Econometric Methods for Expected Shortfall and Value-at-Risk

by

Peter Horvath

Department of Department of Economics
Duke University

Date: _____

Approved:

_____
Andrew J. Patton, Advisor

_____
Tim Bollerslev

_____
Federico Bugni

_____
Jia Li

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Department of Economics
in the Graduate School of Duke University
2020

Abstract

Econometric Methods for Expected Shortfall and Value-at-Risk

by

Peter Horvath

Department of Department of Economics
Duke University

Date: _____

Approved:

_____
Andrew J. Patton, Advisor

_____
Tim Bollerslev

_____
Federico Bugni

_____
Jia Li

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Department of Economics in the Graduate School of Duke University
2020

Copyright © 2020 by Peter Horvath
All rights reserved except the rights granted by the
Creative Commons Attribution-Noncommercial Licence

Abstract

Value-at-Risk (VaR) has been the most prevalent market risk measure in the financial sector. Banks, insurance companies and other financial institutions are required to report their VaR estimates to the regulatory authorities since its introduction to the Basel I Accord in 1996. Acknowledging the theoretical deficiencies of this risk measure, The Basel Committee on Banking Supervision proposed to replace VaR with Expected Shortfall (ES), which overcomes these shortcomings. The practical implementation of this measure is still in process as the literature is lack of simple tools for its estimation and evaluation since by definition, the ES depends on the VaR estimate. This dissertation develops several techniques for estimating and conducting inference on VaR and ES models.

The first chapter, which is a joint paper with Andrew J. Patton, implements a 2-step robust estimation method for estimating the Expected Shortfall. We ease the dependence of the ES estimate from the VaR. To achieve this, in the first step the VaR is estimated by nonparametric methods, which helps us to avoid estimation error in the nuisance process. In the second step, we apply a robust estimation technique which controls for small deviations of the VaR estimates from their theoretically true values. We compare this new method to a 1-step joint estimation when VaR and ES are jointly estimated and to a 2-step non-robust estimation method. We find that with the new 2-step method the estimates are more efficient than applying a non-robust version and it performs better than the joint estimation when we do inference

on the ES model parameters.

The second chapter, which is joint with Jia Li, Zhipeng Liao and Andrew J. Patton, proposes a novel nonparametric specification test for VaR models. We translate the specification test to a conditional moment restriction test. We estimate the conditional moment function via nonparametric series regression and test whether it is identically zero. We use a strong Gaussian approximation theory to characterize the asymptotic behavior of our sup-t test. In addition, we propose an i.i.d. bootstrap method which performs better in finite sample than the asymptotic approximation at more extreme quantiles. As an empirical exercise, we test Conditional VaR models of US financial institutions.

The third chapter builds on this idea and implements the test for multiple conditional moment restrictions to test the correct specification of VaR and ES jointly. In addition, we also propose an average-t test, whose theoretical properties rely on the sup-t test. We find that the average-t test outperforms the sup-t test at more extreme confidence levels. As an empirical exercise, we test several location-scale models in S&P500 data for correct specification of ES and VaR.

Contents

List of Tables

List of Abbreviations and Symbols

Symbols

$\mathbb{E}$	Expectation operator
$\mathbb{P}$	The physical probability measure
$\mathbb{Q}$	The risk-neutral probability measure

Abbreviations

i.i.d.	Independent and identically distributed
GMM	General method of moments
VaR	Value-at-Risk
ES	Expected Shortfall
DGP	Data Generating Process
VIX	Volatility Index
cdf	Cumulative Distribution Function
pdf	Probability Density Function

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my research supervisor, Andrew J. Patton, for providing me the opportunity to conduct research and guiding me through my graduate education. Without his persistence help this dissertation would not have been possible. I am also highly indebted to the members of my committee, Tim Bollerslev, Federico Bugni, Jia Li for their invaluable advice and their continued encouragement throughout my research process. I would also like to express my most sincere gratitude to Peter Reinhard Hansen and George Tauchen for their thought-provoking questions and ideas in the financial econometrics lunch group.

I am grateful to the Department of Economics and the Graduate School for providing me funding to make my graduate education possible. I would also like to thank all the support staff at our department whose relentless help truly made my life better in the past 5 years. I am also thankful for the faculty members who have taught me for the support and education I have received.

In addition, I am grateful to my colleagues and friends at Duke: Scott Abrahams, Junaid Arefeen, Amelia Reid Bell, Federico Bennett, Jackson J. Bunting, Missy Cazer Daffron, Kelsey Evezich, Craig Fratrik, Taylor Michelle Gastineau, Sharon Kim, Andrea Kiss, Ilia Kozis, Olga Kozlova, Margaux Luflade, Jonathan Ronny Moreno Medina, David Min, Linh Nguyen, Gabor Palinko, Leonardo Salim Saker Chaves, Guilherme Salome, Gleb Sinev, Mirjam Szillery, Eugene Tan, Vytautas Valaitis,

Alessandro Tenzin Villa, Andrew Vollmer, Maria Zhu for their help, support and friendship throughout my time at Duke.

Finally, I would like to thank my parents, and my brother for their encouragement and support throughout my time in graduate school.

1

Introduction

One of the most well-known risk measures in the financial industry is the Value-at-Risk (VaR). VaR is a lower (1%-5%) quantile of the return of a portfolio (or a high (95%-99%) quantile of the losses of the same portfolio). Financial institutions are required to report their 1%-VaR estimates to regulatory authorities by the Basel I Accord. Since a quantile is only a point on the distribution, it is not able to capture tail risk and as Artzner et al. (1999) show VaR is also not a coherent risk measures. Acknowledging these deficiencies, The Basel Committee on Banking Supervision proposed to replace the VaR with the Expected Shortfall (ES) (Basel Committee, 2013).

ES is the mean of a return conditional the return being lower of its VaR value. By the definition of this risk measures, it is able to capture the tail risk, and it can be shown that it is also a coherent risk measure. However, the main theoretical shortcoming of this measure, as Fissler and Ziegel (2016) show, it can only be estimated jointly with VaR.

In the first chapter, which is a joint project with Andrew J. Patton, we implement a 2-step locally robust semiparametric GMM estimator, which is based on orthogonal

moment conditions, to ease the dependence of the ES estimator from the VaR. The moment conditions have zero derivative with respect to the first step (in our case it is the VaR estimate), therefore the first step does not affect the asymptotic variance of the ES estimates (see e.g. Ackerberg et al., 2014). Moreover, Chernozhukov et al. (2016) show locally robust moment conditions have small bias property, which leads to better small sample properties. We compare the finite sample properties of the 2-step robust estimation method with a 2-step non-robust and 1-step joint estimation method. As an empirical exercise we estimate the ES of the S&P500 returns.

In the second chapter, which is joint with Jia Li, Zhipeng Liao and Andrew J. Patton, we propose a novel nonparametric specification test for VaR models. In this chapter, my work focuses on the Monte Carlo experiment and the empirical application of the specification test. The theoretical contribution of the chapter is developed by my co-authors. Following Li and Liao (2019), we translate the VaR specification test to testing conditional moment restrictions. To conduct our test, we estimate the conditional moment function via nonparametric series regression and test whether it is identically zero. To characterize the asymptotic behavior of the our sup-t test, we use a strong Gaussian approximation theory. In addition, we propose an i.i.d. bootstrap technique to compute the critical values and examine the finite sample properties of our method in a Monte Carlo simulation study. We test the correct specification of CoVaR models from Tobias and Brunnermeier (2016) for US financial institutions.

In the third chapter, we implement this nonparametric specification test for testing multiple moment conditions jointly. Specifically, we test the correct specification of VaR and ES jointly. As mentioned before, ES depends on the VaR which necessitates the joint testing of these two risk measures. In this chapter, we propose an average-t test based on the theoretical characteristics of the sup-t test. In a Monte Carlo experiment, we compare the finite sample properties of the proposed average-t

test to the sup-t test. We test several location-scale in S&P500 data for correct specification of VaR and ES.

2

Locally Robust Estimation of Expected Shortfall

2.1 Introduction

Value-at-Risk (VaR) has been the main market risk measure for financial institutions. It can be interpreted as a (low) quantile of the distribution of the return, which can be estimated and predicted by numerous methods (see e.g. Koenker and Hallock, 2001, a widely cited article on quantile regressions and the references therein). However, in spite of its simplicity to measure, it has several drawbacks as Artzner et al. (1999) point out. For example, it does not satisfy the sub–additive property, which can have serious consequences when risk is evaluated. Danielsson et al. (2005) argue that if sub-additivity is violated, financial institutions might be under–hedged as they do not recognize the actual level of risk or the lack of sub-additivity can also lead to suboptimal investment choices if VaR is the main measure for risk. Moreover, VaR is not able to capture tail risk as it does not account for the return distribution below the predetermined confidence level. Acknowledging these drawbacks, the Basel Committee proposed to replace the VaR with Expected Shortfall (ES) as a part of the Third Basel Accords (Basel Committee, 2013) .

ES, the average return conditional on the return being below VaR, has been proposed by Artzner et al. (1999), who show that ES satisfy the sub-additive property and contrary to VaR, it is a coherent risk measure. That is, if Y_t is the return on some portfolio with conditional distribution $F(Y_t|\mathcal{F}_{t-1})$, where $\mathcal{F}_{t-1}$ is the information available until period t, then $q \in (0,1)$ level $VaR_{t,q}$ and $ES_{t,q}$ can be defined as:

$$VaR_{t,q} = F^{-1}(q|\mathcal{F}_{t-1})$$
$$ES_{t,q} = \mathbb{E}\left[Y_t|Y_t \leq VaR_{t,q}, \mathcal{F}_{t-1}\right],$$

where we assume that Y_t has finite mean and $F(Y_t|\mathcal{F}_{t-1})$ is strictly increasing and therefore invertible. As we can see from the definition of the ES, it is strongly related to the VaR. Fissler and Ziegel (2016) show that ES is not elicitible, that is, there does not exist any loss function which minimized would result in ES. However, they show that VaR and ES are jointly elicitible and provide a loss function family which minimized jointly with respect to both risk measures result in a consistent ES (and VaR) estimate.

According to the state of the art, it is not possible to estimate ES separately from VaR, that is, the ES estimates depend on the VaR estimate. In this paper, our main goal is to increase the robustness of ES parameter estimates with respect to the VaR estimate. To achieve this goal, as first step we estimate the VaR using non–parametric a sieve polynomial, which means in theory the VaR estimates would not suffer from misspecification error. In the second step using the estimated value of the VaR, we use a multivariate loss function from Fissler and Ziegel (2016), which can be applied to estimate VaR and ES jointly, to estimate the parameters of ES. That is, if we want to estimate VaR and ES jointly, in one step we would minimize one of these parametric multivariate loss functions with respect to VaR and ES and then get the estimated values for both VaR and ES. What we do is: in the first step, we estimate VaR non–parametrically and then in the second step we can plug in this

estimated value of VaR to the loss function and then just minimize with respect to ES to get an estimate only for ES since we already have an estimated VaR process.

To implement our estimation strategy, we use locally robust semiparametric GMM, which is based on orthogonal moment conditions. That is, the moment conditions have zero derivative with respect to the first step (in our case the VaR estimate), therefore the first step does not affect the asymptotic variance (see e.g. Ackerberg et al., 2014). Moreover, Chernozhukov et al. (2016) show locally robust moment conditions have small bias property, which leads to better small sample properties.

The local robustness property makes this estimation method sensible for our problem. Local robustness with respect to the first step means that the value of the moment conditions in the second step does not change if the estimated process from the first step is arbitrarily close to the true VaR process even if it is not equal to the true process. That is, even if we cannot exactly estimate the true VaR but get a close estimate, the estimation of the ES should not be affected by the VaR estimate.

To examine the finite simple properties of the 2-step locally robust GMM method, we use an AR(1)-TS-ARCH(1) and HAR(22) type of models as data generating processes and compare the accuracy of the parameter estimates using the robust GMM method with a non-robust 2-step method and with the 1-step joint method.

Our second objective in this paper is to implement option implied VaR and ES on S&P500 index. The forward-looking nature of the options motivate Barone Adesi (2016) and Mitra (2015) to derive the relation between European put options, VaR and ES. In our empirical study, we apply the locally robust estimation method to investigate which financial measures, such as past absolute returns, range, realized volatility and VIX, have better explanatory power in ES estimation.

The remainder of the paper is structured as follows. In Section 2.2, we provide a short literature review. In Section 2.3, we present the 2-step locally robust GMM estimation method and derive the asymptotic properties of this estimator. In Section

2.4, we examine the finite sample properties of the robust estimation method. In Section 2.5, we implement the option implied VaR/ES and apply the robust estimation method for estimating S&P500 ES using the above mentioned measures and in Section 2.6 we conclude.

2.2 Literature review

This section present the literature review of locally robust 2-step estimators, sieve estimation with dependent data and expected shortfall estimation and prediction.

Orthogonalized/locally robust moment conditions have been used in post-model selection inference in e.g. Belloni et al. (2014). They present robust methods for inference about the effect of treatment variables (low dimensional variables) in models where the number of regressors can be larger than the sample size.

Newey (1994) provides a general form of the adjustment term in the robust moment conditions for first step non-parametric estimators in i.i.d. setting and shows that under regularity conditions the asymptotic variance of semiparametric estimators does not depend on the type of the nonparametric estimate but only on the function which is nonparametrically estimated. Chen et al. (2003) extends this work and provide sufficient conditions, for both i.i.d. and dependent, heterogeneous data, for the consistency and asymptotic normality for 2-step semiparametric estimators, where the first step is infinite dimensional nuisance process.

Chernozhukov et al. (2015) describe a GMM extension of the orthogonality condition in general settings and provide high level assumptions under which valid inference on low dimensional parameters can be conducted in the presence of high dimensional parameters. Chernozhukov et al. (2016) provides a general formula for locally robust/orthogonalized moment condition construction and show that these moment conditions have small bias property with respect to the infinite dimensional first step. Ackerberg et al. (2014) show that these estimators can achieve the semi-

parametric efficiency bound.

A general review of sieve estimators, and their large sample properties can be found in Chen (2007).

Chen and Shen (1998) provide theory on the convergence rate of sieve extremum estimators with stationary β-mixing observations. For strictly stationary ergodic data, Chen and Pouzo (2012) introduce the penalized sieve minimum distance estimation (PSMD) method for estimating the infinite dimensional nuance process. PSMD nests the minimum distance and therefore GMM estimation if the penalization term is 0. Under regularity conditions, they derive the consistency of the estimator and show convergence rate results under Banach and Hilbert norm.

In a sieve semiparametric two-step GMM framework with weakly dependent data, Chen and Liao (2015) show that even if the asymptotic variance matrix of the second-step GMM estimator does not have a closed form solution, we can conduct inference with parametric asymptotic variance matrix. That is, we can conduct inference on the finite dimensional parameters of interest as if the infinite dimensional nuisance function, which is consistently estimated in the first step with some sieve extremum estimator, were estimated parametrically.

There has been some work in the literature to estimate and predict ES. Cai and Wang (2008) propose nonparametric estimate of VaR and ES, which is implement by plugging-in method. Nonparametric estimation method frees their estimator from misspecification error, however, the convergence rate is slower than the parametric, $\sqrt{n}$, rate. Wang and Zhao (2016) investigate a semiparametric estimator of ES. They do not specify the distribution of the noise, which leads to semiparametric models; however, they specify a parametric function of the observables. They also show that their proposed estimator can reach root-n convergence rate.

Fissler and Ziegel (2016) show that ES and VaR are jointly elicitable and provide a class of loss functions which minimized jointly with respect to VaR and ES results

in consistent estimates for both risk measures. Building on their work, Dimitriadis and Bayer (2017) estimates jointly the VaR and ES using i.i.d. data with linear specification. In their paper, they mainly focus on M estimation since they find that the GMM is unstable in estimating the two risk measures jointly. Drawing on the proposed loss function family, Patton et al. (2019) propose new dynamic semiparametric models for ES and VaR jointly, in which they impose the parametric structure for the dynamics of ES and VaR but they do not specify the conditional distributions of returns. In their article, they also provide asymptotic theory for these models, which neither require i.i.d. data nor linear specification of the models.

2.3 Theory

In this section, we present the locally robust semiparametric estimation method (see for example Ackerberg et al. (2014) (henceforth ACHL) or Chernozhukov et al. (2016) (henceforth CEIN)) for our problem. Given a sample of (returns) $Y_t \in \{Y_1, \ldots, Y_T\}$, we want to identify the parameters θ_e^0 of Expected Shortfall of Y_t.

Suppose Y_t is a continuous real-valued random variable with existing conditional distribution function $F(Y_t|\mathcal{F}_{t-1})$ and corresponding density function $f(Y_t|\mathcal{F}_{t-1})$, where $\mathcal{F}_{t-1} = \sigma(Y_{t-1}, X_{t-1}, \ldots, Y_1, X_1)$ is the information set available at time t and X_t is a vector of exogenous or predetermined variables similarly as in Patton et al. (2019).

The conditional VaR and ES of Y_t at confidence level q can be written as

$$VaR_{t,q}(Y_t|\mathcal{F}_{t-1}) = v_t^0$$
$$ES_{t,q}(Y_t|\mathcal{F}_{t-1}) = e_t(\theta_e^0),$$

where we purposefully add the parameter vector for $ES_{t,q}$ to indicate that those are the parameters of interests.

Fissler and Ziegel (2016) show there exists a class of loss function which is con-

sistent for VaR and ES. Drawing on their results, we have moment conditions

$$\mathbb{E}\left[m(Y_t, e_t(\theta_e^0), v_t^0)\right] = 0,$$

which $e_t(\theta_e^0)$ and v_t^0 satisfy uniquely. Therefore, under appropriate assumptions, θ_e^0 can be identified from the correct conditions.

Following Patton et al. (2019), we use the loss function from this family which generates loss differences that are homogeneous of degree zero. As they show assuming $ES_{t,q} < 0$ and $VaR_{t,q} < 0$ a.s. $\forall t$,[1] there exists only one loss function (up to irrelevant location and scale factors) which satisfies the zero degree homogeneity in loss difference, which they name FZ0 loss function:

$$L_{FZ0}(Y_t, e_t(\theta_e), v_t; q) = -\frac{1}{q e_t(\theta_e)} \mathbb{1}\{Y_t \leq v_t\} (v_t - Y_t) + \frac{v_t}{e_t(\theta_e)} + \log(e_t(\theta_e)) - 1. \quad (2.1)$$

Assuming that $\hat{v}_t$ a consistent first step estimate of $VaR_{t,q}$, we can estimate θ_e^0 using GMM estimator with some $\widehat{W}$ positive semidefinite matrix by plugging in $\hat{v}_t$ to the moment conditions:

$$\hat{\theta}_e = \underset{\theta_e}{\operatorname{argmin}}\, m(Y_t, e_t(\theta_e), \hat{v}_t)' \widehat{W}\, m(Y_t, e_t(\theta_e), \hat{v}_t).$$

Following CEIN (p. 6) Definition 1, we can rewrite their definition of locally robustness to our case as:

Definition 1. *The moment functions m(Y, e(θ_e), v) are locally robust if and only if* $v \in \mathcal{V} - v^0$

$$\left.\frac{\partial \mathbb{E}\left[m(Y, e(\theta_e^0), (1-\tau)v^0 + \tau v)\right]}{\partial \tau}\right|_{\tau=0} = 0.$$

Here we used the pathwise derivative condition (see e.g. ACHL p. 922) directly in our definition.

[1] In practice the values of q are ranging from 0.01 and 0.1, therefore assuming negativity of $VaR_{t,q}$ and therefore $ES_{t,q}$, is reasonable.

In 2–step estimation procedure, we can construct locally robust moment conditions by adding an adjustment term, $\phi(\cdot)$ to the original moment conditions which accounts for the first step estimation. That is,

$$T^{-1/2}\sum_{t=1}^{T} m(Y_t, e_t(\theta_e^0), \hat{v}_t) = T^{-1/2}\sum_{t=1}^{T}\left[m(Y_t, e_t(\theta_e^0), v_t^0) + \phi(Y_t, e_t(\theta_e^0), v_t^0)\right] + o_p(1),$$

which is equivalent to equation (2.2) in CEIN.

2.3.1 2–step locally robust GMM estimation

Following ACHL and CEIN, we can derive the 2–step GMM estimator, where we estimate $VaR_{t,q}$ process in the first step using a sieve polynomial. In the second step, we plug in the first step estimate $\hat{v}_t$ into the locally robust moment conditions to estimate the parameters of $ES_{t,q}$, θ_e.

First step: Estimating v_t^0

Since Value–at–Risk at q level is a quantile, we can use the check loss function (Koenker and Bassett Jr, 1978) to estimate the v_t^0 or in our case the parameters in the polynomial, θ_v^0 parameters with GMM.

That is, using

$$\rho(Y_t, v_t(\theta_v); q) = (\mathbb{1}\{Y_t \leq v_t(\theta_v)\} - q)(v_t(\theta_v) - Y_t), \tag{2.2}$$

we can get the following moment conditions

$$\mathbb{E}\left[\frac{\partial \rho(Y_t, v_t(\theta_v^0); q)}{\partial \theta_v}\right] = \mathbb{E}\left[h(Y_t, v_t\left(\theta_v^0\right))\right] = \mathbb{E}\left[\left(\mathbb{1}\{Y_t \leq v_t(\theta_v^0)\} - q\right)\frac{\partial v_t\left(\theta_v^0\right)}{\partial \theta_v}\right] = 0. \tag{2.3}$$

Note that check loss function $\rho(Y_t, v_t(\theta_v); q)$ is not differentiable at $Y_t = v_t(\theta_v)$ point; however, $\mathbb{P}(Y_t = v_t(\theta_v)) = 0$ under the assumption that Y_t has a continuous distribution.

Now using the above moment conditions, we can arrive to the familiar GMM estimator of the parameters θ_v:

$$\hat{\theta}_v = \underset{\theta_v \in \widetilde{\theta}_v}{\operatorname{argmin}} \left(T^{-1} \sum_{t=1}^{T} h(Y_t, v_t(\theta_v)) \right)' \widehat{W}_v \left(T^{-1} \sum_{t=1}^{T} h(Y_t, v_t(\theta_v)) \right),$$

where $\dim(\widetilde{\theta}_v) < \infty$ and $\widehat{W}_v \xrightarrow{p} W$, positive semidefinite matrix, which in our exactly identified case, is the identity matrix. Then plugging in the estimated parameters to our predefined sieve polynomial, we can get the $\hat{v}_t$ process.

Second step: Estimating θ_e^0, the parameters of interest

After we have estimated the $VaR_{t,q}$ process, we can estimate the parameters of interest, θ_e^0 by GMM:

$$\hat{\theta}_e = \underset{\theta_e}{\operatorname{argmin}} \left(T^{-1} \sum_{t=1}^{T} g(Y_t, e_t(\theta_e), \hat{v}_t) \right)' \widehat{W} \left(T^{-1} \sum_{t=1}^{T} g(Y_t, e_t(\theta_e), \hat{v}_t) \right),$$

where

$$g(Y_t, e_t(\theta_e), v_t) = m(Y_t, e_t(\theta_e), v_t) + \phi(Y_t, e_t(\theta_e), v_t), \tag{2.4}$$

where $m(\cdot)$ are moment conditions from the L_{FZ0} loss function and $\phi(\cdot)$, the adjustment term, are defined in the following.

Definition of $m(Y_t, e_t(\theta_e), v_t)$

The function $m(\cdot)$ can be interpreted as moment conditions without adjustment term. That is, $e_t(\theta_e^0)$ and v_t^0 uniquely satisfy the following moment conditions:

$$\mathbb{E}\left[m(Y_t, e_t(\theta_e^0), v_t) \right] = 0.$$

Drawing on the results of Fissler and Ziegel (2016), we know L_{FZ0} loss function (Equation (2.1)) is minimized at the true VaR and ES value. Therefore we can use the first order conditions with respect to θ_e to get the appropriate moment conditions:

$$m(Y_t, e_t(\theta_e), v_t) = \left(\frac{1}{q e_t(\theta_e)^2} \mathbb{1}\{Y_t \leq v_t\} (v_t - Y_t) - \frac{v_t}{e_t(\theta_e)^2} + \frac{1}{e_t(\theta_e)} \right) \frac{\partial e_t(\theta_e)}{\partial \theta_e}. \quad (2.5)$$

Definition of the adjustment term, $\phi(Y_t, e_t(\theta_e), v_t)$

The function $\phi(\cdot)$ is the adjustment term which accounts for the first step estimation error and makes the $g(\cdot)$ moment conditions be locally robust.

In our paper, we build on ACHL, who define the locally robust moment conditions for a case where the first step is a non-parametric quantile regression (p. 926):

$$\phi(Y_t, e_t(\theta_e), v_t) = \left(\frac{1}{T} \sum_{k=1}^{T} \frac{1}{q e_k(\theta_e)^2} (\mathbb{1}(Y_k \leq v_k) - q) \frac{\partial e_t(\theta_e)}{\partial \theta_e} \right) \cdot \left(\frac{1}{T} \sum_{k=1}^{T} f_{Y_k}(v_k) \right)^{-1} (q - \mathbb{1}(Y_t \leq v_t)). \quad (2.6)$$

We can understand the intuition behind the above adjustment term by following ACHL (p. 928) or CEIN (pp. 6–7). Let us consider a parametric case, where $v_t(\theta_v)$ is a function of finite dimensional vector of parameters θ_v^0, which satisfies the following moment condition

$$\mathbb{E}\left[\gamma(Y_t, \theta_v^0)\right] = \mathbb{E}\left[\mathbb{1}\{Y_t \leq v_t(\theta_v^0)\} - q\right] = 0. \quad (2.7)$$

Then it can be shown (see e.g. ACHL p. 928), that

$$T^{-1/2} \sum_{t=1}^{T} m\left(Y_t, e_t(\theta_e^0), v_t(\hat{\theta}_v)\right) =$$

$$T^{-1/2} \sum_{t=1}^{T} m(Y_t, e_t(\theta_e^0), v_t(\theta_v^0))$$

$$- T^{-1/2} \sum_{t=1}^{T} \left(\frac{\partial \mathbb{E}\left[m\left(Y_t, e_t(\theta_e^0), v_t(\theta_v^0)\right)\right]}{\partial \theta_v} \right) \left(\frac{\partial \mathbb{E}\left[\gamma(Y_t, \theta_v^0)\right]}{\partial \theta_v} \right)^{-1} \rho(Y_t, \theta_v^0)$$

$$+ o_p(1).$$

2.3.2 Asymptotic Theory for Locally Robust Moments

In this section we outline the theorems following ACHL and CEIN which are necessary to do inference on the parameters of ES.

Following CEIN, to show that the moment conditions $g(Y_t, e_t(\theta_e), v_t)$, defined as

$$\begin{aligned} g(Y_t, e_t(\theta_e), v_t) = &\left(\frac{1}{q e_t(\theta_e)^2} \mathbb{1}\left(Y_t \leq v_t\right)\left(v_t - Y_t\right) - \frac{v_t}{e_t(\theta_e)^2} + \frac{1}{e(\theta_e)} \right) \frac{\partial e_t(\theta_e)}{\partial \theta_e} \\ &+ \left(\frac{1}{T} \sum_{k=1}^{T} \frac{1}{q e_k(\theta_e)^2} \left(\mathbb{1}\left(Y_k \leq v_k\right) - q\right) \frac{\partial e_t(\theta_e)}{\partial \theta_e} \right) \left(\frac{1}{T} \sum_{k=1}^{T} f_{Y_k}(v_k) \right)^{-1} \\ &\cdot \left(q - \mathbb{1}\left(Y_t \leq v_t\right)\right), \end{aligned}$$

are locally robust, it is sufficient to show that the adjustment term for the first step is zero, that is,

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} g(Y_t, e_t(\theta_e^0), \hat{v}_t) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} g(Y_t, e_t(\theta_e^0), v_t^0) + o_p(1).$$

In the above equations, Y_t is the return/dependent data, v_t is the Value-at-Risk which is estimated non-parametrically, $e_t(\theta_e^0)$ is the Expected Shortfall, θ_e is a vector of parameters of interest, q is the confidence level (5%), and the zero subscripts denote the true value(s).

Assumption 1. *(i) Θ_e is a compact subset of $\mathbb{R}^p$ for $p < \infty$.*

(ii) $\{Y_t\}_{t=1}^{\infty}$ is a strictly stationary and ergodic process. The conditional (on all past information $\mathcal{F}_{t-1}$) distribution of Y_t is $F_t(\cdot|\mathcal{F}_{t-1})$, which $\forall t \geq 1$ is continuously differentiable and belongs to a class of distribution functions on $\mathbb{R}$ with finite first moments and unique q-quantiles.

(iii) $\forall t \geq 1$, $e_t(\theta_e)$ is $\mathcal{F}_{t-1}$ measurable and continuously differentiable in θ_e.

(iv) $W_T \xrightarrow{p} W_0$, where W_0 is a positive semidefinite matrix.

(v) *If* $\forall t \geq 1$, $\mathbb{P}\left[e_t(\theta_e) = e_t(\theta_e^0)\right] = 1$, *then* $\theta_e = \theta_e^0$.

(vi) $\mathbb{E}\left[|Y_t|\right] < \infty$, $\forall t \geq 1 \quad |v_t| < \infty$, $|1/e_t(\theta_e)| < \infty$, $\left|\left|\frac{\partial e(\theta_e)}{\partial \theta_e}\right|\right| < \infty$.

Theorem 1. *If Assumption 1 is satisfied,* $\hat{\theta}_{e,T} \xrightarrow{p} \theta_e^0$ *as* $T \to \infty$.

In the following assumption 2, we follow Theorem 2 of Chen et al. (2003). First, we define $\theta_\delta \equiv \{\theta_e \in \theta : ||\theta_e - \theta_e^0|| \leq \delta\}$ and $\mathcal{V}_\delta = \{v_t \in \mathcal{V} : |v_t - v_t^0| \leq \delta\}$ for some small $\delta > 0$. We also follow their notion of functional derivative. For any $\theta_e \in \theta_\delta$, $\mathbb{E}\left[g(Y_t, \theta_e, v)\right]$ is pathwise differentiable at $v_t \in \mathcal{V}_\delta$ in the direction $[\bar{v}_t - v_t]$ if $\{v_t + \tau(\bar{v}_t - v_t) : \tau \in [0,1]\} \in \mathcal{V}$ and $\lim_{\tau \to 0} \mathbb{E}\left[g(Y_t, \theta_e, v_t + \tau(\bar{v}_t - v_t)) - g(Y_t, \theta_e, v_t)\right]/\tau$ exists and it is denoted as $\frac{\partial \mathbb{E}[g(Y_t, e_t(\theta_e), v_t)]}{\partial v_t}[\bar{v}_t - v_t]$.

Assumption 2. *(vii)* $\left|\left|T^{-1}\sum_{t=1}^{T} g(Y_t, e_t(\hat{\theta}_e), \hat{v}_t)\right|\right| = \inf_{\theta_e \in \theta_\delta}\left|\left|T^{-1}\sum_{t=1}^{T} g(Y_t, e_t(\theta_e), \hat{v}_t)\right|\right| + o_p(n^{-1/2})$

(viii) *The matrix* $\frac{\partial \mathbb{E}\left[g(Y_t, e_t(\theta_e^0), v_t^0)\right]}{\partial \theta_e}$ *is of full rank.*

(ix) *For all* $\theta_e \in \theta_\delta$, *the pathwise derivate derivative* $\frac{\partial \mathbb{E}\left[g(Y_t, e_t(\theta_e^0), v_t^0)\right]}{\partial v_t}[v_t - v_t^0]$ *exists in all directions* $[v_t - v_t^0] \in \mathcal{V}$

(x) *For all* $(\theta_e, v_t) \in \theta_{\delta_n} \times \mathcal{V}_{\delta_n}$ *with a positive sequence* $\delta_n = o(1)$*:*

$$\left|\left|\mathbb{E}\left[g(Y_t, e_t(\theta_e), v_t) - g(Y_t, e_t(\theta_e), v_t^0)\right] - \frac{\partial \mathbb{E}\left[g(Y_t, e_t(\theta_e), v_t^0)\right]}{\partial v_t}[v_t - v_t^0]\right|\right| \leq c\,|v_t - v_t^0|^2$$

for a constant $c \geq 0$

and

$$\left|\left|\frac{\partial \mathbb{E}\left[g(Y_t, e_t(\theta_e), v_t^0)\right]}{\partial v_t}[v_t - v_t^0] - \frac{\partial \mathbb{E}\left[g(Y_t, e_t(\theta_e^0), v_t^0)\right]}{\partial v_t}[v_t - v_t^0]\right|\right| \leq \delta_n$$

(xi) $v_t \in \mathcal{V}$ *with probability 1 and* $|v_t - v_t^0| = o_p(n^{-1/4})$

(xii) With a positive sequence $\delta_n = o(1)$,

$$\sup_{||\theta_e - \theta_e^0|| \leq \delta_n, |v_t - v_t^0| \leq \delta_n} \frac{\sqrt{T}||\bar{g}(y_t, e_t(\theta_e), v_t) - \mathbb{E}\left[g(Y_t, e_t(\theta_e), v_t)\right] - \mathbb{E}\left[g(Y_t, e_t(\theta_e^0), v_t^0)\right]||}{1 + \sqrt{T}\left[||\bar{g}(y_t, e_t(\theta_e), v_t)|| + ||\mathbb{E}\left[g(Y_t, e_t(\theta_e), v_t)\right]||\right]}$$

$$= o_p(1),$$

where $\bar{g}(y_t, e_t(\theta_e), v_t) = T^{-1}\sum_{t=1}^{T} g(Y_t, e_t(\theta_e), v_t)$.

Lemma 1. *If Assumption 1, 2 hold, then*

$$T^{-1/2}\sum_{t=1}^{T} g(Y_t, e_t(\theta_e^0), \hat{v}_t) = T^{-1/2}\sum_{t=1}^{T} g(Y_t, e_t(\theta_e^0), v_t^0) + o_p(1).$$

Assumption 3. *There is a neighborhood* $\mathcal{N}$ *of* θ_e^0 *such that* $\forall t \geq 1$

(xiii) in addition to assumption (iii), $e_t(\theta_e)$ *is two times continuously differentiable in* θ_e.

(xiv) there is a $\zeta > 0$ *and* $d(y_t)$ *with* $\mathbb{E}\left(d(y_t)\right) < \infty$ *such that for* $\theta_e \in \mathcal{N}$ *and* $|v_t - v_t^0|$ *small enough*

$$\left|\left|\frac{\partial g(y_t, e_t(\theta_e), v_t)}{\partial \theta_e} - \frac{\partial g(y_t, e_t(\theta_e^0), v_t^0)}{\partial \theta_e}\right|\right| \leq d(y_t)\left(||\theta_e - \theta_e^0||^{\zeta} + |v_t - v_t^0|^{\zeta}\right)$$

(xv) in addition to assumption (vi) $\left|\frac{\partial e(\theta_e)}{\partial^2 \theta_e}\right| < \infty$, $\left|\frac{\partial L_{FZ0}(\cdot)}{\partial^2 e_t(\theta_e)}\right| < \infty$.

(xvi) $v_t \xrightarrow{p} v_t^0$

(xvii) The deterministic positive sequence c_T *satisfies* $c_T = o_p(1)$ *and* $c_T^{-1} = o_p(T^{1/2})$.

The c_T sequence in our simulation is the bandwidth parameter sequence for density estimation and we set it as $T^{-1/3}$.

Lemma 2. *If Assumption 3 and (ii) are satisfied, then for any* $\bar{\theta}_e \xrightarrow{p} \theta_e^0$, $g(Y_t, e_t(\theta_e^0), \hat{v}_t)$ *differentiable at* $\bar{\theta}_e$ *with probability approaching one and*

$$\frac{\partial g(Y_t, e_t(\bar{\theta}_e), \hat{v}_t)}{\partial \theta_e} \xrightarrow{p} \mathbb{E}\left[\frac{\partial g(Y_t, e_t(\theta_e^0), v_t^0)}{\partial \theta_e}\right] = G$$

Assumption 4. $\forall t \geq 1$, *we have*

(xviii) $\mathbb{E}\left[g(Y_t, e_t(\theta_e^0), v_t^0)\; g(Y_t, e_t(\theta_e^0), v_t^0)'\right]$ *exists and finite.*

(xix) $\lim_{T\to\infty} Var\left[T^{-1/2}\sum_{t=1}^{T} g(Y_t, e_t(\theta_e^0), v_t^0)\right] = \Omega$ *exists and is a finite valued positive definite matrix.*

(xx) $G'WG$ *is nonsingular.*

Theorem 2. *If Assumption 1, 2, 3, 4 are satisfied, then*

$$\sqrt{T}(\hat{\theta}_e - \theta_e^0) \to N(0, V), \text{ where}$$
$$V = (G'WG)^{-1}G'W\Omega WG(G'WG)^{-1}.$$

2.4 Simulation study

In this section we investigate the finite sample properties of the 2-step robust GMM estimator. To analyze the performance of the this estimator, we compare the estimates from the robust estimator to the estimates from similar 2–step GMM estimator without adjustment term, and to the estimates from joint estimator, when the $VaR_{t,q}$ and $ES_{t,q}$ process are estimated jointly. We use a 4^{th} order polynomial to estimate non-parametrically the $VaR_{t,q}$ process and then we use correctly specified form of the $ES_{t,q}$ to estimate θ_e.

In the simulation study, we use 2 different data generating processes to examine the finite sample properties of the robust estimator.

First, we consider a AR(1)-TS-ARCH(1) type of model for the DGP:

$$Y_t = \phi Y_{t-1} + \sigma_t \eta_t$$

$$\sigma_t = \omega + \beta |Y_{t-1}|$$

$$\eta_t \overset{iid}{\sim} N(0,1).$$

We estimate $VaR_{t,q}$ using a fourth order polynomial in the return:

$$\hat{v}_t = \lambda_0 + \lambda_1 Y_{t-1} + \lambda_2 Y_{t-1}^2 + \lambda_3 Y_{t-1}^3 + \lambda_4 Y_{t-1}^4,$$

and we use correct specification for the $ES_{t,q}$ process. In this model, $\theta_e = \{\phi, \alpha, \beta\}$.[2]

Under this DGP, both $VaR_{t,q}$ and $ES_{t,q}$ processes are time varying and they are not proportional to each other like in any (G)ARCH model without AR terms, since by assumption of $\eta_t \overset{iid}{\sim} N(0,1)$ we know

$$VaR_{t,q} = \phi Y_{t-1} + \Phi^{-1}(q)\sigma_t$$

$$ES_{t,q} = \phi Y_{t-1} - \phi(\Phi^{-1}(q))/q\sigma_t.$$

Second, we consider an HAR(22) type of model to account for the long memory property of the volatility:

$$Y_t = \sigma_t \eta_t$$

$$\sigma_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 \frac{1}{4} \sum_{i=2}^{5} Y_{t-i} + \beta_3 \frac{1}{17} \sum_{i=6}^{22} Y_{t-i}$$

$$\eta_t \overset{iid}{\sim} N(0,1).$$

In this model the $VaR_{t,q}$ and $ES_{t,q}$ processes are proportional to each other; however, we do not exploit this property. We estimate the $VaR_{t,q}$ using the first four lags of the return process, that is,

$$\hat{v}_t = \lambda_0 + \lambda_1 Y_{t-1} + \lambda_2 Y_{t-2} + \lambda_3 Y_{t-3} + \lambda_4 Y_{t-4},$$

[2] We make a parametric assumption on the innovation terms so that we can estimate the true ω, β instead of $-\phi(\Phi^{-1}(q))/q\omega$ and $-\phi(\Phi^{-1}(q))/q\beta$. This assumption helps us to make better sanity checks regarding the estimates.

and we use correct specification for the $ES_{t,q}$ process. In this model, $\theta_e = \{\beta_0, \beta_1, \beta_2, \beta_3\}$.

Note that both the AR(1)-TS-ARCH(1) and the HAR(22) type of models are linear in the parameters of interests similarly to our empirical studies. To set the parameters, we estimate the AR(1)-TS-ARCH(1) and HAR(22) type of models on daily S&P500 percentage returns from 2/1/1993 to 2/16/2018, which is also used in the empirical study.

We compare the estimates from the robust estimator to the estimates of a joint GMM estimator and a 2–step GMM estimator without adjustment term. Fissler and Ziegel (2016) show that $VaR_{t,q}$ and $ES_{t,q}$ can be estimated jointly using e.g. the L_{FZ0} as in Patton et al. (2019). Drawing on their results, we use the moment conditions corresponding to $ES_{t,q}$ from the L_{FZ0} loss function:

$$L_{FZ0}(Y_t, v_t, e_t; q) = -\frac{1}{qe_t}\mathbb{1}\left(Y_t \leq v_t\right)(v_t - Y_t) + \frac{v_t}{e_t} + \log(-e_t) - 1,$$

where $e_t = ES_{t,q}$ and $v_t = VaR_{t,q}$. To ease comparison of the 3 estimation methods, we use moment conditions from checkloss function as additional moments.

When we estimate $VaR_{t,q}$ and $ES_{t,q}$ jointly, we use the following GMM estimator:

$$(\hat{\theta}_v, \hat{\theta}_e) = \underset{\boldsymbol{\theta_v}, \boldsymbol{\theta_e}}{\operatorname{argmin}} \left(\frac{1}{T}\sum_{t=1}^{T} l\left(Y_t, Y_{t-1}, e_t(\theta_e), v_t(\boldsymbol{\theta_v})\right)\right)' \widehat{W} \left(\frac{1}{T}\sum_{t=1}^{T} l'\left(Y_t, Y_{t-1}, e_t(\theta_e), v_t(\theta_v)\right)\right),$$

where $\widehat{W}$ is the identity matrix and

$$l\left(Y_t, e_t(\theta_e), v_t(\boldsymbol{\theta_v})\right) = \begin{bmatrix} h(Y_t, v_t(\theta_v)) \\ m(Y_t, e_t(\theta_e), v_t(\theta_v))) \end{bmatrix},$$

where $h(Y_t, v_t(\theta_v))$ is defined in Equation (2.3) and $m(Y_t, e_t(\theta_e), v_t(\theta_v))$ is defined in Equation (2.5).

In the 2–step GMM estimator without adjustment term, the first step is to estimate $VaR_{t,q}$ using GMM:

$$\hat{\theta}_v = \underset{\boldsymbol{\theta_v}}{\operatorname{argmin}} \left(\frac{1}{T}\sum_{t=1}^{T} h(Y_t, v_t(\theta_v))\right)' \widehat{W} \left(\frac{1}{T}\sum_{t=1}^{T} h(Y_t, v_t(\theta_v))\right),$$

where W is the identity matrix and $h(Y_t, v_t(\theta_v))$ is defined in Equation (2.3).

In the second step, we estimate θ_e by plugging in the estimated $VaR_{t,q}$ process to the following GMM estimator:

$$\hat{\theta}_e = \underset{\boldsymbol{\theta_v}, \boldsymbol{\theta_e}}{\operatorname{argmin}} \left(\frac{1}{T} \sum_{t=1}^{T} m(Y_t, e_t(\theta_e), v_t(\theta_v)) \right)' \widehat{W} \left(\frac{1}{T} \sum_{t=1}^{T} m'(Y_t, e_t(\theta_e), v_t(\theta_v)) \right),$$

where $m(Y_t, e_t(\theta_e), v_t(\theta_v))$ is defined in Equation (2.5).

The 2–step robust GMM estimator only differs from this estimator by adding an adjustment term in the second step moment functions, which accounts for the first step specification error (Ackerberg et al. (2014), Chernozhukov et al. (2016)). That is, the moment function in the second step takes the following form:

$$g(Y_t, e_t(\theta_e), v_t) = m(Y_t, e_t(\theta_e), v_t) + \phi(Y_t, e_t(\theta_e), v_t),$$

where $m(\cdot)$ is defined in equation (2.5) and the adjustment term is in equation (2.6).

In the simulation study,[3] as a benchmark case, we consider sample size of $T = 2500$ and confidence level $q = 0.025$, which corresponds to requirements of the Basel III Accords, and repeat all simulations 1000 times to estimate $VaR_{t,0.025}$ and $ES_{t,0.025}$. Table 2.1 includes the summary statistics for the estimated parameters from AR(1)-TS-ARCH(1) model for the benchmark case. The first row in each block presents the true parameter value, which we estimate. We define Joint RMSE as $\sqrt{\sum \text{Bias}_i^2 + \text{Standard Deviation}_i^2}$.

To investigate the inference properties of the different estimation methods in finite sample, we look at several measures. First, we look at the coverage probability corresponding to the 95% confidence level. Second, to illustrate the accuracy of the standard error estimates, which are necessary to run hypothesis tests in the empirical section, of the different methods we compute the "ratio of standard errors (s.e.)". First, we compute the standard deviation of the parameter estimates from the 1000

[3] Appendix includes additional information about the computations.

simulations. Technically this number should be very close to the standard error estimate using the finite sample formula of the asymptotic variance. Therefore, in each simulations, we also compute the standard errors using the finite sample formula of the asymptotic variance/standard deviation. In the row of "Ratio of s.e. (Mean)", we divide the standard deviation of the 1000 parameter estimates by the mean of the 1000 estimates of the standard deviation of the parameters. The row of the "Ratio of s.e. (Median)", we divide by the median, which is less sensitive for outlier values. In an ideal world these numbers would be very close to 1 as the two estimates should be very close to each other.

In the last row, we calculate mean of the RMSE of the ES process, which provide an additional information how well the different techniques estimate the underlying ES process.

The first block presents the measures for the estimates from robust GMM estimator, the second block belongs to the joint estimation, while the last block shows the results for the 2–step GMM estimator without the adjustment term.

As we can see in this table, the 2-step robust GMM is better than the 2-step GMM without the adjustment term both in the accuracy of the parameter estimates and the accuracy of the ES process estimate. In addition, our inference measures are comparable using either of these methods. Comparing the robust estimation method with the joint method, we can conclude that the accuracy of these two methods are similar. However, as we can see from the inference measures, the robust estimation method clearly dominates the joint method, which can be explained by the size of the underlying variance covariance matrix, which is much smaller in the 2-step method since it "ignores" the first step model, therefore it is computationally easier to calculate the standard errors and it is more accurate.

Table 2.1: Simulation results: AR(1)-TS-ARCH(1); T = 2500, $q = 0.025$

Robust GMM			
	ϕ	ω	β
True values:	-0.062	0.859	0.318
Bias	0.002	-0.012	0.000
Standard dev.	0.079	0.058	0.055
RMSE	0.079	0.060	0.055
Joint RMSE		0.113	
Coverage Probability	0.976	0.902	0.931
Ratio of s.e. (Mean)	0.840	1.223	1.029
Ratio of s.e. (Median)	0.852	1.275	1.057
ES RMSE		0.148	

Joint GMM			
	ϕ	ω	β
True values:	-0.062	0.859	0.318
Bias	0.000	0.031	-0.012
Standard dev.	0.044	0.088	0.060
RMSE	0.044	0.094	0.061
Joint RMSE		0.120	
Coverage Probability	0.997	0.980	0.960
Ratio of s.e. (Mean)	0.000	0.000	0.000
Ratio of s.e. (Median)	0.384	1.313	0.643
ES RMSE		0.141	

2-step GMM			
	ϕ	ω	β
True values:	-0.062	0.859	0.318
Bias	0.002	0.035	0.003
Standard dev.	0.093	0.068	0.063
RMSE	0.093	0.077	0.063
Joint RMSE		0.136	
Coverage Probability	0.964	0.874	0.905
Ratio of s.e. (Mean)	0.959	1.327	1.096
Ratio of s.e. (Median)	0.977	1.408	1.125
ES RMSE		0.192	

We also investigate how these 3 estimators perform under different confidence levels ($q = \{0.01, 0.025, 0.05, 0.1\}$) and different time horizons ($T = \{500, 1000, 2500, 5000\}$). In Table 2.2, we present the ratio of the RMSEs. If the ratio is larger than 1, it means that the RMSE from the robust estimator is larger than either from the joint or the 2-step estimator. As we can see from this table, the robust estimator is generally better than the 2-step estimator without the adjustment term and it performs better in terms of RMSE for medium sample sizes ($T = \{1000, 2500\}$) at low confidence levels ($q = \{0.01, 0.025\}$).

Table 2.2: Comparison of Joint and ES RMSE in AR(1)-TS-ARCH(1) type of model

Joint RMSE

	Robust vs Joint T					Robust vs 2-step T			
q	500	1000	2500	5000	q	500	1000	2500	5000
0.010	1.150	0.801	0.815	1.080	0.010	0.887	0.760	0.795	0.812
0.025	1.034	0.911	0.944	1.039	0.025	0.762	0.814	0.834	0.860
0.050	1.044	1.022	1.216	1.107	0.050	0.717	0.823	0.828	0.875
0.100	1.085	1.145	1.059	1.012	0.100	0.755	0.779	0.795	0.859

ES RMSE

	Robust vs Joint T					Robust vs 2-step T			
q	500	1000	2500	5000	q	500	1000	2500	5000
0.010	1.043	0.834	0.899	1.023	0.010	0.851	0.745	0.744	0.715
0.025	1.026	0.974	1.053	1.088	0.025	0.701	0.771	0.774	0.768
0.050	1.086	1.109	1.183	1.102	0.050	0.656	0.814	0.799	0.820
0.100	1.102	1.130	1.109	1.065	0.100	0.766	0.817	0.834	0.878

We repeat the same analysis for the HAR(22) type of model. As we can see in Table 2.3, the robust GMM performs bit worse than either of the competing methods in the benchmark case; however, the difference is not substantial.

Table 2.3: Simulation results: HAR(22); T = 2500, $q = 0.025$

Robust GMM				
	β_0	β_1	β_2	β_3
True values:	0.203	0.073	0.750	1.718
Bias	0.174	0.030	0.322	-1.236
Standard dev.	0.120	0.054	0.428	0.381
RMSE	0.211	0.062	0.536	1.294
Joint RMSE			1.417	
Coverage Probability	0.433	0.981	0.648	0.220
Ratio of s.e. (Mean)	1.272	0.950	2.105	0.790
Ratio of s.e. (Median)	1.397	1.022	2.314	0.843
ES RMSE			0.324	

Joint GMM				
	β_0	β_1	β_2	β_3
True values:	0.203	0.073	0.750	1.718
Bias	0.157	0.018	0.358	-1.122
Standard dev.	0.081	0.035	0.175	0.385
RMSE	0.177	0.040	0.399	1.186
Joint RMSE			1.265	
Coverage Probability	0.679	0.996	0.189	0.334
Ratio of s.e. (Mean)	0.794	0.618	0.806	0.739
Ratio of s.e. (Median)	0.856	0.660	0.864	0.801
ES RMSE			0.319	

2-step GMM				
	β_0	β_1	β_2	β_3
True values:	0.203	0.073	0.750	1.718
Bias	0.173	0.018	0.290	-1.035
Standard dev.	0.113	0.047	0.183	0.698
RMSE	0.207	0.050	0.343	1.248
Joint RMSE			1.312	
Coverage Probability	0.395	0.981	0.648	0.248
Ratio of s.e. (Mean)	1.261	0.854	0.966	1.463
Ratio of s.e. (Median)	1.310	0.903	1.008	1.516
ES RMSE			0.342	

When we compare the estimators across different confidence levels and sample sizes, we can see in Table 2.4, that the robust estimator performs better at higher confidence levels than the joint estimator if we consider the joint RMSE and they are similar in ES RMSE. In contrary to the AR(1)-TS-ARCH(1) type of model, the robust estimator is sometimes worse than the 2-step estimator but in general, the robust estimator is still better in terms of estimation accuracy than the 2-step estimator.

Table 2.4: Comparison of Joint and ES RMSE in HAR(22) type of model

Joint RMSE									
	Robust vs Joint T					Robust vs 2-step T			
q	500	1000	2500	5000	q	500	1000	2500	5000
0.010	2.423	2.251	1.886	1.717	0.010	0.373	1.012	1.661	1.017
0.025	1.545	0.830	1.121	1.137	0.025	0.404	0.626	1.080	1.000
0.050	0.923	0.722	0.972	0.944	0.050	0.629	0.603	0.998	0.998
0.100	1.144	0.858	0.785	0.747	0.100	0.787	0.712	1.002	0.989
ES RMSE									
	Robust vs Joint T					Robust vs 2-step T			
q	500	1000	2500	5000	q	500	1000	2500	5000
0.010	1.381	1.078	0.984	0.877	0.010	0.351	0.539	0.698	0.983
0.025	1.191	0.861	1.013	1.006	0.025	0.500	0.414	0.947	1.001
0.050	1.025	0.879	0.975	0.941	0.050	0.520	0.481	0.977	1.001
0.100	1.149	1.012	0.931	0.861	0.100	0.716	0.674	1.001	0.999

In conclusion, we can see in this simulation study that in terms of estimation accuracy the robust and the joint estimator are comparable (similarly to the findings of Ackerberg et al. (2014)) and the robust estimator is better than the 2-step estimator without adjustment term. In inference, the robust estimator performs more reliably than the joint estimator.

2.5 Empirical application

This section includes the empirical application of the robust estimator. First, we describe the dataset. Second, we describe the option implied Value-at-Risk and Expected Shortfall introduced by Mitra (2015) and Barone Adesi (2016). Third, we investigate which of the proposed explanatory variables, such as absolute return, range, RV and VIX, are statistically significant in the estimation of the one-day, -week and -month ahead ES of the S&P500 index. Fourth, we provide empirical application of the option implied risk measures in ES estimation.

2.5.1 Data

In the empirical application we use five datasets. Our datasets run from 2/1/1993 to 2/16/2018, except the OptionMetrics which starts at 1/2/2009 and runs until 3/31/2016. We use the 5-minute SPY data to construct the daily Realized Variance (RV) variable. In our application, we apply 5 different explanatory variables to explain VaR and ES. We use the lagged daily absolute returns, lagged range, which is defined as the difference between the logarithm of the high and low, lagged VIX and the lagged square root of RV and option implied VaR/ES as explanatory variables. Note that the magnitude of VIX is approximately tenfold comparing to the other explanatory variables, therefore in our analysis we always used VIX/10 so the parameter estimates on all variables are around at the same level.

Table 2.5: Datasets

Data	Source
S&P500 Index Price	yahoo.com
S&P500 Option data	OptionMetrics
VIX	CBOE
High frequency SPY data	TAQ
Risk-free Interest Rate	FRED

2.5.2 Option implied VaR and ES

Mitra (2015) and Barone Adesi (2016) derive model-free, closed-form analytic relation between ES, VaR and European options. The importance of the relation between options, VaR and ES can be motivated by the forward-looking nature of option prices. In this section, we present their derivation.

Let the conditional distribution (under the "objective" or $\mathbb{P}$-measure) of interest be:

$$S_T|\mathcal{F}_t \sim F_t$$

where t is today (the date of the information set), T is the maturity date of the option, and S_T is the value of the stock at time T.

From option pricing theory we know

$$\begin{aligned} P_{t,T} &= \exp\left(-\mu_{t,T}\left(T-t\right)\right)\int_0^K \left(K-s\right) f_t\left(s\right) ds \\ \text{so } \frac{\partial P_{t,T}}{\partial K} &= \exp\left\{-\mu_{t,T}\left(T-t\right)\right\} F_t\left(K\right) \quad \text{using Leibniz's rule} \\ &\equiv \exp\left(-\mu_{t,T}\left(T-t\right)\right) q_{t,T} \\ \text{where } q_{t,T} &\equiv F_t\left(K\right) \in [0,1] \Leftrightarrow K = F_t^{-1}\left(q_{t,T}\right) \equiv \text{VaR}_{t,q}^T, \end{aligned}$$

where $P_{t;T}$ is the price of a put option at period t with maturity at period T, $\mu_{t,T}$ is the risk-premium (SDF) for maturity T. That is, we can think of K as the $q_{t,T}$-VaR of S_T.[4] However, to recover $q_{t,T}$ we need to be able to estimate $\partial P_t/\partial K$ and $\mu_{t,T}$. Under the risk-neutral (or $\mathbb{Q}$) measure, we use $r_{t,T}$, the risk-free rate, which is (essentially) directly observable, in place of $\mu_{t,T}$, which requires a model or further assumptions, meaning we just have to estimate $\partial P_{t,T}/\partial K$.

Barone Adesi (2016) provides a simple method, which does not only eliminate the first-order error in a Taylor expansion but also eliminates the first-order error

[4] Note that in their papers, Mitra (2015) and Barone Adesi (2016) use a different definition of VaR. They define VaR^L as the $1-q$ quantile of portfolio *losses*, where subscript L is our notation for VaR defined on *losses*.

due to the implied volatility changing across strike prices (Aït-Sahalia and Lo, 2000), to estimate $\partial P_{t,T}/\partial K$.

The method works as the following: given three options $\{(P_{i;t,T}, K_i)\}_{i=1}^{3}$ where $K_1 < K_2 < K_3$, we can estimate $\partial P_{t,T}/\partial K$ at K_2 using a simple approximation two-sided estimate of the first derivative. Let

$$\Delta_{2;t,T} \equiv \left.\frac{\partial P_{t,T}}{\partial K}\right|_{K=K_2},$$

then

$$\hat{\Delta}_{2;t,T} = \frac{1}{2}\left(\frac{P_{3;t,T} - P_{2;t,T}}{K_3 - K_2} + \frac{P_{2;t,T} - P_{1;t,T}}{K_2 - K_1}\right).$$

We can then get an estimate of the tail probability (q) associated with a given strike price as:

$$q^{\mathbb{P}}_{2;t,T} = \exp\left(\mu_{t,T}\left(T-t\right)\right)\hat{\Delta}_{2;t,T}.$$

Denote the $\mathbb{Q}$-measure version as

$$q^{\mathbb{Q}}_{2;t,T} = \exp\left(r_{t,T}\left(T-t\right)\right)\hat{\Delta}_{2;t,T}.$$

Note that

$$\begin{aligned}
q^{\mathbb{P}}_{2;t,T} - q^{\mathbb{Q}}_{2;t,T} &= \left[\exp\left(\mu_{t,T}\left(T-t\right)\right) - \exp\left(r_{t,T}\left(T-t\right)\right)\right]\hat{\Delta}_{2;t,T} \\
&\approx \left(T-t\right)\exp\left(r_{t,T}\left(T-t\right)\right)\left(\mu_{t,T} - r_{t,T}\right)\hat{\Delta}_{2;t,T} \quad \text{when } \mu_{t,T} \approx r_{t,T} \\
&\approx \left(T-t\right)\left(1 + r_{t,T}\left(T-t\right)\right)\left(\mu_{t,T} - r_{t,T}\right)\hat{\Delta}_{2;t,T} \quad \text{when } r_{t,T} \approx 0 \\
&\to 0 \text{ as } t \to T \text{ or as } \mu_{t,T} \to r_{t,T} \text{ and assuming no jumps in the underlying.}
\end{aligned}$$

As we are looking at relatively short maturity options, we use $q^{\mathbb{Q}}_{2;t,T}$ instead of $q^{\mathbb{P}}_{2;t,T}$. From the above derivation, we can see that if $\mu_{t,T}$, the risk premium increases while everything remains constant, then $VaR^{T}_{t,q}$ decreases.[5]

[5] Let's say we are interested in 5% VaR. Since $\hat{\Delta}_{2;t,T} > 0$ for put options, $q^{\mathbb{P}}_{2;t,T}$ increases if $\mu_{t,T}$ increases. But then K, the strike price corresponding to $q = 0.05$ has to decrease as $F(\cdot)$ is a non-decreasing function.

Similarly to VaR, Mitra (2015) and Barone Adesi (2016) show that we can extract Expected Shortfall from option prices as follows: as before, note that

$$\begin{aligned} P_{t,T} =& \exp\left(-\mu_{t,T}\left(T-t\right)\right)\int_0^K \left(K-s\right) f_t\left(s\right) ds \\ =& \exp\left(-\mu_{t,T}\left(T-t\right)\right) K q_{t,T} - \exp\left(-\mu_{t,T}\left(T-t\right)\right) q_{t,T} \mathrm{ES}_{t,q}^T \end{aligned}$$

$$\text{where} \quad q_{t,T} \equiv F_t\left(K\right)$$

$$\text{and} \quad ES_{t,q}^T \equiv \frac{1}{q_{t,T}}\int_0^K s f_t\left(s\right) ds.$$

Re-arranging we then obtain:

$$ES_{t,q}^T = K - \frac{1}{q_{t,T}} P_{t,T} \exp\left(\mu_{t,T}\left(T-t\right)\right).$$

To create the 1-day ahead option implied VaR and ES at 2.5% confidence level, first we need to find the strike prices (VaR), which belongs to the 2.5% percentile. Acknowledging the discreteness of the strike prices in our panel, we proceed in 2 steps: first, we follow the method presented in Barone Adesi (2016) and above. That is, we calculate the two-sided estimate of the price derivative with respect to the strike price and then use the estimated interest rate to obtain an estimate for q as

$$\hat{q}_{i;t,T} = \exp(r_{t,T}(T-t))\Delta_{i;t,T}.$$

Second, since generally $\hat{q}_{i;t,T} \neq q \quad \forall i$, we use the weighted average of the two bracketing options to calculate the option implied VaR and ES at q probability. That is, assuming that the options are ordered from lowest to highest strike prices, we can define

$$\begin{aligned} i^* =& \max_i \hat{q}_{i;t,T} \\ \text{s.t.} \quad & \hat{q}_{i;t,T} \leq q, \end{aligned}$$

then $\hat{q}_{i^*;t,T} \leq q \leq \hat{q}_{i^*+1;t,T}$. Then we take a weighted average of the two bracketing options to get VaR and ES at q probability as follows:

$$\widetilde{VaR}^T_{t,q} = \frac{\hat{q}_{i^*+1;t,T} - q}{\hat{q}_{i^*+1;t,T} - \hat{q}_{i^*;t,T}} K_{i^*} + \frac{q - \hat{q}_{i^*;t,T}}{\hat{q}_{i^*+1;t,T} - \hat{q}_{i^*;t,T}} K_{i^*+1}$$

$$\begin{aligned}\widetilde{ES}^T_{t,q} = &\frac{\hat{q}_{i^*+1;t,T} - q}{\hat{q}_{i^*+1;t,T} - \hat{q}_{i^*;t,T}} \left(K_{i^*} - \frac{1}{\hat{q}_{i^*;t,T}} P_{i^*,t} \exp(r_{t,T}(T-t)) \right) \\ &+ \frac{q - \hat{q}_{i^*;t,T}}{\hat{q}_{i^*+1;t,T} - \hat{q}_{i^*;t,T}} \left(K_{i^*+1} - \frac{1}{\hat{q}_{i^*+1;t,T}} P_{i^*+1,t} \exp(r_{t,T}(T-t)) \right).\end{aligned}$$

It is important to mention that a forecaster usually has a target horizon in mind. The general calculations above assumed that $T-t$ matched with this horizon, whereas in general this might be not true as there does not exist options which expire at the exact date. Therefore, we use a similar weighting procedure as with q. That is, we pick $\widetilde{VaR}^{T_1}_{t,q}$ and $\widetilde{VaR}^{T_2}_{t,q}$, where $T_1 \leq T \leq T_2$ and assuming that options are ordered from the shortest to the longest maturity,

$$\begin{aligned} T_1 = &\max_t \ t \\ &\text{s.t.} \quad t \leq T, \end{aligned}$$

similarly for T_2, then we take the weighted average of these two VaR as follow:

$$\widetilde{\text{VaR}}^T_{t,q} = \frac{T_2 - T}{T_2 - T_1} \widetilde{\text{VaR}}^{T_1}_{t,q} + \frac{T - T_1}{T_2 - T_1} \widetilde{\text{VaR}}^{T_2}_{t,q},$$

we can of course repeat this for ES to extract the option implied ES with the exact confidence level, q and expiration, T:

$$\widetilde{ES}^T_{t,q} = \frac{T_2 - T}{T_2 - T_1} \widetilde{\text{ES}}^{T_1}_{t,q} + \frac{T - T_1}{T_2 - T_1} \widetilde{\text{ES}}^{T_2}_{t,q}.$$

However, to implement this bracketing with respect to T, we need to have options expiring before and after the target date. In shorter horizons, there might not be any options in the dataset, which would expire before the target date. In that case,

we cannot use this bracketing and we use the option which expires the closest to the target horizon. To overcome this difference between the expiration date of the option and the target horizon we propose the following transformation to stock returns:

$$q = \mathbb{P}_t\left(S_{t+k} \leq \widetilde{VaR}_{t+k|t}\right) = \mathbb{P}_t\left(\left(\frac{S_{t+k}}{S_t}\right)^{1/k} - 1 \leq \left(\frac{\widetilde{VaR}_{t+k|t}}{S_t}\right)^{1/k} - 1\right),$$

where we only took monotonic transformation and note that S_t is known at time t. Under the assumption that the daily return does not change between t+1 and t+k, that is $S_{t+k} = (1+\mu)^k S_t$,

$$\widehat{\mathrm{VaR}}_{t+1,q|t} \equiv 100 \times \left[\left(\frac{\widetilde{\mathrm{VaR}}_{t+k|t}}{S_t}\right)^{1/k} - 1\right].$$

We transform ES similarly, that is

$$\widehat{\mathrm{ES}}_{t+1,q|t} \equiv 100 \times \left[\left(\frac{\widetilde{\mathrm{ES}}_{t+k,q|t}}{S_t}\right)^{1/k} - 1\right].$$

Note that the above discounting is not mathematically correct for the ES. Therefore, in the empirical part, we use separately the discounted and not discounted measures.

2.5.3 Estimating ES using absolute returns, range, VIX, RV

In this section, we use our dataset spanning the whole horizon from 2/1/1993 until 2/16/2018. First, we use each explanatory variable separately to test whether each of these variables is significant in explaining the 1-, 5- and 22-day ahead ES. Second, we test which of these 5 explanatory variable is statistically significant in estimating 2.5% Expected Shortfall. Third, we repeat this exercise at 5% confidence level.

In the first step, we fit fourth order polynomial to estimate the 2.5% Value-at-Risk, whereas in the ES estimation we only use the first order variable, that is,

$$\mathrm{ES}_{t+1,0.025} = \beta_0 + \beta_1 X_t,$$

where $X_t \in \{\text{Absolute return}_t, \text{Range}_t, \sqrt{\text{RV}_t}, \text{VIX}_t/10\}$. We estimate the standard errors using Newey-West Long Run Variance estimator with Bartlett kernel with $2\times h$ lags, where h is the forecast horizon (Newey and West, 1986). In the multivariate case, we only use the second order cross terms in the first step to ease computation. Note that when we estimate the covariance matrix from the joint and 2-step estimation, we still need to estimate the density function, which is already estimated during the optimization in the robust estimation. It is one of the advantages of the robust estimation method that the density function estimation is part of the optimization, therefore it resulted in more stable standard errors estimate than the other two methods. First, we always tried to use the same density function estimation in the joint and 2-step estimation but when they resulted in explosive standard errors then we always increased or decreased the bandwidth parameter so that the standard error estimates are similar to those from the robust estimation.

Table 2.6 includes our results from the univariate regression using the robust estimation method. The 3 panels belongs to the 3 forecast horizons. Each column shows the results for the separate explanatory variables. The standard errors are in parentheses and the last row in each block shows the average L_{FZ0} loss under the univariate regression.

As we can see from this table all the explanatory variables are statistically significant in the univariate case. As it could have been expected, the coefficients are negative for absolute return, range, RV and VIX as an increase in these variables can be interpreted as higher volatility in the underlying. The RV and VIX models generate smaller losses than the noisier absolute returns or range.

Table 2.6: Parameter estimates for univariate regression at 2.5% confidence level, Robust GMM

Panel A: t+1				
	Abs Ret	Range	RV	VIX/10
β_0	−2.726	−0.961	−0.696	0.222
s.e.	(0.162)	(0.130)	(0.134)	(0.207)
β_1	−0.761	−1.557	−2.478	−1.453
s.e.	(0.164)	(0.137)	(0.207)	(0.121)
Loss	1.181	0.993	0.926	0.873

Panel B: t+5				
	Abs Ret	Range	RV	VIX/10
β_0	−5.173	−2.288	−1.989	−0.066
s.e.	(0.316)	(0.262)	(0.291)	(0.454)
β_1	−1.004	−1.184	−2.056	−1.292
s.e.	(0.177)	(0.098)	(0.160)	(0.112)
Loss	1.911	1.736	1.715	1.670

Panel C: t+22				
	Abs Ret	Range	RV	VIX/10
β_0	−10.739	−0.429	−0.736	5.641
s.e.	(0.644)	(0.664)	(0.782)	(1.052)
β_1	−0.977	−1.926	−3.044	−2.022
s.e.	(0.204)	(0.154)	(0.268)	(0.158)
Loss	2.624	2.475	2.468	2.416

Table 2.7 shows the results of multivariate regressions with the robust estimation method. The 3 columns belongs to the 3 different horizons. We can see from this table that at the shortest horizon (1-day ahead ES), only RV and VIX are statistically significant, while in the 5-day ahead estimation variable RV loses its significance. At the longest horizon (22-day or 1-month ahead estimate) RV and Range becomes significant along the VIX measure. The positive signs on absolute return and RV can be explained by the high correlation between the explanatory variables.

When we repeat the same estimation at 5% confidence level, our conclusion re-

Table 2.7: Multivariate regression results at 2.5% confidence level, Robust GMM

		t+1	t+5	t+22
Constant	β	−0.019	−1.021	5.124
	s.e.	(0.199)	(0.508)	(0.868)
Abs	β	0.007	0.376	0.160
	s.e.	(0.209)	(0.250)	(0.213)
Range	β	−0.017	−0.675	−0.891
	s.e.	(0.270)	(0.332)	(0.271)
RV	β	−0.856	−0.441	0.797
	s.e.	(0.360)	(0.435)	(0.138)
VIX/10	β	−0.948	−0.553	−1.706
	s.e.	(0.152)	(0.229)	(0.225)
Loss		0.866	1.668	2.403

mains the same. As can be seen in Table 2.8, at each forecast horizon the explanatory variables are statistically significant in estimating ES at 5% level.

Table 2.8: Parameter estimates for univariate regression at 5% confidence level, Robust GMM

Panel A: t+1				
	Abs Ret	Range	RV	VIX/10
β_0	−2.161	−0.793	−0.551	0.274
s.e.	(0.099)	(0.093)	(0.092)	(0.136)
β_1	−0.665	−1.307	−2.120	−1.272
s.e.	(0.108)	(0.097)	(0.140)	(0.081)
Loss	0.964	0.812	0.747	0.697

Panel B: t+5				
	Abs Ret	Range	RV	VIX/10
β_0	−4.385	−1.891	−1.477	0.174
s.e.	(0.218)	(0.198)	(0.218)	(0.315)
β_1	−0.675	−0.996	−1.838	−1.143
s.e.	(0.120)	(0.076)	(0.131)	(0.081)
Loss	1.690	1.555	1.537	1.494

Panel C: t+22				
	Abs Ret	Range	RV	VIX/10
β_0	−8.810	−1.361	−0.671	4.099
s.e.	(0.403)	(0.495)	(0.623)	(0.716)
β_1	−0.677	−1.398	−2.454	−1.589
s.e.	(0.127)	(0.109)	(0.203)	(0.106)
Loss	2.393	2.281	2.262	2.218

In Table 2.9, we can see the results for the multivariate regression. As before, at the shortest horizon only RV and VIX are statistically significant, while in the one-week ahead estimation RV loses its significance. In the longest horizon, all variables becomes statistically significant in estimating the 1-month ahead ES while the sign on the RV and absolute return changes due to high correlation between the variables.

Table 2.9: Multivariate regression results at 5% confidence level, Robust GMM

		t+1	t+5	t+22
Constant	β	0.079	−0.463	3.574
	s.e.	(0.140)	(0.359)	(0.667)
Absolute Return	β	0.039	0.295	0.348
	s.e.	(0.159)	(0.168)	(0.149)
Range	β	−0.011	−0.421	−0.801
	s.e.	(0.207)	(0.214)	(0.194)
RV	β	−0.724	−0.280	0.755
	s.e.	(0.273)	(0.319)	(0.075)
VIX/10	β	−0.868	−0.689	−1.395
	s.e.	(0.124)	(0.160)	(0.162)
Loss		0.695	1.489	2.208

2.5.4 Comparing forecasts of ES at 2.5% and 5% level

In this subsection, we conduct out-of-sample estimation for ES at 2.5% and 5% level. We use rolling window estimates with 14-year-long windows by reestimating the parameters at each year. That is, first we estimate the parameters using the sample from 1993 until 2007. Then we estimate the ES for the year of 2008. Then we use the sample from 1994 until 2008 to reestimate the parameters and make a forecast for the year of 2009 etc. As explanatory variables, we use all of the variables, that is, absolute return, range, RV and VIX/10.

To compare the forecast accuracy of the different estimation methods, we conduct Diebold-Mariano tests of differences in average L_{FZ0} losses (Diebold and Mariano, 2002). In Table 2.10 and 2.11 we reported the t-statistics of the Diebold-Mariano tests. If the number is less than -1.96, then we can claim that the forecast using the robust estimation method is significantly better than the forecasts from the other estimation method. If the number is larger than 1.96, then the other estimation method is better than the robust. As we can see from these tables, neither of the estimation method emerges as a winner. For example, at one-month-ahead horizon

the robust estimation method might be better as the joint at 5% level (-2.139) when we include all the explanatory variables but at one-day-ahead horizon at 2.5% level, the joint estimation method results in better forecast (2.104). We can also not announce a winner between the robust and the 2-step (without adjustment term) estimation method in the out-of-sample analysis.

Table 2.10: Diebold-Mariano t-statistics on average out-of-sample loss differences at 2.5% confidence level

	Robust vs Joint			Robust vs 2-step		
	t+1	t+5	t+22	t+1	t+5	t+22
Abs Ret	-4.940	1.280	1.137	0.715	-1.635	-0.816
Range	-1.593	1.483	-0.908	-1.056	2.453	1.091
RV	1.687	2.401	-1.656	1.031	2.329	0.834
VIX/10	-0.823	0.647	-1.909	0.127	-0.030	-2.023
Multivariate	2.104	1.465	0.275	-2.111	1.207	-1.890

Table 2.11: Diebold-Mariano t-statistics on average out-of-sample loss differences at 5% confidence level

	Robust vs Joint			Robust vs 2-step		
	t+1	t+5	t+22	t+1	t+5	t+22
Abs Ret	-1.245	-0.961	-1.583	-5.460	-1.511	-2.022
Range	1.871	-0.936	1.320	-0.382	1.003	-3.415
RV	-0.301	1.145	0.071	-0.311	1.706	0.305
VIX/10	-0.560	0.288	-1.192	-0.169	-0.376	-2.573
Multivariate	-1.716	1.451	-2.139	0.986	1.191	-0.302

2.5.5 Testing the significance of option implied ES in ES estimation

In this subsection we test whether the option implied ES, introduced by Mitra (2015) and Barone Adesi (2016), is statistically significant in ES estimation. Since in our cleaned option dataset we can observe consistently at least 52 expiration dates per year only after 2009, we choose to cut our sample in this analysis.

As we mentioned in Subsection 2.5.2, we do not always observe an option which would expire on the exact date for which we estimate the ES. To overcome this difficulty, we proposed to bracket the estimated ES with respect to the expiration period. If bracketing is not possible, we can still discount the obtained measure. Although, our discounting is mathematically correct for option implied VaR, it is not an appropriate transformation for ES. Therefore, we proceed to use both a discounted (D) and not discounted (ND) ES in this analysis. The drawback of the not discounted measure is that its expiration might be after the targeted horizon.

In Table 2.12 and 2.13, we can see that the option implied measure is statistically significant in estimated the ES at each horizon and the average L_{FZ0} is one of the lowest at the longest horizons.

Similarly to the previous Subsection, we also test whether the either of the option implied measure is significant in estimating the one-day, -week and -month ahead ES. As we can see from Table 2.14 and 2.16, the discounted option implied ES is the only significant measure at the longest horizon both at 2.5% and 5% level. However, the not discounted option implied measure is not statistically significant when the other explanatory variables are included in the analysis.

2.6 Conclusion

In 2013, The Basel Committee on Banking Supervision proposed to replace the current market risk measure, Value-at-Risk (VaR) with Expected Shortfall (ES). Contrary to VaR, ES does not only account for "tail risk" but it is also a coherent risk measure with appealing theoretical properties. Despite its introduction as a market risk measure, the academic literature is not abundant in estimation methods of ES. In this paper, we explored a new estimation method for this risk measure.

Fissler and Ziegel (2016) show that ES is only elicitable jointly with VaR, that is, it is impossible to estimate ES separately from VaR. This implies that the ES

Table 2.12: Parameter estimates for univariate regression at 2.5% confidence level from 2009–2016, Robust GMM

Panel A: t+1						
	Abs Ret	Range	RV	VIX	Implied (D)	Implied (ND)
β_0	−3.112	−0.829	−0.391	0.133	−0.744	−0.723
s.e.	(0.293)	(0.183)	(0.150)	(0.336)	(0.197)	(0.200)
β_1	−0.270	−1.676	−3.086	−1.402	0.934	0.250
s.e.	(0.204)	(0.189)	(0.259)	(0.195)	(0.123)	(0.034)
Loss	1.202	0.998	0.874	0.924	0.864	0.868

Panel B: t+5						
	Abs Ret	Range	RV	VIX	Implied (D)	Implied (ND)
β_0	−5.604	−3.200	−2.772	−2.150	−5.483	−2.167
s.e.	(0.595)	(0.637)	(0.730)	(1.327)	(0.467)	(1.117)
β_1	−0.716	−0.961	−2.043	−0.858	0.235	0.443
s.e.	(0.308)	(0.225)	(0.455)	(0.286)	(0.084)	(0.109)
Loss	1.914	1.808	1.787	1.774	1.937	1.814

Panel C: t+22						
	Abs Ret	Range	RV	VIX	Implied (D)	Implied (ND)
β_0	−10.392	−5.404	−4.838	−0.651	−6.937	−3.191
s.e.	(1.201)	(0.940)	(1.127)	(1.296)	(2.783)	(1.860)
β_1	−0.777	−1.004	−2.073	−1.232	0.450	0.572
s.e.	(0.336)	(0.157)	(0.344)	(0.174)	(0.081)	(0.091)
Loss	2.563	2.493	2.487	2.486	2.468	2.470

estimates depends on the VaR estimates. To ease this dependence on the VaR estimates, we applied a 2-step robust estimation method. In the first step, we estimated VaR with non-parametric sieve polynomial to eliminate the estimation error in VaR, and in the second step we used a robust GMM estimation method to estimate the parameters of the ES model.

In Monte Carlo simulation studies we explored the finite sample properties of the robust estimation method in comparison with a joint (VaR and ES estimated jointly in 1-step) and a "non-robust" 2-step method. Our findings match the findings of the

Table 2.13: Parameter estimates for univariate regression at 5% confidence level from 2009–2016, Robust GMM

Panel A: t+1						
	Abs Ret	Range	RV	VIX/10	Implied (D)	Implied (ND)
β_0	−2.404	−0.743	−0.341	0.214	−0.781	−0.681
s.e.	(0.194)	(0.140)	(0.122)	(0.239)	(0.188)	(0.190)
β_1	−0.387	−1.395	−2.714	−1.274	1.148	0.317
s.e.	(0.151)	(0.150)	(0.212)	(0.137)	(0.152)	(0.040)
Loss	0.994	0.840	0.764	0.783	0.855	0.847

Panel B: t+5						
	Abs Ret	Range	RV	VIX/10	Implied (D)	Implied (ND)
β_0	−4.526	−2.642	−2.103	−1.281	−4.632	−1.767
s.e.	(0.389)	(0.401)	(0.435)	(0.754)	(0.368)	(0.693)
β_1	−0.568	−0.801	−1.849	−0.831	0.208	0.469
s.e.	(0.198)	(0.136)	(0.281)	(0.166)	(0.079)	(0.094)
Loss	1.698	1.619	1.607	1.605	1.721	1.607

Panel C: t+22						
	Abs Ret	Range	RV	VIX/10	Implied (D)	Implied (ND)
β_0	−9.073	−4.909	−4.412	−3.097	−5.765	−3.020
s.e.	(0.884)	(0.829)	(0.998)	(1.202)	(0.425)	(1.099)
β_1	−0.409	−0.779	−1.625	−0.728	0.365	0.445
s.e.	(0.286)	(0.153)	(0.323)	(0.147)	(0.048)	(0.076)
Loss	2.348	2.316	2.309	2.295	2.271	2.271

related literature: in terms of estimation accuracy, the robust estimation method performs similarly to the joint estimation and it has more accurate than the a "non-robust" 2-step method. In statistical inference, the robust estimation dominates the 1-step joint method.

To test the robust estimation method in data, we apply this technique in S&P500 data and examine the significance of different explanatory variables in one-day, one-week and one-month ahead ES. We found that out of the 4 proposed explanatory variables; such as lagged absolute return, lagged range, realized volatility (RV), VIX;

Table 2.14: Multivariate regression results with discounted option implied ES at 2.5% confidence level from 2009–2016, Robust GMM

		t+1	t+5	t+22
Constant	β	0.351	0.207	−4.641
	s.e.	(0.262)	(1.037)	(3.429)
Abs Ret	β	0.026	0.497	0.743
	s.e.	(0.239)	(0.537)	(0.551)
Range	β	0.065	−0.804	−0.061
	s.e.	(0.319)	(0.767)	(0.801)
RV	β	−1.853	−0.608	−2.642
	s.e.	(0.629)	(1.586)	(1.833)
VIX/10	β	−0.620	−0.774	0.200
	s.e.	(0.276)	(0.601)	(0.972)
Implied ES (D)	β	0.285	0.100	0.245
	s.e.	(0.135)	(0.102)	(0.090)
Loss		0.842	1.600	2.467
F test p-val.		0.000	0.000	0.000

Table 2.15: Multivariate regression results with not discounted option implied ES at 2.5% confidence level from 2009–2016, Robust GMM

		t+1	t+5	t+22
Constant	β	−0.285	−0.725	−2.097
	s.e.	(0.340)	(1.053)	(2.163)
Abs Ret	β	−0.546	0.580	0.136
	s.e.	(0.351)	(0.517)	(0.477)
Range	β	0.159	−0.255	0.286
	s.e.	(0.532)	(0.657)	(0.686)
RV	β	−2.370	−2.033	−1.355
	s.e.	(0.889)	(1.341)	(2.401)
VIX/10	β	0.225	0.714	−0.417
	s.e.	(0.348)	(0.682)	(1.088)
Implied ES (ND)	β	0.166	0.680	0.226
	s.e.	(0.047)	(0.246)	(0.231)
Loss		0.843	1.586	2.431
F test p-val.		0.000	0.000	0.000

Table 2.16: Multivariate regression results with discounted option implied ES at 5% confidence level from 2009–2016, Robust GMM

		t+1	t+5	t+22
Constant	β	0.042	−0.282	−5.370
	s.e.	(0.229)	(0.581)	(1.047)
Abs Ret	β	−0.291	0.519	0.231
	s.e.	(0.216)	(0.290)	(0.242)
Range	β	0.493	−0.903	0.185
	s.e.	(0.319)	(0.395)	(0.277)
RV	β	−2.972	−0.475	−1.037
	s.e.	(0.547)	(0.824)	(0.616)
VIX/10	β	−0.139	−0.403	−0.051
	s.e.	(0.239)	(0.362)	(0.231)
Implied ES (D)	β	0.264	0.099	0.151
	s.e.	(0.145)	(0.077)	(0.048)
Loss		0.754	1.602	2.248
F test p-val.		0.000	0.000	0.000

Table 2.17: Multivariate regression results with not discounted option implied ES at 5% confidence level from 2009–2016, Robust GMM

		t+1	t+5	t+22
Constant	β	0.155	−0.627	−0.522
	s.e.	(0.255)	(0.663)	(1.404)
Abs Ret	β	−0.499	0.314	0.694
	s.e.	(0.238)	(0.320)	(0.296)
Range	β	0.553	−0.093	−0.293
	s.e.	(0.333)	(0.409)	(0.429)
RV	β	−2.980	−1.614	−0.558
	s.e.	(0.652)	(0.934)	(0.991)
VIX/10	β	−0.299	−0.210	−0.361
	s.e.	(0.282)	(0.434)	(0.437)
Implied ES (ND)	β	0.040	0.182	0.345
	s.e.	(0.047)	(0.157)	(0.167)
Loss		0.756	1.592	2.248
F test p-val.		0.000	0.000	0.000

RV and VIX are statistically significant of explaining the 1-day and 1-month ahead ES and only VIX remains significant at estimating the one-week ahead ES.

As an additional contribution, we implemented the option implied ES (and VaR), proposed in Mitra (2015) and Barone Adesi (2016). We found that in a univariate regression they have statistical power to explain the ES; however, when we tested their statistical significance along with the previous 4 variables, they lost their significance.

2.7 Appendix

2.7.1 Proofs

Proof of Lemma 1

Chen et al. (2003) show in their proof of Theorem 2 that:

$$\left|\left|\bar{g}(y_t, \theta_e^0, \hat{v}_t) - \bar{g}(y_t, \theta_e^0, v_t^0) - \frac{\partial \mathbb{E}\left[g(Y_t, e_t(\theta_e^0), v_t^0)\right]}{\partial v_t}\left[\hat{v}_t - v_t^0\right]\right|\right| = o_p(n^{-1/2}).$$

The above equality to hold under our Assumption 1 and 2, which satisfy their assumption (2.1)-(2.5).

Now since construction of the locally robust moment condition and the definition of pathwise derivative:

$$\frac{\partial \mathbb{E}\left[g(Y_t, e_t(\theta_e^0), v_t^0)\right]}{\partial v_t}\left[\hat{v}_t - v_t^0\right] = 0,$$

and the result of Lemma 1 follow.

$$\frac{\partial \mathbb{E}\left[g(Y_t, e_t(\theta_e^0), v_t)\right]}{\partial v_t} =$$

$$\frac{\partial \mathbb{E}\left[m(Y_t, e_t(\theta_e^0), v_t)\right]}{\partial v_t} + \left[\frac{\partial \mathbb{E}\left[m(Y_t, e_t(\theta_e^0), v_t^0)\right]}{\partial v_t}\right]\left[\frac{\partial \mathbb{E}\left[\rho(Y_t, v_t^0)\right]}{\partial v_t}\right]^{-1}\frac{\partial \mathbb{E}\left[q - \mathbb{1}\{Y_t \leq v_t\}\right]}{\partial v_t}$$

where $m(\cdot)$ is defined in equation (2.5). Note that $\mathbb{E}\left[q - \mathbb{1}\{Y_t \leq v_t\}\right] = -\mathbb{E}\left[\rho(Y_t, v_t)\right]$ (see the definition of $\rho(Y_t, v_t)$ in equation (2.7)). So if we differentiate at $v_t = v_t^0$, then the condition follows.

Proof of Theorem 1

The proof is based on Theorem 3.1 of Hall (2005). We need to show that $\mathbb{E}\left[g(Y_t, e_t(\theta_e^0), v_t^0)\right] = 0$, $\mathbb{E}\left[g(Y_t, e_t(\bar{\theta}_e), v_t^0)\right] \neq 0$ for all $\bar{\theta}_e \in \Theta_e \setminus \theta_e^0$ and $\mathbb{E}\left[\sup_{\theta_e \in \Theta_e} ||g(Y_t, e_t(\theta_e), v_t^0)||\right] < \infty$ because the other conditions are satisfied.

Patton et al. (2019) show in their proof of Theorem 1 that L_{FZ0} loss function is uniquely minimized at (v_t^0, θ_e^0) under our assumption (iii) and (v). Since $F_t(\cdot|\mathcal{F}_{t-1})$ is continuously differentiable, we find that $\mathbb{E}\left[\frac{\partial L_{FZ0}\left(Y_t, e_t(\theta_e^0), v_t^0\right)}{\partial \theta_e}\right] = 0$. Since $\mathbb{E}\left[\phi(Y_t, e_t(\theta_e^0), v_t^0)\right] = 0$ by definition of the adjustment term, $\mathbb{E}\left[g(Y_t, e_t(\theta_e^0), v_t^0)\right] = 0$ and $\mathbb{E}\left[g(Y_t, e_t(\bar{\theta}_e), v_t^0)\right] \neq 0$ for all $\bar{\theta}_e \in \Theta_e \setminus \theta_e^0$.

Now since,

$$\left|\left|\frac{\partial L_{FZ0}(Y_t, v_t, e_t(\theta_e)}{\partial e_t(\theta_e)}\frac{\partial e_t(\theta_e)}{\partial \theta_e} + \right.\right.$$

$$\left.\left.\left(\frac{1}{T}\sum_{k=1}^{T}\frac{1}{q e_k(\theta_e)^2}\left(\mathbb{1}\left(Y_k \leq v_k\right) - q\right)\frac{\partial e_t(\theta_e)}{\partial \theta_e}\right)\left(\frac{1}{T}\sum_{k=1}^{T} f_{Y_k}(v_k)\right)^{-1}\left(q - \mathbb{1}\left(Y_t \leq v_t\right)\right)\right|\right| \leq$$

$$\left|\frac{\partial L_{FZ0}(Y_t, v_t, e_t(\theta_e)}{\partial e_t(\theta_e)}\right| \quad \left|\left|\frac{\partial e_t(\theta_e)}{\partial \theta_e}\right|\right| +$$

$$\left(\sum_{k=1}^{T}\left|\frac{1}{q e_k(\theta_e)^2}\right| \quad \left|\mathbb{1}\left(Y_k \leq v_k\right) - q\right| \quad \left|\left|\frac{\partial e_t(\theta_e)}{\partial \theta_e}\right|\right|\right)\left(\sum_{k=1}^{T}\left|f_{Y_k}(v_k)\right|\right)^{-1}$$

$$\cdot \left|q - \mathbb{1}\left(Y_t \leq v_t\right)\right| < \infty,$$

$\mathbb{E}\left[\sup_{\theta_e \in \Theta_e} ||g(Y_t, e_t(\theta_e^0), v_t^0)||\right] < \infty$ which completes the proof.

It is shown in Proposition 1 that under assumption (vi), $\left|\frac{\partial L_{FZ0}(Y_t, v_t, e_t(\theta_e)}{\partial e_t(\theta_e)}\right| < \infty$

Proof of Lemma 2

The proof is based on the Lemma 18 in CEIN. We need to show that $\mathbb{E}\left[\left|\left|\frac{\partial g(Y_t,e_t(\theta_e^0),v_t^0)}{\partial\theta_e}\right|\right|\right] < \infty$ because the other assumptions are satisfied. Since

$$\left|\left|\frac{\partial L_{FZ0}(Y_t,v_t,e_t(\theta_e))}{\partial e_t(\theta_e)}\frac{\partial e_t(\theta_e)}{\partial^2\theta_e}+\frac{\partial L_{FZ0}(Y_t,v_t,e_t(\theta_e))}{\partial^2 e_t(\theta_e)}\frac{\partial e_t(\theta_e)}{\partial\theta_e}+\right.\right.$$

$$\left(\frac{1}{T}\sum_{k=1}^{T}\frac{-2}{qe_k(\theta_e)^3}\left(\mathbb{1}\left(Y_k\le v_k\right)-q\right)\frac{\partial e_t(\theta_e)}{\partial\theta_e}\frac{\partial e_t(\theta_e)}{\partial\theta_e}\right)\left(\frac{1}{T}\sum_{k=1}^{T}f_{Y_k}(v_k)\right)^{-1}$$

$$\cdot\left(q-\mathbb{1}\left(Y_t\le v_t\right)\right)+$$

$$\left.\left.\left(\frac{1}{T}\sum_{k=1}^{T}\frac{1}{qe_k(\theta_e)^2}\left(\mathbb{1}\left(Y_k\le v_k\right)-q\right)\frac{\partial e_t(\theta_e)}{\partial^2\theta_e}\right)\left(\frac{1}{T}\sum_{k=1}^{T}f_{Y_k}(v_k)\right)^{-1}\left(q-\mathbb{1}\left(Y_t\le v_t\right)\right)\right|\right|\le$$

$$\left|\frac{\partial L_{FZ0}(Y_t,v_t,e_t(\theta_e))}{\partial e_t(\theta_e)}\right|\quad\left|\left|\frac{\partial e_t(\theta_e)}{\partial^2\theta_e}\right|\right|+\left|\frac{\partial L_{FZ0}(Y_t,v_t,e_t(\theta_e))}{\partial^2 e_t(\theta_e)}\right|\quad\left|\left|\frac{\partial e_t(\theta_e)}{\partial\theta_e}\right|\right|+$$

$$\left(\sum_{k=1}^{T}\left|\frac{-2}{qe_k(\theta_e)^3}\right|\left|\mathbb{1}\left(Y_k\le v_k\right)-q\right|\quad\left|\left|\frac{\partial e_t(\theta_e)}{\partial\theta_e}\right|\right|\left|\left|\frac{\partial e_t(\theta_e)}{\partial\theta_e}\right|\right|\right)\left(\sum_{k=1}^{T}\left|f_{Y_k}(v_k)\right|\right)^{-1}$$

$$\cdot\left|q-\mathbb{1}\left(Y_t\le v_t\right)\right|+$$

$$\left(\sum_{k=1}^{T}\left|\frac{1}{qe_k(\theta_e)^2}\right|\left|\mathbb{1}\left(Y_k\le v_k\right)-q\right|\quad\left|\left|\frac{\partial e_t(\theta_e)}{\partial^2\theta_e}\right|\right|\right)\left(\sum_{k=1}^{T}\left|f_{Y_k}(v_k)\right|\right)^{-1}$$

$$\cdot\left|q-\mathbb{1}\left(Y_t\le v_t\right)\right|<\infty,$$

so $\mathbb{E}\left[\left|\left|\frac{\partial g(Y_t,e_t(\theta_e^0),v_t^0)}{\partial\theta_e}\right|\right|\right]<\infty$. Now since

$$\sum_{k=1}^{T}\frac{1}{2c_T}\mathbb{1}\{|Y_k-v_k|\le c_T\}\xrightarrow{p}\sum_{k=1}^{T}f_{Y_k}(v_k),$$

see e.g. Engle and Manganelli (2004), our proof is complete.

It is shown in Proposition 1 that under assumption (vi) $\left|\frac{\partial L_{FZ0}(Y_t,v_t,e_t(\theta_e))}{\partial^2 e_t(\theta_e)}\right|<\infty$.

Proof of Theorem 2

The GMM estimator $\hat{\theta}_e$ satisfies the first order conditions

$$2 \left(\frac{1}{T} \sum_{t=1}^{T} \frac{\partial g(Y_t, e_t(\hat{\theta}_e), \hat{v}_t)}{\partial \theta_e} \right) W \left(\frac{1}{T} \sum_{t=1}^{T} g(Y_t, e_t(\hat{\theta}_e), \hat{v}_t) \right) = 0 \tag{2.8}$$

By first oder Taylor expansion of $\frac{1}{T} \sum_{t=1}^{T} g(Y_t, e_t(\hat{\theta}_e), \hat{v}_t)$ around θ_e^0,

$$\frac{1}{T} \sum_{t=1}^{T} g(Y_t, e_t(\hat{\theta}_e), \hat{v}_t) = \frac{1}{T} \sum_{t=1}^{T} g(Y_t, e_t(\theta_e^0), \hat{v}_t) + \frac{1}{T} \sum_{t=1}^{T} \frac{\partial g(Y_t, e_t(\tilde{\theta}_e), \hat{v}_t)}{\partial \theta_e} \left(\hat{\theta}_e - \theta_e^0 \right),$$

where $\tilde{\theta}_e$ is between $\hat{\theta}_e$ and θ_e^0.

Substituting this into equation 2.8 and after some algebra we arrive to

$$\sqrt{T}(\hat{\theta}_e - \theta_e^0) = \left[\frac{\partial g(Y_t, e_t(\hat{\theta}_e), \hat{v}_t)}{\partial \theta_e} W \frac{\partial g(Y_t, e_t(\tilde{\theta}_e), \hat{v}_t)}{\partial \theta_e} \right] \frac{\partial g(Y_t, e_t(\hat{\theta}_e), \hat{v}_t)}{\partial \theta_e} W$$

$$T^{-1/2} \sum_{t=1}^{T} g(Y_t, e_t(\theta_e^0), \hat{v}_t).$$

Applying Lemma 1, 2 and Theorem 1:

$$\sqrt{T}(\hat{\theta}_e - \theta_e^0) = (G'WG)^{-1} G^{-1} W T^{-1/2} \sum_{t=1}^{T} g(Y_t, e_t(\theta_e^0), v_t^0) + o_p(1).$$

By Lemma 3.2 in Hall (2005),

$$T^{-1/2} \sum_{t=1}^{T} g(Y_t, e_t(\theta_e^0), v_t^0) \xrightarrow{d} N(0, \Omega),$$

which concludes the proof.

Proposition 1. *Under assumption (vi),* $\left| \frac{\partial L_{FZ0}(Y_t, v_t, e_t(\theta_e)}{\partial e_t(\theta_e)} \right| < \infty$, $\left| \frac{\partial L_{FZ0}(Y_t, v_t, e_t(\theta_e))}{\partial^2 e_t(\theta_e)} \right| < \infty$.

Proof.

$$\left|\frac{\partial L_{FZ0}(Y_t, v_t, e_t(\theta_e)}{\partial e_t(\theta_e)}\right| = \left|\frac{1}{q e_t(\theta_e)^2}\mathbb{1}\left(Y_t \leq v_t\right)\left(v_t - Y_t\right) - \frac{v_t}{e_t(\theta_e)^2} + \frac{1}{e(\theta_e)}\right| \leq$$

$$\left|\frac{1}{q e_t(\theta_e)^2}\right| \left|\mathbb{1}\left(Y_t \leq v_t\right)\right| \left(|v_t| + |Y_t|\right) + \left|\frac{v_t}{e_t(\theta_e)^2}\right| + \left|\frac{1}{e(\theta_e)}\right| < \infty.$$

$$\left|\frac{\partial L_{FZ0}(Y_t, v_t, e_t(\theta_e)}{\partial^2 e_t(\theta_e)}\right| = \left|\frac{-2}{q e_t(\theta_e)^3}\mathbb{1}\left(Y_t \leq v_t\right)\left(v_t - Y_t\right) + 2\frac{v_t}{e_t(\theta_e)^3} - \frac{1}{e(\theta_e)^2}\right| \leq$$

$$\left|\frac{-2}{q e_t(\theta_e)^3}\right| \left|\mathbb{1}\left(Y_t \leq v_t\right)\right| \left(|v_t| + |Y_t|\right) + \left|2\frac{v_t}{e_t(\theta_e)^3}\right| + \left|\frac{1}{e(\theta_e)^2}\right| < \infty.$$

□

2.7.2 Additional information on datasets

Risk-free rate construction

Since we have daily data and the option implied VaR and ES requires to have risk-free interest rate for any maturity (see the derivations in Section 2.5.2 or Mitra (2015) or Barone Adesi (2016)), we need to construct risk-free interest rate.

Following Diebold and Li (2006), we use Nelson-Siegel model (Nelson and Siegel (1987)) to estimate the yield curve for daily maturity. Our dataset includes 1-, 2-, 3-, 6-month, 1-, 2-year Treasury Constant Maturity rate. We omit longer maturity treasury bills because more than 85% of the options have less than 1 year maturity and in our application we are focusing on shorter periods. Moreover, we only used the Nelson-Siegel estimates for interest rates over 1-month. Below 30-day period, we use the 1-month treasury bill as interest rate.

Option data

This subsection describes the option data and the cleaning process. We downloaded the S&P500 Index Options (symbol: SPX) from OptionMetrics database

from 7/31/2001 until 3/31/2016. This includes 7905284 daily, European options.

In the data cleaning process, we closely follow Aït-Sahalia and Lo (1998). We take the averages of best bid and ask prices to calculate the option prices. First, we drop all the options with price less than 1/8, implied volatility greater than 70%, which yields a sample of 6321907.

Since the option implied VaR and ES requires put options with as many strike prices as possible, we derive put option prices from call options with strike prices where put option is not available. To derive these prices, we use the put-call parity relation, which must hold under no arbitrage assumption:

$$\begin{aligned}\hat{P}(S_t, K, T, r_{t,T}, \delta_{t,T}) =& C(S_t, K, T, r_{t,T}, \delta_{t,T}) + K \exp(-r_{t,T}(T-t)) \\ & - F_{t,T} \exp(-r_{t,T}(T-t)),\end{aligned}$$

where S_t is the stock price at period t, K is the strike price, T is the expiration period, $r_{t,T}$ is the risk-free interest rate at period t until period T, $\delta_{t,T}$ is dividend rate, C$(\cdot)$ is the true call price, $F_{t,T}$ is the implied futures prices and $\hat{P}(\cdot)$ is the implied put price.

To obtain the implied future prices we follow Aït-Sahalia and Lo (1998). We use the put-call parity relation with at-the-money[6] puts and calls with the same strike prices and same expiration to derive F$_{t,T}$ for all periods with all expiration. After we get the implied future prices, we can calculate the call implied put prices for those strikes, where put is not available.

Applying the above formula instead of $\hat{P}(S_t, K, T, r_{t,T}, \delta_{t,T}) = C(S_t, K, T, r_{t,T}, \delta_{t,T}) + K \exp(-r_{t,T}(T-t)) - S_t$, which includes stock prices instead of implied futures prices is motivated by 2 related facts. First, when we used the formula with stock prices instead of implied future prices, we got negative "call implied" put prices, which might be explained by the second fact. As Aït-Sahalia and Lo (1998) also argue,

[6] Since in-the-money options are less liquid than at-the-money options, Aït-Sahalia and Lo (1998) argue that at-the-money option pairs have the most reliable prices.

there is no guarantee that the option prices are recorded at the same time as the closing price of the underlying index. And these mismatches can lead to negative "call implied" put prices as in our case.

2.7.3 More words on simulation

Setting starting values

The FZ0 loss function (Equation (2.1)) involves the indicator function, therefore we used fminsearch.m function in MatLab 2017a to minimize the estimator. However, the fminsearch function requires initial starting values to start the search from. In our simulation study, in each simulation we set 20 different starting values drawn from the $\theta_0 \times \mathcal{U}[0.5\,1.5]$ set, where θ_0 are the true parameters. Then we picked those estimates as optima which produced the smallest loss.

Sanity checks in simulation

We also require that the estimates from any starting value satisfy all the sanity checks:

1. The value of GMM is a real number
2. The values of the estimated parameters are real numbers
3. $\min(VaR_{t,q}) \geq -100$, $\forall t$
4. $\min(ES_{t,q}) \geq -100$, $\forall t$
5. $ES_{t,q} \leq VaR_{t,q}$, $\forall t$
6. $\exists Y_t$ such that $Y_t \leq VaR_t$
7. $\exists Y_t$ such that $Y_t \leq ES_t$

In the AR(1)-TS-ARCH(1) model we also require

8. $|\phi| \leq 0.999$

9. $\omega > 0$

10. $0 \leq \beta \geq 1$

If a sanity check is not satisfied from a given starting value, then we drop those starting values and draw a new one. We repeat these checks until we find 20 estimates which satisfy all the sanity checks.

3

A Nonparametric Specification Test of Value-at-Risk Models

3.1 Introduction

The most prevalent market risk measures used in the financial sector is the Value-at-Risk (VaR). VaR can be interpreted as a lower (1%-5%) quantile of the return of a portfolio or as a high (95%-99%) quantile of the losses of the same portfolio. Financial firms are required to report their VaR estimates to the regulatory authorities since its introduction to the Basel I Accords in 1996.

Our main goal in this paper is to provide a nonparametric specification test for this measure. Applying this technique, a modeller could (back)test whether her model can correctly describe the VaR process of the underlying assets/portfolio or a regulator could make a decision whether the reported VaR estimates of a bank are realistic.

A specification test for (conditional) quantiles can be equivalently considered as a test for conditional moment restrictions. Consider a model whose estimates of the q-VaR (or VaR at q confidence level, where q is the 1%-5% or 95%-99% as mentioned

before) are $\widehat{VaR}_{t+1,q}$ for some portfolio with returns, Y_{t+1}. If these estimates satisfy the following conditional moment restriction:

$$\mathbb{E}\left[\mathbb{1}\left(Y_{t+1} \leq \widehat{VaR}_{t+1,q}\right) - q | X_t = x\right] = 0, \quad x \in \mathcal{X},$$

then we can claim that the model is correctly specified for the q-VaR risk measure. Here X_t is a conditioning variable available at time t, such as past return(s) Y_t or volatility etc. Therefore, if we nonparametrically regress the moment, $\mathbb{1}\left(Y_{t+1} \leq \widehat{VaR}_{t+1,q}\right) - q$, on the conditioning variable X_t, the conditional expectation of the resulting function should be uniformly zero. Our goal is to construct a test which examine whether the resulting function is actually zero.

This problem is functional in nature, because we try to do inference on the whole function instead of focusing on some exact points. Li and Liao (2019) propose a uniform nonparametric inference method for time-series data based on series regression. Applying their method we can regress the moment on asymptotically growing number of approximating functions of X_t. However, because of the growing dimension, we cannot apply the regular Functional Central Limit Theorem to do functional inference. That is, this problem is non-Donsker. Li and Liao (2019) overcome this issue by developing a strong Gaussian approximation theory for dependent data, which they use to characterize the asymptotic behavior of the sup-t statistic for functional inference. However, their method cannot be directly applied to our problem because of two issues.

First, the moment condition, $\mathbb{1}\left(Y_{t+1} \leq \widehat{VaR}_{t+1,q}\right) - q$, might depend on some estimated parameters since $\widehat{VaR}_{t+1,q}$ is derived from a possibly parametric model. In this paper, we provide sufficient conditions so that the estimation error in the modelling step becomes asymptotically negligible. Intuitively, the estimated parameters (or estimation error) converges at a parametric, $\sqrt{n}$, rate which is faster than the convergence rate of the nonparametric test. In addition, our conditions also allow

for non-smooth transformation of the estimated parameters, which is needed in our setting, as they enter through the indicator function $\mathbb{1}(\cdot)$. Moreover, we also allow for generated conditioning variables, X_t since they might be also an outcome of a certain model. As in our simulations, it includes GARCH implied volatility.

Second, to compute the critical values of the test statistic, Li and Liao (2019) rely on the limiting distribution of the sup-t test. In our simulation; however, we find that applying the Gaussian approximation can lead to size distortions at higher confidence levels ($q \geq 90\%$). That is, using their method we reject the correct models much more frequently than the nominal rate of the test. To overcome this issue, we propose an i.i.d. bootstrap technique to calculate the critical values. Note, we can use i.i.d. bootstrap even in dependent data as the residual series $\mathbb{1}\left(Y_{t+1} \leq \widehat{VaR}_{t+1,q}\right) - q$ is a martingale difference sequence under the null hypothesis.

The rest of the paper is organized as follows. Section 3.2 provides a short review of the VaR testing literature. Section 3.3 describes the theory of our test. In Section 3.4 we examine the finite sample properties of our test in a Monte Carlo simulation study. Section 3.5 includes an empirical exercise with the proposed test. Section 3.6 concludes. The appendix contains all proofs.[1]

3.2 Literature review

This section gives a short review of the Value-at-Risk testing literature. A big part of the academic literature does not directly test whether the moment condition $\mathbb{E}\left[\mathbb{1}\left(Y_{t+1} \leq VaR_{t+1,q}\right) - q\right] = 0$ holds but it tests an implication of it:
$\mathbb{1}\left(Y_{t+1} \leq VaR_{t+1,q}\right)$ are an i.i.d. Bernoulli(q) random variable.

Applying this testing procedure in their seminal paper, McNeil and Frey (2000) estimate the VaR by combining the quasi maximum likelihood approach with extreme

[1] In this project, I mainly contributed to the Monte Carlo simulation study and empirical exercise. The proofs and underlying theorems are the work of my co-authors.

value theory (EVT) and test 95%-, 99%- and 99.5%-VaR in several different asset classes. Implementing EVT leads better out-of-sample forecast than ignoring the fat tail behavior of the return process.

Escanciano and Olmo (2010) describe a parametric unconditional backtesting procedure for out-of-sample VaR which controls for estimation risk but not for model risk. That is, they work under the assumption that the underlying model is correctly specified and their test explicitly controls for the parameter estimation error in the testing step. Our paper is different in a sense that we test for the correct specification of the model directly and we do not assume it.

In their following work, Escanciano and Olmo (2011) also control for the model risk. That is, they modify the variance-covariance matrix of the limiting distribution of the test statistic such that it includes an additional term which accounts for the possible misspecification of the VaR model. In our test, this is not necessary.

In their seminal paper, Engle and Manganelli (2004) develop the Conditional Autoregressive Value-at-Risk model (CAViaR). The novelty of their method is they directly model the (extreme) quantiles. In our empirical section, we test the CoVaR measure from Tobias and Brunnermeier (2016), who also directly targets the extreme quantiles. Engle and Manganelli (2004) also apply a new test for quantile testing (Dynamic Quantile(DQ) test) by testing whether $\sqrt{n}X_t \left[\mathbb{1}\left(Y_{t+1} \leq VaR_{t+1,q}\right) - q\right]$ is significantly different from zero, where they define X_t as the lagged value of the $\mathbb{1}\left(Y_{t+1} \leq VaR_{t+1,q}\right) - q$ function. They implement the DQ test since testing the above implication is only a necessary but not sufficient condition for the moment restriction to hold. Their tests, however, requires the estimation of the conditional density of the error term at the tail, which is not necessary in our paper.

Kuester et al. (2006) combines these different testing approaches (likelihood ratio test to examine whether $\mathbb{1}\left(Y_{t+1} \leq VaR_{t+1,q}\right)$ is an i.i.d. Bernoulli(q) random variable, and the DQ test) to investigate which models can correctly estimate the VaR in a

NASDAQ market portfolio. Their findings suggest that in parametric model family one should allow for more flexible tail behavior (e.g. by using skew-t GARCH model) and one could improve CaViAR models by estimating the underlying return process simultaneously with the VaR.

3.3 Theory

This section describes the underlying theory of the nonparametric specification test of Value-at-Risk models. Our main goal in this paper is to answer the question whether a VaR model is correctly specified. That is,

$$H_0 : \text{VaR is correctly specified.}$$

We can try to make this theoretical exercise clearer through concrete example. Consider a time series of portfolio returns, Y_{t+1}. A risk manager's task may be to model the q-VaR of this portfolio and she proposes a model, $f_t(\theta)$ with estimated outputs $\widehat{VaR}_{t+1,q}(\hat{\theta})$, which she believes, describes the VaR process of the portfolio. If the outputs of her model are correct, then the following moment condition should hold:

$$\mathbb{E}\left[\mathbb{1}\left(Y_{t+1} \leq \widehat{VaR}_{t+1,q}(\hat{\theta})\right) - q \middle| X_t = x\right] = 0, \qquad \forall x \in \mathcal{X}, \tag{3.1}$$

for some conditioning variable, X_t. Since the true q-VaR process of the portfolio must satisfy the above conditional moment by definition of the conditional quantile. The main question of this section is how to test whether the above moment condition holds.

3.3.1 The testing problem: infeasible case

In more general form, the above conditional moment restriction (Equation (3.1)) can be equivalently written as

$$\mathbb{E}\left[Z^{\star}_{t+1} \middle| X^{\star}_t = x\right] = 0, \qquad \forall x \in \mathcal{X},$$

where x takes values in some space $\mathcal{X}$ and

$$Z_{t+1}^{\star} = \mathbb{1}(Y_{t+1} \leq f_t(\theta^{\star})) - q.$$

Following Li and Liao (2019), we can rewrite this problem as

$$\mathbb{E}\left[Z_{t+1}^{\star} | X_t^{\star} = x\right] = h(x) = 0, \qquad \forall x \in \mathcal{X},$$

that is, we can test whether the $h(x)$ is identically zero.

This problem is functional in nature, because we try to do inference on the whole function instead of focusing on some exact points of this function. To accomplish this, we build on the theory of Li and Liao (2019)'s method based on nonparametric series regression. However, their method is not directly applicable in our setting partly because $Z_{t+1}^{\star}$ and $X_t^{\star}$ might be not directly observable, but we either need to estimate it since the VaR model might depend on some parameters θ or the conditioning variables are model generated (e.g. GARCH implied volatility in our simulations). Therefore, we need to make several extension to their method.

However, it might be easier to go through the building block of our theory in the infeasible case, when both $Z_{t+1}^{\star}$ and $X_t^{\star}$ are observed. First, let's consider a vector of basis functions $P(x) = (p_1(x), p_2(x), \ldots, p_{m_n}(x))^{\top}$, where the number of series term m_n tends to ∞ as n, the sample size, tends to ∞. Technically, we could use simple polynomials, trigonometric functions, splines etc. as basis functions, see e.g. Chen (2007).[2] To estimate $h(x)$, we regress $Z_{t+1}^{\star}$ on $P(X_t^{\star})$ and obtain the regression coefficients:

$$\widehat{b}_n^{\star} = \left(\sum_{t=1}^{n} P(X_t^{\star}) P(X_t^{\star})^{\top}\right)^{-1} \left(\sum_{t=1}^{n} P(X_t^{\star}) Z_{t+1}^{\star}\right).$$

The nonparametric estimator for $h(\cdot)$ is then given by

$$\widehat{h}_n^{\star}(x) = P(x)^{\top} \widehat{b}_n^{\star}, \quad x \in \mathcal{X},$$

[2] In our simulations, we found that Legendre polynomials worked the best for extreme quantiles.

where $\dim(\widehat{b}_n^\star) = m_n$. Since $m_n \to \infty$, the $P(x)^\top \widehat{b}_n^\star$ becomes increasingly flexible. By well-known approximation theory, the unknown function $h(x)$ can be uniformly approximated by $P(x)^\top b_n^\star$. However, as the dimension of the coefficients increases with the sample size, we cannot apply the Functional Central Limit Theorem to approximate the limiting distribution of the $P(x)^\top b_n^\star$ function, and to make statistical inference on the $\widehat{h}_n^\star(x)$ function. That is, this problem is non-Donsker.

Li and Liao (2019) address this issue by establishing a strong Gaussian approximation for the normalized estimator $n^{1/2}(\widehat{b}_n^\star - b_n^\star)$. Under certain regularity conditions, they construct a sequence of m_n-dimensional standard Gaussian random vector ξ_n such that

$$\left\| n^{1/2}(\widehat{b}_n^\star - b_n^\star) - \Sigma_n^{1/2}\xi_n \right\| = o_p(\delta_n)$$

for some sequence $\delta_n = o(1)$. Here, Σ_n is the $m_n \times m_n$ "pre-asymptotic" variance-covariance matrix of the regression coefficient that is given by

$$\Sigma_n = Q_n^{-1} A_n Q_n^{-1},$$

where

$$Q_n = n^{-1} \sum_{t=1}^{n} \mathbb{E}\left[P(X_t^\star)P(X_t^\star)^\top\right],$$

$$A_n = \mathbb{E}\left[\left| n^{-1/2} \sum_{t=1}^{n} P(X_t^\star)\left(Z_{t+1}^\star - h(X_t^\star)\right)\right|^2\right].$$

Therefore, we can strongly approximate $n^{1/2}(\widehat{h}_n(x) - h(x))$ by the Gaussian process $P(x)^\top \Sigma_n^{1/2}\xi_n$, where $\xi_n \sim N(0, I_n)$.

However, we are facing two issues with the above inference method in our paper. First, we need to replace $\left(Z_{t+1}^\star, X_t^\star\right)$ using the corresponding feasible proxies. Second, we find in our simulations that using the Gaussian approximation method to compute the critical values of the test, leads to size distortion at more extreme quantiles as Z_{t+1}

becomes more skewed. This motivates us to develop and apply an i.i.d. bootstrap algorithm to compute the critical values of our test.

3.3.2 Feasible inference via i.i.d bootstrap

In this subsection, we propose our feasible inference method and justify its asymptotic validity. To emphasize the dependence of $\left(Z_{t+1}^{\star}, X_t^{\star}\right)$ on the finite dimensional pseudo-true parameter $\theta^{\star}$, we write

$$Z_{t+1}^{\star} = Z_{t+1}(\theta^{\star}), \quad X_t^{\star} = X_t\left(\theta^{\star}\right),$$

where $Z_{t+1}\left(\theta\right) = \mathbb{1}(Y_{t+1} \leq f_t\left(\theta\right)) - q$ in our problem, and $X_t\left(\theta\right)$ can be e.g. a GARCH model implied volatility. We assume that $\theta^{\star}$ can be estimated with some $n^{1/2}$-consistent estimator, $\hat{\theta}_n$. Therefore, the feasible inference relies on the following generated variables:

$$\widehat{Z}_{t+1} = Z_{t+1}(\hat{\theta}_n), \quad \widehat{X}_t = X_t(\hat{\theta}_n).$$

To construct the test statistic, we proceed as previously: first, we regress $\widehat{Z}_{t+1}$ on $P(\widehat{X}_t)$ and obtain the regression coefficient as

$$\widehat{b}_n \equiv \left(\sum_{t=1}^{n} P(\widehat{X}_t)P(\widehat{X}_t)^{\top}\right)^{-1}\left(\sum_{t=1}^{n} P(\widehat{X}_t)\widehat{Z}_{t+1}\right).$$

The conditional expectation function $h\left(x\right)$ is then estimated by

$$\widehat{h}_n\left(x\right) = P\left(x\right)^{\top}\widehat{b}_n,$$

with residual $\hat{u}_t = \widehat{Z}_{t+1} - \widehat{h}_n(\widehat{X}_t)$. The standard error function is estimated by

$$\hat{\sigma}_n\left(x\right) = \left(P\left(x\right)^{\top}\widehat{\Sigma}_n P\left(x\right)\right)^{1/2},$$

where $\widehat{\Sigma}_n = \hat{Q}_n^{-1}\hat{A}_n\hat{Q}_n^{-1}$ and

$$\hat{Q}_n = n^{-1}\sum_{t=1}^{n} P(\widehat{X}_t)P(\widehat{X}_t)^{\top}, \quad \hat{A}_n = n^{-1}\sum_{t=1}^{n} \hat{u}_t^2 P(\widehat{X}_t)P(\widehat{X}_t)^{\top}.$$

The resulting sup-t statistic is

$$\widehat{T}_n = \sup_{x \in \mathcal{X}} \frac{n^{1/2} \left| \widehat{h}_n(x) \right|}{\hat{\sigma}_n(x)}.$$

In the following, we provide sufficient conditions under which replacing $\theta^\star$ with $\hat{\theta}_n$ results in negligible effect for the nonparametric specification test. Our goal is basically to provide sufficient conditions such that the estimated $\hat{\theta}_n$ converges to the true parameter $\theta^\star$ at a parametric $\sqrt{n}$ rate, which is faster than the nonparametric convergence rate in the second, testing step. We extend the similar assumption in Li and Liao (2019) to allow for non-smooth transformation of the parameter vector and for generated regressors.

Let $B_n(\theta^\star) = \{\theta \in \Theta : \|\theta - \theta^\star\| \leq \delta_{c,n}\}$ where $\delta_{c,n} = c_n^{1/2} n^{-1/2}$ and c_n is a slowly divergent positive sequence. Our first condition is on the first-step parametric estimator $\hat{\theta}_n$.

Assumption 5. *(i)* $\hat{\theta}_n - \theta^\star = O_p(n^{-1/2})$ *where* $\theta^\star$ *lies in* Θ *and* $\Theta \subset \mathbb{R}^{d_\theta}$ *is a compact set.*

(ii) For any t *there exists a random variable* $L_{1,t}$ *such that* $\|X_t(\theta_1) - X_t(\theta_2)\| \leq L_{1,t} \|\theta_1 - \theta_2\|$ *for any* $\theta \in B_n(\theta^\star)$.

(iii) For any t *and any* $y_1, y_2 \in \mathbb{R}$, $\left|F_{t+1|t}(y_1) - F_{t+1|t}(y_2)\right| \leq K |y_1 - y_2|$ *where* $F_{t+1|t}(\cdot)$ *denotes the conditional CDF of* Y_{t+1} *given* $\mathcal{F}_t$.

(iv) For any t *there exists a random variable* $L_{2,t}$ *such that* $|f_t(\theta_1) - f_t(\theta_2)| \leq L_{2,t} \|\theta_1 - \theta_2\|$ *for any* $\theta \in B_n(\theta^\star)$.

(v) $\max_{t \leq n} \mathbb{E}[L_{1,t}^p + L_{2,t}^p] \leq K$ *for some* $p > 2d_\theta$.

Assumption 6. *(i)* $((f_t(\theta^\star), X_t^\star))_t$ *is a strong mixing process with compact support* $\mathcal{X}$ *and mixing coefficients* $(\varphi_s)_{s=0}^\infty$ *satisfying* $\sum_{s=1}^\infty \varphi_s^{1-2/\kappa} \leq K$ *for some finite constant* $\kappa > 2$.

(ii) There exist $\rho_h > 0$ *and* $b_n^\star \in \mathbb{R}^{m_n}$ *such that*

$$\sup_{x \in \mathcal{X}_{\varepsilon_n}} |h_{m_n}(x) - h(x)| = O(m_n^{-\rho_h})$$

where $h(x) = \mathbb{E}[Z_{t+1}^\star | X_t = x]$ *and* $h_{m_n}(x) = P(x)^\top b_n^\star$.

(iii) $h(x)$ *is continuous differentiable.*

(iv) The eigenvalues of Q_n *and* A_n *are between* K^{-1} *and* K *for all* m_n.

(v) $\max_{l \leq m_n} \sup_{x \in \mathcal{X}_{\varepsilon_n}} |\partial^j p_l(x)| \leq \zeta_{j,n}$ *for* $j = 1, 2$, *where* $\zeta_{j,n}$ *is a non-decreasing sequence.*

(vi) $(\zeta_{0,n}^2 m_n^2 + \zeta_{1,n} m_n) n^{-1/2} + n^{1/2} m_n^{-\rho_h} = o((\log(n))^{-1})$.

Theorem 3. *Suppose that Assumptions 5 and 6 hold. Then under the null hypothesis there exists a sequence* ξ_n *of* m_n*-dimensional standard normal random variables such that*

$$\widehat{T}_n - \widetilde{T}_n = o_p\left((\log m_n)^{-1/2}\right).$$

where

$$\widetilde{T}_n = \sup_{x \in \mathcal{X}} \frac{\left| P(x)^\top \Sigma_n^{1/2} \xi_n \right|}{\sigma_n(x)}.$$

Theorem 3 establishes the strong approximation of the sup-t statistic. The theorem also shows that the approximation error vanishes at a rate faster than $(\log m_n)^{1/2}$, which is necessary for analyzing the size property of our test.

We could use this theorem above to compute the critical values similarly to Li and Liao (2019). We would simply compute the quantile of the approximating variable $\widetilde{T}_n$ by simulating the Gaussian process. However, we will see in the simulation that at more extreme quantiles, this approach has larger size distortions. Therefore, we propose the following i.i.d. bootstrap method to compute the critical values:

1. Resample $(\bar{Z}_{t+1}, \bar{X}_t)_{1\leq t\leq n}$ as i.i.d. sample with replacement from $(\widehat{Z}_{t+1}, \widehat{X}_t)_{1\leq t\leq n}$.

2. Compute the sup-t statistic $\left(\bar{T}_n\right)$ for the new sample, $(\bar{Z}_{t+1}, \bar{X}_t)_{1\leq t\leq n}$.

3. Repeat step 1 and 2 many times. Set the critical value, cv, at significant level α, as the $1-\alpha$ quantile of $\bar{T}_n$s in the Monte Carlo sample.

4. Reject the null hypotheses $(h(x)=0)$ if $\widehat{T}_n > cv$.

Theorem 4 establishes the asymptotic property of the bootstrapped test statistic, and shows the size and power properties of the test.

Theorem 4. *Suppose that Assumptions 5 and 6 hold. Then,*

(a) under the null there exists a sequence ξ_n^* of m_n-dimensional standard normal random variables such that

$$\widehat{T}_n^* - \widetilde{T}_n^* = o_p\left((\log m_n)^{-1/2}\right),$$

where

$$\widetilde{T}_n^* = \sup_{x\in\mathcal{X}} \frac{\left|P(x)^\top \Sigma_n^{1/2}\xi_n^*\right|}{\sigma_n(x)}.$$

(b) for $\alpha \in (0, 1/2)$, the level-α test described in Algorithm 1 has size α under the null hypothesis and has asymptotic power one under the alternative (i.e., $h(x) \neq 0$ for some $x \in \mathcal{X}$).

3.4 Simulation study

In this section, we examine the finite sample properties of the proposed test in Monte Carlo studies for VaR models. Section 3.4.1 describes the simulation settings, Section 3.4.2 presents the conditioning variables and the series polynomials and the Section 3.4.3 reports the results.

3.4.1 The simulation settings

We consider three data generating processes (DGP) corresponding to the null hypothesis (Size) and alternative hypotheses (Power 1 and 2). Under the first DGP (Size), we generate a time series Y_t of daily losses from a GARCH(1,1) model with normally distributed errors (Size):

$$Y_t = \sigma_t \varepsilon_t, \quad \varepsilon_t \overset{i.i.d.}{\sim} \mathcal{N}(0,1),$$
$$\sigma_t^2 = \omega + \beta \sigma_{t-1}^2 + \gamma Y_{t-1}^2,$$

where $\{\omega, \beta, \gamma\} = \{0.05, 0.9, 0.05\}$.

The second DGP (Power 1) is Student-t EGARCH(1,1,1) given by

$$Y_t = \sigma_t z_t, \quad z_t \overset{i.i.d.}{\sim} t(0,1,6)$$
$$\ln(\sigma_t^2) = \omega + \gamma \left(|z_{t-1}| - \sqrt{2/\pi} \right) + \delta z_{t-1} + \beta \ln(\sigma_{t-1}^2),$$

where we calibrate the parameters to the S&P500 index data by setting $\{\omega, \beta, \gamma, \delta\} = \{8.9551 \times 10^{-4}, 0.9782, 0.1350, -0.1637\}$.

The last DGP (Power 2) we use a similar DGP to Bontemps (2019), which is also a Student-t EGARCH(1,1,1):

$$Y_t = \sigma_t z_t, \quad z_t \overset{i.i.d.}{\sim} t(0,1,4)$$
$$\ln(\sigma_t^2) = \omega + \gamma \left(|z_{t-1}| - \sqrt{2/\pi} \right) + \delta z_{t-1} + \beta \ln(\sigma_{t-1}^2),$$

where $\{\omega, \beta, \gamma, \delta\} = \{0.0001, 0.9, 0.3, -0.8\}$.

We set the sample size, n, 2000 which is in similar quantity as the number of observations in the empirical section.

To test the size (and power) properties of the test, we estimate GARCH(1,1) model with maximum likelihood and then compute the q-VaR as

$$VaR_{t+1,q} = \Phi^{-1}(q) \sqrt{\hat{\omega} + \hat{\beta} \hat{\sigma}_t^2 + \hat{\gamma} Y_t^2},$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution. That is, q-VaR is correctly specified under the first DGP (Size) but misspecified under the other two DGPs (Power 1 and 2). To test the sensitivity of our test with respect to the confidence level of VaR we set $q \in \{0.75, 0.9, 0.95, 0.99\}$. We run our test at 5% significance level, that is, we expect that the test rejects 5% of the cases under the Size DGP.

3.4.2 The conditioning variables (X_t) and the series polynomials (P(x))

As conditioning variables, we use the previous day returns Y_{t-1} and volatility σ_{t-1} as these two variables may successfully control for the location and the scale of the GARCH models.

To mitigate the effect of possible multicollinearity in the series regression, we use Legendre polynomials of the normalized lagged returns and volatility. To construct the $P(x)$, we proceed in the following steps:

1. Normalize Y_{t-1} and σ_{t-1} to [-1, 1] interval, where Legendre polynomials are orthogonal:
   (a) Rank the observations from 1 to n ($ranked\,X$), where n is the number of observations and X is Y_{t-1} and σ_{t-1} separately
   (b) Normalize them to [-1, 1] as $normalized\,X = 2\frac{rankedX - \min(rankedX)}{\max(rankedX) - \min(rankedX)} - 1$
2. Apply the m_n order Legendre polynomial on the $normalized\,X$

To examine how the number of the terms in the series polynomial might have an effect on the testing procedure, we apply Legendre polynomial with degree of 1, 2 and 3 in the following way:

- $P_1(X_t) = \left(1, \tilde{Y}_t^1, \tilde{\sigma}_t^1, \tilde{Y}_t^1 \tilde{\sigma}_t^1\right)$

- $P_2(X_t) = \left(P_1(X_t), \tilde{Y}_t^2, \tilde{\sigma}_t^2\right)$
- $P_3(X_t) = \left(P_2(X_t), \tilde{Y}_t^3, \tilde{\sigma}_t^3, \tilde{Y}_t^2\tilde{\sigma}_t^1, \tilde{Y}_t^1\tilde{\sigma}_t^2\right)$

where $\tilde{Y}_t^{m_n}$ and $\tilde{\sigma}_t^{m_n}$ denotes the m_n^{th} order Legendre polynomial. Note $P_1(X_t)$ includes 4 terms, $P_2(X_t)$ 6 terms, $P_3(X_t)$ 10 terms.

To compare the bootstrap method with the Gaussian approximation, we use the Algorithm 2 from Li and Liao (2019), which computes the critical values using Gaussian approximation. We draw 1000 random samples/values to compute the critical values using the bootstrap and the Gaussian method.

3.4.3 Results

Table 3.1 reports the rejection rates under the 3 DGPs. The top panel corresponds to the case where the VaR model is correctly specified (Size) and the bottom two panels report the results when the VaR model is misspecified (Power 1 and 2). Both the Gaussian approximation and the bootstrap method have very good size properties when $q = 0.75$. However, as q increases and we estimate VaR closer to the end of the tail, the Gaussian approximation method rejects much more often than the nominal 5% rate while our proposed bootstrap method performs adequately.

As we can see from the bottom two panels, both approaches detect the misspecification correctly especially when $q \leq 0.95$. The Gaussian approximation method might perform better in the most extreme case when $q = 0.99$; however, we could see that this method overrejects under correct specification of the VaR model so these results might be led by the frequent rejection of the Gaussian method.

Table 3.1: Simulation results. Rejection Rates at 5% level

	Gaussian			Bootstrap		
	$P_1(x)$	$P_2(x)$	$P_3(x)$	$P_1(x)$	$P_2(x)$	$P_3(x)$
	Size					
$q = 0.75$	0.057	0.056	0.066	0.052	0.048	0.053
$q = 0.90$	0.066	0.083	0.094	0.055	0.046	0.041
$q = 0.95$	0.074	0.099	0.183	0.044	0.043	0.055
$q = 0.99$	0.211	0.356	0.634	0.023	0.032	0.012
	Power 1					
$q = 0.75$	0.902	0.788	0.662	0.898	0.755	0.602
$q = 0.90$	0.852	0.757	0.717	0.811	0.687	0.528
$q = 0.95$	0.491	0.500	0.554	0.392	0.337	0.232
$q = 0.99$	0.264	0.216	0.366	0.036	0.041	0.038
	Power 2					
$q = 0.75$	1	1	1	0.982	0.982	0.982
$q = 0.90$	1	1	1	0.978	0.978	0.977
$q = 0.95$	1	1	1	0.979	0.977	0.976
$q = 0.99$	0.993	0.990	0.935	0.824	0.577	0.257

3.5 Empirical application

In this section, we use the proposed test to examine specifications of the CoVaR model proposed by Tobias and Brunnermeier (2016). Section 3.5.1 describes the setting and Section 3.5.2 presents the findings.

3.5.1 CoVaR

Conditional Value-at-Risk or CoVaR was introduced by Tobias and Brunnermeier (2016) to measure the systematic risks of financial institutions. They define q-CoVaR of firm i as the q-VaR of the market conditional on that firm i's loss exceeds its q-VaR. To estimate q-CoVaR, they proceed in 2 steps: first, they estimate the conditional

VaR of firm i, then they estimate the conditional VaR of the market using the loss of firm i as an additional control variable. Then they plug in the estimated q-VaR of firm i to get the estimated CoVaR.

In our empirical exercise we use the same dataset as Tobias and Brunnermeier (2016).[3] This dataset includes 1823 publicly traded US commercial banks, broker-dealers, insurance companies, and real estate companies for the period from 1/1/1971 to 6/1/2013 (2209 weeks). On average, a firm has 736 weekly observations available in the dataset. We conduct specification tests for each firm separately and then compute the average rejection rates.

Following Tobias and Brunnermeier (2016), we can define q-VaR of firm i and the q-CoVaR as:

$$VaR^i_{t+1,q} = \alpha^i_q + \beta^i_q \mathcal{M}_t, \tag{3.2}$$

and

$$CoVaR^i_{t+1,q} = \alpha^i_q + \beta^i_q \mathcal{M}_t + \gamma^i_q Y^i_{t+1}, \tag{3.3}$$

where Y^i_{t+1} is the loss of firm i in period $t+1$, $\mathcal{M}_t$ includes the industry-wide control variables such as the weekly real estate sector return (Housing), the weekly market return of the S&P 500 index (MktRet), short-term TED spread (TED), change in the credit spread (Credit), change in three-month yield (Yld3M), change in the slope of the yield curve (TERM), and equity volatility (MktSD) and the CoVaR estimation is based on market losses.[4]

In the following we will refer to Equation (3.2) as VaR and Equation (3.3) as the CoVaR specification.

[3] We downloaded the data from the AER website.

[4] More detail on these variables can be found on page 1718 and 1719 in Tobias and Brunnermeier (2016)

3.5.2 Results

This section presents the results of the specification test for VaR and CoVaR. To capture the economic uncertainty, we use separately the Economic Policy Uncertainty Index (EPU), and macro (MUI) and financial (FUI) uncertainty indices from Jurado et al. (2015) as conditioning variable, X_t. Similarly to the simulation section, we apply Legendre polynomials of the conditioning variable and we conduct the specification tests at 5% significance level.

The rejection rates of the test for the correct specification of VaR and CoVaR (as defined in Equation (3.2) and (3.3)) is in Table 3.2. To examine the sensitivity of the test with respect to the number of terms in the polynomial, we report the results for $m_n = \{4, 5, \ldots, 10\}$. As we can see from this Table, the 95%-VaR model is rejected more frequently than the nominal 5%, which would imply that the above specification is not able to capture the true 95%-VaR. However, the above specification might be correct for the 99%-VaR as the rejection rates are very close to the 5% nominal rate. The conclusion is a bit different for the CoVaR model. The above specification might work well for the 95%-CoVaR; however, our test rarely rejects the 99%-CoVaR model. These conclusions hold regardless of the number of series terms in the Legendre polynomial.

Table 3.2: Empirical Rejection Rates using Different Conditioning Variables

			VaR			
X_t :	EPU		MUI		FUI	
m_n	95%	99%	95%	99%	95%	99%
4	0.134	0.085	0.188	0.093	0.188	0.075
5	0.152	0.082	0.213	0.089	0.194	0.075
6	0.165	0.068	0.206	0.079	0.204	0.072
7	0.166	0.067	0.193	0.077	0.200	0.058
8	0.170	0.064	0.188	0.066	0.205	0.054
9	0.163	0.055	0.183	0.064	0.196	0.051
10	0.157	0.051	0.162	0.055	0.179	0.047

			CoVaR			
X_t :	EPU		MUI		FUI	
m_n	95%	99%	95%	99%	95%	99%
4	0.036	0.037	0.072	0.037	0.032	0.032
5	0.048	0.033	0.071	0.029	0.039	0.029
6	0.064	0.030	0.072	0.031	0.049	0.017
7	0.069	0.017	0.073	0.021	0.046	0.015
8	0.074	0.009	0.075	0.020	0.050	0.014
9	0.078	0.007	0.065	0.018	0.045	0.010
10	0.072	0.007	0.053	0.014	0.035	0.008

As we have seen in Table 3.2, the CoVaR model was rarely rejected. This finding motivates us to find a more parsimonious model for this risk measure. That is, we investigate whether it is possible to find a model which includes less variables as the basic specification in Equation (3.3). To answer this question, we reestimate the CoVaR equation first using neither of the variables from $\mathcal{M}_t$, then including only 1 of the 7 state variables. Table 3.3 presents our findings. If neither of the industry-wide state variables is included in the estimation (None block), the CoVaR model gets rejected much more frequently as the nominal 5% rate, which implies that

some of the state variables carries valuable information regarding this risk measure. If we include e.g. only Housing as a state variable, the 95%-CoVaR model is still rejected for approximately 1/3 of all the firms and the 99%-CoVaR is also rejected at 5%-10% rate. We only observe a larger drop in the rejection rates when the equity volatility (and only that) is included in the $\mathcal{M}_t$ (MktSD). This result might point to the direction that the equity volatility drives the low rejection rate of the CoVaR measure. Therefore, we estimate the CoVaR model using all variables from $\mathcal{M}_t$ except the MktSD. As we can see from the last block in Table 3.3, the rejection rate is almost as high as when neither of the state variables included, which might confirm the importance of the equity volatility in the CoVaR estimation.

Table 3.3: Empirical Rejection Rates of CoVaR with Different Control Variables from $\mathcal{M}_t$

	None		Housing		Mkt Return		TED		Credit	
m_n	95%	99%	95%	99%	95%	99%	95%	99%	95%	99%
4	0.241	0.218	0.321	0.207	0.204	0.229	0.212	0.137	0.185	0.228
5	0.265	0.180	0.325	0.155	0.180	0.201	0.213	0.084	0.210	0.181
6	0.293	0.167	0.346	0.115	0.194	0.159	0.255	0.089	0.246	0.161
7	0.329	0.158	0.351	0.077	0.279	0.139	0.249	0.057	0.292	0.148
8	0.309	0.144	0.340	0.059	0.289	0.147	0.257	0.031	0.276	0.138
9	0.287	0.121	0.325	0.069	0.266	0.150	0.250	0.030	0.267	0.123
10	0.301	0.083	0.324	0.038	0.274	0.127	0.230	0.013	0.267	0.091

	Yield3m		Term		MktSD		All but MktSD	
m_n	95%	99%	95%	99%	95%	99%	95%	99%
4	0.258	0.193	0.270	0.217	0.030	0.032	0.217	0.217
5	0.291	0.165	0.282	0.194	0.041	0.013	0.234	0.173
6	0.332	0.127	0.335	0.166	0.053	0.013	0.252	0.123
7	0.366	0.114	0.371	0.143	0.052	0.009	0.274	0.109
8	0.375	0.081	0.370	0.125	0.055	0.008	0.278	0.084
9	0.342	0.087	0.354	0.110	0.057	0.004	0.268	0.078
10	0.332	0.059	0.339	0.062	0.052	0.005	0.283	0.049

3.6 Conclusion

Value-at-Risk has been the main market risk measure since its introduction to Basel I Accord, which require financial institutions to report their VaR estimates to the regulatory authorities. In this paper, we proposed a nonparametric specification test for VaR models that is based on conditional moment restrictions. We test whether the moment restriction holds on the support of the conditioning variable. We extend the work of Li and Liao (2019) by allowing for non-smooth transformation of the conditioning variable. In addition, we implement an i.i.d. bootstrap method to calculate the critical values of the test which has better size properties in finite

sample than their method. In an empirical exercise, we found that the CoVaR model in Tobias and Brunnermeier (2016) is correctly specified.

3.7 Appendix

3.7.1 Proofs

Assumptions 5(ii, iv) impose Lipschitz-type conditions on $X_t(\theta)$ and $f_t(\theta)$ uniformly over $\theta \in B_n(\theta^\star)$. By Assumptions 5(ii, v), we get

$$\mathbb{E}\left[\max_{t\leq n}|L_{1,t}|\right] \leq n^{1/p}\max_{t\leq n}\|L_{1,t}\|_p \leq Kn^{1/p} \tag{3.4}$$

which together with the Markov inequality and Assumptions 5(i, ii) implies that

$$\max_{t\leq n}\left\|\widehat{X}_t - X_t^\star\right\| = o_p(\varepsilon_n) \tag{3.5}$$

where $\varepsilon_n = c_n n^{1/p-1/2} = o(1)$. Let $\mathcal{X}_{\varepsilon_n} = \left\{x \in \mathbb{R}^{d_\theta} : \inf_{x_1\in\mathcal{X}}\|x - x_1\| \leq \varepsilon_n\right\}$.

Lemma 3. *Under Assumptions 5 and 6, we have*

$$n^{-1}\sum_{l=1}^{m_n}\sum_{t=1}^{n}(p_l(\widehat{X}_t) - p_l(X_t^\star))^2 = O_p(\zeta_{1,n}^2 m_n n^{-1}).$$

Proof of Lemma 3 Consider any $l = 1, \ldots, m_n$. By Assumptions 5(i, ii, vi) and 6(v), (3.5) and the Markov inequality

$$\begin{aligned}
\sum_{l=1}^{m_n} n^{-1}\sum_{t=1}^{n}(p_l(\widehat{X}_t) - p_l(X_t^\star))^2 &= \sum_{l=1}^{m_n} n^{-1}\sum_{t=1}^{n}\left|\partial p_l(\tilde{X}_t)(\widehat{X}_t - X_t^\star)\right|^2 \\
&\leq \sum_{l=1}^{m_n} n^{-1}\sum_{t=1}^{n}||\partial p_l(\tilde{X}_t)||^2\left\|\widehat{X}_t - X_t^\star\right\|^2 \\
&\leq O_p(m_n\zeta_{1,n}^2)\left\|\hat{\theta}_n - \theta^\star\right\|^2 n^{-1}\sum_{t=1}^{n}L_{1,t}^2 = O_p(\zeta_{1,n}^2 m_n n^{-1}),
\end{aligned}$$

where $\tilde{X}_t$ is between $\widehat{X}_t$ and $X_t^\star$ for any $t = 1, \ldots, n$ and hence by (3.5) they lie in $\mathcal{X}_{\varepsilon_n}$ with probability approaching 1. □

Lemma 4. *Under Assumptions 5 and 6, we have*

$$\left\|\hat{Q}_n - Q_n\right\| = O_p(\zeta_{0,n}^{2-2/\kappa} m_n n^{-1/2} + \zeta_{1,n} m_n^{1/2} n^{-1/2}).$$

Proof of Lemma 4 Under Assumptions 6(i, v), we can use Lemma B2 in the Supplemental Appendix of Li and Liao (2019) to get,

$$\|Q_n^\star - Q_n\| = O_p(\zeta_n^{2-2/\kappa} m_n n^{-1/2}) \tag{3.6}$$

where $Q_n^\star = n^{-1} \sum_{t=1}^n P(X_t^\star) P(X_t^\star)^\top$, which together with Assumptions 6(iv, vi) implies that

$$(2K)^{-1} \leq \lambda_{\min}(Q_n^\star) \leq \lambda_{\max}(Q_n^\star) \leq 2K \tag{3.7}$$

with probability approaching 1. By definition

$$\begin{aligned}
\hat{Q}_n - Q_n^\star &= n^{-1} \sum_{t=1}^n \left(P(\widehat{X}_t) P(\widehat{X}_t)^\top - P(X_t^\star) P(X_t^\star)^\top \right) \\
&= n^{-1} \sum_{t=1}^n (P(\widehat{X}_t) - P(X_t^\star))(P(\hat{X}_t) - P(X_t^\star))^\top \\
&\quad + n^{-1} \sum_{t=1}^n P(X_t^\star) (P(\widehat{X}_t) - P(X_t^\star))^\top \\
&\quad + n^{-1} \sum_{t=1}^n (P(\widehat{X}_t) - P(X_t^\star)) P(X_t^\star)^\top. \qquad (3.8)
\end{aligned}$$

We next study the three terms after the second equality of (3.8). By the Cauchy-

Schwarz inequality and Lemma 3,

$$\left\| n^{-1} \sum_{t=1}^{n} (P(\widehat{X}_t) - P\left(X_t^\star\right))(P(\widehat{X}_t) - P\left(X_t^\star\right))^\top \right\|^2$$

$$\leq \left(n^{-1} \sum_{l=1}^{m_n} \sum_{t=1}^{n} (p_l(\hat{X}_t) - p_l(X_t^\star))^2 \right)^2 = O_p(\zeta_{1,n}^4 m_n^2 n^{-2}). \tag{3.9}$$

By (3.7), we deduce that

$$\left\| n^{-1} \sum_{t=1}^{n} P\left(X_t^\star\right) \left(P(\widehat{X}_t) - P\left(X_t^\star\right)\right) \right\|^2$$

$$= \sum_{l=1}^{m_n} \left\| n^{-1} \sum_{t=1}^{n} P\left(X_t^\star\right) \left(p_l(\widehat{X}_t) - p_l\left(X_t^\star\right)\right) \right\|^2$$

$$\leq \lambda_{\max}(Q_n^\star) \sum_{l=1}^{m_n} n^{-1} \sum_{t=1}^{n} (p_l(\hat{X}_t) - p_l\left(X_t^\star\right))^2 = O_p(\zeta_{1,n}^2 m_n n^{-1}) \tag{3.10}$$

where the second equality is by Lemma 3 and (3.7). Collecting the results in (3.8), (3.9) and (3.10), we obtain

$$\begin{aligned} \left\| \hat{Q}_n - Q_n^\star \right\| &\leq \left\| n^{-1} \sum_{t=1}^{n} (P(\widehat{X}_t) - P\left(X_t^\star\right))(P(\widehat{X}_t) - P\left(X_t^\star\right))^\top \right\| \\ &\quad + 2 \left\| n^{-1} \sum_{t=1}^{n} P\left(X_t^\star\right) \left(P(\widehat{X}_t) - P\left(X_t^\star\right)\right)^\top \right\| \\ &= O_p(\zeta_{1,n}^2 m_n n^{-1} + \zeta_{1,n} m_n^{1/2} n^{-1/2}) = O_p(\zeta_{1,n} m_n^{1/2} n^{-1/2}) \end{aligned} \tag{3.11}$$

where the second inequality is by Assumption 6(vi). The claim of the lemma follows by (3.6), (3.11) and the triangle inequality. □

Lemma 5. *Under Assumptions 5 and 6, we have*

$$\sup_{\theta\in B_n(\theta^\star)} n^{-1}\sum_{t=1}^{n}(Z_{t+1}(\theta)-Z_{t+1}^\star)^2 = O_p(\delta_{c,n}).$$

Proof. Let $U_{1,t} = \mathbb{1}(Y_{t+1} \le f_t(\theta^\star) + L_{2,t}\delta_{c,n})$ and $U_{2,t} = \mathbb{1}(Y_{t+1} \le f_t(\theta^\star) - L_{2,t}\delta_{c,n})$. For any $\theta \in B_n(\theta^\star)$, by Assumption 5(iv)

$$\left|Z_{t+1}(\theta) - Z_{t+1}^\star\right| = \left|\mathbb{1}(Y_{t+1} \le f_t(\theta)) - \mathbb{1}(Y_{t+1} \le f_t(\theta^\star))\right| \le U_{1,t} - U_{2,t} \tag{3.12}$$

which implies that

$$\sup_{\theta\in B_n(\theta^\star)} n^{-1}\sum_{t=1}^{n}(Z_{t+1}(\theta)-Z_{t+1}(\theta^\star))^2 \le n^{-1}\sum_{t=1}^{n}(U_{1,t}-U_{2,t})^2 = n^{-1}\sum_{t=1}^{n}(U_{1,t}-U_{2,t}). \tag{3.13}$$

By Assumption 5(iii), $|\mathbb{E}[U_{1,t} - U_{2,t}|\mathcal{F}_t]| \le KL_{2,t}\delta_{c,n}$ which together with Assumption 6(v) implies that

$$n^{-1}\sum_{t=1}^{n}\mathbb{E}[U_{1,t}-U_{2,t}] \le K\delta_{c,n}n^{-1}\sum_{t=1}^{n}\mathbb{E}[L_{2,t}] \le K\delta_{c,n}. \tag{3.14}$$

The claim of the lemma follows by (3.13), (3.14) and the Markov inequality. □

Lemma 6. *Under Assumptions 5 and 6, we have*

$$\sup_{\theta\in B_n(\theta^\star)} n^{-1}\sum_{t\le n} P(X_t(\theta))(Z_{t+1}(\theta)-Z_{t+1}^\star) = O_p(\delta_{c,n} + (\zeta_{1,n}\delta_{c,n}^{3/2} + \zeta_{0,n}\delta_{c,n}^{1/2})m_n^{1/2}n^{-1/2}).$$

Proof of Lemma 6 For any $\theta \in B_n(\theta^\star)$ and any t, let $\tilde{Z}_{t+1}(\theta) = \mathbb{1}(Y_{t+1} \le f_t(\theta)) - F_{t+1|t}(f_t(\theta))$. Then $\tilde{Z}_{t+1}(\theta)$ is a martingale difference array for any $\theta \in B_n(\theta^\star)$. By Assumptions 5(ii, vi) and (3.4)

$$\max_{t\le n}\sup_{\theta\in B_n(\theta^\star)}\|X_t(\theta) - X_t^\star\| = o_p(\varepsilon_n) \tag{3.15}$$

which implies that uniformly over $\theta \in B_n(\theta^\star)$ and for any $t \leq n$, $X_t(\theta) \in \mathcal{X}_{\varepsilon_n}$ with probability approaching 1. By (3.15) and the same arguments in the proof of Lemma 3, we have

$$\sup_{\theta \in B_n(\theta^\star)} n^{-1} \sum_{l=1}^{m_n} \sum_{t=1}^{n} (p_l(X_t(\theta)) - p_l(X_t^\star))^2 = O_p(\zeta_{1,n}^2 m_n \delta_{c,n}^2) \tag{3.16}$$

which together with the same arguments in showing (3.11) implies that

$$\sup_{\theta \in B_n(\theta^\star)} \|Q_n(\theta) - Q_n^\star\| = O_p(\zeta_{1,n} m_n^{1/2} \delta_{c,n}) \tag{3.17}$$

where $Q_n(\theta) = n^{-1} \sum_{i=1}^{n} P(X_t(\theta)) P(X_t(\theta))^\top$. By Assumption 6(vi), (3.7) and (3.17),

$$(2K)^{-1} \leq \lambda_{\min}(Q_n(\theta)) \leq \lambda_{\max}(Q_n(\theta)) \leq 2K \tag{3.18}$$

uniformly over $\theta \in B_n(\theta^\star)$ with probability approaching 1. By Assumptions 5(iii, iv, vi), (3.18) and the Markov inequality

$$\begin{aligned}
&\sup_{\theta \in B_n(\theta^\star)} \left\| n^{-1} \sum_{t \leq n} P(X_t(\theta)) (F_{t+1|t}\left(f_t(\theta)\right) - F_{t+1|t}\left(f_t(\theta^\star)\right)) \right\|^2 \\
\leq \quad & \sup_{\theta \in B_n(\theta^\star)} \lambda_{\max}(Q_n(\theta)) \sup_{\theta \in B_n(\theta^\star)} n^{-1} \sum_{t \leq n} \left(F_{t+1|t}\left(f_t(\theta)\right) - F_{t+1|t}\left(f_t(\theta^\star)\right)\right)^2 \\
\leq \quad & K \sup_{\theta \in B_n(\theta^\star)} \lambda_{\max}(Q_n(\theta)) \sup_{\theta \in B_n(\theta^\star)} \|\theta - \theta^\star\|^2 n^{-1} \sum_{t \leq n} L_{2,t}^2 = O_p(\delta_{c,n}^2)
\end{aligned} \tag{3.19}$$

which implies that

$$\sup_{\theta \in B_n(\theta^\star)} n^{-1} \sum_{t \leq n} P(X_t(\theta)) (F_{t+1|t}\left(f_t(\theta)\right) - F_{t+1|t}\left(f_t(\theta^\star)\right)) = O_p(\delta_{c,n}). \tag{3.20}$$

It remains to show that

$$\sup_{\theta \in B_n(\theta^\star)} n^{-1} \sum_{t \leq n} P(X_t(\theta)) (\tilde{Z}_{t+1}(\theta) - \tilde{Z}_{t+1}(\theta^\star)) = O_p((\zeta_{1,n} \delta_{c,n}^{3/2} + \zeta_{0,n} \delta_{c,n}^{1/2}) m_n^{1/2} n^{-1/2}), \tag{3.21}$$

since the claim of the lemma follows by Assumption 6(vi), (3.20) and (3.21).

By Hölder's inequality,

$$\begin{aligned}
&\mathbb{E}\left[\left\|\sup_{\theta\in B_n(\theta^\star)} n^{-1/2}\sum_{t\le n} P(X_t(\theta))(\tilde{Z}_{t+1}(\theta)-\tilde{Z}_{t+1}(\theta^\star))\right\|^2\right] \\
\le\ & \sum_{l\le m_n}\mathbb{E}\left[\left|\sup_{\theta\in B_n(\theta^\star)} n^{-1/2}\sum_{t\le n} p_l\left(X_t(\theta)\right)\left(\tilde{Z}_{t+1}(\theta)-\tilde{Z}_{t+1}(\theta^\star)\right)\right|^2\right] \\
\le\ & m_n\max_{l\le m_n}\mathbb{E}\left[\left|\sup_{\theta\in B_n(\theta^\star)} n^{-1/2}\sum_{t\le n} p_l\left(X_t(\theta)\right)\left(\tilde{Z}_{t+1}(\theta)-\tilde{Z}_{t+1}(\theta^\star)\right)\right|^2\right] \\
\le\ & m_n\max_{l\le m_n}\left(\mathbb{E}\left[\left|\sup_{\theta\in B_n(\theta^\star)} n^{-1/2}\sum_{t\le n} p_l\left(X_t(\theta)\right)\left(\tilde{Z}_{t+1}(\theta)-\tilde{Z}_{t+1}(\theta^\star)\right)\right|^p\right]\right)^{2/p} \quad (3.22)
\end{aligned}$$

We next show that

$$\left\|\sup_{\theta\in B_n(\theta^\star)} n^{-1/2}\sum_{t\le n} p_l\left(X_t(\theta)\right)\left(\tilde{Z}_{t+1}(\theta)-\tilde{Z}_{t+1}(\theta^\star)\right)\right\|_p \le K(\zeta_{1,n}\delta_{c,n}^{3/2}+\zeta_{0,n}\delta_{c,n}^{1/2}) \quad (3.23)$$

which together with (3.22) and the Markov inequality proves (3.21) and hence the claim of the lemma.

For ease of notations, let $\pi_{l,n}(\theta) = n^{-1/2}\sum_{t\le n} p_l\left(X_t(\theta)\right)\left(\tilde{Z}_{t+1}(\theta)-\tilde{Z}_{t+1}(\theta^\star)\right)$ for any $\theta\in\Theta$. It is clear that $\pi_{l,n}(\theta)$ is a martingale for any $\theta\in\Theta$. For any $\theta\in B_n(\theta^\star)$, by Assumption 5(iii) and (3.12)

$$\mathbb{E}\left[(\tilde{Z}_{t+1}(\theta)-\tilde{Z}_{t+1}(\theta^\star))^2\middle|\mathcal{F}_t\right] \le KL_{2,t}\delta_{c,n}. \quad (3.24)$$

For any θ_1, $\theta_2 \in B_n(\theta^\star)$, by (3.24), Burkholder's inequality and Hölder's inequality,

$$\begin{aligned}
&\mathbb{E}\left[\left|n^{-1/2}\sum_{t\leq n}(p_l\left(X_t(\theta_1)\right)-p_l\left(X_t(\theta_2)\right))(\tilde{Z}_{t+1}(\theta_1)-\tilde{Z}_{t+1}(\theta^\star))\right|^p\right] \\
\leq\ & K\delta_{c,n}^{p/2}\mathbb{E}\left[\left|n^{-1}\sum_{t\leq n}L_{2,t}^2(p_l\left(X_t(\theta_1)\right)-p_l\left(X_t(\theta_2)\right))^2\right|^{p/2}\right] \\
\leq\ & K\zeta_{1,n}^p\delta_{c,n}^{p/2}\left\|\theta_1-\theta_2\right\|^p\mathbb{E}\left[\left|n^{-1}\sum_{t\leq n}L_{2,t}^2\right|^{p/2}\right] \\
\leq\ & K\zeta_{1,n}^p\delta_{c,n}^{p/2}\left\|\theta_1-\theta_2\right\|^p n^{-1}\sum_{t\leq n}\mathbb{E}[L_{2,t}^p]\leq K\zeta_{1,n}^p\delta_{c,n}^p\left\|\theta_1-\theta_2\right\|^{p/2} \qquad (3.25)
\end{aligned}$$

where the second inequality is by Assumption 6(v) and (3.15), the third inequality is by Hölder's inequality and the last inequality is by Assumption 5(v) and $\|\theta_1-\theta_2\|^{1/2}\leq K\delta_{c,n}^{1/2}$ for any θ_1, $\theta_2\in B_n(\theta^\star)$. For any θ_1, $\theta_2\in B_n(\theta^\star)$,

$$\begin{aligned}
(Z_{t+1}(\theta_1)-Z_{t+1}(\theta_2))^2\leq\ & \mathbb{1}(Y_{t+1}\leq f_t\left(\theta_1\right)+\left|f_t\left(\theta_1\right)-f_t\left(\theta_2\right)\right|)-\mathbb{1}(Y_{t+1}\leq f_t\left(\theta_1\right) \\
& -\left|f_t\left(\theta_1\right)-f_t\left(\theta_2\right)\right|)
\end{aligned}$$

which together with Assumptions 5(iii, iv) implies that

$$\mathbb{E}\left[(\tilde{Z}_{t+1}(\theta_1)-\tilde{Z}_{t+1}(\theta_2))^2\middle|\mathcal{F}_t\right]\leq 2K\left|f_t\left(\theta_1\right)-f_t\left(\theta_2\right)\right|\leq 2KL_{2,t}\left\|\theta_1-\theta_2\right\|. \quad (3.26)$$

By (3.26), Burkholder's inequality and Hölder's inequality,

$$\begin{aligned}
&\mathbb{E}\left[\left|n^{-1/2}\sum_{t\leq n} p_l\left(X_t(\theta_1)\right)\left(\tilde{Z}_{t+1}(\theta_1)-\tilde{Z}_{t+1}(\theta_2)\right)\right|^p\right] \\
\leq\ & K\mathbb{E}\left[\left|n^{-1}\sum_{t\leq n} p_l\left(X_t(\theta_1)\right)^2\mathbb{E}\left[\left(\tilde{Z}_{t+1}(\theta_1)-\tilde{Z}_{t+1}(\theta_2)\right)^2\middle|\mathcal{F}_t\right]\right|^{p/2}\right] \\
\leq\ & K\left\|\theta_1-\theta_2\right\|^{p/2}\mathbb{E}\left[\left|n^{-1}\sum_{t\leq n} p_l\left(X_t(\theta_1)\right)^2 L_{2,t}^2\right|^{p/2}\right] \\
\leq\ & K\zeta_{0,n}^p\left\|\theta_1-\theta_2\right\|^{p/2} n^{-1}\sum_{t\leq n}\mathbb{E}[L_{2,t}^p]\leq K\zeta_{0,n}^p\left\|\theta_1-\theta_2\right\|^{p/2} \qquad (3.27)
\end{aligned}$$

where the third inequality is by Assumption 6(v), (3.15) and Hölder's inequality, and the last inequality is by Assumption 5(iv). Combining the results in (3.25) and (3.27), we get

$$\left\|\pi_{l,n}(\theta_1)-\pi_{l,n}(\theta_2)\right\|_p\leq K(\zeta_{1,n}\delta_{c,n}+\zeta_{0,n})\left\|\theta_1-\theta_2\right\|^{1/2} \qquad (3.28)$$

for any θ_1, $\theta_2\in B_n(\theta^\star)$.

We shall use the chaining argument to prove the asserted claim in (3.23). Construct nested sets $\Theta_{0,n}\subset\Theta_{1,n}\cdots\subset B_n(\theta^\star)$ such that $\Theta_{0,n}=\{\theta^\star\}$ and $\Theta_{j,n}$ (for $j>0$) is a maximal set of points in the sense that for every θ_1, $\theta_2\in\Theta_j$ there is $\|\theta_1-\theta_2\|>\delta_{c,n}2^{-j}$. The number of the points in Θ_j is less than $K(2^j)^{d_\theta}$. Link every point $\theta_{j+1}\in\Theta_{j+1}$ to a unique $\theta_j\in\Theta_j$ such that $\|\theta_{j+1}-\theta_j\|\leq\delta_{c,n}2^{-j}$. Consider any positive integer J. Obtain for every θ_{J+1} a chain $\theta_{J+1},\ldots,\theta_0$ that connects it to θ_0. For arbitrary points θ_{J+1} in Θ_{J+1}, by the triangle inequality

$$\left|\pi_{l,n}(\theta_{J+1})\right|=\left|\sum_{j=0}^{J}\left[\pi_{l,n}(\theta_{j+1})-\pi_{l,n}(\theta_j)\right]\right|\leq\sum_{j=0}^{J}\max\left|\pi_{l,n}(\theta_{j+1})-\pi_{l,n}(\theta_j)\right| \qquad (3.29)$$

where $\theta_0=\theta^\star$ and for fixed j the maximum is taken over all links (θ_{j+1},θ_j) from Θ_{j+1} to Θ_j. Thus the jth maximum is taken over at most $K(2^{j+1})^{d_\theta}$ many links. By

(3.28), (3.29), the triangle inequality and Hölder's inequality,

$$\left\|\max|\pi_{l,n}(\theta_{J+1})|\right\|_p \leq \sum_{j=0}^{J}\left\|\max|\pi_{l,n}(\theta_{j+1})-\pi_{l,n}(\theta_j)|\right\|_p$$

$$\leq K\sum_{j=0}^{J}(2^j)^{d_\theta/p}\max\left\|\pi_{l,n}(\theta_{j+1})-\pi_{l,n}(\theta_j)\right\|_p$$

$$\leq K(\zeta_{1,n}\delta_{c,n}^{3/2}+\zeta_{0,n}\delta_{c,n}^{1/2})\sum_{j=0}^{J}(2^{-j})^{1/2-d_\theta/p} \leq K(\zeta_{1,n}\delta_{c,n}^{3/2}+\zeta_{0,n}\delta_{c,n}^{1/2}) \tag{3.30}$$

where the last inequality is by $p > 2d_\theta$. Since the stochastic process $\pi_{l,n}(\theta)$ indexed by $\theta \in B_n(\theta^\star)$ is separable for any $l = 1,\ldots,m_n$ (which can be verified since the sample path of $\pi_{l,n}(\theta)$ is continuous almost surely), letting J go to infinity, we obtain from (3.30) and the triangle inequality, that

$$\left\|\sup_{\theta\in B_n(\theta^\star)}|\pi_{l,n}(\theta)|\right\|_p \leq K(\zeta_{1,n}\delta_{c,n}^{3/2}+\zeta_{0,n}\delta_{c,n}^{1/2}) \tag{3.31}$$

which proves (3.23). □

Lemma 7. *Suppose that Assumptions 5 and 6 hold. Then under the null hypothesis*

$$\sup_{\theta\in B_n(\theta^\star)} n^{-1}\sum_{t\leq n}(P(X_t(\theta))-P(X_t(\theta^\star)))u_t^\star = O_p(\zeta_{1,n}\delta_{c,n}m_n^{1/2}n^{-1/2}).$$

Proof of Lemma 7 Under the null hypothesis, $(u_t^\star)_t$ is a martingale difference sequence. By Hölder's inequality and the same arguments in showing (3.22),

$$\mathbb{E}\left[\left\|\sup_{\theta\in B_n(\theta^\star)} n^{-1/2}\sum_{t\leq n}(P(X_t(\theta))-P(X_t(\theta^\star)))u_t^\star\right\|^2\right]$$

$$\leq \quad m_n\max_{l\leq m_n}\left(\mathbb{E}\left[\left|\sup_{\theta\in B_n(\theta^\star)} n^{-1/2}\sum_{t\leq n}(p_l(X_t(\theta))-p_l(X_t(\theta^\star)))u_t^\star\right|^p\right]\right)^{2/p}. \tag{3.32}$$

We next show that

$$\left\| \sup_{\theta \in B_n(\theta^\star)} n^{-1/2} \sum_{t \le n} (p_l(X_t(\theta)) - p_l(X_t(\theta^\star)))u_t^\star \right\|_p \le K\zeta_{1,n}\delta_{c,n} \tag{3.33}$$

which together with (3.32) and the Markov inequality proves the claim of the lemma.

For ease of notations, let $\pi_{l,n}(\theta) = n^{-1/2}\sum_{t\le n}(p_l(X_t(\theta)) - p_l(X_t(\theta^\star)))u_t^\star$ for any $\theta \in \Theta$. It is clear that $\pi_{l,n}(\theta)$ is a martingale for any $\theta \in \Theta$. For any θ_1, $\theta_2 \in \Theta$, by (3.15), Burkholder's inequality and Hölder's inequality, and similar arguments in deriving (3.25),

$$\begin{aligned}
\mathbb{E}\left[|\pi_{l,n}(\theta_1) - \pi_{l,n}(\theta_2)|^p\right] &= \mathbb{E}\left[\left| n^{-1/2}\sum_{t\le n}(p_l(X_t(\theta_1)) - p_l(X_t(\theta_2)))u_t^\star \right|^p\right] \\
&\le K\mathbb{E}\left[\left| n^{-1}\sum_{t\le n}(p_l(X_t(\theta_1)) - p_l(X_t(\theta_2)))^2 \right|^{p/2}\right] \\
&\le K\zeta_{1,n}^p \|\theta_1 - \theta_2\|^p \,\mathbb{E}\left[\left| n^{-1}\sum_{t\le n} L_{1,t} \right|^{p/2}\right] \\
&\le K\zeta_{1,n}^p \|\theta_1 - \theta_2\|^p \, n^{-1}\sum_{t\le n}\mathbb{E}\left[L_{1,t}^p\right] \le K\zeta_{1,n}^p \|\theta_1 - \theta_2\|^p
\end{aligned}$$

which implies that

$$\|\pi_{l,n}(\theta_1) - \pi_{l,n}(\theta_2)\|_p \le K\zeta_{1,n}\|\theta_1 - \theta_2\| \tag{3.34}$$

for any θ_1, $\theta_2 \in \Theta$. Using the same chaining arguments in the proof of Lemma 6, we

deduce that

$$\|\max|\pi_{l,n}(\theta_{J+1})|\|_p \leq \sum_{j=0}^{J} \|\max|\pi_{l,n}(\theta_{j+1}) - \pi_{l,n}(\theta_j)|\|_p$$

$$\leq K\sum_{j=0}^{J}(2^j)^{d_\theta/p} \max\|\pi_{l,n}(\theta_{j+1}) - \pi_{l,n}(\theta_j)\|_p$$

$$\leq K\zeta_{1,n}\delta_{c,n}\sum_{j=0}^{J}(2^{-j})^{1-d_\theta/p} \leq K\zeta_{1,n}\delta_{c,n} \tag{3.35}$$

where the last inequality is by $p > 2d_\theta$. Letting J go to infinity, we obtain from (3.35) that

$$\left\| \sup_{\theta\in B_n(\theta^\star)} |\pi_{l,n}(\theta)| \right\|_p \leq K\zeta_{1,n}\delta_{c,n} \tag{3.36}$$

which proves (3.33). □

Lemma 8. *Under Assumptions 5 and 6, we have* $\widehat{b}_n - b_n^\star = O_p(\delta_{b,n})$ *where*

$$\delta_{b,n} = \begin{cases} O_p(m_n^{1/2}n^{-1/2}) & \textit{under the null} \\ O_p(\zeta_{1,n}m_n^{1/2}n^{-1/2} + m_n^{-\rho_h}) & \textit{in general} \end{cases}.$$

Proof of Lemma 8 By Assumptions 6(iv, vi) and Lemma 4,

$$(2K)^{-1} \leq \lambda_{\min}(\hat{Q}_n) \leq \lambda_{\max}(\hat{Q}_n) \leq 2K \tag{3.37}$$

with probability approaching 1. By the definition of $\widehat{b}_n$, we can write

$$\begin{aligned} \widehat{b}_n - b_n^\star &= \hat{Q}_n^{-1}\left(n^{-1}\sum_{t=1}^{n} P(X_t^\star)\,u_t^\star\right) + \hat{Q}_n^{-1}\left(n^{-1}\sum_{t=1}^{n} u_t^\star(P(\hat{X}_t) - P(X_t^\star))\right) \\ &\quad + \hat{Q}_n^{-1}n^{-1}\sum_{t=1}^{n} P(\hat{X}_t)(h(X_t^\star) - P(\hat{X}_t)^\top b_n^\star) \\ &\quad + \hat{Q}_n^{-1}\left(n^{-1}\sum_{t=1}^{n} P(\widehat{X}_t)(\widehat{Z}_{t+1} - Z_{t+1}^\star)\right). \end{aligned} \tag{3.38}$$

By Assumption 6(iv) and the Markov inequality,

$$n^{-1}\sum_{t=1}^{n} P\left(X_t^\star\right) u_t^\star = O_p(m_n^{1/2} n^{-1/2}) \tag{3.39}$$

which together with (3.37) implies that

$$\hat{Q}_n^{-1}\left(n^{-1}\sum_{t=1}^{n} P\left(X_t^\star\right) u_t^\star\right) = O_p(m_n^{1/2} n^{-1/2}). \tag{3.40}$$

Since $\left|Z_{t+1}^\star\right| \leq K$ for any t, we have

$$|u_t^\star| + |h(X_t^\star)| \leq K \text{ for any } t. \tag{3.41}$$

By the Cauchy-Schwarz inequality, (3.41) and Lemma 3

$$\left\| n^{-1}\sum_{t=1}^{n} u_t^\star (P(\hat{X}_t) - P\left(X_t^\star\right)) \right\|^2 \leq K n^{-1}\sum_{t=1}^{n} ||P(\hat{X}_t) - P\left(X_t^\star\right)||^2 = O_p(\zeta_{1,n}^2 m_n n^{-1}). \tag{3.42}$$

On the other hand, under the null hypothesis we can use Assumption 6(vi) and Lemma 7 to get

$$n^{-1}\sum_{t=1}^{n} u_t^\star (P(\widehat{X}_t) - P\left(X_t^\star\right)) = O_p(\zeta_{1,n}\delta_{c,n} m_n^{1/2} n^{-1/2}) = o_p(m_n^{1/2} n^{-1/2}). \tag{3.43}$$

By Assumptions 5(i, ii, vi), and 6(ii, iii), (3.5), (3.37), the Cauchy-Schwarz inequality

and the Markov inequality, under the alternative hypothesis we have

$$\begin{aligned}
&\left\|\hat{Q}_n^{-1}n^{-1}\sum_{t=1}^{n}P(\widehat{X}_t)(h(X_t^\star)-P(\widehat{X}_t)^\top b_n^\star)\right\|^2 \\
\leq\ & (\lambda_{\min}(\hat{Q}_n))^{-1}n^{-1}\sum_{t=1}^{n}(h(X_t^\star)-P(\hat{X}_t)^\top b_n^\star)^2 \\
\leq\ & 2(\lambda_{\min}(\hat{Q}_n))^{-1}n^{-1}\sum_{t=1}^{n}(h(\widehat{X}_t)-P(\hat{X}_t)^\top b_n^\star)^2 \\
&+2(\lambda_{\min}(\hat{Q}_n))^{-1}n^{-1}\sum_{t=1}^{n}(h(\widehat{X}_t)-h(X_t^\star))^2 = O_p(m_n^{-2\rho_h}+n^{-1}). \quad (3.44)
\end{aligned}$$

On the other hand, under the null hypothesis, we have $h(\cdot)=0$ and $b_n^\star=0$ which implies that

$$\hat{Q}_n^{-1}n^{-1}\sum_{t=1}^{n}P(\hat{X}_t)(h(X_t^\star)-P(\hat{X}_t)^\top b_n^\star)=0. \quad (3.45)$$

By Assumption 6(vi), (3.37) and Lemma 6

$$\begin{aligned}
\hat{Q}_n^{-1}\left(n^{-1}\sum_{t=1}^{n}P(\hat{X}_t)(\widehat{Z}_{t+1}-Z_{t+1}^\star)\right) &= O_p(\delta_{c,n}+(\zeta_{1,n}\delta_{c,n}^{3/2}+\zeta_{0,n}\delta_{c,n}^{1/2})m_n^{1/2}n^{-1/2}) \\
&= o_p(m_n^{1/2}n^{-1/2}). \quad (3.46)
\end{aligned}$$

Collecting the results in (3.38), (3.40), (3.42), (3.43), (3.44), (3.45) and (3.46), we get

$$\left\|\widehat{b}_n-b_n^\star\right\| = \begin{cases} O_p(m_n^{1/2}n^{-1/2}) & \text{under the null} \\ O_p(\zeta_{1,n}m_n^{1/2}n^{-1/2}+m_n^{-\rho_h}) & \text{in general} \end{cases} \quad (3.47)$$

which finishes the proof. □

Lemma 9. *Under Assumptions 5 and 6,* $\left\|\hat{A}_n-A_n\right\|=\delta_{A,n}$ *where*

$$\delta_{A,n} = \begin{cases} O_p(m_n^{1/2}(\zeta_{1,n}n^{-1/2}+\zeta_{0,n}\delta_{c,n}^{1/2})) & \textit{under the null} \\ O_p(\zeta_{0,n}m_n^{1/2}(\delta_{c,n}^{1/2}+\zeta_{1,n}m_n^{1/2}n^{-1/2}+\zeta_{1,n}^2\zeta_{0,n}m^{3/2}n^{-1})) & \textit{in general} \end{cases}.$$

PROOF OF LEMMA 9 By definition

$$
\begin{aligned}
& n^{-1}\sum_{t=1}^{n}\widehat{u}_t^2 P(\widehat{X}_t)P(\widehat{X}_t)^\top - n^{-1}\sum_{t=1}^{n}u_t^{\star 2}P\left(X_t^\star\right)P\left(X_t^\star\right)^\top \\
= \; & n^{-1}\sum_{t=1}^{n}(\widehat{u}_t^2 - u_t^{\star 2})P(\widehat{X}_t)P(\widehat{X}_t)^\top + n^{-1}\sum_{t=1}^{n}u_t^{\star 2}(P(\widehat{X}_t) - P\left(X_t^\star\right))P\left(X_t^\star\right)^\top \\
& + n^{-1}\sum_{t=1}^{n}u_t^{\star 2}P\left(X_t^\star\right)\left(P(\widehat{X}_t) - P\left(X_t^\star\right)\right)^\top \\
& + n^{-1}\sum_{t=1}^{n}u_t^{\star 2}(P(\widehat{X}_t) - P\left(X_t^\star\right))(P(\widehat{X}_t) - P\left(X_t^\star\right))^\top. \qquad (3.48)
\end{aligned}
$$

Let $A_n^\star = n^{-1}\sum_{t=1}^{n}u_t^{\star 2}P\left(X_t^\star\right)P\left(X_t^\star\right)^\top$. We write

$$
\begin{aligned}
A_n^\star - A_n \; = \; & n^{-1}\sum_{t=1}^{n}(u_t^{\star 2} - \sigma_{u,t}^2)P\left(X_t^\star\right)P\left(X_t^\star\right)^\top \\
& + n^{-1}\sum_{t=1}^{n}\left[\sigma_{u,t}^2 P\left(X_t^\star\right)P\left(X_t^\star\right)^\top - \mathbb{E}\left[\sigma_{u,t}^2 P\left(X_t^\star\right)P\left(X_t^\star\right)^\top\right]\right]
\end{aligned}
$$

where $\sigma_{u,t}^2 = F_{t+1|t}\left(f_t(\theta^\star)\right) - 2F_{t+1|t}\left(f_t(\theta^\star)\right)h(X_t^\star) + h(X_t^\star)^2$ which denotes the conditional variance of $u_t^{\star 2}$ given $\mathcal{F}_t$. Using the same arguments in step 3 and step 4 of the proof of Lemma B3 in the Supplemental Appendix of Li and Liao (2019), we deduce that

$$
\|A_n^\star - A_n\| = O_p(\zeta_{0,n}^{2-2/\kappa}m_n n^{-1/2}). \qquad (3.49)
$$

By Lemma 3, (3.41) and the Cauchy-Schwarz inequality

$$
\begin{aligned}
& \left\| n^{-1}\sum_{t=1}^{n}u_t^{\star 2}(P(\hat{X}_t) - P\left(X_t^\star\right))(P(\widehat{X}_t) - P\left(X_t^\star\right))^\top \right\| \\
\leq \; & Kn^{-1}\sum_{l=1}^{m_n}\sum_{t=1}^{n}(p_l(\widehat{X}_t) - p_l\left(X_t^\star\right))^2 = O_p(\zeta_{1,n}^2 m_n n^{-1}). \qquad (3.50)
\end{aligned}
$$

By Lemma 3, (3.7) and (3.41)

$$\begin{aligned}
&\left\| n^{-1}\sum_{t=1}^{n} u_t^{\star 2} P\left(X_t^\star\right)\left(P(\widehat{X}_t) - P\left(X_t^\star\right)\right)^\top \right\|^2 \\
\leq\ & K\lambda_{\max}(Q_n^\star) n^{-1}\sum_{l=1}^{m_n}\sum_{t=1}^{n}(p_l(\widehat{X}_t) - p_l\left(X_t^\star\right))^2 = O_p(\zeta_{1,n}^2 m_n n^{-1}). \qquad (3.51)
\end{aligned}$$

By (3.5), (3.37) and Assumption 6(v) we have

$$\begin{aligned}
&\left\| n^{-1}\sum_{t=1}^{n}(\widehat{u}_t^2 - u_t^{\star 2}) P(\widehat{X}_t) P(\widehat{X}_t)^\top \right\|^2 \\
=\ & \sum_{l=1}^{m_n}\left\| n^{-1}\sum_{t=1}^{n} P(\hat{X}_t) p_l(\widehat{X}_t)(\widehat{u}_t^2 - u_t^{\star 2})\right\|^2 \\
\leq\ & \lambda_{\max}(\hat{Q}_n)\sum_{l=1}^{m_n} n^{-1}\sum_{t=1}^{n} p_l(\widehat{X}_t)^2(\widehat{u}_t^2 - u_t^{\star 2})^2 \\
\leq\ & m_n\zeta_{0,n}^2\lambda_{\max}(\hat{Q}_n) n^{-1}\sum_{t=1}^{n}(\widehat{u}_t^2 - u_t^{\star 2})^2. \qquad (3.52)
\end{aligned}$$

By definition $\widehat{u}_t = \widehat{Z}_{t+1} - \widehat{h}_n(\hat{X}_t)$, by the triangle inequality we have

$$\begin{aligned}
|\widehat{u}_t - u_t^\star| \leq\ & \left|\widehat{Z}_{t+1} - Z_{t+1}^\star\right| + \left|\widehat{h}_n(\hat{X}_t) - h_{m_n}(\widehat{X}_t)\right| \\
& + \left|h_{m_n}(\widehat{X}_t) - h(\widehat{X}_t)\right| + \left|h(\widehat{X}_t) - h(X_t^\star)\right|. \qquad (3.53)
\end{aligned}$$

Since $\hat{\theta}_n \in B_n(\theta^\star)$ with probability approaching 1 and $|\widehat{Z}_{t+1}| \leq 1$ for any t, by Lemma 4

$$n^{-1}\sum_{t=1}^{n}\left|\widehat{Z}_{t+1} - Z_{t+1}^\star\right|^4 \leq n^{-1}\sum_{t=1}^{n}\left|\widehat{Z}_{t+1} - Z_{t+1}^\star\right|^2 = O_p(\delta_{c,n}). \qquad (3.54)$$

By Lemma 8 and (3.37),

$$\begin{aligned} n^{-1}\sum_{t=1}^{n}\left|\widehat{h}_n(\widehat{X}_t)-h_{m_n}(\widehat{X}_t)\right|^2 &\leq \lambda_{\max}(\hat{Q}_n)\left\|\widehat{b}_n-b_n^\star\right\|^2 \\ &= \begin{cases} O_p(m_n n^{-1}) & \text{under the null} \\ O_p(\zeta_{1,n}^2 m_n n^{-1}+m_n^{-2\rho_h}) & \text{in general} \end{cases} \end{aligned} \tag{3.55}$$

By Assumption 6(v), (3.5), Lemma 8 and (3.55)

$$\begin{aligned} & n^{-1}\sum_{t=1}^{n}\left|\widehat{h}_n(\widehat{X}_t)-h_{m_n}(\widehat{X}_t)\right|^4 \\ \leq\ & \zeta_{0,n}^2 m_n\left\|\widehat{b}_n-b_n^\star\right\|^2 n^{-1}\sum_{t=1}^{n}\left|\widehat{h}_n(\widehat{X}_t)-h_{m_n}(\widehat{X}_t)\right|^2 \\ =\ & \begin{cases} O_p(\zeta_{0,n}^2 m^3 n^{-2}) & \text{under the null} \\ O_p(\zeta_{0,n}^2 m(\zeta_{1,n}^4 m_n^2 n^{-2}+m_n^{-4\rho_h})) & \text{in general} \end{cases}. \end{aligned} \tag{3.56}$$

Since $h(\cdot)=0$ and $h_{m_n}(\cdot)=0$ under the null, by (3.53), (3.54) and (3.55) we have

$$n^{-1}\sum_{t=1}^{n}|\widehat{u}_t-u_t^\star|^2=O_p(\delta_{c,n}+m_n n^{-1}), \tag{3.57}$$

and by (3.53), (3.54) and (3.56) we have

$$n^{-1}\sum_{t=1}^{n}|\widehat{u}_t-u_t^\star|^4=O_p(\delta_{c,n}+\zeta_{0,n}^2 m^3 n^{-2}) \tag{3.58}$$

under the null hypothesis. Combining the results in (3.41), (3.57) and (3.58) and applying Assumption 6(vi), we get

$$\begin{aligned} n^{-1}\sum_{t=1}^{n}(\widehat{u}_t^2-u_t^{\star 2})^2 &\leq Kn^{-1}\sum_{t=1}^{n}|\widehat{u}_t-u_t^\star|^2+Kn^{-1}\sum_{t=1}^{n}|\widehat{u}_t-u_t^\star|^4 \\ &= O_p(\delta_{c,n}+m_n n^{-1}+\zeta_{0,n}^2 m^3 n^{-2})=O_p(\delta_{c,n}) \end{aligned} \tag{3.59}$$

which together with (3.37) and (3.52) implies that

$$\left\| n^{-1} \sum_{t=1}^{n} (\widehat{u}_t^2 - u_t^{\star 2}) P(\widehat{X}_t) P(\widehat{X}_t)^\top \right\| = O_p(\zeta_{0,n} m_n^{1/2} \delta_{c,n}^{1/2}) \tag{3.60}$$

under the null hypothesis. By the triangle inequality, Assumption 6(vi), (3.48), (3.49), (3.50), (3.51) and (3.60), we have

$$\left\| \hat{A}_n - A_n \right\| = O_p(m_n^{1/2}(\zeta_{1,n} n^{-1/2} + \zeta_{0,n} \delta_{c,n}^{1/2}))$$

which proves the lemma under the null.

By the definition of $h(\cdot)$, we have $\left| h(\hat{X}_t) - h(X_t^\star) \right| \leq 2$ for any t which together with Assumptions 5(i, ii) and 6(iii) implies that

$$\begin{aligned} n^{-1} \sum_{t=1}^{n} \left| h(\widehat{X}_t) - h(X_t^\star) \right|^4 &\leq n^{-1} \sum_{t=1}^{n} \left| h(\widehat{X}_t) - h(X_t^\star) \right|^2 \\ &\leq K \left\| \hat{\theta}_n - \theta^\star \right\|^2 n^{-1} \sum_{t=1}^{n} L_{1,t}^2 = O_p(n^{-1}). \end{aligned} \tag{3.61}$$

By Assumption 6(ii) and (3.5),

$$n^{-1} \sum_{t=1}^{n} \left| h_{m_n}(\widehat{X}_t) - h(\hat{X}_t) \right|^2 = O_p(m_n^{-2\rho_h}) \text{ and } n^{-1} \sum_{t=1}^{n} \left| h_{m_n}(\widehat{X}_t) - h(\hat{X}_t) \right|^4 = O_p(m_n^{-4\rho_h}) \tag{3.62}$$

under the alternative hypothesis. By (3.53), (3.54), (3.55), (3.61) and (3.62) we have

$$n^{-1} \sum_{t=1}^{n} |\widehat{u}_t - u_t^\star|^2 = O_p(\delta_{c,n} + m_n^{-2\rho_h} + \zeta_{1,n}^2 m_n n^{-1}), \tag{3.63}$$

and

$$n^{-1} \sum_{t=1}^{n} |\widehat{u}_t - u_t^\star|^4 = O_p(\delta_{c,n} + \zeta_{0,n}^2 m_n (\zeta_{1,n}^4 m_n^2 n^{-2} + m_n^{-4\rho_h})) \tag{3.64}$$

which together with (3.41) and Assumption 6(vi) implies that

$$\begin{aligned} n^{-1}\sum_{t=1}^{n}(\widehat{u}_t^2 - u_t^{\star 2})^2 &\leq Kn^{-1}\sum_{t=1}^{n}|\widehat{u}_t - u_t^{\star}|^2 + Kn^{-1}\sum_{t=1}^{n}|\widehat{u}_t - u_t^{\star}|^4 \\ &= O_p(\delta_{c,n} + \zeta_{1,n}^2 m_n n^{-1} + \zeta_{1,n}^4\zeta_{0,n}^2 m^3 n^{-2}) \end{aligned} \tag{3.65}$$

under the alternative hypothesis. By (3.37), (3.52) and (3.65)

$$\left\| n^{-1}\sum_{t=1}^{n}(\widehat{u}_t^2 - u_t^{\star 2})P(\widehat{X}_t)P(\widehat{X}_t)^{\top} \right\| = O_p(\zeta_{0,n}m_n^{1/2}(\delta_{c,n}^{1/2} + \zeta_{1,n}m_n^{1/2}n^{-1/2} + \zeta_{1,n}^2\zeta_{0,n}m^{3/2}n^{-1})) \tag{3.66}$$

which together with Assumption 6(vi), (3.48), (3.49), (3.50), (3.51) shows that

$$\left\| \hat{A}_n - A_n \right\| = O_p(\zeta_{0,n}m_n^{1/2}(\delta_{c,n}^{1/2} + \zeta_{1,n}m_n^{1/2}n^{-1/2} + \zeta_{1,n}^2\zeta_{0,n}m^{3/2}n^{-1})).$$

This shows the lemma under the alternative. □

Lemma 10. *Suppose that Assumptions 5 and 6 hold. Then under the null hypothesis, there exists a normal random vector* $\tilde{N}_n \sim \mathcal{N}(0, A_n)$ *such that*

$$\left\| n^{-1/2}\sum_{t=1}^{n} P(X_t^{\star})u_t^{\star} - \tilde{N}_n \right\| = O_p(\zeta_{0,n}^{1-1/\kappa}m_n n^{-1/4} + \zeta_{0,n}^{1/3}m_n^{5/6}n^{-1/6}).$$

Proof of Lemma 10 We use Lemma B1 in the Supplemental Appendix of Li and Liao (2019) to show the claim. For this purpose, it is sufficient to verify Assumptions B2(i-vi). Assumption B2(i) is directly assumed in Assumption 6(i). Under the null hypothesis $u_t^{\star} = \mathbb{1}(Y_{t+1} \leq f_t(\theta^{\star})) - q$ and $(u_t^{\star})_t$ is a martingale difference sequence, which implies that

$$\mathbb{E}\left[(u_t^{\star})^2 \middle| \mathcal{F}_t\right] = \mathbb{E}\left[\mathbb{1}(Y_{t+1} \leq f_t(\theta^{\star}))\middle| \mathcal{F}_t\right] - 2q\mathbb{E}\left[\mathbb{1}(Y_{t+1} \leq f_t(\theta^{\star}))\middle| \mathcal{F}_t\right] + q^2 = q(1-q)$$

which implies that Assumptions B2(ii, iii) hold. By (3.41) Assumptions B2(iv) also holds. Assumptions B2(v, vi) are maintained in Assumption 6(iv, v). Therefore, the claim of the lemma follows by Lemma B1 in Li and Liao (2019). □

Below, we denote $\zeta_n^L \equiv \sup_{x_1,x_2\in\mathcal{X}} \|P(x_1) - P(x_2)\| / \|x_1 - x_2\|$. The next assumption is needed.

Assumption 7. *(i) There exists a random vector $L_{F,t}$ with $\mathbb{E}[\|L_{F,t}\|^2] \leq K$ such that*

$$\left|F_{t+1|t}(f_t(\theta)) - F_{t+1|t}\left(f_t(\theta^\star)\right) - L_{F,t}^\top(\theta - \theta^\star)\right| \leq K\|\theta - \theta^\star\|^2$$

for any $\theta \in B_n(\theta^\star)$; (ii) let $g_j(x) \equiv \mathbb{E}\left[L_{F,t}(j) \middle| X_t^\star = x\right]$ then $\sup_{x\in\mathcal{X}} \|g(x)\| \leq K$ and there exist $\rho_g > 0$ and $b_{g_j,n}^\star \in \mathbb{R}^{m_n}$ such that

$$\sup_{x\in\mathcal{X}} |g_{j,m_n}(x) - g_j(x)| = O(m_n^{-\rho_g})$$

where $g_{j,m_n}(\cdot) = P(\cdot)^\top b_{g_j,n}^\star$ for $j = 1, \ldots, d_\theta$; (iii) let $v_{j,t} \equiv L_{F,t}(j) - g_j(X_t^\star)$ then $\lambda_{\max}(\Omega_{v_j,n}) \leq K$ where $\Omega_{v_j,n} = Var(n^{-1/2}\sum_{t=1}^n P(X_t^\star)v_{j,t})$; (iv) $\sup_{x\in\mathcal{X}} \|P(x)\|^{-1} = o_p(((\log(m_n))^{-1/2})$ for all m_n; (v) $\log(\zeta_n^L) = O(\log(n))$ and $\zeta_{0,n}^{1/3} m_n^{5/6} n^{-1/6} = o((\log(m_n))^{-1/2})$.

Proof of Theorem 3 By Assumptions 6(iv, vi), Lemma 4 and Lemma 9, we have

$$\left\|\hat{\Sigma}_n - \Sigma_n\right\| = O_p(\zeta_{0,n} m_n^{1/2}\delta_{c,n}^{1/2} + \zeta_{0,n}^{2-2/\kappa} m_n n^{-1/2} + \zeta_{1,n} m_n^{1/2} n^{-1/2}). \tag{3.67}$$

By Assumption 6(iv),

$$K^{-1} \leq \lambda_{\min}(\Sigma_n) \leq \lambda_{\max}(\Sigma_n) \leq K \tag{3.68}$$

which together with (3.67) and Assumptions 6(vi) implies that

$$(2K)^{-1} \leq \lambda_{\min}(\hat{\Sigma}_n) \leq \lambda_{\max}(\hat{\Sigma}_n) \leq 2K \tag{3.69}$$

with probability approaching 1. By (3.67), (3.68) and (3.69),

$$\begin{aligned}
\sup_{x\in\mathcal{X}}\left|\frac{\sigma_n(x)}{\hat{\sigma}_n(x)}-1\right| &\leq \sup_{x\in\mathcal{X}}\left|\frac{\hat{\sigma}_n(x)-\sigma_n(x)}{\hat{\sigma}_n(x)(\hat{\sigma}_n(x)+\sigma_n(x))}\right| \\
&\leq K(\lambda_{\min}(\hat{\Sigma}_n)+\lambda_{\min}(\Sigma_n))\sup_{x\in\mathcal{X}}\left|\frac{\hat{\sigma}_n^2(x)-\sigma_n^2(x)}{\|P(x)\|^2}\right| \\
&\leq K(\lambda_{\min}(\hat{\Sigma}_n)+\lambda_{\min}(\Sigma_n))\left\|\hat{\Sigma}_n-\Sigma_n\right\| \\
&= O_p(\zeta_{0,n}m_n^{1/2}\delta_{c,n}^{1/2}+\zeta_{0,n}^{2-2/\kappa}m_n n^{-1/2}+\zeta_{1,n}m_n^{1/2}n^{-1/2}). \qquad (3.70)
\end{aligned}$$

By Assumption 6(vi), Lemma 6 and Lemma 11 below, (3.38), (3.43) and (3.45),

$$n^{1/2}(\widehat{b}_n-b_n^\star)=Q_n^{-1}\left(n^{-1/2}\sum_{t=1}^n P(X_t^\star)u_t^\star\right)+o_p((\log m_n)^{-1/2}) \qquad (3.71)$$

under the null hypothesis. Therefore by Lemma 10, (3.69), (3.71), Assumptions 6(ii, iv) and 7(iv, v),

$$\begin{aligned}
\sup_{x\in\mathcal{X}}\frac{n^{1/2}(\widehat{h}_n(x)-h(x))}{\hat{\sigma}_n(x)} &= \sup_{x\in\mathcal{X}}\frac{n^{1/2}(\widehat{h}_n(x)-h_{m_n}(x))}{\hat{\sigma}_n(x)}+O_p(n^{1/2}m_n^{-\rho_h}) \\
&= \sup_{x\in\mathcal{X}}\frac{P(x)^\top Q_n^{-1}}{\hat{\sigma}_n(x)}n^{-1/2}\sum_{t=1}^n P(X_t^\star)u_t^\star+o_p((\log m_n)^{-1/2}) \\
&= \sup_{x\in\mathcal{X}}\frac{P(x)^\top Q_n^{-1}\tilde{N}_n}{\hat{\sigma}_n(x)}+o_p((\log m_n)^{-1/2}) \qquad (3.72)
\end{aligned}$$

where $\tilde{N}_n\sim\mathcal{N}(0,A_n)$. By the same arguments of showing (A.74) in Belloni, Chernozhukov, Chetverikov, and Kato (2015), we have

$$\sup_{x\in\mathcal{X}}\frac{P(x)^\top Q_n^{-1}\tilde{N}_n}{\sigma_n(x)}=O_p((\log(m_n))^{1/2}), \qquad (3.73)$$

which together with Assumption 6(vi), (3.70) and (3.72)

$$\sup_{x\in\mathcal{X}}\frac{n^{1/2}(\widehat{h}_n(x)-h(x))}{\hat{\sigma}_n(x)}=\sup_{x\in\mathcal{X}}\frac{P(x)^\top Q_n^{-1}\tilde{N}_n}{\sigma_n(x)}+o_p((\log(m_n))^{-1/2}) \qquad (3.74)$$

which finishes the proof. □

Lemma 11. *Under Assumptions 5, 6 and 7 we have*

$$\sup_{x\in\mathcal{X}} \frac{\left|P(x)^\top \hat{Q}_n^{-1} \sum_{t=1}^n P(\widehat{X}_t)(\widehat{Z}_{t+1} - Z^\star_{t+1})\right|}{n^{1/2}\sigma_n(x)} = o_p((\log m_n)^{-1/2}).$$

Proof of Lemma 11 For any $x \in \mathcal{X}$, we can write

$$\begin{aligned}
&\frac{P(x)^\top \hat{Q}_n^{-1} \sum_{t=1}^n P(\widehat{X}_t)(\widehat{Z}_{t+1} - Z^\star_{t+1})}{n^{1/2}\sigma_n(x)} \\
= \;& \frac{P(x)^\top \hat{Q}_n^{-1} \sum_{t=1}^n P(\widehat{X}_t)(F_{t+1|t}(f_t(\hat{\theta}_n)) - F_{t+1|t}\left(f_t(\theta^\star)\right))}{n^{1/2}\sigma_n(x)} \\
&+ \frac{P(x)^\top \hat{Q}_n^{-1} \sum_{t=1}^n P(\widehat{X}_t)(\widehat{Z}_{t+1} - Z^\star_{t+1} - F_{t+1|t}(f_t(\hat{\theta}_n)) + F_{t+1|t}\left(f_t(\theta^\star)\right))}{n^{1/2}\sigma_n(x)}
\end{aligned} \tag{3.75}$$

The proof is divided into 5 steps.

Step 1. In this step, we study the second term after the equality in (3.75). By Assumption 6(iv),

$$K^{-1} \le \lambda_{\min}(\Sigma_n) \le \lambda_{\max}(\Sigma_n) \le K \tag{3.76}$$

which together with the Cauchy-Schwarz inequality, (3.21) in the proof of Lemma 6 and (3.37) implies that

$$\begin{aligned}
&\sup_{x\in\mathcal{X}} \frac{\left|P(x)^\top \hat{Q}_n^{-1} \sum_{t=1}^n P(\widehat{X}_t)(\widehat{Z}_{t+1} - Z^\star_{t+1} - F_{t+1|t}(f_t(\hat{\theta}_n)) + F_{t+1|t}\left(f_t(\theta^\star)\right))\right|}{n^{1/2}\sigma_n(x)} \\
\le \;& \sup_{x\in\mathcal{X}} \frac{\left\|n^{-1/2} \sum_{t=1}^n P(\widehat{X}_t)(\widehat{Z}_{t+1} - Z^\star_{t+1} - F_{t+1|t}(f_t(\hat{\theta}_n)) + F_{t+1|t}\left(f_t(\theta^\star)\right))\right\|}{\lambda_{\min}(\hat{Q}_n)(\lambda_{\min}(\Sigma_n))^{1/2}} \\
= \;& O_p((\zeta_{1,n}\delta_{c,n}^{3/2} + \zeta_{0,n}\delta_{c,n}^{1/2})m_n^{1/2}) = o_p((\log m_n)^{-1/2})
\end{aligned} \tag{3.77}$$

where the second equality is by 6(vi).

Step 2. In this step, we show that

$$\sup_{x\in\mathcal{X}}\left|\frac{P(x)^\top\hat{Q}_n^{-1}\sum_{t=1}^n P(\widehat{X}_t)(F_{t+1|t}(f_t(\hat{\theta}_n))-F_{t+1|t}\left(f_t(\theta^\star)\right))}{n^{1/2}\sigma_n(x)}\right.$$
$$\left.-\frac{P(x)^\top\hat{Q}_n^{-1}\sum_{t=1}^n P(X_t^\star)L_{F,t}^\top(\hat{\theta}_n-\theta^\star)}{n^{1/2}\sigma_n(x)}\right|$$
$$=o_p((\log m_n)^{-1/2}). \tag{3.78}$$

By the Cauchy-Schwarz inequality, Assumptions 7(i), (3.37) and (3.76),

$$\sup_{x\in\mathcal{X}}\frac{\left|P(x)^\top\hat{Q}_n^{-1}\sum_{t=1}^n P(\widehat{X}_t)(F_{t+1|t}(f_t(\hat{\theta}_n))-F_{t+1|t}\left(f_t(\theta^\star)\right)-L_{F,t}^\top(\hat{\theta}_n-\theta^\star))\right|}{n^{1/2}\sigma_n(x)}$$
$$\leq\frac{\left\|\hat{Q}_n^{-1}n^{-1/2}\sum_{t=1}^n P(\widehat{X}_t)(F_{t+1|t}(f_t(\hat{\theta}_n))-F_{t+1|t}\left(f_t(\theta^\star)\right)-L_{F,t}^\top(\hat{\theta}_n-\theta^\star))\right\|}{(\lambda_{\min}(\Sigma_n))^{1/2}}$$
$$\leq\frac{\left(\sum_{t=1}^n\left(F_{t+1|t}(f_t(\hat{\theta}_n))-F_{t+1|t}\left(f_t(\theta^\star)\right)-L_{F,t}^\top(\hat{\theta}_n-\theta^\star)\right)^2\right)^{1/2}}{(\lambda_{\min}(\hat{Q}_n)\lambda_{\min}(\Sigma_n))^{1/2}}$$
$$\leq\frac{Kn^{1/2}||\hat{\theta}_n-\theta^\star||^2}{(\lambda_{\min}(\hat{Q}_n)\lambda_{\min}(\Sigma_n))^{1/2}}=O_p(n^{-1/2})=o_p((\log m_n)^{-1/2}). \tag{3.79}$$

By the Cauchy-Schwarz inequality, Lemma 3, (3.37), (3.76), Assumption 7(i) and the Markov inequality, we get

$$\sup_{x\in\mathcal{X}}\frac{\left|P(x)^\top\hat{Q}_n^{-1}n^{-1}\sum_{t=1}^n(P(\hat{X}_t)-P(X_t^\star))L_{F,t}(j)\right|^2}{\sigma_n^2(x)}$$
$$\leq\frac{1}{(\lambda_{\min}(\hat{Q}_n))^2\lambda_{\min}(\Sigma_n)}\sum_{l=1}^{m_n}\left(n^{-1}\sum_{t=1}^n(p_l(\widehat{X}_t)-p_l(X_t^\star))L_{F,t}(j))\right)^2$$
$$\leq\frac{n^{-1}\sum_{t=1}^n L_{F,t}^2(j)}{(\lambda_{\min}(\hat{Q}_n))^2\lambda_{\min}(\Sigma_n)}\sum_{l=1}^{m_n}n^{-1}\sum_{t=1}^n(p_l(\widehat{X}_t)-p_l(X_t^\star))^2=O_p(\zeta_{1,n}^2m_nn^{-1})$$

for $j = 1, \ldots, d_\theta$, which together with Assumption 7(vi) implies that

$$\begin{aligned}
&\sup_{x\in\mathcal{X}} \frac{\left|P(x)^\top \hat{Q}_n^{-1} \sum_{t=1}^n (P(\widehat{X}_t) - P(X_t^\star)) L_{F,t}^\top (\hat{\theta}_n - \theta^\star)\right|}{n^{1/2}\sigma_n(x)} \\
\leq\ & n^{1/2} \left\|\hat{\theta}_n - \theta^\star\right\| \sup_{x\in\mathcal{X}} \frac{\left\|P(x)^\top \hat{Q}_n^{-1} n^{-1} \sum_{t=1}^n (P(\widehat{X}_t) - P(X_t^\star)) L_{F,t}^\top\right\|}{\sigma_n(x)} \\
=\ & O_p(\zeta_{1,n} m_n^{1/2} n^{-1/2}) = o_p((\log(m_n))^{-1/2}). \qquad (3.80)
\end{aligned}$$

The claim in (3.78) follows by (3.79), (3.80) and the triangle inequality.

Step 3. In this step, we show that

$$\begin{aligned}
&\sup_{x\in\mathcal{X}} \left| \frac{P(x)^\top \hat{Q}_n^{-1} \sum_{t=1}^n P(X_t^\star) L_{F,t}^\top (\hat{\theta}_n - \theta^\star)}{n^{1/2}\sigma_n(x)} \right. \\
&\qquad \left. - \frac{P(x)^\top Q_n^{-1} \sum_{t=1}^n P(X_t^\star) g(X_t^\star)^\top (\hat{\theta}_n - \theta^\star)}{n^{1/2}\sigma_n(x)} \right| \\
&= o_p((\log m_n)^{-1/2}). \qquad (3.81)
\end{aligned}$$

By the Cauchy-Schwarz inequality and Assumption 7(iii),

$$\mathbb{E}\left[\left\| n^{-1} \sum_{t=1}^n P(X_t^\star) v_{j,t} \right\|^2\right] = n^{-1} tr(\Omega_{v_j,n}) \leq K m_n n^{-1} \qquad (3.82)$$

which together with the Markov inequality implies that

$$n^{-1} \sum_{t=1}^n P(X_t^\star) v_{j,t} = O_p(m_n^{1/2} n^{-1/2}) \text{ for } j = 1, \ldots, d_\theta. \qquad (3.83)$$

By Assumption 7(vi), (3.37), (3.76) and (3.83),

$$\begin{aligned}
&\sup_{x\in\mathcal{X}} \frac{\left|P(x)^\top \hat{Q}_n^{-1} n^{-1/2} \sum_{t=1}^n P(X_t^\star) v_t^\top (\hat{\theta}_n - \theta^\star)\right|}{\sigma_n(x)} \\
\leq\ & n^{1/2} \left\|\hat{\theta}_n - \theta^\star\right\| \frac{\left\|n^{-1}\sum_{t=1}^n P(X_t^\star) v_t^\top\right\|}{\lambda_{\min}(\hat{Q}_n)(\lambda_{\min}(\Sigma_n))^{1/2}} = O_p(m_n^{1/2} n^{-1/2}) \\
=\ & o_p((\log m_n)^{-1/2}). \qquad (3.84)
\end{aligned}$$

By Assumption 7(i) $\mathbb{E}[g_j(X_t^\star)^2] \leq \mathbb{E}\left[L_{F,t}^2(j)\right] \leq K$ which together with the Markov inequality implies that

$$n^{-1}\sum_{t=1}^n g_j(X_t^\star)^2 = O_p(1) \text{ for } j = 1, \ldots, d_\theta. \qquad (3.85)$$

By the Cauchy-Schwarz inequality, Assumption 6(vi), Lemma 4, (3.37), (3.76) and (3.85),

$$\begin{aligned}
&\sup_{x\in\mathcal{X}} \frac{\left|P(x)^\top (\hat{Q}_n^{-1} - Q_n^{-1}) \sum_{t=1}^n P(X_t^\star) g(X_t^\star)^\top (\hat{\theta}_n - \theta^\star)\right|}{n^{1/2}\sigma_n(x)} \\
\leq\ & n^{1/2} \left\|\hat{\theta}_n - \theta^\star\right\| \sup_{x\in\mathcal{X}} \frac{\left\|P(x)^\top (\hat{Q}_n^{-1} - Q_n^{-1}) n^{-1} \sum_{t=1}^n P(X_t^\star) g(X_t^\star)\right\|}{\sigma_n(x)} \\
\leq\ & n^{1/2} \left\|\hat{\theta}_n - \theta^\star\right\| \frac{n^{-1}\sum_{t=1}^n g_j(X_t^\star)^2 \left\|\hat{Q}_n - Q_n\right\|}{\lambda_{\min}(\hat{Q}_n)\lambda_{\min}(Q_n)(\lambda_{\min}(\Sigma_n))^{1/2}} \\
=\ & O_p(\zeta_{0,n}^{2-2/\kappa} m_n n^{-1/2} + \zeta_{1,n} m_n^{1/2} n^{-1/2}) = o_p((\log m_n)^{-1/2}). \qquad (3.86)
\end{aligned}$$

The claim in (3.81) follows by (3.84), (3.86) and the triangle inequality.

Step 4. In this step, we show that

$$\sup_{x\in\mathcal{X}} \frac{\left|P(x)^\top Q_n^{-1} \sum_{t=1}^n P(X_t^\star) g(X_t^\star)^\top (\hat{\theta}_n - \theta^\star) - g(x)^\top (\hat{\theta}_n - \theta^\star)\right|}{n^{1/2}\sigma_n(x)} = o_p((\log m_n)^{-1/2}). \qquad (3.87)$$

By Assumptions 6(iv, vi), 7(ii) and (3.76)

$$\begin{aligned}
& \sup_{x\in\mathcal{X}} \frac{\left|P(x)^\top Q_n^{-1} n^{-1}\sum_{t=1}^n P(X_t^\star) g_j(X_t^\star) - g_j(x)\right|}{\sigma_n(x)} \\
\leq\ & \sup_{x\in\mathcal{X}} \frac{\left|P(x)^\top Q_n^{-1} n^{-1}\sum_{t=1}^n P(X_t^\star)(g_j(X_t^\star) - g_{j,m_n}(X_t^\star))\right|}{\sigma_n(x)} \\
& + \sup_{x\in\mathcal{X}} \frac{|g_j(x) - g_{j,m_n}(x)|}{\sigma_n(x)} \\
\leq\ & \frac{n^{-1}\sum_{t=1}^n (g_j(X_t^\star) - g_{j,m_n}(X_t^\star))^2}{\lambda_{\min}(Q_n)(\lambda_{\min}(\Sigma_n))^{1/2}} + \sup_{x\in\mathcal{X}} \frac{|g_j(x) - g_{j,m_n}(x)|}{\sigma_n(x)} \\
=\ & O_p(m_n^{-\rho_g}) = o_p((\log(m_n))^{-1/2}) \qquad (3.88)
\end{aligned}$$

for $j = 1, \ldots, d_\theta$. By the Cauchy-Schwarz inequality and (3.88)

$$\begin{aligned}
& \sup_{x\in\mathcal{X}} \frac{\left|P(x)^\top Q_n^{-1}\sum_{t=1}^n P(X_t^\star) g(X_t^\star)^\top(\hat\theta_n - \theta^\star) - g(x)^\top(\hat\theta_n - \theta^\star)\right|}{n^{1/2}\sigma_n(x)} \\
\leq\ & n^{1/2}\left\|\hat\theta_n - \theta^\star\right\| \sup_{x\in\mathcal{X}} \frac{\left\|P(x)^\top Q_n^{-1} n^{-1}\sum_{t=1}^n P(X_t^\star) g(X_t^\star) - g(x)\right\|}{\sigma_n(x)}, \qquad (3.89)
\end{aligned}$$

which together with Assumption 5(i) proves (3.87).

Step 5. By the triangle inequality, (3.75), (3.77), (3.78), (3.81) and (3.87),

$$\begin{aligned}
& \sup_{x\in\mathcal{X}} \frac{\left|P(x)^\top \hat Q_n^{-1}\sum_{t=1}^n P(\widehat{X}_t)(\widehat{Z}_{t+1} - Z_{t+1}^\star)\right|}{n^{1/2}\sigma_n(x)} \\
=\ & \sup_{x\in\mathcal{X}} \frac{\left|g(x)^\top(\hat\theta_n - \theta^\star)\right|}{n^{1/2}\sigma_n(x)} + o_p((\log m_n)^{-1/2}) \\
\leq\ & \frac{n^{1/2}\left\|\hat\theta_n - \theta^\star\right\|}{(\lambda_{\min}(\Sigma_n))^{1/2}} \sup_{x\in\mathcal{X}} \frac{\|g(x)\|}{\|P(x)\|} + o_p((\log m_n)^{-1/2}) \\
=\ & o_p((\log m_n)^{-1/2})
\end{aligned}$$

where the second equality is by Assumptions 5(i), 7(ii, iv) and (3.76). □

Notations:

$$\begin{array}{ll}
\hat{Q}_n = n^{-1}\sum_{t=1}^n P(\hat{X}_t)P(\widehat{X}_t)^\top, & \hat{Q}_n^* = n^{-1}\sum_{t=1}^n P(\hat{X}_t^*)P(\widehat{X}_t^*)^\top, \\
\widehat{u}_t = \widehat{Z}_{t+1} - P(\widehat{X}_t)^\top \widehat{b}_n, & \widehat{u}_t^* = \widehat{Z}_{t+1}^* - P(\widehat{X}_t^*)^\top \widehat{b}_n^*, \\
\hat{A}_n = n^{-1}\sum_{t=1}^n P(\hat{X}_t)P(\widehat{X}_t)^\top \widehat{u}_t^2, & \hat{A}_n^* = n^{-1}\sum_{t=1}^n P(\widehat{X}_t^*)P(\widehat{X}_t^*)^\top (\widehat{u}_t^*)^2, \\
\widehat{\Sigma}_n = \hat{Q}_n^{-1}\hat{A}_n\hat{Q}_n^{-1}, & \widehat{\Sigma}_n^* = (\hat{Q}_n^*)^{-1}\hat{A}_n^*(\hat{Q}_n^*)^{-1}.
\end{array}$$

Lemma 12. *Under Assumptions 5 and 6,* $\left\|\hat{Q}_n^* - \hat{Q}_n\right\|_S = O_p(\zeta_{0,n}(\log(m_n)m_n n^{-1})^{1/2})$.

Proof of Lemma 12 Let $\mathbb{E}^*[\cdot]$ denote the conditional expectation with respect to the bootstrap distribution given the data. Let

$$D_t^* = P(\widehat{X}_t^*)P(\widehat{X}_t^*)^\top - \mathbb{E}^*[P(\widehat{X}_t^*)P(\widehat{X}_t^*)^\top] \text{ for } t = 1, \ldots, n.$$

Let $R_{D,n} = 2\max_{t\le n}||P(\widehat{X}_t)||^2$ and $\sigma_{D,n}^2 = n\lambda_{\max}(\hat{Q}_n)\max_{t\le n}||P(\widehat{X}_t)||^2$. Then we have

$$\mathbb{E}^*[D_t^*] = 0,\ \max_{t\le n}\|D_t^*\|_S \le R_{D,n} \text{ and } \left\|\sum_{t=1}^n \mathbb{E}^*\left[(D_t^*)^2\right]\right\|_S \le \sigma_{D,n}^2. \tag{3.90}$$

Since D_t^* is i.i.d. conditioning on data, by (3.90) we invoke the matrix Bernstein inequality (see, e.g. Theorem 1.4 in Tropp (2012)) to get

$$\begin{aligned}
& \mathbb{P}^*\left(\left\|\hat{Q}_n^* - \hat{Q}_n\right\|_S \ge C(\log(m_n)R_{D,n}n^{-1})^{1/2}\right) \\
= \ & \mathbb{P}^*\left(\left\|\sum_{t=1}^n D_t^*\right\|_S \ge C(\log(m_n)R_{D,n}n)^{1/2}\right) \\
\le \ & 2m_n \exp\left(\frac{-C^2\log(m_n)R_{D,n}n/2}{\sigma_{D,n}^2 + C(\log(m_n))^{1/2}R_{D,n}^{3/2}n^{1/2}/3}\right) \\
\le \ & 2m_n \exp\left(\frac{-C\log(m_n)/2}{\lambda_{\max}(\hat{Q}_n)/C + (\log(m_n)R_n n^{-1})^{1/2}/3}\right)
\end{aligned} \tag{3.91}$$

where the C is any finite constant and the second inequality is by $\sigma_{D,n}^2 \le n\lambda_{\max}(\hat{Q}_n)R_{D,n}$. By Assumption 6(v) and (3.5), $R_{D,n} \le \zeta_{0,n}^2 m_n$ with probability approaching 1. By

(3.37), $\lambda_{\max}(\hat{Q}_n) \leq 2K$ with probability approaching 1. Therefore, by (3.91) and $\zeta_{0,n}^2 m_n n^{-1} = o(1)$ (which is imposed in Assumption 6(vi)), we get

$$\left\| \hat{Q}_n - \hat{Q}_n^* \right\|_S = O_p((\log(m_n) R_{D,n} n^{-1})^{1/2}) = O_p(\zeta_{0,n}(\log(m_n) m_n n^{-1})^{1/2})$$

which finishes the proof. □

Proposition 1. *Under Assumptions 5 and 6, we have* $||\widehat{b}_n^* - \widehat{b}_n|| = O_p(\delta_{b,n}^*)$ *where*

$$\delta_{b,n}^* = \begin{cases} O_p\left(\zeta_{0,n}(\log(m_n) m_n n^{-1})^{1/2}\right) & \textit{in general} \\ O_p(m_n^{1/2} n^{-1/2}) & \textit{under the null} \end{cases}.$$

Proof of Proposition 1 Since $\mathbb{E}^*[P(\widehat{X}_t^*)\widehat{Z}_{t+1}^*] = n^{-1}\sum_{t=1}^n P(\widehat{X}_t)\widehat{Z}_{t+1}$, we have

$$\begin{aligned} \widehat{b}_n^* - \widehat{b}_n &= (\hat{Q}_n^*)^{-1}\left(n^{-1}\sum_{t=1}^n P(\widehat{X}_t^*)\widehat{Z}_{t+1}^*\right) - (\hat{Q}_n)^{-1}\left(n^{-1}\sum_{t=1}^n P(\hat{X}_t)\widehat{Z}_{t+1}\right) \\ &= (\hat{Q}_n^*)^{-1}\left(n^{-1}\sum_{t=1}^n (P(\widehat{X}_t^*)\widehat{Z}_{t+1}^* - \mathbb{E}^*[P(\hat{X}_t^*)\widehat{Z}_{t+1}^*])\right) \\ &\quad + ((\hat{Q}_n^*)^{-1} - (\hat{Q}_n)^{-1})\left(n^{-1}\sum_{t=1}^n P(\widehat{X}_t)\widehat{Z}_{t+1}\right). \end{aligned} \tag{3.92}$$

By Assumption 6(vi), Lemma 12 and (3.37),

$$(2K)^{-1} \leq \lambda_{\min}(\hat{Q}_n^*) \leq \lambda_{\max}(\hat{Q}_n^*) \leq 2K \tag{3.93}$$

with probability approaching 1. Note that $n^{-1}\sum_{t=1}^n (P(\widehat{X}_t^*)\widehat{Z}_{t+1}^* - \mathbb{E}^*[P(\widehat{X}_t^*)\widehat{Z}_{t+1}^*])$

is an average of conditionally independent zero-mean elements. Therefore

$$\begin{aligned}
& \mathbb{E}^* \left[\left\| n^{-1} \sum_{t=1}^{n} (P(\widehat{X}_t^*)\widehat{Z}_{t+1}^* - \mathbb{E}^*[P(\widehat{X}_t^*)\hat{Z}_{t+1}^*]) \right\|^2 \right] \\
= \;& n^{-1}\mathbb{E}^* \left[\left\| P(\widehat{X}_t^*)\widehat{Z}_{t+1}^* - \mathbb{E}^*[P(\widehat{X}_t^*)\widehat{Z}_{t+1}^*] \right\|^2 \right] \\
\leq \;& n^{-1}\mathbb{E}^* \left[\left\| P(\widehat{X}_t^*)\widehat{Z}_{t+1}^* \right\|^2 \right] \\
\leq \;& n^{-2} \sum_{t=1}^{n} \left\| P(\hat{X}_t) \right\|^2 = n^{-1} trace(\hat{Q}_n) = O_p\left(m_n n^{-1}\right) \qquad (3.94)
\end{aligned}$$

where the second inequality is by $\left|\widehat{Z}_{t+1}^*\right| \leq 1$ for any t and $\mathbb{E}^*[||P(\widehat{X}_t^*)||^2] = n^{-1}\sum_{t=1}^{n} ||P(\widehat{X}_t)||^2$, the last equality is by (3.37). Combining the results in (3.93) and (3.94), we get

$$(\hat{Q}_n^*)^{-1} \left(n^{-1} \sum_{t=1}^{n} (P(\widehat{X}_t^*)\widehat{Z}_{t+1}^* - \mathbb{E}^*[P(\hat{X}_t^*)\widehat{Z}_{t+1}^*]) \right) = O_p(m_n^{1/2} n^{-1/2}). \qquad (3.95)$$

Since $|\widehat{Z}_{t+1}^*| \leq 1$ for any t, by (3.37) we obtain

$$\begin{aligned}
& \left\| n^{-1} \sum_{t=1}^{n} P(\hat{X}_t)\widehat{Z}_{t+1} \right\|^2 \\
\leq \;& \frac{\lambda_{\max}(\hat{Q}_n)}{n} \left(\sum_{t=1}^{n} \widehat{Z}_{t+1} P(\widehat{X}_t)^\top \right) \left(\sum_{t=1}^{n} P(\hat{X}_t) P(\widehat{X}_t)^\top \right)^{-1} \left(\sum_{t=1}^{n} P(\widehat{X}_t)\widehat{Z}_{t+1} \right) \\
\leq \;& \lambda_{\max}(\hat{Q}_n) n^{-1} \sum_{t=1}^{n} \widehat{Z}_{t+1}^2 = O_p(1). \qquad (3.96)
\end{aligned}$$

By Lemma 12, (3.37), (3.93) and (3.96),

$$((\hat{Q}_n^*)^{-1} - (\hat{Q}_n)^{-1}) \left(n^{-1} \sum_{t=1}^{n} P(\hat{X}_t)\widehat{Z}_{t+1} \right) = O_p(\zeta_{0,n}(\log(m_n) m_n n^{-1})^{1/2}) \qquad (3.97)$$

which together with (3.95) finishes the proof. □

Lemma 13. *Under Assumptions 5 and 6, we have*

$$\left\|\hat{A}_n^* - \hat{A}_n\right\|_S = O_p(\zeta_{0,n} m_n^{1/2} \delta_{b,n}^* + \zeta_{0,n}^2 m_n \delta_{b,n} \delta_{b,n}^* + \zeta_{0,n}^2 m_n \delta_{b,n}^{*2}).$$

Proof of Lemma 13 We can rewrite

$$\widehat{u}_t^* = \widehat{Z}_{t+1}^* - P(\widehat{X}_t^*)^\top \widehat{b}_n^* = \widehat{Z}_{t+1}^* - P(\widehat{X}_t^*)^\top \widehat{b}_n - P(\widehat{X}_t^*)^\top (\widehat{b}_n^* - \widehat{b}_n).$$

Therefore,

$$\begin{aligned}
\hat{A}_n^* &= n^{-1} \sum_{t=1}^n P(\widehat{X}_t^*) P(\widehat{X}_t^*)^\top (\widehat{u}_t^*)^2 \\
&= n^{-1} \sum_{t=1}^n P(\widehat{X}_t^*) P(\widehat{X}_t^*)^\top (\widehat{Z}_{t+1}^* - P(\hat{X}_t^*)^\top \widehat{b}_n)^2 \\
&\quad -2n^{-1} \sum_{t=1}^n P(\widehat{X}_t^*) P(\widehat{X}_t^*)^\top (\widehat{Z}_{t+1}^* - P(\hat{X}_t^*)^\top \widehat{b}_n) P(\widehat{X}_t^*)^\top (\widehat{b}_n^* - \widehat{b}_n) \\
&\quad +n^{-1} \sum_{t=1}^n P(\widehat{X}_t^*) P(\widehat{X}_t^*)^\top (P(\widehat{X}_t^*)^\top (\widehat{b}_n^* - \widehat{b}_n))^2 \\
&\equiv R_{1n} + 2R_{2n} + R_{3n}. \qquad (3.98)
\end{aligned}$$

We analyze these terms in turn, starting with the (leading) term R_{1n} defined as

$$R_{1n} \equiv n^{-1} \sum_{t=1}^n P(\hat{X}_t^*) P(\widehat{X}_t^*)^\top (\widehat{Z}_{t+1}^* - P(\hat{X}_t^*)^\top \widehat{b}_n)^2.$$

Note that $\widehat{b}_n$ is $\mathcal{D}_n$-measurable where $\mathcal{D}_n$ is the σ-field generated by data and R_{1n} is the average of conditionally i.i.d. variables. The conditional mean of each summand

term is

$$\begin{aligned}
&\mathbb{E}^*\left[P(\widehat{X}_t^*)P(\widehat{X}_t^*)^\top(\widehat{Z}_{t+1}^* - P(\hat{X}_t^*)^\top\widehat{b}_n)^2\right] \\
= \;& n^{-1}\sum_{t=1}^{n} P(\widehat{X}_t)P(\widehat{X}_t)^\top(\widehat{Z}_{t+1} - P(\widehat{X}_t)^\top\widehat{b}_n)^2 \\
= \;& n^{-1}\sum_{t=1}^{n} P(\widehat{X}_t)P(\widehat{X}_t)^\top\widehat{u}_t^2 = \hat{A}_n. \qquad (3.99)
\end{aligned}$$

By (3.99), the conditional variance of each summand term satisfies

$$\begin{aligned}
&\mathbb{E}^*\left[\left\|n^{-1}\sum_{t=1}^{n} P(\widehat{X}_t^*)P(\widehat{X}_t^*)^\top(\widehat{Z}_{t+1}^* - P(\widehat{X}_t^*)^\top\widehat{b}_n)^2 - \hat{A}_n\right\|^2\right] \\
= \;& \sum_{l_1,l_2=1}^{m_n}\mathbb{E}^*\left[\left|n^{-1}\sum_{t=1}^{n} p_{l_1}(\widehat{X}_t^*)p_{l_2}(\widehat{X}_t^*)(\widehat{Z}_{t+1}^* - P(\widehat{X}_t^*)^\top\widehat{b}_n)^2 - \hat{A}_n(l_1,l_2)\right|^2\right] \\
\leq \;& n^{-1}\sum_{l_1,l_2=1}^{m_n}\mathbb{E}^*\left[\left(p_{l_1}(\widehat{X}_t^*)p_{l_2}(\widehat{X}_t^*)(\widehat{Z}_{t+1}^* - P(\widehat{X}_t^*)^\top\widehat{b}_n)^2\right)^2\right] \\
= \;& n^{-2}\sum_{t=1}^{n}\sum_{l_1,l_2=1}^{m_n}\left(p_{l_1}(\widehat{X}_t)p_{l_2}(\hat{X}_t)(\widehat{Z}_{t+1} - P(\widehat{X}_t)^\top\widehat{b}_n)^2\right)^2 \\
= \;& n^{-2}\sum_{t=1}^{n}\left\|P(\hat{X}_t)\right\|^4\widehat{u}_t^4. \qquad (3.100)
\end{aligned}$$

By Assumptions 6(v, vi), (3.41), (3.5), (3.58) and (3.94), we obtain

$$\begin{aligned}
n^{-2}\sum_{t=1}^{n}\left\|P(\hat{X}_t)\right\|^4\widehat{u}_t^4 \leq \;& Kn^{-2}\sum_{t=1}^{n}\left\|P(\hat{X}_t)\right\|^4(u_t^\star)^4 + Kn^{-2}\sum_{t=1}^{n}\left\|P(\hat{X}_t)\right\|^4(\widehat{u}_t - u_t^\star)^4 \\
\leq \;& Kn^{-2}\sum_{t=1}^{n}\left\|P(\hat{X}_t)\right\|^4 + \zeta_{0,n}^4 m_n^2 n^{-2}\sum_{t=1}^{n}(\widehat{u}_t - u_t^\star)^4 \\
&= O_p(\zeta_{0,n}^2 m_n^2 n^{-1}). \qquad (3.101)
\end{aligned}$$

Therefore

$$R_{1n} - \hat{A}_n = n^{-1}\sum_{t=1}^{n} P(\widehat{X}_t^*)P(\widehat{X}_t^*)^\top(\widehat{Z}_{t+1}^* - P(\widehat{X}_t^*)^\top\widehat{b}_n)^2 - \hat{A}_n = O_p(\zeta_{0,n}m_n n^{-1/2}). \tag{3.102}$$

We now deal with

$$R_{2n} \equiv n^{-1}\sum_{t=1}^{n} P(\hat{X}_t^*)P(\widehat{X}_t^*)^\top(\widehat{Z}_{t+1}^* - P(\hat{X}_t^*)^\top\widehat{b}_n)(P(\widehat{X}_t^*)^\top(\widehat{b}_n^* - \widehat{b}_n)).$$

Consider any $a_{m_n} \in \mathbb{R}^{m_n}$ with $a_{m_n}^\top a_{m_n} = 1$. By the Cauchy-Schwarz inequality

$$\begin{aligned}
a_{m_n}^\top R_{2n}^2 a_{m_n} &\leq \lambda_{\max}(\hat{Q}_n^*)n^{-1}\sum_{t=1}^{n}(P(\widehat{X}_t^*)^\top a_{m_n})^2(\widehat{Z}_{t+1}^* - P(\widehat{X}_t^*)^\top\widehat{b}_n)^2(P(\hat{X}_t^*)^\top(\widehat{b}_n^* - \widehat{b}_n))^2 \\
&\leq \lambda_{\max}(\hat{Q}_n^*)\max_{t\leq n}\left\|P(\widehat{X}_t^*)\right\|^2 ||\widehat{b}_n^* - \widehat{b}_n||^2 \\
&\cdot\ n^{-1}\sum_{t=1}^{n}(P(\hat{X}_t^*)^\top a_{m_n})^2(\widehat{Z}_{t+1}^* - P(\widehat{X}_t^*)^\top\widehat{b}_n)^2 \\
&\leq (\lambda_{\max}(\hat{Q}_n^*))^2||\widehat{b}_n^* - \widehat{b}_n||^2\max_{t\leq n}\left\|P(\widehat{X}_t^*)\right\|^2\max_{t\leq n}(\widehat{Z}_{t+1}^* - P(\widehat{X}_t^*)^\top\widehat{b}_n)^2. \qquad (3.103)
\end{aligned}$$

Since $|\widehat{Z}_{t+1}^*| \leq K$ and $|h(\hat{X}_t^*)| \leq K$ for any t, by Assumptions 6(v, vi), (3.5), (3.93) and Proposition 1

$$\begin{aligned}
\max_{t\leq n}\left|\widehat{Z}_{t+1}^* - P(\widehat{X}_t^*)^\top\widehat{b}_n\right| &\leq \max_{t\leq n}\left|\widehat{Z}_{t+1}^* - h(\widehat{X}_t^*)\right| + \max_{t\leq n}\left|h(\widehat{X}_t^*) - P(\widehat{X}_t^*)^\top b_n^\star\right| \\
&\quad + \max_{t\leq n}\left|P(\widehat{X}_t^*)^\top(\widehat{b}_n - b_n^\star)\right| \\
&= O_p(1 + m_n^{-\rho_h} + \zeta_{0,n}m_n^{1/2}\delta_{b,n}) = O_p(1) \qquad (3.104)
\end{aligned}$$

which together with (3.5), (3.93), (3.103) and Proposition 1 implies that

$$\begin{aligned}
\|R_{2n}\|_S &\leq \lambda_{\max}(\hat{Q}_n^*)||\widehat{b}_n^* - \widehat{b}_n||\max_{t\leq n}\left\|P(\hat{X}_t^*)\right\|\max_{t\leq n}\left|\hat{Z}_{t+1}^* - P(\widehat{X}_t^*)^\top\widehat{b}_n\right| \\
&= O_p(\zeta_{0,n}m_n^{1/2}\delta_{b,n}^* + \zeta_{0,n}^2 m_n\delta_{b,n}\delta_{b,n}^*). \qquad (3.105)
\end{aligned}$$

We now deal with

$$R_{3n} \equiv n^{-1} \sum_{t=1}^{n} P(\hat{X}_t^*) P(\widehat{X}_t^*)^\top (P(\widehat{X}_t^*)^\top (\widehat{b}_n^* - \widehat{b}_n))^2.$$

Consider any $a_{m_n} \in \mathbb{R}^{m_n}$ with $a_{m_n}^\top a_{m_n} = 1$. Then by (3.93) and the Cauchy-Schwarz inequality

$$\begin{aligned} a_{m_n}^\top R_{3n}^2 a_{m_n} &\leq \lambda_{\max}(\hat{Q}_n^*) n^{-1} \sum_{t=1}^{n} (P(\widehat{X}_t^*)^\top a_{m_n})^2 (P(\widehat{X}_t^*)^\top (\widehat{b}_n^* - \widehat{b}_n))^4 \\ &\leq (\lambda_{\max}(\hat{Q}_n^*))^2 \max_{t \leq n} \left\| P(\widehat{X}_t^*) \right\|^4 ||\widehat{b}_n^* - \widehat{b}_n||^4 \end{aligned} \tag{3.106}$$

which together with Assumption 6(v), (3.5), (3.93) and Proposition 1 implies that

$$\|R_{3n}\|_S \leq \lambda_{\max}(\hat{Q}_n^*) ||\widehat{b}_n^* - \widehat{b}_n||^2 \max_{t \leq n} \left\| P(\hat{X}_t^*) \right\|^2 = O_p(\zeta_{0,n}^2 m_n \delta_{b,n}^{*2}). \tag{3.107}$$

Collecting the results in (3.98), (3.102), (3.105) and (3.107), we get

$$\begin{aligned} \left\| \hat{A}_n^* - \hat{A}_n \right\|_S &\leq \left\| R_{1n} - \hat{A}_n \right\|_S + \|R_{2n}\|_S + \|R_{3n}\|_S \\ &= O_p(\zeta_{0,n} m_n^{1/2} \delta_{b,n}^* + \zeta_{0,n}^2 m_n \delta_{b,n} \delta_{b,n}^* + \zeta_{0,n}^2 m_n \delta_{b,n}^{*2}) \end{aligned}$$

which finishes the proof. □

Lemma 14. *Suppose that Assumptions 5 and 6 holds. Then under the null hypothesis,*

$$\left\| \hat{H}_n^* - \hat{A}_n \right\|_S = O_p\left(\zeta_{0,n} m_n n^{-1/2}\right).$$

where $\hat{H}_n^* \equiv \mathbb{E}^*\left[(\hat{Z}_{t+1}^*)^2 P(\widehat{X}_t^*) P(\widehat{X}_t^*)^\top\right] - \mathbb{E}^*[\widehat{Z}_{t+1}^* P(\widehat{X}_t^*)] \mathbb{E}^*[P(\widehat{X}_t^*)^\top \widehat{Z}_{t+1}^*]$.

Proof of Lemma 14 Note that

$$\mathbb{E}^*[P(\widehat{X}_t^*) \widehat{Z}_{t+1}^*] = n^{-1} \sum_{t=1}^{n} P(\widehat{X}_t) \widehat{Z}_{t+1} = \hat{Q}_n \widehat{b}_n. \tag{3.108}$$

Under the null hypothesis, $h(x) = 0$ and $b_n^\star = 0$. By (3.37), Lemma 8 and Assumption 6(vi),

$$\hat{Q}_n \widehat{b}_n = O_p(m_n^{1/2} n^{-1/2}) \tag{3.109}$$

which together with (3.108) implies that

$$\mathbb{E}^*[\widehat{Z}^*_{t+1} P(\hat{X}^*_t)]\mathbb{E}^*[P(\widehat{X}^*_t)^\top \widehat{Z}^*_{t+1}] = O_p(m_n n^{-1}). \tag{3.110}$$

Since $\widehat{u}_t = \widehat{Z}_{t+1} - P(\hat{X}_t)^\top \widehat{b}_n$,

$$\begin{aligned}
& \mathbb{E}^* \left[(\widehat{Z}^*_{t+1})^2 (P(\widehat{X}^*_t) P(\widehat{X}^*_t)^\top) \right] - \hat{A}_n \\
= \;& n^{-1} \sum_{t=1}^n (\widehat{Z}_{t+1})^2 (P(\hat{X}_t) P(\widehat{X}_t)^\top) - \hat{A}_n \\
= \;& n^{-1} \sum_{t=1}^n (P(\widehat{X}_t)^\top \widehat{b}_n + \widehat{u}_t)^2 \left(P(\widehat{X}_t) P(\widehat{X}_t)^\top \right) - \hat{A}_n \\
= \;& 2n^{-1} \sum_{t=1}^n \widehat{u}_t (P(\hat{X}_t)^\top \widehat{b}_n) P(\widehat{X}_t) P(\hat{X}_t)^\top \\
+ \;& n^{-1} \sum_{t=1}^n (P(\widehat{X}_t)^\top \widehat{b}_n)^2 P(\widehat{X}_t) P(\widehat{X}_t)^\top .
\end{aligned} \tag{3.111}$$

Let $a_{m_n} \in \mathbb{R}^{m_n}$ be such that $\|a_{m_n}\| \leq 1$. By the Cauchy-Schwarz inequality, Assumption 6(v), (3.5), (3.37) and (3.109),

$$\begin{aligned}
& a_{m_n}^\top \left(n^{-1} \sum_{t=1}^n (P(\widehat{X}_t)^\top \widehat{b}_n)^2 \left(P(\widehat{X}_t) P(\widehat{X}_t)^\top \right) \right)^2 a_{m_n} \\
\leq \;& n^{-1} \sum_{t=1}^n (P(\widehat{X}_t)^\top \widehat{b}_n)^4 (a_{m_n}^\top P(\widehat{X}_t))^2 \\
\leq \;& (\lambda_{\max}(\hat{Q}_n))^2 \max_{1 \leq t \leq n} (P(\widehat{X}_t)^\top \widehat{b}_n)^4 \\
\leq \;& (\lambda_{\max}(\hat{Q}_n))^2 \max_{1 \leq t \leq n} \left\| P(\widehat{X}_t) \right\|^4 ||\widehat{b}_n||^4 = O_p(\zeta_{0,n}^4 m_n^4 n^{-2})
\end{aligned}$$

uniformly over a_{m_n}, which implies that

$$\left\| n^{-1} \sum_{t=1}^{n} (P(\hat{X}_t)^\top \widehat{b}_n)^2 \left(P(\widehat{X}_t) P(\hat{X}_t)^\top \right) \right\|_S = O_p(\zeta_{0,n}^2 m_n^2 n^{-1}). \tag{3.112}$$

By the Cauchy-Schwarz inequality, Assumptions 6(iv, v, vi), (3.5), (3.37), Lemma 9 and (3.109), we have uniformly over a_{m_n}

$$\begin{aligned}
& a_{m_n}^\top \left(n^{-1} \sum_{t=1}^{n} \widehat{u}_t (P(\widehat{X}_t)^\top \widehat{b}_n) \left(P(\widehat{X}_t) P(\widehat{X}_t)^\top \right) \right)^2 a_{m_n} \\
= \; & n^{-1} \sum_{t=1}^{n} (P(\widehat{X}_t)^\top \widehat{b}_n) \left(a_{m_n}^\top P(\widehat{X}_t) \widehat{u}_t \right) P(\widehat{X}_t)^\top \\
& \times n^{-1} \sum_{t=1}^{n} P(\widehat{X}_t) \left(a_{m_n}^\top P(\widehat{X}_t) \widehat{u}_t \right) (P(\widehat{X}_t)^\top \widehat{b}_n) \\
\leq \; & \lambda_{\max}(\hat{Q}_n) n^{-1} \sum_{t=1}^{n} (P(\widehat{X}_t)^\top \widehat{b}_n)^2 \left(a_{m_n}^\top P(\widehat{X}_t) \widehat{u}_t \right)^2 \\
\leq \; & \lambda_{\max}(\hat{Q}_n) \lambda_{\max}(\hat{A}_n) \max_{1 \leq t \leq n} \left\| P(\widehat{X}_t) \right\|^2 ||\widehat{b}_n||^2 = O_p(\zeta_{0,n}^2 m_n^2 n^{-1})
\end{aligned} \tag{3.113}$$

which implies that

$$\left\| n^{-1} \sum_{t=1}^{n} \widehat{u}_t P(\widehat{X}_t)^\top \widehat{b}_n \left(P(\widehat{X}_t) P(\widehat{X}_t)^\top \right) \right\|_S = O_p(\zeta_{0,n} m_n n^{-1/2}). \tag{3.114}$$

The claim of the lemma follows by (3.110), (3.111), (3.112) and (3.114). □

Below, we prove Theorem 4(a) and the size property of the test stated in Theorem 4(b). By the decomposition in (3.92), we can write

$$\begin{aligned}
n^{1/2}(\widehat{b}_n^* - \widehat{b}_n) \; = \; & (\hat{Q}_n^*)^{-1} \left(n^{-1/2} \sum_{t=1}^{n} \left(P(\widehat{X}_t^*) \widehat{Z}_{t+1}^* - \mathbb{E}^*[P(\widehat{X}_t^*) \widehat{Z}_{t+1}^*] \right) \right) \\
& + (\hat{Q}_n^*)^{-1} (\hat{Q}_n - \hat{Q}_n^*)(n^{1/2} \widehat{b}_n).
\end{aligned} \tag{3.115}$$

Under the null hypothesis $h(x) = 0$. Then by (3.37) and (3.109), $n^{1/2}\widehat{b}_n = O_p(m_n^{1/2})$ which together with Assumption 6(vi), Lemma 12 and (3.93) implies that

$$(\hat{Q}_n^*)^{-1}(\hat{Q}_n - \hat{Q}_n^*)(n^{1/2}\widehat{b}_n) = O_p(\zeta_{0,n}(\log(m_n)m_n n^{-1})^{1/2}) = o_p((\log m_n)^{-1/2}). \quad (3.116)$$

By definition, $\hat{H}_n^*$ is the bootstrap variance-covariance matrix. We also denote

$$\hat{B}_n^* \equiv n^{-3/2}\sum_{t=1}^{n}\mathbb{E}^*\left[\left\|P(\widehat{X}_t^*)\widehat{Z}_{t+1}^* - \mathbb{E}^*\left[P(\widehat{X}_t^*)\widehat{Z}_{t+1}^*\right]\right\|^3\right].$$

It is easy to see that

$$\begin{aligned}
\hat{B}_n^* &\leq Kn^{-3/2}\sum_{t=1}^{n}\mathbb{E}^*\left[\left\|P(\widehat{X}_t^*)\widehat{Z}_{t+1}^*\right\|^3\right] \\
&\leq Kn^{-3/2}\sum_{t=1}^{n}\mathbb{E}^*\left[\left\|P(\widehat{X}_t^*)\right\|^3\right] \\
&= Kn^{-3/2}\sum_{t=1}^{n}\left\|P(\hat{X}_t)\right\|^3 = O_p(\zeta_{0,n}^{1/2}m_n^{3/2}n^{-1/2}) \qquad (3.117)
\end{aligned}$$

where the second inequality is by $|\widehat{Z}_{t+1}^*| \leq 1$ for any t and the second equality is by Assumption 6(v) and (3.5). Therefore, by applying Yurinskii's coupling under the $\mathcal{D}_n$-conditional probability, we can construct standard Gaussian vectors ξ_n^* such that

$$\begin{aligned}
&\left\|n^{-1/2}\sum_{t=1}^{n}\left(P(\hat{X}_t^*)\widehat{Z}_{t+1}^* - \mathbb{E}^*\left[P(\widehat{X}_t^*)\widehat{Z}_{t+1}^*\right]\right) - \left(\hat{H}_n^*\right)^{1/2}\xi_n^*\right\| \\
= \; & O_p((\zeta_{0,n}m_n^{5/2}n^{-1/2})^{1/3}) = o_p((\log m_n)^{-1/2}) \qquad (3.118)
\end{aligned}$$

where the second equality is by Assumption 7(v). Combining the results in (3.93),

(3.115), (3.116) and (3.118),

$$\begin{aligned}
&\left\| n^{1/2}(\widehat{b}_n^* - \widehat{b}_n) - (\hat{Q}_n^*)^{-1} \left(\hat{H}_n^*\right)^{1/2} \xi_n^* \right\| \\
\leq\ & \left\| (\hat{Q}_n^*)^{-1}(\hat{Q}_n - \hat{Q}_n^*)(n^{1/2}\widehat{b}_n) \right\| \\
&+ \left\| (\hat{Q}_n^*)^{-1} \left(n^{-1/2} \sum_{t=1}^{n} \left(P(\widehat{X}_t^*)\widehat{Z}_{t+1}^* - \mathbb{E}^* \left[P(\widehat{X}_t^*)\widehat{Z}_{t+1}^* \right] \right) - \left(\hat{H}_n^*\right)^{1/2} \xi_n^* \right) \right\| \\
=\ & o_p((\log m_n)^{-1/2}). \qquad (3.119)
\end{aligned}$$

By the triangle inequality, Assumption 6(vi), Lemma 9 and Lemma 14,

$$\left\| \hat{H}_n^* - A_n \right\|_S \leq \left\| \hat{H}_n^* - \hat{A}_n \right\|_S + \left\| \hat{A}_n - A_n \right\|_S = o_p(m_n^{-1/2} (\log m_n)^{-1/2}) \qquad (3.120)$$

which together with Assumption 6(iv) implies that

$$\left\| (\hat{H}_n^*)^{1/2} - A_n^{1/2} \right\|_S \leq \left\| \hat{H}_n^* - A_n \right\|_S \left\| A_n^{-1} \right\|_S = o_p(m_n^{-1/2} (\log m_n)^{-1/2}). \qquad (3.121)$$

By the triangle inequality, Assumptions 6(iv, vi), Lemma 4, Lemma 12, (3.120) and (3.121), we get

$$\begin{aligned}
&\left\| (\hat{Q}_n^*)^{-1} \left(\hat{H}_n^*\right)^{1/2} - Q_n^{-1} A_n^{1/2} \right\|_S \\
\leq\ & \left\| (\hat{Q}_n^*)^{-1}(\hat{Q}_n^* - Q_n) Q_n^{-1} \left(\hat{H}_n^*\right)^{1/2} \right\|_S \\
&+ \left\| Q_n^{-1} (\left(\hat{H}_n^*\right)^{1/2} - A_n^{1/2}) \right\|_S = O_p(m_n^{-1/2} (\log m_n)^{-1/2}). \qquad (3.122)
\end{aligned}$$

Since $||((\hat{Q}_n^*)^{-1} \left(\hat{H}_n^*\right)^{1/2} - Q_n^{-1} A_n^{1/2})\xi_n^*||^2 \leq \|\xi_n^*\|^2 \, ||(\hat{Q}_n^*)^{-1} \left(\hat{H}_n^*\right)^{1/2} - Q_n^{-1} A_n^{1/2}||_S$, we

have

$$\begin{aligned}
&\mathbb{E}\left[\left\|((\hat{Q}_n^*)^{-1}\left(\hat{H}_n^*\right)^{1/2}-Q_n^{-1}A_n^{1/2})\xi_n^*\right\|^2\middle|\mathcal{D}_n,\mathcal{D}_n^*\right]\\
\leq\ &\mathbb{E}\left[\|\xi_n^*\|^2\right]\left\|(\hat{Q}_n^*)^{-1}\left(\hat{H}_n^*\right)^{1/2}-Q_n^{-1}A_n^{1/2}\right\|_S^2
\end{aligned}$$

where $\mathcal{D}_n^*$ denotes the σ-field generated by boostrapped data, which together with (3.121) and the Markov inequality implies that

$$\left\|(\hat{Q}_n^*)^{-1}\left(\hat{H}_n^*\right)^{1/2}\xi_n^*-Q_n^{-1}A_n^{1/2}\xi_n^*\right\|=o_p((\log m_n)^{-1/2}). \tag{3.123}$$

By (3.119) and (3.123), we have

$$\left\|n^{1/2}(\widehat{b}_n^*-\widehat{b}_n)-Q_n^{-1}A_n^{1/2}\xi_n^*\right\|=o_p((\log m_n)^{-1/2}). \tag{3.124}$$

By Assumption 6(iv), Lemma 12 and Lemma 13

$$\left\|\hat{\mathbf{\Sigma}}_n^*-\mathbf{\Sigma}_n\right\|_S=o_p(m_n^{-1/2}\left(\log m_n\right)^{-1/2}) \tag{3.125}$$

which implies that

$$\begin{aligned}
\sup_{x\in\mathcal{X}}\left|\frac{\sigma_n\left(x\right)}{\widehat{\sigma}_n^*\left(x\right)}-1\right| &= \sup_{x\in\mathcal{X}}\left|\frac{P\left(x\right)^\top(\hat{\mathbf{\Sigma}}_n^*-\mathbf{\Sigma}_n)P\left(x\right)}{\widehat{\sigma}_n^*\left(x\right)\left(\widehat{\sigma}_n^*\left(x\right)+\sigma_n\left(x\right)\right)}\right|\\
&\leq (\lambda_{\min}(\hat{\mathbf{\Sigma}}_n^*))^{-1}(\lambda_{\min}(\mathbf{\Sigma}_n))^{-1/2}\left\|\hat{\mathbf{\Sigma}}_n^*-\mathbf{\Sigma}_n\right\|_S\\
&= o_p(m_n^{-1/2}\left(\log m_n\right)^{-1/2}).
\end{aligned} \tag{3.126}$$

By Assumption 6(iv), $\mathbb{E}[||Q_n^{-1}A_n^{1/2}\xi_n^*||^2]\leq Km_n$ which together with the Markov inequality implies that

$$\left\|Q_n^{-1}A_n^{1/2}\xi_n^*\right\|=O_p(m_n^{1/2}). \tag{3.127}$$

By the triangle inequality, (3.124) and (3.126),

$$\begin{aligned}
& \sup_{x\in\mathcal{X}}\left|\frac{n^{1/2}P\left(x\right)^{\top}\left(\widehat{b}_{n}^{*}-\widehat{b}_{n}\right)}{\widehat{\sigma}_{n}^{*}\left(x\right)}-\frac{P\left(x\right)^{\top}\left(Q_{n}^{-1}A_{n}^{1/2}\xi_{n}^{*}\right)}{\sigma_{n}\left(x\right)}\right| \\
\leq & \sup_{x\in\mathcal{X}}\left|\frac{P\left(x\right)^{\top}\left(n^{1/2}(\widehat{b}_{n}^{*}-\widehat{b}_{n})-Q_{n}^{-1}A_{n}^{1/2}\xi_{n}^{*}\right)}{\widehat{\sigma}_{n}^{*}\left(x\right)}\right| \\
& +\sup_{x\in\mathcal{X}}\left|\frac{P\left(x\right)^{\top}Q_{n}^{-1}A_{n}^{1/2}\xi_{n}^{*}}{\sigma_{n}\left(x\right)}\left(\frac{\sigma_{n}\left(x\right)}{\widehat{\sigma}_{n}^{*}\left(x\right)}-1\right)\right| \\
\leq & (\lambda_{\min}(\hat{\mathbf{\Sigma}}_{n}^{*}))^{-1/2}\left\|n^{1/2}(\widehat{b}_{n}^{*}-\widehat{b}_{n})-Q_{n}^{-1}A_{n}^{1/2}\xi_{n}^{*}\right\| \\
& +(\lambda_{\min}(\mathbf{\Sigma}_{n}))^{-1/2}\left\|Q_{n}^{-1}A_{n}^{1/2}\xi_{n}^{*}\right\|\sup_{x\in\mathcal{X}}\left|\frac{\sigma_{n}\left(x\right)}{\widehat{\sigma}_{n}^{*}\left(x\right)}-1\right| \\
& =o_{p}((\log(m_{n}))^{-1/2}), \qquad (3.128)
\end{aligned}$$

which shows the approximation in Theorem 4(a). Given the results in Theorem 3 and Theorem 4(a), the assertion of the size control can be shown by using similar arguments in the proof of Theorem 5.6 in Belloni, Chernozhukov, Chetverikov, and Kato (2015). We omit the proof for brevity. □

Below, we prove the power property of the test stated in Theorem 4(b). By the triangle inequality and Assumption 6(v)

$$\sup_{x\in\mathcal{X}}\frac{\left|n^{1/2}P\left(x\right)^{\top}\widehat{b}_{n}\right|}{\widehat{\sigma}(x)}\geq n^{1/2}(\lambda_{\max}(\hat{\Sigma}_{n}))^{-1/2}\sup_{x\in\mathcal{X}}\frac{|h(x)|-|\widehat{h}_{n}(x)-h(x)|}{\left\|P\left(x\right)\right\|}. \qquad (3.129)$$

By the triangle inequality, the Cauchy-Schwarz inequality, Assumption 6(ii) and Lemma 8

$$\sup_{x\in\mathcal{X}}\frac{|\widehat{h}_{n}(x)-h(x)|}{\left\|P\left(x\right)\right\|}\leq||\widehat{b}_{n}-b_{n}^{\star}||+\sup_{x\in\mathcal{X}}\frac{|h_{m_{n}}(x)-h(x)|}{\left\|P\left(x\right)\right\|}=O_{p}(\delta_{b,n}) \qquad (3.130)$$

which together with $\sup_{x\in\mathcal{X}}|h(x)|\geq K^{-1}$ under the alternative hypothesis, Assump-

tions 6(v) and 6, (3.69) and (3.129) implies that

$$\widehat{T}_n = \sup_{x \in \mathcal{X}} \frac{\left| n^{1/2} P(x)^\top \widehat{b}_n \right|}{\widehat{\sigma}(x)} \geq n^{1/2} (\zeta_{0,n} m_n^{1/2} K)^{-1} \tag{3.131}$$

with probability approaching 1. By (3.69), Lemma 12 and Lemma 13

$$(2K)^{-1} \leq \lambda_{\min}(\hat{\Sigma}_n^*) \leq \lambda_{\max}(\hat{\Sigma}_n^*) \leq 2K \tag{3.132}$$

with probability approaching 1. Therefore

$$\sup_{x \in \mathcal{X}} \frac{\left| n^{1/2} P(x)^\top (\widehat{b}_n^* - \widehat{b}_n) \right|}{\widehat{\sigma}_n^*(x)} \leq (\lambda_{\min}(\hat{\Sigma}_n^*))^{-1/2} \left\| n^{1/2} (\widehat{b}_n^* - \widehat{b}_n) \right\| = O_p(\zeta_{0,n} (\log(m_n) m_n)^{1/2}) \tag{3.133}$$

where the equality is by Proposition 1 and (3.132). Since $(\log m_n)^{1/2} \zeta_{0,n}^2 m_n n^{-1/2} = o(1)$ under Assumption 6(vi), by (3.131) and (3.133) the test rejects the alternative hypothesis with probability approaching 1. □

4

Joint specification test of Value-at-Risk and Expected Shortfall

4.1 Introduction

In 2013 the Basel Committee on Banking Supervision proposed to replace the current market risk measure, Value-at-Risk (VaR), with Expected Shortfall (ES) (Basel Committee, 2013) by 2019/20. A motivating factor behind this change is the better theoretical properties of ES as it is a coherent risk measure (shown by Artzner et al. (1999)) and it also accounts for tail risk of a portfolio, as applying VaR as a risk measure a researcher is not able to control for the more extreme events. This deficiency can be easily seen from the definition of these two risk measure:

$$\begin{aligned} VaR_{t+1,q} &= F_{t+1}^{-1}(q) \quad q \in (0,1) \\ ES_{t+1,q} &= \mathbb{E}\left[Y_{t+1} | Y_{t+1} \leq VaR_{t+1,q}, \mathcal{F}_t\right] \\ Y_{t+1} | \mathcal{F}_t &\sim F_{t+1}. \end{aligned}$$

That is, q-VaR (or VaR at q confidence level) is simply the q quantile of a specific return, Y_{t+1}, and q-ES (or ES at q confidence level) is the average value of the same return series conditional on that the return is below its q-VaR value. Therefore,

contrary to q-VaR, which is one single point on the distribution curve of the return, q-ES does not only control for that one single point in the distribution but all other points below this distribution ("tail risk").

However, as we can already see from the definition, the ES depends on the VaR estimate, and Fissler and Ziegel (2016) prove this measure is not elicitable separately but only jointly with VaR. This means that there does not exist any loss function which minimized would give a consistent estimate for ES. To give an example for elicitable measure, we can think of the mean as it minimizes the square loss function or the median (or 50% quantile) which minimizes the absolute loss function or any other quantile, such as VaR, which minimizes a specific check loss function.

The lack of elicitability can explain why it has been taking such a long time to implement this measure in practice and why, to this day, VaR is still the main market risk measure. It is theoretically not possible to conduct a specification test for ES independently from VaR. In this paper, our goal is to conduct a specification test for these two risk measures jointly. That is, we answer the question whether the VaR and ES models are correctly specified. To achieve this goal, we implement a new average-t test, based on the idea of sup-t test proposed in Li and Liao (2019) and refined for extremetiles in Horvath et al. (2019) which enables us to test for conditional moment restrictions, such as:

$$\mathbb{E}\left[\mathbb{1}\left(Y_{t+1} \leq VaR_{t+1,q}\right) - q|\mathcal{F}_t\right] = 0 \tag{4.1}$$

$$\mathbb{E}\left[\frac{1}{q}Y_{t+1}\mathbb{1}\left(Y_{t+1} \leq VaR_{t+1,q}\right) - ES_{t+1,q}|\mathcal{F}_t\right] = 0, \tag{4.2}$$

where $\mathcal{F}_t$ is an information set available at time t and $\mathbb{1}(\cdot)$ is the indicator function taking 1 if the condition is satisfied, 0 otherwise. If $VaR_{t+1,q}$ and $ES_{t+1,q}$ models are correctly specified, then the above two moment conditions hold. If either of them is misspecified, then at least one of the moment conditions should be violated. Note, the reason why it is not enough to test the moment condition (4.2) by itself

because it could happen that both $VaR_{t+1,q}$ and $ES_{t+1,q}$ models are misspecified "in different directions", which would lead to the scenario where Equation (4.2) holds but neither of the risk measurement procedures are correctly specified. By testing moment condition (4.2) jointly with (4.1), we can be certain that $VaR_{t+1,q}$ model is correctly specified and therefore $ES_{t+1,q}$ model must be correctly specified to have the second moment condition satisfied.

In this paper, we address the issue how to test multiple moment conditions jointly since the theory in Li and Liao (2019) and Horvath et al. (2019) has been mainly developed and applied to test a single moment condition. Moreover, building on the sup-t statistic we propose an average-t statistic which gives equal weight to all observations in computing the t-statistic and it does not only consider the sup or maximum. As we will see in Section 4.4, the simulation results are more stable using the average-t statistic instead of sup-t statistic.

The rest of the paper is organized as follows: Section 4.2 is a short literature review, Section 4.3 describes the theory behind the sup-t and average-t test, in Section 4.4 we present the size and the power of these tests through Monte Carlo simulation study, in Section 4.5 we apply the average-t test in S&P 500 data for several location-scale models and Section 4.6 concludes.

4.2 Literature review

In this section, we review the (Value-at-Risk and) Expected Shortfall testing literature. The literature on VaR (or quantile) specification tests is abundant comparing to the ES specification test literature as the academic world has only started to give more attention for ES only in the recent years.

Kerkhof and Melenberg (2004) provide a general framework for backtesting risk measurement methods including VaR and ES using the functional delta method. However, to conduct their test, one needs to assume the distribution of the underlying

asset return, which can lead to misspecification error. Neither the sup-t nor the average-t test requires any estimation of the return distribution.

Du and Escanciano (2016) propose a new backtesting procedure for ES based on the idea of VaR backtesting. A general way to test the correct specification of VaR (Escanciano and Olmo, 2010, see e.g.) is to test if Equation 4.1[1] (centered violations/exceedances) is a martingale difference sequence. Since ES can be defined as $ES_{t+1,q} = \int_0^q VaR_{t+1,i}di$ they propose a cumulative violations function and test whether it is a martingale difference sequence to test the correct specification of ES. To conduct their test, one either needs to estimate the cumulative distribution function of the return series in the tail or needs to make an assumption regarding the return distribution.

Nolde and Ziegel (2017) propose a comparative backtest procedure for VaR and ES jointly, based on the loss functions proposed in Fissler and Ziegel (2016). Their goal is to answer the question whether model A is better than model B in forecasting VaR and ES but they do not try to answer whether model A (or B) is correctly specified.

Barendse et al. (2019) investigate the effect of estimation error on VaR and ES backtesting. They propose a test which controls for the estimation error by explicitly including an additional term in the variance-covariance matrix of the test, which includes an estimate of the density function of the return. As we will see in the Theory section (Section 4.3) the sup-t and average-t test implicitly controls for the estimation error and one does not need to compute the density function.

4.3 Theory

This section describes the underlying theory of the joint specification tests. The interested reader is invited to consult with Li and Liao (2019) for more details on

[1] More accurately one needs to test if $\mathbb{1}\left(Y_{t+1} \leq VaR_{t+1,q}\right) - q$ holds.

sup-t test in uniform nonparametric inference for time series data. Refinement of their theory for nonsmooth functions, generated conditinoning variables and VaR testing can be found in Horvath et al. (2019)[2]. In this section we provide an overview of the necessary parts from those two papers and then present the theorem for the average-t test.

Our goal in this paper is to answer whether a sequence of VaR and ES estimates are correctly specified. That is,

$$H_0 : \text{VaR is correctly specified and ES is correctly specified.}$$

To test this null hypothesis, we can test whether the estimated values satisfy a conditional moment restriction. That is, if

$$\mathbb{E}\left[Z_{t+1}|X_t = x\right] = 0, \qquad \forall x \in \mathcal{X}, \tag{4.3}$$

then the measure is correctly specified, where X_t includes the conditioning variable(s) and Z_{t+1} is a function of the risk measure(s). For example, in Horvath et al. (2019), they test whether the $VaR_{t+1,q}$ estimates satisfy

$$\mathbb{E}\left[\underbrace{\mathbb{1}\left(Y_{t+1} \leq VaR_{t+1,q}\right) - q}_{Z_{t+1}} \middle| X_t = x\right] = 0, \qquad \forall x \in \mathcal{X},$$

where Y_{t+1} is the $t+1$ period return and $\mathbb{1}\left(\cdot\right)$ is the indicator function which takes value 1 if the condition is satisfied and 0 otherwise. If one cannot reject the null hypothesis that the above conditional moment condition holds for all values of some conditioning variable(s), then one cannot reject the hypothesis that the estimates $VaR_{t+1,q}$ is correctly specified. One could also test for the correct specification of $ES_{t+1,q}$ by testing whether

$$\mathbb{E}\left[\underbrace{\frac{1}{q}Y_{t+1}\mathbb{1}\left(Y_{t+1} \leq VaR_{t+1,q}\right) - ES_{t+1,q}}_{Z_{t+1}} \middle| X_t = x\right] = 0, \qquad \forall x \in \mathcal{X}$$

[2] Chapter 3 in this thesis is a former version of that project.

holds. Note, however, that the above moment condition also includes $VaR_{t+1,q}$, therefore one needs to test an additional moment condition controlling for $VaR_{t+1,q}$ to avoid the perverse scenario where neither $VaR_{t+1,q}$ nor $ES_{t+1,q}$ are correctly specified but the above single conditional moment restriction holds.

Li and Liao (2019) rewrite the conditional moment testing problem (Equation 4.3) as testing whether $h(x)$ function is equal to 0 everywhere:

$$\mathbb{E}\left[Z_{t+1}|X_t = x\right] = h(x) = 0, \qquad \forall x \in \mathcal{X},$$
$$Z_{t+1} = h(X_t) + u_{t+1}, \quad \mathbb{E}\left[u_{t+1}|X_t = x\right] = 0.$$

They propose to approximate the unknown $h(x)$ function with a series polynomial, $\hat{h}(x)$, by applying the best linear predictor of Z_{t+1} given increasing number of approximating basis functions of X_t, $P(X_t) = (p_1(X_t), p_2(X_t), \ldots, p_{m_n}(X_t))$, where $m_n \to \infty$ as the sample size n increases. That is, $\hat{h}(x) = P(x)^\top \hat{b}_n$, where one can get $\hat{b}_n$ by ordinary least squares:

$$\hat{b}_n = \left(\sum_{t=1}^{n} P(X_t)P(X_t)^\top\right)^{-1} \left(\sum_{t=1}^{n} P(X_t)Z_{t+1}\right).$$

To make uniform inference on $h(x)$, Li and Liao (2019) prove the strong Gaussian approximation theory for series estimators with dependent data. They show that under some regularity condition the "pre-asymptotic" standard error of $\sqrt{n}\left(\hat{h}(x) - h(x)\right)$ is

$$\sigma_n(x) = \left[P(x)^\top \Sigma_n P(x)\right]^{1/2},$$

where

$$\Sigma_n = Q_n^{-1} A_n Q_n^{-1}$$

$$Q_n = n^{-1} \sum_{t=1}^{n} P(X_t) P(X_t)^\top$$

$$A_n = n^{-1/2} \sum_{t=1}^{n} P(X_t) P(X_t)^\top u_{t+1}^2$$

$$u_{t+1} = Z_{t+1} - h(X_t).$$

The resulting sup-t statistic then

$$\widehat{T}_n^{sup} = \sup_{x \in \mathcal{X}} \left| \frac{\sqrt{n} \left(\hat{h}_n(x) - h(x) \right)}{\hat{\sigma}_n(x)} \right| = \sup_{x \in \mathcal{X}} \left| \frac{\sqrt{n} \hat{h}_n(x)}{\hat{\sigma}_n(x)} \right| = \sup_{x \in \mathcal{X}} \left| \frac{\sqrt{n} P(x)^\top \hat{b}_n}{\hat{\sigma}_n(x)} \right|,$$

which can be strongly approximated with the Gaussian process

$$\widetilde{T}_n^{sup} = \sup_{x \in \mathcal{X}} \left| \frac{\sqrt{n} P(x)^\top S_n}{\hat{\sigma}_n(x)} \right| \quad S_n \sim \mathcal{N}(0, \Sigma_n),$$

that is,

$$\widehat{T}_n^{sup} - \widetilde{T}_n^{sup} = o_p \left(\log(n)^{-1/2} \right)$$

where n is the number of observations in the sample.[3] In this paper, we explore a Cramer-von Mises type of test. That is, instead of taking the supremum of the test statistics, we calculate the average with respect to the conditioning variable, which results in the average-t statistic:

$$\widehat{T}_n^{ave} = \int_{\mathcal{X}} \left| \frac{\sqrt{n} P(x)^\top \hat{b}_n}{\hat{\sigma}_n(x)} \right| dx.$$

Similarly to Theorem 2 in Li and Liao (2019), Theorem 5 states the asymptotic property of the average-t statistic.

[3] More details on theory with non-smooth functions is in Horvath et al. (2019).

Theorem 5. *Suppose that Assumption 1 and 2 in Li and Liao (2019) hold. Then under the null hypothesis there exists a sequence of S_n of m_n-dimensional standard normal random variables such that*

$$\widehat{T}_n^{ave} - \widetilde{T}_n^{ave} = o_p\left(\log(n)^{-1/2}\right),$$

where

$$\widetilde{T}_n^{ave} = \int_{\mathcal{X}} \left| \frac{\sqrt{n} P(x)^\top S_n}{\hat{\sigma}_n(x)} \right| dx.$$

Proof.

$$\left| \int_{\mathcal{X}} \left| \frac{\sqrt{n} P(x)^\top \hat{b}_n}{\hat{\sigma}_n(x)} \right| dx - \int_{\mathcal{X}} \left| \frac{\sqrt{n} P(x)^\top S_n}{\hat{\sigma}_n(x)} \right| dx \right| \leq$$

$$\int_{\mathcal{X}} \left| \left| \frac{\sqrt{n} P(x)^\top \hat{b}_n}{\hat{\sigma}_n(x)} \right| - \left| \frac{\sqrt{n} P(x)^\top S_n}{\hat{\sigma}_n(x)} \right| \right| dx \leq$$

$$\int_{\mathcal{X}} \left| \frac{\sqrt{n} P(x)^\top \hat{b}_n}{\hat{\sigma}_n(x)} - \frac{\sqrt{n} P(x)^\top S_n}{\hat{\sigma}_n(x)} \right| dx \leq$$

$$\mu(\mathcal{X}) \sup_{x \in \mathcal{X}} \left| \frac{\sqrt{n} P(x)^\top \hat{b}_n}{\hat{\sigma}_n(x)} - \frac{\sqrt{n} P(x)^\top S_n}{\hat{\sigma}_n(x)} \right| = o_p(\log(n)^{-1/2}),$$

where $\mu(\mathcal{X})$ is the measure of the set $\mathcal{X}$. The last equality follows from A.35. of Li and Liao (2019). □

One difference between the sup-t, $\widehat{T}_n^{sup}$, statistic applied in Li and Liao (2019) and Horvath et al. (2019), and average-t, $\widehat{T}_n^{ave}$, statistic is that the former one only takes into account the largest value among all t statistics while the latter gives equal weight for each $T(x)$ in the support space of the conditioning variable.

Li and Liao (2019) provides an algorithm to compute the critical values of the test based on the Gaussian approximation. However, in a Monte Carlo simulation

study Horvath et al. (2019) show this algorithm leads to bad size properties in finite samples for very high VaRs. To circumvent this issue, they propose the following i.i.d. bootstrap algorithm, which we also follow in this paper:

1. Resample $(\bar{Z}_{t+1}, \bar{X}_t)_{1\leq t\leq n}$ as i.i.d. sample with replacement from $(Z_{t+1}, X_t)_{1\leq t\leq n}$.

2. Compute the t-statistic $\left(\bar{T}_n\right)$ for the new sample, $(\bar{Z}_{t+1}, \bar{X}_t)_{1\leq t\leq n}$.

3. Repeat step 1 and 2 many times. Set the critical value, cv, at significant level α, as the $1-\alpha$ quantile of $\bar{T}_n$s in the Monte Carlo sample.

4. Reject the null hypotheses ($\mathbb{E}\left[Z_{t+1}|X_t = x\right] = h(x) = 0$) if $\widehat{T}_n > cv$.

Theorem 2 Horvath et al. (2019) state the validity of the above bootstrap method for sup-t statistic, which can be transformed into the average-t statistic similarly to Theorem 1.

The above description only allows us to conduct the infeasible inference since Z_{t+1} or X_t might not be directly observed. Horvath et al. (2019) provide sufficient conditions for the case when Z_{t+1} is not directly observed but depends on parameter estimates, that is, $Z_{t+1}(\theta)$. For example, if $VaR_{t+1,q}$ and $ES_{t+1,q}$ are model implied measures as in the simulation and empirical section, then ES moment condition: $Z_{t+1}(\theta) = \frac{1}{q}Y_{t+1}\mathbb{1}\left(Y_{t+1} \leq VaR_{t+1,q}(\theta)\right) - ES_{t+1,q}(\theta)$ naturally depends on the parameter estimates, which could potentially lead to parameter estimation error in the test. However, due to the slower nonparametric convergence rate in the test statistic, the test itself controls for this if the parameters are estimated with parametric, $\sqrt{n}$, rate, and one does not need to explicitly control for this as in e.g. Barendse et al. (2019), which is an advantage to use the above tests. Moreover, Horvath et al. (2019) also provide the theory for the case when the conditioning variable, X_t is generated

by a GARCH model (as in their and our simulations) or estimated as the realized volatility instead of being directly observed in the data.

In the derivation above, Z_{t+1} was assumed to be a scalar (one dimensional variable) since it only included one moment condition. However, as noted above we need to test at least two moment conditions jointly to make a decision about the correct specification of VaR and ES. To aggregate the results from two moment conditions, first we calculate the t-statistics for the two moment conditions separately $(Z_{t+1}^{VaR}, Z_{t+1}^{ES})$ to get T_n^{VaR}, T_n^{ES}, where Z_{t+1}^{ES} denotes the second moment condition which includes both VaR and ES. Then to get the joint t-statistic, we try out two different ways, which we call (sum) and (max) :

- $T_n^{sum} = T_n^{VaR} + T_n^{ES}$ (sum)
- $T_n^{max} = \max(T_n^{VaR}, T_n^{ES})$ (max).

That is, in the (sum) method we simply add up the two t-statistics while in the (max) method we take the maximum of the two t-statistics to get the joint t-statistic value. To compute the critical value of the test, one just needs to calculate the above joint t-statistic value in the bootstrap samples and then take the $1 - \alpha$ percentile of the computed joint t-statistics.

4.4 Simulations

In this section, we examine the finite sample properties of the average-t test in comparison with the sup-t test for GARCH-based models. Section 4.4.1 describe Monte Carlo simulation settings. Section 4.4.2 presents the moments, the conditioning variables (X_t) and the series polynomial $(P(x))$. Section 4.4.3 concludes with the results and it shows that the average-t test has better finite sample properties in this Monte Carlo simulation.

4.4.1 The simulation settings

We consider 4 data generating processes (DGP), the first corresponding to the null-hypothesis (size) while the next 3 (power) correspond to different misspecification of the first DGP (alternative hypotheses, denoted as A-. . . in the following).

To present the size of the test, we consider an AR(1) - GARCH(1, 1) process with the same parametrization as Barendse et al. (2019):

Size:

$$Y_{t+1} = \rho Y_t + \underbrace{\sigma_{t+1}\varepsilon_{t+1}}_{=v_{t+1}} \quad \varepsilon_{t+1} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$$

$$\sigma^2_{t+1} = \omega + \beta\sigma^2_t + \gamma v^2_t,$$

where $\{\rho, \omega, \beta, \gamma\} = \{0.05, 0.05, 0.85, 0.1\}$.

This model has 3 main components, such as conditional mean and variance and the the error distribution. In the following 3 DGPs we deliberately misspecify these 3 attributes to examine how powerful the joint test is to detect these misspecifications.

To misspecify the error distribution (A-Error), we consider an AR(1)-GARCH(1,1) model with Student-t distributed errors similarly to Barendse et al. (2019):

A-Error:

$$Y_{t+1} = \rho Y_t + \underbrace{\sigma_{t+1}\varepsilon_{t+1}}_{=v_{t+1}} \quad \varepsilon_{t+1} \overset{i.i.d.}{\sim} t(0,1,\nu)$$

$$\sigma^2_{t+1} = \omega + \beta\sigma^2_t + \gamma v^2_t,$$

where $\{\rho, \omega, \beta, \gamma, \nu\} = \{0.05, 0.05, 0.85, 0.1, 5\}$.

To misspecify the conditional mean of the process, we consider a TAR(1)-GARCH(1,1) process as in Escanciano and Olmo (2010):

A-Mean:

$$Y_{t+1} = a_t Y_t + \underbrace{\sigma_{t+1}\varepsilon_{t+1}}_{=v_{t+1}} \quad \varepsilon_{t+1} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$$

$$a_t = 0.7 \cdot \mathbb{1}\left(\varepsilon_t < -0.5\right) - 0.7 \cdot \mathbb{1}\left(\varepsilon_t > 0.5\right)$$

$$\sigma^2_{t+1} = \omega + \beta\sigma^2_t + \gamma v^2_t,$$

where $\{\omega, \beta, \gamma\} = \{0.05, 0.05, 0.85, 0.1\}$ and $\mathbb{1}(\cdot)$ is the indicator function, which takes 1 if the condition is satisfied and 0 otherwise.

Lastly, to examine how well the test can detect the misspecification of the conditional variance, we consider an AR(1) - EGARCH(1,1,1) process as in Bontemps (2019):

A-Variance:

$$Y_{t+1} = \rho Y_t + \underbrace{\sigma_{t+1}\varepsilon_{t+1}}_{=v_{t+1}} \quad \varepsilon_{t+1} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$$

$$\ln(\sigma^2_{t+1}) = \omega + \gamma\left(|\varepsilon_t| - \sqrt{2/\pi}\right) + \delta\varepsilon_t + \beta\ln(\sigma^2_t),$$

$\{\rho, \omega, \beta, \gamma, \delta\} = \{0.05, 0.0001, 0.9, 0.3, -0.8\}$.

Note in each of the these three alternative hypotheses, 2 out of the 3 main components are the same as under the null-hypotheses and only one of those are misspecified.

The model estimation and testing goes as follows: in the first, "modeler" step, we estimate an AR(1) process in the return series, Y_t, then we extract the innovation terms $\hat{v}_{t+1} = Y_{t+1} - \hat{\rho}Y_t$ to estimate a GARCH(1, 1) process with quasi maximum likelihood. Under the null hypothesis (or Size), this estimation procedure should result in the correctly specified q-VaR and q-ES:

$$\widehat{VaR}_{q,t+1} = \hat{\rho}Y_t + \Phi^{-1}(q)\sqrt{\hat{\omega} + \hat{\beta}\sigma^2_t + \hat{\gamma}\hat{v}^2_t}$$

$$\widehat{ES}_{q,t+1} = \hat{\rho}Y_t - \frac{\phi\left(\Phi^{-1}(q)\right)}{q}\sqrt{\hat{\omega} + \hat{\beta}\sigma^2_t + \hat{\gamma}\hat{v}^2_t},$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ are the cdf and pdf of the standard normal distribution and q is the confidence level of the VaR and ES. In the second, "testing" step, we test the correct specification of the estimated $\widehat{VaR}_{q,t+1}$ and $\widehat{ES}_{q,t+1}$ both with the sup-t and the proposed average-t test.

Similarly to the empirical study, the sample size in the base case scenario is 5000. We estimate the model parameters using rolling window of 2500 observations (in-sample, R) and to fasten the estimation process we reestimate these parameters after every 100 observations. The objective of the modeling step is to forecast 1-step ahead VaR and ES in the last 2500 observations. That is, n or number of out-of-sample, R, observations, is equal to 2500. To construct the bootstrap confidence values, we resample the data 1000 times. We consider 5%-level test and our main objective is to test $q = 0.025$, which corresponds to the Basel III regulations. To examine how sensitive the results are with respect to the confidence of VaR and ES, we also look at $q = 0.01$ and 0.05. In addition, we also set R and P to 500 and 2500 to investigate how the rejection rates change if the model estimation and/or the testing happens in shorter/longer time window.

4.4.2 The moments, (Z_{t+1}), the conditioning variables (X_t) and the series polynomial ($P(x)$)

In an ideal world, we would only need to use one moment condition to test the correct specification of ES; however, it is only jointly elicitable with VaR, which means there exists no moment functions for ES which is independent of VaR (Fissler and Ziegel, 2016). Therefore, we need to conduct a joint specification test, which requires 2 moment conditions: one controlling for the VaR and one for the ES.

Following Escanciano and Olmo (2010), we use the moment conditions derived from the check loss function (quantile loss function) to test for $VaR_{q,t+1}$:

$$\mathbb{E}\left[\mathbb{1}\left(Y_{t+1} \leq VaR_{t+1,q}\right) - q | X_t\right] = 0,$$

and we use the following moment condition as an "ES moment":

$$\mathbb{E}\left[\frac{1}{q}Y_{t+1}\mathbb{1}\left(Y_{t+1} \leq VaR_{t+1,q}\right) - ES_{t+1,q}|X_t\right] = 0.$$

It may seem natural to use conditioning variables controlling for the location and for the scale of the estimated model in a location-scale model. Therefore, we use the lagged returns Y_{t-1} and the lagged volatility σ_{t-1}, which is generated by the DGP, as conditioning variables in the simulation.

Similarly to Horvath et al. (2019), to mitigate the effect of possible multicollinearity in the series regression, we use Legendre polynomials of the normalized lagged returns and volatility. To construct the $P(x)$, we proceed in the following steps:

1. Normalize Y_{t-1} and σ_{t-1} to [-1, 1] interval, where Legendre polynomials are orthogonal:
   (a) Rank the observations from 1 to n ($ranked\,X$), where n is the number of observations and X is Y_{t-1} and σ_{t-1} separately
   (b) Normalize them to [-1, 1] as $normalized\,X = 2\frac{rankedX - \min(rankedX)}{\max(rankedX) - \min(rankedX)} - 1$
2. Apply the m_n order Legendre polynomial on the $normalized\,X$

To examine how the number of the terms in the series polynomial might have an effect on the testing procedure, we apply Legendre polynomial with degree of 1, 2, 3 and 4 in the following way:

- $P_1(X_t) = \left(1, \tilde{Y}_t^1, \tilde{\sigma}_t^1, \tilde{Y}_t^1\tilde{\sigma}_t^1\right)$
- $P_2(X_t) = \left(P_1(X_t), \tilde{Y}_t^2, \tilde{\sigma}_t^2\right)$
- $P_3(X_t) = \left(P_2(X_t), \tilde{Y}_t^3, \tilde{\sigma}_t^3, \tilde{Y}_t^2\tilde{\sigma}_t^1, \tilde{Y}_t^1\tilde{\sigma}_t^2\right)$

- $P_4(X_t) = \Big(P_3(X_t), \tilde{Y}_t^4, \tilde{\sigma}_t^4, \tilde{Y}_t^2\tilde{\sigma}_t^2, \tilde{Y}_t^1\tilde{\sigma}_t^3, \tilde{Y}_t^3\tilde{\sigma}_t^1\Big)$,

where $\tilde{Y}_t^{m_n}$ and $\tilde{\sigma}_t^{m_n}$ denotes the m_n^{th} order Legendre polynomial. Note $P_1(X_t)$ includes 4 terms, $P_2(X_t)$ 6 terms, $P_3(X_t)$ 10 terms and $P_4(X_t)$ includes 15 terms.

4.4.3 Simulation Results

Table 4.1 reports the rejection rates for the sup-t and average-t test for 2.5%-VaR and -ES. The upper panel presents the results for the (sum) joint test method while the lower panel for the (max) joint test method. The first row in each block reports the rejection rates under the null hypothesis (DGP and estimated models are both an AR(1)-GARCH(1,1)). The next three rows report the rejection rates under the alternative hypotheses. A-Error, when the distribution of the error term is misspecified; A-Mean when the conditional mean is misspecified and A-Volatility corresponds to the case when the conditional volatility is misspecified.

The sup-t statistic tends to overreject under the null-hypothesis and the rejection rate is increasing in the number of polynomial terms. This might be explained by computational difficulties as the largest value of the test statistic carry all the weight in the sup-t statistic. On the other hand, the average-t statistic is a bit liberal as it does not hit the 5% rejection rates under the null-hypothesis; however, the number of polynomial terms does not have substantial effect on the test performance.

Both tests are good at detecting misspefications. However, as we can see it again, the number of polynomial terms has barely any effect on the performance of average-t test and it also detects the misspecification with higher accuracy than the sup-t test.

In conclusion both the size and power results would suggest that the average-t test is better than the sup-t test for testing the correct specification of 2.5%-VaR and -ES.

Table 4.1: Simulation results: Rejection rates at 5% level, q = 0.025, $R = 2500$, $P = 2500$

	(sum) method							
	sup-t				average-t			
	$P_1(x)$	$P_2(x)$	$P_3(x)$	$P_4(x)$	$P_1(x)$	$P_2(x)$	$P_3(x)$	$P_4(x)$
Size	0.054	0.065	0.094	0.138	0.039	0.035	0.040	0.041
A-Error	0.996	0.996	0.972	0.784	0.998	0.998	0.996	0.996
A-Mean	1.000	1.000	0.945	0.407	1.000	1.000	0.997	0.960
A-Volatility	0.988	0.952	0.804	0.673	0.997	0.995	0.993	0.993

	(max) method							
	sup-t				average-t			
	$P_1(x)$	$P_2(x)$	$P_3(x)$	$P_4(x)$	$P_1(x)$	$P_2(x)$	$P_3(x)$	$P_4(x)$
Size	0.037	0.046	0.065	0.083	0.034	0.024	0.030	0.030
A-Error	0.989	0.981	0.892	0.360	0.997	0.996	0.992	0.989
A-Mean	1.000	0.998	0.600	0.077	1.000	1.000	0.994	0.936
A-Volatility	0.891	0.764	0.548	0.397	0.994	0.992	0.989	0.983

A natural question that could arise is how much power can be attributed to the joint test (VaR and ES together) instead of just testing the VaR moment condition by itself. Table 4.2 might answer this question. As we can see from this table, the separate test has close to nominal 5% rejection rates under the null hypothesis, where we can observe again that the sup-t test rejects more frequently as the number of terms in the series polynomial increases and the average-t test has a rejection rate smaller than the nominal rate. The rejection rate is close to 100% under the alternative hypotheses for both of these tests.

From these results, we could conclude that a modeler first needs to make sure that the VaR model is correctly specified before she would proceed to work on the ES model.

Table 4.2: Simulation results for VaR: Rejection rates at 5% level, q = 0.025, $R = 2500$, $P = 2500$

	sup-t				average-t			
	$P_1(x)$	$P_2(x)$	$P_3(x)$	$P_4(x)$	$P_1(x)$	$P_2(x)$	$P_3(x)$	$P_4(x)$
Size	0.049	0.044	0.059	0.077	0.038	0.037	0.038	0.035
A-Error	1.000	1.000	0.999	0.964	1.000	1.000	1.000	1.000
A-Mean	1.000	1.000	0.987	0.710	1.000	1.000	1.000	0.981
A-Volatility	1.000	1.000	0.999	0.990	1.000	1.000	1.000	1.000

To examine how sensitive the above results are with respect to the confidence of the VaR and ES, Table 4.3 and 4.4 report the rejection rates from the joint tests for the 1%- and 5%-VaR and ES.

As we can see from Table 4.3, the size properties of the tests remain similar to the 2.5% case; however, the sup-t test performs poorly in detecting misspecification so far out in the tail. Contrary to this and the findings of Nolde and Ziegel (2017), the performance of the average-t test does not plummet that much when one examines the VaR and ES at such an extreme point. These findings underpin the previous conclusion: the average-t test might have better finite sample properties than the sup-t test.

Table 4.3: Simulation results: Rejection rates at 5% level, q = 0.01, $R = 2500$, $P = 2500$

	(sum) method							
	sup-t				average-t			
	$P_1(x)$	$P_2(x)$	$P_3(x)$	$P_4(x)$	$P_1(x)$	$P_2(x)$	$P_3(x)$	$P_4(x)$
Size	0.038	0.071	0.090	0.056	0.031	0.027	0.040	0.037
A-Error	0.990	0.974	0.520	0.047	0.995	0.992	0.984	0.977
A-Mean	0.957	0.774	0.130	0.046	0.997	0.970	0.714	0.468
A-Volatility	0.877	0.720	0.600	0.444	0.983	0.957	0.927	0.917

	(max) method							
	sup-t				average-t			
	$P_1(x)$	$P_2(x)$	$P_3(x)$	$P_4(x)$	$P_1(x)$	$P_2(x)$	$P_3(x)$	$P_4(x)$
Size	0.021	0.048	0.058	0.031	0.020	0.020	0.025	0.018
A-Error	0.978	0.910	0.118	0.006	0.993	0.987	0.970	0.958
A-Mean	0.776	0.360	0.041	0.018	0.998	0.980	0.795	0.574
A-Volatility	0.565	0.435	0.439	0.314	0.942	0.919	0.879	0.841

The results in Table 4.4 are align with the findings of Barendse et al. (2019) and Nolde and Ziegel (2017): it is easier to estimate and test VaR and ES farther away from the extreme values. When q = 0.05, the average-t test rejects the null hypothesis around 5% of the times as it is expected and both tests have stronger power properties than at lower q levels.

Table 4.4: Simulation results: Rejection rates at 5% level, q = 0.05, $R = 2500$, $P = 2500$

	(sum) method							
	sup-t				average-t			
	$P_1(x)$	$P_2(x)$	$P_3(x)$	$P_4(x)$	$P_1(x)$	$P_2(x)$	$P_3(x)$	$P_4(x)$
Size	0.048	0.057	0.074	0.105	0.048	0.037	0.049	0.052
A-Error	1.000	1.000	0.991	0.980	1.000	1.000	0.999	0.999
A-Mean	1.000	1.000	1.000	0.995	1.000	1.000	1.000	1.000
A-Volatility	0.996	0.995	0.945	0.852	0.999	0.999	0.997	0.996

	(max) method							
	sup-t				average-t			
	$P_1(x)$	$P_2(x)$	$P_3(x)$	$P_4(x)$	$P_1(x)$	$P_2(x)$	$P_3(x)$	$P_4(x)$
Size	0.031	0.034	0.051	0.074	0.045	0.036	0.047	0.043
A-Error	0.996	0.995	0.980	0.934	1.000	0.997	0.996	0.995
A-Mean	1.000	1.000	1.000	0.810	1.000	1.000	1.000	1.000
A-Volatility	0.975	0.922	0.722	0.523	0.999	0.997	0.996	0.995

Table 4.5 and 4.6 tabulate the rejection rates for the 1%- and 5%-VaR tests. As before, the sup-t test performs worse at the more extreme VaR values while the performance of the average-t test barely deteriorates comparing to the 2.5%. Both tests perform well for the 5%-VaR.

Table 4.5: Simulation results for VaR: Rejection rates at 5% level, q = 0.01, $R = 2500$, $P = 2500$

	sup-t				average-t			
	$P_1(x)$	$P_2(x)$	$P_3(x)$	$P_4(x)$	$P_1(x)$	$P_2(x)$	$P_3(x)$	$P_4(x)$
Size	0.021	0.041	0.031	0.003	0.030	0.028	0.035	0.029
A-Error	0.999	0.995	0.823	0.250	1.000	0.999	0.997	0.994
A-Mean	0.945	0.873	0.290	0.056	0.999	0.982	0.841	0.607
A-Volatility	1.000	0.972	0.852	0.696	1.000	1.000	1.000	1.000

Table 4.6: Simulation results for VaR: Rejection rates at 5% level, q = 0.05, $R = 2500$, $P = 2500$

	sup-t				average-t			
	$P_1(x)$	$P_2(x)$	$P_3(x)$	$P_4(x)$	$P_1(x)$	$P_2(x)$	$P_3(x)$	$P_4(x)$
Size	0.049	0.050	0.052	0.074	0.053	0.040	0.047	0.047
A-Error	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
A-Mean	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
A-Volatility	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

To examine how sensitive the base case results are with respect to the number of in-sample and out-of-sample periods, Table 4.7, 4.8 and 4.9 report the rejection rates for the joint tests when q = 0.025 and $(R, P) = \{(500, 500), (2500, 500), (500, 2500)\}$, respectively.

Table 4.7, 4.8 report the results when the number of out-of-sample observations, where the specification tests are conducted, is only 500. Due to the lower number of observations, there is a bigger drop in the rejection rates under the null and alternative hypotheses. The proposed average-t test seems to suffer less from the decreased number of observations. The number of in-sample periods, where the models are estimated, has barely any effect on these numbers because the underlying DGPs are strongly stationary.

Table 4.7: Simulation results: Rejection rates at 5% level, q = 0.025, $R = 500$, $P = 500$

	(sum) method							
	sup-t				average-t			
	$P_1(x)$	$P_2(x)$	$P_3(x)$	$P_4(x)$	$P_1(x)$	$P_2(x)$	$P_3(x)$	$P_4(x)$
Size	0.032	0.049	0.026	0.001	0.030	0.022	0.022	0.022
A-Error	0.710	0.532	0.106	0.038	0.906	0.771	0.559	0.430
A-Mean	0.255	0.134	0.014	0.004	0.599	0.445	0.231	0.127
A-Volatility	0.373	0.370	0.311	0.146	0.600	0.579	0.519	0.432

	(max) method							
	sup-t				average-t			
	$P_1(x)$	$P_2(x)$	$P_3(x)$	$P_4(x)$	$P_1(x)$	$P_2(x)$	$P_3(x)$	$P_4(x)$
Size	0.022	0.018	0.008	0.000	0.017	0.013	0.013	0.010
A-Error	0.523	0.338	0.042	0.017	0.848	0.692	0.451	0.345
A-Mean	0.107	0.047	0.007	0.002	0.504	0.377	0.174	0.089
A-Volatility	0.226	0.247	0.227	0.141	0.460	0.436	0.381	0.293

Table 4.8: Simulation results: Rejection rates at 5% level, q = 0.025, $R = 2500$, $P = 500$

	(sum) method							
	sup-t				average-t			
	$P_1(x)$	$P_2(x)$	$P_3(x)$	$P_4(x)$	$P_1(x)$	$P_2(x)$	$P_3(x)$	$P_4(x)$
Size	0.030	0.050	0.021	0.002	0.026	0.015	0.020	0.023
A-Error	0.750	0.540	0.082	0.021	0.921	0.800	0.575	0.464
A-Mean	0.251	0.136	0.016	0.003	0.616	0.458	0.213	0.132
A-Volatility	0.349	0.370	0.303	0.145	0.586	0.575	0.508	0.428

	(max) method							
	sup-t				average-t			
	$P_1(x)$	$P_2(x)$	$P_3(x)$	$P_4(x)$	$P_1(x)$	$P_2(x)$	$P_3(x)$	$P_4(x)$
Size	0.017	0.022	0.011	0.001	0.015	0.009	0.012	0.015
A-Error	0.558	0.374	0.043	0.012	0.874	0.696	0.476	0.389
A-Mean	0.112	0.056	0.009	0.001	0.534	0.379	0.150	0.095
A-Volatility	0.204	0.240	0.235	0.127	0.436	0.413	0.363	0.293

Table 4.9 reports the results when the number of out-of-sample periods is the same as in the base case but the modeller has only 500 observations to estimate the risk measures. As we can see from this table, the rejection rates are very close to the base case scenario (Table 4.1). This can be explained by the slower convergence rate of the nonparametric test and the strong stationary models in the simulation. That is, estimating the models in 500 or 2500 periods do not have a significant effect on the estimates; however, both the sup-t and average-t test requires ample of observations to perform adequately.

Table 4.9: Simulation results: Rejection rates at 5% level, q = 0.025, $R = 500$, $P = 2500$

	(sum) method							
	sup-t				average-t			
	$P_1(x)$	$P_2(x)$	$P_3(x)$	$P_4(x)$	$P_1(x)$	$P_2(x)$	$P_3(x)$	$P_4(x)$
Size	0.031	0.046	0.074	0.112	0.032	0.035	0.039	0.035
A-Error	0.997	0.997	0.982	0.777	1.000	1.000	0.997	0.997
A-Mean	1.000	1.000	0.965	0.395	1.000	1.000	0.999	0.957
A-Volatility	0.993	0.954	0.817	0.667	0.999	0.997	0.995	0.995

	(max) method							
	sup-t				average-t			
	$P_1(x)$	$P_2(x)$	$P_3(x)$	$P_4(x)$	$P_1(x)$	$P_2(x)$	$P_3(x)$	$P_4(x)$
Size	0.019	0.023	0.048	0.078	0.026	0.030	0.029	0.026
A-Error	0.996	0.992	0.891	0.303	0.998	0.996	0.995	0.993
A-Mean	1.000	0.999	0.665	0.052	1.000	1.000	0.995	0.930
A-Volatility	0.897	0.762	0.534	0.436	0.995	0.989	0.983	0.979

4.5 Empirics

In this section, we apply the average-t test in S&P 500 data. More specifically, we use the S&P 500 index, Microsoft Corporation (MSFT), Bank of America Corporation (BAC), Exxon Mobil Corporation (XOM) and UnitedHealth Group Inc. (UNH). We fit 7 models from simple rolling window estimates through Generalized Autoregressive Score model and we test which model implied VaR and ES are correctly specified. To conduct these tests, we download the S&P 500 data from yahoo.finance from 1/2/1998 until 12/29/2017 (5032 trading days). We forecast the 1-day ahead 2.5%-VaR and 2.5%-ES[4] applying rolling window estimate with 2500 in-sample days and reestimate the model parameters every 100 days. That is, first we fit the models in the first 2500 days (from 1/2/1998 until 12/11/2007), then we forecast the VaR and

[4] The appendix includes the results for the confidence level of 1% and 5%. Table 4.12, 4.13 present the result for the 1%-VaR and -ES while 4.14, 4.15 for the 5% measures.

ES 1-day ahead. We use the same parameter estimates in the next 99 days to forecast the 1-day ahead VaR and ES. After 100 days, we reestimate the model parameters and forecast the 1-day ahead VaR and ES for the next 100 days etc. This procedure results in 2532 1-day ahead VaR and ES forecasts for the period of 12/12/2007 until 12/29/2017.

Similarly to the Simulation section (Section 4.4), we use 2 conditioning variables: 1 controlling for the location of the models (previous day return, Y_{t-1}) and 1 controlling for the scale of the model (previous day realized volatility, RV_{t-1}). To estimate the realized volatility, we use TAQ high frequency data.

To estimate the models first we fit an ARMA(p, q) model for the conditional mean where we choose p and q such that they minimize the BIC. To model the conditional volatility we fit several models such as:

- GARCH with normally distributed errors (GARCH-N)
- GARCH with skewed t distributed errors (GARCH-skew t)
- GARCH with empirical distribution function (GARCH-EDF)
- GJR-GARCH with normally distributed errors (GJR-GARCH-N)
- GJR-GARCH with skewed t distributed errors (GJR-GARCH-skew t)
- One factor GAS ("Generalized Autoregressive Score") model (GAS-1F).

In addition to the above location-scale models, we also use a rolling window estimates of the 1-day ahead VaR and ES using the previous 500 days as an estimation period. All of these models are estimated with the code from Patton et al. (2019).[5]

In the results section, we report the bootstrap p-values. That is, we compute the $\widehat{T}_n^{ave}$ statistic for the whole sample and for the 1000 bootstrap samples ($\bar{T}_n$), then

[5] We needed to adjust the code for the GJR-GARCH models as Patton et al. (2019) do not estimate those models.

we count how many of the bootstrap $\bar{T}_n$s are larger than $\widehat{T}_n$ to get the p-value. For the sake of brevity, we only report the results for $P_3(x)$. The conclusion remain the same under the other polynomials.

4.5.1 Empirical Results

Table 4.10 reports the p-values for 2.5%-VaR and -ES. P-values below 0.05 indicate that the null hypothesis about the correct specification of both risk measures jointly can be rejected. Therefore, neither of these 7 models can correctly model the 2.5%-VaR and -ES for the S&P500 index, MSFT and BAC. However, there is a borderline 0.05 value for GJR-GARCH model with normally distributed errors for UnitedHealth Group and the GJR-GARCH model with skewed t distributed errors for Exxon Mobil can correctly specify the risk measures.

Table 4.10: Joint test, p-values: q = 0.025

(sum) method					
	S&P 500	MSFT	BAC	XOM	UNH
RW-500	0.000	0.000	0.000	0.000	0.000
GARCH-N	0.013	0.000	0.004	0.000	0.000
GARCH-skew t	0.000	0.002	0.018	0.009	0.000
GARCH-EDF	0.000	0.000	0.000	0.009	0.000
GJR-GARCH-N	0.000	0.000	0.006	0.009	0.050
GJR-GARCH-skew t	0.006	0.000	0.000	0.141	0.020
GAS-1F	0.022	0.001	0.000	0.000	0.000
(max) method					
	S&P 500	MSFT	BAC	XOM	UNH
RW-500	0.000	0.000	0.000	0.000	0.000
GARCH-N	0.028	0.000	0.010	0.000	0.000
GARCH-skew t	0.004	0.001	0.020	0.005	0.000
GARCH-EDF	0.000	0.000	0.002	0.005	0.002
GJR-GARCH-N	0.001	0.000	0.007	0.009	0.050
GJR-GARCH-skew t	0.004	0.000	0.000	0.141	0.020
GAS-1F	0.056	0.001	0.001	0.001	0.000

Similarly to the simulations, the conclusion regarding the joint hypothesis of correct specification of both risk measures might be led by the correct specification of VaR as we could conclude from Table 4.11. That is, if a model correctly estimates the VaR, it can correctly estimate the VaR and ES jointly, too.

Table 4.11: VaR test, p-values: q = 0.025

	S&P 500	MSFT	BAC	XOM	UNH
RW-500	0.000	0.000	0.000	0.000	0.000
GARCH-N	0.002	0.000	0.000	0.000	0.000
GARCH-skew t	0.000	0.000	0.020	0.005	0.000
GARCH-EDF	0.000	0.000	0.000	0.005	0.002
GJR-GARCH-N	0.000	0.000	0.000	0.009	0.050
GJR-GARCH-skew t	0.001	0.000	0.000	0.141	0.020
GAS-1F	0.000	0.001	0.000	0.021	0.000

4.6 Conclusion

In the beginning of 2010's, The Basel Committee proposed to replace the Value-at-Risk as the main market risk measure with Expected Shortfall. In this paper, we implement a new nonparametric test for testing the correct specification of VaR and ES jointly, called average-t test.

The theory behind the proposed average-t test heavily relies on the theory of sup-t test (Li and Liao, 2019). We have shown in Monte Carlo studies that the average-t test had better finite sample properties than the sup-t test, especially in more extreme events, closer to the tail of the return distribution. In addition to the good finite sample properties, this test does not require the estimation of the return distribution which has been statistical challenge for several papers in VaR/ES testing literature. Also it implicitly controls for the estimation error of the modelling step.

In an empirical exercise, we applied the average-t test in S&P500 data, including the index and 4 stocks from different industries to test the correct specification of

these two risk measures jointly. We have found that neither the basic location-scale models (GARCH, GJR-GARCH) nor the more complicated 1-Factor GAS model is able to correctly model these risk measures for the majority of these assets.

4.7 Appendix

4.7.1 Tables

Table 4.12: Joint test, p-values: q = 0.01

(sum) method					
	S&P 500	MSFT	BAC	XOM	UNH
RW-500	0.008	0.008	0.003	0.047	0.038
GARCH-N	0.108	0.080	0.205	0.194	0.079
GARCH-skew t	0.104	0.003	0.101	0.081	0.050
GARCH-EDF	0.048	0.006	0.110	0.081	0.057
GJR-GARCH-N	0.194	0.014	0.184	0.091	0.169
GJR-GARCH-skew t	0.000	0.003	0.000	0.141	0.076
GAS-1F	0.086	0.004	0.482	0.010	0.033

(max) method					
	S&P 500	MSFT	BAC	XOM	UNH
RW-500	0.019	0.020	0.019	0.091	0.034
GARCH-N	0.202	0.046	0.262	0.160	0.064
GARCH-skew t	0.097	0.005	0.117	0.077	0.048
GARCH-EDF	0.064	0.006	0.135	0.078	0.080
GJR-GARCH-N	0.202	0.014	0.180	0.070	0.161
GJR-GARCH-skew t	0.000	0.005	0.000	0.141	0.056
GAS-1F	0.142	0.005	0.570	0.012	0.037

Table 4.13: VaR test, p-values: q = 0.01

	S&P 500	MSFT	BAC	XOM	UNH
RW-500	0.002	0.005	0.000	0.003	0.005
GARCH-N	0.032	0.024	0.072	0.032	0.021
GARCH-skew t	0.043	0.001	0.066	0.037	0.027
GARCH-EDF	0.020	0.001	0.023	0.037	0.026
GJR-GARCH-N	0.114	0.006	0.040	0.031	0.103
GJR-GARCH-skew t	0.000	0.002	0.000	0.141	0.047
GAS-1F	0.038	0.001	0.436	0.010	0.044

Table 4.14: Joint test, p-values: q = 0.05

(sum) method					
	S&P 500	MSFT	BAC	XOM	UNH
RW-500	0.000	0.000	0.000	0.000	0.000
GARCH-N	0.000	0.000	0.002	0.000	0.000
GARCH-skew t	0.000	0.000	0.006	0.000	0.000
GARCH-EDF	0.000	0.000	0.015	0.000	0.000
GJR-GARCH-N	0.000	0.000	0.001	0.000	0.000
GJR-GARCH-skew t	0.000	0.000	0.009	0.000	0.000
GAS-1F	0.001	0.001	0.129	0.001	0.001

(max) method					
	S&P 500	MSFT	BAC	XOM	UNH
RW-500	0.000	0.000	0.000	0.000	0.000
GARCH-N	0.002	0.000	0.000	0.000	0.000
GARCH-skew t	0.001	0.000	0.006	0.000	0.000
GARCH-EDF	0.000	0.000	0.024	0.000	0.000
GJR-GARCH-N	0.001	0.000	0.000	0.000	0.000
GJR-GARCH-skew t	0.000	0.000	0.015	0.000	0.000
GAS-1F	0.006	0.002	0.177	0.001	0.001

Table 4.15: VaR test, p-values: q = 0.05

	S&P 500	MSFT	BAC	XOM	UNH
RW-500	0.000	0.000	0.000	0.000	0.000
GARCH-N	0.000	0.000	0.000	0.000	0.000
GARCH-skew t	0.000	0.000	0.000	0.000	0.000
GARCH-EDF	0.000	0.000	0.000	0.000	0.000
GJR-GARCH-N	0.000	0.000	0.000	0.000	0.000
GJR-GARCH-skew t	0.000	0.000	0.000	0.000	0.000
GAS-1F	0.000	0.001	0.032	0.005	0.001

5

Conclusions

Value-at-Risk (VaR) has been the main market risk measure since its introduction to the Basel I Accord in 1996. It can be interpreted as a low quantile (1%-5%) of the return of a portfolio or as a high quantile (95%-99%) of the losses. Since it is only a point on the distribution, neither does it capture the tail risk nor is it a coherent risk measure (Artzner et al., 1999). Acknowledging these deficiencies, The Basel Committee on Banking Supervision proposed to replace the VaR with the Expected Shortfall (ES) (Basel Committee, 2013).

ES is the mean of the return conditional the return being lower of its VaR value (or mean of the loss conditional the loss being above of its VaR value). Contrary to the VaR, it does not only account for tail risk but it is also a coherent risk measure (Artzner et al., 1999). However, the main theoretical shortcoming of this measure: it can only be estimated jointly with VaR (Fissler and Ziegel, 2016).

This thesis consists of 3 chapters relating to the estimation and evaluation of these risk measures. In the first chapter, we implemented a 2-step robust estimation method to estimate the ES. We found in our simulation study that it performed as good as a 1-step joint estimation technique (VaR and ES estimated jointly) and it

outperformed the joint estimation method in inference. In S&P500 data, we found that VIX and realized volatility might be a good predictor for one-day, -week and -month ahead ES.

In the second chapter, we developed a novel nonparametric specification test for VaR models based on the sup-t test idea of Li and Liao (2019). We implemented a novel i.d.d. bootstrap technique to compute the critical values of the test, which resulted in better size properties in our finite sample simulation study than the Gaussian approximation method proposed in Li and Liao (2019). We found that CoVaR model in Tobias and Brunnermeier (2016) are correctly specified.

In the last chapter, we implemented a joint nonparametric specification test for testing VaR and ES jointly. This chapter introduced an average-t test, whose theoretical foundations are based on the paper of Li and Liao (2019) and Horvath et al. (2019). We found in our simulation study that the average-t test outperformed the sup-t test in lower quantiles (deeper in the tail). We applied this method in S&P500 data and found that simple location-scale models might not be able to correctly model the VaR/ES at conventional confidence levels.

Bibliography

Ackerberg, D., Chen, X., Hahn, J., and Liao, Z. (2014), "Asymptotic efficiency of semiparametric two-step GMM," *Review of Economic Studies*, 81, 919–943.

Aït-Sahalia, Y. and Lo, A. W. (1998), "Nonparametric estimation of state-price densities implicit in financial asset prices," *The Journal of Finance*, 53, 499–547.

Aït-Sahalia, Y. and Lo, A. W. (2000), "Nonparametric risk management and implied risk aversion," *Journal of econometrics*, 94, 9–51.

Artzner, P., Delbaen, F., Eber, J.-M., and Heath, D. (1999), "Coherent measures of risk," *Mathematical finance*, 9, 203–228.

Barendse, S., Kole, E., and van Dijk, D. J. (2019), "Backtesting Value-at-Risk and Expected Shortfall in the Presence of Estimation Error," *Working paper*.

Barone Adesi, G. (2016), "VaR and CVaR implied in option prices," *Journal of Risk and Financial Management*, 9, 2.

Basel Committee (2013), "Fundamental review of the trading book: A revised market risk framework," *Consultative Document, October*.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014), "Inference on treatment effects after selection among high-dimensional controls," *The Review of Economic Studies*, 81, 608–650.

Bontemps, C. (2019), "Moment-based tests under parameter uncertainty," *Review of Economics and Statistics*, 101, 146–159.

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

Chen, X. (2007), "Large sample sieve estimation of semi-nonparametric models," *Handbook of econometrics*, 6, 5549–5632.

Chen, X. and Liao, Z. (2015), "Sieve semiparametric two-step GMM under weak dependence," *Journal of Econometrics*, 189, 163–186.

Chen, X. and Pouzo, D. (2012), "Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals," *Econometrica*, 80, 277–321.

Chen, X. and Shen, X. (1998), "Sieve extremum estimates for weakly dependent data," *Econometrica*, pp. 289–314.

Chen, X., Linton, O., and Van Keilegom, I. (2003), "Estimation of semiparametric models when the criterion function is not smooth," *Econometrica*, 71, 1591–1608.

Chernozhukov, V., Hansen, C., and Spindler, M. (2015), "Valid post-selection and post-regularization inference: An elementary, general approach," .

Chernozhukov, V., Escanciano, J. C., Ichimura, H., and Newey, W. K. (2016), "Locally Robust Semiparametric Estimation," *arXiv preprint arXiv:1608.00033*.

Danielsson, J., Jorgensen, B. N., Mandira, S., Samorodnitsky, G., and De Vries, C. G. (2005), "Subadditivity re–examined: the case for value-at-risk," .

Diebold, F. X. and Li, C. (2006), "Forecasting the term structure of government bond yields," *Journal of econometrics*, 130, 337–364.

Diebold, F. X. and Mariano, R. S. (2002), "Comparing predictive accuracy," *Journal of Business & economic statistics*, 20, 134–144.

Dimitriadis, T. and Bayer, S. (2017), "A joint quantile and Expected Shortfall regression framework," *arXiv preprint arXiv:1704.02213*.

Du, Z. and Escanciano, J. C. (2016), "Backtesting expected shortfall: accounting for tail risk," *Management Science*, 63, 940–958.

Engle, R. F. and Manganelli, S. (2004), "CAViaR: Conditional autoregressive value at risk by regression quantiles," *Journal of Business & Economic Statistics*, 22, 367–381.

Escanciano, J. C. and Olmo, J. (2010), "Backtesting parametric value-at-risk with estimation risk," *Journal of Business & Economic Statistics*, 28, 36–51.

Escanciano, J. C. and Olmo, J. (2011), "Robust backtesting tests for value-at-risk models," *Journal of Financial Econometrics*, 9, 132–161.

Fissler, T. and Ziegel, J. F. (2016), "Higher order elicitability and Osband's principle," *The Annals of Statistics*, 44, 1680–1707.

Hall, A. R. (2005), *Generalized method of moments*, Oxford University Press.

Horvath, P., Li, J., Liao, Z., and Patton, A. J. (2019), "Value-at-Risk Testing via Nonparametric Regressions," *Working paper*.

Jurado, K., Ludvigson, S. C., and Ng, S. (2015), "Measuring uncertainty," *American Economic Review*, 105, 1177–1216.

Kerkhof, J. and Melenberg, B. (2004), "Backtesting for risk-based regulatory capital," *Journal of Banking & Finance*, 28, 1845–1865.

Koenker, R. and Bassett Jr, G. (1978), "Regression quantiles," *Econometrica: journal of the Econometric Society*, pp. 33–50.

Koenker, R. and Hallock, K. F. (2001), "Quantile regression," *Journal of economic perspectives*, 15, 143–156.

Kuester, K., Mittnik, S., and Paolella, M. S. (2006), "Value-at-risk prediction: A comparison of alternative strategies," *Journal of Financial Econometrics*, 4, 53–89.

Li, J. and Liao, Z. (2019), "Uniform Nonparametric Inference for Time Series," *Forthcoming in Journal of Econometrics*.

McNeil, A. J. and Frey, R. (2000), "Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach," *Journal of empirical finance*, 7, 271–300.

Mitra, S. (2015), "The relationship between conditional value at risk and option prices with a closed-form solution," *The European Journal of Finance*, 21, 400–425.

Nelson, C. R. and Siegel, A. F. (1987), "Parsimonious Modeling of Yield Curves," *The Journal of Business*, 60, 473–489.

Newey, W. K. (1994), "The asymptotic variance of semiparametric estimators," *Econometrica: Journal of the Econometric Society*, pp. 1349–1382.

Newey, W. K. and West, K. D. (1986), "A simple, positive semi-definite, heteroskedasticity and autocorrelationconsistent covariance matrix," .

Nolde, N. and Ziegel, J. F. (2017), "Elicitability and backtesting: Perspectives for banking regulation," *The annals of applied statistics*, 11, 1833–1874.

Patton, A. J., Ziegel, J. F., and Chen, R. (2019), "Dynamic semiparametric models for expected shortfall (and value-at-risk)," *Journal of Econometrics*, 211, 388–413.

Tobias, A. and Brunnermeier, M. K. (2016), "CoVaR," *The American Economic Review*, 106, 1705.

Tropp, J. A. (2012), "User-friendly tail bounds for sums of random matrices," *Foundations of computational mathematics*, 12, 389–434.

Wang, C.-S. and Zhao, Z. (2016), "Conditional Value-at-Risk: Semiparametric estimation and inference," *Journal of Econometrics*, 195, 86–103.

Biography

Peter Horvath attended Velinszky Laszlo Elementary School, Hetvezer Elementary School and Toparti High School and Art Vocational School in Szekesfehervar where he grew up. He earned a Bachelor of Science degree in Economic Analysis from Corvinus University of Budapest in 2013 and joint Master of Science degree in Economics from Institute for Advanced Studies in Vienna with Vienna University of Technology in 2015. From 2015 to 2020, he lived in Durham, North Carolina where he attended Duke University. He expects to graduate with a doctorate of philosophy in economics in May 2020.