The Pennsylvania State University
The Graduate School
Eberly College of Science

SEMIPARAMETRIC ESTIMATION AND INFERENCE FOR CONDITIONAL VALUE-AT-RISK AND EXPECTED SHORTFALL

A Dissertation in
Statistics
by
Chuan-Sheng Wang

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

May 2018

The dissertation of Chuan-Sheng Wang was reviewed and approved* by the following:

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

Runze Li
Eberly Family Chair in Statistics

Fuqing Zhang
Professor of Meteorology

Lingzhou Xue
Assistant Professor of Statistics

Naomi S. Altman
Professor of Statistics
Head of Graduate Program

*Signatures are on file in the Graduate School.

Abstract

Conditional Value-at-Risk (hereafter, CVaR) and Expected Shortfall (CES) play an important role in financial risk management. Parametric CVaR and CES enjoy both nice interpretation and capability of multi-dimensional modeling, however they are subject to errors from mis-specification of the noise distribution. On the other hand, nonparametric estimations are robust but suffer from the "curse of dimensionality" and slow convergence rate.

To overcome these issues, we study semiparametric CVaR and CES estimation and inference for parametric model with nonparametric noise distribution. In this dissertation, under a general framework that allows for many widely used time series models, we propose a semiparametric CVaR estimator and a semiparametric CES estimator that both achieve the parametric convergence rate.

Asymptotic properties of the estimators are provided to support the inference. Furthermore, to draw simultaneous inference for CVaR at multiple confidence levels, we establish a functional central limit theorem for CVaR process indexed by the confidence level and use it to study the conditional expected shortfall.

A user-friendly bootstrap approach is introduced to facilitate non-expert practitioners to perform confidence interval construction for CVaR and CES. The methodology is illustrated through both Monte Carlo studies and an application to S&P 500 index.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I am greatly indebted to my advisor, Dr. Zhibiao Zhao, who guided my research and career development with enthusiasm and patience. Throughout my Ph.D. study, Dr. Zhao has been very supportive. This dissertation would not have been written without his insipiration as a great teacher and researcher.

I am very grateful to my committee, Dr. Runze Li, Dr. Lingzhou Xue, and Dr. Fuqing Zhang, who provided insightful suggestions for my research. Their advice and ideas helped me think questions from all aspects.

I wish to thank Dr. William W.S. Wei who has been a mentor to me and introduced me to statistics and time series analysis while I was an undergraduate student in National Taiwan University.

I would also like to thank Dr. Xiaoyue Niu and Dr. Kirsten Eilertson who sharpened my communication skills and my ability to convey statistical ideas to statistician and non-statistician experts.

Special thanks goes to my parents, Li-Chun Chan, Ning-Chun Wang, and my little brother Chuan-Chih Wang, who always warmly supported me in my life.

Part of the research was reproduced from Wang and Zhao (2016) and supported by a NSF grant DMS-1309213 and a NIDA grant P50-DA10075-15. The findings and conclusions do not necessarily reflect the view of the funding agency.

Chapter 1 Introduction

1.1 Background

1.1.1 Value-at-Risk

In financial portfolio management, two most important factors of interest are the average return and its associated risk. While the average return tells the investor the mean value of the return of a particular portfolio, the risk of returns concerns the downside of the portfolio, i.e., the potentially large loss when the market moves in the opposite direction. For example, if the return takes large positive value (110%, say) and negative value (-90%, say) equally likely, then the average return is 10%; however, due to the potentially large negative return, a risk-averse investor may avoid this type of double-or-none portfolio and prefer a portfolio with 5% average return but low risk. In fact, due to the importance of financial risk, financial institutions periodically monitor their risk, which forms the basis for dynamic portfolio management, to meet the supervisory guidance set by regulators.

Among many other risk measures (e.g., the standard deviation, tail conditional expectation, and entropic risk measure), Value-at-Risk (hereafter, VaR) is the most prominent risk measure. For example, VaR is the widely used risk measure by regulators in banking supervision (e.g., Scaillet, 2003); also, VaR can cover the presence of netting agreements frequently found in the banking industry (e.g., Fermanian and Scaillet, 2005). For a given portfolio, denote by Y_i the loss (i.e., negative return) at time i, its VaR is the threshold L such that

$$\mathbb{P}\{Y_i \geq L\} = 1 - \tau, \tag{1.1}$$

where $(1-\tau)$ is the confidence level, often taken to be 1% or 5%. For example, at confidence level 1%, the potential loss exceeds the VaR threshold L with probability 1%. The confidence level reflects the investor's level of tolerance of the worst scenario. A conservative investor may use a small confidence level, whereas a more aggressive investor may prefer a larger level. See Duffie and Pan (1997) and Dowd (1998) for an excellent introduction to VaR. Since (1.1) is derived from pure statistical inference of the underlying data-generating process, it is often termed as statistical VaR (Aït-Sahalia and Lo, 2000). To incorporate other aspects of market risk, Aït-Sahalia and Lo (2000) introduced the state-price density (SPD) based economic VaR. The state-price density is the density under which the price of any asset is the riskless-rate discounted expected payoff. Their economic VaR estimate is based on the Black-Scholes options pricing formula with a nonparametric estimate of the volatility function. In this dissertation we focus on statistical VaR.

Depending on the goal of the portfolio holder, another closely related VaR approach is the conditional VaR (hereafter, CVaR), which evaluates the conditional probability version of (1.1), conditioning on some available information. For example, an active trader may be very sensitive to short-term market information, such as the stock performance in the past week and some current global economic variables, and thus he/she may prefer the CVaR modeling, conditioning on the immediately available information when evaluating the probability in (1.1). We refer the reader to Chernozhukov and Umanstev (2001), Cai (2002), Fan and Gu (2003), Engle and Manganelli (2004), and Cai and Wang (2008) for various CVaR approaches. On the other hand, it may be reasonable for a retirement fund manager with a long-time vision to work under the unconditional VaR framework (1.1) as this marginal approach can avoid the unnecessarily volatile short-term market fluctuation. See Danielsson and de Vries (2000) for more discussions on these two approaches. In this dissertation we focus on the CVaR approach.

1.1.2 Expected Shortfall

Another widely used risk measure to quantify the risk of the portfolio is the expected shortfall (ES) defined as

$$\mathrm{ES}(y) = \mathbb{E}(Y_i | Y_i \geq y). \tag{1.2}$$

Here y is the threshold and is chosen as the τ-th quantile of Y_i (i.e., the VaR of Y_i at confidence level $1-\tau$) in vast literature. Intuitively, ES(y) quantifies the average loss given that the loss exceeds the threshold y, i.e., the average worst loss exceeding y. By monitoring expected shortfalls, portfolio managers can actively balance the portfolio to control the risk.

Since Artzner et al. (1999) which provides a complete definition of risks, it is well-known that the ES enjoys some nice properties that VaR does not have. First of all, ES possess subadditivity while VaR does not. For instance, suppose that we have a portfolio including A, B, and C assets. The ES of the portfolio should be equal or less than the sum of ES of individual assets, i.e. $ES_{portfolio} \leq ES_A + ES_B + ES_C$, while this inequality may not hold when ES is replaced with VaR. Thus an investor may fail to reduce his VaR by diversifying his asset allocations in the portfolio. See Frey and McNeil (2002) for more discussions on the theoretical properties of risk measures. Second, by observing (1.1) and (1.2), one can find that ES tells us more information about the potential size of losses given that it already exceeds VaR.

In practice, portfolio managers often have some covariates information, denoted by $\mathbf{X}_i \in \mathbb{R}^p$, and it is then desirable to incorporate such information into the modeling of the loss Y_i. For example, the covariates $\mathbf{X}_i$ may include the historical loss $Y_{i-1}, Y_{i-2}, \ldots$, and some global economics variables, such as the inflation rates and unemployment rates. From (1.2), define the conditional expected shortfall (CES) of Y_i given the covariates $\mathbf{X}_i = x$ (for some given vector x) as

$$\text{CES}(y|x) = \mathbb{E}(Y_i|Y_i \geq y, \mathbf{X}_i = x). \tag{1.3}$$

The CES quantifies the average conditional loss, given the covariates $\mathbf{X}_i = x$ and that the loss exceeds y. Compared to the (unconditional) ES in (1.2), the CES allows us to incorporate the covariates information $\mathbf{X}_i$ into the modeling of Y_i.

1.2 Motivation

To estimate VaR, ES, or their conditional versions, the parametric approach uses a specific parametric model (e.g., ARCH-type or GARCH-type models) with the noises following some known distribution. For example, the RiskMetrics in J. P. Morgan (1996) uses the Normal distribution; other popular choices include the

Student-t and some distributions that can be transformed to Normal (Hull and White, 1998). On the other hand, Chen, Gerlach and Lu (2012) uses GJR-GARCH volatility model with an asymmetric Laplace distribution error distribution to estimate and forecast ES and VaR. These parametric methods enjoy both nice interpretation and capability of multi-dimensional modeling, however they are subject to errors from mis-specification of the noise distribution. For example, there have been numerous discussions on whether stock returns follow Normal, Student-t, symmetric stable, or other distributions. Ait-Sahalia and Brandt (2001) pointed out the difficulty of modeling the conditional distribution of returns in practice. Hypothetically, suppose returns have mean zero and variance one. From (1.1), at confidence level $1 - \tau = 1\%$, the specification of Normal distribution would give the VaR threshold 2.33, whereas the specification of Student-t distribution with 3 degrees of freedom (normalized to have variance one) would give the quite different VaR threshold 2.62. Thus, it is desirable to develop a distribution-free method.

Nonparametric VaR or ES estimation is a robust alternative over the parametric approach. For unconditional VaR estimation, historical simulation and its variants use empirical sample quantiles or the inverse of some marginal distribution function estimate based on the historical data to predict the future VaR; see Butler and Schachter (1998), Gourieroux, Laurent and Scaillet (2000), and Chen and Tang (2005). For unconditional ES, Scaillet (2004) proposed a smoothed and distribution-free estimation based on a kernel approach. Chen (2008) further developed an unsmoothed estimator of ES based on a weighted sample average of excessive losses greater than a VaR, and compare its estimation accuracy with the smoothed estimator in Scaillet (2004). For CVaR estimation, Cai (2002), and Wu, Yu and Mitra (2007) proposed model-free nonparametric estimates based on kernel smoothing estimates of the conditional distribution function, and Cosma, Scaillet and von Sachs (2007) studied wavelets-based nonparametric estimation.

Cai and Wang (2008) use the combination of Nadaraya-Watson (NW) method of Cai (2002) and the double kernel local linear technique of Yu and Jones (1998) then estimates CES nonparametrically by plugging in the estimated conditional probability function and the estimated CVaR function, as an extension of Scaillet (2005) which uses the unweighted NW method.

As pointed out by Chen and Tang (2005), these nonparametric methods have two major advantages: (i) being distribution-free; and (ii) without imposing parametric

models, such as ARCH or GARCH models. Despite their robustness to model assumptions, nonparametric approaches have some well-known challenging issues. Essentially, nonparametric CVaR methods perform estimation in a small local window of the covariates, which may contain very few or almost no observations for high-dimensional covariates. This is the well-known "curse of dimensionality" issue. Other practically challenging issues include bandwidth selection and slow convergence rate. See Li and Racine (2007) for discussions.

This dissertation has two main contributions. Our first contribution is to propose a semiparametric CVaR estimator and a semiparametric CES estimator and establish their $\sqrt{n}$ asymptotic normality. As discussed above, both the parametric and nonparametric approaches have their strengths and weaknesses, and we propose combining their strengths via a semiparametric approach of parametric model with nonparametric noise distribution. Our semiparametric approach has several appealing features. First, the parametric model structure is capable of modeling the dependence of returns on high-dimensional covariates. This can avoid the "curse of dimensionality" issue of the nonparametric approach. Furthermore, the parametric component has the advantage of including some non-Markovian behavior (e.g., GARCH) as opposed to a pure nonparametric kernel regression approach. Second, adopting nonparametric noise distribution can avoid the error from distributional mis-specification. Third, unlike the ARCH/GARCH VaR or CES modeling, our methodology is developed under a very general framework that allows for many linear and nonlinear processes. Fourth, the proposed CVaR and CES estimators can achieve the parametric $\sqrt{n}$ convergence rate.

Despite the vast literature on VaR and CVaR estimation, little attention has been paid to calculating the standard error of the estimates; Chen and Tang (2005) studied this problem for unconditional VaR estimation. Our second contribution is to develop methodology for statistical inference of CVaR. First, we provide consistent standard error for the semiparametric CVaR estimator, which is useful in confidence interval construction. Second, to draw simultaneous inference for CVaR at multiple confidence levels, we establish a functional central limit theorem (hereafter, CLT). As an application of the functional CLT, we study semiparametric estimation of conditional expected shortfall (hereafter, CES). Third, to facilitate non-expert practitioners to construct confidence intervals for CVaR and CES, we introduce an easy-to-implement bootstrap approach. One major advantage of the

bootstrap approach is that practitioners can choose their own particular model and parameter estimation method to address semiparametric CVaR and CES inference.

The rest of this dissertation is organized as follows. Chapter 2 includes CVaR results. Section 2.1 contains main results on semiparametric CVaR estimation, asymptotic normality, and standard error calculation. Section 2.2 studies CVaR process and an application to CES. Section 2.3 introduces bootstrap inference. Section 2.4 briefly studies semiparametric conditional distribution estimation. Numerical analysis is presented in Sections 2.5 and 2.6. Assumptions and proofs of theorems are in Section 2.7. Chapter 3 includes CES results. Section 3.1 contains main results on semiparametric CES estimation, consistency, and asymptotic normality. Finally, technical conditions and proofs are in Section 3.4.

For a matrix $A = (a_{i,j})$, write $|A| = (\sum_{i,j} a_{i,j}^2)^{1/2}$. For a random vector $\mathbf{Z}$, write $\mathbf{Z} \in \mathcal{L}^q, q > 0$, if $\mathbb{E}(|\mathbf{Z}|^q) < \infty$. Throughout, $\xrightarrow{p}$ stands for convergence in probability.

Chapter 2
Conditional Value-at-Risk

2.1 Semiparametric CVaR Estimation

Let $Y_i \in \mathbb{R}$ be scalar-valued portfolio loss (i.e., negative gain) at time i. Suppose the loss Y_i depends on some p-dimensional covariates $\mathbf{X}_i \in \mathbb{R}^{1\times p}$. In practice, the covariates $\mathbf{X}_i$ may include both the historical market information, such as the past portfolio loss $Y_{i-1}, \ldots, Y_{i-q}$, and some overall exogenous economic variables $\mathbf{U}_i \in \mathbb{R}^{p-q}$, such as the inflation rates and unemployment rates. Similar to (1.1), at confidence level $(1-\tau)$, the CVaR of Y_i given $\mathbf{X}_i = x$, denoted by $\mathrm{CVaR}(1-\tau|x)$, is defined as

$$\mathrm{CVaR}(1-\tau|x) = L \quad \text{such that} \quad \mathbb{P}\{Y_i \geq L|\mathbf{X}_i = x\} = 1-\tau. \tag{2.1}$$

Therefore, conditioning on $\mathbf{X}_i = x$, the loss Y_i exceeds $\mathrm{CVaR}(1-\tau|x)$ with probability $(1-\tau)$. In particular, if $\mathbf{X}_i = (Y_{i-1}, \ldots, Y_{i-p})$, then $\mathrm{CVaR}(1-\tau|x)$ is the predicted CVaR of Y_i given $Y_{i-1}, \ldots, Y_{i-p}$. Our goal is to estimate and make inference about $\mathrm{CVaR}(1-\tau|x)$.

Denote by $Q(\tau|x) := Q(\tau|\mathbf{X}_i = x)$ the conditional τ-th quantile of Y_i given $\mathbf{X}_i = x$. From (2.1), $\mathrm{CVaR}(1-\tau|x) = Q(\tau|x)$. From now on we shall focus on $Q(\tau|x)$.

Remark 1. In this dissertation we focus on the case of fixed level $1-\tau$, even though it may be very small. An alternative approach is the extreme quantile approach, i.e., $1-\tau \to 0$ as sample size $n \to \infty$ so that in (2.1) the threshold $L \to \infty$. If we consider aggregated monthly returns as the sum of daily returns,

then the moderate deviation approach in Wu and Zhao (2008) may be applied here, but the asymptotic theory is more challenging. This is beyond the scope of the current paper and will serve as a direction for future research.

2.1.1 The nonparametric quantile regression approach

The conditional τ-th quantile $Q(\tau|x)$ can be estimated by local linear quantile regression

$$\tilde{Q}(\tau|x) = \hat{a}_0, \quad (\hat{a}_0, \hat{a}_1) = \underset{a_0\in\mathbb{R}, a_1\in\mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau\Big\{Y_i - a_0 - (\mathbf{X}_i - x)a_1\Big\} K\Big(\frac{\mathbf{X}_i - x}{b_n}\Big) \quad (2.2)$$

where $\rho_\tau(v) = v(\tau - \mathbf{1}_{v\le 0})$ is the check function, $K(\cdot)$ is a p-variate kernel function, and $b_n > 0$ is bandwidth; see, e.g., Yu and Jones (1997). Another nonparametric quantile estimation approach is based on the inverse of conditional distribution function estimate (Cai, 2002; Wu, Yu and Mitra, 2007; Cai and Wang, 2008). See Chapter 6 in Li and Racine (2007) for more discussions.

While nonparametric quantile regression is robust against the model structure, it also suffers from several drawbacks. First, due to the "curse of dimensionality", it is generally infeasible to perform nonparametric estimation for $p \ge 3$. As a result, when predicting VaR based on historical loss, nonparametric approach can use only very local recent historical information ($p = 1$ or 2 at most). Second, p-dimensional nonparametric conditional quantile estimation has convergence rate $\sqrt{nb_n^p}$, which can be quite slow as $b_n \to 0$. Third, it is a practically non-trivial issue to select the bandwidth b_n (Li and Racine, 2007).

2.1.2 The proposed semiparametric approach

By Section 2.1.1, nonparametric quantile estimation is not very appealing in CVaR estimation. In this section we propose a semiparametric approach. Specifically, we assume

$$Y_i = G(\theta, \varepsilon_i, \mathbf{X}_i), \quad (2.3)$$

where $G(\theta, \varepsilon, x)$ is a parametric function with unknown k-dimensional parameter $\theta \in \mathbb{R}^k$, and $\{\varepsilon_i\}_{i\in\mathbb{Z}}$ are unobservable i.i.d. noises that may represent unobserved

heterogeneity or technological shocks. We leave the distribution of ε_i completely unspecified, leading to a semiparametric structure. This semiparametric approach can avoid potential mis-specification on the distribution of ε_i. For example, a normal distribution may perform poorly in the presence of Cauchy distributed noises $\{\varepsilon_i\}$. On the other hand, the parametric assumption on $G(\theta, \cdot, \cdot)$ allows us to avoid the "curse of dimensionality" in nonparametric approach and thus has the capability of high-dimensional CVaR estimation.

Model (2.3) assumes that the function G can be well parameterized by some parameter θ with the distribution of ε_i unspecified. In applications, we can use a two-stage procedure to determine the function form G. In the first stage, we may use the sieve nonparametric estimation method (e.g., Chen, 2007) or nonparametric kernel estimation method (e.g., Matzkin, 2003) to estimate the function nonparametrically. In the second stage, we then check whether the estimated nonparametric function can be parameterized by some existing models. The parametric specification stage is equally important in all aspects of parametric modeling in the literature, such as model fitting and forecasting, in addition to VaR risk management. Since our goal is not to develop new specification testing methods but to develop CVaR estimation for given model, in this dissertation we assume that the researcher has decided a specific model prior to CVaR estimation. In fact, almost all existing works on parametric VaR (e.g., EWMA in RiskMetrics of J.P. Morgan, 1994; robust-EWMA in Guermat and Harris, 2001; CAViaR in Engle and Manganelli, 2004; skewed-EWMA in Gerlach, Lu and Huang, 2013) also took the same approach.

Example 1. (**Nonlinear AR models**) Let $\mathbf{X}_i = (Y_{i-1}, \ldots, Y_{i-p})$, then (2.3) becomes

$$Y_i = G(\theta, \varepsilon_i, Y_{i-1}, \ldots, Y_{i-p}), \tag{2.4}$$

a nonlinear autoregressive (AR) model of order p. An important special case of (2.4) is the class of nonlinear ARCH model

$$Y_i = \mu(\theta, Y_{i-1}, \ldots, Y_{i-p}) + \sigma(\theta, Y_{i-1}, \ldots, Y_{i-p})\varepsilon_i, \tag{2.5}$$

for parametric functions $\mu(\theta, \cdot)$ and $\sigma(\theta, \cdot) > 0$ with unknown parameter θ. Model (2.5) includes many popular nonlinear models; see Fan and Yao (2003). Also, (2.5)

includes the Euler-discretization of the continuous-time diffusion model $dY_t = \mu(\theta, Y_t)dt + \sigma(\theta, Y_t)d\mathbb{B}_t$, where $\{\mathbb{B}_t\}_{t\geq 0}$ is a Brownian motion or a general Lévy process. $\diamondsuit$

Example 2. (**Nonlinear ARX models**) A more flexible generalization of (2.4) is the nonlinear AR with exogenous/external inputs (ARX) model with $\mathbf{X}_i = (Y_{i-1}, \ldots, Y_{i-q}, \mathbf{U}_i)$:

$$Y_i = G(\theta, \varepsilon_i, Y_{i-1}, \ldots, Y_{i-q}, \mathbf{U}_i), \tag{2.6}$$

where $\mathbf{U}_i \in \mathbb{R}^{1\times(p-q)}$ are exogenous or external variables. For example, an exogenous variable can be the inflation rates or unemployment rates affecting stock returns Y_i. The classical linear ARX model is $Y_i = \sum_{j=1}^{q} \phi_j Y_{i-j} + \mathbf{U}_i\beta + \varepsilon_i$ for coefficients $\phi_1, \ldots, \phi_q \in \mathbb{R}, \beta \in \mathbb{R}^{p-q}$. Model (2.6) allows flexible nonlinear generalization. For example, (2.6) includes the ARCH model with exogenous inputs:

$$Y_i = \sum_{j=1}^{q} \phi_j Y_{i-j} + \mathbf{U}_i\beta + \varepsilon_i \left(\alpha_0^2 + \sum_{j=1}^{q} \alpha_j^2 Y_{i-j}^2 + \mathbf{U}_i^2\gamma^2\right)^{1/2}, \quad \beta, \gamma \in \mathbb{R}^{p-q}.$$

This model generalizes Engle's ARCH model to allow for exogenous variables. $\diamondsuit$

Example 3. (**Nonlinear GARCH models**) Consider the nonlinear GARCH model

$$Y_i = \sigma_i\varepsilon_i \quad \text{with} \quad \sigma_i = g(\theta, \sigma_{i-1}, \ldots, \sigma_{i-q}, Y_{i-1}, \ldots, Y_{i-r}), \tag{2.7}$$

for a parametric function $g > 0$. By specifying different forms of g, this general model includes many widely used variants of GARCH models, including the classical GARCH model (Bollerslev, 1986), the EGARCH model (Nelson, 1991), the GJR-GARCH model (Glosten, Jagannathan and Runkle, 1993), and the TGARCH model (Zakoian, 1994), just to name a few. Under appropriate conditions (e.g., Wu and Shao, 2004), by recursive iteration, σ_i admits the representation $\sigma_i = g^*(\theta, Y_{i-1}, Y_{i-2}, \ldots)$ for some function g^*. Thus (2.7) becomes the ARCH(∞) model $Y_i = g^*(\theta, Y_{i-1}, Y_{i-2}, \ldots)\varepsilon_i$ and we can take the covariates $\mathbf{X}_i = (Y_{i-1}, Y_{i-2}, \ldots)$. In practice, we need not to know the function form g^*; instead we can recursively compute $\sigma_i = g(\theta, \sigma_{i-1}, \ldots, \sigma_{i-q}, Y_{i-1}, \ldots, Y_{i-r})$ when the parameter θ is known or

can be estimated. Since the GARCH model (2.7) is non-Markovian, the "curse of dimensionality" of g^* makes it infeasible to use the nonparametric approach in Section 2.1.1 to estimate $Q(\tau|\mathbf{X}_i = x)$. $\diamondsuit$

To motivate our semiparametric estimator of $Q(\tau|x) := Q(\tau|\mathbf{X}_i = x)$, we assume that in (2.3) the function $G(\theta, \varepsilon_i, \mathbf{X}_i)$ is strictly increasing in ε_i and that ε_i is independent of $\mathbf{X}_i$. By definition, given $\mathbf{X}_i = x$, $Q(\tau|x)$ is the τ-th quantile of $Y_i = G(\theta, \varepsilon_i, \mathbf{X}_i) = G(\theta, \varepsilon_i, x)$. Note that the τ-th quantile of any strictly increasing transformation of a random variable is the same transformation of the τ-th quantile of that random variable. Therefore,

$$Q(\tau|x) = \tau\text{-th quantile of } G(\theta, \varepsilon_i, x) = G(\theta, Q_\varepsilon(\tau), x), \tag{2.8}$$

where $Q_\varepsilon(\tau)$ is the τ-th quantile function of ε_i. In practice, both θ and $Q_\varepsilon(\tau)$ are unknown, and we propose estimating $Q(\tau|x)$ by plugging some consistent estimates of θ and $Q_\varepsilon(\tau)$ into (2.8). However, the true innovations $\{\varepsilon_i\}$ are not observable. Fortunately, under the above strictly increasing assumption on $G(\theta, \varepsilon_i, \mathbf{X}_i)$ (as a function ε_i), we can invert the function to obtain ε_i. Formally, we impose Assumption 1 below.

Assumption 1. *For any given $(\theta, \mathbf{X}_i)$, the function $G(\theta, \varepsilon_i, \mathbf{X}_i)$ is strictly increasing in ε_i so the inverse $G^{-1}(\theta, \cdot, \mathbf{X}_i)$ exists and*

$$\varepsilon_i = H(\theta, Y_i, \mathbf{X}_i) \quad \textit{with} \quad H(\theta, Y_i, \mathbf{X}_i) = G^{-1}(\theta, Y_i, \mathbf{X}_i). \quad (\textit{the inverse}) \tag{2.9}$$

Assumption 1 is satisfied for many practical models. In fact, in the context of nonparametric estimation of non-additive functions, Matzkin (2003) imposed the same condition. Clearly, Assumption 1 is satisfied for the nonlinear ARCH model in (2.5) and the nonlinear GARCH model in (2.7). In addition, it is satisfied for some transformation models. Let $\Lambda(\cdot)$ be a strictly increasing transformation function. Then Assumption 1 is satisfied for the model $\Lambda(Y_i) = \mathbf{X}_i\theta + \varepsilon_i$ or equivalently $Y_i = \Lambda^{-1}(\mathbf{X}_i\theta + \varepsilon_i)$. The latter model includes the well-known Box-Cox transformation and all the transformation models studied in Horowitz (1996). Under appropriate conditions on the conditional hazard condition, duration models with unobserved heterogeneity also satisfy Assumption 1; we refer the reader to Matzkin (2003) for more details.

Under Assumption 1, in view of (2.8), we propose the following estimation procedure:

(i) Let $\hat{\theta}$ be a consistent estimate of θ. From (2.9), we can estimate the innovation ε_i by

$$\text{generalized residuals:} \quad \hat{\varepsilon}_i = H(\hat{\theta}, Y_i, \mathbf{X}_i). \tag{2.10}$$

(ii) Estimate $Q_\varepsilon(\tau)$ by $\hat{Q}_\varepsilon(\tau)$, the sample τ-th quantile of $\{\hat{\varepsilon}_i\}$. Formally,

$$\hat{Q}_\varepsilon(\tau) = \inf\Big\{z : \hat{F}_\varepsilon(z) \geq \tau\Big\}, \quad \text{where} \quad \hat{F}_\varepsilon(z) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\hat{\varepsilon}_i \leq z}. \tag{2.11}$$

(iii) Plugging $\hat{\theta}$ and $\hat{Q}_\varepsilon(\tau)$ into (2.8), we propose the following estimator:

$$\hat{Q}(\tau|x) = G(\hat{\theta}, \hat{Q}_\varepsilon(\tau), x). \tag{2.12}$$

Compared with the nonparametric quantile estimation approach in Section 2.1.1, the proposed semiparametric estimator (2.12) is easy to implement and does not require any bandwidth. To derive the asymptotic normality, a key step is to study the residual empirical process $\hat{F}_\varepsilon(z)$ in (2.11). In fact, due to the important applications in model diagnostics and hypothesis testing, the topic of residual empirical process itself has attracted much attention; see, e.g., Lee and Wei (1999) and Horváth and Teyssière (2001) for residual empirical process from AR models and ARCH models, respectively. Theorem 1 below establishes the uniform approximation of the generalized-residual empirical process. Given the general form of model (2.3), our result is more general than existing ones.

Denote by $F_\varepsilon(\cdot)$, $f_\varepsilon(\cdot)$, and $Q_\varepsilon(\cdot)$, respectively, the distribution, density, and quantile functions of ε_i. Throughout we assume that $G(\theta, \varepsilon, x)$ is continuously differentiable in θ and ε, with corresponding partial derivatives $\dot{G}_\theta(\theta, \varepsilon, x)$ and $\dot{G}_\varepsilon(\theta, \varepsilon, x)$ with respect to θ and ε, respectively.

Theorem 1. *Recall $\hat{F}_\varepsilon(z)$ in (2.11). Suppose that Assumption 1 and Assumptions 3–4 (in Section 2.7) hold. Further assume that $\hat{\theta} = \theta + O_p(n^{-1/2})$. Then for any*

given $c > 0$,

$$\sup_{|z|\le c}\left|\hat{F}_\varepsilon(z) - \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\varepsilon_i\le z} - f_\varepsilon(z)\mathbb{E}\left[\frac{\dot{G}_\theta(\theta,z,\mathbf{X}_0)}{\dot{G}_\varepsilon(\theta,z,\mathbf{X}_0)}\right]^T(\hat{\theta}-\theta)\right| = o_p(n^{-1/2}). \quad (2.13)$$

Here and hereafter $\mathbf{X}_0$ *has the same distribution as* $\mathbf{X}_i$.

By Theorem 1, the asymptotic expansion of $\hat{F}_\varepsilon(z)$ has two components: the first term $n^{-1}\sum_{i=1}^{n}\mathbf{1}_{\varepsilon_i\le z}$ is the empirical process of the true noises $\{\varepsilon_i\}$, which is the leading term, and the second term $f_\varepsilon(z)\mathbb{E}[\dot{G}_\theta(\theta,z,\mathbf{X}_0)/\dot{G}_\varepsilon(\theta,z,\mathbf{X}_0)]^T(\hat{\theta}-\theta)$ is the bias correction term due to the estimation error of $\hat{\theta}$. Therefore, in order to derive the asymptotic distribution of $\hat{Q}(\tau|x)$, it is necessary to impose some condition on $\hat{\theta}-\theta$.

Assumption 2. The estimator $\hat{\theta}\in\mathbb{R}^k$ of $\theta\in\mathbb{R}^k$ admits the Bahadur-type representation

$$\hat{\theta}-\theta = \frac{1}{n}\sum_{i=1}^{n}D(\theta,\varepsilon_i,\mathbf{X}_i) + o_{\mathbb{P}}(n^{-1/2}), \quad (2.14)$$

for some $D(\theta,\cdot,\cdot)\in\mathbb{R}^k$ satisfying $D(\theta,\varepsilon_i,\mathbf{X}_i)\in\mathcal{L}^2$ and $\mathbb{E}[D(\theta,\varepsilon_i,\mathbf{X}_i)|\mathbf{X}_i]=0$.

Assumption 2 asserts that $\hat{\theta}-\theta$ has a linear leading term plus some negligible error. This type of Bahadur representations has been established for different models in the literature. For example, Hall and Yao (2003) obtained Bahadur representation of quasi-maximum likelihood estimates for ARCH and GARCH models, and Zhao (2010) established Bahadur representation for pseudo-likelihood estimate of stochastic regression models. See Section 2.1.4 below for more discussions.

Theorem 2. *Suppose that Assumption 1 and Assumptions 3–5 (in Section 2.7) hold.*
(i) If $\hat{\theta}=\theta+O_p(n^{-1/2})$, *then* $\hat{Q}(\tau|x)$ *is* $\sqrt{n}$*-consistent, i.e.,* $\hat{Q}(\tau|x)=Q(\tau|x)+O_p(n^{-1/2})$.
(ii) If in addition Assumption 2 holds (recall $D(\theta,\varepsilon_i,\mathbf{X}_i)$ *there), then the CLT holds*

$$\sqrt{n}[\hat{Q}(\tau|x)-Q(\tau|x)] \Rightarrow N\Big(0,\Gamma(\tau)\Big), \quad (2.15)$$

where $\Gamma(\tau) = \dot{G}_\varepsilon(\theta, Q_\varepsilon(\tau), x)^2 \mathbb{E}[W_1(\tau)^2]$ *and*

$$\begin{aligned} W_i(\tau) &= \frac{\tau - \mathbf{1}_{\varepsilon_i < Q_\varepsilon(\tau)}}{f_\varepsilon(Q_\varepsilon(\tau))} \\ &+ \left\{ \frac{\dot{G}_\theta(\theta, Q_\varepsilon(\tau), x)}{\dot{G}_\varepsilon(\theta, Q_\varepsilon(\tau), x)} - \mathbb{E}\left[\frac{\dot{G}_\theta(\theta, Q_\varepsilon(\tau), \mathbf{X}_0)}{\dot{G}_\varepsilon(\theta, Q_\varepsilon(\tau), \mathbf{X}_0)} \right] \right\}^T D(\theta, \varepsilon_i, \mathbf{X}_i). \quad (2.16) \end{aligned}$$

By Theorem 2, the proposed semiparametric CVaR estimator can achieve $\sqrt{n}$ parametric convergence rate, regardless of the dimensionality of the covariates $\mathbf{X}_i$. By contrast, for practical reason, the nonparametric quantile regression approach in (2.2) works only for $p = 1$ or 2 and has convergence rate $\sqrt{nb_n^p}$ for some non-trivial choice of bandwidth b_n.

Denote by $\tilde{Q}_\varepsilon(\tau)$ the sample quantile of the true innovations $\{\varepsilon_i\}$. By the well-known theory for sample quantiles, $\sqrt{n}[\tilde{Q}_\varepsilon(\tau) - Q_\varepsilon(\tau)] \Rightarrow N(0, \tau(1-\tau)/f_\varepsilon(Q_\varepsilon(\tau))^2)$. Note that $\mathbb{E}\{[(\tau - \mathbf{1}_{\varepsilon_i < Q_\varepsilon(\tau)})/f_\varepsilon(Q_\varepsilon(\tau))]^2\} = \tau(1-\tau)/f_\varepsilon(Q_\varepsilon(\tau))^2$. Thus, the first term of $W_i(\tau)$ in (2.16) reflects the variation of the sample quantile of the true innovations $\{\varepsilon_i\}$. On the other hand, the second term of $W_i(\tau)$ reflects the error due to the estimator $\hat{\theta}$. The first term is an intrinsic feature of sample quantiles, which never vanishes; the second term generally does not vanish but may vanish under some special settings. For example, if $G(\theta, \varepsilon_i, \mathbf{X}_i) \equiv G(\varepsilon_i, \mathbf{X}_i)$ is completely known (does not depend on any parameter), then $\dot{G}_\theta(\theta, \cdot, \cdot) = 0$ and consequently the second term vanishes. As a second example, if $G(\theta, \varepsilon_i, \mathbf{X}_i) \equiv G(\theta, \varepsilon_i)$ does not depend on $\mathbf{X}_i$, then the second term also vanishes. Intuitively, in the latter case, $\{Y_i\}$ are i.i.d. and therefore $Q(\tau|x)$ is simply the marginal quantile of $\{Y_i\}$.

2.1.3 Consistent estimate of the limiting variance

In the vast literature on VaR and CVaR estimation, the standard error calculation has been largely ignored. Chen and Tang (2005) studied this problem for unconditional VaR estimation; for nonparametric CVaR estimation in Cai and Wang (2008), they did not provide consistent estimate for the standard error, and any such attempt would involve nonparametric function estimation with properly chosen bandwidth. Here we consider consistent estimate of the limiting variance $\Gamma(\tau)$ in (2.15). We propose the following procedure:

(i) Using the estimated innovations $\{\hat{\varepsilon}_i\}$ in (2.10), we estimate the density $f_\varepsilon(z)$ of ε_i by

$$\hat{f}_\varepsilon(z) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{\hat{\varepsilon}_i - z}{h_n}\right), \tag{2.17}$$

where $h_n > 0$ is the bandwidth. For example, the rule-of-thumb bandwidth choice (Silverman, 1986) is $h_n = 0.9n^{-1/5}\min\{\mathrm{sd}(\hat{\varepsilon}_i), \mathrm{IQR}(\hat{\varepsilon}_i)/1.34\}$, where $\mathrm{sd}(\hat{\varepsilon}_i)$ and $\mathrm{IQR}(\hat{\varepsilon}_i)$ are the sample standard deviation and sample interquartile of $\{\hat{\varepsilon}_i\}$.

(ii) Plugging $\hat{\varepsilon}_i, \hat{f}_\varepsilon, \hat{\theta}, \hat{Q}_\varepsilon(\tau)$ [see (2.11)] into $W_i(\tau)$ in (2.16) to obtain the sample version

$$\begin{aligned}\widehat{W}_i(\tau) &= \frac{\tau - \mathbf{1}_{\hat{\varepsilon}_i < \hat{Q}_\varepsilon(\tau)}}{\hat{f}_\varepsilon(\hat{Q}_\varepsilon(\tau))} \\ &+ \left\{\frac{\dot{G}_\theta(\hat{\theta}, \hat{Q}_\varepsilon(\tau), x)}{\dot{G}_\varepsilon(\hat{\theta}, \hat{Q}_\varepsilon(\tau), x)} - \frac{1}{n}\sum_{i=1}^{n}\frac{\dot{G}_\theta(\hat{\theta}, \hat{Q}_\varepsilon(\tau), \mathbf{X}_i)}{\dot{G}_\varepsilon(\hat{\theta}, \hat{Q}_\varepsilon(\tau), \mathbf{X}_i)}\right\}^T D(\hat{\theta}, \hat{\varepsilon}_i, \mathbf{X}_i).\end{aligned}$$

(iii) Using the sample variance of $\widehat{W}_i(\tau)$ to estimate $\Gamma(\tau)$ by

$$\begin{aligned}\hat{\Gamma}(\tau) &= \dot{G}_\varepsilon(\hat{\theta}, \hat{Q}_\varepsilon(\tau), x)^2 \frac{1}{n-1}\sum_{i=1}^{n}\left[\widehat{W}_i(\tau) - \overline{W}(\tau)\right]^2, \text{where} \\ \overline{W}(\tau) &= \frac{1}{n}\sum_{i=1}^{n}\widehat{W}_i(\tau).\end{aligned} \tag{2.18}$$

Theorem 3. *Suppose that Assumptions 1–2 and Assumptions 3–6 (in Section 2.7) hold. In (2.17), assume that: (i) the kernel $K(\cdot)$ has bounded support and bounded derivative; and (ii) the bandwidth h_n satisfies $nh_n^4 \to \infty$. Recall $\Gamma(\tau)$ defined in (2.15). Then*

$$\hat{\Gamma}(\tau) \xrightarrow{p} \Gamma(\tau).$$

Consequently, from Theorem 2,

$$\frac{\sqrt{n}[\hat{Q}(\tau|x) - Q(\tau|x)]}{\sqrt{\hat{\Gamma}(\tau)}} \Rightarrow N(0,1). \tag{2.19}$$

By Theorem 3, an asymptotic $(1-\alpha)$ confidence interval for $Q(\tau|x)$ is

$$\hat{Q}(\tau|x) \pm q_{1-\alpha}\left(\frac{\hat{\Gamma}(\tau)}{n}\right)^{1/2}, \tag{2.20}$$

where $q_{1-\alpha}$ is the $(1-\alpha)$ quantile of $|N(0,1)|$. In Section 2.3 below we introduce an alternative bootstrap approach that can bypass the estimation of $\Gamma(\tau)$.

2.1.4 Bahadur representation in Assumption 2

From Theorem 2, the $\sqrt{n}$-consistency requires only $\hat{\theta} = \theta + O_p(n^{-1/2})$, but the CLT relies on the Bahadur representation (2.14) in Assumption 2. Also, the variance estimator $\hat{\Gamma}(\tau)$ in (2.18) relies on the Bahadur representation. Such Bahadur representation depends on the specific model structure and parameter estimation method. We briefly discuss this issue.

An important example of (2.3) is the nonlinear model with heteroscedastic errors:

$$Y_i = \mu(\theta, \mathbf{X}_i) + \sigma(\theta, \mathbf{X}_i)\varepsilon_i, \tag{2.21}$$

for i.i.d. noises $\{\varepsilon_i\}$ with $\mathbb{E}(\varepsilon_i) = 0$ and $\mathbb{E}(\varepsilon_i^2) = 1$ and parametric functions $\mu(\theta, \cdot)$ and $\sigma(\theta, \cdot) > 0$. Model (2.21) satisfies Assumption 1. Consider the pseudo-likelihood estimate:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} \left\{ \left[\frac{Y_i - \mu(\theta, \mathbf{X}_i)}{\sigma(\theta, \mathbf{X}_i)} \right]^2 + 2\log \sigma(\theta, \mathbf{X}_i) \right\}. \tag{2.22}$$

Theorem 2 in Zhao (2010) established the Bahadur representation (2.14) with

$$D(\theta, \varepsilon_i, \mathbf{X}_i) = \mathcal{I}(\theta)^{-1} \left[\frac{\varepsilon_i \dot{\mu}(\theta, \mathbf{X}_i)}{\sigma(\theta, \mathbf{X}_i)} + (\varepsilon_i^2 - 1) \frac{\dot{\sigma}(\theta, \mathbf{X}_i)}{\sigma(\theta, \mathbf{X}_i)} \right], \tag{2.23}$$

where $\dot{\mu}(\theta, \mathbf{X}_i)$ and $\dot{\sigma}(\theta, \mathbf{X}_i)$ are the partial derivatives with respect to θ, and

$$\mathcal{I}(\theta) = \mathbb{E}\Big[\frac{\dot{\mu}(\theta, \mathbf{X}_0)\dot{\mu}(\theta, \mathbf{X}_0)^T + 2\dot{\sigma}(\theta, \mathbf{X}_0)\dot{\sigma}(\theta, \mathbf{X}_0)^T}{\sigma^2(\theta, \mathbf{X}_0)} \Big]. \tag{2.24}$$

Other estimation methods lead to different Bahadur representations, depending on specific loss functions. For maximum likelihood estimation, the Bahadur representation depends on the score function. For quantile regression based estimator, consider the special case of (2.3) that $\sigma(\cdot,\cdot) \equiv 1$ and $Q_\varepsilon(0.5) = 0$ (ε_i has median zero), then the median quantile regression estimator is the minimizer of $\sum_{i=1}^n |Y_i - \mu(\theta, \mathbf{X}_i)|$, which satisfies (2.14) with

$$D(\theta, \varepsilon_i, \mathbf{X}_i) = \Big\{\mathbb{E}[\dot{\mu}(\theta, \mathbf{X}_0)\dot{\mu}(\theta, \mathbf{X}_0)^T]\Big\}^{-1} \frac{\dot{\mu}(\theta, \mathbf{X}_i)}{f_\varepsilon(0)}\Big(\frac{1}{2} - \mathbf{1}_{\varepsilon_i<0}\Big). \tag{2.25}$$

See Jurečková and Procházka (1994). Zhao and Xiao (2014) obtained a Bahadur representation for quantile regression estimator of the location-scale model $Y_i = \mathbf{X}_i\beta + (\mathbf{X}_i\gamma)\varepsilon_i$. He and Shao (1996) obtained Bahadur representations for general M-estimators.

We point out that it is up to the practitioner to determine the specific model and parameter estimation method, which are the starting point to carry out any subsequent CVaR estimation and inference. This parallels to what we usually do in time series forecasting based on some estimated model for the data. Our semiparametric CVaR estimator hinges on a preliminary $\sqrt{n}$-consistent estimator $\hat{\theta}$ of θ, and our asymptotic confidence interval relies on the Bahadur representation of $\hat{\theta} - \theta$. In Section 2.3 below, we propose a sieve bootstrap approach, which can bypass such Bahadur representation.

2.2 CVaR Process and CES

In financial risk management, the portfolio manager may be interested in different percentiles (e.g., the top $1, 5, 10, 25$-th percentiles) of the potential loss and draw some simultaneous inference. This type of information provides the basis for dynamically managing the portfolio to control the overall risk at different levels. This motivates us to study the CVaR process or equivalently the conditional quantile process $\{Q(\tau|x)\}_\tau$ on some quantile interval $\tau \in [\delta, 1-\delta]$ with some small $\delta > 0$. Theorem 4 establishes a functional CLT version of Theorem 2.

Theorem 4. *Consider $[\delta, 1-\delta]$ with any small $\delta > 0$. Suppose that Assumptions 1–2 and Assumptions 3–4 and 5^* (in Section 2.7) hold. Then the functional CLT*

holds

$$\left\{\sqrt{n}[\hat{Q}(\tau|x) - Q(\tau|x)]\right\}_{\tau\in[\delta,1-\delta]} \Rightarrow \{Z(\tau)\}_{\tau\in[\delta,1-\delta]}, \tag{2.26}$$

where $\{Z(\tau)\}_{\tau\in[\delta,1-\delta]}$ *is a centered Gaussian process with autocovariance given by [in the expression below,* $W_i(\tau)$ *is defined as in (2.16)]*

$$\begin{aligned}\Gamma(\tau,\tau'): &= \text{cov}\{Z(\tau), Z(\tau')\} \\ &= \dot{G}_\varepsilon(\theta, Q_\varepsilon(\tau), x)\dot{G}_\varepsilon(\theta, Q_\varepsilon(\tau'), x)\text{cov}\{W_1(\tau), W_1(\tau')\}. \end{aligned} \tag{2.27}$$

Remark 2. For the quantile interval $[\delta, 1-\delta]$ in Theorem 4, $\delta > 0$ is assumed to be a given small number to avoid the boundary issue. We conjecture that, using more sophisticated arguments, it may be possible to extend the interval to $(0, 1)$. The main technical issue is to establish the Bahadur representation (2.78) (see Section 2.7.3 in the proof section) uniformly for τ in some expanding interval $[\delta_n, 1-\delta_n]$ with $\delta_n \to 0$. For example, Portnoy and Koenker (1989) obtained such results for linear models. To avoid technical difficulties, we shall not pursue this direction. Also, in practice the most extreme VaR level we normally consider is $1-\tau = 1\%$, substantially below which any estimator may become unstable due to scarce observations in the extreme tail. ♢

As in Section 2.1.3, we can estimate the covariance function $\Gamma(\tau, \tau')$ in (2.27) by

$$\begin{aligned}\hat{\Gamma}(\tau,\tau') &= \dot{G}_\varepsilon(\hat{\theta}, \hat{Q}_\varepsilon(\tau), x)\dot{G}_\varepsilon(\hat{\theta}, \hat{Q}_\varepsilon(\tau'), x) \\ &\times \frac{1}{n-1}\sum_{i=1}^{n}\left[\widehat{W}_i(\tau) - \overline{W}(\tau)\right]\left[\widehat{W}_i(\tau') - \overline{W}(\tau')\right], \end{aligned} \tag{2.28}$$

where $\widehat{W}_i(\tau)$ and $\overline{W}(\tau)$ are defined in (2.18). Similar to Theorem 3, the uniform consistency of $\hat{\Gamma}(\tau, \tau')$ can be established along the similar line of argument. We omit the details.

The functional CLT in Theorem 4 provides a theoretical basis for simultaneous inference of CVaR at multiple levels. Here we consider an application to the CES. Recall that $\text{CVaR}(1-\tau|x)$ is the $(1-\tau)$ worst scenario portfolio loss, given the

covariates $\mathbf{X}_i = x$. At level γ, conditioning on $\mathbf{X}_i = x$, the CES is

$$\mathrm{CES}(\gamma|x) = \frac{1}{\gamma}\int_0^\gamma \mathrm{CVaR}(\tau|x)d\tau = \frac{1}{\gamma}\int_{1-\gamma}^1 \mathrm{CVaR}(1-\tau|x)d\tau. \tag{2.29}$$

$\mathrm{CES}(\gamma|x)$ can be interpreted as the average conditional loss given that $\mathbf{X}_i = x$ and that the loss is at or even more extreme than the 100γ-th percentile worst scenario. Some recent nonparametric approaches include Scaillet (2004) and Chen (2008) for nonparametric ES estimation and Scaillet (2005) and Cai and Wang (2008) for nonparametric CES estimation.

In (2.29), the CVaR at all confidence levels are equally weighted, however practitioners may favor some confidence levels more than other levels. Let $\mathcal{T}$ be an interval of confidence level and $w(\cdot)$ a square-integrable weight function (depending on the practitioner's preference). We generalize (2.29) to the the weighted version on $\mathcal{T}$:

$$\mathrm{CES}(\mathcal{T}|x) = \int_{\mathcal{T}} w(1-\tau)\mathrm{CVaR}(1-\tau|x)d\tau, \quad \text{where} \quad \int_{\mathcal{T}} w(1-\tau)d\tau = 1. \tag{2.30}$$

Clearly, (2.29) is a special case of (2.30) with $w(\cdot) \equiv 1/\gamma$ and $\mathcal{T} = [1-\gamma, 1]$. Since $\mathrm{CVaR}(1-\tau|x) = Q(\tau|x)$, plugging in the estimator $\hat{Q}(\tau|x)$ in (2.12), we estimate $\mathrm{CES}(\mathcal{T}|x)$ by

$$\widehat{\mathrm{CES}}(\mathcal{T}|x) = \int_{\mathcal{T}} w(1-\tau)\hat{Q}(\tau|x)d\tau. \tag{2.31}$$

Remark 3. All results stated here also hold when $\mathcal{T}$ is a discrete set of confidence levels. In this case, we simply replace the integrals in (2.30)–(2.31) by summation over $\mathcal{T}$.

By Theorem 4 and the continuous mapping theorem, we can immediately obtain

Theorem 5. *Assume the same conditions in Theorem 4. For any interval* $\mathcal{T} \subset [\delta, 1-\delta]$,

$$\sqrt{n}[\widehat{\mathrm{CES}}(\mathcal{T}|x) - \mathrm{CES}(\mathcal{T}|x)] \;\Rightarrow\; \int_{\mathcal{T}} w(1-\tau)Z(\tau)d\tau, \tag{2.32}$$

where $\{Z(\tau)\}$ *is the Gaussian process in Theorem 4.*

It is easy to see that the limiting distribution in (2.32) is a centered normal distribution with variance

$$\int\int_{\mathcal{T}\times\mathcal{T}} w(1-\tau)w(1-\tau')\Gamma(\tau,\tau')d\tau d\tau'. \tag{2.33}$$

Here $\Gamma(\tau,\tau')$ is the covariance function defined in (2.27). This variance can be estimated by plugging the estimator $\hat{\Gamma}(\tau,\tau')$ in (2.28).

2.3 User-friendly Sieve Bootstrap Inference

As discussed in Section 2.1.4, in order to implement the variance estimator $\hat{\Gamma}(\tau)$ in (2.18) for the confidence interval (2.20), we need to know the Bahadur representation (2.14); the same requirement is also needed for the covariance estimator $\hat{\Gamma}(\tau,\tau')$ in (2.28) [however, the CVaR estimator $\hat{Q}(\tau|x)$ in (2.12) does not require this]. It may be non-trivial for a non-expert practitioner to derive a Bahadur representation for their specific model and parameter estimation. In this section we provide a user-friendly bootstrap approach.

For time series, two popular bootstrap methods are the block bootstrap (Lahiri, 2003) and the sieve bootstrap (Bühlmann, 1997). The block bootstrap requires the challenging issue of block length selection. For a given time series model subject to unknown parameters, the sieve bootstrap creates bootstrap samples by recursively using the model with estimated parameters and resampled residuals, and thus the bootstrap data can preserve the dependence structure of the original data. Here we adopt the sieve bootstrap.

Assume that the covariates $\mathbf{X}_i = (Y_{i-1},\ldots,Y_{i-q},\mathbf{U}_i)$ consist of both lagged Y's and some other covariates $\mathbf{U}_i$. We propose the following sieve bootstrap procedure:

(i) Use some parameter estimation method to obtain the estimate $\hat{\theta}$ and then compute $\{\hat{\varepsilon}_i\}$ [see (2.10)] and $\hat{Q}(\tau|x)$ [see (2.12)] based on the original data.

(ii) Obtain the bootstrap samples $\{(\mathbf{X}_i^*, Y_i^*)\}$ recursively (with same initial values as $\mathbf{X}_i$)

$$Y_i^* = G(\hat{\theta}, \varepsilon_i^*, \mathbf{X}_i^*) \quad \text{with} \quad \mathbf{X}_i^* = (Y_{i-1}^*,\ldots,Y_{i-q}^*,\mathbf{U}_i), \tag{2.34}$$

where $\{\varepsilon_i^*\}$ are i.i.d. random samples (with replacement) from $\{\hat{\varepsilon}_i\}$. Use the bootstrap data $\{(\mathbf{X}_i^*, Y_i^*)\}$ and the same parameter estimation method in step (i) to obtain new parameter estimate $\hat{\theta}^*$ and new conditional quantile estimate $\hat{Q}^*(\tau|x)$.

(iii) Repeat (ii) to obtain a large number (M, say) of realizations of $\hat{Q}^*(\tau|x)$, denoted by $\hat{Q}^{*(1)}(\tau|x), \ldots, \hat{Q}^{*(M)}(\tau|x)$.

We make one important comment about the bootstrap procedure in prediction setting. Suppose we wish to construct bootstrap interval for the predictive quantile $Q(\tau|\{Y_i\}_{i\le n})$ for Y_{n+1} based on data $\{Y_i\}_{i=1}^n$ from the GARCH model $Y_i = \sigma_i\varepsilon_i, \sigma_i^2 = \omega + \alpha Y_{i-1}^2 + \beta\sigma_{i-1}^2$. In step (i) above, we fit GARCH model to obtain estimates $(\hat{\omega}, \hat{\alpha}, \hat{\beta})$, $\hat{\sigma}_i$, and $\hat{\varepsilon}_i = Y_i/\hat{\sigma}_i$. In step (ii), first we generate bootstrap samples $\{Y_i^*\}_{i=1}^n$ from $Y_i^* = \sigma_i^*\varepsilon_i^*, \sigma_i^{*2} = \hat{\omega} + \hat{\alpha}Y_{i-1}^{*2} + \hat{\beta}\sigma_{i-1}^{*2}$, then fit GARCH model to $\{Y_i^*\}_{i=1}^n$ to obtain estimates $\hat{\sigma}_i^*$ and $\hat{\varepsilon}_i^* = Y_i^*/\hat{\sigma}_i^*$, and finally compute $\hat{Q}^*(\tau|\{Y_i\}_{i\le n})$ as $\hat{\sigma}_{n+1}$ multiplied by the sample τ-th quantile of $\{\hat{\varepsilon}_i^*\}_{i=1}^n$. It is important to use $\hat{\sigma}_{n+1}$ instead of $\hat{\sigma}_{n+1}^*$. This is because $Q(\tau|\{Y_i\}_{i\le n})$ is the τ-th quantile of Y_{n+1} given fixed covariates $\{Y_i\}_{i\le n}$ and $\hat{\sigma}_{n+1}$ reflects such fixed covariates. By contrast, using $\hat{\sigma}_{n+1}^*$ would mean that we are estimating $Q(\tau|\{Y_i^*\}_{i\le n})$ instead of $Q(\tau|\{Y_i\}_{i\le n})$.

We discuss some bootstrap inference below.

Bootstrap confidence interval for CVaR

For the realizations $\hat{Q}^{*(1)}(\tau|x), \ldots, \hat{Q}^{*(M)}(\tau|x)$ in step (iii) above, denote by $q_{1-\alpha}^*$ the $(1-\alpha)$ sample quantile of $\sqrt{n}|\hat{Q}^{*(1)}(\tau|x) - \hat{Q}(\tau|x)|, \ldots, \sqrt{n}|\hat{Q}^{*(M)}(\tau|x) - \hat{Q}(\tau|x)|$. Then the $(1-\alpha)$ bootstrap confidence interval for $Q(\tau|x)$ is

$$\hat{Q}(\tau|x) \pm q_{1-\alpha}^*/\sqrt{n}. \tag{2.35}$$

Bootstrap confidence interval for CES

Plugging the realizations $\hat{Q}^{*(1)}(\tau|x), \ldots, \hat{Q}^{*(M)}(\tau|x)$ into (2.31) to obtain the bootstrap-data-based CES estimates $\widehat{\mathrm{CES}}^{*(1)}(\mathcal{T}|x), \ldots, \widehat{\mathrm{CES}}^{*(M)}(\mathcal{T}|x)$. Denote the $(1-\alpha)$ sample quantile of $\sqrt{n}|\widehat{\mathrm{CES}}^{*(1)}(\mathcal{T}|x) - \widehat{\mathrm{CES}}(\mathcal{T}|x)|, \ldots, \sqrt{n}|\widehat{\mathrm{CES}}^{*(M)}(\mathcal{T}|x) - \widehat{\mathrm{CES}}(\mathcal{T}|x)|$ by $r_{1-\alpha}$. Then the $(1-\alpha)$ bootstrap confidence interval for $\mathrm{CES}(\mathcal{T}|x)$ is

$$\widehat{\mathrm{CES}}(\mathcal{T}|x) \pm r_{1-\alpha}/\sqrt{n}. \tag{2.36}$$

The easy-to-implement bootstrap confidence intervals (2.35)–(3.19) only require some parameter estimation method for $\hat{\theta}$, and thus practitioners can construct CVaR and CES confidence intervals using our proposed semiparametric CVaR and CES estimator along with their favorite parameter estimation methods. Since the bootstrap model (2.34) inherits the same structure of the original model (2.3), the bootstrap data can closely mimic the dependence structure of the original data. Our simulation study in Section 2.5 suggests that the bootstrap confidence intervals have better finite sample performance than the asymptotic confidence intervals based on estimated limiting variances.

2.4 Semiparametric conditional distribution estimation

In this section we adopt the semiparametric approach in Section 2.1 to study conditional distribution estimation. Denote by $F(y|x) = \mathbb{P}\{Y_i \leq y|\mathbf{X}_i = x\}$ the conditional distribution function of Y_i given $\mathbf{X}_i = x$. The conditional distribution can fully characterize the distributional dependence of the loss Y_i on the covariates $\mathbf{X}_i$.

To estimate $F(y|x)$, the usual nonparametric kernel regression approach is

$$\tilde{F}(y|x) = \frac{\sum_{i=1}^{n} \mathbf{1}_{Y_i \leq y} K\{(\mathbf{X}_i - x)/b_n\}}{\sum_{i=1}^{n} K\{(\mathbf{X}_i - x)/b_n\}}, \tag{2.37}$$

where $K(\cdot)$ and b_n are the kernel and bandwidth as in Section 2.1.1; see Chapter 6 in Li and Racine (2007). This nonparametric conditional distribution estimate has the same drawback as the nonparametric conditional quantile estimation in Section 2.1.1.

We can easily adapt our method in Section 2.1.2 to construct a $\sqrt{n}$-consistent estimate of $F(y|x)$. Under Assumption 1 and by the independence between ε_i and $\mathbf{X}_i$,

$$\begin{aligned} \mathbb{P}\{Y_i \leq y|\mathbf{X}_i = x\} &= \mathbb{P}\{G(\theta, \varepsilon_i, x) \leq y\} \\ &= \mathbb{P}\{\varepsilon_i \leq H(\theta, y, x)\} \\ &= F_\varepsilon\{H(\theta, y, x)\}. \end{aligned} \tag{2.38}$$

Therefore we propose the following semiparametric estimate of $F(y|x)$:

$$\hat{F}(y|x) = \hat{F}_\varepsilon\{H(\hat{\theta}, y, x)\}, \tag{2.39}$$

where $\hat{\theta}$ is a consistent estimate of θ, and $\hat{F}_\varepsilon(z)$ [defined in (2.11)] is the sample empirical distribution of the generalized residuals $\{\hat{\varepsilon}_i\}$. Theorem 6 below presents a functional CLT, which implies the pointwise CLT.

Theorem 6. *Suppose that the same conditions in Theorem 4 hold. Let $\mathcal{Y} = [\mathcal{Y}_1, \mathcal{Y}_2]$ be any bounded interval. Then the functional CLT holds*

$$\left\{\sqrt{n}[\hat{F}(y|x) - F(y|x)]\right\}_{y\in\mathcal{Y}} \Rightarrow \{S(y)\}_{y\in\mathcal{Y}}, \tag{2.40}$$

where $\{S(y)\}_{y\in\mathcal{Y}}$ is a centered Gaussian process with autocovariance

$$\Sigma(y, y') := \mathrm{cov}\{S(y), S(y')\} = \mathrm{cov}\{V_1(y), V_1(y')\}, \tag{2.41}$$

and

$$\begin{aligned} V_i(y) &= [\mathbf{1}_{\varepsilon_i \le H(\theta,y,x)} - \mathbb{E}(\mathbf{1}_{\varepsilon_i \le H(\theta,y,x)})] \\ &+ f_\varepsilon(H(\theta, y, x))\left\{\mathbb{E}\left[\frac{\dot{G}_\theta(\theta, H(\theta, y, x), \mathbf{X}_0)}{\dot{G}_\varepsilon(\theta, H(\theta, y, x), \mathbf{X}_0)}\right] + \dot{H}(\theta, y, x)\right\}^T D(\theta, \varepsilon_i, \mathbf{X}_i). \end{aligned} \tag{2.42}$$

In (2.42), the first component is from the empirical distribution of the true innovations $\{\varepsilon_i\}$, and the second component is due to the estimation error of $\hat{\theta}$. Compared to the nonparametric kernel smoothing estimator in (2.37), the proposed semiparametric estimator in (2.39) is easy to implement and attains $\sqrt{n}$ parametric convergence rate.

Similar to the estimation of $W_i(\tau)$ in Section 2.1.2, we can estimate $V_i(y)$ by

$$\begin{aligned} \widehat{V}_i(y) &= [\mathbf{1}_{\varepsilon_i \le H(\hat{\theta},y,x)} - \hat{F}_\varepsilon(H(\hat{\theta}, y, x))] \\ &+ \hat{f}_\varepsilon(H(\hat{\theta}, y, x))\left\{\frac{1}{n}\sum_{i=1}^n \frac{\dot{G}_\theta(\hat{\theta}, H(\hat{\theta}, y, x), \mathbf{X}_i)}{\dot{G}_\varepsilon(\hat{\theta}, H(\hat{\theta}, y, x), \mathbf{X}_i)} + \dot{H}(\hat{\theta}, y, x)\right\}^T D(\hat{\theta}, \hat{\varepsilon}_i, \mathbf{X}_i). \end{aligned}$$

Then the covariance in (2.41) can be estimated by the sample covariance

$$\hat{\Sigma}(y,y') = \frac{1}{n-1}\sum_{i=1}^{n}\Big[\widehat{V}_i(y) - \overline{V}(y)\Big]\Big[\widehat{V}_i(y') - \overline{V}(y')\Big], \quad \overline{V}(y) = \frac{1}{n}\sum_{i=1}^{n}\widehat{V}_i(y). \tag{2.43}$$

Similar to Theorem 3, we can establish the consistency of $\hat{\Sigma}(y,y')$. Also, the sieve bootstrap procedure in Section 2.3 can be applied here to avoid the issue of Bahadur representation.

2.5 Monte Carlo Studies

2.5.1 MISE comparison with nonparametric method

For the nonparametric estimator $\tilde{Q}(\tau|x)$ in (2.2), let $K(\cdot)$ be the p-variate standard normal density. Using 300 realizations $\tilde{Q}^{(1)}(\tau|x), \ldots, \tilde{Q}^{(300)}(\tau|x)$ of $\tilde{Q}(\tau|x)$, we measure the performance of $\tilde{Q}(\tau|x)$ by the empirical mean integrated squared error (MISE) on a set $\mathcal{X}$:

$$\text{MISE}\{\tilde{Q}(\tau|\cdot); b_n\} = \frac{1}{300}\sum_{\ell=1}^{300}\int_{\mathcal{X}}[\tilde{Q}^{(\ell)}(\tau|x) - Q(\tau|x)]^2 dx. \tag{2.44}$$

Since $\text{MISE}\{\tilde{Q}(\tau|\cdot); b_n\}$ depends on bandwidth b_n, we consider its best-case scenario:

$$\text{MISE}\{\tilde{Q}(\tau|\cdot)\} = \min_{b_n}\text{MISE}\{\tilde{Q}(\tau|\cdot); b_n\}, \tag{2.45}$$

which is the theoretical minimum MISE of the nonparametric method. For the proposed semiparametric estimator $\hat{Q}(\tau|x)$ in (2.12), its MISE is defined as in (2.44), and we further define its relative MISE (RMISE), relative to the nonparametric estimator $\tilde{Q}(\tau|x)$ under the best-case scenario, as

$$\text{RMISE} = \frac{\text{MISE}\{\tilde{Q}(\tau|\cdot)\}}{\text{MISE}\{\hat{Q}(\tau|\cdot)\}} = \frac{\min_{b_n}\text{MISE}\{\tilde{Q}(\tau|\cdot); b_n\}}{\text{MISE}\{\hat{Q}(\tau|\cdot)\}}. \tag{2.46}$$

A value of RMISE ≥ 1 indicates better MISE performance of the proposed method. The comparison will largely favor the nonparametric method as the choice of bandwidth for $\tilde{Q}(\tau|x)$ is done under the best-case scenario, which is generally

unavailable in practice.

We consider the following four increasingly more complicated models:

$$\begin{aligned}
\text{Model 1:}\quad Y_i &= \theta_0 + \theta_1 Y_{i-1} + \sigma\varepsilon_i, \quad (\theta_0, \theta_1, \sigma) = (0.3, 0.4, 0.5);\\
\text{Model 2:}\quad Y_i &= \theta_0 + \theta_1 Y_{i-1} + \theta_2 Y_{i-2} + \sigma_i\varepsilon_i, \quad (\theta_0, \theta_1, \theta_2, \sigma) = (0.3, 0.4, -0.5, 0.5);\\
\text{Model 3:}\quad Y_i &= \theta_0 + \theta_1 Y_{i-1} + \varepsilon_i\sqrt{\theta_2^2 + \theta_3^2 Y_{i-1}^2}, \quad (\theta_0, \theta_1, \theta_2, \theta_3) = (0.3, 0.4, 0.3, 0.5);\\
\text{Model 4:}\quad Y_i &= \theta_0 + \theta_1 Y_{i-1} + \gamma_1 U_i + \varepsilon_i\sqrt{\theta_2^2 + \theta_3^2 Y_{i-1}^2 + \gamma_2^2 U_i^2}, \quad U_i : \text{uniform } [0,1],\\
&\qquad\qquad \text{with} \quad (\theta_0, \theta_1, \theta_2, \theta_3, \gamma_1, \gamma_2) = (0.3, 0.4, 0.3, 0.5, -0.4, 0.3).
\end{aligned}$$

Model 1 and 2 are simple AR(1) and AR(2) models, Model 3 is an AR(1)-ARCH(1) model, and Model 4 is an AR(1)-ARCH(1) with exogenous variable U_i. The AR-ARCH model with exogenous input allows us to model how stock returns depend on the past returns as well as other external variables. The noise ε_i is from two distributions: (i) standard normal $N(0,1)$, and (ii) $t_3/\sqrt{3}$ (Student-t distribution with 3 degrees of freedom with the normalizer $\sqrt{3}$ making the variance one). In all settings we use sample size $n = 200$.

For Model 1 and 3, we estimate the conditional τ-th quantile of Y_i given $Y_{i-1} = x$, and we take $\mathcal{X}$ in (2.44) to be the range of 2.5-th and 97.5-th percentiles of $\{Y_{i-1}\}$; for Model 2 (resp. Model 4), we estimate the conditional τ-th quantile of Y_i given the bivariate $\mathbf{X}_i := (Y_{i-1}, Y_{i-2}) = (x_1, x_2)$ (resp. $\mathbf{X}_i = (Y_{i-1}, U_i)$ for Model 4), and we take $\mathcal{X}$ in (2.44) to be $\mathcal{X}_1 \times \mathcal{X}_2$, where $\mathcal{X}_1$ and $\mathcal{X}_2$ are, respectively, the range of 2.5-th and 97.5-th percentiles of each of the two coordinates of $\mathbf{X}_i$. The integral in (2.44) is approximated by 20 evenly spaced grid points in the univariate case (Model 1 and 3) or 10×10 evenly spaced grid points in the bivariate case (Model 2 and 4). To implement the semiparametric method, we use (2.22) to estimate the unknown parameters. The procedure is repeated for 13 different quantiles $\tau = 1\%, 5\%, 10\%, \ldots, 90\%, 95\%, 99\%$.

Table 1: RMISE [see (2.46)] of the proposed semiparametric estimate of $\text{CVaR}(1-\tau|x) = Q(\tau|x)$ relative to the nonparametric method in (2.2) with theoretical optimal bandwidth, at different quantiles τ. Numbers ≥ 1 indicate better performance of the proposed method.

		Quantile τ in $\text{CVaR}(1-\tau\|x) = Q(\tau\|x)$												
	noise	1%	5%	10%	20%	30%	40%	50%	60%	70%	80%	90%	95%	99%
Model 1	$N(0,1)$	2.45	2.09	1.62	1.54	1.47	1.42	1.35	1.32	1.34	1.38	1.72	2.03	2.81
	$t_3/\sqrt{3}$	3.01	2.39	2.11	1.50	1.14	0.96	0.94	0.98	1.05	1.34	1.80	2.18	3.38
Model 2	$N(0,1)$	3.54	2.92	2.14	1.67	1.50	1.43	1.39	1.38	1.46	1.69	2.19	2.61	3.61
	$t_3/\sqrt{3}$	5.16	3.22	2.25	1.40	1.09	1.01	0.97	0.99	1.15	1.45	2.67	3.67	4.03
Model 3	$N(0,1)$	2.58	2.01	1.80	1.48	1.27	1.30	1.37	1.54	1.86	2.11	2.15	2.32	2.93
	$t_3/\sqrt{3}$	1.85	1.01	0.84	0.80	0.86	0.87	0.88	0.91	0.90	1.00	1.17	1.36	2.24
Model 4	$N(0,1)$	3.65	2.61	2.28	2.15	1.79	1.45	1.35	1.46	1.71	1.97	2.19	2.60	3.61
	$t_3/\sqrt{3}$	2.71	1.13	1.01	0.95	0.87	0.81	0.78	0.77	0.86	1.00	1.02	1.17	3.09

Table 1 summarizes the RMISE [see (2.46)]. The results show that, for almost all cases considered, a substantial MISE improvement can be achieved by using the semiparametric CVaR estimator. The MISE improvement is more significant for the extreme quantiles $\tau = 90\%, 95\%, 99\%$, which correspond to the most widely used confidence levels $1-\tau = 10\%, 5\%, 1\%$ in the VaR literature. For the practically less interesting middle-range quantiles $\tau = 20\%, \ldots, 80\%$, the semiparametric estimator significantly outperforms the nonparametric estimator for $N(0,1)$ noise, whereas the two methods have comparable performance for Student-t distributed noise. However, we emphasize that the MISE comparison is done between the proposed method and the nonparametric method under the best-case scenario. In practice, the optimal bandwidth is generally unknown, therefore the proposed estimator can deliver relatively even more remarkable performance.

2.5.2 Coverage rate evaluation

To evaluate VaR estimator, another criterion is the empirical coverage rate, i.e., the empirical proportion, denoted by $\hat{\tau}$, of realizations such that $Y_i \leq \hat{Q}(\tau|\mathbf{X}_i)$. Specifically, our empirical coverage rate is calculated as follows:

(i) For each realization $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$, we use observations $\{(\mathbf{X}_i, Y_i)\}_{i=1}^{n-1}$ up to time $n-1$ to fit the model and estimate the predicted τ-th quantile $\hat{Q}(\tau|\mathbf{X}_n)$ for Y_n. We then check whether $Y_n \leq \hat{Q}(\tau|\mathbf{X}_n)$.

(ii) Repeat (i) 1000 times and compute the empirical coverage rate $\hat{\tau}$ as the proportion of realizations such that $Y_n \leq \hat{Q}(\tau|\mathbf{X}_n)$.

By definition, the closer $\hat{\tau}$ to τ, the better performance of $\hat{Q}(\tau|\mathbf{X}_i)$.

The four models in Section 2.5.1 are ARCH-type models, and in this section we examine some GARCH-type models of the form (2.7). We consider three GARCH models:

$$\begin{aligned}
&\text{Model 5:} \quad Y_i = \sigma_i \varepsilon_i, \quad \sigma_i^2 = \omega + \alpha Y_{i-1}^2 + \beta \sigma_{i-1}^2; \\
&\text{Model 6:} \quad Y_i = \sigma_i \varepsilon_i, \quad \log(\sigma_i^2) = \omega + \alpha \frac{Y_{i-1}}{\sigma_{i-1}} + \beta \log(\sigma_{i-1}^2) + \gamma \left[\frac{|Y_{i-1}|}{\sigma_{i-1}} - \mathbb{E}\Big(\frac{|Y_{i-1}|}{\sigma_{i-1}} \Big) \right]; \\
&\text{Model 7:} \quad Y_i = \sigma_i \varepsilon_i, \quad \sigma_i^2 = \omega + \alpha Y_{i-1}^2 + \beta \sigma_{i-1}^2 + \gamma Y_{i-1}^2 \mathbf{1}_{Y_{i-1}<0}.
\end{aligned}$$

Model 5 is the standard GARCH model (Bollerslev, 1986), Model 6 is the EGARCH model (Nelson, 1991), and Model 7 is the GJR-GARCH model (Glosten, Jagannathan and Runkle, 1993). In Model 5, $(\omega, \alpha, \beta) = (0.1, 0.3, 0.5)$; In Model 6, $(\omega, \alpha, \beta, \gamma) = (-3, -0.4, 0.5, 0.3)$; in Model 7, $(\omega, \alpha, \beta, \gamma) = (0.1, 0.3, 0.5, 0.2)$. As in Models 1–4, we consider two distributions, $N(0,1)$ and $t_3/\sqrt{3}$, for the noise ε_i.

As discussed in Example 3, GARCH models are non-Markovian and it is infeasible to use the nonparametric approach. Thus, we only evaluate the coverage rate for the proposed semiparametric method. When using the `R` package `rugarch` (Ghalanos, 2014) to do parameter estimation, we always specify the noise distribution as normal, regardless of the actual noise distribution [$N(0,1)$ or $t_3/\sqrt{3}$]. That is, for $t_3/\sqrt{3}$-distributed noises, the parameter estimation is done under mis-specification of the noise distribution.

Based on 1000 realization for each setting, Table 2 summarizes the empirical coverage rate at two sample sizes $n = 200$ and $n = 500$. Overall, the empirical coverage rate is close to the nominal level, and the larger sample size $n = 500$ leads to better performance.

Table 2: Empirical coverage rate for GARCH models 5–7 of the proposed semiparametric estimate of $\text{CVaR}(1-\tau|x) = Q(\tau|x)$ at different quantiles τ.

			Quantile τ in $\text{CVaR}(1-\tau\|x) = Q(\tau\|x)$												
	noise	n	1%	5%	10%	20%	30%	40%	50%	60%	70%	80%	90%	95%	99%
Model 5	$N(0,1)$	200	0.014	0.046	0.094	0.195	0.293	0.388	0.498	0.610	0.713	0.800	0.900	0.950	0.983
		500	0.012	0.049	0.098	0.198	0.303	0.399	0.497	0.597	0.692	0.777	0.884	0.938	0.983
	$t_3/\sqrt{3}$	200	0.017	0.045	0.087	0.171	0.273	0.386	0.489	0.590	0.685	0.804	0.904	0.946	0.984
		500	0.012	0.056	0.111	0.219	0.314	0.405	0.503	0.604	0.698	0.811	0.896	0.942	0.992
Model 6	$N(0,1)$	200	0.019	0.055	0.103	0.202	0.301	0.399	0.502	0.605	0.703	0.801	0.898	0.948	0.982
		500	0.012	0.051	0.101	0.196	0.299	0.404	0.506	0.604	0.705	0.805	0.904	0.954	0.990
	$t_3/\sqrt{3}$	200	0.023	0.063	0.110	0.208	0.302	0.408	0.503	0.601	0.695	0.794	0.893	0.938	0.978
		500	0.012	0.048	0.096	0.200	0.302	0.402	0.504	0.600	0.711	0.808	0.904	0.955	0.988
Model 7	$N(0,1)$	200	0.019	0.061	0.111	0.207	0.304	0.409	0.505	0.601	0.698	0.801	0.902	0.949	0.982
		500	0.011	0.056	0.106	0.207	0.303	0.402	0.497	0.597	0.698	0.797	0.894	0.945	0.985
	$t_3/\sqrt{3}$	200	0.019	0.062	0.112	0.213	0.309	0.414	0.517	0.609	0.707	0.802	0.899	0.953	0.986
		500	0.015	0.053	0.094	0.189	0.283	0.385	0.481	0.581	0.683	0.790	0.885	0.939	0.986

2.5.3 Comparison with parametric distribution based competitors

Many existing VaR methods use parametric distribution for the noise ε_i in the model. For example, the EWMA in RiskMetrics of J.P. Morgan (1994) uses standard Normal distribution. Different approaches have been proposed to model the observed heavy-tail and asymmetric returns, including the Student-t distribution, Laplace-distribution based robust-EWMA in Guermat and Harris (2001), and the asymmetric-Laplace distribution based skewed-EWMA in Gerlach, Lu and Huang (2013). Both the proposed semiparametric approach and the aforementioned existing methods require a parametric specification on the model structure, but our semiparametric approach does not impose any parametric-distribution on the noises.

To appreciate the advantage of the nonparametric distribution approach (i.e., using the sample quantile in (2.11)), we compare it with three parametric distribution based methods:

(i) Fit Normal distribution using maximum likelihood method.

(ii) Fit Student-t distribution using maximum likelihood method.

(iii) Fit asymmetric Laplace distribution using maximum likelihood method (Gerlach, Lu and Huang, 2013). This includes the Laplace-distribution (Guermat and Harris, 2001) as a special case.

We compare the performance of these methods in estimating the τ-th quantile, denoted by $Q_\varepsilon(\tau)$, of the noise distribution when the noise ε_i comes from five distributions: $N(0,1)$, $t_3/\sqrt{3}$, Laplace/$\sqrt{2}$ (the scaled Laplace with the factor $1/\sqrt{2}$ making the variance one), Normal mixture $0.5N(0,0.5)+0.5N(0,1.5)$, and standard exponential minus 1. All five distributions have variance one. In all settings we use sample size $n = 1000$.

For the nonparametric sample quantile $\hat{Q}_\varepsilon(\tau)$ in (2.11), we define its empirical mean squared error (MSE) as

$$\text{MSE}\{\hat{Q}_\varepsilon(\tau)\} = \frac{1}{1000}\sum_{\ell=1}^{1000}[\hat{Q}_\varepsilon^{(\ell)}(\tau) - Q_\varepsilon(\tau)]^2,$$

where $\hat{Q}_\varepsilon^{(\ell)}(\tau)$ is the estimate based on the ℓ-th realization. The MSE for the three parametric-distribution methods is calculated similarly. As in (2.46), we define the relative MSE (RMSE) of $\hat{Q}_\varepsilon(\tau)$ relative to a parametric method as

$$\text{RMSE} = \frac{\text{MSE}\{\text{parametric method}\}}{\text{MSE}\{\hat{Q}_\varepsilon(\tau)\}}. \tag{2.47}$$

A value of RMSE ≥ 1 indicates better MSE performance of the nonparametric distribution method.

Table 3 summarizes the RMSE results for different τ. When the noise distribution is correctly specified, the parametric-distribution method works well; however, for mis-specified noise distribution, the parametric-distribution methods may suffer from seriously poor performance. By contrast, the nonparametric distribution method yields reasonable performance in all cases.

Table 3: RMSE [see (2.47)] of the proposed nonparametric distribution method relative to three parametric-distribution (Normal, Student-t, and asymmetric-Laplace (ALD)) based competitors in the presence of different noise distributions: $N(0,1)$, $t_3/\sqrt{3}$, standard Laplace/$\sqrt{2}$ with variance one, Normal mixture $0.5N(0,0.5)+0.5N(0,1.5)$, and standard exponential minus 1. Numbers ≥ 1 indicate better performance of the proposed nonparametric distribution method. For convenience, numbers ≥ 100 are marked as ∞.

		Quantile τ in CVaR$(1-\tau\|x) = Q(\tau\|x)$												
noise	Method	1%	5%	10%	20%	30%	40%	50%	60%	70%	80%	90%	95%	99%
$N(0,1)$	Normal	0.27	0.55	0.62	0.66	0.68	0.66	0.66	0.66	0.68	0.65	0.63	0.53	0.27
	Student-t	0.36	0.54	0.68	0.75	0.72	0.67	0.63	0.66	0.69	0.73	0.67	0.53	0.36
	ALD	45.66	9.15	0.90	6.68	8.35	3.84	0.16	3.80	8.21	6.58	0.91	9.24	46.80
$t_3/\sqrt{3}$	Normal	2.50	16.72	45.77	70.69	47.36	14.17	1.56	14.18	46.97	70.20	47.17	16.28	2.62
	Student-t	0.34	0.36	0.48	0.71	0.80	0.81	0.81	0.82	0.81	0.72	0.50	0.37	0.34
	ALD	0.35	2.21	3.34	1.50	1.10	1.08	0.42	1.07	1.07	1.45	3.40	2.24	0.35
Laplace/$\sqrt{2}$	Normal	4.19	0.49	4.94	19.24	23.36	12.29	1.83	12.19	22.71	18.35	4.96	0.52	4.22
	Student-t	0.81	1.05	1.01	0.67	1.68	2.05	1.21	2.08	1.67	0.66	1.02	1.09	0.82
	ALD	0.28	0.55	0.66	0.67	0.61	0.54	0.49	0.54	0.61	0.66	0.68	0.59	0.29
Normal mixture	Normal	1.77	0.44	1.24	2.95	2.31	1.22	0.79	1.31	2.56	3.21	1.51	0.43	1.74
	Student-t	0.40	0.49	0.59	0.82	0.83	0.77	0.75	0.76	0.83	0.82	0.59	0.49	0.40
	ALD	10.93	3.40	0.86	3.11	4.89	2.67	0.20	2.62	4.89	3.12	0.86	3.44	10.77
Exponential -1	Normal	∞	∞	∞	19.35	34.41	84.45	95.26	76.85	46.08	15.23	0.81	7.17	16.84
	Student-t	∞	∞	∞	36.11	2.11	4.53	2.27	2.09	9.13	24.43	38.29	35.09	7.82
	ALD	∞	∞	∞	43.73	∞	∞	∞	41.36	13.43	1.67	2.09	5.36	5.59

2.5.4 Performance under model mis-specification

The proposed semiparametric approach requires a parametrization on the model structure, and below we examine its performance under model mis-specification.

We consider the following model mis-specification:

true model: GJR-GARCH Model 7 with $(\omega, \alpha, \beta) = (0.1, 0.3, 0.5)$ and different γ;
mis-specified model: standard GARCH Model 5.

If $\gamma = 0$, then GJR-GARCH reduces to the standard GARCH model; in general, the parameter γ measures the deviation between the true model and the mis-specified model. As in Section 2.5.2, we examine the empirical coverage rate under the mis-specified model for different γ, noise distribution (normal, $t_3/\sqrt{3}$), and sample size ($n = 200, 500$). Since the results for other settings were similar, Table 4 presents the results for $\gamma = 0.0, 0.2, 0.4, 0.6, 0.8$, normal noise, and sample size $n = 200$. We see that the method performs reasonably well under model mis-specification.

Table 4: Empirical coverage rate of the proposed semiparametric estimate of $\text{CVaR}(1-\tau|x) = Q(\tau|x)$ under model mis-specification. True data-generating model is the GJR-GARCH Model 7 with $(\omega, \alpha, \beta) = (0.1, 0.3, 0.5)$ and different choices of γ, the mis-specified model is the GARCH Model 5, and γ is the deviation parameter. The row $\gamma = 0.0$ is copied from Table 2.

	Quantile τ in $\text{CVaR}(1-\tau\|x) = Q(\tau\|x)$												
γ	1%	5%	10%	20%	30%	40%	50%	60%	70%	80%	90%	95%	99%
0.0	0.014	0.046	0.094	0.195	0.293	0.388	0.498	0.610	0.713	0.800	0.900	0.950	0.983
0.2	0.016	0.057	0.099	0.198	0.302	0.395	0.475	0.587	0.693	0.783	0.890	0.939	0.984
0.4	0.017	0.041	0.095	0.189	0.281	0.401	0.500	0.601	0.702	0.790	0.918	0.945	0.981
0.6	0.021	0.061	0.115	0.212	0.315	0.406	0.506	0.605	0.714	0.802	0.903	0.956	0.977
0.8	0.019	0.062	0.113	0.206	0.295	0.394	0.501	0.613	0.710	0.798	0.903	0.949	0.971

In our numerical studies, we also examined other model mis-specifications. For example, the true data-generating model is

$$\text{AR-GARCH:}\ \ Y_i = \gamma Y_{i-1} + \sigma_i \varepsilon_i \quad \sigma_i^2 = \omega + \alpha Y_{i-1}^2 + \beta \sigma_{i-1}^2.$$

The mis-specified model is the standard GARCH Model 5. Using different choices of γ, we found that the method still yields satisfactory performance. To keep the length, we did not include the results here.

2.5.5 Asymptotic versus bootstrap confidence intervals

The finite sample performance of CVaR confidence intervals has not been examined in the literature; in this section we compare the performance of the asymptotic confidence interval and the bootstrap confidence interval for CVaR: the asymptotic confidence interval is based on the asymptotic normality in Theorem 2 with estimated limiting variance in Section 2.1.3, and the bootstrap confidence interval is constructed using the procedure in Section 2.3 with $M = 1000$ bootstrap replications. The empirical coverage probability is the proportion of confidence intervals among 1000 realizations of $(1-\alpha)$ confidence intervals that cover the true $\text{CVaR}(1-\tau|x)$. Table 5 presents the results for the most typical setting $1-\tau = 5\%$ and $1-\alpha = 90\%, 95\%, 99\%$, at different values of x. Overall, the bootstrap confidence interval delivers much better performance and has empirical coverage probabilities close to the nominal levels. Given the easy implementation and superior empirical performance, we recommend the bootstrap confidence interval in practice.

Table 5: Empirical coverage probability of asymptotic and bootstrap confidence intervals (CI) for CVaR$(0.05|x)$.

(Model 1) CVaR$(0.05|x)$: x at different percentiles of covariates $\{Y_{i-1}\}$

		Asymptotic CI for CVaR$(0.05\|x)$ x at different percentiles						Bootstrap CI for CVaR$(0.05\|x)$ x at different percentiles					
	CI level												
noise	$1-\alpha$	25th	50th	75th	90th	95th	99th	25th	50th	75th	90th	95th	99th
$N(0,1)$	90%	85.8	83.3	82.5	82.6	84.4	85.2	90.4	88.7	88.5	87.8	88.2	88.2
	95%	91.7	89.3	88.7	89.2	90.0	90.7	94.6	94.1	93.9	94.1	94.9	95.1
	99%	96.0	95.7	95.5	96.2	96.8	97.6	98.2	97.8	97.6	98.7	98.6	98.9
$t_3/\sqrt{3}$	90%	83.2	82.2	82.8	83.2	83.2	86.2	90.0	89.8	90.4	91.3	92.0	92.2
	95%	87.6	88.0	87.6	88.6	90.0	91.4	94.1	94.4	95.2	94.8	95.8	95.6
	99%	94.2	94.2	94.6	96.0	96.6	97.2	98.8	98.4	98.4	98.6	98.8	98.8

(Model 2) CVaR$(0.05|(x_1,x_2))$: x_1 at different percentiles of $\{Y_{i-1}\}$, x_2 at median of $\{Y_{i-2}\}$

		Asymptotic CI for CVaR$(0.05\|(x_1,x_2))$ x_1 at different percentiles, x_2 at median						Bootstrap CI for CVaR$(0.05\|(x_1,x_2))$ x_1 at different percentiles, x_2 at median					
	CI level												
noise	$1-\alpha$	25th	50th	75th	90th	95th	99th	25th	50th	75th	90th	95th	99th
$N(0,1)$	90%	90.3	89.6	88.7	88.2	88.2	88.7	91.8	91.7	91.4	91.3	91.9	91.0
	95%	94.7	93.5	92.6	93.1	93.3	94.4	95.9	95.8	96.3	95.6	95.6	95.8
	99%	98.6	97.2	96.5	97.5	98.1	97.9	99.3	98.8	98.1	98.1	98.6	98.6
$t_3/\sqrt{3}$	90%	83.2	82.8	82.8	84.4	85.0	85.6	87.4	87.1	87.0	88.4	90.2	89.8
	95%	88.2	87.0	86.6	89.0	89.2	91.4	93.2	93.8	93.6	92.8	92.4	93.6
	99%	94.2	93.0	92.6	93.2	94.4	96.6	98.0	97.0	97.4	97.6	98.2	98.8

(Model 3) CVaR$(0.05|x)$: x at different percentiles of $\{Y_{i-1}\}$

		Asymptotic CI for CVaR$(0.05\|x)$ x at different percentiles						Bootstrap CI for CVaR$(0.05\|x)$ x at different percentiles					
	CI level												
noise	$1-\alpha$	25th	50th	75th	90th	95th	99th	25th	50th	75th	90th	95th	99th
$N(0,1)$	90%	100	100	99.8	97.7	94.2	85.0	89.0	87.3	88.3	89.1	88.3	89.0
	95%	100	100	100	99.8	98.5	91.7	93.6	93.6	94.2	93.3	93.6	95.0
	99%	100	100	100	100	100	97.7	97.7	98.8	98.3	98.1	98.3	98.8
$t_3/\sqrt{3}$	90%	99.3	99.3	98.3	96.3	94.1	86.6	85.6	90.7	90.2	86.3	86.1	85.9
	95%	99.8	99.8	99.0	98.0	97.6	91.0	91.4	94.1	94.2	91.2	89.7	89.9
	99%	100	100	99.5	99.0	98.5	95.8	97.8	98.0	97.3	97.3	96.8	96.1

(Model 4) CVaR$(0.05|(x_1,x_2))$: x_1 at different percentiles $\{Y_{i-1}\}$, x_2 at median of $\{U_i\}$

		Asymptotic CI for CVaR$(0.05\|(x_1,x_2))$ x_1 at different percentiles, x_2 at median						Bootstrap CI for CVaR$(0.05\|(x_1,x_2))$ x_1 at different percentiles, x_2 at median					
	CI level												
noise	$1-\alpha$	25th	50th	75th	90th	95th	99th	25th	50th	75th	90th	95th	99th
$N(0,1)$	90%	100	100	99.7	99.2	98.3	95.7	88.8	89.6	88.5	87.5	87.3	88.2
	95%	100	100	99.8	99.7	100	98.8	95.3	94.1	94.4	92.8	93.1	93.2
	99%	100	100	100	100	100	100	99.2	98.9	98.2	97.2	96.0	96.0
$t_3/\sqrt{3}$	90%	100	100	100	98.5	95.5	86.2	88.2	89.1	89.9	83.4	80.3	75.7
	95%	100	100	100	99.3	97.2	89.9	93.7	95.9	95.1	91.7	87.4	83.8
	99%	100	100	100	99.8	99.4	95.5	98.5	98.9	98.3	97.4	95.7	92.0

2.6 An Empirical Application to S&P 500 Index

As an illustration, we consider S&P 500 index daily loss defined as $Y_i = -[\log(S_i) - \log(S_{i-1})]$, where S_i is the index at day i. Figure 2.1 is a plot of the loss series $\{Y_i\}$ over the ten-year time period January 2004–December, 2013. There are $n = 2516$ observations. The plot clearly shows volatility clustering, so GARCH models are natural choices.

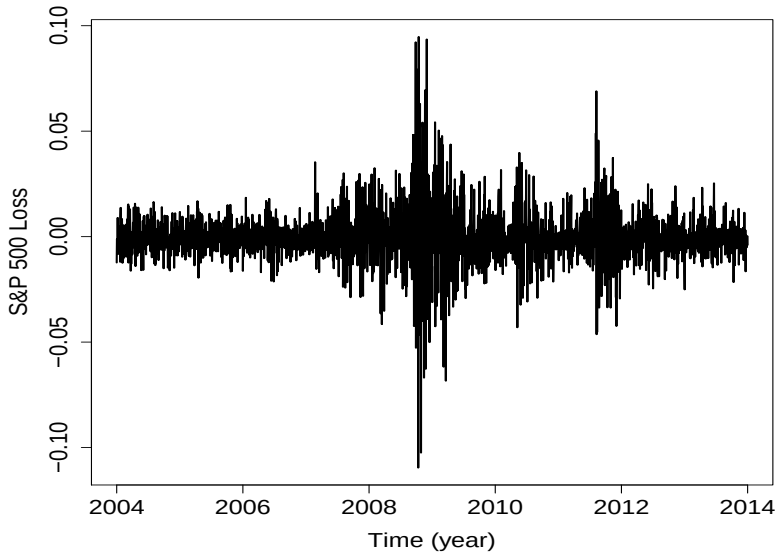

Figure 2.1. Time series plot of daily S&P 500 index loss $\{Y_i\}_{i=1}^n$ (i.e., negative logarithm return) during the ten-year period January 2004–December, 2013.

2.6.1 Comparison under different GARCH models

We consider sequential predictions of CVaR using three GARCH models: standard GARCH, EGARCH, and GJR-GARCH, as described in Model 5–7 in Section 2.5.2. For a given time i, based on the historical data $\mathbf{X}_i = \{Y_j\}_{j \le i-1}$, we apply our semiparametric CVaR method to obtain the estimate $\widehat{\text{CVaR}}(1-\tau|Y_j, j \le i-1)$ for the "unobservable" loss Y_i. Repeating the procedure for $i = n-(J-1), n-(J-2), \ldots, n$, we obtain the sequentially predicted CVaR for the last $J = 1000$ daily losses, which roughly corresponds to the daily losses during the last four years 2010–2013. Cai and Wang (2008) studied daily loss over the period 1998–2006 and nonparametrically estimated the CVaR curve for Y_i conditioning on $Y_{i-1} = x$. In our setting, due to the non-Markovian structure of GARCH models, it is infeasible to use their nonparametric approach.

Using the three GARCH models, Figure 2.2 plots the corresponding sequential CVaR predictions at level $1-\tau = 10\%$ (top plot), 5% (middle plot), and 1% (bottom plot). From Figure 2.2, at each level, the three CVaR curves based on standard GARCH, EGARCH, and GJR-GARCH exhibit quite similar pattern, indicating the robustness of our method.

Despite the vast literature on VaR/CVaR estimation, their confidence interval

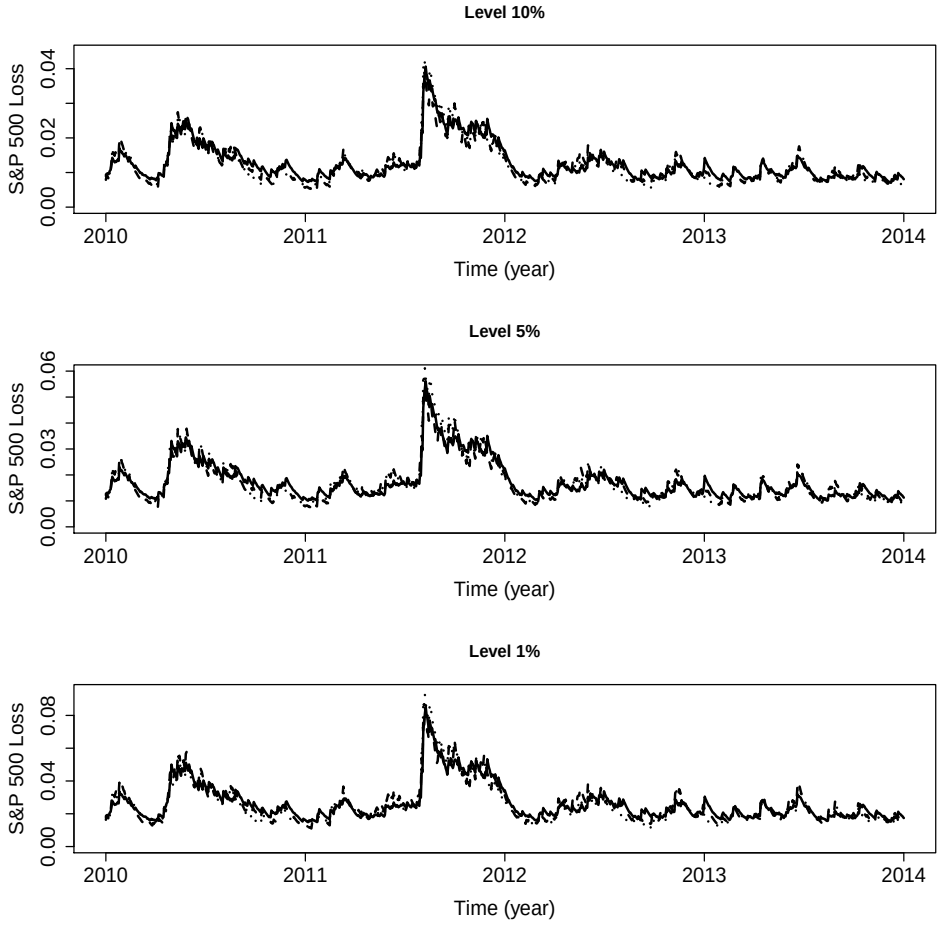

Figure 2.2. Sequentially predicted semiparametric CVaR for daily losses during 2010–2013, using standard GARCH (solid curve), EGARCH (dashed curve), and GJR-GARCH (dotted curve) models. Top, middle, and bottom plots correspond to level $1 - \tau = 10\%, 5\%, 1\%$, respectively.

construction has been largely ignored. Using the bootstrap procedure in Section 2.3, Figure 2.3 presents the semiparametrically estimated CVaR at level 5% and the corresponding pointwise 95% confidence interval. Due to the quite similar pattern of CVaR using different GARCH models, we only report the result for standard GARCH.

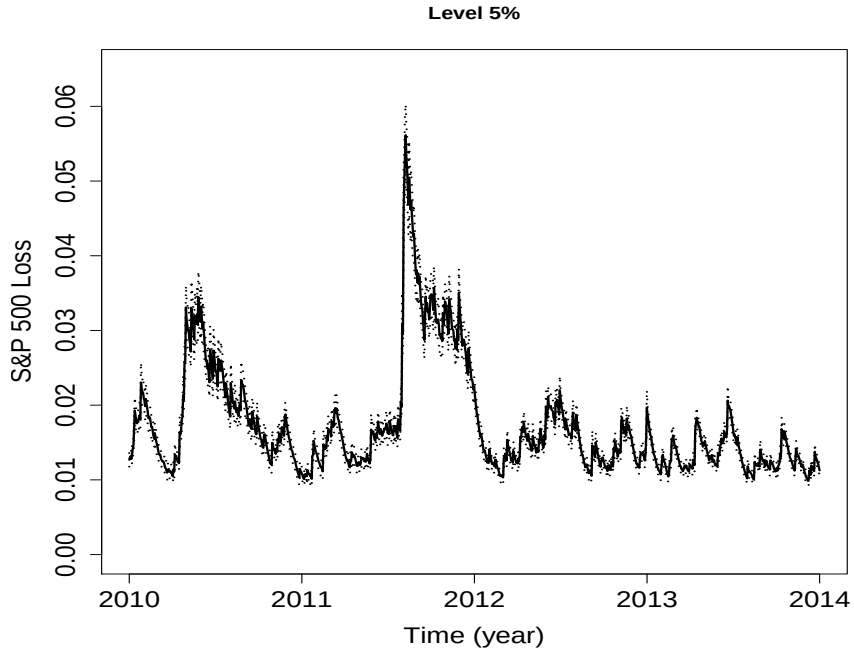

Figure 2.3. Sequentially predicted semiparametric CVaR (solid curve) at level 5% for daily losses during 2010–2013 using standard GARCH. The dotted curves are the pointwise bootstrap 95% confidence interval.

2.6.2 Comparison with some existing methods

We compare our semiparametric CVaR predictions with three parametric-distribution based approaches: the EWMA in RiskMetrics of J.P. Morgan (1994), the robust-EWMA in Guermat and Harris (2001), and the skewed-EWMA in Gerlach, Lu and Huang (2013). The EWMA is based on IGARCH model and RiskMetrics recommends the decay factor 0.94 for daily observations. For the robust-EWMA, Guermat and Harris (2001) found that a decay factor in the range $[0.92, 0.95]$ performs well, so we use the same decay factor 0.94 as the EWMA. To implement the skewed-EWMA in Gerlach, Lu and Huang (2013), we use their procedure with daily re-estimated parameters, including both the time-varying parameters in the asymmetric-Laplace distribution and the time-independent decay factors. Figure 2.4 plots the sequential CVaR predictions using the aforementioned four methods: the semiparametric method with standard GARCH (as shown in Figure 2.2, EGARCH and GJR-GARCH lead to similar curves and are omitted), EWMA, robust-EWMA, and skewed-EWMA. The four methods lead to quite similar CVaR

curves.

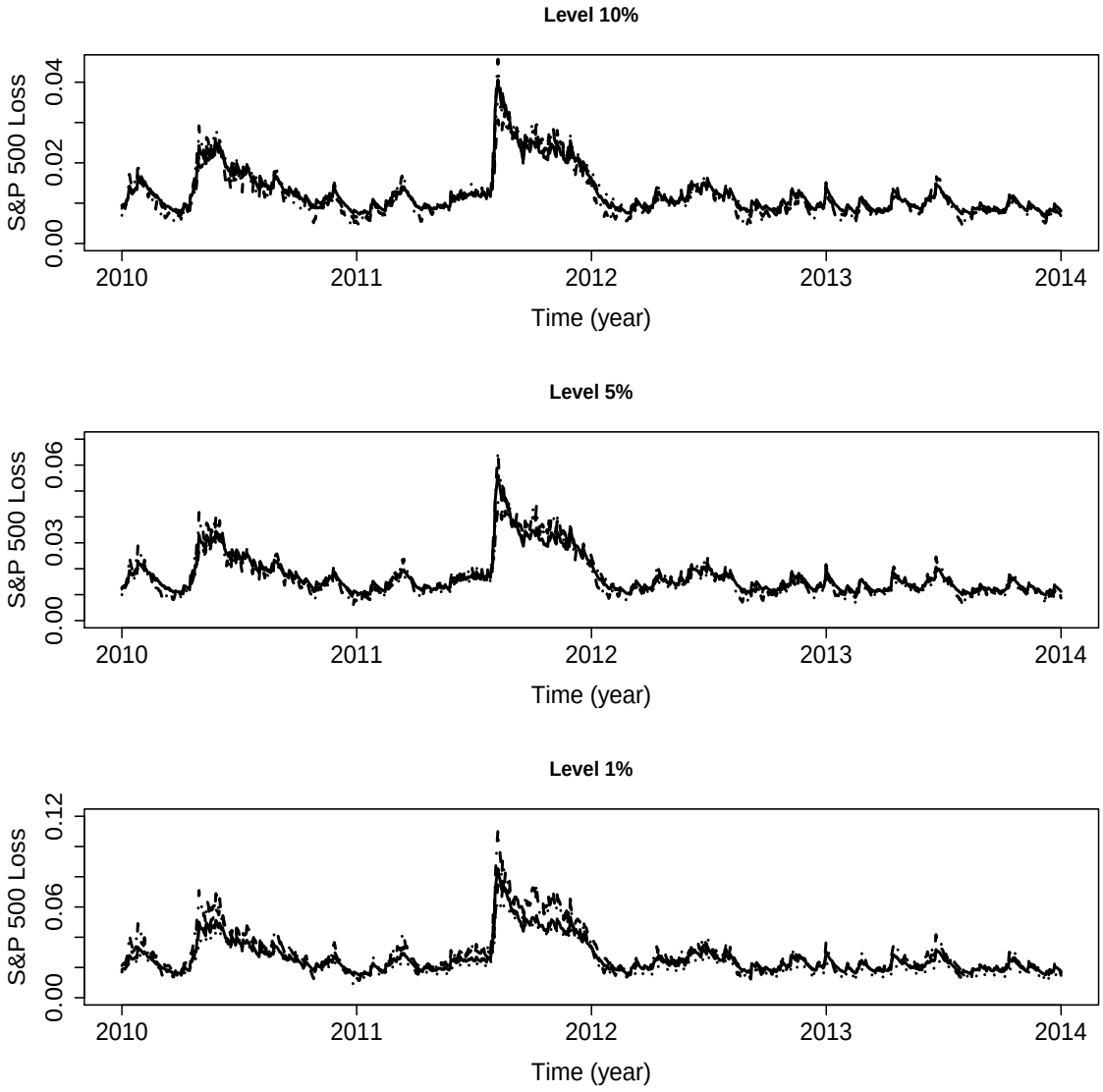

Figure 2.4. Comparison of sequentially predicted CVaR for daily losses during 2010–2013, using four methods: semiparametric method with standard GARCH (solid curve), the EWMA method (dotted curve), the robust-EWMA method (dashed curve), and the skewed-EWMA method (dotdashed curve). Top, middle, and bottom plots correspond to $1-\tau = 10\%, 5\%, 1\%$, respectively.

To numerically examine the accuracy of the four methods in predicting $\mathrm{CVaR}(1-\tau|Y_j, j \le i-1)$ at times $n-999 \le i \le n$, first we consider the empirical violation

rates (i.e., the empirical proportion that the observed loss exceeds the predicted CVaR):

$$\frac{\text{the number of } n-999 \le i \le n \text{ with } Y_i \ge \widehat{\text{CVaR}}(1-\tau|Y_j, j \le i-1)}{1000} \times 100\%.$$

Table 6 presents the empirical violation rates for the four methods at nominal levels $1-\tau = 10\%, 5\%, 1\%$. We can see: (i) The semiparametric method with EGARCH and the skewed-EWMA have comparable and top performance; (ii) The EWMA has the worst performance and substantially underestimates the risk at level 1% and 5%, i.e., the empirical violation rate is much higher than the nominal level; and (iii) the other methods rank in the middle. For the semiparametric method, the empirical violation rates are generally quite close to the nominal levels. Therefore, although the true CVaR is unknown, we conclude that the predicted CVaR should be reasonably close to the true CVaR.

Table 6: Empirical violation rates for four methods: the proposed semiparametric method with different GARCH models (standard GARCH, EGARCH, GJR-GARCH), the EWMA method, the robust-EWMA method, and the skewed-EWMA method. The bracketed number (2.5%) means that the violation rate is different from the nominal level, according to the unconditional coverage test at significance level 5%.

		Level $1-\tau$		
method		1%	5%	10%
semiparametric	standard GARCH	1.3%	5.0%	9.0%
	EGARCH	1.1%	5.1%	9.5%
	GJR-GARCH	1.6%	4.8%	9.2%
parametric	EWMA	(2.5%)	6.1%	9.0%
	robust-EWMA	0.8%	5.6%	9.7%
	skewed-EWMA	1.1%	5.6%	10.0%

Next, as in Gerlach, Lu and Huang (2013), we consider two statistical tests:

- (Unconditional coverage test). Denote by J_v the number of violations $Y_i \ge \widehat{\text{CVaR}}(1-\tau|Y_j, j \le i-1)$. Under the null hypothesis that the true violation rate is $1-\tau$ and the violations are independent, the binomial-distribution induced likelihood ratio test

  $$2[(J-J_v)\log(1-J_v/J) + J_v\log(J_v/J) - (J-J_v)\log(\tau) - J_v\log(1-\tau)]$$

  is asymptotically $\chi^2(1)$-distributed. In Table 6, bracketed number indicates rejection of the null hypothesis at significance level 5%. Thus, EWMA fails the test at nominal level $1-\tau = 1\%$ while other methods all pass the test.

- (Dynamic quantile test.) This test simultaneously tests the joint null hypothesis of correct violation rate and that the violations are uncorrelated over time. The idea is to regress the de-meaned "hit" variables on lagged "hit" variables and test for zero coefficients, i.e., testing for $\beta_0 = \cdots = \beta_p = 0$ in the linear regression

$$\mathbf{Hit}_i \sim \beta_0 + \sum_{r=1}^{p} \beta_r \mathbf{Hit}_{i-r}, \quad \mathbf{Hit}_i := \mathbf{1}_{Y_i \geq \widehat{\mathrm{CVaR}}(1-\tau|Y_j, j\leq i-1)} - (1-\tau). \tag{2.48}$$

  Here p is the order of the lagged regression. Denote by Y and X the corresponding response vector and covariate matrix of the above linear regression. Under the null hypothesis, $[\tau(1-\tau)]^{-1} Y^T X (X^T X)^{-1} X^T Y$ is asymptotically $\chi^2(p+1)$ distributed. See Engle and Manganelli (2004) for more details.

  At significance level 5%, Table 7 presents the results of the dynamic quantile test with different choices of order p in (2.48), where "N" indicates rejection of the null hypothesis. The semiparametric method with EGARCH or GJR-GARCH and the skewed-EWMA have the best performance and pass the test for all choices of order $p = 1, 2, 3, 4$ and nominal levels $1 - \tau = 10\%, 5\%, 1\%$, while the EWMA has the worst performance.

Combining the above analysis from violation rates, unconditional coverage test, and dynamic quantile test, we conclude that the semiparametric method with EGARCH and the skewed-EWMA have the best performance among the methods considered.

Table 7: Dynamic quantile test for the accuracy of the predicted $\widehat{\mathrm{CVaR}}(1-\tau|Y_j, j \leq i-1)$ at times $n - 999 \leq i \leq n$ using the proposed semiparametric method with different GARCH models (standard GARCH, EGARCH, GJR-GARCH), the EWMA method, the robust-EWMA method, and the skewed-EWMA method. "N" represents rejection of the joint null hypothesis of correct violation rate and that the violations are not correlated over time, at significance level 5%.

		Level $1-\tau$											
		order $p=1$			order $p=2$			order $p=3$			order $p=4$		
method		1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
semiparametric	standard GARCH				N			N			N		
	EGARCH												
	GJR-GARCH												
parametric	EWMA	N			N	N		N	N		N		
	robust-EWMA				N			N			N		
	skewed-EWMA												

2.6.3 Conditional VaR vs unconditional VaR

Another popular approach is the unconditional or marginal VaR, which uses the quantile of the marginal distribution of the losses. To estimate unconditional VaR, one can use parametric or nonparametric methods. For parametric methods, one imposes some parametric distribution, denoted by $F_\theta(x)$, for the losses, then the $(1-\tau)$ unconditional VaR is $F^{-1}_{\hat{\theta}}(\tau)$, where $\hat{\theta}$ is an estimate of θ. For example, a simple choice is the $N(\mu, \sigma^2)$ distribution with μ an σ^2 estimated from the sample. For nonparametric or historical simulation method, one uses the sample τ-th quantile of the losses to estimate the $(1-\tau)$ unconditional VaR.

Table 8 summarizes the empirical violation rates for the unconditional VaR using the aforementioned four methods. Comparing Table 6 and Table 8, we can clearly see the overall better performance of the conditional VaR over the unconditional VaR. It is generally believed that financial returns are uncorrelated but not independent. The marginal distribution ignores such dependence and thus leads to inferior performance.

Table 8: Empirical violation rates for unconditional VaR using four methods: Normal distribution, Student-t distribution, asymmetric Laplace distribution, and nonparametric estimate (historical simulation). The bracketed numbers mean that the violation rate is significantly (significance level 5%) different from the nominal level.

	Level $1-\tau$		
method	1%	5%	10%
Normal	0.9%	(2.9%)	(4.6%)
Student-t	0.7%	4.6%	10.2%
asymmetric-Laplace	0.7%	3.7%	8.2%
nonparametric estimate	0.4%	(3.5%)	8.3%

2.7 Assumptions and Proofs of Theorems

Throughout $C_1, C_2, c, c_1, c_2, \ldots$, are generic constants that may vary from line to line.

2.7.1 Technical conditions and some preliminary results

We list some technical conditions and preliminary results used in the proof of our theorems.

Assumption 3. *In model (2.3): (i) the innovations* $\{\varepsilon_i\}_{i\in\mathbb{Z}}$ *are i.i.d.. Denote by* $\mathcal{F}_i$ *the sigma-algebra generated by* $\{\mathbf{X}_{i+1}, \mathbf{X}_i, \ldots; \varepsilon_i, \varepsilon_{i-1}, \ldots\}$. *For each* i, ε_i *is independent of* $\mathcal{F}_{i-1}$. *(ii)* $\{(\mathbf{X}_i, \varepsilon_i)\}_{i\in\mathbb{Z}}$ *is stationary and* α*-mixing with mixing coefficient* $\alpha_j \le C_1 j^{-\alpha}$ *for some constants* $0 < C_1 < \infty$ *and* $\alpha > 8 + 4k$, *where* k *is the dimension of the parameter* θ.

In Assumption 3(ii), the α-mixing condition is the most commonly used dependence assumption [Fan and Yao (2003)].

Assumption 4. *Recall* $H(\theta, Y_i, \mathbf{X}_i)$ *in (2.9). Let* $\epsilon > 0$ *be some small constant.*
(i) $G(\theta, \varepsilon, x)$ *is continuously differentiable in* θ *and* ε.
(ii) There exists $C_2 > 0$ *such that*

$$\mathbb{E}\left[\sup_{|\vartheta|\le\epsilon, z\in\mathbb{R}} \mathbf{1}_{|H(\theta+\vartheta, Y_0, \mathbf{X}_0)-z|\le v}\right] \le C_2 v, \quad \text{for all } v > 0. \tag{2.49}$$

(iii) Let $\dot{H}(\theta, Y_i, \mathbf{X}_i)$ *be the partial derivative with respect to* θ. *There exists* $L(Y_i, \mathbf{X}_i)$,

$$\sup_{|\vartheta|\le\epsilon} |\dot{H}(\theta+\vartheta, Y_i, \mathbf{X}_i)| \le L(Y_i, \mathbf{X}_i) \quad \text{and} \quad L(Y_i, \mathbf{X}_i) \in \mathcal{L}^1. \tag{2.50}$$

(iv) Define

$$J(\vartheta, z) = \mathbb{P}\{H(\theta+\vartheta, Y_0, \mathbf{X}_0) \le z\}, \quad \vartheta \in \mathbb{R}^k, z \in \mathbb{R}. \tag{2.51}$$

Let $\dot{J}(\vartheta, z)$ *and* $\ddot{J}(\vartheta, z)$ *be the gradient vector and Hessian matrix of* $J(\vartheta, z)$ *with respect to* ϑ. $\dot{J}(0, z)$ *is continuous in* z *and* $|\ddot{J}(\vartheta, z)|$ *is bounded on* $|\vartheta| \le \epsilon, z \in \mathbb{R}$.

Intuitively, (2.49) asserts that the probability mass of $H(\theta+\vartheta, Y_0, \mathbf{X}_0)$ on $[z-v, z+v]$ is $O(v)$ uniformly, which is reasonable if the density is uniformly bounded. The condition (2.50) is used to control errors in Taylor's expansions. In view of $\hat{\varepsilon}_i = H(\hat{\theta}, Y_i, \mathbf{X}_i)$ in (2.10), $J(\vartheta, z)$ in (2.51) measures how the distribution function of $\hat{\varepsilon}_i$ changes in response to the departure ϑ from the true parameter θ. In particular, $J(0, z) = \mathbb{P}\{\varepsilon_0 \le z\}$.

Assumption 5. *Density* $f_\varepsilon(z)$ *of* ε_i *is continuous, bounded, positive at* $z = Q_\varepsilon(\tau)$.

Assumption 5*. *Density $f_\varepsilon(z)$ of ε_i is continuous, bounded, positive on the interval $[Q_\varepsilon(\delta), Q_\varepsilon(1-\delta)]$.*

Definition 1. *Recall $\mathcal{F}_i$ in Assumption 3. We say that a function $g(z, \mathcal{F}_i)$ (may be vector or matrix valued) is stochastically continuous at a point z^* if*

$$\lim_{\epsilon \to 0} \mathbb{E}[\mathcal{G}(\epsilon, \mathcal{F}_i)] = 0, \quad \textit{where} \quad \mathcal{G}(\epsilon, \mathcal{F}_i) = \sup_{|z - z^*| \le \epsilon} |g(z, \mathcal{F}_i) - g(z^*, \mathcal{F}_i)|. \tag{2.52}$$

In (2.52), $\mathcal{G}(\epsilon, \mathcal{F}_i)$ quantifies the maximal fluctuation of $g(z, \mathcal{F}_i)$ in the ϵ-neighborhood of z^*, and the condition $\lim_{\epsilon \to 0} \mathbb{E}[\mathcal{G}(\epsilon, \mathcal{F}_i)] = 0$ asserts that the maximal fluctuation asymptotically vanishes under expectation, which suggests "stochastic continuity". The stochastic continuity extends the continuity of deterministic functions to that of stochastic functions.

Assumption 6. *$\dot{G}_\theta(\theta, \varepsilon, \mathbf{X}_i)/\dot{G}_\varepsilon(\theta, \varepsilon, \mathbf{X}_i)$ is stochastically continuous at $(\theta, Q_\varepsilon(\tau))$. For $H(\theta, Y_i, \mathbf{X}_i)$ in (2.9) and $D(\theta, \varepsilon_i, \mathbf{X}_i)$ in Assumption 2, write*

$$D_i(\vartheta) = D(\vartheta, H(\vartheta, Y_i, \mathbf{X}_i), \mathbf{X}_i),$$

with $D_i(\vartheta)$ and $D_i(\vartheta) D_i(\vartheta)^T$ being stochastically continuous at $\vartheta = \theta$.

Lemma 1. *Suppose that $g(z, \mathcal{F}_i)$ is stochastically continuous at $z = z^*$ and that $g(z^*, \mathcal{F}_i) \in \mathcal{L}^1$. Then for any random sequence $z_n \xrightarrow{p} z^*$, $n^{-1} \sum_{i=1}^n g(z_n, \mathcal{F}_i) = \mathbb{E}[g(z^*, \mathcal{F}_0)] + o_p(1)$.*

Proof. Let $\epsilon > 0$ be any given small number. Since $z_n \xrightarrow{p} z^*$, with probability tending to one, $|z_n - z^*| \le \epsilon$. On the event $\{|z_n - z^*| \le \epsilon\}$ [recall $\mathcal{G}(\epsilon, \mathcal{F}_i)$ defined in (2.52)],

$$\left| \frac{1}{n} \sum_{i=1}^n g(z_n, \mathcal{F}_i) - \frac{1}{n} \sum_{i=1}^n g(z^*, \mathcal{F}_i) \right| \le \frac{1}{n} \sum_{i=1}^n \mathcal{G}(\epsilon, \mathcal{F}_i) \quad \xrightarrow{p} \quad \mathbb{E}[\mathcal{G}(\epsilon, \mathcal{F}_0)], \tag{2.53}$$

where the last convergence follows from the ergodic theorem (the mixing condition in Assumption 3 implies ergodicity). The result then follows from (2.52)–(2.53) and the ergodic theorem $n^{-1} \sum_{i=1}^n g(z^*, \mathcal{F}_i) \xrightarrow{p} \mathbb{E}[g(z^*, \mathcal{F}_0)]$. ♢

Lemma 2. *Let z and z' be any real numbers. Then for any $c > 0$,*

$$|\mathbf{1}_{z \le 0} - \mathbf{1}_{z' \le 0}| \le 2 \mathbf{1}_{|z - z'| \ge c} + \mathbf{1}_{|z'| < c}. \tag{2.54}$$

Proof. Notice that $|\mathbf{1}_{z\leq 0} - \mathbf{1}_{z'\leq 0}| = \mathbf{1}_{z\leq 0<z'} + \mathbf{1}_{z'\leq 0<z}$. The result then follows from

$$\mathbf{1}_{z\leq 0<z'} \;=\; \mathbf{1}_{z\leq 0<z',|z-z'|\geq c} + \mathbf{1}_{z\leq 0<z',|z-z'|<c} \leq \mathbf{1}_{|z-z'|\geq c} + \mathbf{1}_{0<z'<c},$$

and similarly $\mathbf{1}_{z'\leq 0<z} \leq \mathbf{1}_{|z-z'|\geq c} + \mathbf{1}_{-c<z'\leq 0}$. ♢

Lemma 3. *let $\{\xi_i\}_{i\in\mathbb{Z}}$ be a stationary α-mixing process with mixing coefficient $\alpha_j, j \in \mathbb{N}$. Assume $\mathbb{E}(\xi_0) = 0$ and $|\xi_i| \leq c$ for some c. Then, for $\ell = 1, \ldots, \lfloor n/2 \rfloor$ and $z > 0$,*

$$\mathbb{P}\left\{\left|\sum_{i=1}^{n}\xi_i\right| > z\right\} \;\leq\; 4\exp\left\{-\frac{z^2\ell}{144n^2\mathbb{E}(\xi_0^2) + 4czn}\right\} + 22\ell\alpha_{\lfloor n/(2\ell)\rfloor}\sqrt{1+\frac{4cn}{z}}. \tag{2.55}$$

Proof. Theorem 2.18 in Fan and Yao (2003) presents a slightly different version of (2.55) with the term $144n^2\mathbb{E}(\xi_0^2)$ replaced by $16n^2\Gamma_r/r^2$, where $r = n/(2\ell)$,

$$\Gamma_r = \max_{0\leq j\leq 2\ell-1}\mathbb{E}\Big\{(\lfloor jr\rfloor + 1 - jr)\xi_1 + \xi_2 + \cdots + \xi_s + (jr + r - \lfloor jr + r\rfloor)\xi_{s+1}\Big\}^2,$$

and $s = \lfloor (j+1)r\rfloor - \lfloor jr\rfloor$. The result then follows from the Cauchy-Schwarz inequality $\Gamma_r \leq (s+1)\mathbb{E}(\xi_1^2 + \cdots + \xi_{s+1}^2) \leq \lfloor r+2\rfloor^2\mathbb{E}(\xi_0^2) \leq 9r^2\mathbb{E}(\xi_0^2)$. ♢

2.7.2 Proof of Theorem 1

Lemma 4. *Recall $J(\vartheta, z)$ and $\dot{J}(\vartheta, z)$ in Assumption 4(iv). Then*

$$\dot{J}(0,z) = f_\varepsilon(z)\mathbb{E}\left[\frac{\dot{G}_\theta(\theta, z, \mathbf{X}_0)}{\dot{G}_\varepsilon(\theta, z, \mathbf{X}_0)}\right]. \tag{2.56}$$

Proof. Under Assumption 1,

$$\begin{aligned}\{H(\theta+\vartheta, Y_0, \mathbf{X}_0) \leq z\} \;&\Leftrightarrow\; \{Y_0 \leq G(\theta+\vartheta, z, \mathbf{X}_0)\}\\ &\Leftrightarrow\; \{G(\theta, \varepsilon_0, \mathbf{X}_0) \leq G(\theta+\vartheta, z, \mathbf{X}_0)\}\\ &\Leftrightarrow\; \{\varepsilon_0 \leq H(\theta, G(\theta+\vartheta, z, \mathbf{X}_0), \mathbf{X}_0)\}.\end{aligned}$$

Thus,

$$J(\vartheta, z) = \mathbb{P}\{\varepsilon_0 \le H(\theta, u, \mathbf{X}_0)\} = \mathbb{E}[F_\varepsilon(H(\theta, u, \mathbf{X}_0))], \quad u = G(\theta + \vartheta, z, \mathbf{X}_0).$$

By the chain rule,

$$\frac{\partial J(\vartheta, z)}{\partial \vartheta} = \mathbb{E}\left[f_\varepsilon(H(\theta, u, \mathbf{X}_0)) \frac{\partial H(\theta, u, \mathbf{X}_0)}{\partial u} \frac{\partial u}{\partial \vartheta} \right].$$

Note that

$$\begin{aligned} \partial H(\theta, u, \mathbf{X}_0)/\partial u &= 1/\dot{G}_\varepsilon(\theta, H(\theta, u, \mathbf{X}_0), \mathbf{X}_0) \text{ and} \\ \partial u/\partial \vartheta &= \dot{G}_\theta(\theta + \vartheta, z, \mathbf{X}_0). \end{aligned}$$

When $\vartheta = 0$, $H(\theta, u, \mathbf{X}_0) = H(\theta, G(\theta, z, \mathbf{X}_0), \mathbf{X}_0) = z$. This completes the proof. $\diamondsuit$

Proof of Theorem 1. To reflect the dependence of $\hat{F}_\varepsilon(z)$ on $\hat{\theta}$ and in view of $\hat{\varepsilon}_i = H(\hat{\theta}, Y_i, \mathbf{X}_i)$, write ϑ as the departure from the true parameter θ and define

$$\hat{F}_\varepsilon(\vartheta, z) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{H(\theta+\vartheta, Y_i, \mathbf{X}_i) \le z}. \tag{2.57}$$

Then $\hat{F}_\varepsilon(z) = \hat{F}_\varepsilon(\hat{\theta} - \theta, z)$. By the expression for $\dot{J}(0, z)$ in (2.56) and the assumption $\hat{\theta} - \theta = O_p(n^{-1/2})$, in order to prove (2.13), it suffices to prove that, for all given $c_1 > 0$,

$$\sup_{|\vartheta| \le c_1/\sqrt{n}, |z| \le c} \left| \hat{F}_\varepsilon(\vartheta, z) - \hat{F}_\varepsilon(0, z) - \dot{J}(0, z)^T \vartheta \right| = o_p(n^{-1/2}). \tag{2.58}$$

For simplicity we assume that: (i) $k = 1$, i.e., ϑ is scalar-valued, (ii) $c_1 = c = 1$, and (iii) $\vartheta \in [0, 1/\sqrt{n}]$ and $z \in [0, 1]$, since the general k-dimensional case follows similarly. Let

$$\xi_i(\vartheta, z) = \mathbf{1}_{H(\theta+\vartheta, Y_i, \mathbf{X}_i) \le z} - \mathbf{1}_{H(\theta, Y_i, \mathbf{X}_i) \le z}. \tag{2.59}$$

Recall $J(\vartheta, z)$ in (2.51). By Taylor's expansion, $\mathbb{E}[\xi_i(\vartheta, z)] = J(\vartheta, z) - J(0, z) = \dot{J}(0, z)^T \vartheta + O(n^{-1})$ uniformly on $\vartheta \in [0, 1/\sqrt{n}], z \in [0, 1]$. Thus, to prove (2.58), it

suffices to prove

$$\sup_{(\vartheta,z)\in[0,1/\sqrt{n}]\times[0,1]} |M(\vartheta,z)| \;=\; o_p(\sqrt{n}), \text{ where } M(\vartheta,z)=\sum_{i=1}^{n}\{\xi_i(\vartheta,z)-\mathbb{E}[\xi_i(\vartheta,z)]\}. \tag{2.60}$$

To prove (2.60), we adopt a chain argument. Let $N=\lfloor n^{1+\epsilon}\rfloor$ with $\epsilon>0$ to be determined later, and consider the evenly spaced points

$$\vartheta_j=j\omega_1 \quad \text{with} \quad \omega_1=\frac{1}{\sqrt{n}N}, \quad z_j=j\omega_2 \quad \text{with} \quad \omega_2=\frac{1}{N}, \qquad j=0,1,\ldots,N.$$

The $(N+1)^2$ grid points $\{(\vartheta_j,z_{j'})\}_{j,j'=0}^{N}$ partition $[0,1/\sqrt{n}]\times[0,1]$ into N^2 cells. For each $(\vartheta,z)\in[0,1/\sqrt{n}]\times[0,1]$, there exists one grid point $(\vartheta_j,z_{j'})$ such that $|\vartheta-\vartheta_j|\le\omega_1$ and $|z-z_{j'}|\le\omega_2$. Thus, by $|M(\vartheta,z)|\le|M(\vartheta_j,z_{j'})|+|M(\vartheta,z)-M(\vartheta_j,z_{j'})|$, we have

$$\sup_{(\vartheta,z)\in[0,1/\sqrt{n}]\times[0,1]} |M(\vartheta,z)| \le \max_{0\le j,j'\le N}|M(\vartheta_j,z_{j'})|+R_n, \tag{2.61}$$

where

$$R_n=\sup_{\Omega}|M(\vartheta,z)-M(\vartheta',z')| \quad \text{with} \quad \Omega=\{|\vartheta-\vartheta'|\le\omega_1,|z-z'|\le\omega_2\}. \tag{2.62}$$

It is easy to see that

$$\sup_{\Omega}|\xi_i(\vartheta,z)-\xi_i(\vartheta',z')|\le 2\sup_{\Omega}\Big[\big|\mathbf{1}_{H(\theta+\vartheta,Y_i,\mathbf{X}_i)\le z}-\mathbf{1}_{H(\theta+\vartheta',Y_i,\mathbf{X}_i)\le z'}\big|\Big]. \tag{2.63}$$

Therefore, by (2.60), (2.62) and (2.63),

$$\begin{aligned}\mathbb{E}(R_n) &\le 4n\mathbb{E}\left\{\sup_{\Omega}\Big[\big|\mathbf{1}_{U(\vartheta)-z\le 0}-\mathbf{1}_{U(\vartheta')-z'\le 0}\big|\Big]\right\} \quad \text{with} \quad U(\vartheta)=H(\theta+\vartheta,Y_0,\mathbf{X}_0)\\ &\le 4n\mathbb{E}\left\{\sup_{\Omega}\Big[2\mathbf{1}_{|[U(\vartheta)-U(\vartheta')]-(z-z')|\ge\lambda}+\mathbf{1}_{|U(\vartheta)-z|<\lambda}\Big]\right\},\end{aligned} \tag{2.64}$$

where the second "$\le$" follows from Lemma 2 and $\lambda>0$ is any given number. By Assumption 4(iii), on Ω, $|[U(\vartheta)-U(\vartheta')]-(z-z')|\le\omega_1 L(Y_0,\mathbf{X}_0)+\omega_2$. Thus, by

(2.64),

$$\begin{aligned}\mathbb{E}(R_n) &\leq 8n\mathbb{P}\{\omega_1 L(Y_0, \mathbf{X}_0) + \omega_2 \geq \lambda\} + 4n\mathbb{E}\left\{\sup_{\Omega} \mathbf{1}_{|U(\vartheta)-z|<\lambda}\right\} \\ &\leq 8n\frac{\omega_1 \mathbb{E}|L(Y_0, \mathbf{X}_0)| + \omega_2}{\lambda} + O(n\lambda), \end{aligned} \tag{2.65}$$

where, in the second "$\leq$", the first term follows from Markov inequality and the second term follows from Assumption 4(ii). Letting $\lambda = \sqrt{\omega_2}$, we have $R_n = O_p(n\sqrt{\omega_2}) = o_p(\sqrt{n})$, where the last equality follows from $\omega_2 = N^{-1}$ and $N = \lfloor n^{1+\epsilon} \rfloor$ with $\epsilon > 0$.

By (2.61), it remains to prove $\max_{0\leq j,j'\leq N} |M(\vartheta_j, z_{j'})| = o_p(\sqrt{n})$. Note that, in (2.59), $\xi_1(\vartheta, z) = -1, 0$, or $+1$, and we in each of the three cases always have $\xi_1^2(\vartheta, z) = |\xi_1(\vartheta, z)| = |\mathbf{1}_{U(\vartheta)-z\leq 0} - \mathbf{1}_{U(0)-z\leq 0}|$ with $U(\vartheta)$ defined in (2.64). Therefore, by Lemma 2,

$$\begin{aligned}\mathrm{var}\{\xi_1(\vartheta, z)\} &\leq \mathbb{E}|\mathbf{1}_{U(\vartheta)-z\leq 0} - \mathbf{1}_{U(0)-z\leq 0}| \\ &\leq \mathbb{E}\Big[2\mathbf{1}_{|U(\vartheta)-U(0)|\geq n^{-1/4}} + \mathbf{1}_{|U(\vartheta)-z|<n^{-1/4}}\Big]. \end{aligned} \tag{2.66}$$

By Assumption 4(iii), $|U(\vartheta) - U(0)| \leq n^{-1/2} L(Y_0, \mathbf{X}_0)$ on $|\vartheta| \leq 1/\sqrt{n}$. Applying the latter inequality to (2.66) and by the same argument in (2.65) [i.e., Markov inequality and Assumption 4(ii)], we obtain that, there exists some constant c_2 such that

$$\mathrm{var}\{\xi_1(\vartheta, z)\} \leq c_2 n^{-1/4}, \quad \text{uniformly on } |\vartheta| \leq 1/\sqrt{n}, z \in \mathbb{R}. \tag{2.67}$$

Note that $|\xi_i(\vartheta, z) - \mathbb{E}[\xi_i(\vartheta, z)]| \leq 2$. By Lemma 3, for any $c_3 > 0$ and $\ell = 1, \ldots, \lfloor n/2 \rfloor$,

$$\begin{aligned}\mathbb{P}\{|M(\vartheta, z)| \geq c_3\sqrt{n}\} &\leq 4\exp\left(-\frac{c_3^2 n\ell}{144c_2 n^{7/4} + 8c_3 n^{3/2}}\right) \\ &+ 22\ell\alpha_{\lfloor n/\ell \rfloor}\sqrt{1 + \frac{8n}{c_3\sqrt{n}}}. \end{aligned} \tag{2.68}$$

Let $\ell = \lfloor n^\beta \rfloor$ with some $\beta \in (3/4, 1)$ to be determined later. Recall $\alpha_j \leq C_1 j^{-\alpha}$ in

Assumption 3(ii). Thus, from (2.68), there exists constants c_4 and c_5 such that

$$\mathbb{P}\{|M(\vartheta, z)| \geq c_3\sqrt{n}\} \leq c_5\Big[\exp\{-c_4 n^{\beta-\frac{3}{4}}\} + n^{\beta(1+\alpha)+\frac{1}{4}-\alpha}\Big], \tag{2.69}$$

uniformly over $|\vartheta| \leq 1/\sqrt{n}, z \in \mathbb{R}$. Recall that $N = \lfloor n^{1+\epsilon} \rfloor$. By (2.69),

$$\begin{aligned}\mathbb{P}\left\{\sup_{0\leq j,j'\leq N} |M(\vartheta_j, z_{j'})| \geq c_3\sqrt{n}\right\} &\leq \sum_{j,j'=0}^{N} \mathbb{P}\{|M(\vartheta_j, z_{j'})| \geq c_3\sqrt{n}\} \\ &= O\Big[n^{2+2\epsilon}\exp\{-c_4 n^{\beta-\frac{3}{4}}\} + n^{2\epsilon+\frac{9}{4}+\beta(1+\alpha)-\alpha}\Big].\end{aligned} \tag{2.70}$$

From Assumption 3(ii), $\alpha > 12$ ($k = 1$), which implies $3/4 < (\alpha - 9/4)/(1+\alpha)$. Take any

$$\frac{3}{4} < \beta < \frac{\alpha - 9/4}{1+\alpha} \qquad \text{and} \qquad \epsilon = \frac{\alpha - 9/4 - \beta(1+\alpha)}{3} > 0.$$

Then it is easy to see that the right hand side of (2.70) goes to zero. Since c_3 is arbitrary, we conclude $\sup_{0\leq j,j'\leq N} |M(\vartheta_j, z_{j'})| = o_p(\sqrt{n})$. This completes the proof. ♢

2.7.3 Proof of Theorem 2

Lemma 5. *Let $d(\cdot,\cdot)$ be a measurable function such that $d(\varepsilon_i, \mathbf{X}_i) \in \mathcal{L}^2$ and $\mathbb{E}[d(\varepsilon_i, \mathbf{X}_i)|\mathbf{X}_i] = 0$. Suppose Assumption 3 holds. Then $n^{-1/2}\sum_{i=1}^{n} d(\varepsilon_i, \mathbf{X}_i) \Rightarrow N(0, \mathbb{E}[d^2(\varepsilon_0, \mathbf{X}_0)])$.*

Proof. Let $\mathcal{F}_{i-1}$ be defined as in Assumption 3. By the independence between ε_i and $\mathcal{F}_{i-1}$ (Assumption 3) and the condition $\mathbb{E}[d(\varepsilon_i, \mathbf{X}_i)|\mathbf{X}_i] = 0$, we have

$$\mathbb{E}[d(\varepsilon_i, \mathbf{X}_i)|\mathcal{F}_{i-1}] = 0.$$

Thus, $\{d(\varepsilon_i, \mathbf{X}_i)\}$ are stationary martingale differences with respect to $\{\mathcal{F}_i\}$. The mixing condition in Assumption 3(ii) implies the ergodicity of $\{d(\varepsilon_i, \mathbf{X}_i)\}$. The result then follows from the CLT for martingales with stationary and ergodic increments. ♢

Proof of Theorem 2. In what follows we shall prove only the CLT in (ii) under Assumption 2. The same argument can be used to prove the weaker assertion $\hat{Q}(\tau|x) = Q(\tau|x) + O_p(n^{-1/2})$ in (i) under the weaker condition $\hat{\theta} = \theta + O_p(n^{-1/2})$.

The empirical quantile function $\hat{Q}_\varepsilon(\tau)$ is a solution to

$$\min_\nu \sum_{i=1}^n \rho_\tau(\hat{\varepsilon}_i - \nu).$$

Here $\rho_\tau(v) = v(\tau - \mathbf{1}_{v\leq 0})$ is the check function. Thus, $\hat{Q}_\varepsilon(\tau)$ is also a minimizer of the function (as a function of ν) $\sum_{i=1}^n [\rho_\tau(\hat{\varepsilon}_i - \nu) - \rho_\tau(\hat{\varepsilon}_i - Q_\varepsilon(\tau))]$. Let the transformation $\Delta = \sqrt{n}[\nu - Q_\varepsilon(\tau)]$. Then $\hat{\varepsilon}_i - \nu = \hat{\varepsilon}_i - Q_\varepsilon(\tau) - \Delta/\sqrt{n}$. Thus, $\hat{\Delta} := \sqrt{n}[\hat{Q}_\varepsilon(\tau) - Q_\varepsilon(\tau)]$ is a minimizer of the re-parametrized minimization problem $\min_\Delta S(\Delta)$, where

$$S(\Delta) = \sum_{i=1}^n \Big[\rho_\tau\{\hat{\varepsilon}_i - Q_\varepsilon(\tau) - \Delta/\sqrt{n}\} - \rho_\tau\{\hat{\varepsilon}_i - Q_\varepsilon(\tau)\}\Big].$$

By Knight's identity $\rho_\tau(u - v) - \rho_\tau(u) = -v(\tau - \mathbf{1}_{u<0}) + \int_0^v (\mathbf{1}_{u\leq s} - \mathbf{1}_{u\leq 0})ds$, we can rewrite

$$S(\Delta) = -\frac{\Delta}{\sqrt{n}} \sum_{i=1}^n [\tau - \mathbf{1}_{\hat{\varepsilon}_i < Q_\varepsilon(\tau)}] + \int_0^{\frac{\Delta}{\sqrt{n}}} \sum_{i=1}^n [\mathbf{1}_{\hat{\varepsilon}_i \leq Q_\varepsilon(\tau)+s} - \mathbf{1}_{\hat{\varepsilon}_i \leq Q_\varepsilon(\tau)}]ds. \quad (2.71)$$

For the second term, by the uniform approximation of $\hat{F}_\varepsilon(z)$ in Theorem 1, for fixed Δ,

$$\int_0^{\frac{\Delta}{\sqrt{n}}} \sum_{i=1}^n [\mathbf{1}_{\hat{\varepsilon}_i \leq Q_\varepsilon(\tau)+s} - \mathbf{1}_{\hat{\varepsilon}_i \leq Q_\varepsilon(\tau)}]ds \quad = \quad I_1 + I_2 + o_p(1), \quad (2.72)$$

where

$$I_1 \;=\; \sum_{i=1}^n \int_0^{\frac{\Delta}{\sqrt{n}}} [\mathbf{1}_{\varepsilon_i \leq Q_\varepsilon(\tau)+s} - \mathbf{1}_{\varepsilon_i \leq Q_\varepsilon(\tau)}]ds, \quad (2.73)$$

$$I_2 \;=\; n \int_0^{\frac{\Delta}{\sqrt{n}}} [\dot{J}(0, Q_\varepsilon(\tau) + s) - \dot{J}(0, Q_\varepsilon(\tau))]^T (\hat{\theta} - \theta)ds. \quad (2.74)$$

For I_1, we have

$$\mathbb{E}(I_1) = n\int_0^{\frac{\Delta}{\sqrt{n}}}[F_\varepsilon(Q_\varepsilon(\tau)+s) - F_\varepsilon(Q_\varepsilon(\tau))]ds \quad \to \quad \frac{f_\varepsilon(Q_\varepsilon(\tau))}{2}\Delta^2, \qquad (2.75)$$

where the last convergence follows from the Taylor expansion $F_\varepsilon(Q_\varepsilon(\tau)+s) - F_\varepsilon(Q_\varepsilon(\tau)) = sf_\varepsilon(Q_\varepsilon(\tau)) + o(s)$. Note that $|\mathbf{1}_{\varepsilon_i \le Q_\varepsilon(\tau)+s} - \mathbf{1}_{\varepsilon_i \le Q_\varepsilon(\tau)}| \le \mathbf{1}_{|\varepsilon_i - Q_\varepsilon(\tau)| \le \frac{|\Delta|}{\sqrt{n}}}$ uniformly on $|s| \le |\Delta|/\sqrt{n}$. Thus, by the i.i.d. assumption of $\{\varepsilon_i\}$,

$$\begin{aligned} \mathrm{var}(I_1) &= \sum_{i=1}^n \mathrm{var}\left\{\int_0^{\frac{\Delta}{\sqrt{n}}}[\mathbf{1}_{\varepsilon_i \le Q_\varepsilon(\tau)+s} - \mathbf{1}_{\varepsilon_i \le Q_\varepsilon(\tau)}]ds\right\} \\ &\le \sum_{i=1}^n \mathbb{E}\left\{\left[\frac{\Delta}{\sqrt{n}}\mathbf{1}_{|\varepsilon_i - Q_\varepsilon(\tau)| \le |\Delta|/\sqrt{n}}\right]^2\right\} \quad = \quad O(n^{-1/2}). \qquad (2.76) \end{aligned}$$

Here the last convergence follows from the continuity and boundedness of $f_\varepsilon(z)$ at $z = Q_\varepsilon(\tau)$, which implies $\mathbb{P}\{|\varepsilon_i - Q_\varepsilon(\tau)| \le |\Delta|/\sqrt{n}\} = O(n^{-1/2})$. By (2.75)–(2.76), we have $I_1 = \Delta^2 f_\varepsilon(Q_\varepsilon(\tau))/2 + o_p(1)$. By Assumption 2 and Lemma 5, $\hat\theta - \theta = O_p(n^{-1/2})$. By the continuity of $\dot J(0,z)$, $I_2 = o_p(1)$ for each fixed Δ. Therefore, in view of (2.71)–(2.72), the following quadratic approximation holds for each fixed Δ:

$$S(\Delta) = \tilde S(\Delta) + o_p(1), \quad \tilde S(\Delta) = -\frac{\Delta}{\sqrt{n}}\sum_{i=1}^n[\tau - \mathbf{1}_{\hat\varepsilon_i < Q_\varepsilon(\tau)}] + \frac{f_\varepsilon(Q_\varepsilon(\tau))}{2}\Delta^2. \qquad (2.77)$$

By the quadratic approximation and convexity lemma [Pollard (1991)], the minimizer $\hat\Delta = \sqrt{n}[\hat Q_\varepsilon(\tau) - Q_\varepsilon(\tau)]$ of $S(\Delta)$ has the approximation

$$\hat\Delta = \operatorname*{argmin}_\Delta \tilde S(\Delta) + o_p(1) = \frac{1}{\sqrt{n}f_\varepsilon(Q_\varepsilon(\tau))}\sum_{i=1}^n[\tau - \mathbf{1}_{\hat\varepsilon_i < Q_\varepsilon(\tau)}] + o_p(1). \qquad (2.78)$$

Recall that $\hat\Delta = \sqrt{n}[\hat Q_\varepsilon(\tau) - Q_\varepsilon(\tau)]$. In (2.78), by the uniform approximation for $\sum_{i=1}^n \mathbf{1}_{\hat\varepsilon_i < Q_\varepsilon(\tau)}$ in Theorem 1 and the Bahadur representation for $\hat\theta - \theta$ in Assumption 2, we obtain

$$\hat Q_\varepsilon(\tau) - Q_\varepsilon(\tau) = \frac{1}{n}\sum_{i=1}^n\left\{\frac{\tau - \mathbf{1}_{\varepsilon_i < Q_\varepsilon(\tau)}}{f_\varepsilon(Q_\varepsilon(\tau))} - \mathbb{E}\left[\frac{\dot G_\theta(\theta, Q_\varepsilon(\tau), \mathbf{X}_0)}{\dot G_\varepsilon(\theta, Q_\varepsilon(\tau), \mathbf{X}_0)}\right]^T D(\theta, \varepsilon_i, \mathbf{X}_i)\right\}$$

$$+o_p(n^{-1/2}). \tag{2.79}$$

By Lemma 5, $\hat{Q}_\varepsilon(\tau) = Q_\varepsilon(\tau) + O_p(n^{-1/2})$. Thus, by (2.79) and Assumptions 2 and 4(i),

$$\begin{aligned} & G(\hat{\theta}, \hat{Q}_\varepsilon(\tau), x) - G(\theta, Q_\varepsilon(\tau), x) \\ = \; & \dot{G}_\theta(\theta, Q_\varepsilon(\tau), x)^T(\hat{\theta} - \theta) + \dot{G}_\varepsilon(\theta, Q_\varepsilon(\tau), x)[\hat{Q}_\varepsilon(\tau) - Q_\varepsilon(\tau)] + o_p(n^{-1/2}) \\ = \; & \dot{G}_\varepsilon(x, Q_\varepsilon(\tau), x)\frac{1}{n}\sum_{i=1}^{n} W_i(\tau) + o_p(n^{-1/2}), \end{aligned} \tag{2.80}$$

where $W_i(\tau)$ is defined in (2.16). The result then follows from (2.80) and Lemma 5. $\diamondsuit$

2.7.4 Proof of Theorem 3

Lemma 6. *Recall $\hat{f}_\varepsilon(z)$ in (2.17). Under the conditions in Theorem 3, $\hat{f}_\varepsilon(z) \xrightarrow{p} f_\varepsilon(z)$ uniformly in the neighborhood of $z = Q_\varepsilon(\tau)$.*

Proof. Let $\tilde{f}_\varepsilon(z) = (nh_n)^{-1}\sum_{i=1}^{n} K\{(\varepsilon_i - z)/h_n\}$ be the density estimator based on the true innovations. By the bounded derivative of $K(\cdot)$, there exists a constant c_1 such that

$$|\hat{f}_\varepsilon(z) - \tilde{f}_\varepsilon(z)| \le \frac{1}{nh_n}\sum_{i=1}^{n}\left|K\left(\frac{\hat{\varepsilon}_i - z}{h_n}\right) - K\left(\frac{\varepsilon_i - z}{h_n}\right)\right| \le \frac{c_1}{nh_n^2}\sum_{i=1}^{n}|\hat{\varepsilon}_i - \varepsilon_i|. \tag{2.81}$$

Under Assumption 4(iii) and $\hat{\theta} - \theta = O_p(n^{-1/2})$, we have

$$|\hat{\varepsilon}_i - \varepsilon_i| = |H(\hat{\theta}, Y_i, \mathbf{X}_i) - H(\theta, Y_i, \mathbf{X}_i)| = O_p(n^{-1/2})L(Y_i, \mathbf{X}_i). \tag{2.82}$$

From (2.81)–(2.82), under condition $L(Y_i, \mathbf{X}_i) \in \mathcal{L}^1$ and $nh_n^4 \to \infty$, we have $|\hat{f}_\varepsilon(z) - \tilde{f}_\varepsilon(z)| = o_p(1)$ uniformly on $z \in \mathbb{R}$. By the well-known theory for kernel density estimator (Silverman, 1986), $\tilde{f}_\varepsilon(z) \xrightarrow{p} f_\varepsilon(z)$ uniformly on compact interval, completing the proof. $\diamondsuit$

Lemma 7. *Under the conditions in Theorem 3, we have $n^{-1}\sum_{i=1}^{n}[\tau - \mathbf{1}_{\hat{\varepsilon}_i < \hat{Q}_\varepsilon(\tau)}] \xrightarrow{p} 0$.*

Proof. By Lemma 2,

$$\begin{aligned} |\mathbf{1}_{\hat{\varepsilon}_i - \hat{Q}_\varepsilon(\tau)<0} - \mathbf{1}_{\varepsilon_i - Q_\varepsilon(\tau)<0}| &\leq 2\mathbf{1}_{|(\hat{\varepsilon}_i-\varepsilon_i)-[\hat{Q}_\varepsilon(\tau)-Q_\varepsilon(\tau)]|\geq 2n^{-1/4}} + \mathbf{1}_{|\varepsilon_i - Q_\varepsilon(\tau)|\leq 2n^{-1/4}} \\ &:= J_i. \end{aligned} \tag{2.83}$$

By (2.82), $|\hat{\varepsilon}_i - \varepsilon_i| = o_p(n^{-1/2}\log n)L(Y_i, \mathbf{X}_i)$ uniformly in i. From the proof of Theorem 2, $\hat{Q}_\varepsilon(\tau) - Q_\varepsilon(\tau) = O_p(n^{-1/2}) = o_p(n^{-1/4})$. Hence, with probability tending to one, $|(\hat{\varepsilon}_i - \varepsilon_i) - [\hat{Q}_\varepsilon(\tau) - Q_\varepsilon(\tau)]| \leq (n^{-1/2}\log n)L(Y_i, \mathbf{X}_i) + n^{-1/4}$, and consequently,

$$J_i \leq 2\mathbf{1}_{L(Y_i,\mathbf{X}_i)|\geq n^{1/4}/\log n} + \mathbf{1}_{|\varepsilon_i - Q_\varepsilon(\tau)|\leq 2n^{-1/4}} \quad := \quad \overline{J}_i. \tag{2.84}$$

Note that $\mathbb{E}[\mathbf{1}_{L(Y_i,\mathbf{X}_i)|\geq n^{1/4}/\log n}] \leq \mathbb{E}L(Y_i, \mathbf{x}_i)/(n^{1/4}\log n) \to 0$. Also, by the continuity and boundedness of $f_\varepsilon(z)$ at $Q_\varepsilon(\tau)$, $\mathbb{E}[\mathbf{1}_{|\varepsilon_1 - Q_\varepsilon(\tau)|\leq 2n^{-1/4}}] \to 0$. Thus, $\mathbb{E}(\overline{J}_i) \to 0$. By (2.83)–(2.84),

$$\frac{1}{n}\sum_{i=1}^n [\tau - \mathbf{1}_{\hat{\varepsilon}_i - \hat{Q}_\varepsilon(\tau)<0}] = \frac{1}{n}\sum_{i=1}^n [\tau - \mathbf{1}_{\varepsilon_i - Q_\varepsilon(\tau)<0}] + o_p(1) \xrightarrow{p} 0, \tag{2.85}$$

via the law of large numbers $n^{-1}\sum_{i=1}^n \mathbf{1}_{\varepsilon_i < Q_\varepsilon(\tau)} \xrightarrow{p} \mathbb{E}[\mathbf{1}_{\varepsilon_1 < Q_\varepsilon(\tau)}] = \tau$. ♢

Proof of Theorem 3. First, we prove $\overline{W}(\tau) = o_p(1)$. Since $\hat{\theta} \xrightarrow{p} \theta$ and $\hat{Q}_\varepsilon(\tau) \xrightarrow{p} Q_\varepsilon(\tau)$, by the continuity of $\dot{G}_\theta(\theta, \varepsilon, x)$ and $\dot{G}_\varepsilon(\theta, \varepsilon, x)$ [Assumption 4(i)] and the stochastic continuity condition in Assumption 6, from Lemma 1 we obtain

$$\begin{aligned} & \frac{\dot{G}_\theta(\hat{\theta}, \hat{Q}_\varepsilon(\tau), x)}{\dot{G}_\varepsilon(\hat{\theta}, \hat{Q}_\varepsilon(\tau), x)} - \frac{1}{n}\sum_{i=1}^n \frac{\dot{G}_\theta(\hat{\theta}, \hat{Q}_\varepsilon(\tau), \mathbf{X}_i)}{\dot{G}_\varepsilon(\hat{\theta}, \hat{Q}_\varepsilon(\tau), \mathbf{X}_i)} \\ \xrightarrow{p} \quad & \frac{\dot{G}_\theta(\theta, Q_\varepsilon(\tau), x)}{\dot{G}_\varepsilon(\theta, Q_\varepsilon(\tau), x)} - \mathbb{E}\left[\frac{\dot{G}_\theta(\theta, Q_\varepsilon(\tau), \mathbf{X}_0)}{\dot{G}_\varepsilon(\theta, Q_\varepsilon(\tau), \mathbf{X}_0)}\right], \end{aligned} \tag{2.86}$$

$$\frac{1}{n}\sum_{i=1}^n D(\hat{\theta}, \hat{\varepsilon}_i, \mathbf{X}_i) = \frac{1}{n}\sum_{i=1}^n D(\hat{\theta}, H(\hat{\theta}, Y_i, \mathbf{X}_i), \mathbf{X}_i) \xrightarrow{p} \mathbb{E}[D(\theta, \varepsilon_0, \mathbf{X}_0)] = 0. \tag{2.87}$$

By (2.86)–(2.87) and Lemmas 6–7, $\overline{W}(\tau) = o_p(1)$.

It remains to prove $n^{-1}\sum_{i=1}^n \widehat{W}_i(\tau)^2 \xrightarrow{p} \mathbb{E}[W_1(\tau)^2]$. Recall that $D_i(\vartheta) = D(\vartheta, H(\vartheta, Y_i, \mathbf{X}_i), \mathbf{X}_i)$ in Assumption 6. Then we have $D(\hat{\theta}, \hat{\varepsilon}_i, \mathbf{X}_i) = D_i(\hat{\theta})$ and

$D(\theta, \varepsilon_i, \mathbf{X}_i) = D_i(\theta)$. By Lemma 6 and (2.86), it suffices to prove

$$\frac{1}{n}\sum_{i=1}^{n}[\tau - \mathbf{1}_{\hat{\varepsilon}_i<\hat{Q}_\varepsilon(\tau)}]^2 - \frac{1}{n}\sum_{i=1}^{n}[\tau - \mathbf{1}_{\varepsilon_i<Q_\varepsilon(\tau)}]^2 \xrightarrow{p} 0, \tag{2.88}$$

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\hat{\varepsilon}_i<\hat{Q}_\varepsilon(\tau)}D_i(\hat{\theta}) - \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\varepsilon_i<Q_\varepsilon(\tau)}D_i(\theta) \xrightarrow{p} 0, \tag{2.89}$$

$$\frac{1}{n}\sum_{i=1}^{n}D_i(\hat{\theta})D_i(\hat{\theta})^T - \frac{1}{n}\sum_{i=1}^{n}D_i(\theta)D_i(\theta)^T \xrightarrow{p} 0. \tag{2.90}$$

Note that $|[\tau - \mathbf{1}_{\hat{\varepsilon}_i<\hat{Q}_\varepsilon(\tau)}]^2 - [\tau - \mathbf{1}_{\varepsilon_i<Q_\varepsilon(\tau)}]^2| \le 2|\mathbf{1}_{\hat{\varepsilon}_i<\hat{Q}_\varepsilon(\tau)} - \mathbf{1}_{\varepsilon_i<Q_\varepsilon(\tau)}|$. Thus, by (2.83),

$$\left|\frac{1}{n}\sum_{i=1}^{n}[\tau - \mathbf{1}_{\hat{\varepsilon}_i<\hat{Q}_\varepsilon(\tau)}]^2 - \frac{1}{n}\sum_{i=1}^{n}[\tau - \mathbf{1}_{\varepsilon_i<Q_\varepsilon(\tau)}]^2\right| \le \frac{2}{n}\sum_{i=1}^{n}|\mathbf{1}_{\hat{\varepsilon}_i<\hat{Q}_\varepsilon(\tau)} - \mathbf{1}_{\varepsilon_i<Q_\varepsilon(\tau)}| \xrightarrow{p} 0.$$

This proves (2.88).

To prove (2.89), assume without loss of generality that $D(\theta, \varepsilon_i, \mathbf{X}_i)$ is scalar-valued. By the triangle inequality and (2.83)–(2.84),

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\hat{\varepsilon}_i<\hat{Q}_\varepsilon(\tau)}D_i(\hat{\theta}) - \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\varepsilon_i<Q_\varepsilon(\tau)}D_i(\theta)\right| \\
\le\ & \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\hat{\varepsilon}_i<\hat{Q}_\varepsilon(\tau)}|D_i(\hat{\theta}) - D_i(\theta)| + \frac{1}{n}\sum_{i=1}^{n}|\mathbf{1}_{\hat{\varepsilon}_i<\hat{Q}_\varepsilon(\tau)} - \mathbf{1}_{\varepsilon_i<Q_\varepsilon(\tau)}||D_i(\theta)| \\
\le\ & \frac{1}{n}\sum_{i=1}^{n}|D_i(\hat{\theta}) - D_i(\theta)| + \frac{1}{n}\sum_{i=1}^{n}\overline{J}_i|D_i(\theta)|, \quad \text{with probability tending to one,}
\end{aligned} \tag{2.91}$$

where $\overline{J}_i$ is defined in (2.84). By the stochastic continuity of $D_i(\vartheta)$ at $\vartheta = \theta$, the argument in Lemma 1 shows $n^{-1}\sum_{i=1}^{n}|D_i(\hat{\theta}) - D_i(\theta)| \xrightarrow{p} 0$. Also, by the Cauchy-Schwarz inequality,

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[J_i|D_i(\theta)|] \le \sqrt{\mathbb{E}[D_1(\theta)^2]}\frac{1}{n}\sum_{i=1}^{n}\sqrt{\mathbb{E}(\overline{J}_i^2)} \le \sqrt{\mathbb{E}[D_1(\theta)^2]}\frac{1}{n}\sum_{i=1}^{n}\sqrt{3\mathbb{E}(\overline{J}_i)} \to 0,$$

where the second "$\le$" follows from $\overline{J}_i \le 3$ so that $\overline{J}_i^2 \le 3\overline{J}_i$, and the last convergence "$\to 0$" follows from $\mathbb{E}(\overline{J}_i) \to 0$ (see the proof of Lemma 7). Thus, in view of

(2.91), (2.89) is verified. Finally, (2.90) follows from the stochastic continuity of $D_i(\vartheta)D_i(\vartheta)^T$ at point $\vartheta = \theta$ (Assumption 6) and Lemma 1. This completes the proof. $\diamondsuit$

2.7.5 Proof of Theorem 4

Lemma 8. *Suppose that Assumption 5* holds. Let Δ be any given number. Define*

$$I(z) = \sum_{i=1}^{n} \zeta_i(z), \quad \zeta_i(z) = \int_0^{\frac{\Delta}{\sqrt{n}}} \Big\{ [\mathbf{1}_{\varepsilon_i \le z+s} - \mathbf{1}_{\varepsilon_i \le z}] - [F_\varepsilon(z+s) - F_\varepsilon(z)] \Big\} ds.$$

Then $I(z) = o_p(1)$ uniformly on $z \in [Q_\varepsilon(\delta), Q_\varepsilon(1-\delta)]$.

Proof. We adopt the same chain argument in Theorem 1. Assume without loss of generality that $\Delta > 0$ and $[Q_\varepsilon(\delta), Q_\varepsilon(1-\delta)] = [0,1]$. Consider the evenly spaced grid points $z_j = j/n, j = 0, \ldots, n$, which partition $[0,1]$ into n equal intervals $[z_{j-1}, z_j], j = 1, \ldots, n$. By the boundedness of f_ε, there exists some universal $c_1 > 0$ such that for all $z \in [z_{j-1}, z_j]$,

$$F_\varepsilon(z_j + s) - F_\varepsilon(z_{j-1}) - \frac{c_1}{n} \le F_\varepsilon(z+s) - F_\varepsilon(z) \le F_\varepsilon(z_{j-1}+s) - F_\varepsilon(z_j) + \frac{c_1}{n}. \tag{2.92}$$

Also, observe the following inequalities

$$\mathbf{1}_{\varepsilon_i \le z_{j-1}+s} \;-\; \mathbf{1}_{\varepsilon_i \le z_j} \le \mathbf{1}_{\varepsilon_i \le z+s} - \mathbf{1}_{\varepsilon_i \le z} \le \mathbf{1}_{\varepsilon_i \le z_j+s} - \mathbf{1}_{\varepsilon_i \le z_{j-1}}, \text{ for all } z \in [z_{j-1}, z_j]. \tag{2.93}$$

Combining (2.92)–(2.93), we obtain

$$\underline{\zeta}_{ij} - \frac{\Delta c_1}{n^{3/2}} \;\le\; \zeta_i(z) \;\le\; \overline{\zeta}_{ij} + \frac{\Delta c_1}{n^{3/2}}, \quad \text{for all } z \in [z_{j-1}, z_j], \tag{2.94}$$

where

$$\begin{aligned} \overline{\zeta}_{ij} &= \int_0^{\frac{\Delta}{\sqrt{n}}} \Big\{ [\mathbf{1}_{\varepsilon_i \le z_j+s} - \mathbf{1}_{\varepsilon_i \le z_{j-1}}] - [F_\varepsilon(z_j+s) - F_\varepsilon(z_{j-1})] \Big\} ds, \\ \underline{\zeta}_{ij} &= \int_0^{\frac{\Delta}{\sqrt{n}}} \Big\{ [\mathbf{1}_{\varepsilon_i \le z_{j-1}+s} - \mathbf{1}_{\varepsilon_i \le z_j}] - [F_\varepsilon(z_{j-1}+s) - F_\varepsilon(z_j)] \Big\} ds. \end{aligned}$$

By (2.94), $|I(z)| \le |\sum_{i=1}^n \overline{\zeta}_{ij}| + |\sum_{i=1}^n \underline{\zeta}_{ij}| + \Delta c_1/\sqrt{n}$ for all $z \in [z_{j-1}, z_j]$. Hence,

$$\begin{aligned} \sup_{z\in[0,1]} |I(z)| &= \max_{1\le j\le n} \sup_{z\in[z_{j-1},z_j]} |I(z)| \\ &\le \max_{1\le j\le n} \left|\sum_{i=1}^n \overline{\zeta}_{ij}\right| + \max_{1\le j\le n} \left|\sum_{i=1}^n \underline{\zeta}_{ij}\right| + o(1). \end{aligned} \tag{2.95}$$

Note that $|\overline{\zeta}_{ij}| \le 2\Delta/\sqrt{n}$. Furthermore, by the same argument in (2.76), we can obtain $\mathrm{var}(\sum_{i=1}^n \overline{\zeta}_{ij}) \le c_2 n^{-1/2}$ for some constant c_2 independent of j. Thus, by Berstein's exponential inequality (Bennett, 1962) for the sum of bounded and independent random variables,

$$\begin{aligned} \mathbb{P}\left\{\left|\sum_{i=1}^n \overline{\zeta}_{ij}\right| \ge n^{-1/4}\log n\right\} &\le 2\exp\left\{\frac{-(n^{-1/4}\log n)^2}{2c_2 n^{-1/2} + 4\Delta n^{-1/2}(n^{-1/4}\log n)}\right\} \\ &= O[\exp(-2\log n)] = O(n^{-2}), \end{aligned}$$

for large enough n and all j. Therefore,

$$\mathbb{P}\left\{\max_{1\le j\le n}\left|\sum_{i=1}^n \overline{\zeta}_{ij}\right| \ge n^{-1/4}\log n\right\} \le \sum_{j=1}^n \mathbb{P}\left\{\left|\sum_{i=1}^n \overline{\zeta}_{ij}\right| \ge n^{-1/4}\log n\right\} = O(n^{-1}),$$

which gives $\max_{1\le j\le n} |\sum_{i=1}^n \overline{\zeta}_{ij}| = o_p(1)$. Similarly, $\max_{1\le j\le n} |\sum_{i=1}^n \underline{\zeta}_{ij}| = o_p(1)$. The proof is completed in view of (2.95). ♢

Proof of Theorem 4. First, we show that the asymptotic representation (2.79) holds uniformly on $\tau \in [\delta, 1-\delta]$. It suffices to show that the quadratic approximation (2.77) holds uniformly on $[\delta, 1-\delta]$. Recall I_1 and I_2 in (2.73) and (2.74). In the proof of Theorem 2, it is shown that $I_1 = \Delta^2 f_\varepsilon(Q_\varepsilon(\tau))/2 + o_p(1)$ and $I_2 = o_p(1)$ for fixed quantile τ. Now we shall prove that they also hold uniformly on $[\delta, 1-\delta]$. By Assumption 4(iv), the continuity of $\dot{J}(0, z)$ on $z \in \mathbb{R}$ implies uniform continuity on compact interval. Thus, $\dot{J}(0, Q_\varepsilon(\tau) + s) - \dot{J}(0, Q_\varepsilon(\tau)) = o(1)$ uniformly on $[\delta, 1-\delta]$ and $|s| \le \Delta/\sqrt{n}$. This shows $I_2 = o_p(1)$ uniformly. Similarly, by the uniform continuity of f_ε, (2.75) holds uniformly. Thus, by Lemma 8, we conclude that $I_1 = \Delta^2 f_\varepsilon(Q_\varepsilon(\tau))/2 + o_p(1)$ uniformly on $[\delta, 1-\delta]$.

By the above argument, (2.79) and hence (2.80) hold uniformly on $[\delta, 1-\delta]$. To prove the functional CLT, it suffices to prove the tightness and finite-dimensional

convergence of the leading term $\dot{G}_\varepsilon(x, Q_\varepsilon(\tau), x)n^{-1}\sum_{i=1}^n W_i(\tau)$ in (2.80). The tightness follows from two facts: (i) after normalization, the empirical process $n^{-1}\sum_{i=1}^n \mathbf{1}_{\varepsilon_i \le Q_\varepsilon(\tau)}$ is tight [Chapter 14 in Billingsley (1999)]; and (ii) if $\eta_n \Rightarrow \eta$, then for any continuous function $g(\tau)$ the process $\{g(\tau)\eta_n\}_\tau$ is tight on $[\delta, 1-\delta]$ [Theorem 7.3 in Billingsley (1999)]. Using (2.80), the finite-dimensional convergence follows from the Cramér-Wold device and Lemma 5. ♢

2.7.6 Proof of Theorem 6

Lemma 9 below follows from the same chain argument in the proof of Lemma 8.

Lemma 9. *Assume that f_ε is continuous and bounded on $[-c_1, c_1]$ for some constant $c_1 > 0$. Then for any given constant $c_2 > 0$,*

$$\sup_{|z|\le c_1, |v|\le c_2/\sqrt{n}} \left| \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\varepsilon_i \le z+v} - \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\varepsilon_i \le z} - v f_\varepsilon(z) \right| = o_p(n^{-1/2}).$$

Proof of Theorem 6. Recall $\hat{F}(y|x)$ in (2.39). Write $z = H(\theta, y, x)$ and $\hat{z} = H(\hat{\theta}, y, x)$. Since by Assumption 1, $H(\hat{\theta}, y, x)$ is an increasing function in y, $H(\hat{\theta}, \mathcal{Y}_1, x) \le H(\hat{\theta}, y, x) \le H(\hat{\theta}, \mathcal{Y}_2, x)$ uniformly on $y \in \mathcal{Y}$. Thus, using $\hat{\theta} = \theta + o_p(1)$ and the continuity assumption, $|\hat{z}| = O_p(1)$ uniformly on $y \in \mathcal{Y}$. Note that $\hat{F}(y|x) = \hat{F}_\varepsilon(\hat{z})$. By the uniform approximation of $\hat{F}_\varepsilon(z)$ in Theorem 1 and the equivalent expression for $\dot{J}(0, z)$ in Lemma 4, we have

$$\sup_{y\in\mathcal{Y}} \left| \hat{F}(y|x) - \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\varepsilon_i \le \hat{z}} - \dot{J}(0, \hat{z})^T(\hat{\theta} - \theta) \right| = o_p(n^{-1/2}). \tag{2.96}$$

Furthermore, by $\hat{\theta} - \theta = O_p(n^{-1/2})$ and Assumption 4(iii), we have $\hat{z} - z = O_p(n^{-1/2})L(y, x) = O_p(n^{-1/2})$ uniformly on $y \in \mathcal{Y}$. Therefore, by Lemma 9,

$$\sup_{y\in\mathcal{Y}} \left| \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\varepsilon_i \le \hat{z}} - \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\varepsilon_i \le z} - (\hat{z} - z) f_\varepsilon(z) \right| = o_p(n^{-1/2}). \tag{2.97}$$

The continuity of $\dot{J}(0, z)$ in Assumption 4(iv) implies the uniform continuity of $\dot{J}(0, z)$ on compact intervals. Thus, $\dot{J}(0, \hat{z}) - \dot{J}(0, z) = o_p(1)$ uniformly on $y \in \mathcal{Y}$. Combining the latter with (2.96)–(2.97) and $\mathbb{E}(\mathbf{1}_{\varepsilon_i \le z}) = F(y|x)$ [see (2.38)], we

obtain the uniform approximation

$$\begin{aligned}\hat{F}(y|x) - F(y|x) &= \frac{1}{n}\sum_{i=1}^{n}[\mathbf{1}_{\varepsilon_i \le z} - \mathbb{E}(\mathbf{1}_{\varepsilon_i \le z})] \\ &+ \dot{J}(0,z)^T(\hat{\theta}-\theta) + (\hat{z}-z)f_\varepsilon(z) + o_p(n^{-1/2}).\end{aligned}$$

Hence, by Taylor's expansion $\hat{z} - z = \dot{H}(\theta, y, x)^T(\hat{\theta} - \theta) + o_p(n^{-1/2})$, the Bahadur representation for $\hat{\theta} - \theta$ in Assumption 2, and the equivalent expression for $\dot{J}(0, z)$ in (2.56) (see Lemma 4), we can further obtain the uniform approximation on $y \in \mathcal{Y}$

$$\hat{F}(y|x) - F(y|x) = \frac{1}{n}\sum_{i=1}^{n} V_i(y) + o_p(n^{-1/2}), \tag{2.98}$$

where $V_i(y)$ is defined in (2.42). From (2.98), the tightness follows from the tightness of the empirical process of $\{\varepsilon_i\}$ and the differentiability of $H(\theta, y, x)$ in y, and the finite-dimensional convergence follows from the Cramér-Wold device and Lemma 5. $\diamondsuit$

Chapter 3
Conditional Expected Shortfall

3.1 Main Results

Assume that we have stationary observations $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$, where $\mathbf{X}_i \in \mathbb{R}^p$ and $Y_i \in \mathbb{R}$ are, respectively, the covariates and stock loss at time i. Our goal is to estimate and draw inference for $\mathrm{CES}(y|x)$ in (1.3). In Sections 3.1.1 and 3.1.2 below, we consider nonparametric approach and semiparametric approach, respectively.

3.1.1 Nonparametric Estimation

Denote by $\mathbf{1}_A$ the usual indicator function for A. From (1.3), we can rewrite

$$\mathrm{CES}(y|x) = \frac{\mathbb{E}(Y_i \mathbf{1}_{Y_i \geq y} | \mathbf{X}_i = x)}{\mathbb{E}(\mathbf{1}_{Y_i \geq y} | \mathbf{X}_i = x)}. \tag{3.1}$$

Note that both the numerator and denominator are of the form $\mathbb{E}[g(Y_i)|\mathbf{X}_i = x]$ for some function g. Thus, we propose the following nonparametric kernel smoothing estimate of $\mathrm{CES}(y|x)$:

$$\widetilde{\mathrm{CES}}(y|x) = \frac{\sum_{i=1}^n Y_i \mathbf{1}_{Y_i \geq y} K_{b_n}(\mathbf{X}_i - x)}{\sum_{i=1}^n \mathbf{1}_{Y_i \geq y} K_{b_n}(\mathbf{X}_i - x)}, \tag{3.2}$$

where $K_{b_n}(u) = K(u/b_n)$ for a p-variate kernel function $K(\cdot)$ and a bandwidth $b_n > 0$. By Li and Racine (2007), under appropriate regularity and mixing conditions, $\widetilde{\mathrm{CES}}(y|x)$ has asymptotic normality with convergence rate $\sqrt{nb_n^p}$. Formally, we impose the regularity and mixing condition below.

Assumption 7. *(i)* $\{(\mathbf{X}_i, Y_i)\}_{i\in\mathbb{Z}}$ *is stationary and* α*-mixing with mixing coefficient* $\alpha_j \le C_1\alpha^j, j \ge 0$, *for some constants* $0 < C_1 < \infty$ *and* $\alpha \in (0,1)$. *(ii)* $\{\varepsilon_i\}_{i\in\mathbb{Z}}$ *are i.i.d., and for each* i, ε_i *is independent of* $\{X_j\}_{j\le i}$.

3.1.2 Semiparametric Estimation

We consider the nonlinear heteroscedastic model:

$$Y_i = \mu(\theta, \mathbf{X}_i) + \sigma(\theta, \mathbf{X}_i)\varepsilon_i, \tag{3.3}$$

for parametric functions $\mu(\theta,\cdot)$ and $\sigma(\theta,\cdot) > 0$ with some unknown column vector of parameter $\theta \in \mathbb{R}^k$, and $\{\varepsilon_i\}$ are independent and identically distributed (i.i.d.) errors.

Example 4. Let $\mathbf{X}_i = (Y_{i-1}, \ldots, Y_{i-p})$, $p \in \mathbb{N}$, be the lagged returns, then (3.3) becomes the nonlinear autoregressive conditional heteroscedastic (NARCH) model $Y_i = \mu(\theta, Y_{i-1}, \ldots, Y_{i-p}) + \sigma(\theta, Y_{i-1}, \ldots, Y_{i-p})\varepsilon_i$.

Example 5. Stock returns Y_i may depend on the lagged returns $(Y_{i-1}, \ldots, Y_{i-p})$ as well as some overall exogenous economic variables $\mathbf{U}_i \in \mathbb{R}^q$, such as the inflation rates and unemployment rates. To incorporate such exogenous variables, let $\mathbf{X}_i = (Y_{i-1}, \ldots, Y_{i-p}, \mathbf{U}_i)$, then (3.3) becomes the NARCH model with exogenous variables.

Example 6. Consider the continuous-time diffusion model $dY_t = \mu(\theta, Y_t)dt + \sigma(\theta, Y_t)d\mathbb{B}_t$, where $\{\mathbb{B}_t\}_{t\ge 0}$ is a Brownian motion or a general Lévy process. With different specifications of μ and σ, the latter model includes many popular models; see Zhao (2008). Let $\Delta > 0$ be the sampling interval. Then the Euler-discretization

$$Y_{i\Delta} - Y_{(i-1)\Delta} = \mu(\theta, Y_{i\Delta})\Delta + \sigma(\theta, Y_{i\Delta})[\mathbb{B}_{i\Delta} - \mathbb{B}_{(i-1)\Delta}]$$

is of the form (3.3) with $\varepsilon_i = \mathbb{B}_{i\Delta} - \mathbb{B}_{(i-1)\Delta}$.

As elaborated below, this NARCHX model possesses several appealing features.

First, it allows exogenous/external variables U_i to affect the main time series $\{Y_i\}$, thus providing a more flexible modeling framework. If we drop U_i from (3.3), the model reduces to the nonlinear autoregressive conditional heteroscedastic

(NARCH) model $Y_i = \mu(\theta, Y_{i-1}, \ldots, Y_{i-p}) + \sigma(\theta, Y_{i-1}, \ldots, Y_{i-p})\varepsilon_i$, which includes many popular nonlinear time series models, such as the linear AR model, threshold AR models [Tong (1990)], exponential AR models [Haggan and Ozaki (1981)], and Engle's ARCH models, among others; see Fan and Yao (2003). On the other hand, the general formulation (3.3) allows user-defined models. For example, (3.3) includes the ARCH model with exogenous (ARCHX) inputs:

$$Y_i = \sum_{j=1}^{p} \phi_j Y_{i-j} + U_i\beta + \varepsilon_i \left(\alpha_0^2 + \sum_{j=1}^{p} \alpha_j^2 Y_{i-j}^2 + U_i^2\gamma^2 \right)^{1/2}, \quad \phi_j, \alpha_j \in \mathbb{R}, \beta, \gamma \in \mathbb{R}^q.$$

This model generalizes Engle's ARCH model to allow for exogenous variables.

Second, the distribution of ε_i is completely unspecified, and thus this semiparametric approach can avoid potential mis-specification on the distribution of ε_i. For example, if $\varepsilon_i \sim N(0,1)$ has the standard normal distribution, then $\mathbb{E}(\varepsilon_i|\varepsilon_i > 2) = 2.37$; if $\varepsilon_i \sim t_3/\sqrt{3}$, i.e., the normalized student-t distribution with 3 degrees of freedom (the normalizer $\sqrt{3}$ makes the variance one), then $\mathbb{E}(\varepsilon_i|\varepsilon_i > 2) = 2.99$, representing a 26% increase from that of the $N(0,1)$ specification. Therefore, it is desirable to construct a distribution-free estimator of the conditional expected shortfall.

Third, using parametric specifications $\mu(\theta, \cdot)$ and $\sigma(\theta, \cdot)$ can avoid the "curse of dimensionality". For nonparametric models, it is practically infeasible to consider three or higher dimensional covaraites, and the convergence rate decreases with dimension. By contrast, as shown below, our semiparametric approach has the parametric $\sqrt{n}$ convergence rate, regardless of the dimensionality of the covariates $\mathbf{X}_i$. This is particularly useful as we can incorporate more variables in $\mathbf{X}_i$ for predicting the CES of Y_i.

In (3.3), we assume that the innovation ε_i is independent of $\mathbf{X}_i$. Given $\mathbf{X}_i = x$, we have: (i) $Y_i = \mu(\theta, x) + \sigma(\theta, x)\varepsilon_i$ is independent of $\mathbf{X}_i$; and (ii) $Y_i \geq y$ is equivalent to $\varepsilon_i \geq [y - \mu(\theta, x)]/\sigma(\theta, x)$. Thus, by the definition in (1.3), we can easily obtain

$$\begin{aligned} \mathrm{CES}(y|x) &= \mu(\theta, x) + \sigma(\theta, x)\mathbb{E}[\varepsilon_i|\varepsilon_i \geq \ell(\theta)] \\ &= \mu(\theta, x) + \sigma(\theta, x)\frac{\mathbb{E}[\varepsilon_i \mathbf{1}_{\varepsilon_i \geq \ell(\theta)}]}{\mathbb{E}[\mathbf{1}_{\varepsilon_i \geq \ell(\theta)}]} \text{ with } \quad \ell(\theta) = \frac{y - \mu(\theta, x)}{\sigma(\theta, x)}. \quad (3.4) \end{aligned}$$

Here and hereafter $\mathbf{1}$ stands for the indicator function. If the parameter θ and the innovations ε_i were known, then we can replace the expectation and probability in (3.1) by their empirical version, leading to the estimate of $\mathrm{CES}(y|x)$:

$$\overline{\mathrm{CES}}(y|x) = \mu(\theta, x) + \sigma(\theta, x)\frac{\sum_{i=1}^n \varepsilon_i \mathbf{1}_{\varepsilon_i \geq \ell(\theta)}}{\sum_{i=1}^n \mathbf{1}_{\varepsilon_i \geq \ell(\theta)}}. \tag{3.5}$$

Assume $\varepsilon_i \in \mathcal{L}^2$. Since ε_i are iid, by the delta-method, the asymptotic normality holds

$$\sqrt{n}\Big[\overline{\mathrm{CES}}(y|x) - \mathrm{CES}(y|x)\Big] \Rightarrow N\Big(0, \mathrm{var}(\eta_0)\Big), \tag{3.6}$$

where

$$\eta_i = \frac{\sigma(\theta, x)}{\mathbb{P}\{\varepsilon_i \geq \ell(\theta)\}^2}\Big[\mathbb{P}\{\varepsilon_i \geq \ell(\theta)\}\varepsilon_i \mathbf{1}_{\varepsilon_i \geq \ell(\theta)} - \mathbb{E}[\varepsilon_i \mathbf{1}_{\varepsilon_i \geq \ell(\theta)}]\mathbf{1}_{\varepsilon_i \geq \ell(\theta)}\Big]. \tag{3.7}$$

In practice, since the parameter θ is unknown and the innovations $\{\varepsilon_i\}$ are not observable, $\overline{\mathrm{CES}}(y|x)$ is an infeasible estimator. Nevertheless, this infeasible estimator serves as a standard against which we can measure other estimators.

Motivated by (3.5), we propose replacing θ and ε_i in (3.5) by their consistent estimates. Specifically, we adopt the following procedure:

(i) Let $\hat{\theta}$ be a consistent estimate of θ. We can estimate the innovation ε_i by

$$\hat{\varepsilon}_i = \frac{Y_i - \mu(\hat{\theta}, \mathbf{X}_i)}{\sigma(\hat{\theta}, \mathbf{X}_i)}. \tag{3.8}$$

(ii) In view of (3.5), we propose estimating $\mathrm{CES}(y|x)$ by

$$\widehat{\mathrm{CES}}(y|x) = \mu(\hat{\theta}, x) + \sigma(\hat{\theta}, x)\frac{\sum_{i=1}^n \hat{\varepsilon}_i \mathbf{1}_{\hat{\varepsilon}_i \geq \ell(\hat{\theta})}}{\sum_{i=1}^n \mathbf{1}_{\hat{\varepsilon}_i \geq \ell(\hat{\theta})}}. \tag{3.9}$$

In practice, if we choose y to be the τ-th quantile of Y_i given $X_i = x$, (3.9) becomes

$$\widehat{\mathrm{CES}}(y|x) = \mu(\hat{\theta}, x) + \sigma(\hat{\theta}, x)\frac{\sum_{i=1}^n \hat{\varepsilon}_i \mathbf{1}_{\hat{\varepsilon}_i \geq \ell(\hat{\theta})}}{n(1-\tau)}. \tag{3.10}$$

The estimator $\widehat{\text{CES}}(y|x)$ has the nice feature that it does not depend on the underlying distribution of innovations ε_i. If $\hat{\theta} = \theta$ and $\hat{\varepsilon}_i = \varepsilon_i$, then by the law of large numbers and Slutsky's Theorem,

$$\widehat{\text{CES}}(y|x) \xrightarrow{p} \mu(\theta, x) + \sigma(\theta, x)\mathbb{E}[\varepsilon_i \mathbf{1}_{\varepsilon_i \geq \ell(\theta)}]/\mathbb{P}\{\varepsilon_i \geq \ell(\theta)\} = \text{CES}(y|x).$$

In general, when $\hat{\theta}$ and $\hat{\varepsilon}_i$ are subject to estimation errors, whether and how fast $\widehat{\text{CES}}(y|x)$ converges to $\text{CES}(y|x)$ depend on the accuracy of the estimates $\hat{\theta}$ and $\hat{\varepsilon}_i$. Theorem 7 below shows that the proposed estimator $\widehat{\text{CES}}(y|x)$ can achieve the parametric $\sqrt{n}$ convergence rate.

Theorem 7. *Suppose Assumptions 7 and 9-10 (in Section 3.3) hold. Further assume* $\hat{\theta} - \theta = O_p(n^{-1/2})$. *Then*

$$\widehat{\text{CES}}(y|x) = \text{CES}(y|x) + O_p(n^{-1/2}).$$

3.1.3 Asymptotic Normality

To implement the estimator $\widehat{\text{CES}}(y|x)$, it is necessary to construct a consistent estimate $\hat{\theta}$ of θ. Using $\mathbf{X}_i$ in (1.3), we can rewrite model (3.3) as $Y_i = \mu(\theta, \mathbf{X}_i) + \sigma(\theta, \mathbf{X}_i)\varepsilon_i$. Assume without loss of generality that the innovations ε_i are standardized so that $\mathbb{E}(\varepsilon_i) = 0$ and $\text{var}(\varepsilon_i) = 1$. Consider the pseudo-likelihood estimate, recall (2.22):

$$\hat{\theta} = \underset{\theta}{\text{argmin}} \sum_{i=1}^{n} \left\{ \left[\frac{Y_i - \mu(\theta, \mathbf{X}_i)}{\sigma(\theta, \mathbf{X}_i)} \right]^2 + 2 \log \sigma(\theta, \mathbf{X}_i) \right\}.$$

Due to the nonlinearity structure, this estimator $\hat{\theta}$ generally does not have a closed form. Under appropriate regularity conditions, Theorem 2 in Zhao (2010) established the following asymptotic Bahadur representation which was discussed in (2.23) and Assumption 2:

$$\hat{\theta} - \theta = \frac{1}{n} \sum_{i=1}^{n} \zeta_i + o_p(n^{-1/2}) \text{ with } \zeta_i = \mathcal{I}(\theta)^{-1} \left[\frac{\varepsilon_i \dot{\mu}(\theta, \mathbf{X}_i)}{\sigma(\theta, \mathbf{X}_i)} + (\varepsilon_i^2 - 1) \frac{\dot{\sigma}(\theta, \mathbf{X}_i)}{\sigma(\theta, \mathbf{X}_i)} \right], \tag{3.11}$$

where $\dot{\mu}(\theta, \mathbf{X}_i)$ and $\dot{\sigma}(\theta, \mathbf{X}_i)$ are the partial derivatives with respect to θ, and

$$\mathcal{I}(\theta) = \mathbb{E}\Big[\frac{\dot{\mu}(\theta, \mathbf{X}_0)\dot{\mu}(\theta, \mathbf{X}_0)^T + 2\dot{\sigma}(\theta, \mathbf{X}_0)\dot{\sigma}(\theta, \mathbf{X}_0)^T}{\sigma^2(\theta, \mathbf{X}_0)}\Big]. \tag{3.12}$$

Assumption 8. The estimator $\hat{\theta} \in \mathbb{R}^k$ of $\theta \in \mathbb{R}^k$ admits the Bahadur-type representation

$$\hat{\theta} - \theta = \frac{1}{n}\sum_{i=1}^{n} \zeta(\theta, \varepsilon_i, \mathbf{X}_i) + o_{\mathbb{P}}(n^{-1/2}), \tag{3.13}$$

for some $\zeta(\theta, \cdot, \cdot) \in \mathbb{R}^k$ satisfying $\zeta(\theta, \varepsilon_i, \mathbf{X}_i) \in \mathcal{L}^2$ and $\mathbb{E}[\zeta(\theta, \varepsilon_i, \mathbf{X}_i)|\mathbf{X}_i] = 0$.

In practice, it is often desirable to derive the asymptotic distribution of (3.10) in order to make statistical inference, such as confidence interval construction and hypothesis testing. Since the accuracy of $\widehat{\mathrm{CES}}(y|x)$ depends on $\hat{\theta}$, it is necessary to study the effect of $\hat{\theta}$. Recall $\hat{\varepsilon}_i$ in (3.8). To reflect the dependence of $\hat{\varepsilon}_i$ on the accuracy of $\hat{\theta}$, define

$$\varepsilon_i(\delta) = \frac{Y_i - \mu(\theta + \delta, \mathbf{X}_i)}{\sigma(\theta + \delta, \mathbf{X}_i)}, \quad \delta \in \mathbb{R}^k. \tag{3.14}$$

Define

$$J_1(\delta, z) = \mathbb{E}[\varepsilon_i(\delta)\mathbf{1}_{\varepsilon_i(\delta)\geq z}], \text{ and } J_2(\delta, z) = \mathbb{E}\mathbf{1}_{\varepsilon_i(\delta)\geq z}. \tag{3.15}$$

Here δ measures the departure of the estimator $\hat{\theta}$ from the true parameter θ. Clearly, $\varepsilon_i(\hat{\theta} - \theta) = \hat{\varepsilon}_i$ and $\varepsilon_i(0) = \varepsilon_i$. Intuitively, $J_1(\delta, z)$ and $J_2(\delta, z)$ measures how the tail expectation and tail probability of $\varepsilon_i(\delta)$ changes in response to the departure δ from the true θ.

Theorem 8. *Suppose Assumptions 7-8 and 9-10 (in Section 3.3) hold. Further assume that (3.11) holds. Denote by $\dot{J}(\delta)$ the gradient vector of $J(\delta)$. Then the CLT holds*

$$\sqrt{n}[\widehat{\mathrm{CES}}(y|x) - \mathrm{CES}(y|x)] \Rightarrow N\Big(0, \mathrm{var}(\eta_0 + H^T\zeta_0)\Big), \tag{3.16}$$

where η_i is defined in (3.7), ζ_i is defined as in (3.11), and

$$H = \dot{\mu}(\theta,x) + \dot{\sigma}(\theta,x)\frac{\mathbb{E}[\varepsilon_0 \mathbf{1}_{\varepsilon_0 \geq \ell(\theta)}]}{\mathbb{E}\mathbf{1}_{\varepsilon_0 \geq \ell(\theta)}} + \frac{\sigma(\theta,x)}{\mathbb{P}\{\varepsilon_0 \geq \ell(\theta)\}^2}\Big\{\mathbb{P}\{\varepsilon_0 \geq \ell(\theta)\}[\dot{J}_1(0,\ell(\theta))^T$$
$$-\dot{\ell}(\theta)\ell(\theta)f_\varepsilon(\ell(\theta))] - \mathbb{E}[\varepsilon_0 \mathbf{1}_{\varepsilon_0 \geq \ell(\theta)}][\dot{J}_2(0,\ell(\theta))^T - \dot{\ell}(\theta)f_\varepsilon(\ell(\theta))]\Big\} \quad (3.17)$$

For the limiting variance $\mathrm{var}(\eta_0 + H^T\zeta_0)$ in (3.16), the two terms represent two sources of randomness: η_0 is the same as in (3.6) and represents the randomness of the estimator when θ and ε_i are known; $H^T\zeta_0$ represents the effect from the randomness of the estimator $\hat{\theta}$ [see the Bahadur representation (3.11)], and the constant factor H^T is analogous to the derivative in the first-order Taylor's expansion of $\widehat{\mathrm{CES}}(y|X)$ at $\hat{\theta} \approx \theta$.

3.2 Monte Carlo Studies

3.2.1 MISE comparison with nonparametric method

In this section, we compare MISE performance between nonparametric CES estimator and our proposed semiparametric CES estimator. Recall the definitions of MISE and RMISE in (2.44)-(2.46), denote $\mathrm{MISE}\{\widehat{\mathrm{CES}}(y|\cdot)\}$ by the MISE of the proposed semiparametric estimator $\widehat{\mathrm{CES}}(y|x)$ in (3.10), its RMISE, relative to the nonparametric estimator $\widetilde{\mathrm{CES}}(y|x)$ under the best-case scenario, is

$$\mathrm{RMISE} = \frac{\mathrm{MISE}\{\widetilde{\mathrm{CES}}(y|\cdot)\}}{\mathrm{MISE}\{\widehat{\mathrm{CES}}(y|\cdot)\}} = \frac{\min_{b_n}\mathrm{MISE}\{\widetilde{\mathrm{CES}}(y|\cdot);b_n\}}{\mathrm{MISE}\{\widehat{\mathrm{CES}}(y|\cdot)\}}. \quad (3.18)$$

As remarked in (2.44)-(2.46), a value of RMISE ≥ 1 indicates better MISE performance of the proposed method.

We consider the four ARCH-type models, Model 1 - Model 4, as defined in Section 2.5.1. In all of the four models, the thresholds y are taken to be τ-th quantile of Y_i given $Y_{i-1} = x$, where $\tau = 2.5\%, 5\%, 10\%, 20\% \ldots, 80\%, 90\%, 95\%, 97.5\%$. The noise ε_i is from two distributions: (i) standard normal $N(0,1)$, and (ii) $t_3/\sqrt{3}$ (Student-t distribution with 3 degrees of freedom with the normalizer $\sqrt{3}$ making the variance one). In all settings we use sample size $n = 200$.

For Model 1 and 3, we estimate the conditional expected shortfall of Y_i given $Y_i \geq y$ and $Y_{i-1} = x$ where $\mathcal{X}$ in (2.44) is taken to be the range of 2.5-th and 97.5-th percentiles of $\{Y_{i-1}\}$; for Model 2 (resp. Model 4), we estimate the conditional τ-th quantile of Y_i given the bivariate $\mathbf{X}_i := (Y_{i-1}, Y_{i-2}) = (x_1, x_2)$ (resp. $\mathbf{X}_i = (Y_{i-1}, U_i)$ for Model 4), and we take $\mathcal{X}$ in (2.44) to be $\mathcal{X}_1 \times \mathcal{X}_2$, where $\mathcal{X}_1$ and $\mathcal{X}_2$ are, respectively, the range of 2.5-th and 97.5-th percentiles of each of the two coordinates of $\mathbf{X}_i$. The integral in (2.44) is approximated by 20 evenly spaced grid points in the univariate case (Model 1 and 3) or 10×10 evenly spaced grid points in the bivariate case (Model 2 and 4). To implement the semiparametric method, we use (2.22) to estimate the unknown parameters.

Table 8: RMISE [see (3.18)] of the proposed semiparametric estimate of CES$(y|x)$ relative to the nonparametric method in (3.1) with theoretical optimal bandwidth, at different quantiles τ. Numbers ≥ 1 indicate better performance of the proposed method.

		Quantile τ in $\mathbb{P}\{Y_i \geq y\} = 1 - \tau$												
	noise	2.5%	5%	10%	20%	30%	40%	50%	60%	70%	80%	90%	95%	97.5%
Model 1	$N(0,1)$	2.52	2.61	3.01	4.05	5.49	8.00	10.60	13.69	16.21	16.69	12.74	8.74	5.46
	$t_3/\sqrt{3}$	3.94	3.84	4.03	5.11	7.24	10.73	12.80	14.76	14.31	12.72	9.55	4.82	2.08
Model 2	$N(0,1)$	8.56	8.78	9.51	12.55	17.04	22.73	28.97	34.98	39.32	40.73	34.32	25.71	17.65
	$t_3/\sqrt{3}$	12.55	12.74	13.91	18.54	25.04	33.60	39.95	46.50	46.70	41.46	26.28	14.68	6.99
Model 3	$N(0,1)$	5.65	5.65	5.53	5.68	6.03	6.91	9.02	13.02	19.20	26.34	27.35	19.27	11.78
	$t_3/\sqrt{3}$	6.38	6.06	5.52	5.93	6.84	7.82	10.39	12.24	15.17	15.15	10.32	6.19	3.66
Model 4	$N(0,1)$	5.45	5.64	5.79	6.27	7.09	7.98	8.91	9.97	11.28	12.84	14.16	12.32	9.03
	$t_3/\sqrt{3}$	8.68	8.47	8.68	8.71	9.38	10.13	11.14	10.72	9.71	7.97	5.81	3.52	1.89

Table 8 summarizes the RMISE [see (3.18)]. The results show that, for almost all cases considered, a substantial MISE improvement can be achieved by using the semiparametric CES estimator.

Comparing with RMISE performance of CVaR in Table 1, nonparametric CES estimates no longer have the advantages in middle-range quantiles $\tau = 20\%, \ldots, 80\%$. The semiparametric estimator significantly outperforms the nonparametric estimator for both $N(0,1)$ and Student-t noise, whereas the two methods have comparable performance for Student-t distributed noise at extreme quantiles $\tau = 90\%, 95\%$, which correspond to the widely used confidence levels $1 - \tau = 10\%, 5\%$ in the ES literature.

3.2.2 Bootstrap confidence intervals

The finite sample performance of CES confidence intervals has not been examined in the literature; in this section we evaluate the performance of the bootstrap confidence interval for CES: the bootstrap confidence interval is constructed using the procedure in Section 2.3. Formally, denote by $r_{1-\alpha}$ the $(1-\alpha)$ sample quantile of

$$\sqrt{n}|\widehat{\mathrm{CES}}^{*(1)}(y|x) - \widehat{\mathrm{CES}}(y|x)|, \ldots, \sqrt{n}|\widehat{\mathrm{CES}}^{*(M)}(y|x) - \widehat{\mathrm{CES}}(y|x)|,$$

with $M = 1000$ bootstrap replications. Then the $(1-\alpha)$ bootstrap confidence interval for $\mathrm{CES}(y|x)$ is

$$\widehat{\mathrm{CES}}(y|x) \pm r_{1-\alpha}/\sqrt{n}.$$

The empirical coverage probability is the proportion of confidence intervals among 1000 realizations of $(1-\alpha)$ confidence intervals that cover the true $\mathrm{CES}(y|x)$. Table 9 presents the results for the most typical setting $1-\tau = 5\%$ and $1-\alpha = 90\%, 95\%, 99\%$, at different values of x. Overall, the bootstrap confidence interval delivers nice performance and has empirical coverage probabilities close to the nominal levels. Given the fact that the limiting variance of asymptotic normality in (3.6) and (3.7) is difficult to estimate and computationally expensive, we recommend the bootstrap confidence intervals that can be easily implemented and possess superior performance over asymptotic confidence intervals in practice.

Table 9: Empirical coverage probability of bootstrap confidence intervals (CI) for CES($y|x$).

(Model 1) CES($y|x$): x at different percentiles of covariates $\{Y_{i-1}\}$

		x at different percentiles											
noise	$1-\alpha$	2.5th	5th	10th	20th	30th	40th	50th	60th	70th	80th	90th	95th	99th
$N(0,1)$	90%	85.7	84.3	89.0	86.3	84.7	87.0	87.0	87.3	88.0	87.0	87.7	89.0	88.0
	95%	94.0	92.0	93.7	93.3	92.0	92.0	92.7	94.0	94.0	93.3	93.7	92.7	93.0
	99%	98.7	98.3	98.3	98.7	98.0	99.0	99.0	97.3	98.3	98.7	98.3	98.7	98.7
$t_3/\sqrt{3}$	90%	85.3	85.3	86.3	85.7	87.0	87.3	87.3	87.3	86.7	85.7	87.7	87.7	87.0
	95%	92.0	91.7	91.3	90.3	92.3	91.7	91.7	91.7	91.3	91.3	92.3	93.3	93.7
	99%	97.3	97.7	97.0	97.0	96.3	95.3	95.3	94.7	95.3	94.7	96.0	97.7	98.7

(Model 2) CES($y|(x_1,x_2)$): x_1 at different percentiles of $\{Y_{i-1}\}$, x_2 at median of $\{Y_{i-2}\}$

		x_1 at different percentiles, x_2 at median											
noise	$1-\alpha$	2.5th	5th	10th	20th	30th	40th	50th	60th	70th	80th	90th	95th	99th
$N(0,1)$	90%	87.0	80.3	70.3	78.7	77.0	83.3	85.0	86.3	89.7	90.0	89.7	92.0	92.0
	95%	92.7	87.3	80.3	82.7	85.0	87.3	88.7	92.3	93.7	93.3	94.0	94.3	94.7
	99%	98.0	94.7	91.7	90.0	93.7	94.0	95.7	96.3	97.3	99.0	98.3	99.0	98.3
$t_3/\sqrt{3}$	90%	77.7	76.3	79.0	79.0	81.7	83.3	81.3	82.7	83.3	84.3	87.0	87.0	89.0
	95%	83.7	82.7	86.0	83.0	85.0	85.7	87.3	88.0	89.3	88.7	88.7	91.0	92.0
	99%	89.3	92.3	92.7	94.0	94.3	93.0	94.0	93.7	94.7	95.3	96.7	96.3	97.0

(Model 3) CES($y|x$): x at different percentiles of $\{Y_{i-1}\}$

		x at different percentiles											
noise	$1-\alpha$	2.5th	5th	10th	20th	30th	40th	50th	60th	70th	80th	90th	95th	99th
$N(0,1)$	90%	90.3	90.7	91.0	92.7	85.0	67.0	81.7	80.7	86.0	87.7	87.7	89.3	90.0
	95%	94.7	94.3	94.7	96.3	91.3	77.7	86.0	89.7	91.3	91.7	93.0	94.3	96.0
	99%	99.3	99.7	99.3	100.0	98.3	90.7	89.7	96.7	97.0	96.7	98.3	99.0	99.7
$t_3/\sqrt{3}$	90%	76.7	72.0	78.0	76.0	80.7	84.3	85.0	88.3	88.7	87.3	85.3	83.7	83.3
	95%	83.3	80.7	84.0	87.7	88.7	88.0	89.7	92.7	93.0	94.0	89.3	89.3	88.7
	99%	90.0	89.3	90.3	93.0	94.0	95.7	96.7	95.3	96.7	97.0	96.0	95.7	96.3

(Model 4) CES($y|(x_1,x_2)$): x_1 at different percentiles $\{Y_{i-1}\}$, x_2 at median of $\{U_i\}$

		x_1 at different percentiles, x_2 at median											
noise	$1-\alpha$	2.5th	5th	10th	20th	30th	40th	50th	60th	70th	80th	90th	95th	99th
$N(0,1)$	90%	84.7	84.3	78.0	77.0	76.0	79.7	83.0	85.0	87.3	91.3	86.3	87.3	89.7
	95%	91.3	90.3	87.0	87.7	86.0	85.0	87.7	89.0	93.3	93.7	92.3	93.0	93.3
	99%	98.7	97.7	96.0	95.3	94.0	94.7	94.7	95.7	96.3	97.0	96.7	95.7	96.3
$t_3/\sqrt{3}$	90%	79.7	80.0	80.3	78.3	75.7	78.7	80.7	84.3	87.7	86.7	82.3	76.7	73.7
	95%	86.3	87.7	86.0	84.7	85.3	85.3	87.7	87.3	91.0	92.0	90.7	86.0	84.3
	99%	96.3	96.0	93.3	92.0	93.7	93.7	94.3	95.0	95.3	97.3	96.0	95.0	95.7

3.3 An Empirical Application to S&P 500 Index

We use the same S&P 500 index data as in Section 2.6. In Figure 2.1, there are a few peaks during the middle of year 2011 when the US credit rating was downgraded from AAA to AA+ and European sovereign debt crisis has been taken place in the European Union. To visualize the robustness of our proposed semiparametric CES estimation, we focus our analysis on the daily losses during this period 2010–2013. Moreover, we compare the performance of our semiparametric method and some existing parametric methods. In the following sections, we use our proposed semiparametric CVaR estimates from Section 2.6 as the thresholds of loss to predict CES and its bootstrap confidence intervals semiparametrically.

3.3.1 Comparison under different GARCH models

In Section 2.6.1, we sequentially predicted CVaR using three GARCH models: standard GARCH, EGARCH, and GJR-GARCH, as described in Model 5–7 in Section 2.5.2. To predict CES sequentially, we use a similar approach. For a given time i, based on the historical data $\mathbf{X}_i = \{Y_j\}_{j \le i-1}$, we apply our proposed semiparametric method to obtain the estimate $\widehat{\mathrm{CVaR}}(1-\tau|Y_j, j \le i-1)$ for the "unobservable" loss Y_i. We then use this semiparametric CVaR estimate as threshold loss y and predict $\widehat{\mathrm{CES}}(y|Y_j, j \le i-1)$ Repeating the procedure for $i = n-(J-1), n-(J-2), \ldots, n$, we obtain the sequentially predicted CES for the last $J = 1000$ daily losses, which is able to capture the pattern of daily losses and not overestimate the daily losses during the last four years 2010–2013. Since GARCH models are non-Markovian, it is infeasible to use their nonparametric approach.

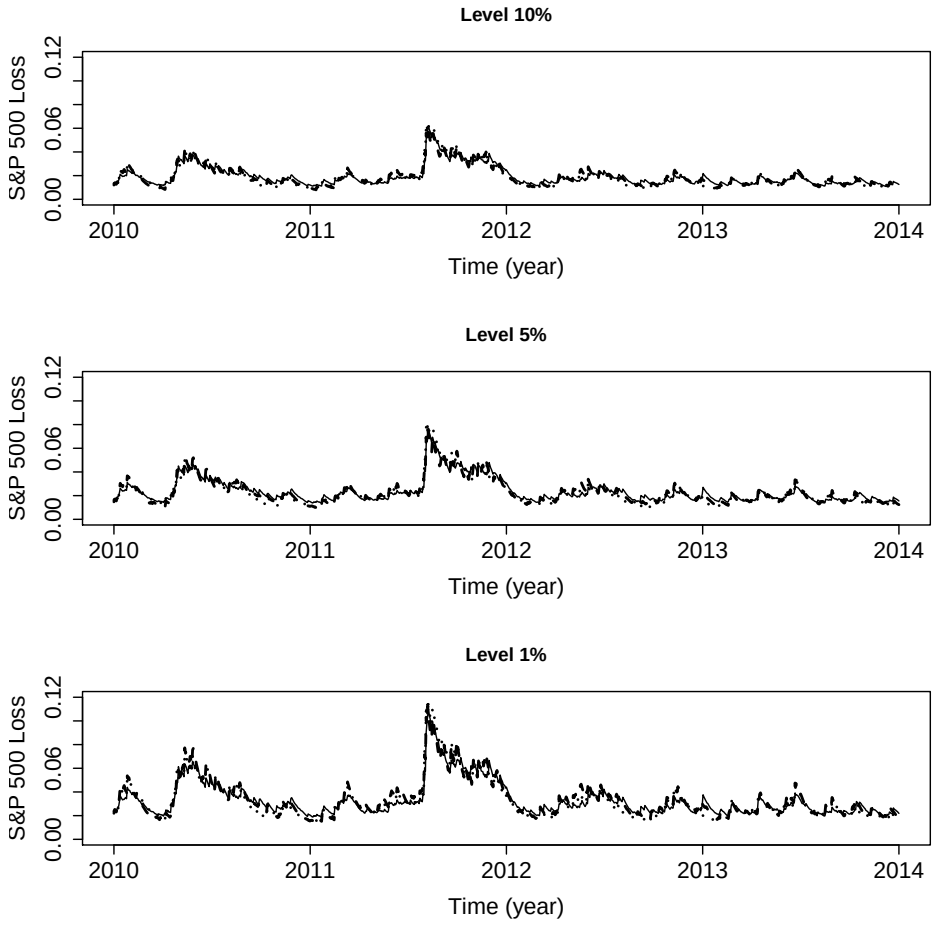

Figure 3.1. Sequentially predicted semiparametric CES for daily losses during 2010–2013, using standard GARCH (solid curve), EGARCH (dashed curve), and GJR-GARCH (dotted curve) models. Top, middle, and bottom plots correspond to level $1-\tau = 10\%, 5\%, 1\%$, respectively.

Using the three GARCH models, Figure 3.1 plots the corresponding sequential CES predictions at level $1-\tau = 10\%$ (top plot), 5% (middle plot), and 1% (bottom plot). From Figure 3.1, at each level, the three CES curves based on standard GARCH, EGARCH, and GJR-GARCH exhibit quite similar pattern, indicating the robustness of our method.

Despite the vast literature on CES estimation, their confidence interval construction has been largely ignored. Using the bootstrap procedure in Section 2.3, Figure 3.2 presents the semiparametrically estimated CES at level 5% and the

corresponding pointwise 95% confidence interval. Due to the quite similar pattern of CES using different GARCH models, we only report the result for standard GARCH.

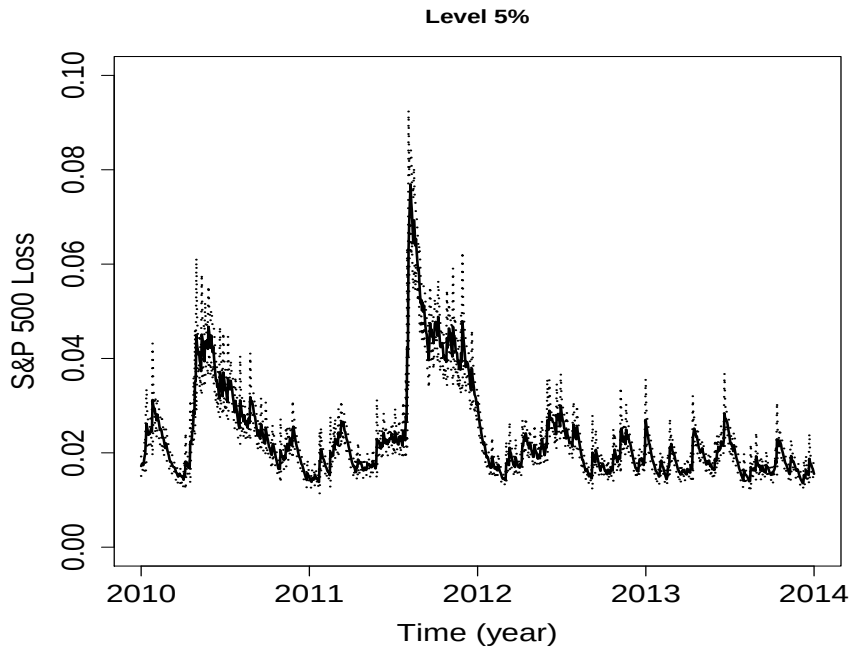

Figure 3.2. Sequentially predicted semiparametric CES (solid curve) at level 5% for daily losses during 2010–2013 using standard GARCH. The dotted curves are the pointwise bootstrap 95% confidence interval.

3.3.2 Comparison with some existing methods

We compare our semiparametric CES predictions with three parametric-distribution based approaches which were introduced in Section 2.6.2. Again, we use the decay factor 0.94 for EWMA and robust-EWMA and the procedure in Gerlach, Lu and Huang (2013) for skewed-EWMA. Figure 3.3 plots the sequential CES predictions using the aforementioned four methods: the semiparametric method with standard GARCH, EWMA, robust-EWMA, and skewed-EWMA. The four methods lead to quite similar CES curves.

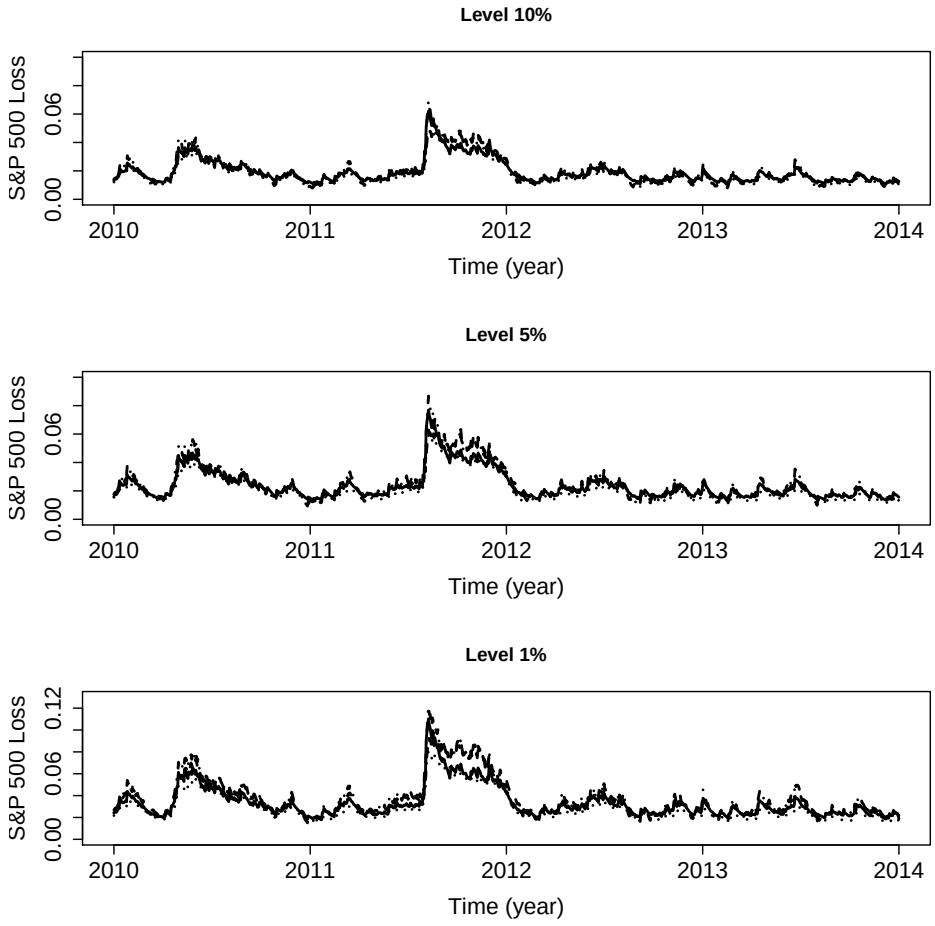

Figure 3.3. Comparison of sequentially predicted CES for daily losses during 2010–2013, using four methods: semiparametric method with standard GARCH (solid curve), the EWMA method (dotted curve), the robust-EWMA method (dashed curve), and the skewed-EWMA method (dotdashed curve). Top, middle, and bottom plots correspond to level $1 - \tau = 10\%, 5\%, 1\%$, respectively.

3.4 Assumptions and Proofs of Theorems

Throughout $C_1, C_2, \ldots, c_1, c_2, \ldots,$ are generic constants that may vary from line to line.

Assumption 9. *Recall $\varepsilon_i(\delta)$ in (3.14). Let $\epsilon > 0$ be some small constant. For all $i \in \{1, 2, ..., n\}$*

(i) There exist b_1, b_2, b_3, and $b_4 > 0$ such that

$$\mathbb{E}\Big[\sup_{|\delta|\le\epsilon, z\in\mathbb{R}} |\varepsilon_i(\delta)|^r \mathbf{1}_{|\varepsilon_i(\delta)-z|\le v}\Big] \le b_r v, \text{ for } r = 1, 2, 3, 4. \tag{3.19}$$

(ii) Let $\dot{\varepsilon}_i(\delta)$ be the partial derivative with respect to δ. There exists $L_1(\mathbf{Y}_i, \mathbf{X}_i)$ and $L_2(\mathbf{Y}_i, \mathbf{X}_i)$, both being independent of $\varepsilon_i(\delta)$ such that

$$\sup_{|\delta|\le\epsilon}[\varepsilon_i(\delta)] \le L_1(\mathbf{Y}_i, \mathbf{X}_i) \in \mathcal{L}^4, \text{ and } \sup_{|\delta|\le\epsilon}[\dot{\varepsilon}_i(\delta)] \le L_2(\mathbf{Y}_i, \mathbf{X}_i) \in \mathcal{L}^4. \tag{3.20}$$

Assumption 10. *Recall $J(\delta, z)$ from (3.15).Denote by $\dot{J}(\delta, z)$ and $\ddot{J}(\delta, z)$ the gradient vector and Hessian matrix of $J(\delta, z)$. Let $\epsilon > 0$ be the small constant in Assumption 9. Assume that $|\dot{J}(0, z)|$ is continuous in z and that $|\ddot{J}(\delta, z)|$ is bounded on $|\delta| \le \epsilon$.*

Lemma 10. *Suppose Assumption7, and 9-10 hold. Then for any given $c > 0$,*

$$\sup_{|z|\le c}\Big|\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i\mathbf{1}_{\hat{\varepsilon}_i\ge z} - \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\mathbf{1}_{\varepsilon_i\ge z} - \dot{J}_1(0, z)(\hat{\theta}-\theta)\Big| \;=\; o_p(n^{-1/2}), \tag{3.21}$$

and

$$\sup_{|z|\le c}\Big|\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\hat{\varepsilon}_i\ge z} - \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\varepsilon_i\ge z} - \dot{J}_2(0, z)(\hat{\theta}-\theta)\Big| \;=\; o_p(n^{-1/2}). \tag{3.22}$$

Proof. We will prove (3.21), (3.22) can be shown using similar arguments. Define

$$\phi_i(\delta, z) = \varepsilon_i(\delta)\mathbf{1}_{\varepsilon(\delta)\ge z} - \varepsilon_i\mathbf{1}_{\varepsilon\ge z}.$$

To prove (3.21), it suffices to prove that, for any given $c_1 > 0$,

$$\sup_{|\delta|\le c_1/\sqrt{n}, |z|\le c}\Big|\frac{1}{n}\sum_{i=1}^{n}\phi_i(\delta, z) - \dot{J}_1(0, z)(\hat{\theta}-\theta)\Big| = o_p(n^{-1/2}) \tag{3.23}$$

For simplicity we assume that: (i) $c_1 = c = 1$, and (ii) $z \in [0, 1], \delta \in [0, 1/\sqrt{n}]$. By Taylor's expansion and $\hat{\theta} = \theta + O_p(n^{-1/2})$,

$$\mathbb{E}[\phi_i(\delta, z)] = J_1(\delta, z) - J_1(0, z) = \delta\dot{J}_1(0, z) + O(n^{-1}), \tag{3.24}$$

uniformly on $\delta \in [0, 1/\sqrt{n}], z \in [0, 1]$. Thus to prove 3.23, it suffices to prove

$$\sup_{(\delta,z)\in[0,1/\sqrt{n}]\times[0,1]} \Big|M(\delta, z)\Big| = o_p(n^{-\frac{1}{2}}), \text{with } M(\delta, z) = \sum_{i=1}^{n} \Big(\phi_i(\delta, z) - \mathbb{E}[\phi_i(\delta, z)]\Big) \tag{3.25}$$

Let $N = \lfloor n^{1+\epsilon} \rfloor$ with ϵ being determined later. Consider the evenly spaced $(N+1)^2$ grid points

$$\delta_j = \sqrt{\frac{j}{n}} w_1 \text{ with } w_1 = \sqrt{\frac{1}{N}}, \text{ and } z_j = j w_2 \text{ with } w_2 = \frac{1}{N}, j = 0, 1, ..., N$$

partitioning $[0, 1/\sqrt{n}] \times [0, 1]$ into N^2 cells. For each $(\delta, z) \in [0, 1/\sqrt{n}] \times [0, 1]$ there exists one grid point $(\delta_j, z_{j'})$ such that $|\delta - \delta_j| \leq w_1$ and $|z - z_{j'}| \leq w_2$. Thus $|M(\delta, z)| \leq |M(\delta_j, z_{j'})| \leq |M(\delta_j, z_{j'})| + |M(\delta, z) - M(\delta_j, z_{j'})|$ and we have

$$\sup_{(\delta,z)\in[0,1/\sqrt{n}]\times[0,1]} |M(\delta, z)| \leq \max_{0\leq j,j'\leq N} |M(\delta_j, z_{j'})| + R_n, \tag{3.26}$$

where

$$R_n = \sup_{\Omega} |M(\delta, z) - M(\delta', z)| \text{ with } \Omega = \{|\delta - \delta'| \leq w_1, |z - z'| \leq w_2\}. \tag{3.27}$$

To show $R_n = o_p(\sqrt{n})$, note that for all $i = 1, 2, ..., n$,

$$\sup_{\Omega} |\phi_i(\delta, z) - \phi_i(\delta', z')| \leq 2 \sup_{\Omega} |\varepsilon_i(\delta) \mathbf{1}_{\varepsilon_i(\delta)\geq z} - \varepsilon_i(\delta') \mathbf{1}_{\varepsilon_i(\delta')\geq z'}|. \tag{3.28}$$

Therefore

$$\mathbb{E}[R_n] \leq 4n\mathbb{E}[\sup_{\Omega} |\varepsilon_1(\delta) \mathbf{1}_{\varepsilon_1(\delta)\geq z} - \varepsilon_1(\delta') \mathbf{1}_{\varepsilon_1(\delta')\geq z'}|] \tag{3.29}$$

Using the fact that $\varepsilon_i(\delta) \geq z$ is equivalent to

$$\varepsilon_i \geq \frac{\sigma(\theta + \delta, \mathbf{X}_i) z + [\mu(\theta + \delta, \mathbf{X}_i) - \mu(\theta, \mathbf{X}_i)]}{\sigma(\theta, \mathbf{X}_i)} \quad := \quad \xi(\delta, z, \mathbf{X}_i),$$

and the inequality that, for any $\lambda > 0$, $|\mathbf{1}_{z\leq 0} - \mathbf{1}_{z'\leq 0}| \leq 2\mathbf{1}_{|z-z'|\geq\lambda} + \mathbf{1}_{|z'|<\lambda}$,we can

write

$$
\begin{aligned}
& \sup_{\Omega} |\varepsilon_1(\delta)\mathbf{1}_{\varepsilon_1(\delta)\geq z} - \varepsilon_1(\delta')\mathbf{1}_{\varepsilon_1(\delta')\geq z'}| \\
= \; & \sup_{\Omega} |[\varepsilon_1(\delta) - \varepsilon_1(\delta)\mathbf{1}_{\varepsilon_1(\delta)<z}] - [\varepsilon_1(\delta') - \varepsilon_1(\delta')\mathbf{1}_{\varepsilon_1(\delta)<z'}]| \\
\leq \; & \sup_{\Omega} |\varepsilon_1(\delta) - \varepsilon_1(\delta')| + \sup_{\Omega} |\varepsilon_1(\delta)\mathbf{1}_{\varepsilon_1<\xi(\delta,z,\mathbf{X}_1)} - \varepsilon_1(\delta')\mathbf{1}_{\varepsilon_1<\xi(\delta',z',\mathbf{X}_1)}| \\
\leq \; & 2\sup_{\Omega} |\varepsilon_1(\delta) - \varepsilon_1(\delta')| + \sup_{\Omega}[|\varepsilon_1(\delta')||\mathbf{1}_{\varepsilon_1<\xi(\delta,z,\mathbf{X}_1)} - \mathbf{1}_{\varepsilon_1<\xi(\delta',z',\mathbf{X}_1)}|] \\
\leq \; & 2\sup_{\Omega} |\varepsilon_1(\delta) - \varepsilon_1(\delta')| + \sup_{\Omega}[|\varepsilon_1(\delta')|(2\mathbf{1}_{|\xi(\delta',z',\mathbf{X}_1)-\xi(\delta,z,\mathbf{X}_1)|\geq\lambda} + \mathbf{1}_{|\varepsilon_1-\xi(\delta',z',\mathbf{X}_1)|<\lambda})] \\
= \; & 2\sup_{\Omega} |(\delta-\delta')\dot{\varepsilon}_1(\delta')| + \sup_{\Omega}[|\varepsilon_1(\delta')|(2\mathbf{1}_{|(\delta-\delta')\dot{\varepsilon}_1(\delta')+(z'-z)|\geq\lambda} + \mathbf{1}_{|\varepsilon_1(\delta')-z'|<\lambda})] \quad (3.30)
\end{aligned}
$$

By Assumption 9(ii), on Ω, $|(\delta-\delta')\dot{\varepsilon}_1(\delta')| \leq w_1|L_2(\mathbf{Y}_1,\mathbf{X}_1)|$. Thus by (3.30) and Assumption 9(i),

$$
\begin{aligned}
\mathbb{E}[R_n] \; \leq \; & 4n\Big\{2w_1\mathbb{E}|L_2(\mathbf{Y}_1,\mathbf{X}_1)| \\
& +2\mathbb{E}|L_1(\mathbf{Y}_1,\mathbf{X}_1)|\mathbb{P}\{w_1|L_2(\mathbf{Y}_1,\mathbf{X}_1)| + w_2 > \lambda\} + b_1\lambda\Big\} \\
\leq \; & 4n\Big\{2w_1\mathbb{E}|L_2(\mathbf{Y}_1,\mathbf{X}_1)| + 2w_1\frac{\mathbb{E}|L_1(\mathbf{Y}_1,\mathbf{X}_1)|\mathbb{E}|L_2(\mathbf{Y}_1,\mathbf{X}_1)+w_2|}{\lambda} + b_1\lambda\Big\}
\end{aligned} \quad (3.31)
$$

The second inequality in (3.31) is followed by Markov inequality. By taking $\lambda = w_1 = \sqrt{\frac{1}{N}}$, $\mathbb{E}[R_n] = O(\frac{n}{\sqrt{n^{1+\epsilon}}}) = o_p(\sqrt{n})$. By (3.26), it remains to show that $\max_{0\leq j,j'\leq N}|M(\delta_j,z'_j)| = o_p(\sqrt{n})$. Define

$$
A = \max_{1\leq i\leq n}[\phi_i(\delta,z) - \mathbb{E}\phi_i(\delta,z)] = \max_{1\leq i\leq n}[\varepsilon_i(\delta)\mathbf{1}_{\varepsilon_i(\delta)\geq z} - \varepsilon_i\mathbf{1}_{\varepsilon_i\geq z}] + O(n^{-1/2}) \quad (3.32)
$$

By Markov's inequality, Assumption 9, and similar argument from (3.30) and (3.31),

$$
\mathbb{P}\{|A| \geq n^{5/12+5\epsilon/8}\} \leq n^{-(5/3+5\epsilon/2)}\mathbb{E}[A]^4
$$

$$
\begin{aligned}
&\leq \quad 8n^{-(5/3+5\epsilon/2)}\mathbb{E}[\max_{1\leq i\leq n}\{[\varepsilon_i(\delta)-\varepsilon_i]\mathbf{1}_{\varepsilon_i\geq\xi(\delta,z,\mathbf{X}_i)}+\varepsilon_i[\mathbf{1}_{\varepsilon_i-\xi(\delta,z,\mathbf{X}_i)\geq 0}-\mathbf{1}_{\varepsilon_i-z\geq 0}]\}^4 \\
&\quad +O(n^{-2})] \\
&\leq \quad 64n^{-(5/3+5\epsilon/2)}\mathbb{E}[\max_{1\leq i\leq n}\{[\delta\dot{\varepsilon}_i(0)]^4+\varepsilon_i^4|\mathbf{1}_{\varepsilon_i-\xi(\delta,z,\mathbf{X}_i)\geq 0}-\mathbf{1}_{\varepsilon_i-z\geq 0}|+O(n^{-2})\}] \\
&\leq \quad 64n^{-(5/3+5\epsilon/2)}\mathbb{E}[\max_{1\leq i\leq n}\{\varepsilon_i^4(2\mathbf{1}_{|\delta\dot{\varepsilon}_i(0)|>n^{-1/3}}+\mathbf{1}_{|\varepsilon_i-z|\leq n^{-1/3}})+O(n^{-2})\} \\
&\leq \quad 64n^{-(5/3+5\epsilon/2)}\mathbb{E}[\max_{1\leq i\leq n}\varepsilon_i^4\frac{2\delta^4\mathbb{E}[L_2(\mathbf{Y}_i,\mathbf{X}_i)]^4}{n^{-4/3}}+O(n^{-1/3})] \\
&= \quad O[n^{-2-5\epsilon/2}] \qquad (3.33)
\end{aligned}
$$

The first and the fifth inequality in (3.33) follows from Markov's inequality; And the second and the third inequality in (3.33) follows from $(a+b)^4 \leq 8(a^4+b^4)$. Similarly we have

$$
Var[\phi_i(\delta,z), -\mathbb{E}\phi_i(\delta,z)] \leq \mathbb{E}[A]^2 \leq c_2 n^{-1/3}, \text{ for some } c_2 > 0. \qquad (3.34)
$$

Thus by exponential inequality for stationary α-mixing process with mixing coefficient recalling from Assumption 7 (i), $\alpha_j < C_1\alpha^j$ with $0 < C_1 < \infty$ and $\alpha \in (0,1)$, for all $c_3 > 0$, take $\ell = \lfloor n^\beta \rfloor$, with some β being determined later, we have

$$
\begin{aligned}
&\mathbb{P}\{|M(\delta,z)| \geq c_3\sqrt{n}\} = \mathbb{P}\{|\sum_{i=1}^n[\phi_i(\delta,z)-\mathbb{E}\phi_i(\delta,z)]| \geq c_3\sqrt{n}\} \\
\leq \quad &\mathbb{P}\{|\sum_{i=1}^n[\phi_i(\delta,z)-\mathbb{E}\phi_i(\delta,z)]| \geq c_3\sqrt{n}, |A| \leq n^{5/12+5\epsilon/8}\} + \mathbb{P}\{|A| > n^{5/12+5\epsilon/8}\} \\
\leq \quad &4exp\Big\{\frac{-c_3^2 n^{\beta+1}}{144n^2c_2n^{-1/3}+4n^{5/12+5\epsilon/8}(c_3\sqrt{n})n}\Big\} \\
&+22n^\beta\alpha^{(n/2n^\beta)}\sqrt{1+\frac{4(n^{5/12+5\epsilon/8})n}{c_3\sqrt{n}}}+O(n^{-2-5\epsilon/2}) \\
= \quad &O\Big[exp\Big\{\frac{-n^\beta}{n^{2/3}+n^{11/12+5\epsilon/8}}\Big\}+n^{\beta+11/24+5\epsilon/16}\alpha^{n^{1-\beta}}+n^{-2-5\epsilon/2}\Big], \qquad (3.35)
\end{aligned}
$$

uniformly on $|\delta| \leq 1/\sqrt{n}, z \in \mathbb{R}$. Recall $N = \lfloor n^{1+\epsilon} \rfloor$, by (3.35),

$$
\mathbb{P}\{\sup_{0\leq j,j'\leq N}|M(\delta_j,z_{j'})| \geq c_3\sqrt{n}\}
$$

$$\begin{aligned} &\leq \sum_{j,j'=0}^{N} \mathbb{P}\{|M(\delta_j, z_{j'})| > c_3\sqrt{n}\} \\ &= O[n^{2+2\epsilon} exp\{-n^{\beta-11/12-5\epsilon/8}\} + n^{\beta+59/24+37\epsilon/16}\alpha^{n^{1-\beta}} + n^{-\epsilon/2}] \end{aligned} \quad (3.36)$$

Take any

$$\frac{11}{12} + \frac{5}{8}\epsilon < \beta < 1, \text{ and } 0 < \epsilon < \frac{2}{15}.$$

Then the right-hand-side of (3.36) becomes $o_p(\sqrt{n})$. Since c_3 is arbitrary, we conclude that $\sup_{0\leq j,j'\leq N} |M(\delta_j, z_{j'})| = o_p(\sqrt{n})$. ♢

Lemma 11. *Suppose f_ε is continuous, bounded, and positive on $[-c_1, c_1]$, for some constant $c_1 > 0$. Then for any given constant $c_2 > 0$,*

$$\sup_{|z|\leq c_1, |v|\leq c_2/\sqrt{n}} \Big| \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \mathbf{1}_{\varepsilon_i \geq z+v} - \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \mathbf{1}_{\varepsilon_i \geq z} + vzf_\varepsilon(z) \Big| = o_p(n^{-\frac{1}{2}}), \quad (3.37)$$

and

$$\sup_{|z|\leq c_1, |v|\leq c_2/\sqrt{n}} \Big| \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\varepsilon_i \geq z+v} - \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\varepsilon_i \geq z} + vf_\varepsilon(z) \Big| = o_p(n^{-1/2}). \quad (3.38)$$

Proof. We shall prove (3.37), (3.38) can be shown similarly. For simplicity we assume that: (i) $c_1 = c_2 = 1$, and (ii) $z \in [0, 1], v \in [-1/\sqrt{n}, 0]$. Let

$$\kappa_i(z, v) = \varepsilon_i \mathbf{1}_{\varepsilon_i \geq z+v} - \varepsilon_i \mathbf{1}_{\varepsilon_i \geq z} = \varepsilon_i \mathbf{1}_{z+v\leq \varepsilon_i < z}. \quad (3.39)$$

Then by Taylor's expansion,

$$\mathbb{E}[\kappa_i(z, v)] = \mathbb{E}[\varepsilon_i \mathbf{1}_{z+v\leq\varepsilon_i<z}] = G(z) - G(z + v) = -v\dot{G}(z) + o_p(n^{-1/2}),$$

where

$$G(z) = zF_\varepsilon(z) - \int_0^z F_\varepsilon(s)ds, \text{ and } \dot{G}(z) = zf_\varepsilon(z).$$

Note that $\dot{G}(z)$ is bounded and continuous, for all $z \in [0, 1]$. To prove (3.37), it

suffice to show

$$\sup_{(z,v)\in[0,1]\times[-1/\sqrt{n}]} |\sum_{i=1}^{n}\{\kappa_i(z,v) - \mathbb{E}[\kappa_i(z,v)]\}| = o_p(\sqrt{n}). \tag{3.40}$$

Consider the evenly-spaced points

$z_j = j/n, v_{j'} = j'/n^{3/2}, j, j' = 0, 1, ..., n$, partitioning $[0,1] \times [-1/\sqrt{n}]$ into n^2 cells.

By the bounded derivative of $G(z)$, there exists an universal $b_3 > 0$ such that for all $z \in [z_{j-1}, z_j]$ and all $v \in [v_{j'}, v_{j'-1}]$,

$$G(z_j) - G(z_{j-1} + v_{j'}) - \frac{b_3}{n} \;\; \leq \;\; G(z) - G(z+v) \leq G(z_{j-1}) - G(z_j + v_{j'-1}) + \frac{b_3}{n} \tag{3.41}$$

Observe that, for $\varepsilon_i > 0$,

$$\varepsilon_i(\mathbf{1}_{\varepsilon_i<z_{j-1}} - \mathbf{1}_{\varepsilon_i<z_j+v_{j'-1}}) \leq \varepsilon_i(\mathbf{1}_{\varepsilon_i<z} - \mathbf{1}_{\varepsilon_i<z+v}) \leq \varepsilon_i(\mathbf{1}_{\varepsilon_i<z_j} - \mathbf{1}_{\varepsilon_i<z_{j-1}+v_{j'}}). \tag{3.42}$$

Define

$$\bar{\lambda}_{ijj'} = \quad \varepsilon_i(\mathbf{1}_{\varepsilon_i<z_j} - \mathbf{1}_{\varepsilon_i<z_{j-1}+v_{j'}}) - [G(z_j) - G(z_{j-1} + v_{j'})] \tag{3.43}$$

$$\lambda_{ijj'} = \quad \varepsilon_i(\mathbf{1}_{\varepsilon_i<z_{j-1}} - \mathbf{1}_{\varepsilon_i<z_j+v_{j'-1}}) - [G(z_{j-1}) - G(z_j + v_{j'-1})] \tag{3.44}$$

Combine (3.41) and (3.42), we have

$$\lambda_{ijj'} - \frac{b_3}{n} \leq \kappa_i(z,v) - \mathbb{E}\kappa_i(z,v) \leq \bar{\lambda}_{ijj'} + \frac{b_3}{n} \tag{3.45}$$

Thus $|\sum_{i=1}^n [\kappa_i(z,v) - \mathbb{E}\kappa_i(z,v)]| \leq |\sum_{i=1}^n \bar{\lambda}_{ijj'}| + |\sum_{i=1}^n \lambda_{ijj'}| + b_3$.
Hence

$$\begin{aligned}
& \sup_{(z,v)\in[0,1]\times[-1/\sqrt{n},0]} |\sum_{i=1}^{n}[\kappa_i(z,v) - \mathbb{E}\kappa_i(z,v)]| \\
= \quad & \max_{1\leq j,j'\leq n} \sup_{z\in[z_{j-1},z_j],v\in[v_{j'},v_{j'-1}]} |\sum_{i=1}^{n}[\kappa_i(z,v) - \mathbb{E}\kappa_i(z,v)]| \\
\leq \quad & \max_{1\leq j,j'\leq n} |\sum_{i=1}^{n} \bar{\lambda}_{ijj'}| + \max_{1\leq j,j'\leq n} |\sum_{i=1}^{n} \lambda_{ijj'}| + o(\sqrt{n})
\end{aligned} \tag{3.46}$$

For all $\varepsilon_i > 0, i = 1, 2, ..., n$,

$$\begin{aligned} |\bar{\lambda}_{ijj'}| = & \quad |\varepsilon_i(\mathbf{1}_{z_{j-1}+v_{j'}<\varepsilon_i<z_j}) - [G(z_j) - G(z_{j-1}+v_{j'})]| \\ \leq & \quad |\varepsilon_i \mathbf{1}_{\varepsilon_i<z_j}| + |z_j - z_{j-1} - v_{j'}||z_{j-1}+v_{j'}|f_\varepsilon(z_{j-1}+v_{j'}) \\ \leq & \quad z_j + [|z_j - z_{j-1}| + |v_{j'}|]|z_{j-1}+v_{j'}|f_\varepsilon(z_{j-1}+v_{j'}) \\ \leq & \quad b_4, \text{ for some } b_4 > 0. \end{aligned} \tag{3.47}$$

The last inequality in (3.47) holds by the positivity and boundedness of f_ε. On the other hand, by Hölder's inequality and Assumption 9(ii),

$$\begin{aligned} Var[\sum_{i=1}^n \bar{\lambda}_{ijj'}] = & \quad nVar[\bar{\lambda}_{1jj'}] = n\mathbb{E}[\bar{\lambda}_{1jj'}^2] \\ \leq & \quad n\mathbb{E}[\varepsilon_1 \mathbf{1}_{z_{j-1}+v_{j'}<\varepsilon_i<z_j}]^2 \\ \leq & \quad n\sqrt{\mathbb{E}[\varepsilon_1^4]\mathbb{E}[\mathbf{1}_{z_{j-1}+v_{j'}<\varepsilon_i<z_j}]} \\ = & \quad n\sqrt{\mathbb{E}[\varepsilon_1^4][F_\varepsilon(z_j) - F_\varepsilon(z_{j-1}+v_{j'})]} \\ = & \quad O[n\sqrt{n^{-1}+n^{-1/2}}] = O(n^{3/4}). \end{aligned} \tag{3.48}$$

Therefore, we have $\sqrt{Var[\sum_{i=1}^n \bar{\lambda}_{ijj'}]} \leq b_5 n^{3/8}$, for some $b_5 > 0$. Using (3.47), (3.48), and Bernstein's exponential inequality (Bennett, 1962) for the sum of bounded and independent random variables, we obtain

$$\begin{aligned} \mathbb{P}\{|\sum_{i=1}^n \bar{\lambda}_{ijj'}| \geq b_6\sqrt{n}\} \leq & \quad 2\exp\left\{\frac{-(\frac{b_6 n^{1/8}}{b_5})^2}{2 + \frac{2}{3}(\frac{b_4}{b_5 n^{3/8}})(\frac{b_6 n^{1/8}}{b_5})}\right\} \\ = & \quad O[\exp\{-n^{1/2}\}], \end{aligned}$$

for large enough n and all j, j', where $b_6 > 0$ is some given constant. Therefore,

$$\mathbb{P}\{\max_{0\leq j,j'\leq n} |\sum_{i=1}^n \bar{\lambda}_{ijj'}| \geq b_6\sqrt{n}\} \leq \sum_{j,j'=0}^n \mathbb{P}\{|\sum_{i=1}^n \bar{\lambda}_{ijj'}| \geq b_6\sqrt{n}\} = O[n^2 \exp\{-n^{1/2}\}].$$

Since b_6 is arbitrary, (3.49) implies that $\max_{0\leq j,j'\leq n} |\sum_{i=1}^n \bar{\lambda}_{ijj'}| = o_p(\sqrt{n})$. Similarly we have $\max_{0\leq j,j'\leq n} |\sum_{i=1}^n \lambda_{ijj'}| = o_p(\sqrt{n})$. The results in (3.37) then follows in

view of (3.46).

$\diamondsuit$

Lemma 12. *Suppose the assumptions in Lemma 10 and Lemma 11 hold. Then*

$$\begin{aligned} & \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i \mathbf{1}_{\hat{\varepsilon}_i \geq \ell(\hat{\theta})} - \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i \mathbf{1}_{\varepsilon_i \geq \ell(\theta)} \\ = \quad & [\dot{J}_1(0,\ell(\theta))^T - \dot{\ell}(\theta)\ell(\theta) f_\varepsilon(\ell(\theta))](\hat{\theta}-\theta) + o_p(n^{-1/2}), \end{aligned} \tag{3.49}$$

and

$$\begin{aligned} & \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\hat{\varepsilon}_i \geq \ell(\hat{\theta})} - \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\varepsilon_i \geq \ell(\theta)} \\ = \quad & [\dot{J}_2(0,\ell(\theta))^T - \dot{\ell}(\theta) f_\varepsilon(\ell(\theta))](\hat{\theta}-\theta) + o_p(n^{-1/2}). \end{aligned} \tag{3.50}$$

Proof. We will prove (3.49), (3.50) can be shown similarly. Let $\mathcal{Y} = [\mathcal{Y}_1, \mathcal{Y}_2]$ be any bounded interval. Recall $\ell(\theta)$ from (3.4). Write $z = \ell(\theta) := \ell(\theta, y, x)$ and $\hat{z} = \ell(\theta+\delta) := \ell(\hat{\theta}, y, x)$. Since $\ell(\hat{\theta}, y, x)$ is clearly increasing in y, $\ell(\hat{\theta}, \mathcal{Y}_1, x) \leq \ell(\hat{\theta}, y, x) \leq \ell(\hat{\theta}, \mathcal{Y}_2, x)$, uniformly on $y \in \mathcal{Y}$. By Lemma 10,

$$\sup_{y\in\mathcal{Y}} \Big| \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i \mathbf{1}_{\hat{\varepsilon}_i \geq \hat{z}} - \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i \mathbf{1}_{\varepsilon_i \geq \hat{z}} - \dot{J}_1(0,\hat{z})(\hat{\theta}-\theta) \Big| = o_p(n^{-1/2}). \tag{3.51}$$

By $\hat{\theta} - \theta = O_p(n^{-1/2})$ and Assumption 9(ii),

$$\hat{z} - z = \ell(\hat{\theta}) - \ell(\theta) = \dot{\ell}(\theta)(\hat{\theta}-\theta) + o_p(n^{-1/2}). \tag{3.52}$$

Thus, by (3.52) and Lemma 11,

$$\sup_{y\in\mathcal{Y}} \Big| \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i \mathbf{1}_{\varepsilon_i \geq \hat{z}} - \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i \mathbf{1}_{\varepsilon_i \geq z} - [-\dot{\ell}(\theta)(\hat{\theta}-\theta)\ell(\theta) f_\varepsilon(\ell(\theta))] \Big| = o_p(n^{-1/2}). \tag{3.53}$$

The continuity of $\dot{J}_1(0,z)$ in Assumption 10 implies $\dot{J}_1(0,\hat{z}) - \dot{J}_1(0,z) = o_p(1)$ uniformly on $y \in \mathcal{Y}$. Combine with (3.51) and (3.53), we obtain the uniform

approximation

$$\begin{aligned} & \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i\mathbf{1}_{\hat{\varepsilon}_i\geq\hat{z}} - \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\mathbf{1}_{\varepsilon_i\geq z} \\ = \quad & \dot{J}_1(0,z)(\hat{\theta}-\theta) + [-\dot{\ell}(\theta)\ell(\theta)f_\varepsilon(\ell(\theta))(\hat{\theta}-\theta)] + o_p(n^{-1/2}). \end{aligned} \tag{3.54}$$

♢

3.4.1 Proof of Theorem 7

Proof. By Lemma 12,

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i\mathbf{1}_{\hat{\varepsilon}_i\geq\ell(\hat{\theta})} = \mathbb{E}[\varepsilon_0\mathbf{1}_{\varepsilon_0\geq\ell(\theta)}] + A + o_p(n^{-1/2}), \tag{3.55}$$

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\hat{\varepsilon}_i\geq\ell(\hat{\theta})} = \mathbb{E}[\mathbf{1}_{\varepsilon_0\geq\ell(\theta)}] + B + o_p(n^{-1/2}). \tag{3.56}$$

where

$$\begin{aligned} A &= \frac{1}{n}\sum_{i=1}^{n}\Big\{\varepsilon_i\mathbf{1}_{\varepsilon_i\geq\ell(\theta)} - \mathbb{E}[\varepsilon_i\mathbf{1}_{\varepsilon_i\geq\ell(\theta)}]\Big\} + [\dot{J}_1(0,\ell(\theta))^T - \dot{\ell}(\theta)\ell(\theta)f_\varepsilon(\ell(\theta))](\hat{\theta}-\theta), \\ B &= \frac{1}{n}\sum_{i=1}^{n}\Big\{\mathbf{1}_{\varepsilon_i\geq\ell(\theta)} - \mathbb{E}[\mathbf{1}_{\varepsilon_i\geq\ell(\theta)}]\Big\} + [\dot{J}_2(0,\ell(\theta))^T - \dot{\ell(\theta)}f_\varepsilon(\ell(\theta))](\hat{\theta}-\theta) \end{aligned} \tag{3.57}$$

Since $\{\varepsilon_i\}$ are iid, the first term in A and the first term in B are of the order $O_p(n^{-1/2})$. Under the condition $\hat{\theta}-\theta = O_p(n^{-1/2})$, $A = O_p(n^{-1/2})$, $B = O_p(n^{-1/2})$. By the condition $\hat{\theta} = \theta + O_p(n^{-1/2})$ and Taylor's expansion, $\mu(\hat{\theta},x) = \mu(\theta,x) + \dot{\mu}(\theta,x)^T(\hat{\theta}-\theta) + o_p(n^{-1/2})$ and $\sigma(\hat{\theta},x) = \sigma(\theta,x) + \dot{\sigma}(\theta,x)^T(\hat{\theta}-\theta) + o_p(n^{-1/2})$. Substituting the latter expansions and (3.56) into (3.10) and using the expression (3.4) for $\mathrm{CES}(y|x)$, we can obtain

$$\begin{aligned} \widehat{\mathrm{CES}}(y|x) &= \mu(\hat{\theta},x) + \sigma(\hat{\theta},x)\frac{\mathbb{E}[\varepsilon_0\mathbf{1}_{\varepsilon_0\geq\ell(\theta)}] + A}{\mathbb{E}[\mathbf{1}_{\varepsilon_0\geq\ell(\theta)}] + B} + o_p(n^{-1/2}), \\ &= \mathrm{CES}(y|x) + \Big[\dot{\mu}(\theta,x)^T + \dot{\sigma}(\theta,x)^T\frac{\mathbb{E}[\varepsilon_0\mathbf{1}_{\varepsilon_0\geq\ell(\theta)}]}{\mathbb{E}\mathbf{1}_{\varepsilon_0\geq\ell(\theta)}}\Big](\hat{\theta}-\theta) \\ &\quad + \sigma(\theta,x)\Big[\frac{\mathbb{E}[\varepsilon_0\mathbf{1}_{\varepsilon_0\geq\ell(\theta)}] + A}{\mathbb{E}[\mathbf{1}_{\varepsilon_0\geq\ell(\theta)}] + B} - \frac{\mathbb{E}[\varepsilon_0\mathbf{1}_{\varepsilon_0\geq\ell(\theta)}]}{\mathbb{E}\mathbf{1}_{\varepsilon_0\geq\ell(\theta)}}\Big] + o_p(n^{-1/2}) \end{aligned}$$

$$= \quad \text{CES}(y|x) + W_n + o_p(n^{-1/2}), \tag{3.58}$$

where

$$\begin{aligned} W_n &= \Big[\dot{\mu}(\theta,x)^T + \dot{\sigma}(\theta,x)^T \frac{\mathbb{E}[\varepsilon_0 \mathbf{1}_{\varepsilon_0 \geq \ell(\theta)}]}{\mathbb{E}\mathbf{1}_{\varepsilon_0 \geq \ell(\theta)}}\Big](\hat{\theta} - \theta) \\ &\quad + \frac{\sigma(\theta,x)}{(\mathbb{E}\mathbf{1}_{\varepsilon_0 \geq \ell(\theta)})^2}\Big[\mathbb{E}\mathbf{1}_{\varepsilon_0 \geq \ell(\theta)} A - \mathbb{E}[\varepsilon_0 \mathbf{1}_{\boldsymbol{\varepsilon_0} \geq \ell(\theta)}]B\Big]. \end{aligned} \tag{3.59}$$

The result then follows from (3.58) in view of $\hat{\theta} - \theta = O_p(n^{-1/2})$, $A = O_p(n^{-1/2})$, and $B = O_p(n^{-1/2})$. ♢

3.4.2 Proof of Theorem 8

Proof. Substituting the expressions for $\hat{\theta} - \theta$ in (3.11) and A and B in (3.57) into (3.59), after some calculations we can rewrite

$$W_n = n^{-1}\sum_{i=1}^{n} \psi_i + o_p(n^{-1/2}), \text{ where } \psi_i \quad = \quad H^T \zeta_i + \eta_i. \tag{3.60}$$

η_i, $\zeta_i = \zeta(\theta, \varepsilon_i, \mathbf{X}_i)$, and H in (3.60) are defined in (3.7), (3.11), and (3.17), respectively. By Assumption 8, $\mathbb{E}\left\{ H^T \zeta_i | \mathbf{X}_i \right\} = 0$. Thus $\mathbb{E}[\psi_i | \mathbf{X}_i] = 0$. On the other hand,

$$\begin{aligned} \mathbb{E}[\psi_i^2] &= (H^T H)\mathbb{E}[\zeta^T(\theta, \varepsilon_i, \mathbf{X}_i)\zeta(\theta, \varepsilon_i, \mathbf{X}_i)] + 2H^T \mathbb{E}[\eta_i \zeta(\theta, \varepsilon_i, \mathbf{X}_i)] + \mathbb{E}[(\eta_i^2] \\ &< \infty. \end{aligned} \tag{3.61}$$

The first term in (3.61) is finite because $\zeta_i \in \mathcal{L}^2$. For the third term in (3.61), by Assumption 9,

$$\mathbb{E}[\eta_i^2] = (\mathbb{E}[\eta_i])^2 + Var(\eta_i) < \infty.$$

The second term in (3.61) then followed by Hölder's inequality. Thus by Lemma 5, (3.16) is proved. ♢

Bibliography

Aït-Sahalia, Y. and A.W. Lo, 2000, Nonparametric risk management and implied risk aversion. Journal of Econometrics 94, 9–51.

Bennett, G., 1962, Probability inequalities for the sum of independent random variables. Journal of the American Statistical Association 57, 33–45.

Billingsley, P., 1999, Convergence of Probability Measures, 2nd edition, John Wiley & Sons.

Bollerslev, T., 1986, Generalized autoregressive conditional heteroscedasticity. Journal of Econometrics 31, 307–327.

Bühlmann, P., 1997, Sieve bootstrap for time series. Bernoulli 3, 123–148.

Butler, J.S. and B. Schachter, 1998, Estimating Value-at-Risk with a precision measure by combining kernel estimation with historical simulation. Review of Derivatives Research 1, 371–390.

Cai, Z., 2002, Regression quantiles for time series. Econometric Theory 18, 169–192.

Cai Z. and X. Wang, 2008, Nonparametric estimation of conditional VaR and expected shortfall. Journal of Econometrics 147, 120–130.

Chen, S., 2008, Nonparametric estimation of expected shortfall. Journal of Financial Econometrics 6, 87–107.

Chen, S. and C. Tang, 2005, Nonparametric inference of value at risk for dependent financial returns. Journal of Financial Econometrics 3, 227–255.

Chen, X., 2007, Large sample sieve estimation of semi-nonparametric models. In: J.J. Heckman and E.L. Edward (editors), Handbook of Econometrics Vol. 6, Part 2, 5549–5632.

Chernozhukov, V. and L. Umanstev, 2001, Conditional value-at-risk: Aspects of modeling and estimation. Empirical Economics 26, 271–292.

Cosma, A., O. Scaillet and R. von Sachs, 2007, Multivariate wavelet–based shape preserving estimation for dependent observations. Bernoulli 13, 301–329.

Danielsson, J. and C.G. Vries, de, 2000, Value-at-risk and extreme returns. Annales déconomie et de statistique 60, 236–269.

Dowd, K., 1998, Beyond Value at Risk: The New Science of Risk Management, Wiley, New York.

Duffie, D. and J. Pan, 1997, An overview of value at risk. Journal of Derivatives 4, 7–49.

Engle, R. and S. Manganelli, 2004, CAViaR: Conditional autoregressive value at risk by regression quantile. Journal of Business and Economics Statistics 22, 367–381.

Fan, J. and Q. Yao, 2003, Nonlinear Time Series: Nonparametric and Parametric Methods, Springer, New York.

Fermanian, J.D. and O. Scaillet, 2005, Sensitivity analysis of VaR and Expected Shortfall for portfolios under netting agreements. Journal of Banking and Finance 29, 927–958.

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, 1317–1334.

Gerlach, R., Z. Lu and H. Huang, 2013, Exponentially smoothing the skewed laplace distribution for value-at-risk forecasting. Journal of Forecasting 32 534–550.

Ghalanos, A., 2014, rugarch: Univariate GARCH models. R package.

Glosten, L.R., R. Jagannathan and D. Runkle, 1993, On the relation between the expected value and the volatility of the nominal excess return on stocks. Journal of Finance 48, 1779–1801.

Gourieroux, X., J.P. Laurent and O. Scaillet, 2000, Sensitivity analysis of Values at Risk. Journal of Empirical Finance 7, 225–245.

Guermat C. and R.D.D. Harris, 2001, Robust conditional variance estimation and value-at-risk. Journal of Risk 4, 25–41.

Hall, P. and Q. Yao, 2003, Inference in ARCH and GARCH models with heavy-tailed errors. Econometrica 71, 285–317.

He, X. and Q.M. Shao, 1996, A general Bahadur representation of M-estimators and its application to linear regression with nonstochastic designs. The Annals of Statistics 24, 2608–2630.

Horowitz, J.L., 1996, Semiparametric estimation of a regression model with an unknown transformation of the dependent variable. Econometrica 64, 103–137.

Horváth, L. and G. Teyssière, 2001, Empirical process of the squared residuals of an arch sequence. The Annals of Statistics 29, 445–469.

Hull, J. and A. White, 1998, Value at Risk when daily changes are not normally distributed. Journal of Derivatives 5, 9–19.

Jurečková, J. and B. Procházka, 1994, Regression quantiles and trimmed least squares estimator in nonlinear regression model. Journal of Nonparametric Statistics 3, 201–222.

Lahiri, S.N., 2003, Resampling Methods for Dependent Data, Springer Series in Statistics, Springer, New York.

Lee, S. and C.Z. Wei, 1999, On residual empirical processes of stochastic regression models with applications to time series. The Annals of Statistics 27, 237–261.

Li, Q. and J. Racine, 2007, Nonparametric Econometrics, Princeton University Press, Princeton, New Jersey.

Matzkin, R.L., 2003, Nonparametric estimation of nonadditive random functions. Econometrica 71, 1339–1375.

Morgan, J.P., 1996, RiskMetrics Technical Document, 4th edn, New York.

Nelson, D.B., 1991, Conditional heteroskedasticity in asset returns: A new approach. Econometrica 59, 347–370.

Pollard, D., 1991, Asymptotic for least absolute deviation regression estimators. Econometric Theory 7, 186–199.

Portnoy, S. and R. Koenker, 1989, Adaptive L-estimation of Linear Models. The Annals of Statistics 17, 362–381.

Scaillet, O., 2003, The origin and development of VaR. In Modern Risk Management: A History, 15th Anniversary of Risk Magazine, Risk Publications, London, 151–158.

Scaillet, O., 2004, Nonparametric estimation and sensitivity analysis of expected shortfall. Mathematical Finance 14, 115–129.

Scaillet, O., 2005, Nonparametric estimation of conditional expected shortfall. Revue Assurances et Gestion des Risques/Insurance and Risk Management Journal 74, 639–660.

Wang, C. and Zhao, Z., 2016, Conditional Value-at-Risk: Semiparametric estimation and inference. Journal of Econometrics 195, 86-103.

Wu, W.B., K. Yu and G. Mitra, 2007, Kernel conditional quantile estimation for stationary processes with application to conditional Value-at-Risk. Journal of

Financial Econometrics 6, 253-ï£¡270.

Wu, W.B. and X. Shao, 2004, Limit theorems for iterated random functions. Journal of Applied Probability 41, 425–436.

Wu, W.B. and Z. Zhao, 2008, Moderate deviations for stationary processes. Statistica Sinica 18, 769–782.

Yu, K. and M.C. Jones, 1998, Local linear quantile regression. Journal of the American Statistical Association 93, 228–237.

Zakoian, J.M., 1994, Threshold heteroskedastic models. Journal of Economic Dynamics and Control 18, 931–955.

Zhao, Z., 2010, Density estimation for nonlinear parametric models with conditional heteroscedasticity. Journal of Econometrics 155, 71–82.

Zhao, Z. and Z. Xiao, 2014, Efficient regressions via optimally combining quantile information. To appear, Econometric Theory.

Vita

Chuan-Sheng Wang

1. Education

   - Ph.D. in Statistics, The Pennsylvania State University Aug 2012 - May 2018
   - B.A. in Economics, National Taiwan University Sep 2006 - Jan 2011

2. Skills

   - Proficient with: time series analysis, analysis of stochastic process, data mining, regression models, nonparametric and semiparametric estimation of risk measures, ANOVA, categorical data analysis
   - Proficient with: R, SAS, SQL, SPSS, LaTex

3. Selected Courses

   - Time series analysis(R, SAS), data mining(R), statistical consulting practicum(R, SAS, SPSS, minitab), stochastic process Monte Carlo method(R), categorical data analysis(R), ANOVA and experimental design(R, SAS, SPSS), regression models(R, SAS, SPSS), spatial models(R), functional data analysis(R), clinical trials(SAS)
   - Mathematical analysis, probability theory, statistical inference, linear models, asymptotic tools, multivariate analysis
   - Investment, financial management, microeconomics, macroeconomics, financial economics, tax theory and policy, business law, international finance