

where $f'_{\mathbf{X}}(\mathbf{x})$ is the first order derivative of $f_{\mathbf{X}}(\mathbf{x})$, which, clearly, leads to a consistent estimate of δ, given by

$$\widehat{\boldsymbol{\delta}} = -\frac{2}{n-1}\sum_{t=1}^{n} Y_t \widehat{f}'_{\mathbf{X},-t}(\mathbf{X}_t)$$

where $\widehat{f}'_{\mathbf{X},-t}(\mathbf{x})$ is the first derivative of the leave-one-out density estimator of $f_{\mathbf{X}}(\mathbf{x})$. One can see that this is a convenient estimator. There is no denominator messing with uniform convergence. There is only a density estimator and no conditional mean is needed. Powell et al. (1989) showed that $\widehat{\boldsymbol{\delta}}$ is $n^{1/2}$-consistent and asymptotically normal, with a convenient covariance matrix.

Now, for the single-index model, it is easy to see $m'(\mathbf{x}) = \boldsymbol{\beta} g'\left(\boldsymbol{\beta}^\top \mathbf{x}\right)$ so that

$$\boldsymbol{\delta} = \boldsymbol{\beta} \int g'\left(\boldsymbol{\beta}^\top \mathbf{x}\right) f_{\mathbf{X}}^2(\mathbf{x}) = c\,\boldsymbol{\beta},$$

where $c = \mathbb{E}\left[g'\left(\boldsymbol{\beta}^\top \mathbf{X}_t\right) f_{\mathbf{X}}\left(\mathbf{X}_t\right)\right]$, from which, one can obtain the average derivative estimator for β by

$$\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\delta}}/\widehat{c},$$

where $\widehat{c}$ is a consistent estimate of c. However, the problem goes back to estimating a density function with a possible a high dimension.

D. MAVE Estimator

Due to the fact that the single index model shares a close connection with the central mean subspace in the sufficient dimension reduction, Xia et al. (2002) proposed the (conditional) minimum average variance estimation (MAVE) method for the dimension reduction problem and later, Xia (2006) showed that this method can be applied to the single index model. Therefore, the MAVE method proposed in Xia (2006) is employed in our setting to estimate β and also the penalized MAVE considered in Wang et al. (2013) is utilized for selecting X, described as follows.

Note that under the least squares loss,

$$\boldsymbol{\beta} = \arg \min_{\tilde{\boldsymbol{\beta}} \in \mathbb{R}^k} \mathbb{E}\left[Y - \mathbb{E}(Y|\tilde{\boldsymbol{\beta}}^\top \mathbf{X})\right]^2. \tag{2.65}$$

In our setting, the index is estimated by the observed data for the control units, $\{(Y_t,\, \mathbf{X}_t)\}_{t=1}^n$. Motivated by the local linear smoothing technique, the sample analogue of (2.65) can be written as

$$\begin{aligned}\boldsymbol{\beta} &= \arg \min_{\tilde{\boldsymbol{\beta}} \in \mathbb{R}^k : \tilde{\boldsymbol{\beta}}^\top \tilde{\boldsymbol{\beta}} = 1} \sum_{j=1}^n \left\{ \min_{a_j, b_j} \sum_{i=1}^n \left[Y_i - a_j - b_j \tilde{\boldsymbol{\beta}}^\top (\mathbf{X}_i - \mathbf{X}_j)\right]^2 w_{ij} \right\} \\ &= \arg \min_{\substack{\tilde{\boldsymbol{\beta}} \in \mathbb{R}^k : \tilde{\boldsymbol{\beta}}^\top \tilde{\boldsymbol{\beta}} = 1 \\ a_j, b_j}} \sum_{j=1}^n \sum_{i=1}^n \left[Y_i - a_j - b_j \tilde{\boldsymbol{\beta}}^\top (\mathbf{X}_i - \mathbf{X}_j)\right]^2 w_{ij}, \end{aligned} \tag{2.66}$$

where $a_j = g(\boldsymbol{\beta}^\top \mathbf{X}_j)$, $b_j = \partial g(u)/\partial u|_{u=\boldsymbol{\beta}^\top \mathbf{X}_j}$, and $w_{ij} = K_h(\boldsymbol{\beta}^\top(\mathbf{X}_i - \mathbf{X}_j))$. Xia (2006) proposed the following algorithm for estimating $\boldsymbol{\beta}$:

Step 1. Set an initial value $\boldsymbol{\beta}^{(0)}$.

Step 2. For $\ell \geq 1$, calculate

$$\begin{pmatrix} \widehat{a}_j^{\boldsymbol{\beta}^{(\ell-1)}} \\ \widehat{d}_j^{\boldsymbol{\beta}^{(\ell-1)}} h \end{pmatrix} = \left\{ \sum_{j=1}^n K_h\left(\boldsymbol{\beta}^{(\ell-1)^\top} \mathbf{X}_{i,j}\right) Z_{ij}^{(k-1)} Z_{ij}^{(\ell-1)^\top} \right\}^{-1} \sum_{j=1}^n K_h\left(\boldsymbol{\beta}^{(\ell-1)^\top} \mathbf{X}_{i,j}\right) \mathbf{Z}_{ij}^{(\ell-1)} Y_j,$$

where $\mathbf{Z}_{ij}^{(\ell-1)} = \left(1, \boldsymbol{\beta}^{(\ell-1)^\top} X_{ij}/h\right)^\top$ with $\mathbf{X}_{i,j} = \mathbf{X}_i - \mathbf{X}_j$, and also, obtain

$$\widehat{f}_{\boldsymbol{\beta}^{(\ell-1)}}(\boldsymbol{\beta}^{(\ell-1)^\top} \mathbf{X}_j) = \frac{1}{n} \sum_{i=1}^n K_h(\boldsymbol{\beta}^{(\ell-1)^\top} \mathbf{X}_{i,j}), \quad \text{and} \quad \widehat{\rho}_j^{\boldsymbol{\beta}^{(\ell-1)^\top}} = \rho_n\left(\widehat{f}_{\boldsymbol{\beta}^{(\ell-1)}}(\boldsymbol{\beta}^{(\ell-1)^\top} \mathbf{X}_j)\right),$$

where $\rho_n(\cdot)$ is a trimming function for the boundary points. Following the suggestion from Xia (2006), $\rho_n(v)$ is chosen as a bounded function with bounded derivative on $\mathbb{R}$ such that $\rho_n(v) = I(v > 2c_0 n^{-\varepsilon})$.

Step 3. Calculate

$$\begin{aligned}\boldsymbol{\beta}^{(\ell)} = &\left\{ \sum_{i=1}^n \sum_{j=1}^n K_h\left(\boldsymbol{\beta}^{(\ell-1)^\top} \mathbf{X}_{i,j}\right) \widehat{\rho}_j^{\boldsymbol{\beta}^{(k-1)}} \left(\widehat{d}_j^{\boldsymbol{\beta}^{(\ell-1)}}\right)^2 \mathbf{X}_{i,j} \mathbf{X}_{i,j}^\top / \widehat{f}_{\boldsymbol{\beta}^{(\ell-1)}}\left(\boldsymbol{\beta}^{(\ell-1)^\top} \mathbf{X}_j\right) \right\}^{-1} \\ &\times \sum_{i=1}^n \sum_{j=1}^n K_h\left(\boldsymbol{\beta}^{(\ell-1)^\top} \mathbf{X}_{i,j}\right) \widehat{\rho}_j^{\boldsymbol{\beta}^{(\ell-1)}} \widehat{d}_j^{\boldsymbol{\beta}^{(\ell-1)}} \mathbf{X}_{i,j} \left(Y_i - \widehat{a}_j^{\boldsymbol{\beta}^{(\ell-1)}}\right) / \widehat{f}_{\boldsymbol{\beta}^{(\ell-1)}}\left(\boldsymbol{\beta}^{(\ell-1)^\top} \mathbf{X}_j\right). \end{aligned}$$

Step 4. Set $\boldsymbol{\beta}^{(\ell)} = \text{sign}(\boldsymbol{\beta}^{(\ell)}) \boldsymbol{\beta}^{(\ell)} / \|\boldsymbol{\beta}^{(\ell)}\|$. Then, repeat Steps 2 and 3 until convergence reaches.

Denote the ultimate estimator for $\boldsymbol{\beta}$ as $\widehat{\boldsymbol{\beta}}_{\text{MAVE}}$. Theoretically, Xia (2006) derived the asymptotic normality for $\widehat{\boldsymbol{\beta}}_{\text{MAVE}}$ and showed that the asymptotic covariance matrix of $\widehat{\boldsymbol{\beta}}_{\text{MAVE}}$ can achieve the information lower bound in the semiparametric sense. From Xia (2006), one can see that under some regularity conditions, $\widehat{\boldsymbol{\beta}}_{\text{MAVE}}$ has the following asymptotic behavior

$$\sqrt{n}\,\left[\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right]=\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\phi(\mathbf{X}_j,Y_j)+o_p(1)\ \xrightarrow{d}\ N(0,\Sigma_{\boldsymbol{\beta}}), \tag{2.67}$$

where $\phi(\mathbf{X}_j,Y_j)=W_g^{+}g'(\boldsymbol{\beta}^{\top}\mathbf{X}_j)v_{\beta}(\mathbf{X}_j)\,e_j$, $W_g=\mathbb{E}\left\{g'(\boldsymbol{\beta}^{\top}\mathbf{X})^2v_{\beta}(\mathbf{X})v_{\beta}^{\top}(\mathbf{X})\right\}$, $W_{m_0}^{+}$ is the Moore-Penrose inverse of W_g, and $v_{\beta}(\mathbf{x})=\mathbb{E}(\mathbf{X}|\boldsymbol{\beta}^{\top}\mathbf{X}=\boldsymbol{\beta}^{\top}\mathbf{x})-\mathbf{x}$, while the asymptotic variance is given by

$$\Sigma_{\boldsymbol{\beta}}=\left[E\{g'(\boldsymbol{\beta}^{\top}\mathbf{X})^2W(\mathbf{X})\}\right]^{+}\mathbb{E}\left\{g'(\boldsymbol{\beta}^{\top}\mathbf{X})^2W_0(\mathbf{X})\epsilon^2\right\}\left[E\{g'(\boldsymbol{\beta}^{\top}\mathbf{X})^2W(\mathbf{X})\}\right]^{+},$$

where $W(\mathbf{x})=\mathbb{E}(\mathbf{X}\mathbf{X}^{\top}|\boldsymbol{\beta}^{\top}\mathbf{X}=\boldsymbol{\beta}^{\top}\mathbf{x})-\mathbb{E}(\mathbf{X}|\boldsymbol{\beta}^{\top}\mathbf{X}=\boldsymbol{\beta}^{\top}\mathbf{x})E^{\top}(\mathbf{X}|\boldsymbol{\beta}^{\top}\mathbf{X}=\boldsymbol{\beta}^{\top}\mathbf{x})$ and $W_0(\mathbf{x})=v_{\beta}(\mathbf{x})v_{\beta}^{\top}(\mathbf{x})$.

From the above discussions, we know that the MAVE estimate of $\boldsymbol{\beta}$ is obtained by solving the minimization problem (2.66). Generally, to select the relevant variables, we can add a penalty term to the least-squares-form loss function in (2.66):

$$\sum_{j=1}^{n}\sum_{i=1}^{n}[Y_i-a_j-b_j\tilde{\boldsymbol{\beta}}^{\top}(\mathbf{X}_i-\mathbf{X}_j)]^2w_{ij}+n\sum_{l=1}^{k}p_{\lambda_n}(|\tilde{\beta}_l|),$$

where $p_{\lambda}(\cdot)$ denotes a penalty function and λ_n denotes the penalty parameter. Different choices of $p_{\lambda}(\cdot)$ can lead to different variable selection methods.

The simplest choice is to set $p_{\lambda_n}(|\tilde{\beta}_l|)=\lambda_n|\tilde{\beta}_l|$, which corresponds to the well-known LASSO. Indeed, Wang and Yin (2008) adopted this L_1 norm penalty and proposed the sparse MAVE method and Zeng et al. (2012) further explored the idea of combining MAVE and LASSO, and proposed the sim-LASSO method. The sim-LASSO method not only penalizes the L_1 norm of the index parameter $\boldsymbol{\beta}$, but also penalizes the terms $\{b_j\}_{j=1}^{n}$ in (2.66). Since $b_j=\partial g(u)/\partial u|_{u=\boldsymbol{\beta}^{\top}\mathbf{X}_j}$, adding this penalty contributes to excluding the data points with less information on estimating $\boldsymbol{\beta}$, which stabilizes and improves the estimation of $\boldsymbol{\beta}$. Finally, Wang et al. (2013) proposed the penalized MAVE method, combining the bridge regression with MAVE. In the case of the single-index-model, the penalized MAVE estimator has the oracle property.

It is widely accepted that a good penalty function should lead to an unbiased, sparse and continuous estimator. However, the LASSO estimator is biased for large parameters. Alternatively, Fan and Li (2001) proposed the smoothly clipped absolute deviation (SCAD) penalty. The SCAD penalty is defined via its first derivative as

$$p'_\lambda(\beta_l) = \lambda\{I(\beta_l \leq \lambda) + \frac{(a\lambda - \beta_l)_+}{(a-1)\lambda} I(\tilde{\beta}_l > \lambda)\}.$$

Due to the oracle property of the SCAD penalty justified by Fan and Li (2001), Peng and Huang (2011) explored the idea of introducing the SCAD penalty into the single index model. Given that the dimension of β is a fixed constant, the SCAD estimator has the oracle property. Hence, we can also combine the SCAD penalty with MAVE, and modify the objective function in (2.66) as:

$$\boldsymbol{\beta} = \arg \min_{\substack{\tilde{\beta}\in\mathbb{R}^k:\tilde{\beta}^\top\tilde{\beta}=1 \\ a_j,b_j}} \left\{ \sum_{j=1}^n \sum_{i=1}^n \left[Y_i - a_j - b_j \tilde{\boldsymbol{\beta}}^\top (\mathbf{X}_i - \mathbf{X}_j) \right]^2 w_{ij} + n \sum_{l=1}^k p_{\lambda_n}^{\text{SCAD}}(|\tilde{\beta}_l|) \right\} \quad (2.68)$$

Similarly, the optimization problem in (2.68) can be solved alternatively and iteratively and the SCAD-MAVE algorithm can be summarized as follows:

Step 1. Given data $\{(Y_t, \mathbf{X}_t)\}_{t=1}^n$, calculate the initial estimator $\widehat{\boldsymbol{\beta}}_{(0)}$ by the MAVE method. Set $\ell = 1$.

Step 2. Given $\widehat{\boldsymbol{\beta}}_{(\ell-1)}$, calculate the refined weights as

$$w_{ij}^{(\ell-1)} = K_{h_1}\left[\widehat{\boldsymbol{\beta}}_{(\ell-1)}^\top (\mathbf{X}_i - \mathbf{X}_j)\right] \Big/ \sum_{l=1}^n K_{h_1}\left[\widehat{\boldsymbol{\beta}}_{(\ell-1)}^\top (\mathbf{X}_l - \mathbf{X}_j)\right].$$

Then, solve the inner optimization problem for $j = 1, \cdots, n$:

$$\min_{a_j,b_j} \sum_{i=1}^n \left[Y_i - a_j - b_j \widehat{\boldsymbol{\beta}}_{(\ell-1)}^\top (\mathbf{X}_i - \mathbf{X}_j) \right]^2 w_{ij}^{(\ell-1)}$$

Clearly, this problem is analogous to the weighted least squares problem. We can easily derive the analytical solutions and denote them as $\widehat{a}_j^{(\ell-1)}$ and $\widehat{b}_j^{(\ell-1)}$.

Step 3. Given $\widehat{a}_j^{(\ell-1)}$ and $\widehat{b}_j^{(\ell-1)}$, we solve the outer optimization problem:

$$\min_{\tilde{\boldsymbol{\beta}}\in\mathbb{R}^k:\tilde{\boldsymbol{\beta}}^\top\tilde{\boldsymbol{\beta}}=1} \left\{ \sum_{j=1}^n \sum_{i=1}^n \left[Y_i - \widehat{a}_j^{(\ell-1)} - \widehat{b}_j^{(\ell-1)} \tilde{\boldsymbol{\beta}}^\top (\mathbf{X}_i - \mathbf{X}_j) \right]^2 w_{ij}^{(\ell-1)} + n \sum_{l=1}^k p_{\lambda_n}^{\text{SCAD}}(|\tilde{\boldsymbol{\beta}}_l|) \right\}$$

Obviously, regardless of the constraint $\tilde{\boldsymbol{\beta}}^\top \tilde{\boldsymbol{\beta}} = 1$, we can rewrite the first part in least squares form, then, we can use the *ncvreg* package in **R** to optimize it and obtain the estimator $\widehat{\boldsymbol{\beta}}_{(\ell)}$.

Let $\widehat{\boldsymbol{\beta}}_{(\ell)} = \text{sign}(\widehat{\boldsymbol{\beta}}_{(\ell)})\widehat{\boldsymbol{\beta}}_{(\ell)}/\|\widehat{\boldsymbol{\beta}}_{(\ell)}\|$.

Step 4. Check whether $\|\widehat{\boldsymbol{\beta}}_{(\ell)} - \widehat{\boldsymbol{\beta}}_{(\ell-1)}\|^2 < c$, where c is an arbitrarily small positive constant, if not, set $\ell = \ell + 1$ and go to Step 2. Denote the final estimator as $\widehat{\boldsymbol{\beta}}_{\text{scad-MAVE}}$.

Based on the above discussions, we can use the SCAD-MAVE method to select relevant variables or control units at first, then, construct the index $\widehat{Z}_j = \widehat{\boldsymbol{\beta}}_{\text{scad-MAVE}}^{\top}\mathbf{X}_j$ for $j = 1, \cdots, n$. Finally, from Peng and Huang (2011), one can show that under some regularity conditions, $\widehat{\boldsymbol{\beta}}_{\text{scad-MAVE}}$ satisfies (2.67).

2.7.3 Functional Coefficient Index Models

Fan et al. (2003) proposed the functional coefficient index model for the iid data without the asymptotic theory, which also was studied by Cai et al. (2015) for time series data with the asymptotic theory. The the functional coefficient index model is given by

$$\mathbb{E}\left(Y_t \mid \mathbf{X}_t, \mathbf{Z}_t\right) = \boldsymbol{\beta}(\gamma^{\top}\mathbf{Z}_t)^{\top}\mathbf{X}_t, \tag{2.69}$$

where $\boldsymbol{\beta}(\cdot)$ is an unknown coefficient function and γ is an unknown parameter, which was applied to a financial application by Cai et al. (2015); see (2.48) for details. The reader is referred to the papers by Fan et al. (2003), Cai et al. (2015), and Cai et al. (2015) for the detailed theory and applications as well as the related model selection issues.

To estimate both functional $\boldsymbol{\beta}(\cdot)$ and parameter γ in (2.69), Cai et al. (2015) proposed a two-stage method to estimate $\boldsymbol{\beta}(\cdot)$ nonparametrically and estimate γ parametrically. Further, Cai et al. (2015) used the LASSO type method to select $\mathbf{Z}$ and $\mathbf{X}$ under the uniform framework. The read is referred to the paper by Cai et al. (2015) for the detailed steps for estimation and model selection procedures.

2.7.4 Distributional Index Models

In this section, we study the following distributional index model

$$F_y(y|\mathbf{X}_t) = F_y(y|\boldsymbol{\beta}^{\top}\mathbf{X}_t), \tag{2.70}$$

where $F_g(y|\mathbf{X}_t)$ is the conditional distribution of Y_t given $\mathbf{X}_t$, which is called the distributional index model. Indeed, the model in (2.70) was successfully applied to a financial application by Aït-Sahalia and Brant (2001). Also, the model in (2.70) implies that the conditional

quantile of Y_t given $\mathbf{X}_t$ is for any $0 < \tau < 1$,

$$q_\tau(\mathbf{X}_t) = F_y^{-1}(\tau|\mathbf{X}_t) = q_\tau(\boldsymbol{\beta}^\top \mathbf{X}_t), \tag{2.71}$$

which is a special case of single-index quantile regression model; see Wu et al. (2010) for details. Also, $F_y(y|\mathbf{x})$ and $q_\tau(\mathbf{x})$ have the following relationship

$$F_y(y|\mathbf{x}) = \int_0^1 I(q_\tau(\mathbf{x}) \le y) d\tau,$$

which can be used to approximate $F_y(y|\mathbf{x})$ if $q_\tau(\mathbf{x})$ is estimable, as follows,

$$F_y(y|\mathbf{x}) \approx \eta + \int_\eta^{1-\eta} I(q_\tau(\mathbf{x}) \le y) d\tau \approx \eta + \sum_{j=2}^{S} (\tau_j - \tau_{j-1}) I(q_{\tau_j}(\mathbf{x}) \le y), \tag{2.72}$$

where the trimming by η (a very small number) avoids estimation of tail quantiles and an equally spaced mesh $\eta = \tau_1 < \cdots < \tau_S = 1 - \eta$ is used for a very large S so that $\tau_j - \tau_{j-1}$ is very small. One of the popular applications is to assume that $q_\tau(\mathbf{x})$ is a linear model as $q_\tau(\mathbf{x}) = \beta_\tau^\top \mathbf{x}$, which provides an easy estimation of β_τ. The idea in (2.72) was used by Chernozhukov et al. (2013) and Cai et al. (2026, 2025a) in estimating counterfactual distributions, which can be used for estimating quantile treatment effects (QTE); see, for instance, the papers by Cai et al. (2026, 2025a) for detailed the methodology and its theory with applications.

Alternatively, one can use the matching method proposed by Hall and Yao (2005). The reader is referred to the paper by Hall and Yao (2005) for details.

2.8 Time-Varying Coefficient Models

Now, consider the popular model, considered by Cai (2007) for time series, as follows

$$Y_t = \boldsymbol{\beta}(t)^\top \mathbf{X}_t + \epsilon_t, \tag{2.73}$$

where $\boldsymbol{\beta}(t) \in \mathbb{R}^p$ is a vector of unknown functionals, which can be regarded as a special case of model (2.18) with $U_t = t$, and a special case of the model proposed in Cai et al. (2000) for the iid data. Therefore, the reader is referred to the paper by Cai et al. (2000) and Cai (2007) for details.

Clearly, if model (2.73) includes only the deterministic time trend function (there is no X_t), the time series $\{Y_t\}$ is not stationary and the model becomes the so-called fixed design

regression model as studied in Gasser and Müller (979) in biomedical applications. The deterministic time trend function $\boldsymbol{\beta}(t)$ might be an important ingredient in modeling economic and financial data and it might not be a polynomial as pointed out by Phillips (2001), who gave a review on some present developments and future challenges about trending time series models. Also, if all functions including $\boldsymbol{\beta}(t)$ and $\sigma^2(t)$ do not depend on t and the time series $\{(X_t, \epsilon_t)\}$ is stationary, then the time series $\{Y_t\}$ generated by the above model is stationary.

The object of (2.73) is to estimate $\boldsymbol{\beta}(t)$ although $\sigma^2(t) = \text{Var}(\epsilon_t)$ is also of interest in many applications. To estimate $\beta(t)$ nonparametrically, Cai (2007) demonstrated that making $\boldsymbol{\beta}(t)$ depend on the sample size n

$$\boldsymbol{\beta}(t) = \boldsymbol{\beta}_0(s_t),$$

where $s_t = t/n$, is necessary to provide the asymptotic justification for any nonparametric smoothing estimators, where $\boldsymbol{\beta}_0(\dot{})$ is an unknown smooth function. The intuitive explanation to this "intensity" assumption is that it is an increasingly intense sampling of data points that derive the consistent estimation; see the paper by Cai (2007) for more discussions on this point. This intensity assumption is applied to the disturbance process $\{\epsilon_t\}_{t=1}^n$ as well, such as $\sigma^2(t) = \sigma_0^2(s_t)$.

To study model (2.73) with heteroscedasticity theoretically and empirically, one can develop a local linear estimation procedure as discussed in Section 2.3 to estimate the coefficient functions. It is shown in Cai (2007) that the estimators based on both local linear fitting and the Nadaraya–Watson methods share the exact same asymptotic behavior at the interior points but they have different asymptotic behaviors at the boundaries. Also, it is shown in Cai (2007) that the consistency of the proposed estimators can be obtained without specifying the error distribution and the asymptotic variance of the proposed estimator depends not only on the variance of the error but also the autocorrelations. This property is shared by parametric estimators. Further, an easily implemented bandwidth selector and a consistent estimate of the standard errors were provided by Cai (2007) for a practical purpose. Finally, an important econometric question in fitting model (2.73), when $\boldsymbol{\beta}_0(\cdot)$ is a constant, arises whether all coefficient functions are actually varying (namely, if a linear model is adequate or the time series $\{Y_t\}$ is stationary), or more generally if a parametric model fits the given data, or there is no time trend at all, or there are some exogenous variables statistically

insignificant. This amounts to testing whether some or all coefficient functions are constant or zero or in a certain parametric form. This is an important issue in testing misspecification and stationarity. An existing testing procedure based on the comparison of the residual sum of squares under the null and alternative models, the so-called generalized likelihood ratio test described in Section 2.5.4 or the generalized F-test as in Cai and Tiwari (2000), is adapted to the current problem and the null distribution of the proposed test statistic is estimated by using a simple nonparametric version of a Bootstrap sampling scheme (i.e. wild Bootstrap). For all details, the reader is referred to the aforementioned papers.

Remark 2.6: *In the model given by* (2.73), $\mathbf{X}_t$ *does not include any lags of* Y_t. *If* $\mathbf{X}_t$ *contains some lags of* Y_t, *the model in* (2.73) *becomes the well known locally stationary model if* $\boldsymbol{\beta}(t)$ *satisfies some conditions either in time domain or frequency domain, introduced by Dahlhaus (1996). There is a vast literature on the theory and applications of locally stationary processes. See, for example, the paper by Vogt (2012) on the kernel-based method to estimate the time-varying regression function and the asymptotic theory for the estimates as well as showing that the main conditions of the theory are satisfied for a large class of nonlinear autoregressive processes with a time-varying regression function.*

2.9 Computer Codes

2.9.1 Codes for Example 2.1

```
# 12-03-2025
graphics.off() # clean the previous graphs on the screen

###############
# Example 2.1
##############

#############################################################################
z1=read.table(file="/NP_lecture_note/data/ex4-1.txt")
# dada: weekly 3-month Treasury bill from 1970 to 1997
x=z1[,4]/100
```

```
n=length(x)
y=diff(x)           # Delta x_t=x_t-x_{t-1}
x=x[1:(n-1)]
n=n-1
x_star=(x-mean(x))/sqrt(var(x))
z=seq(min(x),max(x),length=50)

win.graph()
#postscript(file="/NP_lecture_note/figs/fig-4.1.eps",
# horizontal=F,width=6,height=6)
par(mfrow=c(2,2),mex=0.4,bg="light blue")
scatter.smooth(x,y,span=1/10,ylab="",xlab="x(t-1)",evaluation=60)
title(main="(a) y(t) vs x(t)",col.main="red")
scatter.smooth(x,abs(y),span=1/10,ylab="",xlab="x(t-1)",evaluation=60)
title(main="(b) |y(t)| vs x(t)",col.main="red")
scatter.smooth(x,y^2,span=1/10,ylab="",xlab="x(t-1)",evaluation=60)
title(main="(c) y(t)^2 vs x(t)",col.main="red")
#dev.off()
#########################################################################

#########################
# Nonparametric Fitting #
#########################

#########################################################
# Define the Epanechnikov kernel function
kernel<-function(x){0.75*(1-x^2)*(abs(x)<=1)}

#############################################################
# Define the kernel density estimator
kernden=function(x,z,h,ker){
 # parameters: x=variable; h=bandwidth; z=grid point; ker=kernel
```

```
 nz<-length(z)
 nx<-length(x)
 x0=rep(1,nx*nz)
 dim(x0)=c(nx,nz)
 x1=t(x0)
 x0=x*x0
 x1=z*x1
 x0=x0-t(x1)
 if(ker==1){x1=kernel(x0/h)}        # Epanechnikov kernel
 if(ker==0){x1=dnorm(x0/h)}         # normal kernel
 f1=apply(x1,2,mean)/h
 return(f1)
}
#################################################################
# Define the local constant estimator
local.constant=function(y,x,z,h,ker){
 # parameters: x=variable; h=bandwidth; z=grid point; ker=kernel
 nz<-length(z)
 nx<-length(x)
 x0=rep(1,nx*nz)
 dim(x0)=c(nx,nz)
 x1=t(x0)
 x0=x*x0
 x1=z*x1
 x0=x0-t(x1)
 if(ker==1){x1=kernel(x0/h)}        # Epanechnikov kernel
 if(ker==0){x1=dnorm(x0/h)}         # normal kernel
 x2=y*x1
 f1=apply(x1,2,mean)
 f2=apply(x2,2,mean)
 f3=f2/f1
 return(f3)
```

```
}

#######################################################################
# Define the local linear estimator
local.linear<-function(y,x,z,h){
 # parameters: y=response, x=design matrix; h=bandwidth; z=grid point
 nz<-length(z)
 ny<-length(y)
 beta<-rep(0,nz*2)
 dim(beta)<-c(nz,2)
 for(k in 1:nz){
  x0=x-z[k]
  w0<-kernel(x0/h)
  beta[k,]<-glm(y~x0,weight=w0)$coeff
 }
 return(beta)
}
#######################################################################

h=0.02

# Local constant estimate

mu_hat=local.constant(y,x,z,h,1)
sigma_hat=local.constant(abs(y),x,z,h,1)
sigma2_hat=local.constant(y^2,x,z,h,1)

#win.graph()
postscript(file="/NP_lecture_note/figs/fig-2.1.eps",
horizontal=F,width=6,height=6)
par(mfrow=c(2,2),mex=0.4,bg="light yellow")
scatter.smooth(x,y,span=1/10,ylab="",xlab="x(t-1)")
```

```
points(z,mu_hat,type="l",lty=1,lwd=3,col=2)
title(main="(a) y(t) vs x(t)",col.main="red")
legend(0.04,0.0175,"Local Constant Estimate")
scatter.smooth(x,abs(y),span=1/10,ylab="",xlab="x(t-1)")
points(z,sigma_hat,type="l",lty=1,lwd=3,col=2)
title(main="(b) |y(t)| vs x(t)",col.main="red")
scatter.smooth(x,y^2,span=1/10,ylab="",xlab="x(t-1)")
title(main="(c) y(t)^2 vs x(t)",col.main="red")
points(z,sigma2_hat,type="l",lty=1,lwd=3,col=2)
dev.off()

# Local Linear Estimate

fit2=local.linear(y,x,z,h)
mu_hat=fit2[,1]
fit2=local.linear(abs(y),x,z,h)
sigma_hat=fit2[,1]
fit2=local.linear(y^2,x,z,h)
sigma2_hat=fit2[,1]

#win.graph()
postscript(file="/NP_lecture_note/figs/fig-2.2.eps",
horizontal=F,width=6,height=6)
par(mfrow=c(2,2),mex=0.4,bg="light green")
scatter.smooth(x,y,span=1/10,ylab="",xlab="x(t-1)")
points(z,mu_hat,type="l",lty=1,lwd=3,col=2)
title(main="(a) y(t) vs x(t)",col.main="red")
legend(0.04,0.0175,"Local Linear Estimate")
scatter.smooth(x,abs(y),span=1/10,ylab="",xlab="x(t-1)")
points(z,sigma_hat,type="l",lty=1,lwd=3,col=2)
title(main="(b) |y(t)| vs x(t)",col.main="red")
scatter.smooth(x,y^2,span=1/10,ylab="",xlab="x(t-1)")
```

```
title(main="(c) y(t)^2 vs x(t)",col.main="red")
points(z,sigma2_hat,type="l",lty=1,lwd=3,col=2)
dev.off()
#########################################################################
```

2.9.2 Codes for Additive Modeling Analysis of Boston Data

The following is the R code for making figures in Figures 2.4 and 2.5, respectively.

```
data=read.table("file="/NP_lecture_note/data/ex4-2.txt")
y=data[,14]
x1=data[,1]
x6=data[,6]
x10=data[,10]
x11=data[,11]
x13=data[,13]
y_log=log(y)
library(gam)
fit_gam=gam(y_log~lo(x6)+lo(x10)+lo(x11)+lo(x13))
resid=fit_gam$residuals
y_hat=fit_gam$fitted

postscript(file="/NP_lecture_note/figs/fig-2.3.eps",
horizontal=F,width=6,height=6,bg="light grey")
par(mfrow=c(2,2),mex=0.4)
plot(fit_gam)
title(main="Component of X_13",col.main="red",cex=0.6)
dev.off()
fit_gam1=gam(y_log~lo(x6)+x10+x11+x13)
s1=fit_gam1$smooth[,1]           # obtain the smoothed component
resid1=fit_gam1$residuals
y_hat1=fit_gam1$fitted
print(summary(fit_gam1))
```

```
postscript(file="/NP_lecture_note/figs/fig-2.4.eps",
horizontal=F,width=6,height=6,bg="light green")
par(mfrow=c(2,2),mex=0.4)
plot(y_hat,resid,type="p",pch="o",ylab="",xlab="y_hat")
title(main="Residual Plot of Additive Model",col.main="red",cex=0.6)
abline(0,0)
plot(x6,s1,type="p",pch="o",ylab="s1(x6)",xlab="x6")
title(main="Component of X_6",col.main="red",cex=0.6)
plot(y_hat1,resid1,type="p",pch="o",ylab="",xlab="y_hat")
title(main="Residual Plot of Model II",col.main="red",cex=0.5)
abline(0,0)
plot(density(y),ylab="",xlab="",main="Density of Y")
dev.off()
```

Chapter 3

Quantile Regression Models

3.1 Introduction

Over the last three decades, quantile regression, also called conditional quantile or regression quantile, introduced by Koenker and Bassett (1978), has been used widely in various disciplines, such as finance, economics, medicine, and biology. It is well-known that when the distribution of data is typically skewed or data contains some outliers, the median regression, a special case of quantile regression, is more explicable and robust than the mean regression. Also, regression quantiles can be used to test heteroscedasticity formally or graphically; see, for example, Koenker and Bassett (1982), Efron (1991), Koenker and Zhao (1996), Koenker and Xiao (2002), and references therein. Although some individual quantiles, such as the conditional median, are sometimes of interest in practice, more often one wishes to obtain a collection of conditional quantiles which can characterize the entire conditional distribution. More importantly, another application of conditional quantiles is the construction of prediction intervals for the next value given a small section of the recent past values in a stationary time series; see, for example, Granger et al. (1989), Koenker (1994), Zhou and Portnoy (1996), Koenker and Zhao (1996), Taylor and Bunn (1999), and references therein. Also, Granger et al. (1989), Koenker and Zhao (1996), and Taylor and Bunn (1999) considered an interval forecasting for parametric autoregressive conditional heteroscedastic (ARCH) type models. For more details about the historical and recent developments of quantile regression with applications for time series data, particularly in finance, see, for example, the papers and books by Morgan (1995), Duffie and Pan (1997), Koenker (2000), Koenker and Hallock (2000), Tsay (2002), Khindanova and Rachev (2000), and Bao et al. (2006), and references therein.

Interestingly, the quantile regression technique has been successfully applied to politics. For example, in the 1992 presidential selection, the Democrats used the yearly Current Population Survey data to show that between 1980 and 1992 there was an increase in the number of people in the high-salary category as well as an increase in the number of people in the low-salary category. This phenomena could be illustrated by using the quantile regression method as follows: computing 90% and 10% quantile regression functions of salary as a function of time. An increasing 90% quantile regression function and a decreasing 10% quantile regression function corresponded to the Democrats' claim that "the rich got richer and the poor got poorer" during the Republican administrations; see Figure 6.4 in Fan and Gijbels (1996).

As elaborated in Serfling (1980), in some instances, the quantile approach is feasible and useful when other approaches are out of the question. For example, to estimate the parameter of a Cauchy distribution with density $f(x) = 1/\pi[1 + (x - \mu)^2]$, $-\infty x < \infty$, the sample mean $\bar{X} = \sum_{t=1}^{n} X_t/n$ based on the sample $\{X_t\}_{t=1}^{n}$ is not a consistent estimate of the location parameter μ. However, the sample median $\widehat{\xi}_{1/2}$ follows asymptotically $N(0, \pi^2/4n)$ and thus quite well behaved (robust against outliers). When both the quantile and moment approaches are feasible, it is of interest to examine their relative efficiency. For example, consider a symmetric distribution $F(\cdot)$ having finite variance σ^2 and mean μ. So, the median $\xi_{1/2} = \mu$. In this case, both $\bar{X}$ and $\widehat{\xi}_{1/2}$ are competitors for estimation of μ. Assume that $F(\cdot)$ has a density $f(\cdot)$ positive and continuous at μ. Then, according to the theorems in Serfling (1980), it is easy to establish the following,

$$\sqrt{n}\left[\bar{X} - \mu\right] \xrightarrow{d} N\left(0, \sigma^2\right) \quad \text{and} \quad \sqrt{n}\left[\widehat{\xi}_{1/2} - \mu\right] \xrightarrow{d} N\left(0, 4/f^2(\mu)\right).$$

If we consider asymptotic relative efficiency (ARE) in the sense of the criterion of small asymptotic variance or mean squares errors (MSE) in the normal approximation, then, the asymptotic relative efficiency of $\widehat{\xi}_{1/2}$, relative to $\bar{X}$ is

$$\text{ARE}\left(\widehat{\xi}_{1/2}, \bar{X}\right) = 4\sigma^2 f(\mu).$$

For a normal distribution $F(\cdot)$, this relative efficiency is $2/\pi$, indicating the degree of superiority of $\bar{X}$ over $\widehat{\xi}_{1/2}$. For some other distributions $F(\cdot)$, say, the Laplace distribution, $\widehat{\xi}_{1/2}$ is superior to $\bar{X}$.

More importantly, by following the regulations of the Bank for International Settlements, many of financial institutions have begun to use a uniform measure of risk to measure the

market risks called Value-at-Risk, which can be defined as the maximum potential loss of a specific portfolio for a given horizon in finance. In essence, the interest is to compute an estimate of the lower tail quantile (with a small probability) of future portfolio returns, conditional on current information. Therefore, the VaR can be regarded as a special application of the quantile regression. There is a vast amount of literature in this area; see, to name just a few, Morgan (1995), Duffie and Pan (1997), Engle and Manganelli (2004), Jorion (2001), Tsay (2002), Khindanova and Rachev (2000), and Bao et al. (2006), and references therein.

In this chapter, we assume that $\{\mathbf{X}_t, Y_t\}_{t=-\infty}^{\infty}$ is a stationary sequence. Denote $F(y \mid \mathbf{x})$ the conditional distribution of Y given $\mathbf{X} = \mathbf{x}$, where $\mathbf{X}_t = (X_{t1}, \cdots, X_{tp})^\top$ is the associated covariate vector in $\mathbb{R}^p$ with $p \geq 1$, which might be a function of exogenous (covariate) variables or some lagged (endogenous) variables or time t. The regression (conditional) quantile function $q_\tau(\mathbf{x})$ is defined as, for any $0 < \tau < 1$,

$$q_\tau(\mathbf{x}) = \inf\left\{y \in \mathbb{R}^1 : F(y \mid \mathbf{x}) \geq \tau\right\} = F^{-1}(\tau \mid \mathbf{x}),$$

which can be re-expressed as

$$q_\tau(\mathbf{x}) = \text{argmin}_{a \in \mathbb{R}^1} \mathbb{E}\left\{\rho_\tau\left(Y_t - a\right) \mid \mathbf{X}_t = \mathbf{x}\right\}. \tag{3.1}$$

Alternatively, $q_\tau(\mathbf{x})$ can be re-expressed as a regression type expression as

$$Y_t = q_\tau(\mathbf{X}_t) + \varepsilon_{t,\tau},$$

which is the so-called quantile residual, where $\varepsilon_{t,\tau}$ satisfies $\mathbb{P}(\varepsilon_{t,\tau} \leq 0) = \tau$. Specially, $q_{1/2}(\mathbf{X}_t)$ is called the median function. There are several advantages of using a quantile regression:

1. Similar to mean regression, a quantile regression does not require knowing the distribution of the dependent variable. But, different from a mean model, it does not require the symmetry of the measurement error. One of the most important features of quantile is that it is a robust procedure against outliers. Also, it does not impose any moment conditions, even the first moment does not exist, for example, the Cauchy distribution.

2. It can characterize the whole distribution so that it can depict all moments. Therefore, it can estimate the mean and variance simultaneously. For example, let us look at the location-scale model; that is

$$F_{y|\mathbf{x}}(y|\mathbf{X}_t) = F_\epsilon\left(\frac{y - m(\mathbf{X}_t)}{\sigma(\mathbf{X}_t)}\right), \quad \text{alternatively,} \quad Y_t = m(\mathbf{X}_t) + \sigma(\mathbf{X}_t)\,\epsilon_t, \tag{3.2}$$

where $F_\epsilon(\cdot)$ is the CDF of ϵ_t, which implies that

$$q_\tau(\mathbf{X}_t) = m(\mathbf{X}_t) + F_\epsilon^{-1}(\tau)\sigma(\mathbf{X}_t),$$

which contains both the mean function $m(\mathbf{X}_t)$ and the variance function or conditional variance $\sigma^2(\mathbf{X}_t)$, from which, one can see clearly that if $\sigma^2(\mathbf{X}_t) = \sigma_0^2$ is constant, $q_\tau(\mathbf{X}_t) = m(\mathbf{X}_t) + F_\epsilon^{-1}(\tau)\sigma_0$. This implies that for two different τ values, say $\tau_1 \neq \tau_2$, two quantile functions $q_{\tau_1}(\mathbf{x})$ and $q_{\tau_2}(\mathbf{x})$ are parallel (no crossing). This important feature can be used to detect if $\sigma^2(\mathbf{X}_t)$ is constant or not.

3. For a fixed $\mathbf{x}$, $q_\tau(\mathbf{x})$ is an increasing function of τ; that is, $q_{\tau_1}(\mathbf{x}) \leq q_{\tau_2}(\mathbf{X})$ for all $\mathbf{x}$ and $\tau_1 \leq \tau_2$. In other words, if $q_\tau(\mathbf{x})$ is a quantile function, then, $q_{\tau_1}(\mathbf{x}) \leq q_{\tau_2}(\mathbf{X})$ for all $\mathbf{x} \iff \tau_1 \leq \tau_2$. Therefore, in a real application, it should be cautious to specify a particular quantile function. For example, a linear specification cannot ensure 100% that $q_\tau(\mathbf{X}_t) = \boldsymbol{\beta}_\tau^\top \mathbf{X}_t$ is a real quantile.

4. More importantly, another application of conditional quantiles is the direct construction of prediction intervals for the next value given a small section of the recent past values in a stationary time series. See Example 3.3 in Section 3.5.2 for a real example. As argued in Koenker and Zhao (1996), the prediction interval based on the quantile approach is preferred in applications due to its shortest interval length and less assumptions.

Having conditioned on the observed characteristics $\mathbf{X}_t = \mathbf{x}$, based on the Skorohod representation[1], Y_t and the quantile function $q_\tau(\mathbf{x})$ have a following relationship as

$$Y_t = q\left(\mathbf{X}_t, v_t\right), \tag{3.3}$$

where v_t conditional on $\mathbf{X}_t$ follows $U(0,1)$. We refer to v_t as the rank variable, and note that representation (3.3) is essential to what follows. The rank variable v_t is responsible for heterogeneity of outcomes among individuals with the same observed characteristics $\mathbf{X}_t$. It also determines their relative ranking in terms of potential outcomes; hence one may think of rank v_t as representing some unobserved characteristic. This interpretation makes quantile analysis an interesting tool for describing and learning the structure of heterogeneous effects and controlling for unobserved heterogeneity. Finally, note that the representation in (3.3)

[1]For the definition, please see the book by Durrett (2019).

provides a great tool for us to generate a random variable based on a quantile model; see, for example, Example 3.1 in Section 3.5.1 for simulation studies.

Clearly, the simplest form of model (3.1) is $q_\tau(\mathbf{x}) = \boldsymbol{\beta}_\tau^\top \mathbf{x}$, which is called the linear quantile regression model well studied by many authors. For details, see the papers by Duffie and Pan (1997), Koenker (2000), Tsay (2002), Koenker and Hallock (2000), Khindanova and Rachev (2000), and Bao et al. (2006), Engle and Manganelli (2004), and references therein.

In many practical applications, however, the linear quantile regression model might not be "rich" enough to capture the underlying relationship between the quantile of response variable and its covariates. Indeed, some components may be highly nonlinear or some covariates may be interactive. To make the quantile regression model more flexible, there is a swiftly growing literature on nonparametric quantile regression. Various smoothing techniques, such as kernel methods, splines, and their variants, have been used to estimate the nonparametric quantile regression for both the independent and time series data. For the recent developments and the detailed discussions on theory, methodologies, and applications, see, for example, the papers by He et al. (1998), Yu and Jones (1998), He and Ng (1999), He and Ng (1999), He and Portnoy (2000), Honda (2000, 2004), Tsay (2002), Lu et al. (2000), Khindanova and Rachev (2000), Bao et al. (2006), Cai (2002a), De Gooijer and Zerom (2003), Horowitz and Lee (2005), Yu and Lu (2004), and Li and Racine (2008), and references therein. In particular, for the univariate case, Honda (2000) and Lu et al. (2000) derived the asymptotic properties of the local linear estimator of the quantile regression function under α-mixing condition.

For the high dimensional case, however, the aforementioned methods encounter some difficulties such as the so-called "curse of dimensionality" and their implementation in practice is not easy as well as the visual display is not so useful for the exploratory purposes. To attenuate the above problems, De Gooijer and Zerom (2003), Horowitz and Lee (2005), and Yu and Lu (2004) considered an additive quantile regression model $q_\tau(\mathbf{X}_t) = \sum_{j=1}^{p} g_{j,\tau}(X_{jt})$. To estimate each component, for the time series case, De Gooijer and Zerom (2003) first estimated a high dimensional quantile function by inverting the conditional distribution function estimated by using a weighted Nadaraya-Watson approach, proposed by Cai (2002a), and then used a projection method to estimate each component, as discussed in Cai and Masry (2000), while Yu and Lu (2004) focused on the independent data and used a back-fitting algorithm method to estimate each component. On the other hand, to estimate each additive

component for the independent data, Horowitz and Lee (2005) used a two-stage approach consisting of the series estimation in the first step and a local polynomial fitting in the second step. For the independent data, the above model was extended by He et al. (1998), He and Ng (1999), and He and Portnoy (2000) to include interaction terms by using spline methods. See Section 3.3.5 for details.

In this chapter, we adapt another dimension reduction modeling method to analyze dynamic time series data, termed as the smooth (functional or varying) coefficient modeling approach. This approach allows appreciable flexibility on the structure of fitted models. It allows for linearity in some continuous or discrete variables which can be exogenous or lagged and nonlinear in other variables in the coefficients. In such a way, the model has the ability of capturing the individual variations. More importantly, it can ease the so-called "curse of dimensionality" and combines both additivity and interactivity. A smooth coefficient quantile regression model for time series data takes the following form

$$q_\tau\left(\mathbf{U}_t, \mathbf{X}_t\right) = \sum_{j=1}^{p} a_{j,\tau}\left(\mathbf{U}_t\right) X_{tj} = \mathbf{a}_\tau\left(\mathbf{U}_t^\top \mathbf{X}_t\right), \tag{3.4}$$

where $\mathbf{U}_t \in \mathbb{R}^k$ is called the smoothing variable, which might be one part of $X_{t1}, \cdots, X_{tp}$ or just time or other exogenous variables or the lagged variables, $\mathbf{X}_t = (X_{t1}, \cdots, X_{td})$ with $X_{t1} \equiv 1, \{a_k(\cdot)\}$ are smooth coefficient functions, and $\mathbf{a}_\tau(\cdot) = (a_{1,\tau}(\cdot), \cdots, a_{p,\tau}(\cdot))$. Here, some of $\{a_{k,\tau}(\cdot)\}$ are allowed to depend on τ. For simplicity, we drop τ from $\{a_{j,\tau}(\cdot)\}$ in what follows. It is our interest here to estimate the coefficient functions $\mathbf{a}(\cdot)$ rather than the quantile regression surface $q_\tau(\cdot,\cdot)$ itself. Note that model (3.4) was studied by Honda (2004) for the independent sample, but our focus here is on the dynamic model for nonlinear time series, which is more appropriate for economic and financial applications.

The general setting in (3.4) covers many familiar quantile regression models, including the quantile autoregressive model (QAR) proposed by Koenker and Xiao (2004) by applying the QAR model for the unit root inference. In particular, it includes a specific class of ARCH models, such as heteroscedastic linear models considered by Koenker and Zhao (1996). Also, if there is no $\mathbf{X}_t$ in the model ($d = 0$), $q_\tau\left(\mathbf{U}_t, \mathbf{X}_t\right)$ becomes $q_\tau\left(\mathbf{U}_t\right)$ so that model (3.4) reduces to the ordinary nonparametric quantile regression model which has been studied extensively. For more developments, refer to the papers by He et al. (1998), Yu and Jones (1998), He and Ng (1999), He and Portnoy (2000), Honda (2000), Lu et al. (2000), Cai (2002a), De Gooijer and Zerom (2003), Horowitz and Lee (2005), Yu and Lu (2004), and Li

and Racine (2008). If $\mathbf{U}_t$ is just time, then the model is called the time-varying coefficient quantile regression model, which is potentially useful to see whether the quantile regression changes over time and in a case with a practical interest is, for example, the aforementioned illustrative example for the 1992 presidential election and the analysis of the reference growth data by Cole (1994), Wei et al. (2006), and Wei and He (2006), and the references therein. However, if $\mathbf{U}_t$ is time, the observed time series might not be stationary. Therefore, the treatment for non-stationary case would require a different approach so that it is beyond the scope of this chapter and deserves a further investigation. For more applications, see the work in Xu (2005). Finally, note that the smooth coefficient mean regression model is one of the most popular nonlinear time series models in mean regression and has various applications. For more discussions, refer to the papers by Chen and Tsay (1993), Cai et al. (2000), Cai and Tiwari (2000), Cai (2007), Hong and Lee (2003), and Wang (2003), and the book by Tsay (2002), and references therein.

The motivation of this study comes from an analysis of the well known Boston housing price data, consisting of several variables collected on each of 506 different houses from a variety of locations. The interest is to identify the factors affecting the house price in Boston area. As argued by Şentürk and Müller (2006), the correlation between the house price and the crime rate can be adjusted by the confounding variable which is the proportion of population of lower educational status through a varying coefficient model and the expected effect of increasing crime rate on declining house prices seems to be only observed for lower educational status neighborhoods in Boston. The interesting features of this dataset are that the response variable is the median price of a home in a given area and the distributions of the price and the major covariate (the confounding variable) are left skewed. Therefore, quantile methods are suitable for the analysis of this dataset. Therefore, such a problem can be tackled by using model (3.4). In another example, one is interested in exploring the possible nonlinearity feature, heteroscedasticity, and predictability of the exchange rates such as the Japanese Yen per US dollar. The detailed analysis of these data sets is reported in Section 3.5.

3.2 Parametric Quantile Models

If $\mathbf{a}_\tau(\cdot)$ in (3.4) is constant, say, $\mathbf{a}_\tau(\mathbf{U}_t) = \boldsymbol{\beta}_\tau$, then, (3.4) becomes the following linear quantile regression, popular in economics and finance, see, e.g., the book by Koenker et al.

(2017),

$$q_\tau(\mathbf{X}_t) = \boldsymbol{\beta}_\tau^\top \mathbf{X}_t.$$

If the above linear model holds, then, $\boldsymbol{\beta}_\tau$ can be estimated by the following sample version of the loss function

$$\widehat{\boldsymbol{\beta}}_\tau = \arg\min_{\beta_\tau} \sum_{t=1}^{n} \rho_\tau \left(Y_t - \boldsymbol{\beta}_\tau^\top \mathbf{X}_t\right), \tag{3.5}$$

which includes the least absolute deviation[2] estimation of $\boldsymbol{\beta}_{1/2}$, when $\tau = 1/2$; see, for example, the paper by Chen et al. (2008) and the book by Dodge (2008) for details. It was introduced by Roger Joseph Boscovich in 1757; see, e.g., the book by Dodge (2008) for more information. Evidently, $\widehat{\boldsymbol{\beta}}_\tau$ does not have a close form although $\boldsymbol{\beta}_\tau$ is a consistent estimate of $\boldsymbol{\beta}_\tau$. For the asymptotic theory of $\boldsymbol{\beta}_\tau$, the reader is referred to the book by Koenker (2005). To find the numerical solution of (3.5), Koenker (2004, 2005) suggested using the linear programming technique to compute the numerical solution of (3.5). To fit a linear quantile regression in practice, one can use the command **rq()** in the package **quantreg** in **R**.

For a nonlinear parametric model as $q_\tau(\mathbf{X}_t) = q_{\tau,0}(\mathbf{X}_t, \boldsymbol{\beta}_\tau)$, where $q_{\tau,0}(\cdot)$ is a known function with unknown parameter $\boldsymbol{\beta}_\tau$, then, (3.5) becomes

$$\widehat{\boldsymbol{\beta}}_\tau = \arg\min_{\beta_\tau} \sum_{t=1}^{n} \rho_\tau \left(Y_t - q_{\tau,0}(\mathbf{X}_t, \boldsymbol{\beta}_\tau)\right),$$

which can be empirically obtained by the command is **nlrq()** in **R**. By the way, for a nonparametric quantile model for univariate case, one can use the command **lprq()** for implementing the local polynomial estimation. For an additive quantile regression, one can use the commands **rqss()** and **qss()**. Before proceeding how to model nonparametric quantile regression models, let us talk about some implementation issues in practice for linear or nonlinear parametric quantile regressions.

When p is large, the computing burden is a big concern. Also, it is easy to that the check function $\rho_\tau(u)$ is not differentiable at $u = 0$, and this prevents the gradient based optimization methods from real applications. To overcome these problems, the gradient descent algorithm for quantile regression with smooth approximation (or the smoothed check

[2]Least absolute deviation (LAD) is a regression method that minimizes the sum of the absolute values of the residuals, or errors, between observed and predicted values. Unlike Ordinary Least Squares (OLS), which minimizes the sum of squared errors, LAD is more robust to outliers because it does not give excessive weight to extreme data points. This makes LAD a valuable alternative when a dataset contains outliers, though it is generally more computationally intensive than OLS, which is not a big concern in the current computing capacity.

function) was proposed by Zheng (2011), termed as GDS-QReg algorithm, described as follows. The smoothed check function to approximate the check function $\rho_\tau(u)$ is given by

$$\rho_{\tau,\zeta}(u) = \tau\, u + \zeta\, \log\left(1 + \exp(-u/\zeta)\right), \tag{3.6}$$

where ζ is called as the smooth parameter or the approximation error. Also, it is easy to verify that $\rho_{\tau,\zeta}(u)$ satisfies the following the properties: (1) For any ζ, $\rho_{\tau,\zeta}(u)$ is a convex function of u, and (2) For any ζ, $0 < \rho_{\tau,\zeta}(u) - \rho_\tau(u) < \zeta\, \log 2$, which implies that $\lim_{\zeta \downarrow 0} \rho_{\tau,\zeta}(u) = \rho_\tau(u)$. Furthermore, Zheng (2011) showed that for each $\tau \in (0,1)$, $\widehat{\beta}_{\tau,\zeta} - \widehat{\beta}_\tau \to 0$, where $\widehat{\beta}_{\tau,\zeta}$ is the smoothed estimate of β_τ via replacing the check function in (3.1) by the smoothed check function $\rho_{\tau,\zeta}(u)$ in (3.6). This implies that one can use the solution of smooth quantile regression model with small ζ to approximate the original quantile regression function. Note that this approximation is for computational convenience. Indeed, the same idea is used in the literature, see, for example, Chernozhukov et al. (2010) and Chernozhukov et al. (2013). To see the difference, Figure 3.1 displays the check function with $\tau = 0.15$ and 0.90 and the corresponding smoothed version with the smooth parameter $\zeta = 0.10$. From Figure

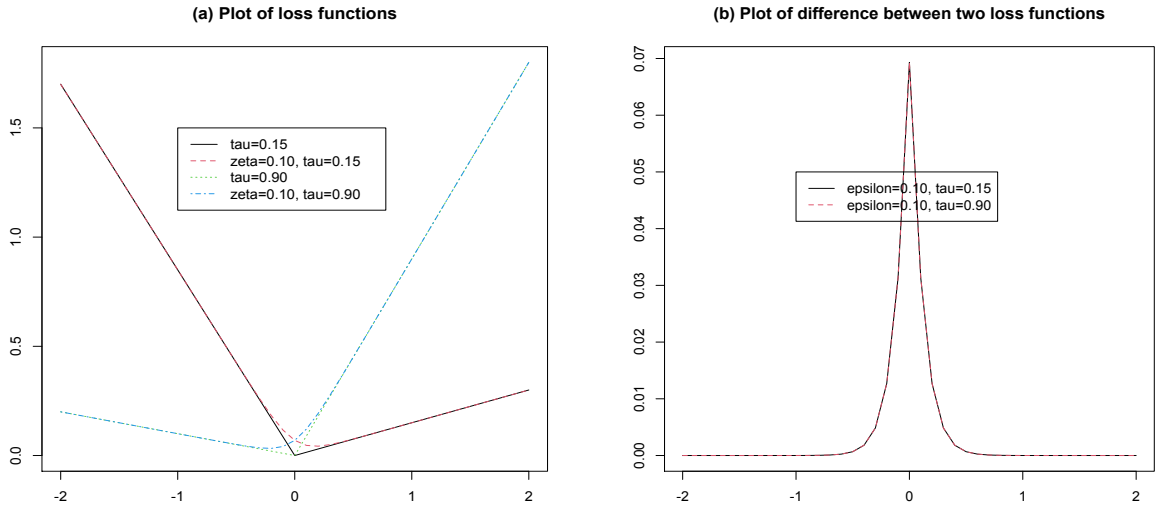

Figure 3.1: For $\tau = 0.15$ and 0.90 and $\zeta = 0.10$, the plot of the check function and the smoothed check function in (a) and the plot of the difference of the smoothed check function and the check function in (b).

3.1(a), one can see that the smoothed check function is always positive, smooth and convex, and dominates the original check function, whereas Figure 3.1(b) illustrates the difference between the check function and its smoothed version, and it is obviously observed that the

two functions approach each other quickly for both $\tau = 0.15$ and 0.90.

3.3 Nonparametric Modeling Procedures

3.3.1 Local Linear Quantile Estimate

Now, we apply the local polynomial method to the smooth coefficient quantile regression model as follows. For the sake of brevity, we only consider the case where $\mathbf{U}_t$ in (3.4) is one-dimensional ($k = 1$), denoted by U_t in what follows. Extension to multivariate $\mathbf{U}_t$ involves fundamentally no new ideas although the theory and procedure continue to hold. Note that the models with high dimension might not be practically useful due to the curse of dimensionality. A local polynomial fitting has several nice properties such as high statistical efficiency in an asymptotic minimax sense, design-adaptation, and automatic edge correction; see, e.g., Fan and Gijbels (1996).

We estimate the functions $\{a_j(\cdot)\}$ using the local polynomial regression method from observations $\{(U_t, \mathbf{X}_t, Y_t)\}_{t=1}^n$. We assume throughout the chapter that the coefficient functions $\mathbf{a}(\cdot)\}$ have the $(q+1)^{th}$ derivative, so that for any given gird point $u_0, a_k(\cdot)$ can be approximated by a polynomial function in a neighborhood of the given grid point u_0 as $\mathbf{a}(U_t) \approx \mathbf{a}(u_0) + \mathbf{a}^\top(u_0)(U_t - u_0) + \cdots + \mathbf{a}^{(q)}(u_0)(U_t - u_0)^q/q!$ and

$$q_\tau(U_t, \mathbf{X}_t) \approx \sum_{j=0}^{q} \mathbf{X}_t^\top \boldsymbol{\beta}_j (U_t - u_0)^j,$$

where $\boldsymbol{\beta}_j = \mathbf{a}^{(j)}(u_0)/j!$, which is a linear quantile regression model. Then, the locally weighted loss function is

$$\sum_{t=1}^{n} \rho_\tau \left(Y_t - \sum_{j=0}^{q} \mathbf{X}_t^\top \boldsymbol{\beta}_j (U_t - u_0)^j \right) K_h(U_t - u_0). \tag{3.7}$$

Solving the minimization problem in (3.7) gives $\widehat{\mathbf{a}}(u_0) = \widehat{\boldsymbol{\beta}}_0$, the local polynomial estimate of $\mathbf{a}(u_0)$, and $\widehat{\mathbf{a}}^{(j)}(u_0) = j!\widehat{\boldsymbol{\beta}}_j$ $(j \geq 1)$, the local polynomial estimate of the j^{th} derivative $\mathbf{a}^{(j)}(u_0)$ of $\mathbf{a}(u_0)$. By moving u_0 along with the real line, one obtains the estimate for the entire curve. For various practical applications, Fan and Gijbels (1996) recommended using the local linear fit ($q = 1$). Therefore, for the expositional purpose, in what follows, we only consider the case $q = 1$ (local linear fitting).

The programming involved in the local (polynomial) linear quantile estimation is relatively simple and can be modified with few efforts from the existing programs for a linear quantile model. For example, for each grid point u_0, the local linear quantile estimation can be implemented in the **R** package **quantreg**, of Koenker (2004) by setting covariates as $\mathbf{X}_t$ and $\mathbf{X}_t (U_t - u_0)$ and the weight as $K_h (U_t - u_0)$.

Although some modifications are needed, the method developed here for the local linear quantile estimation is applicable to a general local polynomial quantile estimation. In particular, we note that the local constant (Nadaraya-Watson type) quantile estimation of $\mathbf{a}(u_0)$, denoted by $\widetilde{\mathbf{a}}(u_0)$, is $\widetilde{\boldsymbol{\beta}}$ minimizing the following subjective function

$$\sum_{t=1}^{n} \rho_\tau \left(Y_t - \mathbf{X}_t^\top \boldsymbol{\beta} \right) K_h (U_t - u_0), \tag{3.8}$$

which is a special case of (3.7) with $q = 0$ and is the weighted version of (3.5). We compare $\widehat{\mathbf{a}}(u_0)$ and $\widetilde{\mathbf{a}}(u_0)$ theoretically in (3.11) later, which leads to suggest that one should use the local linear approach in practice.

3.3.2 Asymptotic Results

We first give some regularity conditions that are sufficient for the consistency and asymptotic normality of the proposed estimators, although they might not be the weakest possible. We introduce the following notations. Denote

$$\Omega(u_0) \equiv \mathbb{E}\left[\mathbf{X}_t \mathbf{X}_t^\top \mid \mathbf{U}_t = u_0 \right] \quad \text{and} \quad \Omega^*(u_0) \equiv \mathbb{E}\left[\mathbf{X}_t \mathbf{X}_t^\top f_{y|u,x}(q_\tau(u_0, \mathbf{X}_t)) \mid \mathbf{U}_t = u_0 \right],$$

where $f_{y|u,x}(y)$ is the conditional density of Y given U and $\mathbf{X}$. Let $f_u(u)$ present the marginal density of U.

Assumptions:

(C1) $\mathbf{a}(u)$ is twice continuously differentiable in a neighborhood of u_0 for any u_0.

(C2) $f_u(u)$ is continuous and $f_u(u_0) > 0$.

(C3) $f_{y|u,x}(y)$ is bounded and satisfies the Lipschitz condition.

(C4) The kernel function $K(\cdot)$ is symmetric and has a compact support, say $[-1, 1]$.

(C5) $\{(\mathbf{X}_t, Y_t, \mathbf{U}_t)\}$ is a strictly α-mixing stationary process with mixing coefficient $\alpha(t)$ satisfies $\sum_{t\geq 1}^{\infty} t^l \alpha^{(\delta-2)/\delta}(t) < \infty$ for some positive real number $\delta > 2$ and $l > (\delta-2)/\delta$.

(C6) $E\left\|\mathbf{X}_t\right\|^{2\delta^*} < \infty$ with $\delta^* > \delta$.

(C7) $\Omega\left(u_0\right)$ is positive-definite and continuous in a neighborhood of u_0

(C8) $\Omega^*\left(u_0\right)$ is continuous and positive-definite in a neighborhood of u_0.

(C9) The bandwidth h satisfies $h \to 0$ and $nh \to \infty$.

(C10) $f\left(u, v \mid \mathbf{x}_0, \mathbf{x}_s; s\right) \leq M < \infty$ for $s \geq 1$, where $f\left(u, v \mid \mathbf{x}_0, \mathbf{x}_s; s\right)$ is the conditional density of (U_0, U_s) given $(\mathbf{X}_0 = \mathbf{x}_0, \mathbf{X}_s = \mathbf{x}_s)$.

(C11) $n^{1/2-\delta/4} h^{\delta/\delta^*-1/2-\delta/4} = O(1)$.

Remark 3.1: *(Discussion of Conditions) Assumptions C1 - C3 include some smoothness conditions on functionals involved. The requirement in C4 that $K(\cdot)$ be compactly supported is imposed for the sake of brevity of proofs, and can be removed at the cost of lengthier arguments. In particular, the Gaussian kernel is allowed. The α-mixing is one of the weakest mixing conditions for weakly dependent stochastic processes. Stationary time series or Markov chains fulfilling certain (mild) conditions are α-mixing with exponentially decaying coefficients; see the discussions in Section 1.1 and Cai (2002a) for more examples. On the other hand, the assumption on the convergence rate of $\alpha(\cdot)$ in C5 might not be the weakest possible and is imposed to simplify the proof. Further, Assumption C10 is just a technical assumption, which is also imposed by Cai (2002a). Assumptions C6 - C8 require some standard moments. Clearly, Assumption C11 allows the choice of a wide range of smoothing parameter values and is slightly stronger than the usual condition of $nh \to \infty$. However, for the bandwidths of optimal size (i.e., $h = O\left(n^{-1/5}\right)$), Assumption C11 is automatically satisfied for $\delta \geq 3$ and it is still fulfilled for $2 < \delta < 3$ if δ^* satisfies $\delta < \delta^* \leq 1 + 1/(3-\delta)$, so that we do not concern ourselves with such refinements. Indeed, this assumption is also imposed by Cai et al. (2000) for the mean regression. Finally, if there is no $\mathbf{X}_t$ in model (3.4), Assumption C5 can be replaced by Assumption C(5)* : $\alpha(t) = O\left(t^{-\delta}\right)$ for some $\delta > 2$ and Assumption C(11) can be substituted by Assumption C(11)* : $nh^{\delta/(\delta-2)} \to \infty$; see Cai (2002a) for details.*

Remark 3.2: *(Identification) It is clear from* (3.4) *that*

$$\Omega(u_0)\,\mathbf{a}(u_0) = \mathbb{E}\left[q_\tau(u_0, \mathbf{X}_t)\,\mathbf{X}_t \mid U_t = u_0\right].$$

Then, $\mathbf{a}(u_0)$ *is identified (uniquely determined) if and only if* $\Omega(u_0)$ *is positive definite for any* u_0*. Therefore, Assumption C(7) is the necessary and sufficient condition for the model identification.*

To establish the asymptotic normality of the proposed estimator, similar to Chaudhuri (1991), we first derive the local Bahadur representation for the local linear estimator. To this end, our analysis follows the approach of Koenker and Zhao (1996), which can simplify the theoretical proofs. Define $\psi_\tau(x) = \tau - I_{\{x<0\}}, U_{th} = (U_t - u_0)/h, \mathbf{X}_t^* = \begin{pmatrix} \mathbf{X}_t \\ U_{th}\mathbf{X}_t \end{pmatrix}, Y_t^* = Y_t - \mathbf{X}_t^\top\left[\mathbf{a}(u_0) + \mathbf{a}^\top(u_0)(U_t - u_0)\right]$, and $\boldsymbol{\theta} = \sqrt{nh}\mathbf{H}\begin{pmatrix} \boldsymbol{\beta}_0 - \mathbf{a}(u_0) \\ \boldsymbol{\beta}_1 - \mathbf{a}^\top(u_0) \end{pmatrix}$ with $\mathbf{H} = \mathrm{diag}\{\mathbf{I}, h\mathbf{I}\}$.

Theorem 3.1: *(Local Bahadur Representation) Under Assumptions C1 - C9, we have*

$$\widehat{\boldsymbol{\theta}} = \frac{[\Omega_1^*(u_0)]^{-1}}{\sqrt{nh} f_u(u_0)} \sum_{t=1}^{n} \psi_\tau(Y_t^*)\,\mathbf{X}_t^* K(U_{th}) + o_p(1), \tag{3.9}$$

where $\Omega_1^*(u_0) = diag\{\Omega^*(u_0), \mu_2(K)\Omega^*(u_0)\}$.

Remark 3.3: *From Theorem 3.1 and Lemma 3.1 (in Section 3.6), it is easy to see that the local linear estimator* $\widehat{\mathbf{a}}(u_0)$ *is consistent with the optimal nonparametric convergence rate* $\sqrt{nh}$

Theorem 3.2: *(Asymptotic Normality) Under Assumptions C1- C11, we have the following asymptotic normality*

$$\sqrt{nh}\left[\mathbf{H}\begin{pmatrix} \widehat{\mathbf{a}}(u_0) - \mathbf{a}(u_0) \\ \widehat{\mathbf{a}}^\top(u_0) - \mathbf{a}^\top(u_0) \end{pmatrix} - \frac{h^2}{2}\begin{pmatrix} \mathbf{a}''(u_0)\,\mu_2(K) \\ 0 \end{pmatrix} + o_p(\left(h^2\right)\right] \quad \xrightarrow{d} \quad N(0, \boldsymbol{\Sigma}(u_0))$$

where $\boldsymbol{\Sigma}(u_0) = diag\{\tau(1-\tau)\nu_0(K)\boldsymbol{\Sigma}_a(u_0), \tau(1-\tau)\nu_2(K)\boldsymbol{\Sigma}_a(u_0)\}$ *with*

$$\boldsymbol{\Sigma}_a(u_0) = [\Omega^*(u_0)]^{-1}\,\Omega(u_0)\,[\Omega^*(u_0)]^{-1}/f_u(u_0) \tag{3.10}$$

In particular,

$$\sqrt{nh}\left[\widehat{\mathbf{a}}(u_0) - \mathbf{a}(u_0) - \frac{h^2\mu_2(K)}{2}\mathbf{a}''(u_0) + o_p(\left(h^2\right)\right] \quad \to \quad N\{0, \tau(1-\tau)\nu_0(K)\boldsymbol{\Sigma}_a(u_0)\}$$

Remark 3.4: *From Theorem 3.2, the asymptotic mean squares error (AMSE) of* $\widehat{\mathbf{a}}(u_0)$ *is given by*

$$AMSE = \frac{h^4\mu_2^2(K)}{4}\left\|\mathbf{a}''(u_0)\right\|^2 + \frac{\tau(1-\tau)\nu_0(K)}{nhf_u(u_0)}tr\left(\mathbf{\Sigma}_a(u_0)\right),$$

which gives the optimal bandwidth h_{opt} *by minimizing the AMSE*

$$h_{opt} = \left(\frac{\tau(1-\tau)\nu_0(K)tr\left(\mathbf{\Sigma}_a(u_0)\right)}{f_u(u_0)\left\|\mathbf{a}''(u_0)\right\|^2}\right)^{1/5} n^{-1/5},$$

and the optimal AMSE is

$$AMSE_{opt} = \frac{5}{4}\left(\frac{\tau(1-\tau)\nu_0(K)tr\left(\mathbf{\Sigma}_a(u_0)\right)}{f_u(u_0)}\right)^{4/5}\left\|\mathbf{a}''(u_0)\right\|^{2/5} n^{-4/5}.$$

Further, notice that the similar results in Theorem 3.2 were obtained by Honda (2004) for the independent data. Finally, it is interesting to note that the asymptotic bias in Theorem 3.2 is the same as that for the mean regression case but the two asymptotic variances are different; see, for example, Cai et al. (2000).

If model (3.4) does not have $\mathbf{X}$, it becomes the nonparametric quantile regression model $q_\tau(\cdot)$. Then, we have the following asymptotic normality for the local linear estimator of the nonparametric quantile regression function $q_\tau(\cdot)$, which covers the results in Yu and Jones (1998), Honda (2000), Lu et al. (2000), and Cai (2002a) for both the independent and time series data.

Corollary 3.2.1: *If there is no* $\mathbf{X}_t$ *in* (3.4)*, then,*

$$\sqrt{nh}\left[\widehat{q}_\tau(u_0) - q_\tau(u_0) - \frac{h^2\mu_2(K)}{2}q_\tau''(u_0) + o_p(\left(h^2\right)\right] \xrightarrow{d} N\left\{0, \sigma_\tau^2(u_0)\right\},$$

where $\sigma_\tau^2(u_0) = \tau(1-\tau)\nu_0(K)f_u^{-1}(u_0)\, f_{y|u}^{-2}(q_\tau(u_0))$.

Now we consider the comparison of the performance of the local linear estimation $\widehat{\mathbf{a}}(u_0)$ obtained in (3.7) with that of the local constant estimation $\widetilde{\mathbf{a}}(u_0)$ given in (3.8). To this effect, first, we derive the asymptotic results for the local constant estimator but the proof is omitted since it is along the same line with the proof of Theorems 3.1 and 3.2; see Xu (2005) for details. Under some regularity conditions, it can be shown that

$$\sqrt{nh}\left[\widetilde{\mathbf{a}}(u_0) - \mathbf{a}(u_0) - \widetilde{\mathbf{b}} + o_p(\left(h^2\right)\right] \quad \to \quad N\left\{0, \tau(1-\tau)\nu_0(K)\Sigma_a(u_0)\right\},$$

where

$$\widetilde{\mathbf{b}} = \frac{h^2\mu_2(K)}{2}\left[\mathbf{a}''(u_0) + 2\mathbf{a}^\top(u_0)\, f_u'(u_0)/f_u(u_0) + 2\left\{\Omega^*(u_0)\right\}^{-1}\Omega^{*\prime}(u_0)\,\mathbf{a}^\top(u_0)\right], \quad (3.11)$$

which implies that the asymptotic bias for $\widetilde{\mathbf{a}}(u_0)$ is different from that for $\widehat{\mathbf{a}}(u_0)$ but both have the same asymptotic variance. Therefore, the local constant quantile estimator does not adapt to nonuniform designs: the bias can be large when $f_u'(u_0)/f_u(u_0)$ or $\{\Omega^*(u_0)\}^{-1}\Omega^{*\prime}(u_0)$ is large even when the true coefficient functions are linear. It is surprising that to the best of our knowledge, this finding seems to be new for the nonparametric quantile regression setting although it is well documented in literature for the ordinary regression case; see Fan and Gijbels (1996) for details.

Finally, to examine the asymptotic behaviors of the local linear and local constant quantile estimators at the boundaries, we offer Theorem 3.3 below but its proofs are omitted due to their similarity to those for Theorem 3.2 with some modifications and for the ordinary regression setting; see, e.g., Fan and Gijbels (1996) and also, see Xu (2005) for the detailed proofs. Without loss of generality, we consider only the left boundary point $u_0 = ch, 0 < c < 1$, if U_t takes values only from $[0, 1]$. A similar result in Theorem 3.3 holds for the right boundary point $u_0 = 1 - ch$. Define $\mu_{j,c} = \int_{-c}^{1} u^j K(u)du$ and $\nu_{j,c} = \int_{-c}^{1} u^j K^2(u)du$.

Theorem 3.3: *(Asymptotic Normality) Under the assumptions in Theorem 3.2 , we have the following asymptotic normality of the local linear quantile estimator at the left boundary point,*

$$\sqrt{nh}\left[\widehat{\mathbf{a}}(ch) - \mathbf{a}(ch) - \frac{h^2 b_c}{2}\mathbf{a}''(0+) + o_p(\left(h^2\right)\right] \xrightarrow{d} N\left\{0, \tau(1-\tau)v_c\boldsymbol{\Sigma}_a(0+)\right\},$$

where

$$b_c = \frac{\mu_{2,c}^2 - \mu_{1,c}\mu_{3,c}}{\mu_{2,c}\mu_{0,c} - \mu_{1,c}^2} \quad \textit{and} \quad v_c = \frac{\mu_{2,c}^2\nu_{0,c} - 2\mu_{1,c}\mu_{2,c}\nu_{1,c} + \mu_{1,c}^2\nu_{2,c}}{\left[\mu_{2,c}\mu_{0,c} - \mu_{1,c}^2\right]^2}.$$

Further, we have the following asymptotic normality of the local constant quantile estimator at the left boundary point $u_0 = ch$ for $0 < c < 1$,

$$\sqrt{nh}\left[\widetilde{\mathbf{a}}(ch) - \mathbf{a}(ch) - \widetilde{\mathbf{b}}_c + o_p(\left(h^2\right)\right] \quad \to \quad N\left\{0, \tau(1-\tau)\nu_{0,c}\boldsymbol{\Sigma}_a(0+)/\mu_{0,c}^2\right\}.$$

where

$$\widetilde{\mathbf{b}}_c = \left[h\mu_{1,c}\mathbf{a}^\top(0+) + \frac{h^2\mu_{2,c}}{2}\left\{\mathbf{a}''(0+) + \frac{2\mathbf{a}^\top(0+)f_u'(0+)}{f_u(0+)} + 2\Omega^{*-1}(0+)\Omega^{*\prime}(0+)\mathbf{a}^\top(0+)\right\}\right]/\mu_{0,c}.$$

Similar results hold for the right boundary point $u_0 = 1 - ch$.

Remark 3.5: *We remark that if the point 0 were an interior point, then, Theorem 3.3 would hold with* $c = 1$*, which becomes Theorem 3.2. Also, as* $c \to 1, b_c \to \mu_2(K)$*, and* $v_c \to \nu_0(K)$ *and these limits are exactly the constant factors appearing respectively in the asymptotic bias and variance for an interior point. Therefore, Theorem 3.3 shows that the local linear estimation has the automatic good behavior at boundaries without the need of boundary correction. Further, one can see from Theorem 3.3 that at the boundaries, the asymptotic bias term for the local constant quantile estimate is of the order* h *by comparing to the order* h^2 *for the local linear quantile estimate. This shows that the local linear quantile estimate does not suffer from boundary effects but the local constant quantile estimate does, which is another advantage of the local linear quantile estimator over the local constant quantile estimator. This suggests that one should use the local linear approach in practice.*

As a special case, Theorem 3.3 includes the asymptotic properties for the local constant quantile estimator of the nonparametric quantile function $q_\tau(\cdot)$ at both the interior and boundary points, stated as follows.

Corollary 3.3.1: *If there is no* $\mathbf{X}_t$ *in* (3.4)*, then, the asymptotic normality of the local constant quantile estimator is given by*

$$\sqrt{nh}\left[\widetilde{q}_\tau(u_0) - q_\tau(u_0) - \frac{h^2\mu_2(K)}{2}\left\{q''_\tau(u_0) + \frac{2q'_\tau(u_0)\,f'_u(u_0)}{f_u(u_0)}\right\} + o_p(\left(h^2\right)\right] \xrightarrow{d} N\left\{0, \sigma^2_\tau(u_0)\right\}.$$

Further, at the left boundary point, we have

$$\sqrt{n\,h}\left[\widetilde{q}_\tau(c\,h) - q_\tau(ch) - \widetilde{b}^*_c + o_p(h^2)\right] \xrightarrow{d} N\left\{0, \sigma^2_c\right\},$$

where

$$\widetilde{b}^*_c = \left[h\mu_{1,c}q'_\tau(0+) + \frac{h^2\mu_{2,c}}{2}\left\{q''_\tau(0+) + 2q'_\tau(0+)f'_u(0+)/f_u(0+)\right\}\right]/\mu_{0,c}$$

and $\sigma^2_c = \tau(1-\tau)\nu_{0,c}f_u^{-1}(0+)f^{-2}_{y|u}\left(q_\tau(0+)\right)/\mu^2_{0,c}$.

3.3.3 Bandwidth Selection

It is well known that the bandwidth plays an essential role in the trade-off between reducing bias and variance. To the best of our knowledge, there has been almost nothing done about selecting the bandwidth in the context of estimating the coefficient functions in the quantile regression even though there is a rich amount of literature on this issue in the

mean regression setting; see, for example, Cai et al. (2000). In practice, it is desirable to have a quick and easily implemented data-driven fashioned method. Based on this spirit, Yu and Jones (1998) or Yu and Lu (2004) proposed a simple and convenient method for the nonparametric quantile estimation. Their approach assumes that the second derivatives of the quantile function are parallel. However, this assumption might not be valid for many applications in economics and finance due to (nonlinear) heteroscedasticity. Further, the mean regression approach cannot directly estimate the variance function. To attenuate these problems, we propose a method of selecting bandwidth for the foregoing estimation procedure, based on the nonparametric version of the Akaike information criterion, which can attend to the structure of time series data and the over-fitting or under-fitting tendency. This idea is motivated by its analogue of Cai and Tiwari (2000) and Cai (2002b) for nonlinear time series models. The basic idea is described below.

By recalling the classical AIC for linear models under the likelihood setting

$$-2(\text{maximized log quasi-likelihood}) + 2(\text{number of estimated parameters}),$$

we propose the following nonparametric version of the bias-corrected AIC, due to Hurvich and Tsai (1989) for parametric models and Hurvich et al. (1998) for nonparametric regression models, to select h by minimizing

$$\text{AIC}(h) = \log\left\{\widehat{\sigma}_\tau^2\right\} + 2\left(p_h + 1\right) / \left[n - \left(p_h + 2\right)\right], \tag{3.12}$$

where $\widehat{\sigma}_\tau^2$ and p_h are defined later. This criterion may be interpreted as the AIC for the local quantile smoothing problem and seems to perform well in some limited applications. Note that similar to (3.12), Koenker et al. (1994) considered the Schwarz information criterion (SIC) of Schwarz (1978) with the second term on the right-hand side of (3.12) replayed by $2n^{-1}p_h \log n$, where p_h is the number of "active knots" for the smoothing spline quantile setting, and Machado (1993) studied similar criteria for parametric quantile regression models and more general M-estimators of regression.

Now the question is how to define $\widehat{\sigma}_\tau^2$ and p_h in this setting. In the mean regression setting, $\widehat{\sigma}_\tau^2$ is just the estimate of the variance σ^2. In the quantile regression, we define $\widehat{\sigma}_\tau^2$ as $n^{-1}\sum_{t=1}^n \rho_\tau\left(Y_t - \mathbf{X}_t^\top \widehat{\mathbf{a}}\left(U_t\right)\right)$, which may be interpreted as the mean square error in the least square setting and was also used by Koenker et al. (1994). In nonparametric models, p_h is the nonparametric version of degrees of freedom, called the effective number of parameters,

and it is usually based on the trace of various quasi-projection (hat) matrices in the least square theory (linear estimators); see, for example, Hastie and Tibshirani (1990), Cai and Tiwari (2000), and Cai (2002b) for a cogent discussion for nonparametric regression models and nonlinear time series models. For the quantile smoothing setting, the explicit expression for the quasi-projection matrix does not exist due to its nonlinearity. However, we can use the first order approximation (the local Bahadur representation) given in (3.9) to derive an explicit expression, which may be interpreted as the quasi-projection matrix in this setting. To this end, define

$$\mathbf{S}_n = \mathbf{S}_n(u_0) = a_n \sum_{t=1}^{n} \xi_t \mathbf{X}_t^* \mathbf{X}_t^{*\prime} K(U_{th}),$$

where $\xi_t = I\left(Y_t \le \mathbf{X}_t^\top \mathbf{a}(u_0) + a_n\right) - I\left(Y_t \le \mathbf{X}_t^\top \mathbf{a}(u_0)\right)$ and $a_n = (n\,h)^{-1/2}$. It is shown in Section 3.6 that

$$\mathbf{S}_n(u_0) = f_u(u_0)\,\Omega_1^*(u_0) + o_p(1). \tag{3.13}$$

From (3.9), it is easy to verify that $\widehat{\boldsymbol{\theta}} \approx a_n \mathbf{S}_n^{-1} \sum_{t=1}^{n} \psi_\tau(Y_t^*)\,\mathbf{X}_t^* K(U_{th})$. Then, we have

$$\widehat{q}_\tau(U_t, \mathbf{X}_t) - q_\tau(U_t, \mathbf{X}_t) \approx \frac{1}{n} \sum_{s=1}^{n} \psi_\tau\left(Y_s^*(U_t)\right) K_h\left((U_s - U_t)/h\right) \mathbf{X}_t^{0^\top} \mathbf{S}_n^{-1}(U_t)\,\mathbf{X}_s^*$$

where $\mathbf{X}_t^0 = \begin{pmatrix} \mathbf{X}_t \\ \mathbf{0} \end{pmatrix}$. The coefficient of $\psi_\tau(Y_s^*(U_s))$ on the right-hand side of the above expression is $\gamma_s = a_n^2 K(0) \mathbf{X}_s^{0^\top} \mathbf{S}_n^{-1}(U_s)\,\mathbf{X}_s^0$. Now, we have that $p_h = \sum_{s=1}^{n} \gamma_s$, which can be regarded as an approximation to the trace of the quasi-projection (hat) matrix for linear estimators. In the practical implementation, we need to estimate $\mathbf{a}(u_0)$ first since $\mathbf{S}_n(u_0)$ involves $\mathbf{a}(u_0)$. We recommend using a pilot bandwidth which can be chosen as the one proposed by Yu and Jones (1998). Similar to the least square theory, as expected, the criterion proposed in (3.9) counteracts the over-fitting tendency of the generalized cross-validation due to its relatively weak penalty and the under-fitting of the SIC of Schwarz (1978) studied by Koenker et al. (1994) because of the heavy penalty.

3.3.4 Covariance Estimate

For the purpose of statistical inference, we next consider the estimation of the asymptotic covariance matrix to construct the point-wise confidence intervals. In practice, a quick and simple way to estimate the asymptotic covariance matrix is desirable. In view of (3.10), the explicit expression of the asymptotic covariance provides a direct estimator. Therefore, we

can use the so-called "sandwich" method. In other words, we need to obtain a consistent estimate for both $\Omega(u_0)$ and $\Omega^*(u_0)$. To this effect, define,

$$\widehat{\Omega}_{n,0} = \frac{1}{n}\sum_{t=1}^{n} \mathbf{X}_t\mathbf{X}_t^\top K_h(U_t - u_0) \quad \text{and} \quad \widehat{\Omega}_{n,1} = \frac{1}{n}\sum_{t=1}^{n} w_t\mathbf{X}_t\mathbf{X}_t^\top K_h(U_t - u_0)$$

where $w_t = I\left(\mathbf{X}_t^\top\widehat{\mathbf{a}}(u_0) - \delta_n < Y_t \le \mathbf{X}_t^\top\widehat{\mathbf{a}}(u_0) + \delta_n\right)/(2\delta_n)$ for any $\delta_n \to 0$ as $n \to \infty$. It is shown in Section 3.6 that

$$\widehat{\Omega}_{n,0} = f_u(u_0)\,\Omega(u_0) + o_p(1) \quad \text{and} \quad \widehat{\Omega}_{n,1} = f_u(u_0)\,\Omega^*(u_0) + o_p(1). \tag{3.14}$$

Therefore, the consistent estimate of $\mathbf{\Sigma}_a(u_0)$ is given by

$$\widehat{\Sigma}_a(u_0) = \left[\widehat{\Omega}_{n,1}(u_0)\right]^{-1}\widehat{\Omega}_{n,0}(u_0)\left[\widehat{\Omega}_{n,1}(u_0)\right]^{-1}.$$

Note that $\widehat{\Omega}_{n,1}(u_0)$ might be close to singular for some sparse regions. To avoid this computational difficulty, there are two alternative ways to construct a consistent estimate of $f_u(u_0)\,\Omega^*(u_0)$ through estimating the conditional density of Y, $f_{y|u,x}(q_\tau(u,\mathbf{x}))$. The first method is the Nadaraya-Watson type (or local linear) double kernel method of Fan et al. (1996) defined as,

$$\widehat{f}_{y|u,x}(q_\tau(u,\mathbf{x})) = \sum_{t=1}^{n} K_{h_2}(U_t - u, \mathbf{X}_t - \mathbf{x})\,L_{h_1}(Y_t - q_\tau(u,\mathbf{x}))\,/\sum_{t=1}^{n} K_{h_2}(U_t - u, \mathbf{X}_t - \mathbf{x}),$$

where $L(\cdot)$ is a kernel function, and the second one is the difference quotients method of Koenker and Xiao (2004) such as

$$\widehat{f}_{y|u,x}(q_\tau(u,\mathbf{x})) = (\tau_j - \tau_{j-1})\,/\left[q_{\tau_j}(u,\mathbf{x}) - q_{\tau_{j-1}}(u,\mathbf{x})\right],$$

for some appropriately chosen sequence of $\{\tau_j\}$; see Koenker and Xiao (2004) for more discussions. Then, in view of the definition of $f_u(u_0)\,\Omega^*(u_0)$, the estimator $\widetilde{\Omega}_{n,1}$ can be constructed as,

$$\widetilde{\Omega}_{n,1} = \frac{1}{n}\sum_{t=1}^{n} \widehat{f}_{y|u,x}\left(\widehat{q}_\tau(U_t, \mathbf{X}_t)\right)\mathbf{X}_t\mathbf{X}_t^\top K_h(U_t - u_0).$$

By an analogue of (3.14), one can show that under some regularity conditions, both estimators are consistent.

3.3.5 Additive Quantile Regression Model

Similar the additive model discussed in Section 2.6 for mean models, De Gooijer and Zerom (2003) and Horowitz and Lee (2005) considered the additive quantile regression as

$$q_\tau(\mathbf{X}_t) = g_{1,\tau} + \sum_{j=2}^{p} g_{j,\tau}(X_{jt}).$$

Again, to identify the model, it is assumed that $\mathbb{E}(g_{j,\tau}(X_{jt})) = 0$ for $2 \le h \le p$. De Gooijer and Zerom (2003) used the projection method to estimate $q_\tau(\mathbf{X}_t)$ first and then use the projection method as in Section 2.6.3 to estimate each component $g_{j,\tau}(\cdot)$. As pointed out by Horowitz and Lee (2005), when $p \ge 5$, the projection method as in De Gooijer and Zerom (2003) needs a bias reduction in the nuisance directions. Therefore, to overcome this difficulty, Horowitz and Lee (2005) still used a two-stage approach: First, use the B-spline approach to approximate each component, and then employ (3.5) to estimate the relevant parameters. As a result, the estimate of each component is regarded an initial estimate, denoted by $\tilde{g}_{j,\tau}(\cdot)$. In the second stage, Horowitz and Lee (2005) suggested to estimating $g_{j,\tau}(\cdot)$ $(2 \le j \le p)$ by running the partial residual,

$$Y_{jt}^* = Y_t - \left[\tilde{g}_{1,0} + \sum_{i=2, i\neq j}^{p} \tilde{g}_{i,\tau}(\mathbf{X}_t)\right],$$

versus X_{jt} as in (3.7), so that $\widehat{g}_{j,\tau}(\cdot)$ is obtained. Finally, note that Yu and Lu (2004) proposed using a backfitting algorithm as in Section 2.6.2 to estimate each nonparametric component; see, e.g., Yu and Lu (2004) for details.

3.4 Semiparametric Models

In this section, if the dimension of $\mathbf{U}_t$ is large, similar to (2.19) in Section 2.5.1 or (2.70) in Section 2.7.4, we can generalize the model in (3.4) into the following functional coefficient index quantile regression model

$$q_\tau(\mathbf{U}_t, \mathbf{X}_t) = \sum_{j=1}^{p} a_{j,\tau}\left(\gamma^\top \mathbf{U}_t\right) X_{tj} = \mathbf{X}_t^\top \mathbf{a}_\tau\left(\gamma^\top \mathbf{U}_t\right), \tag{3.15}$$

which was investigated by Lv and Li (2022) by proposing the following estimation procedure to estimate $\mathbf{a}_\tau(\cdot)$ and γ. First, by the B-spline approximation, $a_{,\tau}(\gamma^\top \mathbf{U}_t) \approx \boldsymbol{\theta}_j^\top \mathbf{B}_j(\gamma^\top \mathbf{U}_t)$,

where $\mathbf{B}_j(\cdot)$ is the vector of the B-spline basis known functions. Then,

$$q_\tau(\mathbf{U}_t, \mathbf{X}_t) \approx \sum_{j=1}^{p} \boldsymbol{\theta}_j^\top \mathbf{B}_j(\gamma^\top \mathbf{U}_t)\mathbf{X}_{tj},$$

so that minimizing the following sample version of the loss function for nonlinear parametric quantile regression

$$\sum_{t=1}^{n} \rho_\tau \left(Y_t - \sum_{j=1}^{p} \boldsymbol{\theta}_j^\top \mathbf{B}_j(\gamma^\top \mathbf{U}_t)\mathbf{X}_{tj} \right)$$

gives the estimate of $\{\boldsymbol{\theta}_j\}$ and γ. Lv and Li (2022) derived the asymptotic theory for the proposed estimators. Clearly, when $\mathbf{X}_t = 1$, the model in (3.15) reduces to the quantile index model as in (2.71), explored by Wu et al. (2010).

Moreover, Cai and Xiao (2012) explored the following semiparametric quantile model, including partially linear quantile model as a special case

$$q_\tau(\mathbf{U}_t, \mathbf{X}_t) = \beta_\tau^\top \mathbf{X}_{1t} + \alpha_\tau(\mathbf{U}_t)^\top \mathbf{X}_{2t}, \tag{3.16}$$

which is the quantile version of the model in (2.60), where $\mathbf{X}^\top = (\mathbf{X}_{1t}^\top, \mathbf{X}_{2t}^\top)$, $\alpha_\tau(\cdot)$ is an unknown function, and β_τ is a unknown parameter. Of interest is to estimate β_τ and $\alpha_\tau(\cdot)$. As stated in Cai and Xiao (2012), to estimate β_τ, the classical profile least squares as in Speckman (1988) for mean models, described in Section 2.6.1, fails for semiparametric quantile models like the model in (3.16), due to the reason that one usually multiplies a projection matrix to remove the nonparametric component $\alpha_\tau(\mathbf{U}_t)^\top \mathbf{X}_{2t}$ and then fit a linear model; see Fan and Huang (2005) for details.

To estimate both the parameter vector β_τ and the functional coefficients $\alpha_\tau(\cdot)$, Cai and Xiao (2012) proposed a three-stage approach as follows. First, β_τ is regarded as a function of $\mathbf{U}_t$, so that $\beta_\tau(\mathbf{U}_t)$. Thus, the model becomes a pure functional coefficient model and all coefficient functions can be estimated by a nonparametric fitting scheme as in (3.7). Second, an average method is used to obtain a root-n consistent estimator for β_τ. To estimate $\alpha_\tau(\cdot)$, for any root-n consistent estimate $\widehat{\beta}_\tau$ of β_τ, we construct the partial quantile residual $Y_t = Y_t - \widehat{\beta}_\tau^\top \mathbf{X}_{1t}$, and a nonparametric approach can be applied to estimate $\alpha_\tau(\cdot)$ based on the partial quantile residuals. Cai and Xiao (2012) showed that our three-stage nonparametric estimator $\alpha_\tau(\cdot)$ is asymptotically consistent and is indeed "oracle" in the sense that the asymptotic properties of this nonparametric estimator are not affected by knowing β_τ or not. Further, Cai and Xiao (2012) addressed the semiparametric efficiency issue for the data

observed from a martingale difference sequence and proposed a simple efficient estimator to estimate β_τ by using the weighted average approach and choosing the optimal weighting function via minimizing the asymptotic variance. An important statistical question in fitting model (3.16) arises if the coefficient functions $\alpha_\tau(\cdot)$ are actually varying. This amounts to testing whether the coefficient functions are constant or in a certain parametric form. A simple and easily implemented testing procedure is proposed based on the asymptotic theory derived in this paper. A simulation conducted by Cai and Xiao (2012) showed that the proposed estimators by Cai and Xiao (2012) have reasonably good sampling properties and the testing procedure is indeed powerful.

3.5 Empirical Examples

In this section we report a Monte Carlo simulation to examine the finite sample property of the proposed estimator and to further explore the possible nonlinearity feature, heteroscedasticity, and predictability of the exchange rate of the Japanese Yen per US dollar and to identify the factors affecting the house price in Boston. In our computation, we use the Epanechnikov kernel $K(u) = 0.75\left(1-u^2\right) I(|u| \leq 1)$ and construct the point-wise confidence intervals based on the consistent estimate of the asymptotic covariance described in Section 3.3.4 without the bias correction. For a predetermined sequence of h's from a wide range, say from h_a to h_b with an increment h_δ, based on the AIC bandwidth selector described in Section 3.3.3, we compute AIC(h) for each h and choose h_{opt} to minimize AIC(h). Note that the computer codes for the following examples are available upon request.

3.5.1 A Simulated Example

Example 3.1: We consider the following data generating process

$$Y_t = a_1\left(U_t\right) Y_{t-1} + a_2\left(U_t\right) Y_{t-2} + \sigma\left(U_t\right) e_t, \quad t = 1, \cdots, n, \tag{3.17}$$

where $a_1\left(U_t\right) = \sin\left(\sqrt{2}\pi U_t\right), a_2\left(U_t\right) = \cos\left(\sqrt{2}\pi U_t\right)$,

$$\sigma\left(U_t\right) = 3\exp\left(-4\left(U_t-1\right)^2\right) + 2\exp\left(-5\left(U_t-2\right)^2\right),$$

U_t is generated from uniform $(0,3)$ independently, and $e_t \sim N(0,1)$. Then it is easy to show from (3.17) that the quantile regression of Y_t given U_t and X_t is given by

$$q_\tau\left(U_t, Y_{t-1}, Y_{t-2}\right) = a_0\left(U_t\right) + a_1\left(U_t\right) Y_{t-1} + a_2\left(U_t\right) Y_{t-2},$$

where $a_0(U_t) = \Phi^{-1}(\tau)\sigma(U_t)$ and $\Phi^{-1}(\tau)$ is the τ-th quantile of the standard normal. Therefore, only $a_0(\cdot)$ is a function of τ. Note that $a_0(\cdot) = 0$ when $\tau = 0.5$. To assess the performance of finite samples, we compute the mean absolute deviation errors (MADE) for $\widehat{a}_j(\cdot)$, which is defined as

$$\text{MADE}_j = n_0^{-1} \sum_{k=1}^{n_0} |\widehat{a}_j(u_k) - a_j(u_k)|,$$

where $\widehat{a}_j(\cdot)$ is either the local linear or local constant quantile estimate of $a_j(\cdot)$ and $\{z_k = 0.1(k-1) + 0.2 : 1 \leq k \leq n_0 = 27\}$ are the grid points. The Monte Carlo simulation is repeated 500 times for each sample size $n = 200, 500$, and 1000 and for each $\tau = 0.05, 0.50$ and 0.95. We compute the optimal bandwidth for each replication, sample size, and τ. We compute the median and standard deviation (in parentheses) of 500MADE values for each scenario and summarize the results in Table 3.1.

From Table 3.1, we can observe that the MADE values for both the local linear and local constant quantile estimates decrease when n increases for all three values of τ and the local linear estimate outperforms the local constant estimate. This is another example to show that the local linear method is superior over the local constant even in the quantile setting. Also, the performance for the median quantile estimate is slightly better than that for two tails ($\tau = 0.05$ and 0.95). This observation is not surprising because of the sparsity of data in the tailed regions. Moreover, another benefit of using the quantile method is that we can obtain the estimate of $a_0(\cdot)$ (conditional standard deviation) simultaneously with the estimation of $a_1(\cdot)$ and $a_2(\cdot)$ (functions in the conditional mean), which, in contrast, avoids a two-stage approach needed to estimate the variance function in the mean regression; see Fan and Yao (1998) for details. However, it is interesting to see that due to the larger variation, the performance for $a_0(\cdot)$, although it is reasonably good, is not as good as that of $a_1(\cdot)$ and $a_2(\cdot)$. This can be further evidenced from Figure 3.2. The results in this simulated experiment show that the proposed procedure is reliable and they are along the line of our asymptotic theory.

Finally, Figure 3.2 plots the local linear estimates for all three coefficient functions with their true values (solid line): $\sigma(\cdot)$ in Figure 3.2 (a), $a_1(\cdot)$ in Figure 3.2(b), and $a_2(\cdot)$ in Figure 3.2(c), for three quantiles $\tau = 0.05$ (dashed line), 0.50 (dotted line) and 0.95 (dotted-dashed line), for $n = 500$ based on a typical sample which is chosen based on its MADE value equal to the median of the 500 MADE values. The selected optimal bandwidths are $h_{\text{opt}} = 0.10$

Table 3.1: The Median and Standard Deviation of 500 MADE Values

The Local Linear Estimator

	$\tau = 0.05$			$\tau = 0.5$			$\tau = 0.95$		
n	MADE_0	MADE_1	MADE_2	MADE_0	MADE_1	MADE_2	MADE_0	MADE_1	MADE_2
200	0.911	0.186	0.177	0.401	0.092	0.089	0.920	0.187	0.175
	(0.520)	(0.041)	(0.041)	(0.091)	(0.032)	(0.032)	(0.517)	(0.042)	(0.039)
500	0.510	0.085	0.083	0.311	0.055	0.055	0.517	0.085	0.083
	(0.414)	(0.023)	(0.02)	(0.056)	(0.019)	(0.018)	(0.390)	(0.023)	(0.023)
1000	0.419	0.060	0.059	0.311	0.050	0.049	0.416	0.060	0.059
	(0.071)	(0.018)	(0.017)	(0.051)	(0.014)	(0.014)	(0.072)	(0.017)	(0.017)

The Local Linear Estimator

	$\tau = 0.05$			$\tau = 0.5$			$\tau = 0.95$		
n	MADE_0	MADE_1	MADE_2	MADE_0	MADE_1	MADE_2	MADE_0	MADE_1	MADE_2
200	3.753	0.285	0.290	0.501	0.144	0.147	3.763	0.287	0.287
	(2.937)	(0.050)	(0.051)	(0.115)	(0.027)	(0.028)	(3.188)	(0.052)	(0.051)
500	2.201	0.147	0.146	0.355	0.084	0.085	2.223	0.147	0.147
	(3.025)	(0.024)	(0.025)	(0.062)	(0.016)	(0.015)	(3.320)	(0.025)	(0.025)
1000	0.883	0.086	0.086	0.322	0.060	0.061	0.882	0.086	0.087
	(0.462)	(0.015)	(0.014)	(0.054)	(0.012)	(0.011)	(0.427)	(0.015)	(0.015)

for $\tau = 0.05, 0.075$ for $\tau = 0.50$, and 0.10 for $\tau = 0.95$. Note that the estimate of $\sigma(\cdot)$ for $\tau = 0.50$ can not be recovered from the estimate of $a_0(\cdot) = 0$ and it is not presented in Figure 3.2(a). The 95% point-wise confidence intervals without the bias correction are depicted in Figure 3.2 in thick lines for the $\tau = 0.05$ quantile estimate. By the same token, we can compute the point-wise confidence intervals (not shown here) for the rest. Basically, all confidence intervals cover the true values. Also, we can see that the confidence interval for $\widehat{a}_0(\cdot)$ is wider than that for $\widehat{a}_1(\cdot)$ and $\widehat{a}_2(\cdot)$ due to the larger variation. Similar plots are obtained (not shown here) for the local constant estimates due to the space limitations. Overall, the proposed modeling procedure performs fairly well.

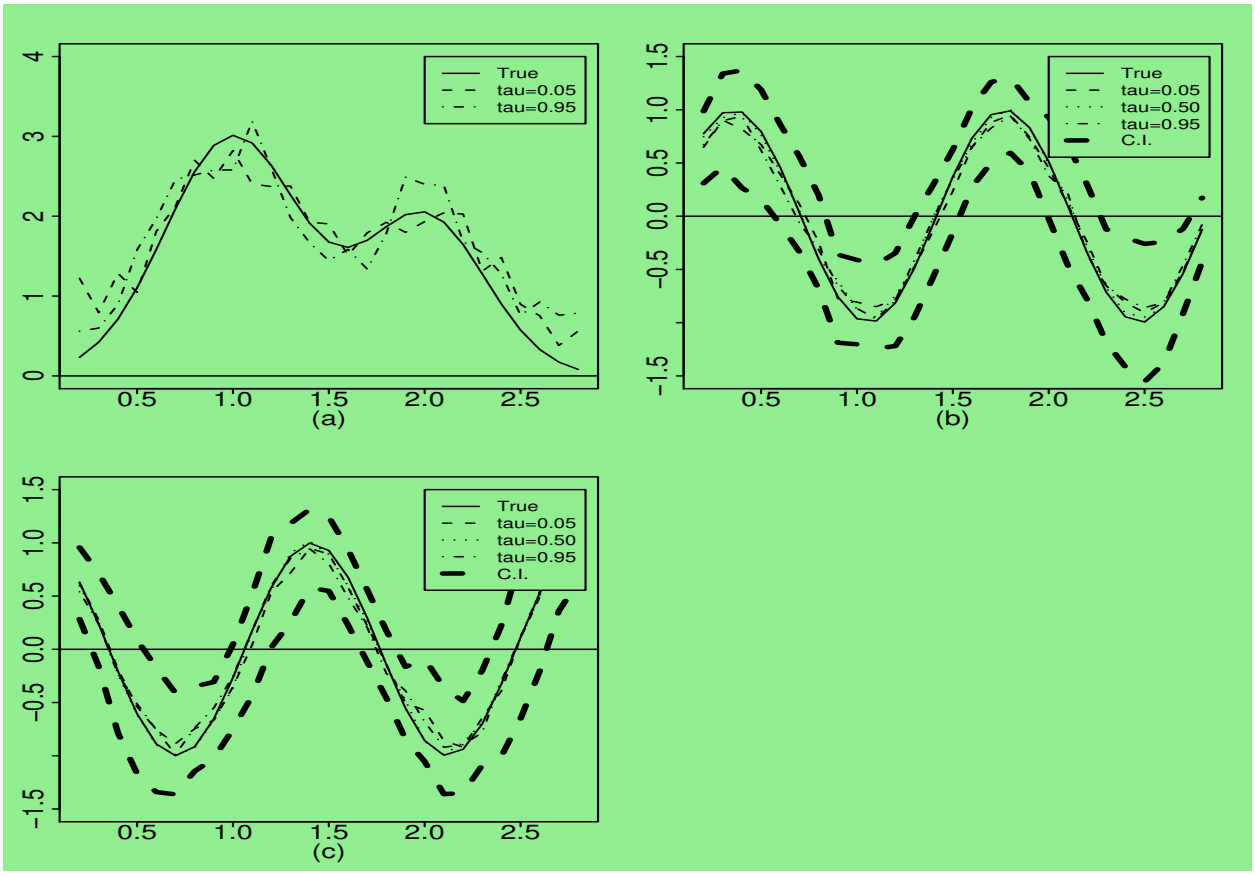

Figure 3.2: *Simulated Example*: The plots of the estimated coefficient functions for three quantiles $\tau = 0.05$ (dashed line), $\tau = 0.50$ (dotted line), and $\tau = 0.95$ (dot-dashed line) with their true functions (solid line): $\sigma(u)$ versus u in (a), $a_1(u)$ versus u in (b), and $a_2(u)$ versus u in (c), together with the 95% point-wise confidence interval (thick line) with the bias ignored for the $\tau = 0.5$ quantile estimate.

3.5.2 Real Data Examples

Example 3.2: (*Boston House Price Data*) We analyze a subset of the Boston house price data (available at http://lib.stat.cmu.edu/datasets/boston) of Harrison and Rubinfeld (1978). This dataset consists of 14 variables collected on each of 506 different houses from a variety of locations. The dependent variable is Y, the median value of owner-occupied homes in $\$1,000$'s (house price); some major factors affecting the house prices used are: proportion of population of lower educational status (i.e. proportion of adults with high school education and proportion of male workers classified as labors), denoted by U, the average number of rooms per house in the area, denoted by X_1, the per capita crime rate by town, denoted by X_2, the full property tax rate per $\$10,000$, denoted by X_3, and the pupil/teacher ratio by town school district, denoted by X_4. For the complete description of all 14 variables,

see Harrison and Rubinfeld (1978). Gilley et al. (1996) provided corrections and examined censoring. So far, there have been several papers devoted to the analysis of this dataset. For example, Breiman and Friedman (1985), CChaudhuri et al. (1997), and Opsomer and Ruppert (1998) used four covariates: X_1, X_3, X_4 and U or their transformations to fit the data through a mean additive regression model whereas Yu and Lu (2004) employed the additive quantile technique to analyze the data. Further, Pace and Gilley (1997) added the geo-referencing factor to improve estimation by a spatial approach. Also, Şentürk and Müller (2006) studied the correlation between the house price Y and the crime rate X_2 adjusted by the confounding variable U through a varying coefficient model and they concluded that the expected effect of increasing crime rate on declining house prices seems to be only observed for lower educational status neighborhoods in Boston. Some existing analyses; see, e.g., Breiman and Friedman (1985) and Yu and Lu (2004), in both mean and quantile regressions concluded that most of the variation seen in housing prices in the restricted data set can be explained by two major variables: X_1 and U. Indeed, the correlation coefficients between Y and U and X_1 are -0.7377 and 0.6954 respectively. The scatter plots of Y versus U and X_1 are displayed in Figures 3.3(a) and 3.3(b), respectively. The interesting features of this data set are that the response variable is the median price of a home in a given area and the distributions of Y and the major covariate U are left skewed (the density estimates are not presented). Therefore, quantile methods are particularly well suited to the analysis of this dataset. Finally, it is surprising that all the existing nonparametric models aforementioned above did not include the crime rate X_2, which may be an important factor affecting the housing price, and did not consider the interaction terms such as U and X_2.

Based on the above discussions, it concludes that the model studied in this chapter might be well suitable to the analysis of this dataset. Therefore, we analyze this dataset by the following quantile smooth coefficient mode[3]

$$q_\tau\left(U_t, \mathbf{X}_t\right) = a_{0,\tau}\left(U_t\right) + a_{1,\tau}\left(U_t\right) X_{t1} + a_{2,\tau}\left(U_t\right) X_{t2}^*, \quad 1 \le t \le n = 506, \tag{3.18}$$

where $X_{t2}^* = \log\left(X_{t2}\right)$. The reason for using the logarithm of X_{t2} in (3.18), instead of X_{t2} itself, is that the correlation between Y_t and X_{t2}^* (the correlation coefficient is -0.4543) is slightly stronger than that for Y_t and $X_{t2}(-0.3883)$, which can be witnessed as well from

[3]We do not include the other variables such as X_3 and X_4 in model (3.18), since we found that the coefficient functions for these variables seem to be constant. Therefore, a semiparametric model would be appropriate if the model includes these variables. It of course deserves a further investigation.

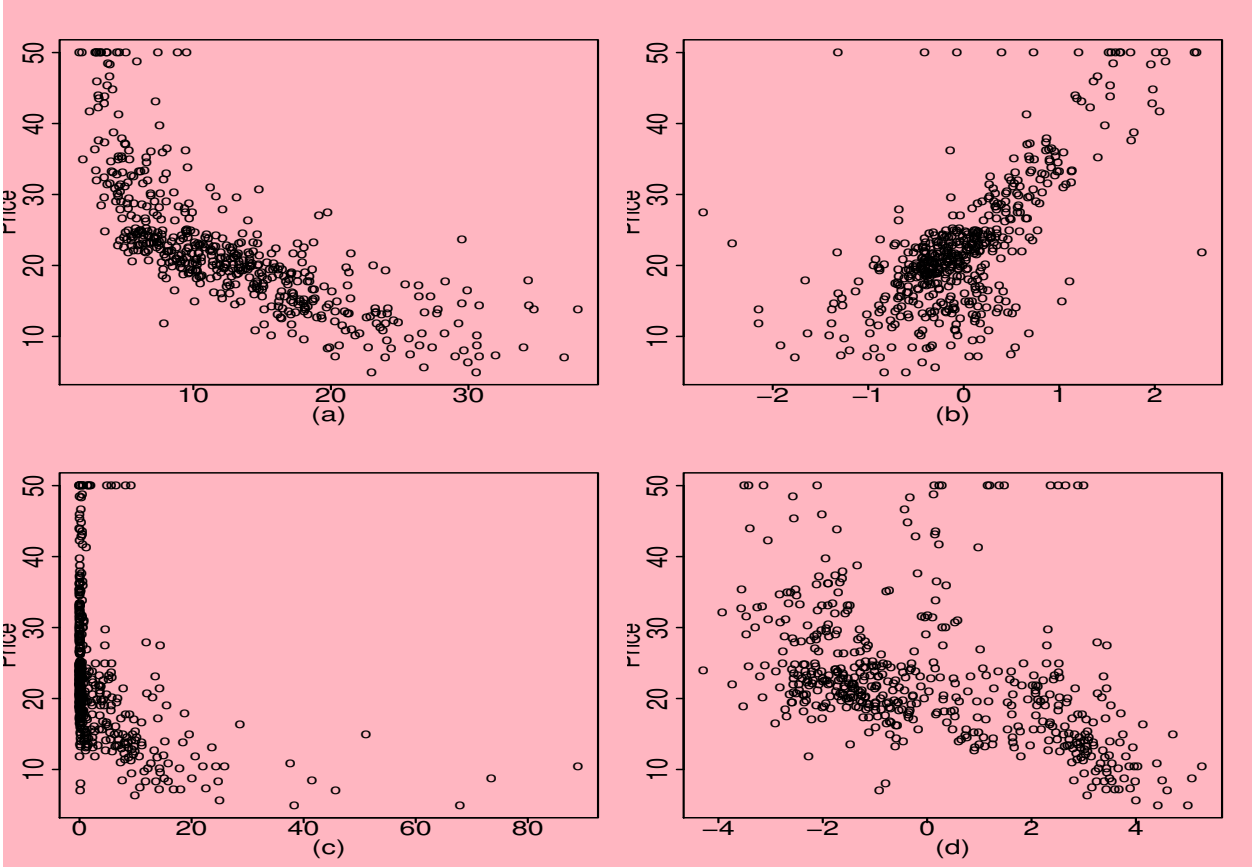

Figure 3.3: *Boston Housing Price Data*: Displayed in (a)-(d) are the scatter plots of the house price versus the covariates U, X_1, X_2 and $\log(X_2)$, respectively.

Figures 3.3(c) and 3.3(d). In the model fitting, covariates X_1 and X_2 are centralized. For the purpose of comparison, we also consider the following functional coefficient model in the mean regression

$$Y_t = a_0(U_t) + a_1(U_t) X_{t1} + a_2(U_t) X_{t2}^* + e_t, \tag{3.19}$$

and we employ the local linear fitting technique to estimate the coefficient functions $\{a_j(\cdot)\}$, denoted by $\{\widehat{a}_j(\cdot)\}$; see Cai et al. (2000) for details.

The coefficient functions are estimated through the local linear quantile approach by using the bandwidth selector described in Section 3.3.3. The selected optimal bandwidths are $h_{\text{opt}} = 2.0$ for $\tau = 0.05, 1.5$ for $\tau = 0.50$, and 3.5 for $\tau = 0.95$. Figures 3.4(e), 3.4(f) and 3.4(g) present the estimated coefficient functions $\widehat{a}_{0,\tau}(\cdot), \widehat{a}_{1,\tau}(\cdot)$, and $\widehat{a}_{2,\tau}(\cdot)$ respectively, for three quantiles $\tau = 0.05$ (solid line), 0.50 (dashed line) and 0.95 (dotted line), together with the estimates $\{\widehat{a}_j(\cdot)\}$ from the mean regression model (dot-dashed line). Also, the 95% point-wise confidence intervals for the median estimate are displayed by the thick dashed

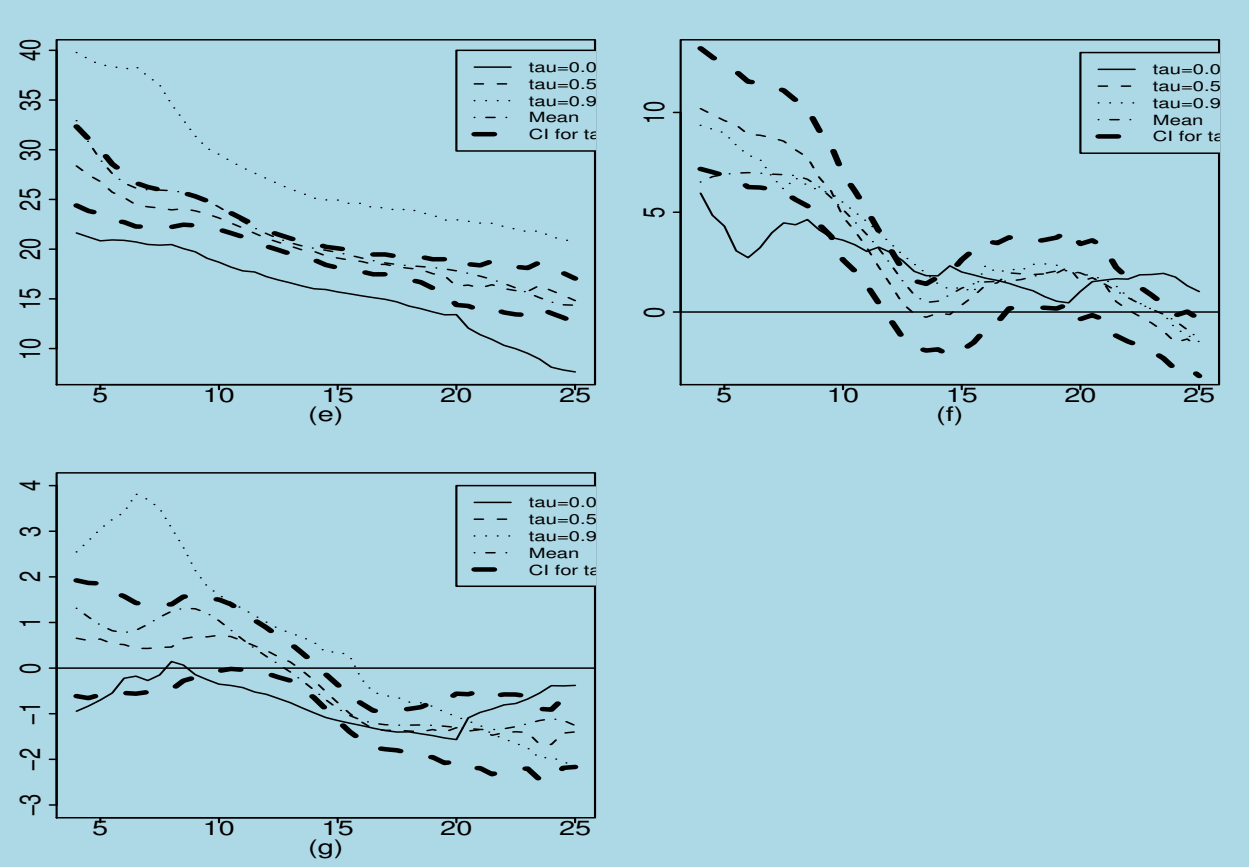

Figure 3.4: *Boston Housing Price Data*: The plots of the estimated coefficient functions for three quantiles $\tau = 0.05$ (solid line), $\tau = 0.50$ (dashed line), and $\tau = 0.95$ (dotted line), and the mean regression (dot-dashed line): $\widehat{a}_{0,\tau}(u)$ and $\widehat{a}_0(u)$ versus u in (e), $\widehat{a}_{1,\tau}(u)$ and $\widehat{a}_1(u)$ versus u in (f), and $\widehat{a}_{2,\tau}(u)$ and $\widehat{a}_2(u)$ versus u in (g). The thick dashed lines indicate the 95% point-wise confidence interval for the median estimate with the bias ignored.

lines without the bias correction. First, from these three figures, one can see that the median estimates are quite close to the mean estimates and the estimates based on the mean regression are always within the 95% confidence interval of the median estimates. It can be concluded that the distribution of the measurement error e_t in (3.19) might be symmetric and $\widehat{a}_{j,0.5}(\cdot)$ in (3.18) is almost same as $\widehat{a}_j(\cdot)$ in (3.19). Also, one can observe from Figure 3.4(e) that three quantile curves are parallel, which implies that the intercept in $\widehat{a}_{0,\tau}(\cdot)$ depends on τ, and they decrease exponentially, which can support that the logarithm transformation may be needed as argued in Yu and Lu (2004). More importantly, one can observe from Figures 3.4(f) and 3.4(g) that three quantile estimated coefficient curves are intersect. This reveals that the structure of quantiles is complex and the lower and upper quantiles have different behaviors and the heteroscedasticity might exist. But unfortunately,

this phenomenon was not observed in any previous analyses in the aforementioned papers.

From Figure 3.4(f), first, we can observe that $\widehat{a}_{1,0.50}(\cdot)$ and $\widehat{a}_{1,0.95}(\cdot)$ are almost same but $\widehat{a}_{1,0.05}(\cdot)$ is different. Secondly, we can see that the correlation between the house price and the number of rooms per house is almost positive except for houses with the median price and/or higher than ($\tau = 0.50$ and 0.95) in very low educational status neighborhoods ($U > 23$). Thirdly, for the low price houses ($\tau = 0.05$), the correlation is always positive and it deceases when U is between 0 and 14 and then keeps almost constant afterwards. This implies that the expected effect of increasing the number of rooms can make the house price slightly higher in any low educational status neighborhoods but much higher in relatively high educational status neighborhoods. Finally, for the median and/or higher price houses, the correlation deceases when U is between 0 and 14 and then keeps almost constant until U up to 20 and finally deceases again afterwards, and it becomes negative for U larger than 23 . This means that the number of room has a positive effect on the median and/or higher price houses in relatively high and low educational status neighborhoods but increasing the number of rooms might not increase the house price in very low educational status neighborhoods. In other words, it is very difficult to sell high price houses with high number of rooms at a reasonable price in very low educational status neighborhoods.

Finally, from Figure 3.4(g), first, one can conclude that the overall trend for all curves is decreasing with $\widehat{a}_{3,0.95}(\cdot)$ deceasing faster than the others, and that $\widehat{a}_{3,0.05}(\cdot)$ and $\widehat{a}_{3,0.50}(\cdot)$ tend to be constant for U larger than 16. Secondly, the correlation between the housing prices ($\tau = 0.50$ and 0.95) and the crime rate seems to be positive for smaller U values (about $U \leq 13$) and becomes negative afterwards. This positive correlation between the housing prices ($\tau = 0.50$ and 0.95) and the crime rate for relatively high educational status neighborhoods seems against intuitive. However, the reason for this positive correlation is the existence of high educational status neighborhoods close to central Boston where high house prices and crime rate occur simultaneously. Therefore, the expected effect of increasing crime rate on declining house prices for $\tau = 0.50$ and 0.95 seems to be observed only for lower educational status neighborhoods in Boston. Finally, it can be seen that the correlation between the housing prices for $\tau = 0.05$ and the crime rate is almost negative although the degree depends on the value of U. This implies that increasing crime rate slightly decreases relatively the house prices for the cheap houses ($\tau = 0.05$).

In summary, it concludes that there is a nonlinear relationship between the conditional

quantiles of the housing price and the affecting factors. It seems that the factors U, X_1 and X_2 do have different effects on the different quantiles of the conditional distribution of the housing price. Overall, the housing price and the proportion of population of lower educational status have a strong negative correlation, and the number of rooms has a mostly positive effect on the housing price whereas the crime rate has the most negative effect on the housing price. In particular, by using the proportion of population of lower educational status U as the confounding variable, we demonstrate the substantial benefits obtained by characterizing the affecting factors X_1 and X_2 on the housing price based on the neighborhoods.

Example 3.3: (*Exchange Rate Data*) This example concerns the closing bid prices of the Japanese Yen (JPY) in terms of US dollar. There is a vast amount of literature devoted to the study of the exchange rate time series; see Sercu et al. (2006) and references therein for details. Here we use the proposed model and its modeling approaches to explore the possible nonlinearity feature, heteroscedasticity, and predictability of the exchange rate series. The data is a weekly series from January 1, 1974 to December 31, 2003. The daily noon buying rates in New York City certified by the Federal Reserve Bank of New York for customs and cable transfers purposes were obtained from the Chicago Federal Reserve Board (http://www.chicagofed.org). The weekly series is generated by selecting the Wednesdays series (if a Wednesday is a holiday then the following Thursday is used), which has 1566 observations. The use of weekly data avoids the so-called weekend effect as well as other biases associated with non-trading, bid-ask spread, asynchronous rates and so on, which are often present in higher frequency data. The previous analysis of this "particularly difficult" data set can be found in Gallant et al. (1991), Fan et al. (2003), and Hong and Lee (2003), and the references within. We model the return series $Y_t = 100 \log(\xi_t/\xi_{t-1})$, plotted in Figure 3.5(a), using the techniques developed in this chapter, where ξ_t is an exchange rate level on the t-th week. Typically the classical financial theory would treat $\{Y_t\}$ as a martingale difference process. Therefore, Y_t would be unpredictable. But this assumption was strongly rejected by Hong and Lee (2003) by examining five major currencies and applying several testing procedures. Note that the return series $\{Y_t\}$ has 1565 observations. Figure 3.5(b) shows that there exists almost no significant autocorrelation in $\{Y_t\}$, which also was confirmed by Tsay (2002) and Hong and Lee (2003) by using several statistical testing procedures.

Based on the evidence from Fan et al. (2003) and Hong and Lee (2003), the exchange

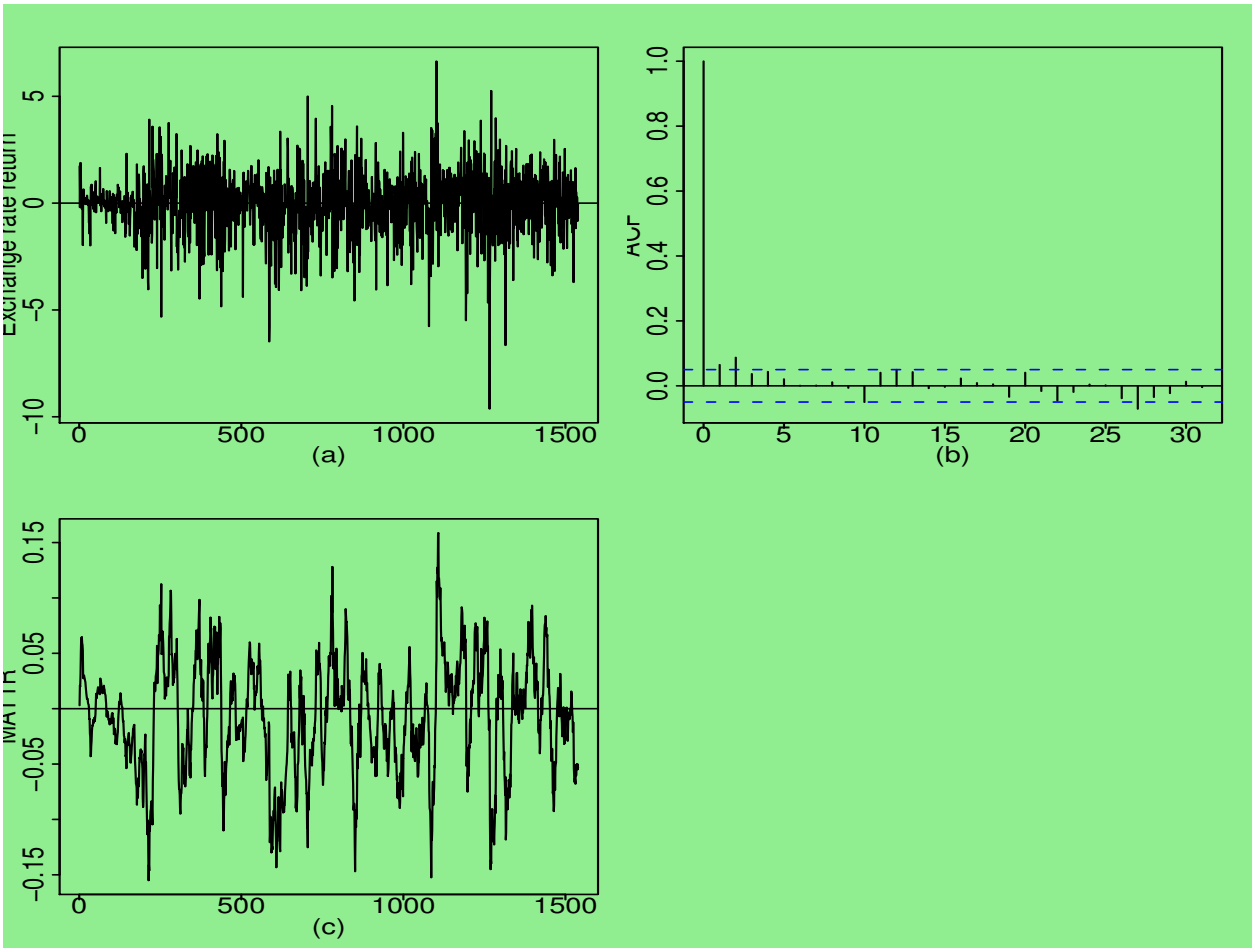

Figure 3.5: *Exchange Rate Series:* (a) Japanese-dollar exchange rate return series $\{Y_t\}$; (b) autocorrelation function of $\{Y_t\}$; (c) moving average trading technique rule.

rate series is predictable by using the functional coefficient autoregressive model

$$Y_t = a_0\left(U_t\right) + \sum_{j=1}^{d} a_j\left(U_t\right) Y_{t-j} + \sigma_t e_t, \tag{3.20}$$

where U_t is the smooth variable defined later and σ_t is a function of U_t and the lagged variables. If $\{U_t\}$ is observable, $a_j(\cdot)$ can be estimated by a local linear fitting; see Cai et al. (2000) for details, denoted by $\widehat{a}_j(\cdot)$. Here, σ_t is the stochastic volatility which may depend on U_t and the lagged variables $\{Y_{t-j}\}$. Now the question is how to choose U_t. Usually, U_t can be chosen based on the knowledge of data or economic theory. However, if no prior information is available, U_t may be chosen as a function of explanatory vector $\{\xi_{t-j}\}$ or through the use of data-driven methods such as AIC or cross-validation. Moreover, Fan et al. (2003) proposed a data-driven method to the choice of U_t by a linear combination of $\{\xi_{t-j}\}$ and the lagged variables $\{Y_{t-j}\}$. By following the analysis of Fan et al. (2003) and Hong and Lee (2003),

we choose the smooth variable U_t as a moving average technical trading rule (MATTR) in finance so that the autoregressive coefficients vary with investment positions. U_t is defined as $U_t = \xi_{t-1}/M_t - 1$, where $M_t = \sum_{j=1}^{L} \xi_{t-j}/L$, which is the moving average and can be regarded as a proxy for the trend at the time $t-1$. Similar to Hong and Lee (2003), We choose $L = 26$ (half a year). $U_t + 1$ is the ratio of the exchange rate at the time $t-1$ to the average rate of the most recent L periods of exchange rates at time $t-1$. The time series plot of $\{U_t\}$ is given in Figure 3.5(c). As pointed out by Hong and Lee (2003), U_t is expected to reveal some useful information on the direction of changes. The MATTR signals 1 (the position to buy JPY) when $U_t > 0$ and -1 (the position to sell JPY) when $U_t < 0$. For the detailed discussions of the MATTR, see (for example) the papers by LeBaron (1999), Hong and Lee (2003), Fan et al. (2003), and the reference therein. Note that model (3.18) was studied by Fan et al. (2003) for the daily data and Hong and Lee (2003) for the weekly data under the homogenous assumption (assume that $\sigma_t = \sigma$) based on the least square theory. In particular, Hong and Lee (2003) provided some empirical evidences to conclude that model (3.20) outperforms the martingale model and autoregressive models.

We analyze this exchange rate series by using the smooth coefficient model under the quantile regression framework with only two lagged variables[4] as follows

$$q_\tau\left(U_t, Y_{t-1}, Y_{t-2}\right) = a_{0,\tau}\left(U_t\right) + a_{1,\tau}\left(U_t\right) Y_{t-1} + a_{2,\tau}\left(U_t\right) Y_{t-2}. \tag{3.21}$$

The first 1540 observations of $\{Y_t\}$ are used for estimation and the last 25 observations are left for prediction. The coefficient functions $\{a_{j,\tau}(\cdot)\}$ are estimated through the local linear quantile approach, denoted by $\{\widehat{a}_{j,\tau}(\cdot)\}$. The previous analysis of this "particularly difficult" data set can be found in optimal bandwidths are $h_{\text{opt}} = 0.03$ for $\tau = 0.05, 0.025$ for $\tau = 0.50$, and 0.03 for $\tau = 0.95$. Figures 3.6(d) - 3.6(g) depict the estimated coefficient functions $\widehat{a}_{0,\tau}(\cdot), \widehat{a}_{1,\tau}(\cdot)$, and $\widehat{a}_{2,\tau}(\cdot)$ respectively, for three quantiles $\tau = 0.05$ (solid line), 0.50 (dashed line) and 0.95 (dotted line), together with the estimates $\{\widehat{a}_j(\cdot)\}$ (dot-dashed line) from the mean regression model in (3.20). Also, the 95% point-wise confidence intervals for the median estimate are displayed by the thick dashed lines without the bias correction.

First, from Figures 3.6(d), 3.6(f) and 3.6(g), we see clearly that the median estimates $\widehat{a}_{j,0.50}(\cdot)$ in (3.21) are almost parallel with or close to the mean estimates $\widehat{a}_j(\cdot)$ in (3.20) and

[4]We also considered the models with more than two lagged variables and we found that the conclusions are similar and not reported here.

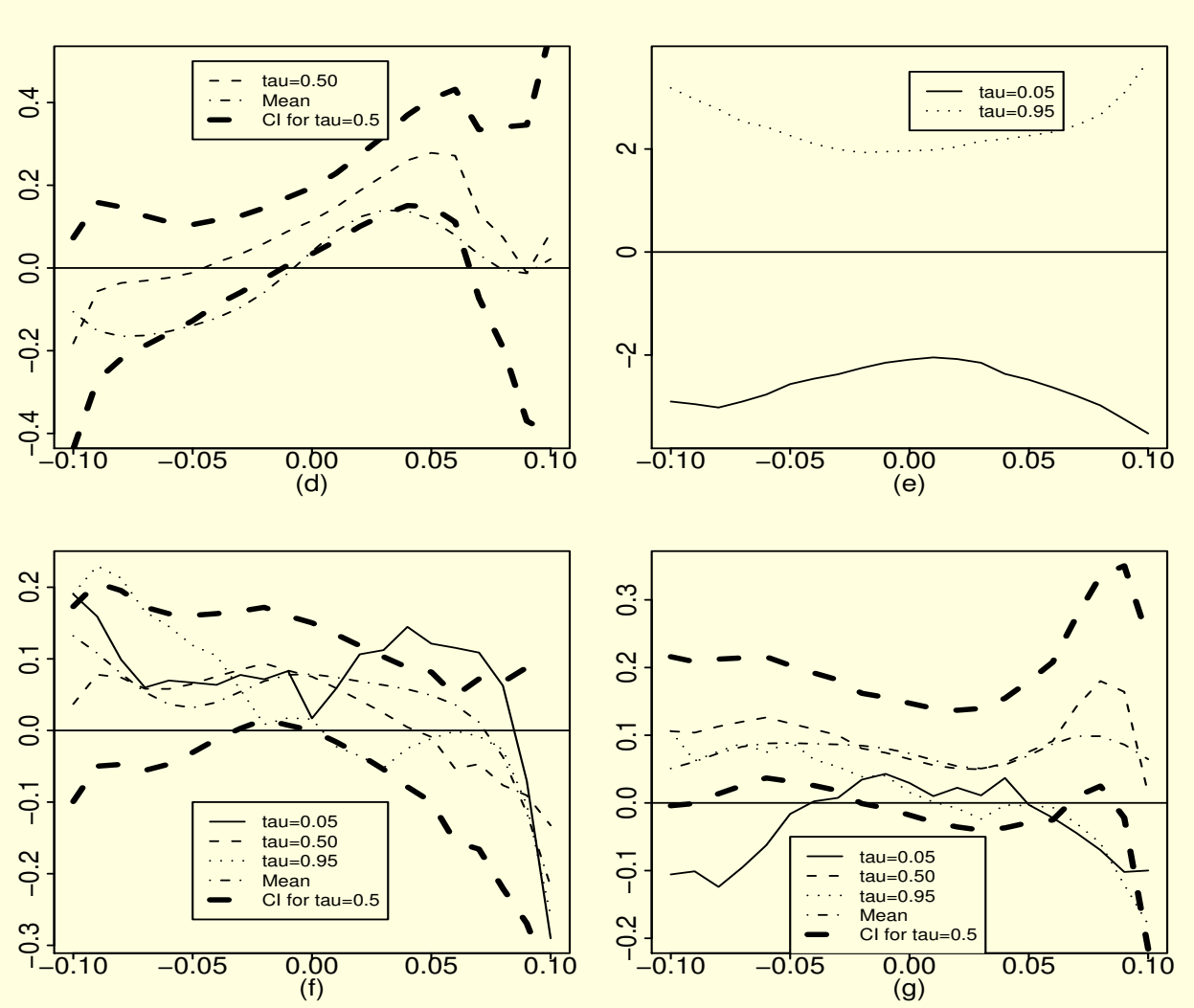

Figure 3.6: *Exchange Rate Series:* The plots of the estimated coefficient functions for three quantiles $\tau = 0.05$ (solid line), $\tau = 0.50$ (dashed line), and $\tau = 0.95$ (dotted line), and the mean regression (dot-dashed line): $\widehat{a}_{0,0.50}(u)$ and $\widehat{a}_0(u)$ versus u in (d), $\widehat{a}_{0,0.05}(u)$ and $\widehat{a}_{0,0.95}(u)$ versus u in (e), $\widehat{a}_{1,\tau}(u)$ and $\widehat{a}_1(u)$ versus u in (f), and $\widehat{a}_{2,\tau}(u)$ and $\widehat{a}_2(u)$ versus u in (g). The thick dashed lines indicate the 95% point-wise confidence interval for the median estimate with the bias ignored.

the mean estimates are almost within the 95% confidence interval of the median estimates. Secondly, $\widehat{a}_{0,0.50}(\cdot)$ in Figure 3.6(d) shows a nonlinear pattern (increasing and then decreasing) and $\widehat{a}_{0,0.05}(\cdot)$ and $\widehat{a}_{0,0.95}(\cdot)$ in Figure 3.6(e) exhibit nonlinearly (slightly U-shape) and symmetrically. More importantly, one can observe from Figures 3.6(f) and 3.6(g) that the lower and upper quantile estimated coefficient curves are intersect and they behave slightly differently. Particularly, from Figure 3.6(g), we observe that $\widehat{a}_{2,0.05}(U_t)$ seems to be nonlinear but $\widehat{a}_{2,0.95}(U_t)$ looks like constant when $U_t < 0.06$, and both $\widehat{a}_{2,0.05}(U_t)$ and $\widehat{a}_{2,0.95}(U_t)$

decrease when $U_t > 0.06$. One might conclude that the distribution of the measurement error e_t in (3.20) might not be symmetric about 0 and there exists a nonlinearity in $a_{j,\tau}(\cdot)$. This supports the nonlinearity test of Hong and Lee (2003). Also, our findings lead to the conclusions that the quantile has a complex structure and the heteroscedasticity exists. This observation supports the existing conclusion in literature that the GARCH (generalized ARCH) effects occur in the exchange rate time series; see Engle et al. (1990) and Tsay (2002).

Finally, we consider the post-sample forecasting for the last 25 observations based on the local linear quantile estimators which are computed by using the same bandwidths as those used in the model fitting. The 95% nonparametric prediction interval is constructed as $(\widehat{q}_{0.025}(\cdot), \widehat{q}_{0.975}(\cdot))$ and the prediction results are reported in Table 3.2, which shows that 24 out of 25 (96%) predictive intervals contain the corresponding true values. The average length of the intervals is 5.77, which is about 35.5% of the range of the data. Therefore, we can conclude that under the dynamic smooth coefficient quantile regression model assumption, the prediction intervals based on the proposed method work reasonably well.

3.6 Mathematical Derivations

3.6.1 Proofs of Main Results

In this section, we give the derivations of the theorems and present certain lemmas with their detailed proofs relegated to Section 3.6. First, we need the following two lemmas.

Lemma 3.1: Let $V_n(\Delta)$ be a vector function that satisfies
(i) $-\Delta^\top V_n(\lambda\Delta) \geq -\Delta^\top V_n(\Delta)$ for $\lambda \geq 1$
and
(ii) $\sup_{\|\Delta\|\leq M} \|V_n(\Delta) + \mathbf{D}\Delta - \mathbf{A}_n\| = o_p(1)$, where $\|\mathbf{A}_n\| = o_p(1), 0 < M < \infty$, and $\mathbf{D}$ is a positive-definite matrix. Suppose that Δ_n is a vector such that $\|V_n(\Delta_n)\| = o_p(1)$, then, we have

$$(1)\,\|\Delta_n\| = o_p(1) \quad \text{and} \quad (2)\Delta_n = \mathbf{D}^{-1}\mathbf{A}_n + o_p(1).$$

Proof : The proof follows from Jurekova (1977) and Koenker and Zhao (1996). □

Table 3.2: The Post-Sample Predictive Intervals For Exchange Rate Data

Observation	True Value	Prediction Interval
Y_{1541}	0.3920	$(-2.891, 2.412)$
Y_{1542}	0.5090	$(-3.099, 2.405)$
Y_{1543}	1.5490	$(-2.943, 2.446)$
Y_{1544}	-0.121	$(-2.684, 2.525)$
Y_{1545}	-0.991	$(-2.677, 2.530)$
Y_{1546}	-0.646	$(-3.110, 2.401)$
Y_{1547}	-0.354	$(-3.178, 2.365)$
Y_{1548}	-1.393	$(-3.083, 2.372)$
Y_{1549}	0.9970	$(-3.110, 2.230)$
Y_{1550}	-0.916	$(-3.033, 2.431)$
Y_{1551}	$\mathbf{-3.707}$	$(\mathbf{-3.021, 2.286})$
Y_{1552}	-0.919	$(-3.841, 2.094)$
Y_{1553}	-0.901	$(-3.603, 2.770)$
Y_{1554}	0.0710	$(-3.583, 2.821)$
Y_{1555}	-0.497	$(-3.351, 2.899)$
Y_{1556}	-0.648	$(-3.436, 2.783)$
Y_{1557}	1.6480	$(-3.524, 2.866)$
Y_{1558}	-1.184	$(-3.121, 2.810)$
Y_{1559}	0.5300	$(-3.529, 2.531)$
Y_{1560}	0.1070	$(-3.222, 2.648)$
Y_{1561}	-0.804	$(-3.294, 2.651)$
Y_{1562}	0.2740	$(-3.419, 2.534)$
Y_{1563}	-0.847	$(-3.242, 2.640)$
Y_{1564}	-0.060	$(-3.426, 2.532)$
Y_{1565}	-0.088	$(-3.300, 2.576)$

Lemma 3.2: Let $\widehat{\boldsymbol{\beta}}$ be the minimizer of the function

$$\sum_{t=1}^{n} w_t \rho_\tau \left(y_t - \mathbf{X}_t^\top \boldsymbol{\beta} \right),$$

where $w_t > 0$. Then,

$$\left\| \Sigma_{t=1}^{n} w_t \mathbf{X}_t \psi_\tau \left(y_t - \mathbf{X}_t^\top \widehat{\boldsymbol{\beta}} \right) \right\| \leq \dim(\mathbf{X}) \max_{t \leq n} \left\| w_t \mathbf{X}_t \right\|.$$

Proof : The proof follows from Ruppert and Carroll (1980). From the definition of $\boldsymbol{\theta}$, we

have

$$\boldsymbol{\beta} = \begin{pmatrix} \mathbf{a}(u_0) \\ \mathbf{a}^\top(u_0) \end{pmatrix} + a_n \mathbf{H}^{-1}\boldsymbol{\theta},$$

where a_n is defined in (3.14). Then, $Y_t - \sum_{j=0}^{q} \mathbf{X}_t^\top \boldsymbol{\beta}_j (U_t - u_0)^j = Y_t^* - a_n \boldsymbol{\theta}^\top \mathbf{X}_t^*$. Therefore,

$$\widehat{\boldsymbol{\theta}} = \text{argmin} \sum_{t=1}^{n} \rho_\tau \left[Y_t^* - a_n \boldsymbol{\theta}^\top \mathbf{X}_t^* \right] K(U_{th}) \equiv \text{argmin} G(\boldsymbol{\theta}).$$

Now, define $V_n(\boldsymbol{\theta})$ as

$$V_n(\boldsymbol{\theta}) = a_n \sum_{t=1}^{n} \psi_\tau \left[Y_t^* - a_n \boldsymbol{\theta}^\top \mathbf{X}_t^* \right] \mathbf{X}_t^* K(U_{th}).$$

To establish the asymptotic properties of $\widehat{\boldsymbol{\theta}}$, in the next three lemmas, we show that $V_n(\boldsymbol{\theta})$ satisfies Lemma 3.1 so that we can derive the local Bahadur representation for $\widehat{\boldsymbol{\theta}}$. The results are stated here and their detailed proofs are given in Section 3.6 For the notational convenience define $A_m = \{\boldsymbol{\theta} : \|\boldsymbol{\theta}\| \leq M\}$ for some $0 < M < \infty$. □

Lemma 3.3: Under the assumptions in Theorem 3.1, we have

$$\sup_{\boldsymbol{\theta} \in A_m} \|V_n(\theta) - V_n(0) - \mathbb{E}[V_n(\theta) - V_n(0)]\| = o_p(1).$$

Lemma 3.4: Under then assumptions in Theorem 3.1, we have

$$\sup_{\boldsymbol{\theta} \in A_m} \|\mathbb{E}[V_n(\boldsymbol{\theta}) - V_n(0)] + f(u_0) \Omega_1^*(u_0) \boldsymbol{\theta}\| = o(1).$$

Lemma 3.5: Let $\mathbf{Z}_t = \psi_\tau(Y_t^*) \mathbf{X}_t^* K(U_{th})$. Under the assumptions in Theorem 3.1, we have

$$\mathbb{E}[\mathbf{Z}_1] = \frac{h^3 f(u_0)}{2} \begin{pmatrix} \mu_2(K) \Omega^*(u_0) \mathbf{a}''(u_0) \\ \mathbf{0} \end{pmatrix} \{1 + o(1)\}$$

and

$$\text{Var}[\mathbf{Z}_1] = h\tau(1-\tau) f(u_0) \Omega_1(u_0) \{1 + o(1)\},$$

where

$$\Omega_1(u_0) = \begin{pmatrix} \nu_0(K) \Omega(u_0) & \mathbf{0} \\ \mathbf{0} & \nu_2 \Omega(u_0) \end{pmatrix}.$$

Further,

$$\text{Var}[V_n(0)] \quad \to \quad \tau(1-\tau) f(u_0) \Omega_1(u_0).$$

Therefore, $\|V_n(0)\| = o_p(1)$.

Now, we can embrace on the proofs of the theorems.

Proof of Theorem 3.1. By Lemmas 3.5, 3.3, and 3.4, $V_n(\boldsymbol{\theta})$ satisfies the condition (ii) of Lemma 3.1; that is, $\|\mathbf{A}_n\| = o_p(1)$ and $\sup_{\boldsymbol{\theta}\in A_m} \|V_n(\boldsymbol{\theta}) + \mathbf{D}\boldsymbol{\theta} - \mathbf{A}_n\| = o_p(1)$ with $\mathbf{D} = f_u(u_0)\,\Omega_1^*(u_0)$ and $\mathbf{A}_n = V_n(0)$. It follows Lemma 3.2 that $\left\|V_n(\widehat{\boldsymbol{\theta}})\right\| = o_p(1)$, where $\widehat{\boldsymbol{\theta}}$ is the minimizer of $G(\boldsymbol{\theta})$. Finally, since $\psi_\tau(x)$ is an increasing function of x, then,

$$\begin{aligned}-\boldsymbol{\theta}^\top V_n(\lambda\boldsymbol{\theta}) &= a_n \sum_{t=1}^{n} \left(-\boldsymbol{\theta}^\top\right)\left(\psi_\tau\left(Y_t^* - \lambda a_n \boldsymbol{\theta}^\top \mathbf{X}_t^*\right)\mathbf{X}_t^* K\left(U_{th}\right)\right.\\ &= a_n \sum_{t=1}^{n} \psi_\tau\left[Y_t^* + \lambda a_n\left(-\boldsymbol{\theta}^\top \mathbf{X}_t^*\right)\right]\left(-\boldsymbol{\theta}^\top \mathbf{X}_t^*\right) K\left(U_{th}\right)\end{aligned}$$

is an increasing function of λ. Thus, the condition (i) of Lemma 3.1 is satisfied. Therefore, it follows that

$$\widehat{\boldsymbol{\theta}} = \mathbf{D}^{-1}\mathbf{A}_n + o_p(1) = \frac{\left(\Omega_1^*\right)^{-1}}{\sqrt{nh} f_u(u_0)} \sum_{t=1}^{n} \psi_\tau\left(Y_t^*\right)\mathbf{X}_t^* K\left(U_{th}\right) + o_p(1). \tag{3.22}$$

This proves (3.9). □

Proof of Theorem 3.2. Let $\varepsilon_t = \psi_\tau\left(Y_t - \mathbf{X}_t^\top \mathbf{a}\left(U_t\right)\right)$. Then, $\mathbb{E}\left(\varepsilon_t\right) = 0$ and $\mathrm{Var}\left(\varepsilon_t\right) = \tau(1-\tau)$. From (3.22),

$$\widehat{\boldsymbol{\theta}} \approx \frac{\left(\Omega_1^*\right)^{-1}}{\sqrt{nh} f_u(u_0)} \sum_{t=1}^{n} \left[\psi_\tau\left(Y_t^*\right) - \varepsilon_t\right]\mathbf{X}_t^* K\left(U_{th}\right) + \frac{\left(\Omega_1^*\right)^{-1}}{\sqrt{nh} f_u(u_0)} \sum_{t=1}^{n} \varepsilon_t \mathbf{X}_t^* K\left(U_{th}\right) \equiv \mathbf{B}_n + \boldsymbol{\xi}_n$$

Similar to the proof of Theorem 2 in Cai et al. (2000), by using the small-block and large-block technique and the CramÃ©r-Wold device, one can show that

$$\boldsymbol{\xi}_n \quad \to \quad N\left(\mathbf{0}, \boldsymbol{\Sigma}\left(u_0\right)\right). \tag{3.23}$$

By the stationarity and Lemma 3.5,

$$\mathbb{E}\left[\mathbf{B}_n\right] = \frac{\left(\Omega_1^*\right)^{-1}}{\sqrt{nh} f_u(u_0)} n\mathbb{E}\left[\mathbf{Z}_1\right]\{1 + o(1)\} = a_n^{-1}\frac{h^2}{2}\begin{pmatrix} \mathbf{a}''(u_0)\,\mu_2(K) \\ \mathbf{0} \end{pmatrix}\{1 + o(1)\}. \tag{3.24}$$

Since $\psi_\tau\left(Y_t^*\right) - \varepsilon_t = I\left(Y_t \le \mathbf{X}_t^\top \mathbf{a}\left(U_t\right)\right) - I\left(Y_t \le \mathbf{X}_t^\top\left(\mathbf{a}(u_0) + \mathbf{a}^\top(u_0)\left(U_t - u_0\right)\right)\right)$, then,

$$\left[\psi_\tau\left(Y_t^*\right) - \varepsilon_t\right]^2 = I\left(d_{1t} < Y_t \le d_{2t}\right) \tag{3.25}$$

where $d_{1t} = \min\left(c_{1t}, c_{2t}\right)$ and $d_{2t} = \max\left(c_{1t}, c_{2t}\right)$ with $c_{1t} = \mathbf{X}_t^\top \mathbf{a}\left(U_t\right)$ and $c_{2t} = \mathbf{X}_t^\top\left[\mathbf{a}(u_0) + \mathbf{a}^\top(u_0)\left(U_t - u_0\right)\right]$. Further,

$$\mathbb{E}\left[\left\{\psi_\tau\left(Y_t^*\right) - \varepsilon_t\right\}^2 K^2\left(U_{th}\right)\mathbf{X}_t^* \mathbf{X}_t^{*\prime}\right] = \mathbb{E}\left[\left\{F_{y|u,x}\left(d_{2t}\right) - F_{y|u,x}\left(d_{1t}\right)\right\} K^2\left(U_{th}\right)\mathbf{X}_t^* \mathbf{X}_t^{*\prime}\right] = O\left(h^3\right).$$

Thus, $\mathrm{Var}\left(\mathbf{B}_n\right) = o(1)$. This, in conjunction with (3.23) and (3.24) and the Slutsky Theorem, proves the theorem. □

3.6.2 Proofs of Lemmas

Note that the same notations in Sections 3.3 and 3.6 are used here. Throughout this section, we denote a generic constant by C, which may take different values at different appearances. Let $F_{y|u,x}(y)$ denote the conditional distribution of Y given U and $\mathbf{X}$.

Proof of Lemma 3.3. First, for any $\boldsymbol{\theta} \in A_m$, we consider the following term

$$V_n(\boldsymbol{\theta}) - V_n(0) = a_n \sum_{t=1}^{n} \left[\psi_\tau\left(Y_{nt}^*\right) - \psi_\tau\left(Y_t^*\right)\right] \mathbf{X}_t^* K\left(U_{th}\right) \equiv a_n \sum_{i=1}^{n} V_{nt}(\boldsymbol{\theta}),$$

where $Y_{nt}^* = Y_t^* - a_n \boldsymbol{\theta}^\top \mathbf{X}_t^*$ and $V_{nt}(\boldsymbol{\theta}) = V_{nt} = \left[\psi_\tau\left(Y_{nt}^*\right) - \psi_\tau\left(Y_t^*\right)\right] \mathbf{X}_t^* K\left(U_{th}\right) = \left(V_{nt1}^\top, V_{nt2}^\top\right)^\top$ with

$$V_{nt1} = \left[\psi_\tau\left(Y_{nt}^*\right) - \psi_\tau\left(Y_t^*\right)\right] \mathbf{X}_t K\left(U_{th}\right) \quad \text{and} \quad V_{nt2} = \left[\psi_\tau\left(Y_{nt}^*\right) - \psi_\tau\left(Y_t^*\right)\right] \mathbf{X}_t U_{th} K\left(U_{th}\right).$$

Thus,

$$\begin{aligned} &\left\|V_n(\boldsymbol{\theta}) - V_n(0) - \mathbb{E}\left[V_n(\boldsymbol{\theta}) - V_n(0)\right]\right\| \\ \leq & a_n \left\|\sum_{t=1}^{n} \left(V_{nt1} - EV_{nt1}\right)\right\| + a_n \left\|\sum_{t=1}^{n} \left(V_{nt2} - EV_{nt2}\right)\right\| \equiv V_n^{(1)} + V_n^{(2)}. \end{aligned}$$

Clearly,

$$V_n^{(1)} \equiv a_n \left\|\sum_{t=1}^{n} \left(V_{nt1} - EV_{nt1}\right)\right\| \leq \sum_{i=0}^{d} \left\|V_n^{(1i)}\right\|,$$

where $V_n^{(1i)} = a_n \sum_{t=1}^{n} \left(V_{nt1}^{(i)} - EV_{nt1}^{(i)}\right)$ and $V_{nt1}^{(i)} = \left[\psi_\tau\left(Y_{nt}^*\right) - \psi_\tau\left(Y_t^*\right)\right] X_{ti} K\left(U_{th}\right)$, which is the i-th component of V_{nt1}. Then,

$$\begin{aligned} \mathrm{Var}\left(V_n^{(1i)}\right) &= a_n^2 \mathbb{E}\left\{\sum_{t=1}^{n} \left(V_{nt1}^{(i)} - EV_{nt1}^{(i)}\right)\right\}^2 \\ &= a_n^2 \left[\sum_{t=1}^{n} \mathrm{Var}\left(V_{nt1}^{(i)}\right) + 2\sum_{s=1}^{n-1} \left(1 - \frac{s}{n}\right) \mathrm{Cov}\left(V_{n11}^{(i)}, V_{n(s+1)1}^{(i)}\right)\right] \\ &\leq \frac{1}{h}\left[\mathrm{Var}\left(V_{n11}^{(i)}\right) + 2\sum_{s=1}^{d_n - 1} \left|\mathrm{Cov}\left(V_{n11}^{(i)}, V_{n(s+1)1}^{(i)}\right)\right| + 2\sum_{s=d_n}^{\infty} \left|\mathrm{Cov}\left(V_{n11}^{(i)}, V_{n(s+1)1}^{(i)}\right)\right|\right] \\ &\equiv J_1 + J_2 + J_3 \end{aligned}$$

for some $d_n \to \infty$ specified later. For J_3, use the Davydov's inequality (see Lemma 1.1) to obtain

$$\left|\text{Cov}\left(V_{n11}^{(i)}, V_{n(s+1)1}^{(i)}\right)\right| \leq C\alpha^{1-2/\delta}(s)\left[E\left|V_{n11}^{(i)}\right|^{\delta}\right]^{2/\delta}.$$

Similar to (3.25), for any $k > 0$,

$$\left|\psi_\tau\left(Y_{nt}^*\right) - \psi_\tau\left(Y_t^*\right)\right|^k = I\left(d_{3t} < Y_t \leq d_{4t}\right),$$

where $d_{3t} = \min\left(c_{2t}, c_{2t} + c_{3t}\right)$ and $d_{4t} = \max\left(c_{2t}, c_{2t} + c_{3t}\right)$ with $c_{3t} = a_n\boldsymbol{\theta}^\top\mathbf{X}_t^*$. Therefore, by Assumption C3, there exists a $C > 0$ independent of $\boldsymbol{\theta}$ such that

$$\mathbb{E}\left\{\left|\psi_\tau\left(Y_{nt}^*\right) - \psi_\tau\left(Y_t^*\right)\right|^k \mid \mathbf{U}_t, \mathbf{X}_t\right\} = F_{y|u,x}\left(c_{4t}\right) - F_{y|u,x}\left(c_{3t}\right) \leq Ca_n\left|\boldsymbol{\theta}^\top\mathbf{X}_t^*\right|,$$

which implies that

$$\begin{aligned} E\left|V_{n11}^{(i)}\right|^{\delta} &= \mathbb{E}\left[\left|\psi_\tau\left(Y_{n1}^*\right) - \psi_\tau\left(Y_1^*\right)\right|^{\delta}\left|X_{1i}\right|^{\delta}K^{\delta}\left(U_{1h}\right)\right] \\ &\leq Ca_n\mathbb{E}\left[\left|\boldsymbol{\theta}^\top\mathbf{X}_t^*\right|\left|X_{1i}\right|^{\delta}K^{\delta}\left(U_{1h}\right)\right] \leq Ca_nh \end{aligned}$$

uniformly in $\boldsymbol{\theta}$ over A_m by Assumption C6. Then,

$$J_3 \leq Ca_n^{2/\delta}h^{2/\delta-1}\sum_{s=d_n}^{\infty}[\alpha(s)]^{1-2/\delta} \leq Ca_n^{2/\delta}h^{2/\delta-1}d_n^{-l}\sum_{s=d_n}^{\infty}s^l[\alpha(s)]^{1-2/\delta} = o\left(a_n^{2/\delta}h^{2/\delta-1}d_n^{-l}\right)$$

uniformly in $\boldsymbol{\theta}$ over A_m. As for J_2, we use Assumption C10 to get

$$\left|\text{Cov}\left(V_{n11}^{(i)}, V_{n(s+1)1}^{(i)}\right)\right| \leq C\left[\mathbb{E}\left\{\left|X_{1i}X_{(s+1)i}\right|K\left(U_{1h}\right)K\left(U_{(s+1)h}\right)\right\} + a_n^2h^2\right] = O\left(h^2\right)$$

uniformly in $\boldsymbol{\theta}$ over A_m. It follows that $J_2 = O\left(d_nh\right)$ uniformly in $\boldsymbol{\theta}$ over A_m. Analogously,

$$J_1 = h^{-1}\text{Var}\left(V_{n11}^{(i)}\right) \leq h^{-1}\mathbb{E}\left(V_{n11}^{(i)}\right)^2 = O\left(a_n\right)$$

uniformly in $\boldsymbol{\theta}$ over A_m. By choosing d_n such that $d_n^lh^{1-2/\delta} = c$, then, $d_nh \to 0$ and $\text{Var}\left(V_n^{(1i)}\right) = o(1)$. Therefore, $V_n^{(1i)} = o_p(1)$ so that $V_n^{(1)} = o_p(1)$ uniformly in $\boldsymbol{\theta}$ over A_m. By the same token, we can show that $V_n^{(2)} = o_p(1)$ uniformly in $\boldsymbol{\theta}$ over A_m. This completes the proof of the lemma. □

Proof of Lemma 3.4. It is easy to justify that

$$\begin{aligned} \mathbb{E}\left[V_n(\theta) - V_n(0)\right] &= na_n\mathbb{E}\left[\left(\psi_\tau\left(Y_t^* - a_n\boldsymbol{\theta}^\top\mathbf{X}_t^*\right) - \psi_\tau\left(Y_t^*\right)\right)\mathbf{X}_t^*K\left(U_{th}\right)\right] \\ &= na_n\mathbb{E}\left[\left\{F_{y|u,x}\left(c_{2t}\right) - F_{y|u,x}\left(c_{2t} + a_n\boldsymbol{\theta}^\top\mathbf{X}_t^*\right)\right\}\mathbf{X}_t^*K\left(U_{th}\right)\right] \\ &\approx -\frac{1}{h}\mathbb{E}\left[f_{y|u,x}\left(c_{2t}\right)\mathbf{X}_t^*\mathbf{X}_t^{*\prime}K\left(U_{th}\right)\right]\boldsymbol{\theta} \\ &\approx -f_u\left(u_0\right)\Omega_1^*\left(u_0\right)\boldsymbol{\theta} \end{aligned}$$

uniformly in $\boldsymbol{\theta}$ over A_m by Assumption C3. The proof of the lemma is complete. □

Proof of Lemma 3.5. Observe by Taylor expansions and Assumption C3 that

$$\begin{aligned}
\mathbb{E}\left[\mathbf{Z}_t\right] &= \mathbb{E}\left[\left\{\tau - F_{y|u,x}\left(c_{2t}\right)\right\}\mathbf{X}_t^* K\left(U_{th}\right)\right] \\
&\approx \mathbb{E}\left[\left\{F_{y|u,x}\left(c_{2t} + \mathbf{X}_t^\top \mathbf{a}''\left(u_0\right) h^2 U_{th}^2/2\right) - F_{y|u,x}\left(c_{2t}\right)\right\}\mathbf{X}_t^* K\left(U_{th}\right)\right] \\
&\approx \frac{h^2}{2}\mathbb{E}\left[f_{y|u,x}\left(c_{2t}\right)\mathbf{X}_t^*\mathbf{X}_t^\top \mathbf{a}''\left(u_0\right) U_{th}^2 K\left(U_{th}\right)\right] \\
&\approx \frac{h^2}{2}\mathbb{E}\left[f_{y|u,x}\left(q_\tau\left(u_0, \mathbf{X}_t\right)\right)\mathbf{X}_t^*\mathbf{X}_t^\top \mathbf{a}''\left(u_0\right) U_{th}^2 K\left(U_{th}\right)\right] \\
&\approx \frac{h^3 f_u\left(u_0\right)}{2}\begin{pmatrix} \mu_2(K)\Omega^*\left(u_0\right)\mathbf{a}''\left(u_0\right) \\ \mathbf{0} \end{pmatrix}. \qquad (3.26)
\end{aligned}$$

Also, we have

$$\begin{aligned}
\mathrm{Var}\left[\mathbf{Z}_t\right] &= \mathbb{E}\left[\left\{\tau - I\left(Y_t < c_{2t}\right)\right\}^2 \mathbf{X}_t^*\mathbf{X}_t^{*\prime} K^2\left(U_{th}\right)\right] \\
&\approx \mathbb{E}\left[\left\{\tau^2 - 2\tau F_{y|u,x}\left(c_{2t}\right) + F_{y|u,x}\left(c_{2t}\right)\right\}\mathbf{X}_t^*\mathbf{X}_t^{*\prime} K^2\left(U_{th}\right)\right] \\
&\approx \tau(1-\tau)\mathbb{E}\left[\mathbf{X}_t^*\mathbf{X}_t^{*\prime} K^2\left(U_{th}\right)\right] \\
&\approx \tau(1-\tau)h f_u\left(u_0\right)\Omega_1\left(u_0\right). \qquad (3.27)
\end{aligned}$$

Next, we show that the last part of lemma holds true. Clearly, $V_n(0) = a_n \sum_{t=1}^n \mathbf{Z}_t$. Similar to the proof of Lemma 3.3, we have

$$\begin{aligned}
\mathrm{Var}\left[V_n(0)\right] &= \frac{1}{h}\mathrm{Var}\left(\mathbf{Z}_1\right) + \frac{2}{h}\sum_{s=1}^{d_n-1}\left(1 - \frac{s}{n}\right)\mathrm{Cov}\left(\mathbf{Z}_1, \mathbf{Z}_{s+l}\right) + \frac{2}{h}\sum_{s=d_n}^{n}\left(1 - \frac{s}{n}\right)\mathrm{Cov}\left(\mathbf{Z}_1, \mathbf{Z}_{s+l}\right) \\
&\equiv J_4 + J_5 + J_6
\end{aligned}$$

for some $d_n \to \infty$ specified later. By (3.27),

$$J_4 \quad \to \quad \tau(1-\tau) f_u\left(u_0\right)\Omega_1\left(u_0\right).$$

Therefore, it suffices to show that $|J_5| = o(1)$ and $|J_6| = o(1)$. For J_6, using the Davydov's inequality (see, e.g., Lemma 1.1) and the boundedness of $\psi_\tau(\cdot)$ to obtain

$$\left|\mathrm{Cov}\left(\mathbf{Z}_1, \mathbf{Z}_{s+1}\right)\right| \le C\alpha^{1-2/\delta}(s)\left[E\left|\mathbf{Z}_1\right|^\delta\right]^{2/\delta} \le Ch^{2/\delta}\alpha^{1-2/\delta}(s),$$

which gives

$$J_6 \le Ch^{2/\delta-1}\sum_{s=d_n}^{\infty}\left[\alpha(s)\right]^{1-2/\delta} \le Ch^{2/\delta-1}d_n^{-l}\sum_{s=d_n}^{\infty} s^l\left[\alpha(s)\right]^{1-2/\delta} = o\left(h^{2/\delta-1}d_n^{-l}\right) = o(1)$$

by choosing d_n to satisfy $d_n^l h^{1-2/\delta} = c$. As for J_5, we use Assumption C10 and (3.26) to get

$$\left|\mathrm{Cov}\left(\mathbf{Z}_1, \mathbf{Z}_{s+1}\right)\right| \le C\left[\mathbb{E}\left\{\left|\mathbf{X}_1^*\mathbf{X}_{s+1}^*\right| K\left(U_{1h}\right) K\left(U_{(s+1)h}\right)\right\} + h^6\right] = O\left(h^2\right)$$

so that $J_5 = O\left(d_n h\right) = o(1)$ by the choice of d_n. We finish the proof of this lemma. □

Proof of (3.13) **and** (3.14). By the Taylor expansion,

$$\mathbb{E}\left[\xi_t \mid U_t, \mathbf{X}_t\right] = F_{y|u,x}\left(\mathbf{X}_t^\top \mathbf{a}\left(u_0\right) + a_n\right) - F_{y|u,x}\left(\mathbf{X}_t^\top \mathbf{a}\left(u_0\right)\right) \approx f_{y|u,x}\left(\mathbf{X}_t^\top \mathbf{a}\left(u_0\right)\right) a_n.$$

Therefore,

$$\mathbb{E}\left[\mathbf{S}_n\right] \approx h^{-1}\mathbb{E}\left[f_{y|u,x}\left(\mathbf{X}_t^\top \mathbf{a}\left(u_0\right)\right) \mathbf{X}_t^* \mathbf{X}_t^{*\prime} K\left(U_{th}\right)\right] \approx f_u\left(u_0\right) \Omega_1^*\left(u_0\right).$$

Similar to the proof of $\text{Var}\left[V_n(0)\right]$ in Lemma 3.5, one can show that $\text{Var}\left(\mathbf{S}_n\right) \to 0$. Therefore, $\mathbf{S}_n \to f_u\left(u_0\right) \Omega_1^*\left(u_0\right)$ in probability. This proves (3.13). Clearly,

$$\mathbb{E}\left[\widehat{\Omega}_{n,0}\right] = \mathbb{E}\left[\mathbf{X}_t \mathbf{X}_t^\top K_h\left(U_t - u_0\right)\right] = \int \Omega\left(u_0 + hv\right) f_u\left(u_0 + hv\right) K(v) dv \approx f_u\left(u_0\right) \Omega\left(u_0\right).$$

Similarly, one can show that $\text{Var}\left(\widehat{\Omega}_{n,0}\right) \to 0$. This proves the first part of (3.14). By the same token, one can show that $\mathbb{E}\left[\widehat{\Omega}_{n,1}\right] \approx f_u\left(u_0\right) \Omega^*\left(u_0\right)$ and $\text{Var}\left(\widehat{\Omega}_{n,1}\right) \to 0$. Thus, $\widehat{\Omega}_{n,1} = f_u\left(u_0\right) \Omega^*\left(u_0\right) + o_p(1)$. We prove (3.14). □

3.7 Composite Quantile Regression

It is very interesting to note that we can use quantile technique to estimate $m(\mathbf{X}_t)$ and $\sigma^2(\mathbf{X}_t)$ in (3.2), which is the so-called composite quantile regression (CQR), although both $m(\mathbf{X}_t)$ and $\sigma^2(\mathbf{X}_t)$ do not depend τ. Instead of using the classical least squares estimation method, the alternative estimator considered in this section is based on the composite quantile regression (CQR) approach of Zou and Yuan (2008), who showed that the CQR estimator for regression betas can be more efficient than the least squares estimator when regression errors are non-normal. The efficiency gain of CQR results from combining the information from multiple quantiles, and can be noticeably large. Kai et al. (2010) and Huang and Zhan (2022) also adopted the CQR approach to nonparametric functionals. Consistent with Zou and Yuan (2008), both Kai et al. (2010) and Huang and Zhan (2022) found that there are efficiency gains to use the CQR estimator instead of the least squares estimator under non-normality. Particularly, Kai et al. (2010) considered nonparametric estimation through local composite quantile regression (LCQR), while Huang and Zhan (2022) explored the boundary points of LCQR around the cutoff of regression discontinuity designs.

For a classical linear model $Y_t = \boldsymbol{\beta}^\top \mathbf{X}_t + \epsilon_t$, to estimate β, the quantile regression is $q_\tau(\mathbf{X}_t) = b_\tau + \boldsymbol{\beta}^\top \mathbf{X}_t$, where $b_\tau = F_\epsilon^{-1}(\tau)$, so that the CQR is given by

$$(\widehat{b}_1, \cdots, \widehat{b}_M, \widehat{\boldsymbol{\beta}}^{\text{CQR}}) = \arg_{b_1,\cdots,b_M} \min_{\beta} \sum_{j=1}^{M} \sum_{t=1}^{n} \rho_{\tau_j}\left(Y_t - b_j - \boldsymbol{\beta}^\top \mathbf{X}_t\right) w_{1,j} \tag{3.28}$$

for any $0 < \tau_1 < \cdots < \tau_M < 1$ and some M, where $\{w_{1,j}\}_{j=1}^M$ are some weights, which is also called the "H-estimator" as in Koenker (2005) and can be computationally implemented via the Package "cqrReg" in **R**, see, for instance, Pietrosanu et al. (2017). In Zou and Yuan (2008), $w_{1,j} = 1$ and $\tau_j = j/(M+1)$. As for how to choose the optimal weights $\{w_{1,j}\}_{j=1}^M$, the reader is referred to the book by Koenker (2005). Furthermore, under some regularity conditions, Zou and Yuan (2008) showed that $\widehat{\boldsymbol{\beta}}^{\mathrm{CQR}}$ has the following asymptotic property for the iid data

$$\sqrt{n}\left[\widehat{\boldsymbol{\beta}}^{\mathrm{CQR}} - \boldsymbol{\beta}\right] \xrightarrow{d} N\left(0, \Sigma_{\mathrm{CQR}}\right),$$

where

$$\Sigma_{\mathrm{CQR}} = \Sigma_{\mathbf{X}}^{-1} \sum_{j_1,j_2=1}^{M} \tau_{j_1,j_2} \left[\sum_{j=1}^{M} f_\epsilon(b_{\tau_j})\right]^{-2},$$

$\tau_{j_1,j_2} = \min(\tau_{j_1}, \tau_{j_2}) - \tau_{j_1}\tau_{j_2}$, $\Sigma_{\mathbf{X}} = \lim_{n\to\infty} \sum_{t=1}^n \mathbf{X}_t\mathbf{X}_t^\top/n > 0$, and $f_\epsilon(\cdot)$ is the density of ϵ_t. Also, as addressed in Section 3.1, Zou and Yuan (2008) investigated the asymptotic relative efficiency by comparing the OLS estimate and obtained the lower bound at 70.26%. For example, when ϵ_t follows the double exponential (Laplace) distribution, the ARE is 150% and it is 95.5% for the normal case.

Kai et al. (2010) extended a linear model to nonparametric setting as $Y_t = m(\mathbf{X}_t) + \epsilon_t$. They proposed a new nonparametric regression technique called local composite quantile regression smoothing to improve local polynomial regression further. Sampling properties of the estimation procedure proposed were studied. They derived the asymptotic bias, variance and normality of the estimate proposed. The asymptotic relative efficiency of the estimate with respect to local polynomial regression was investigated. It was shown that the estimate can be much more efficient than the local polynomial regression estimate for various non-normal errors, while being almost as efficient as the local polynomial regression estimate for normal errors. For simplicity, we present only local linear fitting, namely, $m(\mathbf{X}_t) \approx m(\mathbf{x}) + m'(\mathbf{x})^\top(\mathbf{X}_t - \mathbf{x})$, when $\mathbf{X}_t$ in the neighborhood of $\mathbf{x}$ for any given grid point $\mathbf{x}$, where $m'(\mathbf{x})$ is the first order derivative of $m(\mathbf{x})$. By adding the kernel weight into (3.28), similar to (3.7), we have the following locally weighted CQR loss

$$(\widehat{b}_1, \cdots, \widehat{b}_M, \widehat{\boldsymbol{\beta}}) = \arg_{b_1,\cdots,b_M} \min_{\beta} \sum_{j=1}^{M}\sum_{t=1}^{n} \rho_{\tau_j}\left(Y_t - b_j - \boldsymbol{\beta}^\top(\mathbf{X}_t - \mathbf{x})\right) K((\mathbf{X}_t - \mathbf{x})/h). \quad (3.29)$$

Then, define

$$\widehat{m}(\mathbf{x})_{\mathrm{CQR}} = \frac{1}{M}\sum_{j=1}^{M}\widehat{b}_j \quad \text{and} \quad \widehat{m}'(\mathbf{x})_{\mathrm{CQR}} = \widehat{\beta},$$

which are the so-called local linear CQR estimate of $m(\mathbf{x})$ and $m'(\mathbf{x})$, respectively. See Kai et al. (2010) for a detailed discussion. Kai et al. (2010) derived the asymptotic normality of $\widehat{m}(\mathbf{x})^{\mathrm{CQR}}$, given by

$$\sqrt{n}\left[\widehat{m}(\mathbf{x})_{\mathrm{CQR}} - m(\mathbf{x}) - B(\mathbf{x}) + o_p(h^2)\right] \xrightarrow{d} N\left(0, \sigma^2_{\mathrm{CQR}}\right),$$

where $B(\mathbf{x}) = \frac{1}{2}m''(\mathbf{x})\mu_2\, h^2$ is the asymptotic bias term and $\sigma^2_{\mathrm{CQR}} = \nu_0(K)\sigma^2_\epsilon(\mathbf{x})R_1(M)/f_\epsilon(\mathbf{x})$ with $\sigma^2_\epsilon(\mathbf{x}) = \mathrm{Var}(\epsilon_t|\,\mathbf{X}_t = \mathbf{x})$ and

$$R_1(M) = \frac{1}{M^2}\sum_{j_1=}^{M}\sum_{j_2=1}^{M}\frac{\tau_{j_1,j_2}}{f_\epsilon(b_{\tau_{j_1}})f_\epsilon(b_{\tau_{j_2}})},$$

which depends on only $\{\tau_j\}$ and the density of the error term ϵ_t. By comparing the asymptotic results for the local linear estimator, one can see clearly from Section 2.3.3 that the asymptotic bias term is exactly same and the only difference is the asymptotic variance with the extra term $R_1(M)$, so that the optimal bandwidth $h^{\mathrm{opt}}_{\mathrm{CQR}}$ is proportional to the one for the local linear least squares estimator, $h^{\mathrm{opt}}_{\mathrm{LS}}$; that is $h^{\mathrm{opt}}_{\mathrm{CQR}} = R_1(M)^{1/5}\, h^{\mathrm{opt}}_{\mathrm{LS}}$, as in (2.8) in Kai et al. (2010). As argued in Kai et al. (2010), the asymptotic relative efficiency of the local linear CQR estimator with respect to the local linear least squares estimator, as

$$\mathrm{ARE}\left(\widehat{m}_{\mathrm{CQR}}, \widehat{m}_{\mathrm{LS}}\right) = R_1(M)^{-4/5},$$

where $\widehat{m}_{\mathrm{LS}}$ is the local linear least squares estimator defined in (2.5), which is (2.9) in Kai et al. (2010). ARE depends only on the error distribution, although the dependence could be rather complex. However, for many commonly seen error distributions, we can directly compute the value of ARE.

Remark 3.6: *First, Koenker (2005) did not discuss relative efficiency of the Hogg estimator relative to the least squares estimator. Second, it is worth mentioning here that, although the check loss function is typically used to estimate the conditional quantile function of Y_t given $\mathbf{X}_t$; see, e.g., Koenker (2005) and references therein, several check functions to estimate the regression (mean) function are simultaneously employed in both* (3.28) *and* (3.29)*. Therefore, the local CQR smoother is conceptually different from nonparametric quantile regression by local fitting which has been studied in Yu and Jones (1998) and Chapter 5 of Fan and Gijbels (1996).*

Chapter 4

Nonparametric Measures of Risk

4.1 Introduction

Value-at-Risk (VaR), Expected Shortfall (ES), and Expectile have indeed become prominent measures for quantifying market risk associated with assets and portfolios within the finance industry over recent decades. In particular, VaR has been chosen by the Basle Committee on Banking Supervision as the benchmark of risk measures for capital requirements and both of them have been used by financial institutions for asset managements and minimization of risk as well as have been developed rapidly as analytic tools to assess riskiness of trading activities. See, to name just a few, Morgan (1996), Duffie and Pan (1997), Jorion (2001, 2003), and Duffie and Singleton (2003) for the financial background, statistical inferences, and various applications. In terms of the formal definition, VaR is simply a quantile of the loss distribution (future portfolio values) over a prescribed holding period (e.g., 2 weeks) at a given confidence level, while ES is the expected loss, given that the loss is at least as large as some given quantile of the loss distribution (e.g., VaR). It is well known from Artzner et al. (1999) that ES is a coherent risk measure such as it satisfies the following four axioms:

- homogeneity: increasing the size of a portfolio by a factor should scale its risk measure by the same factor,
- monotonicity: a portfolio must have greater risk if it has systematically lower values than another,
- risk-free condition or translation invariance: adding some amount of cash to a portfolio should reduce its risk by the same amount, and

- subadditivity: the risk of a portfolio must be less than the sum of separate risks or merging portfolios cannot increase risk.

VaR satisfies homogeneity, monotonicity, and risk-free condition but is not sub-additive. See AArtzner et al. (1999) for details. As advocated by Artzner et al. (1999), ES is preferred due to its better properties although VaR is widely used in applications. However, estimating ES is not straightforward so that it demands a heavy computing as we will see this later.

Measures of risk might depend on the state of the economy since economic and market conditions vary from time to time. This requires risk managers should focus on the conditional distributions of profit and loss, which take full account of current information about the investment environment (macroeconomic and financial as well as political) in forecasting future market values, volatilities, and correlations. As pointed out by Duffie and Singleton (2003), not only are the prices of the underlying market indices changing randomly over time, the portfolio itself is changing, as are the volatilities of prices, the credit qualities of counter-parties, and so on. On the other hand, one would expect the VaR to increase as the past returns become very negative, because one bad day makes the probability of the next somewhat greater. Similarly, very good days also increase the VaR, as would be the case for volatility models. Therefore, VaR could depend on the past returns in someway. Hence, an appropriate risk analytical tool or methodology should be allowed to adapt to varying market conditions and to reflect the latest available information in a time series setting rather than the iid framework. Most of the existing risk management literature has concentrated on unconditional distributions and the iid setting although there have been some studies on the conditional distributions and time series data. For more background, see Chernozhukov and Umantsev (2001), Cai (2002a), Fan and Gu (2003), Engle and Manganelli (2004), Cai and Xu (2008), Scaillet (2005), and Cosma et al. (2007), and references therein for conditional models, and Duffie and Pan (1997), Artzner et al. (1999), Rockafellar et al. (2000), Acerbi and Tasche (2002), Frey and McNeil (2002), Scaillet (2004), Chen and Tang (2005), Chen (2008), and among others for unconditional models. Also, most of studies in the literature and applications are limited to parametric models, such as all standard industry models like CreditRisk$^+$, CreditMetrics, CreditPortfolio View and the model proposed by the KMV corporation. See Chernozhukov and Umantsev (2001), Frey and McNeil (2002), Engle and Manganelli (2004), and references therein on parametric models in practice and Fan and Gu (2003) and references therein for semiparametric models.

The main focus of this chapter is on studying the conditional value-at-risk (CVaR) and conditional expected shortfall (CES) as well as conditional expectile, and proposing some new nonparametric estimation procedures to estimate CVaR function and CES function as well as conditional expectile function, where the conditional information is allowed to contain economic and market (exogenous) variables and the past observed returns. Parametric models for CVaR and CES can be most efficient if the underlying functions are correctly specified. See Chernozhukov and Umantsev (2001) for a polynomial type regression model and Engle and Manganelli (2004) for a GARCH type parametric model for CVaR based on regression quantile. However, a misspecification may cause serious bias and model constraints may distort the underlying distributions. A nonparametric modeling is appealing in several aspects. One of the advantages for nonparametric modeling is that little or no restrictive prior information on functionals is needed. Further, it may provide a useful insight for further parametric fitting.

The approach proposed by Cai and Wang (2008) has several advantages. The first one is to propose a new nonparametric approach to estimate CVaR and CES. In essence, our estimator for CVaR is based on inverting a newly proposed estimator of the conditional distribution function for time series data and the estimator for CES is by a plugging-in method based on plugging in the estimated conditional probability density function and the estimated CVaR function. Note that they are analogous to the estimators studied by Scaillet (2005) by using the Nadaraya-Watson (NW) type double kernel (smoothing in both the y and x directions) estimation, and Cai (2002a) by utilizing the weighted NadarayaWatson (WNW) kernel type technique to avoid the so-called boundary effects as well as Yu and Jones (1998) by employing the double kernel local linear method. More precisely, our newly proposed estimator combines the WNW method of Cai (2002a) and the double kernel local linear technique of Yu and Jones (1998), termed as *weighted double kernel local linear* (WDKLL) estimator.

The second merit is to establish the asymptotic properties for the WDKLL estimators of the conditional probability density function and cumulative distribution function for the α-mixing time series at both boundary and interior points. It is therefore shown that the WDKLL method enjoys the same convergence rates as those of the double kernel local linear estimator of Yu and Jones (1998) and the WNW estimator of Cai (2002a). It is also shown that the WDKLL estimators have desired sampling properties at both boundary and

interior points of the support of the design density, which seems to be seminal. Finally, we derive the WDKLL estimator of CVaR by inverting the WDKLL conditional distribution estimator and the WDKLL estimator of CES by plugging in the WDKLL estimators of PDF and CVaR. We show that the WDKLL estimator of CVaR exists always due to the WDKLL estimator of CDF being a distribution function itself, and that it inherits all better properties from the WDKLL estimator of CDF; that is, the WDKLL estimator of CDF is a CDF and differentiable, and it possess the asymptotic properties such as design adaption, avoiding boundary effects, and mathematical efficiency. Note that to preserve shape constraints, Cosma et al. (2007) used a wavelet method to estimate conditional probability density and cumulative distribution functions and then to estimate conditional quantiles.

Note that CVaR defined here is essentially the conditional quantile or quantile regression of Koenker and Bassett (1978), based on the conditional distribution, rather than CVaR defined in some risk management literature such as Rockafellar et al. (2000) and Jorion (2001, 2003), which is the so-called ES. Also, note that the ES here is called TailVaR in Artzner et al. (1999). Moreover, as aforementioned, CVaR can be regarded as a special case of quantile regression. See Cai and Xu (2008) for the state-of-the-art about the research on nonparametric quantile regression, including CVaR. Further, note that both ES and CES have been known for decades among actuary sciences and they are very popular in insurance industry. Indeed, they have been used to assess risk on a portfolio of potential claims, and to design reinsurance treaties. See the book by Embrechts et al. (1997) for the excellent review on this subject and the papers by McNeil (1997), Hürlimann (2003), Scaillet (2005), and Chen (2008). Finally, ES or CES is also closely related to other applied fields such as the mean residual life function in reliability and the biometric function in biostatistics. See Oakes and Dasu (1990) and Cai and Qian (2000) and references therein.

4.2 Nonparametric Model Setup

Assume that the observed data $\{(\mathbf{X}_t, Y_t); 1 \leq t \leq n\}, \mathbf{X}_t \in \mathbb{R}^p$, are available and they are observed from a stationary time series model. Here Y_t is the risk or loss variable which can be the negative logarithm of return (log loss) and $\mathbf{X}_t$ is allowed to include both economic and market (exogenous) variables and the lagged variables of Y_t and also it can be a vector. But, for the expositional purpose, we consider only the case when X_t is a scalar ($p = 1$). Note that the proposed methodologies and their theory for the univariate case ($p = 1$) continue to hold

for multivariate situations ($p > 1$). Extension to the case $p > 1$ involves no fundamentally new ideas. Note that models with large p are often not practically useful due to the curse of dimensionality.

We now turn to considering the nonparametric estimation of the conditional expected shortfall $\mu_\tau(x)$, which is defined as

$$\mu_\tau(x) = \mathbb{E}\left[Y_t \mid Y_t \geq \nu_\tau(x), X_t = x\right],$$

where $\nu_\tau(x)$ is the conditional value-at-risk, which is defined as the solution of

$$\mathbb{P}\left(Y_t \geq \nu_\tau(x) \mid X_t = x\right) = S\left(\nu_\tau(x) \mid x\right) = \tau$$

or expressed as $\nu_\tau(x) = S^{-1}(\tau \mid x)$, where $S(y \mid x)$ is the conditional survival function of Y_t given $X_t = x$, $S(y \mid x) = 1 - F(y \mid x)$, and $F(y \mid x)$ is the conditional cumulative distribution function. It is easy to see that

$$\mu_\tau(x) = \int_{\nu_\tau(x)}^{\infty} y f(y \mid x) dy / \tau,$$

where $f(y \mid x)$ is the conditional probability density function of Y_t given $X_t = x$. To estimate $\mu_\tau(x)$, one can use the plugging-in method as

$$\widehat{\mu}_\tau(x) = \int_{\widehat{\nu}_\tau(x)}^{\infty} y \widehat{f}(y \mid x) dy / \tau, \tag{4.1}$$

where $\widehat{\nu}_\tau(x)$ is a nonparametric estimation of $\nu_\tau(x)$ and $\widehat{f}(y \mid x)$ is a nonparametric estimation of $f(y \mid x)$. But the bandwidths for $\widehat{\nu}_\tau(x)$ and $\widehat{f}(y \mid x)$ are not necessary to be same.

Note that Scaillet (2005) used the NW type double kernel method to estimate $f(y \mid x)$ first, due to Roussas (1969b), denoted by $\widetilde{f}(y \mid x)$, and then estimated $\nu_\tau(x)$ by inverting the estimated conditional survival function, denoted by $\widetilde{\nu}_\tau(x)$, and finally estimated $\mu_\tau(x)$ by plugging $\widetilde{f}(y \mid x)$ and $\widetilde{\nu}_\tau(x)$ into (4.1), denoted by $\widetilde{\mu}_\tau(x)$, where $\widetilde{\nu}_\tau(x) = \widetilde{S}^{-1}(y \mid x)$ and $\widetilde{S}(y \mid x) = \int_y^\infty \widetilde{f}(u \mid x) du$. But, it is well documented; see, e.g., Fan and Gijbels (1996) that the NW kernel type procedures have serious drawbacks: the asymptotic bias involves the design density so that they can not be adaptive, and boundary effects exist so that they require boundary modifications. In particular, boundary effects might cause a serious problem for estimating $\nu_\tau(x)$ since it is only concerned with the tail probability. The question is now how to provide a better estimate for $f(y \mid x)$ and $\nu_\tau(x)$ so that we have a good estimate for $\mu_\tau(x)$. Therefore, we address this issue in the next section.

4.3 Nonparametric Estimating Procedures

We start with the nonparametric estimators for the conditional density function and its distribution function first and then turn to discussing the nonparametric estimators for the conditional VaR and ES functions.

There are several methods available for estimating $\nu_\tau(x), f(y \mid x)$, and $F(y \mid x)$ in the literature, such as kernel and nearest-neighbor[1]. To attenuate these drawbacks of the kernel type estimators mentioned in Section 4.2, some new methods have been proposed to estimate conditional quantiles. The first one, a more direct approach, by using the "check" function such as the robustified local linear smoother, was provided by Fan et al. (1994) and further extended by Yu and Jones (1998) for the iid data. A more general nonparametric setting was explored by Cai and Xu (2008) for time series data. This modeling idea was initialed by Koenker and Bassett (1978) for linear regression quantiles and Fan et al. (1994) for nonparametric models. See Cai and Xu (2008) and references therein for more discussions on models and applications. An alternative procedure is first to estimate the conditional distribution function by using double kernel local linear technique of Fan et al. (1996) and then to invert the conditional distribution estimator to produce an estimator of a conditional quantile or CVaR. Yu and Jones (1998) compared these two methods theoretically and empirically and suggested that the double kernel local linear would be better.

4.3.1 Estimation of Conditional PDF and CDF

To make a connection between the conditional density (distribution) function and nonparametric regression problem, it is noted by the standard kernel estimation theory; see, e.g., Fan and Gijbels (1996), that for a given symmetric density function $K(\cdot)$,

$$\mathbb{E}\left\{K_{h_0}\left(y - Y_t\right) \mid X_t = x\right\} = f(y \mid x) + \frac{h_0^2}{2}\mu_2(K) f^{2,0}(y \mid x) + o\left(h_0^2\right) \approx f(y \mid x), \text{ as } h_0 \ \rightarrow \ 0, \tag{4.2}$$

where $K_{h_0}(u) = K\left(u/h_0\right)/h_0$, $f^{2,0}(y \mid x) = \partial^2/\partial y^2 f(y \mid x)$, and $\approx$ denotes an approximation by ignoring the higher terms. Note that $Y_t^*(y) = K_{h_0}\left(y - Y_t\right)$ can be regarded as an initial estimate of $f(y \mid x)$ smoothing in the y direction. Also, note that this approximation ignores the higher order terms $O\left(h_0^j\right)$ for $j \geq 2$, since they are negligible if $h_0 = o(h)$, where h is the

[1]To name just a few, see Lejeune and Sarda (1988), Truong (1989), Samanta (1989), and Chaudhuri (1991) for the iid errors, Roussas (1969b, 1991) for Markovian processes, and Truong and Stone (1992) and Boente and Fraiman (1995) for mixing sequences.

bandwidth used in smoothing in the x direction (see (4.3) below). Therefore, the smoothing in the y direction is not important in the context of this subject so that intuitively, it should be under-smoothed. Thus, the left hand side of (4.2) can be regraded as a nonparametric regression of the observed variable $Y_t^*(y)$ versus X_t and the local linear (or polynomial) fitting scheme of Fan and Gijbels (1996) can be applied to here. This leads us to consider the following locally weighted least squares regression problem:

$$\sum_{t=1}^{n} \{Y_t^*(y) - a - b\,(X_t - x)\}^2 W_h\,(X_t - x)\,, \tag{4.3}$$

where $W(\cdot)$ is a kernel function and $h = h(n) > 0$ is the bandwidth satisfying $h \to 0$ and $nh \to \infty$ as $n \to \infty$, which controls the amount of smoothing used in the estimation. Note that (4.3) involves two kernels $K(\cdot)$ and $W(\cdot)$. This is the reason of calling "double kernel".

Minimizing the above locally weighted least squares in (4.3) with respect to a and b, we obtain the locally weighted least squares estimator of $f(y \mid x)$, denoted by $\widehat{f}(y \mid x)$, which is $\widehat{a}$. From Fan and Gijbels (1996) or Fan et al. (1996), $\widehat{f}(y \mid x)$ can be re-expressed as a linear estimator form as

$$\widehat{f}_{ll}(y \mid x) = \sum_{t=1}^{n} W_{ll,t}(x,h) Y_t^*(y),$$

where with $S_{n,j}(x) = \sum_{t=1}^{n} W_h\,(x - X_t)\,(X_t - x)^j$, the weights $\{W_{ll,t}(x,h)\}$ are given by

$$W_{ll,t}(x,h) = \frac{[S_{n,2}(x) - (x - X_t)\,S_{n,1}(x)]\,W_h\,(x - X_t)}{S_{n,0}(x)S_{n,2}(x) - S_{n,1}^2(x)}.$$

Clearly, $\{W_{ll,t}(x,h)\}$ satisfy the so-called discrete moments conditions like (2.6) and (2.14), as follows: for $0 \le j \le 1$,

$$\sum_{t=1}^{n} W_{ll,t}(x,h)\,(X_t - x)^j = \delta_{0,j} = \begin{cases} 1 & \text{if } j = 0 \\ 0 & \text{otherwise} \end{cases} \tag{4.4}$$

based on the least squares theory; see (3.12) of Fan and Gijbels (1996). Note that the estimator $\widehat{f}_{ll}(y \mid x)$ can range outside $[0, \infty)$. The double kernel local linear estimator of $F(y \mid x)$ is constructed (see (8) of Yu and Jones (1998)) by integrating $\widehat{f}_{ll}(y \mid x)$

$$\widehat{F}_{ll}(y \mid x) = \int_{-\infty}^{y} \widehat{f}_{ll}(y \mid x)dy = \sum_{t=1}^{n} W_{ll,t}(x,h) G_{h_0}\,(y - Y_t)\,,$$

where $G(\cdot)$ is the distribution function of $K(\cdot)$ and $G_{h_0}(u) = G\,(u/h_0)$. Clearly, $\widehat{F}_{ll}(y \mid x)$ is continuous and differentiable with respect to y with $\widehat{F}_{ll}(-\infty \mid x) = 0$ and $\widehat{F}_{ll}(\infty \mid x) = 1$.

Note that the differentiability of the estimated distribution function can make the asymptotic analysis much easier for the nonparametric estimators of CVaR and CES (see later).

Although Yu and Jones (1998) showed that the double kernel local linear estimator has some attractive properties such as no boundary effects, design adaptation, and mathematical efficiency; see, e.g., Fan and Gijbels (1996), it has the disadvantage of producing conditional distribution function estimators that are not constrained either to lie between zero and one or to be monotone increasing, which is not good for estimating CVaR if the inverting method is used. In both these respects, the NW method is superior, despite its rather large bias and boundary effects. The properties of positivity and monotonicity are particularly advantageous if the method of inverting conditional distribution estimator is applied to produce the estimator of a conditional quantile or CVaR. To overcome these difficulties, Hall et al. (1999) and Cai (2002a) proposed the WNW estimator based on an empirical likelihood principle, which is designed to possess the superior properties of local linear methods such as bias reduction and no boundary effects, and to preserve the property that the NW estimator is always a distribution function, although it might require more computational efforts since it requires estimating and optimizing additional weights aimed at the bias correction. Cai (2002a) discussed the asymptotic properties of the WNW estimator at both interior and boundary points for the mixing time series under some regularity assumptions and showed that the WNW estimator has a better performance than other competitors. See Cai (2002a) for details. Finally, Cosma et al. (2007) proposed a shape preserving estimation method to estimate cumulative distribution functions and probability density functions using the wavelet methodology for multivariate dependent data and then to estimate a conditional quantile or CVaR.

The WNW estimator of the conditional distribution $F(y \mid x)$ of Y_t given $X_t = x$ is defined by

$$\widehat{F}_{c1}(y \mid x) = \sum_{t=1}^{n} W_{c,t}(x,h) I\left(Y_t \leq y\right), \tag{4.5}$$

where the weights $\{W_{c,t}(x,h)\}$ are given by

$$W_{c,t}(x,h) = \frac{p_t(x) W_h\left(x - X_t\right)}{\sum_{t=1}^{n} p_t(x) W_h\left(x - X_t\right)}, \tag{4.6}$$

and $\{p_t(\mathbf{x})\}$ is chosen to be $p_t(x) = n^{-1}\left\{1 + \lambda\left(X_t - x\right) W_h\left(x - X_t\right)\right\}^{-1} \geq 0$ with λ, a function of data and x, uniquely defined by maximizing the logarithm of the empirical

likelihood

$$L_n(\lambda) = -\sum_{t=1}^{n} \log\left\{1 + \lambda\left(X_t - x\right) W_h\left(x - X_t\right)\right\}$$

subject to the constraints $\sum_{t=1}^{n} p_t(x) = 1$ and the discrete moments conditions in (4.4); that is,

$$\sum_{t=1}^{n} W_{c,t}(x,h)\left(X_t - x\right)^j = \delta_{0,j} \tag{4.7}$$

for $0 \leq j \leq 1$. Also, see Cai (2002a) for details on this aspect. In implementation, Cai (2002a) recommended using the Newton-Raphson scheme to find the root of equation $L_n^{\top}(\lambda) = 0$. Note that $0 \leq \widehat{F}_{c1}(y \mid x) \leq 1$ and it is monotone in y. But $\widehat{F}_{c1}(y \mid x)$ is not continuous in y and of course, not differentiable in y either. Note that under regression setting, Cai (2001) provided a comparison of the local linear estimator and the WNW estimator and discussed the asymptotic minimax efficiency of the WNW estimator.

To accommodate all nice properties (monotonicity, continuity, differentiability, and lying between zero and one) and the attractive asymptotic properties (design adaption, avoiding boundary effects, and mathematical efficiency, see Cai (2002a) for detailed discussions) of both estimators $\widehat{F}_{ll}(y \mid x)$ and $\widehat{F}_{c1}(y \mid x)$ under a unified framework, we propose the following nonparametric estimators for the conditional density function $f(y \mid x)$ and its conditional distribution function $F(y \mid x)$, termed as *weighted double kernel local linear estimation*,

$$\widehat{f}_c(y \mid x) = \sum_{t=1}^{n} W_{c,t}(x,h) Y_t^*(y),$$

where $W_{c,t}(x,h)$ is given in (4.6), and

$$\widehat{F}_c(y \mid x) = \int_{-\infty}^{y} \widehat{f}_c(y \mid x) dy = \sum_{t=1}^{n} W_{c,t}(x,h) G_{h_0}\left(y - Y_t\right). \tag{4.8}$$

Note that if $p_t(x)$ in (4.6) is a constant for all t, or $\lambda = 0$, then $\widehat{f}_c(y \mid x)$ becomes the classical NW type double kernel estimator used by Scaillet (2005). However, Scaillet (2005) adopted a single bandwidth for smoothing in both the y and x directions. Clearly, $\widehat{f}_c(y \mid x)$ is a probability density function so that $\widehat{F}_c(y \mid x)$ is a cumulative distribution function (monotone, $0 \leq \widehat{F}_c(y \mid x) \leq 1, \widehat{F}_c(-\infty \mid x) = 0$, and $\widehat{F}_c(\infty \mid x) = 1\Big)$. Also, $\widehat{F}_c(y \mid x)$ is continuous and differentiable in y. Further, as expected, it will be shown that like $\widehat{F}_{c1}(y \mid x), \widehat{F}_c(y \mid x)$ has the attractive properties such as no boundary effects, design adaptation, and mathematical efficiency.

4.3.2 Estimation of Conditional VaR and ES

We now are ready to formulate the nonparametric estimators for $\nu_\tau(x)$ and $\mu_\tau(x)$. To this end, from (4.8), $\nu_\tau(x)$ is estimated by inverting the estimated conditional survival distribution $\widehat{S}_c(y \mid x) = 1 - \widehat{F}_c(y \mid x)$, denoted by $\widehat{\nu}_\tau(x)$ and defined as $\widehat{\nu}_\tau(x) = \widehat{S}_c^{-1}(\tau \mid x)$. Note that $\widehat{\nu}_\tau(x)$ always exists since $\widehat{S}_c(\tau \mid x)$ is a survival function itself. Plugging-in $\widehat{\nu}_\tau(x)$ and $\widehat{f}_c(y \mid x)$ into (4.1), we obtain the nonparametric estimation of $\mu_\tau(x)$,

$$\begin{aligned}\widehat{\mu}_\tau(x) &= \tau^{-1}\int_{\widehat{\nu}_\tau(x)}^{\infty} y\widehat{f}_c(y \mid x)dy = \tau^{-1}\sum_{t=1}^{n} W_{c,t}(x,h)\int_{\widehat{\nu}_\tau(x)}^{\infty} yK_{h_0}\left(y - Y_t\right)dy \\ &= \tau^{-1}\sum_{t=1}^{n} W_{c,t}(x,h)\left[Y_t\bar{G}_{h_0}\left(\widehat{\nu}_\tau(x) - Y_t\right) + h_0 G_{1,h_0}\left(\widehat{\nu}_\tau(x) - Y_t\right)\right], \qquad (4.9)\end{aligned}$$

where $\bar{G}(u) = 1 - G(u)$ is the survival function of $K(\cdot)$, $G_{1,h_0}(u) = G_1\left(u/h_0\right)$, and $G_1(u) = \int_u^\infty vK(v)dv$. Note that as mentioned earlier, $\widehat{\nu}_\tau(x)$ in (4.9) can be an any consistent estimator.

4.4 Asymptotic Theories

4.4.1 Assumptions

Before we proceed with the asymptotic properties of the proposed nonparametric estimators, we first list all assumptions needed for the asymptotic theory, although some of them might not be the weakest possible. Note that proofs of the asymptotic results presented in this section may be found in Section 4.4.4A with some lemmas and their detailed proofs relegated to Section 4.4.4B. First, we introduce some notation. Let $\alpha(K) = \int_{-\infty}^{\infty} uK(u)\bar{G}(u)du$. Also, for any $j \geq 0$, write

$$l_j(u \mid v) = \mathbb{E}\left[Y_t^j I\left(Y_t \geq u\right) \mid X_t = v\right] = \int_u^\infty y^j f(y \mid v)dy, \quad l_j^{a,b}(u \mid v) = \frac{\partial^{ab}}{\partial u^a \partial v^b} l_j(u \mid v),$$

and $l_j^{a,b}\left(\nu_\tau(x) \mid x\right) = l_j^{a,b}(u \mid v)\Big|_{u=\nu_\tau(x), v=x}$. Clearly, $l_0(u \mid v) = S(u \mid v)$ and $l_1\left(\nu_\tau(x) \mid x\right) = p\mu_\tau(x)$. Finally, $l_j^{1,0}(u \mid v) = -u^j f(u \mid v)$ and $l_j^{2,0}(u \mid v) = -\left[u^j f^{1,0}(u \mid v) + ju^{j-1}f(u \mid v)\right]$. We now list the following regularity conditions.

Assumptions:

(D1) For fixed y and $x, 0 < F(y \mid x) < 1, g(x) > 0$, the marginal density of X_t, and is

continuous at x, and $F(y \mid x)$ has continuous second order derivative with respect to both x and y.

(D2) The kernels $K(\cdot)$ and $W(\cdot)$ are symmetric, bounded, and compactly supported density.

(D3) $h \to 0$ and $n\,h \to \infty$, and $h_0 \to 0$ and $n\,h_0 \to \infty$, as $n \to \infty$.

(D4) Let $f_{1,t}(\cdot,\cdot)$ be the joint density of X_1 and X_t for $t \geq 2$. Assume that $\mid f_{1,t}(u,v) - f(u)f(v) \mid \leq M < \infty$ for all u and v.

(D5) The process $\{(X_t, Y_t)\}$ is a stationary α-mixing with the mixing coefficient satisfying $\alpha(t) = O\left(t^{-(2+\delta)}\right)$ for some $\delta > 0$.

(D6) $n\,h^{1+2/\delta} \to \infty$

(D7) $h_0 = o(h)$.

(D8) Assume that $\mathbb{E}\left(|Y_t|^{\delta} \mid X_t = u\right) \leq M_3 < \infty$ for some $\delta > 2$, in a neighborhood of x.

(D9) Assume that $|g_{1,t}(y_1, y_2 \mid x_1, x_2)| \leq M_1 < \infty$ for all $t \geq 2$, where $g_{1,t}(y_1, y_2 \mid x_1, x_2)$ be the conditional density of Y_1 and Y_t given $X_1 = x_1$ and $X_t = x_2$.

(D10) The mixing coefficient of the α-mixing process $\{(X_t, Y_t)\}_{t=-\infty}^{\infty}$ satisfies $\sum_{t\geq 1} t^a \alpha^{1-2/\delta}(t) < \infty$ for some $a > 1 - 2/\delta$, where δ is given in Assumption D8.

(D11) Assume that there exists a sequence of integers $s_n > 0$ such that $s_n \to \infty, s_n = o\left((nh)^{1/2}\right)$, and $(n/h)^{1/2}\alpha(s_n) \to 0$, as $n \to \infty$.

(D12) There exists $\delta^* > \delta$ such that $\mathbb{E}\left(|Y_t|^{\delta^*} \mid X_t = u\right) \leq M_4 < \infty$ in a neighborhood of $x, \alpha(t) = O\left(t^{-\theta^*}\right)$, where δ is given in Assumption D8, $\theta^* \geq \delta^*\delta / \{2(\delta^* - \delta)\}$, and $n^{1/2-\delta/4}h^{\delta/\delta^*-1/2-\delta/4} = O(1)$.

Remark 4.1: *Note that Assumptions D1 - D5 and D8 - D12 are used commonly in the literature of time series data; see, e.g., Masry and Fan (1997) and Cai (2001). Note that α-mixing imposed in Assumption D5 is weaker than β-mixing in Hall et al. (1999) and ρ-mixing in Fan et al. (1996). Because D6 is satisfied by the bandwidths of optimal size (i.e., $h \approx n^{-1/5}$) if $\delta > 1/2$, we do not concern ourselves with such refinements. Indeed, Assumptions D1 - D6 are also required in Cai (2002a). Assumption A7 means that the initial*

step bandwidth should be chosen as small as possible so that the bias from the initial step can be ignored. Since the common technique - truncation approach for time series data is not applicable to our setting; see, e.g., Masry and Fan (1997), the purpose of Assumption D12 is to use the moment inequality. If $\alpha(t)$ decays geometrically, then Assumptions D11 and D12 are satisfied automatically. Note that Assumptions D10, D11, and D12 are stronger than Assumptions D5 and D6. This is not surprising because the higher moments involved, the faster decaying rate of $\alpha(\cdot)$ is required. Finally, Assumptions D8 - D12 are also imposed in Cai (2001).

4.4.2 Asymptotic Properties for Conditional PDF and CDF

First, we investigate the asymptotic behaviors of $\widehat{f}_c(y \mid x)$, including the asymptotic normality stated in the following theorem.

Theorem 4.1: *Under Assumptions D1 - D6 with h in D3 and D6 replaced by $h_0 h$, we have*

$$\sqrt{n\, h_0 h}\left[\widehat{f}_c(y \mid x) - f(y \mid x) - B_f(y \mid x) + o_p(h^2 + h_0^2)\right] \xrightarrow{d} N\left\{0, \sigma_f^2(y \mid x)\right\},$$

where the asymptotic bias is

$$B_f(y \mid x) = \frac{h^2}{2}\mu_2(W) f^{0,2}(y \mid x) + \frac{h_0^2}{2}\mu_2(K) f^{2,0}(y \mid x),$$

and the asymptotic variance is $\sigma_f^2(y \mid x) = \nu_0(K)\nu_0(W) f(y \mid x)/g(x)$.

Remark 4.2: *The asymptotic results for $\widehat{f}_c(y \mid x)$ in Theorem 4.1 are similar to those for $\widehat{f}_{ll}(y \mid x)$ in Fan et al. (1996) for the ρ-mixing sequence, which is stronger than α-mixing, but as mentioned earlier, $\widehat{f}_{ll}(y \mid x)$ is not always a probability density function. The asymptotic bias and variance are intuitively expected. The bias comes from the approximations in both x and y directions and the variance is from the local conditional variance in the density estimation setting, which is $f(y \mid x)$.*

Next, we study the asymptotic behaviors for $\widehat{S}_c(y \mid x)$ at both interior and boundary points. Similar to Theorem 4.1 for $\widehat{f}_c(y \mid x)$, we have the following asymptotic normality for $\widehat{S}_c(y \mid x)$

Theorem 4.2: *Under Assumptions D1 - D6, we have*

$$\sqrt{nh}\left[\widehat{S}_c(y \mid x) - S(y \mid x) - B_S(y \mid x) + o_p(h^2)\right] \xrightarrow{d} N\left\{0, \sigma_S^2(y \mid x)\right\},$$

where the asymptotic bias is given by

$$B_S(y \mid x) = \frac{h^2}{2}\mu_2(W)S^{0,2}(y \mid x) - \frac{h_0^2}{2}\mu_2(K)f^{1,0}(y \mid x),$$

and the asymptotic variance is $\sigma_S^2(y \mid x) = \nu_0(W)S(y \mid x)[1 - S(y \mid x)]/g(x)$. *In particular, if Assumption D7 holds true, then,*

$$\sqrt{nh}\left[\widehat{S}_c(y \mid x) - S(y \mid x) - \frac{h^2}{2}\mu_2(W)S^{0,2}(y \mid x) + o_p(h^2)\right] \xrightarrow{d} N\left\{0, \sigma_S^2(y \mid x)\right\}.$$

Remark 4.3: *Note that the asymptotic results for* $\widehat{S}_c(y \mid x)$ *in Theorem 4.2 are analogous to those for* $\widehat{S}_{ll}(y \mid x) = 1 - \widehat{F}_{ll}(y \mid x)$ *in Yu and Jones (1998) for the iid data, but as mentioned previously,* $\widehat{F}_{ll}(y \mid x)$ *is not always a distribution function. A comparison of* $B_s(y \mid x)$ *with the asymptotic bias for* $\widehat{S}_{c1}(y \mid x)$ *(see Theorem 1 in Cai (2002a)), it reveals that there is an extra term* $\frac{h_0^2}{2}f^{1,0}(y \mid x)\mu_2(K)$ *in the asymptotic bias expression* $B_s(y \mid x)$ *due to the vertical smoothing in the y direction. Also, there is an extra term in the asymptotic variance (see* (4.20)*). These extra terms are carried over from the initial estimate but they can be ignored if the bandwidth at the initial step is taken to be a higher order than the bandwidth at the smoothing step.*

Remark 4.4: *It is important to examine the performance of* $\widehat{S}_c(y \mid x)$ *by considering the asymptotic mean squared error (AMSE). Theorem 4.2 concludes that the AMSE of* $\widehat{S}_c(y \mid x)$ *is*

$$\begin{aligned} AMSE\left(\widehat{S}_c(y \mid x)\right) =& \frac{\left\{h^2\mu_2(W)S^{0,2}(y \mid x) - h_0^2\mu_2(K)f^{1,0}(y \mid x)\right\}^2}{4} \\ &+ \frac{1}{nh}\frac{\nu_0(W)S(y \mid x)[1 - S(y \mid x)]}{g(x)}. \end{aligned} \tag{4.10}$$

By minimizing AMSE in (4.10) *and taking* $h_0 = o(h)$, *therefore, we obtain the optimal bandwidth given by*

$$h_{opt\ ,S}(y \mid x) = \left[\frac{\nu_0(W)S(y \mid x)[1 - S(y \mid x)]}{\left\{\mu_2(W)S^{0,2}(y \mid x)\right\}^2 g(x)}\right]^{1/5} n^{-1/5}.$$

Therefore, the optimal rate of the AMSE of $\widehat{S}_c(y \mid x)$ *is* $n^{-4/5}$.

As for the boundary behavior of the WDKLL estimator, we can follow Cai (2002a) to establish a similar result for $\widehat{S}_c(y \mid x)$ like Theorem 2 in Cai (2002a). Without loss of generality, we consider the left boundary point $x = ch, 0 < c < 1$. From Fan et al. (1994), we

take $W(\cdot)$ to have support $[-1,1]$ and $g(\cdot)$ to have support $[0,1]$. Then, under Assumptions D1 - D7, by following the same proof as that for Theorem 4.2 and using the second assertion in Lemma 4.1, although not straightforward, we can show that

$$\sqrt{nh}\left[\widehat{S}_c(y \mid ch) - S_c(y \mid ch) - B_{S,c}(y) + o_p(h^2)\right] \xrightarrow{d} N\left(0, \sigma^2_{S,c}(y)\right), \tag{4.11}$$

where the asymptotic bias term is given by $B_{S,c}(y) = h^2\beta_0(c)S^{0,2}(y \mid 0+)/\left[2\beta_1(c)\right]$ and the asymptotic variance is $\sigma^2_{S,c}(y) = \beta_2(0)S(y \mid 0+)[1 - S(y \mid 0+)]/\left[\beta_1^2(c)g(0+)\right]$ with $g(0+) = \lim_{z\downarrow 0} g(z)$

$$\beta_0(c) = \int_{-1}^{c} \frac{u^2 W(u)}{1 - \lambda_c u W(u)} du, \quad \beta_j(c) = \int_{-1}^{c} \frac{W^j(u)}{\{1 - \lambda_c u W(u)\}^j} du, \quad 1 \le j \le 2,$$

and λ_c being the root of equation $L_c(\lambda) = 0$

$$L_c(\lambda) = \int_{-1}^{c} \frac{uW(u)}{1 - \lambda u W(u)} du.$$

Note that the proof of (4.11) is similar to that for Theorem 2 in Cai (2002a) and omitted. Theorem 4.2 and (4.11) reflect two of the major advantages of the WKDLL estimator: (a) the asymptotic bias does not depend on the design density $g(x)$, and indeed it is dependent only on the simple conditional distribution curvature $S^{0,2}(y \mid x)$ and conditional density curvature $f^{1,0}(y \mid x)$; and (b) it has an automatic good behavior at boundaries. See Cai (2002a) for the detailed discussions.

Finally, we remark that if the point 0 were an interior point, then, (4.11) would hold with $c = 1$, which becomes Theorem 4.2. Therefore, Theorem 4.2 shows that the WKDLL estimation has the automatic good behavior at boundaries without the need of the boundary correction.

4.4.3 Asymptotic Theory for CVaR and CES

By the differentiability of $\widehat{S}_c\left(\widehat{\nu}_\tau(x) \mid x\right)$, we use the Taylor expansion and ignore the higher terms to obtain

$$\widehat{S}_c\left(\widehat{\nu}_\tau(x) \mid x\right) = \tau \approx \widehat{S}_c\left(\nu_\tau(x) \mid x\right) - \widehat{f}_c\left(\nu_\tau(x) \mid x\right)\left(\widehat{\nu}_\tau(x) - \nu_\tau(x)\right), \tag{4.12}$$

then, by Theorem 4.1,

$$\widehat{\nu}_\tau(x) - \nu_\tau(x) \approx \left[\widehat{S}_c\left(\nu_\tau(x) \mid x\right) - \tau\right] / \widehat{f}_c\left(\nu_\tau(x) \mid x\right) \approx \left[\widehat{S}_c\left(\nu_\tau(x) \mid x\right) - \tau\right] / f\left(\nu_\tau(x) \mid x\right).$$

As an application of Theorem 4.2, we can establish the following theorem for the asymptotic normality of $\widehat{\nu}_\tau(x)$ but the proof is omitted since it is similar to that for Theorem 4.2.

Theorem 4.3: *Under Assumptions D1 - D6, we have*

$$\sqrt{nh}\left[\widehat{\nu}_\tau(x) - \nu_\tau(x) - B_\nu(x) + o_p(h^2)\right] \xrightarrow{d} N\left\{0, \sigma_\nu^2(x)\right\},$$

where the asymptotic bias is $B_\nu(x) = B_S\left(\nu_\tau(x) \mid x\right)/f\left(\nu_\tau(x) \mid x\right)$ *and the asymptotic variance is* $\sigma_\nu^2(x) = \nu_0(W)\tau(1-\tau)/\left[g(x)f^2\left(\nu_\tau(x) \mid x\right)\right]$. *In particular, if Assumption D7 holds, then,*

$$\sqrt{nh}\left[\widehat{\nu}_\tau(x) - \nu_\tau(x) - \frac{h^2}{2}\frac{S^{0,2}\left(\nu_\tau(x) \mid x\right)}{f\left(\nu_\tau(x) \mid x\right)}\mu_2(W) + o_p(h^2)\right] \xrightarrow{d} N\left\{0, \sigma_\nu^2(x)\right\}.$$

Remark 4.5: *First, as a consequence of Theorem 4.3,* $\widehat{\nu}_\tau(x) - \nu_\tau(x) = O_p(\left(h^2 + h_0^2 + (nh)^{-1/2}\right)$ *so that* $\widehat{\nu}_\tau(x)$ *is a consistent estimator of* $\nu_\tau(x)$ *with a convergence rate. Also, note that the asymptotic results for* $\widehat{\nu}_\tau(x)$ *in Theorem 4.3 are akin to those for* $\widehat{\nu}_{l,p}(x) = \widehat{S}_{ll}^{-1}(\tau \mid x)$ *in Yu and Jones (1998) for the iid data. But in the bias term of Theorem 4.3, the quantity* $S^{0,2}\left(\nu_\tau(x) \mid x\right)/f\left(\nu_\tau(x) \mid x\right)$, *involving the second derivative of the conditional distribution function with respect to* x, *replaces* $\nu_\tau''(x)$, *the second derivative of the conditional VaR function itself, which is in the bias term of the check function type local linear estimator in Yu and Jones (1998) for the iid data and Cai and Xu (2008) for time series. See Cai and Xu (2008) for details. This is not surprising since the bias comes only from the approximation. The former utilizes the approximation of the conditional distribution function but the later uses the approximation of the conditional VaR function. Finally, Theorems 4.2 and 4.3 imply that if the initial bandwidth* h_0 *is chosen small as possible such as* $h_0 = o(h)$, *the final estimates of* $S(y \mid x)$ *and* $\nu_\tau(x)$ *are not sensitive to the choice of* h_0 *as long as it satisfies Assumption D7. This makes the selection of bandwidths much easier in practice, which will be elaborated later (see Section 4.5.1).*

Remark 4.6: *Similar to Remark 4.5 , we can derive the asymptotic mean squared error for* $\widehat{\nu}_\tau(x)$. *By following Yu and Jones (1998), Theorem 4.3, and* (4.20) *(given in Section 4.6 imply that the AMSE of* $\widehat{\nu}_\tau(x)$ *is given by*

$$\begin{aligned} AMSE\left(\widehat{\nu}_\tau(x)\right) =& \frac{\left\{h^2 S^{0,2}\left(\nu_\tau(x) \mid x\right)\mu_2(W) - h_0^2 f^{1,0}\left(\nu_\tau(x) \mid x\right)\mu_2(K)\right\}^2}{4f^2\left(\nu_\tau(x) \mid x\right)} \\ &+ \frac{1}{nh}\frac{\nu_0(W)\left[\tau(1-\tau) + 2h_0 f\left(\nu_\tau(x) \mid x\right)\alpha(K)\right]}{f^2\left(\nu_\tau(x) \mid x\right)g(x)}. \end{aligned} \tag{4.13}$$

Note that the above result is similar to that in Theorem 1 in Yu and Jones (1998) for the double kernel local linear conditional quantile estimator. But, a comparison of (4.13) *with Theorem 3 in Cai (2002a) for the WNW estimator reveals that* (4.13) *has two extra terms*

(negligible if Assumption D7 is satisfied) due to the vertical smoothing in the y direction, as mentioned previously. By minimizing AMSE in (4.13) *and taking* $h_0 = o(h)$*, therefore, we obtain the optimal bandwidth given by*

$$h_{opt\ ,\nu}(x) = \left[\frac{\nu_0(W)\tau(1-\tau)}{\{\mu_2(W)S^{0,2}\left(\nu_\tau(x) \mid x\right)\}^2 g(x)}\right]^{1/5} n^{-1/5}.$$

Therefore, the optimal rate of the AMSE of $\widehat{\nu}_\tau(x)$ *is* $n^{-4/5}$*. By comparing* $h_{opt,\nu}(x)$ *with* $h_{opt\ ,S}(y \mid x)$*, it turns out that* $h_{opt,\nu}(x)$ *is* $h_{opt,\nu}(y \mid x)$ *evaluated at* $y = \nu_\tau(x)$*. Therefore, the best choice of the bandwidth for estimating* $S_c(y \mid x)$ *can be used for estimating* $\nu_\tau(x)$*.*

Remark 4.7: *Similar to* (4.11)*, one can establish the asymptotic result at boundaries for* $\nu_\tau(x)$ *as follows, one can show that under Assumption D7,*

$$\sqrt{nh}\left[\widehat{\nu}_\tau(ch) - \nu_\tau(ch) - B_{\nu,c} + o_p(h^2)\right] \xrightarrow{d} N\left(0, \sigma^2_{\nu,c}\right),$$

where the asymptotic bias is $B_{\nu,c} = h^2\beta_2(c)S^{0,2}\left(\nu_\tau(0+) \mid 0+\right)/\left[2\beta_1(c)f\left(\nu_\tau(0+) \mid 0+\right)\right]$ *and the asymptotic variance is* $\sigma^2_{\nu,c} = \beta_0(0)\tau(1-\tau)/\left[\beta_1^2(c)f^2\left(\nu_\tau(0+) \mid 0+\right)g(0+)\right]$*. Clearly,* $\widehat{\nu}_\tau(x)$ *inherits all good properties from the WDKLL estimator of* $S_c(y \mid x)$*. Note that the above result can be established by using the second assertion in Lemma 4.1 and following the same lines along with those used in the proof of Theorem 4.2 and omitted.*

Finally, we examine the asymptotic behavior for $\widehat{\mu}_\tau(x)$ at both interior and boundary points. First, we establish the following theorem for the asymptotic normality for $\widehat{\mu}_\tau(x)$ when x is an interior point.

Theorem 4.4: *Under Assumptions D1 - D4 and D9 - D12, we have*

$$\sqrt{n\,h}\left[\widehat{\mu}_\tau(x) - \mu_\tau(x) - B_\mu(x) + o_p(h^2 + h_0^2)\right] \xrightarrow{d} N\left\{0, \sigma^2_\mu(x)\right\},$$

where the asymptotic bias is $B_\mu(x) = B_{\mu,0}(x) + \frac{h_0^2}{2}\mu_2(K)\tau^{-1}f\left(\nu_\tau(x) \mid x\right)$ *with*

$$B_{\mu,0}(x) = \frac{h^2}{2}\mu_2(W)\tau^{-1}\left[l_1^{0,2}\left(\nu_\tau(x) \mid x\right) - \nu_\tau(x)S^{0,2}\left(\nu_\tau(x) \mid x\right)\right],$$

and the asymptotic variance is

$$\sigma^2_\mu(x) = \frac{\nu_0(W)}{\tau\, g(x)}\left[\tau^{-1}l_2\left(\nu_\tau(x) \mid x\right) - \tau\mu^2_\tau(x) + (1-\tau)\nu_\tau(x)\left\{\nu_\tau(x) - 2\mu_\tau(x)\right\}\right].$$

In particular, if Assumption D7 holds true, then,

$$\sqrt{nh}\left[\widehat{\mu}_\tau(x) - \mu_\tau(x) - B_{\mu,0}(x) + o_p(h^2 + h_0^2)\right] \xrightarrow{d} N\left\{0, \sigma^2_\mu(x)\right\}.$$

Remark 4.8: *First, Theorem 4.4 concludes that* $\widehat{\mu}_\tau(x) - \mu_\tau(x) = O_p(\left(h^2 + h_0^2 + (nh)^{-1/2}\right)$ *so that* $\widehat{\mu}_\tau(x)$ *is a consistent estimator of* $\mu_\tau(x)$ *with a convergence rate. Also, note that the asymptotic results in Theorem 4.4 imply that* $\widehat{\mu}_\tau(x)$ *is a consistent estimator for* $\mu_\tau(x)$ *with a convergence rate* $(nh)^{-1/2}$*. Further, note that although the asymptotic variance* $\sigma_\mu^2(x)$ *is the same as that in Scaillet (2005) for* $\widetilde{\mu}_\tau(x)$*, Scaillet (2005) did not provide an expression for the asymptotic bias term like* $B_\mu(x)$ *in the first result or* $B_{\mu,0}(x)$ *in the second conclusion in Theorem 4.4. Clearly, the second term in the asymptotic bias expression is carried over from the y direction smoothing at the initial step and it is negligible if Assumption D7 is satisfied. Clearly, Assumption D7 implies that* $B_\mu(x)$ *becomes* $B_{\mu,0}(x)$*.*

Remark 4.9: *Like Remark 4.5, the AMSE for* $\widehat{\mu}_\tau(x)$ *can be derived in the same manner. It follows from Theorem 4.4 that the AMSE of* $\widehat{\mu}_\tau(x)$ *is given by*

$$AMSE(\widehat{\mu}_\tau(x)) = \frac{1}{nh}\sigma_\mu^2(x) + \left\{ B_{\mu,0}(x) + \frac{h_0^2}{2}\mu_2(K)\tau^{-1} f\left(\nu_\tau(x) \mid x\right) \right\}^2. \tag{4.14}$$

Under Assumption D7, minimizing AMSE in (4.14) *with respect to h yields the optimal bandwidth given by*

$$h_{opt\ ,\mu}(x) = \left[\frac{\sigma_\mu(x)}{\mu_2(W)\tau^{-1}\left\{ l_1^{0,2}\left(\nu_\tau(x) \mid x\right) - \nu_\tau(x) S^{0,2}\left(\nu_\tau(x) \mid x\right) \right\}} \right]^{2/5} \quad n^{-1/5}.$$

Therefore, as expected, the optimal rate of the AMSE of $\widehat{\mu}_\tau(x)$ *is* $n^{-4/5}$*.*

Finally, we offer the asymptotic results for $\widehat{\mu}_\tau(x)$ at the left boundary point $x = ch$. By the same fashion, one can show that under Assumption D7,

$$\sqrt{nh}\left[\widehat{\mu}_\tau(ch) - \mu_\tau(ch) - B_{\mu,c} + o_p(h^2)\right] \xrightarrow{d} N\left(0, \sigma_{\mu,c}^2\right),$$

where the asymptotic bias is

$$B_{\mu,c} = h^2\beta_2(c)\tau^{-1}\left[l_1^{0,2}\left(\nu_\tau(0+) \mid 0+\right) - \nu_\tau(0+) S^{0,2}\left(\nu_\tau(0+) \mid 0+\right)\right] / \left[2\beta_1(c)\right],$$

and the asymptotic variance is

$$\sigma_{\mu,c}^2 = \frac{\beta_0(0)}{\tau\beta_1^2(c)g(0+)}\left[\tau^{-1} l_2\left(\nu_\tau(0+) \mid 0+\right) - \tau\mu_\tau^2(0+) + (1-\tau)\nu_\tau(0+)\left\{\nu_\tau(0+) - 2\mu_\tau(0+)\right\}\right].$$

Note that the proof of the above result can be carried over by using the second assertion in Lemma 4.1 and following the same lines along with those used in the proof of Theorem 4.4 and

omitted. Next, we consider the comparison of the performance of the WDKLL estimation $\widehat{\mu}_\tau(x)$ with the NW type kernel estimator $\widetilde{\mu}_\tau(x)$ as in Scaillet (2005). To this effect, it is not very difficult to derive the asymptotic results for the NW type kernel estimator but the proof is omitted since it is along the same line with the proof of Theorem 4.2. See Scaillet (2005) for the results at the interior point. Under some regularity conditions, it can be shown although tediously (see Cai (2002a) for details) that at the left boundary $x = ch$, the asymptotic bias term for the NW type kernel estimator $\widetilde{\mu}_\tau(x)$ is of the order h by comparing to the order h^2 for the WDKLL estimate (see $B_{\mu,c}$ above). This shows that the WDKLL estimate does not suffer from boundary effects but the NW type kernel estimator estimate does. This is another advantage of the WDKLL estimator over the WW type kernel estimator $\widetilde{\mu}_\tau(x)$.

4.4.4 Theoretical Proofs

A. Proofs of Theorems

In this section, we present the proofs of Theorems 4.1 - 4.4. First, we list two lemmas. The proof of Lemma 4.1 can be found in Cai (2002a) and the proof of Lemma 4.2 is relegated to the end of this section.

Lemma 4.1: Under Assumptions D1 - D5, we have

$$\lambda = -h\lambda_0 \{1 + o_p(1)\} \quad \text{and} \quad p_t(x) = n^{-1} b_t(x) \{1 + o_p(1)\},$$

where $\lambda_0 = \mu_2(W) g^\top(x) / [2\mu_2(W^2) g(x)]$ and $b_t(x) = [1 - h\lambda_0 (X_t - x) W_h (x - X_t)]^{-1}$. Further, we have

$$p_t(ch) = n^{-1} b_t^c(ch) \{1 + o_p(1)\},$$

where $b_t^c(x) = [1 + \lambda_c (X_t - x) K_h (x - X_t)]^{-1}$.

Lemma 4.2: Under Assumptions D1 - D5, we have, for any $j \geq 0$,

$$J_j = n^{-1} \sum_{t=1}^{n} c_t(x) \left(\frac{X_t - x}{h} \right)^j = g(x)\mu_j(W) + o_p(\left(h^2\right),$$

where $c_t(x) = b_t(x) W_h (x - X_t)$.

Before we start to provide the main steps for proofs of theorems. First, it follows from Lemmas 4.1 and 4.2 that

$$W_{c,t}(x, h) \approx \frac{b_t(x) W_h (x - X_t)}{\sum_{t=1}^{n} b_t(x) W_h (x - X_t)} \approx n^{-1} g^{-1}(x) b_t(x) W_h (x - X_t) = \frac{c_t(x)}{n g(x)}. \tag{4.15}$$

Now we embark on the proofs of the theorems.

Proof of Theorem 4.1. By (4.7), we decompose $\widehat{f}_c(y \mid x) - f(y \mid x)$ into three parts as follows

$$\widehat{f}_c(y \mid x) - f(y \mid x) \equiv I_1 + I_2 + I_3, \tag{4.16}$$

where with $\varepsilon_{t,1} = Y_t^*(y) - \mathbb{E}\left(Y_t^*(y) \mid X_t\right)$,

$$I_1 = \sum_{t=1}^{n} \varepsilon_{t,1} W_{c,t}(x,h), \quad I_2 = \sum_{t=1}^{n} \left[\mathbb{E}\left(Y_t^*(y) \mid X_t\right) - f\left(y \mid X_t\right)\right] W_{c,t}(x,h),$$

and

$$I_3 = \sum_{t=1}^{n} \left[f\left(y \mid X_t\right) - f(y \mid x)\right] W_{c,t}(x,h).$$

An application of the Taylor expansion, (4.7), (4.15), and Lemmas 4.1 and 4.2 gives

$$\begin{aligned} I_3 &= \sum_{t=1}^{n} \frac{1}{2} f^{0,2}(y \mid x) W_{c,t}(x,h) \left(X_t - x\right)^2 + o_p(\left(h^2\right) \\ &= \frac{1}{2} g^{-1}(x) f^{0,2}(y \mid x) n^{-1} \sum_{t=1}^{n} c_t(x) \left(X_t - x\right)^2 + o_p(\left(h^2\right) \\ &= \frac{h^2}{2} \mu_2(W) f^{0,2}(y \mid x) + o_p(\left(h^2\right). \end{aligned}$$

By (4.2) and following the same steps as in the proof of Lemma 4.2, we have

$$I_2 = \frac{h_0^2 \mu_2(K)}{2g(x)} n^{-1} \sum_{t=1}^{n} f^{2,0}\left(y \mid X_t\right) c_t(x) + o_p(\left(h_0^2 + h^2\right) = \frac{h_0^2}{2} \mu_2(K) f^{2,0}(y \mid x) + o_p(\left(h_0^2 + h^2\right).$$

Therefore,

$$I_2 + I_3 = \frac{h^2}{2} \mu_2(W) f^{0,2}(y \mid x) + \frac{h_0^2}{2} \mu_2(K) f^{2,0}(y \mid x) + o_p(\left(h^2 + h_0^2\right) = B_f(y \mid x) + o_p(\left(h^2 + h_0^2\right).$$

Thus, (4.16) becomes

$$\begin{aligned} &\sqrt{nh_0 h}\left[\widehat{f}_c(y \mid x) - f(y \mid x) - B_f(y \mid x) + o_p(\left(h^2 + h_0^2\right)\right] = \sqrt{nh_0 h} I_1 \\ =& g^{-1}(x) I_4 \left\{1 + o_p(1)\right\} \xrightarrow{d} N\left\{0, \sigma_f^2(y \mid x)\right\} \end{aligned}$$

where $I_4 = \sqrt{h_0 h / n} \sum_{t=1}^{n} \varepsilon_{t,1} c_t(x)$. This, together with Lemma 4.3 in Section 4.4.4B, therefore, proves the theorem. □

Proof of Theorem 4.2. Similar to (4.16), we have

$$\widehat{S}_c(y \mid x) - S(y \mid x) \equiv I_5 + I_6 + I_7, \tag{4.17}$$

where with $\varepsilon_{t,2} = \bar{G}_{h_0}(y - Y_t) - \mathbb{E}\left(\bar{G}_{h_0}(y - Y_t) \mid X_t\right)$,

$$I_5 = \sum_{t=1}^{n} \varepsilon_{t,2} W_{c,t}(x,h), \quad I_6 = \sum_{t=1}^{n} \left[\mathbb{E}\left\{\bar{G}_{h_0}(y - Y_t) \mid X_t\right\} - S(y \mid X_t)\right] W_{c,t}(x,h),$$

and

$$I_7 = \sum_{t=1}^{n} \left[S(y \mid X_t) - S(y \mid x)\right] W_{c,t}(x,h).$$

Similar to the analysis of I_2, by the Taylor expansion, (4.7), and Lemmas Lemmas 4.1 and 4.2, we have

$$\begin{aligned} I_7 &= \sum_{t=1}^{n} \frac{1}{2} S^{0,2}(y \mid x) W_{c,t}(x,h) (X_t - x)^2 + o_p(\left(h^2\right) \\ &= \frac{1}{2} S^{0,2}(y \mid x) g^{-1}(x) n^{-1} \sum_{t=1}^{n} c_t(x) (X_t - x)^2 + o_p(\left(h^2\right) \\ &= \frac{h^2}{2} \mu_2(W) S^{0,2}(y \mid x) + o_p(\left(h^2\right). \end{aligned}$$

To evaluate I_6, first, we consider the following

$$\begin{aligned} \mathbb{E}\left[\bar{G}_{h_0}(y - Y_t) \mid X_t = x\right] &= \int_{-\infty}^{\infty} K(u) S(y - h_0 u \mid x)\, du \\ =& S(y \mid x) + \frac{h_0^2}{2} \mu_2(K) S^{2,0}(y \mid x) + o\left(h_0^2\right) \\ =& S(y \mid x) - \frac{h_0^2}{2} \mu_2(K) f^{1,0}(y \mid x) + o\left(h_0^2\right). \end{aligned} \tag{4.18}$$

By (4.18) and following the same arguments as in the proof of Lemma 4.2, we have

$$I_6 = -\frac{h_0^2 \mu_2(K)}{2g(x)} n^{-1} \sum_{t=1}^{n} f^{1,0}(y \mid X_t) c_t(x) + o_p(\left(h_0^2 + h^2\right) = -\frac{h_0^2}{2} \mu_2(K) f^{1,0}(y \mid x) + o_p(\left(h_0^2 + h^2\right).$$

Therefore,

$$I_6 + I_7 = \frac{h^2}{2} \mu_2(W) S^{0,2}(y \mid x) - \frac{h_0^2}{2} \mu_2(K) f^{1,0}(y \mid x) + o_p(\left(h^2 + h_0^2\right) = B_S(y \mid x) + o_p(\left(h^2 + h_0^2\right),$$

so that by (4.17),

$$\sqrt{nh}\left[\widehat{S}_c(y \mid x) - S(y \mid x) - B_S(y \mid x) + o_p(\left(h^2 + h_0^2\right)\right] = \sqrt{nh} I_5.$$

Clearly, to accomplish the proof of theorem, it suffices to establish the asymptotic normality of $\sqrt{nh}I_5$. To this end, first, we compute $\mathrm{Var}\left(\varepsilon_{t,2} \mid X_t = x\right)$. Note that

$$\begin{aligned}
&\mathbb{E}\left[\bar{G}_{h_0}^2\left(y - Y_t\right) \mid X_t = x\right] = \int_{-\infty}^{\infty} \bar{G}_{h_0}^2(y-u) f(u \mid x) du \\
=&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} K\left(u_1\right) K\left(u_2\right) S\left(\max\left(y - h_0 u_1, y - h_0 u_2\right) \mid x\right) du_1 du_2 \\
=&S(y \mid x) + 2h_0\alpha(K) f(y \mid x) + O\left(h_0^2\right), \qquad (4.19)
\end{aligned}$$

which, in conjunction with (4.18), implies that

$$\mathrm{Var}\left(\varepsilon_{t,2} \mid X_t = x\right) = S(y \mid x)[1 - S(y \mid x)] + 2h_0\alpha(K) f(y \mid x) + o\left(h_0\right).$$

This, together with the fact that

$$\mathrm{Var}\left(\varepsilon_{t,2} c_t(x)\right) = \mathbb{E}\left[c_t^2(x)\mathbb{E}\left\{\varepsilon_{t,2}^2 \mid X_t\right\}\right] = \mathbb{E}\left[c_t^2(x)\mathrm{Var}\left(\varepsilon_{t,2} \mid X_t\right)\right],$$

leads to

$$h\mathrm{Var}\left\{\varepsilon_{t,2} c_t(x)\right\} = \nu_0(W) g(x)\left[S(y \mid x)\{1 - S(y \mid x)\} + 2h_0\alpha(K) f(y \mid x)\right] + o\left(h_0\right).$$

Now, since $|\varepsilon_{t,2}| \leq 1$, by following the same arguments as those used in the proofs of Lemmas 4.2 and 4.3 in Section 4.4.4B (or Lemma 1 and Theorem 1 in Cai (2002a)), we can show although tediously that

$$\mathrm{Var}\left(I_8\right) = \sigma_S^2(y \mid x) g^2(x) + 2\nu_0(W) h_0 \alpha(K) f(y \mid x) g(x) + o\left(h_0\right), \qquad (4.20)$$

where $I_8 = \sqrt{h/n}\sum_{t=1}^{n} \varepsilon_{t,2} c_t(x)$, and

$$\sqrt{nh} I_5 = g^{-1}(x) I_8\left\{1 + o_p(1)\right\} \xrightarrow{d} N\left\{0, \sigma_S^2(y \mid x)\right\}$$

This completes the proof of Theorem 4.2. □

Proof of Theorem 4.4. Similar to (4.12), we use the Taylor expansion and ignore the higher terms to obtain

$$\begin{aligned}
&\int_{\widehat{\nu}_\tau(x)}^{\infty} y K_{h_0}\left(y - Y_t\right) dy \approx \int_{\nu_\tau(x)}^{\infty} y K_{h_0}\left(y - Y_t\right) dy - \nu_\tau(x) K_{h_0}\left(\nu_\tau(x) - Y_t\right)\left[\widehat{\nu}_\tau(x) - \nu_\tau(x)\right] \\
&= Y_t \bar{G}_{h_0}\left(\nu_\tau(x) - Y_t\right) - \nu_\tau(x) K_{h_0}\left(\nu_\tau(x) - Y_t\right)\left[\widehat{\nu}_\tau(x) - \nu_\tau(x)\right] + h_0 G_{1,h_0}\left(\nu_\tau(x) - Y_t\right).
\end{aligned}$$

Plugging the above into (4.9) leads to

$$\tau\widehat{\mu}_\tau(x) \approx \widehat{\mu}_{\tau,1}(x) + I_9, \qquad (4.21)$$

where

$$\widehat{\mu}_{\tau,1}(x) = \sum_{t=1}^{n} W_{c,t}(x,h) Y_t \bar{G}_{h_0}\left(\nu_\tau(x) - Y_t\right) - \nu_\tau(x)\widehat{f}_c\left(\nu_\tau(x) \mid x\right)\left[\widehat{\nu}_\tau(x) - \nu_\tau(x)\right],$$

which will be shown later to be the source of both the asymptotic bias and variance, and

$$I_9 = h_0 \sum_{t=1}^{n} W_{c,t}(x,h) G_{1,h_0}\left(\nu_\tau(x) - Y_t\right),$$

which will be shown to contribute only the asymptotic bias (see Lemma 4.4 in Section (4.7). From (4.12) and (4.8),

$$\widehat{f}_c\left(\nu_\tau(x) \mid x\right)\left[\widehat{\nu}_\tau(x) - \nu_\tau(x)\right] \approx \sum_{t=1}^{n} W_{c,t}(x,h)\left\{\bar{G}_{h_0}\left(\nu_\tau(x) - Y_t\right) - \tau\right\}.$$

Therefore, by (4.15),

$$\begin{aligned}
\widehat{\mu}_{\tau,1}(x) &= \sum_{t=1}^{n} W_{c,t}(x,h)\left[\left\{Y_t - \nu_\tau(x)\right\}\bar{G}_{h_0}\left(\nu_\tau(x) - Y_t\right) - \tau\nu_\tau(x)\right] \\
&= \sum_{t=1}^{n} W_{c,t}(x,h)\varepsilon_{t,3} + \sum_{t=1}^{n} W_{c,t}(x,h)\mathbb{E}\left\{\zeta_t(x) \mid X_t\right\} \\
&\approx g^{-1}(x)n^{-1}\sum_{t=1}^{n}\varepsilon_{t,3}c_t(x) + \sum_{t=1}^{n} W_{c,t}(x,h)\mathbb{E}\left\{\zeta_t(x) \mid X_t\right\} \\
&\equiv \widehat{\mu}_{\tau,2}(x) + \widehat{\mu}_{\tau,3}(x),
\end{aligned}$$

where $\zeta_t(x) = \left[Y_t - \nu_\tau(x)\right]\bar{G}_{h_0}\left(\nu_\tau(x) - Y_t\right) + p\nu_\tau(x)$ and $\varepsilon_{t,3} = \zeta_t(x) - \mathbb{E}\left\{\zeta_t(x) \mid X_t\right\}$. Next, we derive the asymptotic bias and variance for $\widehat{\mu}_{\tau,1}(x)$. Indeed, we will show that asymptotic bias of $\widehat{\mu}_\tau(x)$ comes from both $\widehat{\mu}_{\tau,3}(x)$ and I_9, and the asymptotic variance for $\widehat{\mu}_{\tau,1}(x)$ is only from $\widehat{\mu}_{\tau,2}(x)$. First, we consider $\widehat{\mu}_{\tau,3}(x)$. Now, it is easy to see by the Taylor expansion that

$$\begin{aligned}
&\mathbb{E}\left[Y_t\bar{G}_{h_0}\left(\nu_\tau(x) - Y_t\right) \mid X_t = v\right] = \int_{-\infty}^{\infty} K(u)du \int_{\nu_\tau(x)-h_0u}^{\infty} yf(y \mid v)dy \\
=& \int_{-\infty}^{\infty} l_1\left(\nu_\tau(x) - h_0u \mid v\right)K(u)du = l_1\left(\nu_\tau(x) \mid v\right) + \frac{h_0^2}{2}\mu_2(K)l_1^{2,0}\left(\nu_\tau(x) \mid v\right) + o\left(h_0^2\right) \\
=& l_1\left(\nu_\tau(x) \mid v\right) - \frac{h_0^2}{2}\mu_2(K)\left[\nu_\tau(x)f^{1,0}\left(\nu_\tau(x) \mid v\right) + f\left(\nu_\tau(x) \mid x\right)\right] + o\left(h_0^2\right),
\end{aligned}$$

which, in conjunction with (4.18), leads to

$$\zeta(v) = \mathbb{E}\left[\zeta_t(x) \mid X_t = v\right] = A\left(\nu_\tau(x) \mid v\right) - \frac{h_0^2}{2}\mu_2(K)f\left(\nu_\tau(x) \mid v\right) + o\left(h_0^2\right), \tag{4.22}$$

where $A\left(\nu_\tau(x) \mid v\right) = l_1\left(\nu_\tau(x) \mid v\right) - \nu_\tau(x)\left[S\left(\nu_\tau(x) \mid v\right) - \tau\right]$. It is easy to verify that $A\left(\nu_\tau(x) \mid v\right) = \mathbb{E}\left[\{Y_t - \nu_\tau(x)\} I\left(Y_t \geq \nu_\tau(x)\right) \mid X_t = v\right] + p\nu_\tau(x)$, $A\left(\nu_\tau(x) \mid x\right) = p\mu_\tau(x)$, and $A^{0,2}\left(\nu_\tau(x) \mid x\right) = l_1^{0,2}\left(\nu_\tau(x) \mid x\right) - \nu_\tau(x) S^{0,2}\left(\nu_\tau(x) \mid x\right)$. Therefore, by (4.22), the Taylor expansion, and (4.7), $\widehat{\mu}_{\tau,3}(x)$ becomes

$$\widehat{\mu}_{\tau,3}(x) = \sum_{t=1}^{n} W_{c,t}(x,h)\zeta\left(X_t\right) = \zeta(x) + \frac{1}{2}\zeta''(x)\sum_{t=1}^{n} W_{c,t}(x,h)\left(X_t - x\right)^2 + o_p(\left(h^2\right).$$

Further, by Lemmas 4.1 and 4.2,

$$\begin{aligned}\widehat{\mu}_{\tau,3}(x) &= \zeta(x) + \frac{h^2}{2}\mu_2(W)\zeta''(x) + o_p(\left(h^2\right) \\ &= p\mu_\tau(x) + \frac{h^2}{2}\mu_2(W)A^{0,2}\left(\nu_\tau(x) \mid x\right) - \frac{h_0^2}{2}\mu_2(K) f\left(\nu_\tau(x) \mid x\right) + o_p(\left(h_0^2\right).\end{aligned}$$

This, in conjunction with Lemma 4.4 in Section 4.7 concludes that

$$\widehat{\mu}_{\tau,3}(x) + I_9 = p\left[\mu_\tau(x) + B_\mu(x)\right] + o_p(\left(h^2 + h_0^2\right),$$

so that by (4.21),

$$\widehat{\mu}_{\tau,1}(x) - \tau\left[\mu_\tau(x) + B_\mu(x)\right] = \widehat{\mu}_{\tau,2}(x) + o_p(\left(h^2 + h_0^2\right),$$

and

$$\widehat{\mu}_\tau(x) - \mu_\tau(x) - B_\mu(x) = \tau^{-1}\widehat{\mu}_{\tau,2}(x) + o_p(\left(h^2 + h_0^2\right).$$

Finally, by Lemma 4.5 in Section 4.7, we have

$$\sqrt{nh}\left[\widehat{\mu}_\tau(x) - \mu_\tau(x) - B_\mu(x) + o_p(\left(h^2 + h_0^2\right)\right] = \frac{1}{\tau g(x)} I_{10}\left\{1 + o_p(1)\right\} \xrightarrow{d} N\left\{0, \sigma_\mu^2(x)\right\},$$

where $I_{10} = \sqrt{h/n}\sum_{t=1}^{n}\varepsilon_{t,3}c_t(x)$. Thus, we prove the theorem. □

B. Proofs of Lemmas

In this section, we present the proofs of Lemmas 4.2 - 4.5. Note that we use the same notation as that in Sections 4.2 - 4.4. Also, throughout this section, we denote a generic constant by C, which may take different values at different appearances.

Proof of Lemmas 4.2. Let $\xi_t = c_t(x)\left(X_t - x\right)^j / h^j$. It is easy to verify by the Taylor expansion that

$$\mathbb{E}\left(J_j\right) = \mathbb{E}\left(\xi_t\right) = \int \frac{v^j W(v) g(x - hv)}{1 + h\lambda_0 v W(v)} dv = g(x)\mu_j(W) + O\left(h^2\right), \tag{4.23}$$

and

$$\mathbb{E}\left(\xi_t^2\right) = h^{-1}\int \frac{v^{2j}W^2(v)g(x-hv)}{\left[1+h\lambda_0 vW(v)\right]^2}dv = O\left(h^{-1}\right).$$

Also, by the stationarity, a straightforward manipulation yields

$$n\text{Var}\left(J_j\right) = \text{Var}\left(\xi_1\right) + \sum_{t=2}^{n} l_{n,t}\text{Cov}\left(\xi_1,\xi_t\right), \tag{4.24}$$

where $l_{n,t} = 2(n-t+1)/n$. Now decompose the second term on the right hand side of (4.24) into two terms as follows

$$\sum_{t=2}^{n}\left|\text{Cov}\left(\xi_1,\xi_t\right)\right| = \sum_{t=2}^{d_n}(\cdots) + \sum_{t=d_n+1}^{n}(\cdots) \equiv J_{j1} + J_{j2}, \tag{4.25}$$

where $d_n = O\left(h^{-1/(1+\delta/2)}\right)$. For J_{j1}, it follows by Assumption D4 that $\left|\text{Cov}\left(\xi_1,\xi_t\right)\right| \leq C$, so that $J_{j1} = O\left(d_n\right) = o\left(h^{-1}\right)$. For J_{j2}, Assumption D2 implies that $\left|(X_t-x)^j W_h\left(x-X_t\right)\right| \leq Ch^{j-1}$, so that $|\xi_t| \leq Ch^{-1}$. Then, it follows from the Davydov's inequality (see, e.g., Lemma 1.1) that $\left|\text{Cov}\left(\xi_1,\xi_{t+1}\right)\right| \leq Ch^{-2}\alpha(t)$, which, together with Assumption D5, implies that

$$J_{j2} \leq Ch^{-2}\sum_{t\geq d_n}\alpha(t) \leq Ch^{-2}d_n^{-(1+\delta)} = o\left(h^{-1}\right).$$

This, together with (4.24) and (4.25), therefore implies that $\text{Var}\left(J_j\right) = O\left((nh)^{-1}\right) = o(1)$. This completes the proof of the lemma. □

Lemma 4.3: Under Assumptions D1 - D6 with h in D3 and D6 replaced by $h\,h_0$, we have

$$I_4 = \sqrt{\frac{h_0 h}{n}}\sum_{t=1}^{n}\varepsilon_{t,1}c_t(x) \xrightarrow{d} N\left\{0,\sigma_f^2(y\mid x)g^2(x)\right\}.$$

Proof of Lemmas 4.3. It follows by using the same lines as those used in the proof of Lemma 4.2 and Theorem 1 in Cai (2002a), omitted. The outline is described as follows. First, similar to the proof of Lemma 4.2, it is easy to see that

$$\text{Var}\left(I_4\right) = h_0 h\text{Var}\left(\varepsilon_{t,1}c_t(x)\right) + h_0 h\sum_{t=2}^{n} l_{n,t}\text{Cov}\left(\varepsilon_{1,1}c_1(x),\varepsilon_{t,1}c_t(x)\right). \tag{4.26}$$

Next, we compute $\text{Var}\left(\varepsilon_{t,1}\mid X_t = x\right)$. Note that

$$h_0\mathbb{E}\left[Y_t^*(y)^2 \mid X_t = x\right] = \int_{-\infty}^{\infty} K^2(u)f\left(y-h_0u\mid x\right)du = \nu_0(K)f(y\mid x) + O\left(h_0^2\right),$$

which, together with the fact that

$$\text{Var}\left(\varepsilon_{t,1}c_t(x)\right) = \mathbb{E}\left[c_t^2(x)\mathbb{E}\left\{\varepsilon_{t,1}^2 \mid X_t\right\}\right] = \mathbb{E}\left[c_t^2(x)\text{Var}\left(\varepsilon_{t,1} \mid X_t\right)\right]$$

and (4.2), implies that

$$hh_0\text{Var}\left(\varepsilon_{t,1}c_t(x)\right) = \nu_0(K)\nu_0(W)f(y \mid x)g(x) + O\left(h_0^2\right) = \sigma_f^2(y \mid x)g^2(x) + O\left(h_0^2\right).$$

As for the second term on the right hand side of (4.26), similar to (4.25), it is decomposed into two summons. By using Assumption D4 for the first summon and using the Davydov's inequality and Assumption D5 to the second summon, we can show that the second term on the right hand side of (4.26) goes to zero as n goes to infinity. Thus, $\text{Var}\left(I_4\right) \to \sigma_f^2(y \mid x)g^2(x)$ by (4.26). To show the normality, we employ Doob's small-block and large-block technique; see, e.g., Ibragimov and Linnik (1971). Namely, partition $\{1, \cdots, n\}$ into $2q_n + 1$ subsets with large-block of size $r_n = \left\lfloor (nhh_0)^{1/2} \right\rfloor$ and small-block of size $s_n = \left\lfloor (nhh_0)^{1/2} / \log n \right\rfloor$, where $q_n = \lfloor n/\left(r_n + s_n\right) \rfloor$ with $\lfloor x \rfloor$ denoting the integer part of x. By following the same steps as in the proof of Theorem 1 in Cai (2002a), we can accomplish the rest of proofs: the summands for the large-blocks are asymptotically independent, two summands for the small-blocks are asymptotically negligible in probability, and the standard Lindeberg-Feller conditions hold for the summands for the large-blocks. See Cai (2002a) for details. So, the proof of the lemma is complete. □

Lemma 4.4: Under Assumptions D1 - D6, we have

$$I_9 = h_0 \sum_{t=1}^{n} W_{c,t}(x,h)G_{1,h_0}\left(\nu_\tau(x) - Y_t\right) = h_0^2\mu_2(K)f\left(\nu_\tau(x) \mid x\right) + o_p(\left(h_0^2\right).$$

Proof of Lemmas 4.4. Define $\xi_{t,1} = c_t(x)G_{1,h_0}\left(\nu_\tau(x) - Y_t\right)$. Then, by Lemma 4.1, $I_9 = I_{10}\left\{1 + o_p(1)\right\}$, where $I_{10} = g^{-1}(x)h_0\sum_{t=1}^{n}\xi_{t,1}/n$. Similar to (4.23),

$$\begin{aligned}
\mathbb{E}\left(\xi_{t,1}\right) &= \mathbb{E}\left[c_t(x)\mathbb{E}\left\{G_{1,h_0}\left(\nu_\tau(x) - Y_t\right) \mid X_t\right\}\right] \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{K(u)W(v)uS\left(\nu_\tau(x) - h_0u\right) \mid x)\, g(x - hv)}{1 + h\lambda_0 vW(v)} dudv \\
&= h_0\mu_2(K)f\left(\nu_\tau(x) \mid x\right)g(x) + O\left(h_0h^2\right),
\end{aligned}$$

and

$$\mathbb{E}\left(\xi_{t,1}^2\right) = \mathbb{E}\left[b_t^2(x)W_h^2\left(x - X_t\right)\mathbb{E}\left\{G_{1,h_0}^2\left(\nu_\tau(x) - Y_t\right) \mid X_t\right\}\right] = O\left(h_0/h\right),$$

so that $\text{Var}\left(\xi_{t,1}\right) = O\left(h_0/h\right)$. By following the same arguments in the derivation of $\text{Var}\left(J_j\right)$ in Lemma 4.2, one can show that $\text{Var}\left(I_{10}\right) = O\left((nh)^{-1}\right) = o(1)$. This proves the lemma. □

Lemma 4.5: Under Assumptions D1 - D4 and D9 - D12, we have Under Assumptions D1 - D6, we have

$$I_{10} = \sqrt{\frac{h}{n}} \sum_{t=1}^{n} \varepsilon_{t,3} c_t(x) \xrightarrow{d} N\left\{0, \tau^2 g^2(x) \sigma_\mu^2(x)\right\}.$$

Proof of Lemmas 4.5. It follows by using the same lines as those used in the proof of Lemma 4.1 and Theorem 1 in Cai (2001), omitted. The main idea is as follows. First, similar to the proof of Lemmas 4.2 and 4.3, we will show by Assumptions D8 - D10 that

$$\text{Var}\left(I_{10}\right) \rightarrow \tau^2 \sigma_\mu^2(x) g^2(x). \tag{4.27}$$

Finally, we need to compute $\text{Var}\left(\varepsilon_{t,3} c_t(x)\right)$. Since

$$\text{Var}\left(\varepsilon_{t,3} c_t(x)\right) = \mathbb{E}\left[c_t^2(x) \mathbb{E}\left\{\varepsilon_{t,3}^2 \mid X_t\right\}\right] = \mathbb{E}\left[c_t^2(x) \text{Var}\left(\zeta_t(x) \mid X_t\right)\right],$$

then, we first need to calculate $\text{Var}\left(\zeta_t(x) \mid X_t\right)$. To this effect, by (4.22),

$$\begin{aligned}
&\text{Var}\left(\zeta_t(x) \mid X_t = v\right) = \text{Var}\left[\left(Y_t - \nu_\tau(x)\right) \bar{G}_{h_0}\left(\nu_\tau(x) - Y_t\right) \mid X_t = v\right] \\
=&\mathbb{E}\left[\left(Y_t - \nu_\tau(x)\right)^2 \bar{G}_{h_0}^2\left(\nu_\tau(x) - Y_t\right) \mid X_t = v\right] - \left[l_1\left(\nu_\tau(x) \mid v\right) - \nu_\tau(x) S\left(\nu_\tau(x) \mid v\right)\right]^2 + O\left(h_0^2\right).
\end{aligned}$$

Similar to (4.19),

$$\begin{aligned}
&\mathbb{E}\left[\left(Y_t - \nu_\tau(x)\right)^2 \bar{G}_{h_0}^2\left(\nu_\tau(x) - Y_t\right) \mid X_t = v\right] = \int_{-\infty}^{\infty} G_{h_0}^2\left(\nu_\tau(x) - y\right)\left(y - \nu_\tau(x)\right)^2 f(y \mid v) dy \\
=&\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} K\left(u_1\right) K\left(u_2\right) \tau\left(\max\left(\nu_\tau(x) - h_0 u_1, \nu_\tau(x) - h_0 u_2\right) \mid v\right) du_1 du_2 \\
=&\tau\left(\nu_\tau(x) \mid v\right) - 2h_0 \tau^{1,0}\left(\nu_\tau(x) \mid v\right) \alpha(K) + O\left(h_0^2\right) = \tau\left(\nu_\tau(x) \mid v\right) + O\left(h_0^2\right),
\end{aligned}$$

since $\tau^{1,0}\left(\nu_\tau(x) \mid v\right) = 0$, where $\tau(u \mid v) = l_2(u \mid v) - 2\nu_\tau(x) l_1(u \mid v) + \nu_\tau^2(x) S(u \mid v)$. Therefore,

$$\text{Var}\left(\zeta_t(x) \mid X_t = v\right) = \text{Var}\left[\left(Y_t - \nu_\tau(x)\right) I\left(Y_t \geq \nu_\tau(x)\right) \mid X_t = v\right] + O\left(h_0^2\right),$$

and

$$h\text{Var}\left(\varepsilon_{t,3} c_t(x)\right) = \nu_0(W) \text{Var}\left[\left(Y_t - \nu_\tau(x)\right) I\left(Y_t \geq \nu_\tau(x)\right) \mid X_t = x\right] g(x) + o(1).$$

Similar to Lemmas 4.2 and 4.3, clearly, we have,

$$\text{Var}\left(I_{10}\right) = h\text{Var}\left(\varepsilon_{t,3} c_t(x)\right) + h \sum_{t=2}^{n} l_{n,t} \text{Cov}\left(\varepsilon_{1,3} c_1(x), \varepsilon_{t,3} c_t(x)\right),$$

and the first term on right hand side of the above equation converges to $\tau^2\sigma_\mu^2(x)g^2(x)$. As for the second term on the right hand side of the above equation, similar to (4.25), it is decomposed into two summons. By using Assumptions D4 and D9 for the first summon and using the Davydov's inequality and Assumptions D5 and D10 to the second summon, we can show that the second term on the right hand side of the above equation goes to zero as n goes to infinity. Thus, (4.27) holds. To show the normality, we employ Doob's small-block and large-block technique; see, e.g., Ibragimov and Linnik (1971). Namely, partition $\{1, \cdots, n\}$ into $2q_n+1$ subsets with large-block of size r_n and small-block of size s_n, where s_n is given in Assumption D11, $q_n = \lfloor n/(r_n + s_n)\rfloor$, and $r_n = \left\lfloor (nh)^{1/2}/\gamma_n \right\rfloor$ with γ_n satisfying followings: γ_n is a sequence of positive numbers $\gamma_n \to \infty$ such that $\gamma_n s_n/\sqrt{nh} \to 0$ and $\gamma_n(n/h)^{1/2}\alpha(s_n) \to 0$ by Assumption D11. By following the same steps as in the proof of Theorem 1 in Cai (2001) and using Assumption D12, we can accomplish the rest of proofs: the summands for the large-blocks are asymptotically independent, two summands for the small-blocks are asymptotically negligible in probability, and the standard Lindeberg-Feller conditions hold for the summands for the large-blocks. See Cai (2001) for details. Therefore, the lemma is proved. □

4.5 Empirical Examples

To illustrate the proposed methods, we consider two simulated examples and two real data examples on stock index returns and security returns. Throughout this section, the Epanechnikov kernel $K(u) = 0.75\left(1-u^2\right)_+$is used and bandwidths are selected as described in the next section. Note that the computer codes for the following examples are available upon request.

4.5.1 Bandwidth Selection

With the basic model at hand, one must address the important bandwidth selection issue, as the quality of the curve estimates depends sensitively on the choice of the bandwidth. For practitioners, it is desirable to have a convenient and effective data-driven rule. However, almost nothing has been done so far about this problem in the context of estimating $\nu_\tau(x)$ and $\mu_\tau(x)$ although there are some results available in the literature in other contexts for some specific purposes.

As indicated earlier, the choice of the initial bandwidth h_0 is not very sensitive to the

final estimation but it needs to be specified. First, we use a very simple idea to choose h_0. As mentioned previously, the WNW method involves only one bandwidth in estimating the conditional distribution and VaR. Because the WNW estimate is a linear smoother (see (4.5)), we recommend using the optimal bandwidth selector, the so-called nonparametric AIC proposed by Cai and Tiwari (2000), to select the bandwidth, called $\widetilde{h}$. Then we take $0.1 \times \widetilde{h}$ or smaller as the initial bandwidth h_0. For the given h_0, we can select h as follows. According to (4.8), $\widehat{F}_c(\cdot \mid \cdot)$ is a linear estimator so that the nonparametric AIC selector of Cai and Tiwari (2000) can be applied here to select the optimal bandwidth for $\widehat{F}_c(\cdot\,\cdot\,\cdot)$, denoted by h_S. As mentioned at the end of Remark 6 , the bandwidth for $\widehat{\nu}_\tau(x)$ is the same as that for $\widehat{F}_c(\cdot \mid \cdot)$ so that it is simply to take h_S as h_ν. From (4.9), $\widehat{\mu}_\tau(x)$ is a linear estimator too for given $\widehat{\nu}_\tau(x)$. Therefore, by the same token, the nonparametric AIC selector is applied to selecting h_μ for $\widehat{\mu}_\tau(x)$. This simple approach is used in our implementation in the next sections.

4.5.2 Simulated Examples

In the simulated examples, we demonstrate the finite sample performance of the estimators in terms of the mean absolute deviation error. For example, the MADE for $\widehat{\mu}_\tau(x)$ is defined as

$$\mathcal{E}_{\mu_\tau} = \frac{1}{n_0} \sum_{k=1}^{n_0} \left| \widehat{\mu}_\tau \left(x_k \right) - \mu_\tau \left(x_k \right) \right|,$$

where $\{x_k\}_{k=1}^{n_0}$ are the pre-determined regular grid points. Similarly, we can define the MADE for $\widehat{\nu}_\tau(x)$, denoted by $\mathcal{E}_{\nu_\tau}$.

Example 4.1: We consider an ARCH type model with $X_t = Y_{t-1}$,

$$Y_t = 0.9 \sin \left(2.5 X_t \right) + \sigma \left(X_t \right) \varepsilon_t,$$

where $\sigma^2(x) = 0.8\sqrt{1.2 + x^2}$ and $\{\varepsilon_t\}$ are the iid standard normal random variables. We consider three sample sizes: $n = 250$ and 500, and 1000 and the experiment is repeated 500 times for each sample size. The mean absolute deviation errors are computed for each sample size and each replication.

The 5% WDKLL and NW estimations are summarized in Figure 4.1 for CVaR and in Figure 4.2 for CES. For each n, the Box-plots of $500\mathcal{E}_{\nu_\tau}$-values of the WDKLL and NW estimations are plotted in Figure 4.1(d) for CVaR and in Figure 4.2 (d) for CES.

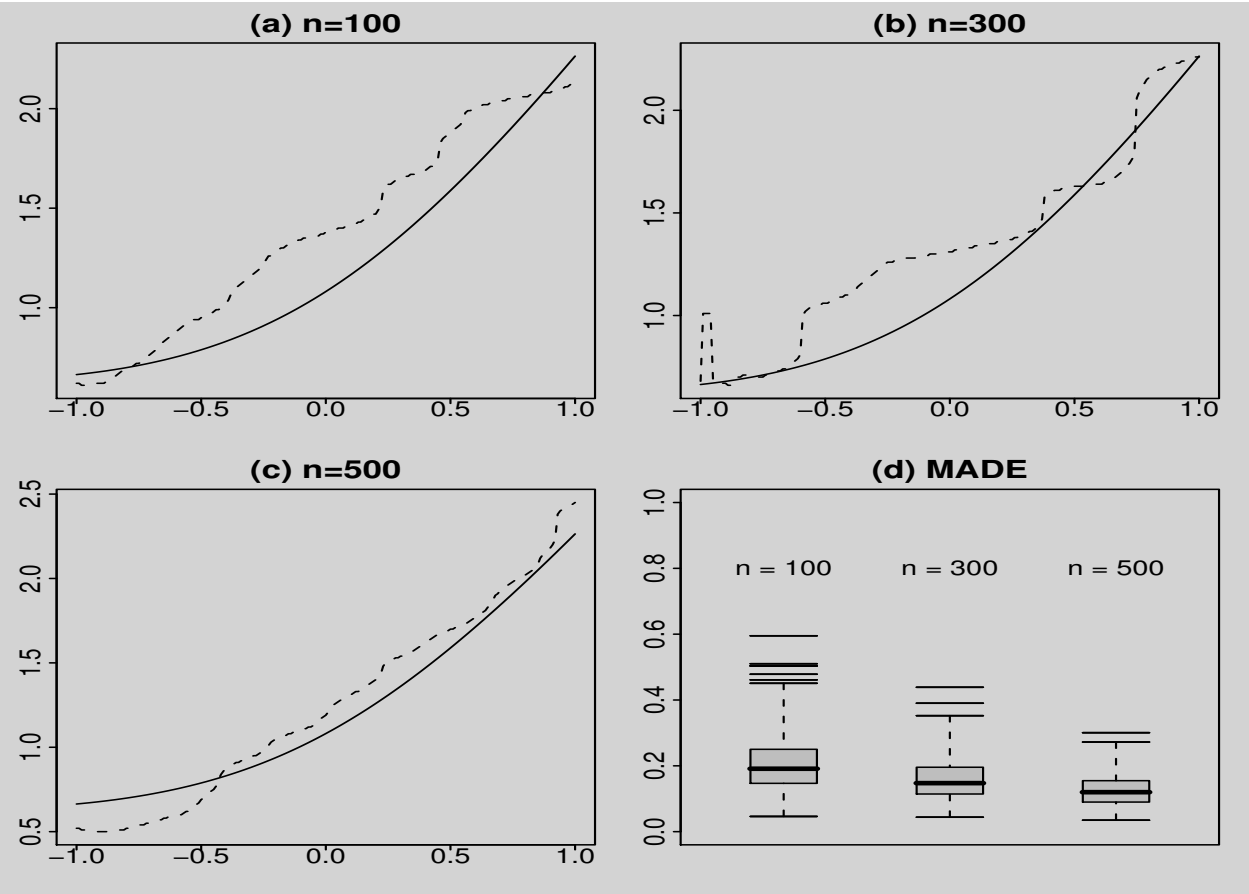

Figure 4.1: Simulation results for Example 4.1 when $p = 0.05$. Displayed in (a) - (c) are the true CVaR functions (solid lines), the estimated WDKLL CVaR functions (dashed lines), and the estimated NW CVaR functions (dotted lines) for $n = 250, 500$ and 1000, respectively. Box-plots of the 500 MADE values for both the WDKLL and NW estimations of CVaR are plotted in (d).

From Figures 4.1(d) and 4.2(d), we can observe that the estimation becomes stable as the sample size increases for both the WDKLL and NW estimators. This is in line with our asymptotic theory that the proposed estimators are consistent. Further, it is obvious that the MADEs of the WDKLL estimator are smaller than those for the NW estimator. This indicates that our WDKLL estimator has smaller bias than that for the NW estimator. This implies that the overall performance of the WDKLL estimator should be better than that for the NW estimator.

Figures $4.1(a) - (c)$ for $n = 250, 500$ and 1000, respectively, display the true CVaR function (solid line) $\nu_\tau(x) = 0.9\sin(2.5x) + \sigma(x)\Phi^{-1}(1-\tau)$, where $\Phi(\cdot)$ is the standard normal distribution function, together with the dashed and dotted lines representing the

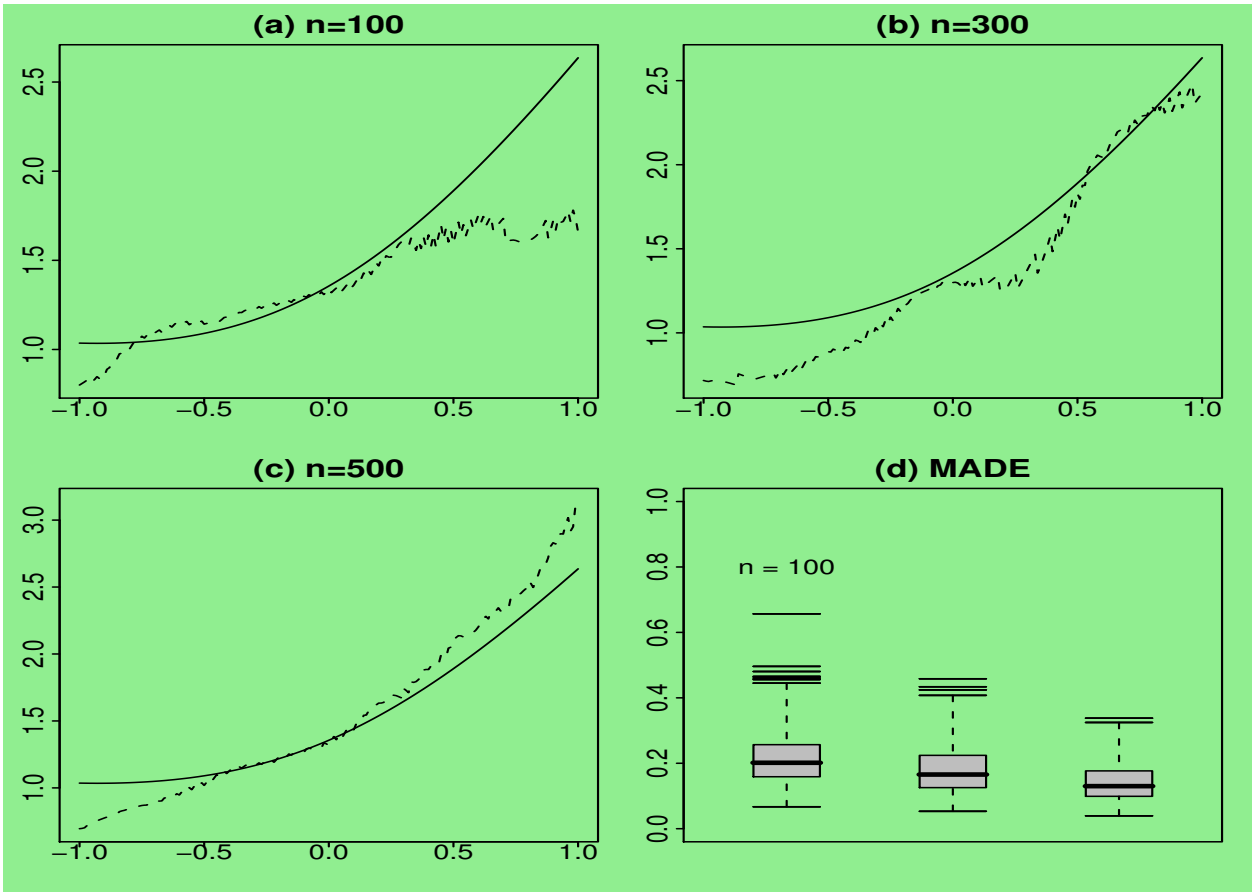

Figure 4.2: Simulation results for Example 4.1 when $p = 0.05$. Displayed in (a) - (c) are the true CES functions (solid lines), the estimated WDKLL CES functions (dashed lines), and the estimated NW CES functions (dotted lines) for $n = 250, 500$ and 1000, respectively. Box-plots of the 500 MADE values for both the WDKLL and NW estimations of CES are plotted in (d).

proposed WDKLL (dashed) and NW (dotted) estimates of CVaR, respectively, which are computed based on a typical sample. The typical sample is selected in such a way that its $\mathcal{E}_{\nu_\tau}$ value is equal to the median in the 500 replications. From Figures 4.1$(a) - (c)$, we can observe that both the estimated curves are closer to the true curve as n increases and the performance of the WDKLL estimator is better than that for the NW estimator, especially at boundaries.

In Figures 4.2(a)-(c), the true CES function $\mu_\tau(x) = 0.9\sin(2.5x)p + \sigma(x)\mu_1\left(\Phi^{-1}(1-\tau)\right)$ is displayed by the solid line, where $\mu_1(t) = \int_t^\infty u\phi(u)du$ and $\phi(\cdot)$ is the standard normal distribution density function, and the dashed and dotted lines present the proposed WDKLL (dashed) and NW (dotted) estimates of CES, respectively, from a typical sample. The

typical sample is selected in such a way that its $\mathcal{E}_{\mu_\tau}$-value is equal to the median in the 500 replications. We can conclude from Figures 4.2$(a)-(c)$ that the CES estimator has a similar performance as that for the CVaR estimator.

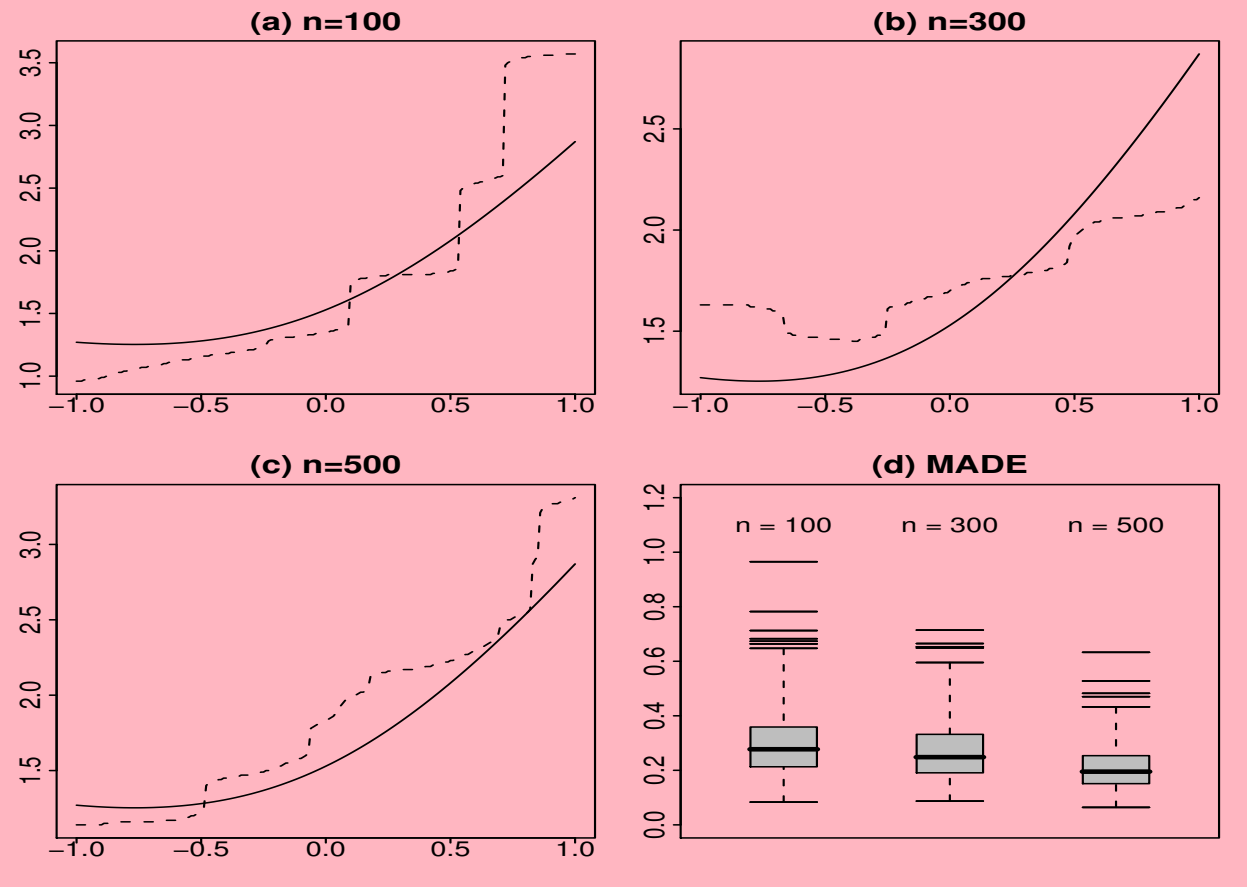

Figure 4.3: Simulation results for Example 4.1 when $p = 0.01$. Displayed in (a) - (c) are the true CVaR functions (solid lines), the estimated WDKLL CVaR functions (dashed lines), and the estimated NW CVaR functions (dotted lines) for $n = 250, 500$ and 1000 , respectively. Box-plots of the 500 MADE values for both WDKLL and NW estimation of the conditional VaR are plotted in (d).

The 1% WDKLL and NW estimates of CVaR and CES are computed under the same setting and they are displayed in Figures 4.3 and 4.4, respectively. Similar conclusions to those for the 5% estimates can be observed. But it is not surprising to see that the performance of the 1% CVaR and CES estimates is not good as that for the 5% estimates due to the sparsity of data.

Example 4.2: In the above example, we consider only the case when X_t is one-dimensional. In this example, we consider the multivariate situation, i.e. X_t consists of two lagged vari-

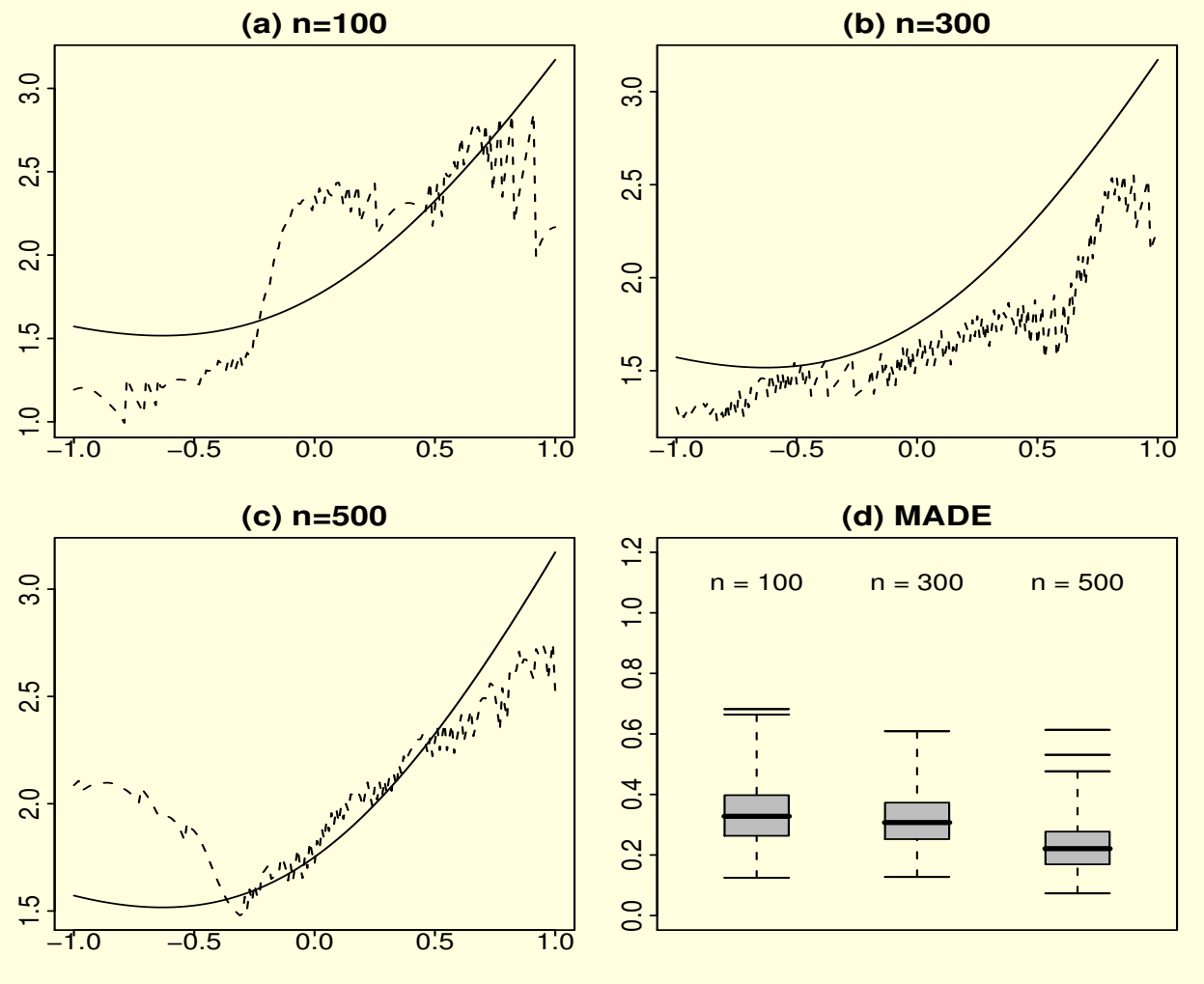

Figure 4.4: Simulation results for 4.1 when $p = 0.01$. Displayed in (a) - (c) are the true CES functions (solid lines), the estimated WDKLL CES functions (dashed lines), and the estimated NW CES functions (dotted lines) for $n = 250, 500$ and 1000, respectively. Box-plots of the 500 MADE values for both the WDKLL and NW estimations of CVaR are plotted in (d).

ables: $X_{t1} = Y_{t-1}$ and $X_{t2} = Y_{t-2}$. The data generating model is given below:

$$Y_t = m\left(X_t\right) + \sigma\left(X_t\right)\varepsilon_t,$$

where $m(x) = 0.63x_1 - 0.47x_2, \sigma^2(x) = 0.5 + 0.23x_1^2 + 0.3x_2^2$, and $\{\varepsilon_t\}$ are iid generated from $N(0,1)$. Three sample sizes: $n = 200, 400$, and 600 , are considered here. For each sample size, we replicate the design 500 times. Here we present only the Box-plots of the 500 MADEs for the CVaR and CES estimates in Figure 4.5. Figure 4.5(a) displays the Box-plots of the $500\mathcal{E}_{\nu_\tau}$-values of the WDKLL and NW estimates of CVaR and the Box-plots of the $500\mathcal{E}_{\mu_\tau}$-values of the WDKLL and NW estimates of CES are given in Figure 4.5(b). From

Figures 4.5(a) and (b), it is visually verified that both WDKLL and NW estimations become stable as the sample size increases and the performance of the WDKLL estimator is better than that for the NW estimator.

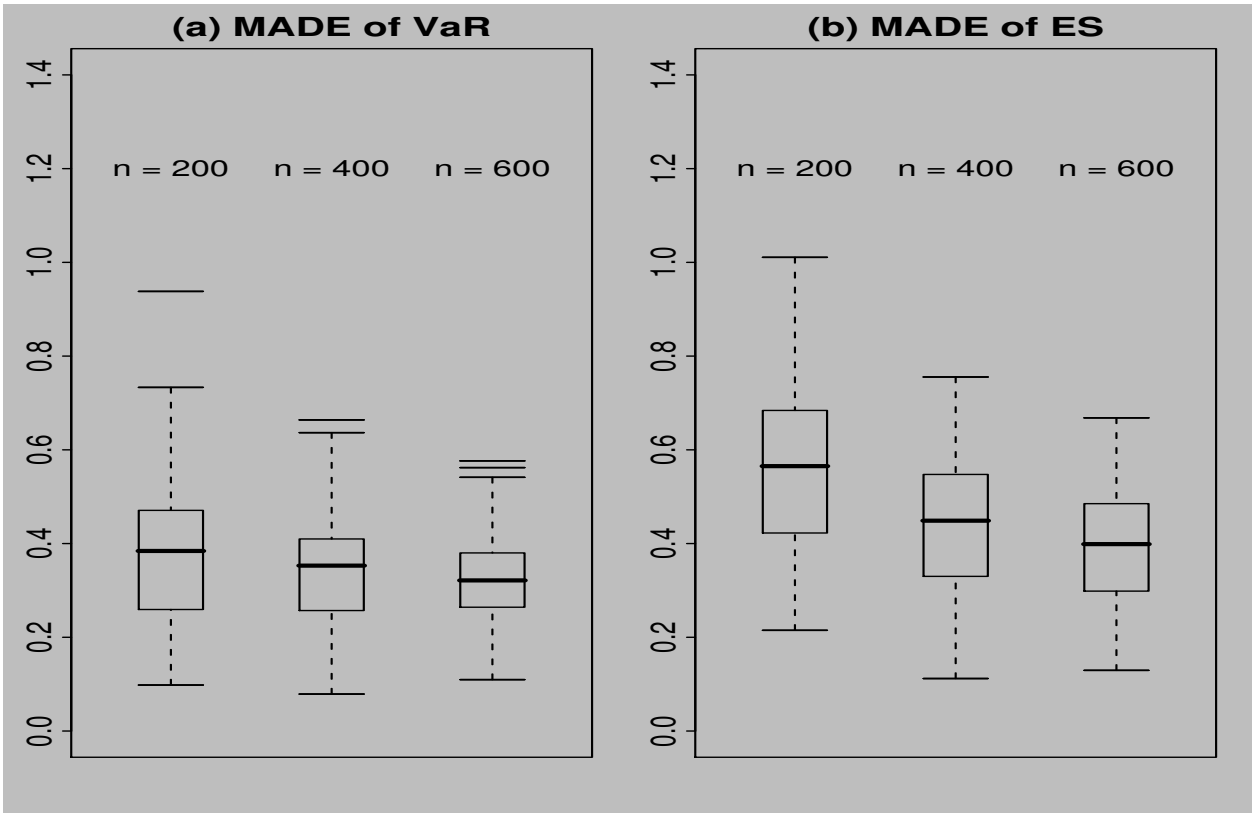

Figure 4.5: Simulation results for Example 4.2 when $p = 0.05$. (a) Box-plots of MADEs for both the WDKLL and NW estimates for CVaR. (b) Box-plots of MADEs for Both the WDKLL and NW estimates for CES.

4.5.3 Real Examples

Example 4.3: Now, we illustrate our proposed methodology by considering a real data set on Dow Jones Industrials (DJI) index returns. We took a sample of 1801 daily prices from DJI index, from November 3, 1998 to January 3, 2006, and computed the daily returns as 100 times the difference of the log of prices. Let Y_t be the daily negative log return (log loss) of DJI and X_t be the first lagged variable of Y_t. The estimators proposed in this chapter are used to estimate the 5% CVaR and CES functions. The estimation results are shown in Figure 4.6 for the 5% CVaR estimate in Figure 4.6(a) and the 5% CES estimate in Figure 4.6(b). Both CVaR and CES estimates exhibit a U-shape, which corresponds to the so-called

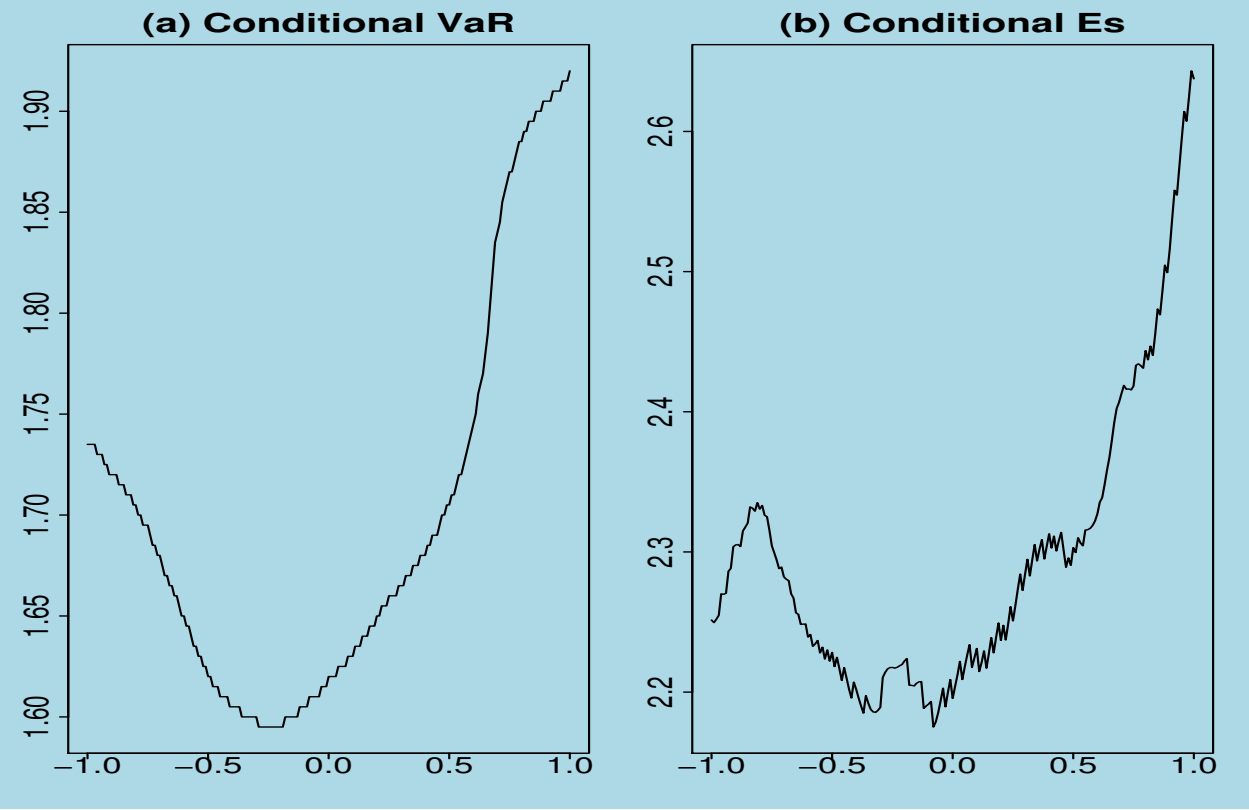

Figure 4.6: (a) 5% CVaR estimate for DJI index. (b) 5% CES estimate for DJI index.

"volatility smile". Therefore, the risk tends to be lower when the lagged log loss of DJI is close to the empirical average and larger otherwise. We can also observe that the curves are asymmetric. This may indicate that the DJI is more likely to fall down if there was a loss within the last day than there was a same amount positive return.

Example 4.4: We apply the proposed methods to estimate the conditional value-at-risk and expected shortfall of the International Business Machine Co. (NYSE: IBM) security returns. The data are daily prices recorded from March 1, 1996 to April 6, 2005. We use the same method to calculate the daily returns as in Example 4.3. In order to estimate the value-at-risk of a stock return, generally, the information set X_t may contain a market index of corresponding capitalization and type, the industry index, and the lagged values of stock return. For this example, Y_t is the log loss of IBM stock returns and only two variables are chosen as information set for the sake of simplicity. Let X_{t1} be the first lagged variable of Y_t and X_{t2} denote the first lagged daily log loss of Dow Jones Industrials (DJI) index. Our main results from the estimation of the model are summarized in Figure 4.7. The surfaces of the estimators of IBM returns are given in Figure 4.7(a) for CVaR and in Figure 4.7(b)

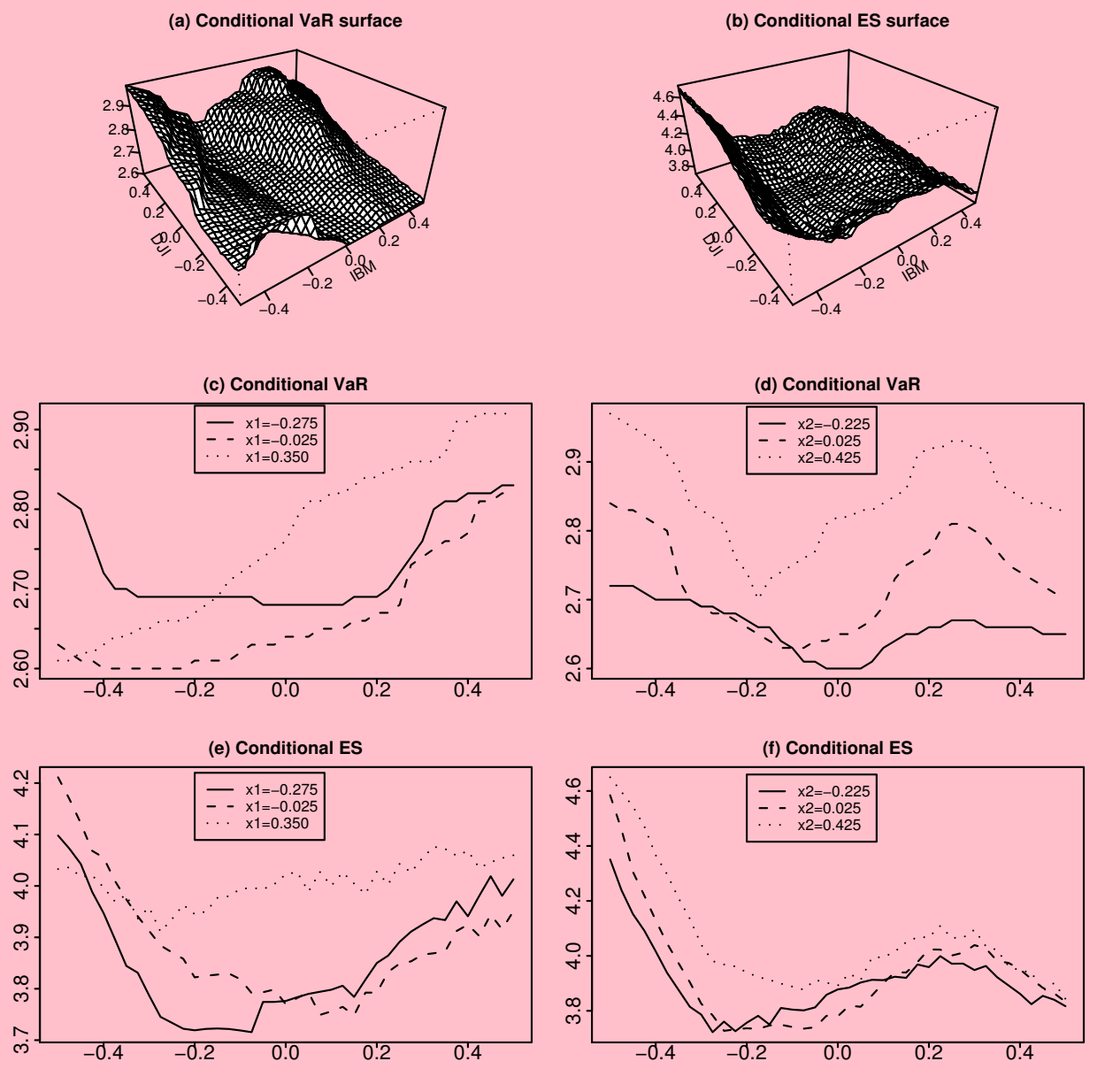

Figure 4.7: (a) 5% CVaR estimates for IBM stock returns. (b) 5% CES estimates for IBM stock returns index. (c) 5% CVaR estimates for three different values of lagged negative IBM returns $(-0.275, -0.025, 0.325)$. (d) 5% CVaR estimates for three different values of lagged negative DJI returns $(-0.225, 0.025, 0.425)$. (e) 5% CES estimates for three different values of lagged negative IBM returns $(-0.275, -0.025, 0.325)$. (f) 5% CES estimates for three different values of lagged negative DJI returns $(-0.225, 0.025, 0.425)$.

for CES. For visual convenience, Figures 4.7(c) and (e) depict the estimated CVaR and CES curves (as function of X_{t2}) for three different values of $X_{t1} = (-0.275, -0.025, 0.325)$ and Figures 4.7(d) and (f) display the estimated CVaR and CES curves (as function of X_{t1}) for

three different values of $X_{t2} = (-0.225, 0.025, 0.425)$.

From Figures 4.7(c) - (f), we can observe that most of these curves are U-shaped. This is consistent with the results observed in Example 4.3. Also, we can see that these three curves in each figure are not parallel. This implies that the effects of lagged IBM and lagged DJI variables on the risk of IBM are different and complex. To be concrete, let us examine Figure 4.7(d). Three curves are close to each other when the lagged IBM log loss is around -0.2 and far away otherwise. This implies that DJI has fewer effects (less information) on CVaR around this value. Otherwise, DJI has more effects when the lagged IBM log loss is far from this value.

4.6 Semiparametric Expectile Regressions

4.6.1 Instruction

As mentioned in Cai et al. (2018), expectile, defined in Section 2.1, as an alternative risk measure to VaR, had received more attentions in the last three decades. It is well known that VaR denotes the loss that is likely to be exceeded by a specified probability level, which is actually the quantile of a portfolio loss distribution. However, in the case that the size of extreme losses matters, for example, the occurrence of catastrophic events, VaR becomes a very conservative tail risk measure because a quantile based risk measure depends only on the probability of the occurrence of extreme losses rather than the magnitude of extreme losses. Expectile, first introduced by Newey and Powell (1987), can rectify such an undesirable situation by defining a risk measure based on the minimization of asymmetrically weighted mean square errors. Moreover, expectile has more merits compared to other popular risk measures in several ways. For example, expectile is considered to be a better alternative to both VaR and expected shortfall because expectile shares the desirable properties of coherence and elicitability; see, for example, the papers by Bellini et al. (2014), Bellini and Valeria (2015), and Ziegel (2016) for details. Another advantage is that expectile is easier to be computed than VaR and expected shortfall, which is attractive in applications. Finally, since there exists an one-to-one mapping between quantiles and expectiles as argued in Efron (1991), Jones (1994), and Yao and Tong (1996), and the link between VaR and expected shortfall as addressed in Taylor (2008), expectile can be used to calculate both VaR and expected shortfall.

Based on the aforementioned advantages of expectile, there had been an increasing number of studies devoted to developing conditional expectile models in the last two decades. For example, Kuan et al. (2009) proposed a class of conditional autoregressive expectile (CARE) models which allow for asymmetric dynamic effects of the magnitude of positive and negative lagged returns on tail expectiles, while De Rossi and Harvey (2009) proposed applying a modified state space signal extraction algorithm to estimate time-varying expectiles, which may offer an alternative method to that in Kuan et al. (2009). Moreover, Xie et al. (2014) enriched the conditional dynamic expectile model by including variables reflecting current information of investment environment and adopting a varying-coefficient setup. In such a way, a varying-coefficient setup allows the conditional expectile model to be linear in some components with their coefficients determined by unknown functions of other variables. Compared to the aforementioned parametric models, a varying-coefficient model can provide more flexibility and capture parameter heterogeneity and nonlinearity. Furthermore, a varying-coefficient model can accommodate structural information by choosing smoothing variables and alleviate the curse of dimensionality problem by adopting an additive structure; see, for example, Cai et al. (2000) or Section 2.5 for more details.

Cai et al. (2018) considered a new conditional dynamic expectile model, which was applied to an empirical study on characterizing heteroskedasticity and nonlinearity as well as asymmetry in assessing the tail risk of asset returns for S&P500. This new model adopts a partially linear form, in which some coefficients are assumed to be constant while other coefficients are allowed to depend on some smoothing variables selected by economic theories or stylized facts, and it is actually quite flexible so that it includes both models in Kuan et al. (2009) and Xie et al. (2014) as special cases. Particularly, it shares not only all merits of a fully varying-coefficient model but also can achieve more efficient estimation for the parametric coefficient part. Different from a fully varying-coefficient model in Xie et al. (2014), the partially linear setup leads itself to a three-stage estimation procedure. The first stage is to fit a fully varying-coefficient model, the second stage helps achieve the estimation of constant parameters with a parametric convergence rate, and the third stage re-estimates the varying coefficients by using the estimates in the second stage. Now, an important statistical question in fitting model (4.33) (see later) arises if the coefficient functions are actually varying or more generally if a parametric model fits the given data. This amounts to testing whether the coefficient functions are constant or in a certain parametric form. To this end,

Cai et al. (2018) developed a simple constancy test on testing varying coefficients to see if they really depend on particular economic variables.

4.6.2 Relationship Between Expectile and ES

Assume that $\{(Y_t, \mathbf{W}_t)\}_{t=1}^n$ is a sequence of strictly stationary random vectors. The η-th conditional expectile of Y_t given $\mathbf{U}_t = \mathbf{u}$ and $\mathbf{X}_t = \mathbf{x}$ is then defined by

$$e_\eta(\mathbf{w}) = \arg\min_{\xi\in\mathbb{R}} \mathbb{E}\{\ell_\eta(Y_t - \xi)|\mathbf{W}_t = \mathbf{w}\},$$

where $\ell_\eta(v)$ is the asymmetric squared loss function, defined in Section 2.1. By taking derivative with respective to ξ, and setting it to zero, expectiles are uniquely identified by the first-order condition as

$$\frac{\mathbb{E}\{|Y_t - e_\eta(\mathbf{w})|I(Y_t \le e_\eta(\mathbf{w}))\}}{\mathbb{E}\{|Y_t - e_\eta(\mathbf{w})|\}} = \eta$$

for a given $\mathbf{w}$, which is (6.16) in Fan and Gijbels (1996). Alternatively, the above formula can be re-expressed as

$$\frac{\mathbb{E}\{|Y_t - e_\eta(\mathbf{w})|I(Y_t > e_\eta(\mathbf{w}))\}}{\mathbb{E}\{|Y_t - e_\eta(\mathbf{w})|I(Y_t \le e_\eta(\mathbf{w}))\}} = \frac{1-\eta}{\eta} \equiv \Omega(e_\eta(\mathbf{w})), \tag{4.28}$$

which is the so-called Omega ratio in Taylor (2022); see, for instance, (12) in Taylor (2022). This provides a good interpretation of $e_\eta(\mathbf{w})$. For each strip near $\mathbf{w}$, the average distance from the data Y_t below $e_\eta(\mathbf{w})$ to $e_\eta(\mathbf{w})$ is $100\eta\%$. Therefore, the interpretation of the expectile is similar to that for the quantile, by replacing the distance by the number of observations. As pointed out by Fan and Gijbels (1996), the key advantage of the expectile is its easy computing algorithm, which was initiated by Efron (1991). Starting from the least squares estimate with $\eta = 1/2$ and iterating only once or twice by using the Newton-Raphson method as $\mathbf{w}$ slowly deviates away from 0.5. However, the expectile does not have a an interpretation as direct as the quantile. Also, the quantity $e_\eta(\mathbf{w})$ is well defined as long as $\mathbb{E}(Y_t)$ is finite. When $\eta = 1/2$, it can be easily seen that $e_{1/2}(\mathbf{w}) = m(\mathbf{w})$, the conditional mean function. Therefore, expectiles can be viewed as an asymmetric generalization of the mean, and the term "expectile" stems from a combination of "expectation" and "quantile".

Interestingly, one can find that $q_{\eta,t} = q_\tau(\mathbf{W}_t)$ is the same as $e_{\eta,t} = e_\eta(\mathbf{W}_t)$, where $q_{\tau,t}$ denotes the conditional quantile of Y_t given $\mathbf{W}_t$, at a probability level τ corresponding to η. Indeed, τ and η have a relationship as

$$\eta = \frac{\mathbb{E}\{(|Y_t - q_\tau|I(Y_t \le q_\tau)\}}{\mathbb{E}\{|Y_t - q_\tau|\}} \equiv \psi(\tau), \tag{4.29}$$

where q_τ is the τth quantile of Y_t; see, for example, Proposition 4 in De Rossi and Harvey (2009). This means that the population τ-quantiles and η-expectiles coincide for η satisfying the above identity. Therefore, via $\widehat{q}_\tau$, we can estimate η using (4.29). Also, a link between conditional expectile and conditional ES can be formulated in the following expression:

$$\mathrm{ES}_{\tau,t} = \left(1 + \frac{\eta}{(1-2\eta)\tau}\right) e_{\eta,t} - \frac{\eta}{(1-2\eta)\tau}\mathbb{E}(Y_t), \tag{4.30}$$

where $\mathrm{ES}_{\eta,t}$ is the ηth conditional ES of Y_t, which can be found in Newey and Powell (1987) and Taylor (2008). If $\mathbb{E}(Y_t) = 0$, it can be simplified to

$$\mathrm{ES}_{\tau,t} = \left(1 + \frac{\eta}{(1-2\eta)\tau}\right) e_{\eta,t}, \tag{4.31}$$

which is also re-expressed as; see, e.g., (13) in Taylor (2022),

$$\mathrm{ES}_{\tau,t} = \left(1 + \frac{1}{(\Omega(e_{\eta,t})-1)\tau}\right) e_{\eta,t} = \left(1 + \frac{1}{(\Omega(q_{\tau,t})-1)\tau}\right) q_{\tau,t}, \tag{4.32}$$

because as the probability of exceeding $e_{\eta,t}$ is τ, $e_{\eta,t}$ is equal to the quantile with probability level τ, $q_{\tau,}$; see, e.g., Taylor (2008). Clearly, either (4.30) or (4.31) or (4.32) provides an easy way to estimate the conditional ES via GCARE(p, q) or quantile, so that a regression technique can be used to estimate $\mathrm{ES}_{\tau,t}$. As recommended by Taylor (2022), the asymmetric slope in the conditional autoregressive value at risk by regression quantile (CAViaR) in Engle and Manganelli (2004) is used for estimating the VaR, and the ES can be modeled using a dynamic Omega ratio as in (4.28).

4.6.3 Estimation Procedures

Now, the question arises is how to model $e_{\eta,t}$ as a random variable or how to estimate $e_\eta(\mathbf{w})$ as a function of $\mathbf{w}$ in practice. There are several ways to model $e_{\eta,t}$.

First, parametrically, Cai et al. (2025) proposed a generalized conditional autoregressive expectile model, termed as GCARE(p, q), like the GARCH type model, including autoregressive components in assessing tail risk, which can be treated as an infinite version of the conditional autoregressive expectile model proposed by Kuan et al. (2009) and can be implemented as a vehicle for estimating CAViaR model proposed in Engle and Manganelli (2004) and studied by Xiao and Koenker (2009). By regarding $e_{\eta,t}$ as a random variable, then it is assumed to follow an ARMAR(p, q) type model. Therefore, the GCARE(p, q) model is

defined as

$$e_{\eta,t} = \alpha_{\eta,0} + \sum_{j=1}^{q} \alpha_{\eta,j}^{\top} \mathbf{W}_{t-j} + \sum_{i=1}^{q} \beta_{\eta,i} e_{\eta,t-i},$$

where $\alpha_{\eta,j}$ for $0 \leq j \leq p$ and $\beta_{\eta,i}$ for $i \leq i \leq q$ should satisfy some conditions, which can be found in Cai et al. (2025). Also, the parameters can be estimated by the quasi-likelihood approach similar to Engle and Manganelli (2004); see Cai et al. (2025) for details. How to choose p and q when a GCARE(p, q) is adopted in practice is an important computing issue. One can follow the idea in Cai et al. (2026) to select both p and q using the adaptive LASSO method.

Next, to estimate $e_\eta(\mathbf{w})$ in a nonparametric or semiparametric way, Cai et al. (2018) considered the following model. By assuming that $\mathbf{W} = (\mathbf{U}_t, \mathbf{X}_t)$, similar to the mean model as in (2.60) in Section 2.7.1 and the quantile model as in (3.16) in Section 3.4, we consider a model with a partially varying-coefficient framework as

$$e_\eta(\mathbf{U}_t, \mathbf{X}_t) = \mathbf{a}_\eta^{\top} \mathbf{X}_{t,1} + \mathbf{b}_\eta^{\top}(\mathbf{U}_t) \mathbf{X}_{t,2}, \tag{4.33}$$

where $\mathbf{X}_t = (\mathbf{X}_{t,1}^{\top}, \mathbf{X}_{t,2}^{\top})^{\top} \in \mathbb{R}^{p_1+p_2}$ and $\mathbf{U}_t$ is a smoothing variable. Here, both $\mathbf{X}_t$ and $\mathbf{U}_t$ are allowed to include the past returns of Y_t so that the model is dynamic. Without loss of generality, we assume $\mathbf{U}_t = U_t$ to be a scalar variable for simplicity. Moreover, $\mathbf{a}_\eta$ denotes a vector of constant coefficients of $\mathbf{X}_{t,1}$ and $\mathbf{b}_\eta(\cdot)$ is a vector of functional coefficients of $\mathbf{X}_{t,2}$. For simplicity, η is dropped in $\mathbf{a}_\eta$ and $\mathbf{b}_\eta(\cdot)$ from now on if without causing any confusion.

The model in (4.33) is general enough to include some existing expectile models as special cases. For example, the CARE model proposed by Kuan et al. (2009) can be regarded as the special case of a partially varying-coefficient expectile model, where the coefficients of the intercept term and past returns are constant but the coefficients of the magnitude of past return, measured either by the square of past returns or by the absolute value of past returns, are varying, depending on whether the past returns are positive or negative. Moreover, if the constant coefficients $\mathbf{a}$ are not included, the model in (4.33) becomes to a fully varying-coefficient expectile model as in Xie et al. (2014).

Similar to the quantile model with partially varying coefficients in Cai and Xiao (2012), the well known estimation method in Robinson (1988) or profile least squares estimation approach in Speckman (1988) for classical semiparametric regression estimation approach cannot be applied to estimating $\mathbf{a}$ and $\mathbf{b}(\cdot)$ due to the fact that the expectile model does not

have explicit normal equations. Therefore, estimation of a partially varying-coefficient model is not trivial compared to a fully varying-coefficient model as in Xie et al. (2014). To estimate $\mathbf{a}$ and $\mathbf{b}(\cdot)$, Cai et al. (2018) proposed the following estimation procedures. First, we regard $\mathbf{a}$ as a function of U_s for $1 \le s \le n$, and then, based on the local constant approximation, $\mathbf{a}(U_s)$ can be estimated by minimizing the following locally weighted loss function

$$\min_{\mathbf{a},\mathbf{b}} \sum_{t=1}^{n} \ell_\eta \Big(Y_t - \mathbf{a}^\top(U_s)\mathbf{X}_{t,1} - \mathbf{b}^\top(U_s)\mathbf{X}_{t,2} \Big) K_{h_1}(U_t - U_s),$$

where $K(\cdot)$ is a kernel function, $K_{h_1}(x) = K(x/h_1)/h_1$, and h_1 denotes the bandwidth used at this step, which controls the smoothness and satisfies that $h_1 = h_1(n) \to 0$ and $n\, h_1^2 \to \infty$. The local constant estimator for $\mathbf{a}(U_s)$ is obtained, denoted by $\widehat{\mathbf{a}}(U_s)$. To improve estimation efficiency for $\mathbf{a}$ by using full sample information, in the second stage, we take a simple average method for $\widehat{\mathbf{a}}(U_s)$, which is given by

$$\tilde{\mathbf{a}} = \tilde{\mathbf{a}}_\eta = \frac{1}{n} \sum_{s=1}^{n} \widehat{\mathbf{a}}(U_s), \tag{4.34}$$

which is shown in Theorem 1 in Cai et al. (2018) that the above estimator is $\sqrt{n}$-consistent and asymptotically normally distributed. Note that the estimator in (4.34) might not be efficient in the semiparametric sense. To obtain an efficient estimate of $\mathbf{a}$, we can follow the idea from Cai and Xiao (2012) by adding an optimal weight to (4.34) as

$$\tilde{\mathbf{a}}_{\eta,w} = \frac{1}{n} \sum_{s=1}^{n} w_s \widehat{\mathbf{a}}(U_s),$$

where $\{w_s\}_{s=1}^n$ is the optimal weight to minimize the asymptotic variance of $\tilde{\mathbf{a}}_{\eta,w}$ as in Cai and Xiao (2012). See Cai and Xiao (2012) or Section 3.4 for details.

Finally, we re-estimate $\mathbf{b}(\cdot)$ by using the partial expectile residual $Y_t^* = Y_t - \tilde{\mathbf{a}}^\top \mathbf{X}_{t,1}$, where $\tilde{\mathbf{a}}$ is a $\sqrt{n}$-consistent estimator of $\mathbf{a}$, obtained possibly from the second stage. Thus, for the given grid point u_0, the estimator of $\mathbf{b}(u_0)$ can be obtained by the following minimization problem using local linear approximation of $\mathbf{b}(U_t)$ at the grid point u_0,

$$\min_{\mathbf{a},\mathbf{b}'} \sum_{t=1}^{n} \ell_\eta \Big(Y_t^* - \mathbf{b}^\top(u_0)\mathbf{X}_{t,2} - \mathbf{b}'(u_0)^\top \mathbf{X}_{t,2}(U_t - u_0) \Big) K_{h_2}(U_t - u_0),$$

where h_2 denotes the bandwidth at this stage and $\mathbf{b}'(\cdot)$ is the first order derivative of $\mathbf{b}(\cdot)$. The local linear estimator of $\mathbf{b}(u_0)$ is denoted by $\tilde{\mathbf{b}}(u_0)$.

Note that Cai et al. (2018) investigated the asymptotic properties of the proposed estimators under time series context. Also, they proposed a new simple and easily implemented test for the goodness of fit of models and a bandwidth selector based on newly defined cross-validation estimation for the expected forecasting expectile errors. The proposed methodology is data-analytic and of sufficient flexibility to analyze complex and multivariate nonlinear structures without suffering from the curse of dimensionality. Finally, the proposed model was applied by Cai et al. (2018) to analyzing the daily data of the S&P500 return series; see, for example, the paper by Cai et al. (2018) for details.

In sum, expectile regression technique is a nice tool for estimating the conditional expectiles of a response variable given a set of covariates, especially in risk management. The **R** package "**erboost**" can implement a regression tree based gradient boosting estimator for nonparametric multiple expectile regression, proposed by Yang and Zou (2015).

Chapter 5

Nonparametric Models with Nonstationary Covariates

5.1 Introduction

Nonparametric estimation techniques have become cornerstone research topics in statistics for the last three decades since they offer numerous advantages relative to parametric techniques and have more flexibility and robustness to functional form misspecification, and have been embraced by applied researchers in many fields; see the books by Fan and Gijbels (1996) and Fan and Yao (2003) as well as Li and Racine (2008). Asymptotic theory underlying various nonparametric estimators and test statistics for many commonly used models have been well established for the iid data and some weak and strong dependent time series. The only nonparametric asymptotic analysis when covariates are integrated or unit root, time series that we are aware of includes the papers, to name just a few, by Phillips and Park (1998), Park and Hahn (1999), Chang and Martinez-Chombo (2003), Chang and Park (2003), Juhl (2005), Cai et al. (2009), Xiao (2009), and Phillips (2009). Particularly, Phillips and Park (1998), Juhl (2005), and Wang and Phillips (2009a,b) considered the case when the true data generating process is a linear unit root process, while Park and Hahn (1999), Chang and Martinez-Chombo (2003), Chang and Park (2003), Cai et al. (2009), Xiao (2009), Cai and Wang (2014), Cai et al. (2015), and Sun et al. (2013) studied the models linearized in the nonstationary variables.

In this chapter, for the observed data $\{(Y_t, Z_t)\}_{t=1}^n$, a nonparametric regression function with integrated covariate is investigated as follows,

$$Y_t = \beta(Z_t) + \varepsilon_t, \tag{5.1}$$

where $\mathbb{E}(\varepsilon_t|Z_t) = 0$, $\{\varepsilon_t\}$ is stationary and $\beta(\cdot)$ is an unknown regression function. Here, Z_t is an integrated process satisfying

$$Z_t = \rho\, Z_{t-1} + u_t, \tag{5.2}$$

where $\rho = 1$ and $\{u_t\}$ is a stationary sequence, which was considered in Cai (2011). Clearly, Z_t is persistent and nonstationary. Indeed, model (5.1) is not new in literature but its asymptotics developed in the present chapter is novel when Z_t is persistent and nonstationary. For example, if Z_t is stationary, model (5.1) has been studied extensively in the literature; see Fan and Gijbels (1996) and Fan and Yao (2003) for details, while it was investigated by Karlsen and Tjøstheim (2001) for Z_t being null recurrent time series and Karlsen et al. (2007) for the ϕ-irreducible Markov chain time series and by Bandi (2002) and Cai et al. (2017) for nearly integrated time series (ρ in (5.2) is assumed to be $\rho = 1 + c/n$ with $c < 0$). A functional coefficient type model with I(1) covariates and nonparametric co-integration is investigated by Cai et al. (2009), Xiao (2009), Cai et al. (2015), and Sun et al. (2013), respectively, and nonlinear parametric co-integration studied by Park and Phillips (2001) by assuming that $\beta(\cdot)$ in (5.1) has a particular parametric form. Finally, note that Wang and Phillips (2009a,b) considered the case to allow $\mathbb{E}(\varepsilon_t|Z_t) \neq 0$ and Cai et al. (2017) extended the situation to allow $\{u_t\}$ in (5.2) to be a long memory process and ρ to be nearly one. For simplicity of notation, we consider only one-dimensional case since extension to multivariate Z_t involves fundamentally no new ideas but complicated notations.

Model (5.1) might have a great potential in many applications. For example, in macroeconomics, a particular parametric form of (5.1) can be used for forecasting inflation rate based on some persistent and nonstationary covariates such as velocity of monetary supply; see Bachmeier et al. (2007), which showed that the velocity is an I(1) process. Also, using a semiparametric regression model with integrated covariates, Sun et al. (2013) considered the purchasing power parity hypothesis using Canadian and US price and exchange rate data. Indeed, they showed that the difference between the two countries' 10-year Treasury bond rates is an I(1) process. Finally, it can be employed for the predictability of stock returns using various lagged financial variables, such as the dividend yield, term and default premia, the dividend-price ratio, the earning-price ratio, the book-to-market ratio, and interest rates; see Elliott and Stock (1994), Cavanagh et al. (1995), Bandi (2002), Torous et al. (2004), Campbell and Yogo (2006), Polk et al. (2006), Rossi (2007), Cai and Wang (2014), Cai et al. (2015), and Cai et al. (2017), and among others. In fact, Campbell and Yogo (2006) showed

that the 95% confidence intervals for ρ in (5.2) are $[0.957, 1.007]$ and $[0.939, 1.000]$ for the log dividend-price ratio and the log earnings-price ratio, respectively; see Panel A in Table 4 of Campbell and Yogo (2006). As advocated by Campbell and Yogo (2006), Bachmeier et al. (2007), and Cai et al. (2009), the predictive power of using integrated or nearly integrated (highly persistent) covariates in a regression model can be improved significantly due to less noise.

The main purpose of the current chapter is to estimate the nonparametric regression $\beta(\cdot)$ by using the local linear (polynomial) and local constant (Nadaraya-Watson) fitting schemes and the main contribution of present chapter to the literature is to derive the asymptotic theory for both estimators. For simplicity, the main results can be summarized as follows. First, the optimal rate of convergence is $n^{1/5}$ slower than the usual $n^{2/5}$ rate for stationary case. Consequently, the order of the asymptotic mean-squared error (AMSE) is $n^{-2/5}$ rather than the standard rate $n^{-4/5}$. The intuitive explanation to this phenomenon is that an I(1) time series takes longer to revisit levels in its range. Second, the asymptotic bias term, similar to the stationary case, is independent of the stationary density of the regressor and is due to the linear approximation, which is typical for a local linear fitting scheme; see, for example, Fan and Gijbels (1996) for details. Third, the limiting distribution is a mixed-normal (conditional normal) with the asymptotic variance depending inversely on the local time of a Brownian motion in which the unit root series can be embedded. Furthermore, the integrated covariate requires the larger bandwidths. Indeed, the optimal (in the AMSE sense) bandwidth is $O_p(n^{-1/10})$ implying a larger optimal bandwidth than in conventional kernel regressions with stationary regressors where the optimal bandwidth is known to be $O(n^{-1/5})$. Clearly, the use of conventional bandwidth has the theoretical potential of under-smoothing in the presence of I(1) covariates. Finally, it is very interesting that both local linear and local constant estimators share exactly same asymptotic properties at both interior and boundary points. The reader is referred to the paper by Cai (2011) for details.

5.2 Statistical Properties

5.2.1 Local Linear Estimation

In this section, we develop the estimation procedure on how $\beta(\cdot)$ is estimated using local linear fitting from observations $\{(Y_t, Z_t)\}_{t=1}^n$. Our motivation of using local linear

fitting is its high statistical efficiency in an asymptotic minimax sense, design adaptation and automatic correction for edge effects, as discussed in Fan and Gijbels (1996). Although a general local polynomial technique is applicable as well, it is well known that the local linear fitting will suffice for many applications; see Fan and Gijbels (1996) for a very comprehensive discussion, and that the theory developed for the local linear estimator continues to hold for the local polynomial estimator with only slight modification. Another virtue of using local polynomials is that both the unknown functions as well as their derivatives can be estimated simultaneously. For simplicity, the focus is only on local linear estimation and leave the generalization for additional research.

It is assumed throughout this chapter that $\beta(\cdot)$ is twice continuously differentiable, so that at any given z, a local approximation is used as $\beta(Z_t) \simeq \beta(z) + \beta'(z)\,(Z_t - z)$, when Z_t is a neighborhood of z, where $\simeq$ denotes the first order Taylor approximation and $\beta'(z)$ is the first derivative of $\beta(z)$. Hence, (5.1) is approximated by

$$Y_t \simeq \theta_0 + (Z_t - z)\,\theta_1 + \varepsilon_t,$$

and it becomes a local linear model. Therefore, the locally weighted sum of squares is

$$\sum_{t=1}^{n} [Y_t - \theta_0 - (Z_t - z)\,\theta_1]^2 \; K_h(Z_t - z). \tag{5.3}$$

By minimizing (5.3) with respect to θ_0 and θ_1, the local linear estimate of $\beta(z)$ is obtained and is denoted by $\widehat{\beta}(z)$, and the local linear estimator of the derivative of $\beta(z)$ is denoted by $\widehat{\beta}'(z)$. It is easy to show that the minimizer of (5.3) is given by

$$\begin{pmatrix} \widehat{\beta}(z) \\ \widehat{\beta}'(z) \end{pmatrix} = \left[\sum_{t=1}^{n} \begin{pmatrix} 1 \\ Z_t - z \end{pmatrix}^{\otimes 2} K_h(Z_t - z)\right]^{-1} \sum_{t=1}^{n} \begin{pmatrix} 1 \\ Z_t - z \end{pmatrix} Y_t\, K_h(Z_t - z), \tag{5.4}$$

where $A^{\otimes 2} = A\,A^{\top}$ ($A^{\otimes 1} = A$) for a vector or matrix A.

5.2.2 Notations and Assumptions

Since Z_t is an I(1) process, it can be re-expressed as $Z_t = Z_0 + \sum_{s=1}^{t} u_s$, where $\{u_s\}$ is a stationary process with mean zero and variance σ_u^2. In what follows, it is assumed that the process $\{u_t\}$ is a stationary linear process as $u_s = \sum_{j=0}^{\infty} c_j\,\omega_{s-j}$, where ω_j is a white noise with mean zero and $\sigma_\omega^2 = \mathrm{Var}(\omega_j) < \infty$, and $\{c_j\}$ satisfies, for some $0 < \tau \leq 1$,

$$\sum_{j=0}^{\infty} |c_j|^{\tau} < \infty, \qquad \text{and} \qquad \sum_{j=0}^{\infty} c_j = 1. \tag{5.5}$$

Then, $\sigma_u^2 = \mathrm{Var}(u_s) = \sigma_\omega^2 \sum_{j=0}^{\infty} c_j^2$ and $\mathrm{Cov}(u_s, u_{s+t}) = \sigma_\omega^2 \sum_{j=0}^{\infty} c_j\, c_{j+t}$ for any s and t. Note that the assumption on $\{u_t\}$ being a linear process is due to an application of some results from Jeganathan (2004). Of course, it can be relaxed at the cost of involving lengthier mathematical proofs. Clearly, one has

$$Z_t/\sqrt{n} = Z_0/\sqrt{n} + \frac{1}{\sqrt{n}} \sum_{s=1}^{[nr]} u_s$$

for $r = t/n$. An application of Donsker's theorem (see, for example, Theorem 14.1 in Billingsley (1999) for iid $\{u_t\}$ with the existence of the second moment of u_t) leads to

$$Z_t/\sqrt{n} \implies W_u(r), \tag{5.6}$$

where " $\Longrightarrow$ " represents weak convergence, $W_u(\cdot) = \sigma_0\, W(\cdot)$ with $W(\cdot)$ being a standard Brownian motion on $[0,\ 1]$ and $\sigma_0^2 = \lim_{n\to\infty} \mathrm{Var}(n^{-1/2} \sum_{t=1}^{n} u_t)$, which is assumed to exist and be finite. In particular, it follows from Merlevéde et al. (2006) that (5.6) holds if $\{u_t\}$ is stationary strong mixing sequence and satisfies, for some $\delta_0 > 0$,

$$E|u_t|^{2+\delta_0} < \infty, \quad \text{and} \quad \sum_{k=1}^{\infty} k^{(2+\delta_0)/\delta_0}\, \alpha(k) < \infty, \tag{5.7}$$

where $\alpha(\cdot)$ is the mixing coefficient; see, e.g., Hall and Heyde (1980) for the definition.

Define $\eta_{t,z} = (Z_t - z)/\sqrt{t}$ for any z and let $f_{t,z}(\cdot)$ denote the density of $\eta_{t,z}$. Also, let $f_{t,s,z}(\cdot,\cdot)$ represent the joint density function of $(\eta_{t,z}, \eta_{s,z})$. Furthermore, let $\mathcal{F}_t$ be the smallest sigma field generated by $\{(Y_s,\ Z_s)\}_{s=-\infty}^{t}$. The following assumptions are listed.

Assumptions:

(E1) $\mathbb{E}(\varepsilon_t | Z_t, \mathcal{F}_{t-1}) = 0$, $\sigma_\epsilon^2(Z_t) = \mathbb{E}(\varepsilon_t^2 | Z_t, \mathcal{F}_{t-1}) = \sigma_\epsilon^2$, $\mathbb{E}(\varepsilon_t^4 | Z_t, \mathcal{F}_{t-1}) < C$ a.s., and $\{u_t\}$ is a stationary and mixing process satisfying constraints as imposed by (5.5) and (5.7), where σ_ϵ^2, C and σ_0^2 are finite positive constants.

(E2) Both $f_{t,z}(\cdot)$ and $f_{t,s,z}(\cdot,\cdot)$ have bounded continuous derivative functions (for all t, s and fixed z).

(E3) $K(\cdot)$ is a kernel function with a finite support, say $[-1,\ 1]$ and it is symmetric.

(E4) $n\,h \to \infty$ and $n\,h^{10} = O(1)$.

Remark 5.1: *Now, we discuss the above conditions. Condition E1 requires that* $\{\varepsilon_t\}$ *is a martingale difference process with conditional homogenous variance and a finite fourth moment. The martingale difference assumption can be relaxed to a mixing process, and the assumption on the conditional homogenous error can be relaxed to the case that* $\sigma_\epsilon^2(Z_t)$ *is non-constant, with a lengthier proof. E2 is a very mild assumption. Indeed, it is satisfied if* $\{u_t\}$ *is commonly assumed to be iid normal. Finally, Assumptions E3 and E4 are commonly imposed in the kernel estimation literature and Assumption E4 is satisfied for the optimal bandwidth* $h = O(n^{-1/10})$ *(see later).*

Finally, the local time $L(t, x)$ for a standard Brownian motion is defined in (1.10), which is

$$L(t,x) = \lim_{\Delta\to 0} \frac{1}{2\Delta} \int_0^t I_{\{|W(s)-x|\le\Delta\}} ds, \quad 0 \le t \le 1, \quad \text{and} \quad x \in \mathbb{R},$$

where I_A is the indicator function of an event A and $W(\cdot)$ is a standard Brownian motion; see, for example, the book by Karatzas and Shreve (1991) or the papers by Phillips and Park (1998) and Park and Phillips (1999) for details. Note that $L(t, z)$ can be consistently estimated by $S_{n,0}(z)$; see Lemma 5.1 in Section 5.7.

5.2.3 Asymptotic Results

Now, main result is stated below and the proof is relegated to Section 5.7.

Theorem 5.1: *Under Assumptions E1 – E4, one has*

$$\sqrt{n^{1/2}\, h}\, \left[\widehat{\beta}(z) - \beta(z) - h^2 B(z) + o_p(h^2)\right] \xrightarrow{d} \mathrm{MN}(\sigma_\beta^2),$$

where $B(z) = \mu_2(K)\beta''(z)/2$ *and* $\xi = \mathrm{MN}(\sigma_\beta^2)$ *is a mixed normal distribution with mean zero and variance* $\sigma_\beta^2 = \sigma_\varepsilon^2\, \sigma_0\, \nu_0(K)\, /L(1,0)$ *with* σ_ε^2 *defined in Assumption E1.*

From Theorem 5.1, one can see that first, ξ_t is called to be a mixed normal with mean μ_t and covariance Σ_t if the conditional distribution of ξ_t given μ_t and Σ_t is $N(\mu_t, \Sigma_t)$; see Phillips and Park (1998) for details. Note that the asymptotic properties for $\widehat{\beta}'(z)$ can be obtained as the same fashion as those in Theorem 5.1 and omitted. By comparing the results in Theorem 5.1 and conventional findings in Fan and Gijbels (1996) and Fan and Yao (2003) for the stationary covariates, the new results can be summarized as follows. Clearly, $h^2\mu_2(K)\beta''(z)/2$ serves as the asymptotic bias, which is the same as that for stationary case

when one uses a local linear estimation method; see Fan and Yao (2003). However, the convergence rate is the order of $n^{1/4}h^{1/2}$ much lower with a factor $n^{1/4}$ by comparing with that for stationary covariates. Also, the stochastic asymptotic variance is independent of the grid point z. Indeed, one can show that the results in Theorem 5.1 hold true as long as any $z = z_n$ satisfies $z_n/\sqrt{n} \to 0$ and $n^{1/4}\, h^{5/2}\, \beta''(z_n) = O(1)$; see Theorem 5.2 later. Furthermore, from the asymptotic bias and variance presented in Theorem 5.1, the stochastic AMSE is given by

$$\text{AMSE} = \text{Var} + \text{bias}^2 = \sigma_\beta^2\, n^{-1/2} h^{-1} + \frac{h^4}{4}\, \mu_2^2(K)\, [\beta''(z)]^2\,.$$

The minimization of the AMSE with respect to h yields the optimal bandwidth

$$h_{opt} = \left(\frac{\sigma_\beta}{\mu_2(K)|\beta''(z)|}\right)^{2/5} n^{-1/10} = O_p(n^{-1/10}),$$

which is stochastic and much larger than the conventional optimal bandwidth $h_{opt,s} = O(n^{-1/5})$ for the stationary case; see Fan and Yao (2003). Therefore, if $h_{opt,s}$ were be used in estimating $\beta(\cdot)$ in (5.1), the nonparametric estimator given in (5.6) would be under-smoothing. Hence, it would be a very interesting future research topic on how to select the data-driven (optimal) bandwidth theoretically and empirically.

Now, the focus is on investigating the asymptotic behaviors at boundaries. When Z_t is $I(1)$, it follows from (5.6) that when $z = a\,\sqrt{n}$ $(a \neq 0)$ and $r = t/n$,

$$\mathbb{P}(Z_t \geq z) = \mathbb{P}(Z_t \geq a\sqrt{n}) \;\to\; \mathbb{P}(W_u(r) \geq a/\sigma_u) = 1 - \Phi(a/\sqrt{r}\sigma_0) > 0,$$

where $\Phi(\cdot)$ is the distribution of the standard normal random variable. This means that there is a great chance that $|Z_t|$ can take large values. In other words, an I(1) time series takes longer to revisit levels in its range. Now, the question is how the asymptotic behaviors of the estimator look like when z is large like $z = a\,\sqrt{n}$ for any fixed a. To this end, the following asymptotic results is obtained at boundary $z = a\,\sqrt{n}$ for any fixed a. However, the detailed proofs are not provided since they follow closely the same arguments as those used in the proof of Theorem 5.1.

Theorem 5.2: *If Assumptions E1 – E4 hold and* $n^{1/4}\, h^{5/2}\, \beta''(a\sqrt{n}) = O(1)$ *for any* a*, then, one has*

$$\sqrt{n^{1/2}\, h}\, \left[\widehat{\beta}(a\sqrt{n}) - \beta(a\sqrt{n}) - h^2\, B(a\sqrt{n}) + o_p(h^2)\right] \xrightarrow{d} MN(\sigma_a^2),$$

where $MN(\sigma_a^2)$ *is a mixed normal distribution with mean zero and variance* $\sigma_a^2 = \sigma_\varepsilon^2\, \sigma_0\, \nu_0(K)/L(1, a/\sigma_0)$.

Remark 5.2: *Comparing Theorem 5.2 with Theorem 5.1, one can observe that the magnitude of the asymptotic variance of $\widehat{\beta}(\cdot)$ at the boundary points ($z = O(n^{1/2})$) differs from that for the interior points ($z = o(n^{1/2})$). This finding is different from its stationary counterpart; see Fan and Gijbels (1996) for the stationary case.*

5.2.4 Nadaraya-Watson Estimation

Now, the turn is to discussing the asymptotic properties for the local constant estimator of $\beta(\cdot)$. It is well documented that the Nadaraya-Watson estimator is given by

$$\widetilde{\beta}(z) = \sum_{t=1}^{n} Y_t \, K_h(Z_t - z) / \sum_{t=1}^{n} K_h(Z_t - z).$$

For $\widetilde{\beta}(z)$, the following theorem can be established.

Theorem 5.3: *Under the assumptions of Theorem 5.1, both $\widetilde{\beta}(z)$ and $\widehat{\beta}(z)$ share the exact same asymptotic properties. That is,*

$$\sqrt{n^{1/2}\, h}\, \left[\widetilde{\beta}(z) - \beta(z) - h^2 B(z) + o_p(h^2)\right] \xrightarrow{d} MN(\sigma_\beta^2),$$

where $B(z) = \mu_2(K)\beta''(z)/2$ and $MN(\sigma_\beta^2)$ is a mixed normal distribution with mean zero and variance $\sigma_\beta^2 = \sigma_\varepsilon^2 \, \sigma_0 \, \nu_0(K) \, / L(1,0)$. Further, Theorem 5.2 holds for $\widetilde{\beta}(z)$.

Remark 5.3: *It is clear that $h^2\mu_2(K)\beta''(z)/2$ serves as the asymptotic bias, which is the same as that case when one uses a local linear estimation method (see Theorem 5.1). However, for the stationary Z_t case with a local constant estimation method, there is an additional leading bias term which has the form of $h^2\mu_2(K)f_z'(z)\beta'(z)/2f_z(z)$, where $f_z(\cdot)$ is the stationary density of Z_t when Z_t is stationary; see Fan and Gijbels (1996). Theorem 5.3 shows that for non-stationary Z_t, the local constant estimator has the same leading bias as that of a local linear method. This is an interesting new finding that is not shared by a local constant estimator if Z_t is stationary. It can be shown that with nonstationary Z_t, the bias term associated with $f_{t,z}'(z)\beta'(z)$ has an order of $h^2 n^{-1/2} \ln(n)$, which is smaller than h^2; see Lemma 5.4 in 5.7. Therefore, the leading bias contains only one term associated with $\beta''(z)$ with the order h^2. Interestingly, as in the case of standard local polynomial methods, the Nadaraya-Watson estimator is design-adaptive too in the sense of Fan and Gijbels (1996). Clearly, this property should be interpreted as follows. The clustered designs are not expected*

to occur in the presence of integrated (highly persistent) processes. Therefore, the theoretical relevance of the design-adaptation property and the theoretical appeal of local polynomial methods over the standard Nadaraya-Watson kernel estimates seem to vanish.

5.3 Functional Coefficient Models

Model (5.1) is indeed a special case of the following functional coefficient model

$$Y_t = \boldsymbol{\beta}(Z_t)^\top \mathbf{X}_t + \varepsilon_t, \tag{5.8}$$

where $\mathbf{X}_t$ can contain both stationary and nonstationary components, Z_t can be stationary or nonstationary or time, $\mathbb{E}(\varepsilon_t|Z_t, \mathbf{X}_t) = 0$, $\{\varepsilon_t\}$ is stationary, and $\beta(\cdot)$ is an unknown regression function. For example, if Z_t is stationary, model (5.8) was explored by Cai et al. (2009), Xiao (2009) and Sun and Li (2011), while it was studied by Sun et al. (2013) if Z_t is I(1). When Z_t is time, it was investigated by Park and Hahn (1999), Phillips et al. (2013), and Cai et al. (2015).

One can use the local linear or constant estimation procedure to estimate the coefficient function $\beta(\cdot)$ in (5.8). The asymptotic theory can be found in Cai et al. (2009) when Z_t is stationary and Sun et al. (2013) for the case that Z_t is I(1). The question arises is how to choose the bandwidth in a data-driven fashion when estimating $\beta(\cdot)$ in (5.8). To answer this question, Sun and Li (2011) extended the asymptotic results of the traditional least squares cross-validation bandwidth selection method to semiparametric regression models with nonstationary data. Their findings can be summarized as follows. First, the CV-selected bandwidth is stochastic even asymptotically and second, the selected bandwidth based on the local constant method converges to 0 at a different speed than that based on the local linear method, although two estimation methods share exactly the same asymptotic bias as discussed in Remark 5.3. Both findings are in sharp contrast to existing results when working with weakly dependent or independent data. Also, they conducted simulations to confirm their theoretical results and show that the automatic data-driven method works well. Finally, note that this data-driven bandwidth selection approach proposed by Sun and Li (2011) can be applied to the model in (5.1) to estimate $\boldsymbol{\beta}(\cdot)$.

5.4 Specification Tests

5.4.1 Nonparametric Tests of Mean Function

In real applications, one might prefer a parametric model, say $\beta(z)$ in (5.1) has a particular form such as $\beta(z) = \beta_0(z, \theta)$, where $\beta_0(\cdot)$ is a known function and θ is an unknown parameter. Then, the econometric issue to about the testing problem

$$H_0 : \mathbb{P}(\beta(Z_t) = \beta_0(Z_t, \theta)) = 1 \quad \text{versus} \quad H_a : \mathbb{P}(\beta(Z_t) \neq \beta_0(Z_t, \theta)) > 0, \tag{5.9}$$

where $\theta \in \Theta$ is a compact subset of $\mathbb{R}^p$. Clearly, the testing problem formulated in (5.9) is similar to (2.22) for the stationary case, which is the test of nonparametric function versus parametric function. Also, (5.9) could be applied to check if $\beta(z)$ is a threshold function or other types of parametric forms to see if some financial/economic theory holds. Furthermore, we can see clearly that Y_t is stationary under H_0 if $\beta_0(z) = 0$. However, under H_a, Y_t might be nonstationary, so that as studied in Park and Phillips (2001), a parametric nonlinear co-integrating exists between Y_t and Z_t.

To test H_0 formulated in (5.9), one might use the test statistic as in (2.23). But, we do not know whether the test statistic in (2.23) for the iid case or the stationary time series case can be applied to the nonstationary case or not, which is a very interesting topic, left as a future research. Instead, we can use the following L_2-type test statistic, which is a routine test statistic for testing nonparametric versus parametric, proposed by Wu (2013),

$$\int \left[\widehat{\beta}(z) - \beta_0(z, \widehat{\theta})\right]^2 D(z)dz, \tag{5.10}$$

where $D(z)$ is a weighting function $D(\cdot)$ to avoid a random denominator and $\widehat{\theta}$ is the estimate of θ under the null hypothesis; see, e.g., Park and Phillips (2001) on how to obtain $\widehat{\theta}$, which can be simplified as a U-statistic as follows; see, e.g., Wu (2013) for details,

$$J_n(n) = \frac{1}{n^{3/4}h^{1/2}} \sum_{t \neq s=1}^{n} \widehat{\epsilon}_t \widehat{\epsilon}_s K_{ts}, \tag{5.11}$$

where $\widehat{\epsilon}_t = Y_t - \beta_0(Z_t, \widehat{\theta})$ is the parametric residual under H_0, and h is the bandwidth, $K_{ts} = K^*((Z_t - Z_s)/h)$ with $K^*(u)$ being the convolution function of $K(\cdot)$ and $K(\cdot)$, defined in (1.14), and $K(\cdot)$ is a kernel function used to estimate $\beta(\cdot)$ in (5.1). Also, Wu (2013) derived the asymptotic distribution of the proposed test statistic $J_n(n)$ under H_0 and showed that it

diverges to ∞ under the alternative hypothesis. In particular, of interest is that Wu (2013) derived the limiting result for a U-statistic as in (5.10) involving nonstationary variables.

As in Kasparis et al. (2015), by assuming that $\beta_0(\cdot) = \mu_0$ with unknown μ_0, a special case of the above hypothesis test by specifying $H_0 : \beta(z) = \mu_0$ for all z. Then, based on the asymptotic distribution of the nonparametric estimation of $\beta(z_j)$, denoted by $\widehat{\beta}(z_j)$ for some grid points $\{z_j\}_{j=1}^m$, they proposed a naive test as follows

$$F_{\text{sum}} = \sum_{j=1}^{m} A(z_j) \left(\widehat{\beta}(z_j) - \widehat{\mu}_0\right)^2, \quad \text{or} \quad F_{,\max} = \max \left[A(z_j) \left(\widehat{\beta}(z_j) - \widehat{\mu}_0\right)^2 \right],$$

where $A(\cdot)$ is a self-normalized function and $\widehat{\mu}_0$ is the estimate of m under the null hypothesis. They argued that that $F_{\text{sum}} \xrightarrow{d} \chi^2_m$ and $F_{\text{sum}} \xrightarrow{d} Z_*$ for some random variable Z. But, as pointed out by Kasparis et al. (2015), the proposed test depends on the choice of $\{z_j\}_{j=1}^m$ and m. Particularly, by assuming that $\mu_0 = 0$, Juhl (2014) considered testing the hypothesis $H_0 : \beta(z) = 0$ for all z, and proposed using the conditional moment testing approach as in Zheng (1996) to construct the test statistic. Different from $J_n(n)$ in (5.11) involving $\{\widehat{\epsilon}_t\}$, the test statistic proposed by Juhl (2014) is given as follows:

$$J_{n,j}(n) = \frac{1}{n^2\, h} \sum_{t \neq s=1}^{n} Y_t Y_s K_{ts},$$

which is a function of $\{Y_t\}$ instead of $\{\widehat{\epsilon}_t\}$, and its limiting distribution under H_0 was derived by Juhl (2014); see Juhl (2014) for details.

Different from the above test from the model in (5.1), Sun et al. (2016) considered testing if the coefficient functions in (5.8) are constant or not, formulated as

$$H_0 : \mathbb{P}(\boldsymbol{\beta}(Z_t) = \boldsymbol{\beta}_0) = 1 \quad \text{versus} \quad H_a : \mathbb{P}(\boldsymbol{\beta}(Z_t) \neq \boldsymbol{\beta}_0) > 0, \tag{5.12}$$

where $\boldsymbol{\beta}_0 \in \mathscr{B}$ is a compact subset of $\mathbb{R}^p$. That is, we test whether the coefficient functions in (5.8), $\boldsymbol{\beta}(\cdot)$, are constant. If the null hypothesis holds true, model (5.8) becomes a linear co-integrating model; otherwise, model (5.8) is a semiparametric varying co-integrating model. Similar to (5.10), we propose the following L_2-type test statistic to the testing problem in (5.12),

$$\int \left(\widehat{\boldsymbol{\beta}}(z) - \widehat{\boldsymbol{\beta}}_0\right)^\top \mathbf{W}_n \left(\widehat{\boldsymbol{\beta}}(z) - \widehat{\boldsymbol{\beta}}_0\right) dz,$$

where $\mathbf{W}_n$ is a weighting matrix to avoid a random denominator and $\widehat{\boldsymbol{\beta}}_0$ is the estimate of $\boldsymbol{\beta}_0$ under the null hypothesis; see, e.g., Cai and Wang (2014) on how to obtain $\widehat{\boldsymbol{\beta}}_0$, which

can be simplified as a U-statistic as follows; see, e.g., Sun et al. (2016) for details,

$$I_n(n) = \frac{1}{n^2 h^{1/2}} \sum_{t \neq s=1}^{n} \mathbf{X}_t \mathbf{X}_s^\top \widehat{\epsilon}_t \widehat{\epsilon}_s K_{ts}, \tag{5.13}$$

where $\widehat{\epsilon}_t = Y_t - \widehat{\boldsymbol{\beta}}_0^\top \mathbf{X}_t$ is the parametric residual under H_0.

Moreover, Sun et al. (2016) derived the asymptotic properties of $I_n(n)$ under both H_0 and H_a by considering two cases: (a) $\mathbf{X}_t$ is a vector of integrated variables and (b) $\mathbf{X}_t = (\mathbf{X}_{1,t}, \mathbf{X}_{2,t})$ contains both stationary and integrated variables, where $\mathbf{X}_{1,t}$ is of dimension p_1 with its first component unity and the remainder I(0) variables, and $\mathbf{X}_{2,t}$ is of dimension p_2 with I(1) variables, so that the model in (5.8) becomes

$$Y_t = \boldsymbol{\beta}_1(Z_t)^\top \mathbf{X}_{1,t} + \boldsymbol{\beta}_2(Z_t)^\top \mathbf{X}_{2,t} + \epsilon_t.$$

The reason of doing so is that although the convergent rates under H_0 for both cases is same; see, for example, Theorem 3.1(i) for the case (a) in Sun et al. (2016) and Theorem 3.2 for the case (b) in Sun et al. (2016), the convergent rates under H_a for both case are totally different; see, for example, Theorem 3.1(ii) and Theorem 3.3 in Sun et al. (2016) for details.

Remark 5.4: *Sun et al. (2016) showed that under H_a, the test statistic $I_n(n)$ given in* (5.13) *diverges to ∞ at different rates depending on whether or not $\boldsymbol{\beta}_2(z) = \boldsymbol{\beta}_{20}$ (a constant vector). Although $I_n(n)$ is a consistent test under both cases, more samples are required for the power of the test statistic to approach one under case (b) than under case (a). Moreover, the proof in Appendix A in Sun et al. (2016) indicates that when $\boldsymbol{\beta}_1(z) = \boldsymbol{\beta}_{10}$ (a constant vector) for all z and $\boldsymbol{\beta}_2(z) = \boldsymbol{\beta}_{20}$ over a nonempty interval of z, the least squares estimator β_1 of the misspecified linear regression model diverges to ∞ at the rate of root-n if $Cov(\mathbf{X}_{1,t}, \boldsymbol{\beta}_2(Z_t)) \neq 0$. This result suggests that it is very important to test the correct model specification when $\mathbf{X}_t$ contains both I(0) and I(1) components in model* (5.8). *The reader is referred to the paper by Sun et al. (2016) for more discussions. This remark rings a bell to the reader who is conducting a test when regressors contain both stationary and nonstationary.*

5.4.2 Nonparametric Test of Heteroskedasticity

Evidently from the previous sections, to obtain the asymptotic theory for the proposed estimators or the proposed test statistics, it is implicit to assume that either $\sigma_\epsilon^2(Z_t)$ in (5.1)

or $\sigma_\epsilon^2(Z_t, \mathbf{X}_t) = \mathbb{E}(\varepsilon_t^2|Z_t, \mathbf{X}_t)$ in (5.8) is constant. However, this assumption might not be satisfied for many applications; see, for example, the papers by Park (2002), Choi et al. (2016) and Hong et al. (2021). Particularly, Park (2002) and Hong et al. (2021) considered the following model

$$Y_t = \mu_t + \epsilon_t$$

with $\epsilon_t = \sigma_t\, e_t$, where μ_t is a mean function, $\sigma_t^2 = \sigma^2(X_{t-1}) = \mathbb{E}(\epsilon_t^2|, X_{t-1})$, and $\{e_t\}$ is iid with mean zero and variance one. Here, it is assume that $\mu_t = 0$ in Park (2002) for estimating $\sigma^2(\cdot)$ and $\mu_t = \alpha + \beta\, X_{t-1}$ in Hong et al. (2021) for testing predictive regression, where X_t is nonstationary such as I(1), like the earning-price ratio or dividend-price; see, e.g., Hong et al. (2021). Furthermore, to derive the asymptotic properties of the proposed estimators, Park (2002) defined two classes of $\sigma^2(\cdot)$, the so-called H-regular and I-regular. By specifying $\sigma^2(X_{t-1}) = \sigma_0^2(X_{t-1}, \theta)$, a nonlinear least squares can be applied to estimate θ; see, e.g., Park (2002) for details.

From from Park (2002), Hong et al. (2021) considered the following testing issue

$$H_0 : \sigma^2(X_{t-1}) = \sigma_0^2 \quad \text{versus} \quad H_a : \sigma^2(X_{t-1}) \neq \sigma_0^2,$$

where σ_0^2 is an unknown parameter. To derive the test statistics, one needs to estimate $\mu_t = \alpha + \beta\, X_{t-1}$. As suggested by Hong et al. (2021), the estimation procedure as in Cai and Wang (2014) is employed here, so that $\widehat{\epsilon}_t$ is obtained, so is $r_t = \widehat{\epsilon}_t^2$. Then, run a regression of r_t versus X_{t-1} in a nonparametric way as

$$r_t = \sigma^2(X_{t-1}) + \xi_t,$$

so that $\widehat{\sigma}^2(X_{t-1})$ is obtained. Under H_0, the sample average is used to estimate σ_0^2; that is, $\widehat{\sigma}_0^2 = \sum_{t=1}^n r_t/n$, so that the residual under H_0 is $\widehat{\xi} = r_t - \widehat{\sigma}_0^2$. Then, the test statistic, similar to (5.10), is given by

$$\int \left[\widehat{\sigma}^2(x) - \widehat{\sigma}_0^2\right]^2 D_\sigma(x) dx,$$

where $D_\sigma(x)$ is a weighting function to avoid a random denominator, which, similar to (5.11), is simplified as follows

$$S_{n,\sigma} = \sum_{t\neq s=2}^{n} \widehat{\xi}_t\, \widehat{\xi}_s\, K_{ts}. \tag{5.14}$$

Under some regularity assumptions, which are similar to Assumptions 1–5 in Wang and Phillips (2012), by following the proof in Wang and Phillips (2012), it is not difficult to show

that

$$S_{n,\sigma}/\sqrt{2\widehat{\Sigma}} \xrightarrow{d} N(0,1),$$

which is similar to (3.10) in Zheng (1996) for iid case, and (3.1) in Wang and Phillips (2012) for nonstationary situations, where

$$\widehat{\Sigma} = \sum_{t\neq s=2}^{n} \widehat{\xi}_t^2 \, \widehat{\xi}_s^2 K_{ts}^2,$$

which is the exact same as that for the iid case in Zheng (1996) and for stationary time series context in Fan and Li (1996) and Li (1999) as well as for nonstationary time series in Wang and Phillips (2012). Also, one can show that the test statistic $S_{n,\sigma}$ in (5.14) is a consistent test.

5.5 Empirical Applications

Example 5.1: As an illustration of the proposed methodology in (5.14), Hong et al. (2021) tested heteroskedasticity for predictive regression of stock return with dividend-price ratio and earning-price ratio as the predictors. Two different series of stock returns are employed, the returns on NYSE/AMEX value weighted index and S&P 500 index from the Center for Research in Security Prices (CRSP). They used monthly data from November 1926 to December 2019. As a result, both null hypotheses are rejected at 5% significance level. This means that $\sigma^2(X_{t-1})$ depends on X_{t-1}. See, e.g., the paper by Hong et al. (2021) for details.

Example 5.2: Sun et al. (2013) investigated the purchasing power parity (PPP) hypothesis using Canadian and US price index and exchange rate data. The PPP theory, although it is still in debate, says that the following setup holds the standard PPP model

$$s_t = \beta_0 + \beta_1 \, p_t + \beta_2 \, p_t^* + u_t, \tag{5.15}$$

where s_t, p_t, and p_t^* are the logarithm of the nominal exchange rate expressed as Canadian dollars per unit of US dollar, the Canadian and US aggregate price levels, respectively. The aggregate price index is measured by the producer price index (PPI) base-weighted to the year 2000. Sun et al. (2013) used monthly data for the period from January 1974 to December 2009 so that there are 432 observations. For the ideal scenario, β_0 should be zero and β_1 should be one and should be $-\beta_2$ according to the PPP theory. Now, the debating

issue hanging in this area is that the ideal case might not hold due to many reasons. Here, we re-visit this issue as follows.

First, when we apply the Engle-Granger (EG) residual-based co-integrating test; see, e.g, the paper by Engle and Granger (1987), to model (5.15), the EG test statistic equals -2.8566 with a p-value of 0.2254, which indicates a failure to reject a spurious regression at any conventional significance level. Second, when we apply the test based on Phillips and Ouliaris (1990), the test statistic equals -0.3006 with a p-value bigger than 0.15, which means that we cannot reject the null hypothesis that the series are not co-integrated. In addition, Johansen's co-integrating rank test statistic; see, e.g., the paper by Johansen (1991), for no co-integrating relation against at least one co-integrating relation yields a value of 21.71, which is smaller than the entire 10%, 5% and 1% critical values. The critical values are 23.11, 25.54 and 30.34, respectively. Therefore, linear model based testing results fail to support the long-run PPP theory.

To overcome the above problem, Sun et al. (2013) argued that based on the sticky-price theory of exchange rate determination, exchange rate movements also respond to monetary shocks. Due to sticky prices, the goods markets adjust to the monetary shocks slower than asset markets. Hence, in addition to the aggregate price levels, some other economic variables, such as interest rate differentials between two nations, also affect exchange rate formation and adjust more quickly to monetary shocks than the aggregate price indexes do. Therefore, to verify this economic theory, one can examine whether exchange rate depends on the interest rate differential between US and Canada. Specifically, $Z_t = T_{\mathrm{US,t}} - T_{\mathrm{CN,t}}$ denotes the difference between the two countries' 10-year Treasury bond rates. The time series plot of Z_t is given in Figure 5.1(a) and its autocorrelation function (ACF) plot is displayed in Figure 5.1(b). Applying the augmented Dickey-Fuller (ADF) test statistic to the interest rate differential, I can not reject the null hypothesis $H_0 : \rho = 1$, so that Z_t follows a unit root process at the 5% significance level. Therefore, Z_t is treated as an integrated series. Indeed, one might evidence visually from Figures 5.1(a) and 5.1(b) that Z_t is a unit root process. Also, one can see from Figure 5.1(c) that $u_t = Z_t - Z_{t-1}$ is autocorrelated. Indeed, the Ljung-Box test rejects the null hypothesis of un-autocorrelation of u_t.

Thus, for simplicity, the following nonparametric model is considered, by ignoring the price indices (p_t and p_t^*) for two countries

$$s_t = \beta(Z_t) + \varepsilon_t. \tag{5.16}$$

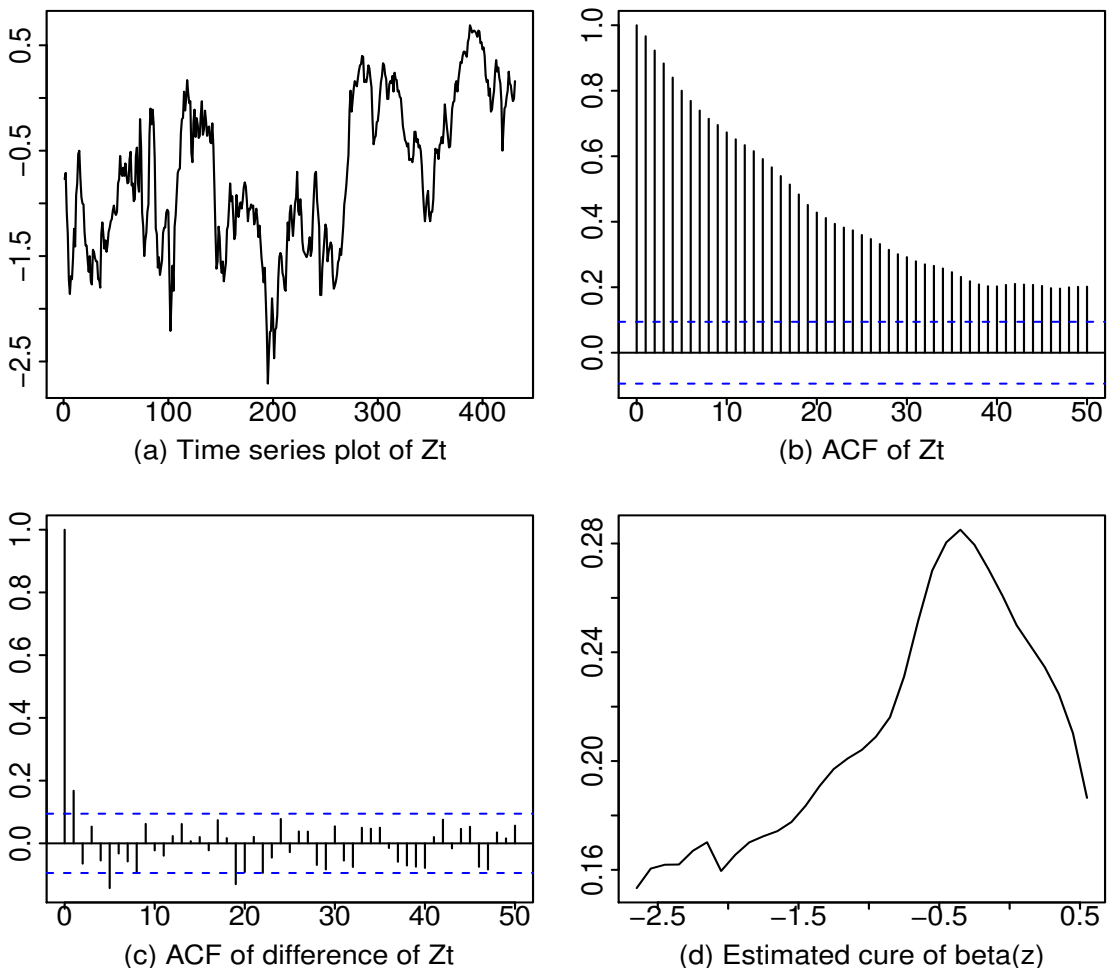

Figure 5.1: (a) Time series plot of Z_t; (b) ACF plot of Z_t; (c) ACF plot of $Z_t - Z_{t-1}$; (d) Estimated curve of $\beta(z)$.

The Epanechnikov kernel $K(u) = 0.75(1-u^2)_+$ is used, and the smoothing parameter h is selected by the least squares cross-validation method so that $h = 0.425$. Figure 5.1(d) depicts the nonparametric estimate of $\beta(z)$. From Figure 5.1(d), it is very interesting to learn that $\widehat{\beta}(z)$ is nonlinear and reaches its max when $z = -0.35$. Further, it is increasing if $z \leq -0.35$ and then it is decreasing when $z > -0.35$. Also, it is asymmetric and the left side has a longer tail. The reaction of exchange rate to the 10-year Treasury bond rate differential between the two nations is different based on the differential value. This means that when the US 10-year Treasury bond rate becomes much lower or higher than that for

the Canadian bond rate, the exchange rate between two nations becomes lower. In other words, the Canadian dollar is appreciated. Therefore, my analysis confirms that exchange rate between US and Canada depends on the interest rate differential between two nations.

Furthermore, the standard PPP model in (5.15) and the nonlinear model in (5.16) are extended to the following functional coefficient model as in Cai et al. (2009),

$$s_t = \beta_0(Z_t) + \beta_1(Z_t)p_t + \beta_2(Z_t)p_t^* + \varepsilon_t, \tag{5.17}$$

The estimation procedure given in (2.21) can be applied to estimating $\beta_1(z)$ and $\beta_2(z)$ for any given grid point z. It is clear that (5.17) is a generalization of (5.15) and (5.16). The

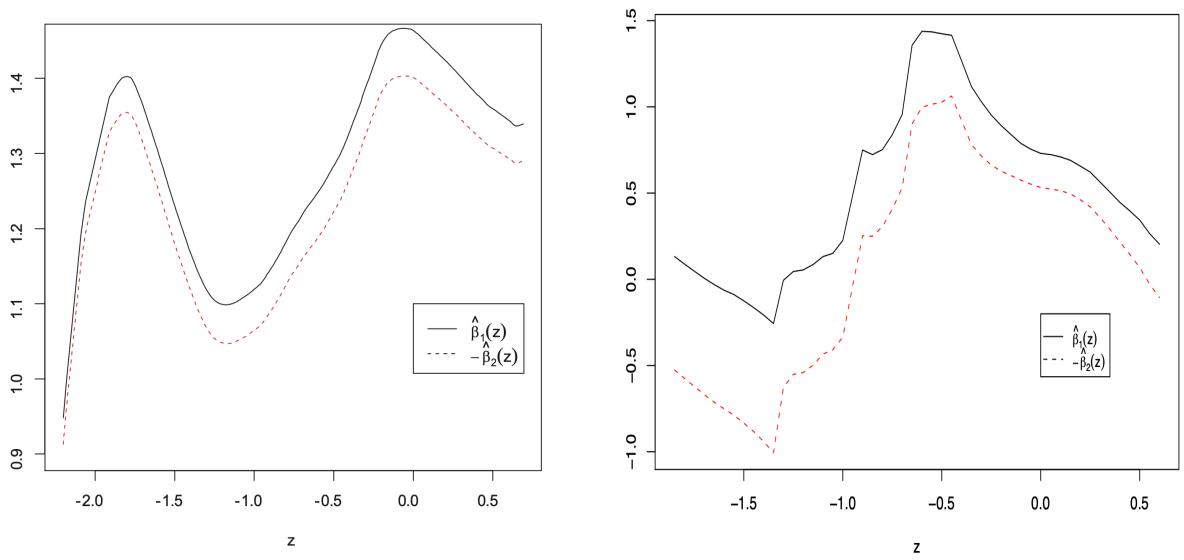

Figure 5.2: Plots for $\widehat{\beta}_1(z)$ in the back solid line and $-\widehat{\beta}_2(z)$ in the red dashed line. The left panel is for the US-Canada case and the right panel is for the US-China case.

left panel in Figure 5.2 plots $\widehat{\beta}_1(z)$ in the back solid line and $-\widehat{\beta}_2(z)$ in the red dashed line. Examining the graph, we find that the estimated coefficient curves $\widehat{\beta}_1(z)$ and $-\widehat{\beta}_2(z)$ have similar shapes for the US-Canada case. Therefore, in contrast to the result obtained from the linear model (5.15), the nonlinear co-integrated relationship of exchange rate and price indexes between Canada and US can be evidently found in the left panel in Figure 5.2, when allowing for the coefficients of the aggregate price indexes to vary with respect to a relevant macroeconomic variable: the 10-year Treasury bond rate differential between the two nations. Now, we compare the results with the long-run PPP theory between China and US using the same model as in (5.17) and estimation method. The estimation results are

displayed in the right panel in Figure 5.2, from which one can see clearly that the estimated coefficient curves $\widehat{\beta}_1(z)$ and $-\widehat{\beta}_2(z)$ also have similar shapes. This result also support our conclusion, the long-run PPP theory exists between China and US, from another way. In other words, the exchange rate between China and US is not depreciated.

5.6 Discussions

This chapter studies a nonparametric regression model for integrated time series data by considering using the local polynomial local constant fitting schemes to estimate the nonparametric function and derives the asymptotic properties of the proposed estimators. The theoretical results show that the asymptotic bias has the same as that for stationary covariates. But, the convergence rate for the nonstationary covariates is slower than that for the stationary covariates by a factor of $n^{-1/4}$. Further, the asymptotic distribution is not normal any more but just a mixed normal associated with the local time of a standard Brownian motion. Moreover, it shows that the asymptotic properties for both the local linear and local constant estimators are exactly same.

It would like to mention some interesting future research topics related to this chapter. First, it would be very useful and important to discuss how to select the data-driven (optimal) bandwidth theoretically and empirically. Second, it should allow the errors $\{\varepsilon_t\}$ to be serially correlated time series, say α-mixing, to be heteroscedastic, and to be correlated with covariates as in Wang and Phillips (2009a,b). Third, the model should include both stationary and nonstationary covariates. Finally, it is warranted to consider some extensions to other types of models like additive models, index models and varying coefficient models, and other types of non-stationarity such as nearly integrated processes; see, e.g., Bandi (2002), Torous et al. (2004), Campbell and Yogo (2006), Polk et al. (2006), Rossi (2007), Cai and Wang (2014), Cai et al. (2015), and Cai et al. (2017), which have a potential application in applied fields like economics and finance.

5.7 Theoretical Proofs

Before proving the main results of this chapter, we first give a few lemmas that will be used frequently in the proofs below. Throughout this section, C denotes a generic positive constant and it may take different values at different appearances.

To prove Theorem 5.1, define $G_j(u) = u^j K(u)$ for any $j \geq 0$. Then, it is easy to verify that $G_j(\cdot)$ is continuous and has a compact support. Also, both $G_j(\cdot)$ and $G_j^2(\cdot)$ are integrable. Also, define $S_n(z)$ as follows

$$S_n(z) = n^{-1/2} \sum_{t=1}^{n} K_h(Z_t - z) \begin{pmatrix} 1 \\ Z_{t,z,h} \end{pmatrix}^{\otimes 2} = \begin{pmatrix} S_{n,0}(z) & S_{n,1}(z) \\ S_{n,1}(z) & S_{n,2}(z \end{pmatrix}$$

where $Z_{t,z,h} = (Z_t - z)/h$ and for $0 \leq j \leq 2$,

$$S_{n,j}(z) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} K_{j,h}(Z_t - z)$$

with $K_{j,h}(u) = G_j(u/h)/h$. Then, re-express $S_{n,j}(z)$ as

$$S_{n,j}(z) = \frac{\beta_n}{n} \sum_{t=1}^{n} G_j(\beta_n(\gamma_n^{-1} Z_t + x_n)),$$

where $\beta_n = \sqrt{n}/h$, $\gamma_n = \sqrt{n}$, and $x_n = -z/\sqrt{n}$. Clearly, $x_n \to 0$ for any fixed z and $x_n = -a$ if $z = a\sqrt{n}$. Finally, let $\phi_\delta(x) = \exp(-x^2/2\,\delta^2)/\sqrt{2\,\pi\,\delta^2}$ for any $\delta > 0$ and $o_{L_2}(1)$ denote the convergence in L_2. Before proving the theorem, I first present some preliminary results. In what follows, it is assumed that Z_t satisfies (5.5).

Lemma 5.1: *Under assumption that the density of $\eta_{t,z}$ is bounded for all t,*

$$(i) \quad S_{n,j}(z) \xrightarrow{p} \begin{cases} \mu_j(K)\, L(1,0)/\sigma_0, & \text{if } z \text{ is fixed,} \\ \mu_j(K)\, L(1, a/\sigma_0)/\sigma_0, & \text{if } z = a\sqrt{n}, \end{cases}$$

and for any $p > 0$ and z,

$$(ii) \quad \mathbb{E}\,[S_{n,j}(z)] = O(1), \qquad \text{and} \qquad (iii) \quad \mathbb{E}\,[|K_{j,h}(Z_t - z)|^p] = O(t^{-1/2}\, h^{1-p}).$$

Note that the above results still hold if fixed z is changed to be any z_n satisfying $z_n/\sqrt{n} \to 0$.

Proof: To establish the first assertion, I use some results from Jeganathan (2004). Indeed, by Proposition 6 and Lemma 7 of Jeganathan (2004), for each $\delta > 0$,

$$S_{n,j}(z) = \frac{\mu_j(K)}{n} \sum_{t=1}^{n} \phi_\delta(\gamma_n^{-1} Z_t + x_n) + o_{L_2}(1).$$

Since $\phi_\delta(z)$ satisfies the Lipschitz condition and $x_n \to 0$,

$$S_{n,j}(z) = \frac{\mu_j(K)}{n} \sum_{t=1}^{n} \phi_\delta(\gamma_n^{-1} Z_t) + o_{L_2}(1) = \frac{\mu_j(K)}{n} \sum_{t=1}^{n} \phi_\delta(W_u(t/n)) + o_{L_2}(1)$$

in view of (5.6) and (5.7). By Lemma 9 of Jeganathan (2004), one has

$$S_{n,j}(z) = \mu_j(K) \int_0^1 \phi_\delta(W_u(s))ds + o_{L_2}(1).$$

An application of Proposition 11 of Jeganathan (2004) gives

$$S_{n,j}(z) = \mu_j(K)\, L(1,0)/\sigma_0 + o_{L_2}(1)$$

as $\delta \downarrow 0$. By the same token, it is easy to show the case of $x_n = -a$ ($z = a\sqrt{n}$). For assertion (ii), one has

$$\begin{aligned}
\mathbb{E}\left[S_{n,j}(z)\right] &= n^{-1/2} \sum_{t=1}^{n} \mathbb{E}\left[K_{j,h}(Z_t - z)\right] \\
&= n^{-1/2}\, h^{-1} \sum_{t=1}^{n} \int G_j(t^{1/2}u/h)\, f_{t,z}(u)du \\
&= n^{-1/2} \sum_{t=1}^{n} t^{-1/2} \int G_j(v)\, f_{t,z}(ht^{-1/2}v)dv \\
&\leq C\, n^{-1/2} \sum_{t=1}^{n} t^{-1/2} = O(1).
\end{aligned}$$

Finally, recall that $K_{j,h}(u) = h^{-1}G_j(u/h)$ and $G_j(u) = u^j\, K(u)$. It can be shown easily by the boundedness of $f_{t,z}(\cdot)$ that

$$\begin{aligned}
&\mathbb{E}\left[|K_{j,h}(Z_t - z)|^p\right] = h^{-p} \int |K_{j,h}(t^{1/2}u/h)|^p f_t(u)du \\
= \;& t^{-1/2}h^{1-p} \int |G_j(v)|^p f_{t,z}(t^{-1/2}hv)dv \leq Ct^{-1/2}h^{1-p}.
\end{aligned}$$

This proves the lemma. □

By Lemma 5.1, one has

$$S_n(z) = \begin{pmatrix} S_{n,0}(z) & S_{n,1}(z) \\ S_{n,1}(z) & S_{n,2}(z) \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & h^2\,\mu_2(K) \end{pmatrix} L(1,0)/\sigma_0\,\{1 + o_p(1)\},$$

which, by replacing Y_t in (5.4) by $Y_t = \beta(Z_t) + \varepsilon_t$, implies that

$$\widehat{\beta}(z) - \beta(z) \equiv \left[L(1,0)/\sigma_0\right]^{-1} \{B_n + C_n\}\,\{1 + o_p(1)\}, \tag{5.18}$$

where $B_n = n^{-1/2} \sum_{t=1}^{n} \left[\beta(Z_t) - \beta(z) - \beta'(z)(Z_t - z)\right] K_h(Z_t - z)$ and $C_n = n^{-1/2} \sum_{t=1}^{n} \varepsilon_t$ $K_h(Z_t - z)$. B_n and C_n are analyzed in Lemma 5.2 and Lemma 5.3 below.

Lemma 5.2: *Under Assumptions given in Theorem 5.1, then,*

$$B_n = h^2 B(z)[S_{n,2}(z)] + o_p(h^2) = h^2 B(z) L(1,0)/\sigma_0 + o_p(h^2).$$

Proof: Note that the proof is similar to that for Lemma 5.1. Similar to Lemma 5.1, one can show that

$$\begin{aligned} B_n &= n^{-1/2} \sum_{t=1}^{n} [\beta(Z_t) - \beta(z) - \beta'(z)(Z_t - z)] K_h(Z_t - z) \\ &= \frac{h^2}{2} \beta''(z) S_{n,2}(z)\{1 + o_p(1)\} \\ &= \frac{h^2}{2} L(1,0) \beta''(z) \mu_2(K)/\sigma_0\{1 + o_p(1)\}. \end{aligned}$$

This completes the proof of Lemma 5.2. □

Lemma 5.3: *Under Assumptions given in Theorem 5.1, then,*

$$n^{3/4} h^{1/2} C_n \xrightarrow{d} MN(\sigma_1^2),$$

where $MN(\sigma_1^2)$ *is a mixed normal with mean zero and covariance matrix* $\sigma_1^2 = \sigma_\varepsilon^2 \nu_0(K)$ $L(1,0)/\sigma_0$.

Proof: Clearly, $\mathbb{E}[C_n] = 0$ since $\mathbb{E}(\varepsilon_t | Z_t) = 0$. Also, by the assumptions that $\{\varepsilon_t\}$ is a martingale difference and $\mathbb{E}(\varepsilon_t^2 | Z_t) = \sigma_\varepsilon^2$, one can conclude that the conditional variance of $n^{1/4} h^{1/2} C_n$, given $\{Z_t\}$, is

$$D_n = \frac{\sigma_\varepsilon^2 h}{\sqrt{n}} \sum_{t=1}^{n} K_h^2(Z_t - z).$$

Similar to the proof of Lemma 5.1, one can show that

$$D_n = \sigma_\varepsilon^2 \nu_0(K) L(1,0)/\sigma_0 + o_p(1).$$

Finally, by the central limit theorem for a martingale difference, see, e.g., Hall and Heyde (1980), one obtains the conditional limiting distribution of C_n given $\{Z_t\}$,

$$n^{1/4} h^{1/2} C_n \xrightarrow{d} MN(\sigma_1^2).$$

This proves the lemma. □

Proof of Theorem 5.1: It is easy to check from Lemmas 5.1 and 5.2 that

$$C_n = h^2 B(z)\, L(1,0)/\sigma_0 + o_p(h^2).$$

Therefore, by (5.18) and Lemma 5.3, one has

$$\begin{aligned} & n^{1/4}h^{1/2}\left[\widehat{\beta}(z) - \beta(z) - h^2 B(z) + o_p(h^2)\right] \\ = \;& \sigma_0\, [L(1,0)]^{-1}\; n^{1/4}h^{1/2}\, C_n\{1+o_p(1)\} \xrightarrow{d} MN(\sigma_\beta^2), \end{aligned}$$

which concludes the proof of the theorem. □

Proof of Theorem 5.3: It is easy to see from Lemma 5.1 that

$$\widetilde{\beta}(z) - \beta(z) \equiv \{E_n + C_n\}\,/S_{n,0}(z) = [L(1,0)/\sigma_0\,]^{-1}\,\{E_n + C_n\}\,\{1+o_p(1)\},$$

where $E_n = n^{-1/2}\sum_{t=1}^n [\beta(Z_t) - \beta(z)]\, K_h(Z_t - z)$. Similar to Lemma 5.2, one has

$$E_n = \left[h\,\beta'(z)\, S_{n,1}(z) + \frac{h^2}{2}\,\beta''(z)\, S_{n,2}(z)\right]\{1+o_p(1)\} = \frac{h^2}{2}\; L(1,0)\,\beta''(z)\,\mu_2(K)/\sigma_0\{1+o_p(1)\}$$

by Lemma 5.4 below. By Lemma 5.3, similar to the proof of Theorem 5.1, Theorem 5.3 is proved. □

Lemma 5.4: *Under Assumptions given in Theorem 5.1, then,*

$$\mathbb{E}[E_n] = O(h^2 n^{-1/2}\ln(n)) + O(h^2).$$

Proof: I first compute the following intermediate quantity. A simple calculation leads to

$$\begin{aligned} & \mathbb{E}[(\beta(Z_t) - \beta(z))K_h(Z_t - z)] \\ = \;& t^{-1/2}\int [\beta(z + h\,v) - \beta(z)]\, K(v)\, f_{t,z}(t^{-1/2}hv)dv. \\ \approx \;& t^{-1/2}\int [\beta'(z)hv + h^2\beta''(z)v^2/2][f_{t,z}(0) + f'_{t,z}(0)t^{-1/2}hv]K(v)dv \\ = \;& h^2 t^{-1}\beta'(z)f'_{t,z}(0)\mu_2(K) + \frac{1}{2}\, h^2 t^{-1/2}\beta''(z)f_{t,z}(0)\mu_2(K), \end{aligned}$$

which implies that the order of the second term dominates the order of the first term. Therefore,

$$\begin{aligned} \mathbb{E}[E_n] \;&\approx\; h^2\beta'(z)\mu_2(K)n^{-1/2}\sum_{t=1}^n t^{-1} f'_{t,z}(0) + \frac{1}{2}\, h^2\mu_2(K)\beta''(z)n^{-1/2}\sum_{t=1}^n t^{-1/2} f_{t,z}(0) \\ &=\; h^2\beta'(z)\mu_2(K)^{-1/2}n^{-1/2}O(\ln(n)) + \frac{1}{2}\, h^2\mu_2(K)\beta''(z)n^{-1/2}O(n^{1/2}) \\ &=\; O(h^2 n^{-1/2}\ln(n)) + O(h^2). \end{aligned}$$

This concludes the proof of the lemma. □

Chapter 6

Nonparametric Estimation Equations

6.1 Introduction

Various methodologies have been used in many areas to estimate nonparametric regression functions in nonlinear time series analysis. For example, to name just a few, see Robinson (1984), Collomb and Härdle (1986), Boente and Fraiman (1989, 1990), Truong (1992), Laïb and Ould-Saïd (2000), and Cai and Ould-Saïd (2003) for robust nonparametric estimation, Robinson (1983), Roussas (1969a,b, 1990), Auestad and Tjøstheim (1990), Härdle and Vieu (1992), Masry and Tjøstheim (1995), Masry (1996a,b), and Masry and Fan (1997) for kernel type estimations, and Cai (2003a) and Cai et al. (2001) for quasi-likelihood estimation, among others. All methods mentioned above can be regarded as a special case of the so-called nonparametric estimation equations (NEE) method, which can be viewed as a generalization of the parametric estimation equations approach, a widely popular general methodology. The NEE method is the topic of this chapter. To the best of my knowledge, the use of such a NEE technique for nonlinear time series analysis was first advocated by Cai (2003b).

The estimation equations method is attractive and popular in parametric problems, especially in biostatistics, due to its yielding consistent estimators with asymptotically valid inferences obtained via the sandwich formula, see, e.g., Carroll et al. (1998). The main aim of this chapter is to discuss NEE and to provide a general formula to estimate unknown functions for both nonlinear discrete and continuous time series data. The idea of the NEE is to utilize the local linear fitting with local weighting of the estimation equations, as proposed in Cai (2003b). The main advantages of using estimation equations are that it does not require a likelihood, but only an unbiased estimating function, and it provides a general formula for the estimation of variance without going through case-by-case. Note that the

NEE idea was introduced by Fan and Gijbels (1996) and explored by Carroll et al. (1998) for the iid samples and that Cai (2003a) gave a detailed discussion on how to apply this methodology to real examples for both continuous and discrete nonlinear time series data.

6.2 Estimating Equations

6.2.1 Nonparametric Estimating Equations

To motivate the NEE, the estimating equations method for a parametric model is recalled. Suppose that $g(\cdot)$ is an unknown parameterized function of interest, say, $g(\cdot) = g(\cdot, \boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \mathbb{R}^k$ is an unknown parameter vector. Let $\{(\mathbf{X}_t, Y_t)\}_{t=1}^n$ be n realizations from an underlying stationary sequence, where $\mathbf{X}_t$ in $\mathbb{R}^p$ consists of either exogenous variables, or the past observations. Assume that $(\mathbf{X}_t, Y_t)$ has the same distribution as $(\mathbf{X}, Y)$. For the given sample $\{(\mathbf{X}_t, Y_t)\}_{t=1}^n$, the estimator of the parameter vector $\boldsymbol{\theta}$, $\widehat{\boldsymbol{\theta}}$, is the solution of the *parametric estimating equations*

$$\sum_{t=1}^{n} q\left\{Y_t, g(\mathbf{X}_t, \boldsymbol{\theta})\right\} g'(\mathbf{X}_t, \boldsymbol{\theta}) = \mathbf{0}, \tag{6.1}$$

where $q(\cdot, \cdot)$ is an objective function and $g'(\mathbf{X}, \boldsymbol{\theta}) = \partial g(\mathbf{X}, \boldsymbol{\theta})/\partial \boldsymbol{\theta}$. Clearly, equation (6.1) is a generalization of a normal equation of the least squares problem for nonlinear parametric model ($Y_t = g(\mathbf{X}_t, \boldsymbol{\theta}) + u_t$), where $q\left\{Y_t, g(\mathbf{X}_t, \boldsymbol{\theta})\right\} = -2(Y_t - g(\mathbf{X}_t, \boldsymbol{\theta}))$. Also, it covers an application of generalized method of moments (GMM) as in Hansen (1982, 2001) and Jagannathan et al. (2002), which is widely popular in economics and finance for the estimation of parameters satisfying moments or orthogonality conditions. In other words, when the number of parameters to be estimated is the same as the number of orthogonality conditions, the GMM estimator is the value that satisfies (6.1), see, e.g., Hamilton (1994). Finally, for more applications, see, for example, Li (2001) and references therein for estimation equations from maximizing log (quasi-)likelihood with the iid samples.

However, different from Hansen (1982, 2001) and Jagannathan et al. (2002) for estimating parameters, we consider the case that the form of $g(\cdot)$ in (6.1) is completely unknown, so that it is of interest to estimate $g(\mathbf{x})$ at any given grid point $\mathbf{x}_0$. Particularly, as in Jagannathan et al. (2002), for applying nonlinear rational expectation models and the assessment of asset pricing models using the stochastic discount factor representation, one ends up the following

very general conditional moment conditions

$$\mathbb{E}\left[Q(Y_t, g(\mathbf{X}_t)) \,|\, \mathbf{X}_t\right] = 0, \tag{6.2}$$

where $Q(\cdot)$ is a known function. See (5) and (11) in Jagannathan et al. (2002) for macroeconomic and financial applications. However, instead of estimating functionals, Jagannathan et al. (2002) converted their (5) and (11) into the unconditional moment conditions and then employed the GMM of Hansen (1982) to estimate parameters. Furthermore, as in Aït-Sahalia and Brant (2001), to update portfolio allocations, one considers the following conditional expected utility problem

$$\boldsymbol{\alpha}(\mathbf{X}_t) = \text{argmax}\, \mathbb{E}\left[U(W_{t+1}) \,|\, \mathbf{X}_t\right],$$

where $U(\cdot)$ is a utility function and $W_{t+1} = \alpha(X_t)^\top \mathbf{R}_{t+1}$ is a portfolio with some risky assets $\mathbf{R}_{t+1}$. Of interest is to estimate the portfolio allocations $\boldsymbol{\alpha}(\mathbf{x})$ nonparametrically. To do so, we have the first order condition as

$$\mathbb{E}\left[U'(W_{t+1})\, \mathbf{R}_{t+1} \,|\, \mathbf{X}_t\right] = \mathbb{E}\left[Q(Y_t, \boldsymbol{\theta}) \,|\, \mathbf{X}_t\right] = 0, \tag{6.3}$$

where $U'(\cdot)$ is the derivative of $U(\cdot)$ if $U(\cdot)$ is assumed to be differentiable, $Q(Y_t, g(\mathbf{X}_t)) = U'(W_{t+1})\, \mathbf{R}_{t+1}$, and $g(\mathbf{X}_t) = \boldsymbol{\alpha}(\mathbf{X}_t)$. Another example application is the nonparametric probit model $Y_t = I[g(\mathbf{X}_t) + e_t \geq 0]$, where $I(\cdot)$ denotes the indicator function that equals one if its argument is true and zero otherwise, $g(\mathbf{x})$ is an unknown function to be estimated, and e_t is a standard normal independent of $\mathbf{X}$, or has some other known distribution. Then, estimation could be based on (6.2), where the function $Q(\cdot)$ is $Q(Y, g) = Y - F_e(g)$, where $F_e(\cdot)$ is the known cumulative distribution function of e_t. Nonparametric censored or truncated regression would have a similar form. From the above aforementioned examples, one can see that it needs to consider the setting in (6.2) with unknown $g(\cdot)$.

Assume that the second order partial derivative of $g(\mathbf{x})$ exists and is continuous at a given grid point $\mathbf{x}_0$, so that in a neighborhood of $\mathbf{x}_0$, $g(\mathbf{X}_i)$ can be approximated by a linear function as $g(\mathbf{X}_t) \approx \beta_1 + \boldsymbol{\beta}_2^\top (\mathbf{X}_t - \mathbf{x}_0) = \boldsymbol{\beta}^\top \mathbf{H}\, \mathbf{Z}_t$, where $\boldsymbol{\beta}^\top = (\beta_1,\, \boldsymbol{\beta}_2^\top)$, $\mathbf{Z}_i^\top = (1,\, (\mathbf{X}_t - \mathbf{x}_0)^\top / h)$, and $\mathbf{H} = \text{diag}\{1,\, h\, \mathbf{I}_p\}$. For the given sample $\{(\mathbf{X}_t,\, Y_t)\}_{t=1}^n$, the *locally linear nonparametric estimating equations* are defined as

$$S_n(\boldsymbol{\beta}) = \sum_{t=1}^{n} Q\left\{Y_t,\, \boldsymbol{\beta}^\top \mathbf{H}\, \mathbf{Z}_t\right\} \mathbf{Z}_t\, K_h(\mathbf{X}_t - \mathbf{x}_0) = \mathbf{0}, \tag{6.4}$$

where $K_h(\cdot) = K(\cdot/h)/h^p$. Clearly, one can see that (6.4) has $p+1$ parameters β_1 and $\boldsymbol{\beta}_2$ and $p+1$ equations (local constant), if $g(\cdot)$ is a scalar function, so that (6.4) should have solutions. Let $\widehat{\beta}_1$ and $\widehat{\boldsymbol{\beta}}_2$ be the roots of (6.4). The NEE estimates of $g(\mathbf{x}_0)$ and $g'(\mathbf{x}_0)$ are $\widehat{g}(\mathbf{x}_0) = \widehat{\beta}_1$ and $\widehat{g}'(\mathbf{x}_0) = \widehat{\boldsymbol{\beta}}_2$, respectively. Note that the counterpart of (6.4) for the iid samples was proposed by Fan and Gijbels (1996) and studied by Carroll et al. (1998). Also note that (6.4) can be easily extended to the local polynomial estimating equations if $g(\cdot)$ has the $(q+1)$th order derivative continuous at $\mathbf{x}_0$. If there is no $\boldsymbol{\beta}_2$, (6.4) is the kernel estimation equations, which was applied by Aït-Sahalia and Brant (2001) to estimate $\boldsymbol{\alpha}(\cdot)$ in (6.3). Finally, the NEE was generalized to a more general setting, termed as the nonparametric generalized method of moments (NPGMM) by Cai and Li (2008), which is formulated as follows. From (6.2), for any function $\boldsymbol{M}(\cdot)$ at an appropriate dimension, one has

$$\mathbb{E}\left[\boldsymbol{M}(\mathbf{X}_t)Q(Y_t, g(\mathbf{X}_t)) \,|\, \mathbf{X}_t\right] = 0.$$

Then, (6.4) is generalized to

$$\sum_{t=1}^{n} \boldsymbol{M}(\mathbf{X}_t)Q\left\{Y_t,\, \boldsymbol{\beta}^\top \mathbf{H}\, \mathbf{Z}_t\right\} K_h(\mathbf{X}_t - \mathbf{x}_0) = \mathbf{0}. \tag{6.5}$$

By choosing an optimal $\boldsymbol{M}(\mathbf{X}_t)$ with an appropriate dimension, one can obtain $\widehat{\beta}_1$ and $\widehat{\boldsymbol{\beta}}_2$, which are the NPGMM estimators for β_1 and $\boldsymbol{\beta}_2$, respectively. The reader is referred to the paper by Cai and Li (2008) for more details on the methodology and theory as well as how to choose the optimal $\boldsymbol{M}(\mathbf{X}_t)$.

Remark 6.1: *In many applications, the unknown function $g(\cdot)$ is a mean regression function or a transformed conditional mean function in generalized linear models. In such a case, if the estimating equations involve a variance function $V(\mathbf{x}_0)$, which is not a function of $g(\mathbf{x}_0)$, the NEE method is still applicable, because of the local homoscedasticity: $V(\mathbf{X}_t) \approx V(\mathbf{x}_0)$ for $\mathbf{X}_t$ in a neighborhood of $\mathbf{x}_0$. An efficient approach is to use a partial-residual method, such as, estimating $g(\cdot)$ regarding $V(\cdot)$ locally as a constant and then applying a local modeling to $V(\cdot)$ by using a different bandwidth.*

If the number of equations in (6.4) is more than the dimension of $\boldsymbol{\beta}$, (6.4) might not have solutions. so that the model is over-identified. The simplest way to overcome this difficulty is just to apply the idea as in Lewbel (2007), the so-called local GMM (LGMM), defined as

$$\widehat{\boldsymbol{\beta}} = \arg\inf_{\boldsymbol{\beta}} \; S_n(\boldsymbol{\beta})^\top \Omega_n S_n(\boldsymbol{\beta}), \tag{6.6}$$

where Ω_n is a finite positive definite matrix for all n, as is $\Omega = \lim_{n\to\infty} \Omega_n > 0$. For details, see the paper by Lewbel (2007). Note that the LGMM might have an ability to take care of the heteroscedasticity issue as mentioned above. The second way is to use the so-called NPGMM as in (6.5). Note that by a comparison of (6.6) with (6.5), one can see that this is difference between the local GMM and nonparametric GMM. In the local GMM as in (6.6), the kernel weight is used twice due to the sandwich formula, but it is used once in (6.5).

Remark 6.2: *Finding roots in (6.4) can be costly by using the iteration algorithm, because one needs to find usually hundreds of roots to (6.4). To save computational cost, the one-step Newton-Raphson method, proposed by Cai et al. (2000), can be applied here. For more details, see the article just cited.*

6.2.2 Examples

Equations such as (6.2) are already commonly used in the time series literature, although not at the level of generality used presently. Here are a few more examples for nonlinear time series:

A. The Nadaraya-Watson kernel regression has $Q(Y, s) = Y - s$ and $\boldsymbol{\beta}_2 = \mathbf{0}$. See Robinson (1983), Roussas (1969a,b, 1990), Auestad and Tjøstheim (1990), Härdle and Vieu (1992), Masry and Tjøstheim (1995), among others.

B. The local linear (polynomial) regression has $Q(Y, s) = Y - s$. See Masry (1996a,b), and Masry and Fan (1997).

C. The nonparametric quasi-likelihood regression is based on

$$Q(Y, s) = (Y - \mu(s))\, \mu'(s)/V(\mu(s)),$$

when the conditional mean and variance of a response Y are related through $\mathbb{E}(Y \mid \mathbf{X} = \mathbf{x}) = \mu(g(\mathbf{x}))$ and $\mathrm{Var}(Y \mid \mathbf{X} = \mathbf{x}) = \sigma^2\, V\{\mu(g(\mathbf{x}))\}$ for some known functions $\mu(\cdot)$ and $V(\cdot)$. See Cai (2003a) and Cai et al. (2001).

D. The robust regression has $Q(Y, s) = \psi(Y - s)$, where $\psi(\cdot)$, say the Huber's function, is the derivative of some nonnegative loss function. See Robinson (1984), Collomb and Härdle (1986), Boente and Fraiman (1989, 1990), Truong (1992), and Laïb and Ould-Saïd (2000) for kernel method, and Cai and Ould-Saïd (2003) for local linear fitting.

E. The expectile regression problem consists in minimizing

$$\sum_{t=1}^{n} \ell_\alpha \left\{ Y_t - \boldsymbol{\beta}^\top \mathbf{H}\, \mathbf{Z}_t \right\} K_h(\mathbf{X}_t - \mathbf{x}_0),$$

where $\ell_\alpha(v) = \alpha\, v^2\, I(v > 0) + (1-\alpha)\, v^2\, I(v \le 0)$ for $0 \le \alpha \le 1$. See Section 6.3 in Fan and Gijbels (1996), Yao and Tong (1996), and Cai et al. (2018) for details.

F. Finally, the NEE method is particularly appealing in the functional-coefficient modeling, in which the conditional mean is assumed to be of the form

$$\mathbb{E}(Y_t \,|\, U_t,\, \mathbf{X}_t) = \mu \left\{ \sum_{j=1}^{p} g_j(U_t)\, X_{jt} \right\} \tag{6.7}$$

with a smoothing variable U_t, where $\mathbf{X}_t = (X_{1t}, \cdots, X_{pt})^\top$ and $\mu(\cdot)$ is a link function. Then, the version of (6.4) for the functional-coefficient models is

$$\sum_{t=1}^{n} Q \left\{ Y_t,\, \mu \left(\boldsymbol{\theta}^{*\top}\, \mathbf{X}_t^* \right) \right\}\, \mu' \left(\boldsymbol{\theta}^{*\top}\, \mathbf{X}_t^* \right) \mathbf{X}_t^*\, K_h(U_t - u_0) = \mathbf{0},$$

where $\mathbf{X}_t^* = (\mathbf{X}_t^\top,\, (U_t - u_0)\, \mathbf{X}_t^\top)^\top$. Model (6.7) has been studied extensively, both theoretically and empirically. For example, see Chen and Tsay (1993), Hong and Lee (2003), Xia and Li (1999), Cai et al. (2000), and Cai and Tiwari (2000) for the Gaussian case, and Cai (2003a) for the non-Gaussian context. Clearly, (6.7) is a natural extension of the parametric Markov regression models for time series proposed by Zeger and Qaqish (1988). Such an extension makes the fitted model more appealing in practice in four directions: In exploring the nonlinear feature, in allowing the appreciable flexibility on the structure of fitted model in avoiding the curse of dimensionality, in reducing the possible modeling bias, and in letting the data select a model describing themselves well. Note that Cai (2003a) applied the model (6.7) to study an environmental problem by using a functional coefficient Poisson time series model.

6.3 Asymptotic Theory

It is well known that among various mixing conditions used in the literature, α-mixing is reasonably weak and is known to be fulfilled for many stochastic processes, including many familiar linear and nonlinear time series models, see, for example, Gorodetskii (1977), Withers (1981), Auestad and Tjøstheim (1990), Chen and Tsay (1993), and Masry and

Tjøstheim (1995, 1997). Therefore, the asymptotic properties are established under α-mixing assumption. The following technical conditions are imposed to facilitate the proofs of the weak consistency and asymptotic normality of the NEE estimators although some of them may not be the weakest possible.

Assumptions:

(F1) The kernel function $K(\cdot)$ is symmetric and has a compact support, say $[-1,\ 1]$.

(F2) The density $f(\cdot)$ of $\mathbf{X}$ and $Q(y,\ g(\mathbf{x}))$ are continuous at $\mathbf{x}_0$ for any y and $f(\mathbf{x}_0) > 0$. Also, the second partial derivative of $g(\mathbf{x})$ exists and is continuous at $\mathbf{x}_0$ and that for any y, $Q'(y,\ s) = (\partial/\partial s)\, Q(y,\ s)$ exists almost everywhere and is continuous at $\mathbf{x}_0$.

(F3) The functions $\mathbb{E}\left[Q(Y,\ s)\middle|\, \mathbf{X} = \mathbf{x}\right]$, $\mathbb{E}\left[Q'(Y,\ s)\middle|\, \mathbf{X} = \mathbf{x}\right]$, and $\mathbb{E}\left[Q^2(Y,\ s)\middle|\, \mathbf{X} = \mathbf{x}\right]$ are continuous at $\mathbf{x}_0$ for s in a neighborhood of $g(\mathbf{x}_0)$, and there exists a constant $\gamma > 0$ such that $\mathbb{E}\left[|Q(Y,\ s)|^{2(1+\gamma)} \,\middle|\, \mathbf{X} = \mathbf{x}\right]$ and $\mathbb{E}\left[|Q'(Y,\ s)|^{2(1+\gamma)} \,\middle|\, \mathbf{X} = \mathbf{x}\right]$ are bounded in a neighborhood of $\mathbf{x}_0$ for s in a neighborhood of $g(\mathbf{x}_0)$.

(F4) The series $\{(\mathbf{X}_t,\ Y_t)\}_{t=1}^{\infty}$ is stationary α-mixing with mixing coefficient $\alpha(n)$ satisfying $\alpha(n) = O\left(n^{-(2+\gamma)(1+\gamma)/\gamma}\right)$, where γ is given in Assumption F3.

(F5) For all $t \geq 2$, the joint density of $\mathbf{X}_1$ and $\mathbf{X}_t$ is bounded, and the functions

$$\mathbb{E}\left\{\left[\{Q(Y_1,\ g(\mathbf{X}_1))\}^2 + \{Q(Y_t,\ g(\mathbf{X}_t))\}^2\right]^2 \,\middle|\, \mathbf{X}_1 = \mathbf{u},\ \mathbf{X}_t = \mathbf{v}\right\}$$

and

$$\mathbb{E}\left\{\left[\{Q'(Y_1,\ g(\mathbf{X}_1))\}^2 + \{Q'(Y_t,\ g(\mathbf{X}_t))\}^2\right]^2 \,\middle|\, \mathbf{X}_1 = \mathbf{u},\ \mathbf{X}_t = \mathbf{v}\right\}$$

are bounded for all $\mathbf{u}$ and $\mathbf{v}$ in a neighborhood of $\mathbf{x}_0$.

(F6) The second derivative $Q''(y,\ g(\mathbf{x}))$ exists almost everywhere and $\mathbb{E}[|Q''(Y,\ s)|\,|\,\mathbf{X} = \mathbf{x}]$ is bounded for all $\mathbf{x}$ in a neighborhood of $\mathbf{x}_0$ and s in a neighborhood of $g(\mathbf{x}_0)$, where $Q''(y,\ s) = (\partial^2/\partial s^2)\, Q(Y,\ s)$.

(F7) $n\, h^{p\,[1+2/(1+\gamma)]} \to \infty$, where γ is given in Assumption F3.

Remark 6.3: *Assumption F1 is imposed for brevity of proofs and it could be removed at the price of lengthier proofs. In particular, the Gaussian kernel is allowed. Because Assumption F7 is satisfied by the bandwidths of optimal size (i.e., $h \approx n^{-1/(p+4)}$) if $\gamma > p/2 - 1$, we do not concern ourselves with such refinements.*

Remark 6.4: *If $Q'(y,\, g(\mathbf{x}))$ is not continuous or $Q''(y,\, g(\mathbf{x}))$ does not exist, one needs some assumptions on $Q'(y,\, g(\mathbf{x}))$ so that the asymptotic results continue to hold with some modifications of the technical proofs; see Cai and Ould-Saïd (2003) for the robust regression setting. In the proofs of the theorems below, it is assumed without loss of generality that $Q''(y,\, g(\mathbf{x}))$ exists, which is satisfied for most regression problems.*

Remark 6.5: *If $\mathbf{X}_t$ is a discrete random variable, it is assumed that for any $\mathbf{j}$, $f(\mathbf{j}) = \mathbb{P}(\mathbf{X}_i = \mathbf{j}) = \int_{\mathbf{j}}^{\mathbf{j+1}} \phi(\mathbf{x}) d\mathbf{x}$ where $\phi(\cdot)$ is a density function on $\mathbb{R}^p$. This assumption reflects the fact that the sparse asymptotics depicts the performance of the estimator when the probabilities of $\mathbf{X}_t$ falling into each cell are small. This is the case when the smoothing is most relevant; see Cai (2003a) and Cai et al. (2001) for the detailed discussions on the sparse asymptotics for discrete time series data. Also, it ensures that $f(\cdot)$, the marginal density of $\mathbf{X}_t$, is continuous.*

Theorem 6.1: *Under Assumptions F1 – F6, $\widehat{g}(\mathbf{x}_0)$ converges to $g(\mathbf{x}_0)$ in probability.*

Theorem 6.2: *Under Assumptions F1 – F7,*

$$\sqrt{n\, h^p} \left[\widehat{g}(\mathbf{x}_0) - g(\mathbf{x}_0) - \frac{h^2}{2}\, tr\{g''(\mathbf{x}_0)\, \boldsymbol{\mu}_2\} + o_p(h^2) \right] \xrightarrow{d} N\left\{0,\, \sigma^2(\mathbf{x}_0)\right\},$$

where

$$\sigma^2(\mathbf{x}_0) = \frac{\nu_0(K)\, \mathbb{E}\left\{Q^2(Y,\, g(\mathbf{X}))\,|\, \mathbf{X} = \mathbf{x}_0\right\}}{f(\mathbf{x}_0)\ \left[E\{Q'(Y,\, g(\mathbf{X}))\,|\, \mathbf{X} = \mathbf{x}_0\}\right]^2}.$$

6.4 Theoretical Proofs

The same notation as in Sections 6.2 and 6.3 is used. To prove the theorems, similar to the proofs in Section 2.5.6, the following lemma is needed. First, define

$$\mathbf{C}_{n,1}(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} Q\left\{Y_i,\, \boldsymbol{\beta}^\top \mathbf{H}\, \mathbf{Z}_i\right\}\ \mathbf{Z}_i\ K_h(\mathbf{X}_i - \mathbf{x}_0)$$

and

$$\mathbf{C}_{n,2}(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} Q\left\{Y_i,\, \boldsymbol{\beta}^\top \mathbf{H}\, \mathbf{Z}_i\right\}\ \mathbf{Z}_i\, \mathbf{Z}_i^\top\ K_h(\mathbf{X}_i - \mathbf{x}_0).$$

Set

$$\mathbf{B}_1 = \begin{pmatrix} \mu_0(K) & \mathbf{0}^\top \\ \mathbf{0} & \boldsymbol{\mu}_2(K) \end{pmatrix} \qquad \text{and} \qquad \mathbf{B}_2 = \begin{pmatrix} \nu_0(K) & \mathbf{0}^\top \\ \mathbf{0} & \boldsymbol{\nu}_2(K) \end{pmatrix}.$$

Lemma 6.1: Under Assumptions F2 – F5,

$$\mathbb{E}[\mathbf{C}_{n,1}(\boldsymbol{\beta})] = \mathbf{M}_1(\boldsymbol{\beta}) + o(1) \qquad \text{and} \qquad n\,h^p\,\mathrm{Var}(\mathbf{C}_{n,1}) = \mathbf{M}_2(\boldsymbol{\beta}) - \mathbf{M}_1(\boldsymbol{\beta})\,\mathbf{M}_1(\boldsymbol{\beta})^\top + o(1),$$

where

$$\mathbf{M}_1(\boldsymbol{\beta}) = f(\mathbf{x}_0)\,\mathbb{E}[Q(Y,\,\beta_1)\,|\,\mathbf{X}=\mathbf{x}_0]\begin{pmatrix}\mu_0\\ \boldsymbol{\mu}_1\end{pmatrix} \quad \text{and} \quad \mathbf{M}_2(\boldsymbol{\beta}) = f(\mathbf{x}_0)\,\mathbb{E}\left[Q^2(Y,\,\beta_1)\,|\,\mathbf{X}=\mathbf{x}_0\right]\,\mathbf{B}_2.$$

Furthermore,

$$\mathbb{E}[\mathbf{C}_{n,2}(\boldsymbol{\beta})] = f(\mathbf{x}_0)\,\mathbb{E}[Q'(Y,\,\beta_1)\,|\,\mathbf{X}=\mathbf{x}_0]\,\mathbf{B}_1 + o(1) \qquad \text{and} \qquad \mathrm{Var}(\mathbf{C}_{n,2}(\boldsymbol{\beta})) = O\left((n\,h^p)^{-1}\right).$$

Proof: Let $\boldsymbol{\xi}_i = Q\left\{Y_i,\,\boldsymbol{\beta}^\top\,\mathbf{H}\,\mathbf{Z}_i\right\}\,\mathbf{Z}_i\,K_h(\mathbf{X}_i-\mathbf{x}_0) = (\xi_{i,1},\,\cdots,\,\xi_{i,p})^\top$. By the stationarity of $\{\boldsymbol{\xi}_i\}$,

$$\begin{aligned}
\mathbb{E}[\mathbf{C}_{n,1}(\boldsymbol{\beta})] &= \mathbb{E}[\boldsymbol{\xi}_1] \\
&= \int \mathbb{E}\left[Q\left\{Y,\,\beta_1 + h\,\boldsymbol{\beta}_2^\top\,\mathbf{u}\right\}\,\middle|\,\mathbf{X}=\mathbf{x}_0+h\,\mathbf{u}\right]\,\begin{pmatrix}1\\ \mathbf{u}\end{pmatrix}\,K(\mathbf{u})\,f(\mathbf{x}_0+h\,\mathbf{u})\,d\mathbf{u} \\
&\to \mathbf{M}_1(\boldsymbol{\beta}) \qquad (6.8)
\end{aligned}$$

by Assumptions F2 and F3 and

$$n\,h^p\,\mathrm{Var}[\mathbf{C}_{n,1}(\boldsymbol{\beta})] = h^p\,\mathrm{Var}(\boldsymbol{\xi}_1) + 2\,h^p\,\sum_{i=2}^{n}\left(1-\frac{i-1}{n}\right)\,\mathrm{Cov}(\boldsymbol{\xi}_1,\,\boldsymbol{\xi}_i). \qquad (6.9)$$

Similar to (6.8), the first term on the right-hand side of (6.9) converges to $\mathbf{M}_2(\boldsymbol{\beta})-\mathbf{M}_1(\boldsymbol{\beta})\,\mathbf{M}_1(\boldsymbol{\beta})^\top$. It suffices to show that the second term on the right-hand side of (6.9) converges to $\mathbf{0}$. To this end, the sum is split into two terms as follows

$$\sum_{i=2}^{n}|\mathrm{Cov}(\boldsymbol{\xi}_1,\,\boldsymbol{\xi}_i)| = \sum_{i=2}^{d_n}|\mathrm{Cov}(\boldsymbol{\xi}_1,\,\boldsymbol{\xi}_i)| + \sum_{i=d_n+1}^{n}|\mathrm{Cov}(\boldsymbol{\xi}_1,\,\boldsymbol{\xi}_i)| \equiv \mathbf{J}_{n,1}+\mathbf{J}_{n,2}$$

for some positive integers $\{d_n\}$ such that $d_n\,h^p \to 0$. By conditioning on $\mathbf{X}_1$ and $\mathbf{X}_i$ and using Assumption F5, one has

$$|\mathrm{Cov}(\boldsymbol{\xi}_1,\,\boldsymbol{\xi}_i)| \le C\,\mathbb{E}\left[\left|\mathbf{Z}_1\,\mathbf{Z}_i^\top\right|\,K_h(\mathbf{X}_1-\mathbf{x}_0)\,K_h(\mathbf{X}_i-\mathbf{x}_0)\right] = O(1),$$

which implies that

$$\mathbf{J}_{n,1} = O(d_n\,h^p) = o(1). \qquad (6.10)$$

For $\mathbf{J}_{n,2}$, an application of the Davydov's inequality, see, e.g., Hall and Heyde (1980), concludes that the (k,l)th element of $\mathrm{Cov}(\boldsymbol{\xi}_1,\, \boldsymbol{\xi}_i)$ is bounded by

$$|\mathrm{Cov}(\boldsymbol{\xi}_1,\, \boldsymbol{\xi}_i)_{k,l}| \leq C\, [\alpha(i-1)]^{\gamma/(2+\gamma)}\, \left[E|\boldsymbol{\xi}_1|_k^{2+\gamma}\right]^{1/(2+\gamma)}\, \left[E|\boldsymbol{\xi}_1|_l^{2+\gamma}\right]^{1/(2+\gamma)}.$$

Condition on $\mathbf{X}_1$ and use Assumption F3 to obtain

$$E|\boldsymbol{\xi}_1|^{2+\gamma} \leq C\, \mathbb{E}\left[|\mathbf{Z}_1|^{2+\gamma}\;\; K_h^{2+\gamma}(\mathbf{X}_1 - \mathbf{x}_0)\right] = O\left(h^{-p(1+\gamma)}\right), \tag{6.11}$$

so that

$$\mathbf{J}_{n,2} = O\left(h^{-d\gamma/(2+\gamma)}\right) \sum_{i \geq d_n} [\alpha(i)]^{\gamma/(\gamma+2)} = o\left(h^{-d\gamma/(\gamma+2)}\; d_n^{-\gamma}\right) = o(1) \tag{6.12}$$

by Assumption F4 and by choosing d_n such that $h^p\; d_n^{2+\gamma} = O(1)$, so that d_n satisfies $d_n\, h^p \to 0$. A combination of (6.8) – (6.12) proves the first result in the lemma. The second result is established similarly. This proves the lemma. □

Proof of Theorem 6.1. Let $\boldsymbol{\beta}^* = \sqrt{n\, h^p}\, \mathbf{H}\, (\boldsymbol{\beta} - \boldsymbol{\beta}_0)$ and let $\gamma_n = (n\, h^p)^{-1/2}$. Then, $\boldsymbol{\beta} = \boldsymbol{\beta}_0 + \gamma_n\, \mathbf{H}^{-1}\, \boldsymbol{\beta}^*$ and for $\widehat{\boldsymbol{\beta}}$ to maximize $U_n(\boldsymbol{\beta})$ is equivalent for $\widehat{\boldsymbol{\beta}}^*$ to maximize

$$U_n^*(\boldsymbol{\beta}^*) = \sum_{i=1}^{n} \left[Q\left(Y_i,\, \zeta_i + \gamma_n\, \boldsymbol{\beta}^{*\top}\, \mathbf{Z}_i\right) - Q(Y_i,\, \zeta_i)\right]\; K((\mathbf{X}_i - \mathbf{x}_0)/h),$$

where $\zeta_i = g(\mathbf{x}_0) + g'(\mathbf{x}_0)^\top (\mathbf{X}_i - \mathbf{x}_0)$. By Assumption F6 and the Taylor expansion,

$$U_n^*(\boldsymbol{\beta}^*) = \mathbf{F}_n\, \boldsymbol{\beta}^* - \frac{1}{2}\, \boldsymbol{\beta}^{*\top}\, \mathbf{D}_n\, \boldsymbol{\beta}^* + \frac{\gamma_n^3}{6} \sum_{i=1}^{n} q''\left(Y_i,\, \widetilde{\zeta}_i\right) \left(\boldsymbol{\beta}^{*\top}\, \mathbf{Z}_i\right)^3\; K((\mathbf{X}_i - \mathbf{x}_0)/h), \tag{6.13}$$

where

$$\mathbf{F}_n = \gamma_n \sum_{i=1}^{n} Q(Y_i,\, \zeta_i)\, \mathbf{Z}_i\, K((\mathbf{X}_i - \mathbf{x}_0)/h) \quad \text{and} \quad \mathbf{D}_n = -\gamma_n^2 \sum_{i=1}^{n} Q'(Y_i,\, \zeta_i)\, \mathbf{Z}_i\, \mathbf{Z}_i^\top\, K((\mathbf{X}_i - \mathbf{x}_0)/h)$$

with $\widetilde{\zeta}_i$ being between ζ_i and $\zeta_i + \gamma_n\, \boldsymbol{\beta}^{*\top}\, \mathbf{Z}_i$. Clearly, $\mathbf{F}_n = \gamma_n^{-1}\, \mathbf{C}_{n,1}(\boldsymbol{\beta}_0)$, $\mathbf{D}_n = -\mathbf{C}_{n,2}(\boldsymbol{\beta}_0)$, and by Lemma 6.1,

$$\mathbf{D}_n = -f(\mathbf{x}_0)\, \mathbb{E}[Q'(Y,\, g(\mathbf{X}))\,|\, \mathbf{X} = \mathbf{x}_0]\, \mathbf{B}_1 + o_p(1) \equiv \mathbf{D} + o_p(1). \tag{6.14}$$

By Assumptions F1 and F6, the expected value of the absolute value of the last term in (6.13) is bounded by

$$O\left[n\, \gamma_n^3\, \mathbb{E}\left\{|q''(Y,\, \widetilde{\zeta}\,)|\, |\mathbf{X}|^3\, K((\mathbf{X} - \mathbf{x}_0)/h)\right\}\right] = O(\gamma_n) = o(1). \tag{6.15}$$

A substitution of (6.14) and (6.15) into (6.13) yields

$$U_n(\boldsymbol{\beta}^*) = \mathbf{F}_n\,\boldsymbol{\beta}^* - \frac{1}{2}\,\boldsymbol{\beta}^{*\top}\,(\mathbf{D} + o_p(1))\,\boldsymbol{\beta}^* + o_p(1).$$

Therefore, an application of the quadratic approximation lemma; see, e.g., Fan and Gijbels (1996) gives

$$\widehat{\boldsymbol{\beta}}^* = \mathbf{D}^{-1}\,\mathbf{F}_n + o_p(1),$$

if $\mathbf{F}_n$ is a sequence of stochastically bounded random vectors. By Assumption F2 and the Taylor's expansion,

$$g(\mathbf{X}_1) = g(\mathbf{x}_0) + g'(\mathbf{x}_0)^\top(\mathbf{X}_1 - \mathbf{x}_0) + \frac{1}{2}\,(\mathbf{X}_1 - \mathbf{x}_0)^\top\,g''(\mathbf{x}_0)\,(\mathbf{X}_1 - \mathbf{x}_0) + o_p(h^2),$$

so that

$$\begin{aligned}
\gamma_n\,\mathbb{E}(\mathbf{F}_n) &= \mathbb{E}\left[Q\left\{Y_1,\,\zeta_1\right\}\,\mathbf{Z}_1\,K_h(\mathbf{X}_1 - \mathbf{x}_0)\right] \\
&= -\frac{h^2}{2}\,f(\mathbf{x}_0)\,\mathbb{E}\left[Q'(Y,\,g(\mathbf{X}))\,|\,\mathbf{X} = \mathbf{x}_0\right]\,\begin{pmatrix} tr\left\{g''(\mathbf{x}_0)\,\boldsymbol{\mu}_2\right\} \\ \mathbf{0} \end{pmatrix} + o(h^2) \\
&= \mathbf{D}\,\begin{pmatrix} \frac{h^2}{2}\,\mathrm{tr}\left\{g''(\mathbf{x}_0)\,\boldsymbol{\mu}_2\right\} \\ \mathbf{0} \end{pmatrix} + o(h^2).
\end{aligned}$$

Then,

$$\sqrt{n\,h^p}\,\left[\widehat{g}(\mathbf{x}_0) - g(\mathbf{x}_0) - \frac{h^2}{2}\,\mathrm{tr}\left\{g''(\mathbf{x}_0)\,\boldsymbol{\mu}_2\right\} + o_p(h^2)\right] = \mathbf{e}^\top\,\mathbf{D}^{-1}\left[\mathbf{F}_n - \mathbb{E}(\mathbf{F}_n)\right] + o_p(1), \quad (6.16)$$

where $\mathbf{e} = \begin{pmatrix} 1 \\ \mathbf{0} \end{pmatrix}$. It follows from Lemma 6.1 that the right hand side of (6.16) is $O_p(1)$. This proves the theorem.

Proof of Theorem 6.2. From (6.16), to accomplish the proof of the asymptotic normality of $\widehat{g}(\mathbf{x}_0)$, it suffices to establish the asymptotic normality of $\mathbf{F}_n - \mathbb{E}(\mathbf{F}_n)$. To this end, the Doob's small-block and large-block technique is used. Namely, partition $\{1, \cdots, n\}$ into $2\,q_n + 1$ subsets with large block of size $r = r_n = \lfloor (n\,h^p)^{1/2} \rfloor$ and small block of size $s = s_n = \lfloor (n\,h^p)^{1/2}/\log n \rfloor$, where

$$q = q_n = \left\lfloor \frac{n}{r_n + s_n} \right\rfloor$$

and $\lfloor x \rfloor$ denotes the integer part of x. Then, it can easily be shown from Assumption F4 that as $n \to \infty$,

$$s_n/r_n \to 0, \qquad r_n/n \to 0, \qquad \text{and} \qquad (n/r_n)\,\alpha(s_n) \to 0. \quad (6.17)$$

Now, the Cramér-Wold device is employed here to derive the asymptotic normality of $\mathbf{F}_n - \mathbb{E}(\mathbf{F}_n)$. For any unit vector $\mathbf{d} \in \mathbb{R}^{p+1}$ and $i = 1, \cdots, n$, define

$$Z_{n,i} = h^{d/2} \left[Q(Y_i, \zeta_i)\, \mathbf{d}^\top \mathbf{Z}_i\, K_h(\mathbf{X}_i - \mathbf{x}_0) - \mathbb{E}\left\{ q(Y_i, \zeta_i)\, \mathbf{d}^\top \mathbf{Z}_i\, K_h(\mathbf{X}_i - \mathbf{x}_0) \right\} \right].$$

Then

$$\mathbf{d}^\top \left[\mathbf{F}_n - \mathbb{E}(\mathbf{F}_n) \right] = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_{n,i},$$

and by Lemma 6.1,

$$\mathrm{Var}(Z_{n,1}) = f(\mathbf{x}_0)\, \mathbb{E}\left[Q^2(Y, g(\mathbf{X})) \,|\, \mathbf{X} = \mathbf{x}_0 \right] \mathbf{d}^\top \mathbf{B}_2\, \mathbf{d}\, (1 + o(1)) \equiv \theta^2(\mathbf{x}_0)(1 + o(1),$$

$$\sum_{i=2}^{n} |\mathrm{Cov}(Z_{n,1},\, Z_{n,i})| = o(1), \qquad \text{and} \qquad \mathrm{Var}(\mathbf{F}_n) = f(\mathbf{x}_0)\, \mathbb{E}\left[Q^2(Y, g(\mathbf{X})) \,|\, \mathbf{X} = \mathbf{x}_0 \right] \mathbf{B}_2\, (1 + o(1)).$$

For $0 \le j \le q - 1$, define

$$V_{j,1} = \sum_{i=j(r+s)+1}^{j(r+s)+r} Z_{n,i}, \qquad V_{j,2} = \sum_{i=j(r+s)+r+1}^{(j+1)(r+s)} Z_{n,i}, \qquad \text{and} \qquad V_q = \sum_{i=q(r+s)+1}^{n} Z_{n,i}.$$

Then,

$$\mathbf{d}^\top \left[\mathbf{F}_n - \mathbb{E}(\mathbf{F}_n) \right] = \frac{1}{\sqrt{n}} \left\{ \sum_{j=0}^{q-1} V_{j,1} + \sum_{j=0}^{q-1} V_{j,2} + V_q \right\} \equiv \frac{1}{\sqrt{n}} \left\{ F_{n,1} + F_{n,2} + F_{n,3} \right\}.$$

It is shown that as $n \to \infty$,

$$\frac{1}{n}\, \mathbb{E}\left[F_{n,2} \right]^2 \to 0, \qquad \frac{1}{n}\, \mathbb{E}\left[F_{n,3} \right]^2 \to 0, \tag{6.18}$$

$$\left| \mathbb{E}\left[\exp(i\, t\, F_{n,1}) \right] - \prod_{j=0}^{q-1} \mathbb{E}\left[\exp(i\, t\, V_{j,1}) \right] \right| \to 0, \tag{6.19}$$

$$\frac{1}{n} \sum_{j=0}^{q-1} \mathbb{E}\left(V_{j,1}^2 \right) \to \theta^2(\mathbf{x}_0), \tag{6.20}$$

and

$$\frac{1}{n} \sum_{j=0}^{q-1} \mathbb{E}\left[V_{j,1}^2 I\left\{ |V_{j,1}| \ge \varepsilon\, \theta(\mathbf{x}_0)\, \sqrt{n} \right\} \right] \to 0 \tag{6.21}$$

for every $\varepsilon > 0$. (6.18) implies that $F_{n,2}$ and $F_{n,3}$ are asymptotically negligible in probability, (6.19) shows that the summands $V_{j,1}$ in $F_{n,1}$ are asymptotically independent and (6.20) and (6.21) are the standard Lindeberg-Feller conditions for asymptotic normality of $F_{n,1}$ for the independent setup.

First, we establish (6.18). Observe that

$$\mathbb{E}\left[F_{n,2}\right]^2 = \sum_{j=0}^{q-1} \mathrm{Var}(V_{j,2}) + 2 \sum_{0 \le i < j \le q-1} \mathrm{Cov}(V_{i,2},\, V_{j,2}) \equiv I_1 + I_2.$$

From stationarity, Lemma 6.1, and (6.17), it follows that

$$I_1 = q_n \,\mathrm{Var}(V_{1,2}) = q_n \,\mathrm{Var}\left(\sum_{i=1}^{s_n} Z_{n,i}\right) = q_n \, s_n \left[\theta^2(\mathbf{x}_0) + o(1)\right] = o(n). \tag{6.22}$$

Next, we consider the second term I_2. Let $r_j^* = j(r_n + s_n)$, then $r_j^* - r_i^* \ge r_n$ for all $j > i$. Thus,

$$|I_2| \le 2 \sum_{0 \le i < j \le q-1} \sum_{j_1=1}^{s_n} \sum_{j_2=1}^{s_n} \left|\mathrm{Cov}(Z_{n,r_i^*+r_n+j_1},\, Z_{n,r_j^*+r_n+j_2})\right| \le 2 \sum_{j_1=1}^{n-r_n} \sum_{j_2=j_1+r_n}^{n} \left|\mathrm{Cov}(Z_{n,j_1},\, Z_{n,j_2})\right|.$$

By stationarity and Lemma 6.1, one obtains

$$|I_2| \le 2n \sum_{j=r_n+1}^{n} \left|\mathrm{Cov}(Z_{n,1},\, Z_{n,j})\right| = o(n),$$

which, in conjunction with (6.22), implies that the first result in (6.18) holds. From stationarity, Lemma 6.1, and (6.17), it follows that

$$\mathrm{Var}\left[F_{n,3}\right] = \mathrm{Var}\left(\sum_{i=1}^{n-q_n(r_n+s_n)} Z_{n,i}\right) = O(n - q_n(r_n + s_n)) = o(n),$$

which establishes (6.18). To establish (6.19), use Lemma 1.1 of Volkonskii and Rozanov (1959) to obtain

$$\left|\mathbb{E}\left[\exp(i\,t\,F_{n,1})\right] - \prod_{j=0}^{q_n-1} \mathbb{E}\left[\exp(i\,t\,V_{j,1})\right]\right| \le 16\,(n/r_n)\,\alpha(s_n).$$

The right-hand side tends to 0 by (6.17) so that (6.19) holds. As for (6.20), by stationarity, (6.17), and Lemma 6.1, it is easily seen that

$$\frac{1}{n} \sum_{j=0}^{q_n-1} \mathbb{E}\left(V_{j,1}^2\right) = \frac{q_n}{n} \mathbb{E}\left(V_{1,1}^2\right) = \frac{q_n\, r_n}{n} \cdot \frac{1}{r_n} \mathrm{Var}\left(\sum_{i=1}^{r_n} Z_{n,i}\right) \to \theta^2(\mathbf{x}_0).$$

It remains to prove (6.21). For this purpose, we use Theorem 4.1 of Shao and Yu (1996) and Assumptions F3 and F4 to obtain

$$\mathbb{E}\left[V_{1,1}^2\, I\left\{|V_{1,1}| \ge \varepsilon\, \theta(\mathbf{x}_0)\, \sqrt{n}\right\}\right] = O\left(n^{-\gamma/2}\, r_n^{(2+\gamma)/2}\, \left\{\mathbb{E}\left(|Z_{n,1}|^{2(1+\gamma)}\right)\right\}^{(2+\gamma)/2(1+\gamma)}\right).$$

As in (6.11),

$$\mathbb{E}\left(|Z_{n,1}|^{2(1+\gamma)}\right) = O\left(h^{-p\,\gamma}\right),$$

which, in conjunction with the above equation, implies that

$$\mathbb{E}\left[V_{1,1}^2\, I\left\{|V_{1,1}| \geq \varepsilon\, \theta(\mathbf{x}_0)\, \sqrt{n}\right\}\right] = O\left(n^{-\gamma/2}\, r_n^{(2+\gamma)/2}\, h^{-p(2+\gamma)\gamma/2(1+\gamma)}\right).$$

Thus, by the definition of q_n and r_n,

$$\frac{1}{n}\sum_{j=0}^{q-1} \mathbb{E}\left[V_{j,1}^2 I\left\{|V_{j,1}| \geq \varepsilon\theta(\mathbf{x}_0)\sqrt{n}\right\}\right] = O\left(r_n^{\gamma/2} n^{-\gamma/2} h^{-p(2+\gamma)\gamma/2(1+\gamma)}\right) = O\left(n^{-\gamma/4} h^{-p[1+2/(1+\gamma)]\gamma/4}\right).$$

The right-hand side tends to 0 by Assumption F7 so that (6.21) holds. The theorem is proved. □

Chapter 7

Nonparametric and Semiparametric Models for Casual Inferences

7.1 Introduction

Causal inference has a wide range of applications in our daily lives. For example, doctors want to know whether a new drug is effective, educators want to know whether students in smaller class sizes can perform better in examinations, marketing teams would like to know whether specific advertising can promote sales, etc. economists and statisticians use causal inference to conduct policy evaluation, i.e., analyzing the impact of implementing a particular policy on some outcome variables of their interest. For example, people might have an interest to see if the escalating trade war between China and the US, initiated in 2018 by the first Donald Trump Administration, has significantly impacted the trade pattern of these two nations. Scholars depict this impact as the treatment effect, and the most common object of study is the average treatment effect (ATE) or quantile treat effect (QTE). For a simple review on how to estimate ATE or QTE in a nonparametric or semiparametric way, the reader is referred to the survey paper by Imbens (2004).

Denote Y as the outcome variable of our interest, $\boldsymbol{X} \in \mathbb{R}^p$ as a vector of covariates, and d as the binary treatment variable, where $d = 1$ indicates being treated (the treated group) and $d = 0$ indicates being untreated (the control group). Under the Rubin causal model, denote $Y(1)$ and $Y(0)$ as the potential outcomes corresponding to the treated and untreated status, respectively. Then, $\{(Y_i, d_i, \mathbf{X}_i)\}_{i=1}^n$ is the observed sample, where $Y_i = d_i Y_i(1) + (1-d_i)Y_i(0)$. The average treatment effect and quantile treatment effect are defined as

$$\Delta = \mathbb{E}[Y_i(1) - Y_i(0)] \quad \text{and} \quad \Delta_\tau = q_\tau(1) - q_\tau(0), \tag{7.1}$$

respectively, where $0 < \tau < 1$, $q_\tau(1)$ is the τth quantile of $Y_i(1)$, and $q_\tau(0)$ is the τth quantile of $Y_i(0)$.

Before discussing how to estimate Δ and Δ_τ, first, define the conditional probability of receiving the treatment as $\pi(\boldsymbol{x}) = \mathbb{P}(d = 1|\boldsymbol{X} = \boldsymbol{x})$, which is called as the propensity score function in the causal inference literature. Then, the following assumptions are commonly imposed for making causal inferences, since $Y(0)$ is not observable.

Assumption 7.1: *(Stable Unit Treatment Value Assumption, SUTVA)*
The potential outcomes for any unit do not vary with the treatment assigned to other units. And for each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes.

Assumption 7.2: *(Strongly Ignorable Treatment Assignment)*
(i) (Unconfoundedness) $(Y(1), Y(0)) \perp\!\!\!\perp d \,|\, \boldsymbol{X}$*;*
(ii) (Overlap) There exists $\varepsilon > 0$*, such that for all* $\boldsymbol{x} \in \mathcal{X}$*,* $\varepsilon < \pi(\boldsymbol{x}) < 1 - \varepsilon$*, where* $\mathcal{X} \subset \mathbb{R}^p$ *is the support of* $\boldsymbol{X}$*.*

Remark 7.1: *Assumption 7.2(i) is commonly called to be conditional independent or conditional unconfoundedness assumption in the causal inference literature, which is not easy to verify or test in practice. For details, see, for example, the papers by Fang et al. (2020) for the iid setting and Cai et al. (2024) for the time series context on how to test this assumption in practice. For some applications (for example, Heckman et al. (1998)), one might only need the conditional mean independence such as* $\mathbb{E}[Y(d) \,|\, d, \mathbf{X}] = \mathbb{E}[Y(d) \,|\, \mathbf{X}]$ *for* $d = 0$ *and* 1*, which is clearly weaker than Assumption 7.2(i). Also, by Lemma 2.1 in Imbens (2004), Assumption 7.2(i) implies that* $(Y(1), Y(0)) \perp\!\!\!\perp d \,|\, \pi(\boldsymbol{X})$*.*

7.2 Estimation Methods of ATE

First, let us consider the estimate of Δ. To do so, under the above assumptions, it is easy show that the ATE can be written as

$$\begin{aligned}\mathbb{E}[Y(1) - Y(0)] &= \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] = \mathbb{E}[Y(1)|\boldsymbol{X}] - \mathbb{E}[Y(0)|\boldsymbol{X}] \\ &= \mathbb{E}[Y(1)|d = 1, \boldsymbol{X}] - \mathbb{E}[Y(0)|d = 0, \boldsymbol{X}] \\ &= \mathbb{E}[Y|d = 1, \boldsymbol{X}] - \mathbb{E}[Y|d = 0, \boldsymbol{X}] = \mu_1(\mathbf{X}) - \mu_0(\mathbf{X}), \qquad (7.2)\end{aligned}$$

where $\mu_1(\mathbf{x}) = \mathbb{E}(Y \,|\, d = 1, \mathbf{X} = \mathbf{x}) = \mathbb{E}[Y(1)\ \mathbf{X} = \mathbf{x}]$ and $\mu_0(\mathbf{x}) = \mathbb{E}(Y \,|\, d = 0, \mathbf{X} = \mathbf{x}) = \mathbb{E}(Y(0) \,|\, \mathbf{X} = \mathbf{x}]$. Suppose that we can observe a random sample $\{Y_i,\, \boldsymbol{X}_i,\, d_i\}_{i=1}^n$ from the population $(Y,\, \boldsymbol{X},\, d)$. If this sample satisfies Assumptions 7.1 and 7.2, we can directly use this sample to identify the ATE by (7.2). However, it is difficult to verify whether the unconfoundedness assumption holds. The random clinical trial (RCT) data usually ensure the unconfoundedness assumption. Nevertheless, the RCT is hard to conduct and even unethical sometimes. The observational data are available in most cases, but they are generally confounding. We define the variables that affect both the treatment assignment and the outcome as confounders. Due to the existence of such confounders, the covariate distribution of the treated group differs from that of the control group, leading to the selection bias problem. To de-confound the observational data, scholars use the sample re-weighting technique to create a pseudo population on which the covariate distributions of the treated and control groups are identical.

There are several methods available in the causal inference literature to estimate Δ such as inverse probability weighting (IPW), covariate balancing propensity score (CBPS) procedures, and synthetic control method (SCM) and its extensions, addressed in the following subsections. Suppose we can observe the iid sample $\{(Y_i,\, d_i,\, \boldsymbol{X}_i)\}_{i=1}^n$ from the population $(Y,\, T,\, \boldsymbol{X})$, where Y_i is the outcome variable of our interest, $d_i = \{0, 1\}$ is the binary treatment variable, and $\boldsymbol{X}_i \in \mathcal{X}$ is the $p \times 1$ vector of covariates. The purpose of this section to introduce some methods to estimate the average treatment effect Δ in (7.1) with the observed data.

7.2.1 Propensity Score Based Approaches

It is easy to show, as in Rosenbaum (1987), that under Assumption 7.2, the ATE can be identified by

$$\Delta = \mathbb{E}\left[\left(\frac{d}{\pi(\boldsymbol{X})} - \frac{1-d}{1-\pi(\boldsymbol{X})}\right) Y\right],$$

which is (6) in Imbens (2004), so that this formulation motivates the popular inverse probability weighting estimator for ATE, defined as

$$\widehat{\Delta}_{\text{ipw}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{d_i}{\widehat{\pi}(\boldsymbol{X}_i)} - \frac{1-d_i}{1-\widehat{\pi}(\boldsymbol{X}_i)}\right) Y_i = \frac{1}{n}\sum_{i=1}^{n} w_i\, Y_i, \tag{7.3}$$

where $w_i = d_i/\widehat{\pi}(\boldsymbol{X}_i) + (1-d_i)/[1-\widehat{\pi}(\boldsymbol{X}_i)]$ and $\widehat{\pi}(\boldsymbol{X}_i)$ is a consistent estimator of $\pi(\boldsymbol{X}_i)$. It is easy to see that the IPW estimator given by (7.3) can be regarded as re-weighting

each unit by $\{w_i\}_{i=1}^n$. Also, if the propensity scores were known, then this estimator will be unbiased for the ATE; see, for instance, Tsiatis (2006). Furthermore, when the propensity scores are estimated consistently, then this estimator is consistent for the ATE. However, the simple IPW estimator is also widely believed to have poor small sample properties when the propensity score gets close to zero or one for some observations. In practice, to achieve a smaller variance, $\{w_i\}_{i=1}^n$ are normalized such that the sum of the weights within one group is 1; see, e.g., Imbens (2004) for details, as

$$\widehat{\Delta}_{\text{ipw},1} = \sum_{i=1}^{n} w_{1,i}\, Y_i - \sum_{i=1}^{n} w_{0,i}\, Y_i, \tag{7.4}$$

where

$$w_{1,i} = \frac{d_i}{\widehat{\pi}(\boldsymbol{X}_i)} \left[\sum_{i=1}^{n} \frac{d_i}{\widehat{\pi}(\boldsymbol{X}_i)}\right]^{-1} \quad \text{and} \quad w_{0,i} = \frac{1-d_i}{1-\widehat{\pi}(\boldsymbol{X}_i)} \left[\sum_{i=1}^{n} \frac{1-d_i}{1-\widehat{\pi}(\boldsymbol{X}_i)}\right]^{-1}.$$

Also, the large sample theory of the estimators defined in (7.3) and (7.4) has been studied thoroughly in the literature such as Hahn (1998), Hirano et al. (2003), Li et al. (2018) and references therein. Although adjusting for the propensity score is enough to remove bias due to all observed covariates, the IPW estimator relies heavily on the quality of the propensity score estimation. Even slight misspecification of the propensity score model can deteriorate the estimator drastically. To overcome the hazard of the possible misspecification, Robins et al. (1994) proposed the augmented IPW (AIPW) estimator for ATE:

$$\widehat{\Delta}_{\text{aipw}} = \frac{1}{n} \sum_{i=1}^{n} \left\{ \widehat{\mu}_1(\boldsymbol{X}_i) - \widehat{\mu}_0(\boldsymbol{X}_i) + \frac{d_i}{\widehat{\pi}(\boldsymbol{X}_i)} \left[Y_i - \widehat{\mu}_1(\boldsymbol{X}_i)\right] - \frac{1-d_i}{1-\widehat{\pi}(\boldsymbol{X}_i)} \left[Y_i - \widehat{\mu}_0(\boldsymbol{X}_i)\right] \right\}, \tag{7.5}$$

where $\widehat{\mu}_1(\boldsymbol{x})$ and $\widehat{\mu}_0(\boldsymbol{x})$ are the regression estimators for $\mu_1(\mathbf{x})$ and $\mu_0(\mathbf{x})$, respectively. The AIPW estimator (7.4) is doubly robust in the sense that it is consistent if either $\widehat{\pi}(\boldsymbol{x})$ is consistent or $\widehat{\mu}_1(\boldsymbol{x})$ and $\widehat{\mu}_0(\boldsymbol{x})$ are consistent. Indeed, Glynn and Quinn (2010) conducted a simulation study to show that the AIPW actually outperforms better in finite sample than some competing estimators.

To obtain a consistent estimate of $\pi(\boldsymbol{x})$ in (7.4) and (7.5), one can use logistic type approach. To avoid a possible misspecification of logistic type setup, Hirano et al. (2003) proposed the Series logist estimator (SLE) to estimate $\pi(\boldsymbol{x})$. Also, one can obtain the estimated outcome regression function in a sample of size n is given by $\widehat{\mu}_j(\boldsymbol{x})$ for $j = 0$ and 1, in an appropriate way, say a parametric method or a kernel approach. See Imbens (2004) and Glynn and Quinn (2010) for more discussions.

7.2.2 Covariate Balancing Methods

In addition to constructing sample weights by propensity score, other methods directly construct weights to achieve the goal of covariate balancing. For example, Hainmueller (2012) proposed the entropy balancing method, which weights each sample to achieve balance on some specific moments of the covariates. The entropy balancing method also minimizes the Kullback–Leibler divergence of the estimated weights from a set of pre-specified base weights. Furthermore, Graham et al. (2012) proposed the inverse propensity tilting (IPT) estimator, which modifies the IPW estimator based on the maximum likelihood estimate by an alternative method-of-moments estimate. Also, Imai and Ratkovic (2014) proposed the covariate balancing propensity score (CBPS) approach, which exploits the dual characteristics of the propensity score as a covariate balancing score and the conditional probability of treatment assignment. Indeed, the CBPS method estimates the parameters of the propensity score model, $\boldsymbol{\beta}$, by solving the following m-dimensional estimating equation

$$G(\beta) = \frac{1}{n}\sum_{i=1}^{n} g_{\boldsymbol{\beta}}(d_i, \mathbf{X}_i) = 0, \quad \text{where} \quad g_{\beta}(d_i, \mathbf{X}_i) = \left[\frac{d_i}{\pi_{\boldsymbol{\beta}}(\mathbf{X}_i)} - \frac{(1-d_i)}{(1-\pi_{\boldsymbol{\beta}}(\mathbf{X}_i))}\right] f(\mathbf{X}_i)$$

for some function $f(\cdot)$ from $\mathbb{R}^p \to \mathbb{R}^m$, when the number of equations m is equal to the number of parameters $d_{\boldsymbol{\beta}}$. This guarantees that the treatment and control groups have an identical sample mean $f(\mathbf{X}_i)$ of after weighting by the estimated propensity score. When $m > d_{\boldsymbol{\beta}}$, the classical GMM of Hansen (1982) can be applied to estimate β by minimizing the following J-statistic

$$J_n(\boldsymbol{\beta}) = n\, G(\boldsymbol{\beta})^{\top} \mathbf{W}_n G(\boldsymbol{\beta}),$$

where $\mathbf{W}_n$ is an $m \times m$ weighting positive matrix, which is assumed to be independent of $\boldsymbol{\beta}$.

Imai and Ratkovic (2014) pointed out that the common practice of fitting a logistic model is equivalent to balancing the score function with $f(\mathbf{X}_i) = \partial \pi_{\boldsymbol{\beta}}(\mathbf{X}_i)/\partial \boldsymbol{\beta}$ and they found that choosing $f(\mathbf{X}_i) = \mathbf{X}_i$, which balances the first moment between the treatment and control groups, significantly reduces the bias of the estimated ATE. Some researchers also include higher moments and/or interactions, e.g., $f(\mathbf{X}_i) = (\mathbf{X}_i, \mathbf{X}_i^2)$, in their applications. In such a way, the CBPS method can improve the robustness of both the IPW estimator and the matching estimators based on the propensity score, while Fong et al. (2018) extended the original CBPS to continuous treatment. Moreover, Zhao (2019) proposed to use tailored loss

functions-covariate balancing scoring rules to estimate the propensity score. Now, one open question remains in this literature: How shall we choose the covariate balancing function $f(\mathbf{X}_i)$? In particular, if the propensity score model is misspecified, this problem becomes even more important. Recently, to answer this question, Fan et al. (2023) derived the optimal choice of the covariate balancing function for the estimation of CBPS and studied the theoretical properties of the proposed method, termed as oCBPS. They also showed that the IPTW estimator based on the oCBPS method retains the double robustness property.

Moreover, Zubizarreta (2015) used the minimum variance weights that adjust the empirical distribution of the covariates up to a pre-specified level, while Chan et al. (2016) defined a general class of calibration weights to achieve an exact three-way balance among the treated, the controls, and the combined group, and this class accommodates several existing methods, such as exponential tilting and empirical likelihood. Furthermore, Wong and Chan (2018) proposed a weighting method that balances covariate functionals nonparametrically via reproducing kernel Hilbert space. Finally, Hazlett (2020) presented the kernel balancing method, which chooses sample weights achieving mean balance on some basis functions related to the kernel of the covariates. Actually, Kernel balancing not only derives an unbiased estimator for the treatment effect but also ensures that the multivariate density of the covariates, estimated by the same kernel, is equal for the treated and the control group. To implement the aforementioned CBPS methods, one can use the **R** package "CBPS".

To briefly summarize the existing literature, one can conclude that in econometrics, existing methods can only attain covariate balance on some moments or some pre-specified functions, which is far from the ideal goal of attaining balance on the distribution of covariates. In machine learning, scholars are more interested in extracting effective representations for predicting the counterfactual outcomes first and then achieving balance on these representations.

7.2.3 CB Approach Based on Machine Learning Method

A. Model Framework

Based on the existing research, we want to construct a weighting estimator as (7.6) using a machine learning method. There is a consensus that the re-weighted sample should satisfy the unconfoundedness assumption, while (7.7) indicates that the key is to balance the covariate distribution of the full sample and the weighted treated (control) group. Different

from (7.3), we consider a new approach to define the estimator of ATE. For the given weight vector $\boldsymbol{w} = (w_1, \cdots, w_n)'$, where $\boldsymbol{w} \in [0,1]^n$, the weighting estimator for ATE is

$$\widehat{\Delta}_{\boldsymbol{w}} = \frac{1}{n_1} \sum_{i:d_i=1} w_i Y_i - \frac{1}{n_0} \sum_{i:d_i=0} w_i Y_i. \tag{7.6}$$

Denote $F(\boldsymbol{x})$ as the distribution function of covariates in the population, and $F_n(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^n I(\boldsymbol{X}_i \leq \boldsymbol{x})$ as the empirical distribution function of covariates in the full sample. After sample re-weighting, denote $F_{n,t,\boldsymbol{w}}(\boldsymbol{x}) = \frac{1}{n_1}\sum_{i:d_i=1} w_i I(\boldsymbol{X}_i \leq \boldsymbol{x})$ and $F_{n,c,\boldsymbol{w}}(\boldsymbol{x}) = \frac{1}{n_0}\sum_{i:d_i=0} w_i I(\boldsymbol{X}_i \leq \boldsymbol{x})$ as the empirical distribution functions of covariates in the weighted treated group and the weighted control group, respectively. The error of the estimator (7.6) can be decomposed as, by a simple algebra,

$$\begin{aligned}
\widehat{\Delta}_{\boldsymbol{w}} - \Delta = & \left(\frac{1}{n_1} \sum_{i:d_i=1} w_i Y_i - \frac{1}{n_0} \sum_{i:d_i=0} w_i Y_i \right) - \int [\mu_1(\boldsymbol{x}) - \mu_0(\boldsymbol{x})] \mathrm{d}F(\boldsymbol{x}) \\
& - \int \mu_1(\boldsymbol{x}) \mathrm{d}F_{n,t,w}(\boldsymbol{x}) + \int \mu_0(\boldsymbol{x}) \mathrm{d}F_{n,c,w}(\boldsymbol{x}) \\
= & \left\{ \frac{1}{n_1} \sum_{i:d_i=1} w_i \left[Y_i - \mu_1(\boldsymbol{X}_i) \right] - \frac{1}{n_0} \sum_{i:d_i=0} w_i \left[Y_i - \mu_0(\boldsymbol{X}_i) \right] \right\} \\
& - \int [\mu_1(\boldsymbol{x}) - \mu_0(\boldsymbol{x})] \mathrm{d} \left(F - F_n \right) (\boldsymbol{x}) \\
& \left\{ - \int \mu_1(\boldsymbol{x}) \mathrm{d} \left(F_n - F_{n,t,w} \right) (\boldsymbol{x}) + \int \mu_0(\boldsymbol{x}) \mathrm{d} \left(F_n - F_{n,c,w} \right) (\boldsymbol{x}) \right\},
\end{aligned} \tag{7.7}$$

from which, we can see clearly that the expectation of the first term is zero, and the second term tends to zero if the sample size n is big enough to ensure that the sample is representative of the population. Now, the third term remains the intractable one, which depends on the discrepancy between two distributions F_n and $F_{n,t,w}(F_{n,c,w})$. Therefore, the ideal goal of covariate balance is to balance the covariate distribution of the full sample and the weighted treated (control) group. In other words, it attempts to make the third term in (7.7) as small as possible. To achieve this goal, we use the idea in Cai et al. (2025b) to generalize the generative adversarial network (GAN) to a framework of distribution augmentation, termed as adversarial covariate balancing network (ACBN), and show how this new framework can facilitate achieving covariate balancing in distribution. Note that this new covariate balancing method is an adaption of vanilla generative adversarial network (GAN), a popular algorithm for data augmentation in deep learning, towards the balance of the distribution of the covariates, so that it can be regarded as extension of vanilla GAN from data augmentation to distribution augmentation.

B. An Introduction of Generative Adversarial Network

Next, we briefly review some literature on GAN. Suppose we want to learn the distribution over the observed real data $\boldsymbol{x}$. Different from traditional methods of trying to derive an explicit formula for the distribution function, the GAN learns the distribution by generating new data sharing the same distribution with $\boldsymbol{x}$. Motivated by a zero-sum game, GAN designs a generator to generate data and a discriminator to decide whether a sample comes from real data or the generator. By playing an adversarial game seeking the Nash equilibrium, the generated data distribution is indistinguishable from real data distribution.

Now, we formulate GAN mathematically. First, we define a noise variable $\boldsymbol{z}$, which is of the same dimension as $\boldsymbol{x}$. Define the generator as a function $G(\boldsymbol{z};\theta_g)$, which maps $\boldsymbol{z}$ to the desired data space. Usually, $\boldsymbol{z}$ comes from a simple distribution, such as normal distribution or uniform distribution, whereas the distribution of real data $\boldsymbol{x}$ is quite complex. A multi-layer perceptron is hence a common choice for G. Meanwhile, we define a function $D(\boldsymbol{x}';\theta_d)$ as the discriminator, which outputs the probability of $\boldsymbol{x}'$ coming from real data rather than the generator. In other words, D solves a binary classification problem, labeling the sample from real data as 1 and the sample from the generator as 0. D is also represented by a multi-layer perceptron. In the game, we train the discriminator to maximize the probability of correctly labeling, while the training goal of the generator is to minimize the probability of correctly labeling. The objective of this game can be summarized as the following function:

$$\min_G \max_D V(D,G) = \min_G \max_D \left\{\mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1-D(G(\boldsymbol{z})))]\right\}, \quad (7.8)$$

where $p_{\text{data}}(\cdot)$ is the distribution of real data and $p_{\boldsymbol{z}}(\cdot)$ is the distribution of the noise variable. Indeed, (7.8) characterizes the vanilla GAN mathematically. Given $\boldsymbol{z}\sim p_{\boldsymbol{z}}$, we denote the distribution of the generated sample $G(\boldsymbol{z})$ as p_g. Then, by (7.8), for a fixed generator G, the optimal discriminator D_G^* satisfies

$$D_G^*(\boldsymbol{x}) = p_{\text{data}}(\boldsymbol{x}) / \left[p_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})\right].$$

Under this optimal discriminator D_G^*, we can rewrite (7.8) as

$$\begin{aligned}\min_G C(G) &= \min_G \left\{\mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}\left[\log D_G^*(\boldsymbol{x})\right] + \mathbb{E}_{\boldsymbol{x}\sim p_g}\left[\log(1-D_G^*(\boldsymbol{x}))\right]\right\} \\ &= \min_G \left\{\mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}\left[\log\left(\frac{p_{\text{data}}(\boldsymbol{x})}{p_{\text{data}}(\boldsymbol{x})+p_g(\boldsymbol{x})}\right)\right] + \mathbb{E}_{\boldsymbol{x}\sim p_g}\left[\log\left(\frac{p_g(\boldsymbol{x})}{p_{\text{data}}(\boldsymbol{x})+p_g(\boldsymbol{x})}\right)\right]\right\}\end{aligned}$$

$$
\begin{aligned}
&= \min_G \left\{ \log(4) + \mathrm{KL}\left(p_{\text{data}} \| \frac{p_{\text{data}} + p_g}{2} \right) + \mathrm{KL}\left(p_g \| \frac{p_{\text{data}} + p_g}{2} \right) \right\} \\
&= \min_G \{ -\log(4) + 2\,\mathrm{JSD}(p_{\text{data}} \| p_g) \}, \qquad (7.9)
\end{aligned}
$$

where $\mathrm{KL}(P\|\frac{P+Q}{2})$ is the Kullback-Leibler divergence between two probability distributions P and Q and $\mathrm{JSD}(P\|Q)$ is the Jensen-Shannon divergence between P and Q. From (7.9), we can see that the minimax problem of vanilla GAN is closely connected with the statistic distance between distributions of real data and the generated data. Since $\mathrm{JSD}(P\|Q)$ is nonnegative and equals to zero if and only if $P = Q$, then the global minimum of $C(G)$ is achieved if and only if $p_g = p_{\text{data}}$. This finding indicates that we are capable of replicating the distribution of real data p_{data} under (7.8). Goodfellow et al. (2020) showed that if both G and D have enough capacity, and or any given generator G, the discriminator D is allowed to achieve its optimum D^*_G, then p_g converges to p_{data}. In practice, we train the generator and the discriminator via an iterative optimization algorithm.

As a powerful generative model, GAN has been widely used in various fields, such as image processing, texture processing, and natural language processing. For example, Radford et al. (2015) introduced convolutional networks into the framework of GAN and propose deep convolutional generative adversarial networks (DCGANs) and showed that DCGAN is capable of generating human faces of higher quality. Also, Ledig et al. (2017) proposed a generative adversarial network for image super-resolution (SRGAN), which can infer photo-realistic natural images for quadruple upscaling factors. Furthermore, Wu et al. (2016) proposed a framework for generating 3D objects by combining GAN and volumetric convolutional networks. Moreover, Zhang et al. (2017) discussed the application of GAN in text generation for the first time, while Yu et al. (2017) used SeqGAN that bypasses the differentiation difficulty for discrete data by employing policy gradient in reinforcement learning, achieving great success in text generation, music generation, etc. Combining with long short-term memory (LSTM), GAN is also competent in generating time series data and especially for financial data; see, for instance, the paper by Cai et al. (2025b) and references therein. Finally, note that here is also growing literature applying GAN to causal inference. The reader is referred to the paper by Cai et al. (2025b) for more discussions.

However, it is found in the literature that the vanilla GAN might not be stable in practice due to some inherent pitfalls of the Kullback-Leibler divergence. In recent years, several variants of the vanilla GAN have been proposed by modifying the Kullback-Leibler divergence

into other statistical distance measures, such as the f-divergences as in Nowozin et al. (2016), the Earth Mover (EM) distance as in Arjovsky et al. (2017), and the Pearson χ^2 divergence as in Mao et al. (2017). Furthermore, some efforts are also devoted to improving the structure of the vanilla GAN. For example, Mirza and Osindero (2014) extended the vanilla GAN to a conditional model and proposed conditional GAN, while Chen et al. (2016) decomposed the input noise vector $\boldsymbol{z}$ into two parts and proposed InfoGAN. Finally, Karras et al. (2018) presented a new framework based on progressive neural networks and Zhang et al. (2019) proposed the self-attention generative adversarial network (SaGAN), which allows attention-driven and long-range dependency modeling for data generation tasks. 8

In conclusion, GAN has received a wide attention since its birth. Vast literature has discussed its variants, applications and theory. Of course, it is still of great potential.

C. Adversarial Covariate Balancing Network

Denote the covariate distributions of the treated and the control group as $p_T(\boldsymbol{x})$ and $p_C(\boldsymbol{x})$ respectively. Also, denote the covariate distribution of the full sample as $p(\boldsymbol{x})$. Now, we re-weight the treated and the control group with weights $w_T(\boldsymbol{x})$ and $w_C(\boldsymbol{x})$, where $0 \leq w_T(\boldsymbol{x}), w_C(\boldsymbol{x}) \leq 1$. Then, after sample re-weighting, the covariate distributions of the weighted treated and the weighted control group can be written as

$$p_{w_T}(\boldsymbol{x}) = p_T(\boldsymbol{x}) w_T(\boldsymbol{x}) \left[\int p_T(\boldsymbol{x}) w_T(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right]^{-1}, \tag{7.10}$$

and

$$p_{w_C}(\boldsymbol{x}) = p_C(\boldsymbol{x}) w_C(\boldsymbol{x}) \left[\int p_C(\boldsymbol{x}) w_C(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right]^{-1}. \tag{7.11}$$

Now, we need to ensure that $p_{w_T}(\boldsymbol{x})$ ($p_{w_C}(\boldsymbol{x})$) stays as close to $p(\boldsymbol{x})$ as possible.

From the GAN literature, we know that GAN can generate new samples of the same distribution as the observed data. Here, we propose an extension of GAN from data augmentation to distribution augmentation, accommodating the goal of covariate balancing well. In the vanilla GAN, we input a noise variable $\boldsymbol{z}$ into the generator and the generator maps $\boldsymbol{z}$ to the desired data space. Different structures of mapping create different distributions. In our setting, the generator is in charge of generating different weighting functions $w_T(\boldsymbol{x})$ and $w_C(\boldsymbol{x})$. These weighting functions further create different distributions via (7.10) and (7.11). Combining (7.8), (7.10) and (7.11), we present our objective function as follows.

$$\min_{w_T} \max_{D_T} V_T(w_T, D_T) = \min_{w_T} \max_{D_T} \left\{ \mathbb{E}_{\boldsymbol{x} \sim p}[\log D_T(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x} \sim p_{w_T}}[\log(1 - D_T(\boldsymbol{x}))] \right\}, \tag{7.12}$$

and

$$\min_{w_C}\max_{D_C} V_C(w_C, D_C) = \min_{w_C}\max_{D_C} \left\{\mathbb{E}_{\boldsymbol{x}\sim p}[\log D_C(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x}\sim p_{w_C}}[\log(1 - D_C(\boldsymbol{x}))]\right\}, \quad (7.13)$$

where $D_T(\boldsymbol{x})$ and $D_C(\boldsymbol{x})$ are two different discriminators outputting the probability of $\boldsymbol{x}$ belongs to $p(\boldsymbol{x})$ rather than $p_{w_T}(\boldsymbol{x})$ or $p_{w_C}(\boldsymbol{x})$. The weighting functions and the discriminators are all represented by multi-layer perceptrons, and all can be regarded as complex functions mapping $\mathcal{X}$ to $[0, 1]$.

One can see from (7.12) and (7.13) that now the discriminators label sample from p as 1 and label sample from p_{w_T} (p_{w_C}) as 0. The weighting functions $w_T(\boldsymbol{x})$ and $w_C(\boldsymbol{x})$ serve as the generators, which indirectly generate different distributions through (7.10) and (7.11). These two minimax games are essentially not different than (7.8). Thus, the connection between the minimax game and the Jensen-Shannon divergence revealed by (7.9) still can be applied to our setting. In other words, the equilibriums in (7.12) and (7.13) imply the equivalences of $p_{w_T}(\boldsymbol{x})$ and $p_{w_C}(\boldsymbol{x})$ with $p(\boldsymbol{x})$, which is basically consistent with the ideal goal of covariate balancing.

Now, for the given data $\{(Y_i,\, d_i,\, \boldsymbol{X}_i)\}_{i=1}^n$, we can write the sample analogues of (7.12) and (7.13) as

$$\min_{\boldsymbol{w}_T}\max_{D_T} \widehat{V}_T(w_T, D_T) = \min_{\boldsymbol{w}_T}\max_{D_T} \left\{\frac{1}{n}\sum_{i=1}^n \log(D_T(\boldsymbol{X}_i)) + \sum_{i:d_i=1}^n w_{T,1}(\boldsymbol{X}_i)\log(1 - D_T(\boldsymbol{X}_i))\right\}, \quad (7.14)$$

where $w_{T,1}(\boldsymbol{X}_i) = w_T(\boldsymbol{X}_i)\left[\sum_{i:d_i=1} w_T(\boldsymbol{X}_i)\right]^{-1}$ and $\boldsymbol{w}_T = \{w_T(\boldsymbol{X}_i) : i \in \{j :\, d_j = 1\}\}$ is the vector of weights for the treated group, and

$$\min_{\boldsymbol{w}_C}\max_{D_C} \widehat{V}_C(w_C, D_C) = \min_{\boldsymbol{w}_C}\max_{D_C} \left\{\frac{1}{n}\sum_{i=1}^n \log(D_C(\boldsymbol{X}_i)) + \sum_{i:d_i=0}^n w_{C,0}(\boldsymbol{X}_i)\log(1 - D_C(\boldsymbol{X}_i))\right\}, \quad (7.15)$$

where $w_{C,0}(\boldsymbol{X}_i) = w_C(\boldsymbol{X}_i)\left[\sum_{i:d_i=0} w_C(\boldsymbol{X}_i)\right]^{-1}$ and $\boldsymbol{w}_C = \{w_C(\boldsymbol{X}_i) : i \in \{j :\, d_j = 0\}\}$ is the vector of weights for the control group. However, Zubizarreta (2015) pointed out that the variance of the weighting estimator (7.6) should be affected by the variance of the weights. To attain a more stable estimator for Δ, we propose the following function to minimize the variance of the estimated weights:

$$\min_{\boldsymbol{w}_T} \quad \|\boldsymbol{w}_T - \bar{\boldsymbol{w}}_T\|_2^2, \quad \text{and} \quad \min_{\boldsymbol{w}_C} \quad \|\boldsymbol{w}_C - \bar{\boldsymbol{w}}_C\|_2^2,$$

where $\bar{\boldsymbol{w}}_T$ and $\bar{\boldsymbol{w}}_C$ denote the mean values of $\boldsymbol{w}_T$ and $\boldsymbol{w}_C$, respectively.

Based on the above discussions, the ACBN is constructed as follows, which consists of two sub-networks responsible for achieving covariate balance between the treated (control) group and the full sample separately. The objective functions of these two sub-networks can be summarized as follows.

A. Balance the treated group and the full sample

$$\max_{D_T} \left\{ \frac{1}{n} \sum_{i=1}^{n} \log D_T(\boldsymbol{X}_i) + \sum_{i:d_i=1}^{n} w_{T,1}(\boldsymbol{X}_i) \log(1 - D_T(\boldsymbol{X}_i)) \right\}, \tag{7.16}$$

and

$$\min_{\boldsymbol{w}_T} \left\{ \frac{1}{n} \sum_{i=1}^{n} \log D_T(\boldsymbol{X}_i) + \sum_{i:d_i=1}^{n} w_{T,1}(\boldsymbol{X}_i) \log(1 - D_T(\boldsymbol{X}_i)) + \alpha \cdot \|\boldsymbol{w}_T - \bar{\boldsymbol{w}}_T\|^2 \right\}. \tag{7.17}$$

B. Balance the control group and the full sample

$$\max_{D_C} \; d \left\{ \frac{1}{n} \sum_{i=1}^{n} \log D_C(\boldsymbol{X}_i) + \sum_{i:d_i=0}^{n} w_{C,0}(\boldsymbol{X}_i) \log(1 - D_C(\boldsymbol{X}_i)) \right\}, \tag{7.18}$$

and

$$\min_{\boldsymbol{w}_C} \left\{ \frac{1}{n} \sum_{i=1}^{n} \log D_C(\boldsymbol{X}_i) + \sum_{i:d_i=0}^{n} w_{C,0}(\boldsymbol{X}_i) \log(1 - D_C(\boldsymbol{X}_i)) + \beta \cdot \|\boldsymbol{w}_C - \bar{\boldsymbol{w}}_C\|^2 \right\}. \tag{7.19}$$

Here, α and β leverage the trade-off between the precision and the stability of our estimator.

D. Implementation Algorithm

Motivated by the powerful fitting capacity of neural networks, we set both generators and discriminators as multi-layer neural networks. Since all the outputs of our neural networks are in $[0, 1]$, we set the activation functions of each output layer as the sigmoid function. Of course, one can choose the number of layers and the number of neurons in each hidden layer flexibly based on the data. We choose Adam[1] as the optimizer for parameters and let the

[1]Adam is a good choice for a model's parameters because it combines the benefits of momentum and adaptive learning rates to converge faster. It is particularly effective for large-scale models and sparse data, and its default settings often work well, requiring minimal hyper-parameter tuning.

learning rate decay at an exponential rate. Besides, set $\alpha = \beta = 1$. The ACBN procedure is described in Algorithms 1 and 2.

Algorithm 1 Sub-network 1: Balance the treated group and the full sample

Input: $\{\boldsymbol{X}_i\}_{i=1}^n$, $\{\boldsymbol{X}_i\}_{i:d_i=1}$
Output: $\boldsymbol{w}_T$
Parameters: parameters for the generator θ_G and the discriminator θ_D, learning rate for the generator lr_G and the discriminator lr_D, number of iterations n_{iter}, number of inner iterations of the discriminator k
Initialize θ_G and θ_D
for n_{iter} **do**
  Input $\{\boldsymbol{X}_i\}_{i:d_i=1}$ into the generator, output $\boldsymbol{w}_T$
  for k steps **do**
    Update the discriminator by the chosen optimizer according to (7.16)
  end for
  Update the generator by the chosen optimizer according to (7.17)
end for

Algorithm 2 Sub-network 2: Balance the control group and the full sample

Input: $\{\boldsymbol{X}_i\}_{i=1}^n$, $\{\boldsymbol{X}_i\}_{i:d_i=0}$
Output: $\boldsymbol{w}_C$
Parameters: parameters for the generator θ_G and the discriminator θ_D, learning rate for the generator lr_G and the discriminator lr_D, number of iterations n_{iter}, number of inner iterations of the discriminator k
Initialize θ_G and θ_D
for n_{iter} **do**
  Input $\{\boldsymbol{X}_i\}_{i:d_i=0}$ into the generator, output $\boldsymbol{w}_C$
  for k steps **do**
    Update the discriminator by the chosen optimizer according to (7.18)
  end for
  Update the generator by the chosen optimizer according to (7.19)
end for

After the training of the ACBN, we plug the weights $\boldsymbol{w}_T$ and $\boldsymbol{w}_C$ into the weighting estimator (7.6) (the weights are normalized), which is similar to (7.4), and propose the estimator for Δ as

$$\widehat{\Delta}_{\text{acbn}} = \sum_{i:d_i=1} w_{T,1}(\boldsymbol{X}_i)Y_i - \sum_{i:d_i=0} w_{C,0}(\boldsymbol{X}_i)Y_i, \tag{7.20}$$

which is the ACBN estimator of Δ. It is clear that by comparing (7.20) with (7.3), (7.4) and (7.5), computing $\widehat{\pi}(\mathbf{X}_i)$ and $\widehat{\mu}_d(\mathbf{X}_i)$ for $d = 0$ and 1 is not needed. Therefore, it avoids a

possible misspecification of $\pi(\mathbf{X}_i)$ and $\mu_d(\mathbf{X}_i)$ for $d = 0$ and 1, so that the ACBN estimator is one of the doubly robust estimators.

E. Monte Carlo Simulations

To see how the finite sample performance of the proposed ACBN estimator, following Cai et al. (2025b), a series of Monte Carlo simulation studies are conducted in this section to evaluate the proposed estimator with a comparison with other existing methods. Similar to Fan et al. (2023), we set the covariate as $\boldsymbol{X}_i = (X_{i,1}, X_{i,2}, X_{i,3}, X_{i,4})$, where $X_{i,1} \sim \mathrm{N}(3, 2)$ and $X_{i,2}$, $X_{i,3}$ and $X_{i,4} \sim \mathrm{N}(0, 1)$ independently. We consider four different data generating process as follows.

M1: Both propensity score model and potential outcome model are correctly specified. The true potential outcome model is $Y_i(1) = 200 + 27.4X_{i,1} + 13.7(X_{i,2} + X_{i,3} + X_{i,4}) + \varepsilon_i$ and $Y_i(0) = 200 + 13.7(X_{i,2} + X_{i,3} + X_{i,4}) + \varepsilon_i$, where $\varepsilon_i \sim \mathrm{N}(0, 1)$ and the true value of the average treatment effect is $\Delta = 82.2$ with the true propensity score model as

$$\mathbb{P}(d_i = 1|\boldsymbol{X}_i = \boldsymbol{x}_i) = \frac{\exp(-\beta_1 x_{i,1} + 0.5x_{i,2} - 0.25x_{i,3} - 0.1x_{i,4})}{1 + \exp(-\beta_1 x_{i,1} + 0.5x_{i,2} - 0.25x_{i,3} - 0.1x_{i,4})},$$

where $\beta_1 = 0$, 0.33, 0.67, and 1.

M2: The potential outcome model is correctly specified but the propensity score model is misspecified. The potential outcome model is same as that in M1, where the true value of the average treatment effect is $\Delta = 82.2$. Set $x^*_{i,1} = \exp(x_{i,1}/3)$, $x^*_{i,2} = x_{i,2}/\{1 + \exp(x_{i,1})\} + 10$, $x^*_{i,3} = x_{i,1}x_{i,3}/25 + 0.6$, and $x^*_{i,4} = x_{i,1} + x_{i,4} + 20$. The true propensity score model is

$$\mathbb{P}(d_i = 1|\boldsymbol{X}_i = \boldsymbol{x}_i) = \frac{\exp(-\beta_1 x^*_{i,1} + 0.5x^*_{i,2} - 0.25x^*_{i,3} - 0.1x^*_{i,4})}{1 + \exp(-\beta_1 x^*_{i,1} + 0.5x^*_{i,2} - 0.25x^*_{i,3} - 0.1x^*_{i,4})}.$$

However, $x^*_{i,1}$, $x^*_{i,2}$, $x^*_{i,3}$ and $x^*_{i,4}$ are unobservable, but we can only observe $x_{i,1}$, $x_{i,2}$, $x_{i,3}$ and $x_{i,4}$. Similarly, $\beta_1 = 0$, 0.33, 0.67, and 1.

M3: The propensity score model is correctly specified but the potential outcome model is misspecified. The propensity score model is same as that in M1, where $\beta_1 = 0$, 0.13, 0.27, and 0.4, and the true potential outcome model is $Y_i(1) = 200 + 27.4X^2_{i,1} + 13.7(X^2_{i,2}+X^2_{i,3}+X^2_{i,4})+\varepsilon_i$ and $Y_i(0) = 200+13.7(X^2_{i,2}+X^2_{i,3}+X^2_{i,4})\varepsilon_i$, where $\varepsilon_i \sim \mathrm{N}(0, 1)$ with the true value of the average treatment effect $\Delta = 301.4$.

M4: Both propensity score model and potential outcome model are misspecified. The propensity score model is same as that in M2, where $\beta_1 = 0$, 0.13, 0.27, and 0.4 and the potential outcome model is same as that in M3. The true value of the average treatment effect is $\Delta = 301.4$.

We conduct 500 Monte Carlo simulations for each data generating process with the sample size $n = 300$ and $n = 1000$, respectively. Furthermore, we compare our estimator as in (7.20) with a range of existing methods:

1. *base*: $\widehat{\Delta}_{\text{base}} = \frac{1}{n_1}\sum_{i:d_i=1} Y_i - \frac{1}{n_0}\sum_{i:d_i=0} Y_i$;
2. *ps*: the IPTW estimator with the propensity score fitted by the generalized maximum likelihood;
3. *gbm*: the IPTW estimator with the propensity score fitted by the gradient boosting machine;
4. *cbps1*: the just-identified covariate balancing propensity score estimator proposed by Imai and Ratkovic (2014);
5. *cbps2*: the over-identified covariate balancing propensity score estimator proposed by Imai and Ratkovic (2014);
6. *npcbps*: the nonparametric covariate balancing propensity score estimator proposed by Fong et al. (2018);
7. *ebal*: the entropy balancing estimator proposed by Hainmueller (2012);
8. *optw*: the stable balancing weights estimator proposed by Zubizarreta (2015);
9. *ebcw*: the empirical balancing calibration weighting estimator proposed by Chan et al. (2016);
10. *energy*: the energy balancing estimator proposed by Huling and Mak (2024).

For 500 simulations, the median of biases and the root mean squares error of biases are computed, denoted by BIAS and RMSE, respectively, which are presented in Tables 7.1 - 7.4. For each setting, it is emphasized the smallest value in magnitude of different metrics among all estimators in boldface.

The simulation result under M1 is shown in Table 7.1, where both the potential outcome model and the propensity score model are correctly specified. Our method outperforms other methods in most cases and is relatively insensitive to the value of β_1, which implies that the performance of our estimator improves as the sample size n increases. Table 7.2 displays the

Table 7.1: The median (in the top panel) of biases and root mean square error (in the bottom panel) of various estimators under M1.

		n=300				n=1000			
	β_1	0	0.33	0.67	1	0	0.33	0.67	1
BIAS	base	1.2961	-9.1492	-25.3396	-39.6559	2.2062	-9.7610	-25.1639	-38.7628
	ps	-0.0400	0.4905	-1.6428	-7.6996	0.1179	0.4403	-0.6050	-0.6730
	gbm	0.2523	-2.6716	-10.4586	-18.5409	0.5139	-2.2627	-7.3940	-12.4089
	cbps1	0.1770	-1.2717	-5.0317	-11.9379	0.2836	-0.4356	-2.4660	-4.3048
	cbps2	-0.0968	0.0226	-1.8654	-8.0433	0.1140	0.2065	-0.8103	-1.9758
	npcbps	-0.1807	0.5684	0.7333	-16.3177	-0.1086	1.1681	2.0590	2.0144
	ebal	-0.1064	0.2432	0.1429	-1.5093	0.1344	0.2321	-0.2581	-0.2546
	optw	-0.2128	**-0.0187**	**-0.0933**	-18.4330	**0.0038**	**-0.0710**	-0.5296	-0.5022
	ebcw	-0.1064	0.2432	0.1429	-17.8369	0.1344	0.2321	**-0.2581**	-0.2545
	energy	0.1899	-1.2865	-5.4640	-13.4037	0.8659	-6.0509	-5.3865	-14.9079
	acbn	**0.0158**	-0.0771	-0.3617	**-0.4302**	-0.2906	-0.2712	-0.2642	**-0.1596**
RMSE	base	4.3671	10.5163	26.2313	40.4526	3.2982	10.1232	25.3883	38.9802
	ps	2.2006	4.1367	10.3688	18.0311	1.6398	2.0238	5.9417	11.7304
	gbm	2.5950	4.5358	11.9229	21.0067	1.7586	2.8627	7.9021	13.3282
	cbps1	2.4020	3.6826	7.9672	16.2282	1.6879	1.8270	4.7555	8.7796
	cbps2	2.2065	2.5061	3.8378	10.0679	1.6246	1.3345	1.7019	3.2791
	npcbps	2.2024	2.3211	3.7552	21.2753	1.5551	1.8458	3.2348	6.8233
	ebal	2.1852	2.2850	2.3451	5.8765	1.5295	1.2464	1.1595	1.3062
	optw	2.2000	2.3035	2.3156	28.5389	1.5191	**1.2347**	1.2489	1.3804
	ebcw	2.1852	2.2850	2.3451	28.6662	1.5295	1.2463	**1.1594**	**1.3062**
	energy	3.2038	3.8907	7.6297	14.8766	4.4616	8.7100	9.6490	23.2526
	acbn	**1.8341**	**1.7075**	**1.7568**	**1.7738**	**1.1149**	1.2560	1.2353	1.3204

simulation result for the case where only the propensity score model is misspecified. We find that the misspecification of the propensity score model may not deteriorate the estimation compared to M1 and our estimator performs relatively well in most cases. Table 7.3 shows the simulation result for the case where only the potential outcome model is misspecified. It is clear that the misspecified potential outcome model exhibits a more negative impact on the estimation than the misspecified propensity score model. Our estimator significantly outperforms other methods. Although the performance of our estimator still improves as the sample size n increases, the improvement is relatively tiny. The simulation result under M4 is shown in Table 7.4, where both the potential outcome model and the propensity score model are misspecified. Our estimator still significantly outperforms other estimators.

Table 7.2: The median (in the top panel) of biases and root mean square error (in the bottom panel) of various estimators under M2.

		n=300				n=1000			
	β_1	0	0.33	0.67	1	0	0.33	0.67	1
BIAS	base	-2.1260	-6.0864	-14.2512	-22.8557	-1.3461	-6.3206	-13.9825	-23.1692
	ps	-0.5340	0.1746	-1.6978	-5.5086	**-0.0082**	0.2013	-1.7661	-5.6043
	gbm	-0.9289	-2.4378	-6.6090	-10.7574	-0.4542	-1.7273	-4.3616	-8.2150
	cbps1	-0.7831	-1.0681	-3.9038	-8.0708	-0.1326	-0.7965	-3.5050	-7.5880
	cbps2	**0.0413**	-0.2810	-2.0062	-4.0073	-0.0095	-0.7802	-2.3342	-4.0380
	npcbps	-0.2376	**0.0658**	0.7327	1.4217	0.1295	0.4225	1.1368	2.5571
	ebal	-0.2148	-0.1010	0.2088	0.0864	-0.0427	**-0.0583**	0.1623	**-0.1374**
	optw	-0.3503	-0.3464	**-0.0491**	-0.1188	-0.2312	-0.3575	**-0.1237**	-0.4110
	ebcw	-0.2149	-0.1010	0.2088	**0.0863**	-0.0426	-0.0584	0.1622	-0.1375
	energy	-1.3292	-0.2073	-0.3587	-2.4505	-1.3487	-3.4626	-2.9450	-1.6811
	acbn	-0.1403	-0.3637	-0.5660	-0.2303	-0.2334	-0.2831	-0.1913	-0.1558
RMSE	base	6.2357	7.5695	14.6868	23.1256	3.4568	6.8034	14.1351	23.2576
	ps	3.5803	3.2323	3.6755	7.0083	1.5742	1.8763	2.4755	5.9610
	gbm	3.7170	3.8625	7.1863	11.1870	2.0538	2.4118	4.6152	8.3745
	cbps1	3.1202	2.8668	4.9320	8.9069	1.6834	1.8670	3.8793	7.8386
	cbps2	2.6102	2.1220	3.2927	5.1491	1.2727	1.6192	2.8236	4.3439
	npcbps	2.4678	1.9556	2.2283	3.3356	1.3048	1.4838	2.4253	4.7894
	ebal	2.4023	1.9835	2.0548	2.3759	**1.2562**	1.3570	1.2917	1.3631
	optw	2.4370	2.0195	2.0613	2.3764	1.3018	1.4290	1.3085	1.4168
	ebcw	2.4024	1.9835	2.0548	2.3759	**1.2562**	1.3570	1.2917	1.3632
	energy	8.7623	2.8737	2.5373	4.2829	4.3693	5.3521	7.0131	3.9265
	acbn	**1.8699**	**1.8017**	**1.8085**	**1.7340**	1.3629	**1.2141**	**1.1689**	**1.2083**

In summary, the proposed estimator performs comparably well for all settings, especially when the outcome model is misspecified. Also, it is insensitive to the value of β_1, which controls the degree of confoundedness. The performance of the proposed estimator improves as the sample size increases, though the improvement is quite faint, suggesting that the convergence rate of the proposed estimator may be relatively slow. Despite the deficient theoretical properties, the proposed econometric method is reasonably practical. It has been noticed that our estimator is not always optimal for cases M1 and M2. This is probably because it is sufficient to balance some lower order moments rather than the whole distribution in finite samples under these settings.

F. An Empirical Example

In this section, we briefly present a study on an empirical example. For details, please read the paper by Cai et al. (2025b). We assess our proposed method in the well-known Twins dataset, which collects data from twin births in the USA between 1989-1991 and is

Table 7.3: The median (in the top panel) of biases and root mean square error (in the bottom panel) of various estimators under M3.

		n=300				n=1000			
	β_1	0	0.33	0.67	1	0	0.33	0.67	1
BIAS	base	**-0.0547**	-24.1660	-53.7813	-85.2450	0.9625	-22.7356	-49.2494	-80.2607
	ps	-0.8655	-1.4628	-3.3118	-7.1340	0.6294	-1.1506	0.2360	**-0.0589**
	gbm	-1.5567	-11.2590	-25.9522	-41.3039	-0.5938	-8.6402	-17.4871	-28.7185
	cbps1	-0.3840	-5.0993	-11.9178	-20.5596	0.5586	-3.4921	-4.0569	-6.6465
	cbps2	-1.0062	-1.5351	-4.2287	-8.7881	0.6477	-0.9057	**0.1927**	-1.1483
	npcbps	-0.6598	**0.6212**	**1.8049**	**2.3110**	0.9812	3.7006	10.2456	16.6889
	ebal	-0.9158	-1.6202	-4.1878	-8.2487	0.8850	**-0.5967**	-0.8491	-3.3163
	optw	-1.1154	-3.9117	-8.9627	-15.7578	0.9151	-2.9258	-5.54734	-11.5992
	ebcw	-0.9159	-1.6201	-4.1877	-8.2492	0.8850	-0.5967	-0.8494	-3.3164
	energy	-2.5611	-5.1347	-11.3633	-18.5425	**-0.2165**	-8.5588	-22.2920	-35.3952
	acbn	-1.3561	-0.8914	-3.2469	-3.0026	-1.3874	-1.2955	-0.2791	-1.6704
RMSE	base	21.1486	32.4145	58.0516	87.5306	11.6375	25.2220	50.6434	81.4115
	ps	17.5513	18.8284	25.1193	31.6156	8.6522	9.5420	13.3271	17.2422
	gbm	17.1320	20.5305	32.0786	45.8038	8.1203	11.7608	20.4453	31.5381
	cbps1	17.1192	19.5551	24.2179	31.4697	8.7822	10.1593	13.7573	16.2315
	cbps2	17.4907	17.5708	20.0307	22.0649	8.5273	8.8625	11.0524	11.1984
	npcbps	15.8371	16.6788	18.7319	17.4638	8.4356	9.3388	15.1657	20.5444
	ebal	15.9373	16.6127	18.0734	18.2692	8.2323	8.1366	9.8629	10.1463
	optw	16.0736	16.8913	19.3864	22.3264	8.4299	8.6531	11.2349	15.0933
	ebcw	15.9374	16.6127	18.0733	18.2695	8.2322	8.1365	9.8628	10.1464
	energy	15.1327	16.9856	21.5986	27.1951	12.1475	18.4524	35.0192	54.0601
	acbn	**6.5086**	**6.9251**	**6.9858**	**7.5057**	**5.7516**	**5.6200**	**6.0959**	**6.0105**

first studied by Almond et al. (2005). For each pair of twins, there are various covariates concerning the pregnancy, the delivery, the twins, and the parents. The outcome variable is the mortality of each of the twins in the first year after birth. We can utilize this dataset to study the impact of the birth weight on the first-year mortality. Controlling for the sex, we can define the treatment $d = 1$ as being the heavier one in each pair of twins and $d = 0$ as being the lighter one in each pair of twins. Notice that we have records for both twins, so that we can assume that the two potential outcomes are both observable. This appealing feature prompts the universal use of the Twins dataset as a benchmark for causal inferences, see, for example, Louizos et al. (2017), Yao et al. (2018), and Yoon et al. (2018). We can simulate an observational study and different degrees of selection bias by designing different treatment assignment mechanisms. In other words, for each pair of twins, we can decide which one is the observed infant by the generated treatment variable d. In the Twins dataset, the factual outcome is generated by simulating the treatment assignment, which is different from being directly observed as usual. If we just generate d at random, then the experiment is no more

Table 7.4: The median (in the top panel) of biases and root mean square error (in the bottom panel) of various estimators under M4.

		n=300				n=1000			
	β_1	0	0.33	0.67	1	0	0.33	0.67	1
BIAS	base	-3.7239	-12.3598	-23.3587	-42.8503	-4.6313	-10.5099	-25.0441	-42.4660
	ps	1.0370	-1.8585	-1.5911	-7.2620	-1.7465	**-0.0942**	-3.2330	-6.9679
	gbm	2.3413	-3.5267	-7.6936	-18.5557	2.3054	0.4619	-5.8666	-13.0168
	cbps1	0.9426	-2.9557	-4.2475	-13.1586	-2.0561	-0.8889	-5.5983	-11.3090
	cbps2	2.9845	**1.1396**	**-0.3412**	-6.9022	-1.3686	-0.1450	-5.1126	-9.9579
	npcbps	1.1555	-1.8040	-1.3154	-5.4529	-2.0578	0.1572	-2.1501	-2.5861
	ebal	1.8006	-1.9863	-2.2789	-7.6795	-1.7765	-0.6517	-4.0660	-7.2239
	optw	1.4323	-2.7878	-3.5196	-9.7951	-2.2424	-1.5822	-5.4457	-9.4888
	ebcw	0.8580	-1.9862	-2.2791	-7.6796	-1.7765	-0.6519	-4.0666	-7.2242
	energy	-9.5723	-7.9186	-5.3200	-4.2981	-4.2827	-10.3243	-15.4538	-15.2420
	acbn	**-0.7377**	-1.5810	-1.4478	**-0.7695**	**-1.2173**	-1.4462	**-2.0633**	**-1.6943**
RMSE	base	16.9462	21.7012	27.4203	45.1989	9.9552	13.0830	26.2393	43.2093
	ps	17.9595	17.3012	16.6683	15.9745	8.8165	7.4856	8.7407	11.4374
	gbm	15.7497	15.9577	15.9822	22.6752	8.7966	7.9084	9.2203	15.6140
	cbps1	17.0589	17.3882	17.1411	19.2636	8.8896	7.6658	10.1111	14.6533
	cbps2	17.6004	18.2754	18.8160	17.3294	8.8353	7.8356	10.1401	13.7505
	npcbps	16.7744	16.9897	16.0459	14.8887	8.9892	7.3999	8.3995	9.6460
	ebal	16.4257	16.9420	15.8236	15.6512	8.8407	7.4480	8.7982	11.0448
	optw	16.5672	17.1011	15.8702	16.7178	8.9925	7.6405	9.5291	12.5681
	ebcw	17.7761	16.9422	15.8236	15.6512	8.8407	7.4480	8.7984	11.0449
	energy	31.3516	33.6812	22.6368	17.6054	10.6934	19.7238	22.2289	28.8980
	acbn	**5.8358**	**5.7089**	**7.0290**	**6.3077**	**5.0374**	**4.5200**	**5.0601**	**4.8748**

than a randomized clinical trial.

Referring to the existing research, the focus here is only on twins of the same sex with born weights less than 2 kilograms, with the sample size of the dataset $n = 11,400$, and 30 covariates are selected, shown in Table 7.5. The reader is referred to the paper by Cai et al. (2025b) for more information about this dataset and the detailed descriptions of the related covariates. In this dataset, the mortality rate of the lighter twin is about 17.7% and of the heavier 16.1%, indicating that the true value of the average treatment effect is about $\Delta = -1.6\%$. To simulate the confoundedness, the treatment assignment mechanism should be related to the observed covariates. Referring to existing literature, we design the treatment assignment mechanism as follows. Indeed, the Twins dataset is a popular benchmark in causal inference since it is semi-synthetic. There is no standard guidance on how to design the treatment assignment mechanism. Louizos et al. (2017) based the treatment assignment on a single variable which has the highest correlation with the outcome and Yoon et al.

Table 7.5: A summary of the chosen covariates.

Class	Covariates
Pregnancy Risk Factors	Anemia, Cardiac disease, Acute or chronic lung disease, Diabetes, Genital herpes, Hydramnios, Hemoglobinopathy, Chronic hypertension, Pregnancy-associated hypertension, Eclampsia, Incompetent cervix, Previous infant 4000+ grams, Previous preterm infant, Renal disease, Rh sensitization, Uterine bleeding, Other medical risk factors (all are binary)
Gestation and Delivery	Gestation period, Weight gain, Place or facility of delivery, Adequacy of care, Total number of prenatal visits, Month prenatal care began, Average number of cigarettes per day, Average number of drinks per week (all are categorical)
Mom	Age (categorical), Education (categorical), Marital status (binary) Resident status (categorical), Detail total birth order (ordinal)

(2018) treated the treatment assignment on all observed covariates. We adopt the treatment assignment mechanism as in Yoon et al. (2018), which constructs a simple logistic model for the treatment variable d as, by following Cai et al. (2025b),

$$P(d = 1|\boldsymbol{X}, \varepsilon) = \left[1 + \exp(\boldsymbol{\beta}^\top \boldsymbol{X} + \varepsilon)\right]^{-1},$$

where $\boldsymbol{\beta} \sim \mathrm{U}(-0.1, 0.1)^{30\times 1}$ and $\varepsilon \sim \mathrm{N}(0, 0.1)$. We repeat this simulation 200 times. We first examine whether the covariates are balanced in the 200 sets of data. We statistically test whether the mean value of each covariate is equal for the treated and control groups in each set of data. The result of t-tests is shown in Table 7.6. From Table 7.6, one can see that the covariates are indeed imbalanced in the 200 sets of data, so that it is necessary to adjust the data for the unconfoundedness assumption holds.

Table 7.6: The number of rejections (Rej.) and acceptances (Accept) among 200 trials.

Covariate	Rej.	Accept.	Covariate	Rej.	Accept.	Covariate	Rej.	Accept.
X1	129	71	X11	93	107	X21	185	15
X2	25	175	X12	49	151	X22	51	149
X3	41	159	X13	45	155	X23	157	43
X4	31	169	X14	42	158	X24	166	34
X5	47	153	X15	42	158	X25	150	50
X6	109	91	X16	66	134	X26	142	58
X7	23	177	X17	105	95	X27	66	134
X8	82	118	X18	34	166	X28	119	81
X9	28	172	X19	175	25	X29	175	25
X10	105	95	X20	86	114	X30	159	41

We compare our method with the aforementioned existing methods and measure the performance by the median of biases and the root mean squares error of these estimators. The results are shown in Table 7.7, from which one can see that our estimator is comparable,

Table 7.7: The performance of various estimators on the Twins dataset.

	$\widehat{\Delta}_{\text{base}}$	$\widehat{\Delta}_{\text{ps}}$	$\widehat{\Delta}_{\text{cbps1}}$	$\widehat{\Delta}_{\text{cbps2}}$	$\widehat{\Delta}_{\text{ebal}}$	$\widehat{\Delta}_{\text{ebcw}}$	$\widehat{\Delta}_{\text{energy}}$	$\widehat{\Delta}_{\text{acbn}}$
BIAS	0.0083	-0.0026	-0.0011	-0.0008	-0.0102	0.0006	-0.0004	-0.0008
RMSE	0.1030	0.0620	0.0445	0.0446	0.1233	0.0956	0.0735	0.0582

although it might not be the best. The covariate balancing propensity score methods proposed by Imai and Ratkovic (2014) seem to perform slightly better. This is probably due to the extreme ratio of the size of the treated group and control groups in some sets of data. In addition, most covariates are discrete which only take values in 0 or 1. In such a situation, it is probably sufficient to balance the first order moment of the covariates, which is why the CBPS method might perform slightly better. This phenomenon is consistent with the observation in the simulation study as in the previous section (Section 7.2.3E).

7.2.4 Synthetic Control Methods

In this section, assume that we observe n units, some of which are exposed to the treatment or intervention of interest. We code the treatment status of unit i using the binary variable d_i, so $d_i = 1$ if i unit is treated and $d_i = 0$ otherwise. To define treatment effects, we adopt a potential outcomes framework, as in Rubin (1974). Let Y_{1i} and Y_{0i} be the be random variables representing potential outcomes under treatment and under no treatment, respectively, for unit i. The effect of the treatment for unit i is $\Delta_i = Y_{1i} - Y_{0i}$. Randomized outcomes $Y_i = d_i Y_{1i} + (1 - d_i)Y_{0i}$. Let $\mathbf{X}_i \in \mathbb{R}^p$ be a $(p \times 1)$ vector of pre-treatment predictors of Y_{0i}, where d might be very large. We assume that the observed data are $(Y_i; \mathbf{X}_i) = (Y_{1i}; \mathbf{X}_i)$ for n_1 treated observations and $(Y_i; \mathbf{X}_i) = (Y_{0i}; \mathbf{X}_i)$ for n_0 untreated observations. Combining data for treated and non-treated, we obtain the pooled data set, $\{(Y_i; d_i; \mathbf{X}_i)\}_{i=1}^n$ with $n = n_0 + n_1$. To simplify notation, we reorder the observations in the data so that the n_0 untreated observations come first. The quantities of interest are the treatment effects on the treated units, Δ_i for $i = n_0+1, \cdots, n$, and/or the average treatment

effect on the treated:

$$\widehat{\Delta} = \frac{1}{n_1}\sum_{i=n_0+1}^{n}\Delta_i = \frac{1}{n_1}\sum_{i=n_0+1}^{n}(Y_{1i}-Y_{0i}), \tag{7.21}$$

the estimated value of $\Delta = \mathbb{E}(\Delta_i)$, the average treatment effect. But, Y_{0i} for $i = n_0+1, \cdots, n$ is not observed. To obtain a feasible version of $\widehat{\Delta}$, one of popular methods proposed in the literature is the synthetic control (SC) method, as argued in Abadie (2021).

A SC method is defined as a weighted average of the units in the pool of observed outcomes. Formally, a synthetic control can be represented by a $n_0 \times 1$ vector of weights, $\mathbf{W}_i^* = (W_{i,1}^*, \cdots, W_{i,n_0}^*)^\top$. Given a set of weights, $\mathbf{W}_i^*$, the synthetic control estimators (linear predictor) of Y_{0i} and Δ_i are $\widehat{Y}_{0i} = \sum_{j=1}^{n_0} W_{i,j}^* Y_j$ for $i = n_0+1, \cdots, n$, and $\widehat{\Delta}_i = Y_{1t} - \widehat{Y}_{0i}$, respectively. As elaborated by Abadie and Gardeazabal (2003) and Abadie et al. (2010), the purpose of choosing $\mathbf{W}_i^*$ is that the resulting synthetic control best resembles the characteristics of the treated unit before the intervention. That is, Abadie and Gardeazabal (2003) and Abadie et al. (2010) proposed to choose the synthetic control, $\mathbf{W}_i^*$ for each treated unit $i = n_0+1, \cdots, n$, to minimize

$$\left(\mathbf{X}_i - \sum_{j=1}^{n_0} W_{i,j}\mathbf{X}_j\right)^\top \mathbf{V}\left(\mathbf{X}_i - \sum_{j=1}^{n_0} W_{i,j}\mathbf{X}_j\right) \tag{7.22}$$

subject to the restriction that $\{W_{i,1}, \cdots, W_{i,n_0}\}$ are non-negative and sum to one, which is the so-called convex hull constraint, where $\mathbf{V}$ is a weighting matrix. Then, the estimated treatment effect for the treated unit $i = n_0+1, \cdots, n$ is

$$\widehat{Y}_{0i} = \sum_{j=1}^{n_0} W_{i,j}^* Y_j, \tag{7.23}$$

so that the synthetic control estimator in (7.21) becomes

$$\begin{aligned}\widehat{\Delta} &= \frac{1}{n_1}\sum_{i=n_0+1}^{n}\widehat{\Delta}_i = \frac{1}{n_1}\sum_{i=n_0+1}^{n}(Y_{1i}-\widehat{Y}_{0i}) = \frac{1}{n_1}\sum_{i=n_0+1}^{n}\left[Y_{1i} - \sum_{j=1}^{n_0} W_{i,j}^* Y_j\right] \\ &= \frac{1}{n_1}\sum_{i=n_0+1}^{n} Y_{1i} - \frac{1}{n_0}\sum_{j=1}^{n_0} W_j^{\text{sc}} Y_j,\end{aligned} \tag{7.24}$$

where $W_j^{\text{sc}} = n_0 \sum_{i=n_0+1}^{n} W_{i,j}^*/n_1$. For many quasi-experimental data used in economics, marketing and other social sciences, the treated unit and the control units may exhibit substantial heterogeneity and the treated units outcome and a weighted average (with weights sum to one) of the control units outcomes may not follow parallel paths in the absent of

treatment. There are two simple modifications advocated by Doudchenko and Imbens (2016). Specifically, one can add an intercept and remove the coefficients sum to one restriction in a standard synthetic control model, i.e., one still keeps the non-negative constraints but drop the add-to-one restriction, as in Li (2020). Without loss of generality, assume that the first component in $\mathbf{X}_i$ is 1 so that an intercept is included.

As argued by Athey and Imbens (2017), the SC method is regarded as "arguably the most important innovation in the evaluation literature in the last 15 years." Also, Wan et al. (2018) explored the pros and cons of the SCM and pointed out that the SCM is a more efficient method, when the constraints are valid. However, when the constraints are invalid, the SCM could lead to biased prediction of counterfactual. As elaborated by Wan et al. (2018), the SCM convex hull constraints might not be needed nor necessarily satisfied in many cases. Also, it is important to note, however, that for any particular data set there are not ex ante guarantees on the size of the differences $\mathbf{X}_i - \sum_{j=1}^{n_0} W_{i,j}\mathbf{X}_j$ in (7.22) for $i = n_0 + 1, \cdots, n$. When these differences are large, Abadie et al. (2010) and Abadie (2021) recommended against the use of synthetic controls because of the potential for substantial biases. The worst case is that if the correlations among X_i for $i = n_0+1, \cdots, n$ and X_j for $j = 1, \cdots, n_0$ are very weak, $W_{i,j}$'s in (7.22) are zero or very close to zero. For such a case, (7.23) might not be good to predict Y_{0i}. Therefore, to avoid producing misleading estimates from synthetic control estimators, Abadie (2021) provided some practical guidances for applied researchers on how to employ synthetic control methods. Also, to avoid interpolation biases, Abadie and L'Hour (2021) proposed a penalized version of the SCM for disaggregated data by adding a penalty term in (7.22).

7.2.5 Quasi-SC Method for Nonlinear Models

To make the conventional synthetic control method more flexible to estimate the average treatment, Cai et al. (2025c) proposed a quasi synthetic control (QSC) method for nonlinear models under the index model framework with possible high-dimensional covariates, together with a suggestion of using the minimum average variance estimation (MAVE) method, as elaborated in Section 2.7.2, to estimate parameters and the LASSO-type procedure to choose high-dimensional covariates. They derived the asymptotic distribution of the proposed ATE estimators for both finite and diverging dimensions of covariates. A properly designed Bootstrap method is proposed to obtain confidence intervals and its theoretical justification is

provided. When the dimension of covariates is greater than the sample size, Cai et al. (2025c) suggested using the robust version of sure independence screening procedure based on the distance correlation to first reduce the dimensionality and then apply the MAVE approach to estimate parameters. The QSC method is briefly described below.

To consider a more general setting, one can consider the prediction function based on the conditional expectation of Y_{0i} given $\mathbf{X}_i$, denoted by $m(x) = \mathbb{E}(Y_{0i}|\mathbf{X}_i = \mathbf{x})$, in an index form as $m(\mathbf{x}) = m(\boldsymbol{\beta}^\top \mathbf{x}) = m(z)$, where $m(\cdot)$ is an unknown function and $z = \boldsymbol{\beta}^\top \mathbf{x} \in \mathbb{R}^1$, which covers the linear model as a special case. For the identification purpose, it is commonly assumed, in what follows, that the first element of $\boldsymbol{\beta}$ is positive and $||\boldsymbol{\beta}||^2 = \sum_{k=1}^p \beta_k^2 = 1$. Then, for $i = n_0 + 1, \cdots, n$, $\mathbb{E}(Y_{0i}) = E\left[\mathbb{E}(Y_{0i}|\mathbf{X}_i)\right] = \mathbb{E}\left[E(Y_{0i}|Z_i)\right]$, where $Z_i = \boldsymbol{\beta}^\top X_i$ for a given $\boldsymbol{\beta}$, so that the estimation of $m(z)$ is one-dimensional and the so-called curse of dimensionality in a nonparametric smoothing can be avoided. Under some regularity conditions, the kernel type (Nadaraya-Watson)[2] estimate of $m(z)$, based on the data $\{(Y_j, \mathbf{X}_j)\}_{j=1}^{n_0}$ of the control group, is given by

$$\tilde{m}(z) = \sum_{j=1}^{n_0} c_{j,h}(z) Y_j,$$

where $c_{j,h}(z) = K_h(Z_j - z) / \sum_{l=1}^{n_0} K_h(Z_l - z)$. Now, the infeasible prediction of Y_{0i} is denoted by $\tilde{Y}_{0i}$

$$\tilde{Y}_{0i} = \tilde{m}(Z_i) = \sum_{j=1}^{n_0} c_{j,h}(Z_i) Y_j \tag{7.25}$$

for $i = n_0 + 1, \cdots, n$. Actually, (7.25) is infeasible since it is based on the unknown quantities $\{Z_j\}_{j=1}^{n_0}$. Accordingly, the infeasible estimate of Δ, $\tilde{\Delta}$ is given by

$$\tilde{\Delta} = \frac{1}{n_1} \sum_{i=n_0+1}^{n} \left[Y_{1i} - \sum_{j=1}^{n_0} c_{j,h}(Z_i) Y_j \right] = \frac{1}{n_1} \sum_{i=n_0+1}^{n} Y_i - \frac{1}{n_0} \sum_{j=1}^{n_0} a_{j,h} Y_j, \tag{7.26}$$

where $a_{j,h} = a_h(Z_j)$ and

$$a_h(z) = \frac{1}{n_1} \sum_{i=n_0+1}^{n} K_h(Z_i - z) \left[\frac{1}{n_0} \sum_{l=1}^{n_0} K_h(Z_l - Z_i) \right]^{-1}.$$

Clearly, under this nonlinear setting, we need to find the weights $\boldsymbol{\beta}$ such that $\boldsymbol{\beta}^\top \mathbf{X}_i$ can be the best to predict Y_{0i} for $i = 1, \cdots, n_0$. In other words, we need to search for $\boldsymbol{\beta}$ such that its

[2]Of course, one can use the kernel smoothing technique such as the local polynomial estimation method as in Fan and Gijbels (1996).

conditional expectation of Y_{0i} given Z_i matches the conditional expectation of Y_{0i} given $\mathbf{X}_i$ as close as possible. Therefore, to do so, we suggest using the index model and its estimation approach such as MAVE, described in Section 2.7.2 or Section 2.4 of Cai et al. (2025c).

Interestingly, our method shares some similarities and differences with the synthetic control method proposed by Abadie and Gardeazabal (2003). Although the SC method originally designed to deal with the panel data setting, Abadie and L'Hour (2021) presented a penalized version of the SC method for disaggregated data. Apparently, $c_{j,h}(Z_i)$ in (7.25) for nonlinear model is identical to the SC method weights $W^*_{i,j}$ in (7.23) for linear model or $\mathbf{W}^{\text{sc}}_i$ defined in (4) in Abadie and L'Hour (2021). However, different from Abadie and L'Hour (2021), the weights $\{c_{j,h}(Z_i)\}$ take care of both the best prediction to resemble the characteristics of the treated unit before the intervention and nonlinearity of prediction function since our model is in a semiparametric nature. Besides, our approach does not require that weights $a_{j,h}$ should satisfy the standard constraints as in Abadie and L'Hour (2021). Instead, our method is similar to that for the panel data approach (PDA) or HCW method as in Hsiao et al. (2012) and Wan et al. (2018) as well as Ouyang and Peng (2015), which will be discussed in the next section (Section 7.2.6), in the sense that it does not have constraints on weights such as nonnegative weights. Therefore, this is the reason that our method is termed as the quasi synthetic control method, although both have different motivations.

From the above discussions, the QSC method estimation procedure for estimating Δ consists of the following two steps. First, use (2.66) given in Section 2.7.2 or (9) in Section 2.4 in Cai et al. (2025c) to obtain $\widehat{\boldsymbol{\beta}}$, and then, set $\widehat{Z}_j = \widehat{\boldsymbol{\beta}}^\top \mathbf{X}_j$ for $j = 1, \cdots, n_0$ and $\widehat{Z}_i = \widehat{\boldsymbol{\beta}}^\top \mathbf{X}_i$ for $i = n_0 + 1, \cdots, n$. Second, compute the feasible estimate of Δ based on (7.26), and $\widehat{\Delta}$ is defined as

$$\widehat{\Delta} = \frac{1}{n_1} \sum_{i=n_0+1}^{n} \left[Y_{1i} - \sum_{j=1}^{n_0} \widehat{c}_{j,h}(\widehat{Z}_i) Y_j \right] = \frac{1}{n_1} \sum_{i=n_0+1}^{n} Y_i - \frac{1}{n_0} \sum_{j=1}^{n_0} \widehat{a}_{j,h} Y_j, \tag{7.27}$$

where $\widehat{a}_{j,h} = \widehat{a}_h(\widehat{Z}_j) = \frac{1}{n_1} \sum_{i=n_0+1}^{n} K_h(\widehat{Z}_i - \widehat{Z}_j) \left[\frac{1}{n_0} \sum_{l=1}^{n_0} K_h(\widehat{Z}_l - \widehat{Z}_i) \right]^{-1}$, which is similar to W^{sc}_j as in (7.24) or (4) in Abadie and L'Hour (2021).

Under some regularity conditions, Cai et al. (2025c) showed that the $\widehat{\Delta}$ in (7.27) has the following asymptotic normality

$$\sqrt{n_1} \left(\widehat{\Delta} - \Delta \right) \xrightarrow{d} N\left(0, \sigma^2_\Delta\right), \tag{7.28}$$

where σ_Δ^2 is the asymptotic variance, which is Theorem 1 in Cai et al. (2025c), from which one can see that σ_Δ^2 has a complicated expression, so that it is not easy to get its consistent estimate due to the complicated structure. To facilitate an easy inference, Cai et al. (2025c) proposed the following hybrid Bootstrap procedure by combining the (conditional) wild Bootstrap similar to that in Zhang et al. (2020) for single index models and the nonparametric Bootstrap, to estimate σ_Δ^2.

Step 1. Given $\{(\mathbf{X}_j, Y_j)\}_{j=1}^{n_0}$ and $\{(\mathbf{X}_i, Y_i)\}_{i=n_0+1}^{n}$, estimate the treatment effect by (7.27) as $\widehat{\Delta}$.

Step 2. Generate the nonparametric Bootstrap sample $\{(\mathbf{X}_i^*, Y_i^*)\}_{i=n_0+1}^{n}$ by drawing with replacement from the original treated group $\{(\mathbf{X}_i, Y_i)\}_{i=n_0+1}^{n}$.

Step 3. Generate the wild Bootstrap sample $\{(\mathbf{X}_j, Y_j^*)\}_{j=1}^{n_0}$ of the control group, where $Y_j^* = \widehat{m}(\widehat{\boldsymbol{\beta}}^\top \mathbf{X}_j) + \varepsilon_j^*$ with $\widehat{m}(\widehat{\boldsymbol{\beta}}^\top \mathbf{X}_j) = \sum_{l=1}^{n_0} K_h(\widehat{\boldsymbol{\beta}}^\top \mathbf{X}_j - \widehat{\boldsymbol{\beta}}^\top \mathbf{X}_l) Y_l / \sum_{l=1}^{n_0} K_h(\widehat{\boldsymbol{\beta}}^\top \mathbf{X}_j - \widehat{\boldsymbol{\beta}}^\top \mathbf{X}_l)$, $\varepsilon_j^* = [Y_j - \widehat{m}(\widehat{\boldsymbol{\beta}}^\top \mathbf{X}_j)]\xi_j$, and $\{\xi_j\}_{j=1}^{n_0}$ being iid. random disturbances with mean zero and unit variance. Using $\{(\mathbf{X}_j, Y_j^*)\}_{j=1}^{n_0}$ to re-estimate the index parameter as $\widehat{\beta}^*$.

Step 4. Set $\widehat{Z}_j^* = \widehat{\boldsymbol{\beta}}^{*\top} \mathbf{X}_j$ for $j = 1, \cdots, n_0$ and $\widehat{Z}_i^* = \widehat{\boldsymbol{\beta}}^{*\top} \mathbf{X}_i^*$ for $i = n_0 + 1, \cdots, n$. Then, obtain the quasi synthetic control estimator $\widehat{\Delta}^*$ as

$$\widehat{\Delta}^* = \frac{1}{n_1} \sum_{i=n_0+1}^{n} Y_i^* - \frac{1}{n_0} \sum_{j=1}^{n_0} \widehat{a}_{j,h}^* Y_j^*,$$

where $\widehat{a}_{j,h}^* = \frac{1}{n_1} \sum_{i=n_0+1}^{n} K_h(\widehat{Z}_i^* - \widehat{Z}_j^*) \left[\frac{1}{n_0} \sum_{l=1}^{n_0} K_h(\widehat{Z}_i^* - \widehat{Z}_l^*) \right]^{-1}$, which is the Bootstrap version of $\widehat{a}_{j,h}$ in (7.27).

Step 5. Repeat steps 2 to 4 a large number of times, say, B times to obtain $\{\widehat{\Delta}_b^*\}_{b=1}^{B}$. Then σ_Δ^2 can be estimated as $\widehat{\sigma}_\Delta^2 = n_1 \sum_{b=1}^{B} (\widehat{\Delta}_b^* - \widehat{\Delta})^2 / (B-1)$.

A (1-α)100% Bootstrap confidence interval for Δ can be constructed as $\widehat{\Delta} \pm z_{\alpha/2} \widehat{\sigma}_\Delta / \sqrt{n_1}$ based on the asymptotic normality of $\widehat{\Delta}$ in (7.28), where $z_{\alpha/2}$ is the $(1-\alpha/2)$th quantile of the standard normal distribution. The theoretical validity of this hybrid Bootstrap procedure can be founded in Cai et al. (2025c). That is, under the conditions imposed in the proof of (7.28), conditional on the original sample $\{(\mathbf{X}_j,\, Y_j)\}_{j=1}^{n_0}$ and $\{(\mathbf{X}_i,\, Y_i)\}_{i=n_0+1}^{n}$ and in probability, Cai et al. (2025c) proved that

$$\sqrt{n_1} \left(\widehat{\Delta}^* - \widehat{\Delta} \right) \xrightarrow{d} N\left(0, \sigma_\Delta^2\right),$$

where σ_Δ^2 is defined in (7.28).

7.2.6 Panel Data Approaches and Modified SC Methods

As pointed out by Li (2020), the synthetic control method, a powerful tool for estimating average treatment effects, is increasingly popular in fields such as statistics, economics, political science, and marketing. The SC is particularly suitable for estimating ATE with a single (or a few) treated unit(s), a fixed number of control units, and large pre and post-treatment periods (which we refer as "long panels"). Then, Li (2020) derived the asymptotic distribution of the SC and modified synthetic control (MSC) ATE estimators using projection theory. Furthermore, Li and Bell (2017) investigated the PDA and accomplished the following: (i) They relaxed some of the distributional assumptions made in Hsiao et al. (2012) and showed that the HCW method works for a much wider range of data generating processes; (ii) They derived the asymptotic distribution of HCW's average treatment effect estimator which facilitates inference; (iii) When there exists a large number of control units, they proposed using the LASSO method to select control units. Also, they showed that the LASSO method is computationally more efficient compared to conventional model selection criteria. Moreover, the LASSO method leads to more accurate out-of-sample prediction results than many commonly adopted approaches such as BIC, AIC, AICC and the leave-many-out cross validation methods. Recently, Hsiao (2025) gave an excellent review on causal and non-causal approaches for panel data, which provide the possibility to simultaneously capturing inter-individual differences and intra-individual dynamics. Compared to the cross-sectional ($T = 1$) or time series ($N = 1$) data sets, panel data possess several advantages:

1. It provides better possibility to control issues arising from observed data subject to selection on observables and/or selection on unobservables in the estimation of ATE with less restrictive assumptions; see, e.g., Bai et al. (2014), Ouyang and Peng (2015), Wan et al. (2018), Li and Bell (2017), Li (2020), and Hsiao (2025).

2. Information on individual's response to policy changes provides the possibility to identify if the differences in individual treatment effects can be considered as due to chance events (i.e., homogeneous) or due to some fundamental differences (i.e., heterogeneous), thus whether it makes sense to consider the estimation of ATE; see, e.g., Hsiao (2025).

3. Information across individuals over time not only provides the possibility of examining if there are "treatment effects," but also provides the possibility of examining whether treatment effects are evolutionary over time or stationary around a common mean;

see, e.g., Hsiao (2025).

4. It provides the possibility to blend the advantages of both the nonparametric and semiparametric approach to estimate the treatment effects with the parametric approach to identify the causal factors. More importantly, it can accommodate both stationary and nonstationary variables into the framework; see, e.g., Bai et al. (2014), Ouyang and Peng (2015), Li (2020), Cai and Li (2025), and Hsiao (2025).

Also, Hsiao (2025) discussed some limitations for the panel data approaches. The reader is referred to this review paper.

We start by introducing some notation. Let $Y_{it}(1)$ and $Y_{it}(0)$ denote unit $i's$ outcome in period t with and without treatment, respectively. The treatment effect from intervention for the ith unit at time t is defined as $\Delta_{it} = Y_{it}(1) - Y_{it}(0)$. However, we do not simultaneously observe $Y_{it}(1)$ and $Y_{it}(0)$. The observed data is in the form $Y_{it} = d_{it}Y_{it}(1) + (1 - d_{it})Y_{it}(0)$, where $d_{it} = 1$ if the ith unit is under the treatment at time t, and $d_{it} = 0$ otherwise. Therefore, the observed panel data are denoted by $\{(Y_{it}, d_{it}); 1 \leq i \leq N\}_{t=1}^{T}$. We consider the case where there is a finite number of treated and control units and the treated units are drawn from heterogeneous distributions (i.e., they are not randomly assigned). Also, the treatment time occurs at different times for different treated units. In this type of situation, it is reasonable to estimate treatment effects for each treated unit separately. Under the assumption that the treatment effects $\Delta_{it} = Y_{it}(1) - Y_{it}(0)$ follow a stationary process, we can define the ATE as $\Delta_i = \mathbb{E}(\Delta_{it})$, where the expectation is with respect to the stationary distribution of Δ_{it}. In this way, we can obtain ATE for each treated unit. To obtain ATE over all the treated units, we can average (possibly with different weights) over all treated units. Hence, we focus on the case where there is one treated unit that receives a treatment at time $T_1 + 1$. Without loss of generality, we assume that it is the first unit. We want to estimate ATE for the first unit: $\Delta_1 = \mathbb{E}(\Delta_{1t})$. The difficulty in estimating the treatment effects is that $y_{it}(0)$ is not observable for $t \geq T_1 + 1$. Specific methods for estimating $Y_{1t}(0)$ are discussed later. For now, let $\widehat{y}_{1t}(0)$ be a generic estimator of $Y_{1t}(0)$. Then, ATE is estimated by averaging over the post-treatment period,

$$\widehat{\Delta}_1 = \frac{1}{T_2} \sum_{t=T_1+1}^{T} \widehat{\Delta}_{1t},$$

where $T_2 = T - T_1$ is the post-treatment sample size.

We examine the scenario where a treatment was administered to the first unit at $t = T_1 + 1$. Thus, the remaining $N - 1$ units are control units. Hsiao et al. (2012) proposed to estimate the cross-sectional correlations between the treatment and control individuals based on the pre-treatment data. In this way, individuals that display strong correlations should share common latent factors with the treatment individual. More importantly, estimating the cross-sectional correlations rather than taking simple differences allows the influence of common latent factors to vary cross-section. To use unified notation to cover both the SC and the modified SC (MSC) methods, we add an intercept to the classical SC method, denoted by OSC. Therefore, utilizing the correlation between Y_{1t} and Y_{jt}, where $2 \leq i \leq N$, we can estimate the SC counterfactual outcome $y_{1t}(0)$ based on the following regression model:

$$Y_{1t} = \boldsymbol{\beta}^\top \mathbf{X}_t + u_{1t}, \quad 1 \leq t \leq T_1, \tag{7.29}$$

where $\mathbf{X}_t = (1, Y_{2t}, \cdots, Y_{Nt})$ is an $N \times 1$ vector of a constant (of one) and the control units' outcome variables, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_N)$ is an $N \times 1$ vector of unknown coefficients, and u_{1t} is a zero mean, finite variance idiosyncratic error term. Also, $\mathbf{X}_t$ contains the relevant covariates if there exist some relevant covariates. Then, after obtaining $\widehat{\boldsymbol{\beta}}$ based on (7.29), we can have the predicted value $\widehat{Y}_{1t}(0) = \widehat{\boldsymbol{\beta}}^\top \mathbf{X}_t$ for $T_1 + 1 \leq t \leq T$, where $\widehat{\boldsymbol{\beta}}$ is obtained based on different methods, described as follows.

Let us define $\mathbb{H}_{\text{OSC}} = \{\boldsymbol{\beta}; \boldsymbol{\beta} \in \mathbb{R}^N, \beta_j \geq 0, \text{ and } \sum_{j=1}^N \beta_j = 1\}$ and $\mathbb{H}_{\text{MSC}} = \{\boldsymbol{\beta}; \boldsymbol{\beta} \in \mathbb{R}^N, \beta_j \geq 0\}$. First, according to Abadie and Gardeazabal (2003), $\widehat{\boldsymbol{\beta}}$ should be obtained by

$$\widehat{\boldsymbol{\beta}}_{\text{OSC}} = \arg_{\boldsymbol{\beta} \in \mathbb{H}_{\text{OSC}}} \sum_{t=1}^{T_1} (Y_{1t} - \boldsymbol{\beta}^\top \mathbf{X}_t)^2.$$

To relax the convex hull constraint, by modifying the convex hull assumption, Doudchenko and Imbens (2016) considered the MSC method via obtaining $\widehat{\boldsymbol{\beta}}$ by

$$\widehat{\boldsymbol{\beta}}_{\text{MSC}} = \arg_{\boldsymbol{\beta} \in \mathbb{H}_{\text{MSC}}} \sum_{t=1}^{T_1} (Y_{1t} - \boldsymbol{\beta}^\top \mathbf{X}_t)^2.$$

To remove all constraints, Hsiao et al. (2012) proposed the ordinary least squares estimate of $\boldsymbol{\beta}$ in (7.29), termed as HCW approach or panel data approach, to obtain $\widehat{\boldsymbol{\beta}}_{\text{HCW}}$. Using the projection theory, Li (2020) derived the asymptotic distribution of the SC and modified synthetic control ATE estimators. More importantly, Bai et al. (2014) extended the HCW method to allow nonstationary case, say I(1) process, termed as BLO, and proved that,

as long as the data follows I(1), the OLS estimator of the correlation coefficient is the unique and consistent estimator that can uncover the post-treatment counterfactuals using the post-treatment observations of the control individuals, as well as explored the influence of property taxes on home prices, taking advantage of a policy experiment of property taxation in Shanghai and Chongqing. Moreover, by relaxing the linear conditional mean assumption in (7.29), Ouyang and Peng (2015) generalized the HCW method to a semiparametric setting for stationary time series as

$$Y_{1t} = \boldsymbol{\beta}_1^\top \mathbf{X}_{1t} + \beta_2(\mathbf{X}_{2t}) + u_{1t}, \quad 1 \le t \le T_1,$$

where $\mathbf{X}_t = (\mathbf{X}_{1t}, \mathbf{X}_{2t})$ and $\beta_2(\mathbf{X}_{2t})$ is an unknown function, and studied the macroeconomic effect of the 2008 Chinese economic stimulus program. For more new developments, see, for instance, Hsiao (2025). Furthermore, when there exists a large number of control units, Li and Bell (2017) proposed using the LASSO method to select control units and showed that the LASSO method is computationally more efficient compared to conventional model selection criteria. Also, Li and Bell (2017) relaxed this linear conditional mean function assumption to allow for the conditional mean function to have any unknown functional form and they used a linear project argument to show that the PDA remains valid, although it may be less efficient than estimating the conditional mean function nonparametrically when the sample size is sufficiently large. Finally, Cai and Li (2025) successfully applied the BLO and HCW approaches with LASSO for evaluating the China–US trade war (started from 2018) effects.

7.3 Estimation of QTE

7.3.1 Unconditional QTE

It is clear that under Assumption 7.2, the distribution of the potential outcomes $Y(1)$ and $Y(0)$ can be identified by

$$F_1(y) = \mathbb{E}\left[\frac{d\, I(Y(1) \le y)}{\pi(\mathbf{X})}\right], \text{ and } F_0(y) = \mathbb{E}\left[\frac{(1-d)\, I(Y_i(0) \le y)}{\{1-\pi(\mathbf{X})\}}\right]. \tag{7.30}$$

Consequently, by definition, the QTE parameter Δ_τ in (7.1) can be identified as the difference between two quantile functions:

$$\Delta_\tau = F_1^{-1}(\tau) - F_0^{-1}(\tau) = q_\tau(1) - q_\tau(0), \tag{7.31}$$

where $F_j^{-1}(\tau) = \inf\{y : F_j(y) \geq \tau\}$ for $j = 0$ and 1. Alternatively, as shown in Firpo (2007), the QTE parameter Δ_τ can be identified directly from Assumption 7.2. Specifically, Firpo (2007) proved that, under Assumption 7.2, the quantile functions of the potential outcome distributions $q_\tau(1)$ and $q_\tau(0)$ can be identified by the following moment conditions,

$$\mathbb{E}\left[\frac{d\, I(Y(1) \leq q_\tau(1))}{\pi(\mathbf{X})} - \tau\right] = 0 \text{ and } \mathbb{E}\left[\frac{(1-d)\, I(Y_i(0) \leq q_\tau(0))}{\{1-\pi(\mathbf{X})\}} - \tau\right] = 0. \tag{7.32}$$

Consequently, the identification of QTE parameter Δ_τ is a straightforward consequence from (7.32). To estimate $q_\tau(d)$ based on the observed data $\{(Y_i, d_i, \mathbf{X}_i)\}_{i=1}^n$, one can use the weighted quantile regression as in Koenker and Bassett (1978), given by

$$\widehat{q}_\tau(j) = \arg\min_q \sum_{i=1}^n \widehat{\omega}_{j,i}\, \rho_\tau(Y_i - q) \tag{7.33}$$

for $j = 0$ and 1, where the check function $\rho_\tau(\cdot)$ is defined in Section 2.1, $\widehat{\omega}_{1,i} = d_i/\widehat{\pi}(\mathbf{X}_i)$, and $\widehat{\omega}_{0,i} = (1-d_i)/\widehat{\pi}(\mathbf{X}_i)$ with $\widehat{\pi}(\mathbf{x})$ being the nonparametric power series estimator of $\pi(\mathbf{x})$ like SLE. Therefore, by plugging (7.33) into (7.31), $\widehat{\Delta}_\tau$ is obtained. Under some regularity conditions, Firpo (2007) showed that the resulting estimator $\widehat{\Delta}_\tau$ is $\sqrt{n}$-consistent and asymptotically normally distributed. Furthermore, Firpo (2007) argued that the asymptotic variance of the estimator $\widehat{\Delta}_\tau$ presented above can attain the semiparametric efficiency bound.

7.3.2 Partially Conditional QTE

A. Introduction

In many applications, researchers may be interesting in estimating the effect of a treatment or policy on outcome of interest in various sub-populations defined by the possible values of some component(s) of the pre-treatment variables $\mathbf{X}$. For example, Abrevaya et al. (2015) and Lee et al. (2017) examined the mean effect of maternal smoking during pregnancy on infant birth weights conditional on mother's age. For the partially conditional average treatment effect model, the reader is referred to the papers by Abrevaya et al. (2015), Lee et al. (2017), Cai et al. (2021) and references therein for details.

Now, the conditional version of (7.30) is

$$F_1(y \,|\, w) = \mathbb{E}\left[\frac{d\, I(Y(1) \leq y)}{\pi(\mathbf{X})}\Big|\, W = w\right], \text{ and } F_0(y \,|\, w) = \mathbb{E}\left[\frac{(1-d)\, I(Y_i(0) \leq y)}{\{1-\pi(\mathbf{X})\}}\Big|\, W = w\right], \tag{7.34}$$

where $W \in \mathbb{R}^{d_w}$ is a part of $\mathbf{X}$ ($d_w < p$) or W can be $\mathbf{X}$. Consequently, the partially conditional quantile treatment effect $\Delta_\tau(w)$, denoted by PCQTE, is defined as

$$\Delta_\tau(w) = F_1^{-1}(\tau \,|\, w) - F_0^{-1}(\tau \,|\, w) = q_{1,\tau}(w) - q_{0,\tau}(w), \tag{7.35}$$

where

$$q_{j,\tau}(w) = \arg\inf_q \mathbb{E}\left[\omega_j(\mathbf{X})\rho_\tau(Y - q)\right]$$

with $\omega_0(\mathbf{X}) = (1-d)/[1-\pi(\mathbf{X})]$ and $\omega_1(\mathbf{X}) = d/\pi(\mathbf{X})$.

B. Parametric Models

In Tang et al. (2021), a parametric form of $q_{j\tau}(w)$ was considered; like a linear quantile regression as

$$q_{j,\tau}(w) = \alpha_{j,\tau} + \boldsymbol{\beta}_{j,\tau}^\top w.$$

For this setting, Tang et al. (2021) proposed estimating $\alpha_{j,\tau}$ and $\boldsymbol{\beta}_{j,\tau}$ by the classical weighted quantile regression

$$(\widehat{\alpha}_{j,\tau}, \widehat{\boldsymbol{\beta}}_{j,\tau}) = \arg\inf_{\alpha_{j,\tau}, \boldsymbol{\beta}_{j,\tau}} \sum_{i=1}^n \widehat{\omega}_j(\mathbf{X}_i)\rho_\tau(Y_i - \alpha_{j,\tau} - \boldsymbol{\beta}_{j,\tau}^\top w), \tag{7.36}$$

where $\widehat{\omega}_j(\mathbf{X}_i)$ is a consistent estimate of $\widehat{\omega}_j(\mathbf{X})$. Then, the estimated the treatment effect $\widehat{\Delta}_\tau(w)$ has the following form

$$\widehat{\Delta}_\tau(w) = \widehat{\alpha}_{1,\tau} - \widehat{\alpha}_{0,\tau} + (\widehat{\boldsymbol{\beta}}_{1,\tau} - \widehat{\boldsymbol{\beta}}_{0,\tau})^\top w.$$

Of curse, it is easy to generalize the linear model to a nonlinear parametric quantile model.

C. Nonparametric Models

If $q_{j,\tau}(w)$ is an unknown function, this leads to the estimation of $q_{j,\tau}(w)$, via modifying (7.36), by using the locally weighted quantile regression technique as

$$\widehat{q}_{j,\tau}(w) = \arg\inf_q \sum_{i=1}^n \widehat{\omega}_j(\mathbf{X})\rho_\tau(Y_i - q)K_h(W_i - w),$$

which implies that the estimation of $\Delta_\tau(w)$ is given by

$$\widehat{\Delta}_\tau(w) = \widehat{q}_{1,\tau}(w) - \widehat{q}_{0,\tau}(w). \tag{7.37}$$

Under some regularity conditions, Cai et al. (2021) showed that the proposed estimator $\Delta_\tau(w)$ is consistent and asymptotically normally distributed. The reader is referred to the paper by Cai et al. (2021) for more details. If d_w is large, similarly, one can use a dimension reduction approach to estimate the conditional quantiles, such as single index quantile method specified as $q_{j,\tau}(w) = q_{j,\tau}(\gamma^\top w)$ as in Wu et al. (2010). Then, the MAVE estimation procedure for the single index model as in Section 2.7.2 can be used to estimating the functional $q_{j,\tau}(\cdot)$ and parameter γ by replacing the quadratic loss function as in (2.66) in Section 2.7.2 by the check function $\rho_\tau(\cdot)$. For details, the reader is referred to the paper by Wu et al. (2010).

D. An Empirical Example

In this section, the proposed parametric and nonparametric quantile regression models for treatment effect are applied to the analysis of the quantile treatment effect of maternal smoking during pregnancy on infant birth weight while allowing for arbitrary treatment effect heterogeneity conditional on the mother's age. We use the same dataset as in Tang et al. (2021) and Cai et al. (2021), collected by the North Carolina State Center Health Services. Similar to the aforementioned two papers, our focus is on the sample for the first-time pregnant white mothers with $433,558$ observations and $157,989$ observations for the blacks group. For the detailed descriptions of this dataset, please read the papers by Tang et al. (2021) and Cai et al. (2021). Since our interest is to estimate how the quantile effect of maternal smoking during pregnancy changes across different age groups of mothers, the conditional variable W is the mother's age. In addition, d denotes the treatment variable which is equal to one if the mother smokes and zero otherwise. The outcome of interest Y is the baby's birth weight measured in grams. In this example, $Y(1)$ denotes the infant birth weight for the treated (smoking) group and $Y(0)$ stands for the infant birth weight for the untreated (no-smoking) group. For whites, Figure 7.1 depicts the kernel density estimations of infant birth weight for the un-treated (solid line) and treated (dotted line) groups, respectively, with the left panel for whites and the right panel for blacks. It looks that the density estimations of infant birth weight for both treated and untreated groups are asymmetric and fat-tailed in the left side. To further confirm these findings, the sample skewness and kurtosis of infant birth weights are computed and also, a symmetry test is conducted to see if the distributions are symmetry. The results are reported in Table 7.8, which support the findings observed from Figure 7.1. Based on the above discussions, it

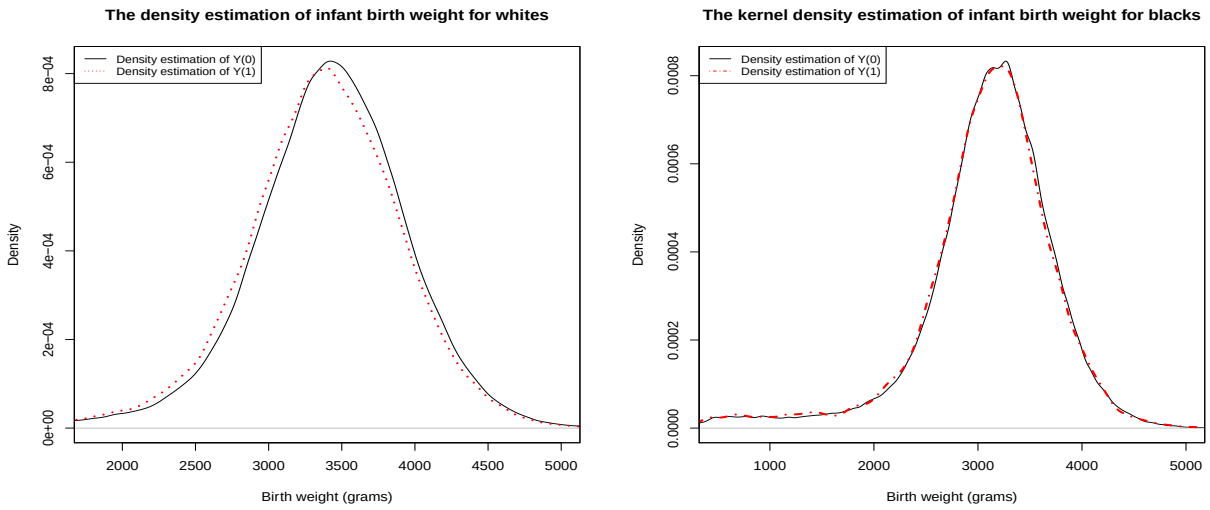

Figure 7.1: The kernel density estimation of infant birth weight for white. The solid line is for $Y(0)$ and the dotted line for $Y(1)$. The left panel is for whites and the right panel is for blacks.

gives us a strong motivation to consider the distributional effect of maternal smoking during pregnancy on infant birth weight.

Table 7.8: Descriptive statistics and symmetry testing results.

	Whites		Blacks	
Variable	$Y(0)$	$Y(1)$	$Y(0)$	$Y(1)$
Mean	3398.681	3346.848	3103.722	3082.726
Skewness	-0.846	-0.840	-1.181	-1.204
Kurtosis	5.931	5.734	6.245	6.164
Symmetry test (p-value)	0.000	0.000	0.000	0.000
Number of observations	359172	74386	146399	11590

First, Tang et al. (2021) considered the parametric model as in (7.36). A polynomial model is explored and it turns out that the following models for $q_{j,\tau}(w)$ are suitable

$$q_{j,\tau}(w) = \alpha_{j,\tau} + \beta_{j,\tau} \cdot \text{age} + \gamma_{j,\tau} \cdot \text{age}^2$$

for $j = 0$ and 1. Therefore,

$$\Delta_\tau(w) = (\alpha_{1,\tau} - \alpha_{0,\tau}) + (\beta_{1,\tau} - \beta_{0,\tau}) \cdot \text{age} + (\gamma_{1,\tau} - \gamma_{0,\tau}) \cdot \text{age}^2,$$

which is a quadratic form of age. Based on these estimated coefficients, $\widehat{\Delta}_\tau(w)$ can be computed easily and it is displayed in Figure 7.2, plotting the estimation results for the

partially conditional quantile treatment effect, $\widehat{\Delta}_\tau(w)$, at three quantile levels $\tau = 0.1$, $\tau = 0.25$, and $\tau = 0.5$, respectively. For example,

$$\widehat{\Delta}_{0.10}(w) = -57.222 - 8.451 \cdot \text{age} + 0.067 \cdot \text{age}^2.$$

However, since $\widehat{\gamma}_{1,\tau} - \widehat{\gamma}_{0,\tau} = 0.067$ for $\tau = 0.10$ (-0.043 for $\tau = 0.25$ and -0.025 for $\tau = 0.50$) is too small, the linear term in the estimated curve $\widehat{\Delta}_\tau(w)$ dominates the whole curve in the range of $(20, 30)$ (look like a linear), which can be seen in Figure 7.2. From Figure 7.2, one

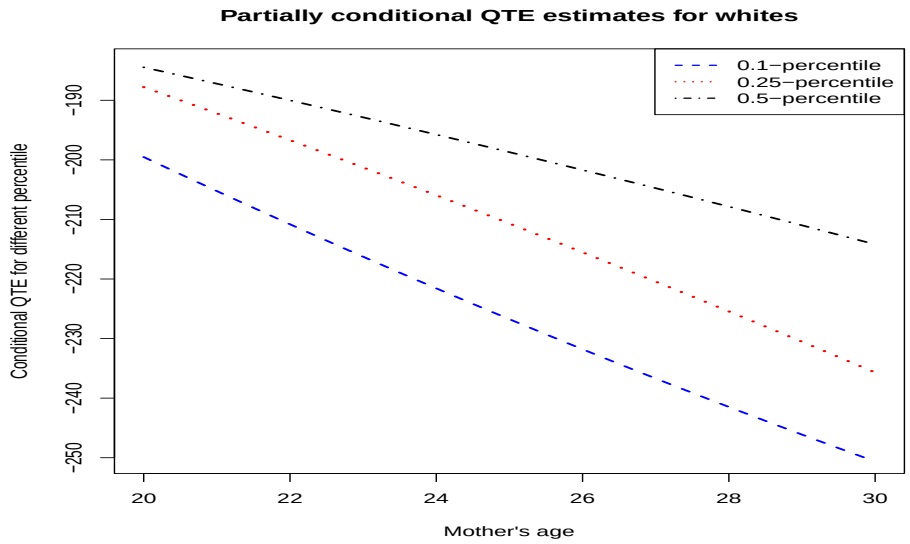

Figure 7.2: Parametric estimation results for the partially conditional quantile treatment effects

can see clearly that the results exhibit a number of striking features. First, one can observe that there is a significant negative effect of smoking on infant birth weight across all ages and quantile levels considered. Second, the estimation results show substantial heterogeneity across different ages. For example, the estimated effect ranges from about -200g to -250g at the quantile level $\tau = 0.1$ as the mother's age increases from 20 to 30 and similar pattern can be observed for other quantile levels considered. Another interesting feature of the results is that for a given age, the numerical values of the quantile treatment effect point estimates at lower quantiles are bigger than those at the higher quantiles, and for a given quantile level, the estimated quantile treatment effects become stronger (more negative) at higher ages. These findings are new in the literature.

Second, Cai et al. (2021) used the nonparametric method as in (7.37) to estimate $\Delta_\tau(w)$ using the same dataset. To do so, it needs to estimate the unknown propensity score function

$\pi(\mathbf{x})$. For this purpose, Cai et al. (2021) suggested using a logistic model to estimate the propensity score function $\pi(\mathbf{x})$. The explanatory variables used in the logistic model consist of all the elements of $\mathbf{X}$, the square of the mother's age, and the interaction terms between the mother's age and all other elements of X. Finally, the partially conditional QTE is estimated for mothers between 20 and 30 years of age, for both whites and blacks. Figure 7.3 plots the estimated PCQTEs for whites (the left panel) and blacks (the right panel) for three quantile levels $\tau = 0.10$, $\tau = 0.25$ and $\tau = 0.50$, respectively, together with the estimated unconditional 0.5-QTEs and their 95% confidence intervals. It can be observed that, as the mother's age increases, the PCQTEs of maternal smoking on infant birth weight decrease quickly for whites, while the PCQTEs for blacks decrease very slowly, compared to the width of the 95% confidence intervals of the corresponding unconditional 0.5-QTEs. This implies that $\Delta_\tau(w)$ for blacks is basically constant. Therefore, this motivates us to consider the testing issues to see if $\Delta_\tau(w)$ changes really over age, which will be discussed in detail in the next section.

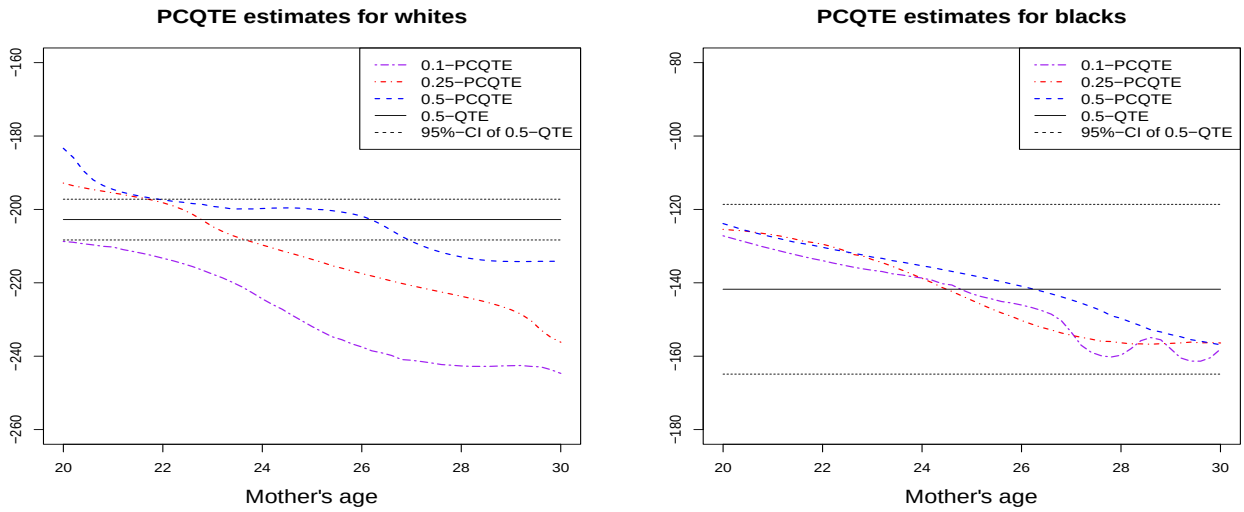

Figure 7.3: Estimated PCQTEs for whites (the left panel) and blacks (th right panel) for three quantile levels $\tau = 0.10$, $\tau = 0.25$ and $\tau = 0.50$, respectively, together with the estimated unconditional 0.5-QTEs and their 95% confidence intervals.

7.3.3 Nonparametric Tests of Heterogeneity

From (7.35), if $\Delta_\tau(w) > 0$ for all τ and w, $q_{1,\tau}(w) > q_{0,\tau}(w)$, which implies that $F_1(y \,|\, w) < F_0(y \,|\, w)$ for all y and w. This is the so-called stochastic dominance; see, for example, the book by Whang (2019) for a general method for testing inequality restrictions

and Cai et al. (2025). As argued in Cai et al. (2025), of interest is to consider the specification test as

$$H_0 : \Delta_\tau(w) = \Delta_\tau \quad \text{for all } w \quad \text{versus} \quad H_a : \Delta_\tau(w) \neq \Delta_\tau \quad \text{for some } w, \tag{7.38}$$

where Δ_τ is a constant. Under the null hypothesis, the partially conditional quantile effect of the treatment is a constant and under the alternative, the PCQTE varies across different sub-populations defined by W_i. In order to test whether the hypothesis testing problem formulated in (7.38) holds, a test statistic is constructed based on the Cramér–von Mises criterion as follows

$$J_{n,\tau} = \int \left(\widehat{\Delta}_\tau(w) - \widehat{\Delta}_\tau\right)^2 J_0(w)dw, \tag{7.39}$$

where $J_0(w)$ is a weighting function, $\widehat{\Delta}_\tau(w)$ is given by (7.37) and $\widehat{\Delta}_\tau = \sum_{i=1}^n \widehat{\Delta}_\tau(W_i)/n$, which also can be the estimate based on (7.31). It is clear that rejecting H_0 is and only if $J_{n,\tau}$ for given τ is large.

Under some regularity conditions, Cai et al. (2025) studied the asymptotic behaviors of the proposed test statistic J_n in (7.38) under both the null and alternative hypotheses. Due to the complicated expression of the asymptotic variance of J_n, Cai et al. (2025) suggested a nonparametric Bootstrap procedure to approximate the finite-sample null distribution of the proposed test. Then, they showed that the asymptotic validity of the proposed Bootstrap test is theoretically justified.

Remark 7.2: *If W is taken to be $\mathbf{X}$ in* (7.38)*, then the hypothesis testing problem in* (7.38) *collapses into testing whether the conditional QTE is a constant for all values of the covariates. Unlike our setting, Crump et al. (2008) tested whether the conditional ATE is a constant or zero for all values of the covariates. Note that even if the conditional ATE is equal to a constant, the conditional QTE may not be a constant. Consequently, our paper complements and extends Crump et al. (2008) by testing whether there exists treatment effect heterogeneity across different subpopulations defined by a subset of the covariates.*

Remark 7.3: *The hypothesis in* (7.38) *can be generalized to*

$$H_0 : \Delta_\tau(w) = \Delta_{0,\tau}(w, \theta) \quad \text{for all } w \quad \text{versus} \quad H_a : \Delta_\tau(w) \neq \Delta_{0,\tau}(w, \theta) \quad \text{for all } w,$$

where $\Delta_{0,\tau}(w, \theta)$ is a known function without unknown parameter θ, say a linear function of w. Then, the test statistic in (7.39) *should be modified to be*

$$J_{n,\tau}(\theta) = \int \left(\widehat{\Delta}_\tau(w) - \Delta_{0,\tau}(w, \widehat{\theta})\right)^2 J_0(w)dw,$$

where $\widehat{\theta}$ is a consistent estimate of θ under H_0.

Finally, let us continue the empirical example studied in the previous section (section 7.3.2D). Now, Cai et al. (2025) used the above testing procedure to test whether the partially conditional QTE changes over mother's age with the testing results depicted in Table 7.9, which shows clearly that one should reject the null hypothesis for whites for all quantiles considered at 5%. This means that PCQTEs do change over mother's age for all quantile levels considered at the significance level $\alpha = 5\%$ for whites. However, for blacks, the change in the PCQTE over mother's age is slight and statistically insignificant for all quantile levels. These testing results support the empirical findings in Cai et al. (2021). As pointed out by Yang et al. (2014), this phenomenon may occur through various interconnected mechanisms. Blacks are exposed to an environment with many risk factors for smoking, such as job loss and economic hardship. Finally, we use the test statistic proposed in Cai et al. (2025) to test whether the partially conditional QTE changes over mother's age for $\tau \in [0.1, 0.9]$. The testing results are displayed in the last row in Table 7.9. The same conclusion can be made.

Table 7.9: Test results for testing if PCQTE function changes over mother's age.

Quantile level	Bootstrap p-values	
τ	Whites	Blacks
0.10	0.022	0.573
0.25	0.002	0.307
0.50	0.033	0.151
0.75	0.035	0.474
0.90	0.042	0.326
[0.1, 0.9]	0.044	0.357

7.4 QTE for Panel Data

7.4.1 Model Setup

Similar to Section 7.2.6, we consider how to estimate the QTE for panel data in this section. For this purpose, we still try to use the same notation as in Section 7.2.6. The setting of panel data is the same as that in HCW. Suppose that we have a panel data set $\{(y_{it}, \mathbf{Z}_t); 1 \leq i \leq N, 1 \leq t \leq T\}$, where $\mathbf{Z}_t \in \mathbb{R}^{d_z}$ is a vector of covariates[3]. For the first

[3]For example, in Cai et al. (2026), it includes 3 macroeconomic variables such as the monthly CPI growth rate, the monthly M1 growth rate, and the monthly M2 growth rate.

unit, a treatment is implemented from $t = T_1 + 1$ to $t = T$ and no treatment occurs before $t = T_1 + 1$. The remaining $N - 1$ units as a control group remain untreated throughout the time. Denote $T_2 = T - T_1$. For simplicity, let $\lim_{T\to\infty} T_2/T_1 = c$ throughout the paper, where $c > 0$ is a constant, so that $\lim_{T\to\infty} T_1/T = \lambda$ and $\lim_{T\to\infty} T_2/T = 1 - \lambda$ with $\lambda = 1/(1 + c)$. To ease notation, the panel data are divided into four parts as follows:

$$\begin{array}{c|c} \mathbf{Y}_1 & \mathbf{X}_1 \\ \hline \mathbf{Y}_2 & \mathbf{X}_2 \end{array},$$

where $\mathbf{Y}_1 = \{y_{1t}, t = 1, \cdots, T_1\}$ represents the first unit from $t = 1$ to $t = T_1$, $\mathbf{X}_1 = \{(y_{it}, \mathbf{Z}_t); i = 2, \cdots, N, \text{ and } t = 1, \cdots, T_1\}$ stands for the remaining units from $t = 1$ to $t = T_1$, $Y_2 = \{y_{1t}, t = T_1+1, \cdots, T\}$ is the information for the first unit from $t = T_1+1$ to $t = T$, and $\mathbf{X}_2 = \{(y_{it}, \mathbf{Z}_t); i = 2, \cdots, N, \text{ and } t = T_1 + 1, \cdots, T\}$ is for the remaining units from $t = T_1 + 1$ to $t = T$. Let $\mathbf{Y}_2^0 = \{y_{1t}^0, t = T_1 + 1, \cdots, T\}$ be the counterfactual outcome of $\mathbf{Y}_2$. For convenience, the observed outcome Y_2 is also denoted as $\mathbf{Y}_2^1 = \{y_{1t}^1, t = T_1 + 1, \cdots, T\}$. Then, the QTE for the first unit after $t = T_1$ is defined as

$$\Delta_\tau = q_{1\tau}^1 - q_{1\tau}^0, \tag{7.40}$$

where is similar to (7.1), where $q_{1\tau}^j$ is the τth quantile of $F_j(y) = P(y_{1t}^j \le y)$ for $j = 0$ and 1 and $\tau \in (0, 1)$. The next section describes details on how to estimate Δ_τ.

7.4.2 Inference Procedures

To estimate the QTE, Δ_τ, it suffices to estimate $q_{1\tau}^1$ and $q_{1\tau}^0$, respectively. Since the outcome for the first unit under treatment with $t > T_1$ is observable, the sample quantile of Y_2 is simply used, denoted as $\widehat{q}_{1\tau}^1$, to estimate $q_{1\tau}^1$. The difficulty in estimating QTE is due to the fact that Y_2^0 is not observable, so that it is not straightforward to estimate $q_{1\tau}^0$. To estimate the counterfactual quantile for the first unit, different from the factor augmented idea in HCW, a new method is proposed by utilizing the relationship between the conditional and unconditional CDFs, described as follows.

To be specific, $q_{1\tau}^0$ is written as

$$q_{1\tau}^0 = \inf\left\{y : F_{Y_2^0}(y) \ge \tau\right\} = \inf\left\{y : \mathbb{E}[F_{Y_2^0|\mathbf{X}_2}(y|\mathbf{X}_{2t})] \ge \tau\right\},$$

where $F_{Y_2^0}(\cdot)$ is the CDF of y_{1t}^0 for $t > T_1$ and $F_{Y_2^0|\mathbf{X}_2}(\cdot|\cdot)$ denotes the conditional CDF of y_{1t}^0 given $\mathbf{X}_2$ for $t > T_1$, which leads to an estimator of $q_{1\tau}^0$ as

$$\overline{q}_{1\tau}^0 = \inf\left\{y : \frac{1}{T_2}\sum_{t=T_1+1}^{T} F_{Y_2^0|X_2}(y|\mathbf{X}_{2t}) \ge \tau\right\}. \tag{7.41}$$

Generally speaking, the conditional CDF $F_{Y_2^0|\mathbf{X}_2}(y|\mathbf{x})$ is unknown, so that the above estimator of $q_{1\tau}^0$ is infeasible. To get a feasible estimate of $q_{1\tau}^0$ from the observed data, it needs to estimate $F_{Y_2^0|\mathbf{X}_2}(y|\mathbf{x})$ first. To this end, the following identification assumption is needed, which indeed, is similar to the assumption for the ATE setting imposed in HCW, see Assumption (7.29), and the aforementioned references and the quantile setting as in Callaway et al. (2018).

Assumption 7.3: *(Identification Condition) The structures of the conditional CDFs of* $Y_1|\mathbf{X}_1$ *and* $Y_2^0|\mathbf{X}_2$ *are the same; that is,* $F_{Y_1|\mathbf{X}_1}(\cdot|\cdot) = F_{Y_2^0|\mathbf{X}_2}(\cdot|\cdot) \equiv F(\cdot|\cdot)$.

Assumption 7.3 postulates some kind of structure invariance, which ensures that given the outcomes of the control group and the covariates, the conditional distribution of the potential outcome of the treated unit without treatment remains the same before and after the treatment. With this identification assumption, it is then possible to estimate the counterfactual conditional CDF in the treated group by the observed data before treatment. This assumption corresponds to Assumption 1 in Rothe (2010) for a nonparametric structural model. Also, Hsu et al. (2022) adopted the same kind of assumption (see their Assumption 2.3) in the counterfactual treatment effects settings.

Clearly, with the identification condition in Assumption 1, a kernel method such as Nadaraya-Watson estimation method or other procedures, can be used to estimate $F(y|\mathbf{x})$ if $d_x = N - 1 + d_z$ is not very large. Specifically, $F(y|\mathbf{x})$ can be estimated using the observed data before treatment as follows

$$\tilde{F}(y|\mathbf{x}) = \frac{\sum_{t=1}^{T_1} I\left(Y_{1t} \leq y\right) K_h\left(\mathbf{X}_{1t} - x\right)}{\sum_{t=1}^{T_1} K_h\left(\mathbf{X}_{1t} - x\right)}, \tag{7.42}$$

where $K_h(\mathbf{X}_{1t} - x) = h^{-d_x} K((\mathbf{X}_{1t} - x)/h)$, $K(\cdot)$ is a higher-order kernel function as defined in Assumption 4[4], and h is bandwidth. Then, we plug the estimated conditional CDF into (7.41) to obtain

$$\tilde{q}_{1\tau}^0 = \inf\left\{y : \frac{1}{T_2}\sum_{t=T_1+1}^{T} \tilde{F}_{Y_2^0|X_2}(y|\mathbf{X}_{2t}) \geq \tau\right\} = \inf\left\{y : \tilde{F}_{Y_2^0}(y) \geq \tau\right\},$$

where $\tilde{F}_{Y_2^0}(y) \equiv \frac{1}{T_2}\sum_{t=T_1+1}^{T} \tilde{F}_{Y_2^0|X_2}(y|\mathbf{X}_{2t})$. Note that when high-order kernels are used in (7.42), $\tilde{F}_{Y_2^0}(y)$ could be non-monotonic or take values outside the $[0, 1]$ interval. One can use

[4]See Gasser et al. (1985) for details on the definition of higher-order kernel.

the re-weighting method in Rothe (2010) or the monotonization method in Hsu et al. (2022) to turn $\tilde{F}_{Y_2^0}(y)$ into a monotonically nondecreasing CDF. Here, we follow Hsu et al. (2022) and let

$$\widehat{F}_{Y_2^0}(y) = \sup_{u \le y} \tilde{F}_{Y_2^0}(u) \left[\sup_{-\infty < u < \infty} \tilde{F}_{Y_2^0}(u) \right]^{-1}.$$

Obviously, $\widehat{F}_{Y_2^0}(y)$ is a CDF with probability one. Then, the estimator of the QTE for the first unit, Δ_τ in (7.40), is given by

$$\widehat{\Delta}_\tau = \widehat{q}^1_{1\tau} - \widehat{q}^0_{1\tau},$$

where $\widehat{q}^0_{1\tau} = \inf \left\{ y : \widehat{F}_{Y_2^0}(y) \ge \tau \right\}$.

Under some regularity conditions[5], Cai et al. (2026) derived the following asymptotic normality

$$\sqrt{T_2} \left(\widehat{\Delta}_\tau - \Delta_\tau \right) \xrightarrow{d} N\left(0, \sigma_\tau^2\right), \tag{7.43}$$

where σ_τ^2 is the asymptotic variance, given by

$$\sigma_\tau^2 = \sum_{t=1}^{\infty} \left[\mathrm{Cov}(\eta_1, \eta_{t+1}) + \mathrm{Cov}(\xi_1, \xi_{t+1}) \right],$$

where $\{\eta_t\}$ and ξ_t are the quantile residuals; see, e.g., Cai et al. (2026) for details on the definition of $\widehat{\xi}_t$ and $\widehat{\eta}_t$. Consequently, this asymptotic result in (7.43) implies that $\widehat{\Delta}_\tau = \Delta_\tau + O_p(T_2^{-1/2})$ so that it is consistent as $T_2 \to \infty$. Furthermore, it shows that, although the kernel method is used to estimate the conditional CDF, the proposed QTE estimators still achieve the $\sqrt{T_2}$ convergence rate. Also, (7.43) demonstrates clearly that it would be easy to construct $(1-\alpha)100\%$ confidence interval (CI) for Δ_τ for given τ as $\widehat{\Delta}_\tau \pm z_{\alpha/2}/\sqrt{T_2}\sigma_\tau$ if σ_τ^2 would be known, where $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$ is the critical value. One way to estimate consistently σ_τ^2 is to employ the heteroskedasticity and autocorrelation consistent estimation of Newey and West (1987). However, due to the complicated structure of σ_τ^2, it might not be easy to obtain a consistent estimate of σ_τ^2 by using a Bootstrap approach instead, as suggested by Cai et al. (2026); see, for example, Cai et al. (2026) for details. Finally, as considered in Cai et al. (2026), this method can be extended to the high-dimensional case that d_x is large, by using the distributional index model as in (2.70) in Section 2.7.4. Then, the conditional quantile becomes the single index quantile model as in (2.71) and the approximation procedure in (2.72) can be used to obtain $\tilde{F}(y|\mathbf{x})$ instead of using (7.42).

[5]See, e.g., Cai et al. (2026) for details on the assumptions

7.4.3 An Empirical Application

This section is devoted to estimating the impact of introducing index futures trading on spot stock volatility (VIX), which is an important policy issue but still in a big debate, by employing the QTE approach proposed in Section 7.3. Commonly, the ATE is used for this regard, as addressed in Chen et al. (2013). But, it would be distorted by the fact that VIX is usually asymmetric and heavily tailed. To gauge this phenomenon, let us look at the stock VIX of the monthly VIX of the stock market in China from January 2002 to February 2021, displayed in Figure 7.4 for the estimated density of the pre-treatment, post-treatment and

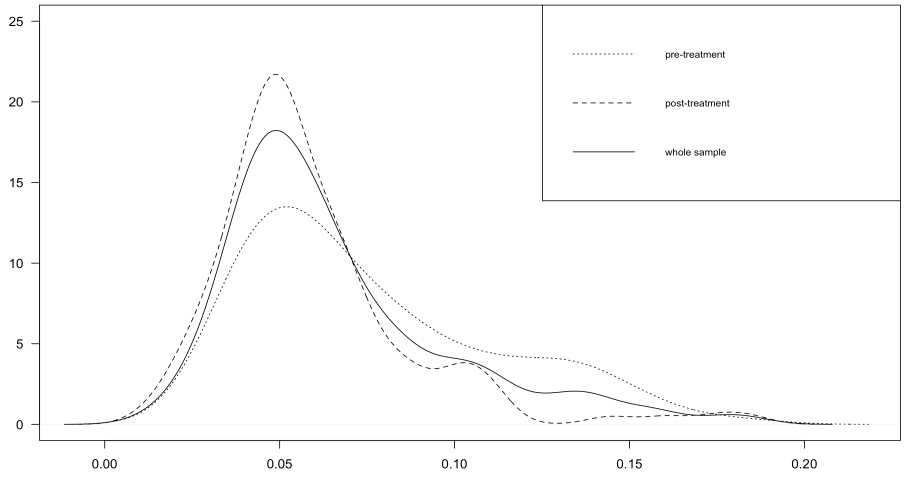

Figure 7.4: The plot of the estimated density for the pre-treatment, post-treatment and whole sample VIX of CSI 300 index.

whole sample VIX of China Securities Index (CSI) 300 index[6]

A. Data and Descriptive Statistics

In this section, our proposed method is illustrated by estimating the QTEs of introducing CSI 300 index futures trading on the spot price volatility of the Chinese stock market and its volatility, respectively. As part of financial reform, the CSI 300 index futures contracts were formally introduced by the China Financial Futures Exchange on April 16, 2010. Since then, China opens her own futures market. Whether the introduction of the futures trading

[6]The CSI 300 is a capitalization-weighted stock market index designed to replicate the performance of the top 300 stocks traded on the Shanghai Stock Exchange and the Shenzhen Stock Exchange.

has a positive impact on the stock market in China or not is a controversial issue in finance literature. Some criticized that the introduction of the index futures trading may shake the spot market due to the excessive speculation while others believed that the index futures market can improve the speed and quality of the information flows and make the financial markets more complete. To identify the impact of introducing CSI 300 index futures trading on the volatility of the stock market in China, similar to Chen et al. (2013) and Cai et al. (2026), we take the geographical tie and trade relations into account, such that 13 major international market indices are selected as the control units, which include the Hang Seng Index (HSI), the Hang Seng China Affiliated Corporation Index (HSCCI), Korean Composite Stock Price Index, Japanese Nikkei 225 Index, Singaporean Strait Times Index, Taiwanese Composite Index, the FTSE 100 Index in KU, the S&P 500 Index in USA, Franch CAC 40 Index, German Frankfurt DAX Index, Brazilian Bovespa Index, Canadian S&P/TSX Composite Index, and Australian All Ordinaries Index. In addition, 3 macroeconomic variables are also included: the monthly CPI growth rate, the monthly M1 growth rate, and the monthly M2 growth rate. The time period of the data is from January 2002 to February 2021. All the market indices are collected from the Resset Financial Research Database[7] and the macroeconomic data are from the CEIC Database[8]. The monthly stock log-returns[9] of the 14 market indices are calculated by the difference of the log-returns between the last day and the first day in a month. Therefore, the total sample size $T = 230$. The descriptive statistics for the log-returns of the 14 indices and the 3 macroeconomic variables are reported in Table 7.10, from which one can see that for most of market indices, their distribution of the log-return is almost symmetric.

Following Chen et al. (2013) and Cai et al. (2026), the monthly stock volatilities of the 14 market indices are calculated as the standard deviation of daily index returns multiplied by the square root of the number of trading days in that month. The descriptive statistics for the volatilities of the 14 indices are reported in Table 7.11, from which it can be observed that the distributions of volatility for all 14 market indices are asymmetric (see Column 8 in Table 7.11) and heavy-tailed (see Column 7 in Table 7.11), which are strongly supported by observing Figure 7.4.

[7]http://www.resset.cn/endatabases

[8]https://www.ceicdata.com

[9]The monthly log-return is computed as $r_t = \log(p_t) - \log(p_{t-1})$, where p_t is the closing price at the last day of the t month, p_{t-1} is the closing price at the first day of the t month.

Table 7.10: Descriptive Statistics of Monthly Return

Index	Mean	Std. Dev.	Median	Min.	Max.	Kurt.	Skew.
CSI 300	-0.052	1.056	-0.064	-3.807	4.544	6.350	0.424
HSI	-0.088	0.867	-0.065	-2.535	2.303	3.295	0.106
HSCCI	-0.113	1.063	-0.100	-3.426	2.993	3.317	-0.022
Korea	-0.051	0.827	-0.065	-2.730	2.498	3.982	-0.189
Japan	-0.077	0.949	-0.097	-2.650	3.146	4.128	0.464
Singapore	-0.023	0.800	-0.090	-2.340	5.368	13.436	1.760
Taiwan	0.037	0.737	-0.037	-2.452	2.921	4.171	0.211
UK	-0.160	0.813	-0.254	-1.686	4.224	7.182	1.325
US	-0.085	0.821	-0.148	-2.656	4.673	9.084	1.371
France	-0.075	0.927	-0.106	-2.262	4.199	5.728	0.972
Germany	-0.106	0.943	-0.122	-2.423	4.337	6.298	1.023
Brazil	-0.141	1.083	-0.280	-3.380	3.600	4.030	0.430
Canada	0.009	0.767	-0.006	-1.513	5.016	12.538	2.042
Australia	-0.012	0.603	-0.012	-1.933	2.187	4.533	0.190
CPI growth rate	0.001	0.028	0.002	-0.095	0.058	4.106	-0.615
M1 growth rate	0.134	0.075	0.127	0.000	0.390	3.053	0.554
M2 growth rate	0.149	0.047	0.142	0.080	0.297	3.861	0.800

The monthly stock log-return is calculated as 100 multiplied by the difference of the log-returns between the last day and the first day in a month. CPI, M1 and M2 growth rates denote the monthly growth rate compared to those in the same month of the previous year.

Table 7.11: Descriptive Statistics of Monthly Volatility

Index	Mean	Std. Dev.	Median	Min.	Max.	Kurt.	Skew.
CSI 300	0.066	0.033	0.057	0.013	0.184	4.568	1.334
HSI	0.056	0.032	0.047	0.020	0.325	26.522	3.777
HSCCI	0.068	0.034	0.060	0.026	0.317	15.987	2.712
Korea	0.053	0.030	0.046	0.017	0.249	12.208	2.381
Japan	0.059	0.031	0.053	0.019	0.318	24.820	3.284
Singapore	0.044	0.030	0.037	0.013	0.256	18.776	3.284
Taiwan	0.049	0.025	0.042	0.016	0.142	4.812	1.401
UK	0.047	0.029	0.039	0.012	0.231	13.735	2.711
USA	0.046	0.034	0.037	0.011	0.276	18.382	3.287
France	0.058	0.034	0.049	0.017	0.248	10.028	2.240
Germany	0.059	0.034	0.050	0.018	0.239	9.239	2.165
Brazil	0.073	0.037	0.065	0.028	0.360	28.565	4.116
Canada	0.040	0.031	0.032	0.011	0.290	29.700	4.398
Australia	0.040	0.024	0.035	0.012	0.222	20.896	3.370

The monthly stock index volatility is calculated as the standard deviation of daily index returns multiplied by the square root of the number of trading days in that month.

For the sample period from January 2002 to June 2011, Chen et al. (2013) employed the panel data policy evaluation approach by HCW to construct counterfactuals of the spot market volatility, mainly based on the correlations between China and international stock markets, and draw the conclusion that the introduction of index futures trading can significantly reduce the volatility of the Chinese stock market. However, different from Chen et al. (2013), following Cai et al. (2026), we consider the QTEs of the index futures trading on both the log-return and volatility of the Chinese stock market with similar datasets but with different time periods. According to the introduction date of the CSI 300 index futures, the whole time period is divided into two sections: the pre-treatment period from January 2002 to April 2010 which consists of the sample size of $T_1 = 100$ observations and the post-treatment period from May 2010 to February 2021 which consists of the sample size of $T_2 = 130$ observations so that $\lambda = 10/23$. Finally, considering the sample size and the number of the control units, the conditional CDF in this application is estimated by the quantile regression method since $d_x = 16$, which is moderate.

B. QTE of Futures Trading on Stock Returns

First, Cai et al. (2026) studied the QTE of introducing the index futures trading on the monthly log-return of the stock market in China. Then, they implemented the method proposed in this section to calculate the estimated QTEs of CSI 300 index futures trading on the log-return (Y_{1t}) of the Chinese stock market. Figure 7.5 presents the estimated QTEs of the CSI 300 index futures trading on the log-return of the Chinese stock market, together with 95% CI (the red shaded area) for each quantile based on the blockwise Bootstrap with $B = 1000$ replications, as proposed in Cai et al. (2026). Also, the ATE $\widehat{\Delta}_1$, calculated by the HCW's approach is plotted by the horizontal (blue) line, together with its 95% CI (the blue shaded area).

From Figure 7.5, it can be seen that first, the 95% CI for $\widehat{\Delta}_1$ contains basically zero, which implies that Δ_1 should be zero so that the average return is not affected by the introduction of the CSI 300 index futures trading, which, as expected, is not surprising. Second, the estimated QTEs changes (decreases almost linearly) with τ and are significantly positive at the lower quantiles (about 0.008 at the 10% quantile), while significantly negative at the higher quantiles (about -0.01 at the 90% quantile), which indicates that the introduction of the CSI 300 index futures trading has different impacts on the log-return of the Chinese stock

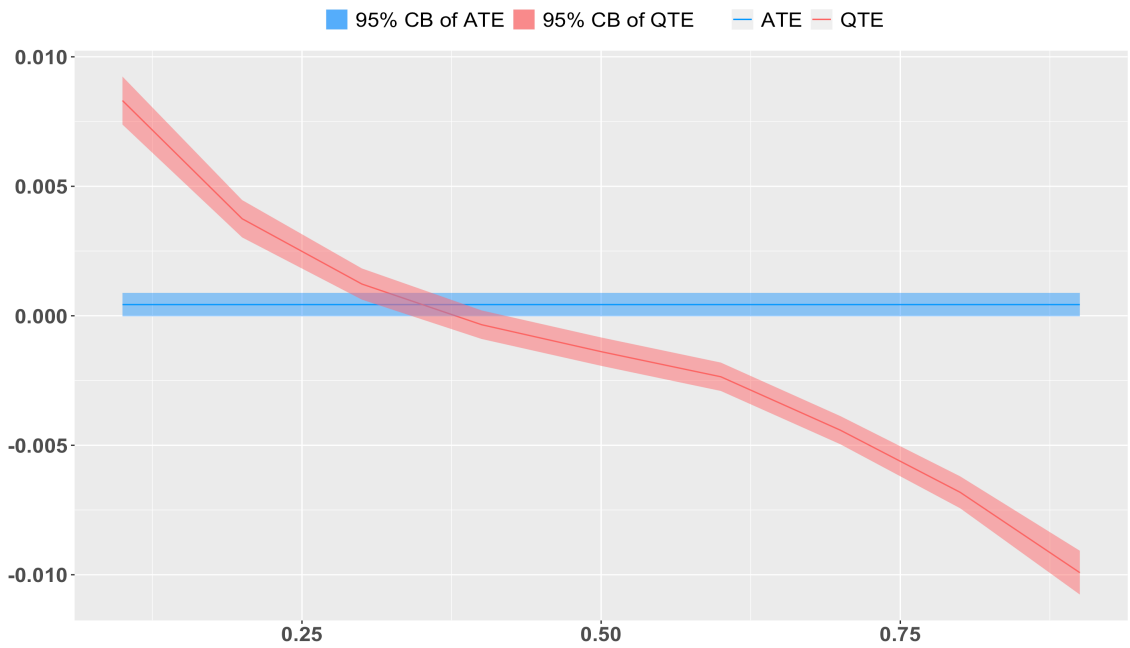

Figure 7.5: The plot of the estimated QTE is in the red line, $\widehat{\Delta}_\tau$ versus τ, together with its 95% CI (the shaded area) based on the blockwise Bootstrap proposed in Cai et al. (2026). The horizontal (blue) line is $\widehat{\Delta}_1$, the ATE calculated by the HCW's approach.

market at different quantiles. Indeed, a quantile of log-return can be used to characterize the risk of log-return, as argued by Xiao and Koenker (2009). For example, a lower quantile corresponds to the Value-at-Risk (VaR), a well-known downside risk measure in finance literature. The positive QTEs at the lower quantiles indicate that the introduction of the CSI 300 futures trading can reduce the VaR by making the negative log-return less negative. Meanwhile, the negative QTEs at the higher quantiles suggest that the introduction of the CSI 300 futures trading can also reduce the VaR by making the positive log-return less positive. In conclusion, similar to Chen et al. (2013), introducing the CSI 300 futures trading makes the stock market in China more stable in terms of the VaR. However, Chen et al. (2013) did not find such an asymmetric effect and then it is hard for them to empirically interpret why introducing the futures trading can stabilize the spot stock market.

Furthermore, Cai et al. (2026) also investigated the QTEs of introducing the index futures trading on the volatility of the stock market in China. Different from the previous section, Y_{1t} in this section is volatility instead of log-return. As discussed earlier, the QTE can be used to evaluate the impact of volatility of volatility (VVIX), see, for example, Huang et al. (2019) for details on its definition and its estimation procedures as well as financial

implications. Next, we implement the proposed method to calculate the QTEs of CSI 300 index futures trading on the volatility of the Chinese stock market. Figure 7.6 depicts the estimated QTEs for the CSI 300 index futures trading on the volatility of the Chinese stock market, together with 95% CI (the red shaded area) for each quantile, which is obtained via the blockwise Bootstrap with $B = 1000$ replications. Also, for a comparison purpose, the ATE $\widehat{\Delta}_1$ calculated by the HCW's approach is plotted by the horizontal (blue) line, together with its 95% CI (the blue shaded area).

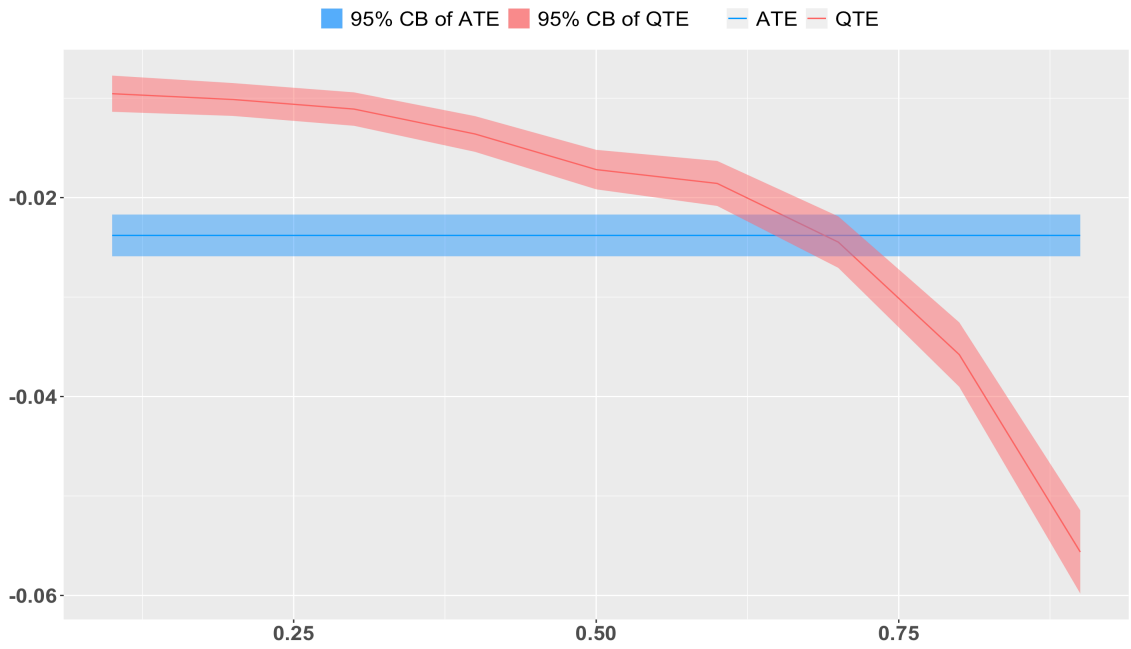

Figure 7.6: The plot of the estimated QTE is in the red line, $\widehat{\Delta}_\tau$ versus τ, together with its 95% CI (the red shaded area) based on the blockwise Bootstrap proposed in Cai et al. (2026). The horizontal (blue) line is $\widehat{\Delta}_1$, the ATE calculated by the HCW's approach.

From Figure 7.6, first, it is not surprising to see that the estimated median effect $\widehat{\Delta}_{\tau=1/2}$ is not the same as the ATE, $\widehat{\Delta}_1$, because the distribution of volatility is asymmetric and heavily tailed, as seen in Figure 7.4. Second, it is clear to see that the estimated QTEs decreases in two phases (two piecewise linear) and are significantly negative and the volatility at the higher quantile is much negative than that at the lower quantile, which decreases about 0.054 at the 90% quantile compared to only 0.01 at the 10% quantile. The ATE by the HCW's approach gives a general effect of the CSI 300 index futures trading on the volatility variation while the QTE by the proposed method can offer more details for the effect of the

CSI 300 index futures trading on the variation of the volatility. Overall, introducing the CSI 300 futures trading can reduce the volatility of the stock market and the higher the volatility is, the more significant the treatment effect demonstrates. The results are consistent with the findings earlier in terms of the Value-at-Risk and both suggest that introducing the CSI 300 futures market can make the stock market more stable in China.

7.5 Estimating Counterfactual Distribution Functions

7.5.1 Estimation Approach

We consider using the weighting method to estimate $F_d(\cdot)$, the CDF of $Y(d)$ for $d = 0$ and 1. Let $w_i \geq 0$ be the weight associated with the observation $(Y_i, d_i, \mathbf{X}_i)$ for $i = 1, \cdots, n$, and define the weight vector $\boldsymbol{w} = (w_1, \cdots, w_n)^\top$. Further, define $w_{di} = 1(d_i = d)w_i$ for $d = 0$ and 1. We estimate $F_d(y)$ using the following weighted empirical CDF

$$\widehat{F}_d(y) = \frac{1}{n}\sum_{i=1}^{n} w_{di} 1\left(Y_i \leq y\right) \tag{7.44}$$

for $d = 0$ and 1. Conventionally, the weights are obtained by first modeling the propensity score function $\pi(\mathbf{x})$ and then inverting the estimated propensity scores, which are called the inverse propensity score weights. More specifically, the IPWs are defined as $w_{0i}^{\text{IPW}} = (1 - d_i)/[1 - \pi(\mathbf{X}_i)]$ and $w_{1i}^{\text{IPW}} = d_i/\pi(\mathbf{X}_i)$. Despite being widely used, the IPW approach suffers from some problems in practice as discussed earlier.

First, we consider the estimation error of $\widehat{F}_d(y)$ defined in (7.44). To this end, let $F_d(\cdot|\mathbf{x})$ denote the conditional CDF of $Y(d)$ given $\mathbf{X} = \mathbf{x}$ for $d = 0$ and 1. Given the weights w_{di}, the estimation error of $\widehat{F}_d(y)$ can be decomposed as

$$\begin{aligned}
&\widehat{F}_d(y) - F_d(y) = \frac{1}{n}\sum_{i=1}^{n} w_{di} 1\left(Y_i \leq y\right) - F_d(y) \\
=&\left[\frac{1}{n}\sum_{i=1}^{n} w_{di} F_d(y \,|\, \mathbf{X}_i) - \frac{1}{n}\sum_{i=1}^{n} F_d(y \,|\, \mathbf{X}_i)\right] + \left[\frac{1}{n}\sum_{i=1}^{n} F_d(y \,|\, \mathbf{X}_i) - F_d(y)\right] \\
&+ \frac{1}{n}\sum_{i=1}^{n} w_{di}\left[1\left(Y_i \leq y\right) - F_d(y \,|\, \mathbf{X}_i)\right] \\
\equiv& B_1 + B_2 + B_3.
\end{aligned}$$

One can see clearly that the second term B_2 and the third term B_3 go to zero by the law of

large numbers under certain regularity conditions. If we choose the weights such that

$$\frac{1}{n}\sum_{i=1}^{n} w_{di} F_d\left(y \,|\, \mathbf{X}_i\right) = \frac{1}{n}\sum_{i=1}^{n} F_d\left(y \,|\, \mathbf{X}_i\right), \tag{7.45}$$

which is termed as *the distribution balancing condition*, then, $F_d(y \,|\, \mathbf{X}_i)$ achieves balance between the treated/untreated group and full population after weighting. Therefore, the first term B_1 is zero and $\widehat{F}_d(y)$ becomes a consistent estimator of $F_d(y)$.

In practice, the balancing condition (7.45) cannot be directly applied. Firstly, the conditional CDF $F_d\left(y \,|\, \mathbf{X}_i\right)$ is unknown and must be estimated. To this end, we propose using the kernel method to estimate $F_d\left(y \,|\, \mathbf{X}_i\right)$. When the dimension of X_i is relatively large, some semiparametric estimation methods can also be employed, as described in Section 2.5. Secondly, to estimate $F_d(y)$ for all $y \in \mathcal{Y}$, where $\mathcal{Y}$ is the support of the outcome variable, it is infeasible to require that the balance condition (7.45) exactly holds for all $y \in \mathcal{Y}$. Without loss of generality, we assume that $\mathcal{Y}$ is a closed interval, denoted as $[y_l, y_u]$. Let $y_l = q_1 < \cdots < q_J = y_u$ be the equally spaced grid points on $[y_l, y_u]$. Then, according to the polynomial interpolation error formula from Süli and Mayers (2003), we have

$$F_d\left(y \,|\, \mathbf{X}_i\right) = \sum_{j=1}^{J} c_j(y) F_d\left(q_j \,|\, \mathbf{X}_i\right) + \frac{F_d^{(J)}\left(\xi_y \,|\, \mathbf{X}_i\right)}{J!} \prod_{j=1}^{J} \left(y - q_j\right) \tag{7.46}$$

for any $y \in [y_l, y_u]$, where $c_j(y) = \prod_{k=1, k\neq j}^{J} \frac{y-q_k}{q_j - q_k}$, $\xi_y \in [y_l, y_u]$ depends on y and $\mathbf{X}_i$, and $F_d^{(J)}\left(\xi_y \,|\, \mathbf{X}_i\right) = \frac{\partial^J F_d(u \,|\, X_i)}{\partial u^J}\big|_{u=\xi_y}$. It is easy to see that $\left|\prod_{j=1}^{J}\left(y - q_j\right)\right| = \prod_{j=1}^{J}\left|y - q_j\right| \leq \frac{(J-1)!}{4}\left(\frac{y_u - y_l}{J-1}\right)^J$. If we assume that the absolute value of $F_d^{(J)}\left(\xi_y \,|\, \mathbf{X}_i\right)$ is bounded by $C_0 > 0$ that does not depend on J, then, equation (7.46) leads to

$$\left| F_d\left(y \,|\, \mathbf{X}_i\right) - \sum_{j=1}^{J} c_j(y) F_d\left(q_j \,|\, \mathbf{X}_i\right) \right| \leq \frac{C_0}{4J}\left(\frac{y_u - y_l}{J-1}\right)^J.$$

Assume that the balance condition (7.45) exactly holds for $y = q_1, \cdots, q_J$. Since $F_d\left(q_J \,|\, \mathbf{X}_i\right) \equiv 1$ by definition, $\frac{1}{n}\sum_{i=1}^{n} w_{di} F_d\left(q_J \,|\, \mathbf{X}_i\right) = \frac{1}{n}\sum_{i=1}^{n} F_d\left(q_J \,|\, \mathbf{X}_i\right)$ implies $\frac{1}{n}\sum_{i=1}^{n} w_{di} = 1$ for $d = 0$ and 1. Therefore, we always assume $\frac{1}{n}\sum_{i=1}^{n} w_{0i} = 1$ and $\frac{1}{n}\sum_{i=1}^{n} w_{1i} = 1$ are in the balance conditions. Notice that $c_j(y)$ does not depend on $\mathbf{X}_i$. Thus, for any $y \in [y_l, y_u]$, the balance

error is

$$\begin{aligned}
& \left| \frac{1}{n}\sum_{i=1}^{n} w_{di} F_d\left(y \,|\, \mathbf{X}_i\right) - \frac{1}{n}\sum_{i=1}^{n} F_d\left(y \,|\, \mathbf{X}_i\right) \right| \\
\leq & \left| \frac{1}{n}\sum_{i=1}^{n} w_{di} \sum_{j=1}^{J} c_j(y) F_d\left(q_j \,|\, \mathbf{X}_i\right) - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J} c_j(y) F_d\left(q_j \,|\, \mathbf{X}_i\right) \right| + \frac{C_0}{2J}\left(\frac{y_u - y_l}{J-1}\right)^J \\
= & \left| \sum_{j=1}^{J} c_j(y) \left[\frac{1}{n}\sum_{i=1}^{n} w_{di} F_d\left(q_j \,|\, \mathbf{X}_i\right)\right] - \sum_{j=1}^{J} c_j(y) \left[\frac{1}{n}\sum_{i=1}^{n} F_d\left(q_j \,|\, \mathbf{X}_i\right)\right] \right| + \frac{C_0}{2J}\left(\frac{y_u - y_l}{J-1}\right)^J \\
= & \frac{C_0}{2J}\left(\frac{y_u - y_l}{J-1}\right)^J.
\end{aligned}$$

It is clear that the balance error for any $y \in [y_l, y_u]$ tends to zero as the number of grid points J approaches infinity. We can control J so that the balance error is negligible relative to the asymptotic performance of $\widehat{F}_d(y)$.

Now, we present a three-step procedure for estimating the counterfactual distribution functions $\widehat{F}_0(y)$ and $\widehat{F}_1(y)$ as follows.

Step 1: We estimate the conditional CDF $F_d\left(y \,|\, \mathbf{x}\right)$, $d = 0, 1$, by the Nadaraya-Watson estimator as

$$\widetilde{F}_d(y \,|\, \mathbf{x}) = \frac{\sum_{i=1}^{n} 1(d_i = d) 1\left(Y_i \leq y\right) K_{h_d}\left(\mathbf{X}_i - \mathbf{x}\right)}{\sum_{i=1}^{n} 1(d_i = d) K_{h_d}\left(\mathbf{X}_i - \mathbf{x}\right)}, \quad d = 0, 1, \tag{7.47}$$

where $K(\cdot)$ is a kernel function, h_d is the bandwidth, and $K_{h_d}\left(\mathbf{X}_i - \mathbf{x}\right) = h_d^{-p} K\left(\left(\mathbf{X}_i - \mathbf{x}\right)/h_d\right)$.

Step 2: Compute the optimal distribution balancing weights $\widehat{\boldsymbol{w}} = (\widehat{w}_1, \cdots, \widehat{w}_n)^\top$ by letting

$$\widehat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} \frac{1}{n}\sum_{i=1}^{n} \phi\left(w_i\right) \tag{7.48}$$

subject to $w_i \geq 0$,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} 1(d_i = 0) w_i \widetilde{F}_0\left(q_j \,|\, \mathbf{X}_i\right) &= \frac{1}{n}\sum_{i=1}^{n} \widetilde{F}_0\left(q_j \,|\, \mathbf{X}_i\right), \quad j = 1, \cdots, J, \\
\frac{1}{n}\sum_{i=1}^{n} 1(d_i = 1) w_i \widetilde{F}_1\left(q_j \,|\, \mathbf{X}_i\right) &= \frac{1}{n}\sum_{i=1}^{n} \widetilde{F}_1\left(q_j \,|\, \mathbf{X}_i\right), \quad j = 1, \cdots, J,
\end{aligned}$$

and for $\iota = 1, \cdots, L$,

$$\frac{1}{n}\sum_{i=1}^{n} 1(d_i = 0) w_i u_\iota\left(\mathbf{X}_i\right) = \frac{1}{n}\sum_{i=1}^{n} 1(d_i = 1) w_i u_\iota\left(\mathbf{X}_i\right) = \frac{1}{n}\sum_{i=1}^{n} u_\iota\left(\mathbf{X}_i\right),$$

where $\phi(w_i)$ is a non-negative, continuously differentiable and strictly convex function, which includes some special cases such as the entropy divergence as in Hainmueller (2012) with $\phi(w_i) = w_i \log(w_i)$, the stable balancing variance considered in Zubizarreta (2015) defined as $\phi(w_i) = 1(d_i = 0)(w_i - n/n_0)^2 + 1(d_i = 1)(w_i - n/n_1)^2$ with $n_0 = \sum_{i=1}^n 1(d_i = 0)$ and $n_1 = \sum_{i=1}^n 1(d_i = 1)$, and other distance measures as in Chan et al. (2016), Wang and Zubizarreta (2020), and Josey et al. (2021). The above objective function $\frac{1}{n}\sum_{i=1}^n \phi(w_i)$ measures the dispersion of the weights $w_1, \cdots, w_n$ and minimizing $\frac{1}{n}\sum_{i=1}^n \phi(w_i)$ tries to control the variance of the estimator. In addition, besides the key constraint developed from the balance condition in (7.45), we also allow other functions $u_\iota(\mathbf{X}_i)$, $\iota = 1, \cdots, L$, to be balanced across the treated group, untreated group, and combined group. For example, taking $u_\iota(\mathbf{X}_i) = \mathbf{X}_i^\iota$ means to balance the ιth moment; see, for example, the papers by Imai and Ratkovic (2014) and Fan et al. (2023) for details. The algorithm to calculate the optimal distribution balancing weights is presented in Section 7.5.2.

Step 3: Let $\widehat{w}_{di} = 1(d_i = d)\widehat{w}_i$ for $d = 0, 1$ and $i = 1, \cdots, n$. Then, the counterfactual distribution functions $F_d(y)$ for $d = 0$ and 1 are estimated by

$$\widehat{F}_d(y) = \frac{1}{n}\sum_{i=1}^n \widehat{w}_{di} 1(Y_i \leq y) \tag{7.49}$$

for all $y \in \mathcal{Y}$.

Remark 7.4: *It is worth to mention that Rothe (2010), Hsu et al. (2022), and Cai et al. (2026) also considered the estimation of the counterfactual distributions $F_d(y)$ using $\widetilde{F}_d(y \,|\, \mathbf{x})$ as in* (7.47)*. Different from our method, they proposed estimating $F_d(y)$ by $\widetilde{F}_d(y) = \frac{1}{n}\sum_{i=1}^n \widetilde{F}_d(y|\mathbf{X}_i)$. Rothe (2010) demonstrated that the estimator $\widetilde{F}_d(y)$ achieves $\sqrt{n}$-consistency by using high-order kernels. However, the performance of $\widetilde{F}_d(y)$ greatly depends on the choice of bandwidth h_d in* (7.47)*. For our method, $\widetilde{F}_d(y \,|\, x)$ is only used within the balance conditions, and the counterfactual distribution $F_d(y)$ is estimated by the weighted empirical CDF. The new estimator $\widehat{F}_d(y)$ is not very sensitive to the choice of bandwidth h_d in the first stage, which is supported by the results of Monte Carlo simulations reported in Section 4 in Cai et al. (2025a).*

To investigate the asymptotic properties of the proposed estimator, under some regularity conditions; see, for instance, Assumptions 1-10 in Cai et al. (2025a), first, Cai et al. (2025a) provided an explicit uniform convergence rate for the first-step estimator based on

the kernel method, $\widetilde{F}_d(y|\mathbf{x})$. Then, they showed the consistency of the optimal distribution balancing weights obtained in the second step; see, for instance, Propositions 1 and 2 in Cai et al. (2025a). Next, to derive the asymptotic properties of $\widehat{F}_d(y)$, they proved that $\sqrt{n}\left[\widehat{F}_d(y) - F_d(y)\right]$ is asymptotically linear with an influence function representation, similar to the Bahadur representation for sample quantile, that is,

$$\sqrt{n}\left[\widehat{F}_d(y) - F_d(y)\right] = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_d^F(y, \mathbf{Z}_i) + o_p(1)$$

uniformly in y, where

$$\psi_d^F(y, \mathbf{Z_i}) = \frac{1\{d_i = d\}\left[1\{Y_i \leq y\} - F_d(y \,|\, \mathbf{X}_i)\right]}{\pi(\mathbf{X}_i)^{d_i}[1 - \pi(\mathbf{X}_i)]^{1-d_i}} + \left[F_d\left(y \,|\, \mathbf{X}_i\right) - F_d(y)\right] \tag{7.50}$$

with $\mathbf{Z}_i = (Y_i, d_i, \mathbf{X}_i)$. Then, the following weak convergence result holds by the functional central limit theorem[10], uniformly for $\mathbf{y} = (y_0, y_1) \in \mathcal{Y} \times \mathcal{Y}$, we have

$$\sqrt{n}\left[\widehat{\mathbf{F}}(\mathbf{y}) - \mathbf{F}(\mathbf{y})\right] \;\Rightarrow\; \mathbb{F}(\mathbf{y}), \tag{7.51}$$

where " $\Rightarrow$ " denotes the weak convergence, $\mathbb{F}(\mathbf{y}) = \left(\mathbb{F}_0\left(y_0\right), \mathbb{F}_1\left(y_1\right)\right)$ is a two-dimensional Gaussian process with zero mean and covariance function

$$\Psi^F\left(\mathbf{y}_1, \mathbf{y}_2\right) = \mathbb{E}\left[\psi^F\left(\mathbf{y}_1, \mathbf{Z}\right)\psi^F\left(\mathbf{y}_2, \mathbf{Z}\right)^\top\right],$$

and the convergence takes place in $\ell^\infty(\mathcal{Y}) \times \ell^\infty(\mathcal{Y})$, where $\ell^\infty(\mathcal{Y})$ is the set of bounded functions over $\mathcal{Y}$. Here, $\psi^F\left(\mathbf{y}, \mathbf{Z}\right) = \left(\psi_0^F\left(y_0, \mathbf{Z}\right), \psi_1^F\left(y_1, \mathbf{Z}\right)\right)$, where $\psi_d^F(y, \mathbf{Z})$ defined in (7.50).

Evidently, (7.51) shows that $\widehat{\mathbf{F}}(\mathbf{y}) - \mathbf{F}(\mathbf{y})$ converges weakly to a mean-zero Gaussian process at the usual parametric rate of $\sqrt{n}$ despite the use of nonparametric estimators in the first step. Also, the result in (7.51) facilitates making inferences of the QTE and testing the stochastic dominance relationship between the distributions of potential outcomes; see, for example, the paper by Cai et al. (2025a) for details. For testing the stochastic dominance, Section 7.5.4 gives the detailed description.

7.5.2 Implementation of Finding Optimal Weights

The constrained optimization problem in (7.48) is a convex separable programming problem with linear constraints. Actually, Tseng and Bertsekas (1987) showed that its dual

[10]For the detailed definitions of Donsker class, the weak convergence, and the functional central limit theorem, the reader is referred to the book by Billingsley (1999).

problem is an unconstrained convex maximization problem that can be solved by efficient and stable numerical algorithms. Therefore, we consider its dual problem in the following. To this end, define $\widetilde{\mathbf{U}}_d(\mathbf{x}) = \left(\widetilde{\mathbf{F}}_d(\mathbf{q}\,|\,\mathbf{x})^\top, \mathbf{u}(\mathbf{x})^\top\right)^\top$ with $\widetilde{\mathbf{F}}_d(\mathbf{q}\,|\,\mathbf{x}) = \left(\widetilde{F}_d(q_1\,|\,\mathbf{x}), \cdots, \widetilde{F}_d(q_J\,|\,\mathbf{x})\right)^\top$ and $\mathbf{u}(\mathbf{x}) = (u_1(\mathbf{x}), \cdots, u_L(\mathbf{x}))^\top$ for $d = 0$ and 1. Then, the Lagrangian of the optimization problem in (7.48) can be written as

$$\begin{aligned}
\widetilde{G}_n(\boldsymbol{w}, \boldsymbol{\lambda}_0, \boldsymbol{\lambda}_1) =& \frac{1}{n}\sum_{i=1}^n \phi(w_i) + \sum_{j=1}^J \lambda_{0,j}\left[\frac{1}{n}\sum_{i=1}^n 1(D_i = 0) w_i \widetilde{F}_0(q_j\,|\,\mathbf{X}_i) - \frac{1}{n}\sum_{i=1}^n \widetilde{F}_0(q_j\,|\,\mathbf{X}_i)\right] \\
&+ \sum_{\iota=1}^L \lambda_{0,J+\iota}\left[\frac{1}{n}\sum_{i=1}^n 1(D_i = 0) w_i u_\iota(\mathbf{X}_i) - \frac{1}{n}\sum_{i=1}^n u_\iota(\mathbf{X}_i)\right] \\
&+ \sum_{j=1}^J \lambda_{1,j}\left[\frac{1}{n}\sum_{i=1}^n 1(D_i = 1) w_i \widetilde{F}_1(q_j\,|\,\mathbf{X}_i) - \frac{1}{n}\sum_{i=1}^n \widetilde{F}_1(q_j\,|\,\mathbf{X}_i)\right] \\
&+ \sum_{\iota=1}^L \lambda_{1,J+\iota}\left[\frac{1}{n}\sum_{i=1}^n 1(D_i = 1) w_i u_\iota(\mathbf{X}_i) - \frac{1}{n}\sum_{i=1}^n u_\iota(\mathbf{X}_i)\right] \\
=& \frac{1}{n}\sum_{i=1}^n \phi(w_i) + \frac{1}{n}\sum_{i=1}^n \sum_{d=0,1} 1(D_i = d) w_i \widetilde{\mathbf{U}}_d(\mathbf{X}_i)^\top \boldsymbol{\lambda}_d - \frac{1}{n}\sum_{i=1}^n \sum_{d=0,1} \widetilde{\mathbf{U}}_d(\mathbf{X}_i)^\top \boldsymbol{\lambda}_d,
\end{aligned} \tag{7.52}$$

where $\boldsymbol{\lambda}_0 = (\lambda_{0,1}, \cdots, \lambda_{0,J+L})^\top$ and $\boldsymbol{\lambda}_1 = (\lambda_{1,1}, \cdots, \lambda_{1,J+L})^\top$ are the Lagrange multipliers. The first order condition $\partial \widetilde{G}_n(\boldsymbol{w}, \boldsymbol{\lambda}_0, \boldsymbol{\lambda}_1)/\partial w_i = 0$ yields

$$\phi'(w_i) = -\sum_{d=0,1} 1(D_i = d) \widetilde{\mathbf{U}}_d(X_i)^\top \boldsymbol{\lambda}_d,$$

where $\phi'(\cdot)$ is the first derivative of $\phi(\cdot)$. Let $(\phi')^{-1}(\cdot)$ be the inverse function of $\phi'(\cdot)$. Then,

$$w_i = (\phi')^{-1}\left(-\sum_{d=0,1} 1(D_i = d) \widetilde{\mathbf{U}}_d(X_i)^\top \boldsymbol{\lambda}_d\right). \tag{7.53}$$

For simplicity, define $\rho(t) = \phi\left\{(\phi')^{-1}(-t)\right\} + t\,(\phi')^{-1}(-t)$. Plugging (7.53) back into (7.52) eliminates the constraints, resulting in an unrestricted dual maximization problem given by

$$\begin{aligned}
\widetilde{G}_n\big[\boldsymbol{w}\big(\boldsymbol{\lambda}_0, \boldsymbol{\lambda}_1\big), \boldsymbol{\lambda}_0, \boldsymbol{\lambda}_1\big] =& \frac{1}{n}\sum_{i=1}^n \sum_{d=0,1} 1(D_i = d) \rho\left(\widetilde{\mathbf{U}}_d(X_i)^\top \boldsymbol{\lambda}_d\right) - \frac{1}{n}\sum_{i=1}^n \sum_{d=0,1} \widetilde{\mathbf{U}}_d(X_i)^\top \boldsymbol{\lambda}_d \\
\equiv& \widetilde{G}_{n,0}(\boldsymbol{\lambda}_0) + \widetilde{G}_{n,1}(\boldsymbol{\lambda}_1),
\end{aligned}$$

where

$$\widetilde{G}_{n,d}(\boldsymbol{\lambda}_d) = \frac{1}{n}\sum_{i=1}^n 1(D_i = d) \rho\left(\widetilde{\mathbf{U}}_d(X_i)^\top \boldsymbol{\lambda}_d\right) - \frac{1}{n}\sum_{i=1}^n \widetilde{\mathbf{U}}_d(X_i)^\top \boldsymbol{\lambda}_d.$$

It is clear that $\rho'(t) = (\phi')^{-1}(-t)$ and $\rho''(t) = -1/\phi''\left((\phi')^{-1}(-t)\right)$. Thus, both $\widetilde{G}_{n,0}(\boldsymbol{\lambda}_0)$ and $\widetilde{G}_{n,1}(\boldsymbol{\lambda}_1)$ are strictly concave due to the strict convexity of $\phi(\cdot)$. Therefore, the solution to the constrained optimization problem in (7.48) is

$$\widehat{w}_i = (\phi')^{-1}\left(-\sum_{d=0,1} 1(D_i = d)\widetilde{\mathbf{U}}_d(X_i)^\top \widehat{\boldsymbol{\lambda}}_d\right) = \rho'\left(\sum_{d=0,1} 1(D_i = d)\widetilde{\mathbf{U}}_d(X_i)^\top \widehat{\boldsymbol{\lambda}}_d\right),$$

where $\widehat{\boldsymbol{\lambda}}_0$ and $\widehat{\boldsymbol{\lambda}}_1$ are the unique maximizers of $\widetilde{G}_{n,0}(\boldsymbol{\lambda}_0)$ and $\widetilde{G}_{n,1}(\boldsymbol{\lambda}_1)$, respectively.

7.5.3 Extension to High-Dimensional Case

It is well known in the nonparametric statistics literature that when the covariate dimension is relatively large but still finite (p does not depend on n), it is not desirable to estimate the conditional CDF by kernel method as in (7.47). To circumvent this problem, some semiparametric estimators can be employed. One can adopt the approach suggested in Aït-Sahalia and Brant (2001) and Hall and Yao (2005), which involves using an index model. Specifically, it assumes that there exists a $p \times 1$ vector $\boldsymbol{\gamma}_d$ so that $F_d(y \,|\, \mathbf{x}) = F_d(y \,|\, \boldsymbol{\gamma}_d^\top \mathbf{x})$. One can first estimate $\boldsymbol{\gamma}_d$ to approximate $F_d(y \,|\, \mathbf{x})$ by $F_d(y \,|\, \boldsymbol{\gamma}_d^\top \mathbf{x})$ under a least-squares criterion, then, use $\boldsymbol{\gamma}_d^\top \mathbf{x}$ as the smooth variable to estimate the conditional CDF[11]. Another approach is to estimate the conditional CDF using quantile regression as in Koenker and Bassett (1978). This idea was also used in Chernozhukov et al. (2013) and Cai et al. (2026). We present this method below in detail.

Let $Q_d(\tau \,|\, \mathbf{x}) = \inf\{y : F_d(y \,|\, \mathbf{x}) \geq \tau\}$ denote the conditional quantile function of $Y(d)$ conditional on $\mathbf{X} = \mathbf{x}$ at the quantile level $\tau \in (0, 1)$. Assume that $Q_d(\tau \,|\, \mathbf{x}) = \boldsymbol{\beta}_d(\tau)^\top \mathbf{x}$ for each quantile level τ. The coefficients $\boldsymbol{\beta}_d(\tau)$ can be estimated by

$$\widehat{\boldsymbol{\beta}}_d(\tau) = \arg\min_{\boldsymbol{\beta}_d(\tau)} \sum_{i=1}^{n} 1(D_i = d)\, \rho_\tau\left(Y_i - \boldsymbol{\beta}_d(\tau)^\top \mathbf{X}_i\right). \tag{7.54}$$

Since $F_d(y \,|\, \mathbf{x}) = \int_0^1 1\left(Q_d(\tau \,|\, \mathbf{x}) \leq y\right) d\tau$, the conditional CDF can be estimated by

$$\widehat{F}_d(y \,|\, \mathbf{x}) = \varepsilon + \int_\varepsilon^{1-\varepsilon} 1\left(\widehat{\boldsymbol{\beta}}_d(\tau)^\top \mathbf{x} \leq y\right) d\tau \approx \varepsilon + \sum_{j=2}^{S} (\tau_j - \tau_{j-1})\, 1\left(\widehat{\boldsymbol{\beta}}_d(\tau_j)^\top \mathbf{x} \leq y\right), \tag{7.55}$$

where the trimming by ε avoids estimation of tail quantiles, and $\widehat{\boldsymbol{\beta}}_d(\tau)$ is estimated by (7.54) on a equally spaced mesh $\varepsilon = \tau_1 < \cdots < \tau_S = 1 - \varepsilon$.

[11] For details, the reader is referred to the paper by Hall and Yao (2005).

Remark 7.5: *Based on Proposition 5 in Chernozhukov et al. (2010), the conditional CDF estimators obtained by* (7.55) *are* $\sqrt{n}$*-consistent when the linear conditional quantile models are correctly specified and the mesh width is* $o(n^{-1/2})$*. In this case, the convergence rate of* $\widehat{F}_d(y\,|\,\mathbf{x})$ *is faster than that of the kernel estimator* $\widetilde{F}_d(y\,|\,\mathbf{x})$ *as stated in Proposition 1 in Cai et al. (2025a). Consequently, the conclusion in* (7.51) *still holds if we use the estimator* $\widehat{F}_d(y\,|\,x)$ *to replace* $\widetilde{F}_d(y\,|\,\mathbf{x})$ *in the first stage.*

Remark 7.6: *In the above, we only consider the case that* p *is finite. In some applications,* p *might be allowed to depend on the sample size so that the model setting in this section becomes the case with either high-dimensional (*$p \to \infty$ *but* $p/n \to 0$*) or ultra-high dimensional (*$p \gg n$*) covariates. For such cases, one might follow the idea in Cai et al. (2025c) for a mean model to do some extensions, which are not straightforward and can be warranted as future research topics.*

7.5.4 Testing Stochastic Dominance

Making inferences regarding stochastic dominance relationship plays an important role in social sciences, with a vast amount of literature in economics, including but not limited to, Anderson (1996), Barrett and Donald (2003), Linton et al. (2023), and references therein. Different from the existing literature, the focus is on testing the stochastic dominance relationship between the counterfactual distributions, which are not derived from any observable populations. Interestingly, Rothe (2010), Maier (2011), and Donald and Hsu (2014) also considered such a test in a similar scenario, but they used different methods to estimate the counterfactual distributions. Indeed, Rothe (2010) used the estimation method outlined in Remark 7.4, while Maier (2011) and Donald and Hsu (2014) estimated the counterfactual distributions by inverse propensity score weights. In this section, we use $\widehat{F}_0(\cdot)$ and $\widehat{F}_1(\cdot)$ obtained in Section 7.5.1 to test stochastic dominance.

We only discuss test for the first order stochastic dominance (SD1) between the potential outcomes $Y(0)$ and $Y(1)$. To test if $Y(1)$ SD1 $Y(0)$, the hypothesis is formulated as

$$H_0 : F_1(y) \le F_0(y) \quad \text{for all } y \in \mathcal{Y} \quad \text{versus} \quad H_1 : F_1(y) > F_0(y) \quad \text{for some } y \in \mathcal{Y}. \tag{7.56}$$

A commonly used statistic for testing the first order stochastic dominance is the Kolmogorov-Smirnov statistic, given by

$$\widehat{\mathrm{KS}}_{\mathrm{SC}} = \sqrt{n} \sup_{y \in \mathcal{Y}} \left[\widehat{F}_1(y) - \widehat{F}_0(y) \right] = \sqrt{n} \max_{y \in \{Y_1, \cdots, Y_n\}} \left[\widehat{F}_1(y) - \widehat{F}_0(y) \right].$$

The second equality follows from the fact that both $\widehat{F}_1(y)$ and $\widehat{F}_0(y)$ are step function and their values change only at the observed Y_i, $i = 1, \cdots, n$. For more discussions about SD1, the reader is referred to the book by Whang (2019) and the paper by Linton et al. (2023).

Note that we are testing a composite null hypothesis. For this case, it is challenging to find the limit null distribution since the limit null distribution depends on the underlying distributions, while there are infinitely many different combinations of $F_1(\cdot)$ and $F_0(\cdot)$ satisfying the null hypothesis. The typical way to solve this problem is to find the least favorable configuration (LFC)[12] to construct an asymptotically valid test procedure based on Bootstrapping the test statistic similar to that in Barrett and Donald (2003). It is easy to see that the LFC in this context corresponds to $F_1(y) = F_0(y)$ for all $y \in \mathcal{Y}$. Let $\widehat{F}_1^b$ and $\widehat{F}_0^b$ be the Bootstrap estimates of the potential outcomes' distributions based on the classical nonparametric Bootstrap. Then, the Bootstrap p-value can be calculated as

$$\widehat{p} = B^{-1} \sum_{b=1}^{B} 1\left(\widehat{\mathrm{KS}}_{\mathrm{SC}}^b > \widehat{\mathrm{KS}}_{\mathrm{SC}}\right),$$

where $\widehat{\mathrm{KS}}_{\mathrm{SC}}^b$ is the Bootstrap version of Kolmogorov-Smirnov statistic, similar to that in Section 1.2, given by

$$\widehat{\mathrm{KS}}_{\mathrm{SC}}^b = \sqrt{n} \max_{y \in \{Y_1, \ldots Y_n\}} \left\{ \left[\widehat{F}_1^b(y) - \widehat{F}_0^b(y)\right] - \left[\widehat{F}_1(y) - \widehat{F}_0(y)\right] \right\}, \tag{7.57}$$

which is similar to $D_{n,\theta}^{(b)}$ in (1.7) in Section 1.2. Thus, we reject H_0 if $\widehat{p}$ is less than the significance level α. Cai et al. (2025a) delivered theoretical justification for this Bootstrap approach. If we reject H_0 when $\widehat{p} < \alpha$, where α is the significance level, then, (i) under H_0 defined in (7.56), $\lim_{n\to\infty} P(\widehat{p} < \alpha) \leq \alpha$, and (ii) under a fixed alternative hypothesis defined in (7.56), $\lim_{n\to\infty} P(\widehat{p} < \alpha) = 1$. This theoretical result implies that the size of the proposed test is asymptotically no larger than the pre-specified significance level α, and the power of the proposed test is asymptotically approaching 1 under the alternative hypothesis. Cai et al. (2025a) conducted a simulation study to verify that this proposed test in (7.57) performs reasonably well in the finite sample evaluation.

[12]Under a composite null hypothesis, the least favorable case is the distribution for which the null holds, but which is most difficult to distinguish from any distribution in the alternative hypothesis. See Section 3 in Lehmann and Romano (2005).

Bibliography

Abadie, A. (2021). Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature* **59**(2), 391–425.

Abadie, A., A. Diamond, and J. Hainmueller (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association* **105**(490), 493–505.

Abadie, A. and J. Gardeazabal (2003). The economic costs of conflict: A case study of the Basque country. *American Economic Review* **93**(1), 113–132.

Abadie, A. and J. L'Hour (2021). A penalized synthetic control estimator for disaggregated data. *Journal of Business & Economic Statistics* **116**(536), 1817–1835.

Abrevaya, J., Y. C. Hsu, and R. P. Lieli (2015). Estimating conditional average treatment effects. *Journal of Business & Economic Statistics* **33**(4), 485–505.

Acerbi, C. and D. Tasche (2002). On the coherence of expected shortfall. *Journal of Banking and Finance* **26**(7), 1487–1503.

Aït-Sahalia, Y. (1996). Nonparametric pricing of interest rate derivative securities. *Econometrica* **64**(3), 527–560.

Aït-Sahalia, Y. and M. W. Brant (2001). Variable selection for portfolio choice. *Journal of Finance* **56**(4), 1297–1351.

Aït-Sahalia, Y. and A. W. Lo (1998). Nonparametric estimation of state-price densities implicit in financial asset prices. *Journal of Finance* **53**(2), 499–547.

Aït-Sahalia, Y. and A. W. Lo (2000). Nonparametric risk management and implied risk aversion. *Journal of Econometrics* **94**(1-2), 9–51.

Almond, D., K. Y. Chay, and D. S. Lee (2005). The costs of low birth weight. *Quarterly Journal of Economics* **120**(3), 1031–1083.

Anderson, G. (1996). Nonparametric tests of stochastic dominance in income distributions. *Econometrica* **65**(4), 1183–1193.

Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica* **59**(3), 817–58.

Arjovsky, M., S. Chintala, and L. Bottou (2017). Wasserstein generative adversarial networks. *Proceedings of the 34th International Conference on Machine Learning* **70**(PMLR), 214–223.

Artzner, P., F. Delbaen, J.-M. Eber, and D. Heath (1999). Coherent measures of risk. *Mathematical Finance* **9**(3), 203–228.

Athey, S. and G. W. Imbens (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives* **31**(1), 3–32.

Auestad, B. and D. Tjøstheim (1990). Identification of nonlinear time series: First order characterization and order determination. *Biometrika* **77**(4), 669–687.

Babu, G. J. and C. R. Rao (2004). Goodness-of-fit tests when parameters are estimated. *Sankhyā* **66**(1), 63–74.

Bachmeier, L., S. Leelahanon, and Q. Li (2007). Money growth and inflation in the United States. *Macroeconomic Dynamics* **11**(1), 113–127.

Bai, C., Q. Li, and M. Ouyang (2014). Property taxes and home prices: A tale of two cities. *Journal of Econometrics* **180**(1), 1–15.

Bandi, F. M. (2002). On persistence and nonparametric estimation (with an application to stock return predictability). Working paper, Graduate School of Business, University of Chicago.

Bao, Y., T.-H. Lee, and B. Saltoglu (2006). Evaluating predictive performance of value-at-risk models in emerging markets: A reality check. *Journal of Forecasting* **25**(2), 101–128.

Barrett, G. F. and S. G. Donald (2003). Consistent tests for stochastic dominance. *Econometrica* **71**(1), 71–104.

Bellini, F., B. Klar, A. Müller, and E. R. Gianin (2014). Generalized quantiles as risk measures. *Insurance: Mathematics and Economics* **54**(1), 41–48.

Bellini, F. and B. Valeria (2015). On elicitable risk measures. *Quantitative Finance* **15**(5), 725–733.

Belsley, D. A., E. Kuh, and R. E. Welsch (1980). *Regression Diagnostic: Identifying Influential Data and Sources of Collinearity*. New York: John Wiley & Sons.

Billingsley, P. (1999). *Convergence of Probability Measures* (2nd ed.). New York: John Wiley & Sons.

Boente, G. and R. Fraiman (1989). Robust nonparametric for dependent variables. *Annals of Statistics* **17**(3), 1424–1456.

Boente, G. and R. Fraiman (1990). Asymptotic distribution of robust estimator for nonparametric models from mixing processes. *Annals of Statistics* **18**(2), 891–906.

Boente, G. and R. Fraiman (1995). Asymptotic distribution of smoothers based on local means and local medians under dependence. *Journal of Multivariate Analysis* **54**(1), 77–90.

Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika* **71**(2), 353–360.

Breiman, L. and J. H. Friedman (1985). Estimating optimal transformation for multiple regression and correlation. *Journal of the American statistical Association* **80**(391), 580–598.

Cai, Y. (2010a). Multivariate quantile function models. *Statistica Sinica* **20**(3), 481–496.

Cai, Z. (2001). Weighted nadaraya-watson regression estimation. *Statistics & Probability Letters* **51**(3), 307–318.

Cai, Z. (2002a). Regression quantiles for time series. *Econometric Theory* **18**(1), 169–192.

Cai, Z. (2002b). A two-stage approach to additive time series models. *Statistica Neerlandica* **56**(4), 415–433.

Cai, Z. (2003a). Local quasi-likelihood approach to varying-coefficient discrete-valued time series models. *Journal of Nonparametric Statistics* **15**(6), 693–711.

Cai, Z. (2003b). Nonparametric estimation equations for time series data. *Statistics & Probability Letters* **62**(4), 379–390.

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

Cai, Z. (2010b). Functional coefficient models for economic and financial data. In F. Ferraty and Y. Romain (Eds.), *Oxford Handbook of Functional Data Analysis*, pp. 166–186. Oxford, UK: Oxford University Press.

Cai, Z. (2011). Nonparametric regression models with integrated covariates. In J. Jiang, G. Roussas, and F. Samaniego (Eds.), *Nonparametric Statistical Methods and Related Topics:* A Festschrift in Honor of Professor P.K. Bhattacharya on his 80th Birthday, pp. 257–275. Singapore: World Scientific.

Cai, Z., M. Das, H. Xiong, and X. Wu (2006). Functional coefficient instrumental variables models. *Journal of Econometrics* **133**(1), 207–241.

Cai, Z. and J. Fan (2000). Average regression surface for dependent data. *Journal of Multivariate Analysis* **75**(1), 112–142.

Cai, Z., J. Fan, and R. Li (2000). Efficient estimation and inferences for varying-coefficient models. *Journal of the American Statistical Association* **95**(451), 888–902.

Cai, Z., J. Fan, and Q. Yao (2000). Functional-coefficient regression models for nonlinear time series. *Journal of the American Statistical Association* **95**(451), 941–956.

Cai, Z., Y. Fang, M. Lin, and S. Tang (2021). Estimation of partially conditional quantile treatment effects. *China Journal of Econometrics* **1**(4), 741–762.

Cai, Z., Y. Fang, M. Lin, and S. Tang (2024). Testing conditional independence in casual inference for time series data. *Statistica Neerlandica* **78**(2), 397–426.

Cai, Z., Y. Fang, M. Lin, and S. Tang (2025). A nonparametric test of heterogeneity in conditional quantile treatment effects. *Econometric Theory* **41**(3), 660–687.

Cai, Z., Y. Fang, M. Lin, and Y. Wu (2025a). Estimating counterfactual distribution functions via optimal distribution balancing with applications. Working Paper, Department of Economics, University of Kansas.

Cai, Z., Y. Fang, M. Lin, and Z. Wu (2025b). A new adversarial covariate balancing network for causal inference. Working Paper, Department of Economics, University of Kansas.

Cai, Z., Y. Fang, M. Lin, and Z. Wu (2025c). A quasi synthetic control method for nonlinear models with high-dimensional covariates. *Statistica Sinica,* DOI: https://doi.org/10.5705/ss.202023.0271.

Cai, Z., Y. Fang, M. Lin, and M. Zhan (2026). Estimating quantile treatment effects for panel data. *Scientia Sinica Mathematica* **56**(1), 33–54.

Cai, Z., Y. Fang, and D. Tian (2018). Assessing tail risk using expectile models with partially varying coefficients. *Journal of Management Science and Engineering, Series B* **3**(4), 179–209.

Cai, Z., Y. Fang, and D. Tian (2025). Assessing tail risk via a generalized conditional autoregressive expectile model. *Journal of Financial Econometrics* **23**(2), nbaf010.

Cai, Z., Y. Fang, and D. Tian (2026). CAViaR model selection via adaptive LASSO. *Journal of Time Series Analysis*, DOI: https://doi.org/10.1111/jtsa.12804.

Cai, Z. and Y. Hong (2009). Some recent developments in nonparametric finance. *Advances in Econometrics* **25**, 379–432.

Cai, Z., B. Y. Jing, X. K. Kong, and Z. Liu (2017). Nonparametric regression with nearly integrated regressors under long run dependence. *Econometrics Journal* **20**(1), 118–138.

Cai, Z., T. Juhl, and B. Yang (2015). Functional index coefficient models with variable selection. *Journal of Econometrics* **189**(2), 272–284.

Cai, Z. and J. Li (2025). Econometric evaluation of the China–US trade war effects. *China Economic Review* **94**(102567), 1–17.

Cai, Z. and Q. Li (2008). Nonparametric estimation of varying coefficient dynamic panel models. *Econometric Theory* **24**(5), 1321–1342.

Cai, Z., Q. Li, and J. Y. Park (2009). Functional-coefficient models for nonstationary time series data. *Journal of Econometrics* **148**(1), 101–113.

Cai, Z. and E. Masry (2000). Nonparametric estimation of additive nonlinear ARX time series: Local linear fitting and projection. *Econometric Theory* **16**(4), 465–501.

Cai, Z. and E. Ould-Saïd (2003). Local M-estimator for nonparametric time series. *Statistics & Probability Letters* **65**(4), 433–449.

Cai, Z. and L. Qian (2000). Local estimation of a biometric function with covariate effects. In M. Puri (Ed.), *Asymptotics in Statistics and Probability*, pp. 47–70.

Cai, Z., Y. Ren, and B. Yang (2015). A semiparametric conditional capital asset pricing model. *Journal of Banking and Finance* **61**(1), 117–126.

Cai, Z. and G. G. Roussas (1997). Smooth estimate of quantiles under association. *Statistics & Probability Letters* **36**(3), 275–287.

Cai, Z. and G. G. Roussas (1998). Efficient estimation of a distribution function under quadrant dependence. *Scandinavian Journal of Statistics* **25**(1), 211–224.

Cai, Z. and R. C. Tiwari (2000). Application of a local linear autoregressive model to BOD time series. *Environmetrics* **11**(3), 341–350.

Cai, Z. and X. Wang (2008). Nonparametric methods for estimating conditional VaR and expected shortfall. *Journal of Econometrics* **147**(1), 120–130.

Cai, Z. and Y. Wang (2014). Testing predictive regression models with nonstationary regressors. *Journal of Econometrics* **178**(1), 4–14.

Cai, Z., Y. Wang, and Y. Wang (2015). Testing instability in predictive regression model with nonstationary regressors. *Econometric Theory* **31**(5), 953–980.

Cai, Z. and Z. Xiao (2012). Semiparametric quantile regression estimation in dynamic models with partially varying coefficients. *Journal of Econometrics* **167**(2), 413–425.

Cai, Z. and X. Xu (2008). Nonparametric quantile estimations for dynamic smooth coefficient models. *Journal of the American Statistical Association* **103**(484), 1595–1608.

Cai, Z., Q. Yao, and W. Zhang (2001). Smoothing for discrete-valued time series. *Journal of the Royal Statistical Society, Series B* **63**(2), 357–375.

Callaway, B., T. Li, and T. Oka (2018). Quantile treatment effects in difference in differences models under dependence restrictions and with only two time periods. *Journal of Econometrics* **206**(2), 395–413.

Cameron, A. C. and P. K. Trivedi (2005). *Microeconometrics: Methods and Applications*. New York: Cambridge University Press.

Campbell, J. Y. and M. Yogo (2006). Efficient tests of stock return predictability. *Journal of Financial Economics* **81**(1), 27–60.

Carrasco, M. and X. Chen (2002). Mixing and moment properties of various GARCH and stochastic volatility models. *Econometric Theory* **18**(1), 17–39.

Carroll, R. J., D. Ruppert, and A. H. Welsh (1998). Local estimation equations. *Journal of the American Statistical Association* **93**(441), 214–227.

Cavanagh, C. L., D. Elliott, and J. H. Stock (1995). Inference in models with nearly integrated regressors. *Econometric Theory* **11**(5), 1131–1147.

Chan, K. C. G., S. C. P. Yam, and Z. Zhang (2016). Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of the Royal Statistical Society, Series B* **78**(3), 673–700.

Chandy, M., E. D. Schifano, J. Yan, and X. Zhang (2025). Nonparametric block bootstrap kolmogorov-smirnov goodness-of-fit test. URL: https://arxiv.org/pdf/2511.05733.

Chang, Y. and E. Martinez-Chombo (2003). Electricity demand analysis using cointegration and error-correction models with time varying parameters: The Mexican case. Working paper, Department of Economics, Indiana University.

Chang, Y. and J. Y. Park (2003). Index models with integrated time series. *Journal of Econometrics* **114**(1), 73–106.

Chapman, D., J. Long, and N. Pearson (1999). Using proxies for the short rate: When are three months like an instant. *Review of Financial Studies* **12**(4), 763–807.

Chapman, D. and N. Pearson (2000). Is the short rate drift actually nonlinear? *Journal of Finance* **55**(1), 355–388.

Chaudhuri, P. (1991). Nonparametric estimates of regression quantiles and their local bahadur representation. *Annals of Statistics* **19**(2), 760–777.

Chaudhuri, P., K. Doksum, and A. Samarov (1997). On average derivative quantile regression. *Annuals of Statistics* **25**(2), 715–744.

Chen, H., Q. Han, Y. Li, and K. Wu (2013). Does index futures trading reduce volatility in the Chinese stock market? A panel data evaluation approach. *Journal of Futures Markets* **33**(12), 1167–1190.

Chen, K., Z. Ying, H. Zhang, and L. Zhao (2008). Analysis of least absolute deviation. *Biometrika* **95**(1), 107–122.

Chen, R. and R. S. Tsay (1993). Functional-coefficient autoregressive models. *Journal of the American Statistical Association* **88**(421), 298–308.

Chen, S. X. (2008). Nonparametric estimation of expected shortfall. *Journal of Financial Econometrics* **6**(1), 87–107.

Chen, S. X. and C. Y. Tang (2005). Nonparametric inference of value-at-risk for dependent financial returns. *Journal of Financial Econometrics* **3**(2), 227–255.

Chen, X., Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel (2016). InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. *Proceedings of The 30th International Conference on Neural Information Processing Systems* **29**(NIPS2016), 2180–2188.

Chernozhukov, V., I. Fernández-Val, and A. Galichon (2010). Quantile and probability curves without crossing. *Econometrica* **78**(3), 1093–1125.

Chernozhukov, V., I. Fernández-Val, and B. Melly (2013). Inference on counterfactual distributions. *Econometrica* **81**(6), 2205–2268.

Chernozhukov, V. and L. Umantsev (2001). Conditional Value-at-Risk: Aspects of modeling and estimation. *Empirical Economics* **26**(1), 271–292.

Chiu, S. T. (1991). Bandwidth selection for kernel density estimation. *Annals of Statistics* **19**(4), 1883–1905.

Choi, Y., S. Jacewitz, and J. Y. Park (2016). A reexamination of stock return predictability. *Journal of Econometrics* **192**(1), 168–189.

Cole, T. J. (1994). Growth charts for both cross-sectional and longitudinal data. *Statistics in Medicine* **13**(23-24), 2477–2492.

Collomb, G. and W. Härdle (1986). Strong uniform convergence rates in robust nonparametric time series analysis and prediction: Kernel regression estimation from dependent observations. *Stochastic Processes and Their Applications* **23**(1), 77–89.

Coppejansa, M. and R. A. Gallant (2002). Cross-validated SNP density estimates. *Journal of Econometrics* **110**(1), 27–65.

Cosma, A., O. Scaillet, and R. Von Sachs (2007). Multivariate wavelet-based shape- preserving estimation for dependent observations. *Bernoulli* **13**(2), 301–329.

Craven, P. and G. Wahba (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik* **31**(4), 377–403.

Crump, R. K., V. J. Hotz, G. W. Imbens, and O. A. Mitnik (2008). Nonparametric tests for treatment effect heterogeneity. *Review of Economics and Statistics* **90**(3), 389–405.

Dahlhaus, R. (1996). On the Kullback-Leibler information divergence of locally stationary processes. *Stochastic Processes and Their Applications* **62**(1), 139–168.

De Gooijer, J. G. and D. Zerom (2003). On additive conditional quantiles with high- dimensional covariates. *Journal of the American Statistical Association* **98**(461), 135–146.

De Rossi, G. and A. Harvey (2009). Quantiles, expectiles and splines. *Journal of Econometrics* **152**(2), 179–185.

Del Brio, E. B., T.-M. Ñíguez, and J. Perote (2010). Multivariate semi-nonparametric distributions with dynamic conditional correlations. *International Journal of Forecasting* **27**(2), 347–364.

Deng, H. and H. Wickham (2011). Density estimation in R. URL: https://vita.had.co.nz/papers/density-estimation.pdf.

Dodge, Y. (2008). *The Concise Encyclopedia of Statistics*. pp. 299, Berlin: Springer-Verlag.

Donald, S. G. and Y.-C. Hsu (2014). Estimation and inference for distribution functions and quantile functions in treatment effect models. *Journal of Econometrics* **178**, 383–397.

Doudchenko, N. and G. W. Imbens (2016). Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. NBER Working Papers, No. 22791.

Duffie, D. and J. Pan (1997). An overview of value at risk. *Journal of Derivatives* **4**(3), 7–49.

Duffie, D. and K. J. Singleton (2003). *Credit Risk: Pricing, Measurement, and Management*. Princeton, NJ: Princeton University Press.

Durrett, R. (2019). *Probability: Theory and Examples* (Fifth ed.). New York: Cambridge University Press.

Dvidson, R. and J. G. MacKinnon (2004). *Econometric Theory and Methods*. New York: Oxford University Press.

Efron, B. (1991). Regression percentiles using asymmetric squared error loss. *Statistica Sinica* **1**(1), 93–125.

Elliott, G. and J. H. Stock (1994). Inference in time series regression when the order of integration of a regressor is unknown. *Econometric Theory* **10**(3-4), 672–700.

Embrechts, P., C. Klüppelberg, and T. Mikosch (1997). *Modeling Extremal Events For Finance and Insurance*. New York: Springer-Verlag.

Engle, R. F. and C. W. J. Granger (1987). Co-integration and error correction: Representation, estimation and testing. *Econometrica* **55**(2), 251–276.

Engle, R. F., C. W. J. Granger, J. Rice, and A. Weiss (1986). Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American statistical Association* **81**(394), 310–320.

Engle, R. F., T. Ito, and W.-L. Lin (1990). Meteor showers or heat waves? heteroskedastic intra-daily volatility in the foreign exchange market. *Econometrica* **58**(3), 525–542.

Engle, R. F. and S. Manganelli (2004). Caviar: Conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics* **22**(4), 367–381.

Epstein, L. G. and S. E. Zin (1989). Substitution, risk aversion, and the temporal behavior of consumption and asset returns: A theoretical framework. *Econometrica* **57**(4), 937–969.

Falk, M. (1983). Relative efficiency and deficiency of kernel type estimators of smooth distribution functions. *Statistica Neerlandica* **37**(2), 73–83.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiency. *Annals of Statistics* **21**(1), 196–216.

Fan, J., T. Gasser, I. Gijbels, M. Brockmann, and J. Engel (1997). Local polynomial regression: Optimal kernels and asymptotic minimax efficiency. *Annals of the Institute of Statistical Mathematics* **49**(1), 79–99.

Fan, J. and I. Gijbels (1996). *Local Polynomial Modeling and Its Applications*. London: Chapman & Hall.

Fan, J. and J. Gu (2003). Semiparametric estimation of value at risk. *Econometrics Journal* **6**(2), 261–290.

Fan, J., W. Härdle, and E. Mammen (1998). Direct estimation of low dimensional components in additive models. *Annals of Statistics* **26**(3), 943–971.

Fan, J., N. E. Heckman, and M. P. Wand (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *Journal of the American Statistical Association* **90**(429), 141–150.

Fan, J., T.-C. Hu, and Y. K. Truong (1994). Robust non-parametric function estimation. *Scandinavian Journal of Statistics* **21**(4), 433–446.

Fan, J. and T. Huang (2005). Profile likelihood inferences on semiparametric varying- coefficient partially linear models. *Bernoulli* **11**(6), 1031–1057.

Fan, J., K. Imai, I. Lee, H. Liu, Y. Ning, and X. Yang (2023). Optimal covariate balancing conditions in propensity score estimation. *Journal of Business & Economic Statistics* **41**(1), 97–110.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**(456), 1348–1360.

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

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

Fan, J., Q. Yao, and Z. Cai (2003). Adaptive varying-coefficient linear models. *Journal of the Royal Statistical Society: Series B* **65**(1), 57–80.

Fan, J., Q. Yao, and H. Tong (1996). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika* **83**(1), 189–206.

Fan, J. and C. Zhang (2003). A re-examination of diffusion estimators with applications to financial model validation. *Journal of the American Statistical Association* **98**(461), 118–134.

Fan, J., C. Zhang, and J. Zhang (2001). Generalized likelihood ratio statistics and wilks phenomenon. *Annals of Statistics* **29**(1), 153–193.

Fan, Y. and Q. Li (1996). Consistent model specification tests: Omitted variables and semiparametric functional forms. *Econometrica* **64**(4), 865–890.

Fang, Y., S. Tang, Z. Cai, and M. Lin (2020). An alternative testing for conditional unconfoundedness using auxiliary variables. *Economics Letters* **194**(109310), 1–5.

Firpo, S. (2007). Efficient semiparametric estimation of quantile treatment effects. *Econometrica* **75**(1), 259–276.

Fong, C., C. Hazlett, and K. Imai (2018). Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *Annals of Applied Statistics* **12**(1), 156–177.

Frey, R. and A. J. McNeil (2002). VaR and expected shortfall in portfolios of dependent credit risks: Conceptual and practical insights. *Journal of Banking and Finance* **26**(7), 1317–1334.

Gallant, A. R., D. A. Hsieh, and G. E. Tauchen (1991). On fitting a recalcitrant series: The pound/dollar exchange rate, 1974-1983. In W. Barnett, J. Powell, and G. Tauchen (Eds.), *Nonparametric And Semiparametric Methods in Econometrics and Statistics*, pp. 199–240. Cambridge, UK: Cambridge University Press.

Gallant, A. R. and D. W. Nychka (1987). Seminonparametric maximum likelihood estimation. *Econometrica* **55**(2), 363–390.

Galvao, A. F. and G. Montes-Rojas (2025). Multivariate quantile regression. Working Paper, Department of Economics, Michigan State University.

Gasser, T. and H. G. Müller (979). Kernel estimation of regression functions. In *Smoothing Techniques for Curve Estimation*, pp. 23–68. Berlin: Springer-Verlag.

Gasser, T., H. G. Müller, and V. Mammitzsch (1985). Kernels for nonparametric curve estimation. *Journal of the Royal Statistical Society, Series B* **47**(2), 238–252.

Genon-Catalot, V., T. Jeantheau, and C. Larédo (2000). Stochastic volatility models as hidden Markov models and statistical applications. *Bernoulli* **6**(6), 1051–1079.

Gilley, O. W., R. K. Pace, and at al. (1996). On the harrison and rubinfeld data. *Journal of Environmental Economics and Management* **31**(3), 403–405.

Glynn, A. N. and K. M. Quinn (2010). An introduction to the augmented inverse propensity weighted estimator. *Political Analysis* **18**(1), 36–56.

Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio (2020). Generative adversarial networks. *Communications of the ACM* **63**(11), 139–144.

Gorodetskii, V. V. (1977). On the strong mixing property for linear sequences. *Theory of Probability and Its Applications* **22**(2), 411–413.

Graham, B. S., C. C. de Xavier Pinto, and D. Egel (2012). Inverse probability tilting for moment condition models with missing data. *Review of Economic Studies* **79**(3), 1053–1079.

Granger, C. W. J. and T. Teräsvirta (1993). *Modeling Nonlinear Economic Relationships*. Oxford, UK: Oxford University Press.

Granger, C. W. J., H. White, and M. Kamstra (1989). Interval forecasting: An analysis based upon ARCH-quantile estimators. *Journal of Econometrics* **40**(1), 87–96.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* **66**(2), 315–331.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis* **20**(1), 25–46.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. New York: Springer-Verlag.

Hall, P. and C. C. Heyde (1980). *Martingale Limit Theory and its Applications*. New York: Academic Press.

Hall, P., J. S. J. S. Racine, and Q. Li (2004). Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association* **99**(468), 1015–1026.

Hall, P. and I. Johnstone (1992). Empirical functionals and efficient smoothing parameter selection. *Journal of the Royal Statistical Society: Series B* **54**(2), 475–509.

Hall, P. and T. E. Wehrly (1991). A geometrical method for removing edge effects from kernel-type nonparametric regression estimators. *Journal of the American Statistical Association* **86**(415), 665–672.

Hall, P., R. C. Wolff, and Q. Yao (1999). Methods for estimating a conditional distribution function. *Journal of the American Statistical Association* **94**(445), 154–163.

Hall, P. and Q. Yao (2005). Approximating conditional distribution functions using dimension reduction. *Annals of Statistics* **33**(3), 1404–1421.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton, NJ: Princeton University Precess.

Hansen, P. L. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50**(4), 1029–1054.

Hansen, P. L. (2001). Generalized method of moments estimation: a time series perspective. Working paper, Department of Economics, University of Chicago.

Härdle, W., P. Hall, and H. Ichimura (1993). Optimal smoothing in single-index models. *Annals of Statistics* **21**(1), 157–178.

Härdle, W. and P. Vieu (1992). Kernel regression smoothing of time series. *Journal of Time Series Analysis* **13**(3), 209–232.

Harrison, D. and D. L. Rubinfeld (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management* **5**(1), 81–102.

Hastie, T. and R. Tibshirani (1990). *Generalized Additive Models*. London: Chapman and Hall.

Hazlett, C. (2020). Kernel balancing. *Statistica Sinica* **30**(3), 1155–1189.

He, X. and P. Ng (1999). Quantile splines with several covariates. *Journal of Statistical Planning and Inference* **75**(2), 343–352.

He, X., P. Ng, and S. Portnoy (1998). Bivariate quantile smoothing splines. *Journal of the Royal Statistical Society: Series B* **60**(3), 537–550.

He, X. and S. Portnoy (2000). Some asymptotic results on bivariate quantile splines. *Journal of Statistical Planning and Inference* **91**(2), 341–349.

Heckman, J., H. Ichimura, and P. Todd (1998). Matching as an econometric evaluation estimator. *Review of Economic Studies* **65**(2), 261–294.

Hirano, K., G. W. Imbens, and G. Ridder (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71**(4), 1161–1189.

Hjort, N. L. and M. C. Jones (1996a). Better rules of thumb for choosing bandwidth in density estimation. Working Paper, Department of Mathematics, University of Oslo, Norway.

Hjort, N. L. and M. C. Jones (1996b). Locally parametric nonparametric density estimation. *Annals of Statistics* **24**(4), 1619–1647.

Honda, T. (2000). Nonparametric estimation of a conditional quantile for α-mixing processes. *Annals of the Institute of Statistical Mathematics* **52**(3), 459–470.

Honda, T. (2004). Quantile regression in varying coefficient models. *Journal of Statistical Planning and Inferences* **121**(1), 113–125.

Hong, S., Z. Zhang, and Z. Cai (2021). Testing heteroskedasticity for predictive regressions with nonstationary regressors. *Economics Letters* **201**(109781), 1–4.

Hong, Y. and T. H. Lee (2003). Inference and forecast of exchange rates via generalized spectrum and nonlinear times series models. *Reviews of Economics and Statistics* **85**(4), 1048–1062.

Hong, Y. and H. Li (2005). Nonparametric specification testing for continuous-time models with applications to term structure of interest rates. *Review of Financial Studies* **18**(1), 37–84.

Horowitz, J. L. and S. Lee (2005). Nonparametric estimation of an additive quantile regression model. *Journal of the American Statistical Association* **100**(472), 1238–1249.

Hsiao, C. (2025). A selective review of panel approaches to construct counterfactuals. *Empirical Economics* **69**(5), 2589–2608.

Hsiao, C., S. Ching, and S. Wan (2012). A panel data approach for program evaluation: Measuring the benefits of political and economic integration of Hong kong with Mainland China. *Journal of Applied Econometrics* **27**(5), 705–740.

Hsu, Y. C., T. C. Lai, and R. P. Lieli (2022). Counterfactual treatment effects: Estimation and inference. *Journal of Business & Economic Statistics* **40**(1), 240–255.

Huang, D., C. Schlag, I. Shaliastovich, and J. Thimme (2019). Volatility-of-volatility risk. *Journal of Financial and Quantitative Analysis* **54**(6), 2432–2452.

Huang, X. and Z. Zhan (2022). Local composite quantile regression for regression discontinuity. *Journal of Business & Economic Statistics* **40**(4), 1863–1875.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics* **35**(4), 73–101.

Huling, J. and S. Mak (2024). Energy balancing of covariate distributions. *Journal of Causal Inference* **12**(20220029), 1–22.

Hürlimann, W. (2003). A Gaussian exponential approximation to some compound poisson distributions. *ASTIN Bulletin: The Journal of the IAA* **33**(1), 41–55.

Hurvich, C. M., J. S. Simonoff, and C.-L. Tsai (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society: Series B* **60**(2), 271–293.

Hurvich, C. M. and C.-L. Tsai (1989). Regression and time series model selection in small samples. *Biometrika* **76**(2), 297–307.

Ibragimov, I. A. and Y. V. Linnik (1971). *Independent and Stationary Sequences of Random Variables*. Walters-Noordhoff, Netherlands: Groningen.

Ichimuara, H. (1993). Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of Econometrics* **58**, 71–120.

Imai, K. and M. Ratkovic (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society, Series B* **76**(1), 243–263.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics* **86**(1), 4–29.

Jagannathan, R., G. Skoulajis, and Z. Wang (2002). Generalized method of moments: Applications in finance. *Journal of Business & Economic Statistics* **20**(4), 470–481.

Jeganathan, P. (2004). Convergence of functionals of sums of random variables to local times of fractional stable motions. *Annals of Probability* **32**(3), 1771–1795.

Johansen, S. (1991). Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. *Econometrica* **59**(6), 1551–1580.

Jondeau, E., E.-H. Ser-Huang Poon, and M. Rockinger (2007). *Financial Modeling under Non-Gaussian Distributions*. New York: Springer-Verlag.

Jones, M. C. (1994). Expectiles and M-quantiles are quantiles. *Statistics & Probability Letters* **20**(2), 149–153.

Jones, M. C., J. S. Marron, and S. J. Sheather (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association* **91**(433), 401–407.

Jorion, P. (2001). *Value at Risk* (2nd ed.). New York: McGraw-Hill.

Jorion, P. (2003). *Financial Risk Manager Handbook* (2nd ed.). New York: John Wiley & Sons.

Josey, K. P., E. Juarez-Colunga, F. Yang, and D. Ghosh (2021). A framework for covariate balance using Bregman distances. *Scandinavian Journal of Statistics* **48**(3), 790–816.

Juhl, T. (2005). Functional coefficient models under unit root behavior. *Econometrics Journal* **8**(2), 197–213.

Juhl, T. (2014). Nonparametric test of the predictive regression model. *Journal of Business & Economic Statistics* **32**(3), 387–394.

Jurekova, J. (1977). Asymptotic relations of m-estimates and r-estimates in linear regression model. *Annals of Statistics* **5**(3), 464–472.

Kai, B., R. Li, and H. Zou (2010). Local composite quantile regression smoothing: An efficient and safe alternative to local polynomial regression. *Journal of the Royal Statistical Society Series B* **72**(1), 49–69.

Karatzas, I. and S. E. Shreve (1991). *Brownian Motion and Stochastic Calculus* (2nd ed.). New York: Springer-Verlag.

Karlsen, H. A., T. Myklebust, and D. Tjøstheim (2007). Nonparametric estimation in a nonlinear cointegration type model. *Annals of Statistics* **35**(1), 252–299.

Karlsen, H. A. and D. Tjøstheim (2001). Nonparametric estimation in null recurrent time series. *Annals of Statistics* **29**(2), 372–416.

Karras, T., T. Aila, S. Laine, and J. Lehtinen (2018). Progressive growing of GANs for improved quality, stability, and variation. URL: https://arxiv.org/abs/1710.10196.

Karunamuni, R. J. and T. Alberts (2005). On boundary correction in kernel density estimation. *Statistical Methodology* **2**(3), 191–212.

Kasparis, I., E. Andreou, and P. C. B. Phillips (2015). Nonparametric predictive regression. *Journal of Econometrics* **185**(2), 468–494.

Khindanova, I. N. and S. T. Rachev (2000). *Value at risk: Recent advances*. Handbook on Analytic-Computational Methods in Applied Mathematics: CRC Press LLC.

Klein, R. W. and R. H. Spady (1993). An efficient semiparametric estimator for binary response models. *Econometrica* **61**(2), 387–421.

Koenker, R. (1994). Confidence intervals for regression quantiles. In P. Mandl and M. Huskova (Eds.), *Proceedings of the Fifth Prague Symposium on Asymptotic Statistics*, Heidelberg, pp. 349–359. Physica.

Koenker, R. (2000). Galton, edgeworth, frisch and prospects for quantile regression in econometrics. *Journal of Econometrics* **9**(5), 347–374.

Koenker, R. (2004). *Quantreg: An R package for quantile regression and related methods*. The Comprehensive R Archive Network website.

Koenker, R. (2005). *Quantile Regression*. New York: Cambridge University Press.

Koenker, R. and G. Bassett (1978). Regression quantiles. *Econometrica* **46**(1), 33–50.

Koenker, R. and G. Bassett (1982). Robust tests for heteroscedasticity based on regression quantiles. *Econometrica* **50**(1), 43–61.

Koenker, R., V. Chernozhukov, X. He, and L. Peng (2017). *Handbook of Quantile Regression*. Boca Raton, FL: Chapman and Hall/CRC.

Koenker, R. and K. Hallock (2000). Quantile regression: An introduction. *Journal of Economic Perspectives* **15**(1), 143–157.

Koenker, R., P. Ng, and S. Portnoy (1994). Quantile smoothing splines. *Biometrika* **81**(4), 673–680.

Koenker, R. and Z. Xiao (2002). Inference on the quantile regression process. *Econometrica* **70**, 1583–1612.

Koenker, R. and Z. Xiao (2004). Unit root quantile autoregression inference. *Journal of the American Statistical Association* **99**(467), 775–787.

Koenker, R. and Q. Zhao (1996). Conditional quantile estimation and inference for ARCH models. *Econometric Theory* **12**(5), 793–813.

Kreiss, J.-P., M. H. Neumann, and Q. Yao (1998). Bootstrap tests for simple structures in nonparametric time series regression. *Statistics and Its Interface* **1**(2), 367–380.

Kuan, C.-M., J.-H. Yeh, and Y.-C. Hsu (2009). Assessing value at risk with CARE, the conditional autoregressive expectile models. *Journal of Econometrics* **150**(2), 261–270.

Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *Annals of Statistics* **17**(3), 1217–1241.

Laïb, N. and E. Ould-Saïd (2000). A robust nonparametric estimation of autoregression function under an ergodic hypothesis. *Canadian Journal of Statistics* **28**(4), 817–828.

LeBaron, B. (1999). Technical trading rule profitability and foreign exchange intervention. *Journal of International Economics* **49**(1), 125–143.

Ledig, C., L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi (2017). Photo-realistic single image super-resolution using a generative adversarial network. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 105–114.

Lee, S., R. Okui, and Y. J. Whang (2017). Doubly robust uniform confidence band for the conditional average treatment effect function. *Journal of Applied Econometrics* **32**(7), 1207–1225.

Lehmann, E. L. (1966). Some concepts of dependence. *Annals of Mathematical Statistics* **37**(5), 1137–1153.

Lehmann, E. L. and J. P. Romano (2005). *Testing Statistical Hypotheses*. Berlin: Springer-Verlag.

Lejeune, M. G. and P. Sarda (1988). Quantile regression: A nonparametric approach. *Computational Statistics & Data Analysis* **6**(3), 229–239.

Lewbel, A. (2007). A local generalized method of moments estimator. *Economic Letters* **94**(1), 124–128.

Li, B. (2001). On quasi likelihood equations with nonparametric weights. *Scandinavian Journal of Statistics* **28**(4), 577–602.

Li, F., K. L. Morgan, and A. M. Zaslavsky (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association* **113**(521), 390–400.

Li, K. T. (2020). Statistical inference for average treatment effects estimated by synthetic control methods. *Journal of the American Statistical Association* **115**(532), 2068–2083.

Li, K. T. and D. R. Bell (2017). Estimation of average treatment effects with panel data: Asymptotic theory and implementation. *Journal of Econometrics* **197**(1), 65–75.

Li, Q. (1999). Consistent model specification tests for time series econometric models. *Journal of Econometrics* **92**(1), 101–147.

Li, Q., C. J. Huang, D. Li, and T.-T. Fu (2002). Semiparametric smooth coefficient models. *Journal of Business & Economic Statistics* **20**(3), 412–422.

Li, Q. and J. S. Racine (2003). Nonparametric estimation of distributions with both categorical and continuous data. *Journal of Multivariate Analysis* **86**(2), 266–292.

Li, Q. and J. S. Racine (2008). Nonparametric estimation of conditional cdf and quantile functions with mixed categorical and continuous data. *Journal of Business & Economic Statistics* **26**(4), 423–434.

Li, Q. and R. S. Racine (2007). *Nonparametric Econometrics, Theory and Practice*. New York: Princeton University Press.

Linton, O. and J. P. Nielsen (1995). A kernel method of estimating structured non- parametric regression based on marginal integration. *Biometrika* **82**(1), 93–100.

Linton, O., M. H. Seo, and Y.-J. Whang (2023). Testing stochastic dominance with many conditioning variables. *Journal of Econometrics* **235**(2), 507–527.

Linton, O. B. (1997). Miscellanea efficient estimation of additive nonparametric regression models. *Biometrika* **84**(2), 469–473.

Linton, O. B. (2000). Efficient estimation of generalized additive nonparametric regression models. *Econometric Theory* **16**(4), 502–523.

Loader, C. R. (1996). Local likelihood density estimation. *Annals of Statistics* **24**(4), 1602–1618.

Louizos, C., U. Shalit, J. M. Mooij, D. Sontag, R. Zemel, and M. Welling (2017). Causal effect inference with deep latent-variable models. *Proceedings of the 31st International Conference on Neural Information Processing Systems* **30**(NIPS'17), 6449–6459.

Lu, Z., Y. Hui, and Q. Zhao (2000). Local linear quantile regression under dependence: Bahadur representation and application. Working Paper, Department of Management Sciences, City University of Hong Kong.

Lv, J. and J. Li (2022). High-dimensional varying index coefficient quantile regression model. *Statistica Sinica* **32**(4), 673–694.

Machado, J. A. F. (1993). Robust model selection and m-estimation. *Econometric Theory* **9**(3), 478–493.

Maier, M. (2011). Tests for distributional treatment effects under unconfoundedness. *Economics Letters* **110**(1), 49–51.

Mammen, E., O. Linton, and J. Nielsen (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Annals of Statistics* **27**(5), 1443–1490.

Mammitzsch, V. (1984). On the asymptotically optimal solution within a certain class of kernel type estimators. *Statistics and Risk Modeling* **2**(3-4), 247–256.

Mao, X., Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley (2017). Least squares generative adversarial networks. *Proceedings of the IEEE International Conference on Computer Vision*, 2794–2802.

Marcus, M. and J. Rosen (2006). *Markov Processes, Gaussian Processes, and Local Times*. New York: Cambridge University Press.

Marron, J. S. and D. Ruppert (1994). Transformations to reduce boundary bias in kernel density estimation. *Journal of the Royal Statistical Society: Series B* **56**(4), 653–671.

Masry, E. (1996a). Multivariate local polynomial regression estimation for time series: Uniform strong consistency and rates. *Journal of Time Series Analysis* **17**(6), 571–599.

Masry, E. (1996b). Multivariate regression estimation: Local polynomial fitting for time series. *Stochastic Processes and Their Applications* **65**(1), 81–101.

Masry, E. and J. Fan (1997). Local polynomial estimation of regression functions for mixing processes. *Scandinavian Journal of Statistics* **24**(1), 165–179.

Masry, E. and D. Tjøstheim (1995). Nonparametric estimation and identification of nonlinear ARCH time series: Strong convergence and asymptotic normality. *Econometric Theory* **11**(1), 258–289.

Masry, E. and D. Tjøstheim (1997). Additive nonlinear arx time series and projection estimates. *Econometric Theory* **13**(1), 214–252.

McLeish, D. L. (1975). A maximal inequality and dependent strong laws. *Annals of probability* **3**(5), 829–839.

McNeil, A. J. (1997). Estimating the tails of loss severity distributions using extreme value theory. *ASTIN Bulletin: The Journal of the IAA* **27**(1), 117–137.

Merlevéde, F., M. Peligrad, and S. Utev (2006). Recent advances in invariance principles for stationary sequences. *Probability Surveys* **3**(1), 1–36.

Mirza, M. and S. Osindero (2014). Conditional generative adversarial nets. URL: https://arxiv.org/abs/1411.1784.

Morgan, J. P. (1995). *Riskmetrics – Technical Manual* (3rd ed.). New York.

Morgan, J. P. (1996). *Risk Metrics – Technical Documents* (4th ed.). New York.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and Its Applications* **9**(1), 141–142.

Newey, W. (1994). The asymptotic variance of semiparametric estimators. *Econometrica* **62**(6), 349–1382.

Newey, W. K. and J. L. Powell (1987). Asymmetric least squares estimation and testing. *Econometrica* **55**(4), 819–847.

Newey, W. K. and K. D. West (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* **55**(3), 703–708.

Nowozin, S., B. Cseke, and R. Tomioka (2016). f-GAN: Training generative neural samplers using variational divergence minimization. *Proceedings of the 30th International Conference on Neural Information Processing Systems* **29**(NIPS'16), 271–279.

Oakes, D. and T. Dasu (1990). A note on residual life. *Biometrika* **77**(2), 409–410.

Øksendal, B. (1985). *Stochastic Differential Equations: An Introduction with Applications* (3rd ed.). New York: Springer-Verlag.

Opsomer, J. D. and D. Ruppert (1998). A fully automated bandwidth selection method for fitting additive models. *Journal of the American Statistical Association* **93**(442), 605–619.

Ouyang, M. and Y. Peng (2015). The treatment-effect estimation: A case study of the 2008 economic stimulus package of China. *Journal of Econometrics* **188**(2), 545–557.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75**(2), 237–249.

Pace, R. K. (1993). Nonparametric methods with applications to hedonic models. *Journal of Real Estate Finance and Economics* **7**(3), 185–204.

Pace, R. K. and O. W. Gilley (1997). Using the spatial configuration of the data to improve estimation. *The Journal of Real Estate Finance and Economics* **14**(3), 333–340.

Park, J. Y. (2002). Nonstationary nonlinear heteroskedasticity. *Journal of Econometrics* **110**(2), 383–415.

Park, J. Y. and S. B. Hahn (1999). Cointegrating regressions with time varying coefficients. *Econometric Theory* **15**(5), 664–703.

Park, J. Y. and P. C. B. Phillips (1999). Asymptotics for nonlinear transformations of integrated time series. *Econometric Theory* **15**(3), 269–298.

Park, J. Y. and P. C. B. Phillips (2001). Nonlinear regressions with integrated time series. *Econometrica* **69**(1), 117–161.

Parzen, E. (1962). On estimation of a probability of density function and mode. *Annals of Mathematical Statistics* **33**(3), 1065–1076.

Peng, H. and T. Huang (2011). Penalized least squares for single index models. *Journal of Statistical Planning and Inference* **141**(4), 1362–1379.

Phillips, P. C. B. (2001). Trending time series and macroeconomic activity: Some present and future challenges. *Journal of Econometrics* **100**(1), 21–27.

Phillips, P. C. B. (2009). Local limit theory and spurious nonparametric regression. *Econometric Theory* **25**(6), 1466–1497.

Phillips, P. C. B., D. Li, and J. Gao (2013). Estimating smooth structural change in cointegration models. Discussion Paper No. 1910, Cowles Foundations, Yale University.

Phillips, P. C. B. and S. Ouliaris (1990). Asymptotic properties of residual based tests for cointegration. *Econometrica* **58**(1), 165–193.

Phillips, P. C. B. and J. Y. Park (1998). Nonstationary density and kernel autoregression. Discussion Paper, No. 1181, Cowles Foundation, Yale University.

Pietrosanu, M., G. J., L. Kong, B. Jiang, and D. Niu (2017). cqrReg: An R Package for quantile and composite quantile regression and variable selection. arXiv:1709.04126v1.

Politis, D. N. and J. P. Romano (1992). A circular block-resampling procedure for stationary data. In R. Lepage and L. Billard (Eds.), *Exploring the Limits of Bootstrap*, pp. 263–270. New York: John Wiley & Sons.

Politis, D. N. and H. White (2004). Automatic block-length selection for the dependent bootstrap. *Econometric Reviews* **23**(1), 53–70.

Polk, C., S. Thompson, and T. Vuolteenaho (2006). Cross-sectional forecasts of the equity premium. *Journal of Financial Economics* **81**(1), 101–141.

Powell, J. L., J. H. Stock, and T. M. Stoker (1989). Semiparametric estimation of index coefficients. *Econometrica* **34**(3), 1403–1430.

Priestley, M. B. and M. Chao (1972). Non-parametric function fitting. *Journal of the Royal Statistical Society: Series B* **34**(3), 385–392.

Pritsker, M. (1998). Nonparametric density estimation and tests of continuous time interest rate models. *Review of Financial Studies* **11**(3), 449–487.

Radford, A., L. Metz, and S. Chintala (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. URL: https://arxiv.org/abs/1511.06434.

Reiss, R. D. (1981). Nonparametric estimation of smooth distribution functions. *Scandinavia Journal of Statistics* **8**(2), 116–119.

Rice, J. (1984). Bandwidth choice for nonparametric regression. *Annals of Statistics* **12**(4), 1215–1230.

Robins, J. M., R. Rotnitzky, and L. P. Zhao (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* **89**(427), 846–66.

Robinson, P. M. (1983). Nonparametric estimators for time series. *Journal of Time Series Analysis* **4**(3), 185–297.

Robinson, P. M. (1984). Robust nonparametric autoregression. In *Lecture Notes in Statistics, No. 26*, pp. 247–255. New York: Springer-Verlag.

Robinson, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica* **56**(4), 931–954.

Rockafellar, R. T., S. Uryasev, et al. (2000). Optimization of conditional value-at-risk. *Journal of Risk* **2**(1), 21–42.

Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American statistical Association* **82**(398), 387–394.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics* **27**(3), 832–837.

Rossi, B. (2007). Expectation hypothesis tests and predictive regressions at long horizons. *Econometrics Journal* **10**(1), 1–26.

Rothe, C. (2010). Nonparametric estimation of distributional policy effects. *Journal of Econometrics* **155**(1), 56–70.

Roussas, G. G. (1969a). Nonparametric estimation in Markov processes. *Annals of the Institute of Mathematical Statistics* **21**(1), 72–87.

Roussas, G. G. (1969b). Nonparametric estimation of the transition function of a Markov process. *Annals of Mathematical Statistics* **40**(4), 1386–1400.

Roussas, G. G. (1990). Nonparametric regression estimation under mixing conditions. *Stochastic Processes and Their Applications* **36**(1), 107–116.

Roussas, G. G. (1991). Estimation of transition distribution function and its quantiles in markov processes: Strong consistency and asymptotic normality. In *Nonparametric Functional Estimation and Related Topics*, pp. 443–462. Berlin: Springer-Verlag.

Rousseeuw, P. J. and A. M. Leroy (1987). *Robust regression and outlier detection*. New York: John Wiley & sons.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**(5), 688–701.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavia Journal of Statistics* **9**(2), 65–78.

Ruppert, D. and R. J. Carroll (1980). Trimmed least squares estimation in the linear model. *Journal of the American Statistical Association* **75**(372), 828–838.

Ruppert, D., S. J. Sheather, and M. P. Wand (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association* **90**(432), 1257–1270.

Ruppert, D. and M. P. Wand (1994). Multivariate locally weighted least squares regression. *Annals of Statistics* **22**(4), 1346–1370.

Samanta, M. (1989). Non-parametric estimation of conditional quantiles. *Statistics & Probability Letters* **7**(5), 407–412.

Scaillet, O. (2004). Nonparametric estimation and sensitivity analysis of expected short- fall. *Mathematical Finance* **14**(1), 115–129.

Scaillet, O. (2005). Nonparametric estimation of conditional expected shortfall. *Insurance and Risk Management Journal* **74**(1), 639–660.

Schuster, E. F. (1985). Incorporating support constraints into nonparametric estimators of densities. *Communications in Statistics - Theory and methods* **14**(5), 1123–1136.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464.

Şentürk, D. and H.-G. Müller (2006). Inference for covariate adjusted regression via varying coefficient models. *Annals of Statistics* **34**(2), 654–679.

Sercu, P., R. Uppal, et al. (2006). *Exchange Rate Volatility, Trade, and Capital Flows under Alternative Rate Regimes*. New York: Cambridge University Press.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. New York: John Wiley & Sons.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American statistical Association* **88**(422), 486–494.

Shao, Q. and H. Yu (1996). Weak convergence for weighted empirical processes of dependent sequences. *Annals of Probability* **24**(6), 2098–2127.

Sheather, S. J. and M. C. Jones (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society: Series B* **5**(3), 683–690.

Shiller, R. J. (1984). Smoothness priors and nonlinear regression. *Journal of the American Statistical Association* **72**(376), 420–423.

Speckman, P. (1988). Kernel smoothing partial linear models. *Journal of Royal Statistical Society, Series B* **50**(3), 413–426.

Sperlich, S., D. Tjøstheim, and L. Yang (2002). Nonparametric estimation and testing of interaction in additive models. *Econometric Theory* **18**(2), 197–251.

Stanton, R. (1997). A nonparametric model of term structure dynamics and the market price of interest rate risk. *Journal of Finance* **52**(5), 1973–2002.

Stone, C. J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *Annals of Statistics* **12**(4), 1285–1297.

Süli, E. and D. F. Mayers (2003). *An Introduction to Numerical Analysis*. New York: Cambridge University Press.

Sun, Y., Z. Cai, and Q. Li (2013). Semiparametric functional coefficient models with integrated covariates. *Econometric Theory* **29**(3), 659–672.

Sun, Y., Z. Cai, and Q. Li (2016). A consistent nonparametric test on semiparametric smooth coefficient models with integrated variables. *Econometric Theory* **32**(4), 988–1022.

Sun, Y. and Q. Li (2011). Data-driven method selecting smoothing parameters in semiparametric models with integrated time series data. *Journal of Business & Economic Statistics* **29**(4), 541–551.

Sun, Z. (1984). Asymptotic unbiased and strong consistency for density function estimator. *Acta Mathematica Sinica* **27**(4), 769–782.

Tang, S., Z. Cai, Y. Fang, and M. Lin (2021). A new quantile treatment effect model to study smoking effect on birth weight during mother's pregnancy. *Journal of Management Science and Engineering, Series B* **6**(3), 336–343.

Taylor, J. W. (2008). Estimating value at risk and expected shortfall using expectiles. *Journal of Financial Econometrics* **6**(2), 231–252.

Taylor, J. W. (2022). Forecasting Value at Risk and expected shortfall using a model with a dynamic omega ratio. *Journal of Banking and Finance* **140**(C), 106519.

Taylor, J. W. and D. W. Bunn (1999). A quantile regression approach to generating prediction intervals. *Management Science* **45**(2), 225–237.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society: Series B* **58**(1), 267–288.

Tibshirani, R. and T. Hastie (1987). Local likelihood estimation. *Journal of the American Statistical Association* **82**(398), 559–567.

Tjøstheim, D. and B. H. Auestad (1994a). Nonparametric identification of nonlinear time series: Projections. *Journal of the American Statistical Association* **89**(428), 1398–1409.

Tjøstheim, D. and B. H. Auestad (1994b). Nonparametric identification of nonlinear time series: Selecting significant lags. *Journal of the American Statistical Association* **89**(428), 1410–1419.

Torous, W., R. Valkanov, and S. Yan (2004). On predicting stock returns with nearly integrated explanatory variables. *Journal of Business* **77**(4), 937–966.

Truong, Y. (1992). Robust nonparametric time series regression. *Journal of Multivariate Analysis* **41**(2), 163–177.

Truong, Y. K. (1989). Asymptotic properties of kernel estimators based on local medians. *Annals of Statistics* **17**(2), 606–617.

Truong, Y. K. and C. J. Stone (1992). Nonparametric function estimation involving time series. *Annals of Statistics* **20**(1), 77–97.

Tsay, R. S. (2002). *Analysis of Financial Time Series* (2nd ed.). New York: John Wiley & Sons.

Tseng, P. and D. P. Bertsekas (1987). Relaxation methods for problems with strictly convex separable costs and linear constraints. *Mathematical Programming* **38**(3), 303–321.

Tsiatis, A. (2006). *Semiparametric Theory and Missing Data*. New York: Springer-Verlag.

van Dijk, D., T. Teräsvirta, and P. H. Franses (2002). Smooth transition autoregressive models: A survey of recent developments. *Econometric Reviews* **21**(1), 1–47.

Vogt, M. (2012). Nonparametric regression for locally stationary time series. *Annals of Statistics* **40**(5), 2601–2633.

Volkonskii, V. A. and Y. A. Rozanov (1959). Some limit theorems for random functions. I. *Theory of Probability and Its Applications* **4**(1), 178–197.

Wan, S. K., Y. Xie, and C. Hsiao (2018). Panel data approach versus synthetic control method. *Economics Letters* **164**(C), 121–123.

Wand, M. P. and M. C. Jones (1994). *Kernel Smoothing*. London: Chapman and Hall.

Wand, M. P., J. S. Marron, and D. Ruppert (1991). Transformations in density estimation (with discussion). *Journal of the American Statistical Association* **86**(414), 343–353.

Wang, K. Q. (2003). Asset pricing with conditioning information: A new test. *Journal of Finance* **58**(1), 161–196.

Wang, Q. and P. C. B. Phillips (2009a). Asymptotic theory for local time density estimation and nonparametric cointegrating regression. *Econometric Theory* **25**(3), 710–738.

Wang, Q. and P. C. B. Phillips (2009b). Structural nonparametric cointegrating regression. *Econometrica* **77**(6), 1901–1948.

Wang, Q. and P. C. B. Phillips (2012). A specification test for nonlinear nonstationary models. *Annuals of Statistics* **40**(2), 727–758.

Wang, Q. and X. Yin (2008). A nonlinear multi-dimensional variable selection method for high dimensional data: Sparse MAVE. *Computational Statistics & Data Analysis* **52**(9), 4512–4520.

Wang, T., P. Xu, and L. Zhu (2013). Penalized minimum average variance estimation. *Statistica Sinica* **23**(2), 543–569.

Wang, Y. and J. R. Zubizarreta (2020). Minimal dispersion approximately balancing weights: asymptotic properties and practical considerations. *Biometrika* **107**(1), 93–105.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā* **26**(4), 359–372.

Wei, Y. and X. He (2006). Conditional growth charts (with discussions). *Annals of Statistics* **34**(5), 2069–2097.

Wei, Y., A. Pere, R. Koenker, and X. He (2006). Quantile regression methods for reference growth charts. *Statistics in Medicine* **25**(8), 1369–1382.

Whang, Y. J. (2019). *Econometric Analysis of Stochastic Dominance: Concepts, Methods, Tools, and Applications*. London: Cambridge University Press.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* **48**(4), 817–838.

Withers, C. S. (1981). Conditions for linear processes to be strong-mixing. *Zeitschrift fur Wahrscheinlichkeitstheorie verwandte Gebiete* **57**(4), 477–480.

Wong, R. K. and K. C. G. Chan (2018). Kernel-based covariate functional balancing for observational studies. *Biometrika* **105**(1), 199–213.

Wu, J., C. Zhang, T. Xue, B. Freeman, and J. Tenenbaum (2016). Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. *Proceedings of the 30th International Conference on Neural Information Processing System* **29**(NIPS'16), 82–90.

Wu, L. (2013). *A Consistent Nonparametric Test When Regressors Are Nonstationary*. Ph. D. thesis, Department of Mathematics and Statistics, University of North Carolina at Charlotte.

Wu, T. Z., K. Yu, and Y. Yu (2010). Single-index quantile regression. *Journal of Multivariate Analysis* **101**(7), 1607–1621.

Xia, Y. (2006). Asymptotic distributions for two estimators of the single-index model. *Econometric Theory* **22**(6), 1112–1137.

Xia, Y. and W. K. Li (1999). On the estimation and testing of functional-coefficient linear models. *Statistica Sinica* **9**(4), 735–757.

Xia, Y., H. Tong, W. K. Li, and L. Zhu (2002). An adaptive estimation of dimension reduction space (with discussions). *Journal of the Royal Statistical Society, Series B* **64**(2), 363–410.

Xiao, Z. (2009). Functional-coefficient cointegration models. *Journal of Econometrics* **152**(1), 81–92.

Xiao, Z. and R. Koenker (2009). Conditional quantile estimation for generalized autoregressive conditional heteroscedasticity models. *Journal of American Statistical Association* **104**(488), 1696–1712.

Xie, S., Y. Zhou, and A. T. Wan (2014). A varying-coefficient expectile model for estimating Value at Risk. *Journal of Business & Economic Statistics* **32**(4), 576–592.

Xu, X. (2005). *Semiparametric Quantile Dynamic Time Series Models and Their Applications*. Ph. D. thesis, Department of Mathematics and Statistics, University of North Carolina at Charlotte.

Yang, T. C., C. Shoff, A. J. Noah, N. Black, and C. S. Sparks (2014). Racial segregation and maternal smoking during pregnancy: A multilevel analysis using the racial segregation interaction index. *Social Science and Medicine* **107**(April), 26–36.

Yang, Y. and H. Zou (2015). Nonparametric multiple expectile regression via ER-Boost. *Journal of Statistical Computation and Simulation* **85**(4), 1442–1458.

Yao, L., S. Li, Y. Li, M. Huai, J. Gao, and A. Zhang (2018). Representation learning for treatment effect estimation from observational data. *Proceedings of the 32nd1 International Conference on Neural Information Processing Systems* **31**(NIPS'18), 2638–2648.

Yao, Q. and H. Tong (1996). Asymmetric least squares regression estimation: A nonparametric approach. *Journal of Nonparametric Statistics* **6**(6), 273–292.

Yoon, J., J. Jordon, and M. Van Der Schaar (2018). GANITE: Estimation of individualized treatment effects using generative adversarial nets. *International Conference on Learning Representations*, ICLR.

Yoshihara, K.-I. (1995). The Bahadour representation of sample quantiles for sequences of strongly mixing random variables. *Statistics & Probability Letters* **24**(4), 299–304.

Yu, . K. and M. C. Jones (1998). Local linear quantile regression. *Journal of the American Statistical Association* **93**(411), 228–237.

Yu, . K. and Z. Lu (2004). Local linear additive quantile regression. *Scandinavian Journal of Statistics* **31**(3), 333–346.

Yu, L., W. Zhang, J. Wang, and Y. Yu (2017). SeqGAN: Sequence generative adversarial nets with policy gradient. *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence* **31**(AAAI'17), 2852–2858.

Zeger, S. L. and B. Qaqish (1988). Markov regression models for time series: A quasi-likelihood approach. *Biometrika* **44**(4), 1019–1031.

Zeng, P., T. He, and Y. Zhu (2012). A LASSO-type approach for estimation and variable selection in single index models. *Journal of Computational and Graphical Statistics* **21**(1), 92–109.

Zhang, H., I. Goodfellow, D. Metaxas, and A. Odena (2019). Self-attention generative adversarial networks. *Proceedings of the 36th International Conference on Machine Learning* **72**(PMLR97), 7354–7363.

Zhang, H., L. Huang, and L. L. Liu (2020). On Bootstrap consistency of MAVE for single index models. *Computational Statistics & Data Analysis* **141**(C), 28–39.

Zhang, S. and R. J. Karunamuni (1998). On kernel density estimation near endpoints. *Journal of Statistical Planning and Inference* **70**(2), 301–316.

Zhang, Y., Z. Gan, K. Fan, Z. Chen, R. Henao, D. Shen, and L. Carin (2017). Adversarial feature matching for text generation. *Proceedings of the 34th International Conference on Machine Learning*.

Zhao, Q. (2019). Covariate balancing propensity score by tailored loss functions. *Annals of Statistics* **47**(2), 965–993.

Zheng, J. X. (1996). A consistent test of functional form via nonparametric estimation techniques. *Journal of Econometrics* **75**(2), 263–289.

Zheng, S. (2011). Gradient descent algorithms for quantile regression with smooth approximation. *International Journal of Machine Learning and Cybernetics* **2**(2), 191–207.

Zhou, K. Q. and S. L. Portnoy (1996). Direct use of regression quantiles to construct confidence sets in linear models. *Annals of Statistics* **24**(1), 287–306.

Zhu, F., M. Liu, S. Ling, and Z. Cai (2023). Testing for structural change of predictive regression model to threshold predictive regression model. *Journal of Business & Economic Statistics* **41**(1), 228–240.

Ziegel, J. F. (2016). Coherence and elicitability. *Mathematical Finance* **26**(4), 901–918.

Zou, H. and M. Yuan (2008). Composite quantile regression and the oracle model selection theory. *Annals of Statistics* **36**(3), 1108–1126.

Zubizarreta, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association* **110**(511), 910–922.