

Essays on Financial Econometrics

by

Rui Chen

Department of Economics

Duke University

Date: _____

Approved:

Andrew Patton, Advisor

Jia Li

Tim Bollerslev

Federico Bugni

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

2020

ABSTRACT

Essays on Financial Econometrics

by

Rui Chen

Department of Economics
Duke University

Date: _____

Approved:

Andrew Patton, Advisor

Jia Li

Tim Bollerslev

Federico Bugni

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

2020

Copyright © 2020 by Rui Chen
All rights reserved

Abstract

This dissertation contains five chapters. Chapter 1 gives an overview of this dissertation. The second chapter, which is joint work with Jia Li, Viktor Todorov and George Tauchen, develops an efficient mixed-scale estimator for jump regressions using high-frequency asset returns. A novel bootstrap procedure is proposed to make inference about our estimator, which has a non-standard asymptotic distribution that cannot be made asymptotically pivotal via studentization. The Monte Carlo analysis indicates good finite-sample performance of the general specification test and confidence intervals based on the bootstrap. When the method is applied to a high-frequency panel of Dow stock prices together with the market index defined by the S&P 500 index futures over the period 2007–2014, we observe remarkable temporal stability in the way that stocks react to market jumps.

Chapter 3 is co-authored with Andrew J. Patton and Johanna F. Ziegel. We use recent results from statistical decision theory to overcome the problem of “elicitability” for ES by *jointly* modelling ES and VaR, and propose new time series models for these risk measures. Estimation and inference methods are provided for the proposed models and confirmed via simulation studies to have good finite-sample properties. We apply these models to daily returns on four international equity indices, and find the proposed new ES-VaR models outperform forecasts based on GARCH or rolling window models.

Chapter 4 is my single-authored paper which proposes a consistent specification test of dynamic joint models for VaR and ES. To overcome the intractability problem of the asymptotic distribution of the test statistics under the null hypothesis, we use subsampling approximation to get the asymptotic critical values. The proposed test is confirmed via Monte Carlo studies to have better empirical size and power

performance in finite samples than other existing tests.

Finally, Chapter 5 concludes.

Contents

Abstract	iv
List of Figures	ix
List of Tables	x
Acknowledgements	xii
1 Introduction	xiii
2 Mixed-scale Jump Regressions with Bootstrap Inference	1
2.1 Introduction	1
2.2 The setting for mixed-scale jump regressions	6
2.2.1 The formal setup	6
2.2.2 Mixed-Scale Jump Regressions	8
2.3 Asymptotic theory	13
2.3.1 The efficient estimation of jump beta	13
2.3.2 Higher-order refinement and bootstrap inference	17
2.3.3 Specification testing and its bootstrap implementation	23
2.4 Simulation Results	25
2.5 Empirical application	30
2.6 Conclusion	41
3 Dynamic Semiparametric Models for ES (and VaR)	43
3.1 Introduction	43
3.2 Dynamic models for ES and VaR	47
3.2.1 A consistent scoring rule for ES and VaR	47

3.2.2	A GAS model for ES and VaR	52
3.2.3	A one-factor GAS model for ES and VaR	57
3.2.4	Existing dynamic models for ES and VaR	58
3.2.5	GARCH and ES/VaR estimation	60
3.3	Estimation of dynamic models for ES and VaR	62
3.4	Simulation study	67
3.5	Forecasting equity index ES and VaR	78
3.5.1	In-sample estimation	80
3.5.2	Out-of-sample forecasting	83
3.6	Conclusion	93
4	A Consistent Joint Test of Dynamic Models for VaR and ES	94
4.1	Introduction	94
4.2	The Test Statistics and Its Asymptotic Theory	97
4.2.1	The Test Statistics	97
4.2.2	The Limiting Distribution of the Test Statistics under the Null	100
4.2.3	Consistency against all Fixed Alternatives	104
4.3	Subsampling Approximation	105
4.4	Finite Sample Performance	108
4.4.1	Comparative Test Statistics	109
4.4.2	DGP	110
4.4.3	Discussion on the Results	112
4.4.4	Regression-based Test	118
4.5	Proofs	122
4.6	Conclusion	136

5	Conclusions	138
A	Appendix to Chapter 3	140
A.1	Proofs	140
A.2	Derivations	143
A.2.1	Generic calculations for the FZ0 loss function	143
A.2.2	Derivations for the one-factor GAS model for ES and VaR	145
A.2.3	ES and VaR in location-scale models	146
A.2.4	VaR and ES for Hansen’s skew t random variables	147
A.3	Estimation using the FZ0 loss function	148
A.4	Dynamics in the Skew t distribution	150
B	Supplemental Appendix to Chapter 3	152
B.1	Detailed proofs	152
B.2	Estimating a GARCH(1,1) model by FZ loss minimization	172
C	Additional Tables for Chapter 3	191
	Bibliography	199
	Biography	210

List of Figures

2.1	Market Jump Events and Stock Reaction.	32
2.2	Scatter of Stock versus Market Returns at Market Jump Times of the Full Sample.	33
2.3	Time Series of Yearly Jump Betas, 2007–2014.	37
2.4	Scatter of Stock versus Market Returns at Sector-Specific Jump Times.	40
3.1	Plot of FZ0 loss function	50
3.2	Contours of expected FZ0 loss when the target variable is standard Normal.	51
3.3	Plot of values of VaR and ES as a function of the lagged return.	56
3.4	Plot of the estimated 5% VaR and ES or daily returns on the S&P 500 index, over the period January 1990 to December 2016.	86
3.5	Plot of the estimated 5% VaR and ES for daily returns on the S&P 500 index, over the period January 2015 to December 2016.	87

List of Tables

2.1	Monte Carlo Rejection Rates of Specification Tests	28
2.2	Summary of Estimation and Coverage Results	29
2.3	Full sample WLS beta estimates	34
2.4	Specification testing results for 30 DJIA stocks	36
2.5	R^2 of the market factor for two types of jumps.	38
3.1	Simulation results for Normal innovations	71
3.2	Simulation results for skew t innovations	72
3.3	Simulation results for $T=500$	73
3.4	Sampling variation of FZ estimation relative to (Q)MLE and CAViaR	74
3.5	Mean absolute errors for VaR and ES estimates	77
3.6	Summary statistics on the four daily equity return series.	79
3.7	ARMA, GARCH, and Skew t results	79
3.8	Estimated paramters of GAS models for VaR and ES	82
3.9	Out-of-sample average losses and goodness-of-fit tests ($\alpha=0.05$) .	88
3.10	Diebold-Mariano t-statistics on average out-of-sample loss differences	88
3.11	Out-of-sample performance rankings for various alpha	92
4.1	Empirical rejection rates for jointly testing 1% VaR and ES at 5% significant level based on 1000 simulations	114
4.2	Empirical rejection rates for jointly testing 2.5% VaR and ES at 5% significant level based on 1000 simulations	115

4.3	Empirical rejection rates for jointly testing 5% VaR and ES at 5% significant level based on 1000 simulations	116
4.4	Empirical rejection rates for jointly testing 10% VaR and ES at 5% significant level based on 1000 simulations	117
4.5	Empirical size for regression-based test of $\alpha = 1\%, 2.5\%, 5\%, 10\%$ VaR and ES at 5% significant level based on 1000 simulations	121
C.1	Finite-sample performance of (Q)MLE	192
C.2	Simulation results for Normal innovations, estimation by CAViaR . . .	193
C.3	Simulation results for skew t innovations, estimation by CAViaR . . .	194
C.4	Diebold-Mariano t-statistics on average out-of-sample loss differences for the DJIA, NIKKEI and FTSE100 ($\alpha = 0.05$)	195
C.5	(cont'd) Diebold-Mariano t-statistics on average out-of-sample loss dif- ferences for the DJIA, NIKKEI and FTSE100 ($\alpha = 0.05$)	196
C.6	Out-of-sample average losses and goodness-of-fit tests ($\alpha=0.025$) .	196
C.7	Diebold-Mariano t-statistics on average out-of-sample loss differences for the S&P 500, DJIA, NIKKEI and FTSE100 ($\alpha=0.025$)	197
C.8	(cont'd): Diebold-Mariano t-statistics on average out-of-sample loss differences for the S&P 500, DJIA, NIKKEI and FTSE100 ($\alpha=0.025$)	198

Acknowledgements

I am deeply indebted to my advisor, Prof. Andrew Patton, for his continued guidance and support throughout my entire Ph.D journey. The generosity with which he was willing to share his time is truly encouraging. I could not have imagined having a better advisor and mentor for my Ph.D study.

I would also like to express my sincere thanks and appreciation to Prof. Jia Li and Prof. Tim Bollerslev, for their insightful advice and for their mental support, which helped me grow both in research ability, but also in wisdom that would nourish my whole life. I am also grateful to Prof. Federico Bugni for his enormous help and support on both my research and job looking process.

I would also like to thank Charles Becker who gave me a lot of help and support during my master's study. Special thanks also go to James Robert who provided me teaching opportunities from which I learnt a lot.

The amazing and various fitness classes provided by Duke University also have made my years in Durham much more colourful. Thanks to Casey and Kyra who introduced me to the Zumba world, and helped me find this life-long hobby.

Last but not the least, I would like to thank my parents for their immeasurable love and encouragement, and I dedicate this work to them.

Chapter 1

Introduction

This dissertation contains my research results on two topics of financial econometrics. The first topic is jump regression where the first step – jump selection step can be viewed as the analogy of dimension reduction for the classical big "P" problem in statistics to the big "N" problem in financial econometrics. The second step – the regression step leads to an estimator which shares the same formula as the OLS coefficient in an univariate regression. However, the jump regression is completely different from OLS in theory. The asymptotic theory for OLS is based on the assumption that the sample size goes to infinity. However, in any given fixed span of time, there are only a finite number of jumps. Therefore, the common intuition underlying the law of large numbers does not apply there. The asymptotics of our estimator is actually based on the fact that the error term in our regression is the Brownian motion continuous returns of order $\sqrt{\Delta_n}$, where Δ_n is the size of sampling interval, and shrinks to 0 as the interval size goes to 0.

Chapter 2, which is joint work with Jia Li, Viktor Todorov and George Tauchen, contributes to this topic by extending the pioneering work of Li, Todorov, and Tauchen (2016) to allow for different sampling frequency of the explanatory and dependent variables. This mixed-scale adaption was motivated by observation in the data that market return (approximated by the S&P 500 E-mini futures) usually responds to the market-wide shocks ¹ very quickly, within 1 minute usually, while the less liquid individual stocks (we focus on Dow 30 stocks) may take longer time

¹The market-wide shocks include, but are not limited to, macro announcements, for example, the Fed announcements followed by Fed meetings, geopolitical events and natural disasters etc.

than the market to fully incorporate the new information. This extension not only provides a flexible way of using data for empirical studies, but also leads to novel asymptotic results (cf. Li, Todorov, and Tauchen (2016)).

The second topic is about estimation and testing of time series models for Value-at-Risk (VaR) and Expected Shortfall (ES). This research topic was motivated by the Third Basel Accord (Basel Committee, 2010), where new emphasis is placed on ES as a measure of risk, complementing, and in parts substituting, the more-familiar VaR measure. ES is the expected return on an asset conditional on the return being below a given quantile of its distribution, namely its VaR. As Basel III is implemented worldwide (implementation is expected to occur in the period leading up to January 1st, 2019), ES will inevitably gain, and require, increasing attention from risk managers and banking supervisors and regulators. There is, however, a paucity of empirical models for ES. This dearth is perhaps in part because regulatory interest in this risk measure is only recent, and may also be due to the fact that this measure is not “elicitable.”² A recent result from Fissler and Ziegel (2016), who show that ES is *jointly elicitable* with VaR, opens a new channel to model ES.

Chapter 3, which is co-authored with Andrew J. Patton and Johanna F. Ziegel, use this result to propose new time series models for ES. Asymptotic theory is developed for a general class of dynamic semiparametric models for ES and VaR. We apply our new models and estimation methods to an out-of-sample analysis of forecast of ES and VaR for four international equity indices over the period January 1990 to December 2016 and find the proposed new ES-VaR models outperform forecasts based on GARCH or rolling window models.

Continuing this line of research Chapter 4, which is my single-authored paper,

²A risk measure (or statistical functional more generally) is said to be “elicitable” if there exists a loss function such that the risk measure is the solution to minimizing the expected loss.

proposes a consistent specification test of dynamic joint models for VaR and ES. To overcome the intractability problem of the asymptotic distribution of the test statistics under the null hypothesis, we use subsampling approximation to get the asymptotic critical values. The proposed test is confirmed via Monte Carlo studies to have better empirical size and power performance in finite samples than other existing tests.

Chapter 2

Mixed-scale Jump Regressions with Bootstrap Inference

2.1 Introduction

The availability of high-frequency data has led to new ways of estimating an asset's exposures to systematic risks such as the aggregate stock market return in the standard CAPM. The high-frequency estimation approach ((Andersen, Bollerslev, Diebold, and Vega, 2003); Barndorff-Nielsen and Shephard (2004a); Andersen, Bollerslev, Diebold, and Wu (2006); Mykland and Zhang (2009)) uses realized variation measures to infer beta over a fixed period of time, usually a day or a month, and then tracks these estimates over non-overlapping sample periods. More recent practice is to conduct estimation using jump-robust measures of variation and covariation (Todorov and Bollerslev (2010); Gobbi and Mancini (2012)). All of the above mentioned beta measures (with or without truncation) mainly pertain to the locally Gaussian diffusive moves in the market, because the large number of small diffusive moves are known to account for a major part of the market variation. Economically speaking, these small moves in part reflect the market's gradual price discovery process of distilling minor news on fundamentals from noise trading (Kyle (1985)) which can lead to a situation with low signal to noise ratio and temporal instability.¹ Li, Todorov, and Tauchen (2016), on the other hand, suggest an opposite approach that

¹Indeed, Kalnina (2013) and Reiss, Todorov, and Tauchen (2015) document that spot betas remain constant only over very short periods of time, usually a week or, at best, a month.

mainly uses abrupt and locally large jump moves to generate an effective measure of beta.²³ Such moves are typically related to important market-wide shocks which include, but are not limited to, macro announcements, geopolitical events and natural disasters. [See Chapter 8 of Hasbrouck (2015) for more discussion.]

The use of large rare jumps in a regression setting requires new ways of thinking about regression and inference. On the one hand, in any given fixed span of time, there are only a finite number of jumps. This means that the number of informative observations in a jump regression is finite and does not increase to infinity asymptotically.⁴ Therefore, the common intuition underlying the law of large numbers does not apply here. On the other hand, we recognize that the jumps are of fixed size regardless of the sampling frequency, whereas the diffusive moves are on the order $\Delta_n^{1/2}$, where Δ_n is the sampling interval which goes to zero asymptotically. The diffusive moves in the vicinity of jumps can be viewed as measurement errors induced by discrete sampling, and they play the role of random disturbances in classical regressions. The magnitude of such measurement error shrink at the parametric rate with well-behaved asymptotic properties, which can be further used for studying the asymptotics of our estimators. In the same vein, the correct specification of a linear jump regression model amounts to a perfect fitting (i.e., $R^2 = 1$) of the dependent jumps in the continuous-time limit. This test can be carried out by examining whether the observed R^2 is statistically significantly below unity.

This chapter develops a new mixed-scale strategy for jump regressions, which

²Jump betas have been first introduced in Todorov and Bollerslev (2010). Todorov and Bollerslev (2010) use higher order power variations to identify the jump betas from the high-frequency data. This approach, unlike Li, Todorov, and Tauchen (2016), makes use of all of the high-frequency increments. Of course, the role of the increments without jumps vanishes asymptotically in the higher order power variations.

³Theoretically, the betas at jump and non-jump times do not need to coincide.

⁴Even if the asset price process has infinitely active jumps, the number of jumps that have sizes greater than any fixed level remains finite.

addresses a natural asymmetry between the explanatory and dependent variables seen in applications.⁵ On the one hand, the explanatory variables are often returns of highly liquid assets such as market index futures. We sample these variables at a fine scale, which greatly improves the accuracy of jump detection. On the other hand, the dependent variables are typically returns of less liquid assets such as individual stocks, which are subject to a slower price discovery process for incorporating new information. Realistically speaking, due to the trading mechanisms on the exchanges, a jump typically cannot be observed instantly. Rather, it is often realized through a sequence of transactions. See Barndorff-Nielsen, Hansen, Lunde, and Shephard (2009) for a discussion of what they term “gradual jumps.” It is therefore prudent to sample these asset prices at a coarse scale when estimating the jump regression model, at the cost of statistical efficiency. The mixed-scale approach provides a flexible way of using data that play distinct roles in the jump regression. The fact that the jump detection step and the jump regression step are performed under two (possibly) distinct scales also leads to novel asymptotic results (cf. Li, Todorov, and Tauchen (2016)). In addition, we present all theory here in a multivariate setting so as to facilitate applications to multi-factor models of risk exposure.

We extend the analysis of Li, Todorov, and Tauchen (2016) by providing a refined inference for the mixed-scale jump regression which is beneficial, particularly when sampling at coarser frequencies. We first derive a higher-order asymptotic approximation for the jump regression estimates. This expansion accounts for the error in the volatility estimation around the jump times (which is of higher order). We then

⁵Our mixed-scale strategy is designed to improve the accuracy of jump detection for a subvector of a multivariate semimartingale process, so the goal here is to reduce the misclassification (i.e., jump or non-jump) error. This is fundamentally different from the multi-scale method of Zhang, Mykland, and Ait-Sahalia (2005), which conducts a jackknife bias-correction using realized variances computed at subsamples with different frequencies in the estimation of integrated volatility.

propose a bootstrap method which we show is asymptotically valid. The bootstrap provides a conceptually different alternative to the higher-order asymptotic expansion. The latter is based on direct higher-order asymptotic approximations while the current bootstrap method is aimed at approximating the finite sample distribution of the estimator using simulated data. Our motivation for using the bootstrap is that the asymptotic distribution of the estimator of jump beta is non-standard because volatility may co-jump at the jump times of the explanatory variable(s); see, for example, Jacod and Todorov (2010), Todorov and Tauchen (2011) and Bandi and Renó (2015). In fact, the limiting distribution of the estimator is not Gaussian even conditional on the underlying information set. The asymptotic covariance matrix alone is thereby insufficient for asymptotically valid inference; in particular, the conventional t -statistic is not asymptotically pivotal. We therefore propose a bootstrap method that is very simple to implement. The user only needs to repeatedly compute the estimator in a bootstrap sample that consists of small sub-samples within local windows of the detected jump times. The bootstrap sample size is much smaller than the original sample size, resulting in a significant reduction in computational time. The same bootstrap sample can also be used to compute critical values for the specification test. The bootstrap procedure achieves a higher-order refinement over the asymptotic approximations to the usual order. Our bootstrap refinement is atypical because it does not concern Edgeworth expansions for asymptotically pivotal statistics; instead, here, the refinement accounts for the higher-order sampling variability in the weights of the efficient regression procedure. Monte Carlo evidence shows good finite-sample performance of the bootstrap.

The bootstrap has been first introduced to the high-frequency literature by Gonçalves and Meddahi (2009) in the context of estimating integrated volatility. Since we focus on the inference about jumps, which is well known to be very different from the

inference about volatility, the proposed bootstrap method and the associated asymptotic theory deviate significantly from prior work. To the best of our knowledge, the paper that this chapter is based on is the first to study the bootstrap inference for jumps using high-frequency data. Although the bootstrap method is presented in the context of jump regressions, it can be readily extended to many other contexts concerning jumps as well.

We apply the mixed-scale jump regression method to a high-frequency one-minute panel of Dow stock prices together with the S&P 500 E-mini futures price for the market index over the period 2007–2014. We start with concrete examples of how individual asset prices react, either promptly or gradually, to news events generating market jumps, so as to illustrate the empirical relevance of the mixed-scale approach. We further provide evidence that using a coarse scale of 3–5 minutes is sufficiently conservative in the jump regression step for these blue-chip stocks. We then proceed to conduct stock-by-stock tests of the key hypothesis that $R^2 = 1$.⁶ A striking finding is that by sampling the data on a slightly coarse scale in the regression step, the null hypothesis is rejected much less frequently. This reduction cannot be fully explained by pure statistical reasons. Instead, it reaffirms the usefulness of the mixed-scale approach in the testing context. Using the efficient estimator, we document how the market jump risk exposure varies across stocks and over time. Lastly, we study the sensitivity of various stocks to market risk at alternative jump times defined by sector-specific jumps in the nine industry ETFs for the S&P 500 composite index. For many of the stocks in our sample, we find the relationship between individual stocks and the market to be significantly noisier and more unstable at the sector-specific jump times than it is at the market-wide jump times.

⁶Earlier work by Roll (1987) have documented relatively low R^2 -s of time series regressions of stocks' returns on their systematic risk exposures, even after excluding days with firm-specific news (and hence more idiosyncratic noise).

The rest of this chapter is organized as follows. Section 2.2 describes the econometric framework and Section 2.3 presents the main theorems. Section 2.4 contains the Monte Carlo evaluation and Section 2.5 shows the empirical results. Section 2.6 concludes.

2.2 The setting for mixed-scale jump regressions

We describe the formal high-frequency asymptotic setting in Section 2.2.1 and the mixed-scale jump regression setting in Section 2.2.2. The following notations are used in the sequel. We denote the transpose of a matrix A by A^\top and denote its (j, k) component by A^{jk} . All vectors are column vectors. For notational simplicity, we write (a, b) in place of $(a^\top, b^\top)^\top$. For two vectors a and b , we write $a \leq b$ if the inequality holds component-wise. The Euclidean norm of a linear space is denoted by $\|\cdot\|$. The cardinality of a (possibly random) set \mathcal{P} is denoted by $|\mathcal{P}|$. The largest smaller integer function is $\lfloor \cdot \rfloor$. For two sequences of positive real numbers a_n and b_n , we write $a_n \asymp b_n$ if $b_n/c \leq a_n \leq cb_n$ for some constant $c \geq 1$ and all n . All limits are for $n \rightarrow \infty$. We use $\xrightarrow{\mathbb{P}}$ and $\xrightarrow{\mathcal{L}\text{-}s}$ to denote convergence in probability and stable convergence in law, respectively.

2.2.1 The formal setup

We proceed with the formal setup. Let Y and Z be defined on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, \mathbb{P})$ which take values in \mathbb{R} and \mathbb{R}^{d_z} , respectively. Throughout the chapter, all processes are assumed to be càdlàg (i.e., right continuous with left limit) adapted. Let $X \equiv (Y, Z)$ and $d \equiv d_z + 1$. The d -dimensional process X is observed at discrete times $i\Delta_n$, for $i \in \{0, \dots, \lfloor T/\Delta_n \rfloor\}$, within the fixed time interval $[0, T]$, where the sampling interval $\Delta_n \rightarrow 0$ asymptotically. We denote the increments of X

by

$$\Delta_i^n X \equiv X_{i\Delta_n} - X_{(i-1)\Delta_n}, \quad i \in \mathcal{I}_n \equiv \{1, \dots, \lfloor T/\Delta_n \rfloor\}. \quad (2.1)$$

Our basic assumption is that X is a d -dimensional Itô semimartingale (see, e.g., Jacod and Protter (2012), Section 2.1.4) of the form

$$\begin{cases} X_t = X_t^c + J_t, \\ X_t^c = X_0 + \int_0^t b_s ds + \int_0^t \sigma_s dW_s & \text{(continuous component),} \\ J_t = \int_0^t \int_{\mathbb{R}} \delta(s, u) \mu(ds, du) & \text{(jump component),} \end{cases} \quad (2.2)$$

where the drift b_t takes value in \mathbb{R}^d ; the volatility process σ_t takes value in \mathcal{M}_d , the space of d -dimensional positive definite matrices; W is a d -dimensional standard Brownian motion; $\delta(\cdot) \equiv (\delta_Y(\cdot), \delta_Z(\cdot)) : \Omega \times \mathbb{R}_+ \times \mathbb{R} \mapsto \mathbb{R}^d$ is a predictable function; μ is a Poisson random measure on $\mathbb{R}_+ \times \mathbb{R}$ with its compensator $\nu(dt, du) = dt \otimes \lambda(du)$ for some measure λ on \mathbb{R} . The jump of X at time t is denoted by

$$\Delta X_t \equiv X_t - X_{t-}, \quad \text{where } X_{t-} \equiv \lim_{s \uparrow t} X_s. \quad (2.3)$$

We denote the spot covariance matrix of X at time t by $c_t \equiv \sigma_t \sigma_t^\top$. Our basic regularity condition for X is the following.

Assumption 1. (a) The process $(b_t)_{t \geq 0}$ is locally bounded; (b) c_t is nonsingular for $t \in [0, T]$ almost surely; (c) $\nu([0, T] \times \mathbb{R}) < \infty$.

The only nontrivial restriction in Assumption 1 is the assumption of finite-activity jumps in X . This assumption is used mainly to simplify our technical exposition because our empirical focus in this chapter are the big jumps. Technically speaking, this means that we can drop Assumption 1(c) and focus on jumps with size bounded away from zero. Doing so automatically verifies the finite-activity assumption, but

with very little effect on the empirical investigation in the current chapter.

2.2.2 Mixed-Scale Jump Regressions

The jump regression is based on the following (population) relationship between the jumps of Y and Z :

$$\Delta Y_\tau = \beta^{*\top} g(\Delta Z_\tau), \quad \tau \in \mathcal{T}, \quad (2.4)$$

where $g(\cdot) : \mathbb{R}^{d_z} \mapsto \mathbb{R}^q$ is a deterministic function, τ is a jump time of the process Z , and \mathcal{T} collects these jump times. We stress that the restriction (2.4) is only postulated at the jump times of Z . In particular, we allow Y to have idiosyncratic jumps, i.e., jumps that do not occur at the same times as those of Z . Therefore, in general (provided $g(\mathbf{0}) = \mathbf{0}$) we have

$$\Delta Y_t = \beta^{*\top} g(\Delta Z_t) + \Delta \epsilon_t, \quad \Delta Z_t \Delta \epsilon_t = \mathbf{0}, \quad t \in [0, T], \quad (2.5)$$

with ϵ_t capturing the idiosyncratic jump risk in the asset Y . We note that this type of model for the jump parts of assets naturally arises in economies in which the market-wide pricing kernel is specified as a function of systematic factors (containing jumps) and the cash flows of the assets contain in addition idiosyncratic jump shocks in the sense of Merton (1976). We refer to Li, Todorov, and Tauchen (2016) for more discussion of our deterministic jump model.

We refer to the coefficient β^* as the jump beta, which is the parameter of interest in our econometric analysis. As in Li, Todorov, and Tauchen (2016), we are mainly interested in the linear specification $g(\Delta Z_\tau) = \Delta Z_\tau$ because it turns out to deliver quite good fitting in practice. That being said, the general form (2.4) is also of economic interest. For example, with $g(\Delta Z_\tau) = (\Delta Z_\tau 1_{\{\Delta Z_\tau > 0\}}, \Delta Z_\tau 1_{\{\Delta Z_\tau < 0\}})$, (2.4)

conveniently allows for asymmetric response of Y with respect to positive and negative jumps in Z . Assumption 2, below, ensures the identification of the jump beta. It also imposes some mild smoothness condition on $g(\cdot)$ that facilitates the asymptotic analysis.

Assumption 2. (a) The matrix $\sum_{\tau \in \mathcal{T}} g(\Delta Z_\tau) g(\Delta Z_\tau)^\top$ is nonsingular almost surely.

(b) For each t , the measure defined by $A \mapsto \lambda(\{u : \delta_Z(t, u) \in A \setminus \{0\}\})$ is atomless.

Moreover, $g(\cdot)$ is twice continuously differentiable almost everywhere.

In finite samples, neither the times nor the magnitudes of jumps are directly observable. Empirically, we need to use discretely sampled data $\Delta_i^n X = (\Delta_i^n Y, \Delta_i^n Z)$ to make statistical inference based on model (2.4). Since (2.4) only concerns the jump moves of the asset prices, it is conceptually natural to first select observed returns that contain jumps. We do so using a thresholding method (Mancini (2001)) as follows. We consider a sequence of thresholds $(u_n)_{n \geq 1} \subset \mathbb{R}^{d_z}$ such that

$$u_{j,n} \asymp \Delta_n^\varpi, \quad \text{for some } \varpi \in (0, 1/2) \text{ and all } 1 \leq j \leq d_z.$$

We then collect the jump returns using

$$\mathcal{J}_n \equiv \mathcal{I}_n \setminus \{i : -u_n \leq \Delta_i^n Z \leq u_n\}. \quad (2.6)$$

Time-invariant choice for u_n , although asymptotically valid, leads to very bad results in practice as it does not account for the time-varying diffusive spot covariance matrix c_t . Hence, a sensible choice for u_n should take into account the variation of c_t in an adaptive, data-driven way. We refer to our application in Sections 2.4 and 2.5 for the details of such a way of constructing u_n using the bipower variation estimator (Barndorff-Nielsen and Shephard (2004c)).

Under Assumption 1, it can be shown that \mathcal{J}_n consistently locates the sampling intervals that contain jumps.⁷ That is,

$$\mathbb{P}(\mathcal{J}_n = \mathcal{J}_n^*) \rightarrow 1, \quad \text{where } \mathcal{J}_n^* \equiv \{i : \tau \in ((i-1)\Delta_n, i\Delta_n] \text{ for some } \tau \in \mathcal{T}\}. \quad (2.7)$$

Parallel to (2.4), the jump regression equation is then given by

$$\Delta_i^n Y = \beta^{*\top} g(\Delta_i^n Z) + \varepsilon_i^n, \quad i \in \mathcal{J}_n, \quad (2.8)$$

with the error term ε_i^n being implicitly defined by (2.8).

Despite the apparent similarity between the jump regression equation (2.8) and the classical regression, there are fundamental differences. We first observe that (2.8) only concerns a finite number of large jump returns even asymptotically (recall (2.7)). This means, the intuition underlying the classical law of large numbers and the central limit theorem does not apply here. The reason is that the finite number of error terms $(\varepsilon_i^n)_{i \in \mathcal{J}_n}$ would not “average out.” However, we observe that these error terms are actually asymptotically small. Indeed, under (2.4), we have for each $i \in \mathcal{J}_n^*$,

$$\varepsilon_i^n = \Delta_i^n Y^c - \beta^{*\top} (g(\Delta Z_\tau + \Delta_i^n Z^c) - g(\Delta Z_\tau)),$$

where τ is the unique (which holds true at high frequency) jump time that occurs in $((i-1)\Delta_n, i\Delta_n]$. Since the diffusive moves $(\Delta_i^n Y^c, \Delta_i^n Z^c)$ are of order $O_p(\Delta_n^{1/2})$, so is ε_i^n . In addition, these small error terms have well-behaved asymptotic properties, which we use to derive the asymptotic property of our inference procedures.

In empirical work, the use of high-frequency data is confounded by various trading frictions that make the transaction price deviate from the efficient price. The

⁷See, for example, Proposition 1 of Li, Todorov, and Tauchen (2016).

deviation of the observed from efficient (fundamental) price is commonly referred to as microstructure noise. Typical sources of microstructure noise are bid-ask bounces and rounding error. A standard assumption in the literature is to assume that the noise is centered at zero and it has some form of weak dependence across observation times. There is a large body of work dealing with microstructure noise of this type, see e.g., chapter 16 of Jacod and Protter (2012) and the many references therein. The earlier literature has “dealt” with the potential presence of noise by sampling sparsely, with the idea being that at the coarser frequencies the importance of the noise in relative terms is rather small and can be ignored. Subsequent work has developed formal statistical methods for dealing with the microstructure noise. Although the methods differ, they are all based on averaging the noise in some way. In other words, the existing methods all rely on weak dependence of the noise at observation times and apply law of large numbers to purge the high frequency based measures from it.

There is another type of microstructure noise, which stems from staleness and infrequent trading. Mainly, for assets which are not very liquid, the price can be relatively slow to react to news. In particular, when there is a big jump on the stock market, less liquid individual assets can be slow to react and adjust fully to the new (latent) efficient price level that corresponds to the new information. There can be various sources for this type of price staleness. One typical example is the presence of stale limit orders in the limit order book which get “hit” as the price is adjusting to the new equilibrium level. This type of noise causes a phenomenon referred to by Barndorff-Nielsen, Hansen, Lunde, and Shephard (2009) as gradual jumps. Obviously this type of noise is very difficult to deal with formally as it by its very nature has a lot of dependence across observation times and also it depends very strongly on the actual fundamental price. Hence local averaging type procedures will not work for

it. Also, it is clear that this type of trading friction has a rather nontrivial impact on the analysis of jumps since, by their very nature, jumps are rare events.

In this chapter we are mainly concerned with the second type of noise, i.e., the one that is due to staleness.⁸ To mitigate its impact, we will sample sparsely. The proper sampling scheme of course depends on the asset of interest as staleness and liquidity are asset specific. For example, in our applications, we take Y to be the price of a blue-chip stock and take Z to be the price of a futures contract on a major market index. It is common to sparsely sample the stocks at, say, every 3–5 minutes, while the highly liquid index futures can be safely sampled at shorter intervals such as every minute.

The difference in liquidity of the left- and the right-hand side assets hence creates an interesting tradeoff in the choice of the sampling scheme. On the one hand, sampling at high frequency (e.g., 1 minute) greatly increases the accuracy for jump detection and, hence, reduces jump-misclassification bias in finite samples. On the other hand, sampling at such frequency is unlikely to be conservative enough for mitigating microstructure effects in Y . Indeed, as we shall illustrate with concrete examples in Section 2.5, individual stocks may take longer time than the market to fully incorporate new information that leads to a visible jump in the market index. See Barndorff-Nielsen, Hansen, Lunde, and Shephard (2009) for additional discussions on this type of gradual jumps.

We propose to break the tension between these two conflicting effects using a mixed-scale jump regression procedure: we maintain the jump detection (2.6) at the fine sampling scale Δ_n , but implement the jump regression at a (possibly) coarser scale $k\Delta_n$ for some $k \geq 1$. By doing so, we maintain high precision in detecting

⁸For the frequencies we use in our empirical work, the first type of noise has relatively small impact, see Section 2.5 for further details.

market jumps and reduce the concern of “breaking” gradual jumps. More precisely, we denote $\Delta_{i,k}^n X = (\Delta_{i,k}^n Y, \Delta_{i,k}^n Z)$, where

$$\Delta_{i,k}^n X = X_{(i-1+k)\Delta_n} - X_{(i-1)\Delta_n}.$$

The mixed-scale jump regression is then given by, with $\varepsilon_{i,k}^n$ implicitly defined below,

$$\Delta_{i,k}^n Y = \beta^{*\top} g(\Delta_{i,k}^n Z) + \varepsilon_{i,k}^n, \quad i \in \mathcal{J}_n. \quad (2.9)$$

Clearly, (2.8) is a special case of (2.9) with $k = 1$. The fact that the jump detection and the jump regression are performed at different sampling scales leads to notable differences between the inference procedures proposed below and those in the single-scale setting of Li, Todorov, and Tauchen (2016), mainly because of the presence of volatility-price co-jumps. We now turn to the details.

2.3 Asymptotic theory

2.3.1 The efficient estimation of jump beta

In this subsection, we describe a class of mixed-scale estimators for the jump beta and derive their asymptotic properties. In view of (2.9), a natural estimator of β^* is the mixed-scale ordinary least square (OLS) estimator given by

$$\hat{\beta}_n \equiv \left(\sum_{i \in \mathcal{J}_n} g(\Delta_{i,k}^n Z) g(\Delta_{i,k}^n Z)^\top \right)^{-1} \left(\sum_{i \in \mathcal{J}_n} g(\Delta_{i,k}^n Z) \Delta_{i,k}^n Y \right).$$

However, since the error terms $(\varepsilon_{i,k}^n)_{i \in \mathcal{J}_n}$ can exhibit arbitrary heteroskedasticity due to both time-varying volatility and jump size, the mixed-scale OLS estimator is not

efficient. Following Li, Todorov, and Tauchen (2016), we consider efficient estimation using a semiparametric two-step weighted estimator.

To construct the weights, we first nonparametrically estimate the spot covariance matrices before and after each detected jump. To this end, we pick an integer sequence k_n of block sizes such that

$$k_n \rightarrow \infty \quad \text{and} \quad k_n \Delta_n \rightarrow 0. \quad (2.10)$$

We also pick a \mathbb{R}^d -valued sequence u'_n of truncation thresholds that satisfies

$$u'_{j,n} \asymp \Delta_n^{-\varpi}, \quad \text{for some } \varpi \in (0, 1/2) \text{ and all } 1 \leq j \leq d.$$

We then set the index set of the diffusive returns to be

$$\mathcal{C}_n = \{i \in \mathcal{I}_n : -u'_n \leq \Delta_i^n X \leq u'_n\}. \quad (2.11)$$

For each $i \in \mathcal{J}_n$, we estimate the pre-jump and the post-jump spot covariance matrices using

$$\left\{ \begin{array}{l} \hat{c}_{n,i-} \equiv \frac{\sum_{j=0}^{k_n-1} (\Delta_{i-k_n+j}^n X)(\Delta_{i-k_n+j}^n X)^\top \mathbf{1}_{\{i-k_n+j \in \mathcal{C}_n\}}}{\Delta_n \sum_{j=0}^{k_n-1} \mathbf{1}_{\{i-k_n+j \in \mathcal{C}_n\}}}, \\ \hat{c}_{n,i+} \equiv \frac{\sum_{j=0}^{k_n-1} (\Delta_{i+k+j}^n X)(\Delta_{i+k+j}^n X)^\top \mathbf{1}_{\{i+k+j \in \mathcal{C}_n\}}}{\Delta_n \sum_{j=0}^{k_n-1} \mathbf{1}_{\{i+k+j \in \mathcal{C}_n\}}}. \end{array} \right. \quad (2.12)$$

We note that these spot covariance estimates are constructed using returns sampled at the “fine” scale which, in our empirical analysis in Section 2.5, is set to be 1 minute. At such a frequency, “usual” microstructure noises such as bid-ask bounces have negligible impact on volatility estimation for liquid stocks. That being said, one may also estimate spot volatilities at coarser sampling intervals to further guard against microstructure noise but at the cost of higher sampling variability in finite

samples. This only results in notational changes in the theory that follows and we omit the details for brevity.

We consider a weight function $w : \mathcal{M}_d \times \mathcal{M}_d \times \mathbb{R}^{d_z} \times \mathbb{R}^q \mapsto (0, \infty)$ such that $w(c_-, c_+, z, \beta)$ is continuously differentiable at $\beta = \beta^*$, all $c_-, c_+ \in \mathcal{M}_d$ and almost every $z \in \mathbb{R}^{d_z}$. To simplify notation, we denote

$$\hat{w}_{n,i} = w\left(\hat{c}_{n,i-}, \hat{c}_{n,i+}, \Delta_{i,k}^n Z, \hat{\beta}_n\right).$$

The mixed-scaled WLS estimator is then given by

$$\hat{\beta}_n(w) \equiv \left(\sum_{i \in \mathcal{J}_n} \hat{w}_{n,i} g(\Delta_{i,k}^n Z) g(\Delta_{i,k}^n Z)^\top \right)^{-1} \left(\sum_{i \in \mathcal{J}_n} \hat{w}_{n,i} g(\Delta_{i,k}^n Z) \Delta_{i,k}^n Y \right). \quad (2.13)$$

In order to describe the asymptotic behavior of $\hat{\beta}_n(w)$, we introduce some auxiliary random variables. Let $(\tau_p)_{p \geq 1}$ denote the successive jump times of Z . We consider random variables $(\kappa_p, \xi_{p-}, \xi_{p+})_{p \geq 1}$ that are mutually independent and are independent of \mathcal{F} such that κ_p is uniformly distributed on $[0, 1]$ and the variables (ξ_{p-}, ξ_{p+}) are d -dimensional standard normal. We then denote, for $p \geq 1$,

$$\begin{cases} \varsigma_p \equiv (1, -\beta^{*\top} \partial g(\Delta Z_{\tau_p})) \left(\sqrt{\kappa_p} \sigma_{\tau_p} \xi_{p-} + \sqrt{k - \kappa_p} \sigma_{\tau_p} \xi_{p+} \right), \\ w_p \equiv w(c_{\tau_{p-}}, c_{\tau_p}, \Delta Z_{\tau_p}, \beta^*). \end{cases} \quad (2.14)$$

Finally, we set

$$\Xi(w) \equiv \sum_{p \in \mathcal{P}} w_p g(\Delta Z_{\tau_p}) g(\Delta Z_{\tau_p})^\top, \quad \Lambda(w) \equiv \sum_{p \in \mathcal{P}} w_p g(\Delta Z_{\tau_p}) \varsigma_p.$$

Theorem 1, below, describes the stable convergence in law of $\hat{\beta}_n(w)$.

Theorem 1. *Under Assumptions 1 and 2, $\Delta_n^{-1/2}(\hat{\beta}_n(w) - \beta^*) \xrightarrow{\mathcal{L}\text{-}s} \Xi(w)^{-1} \Lambda(w)$.*

Theorem 1 shows that $\hat{\beta}_n(w)$ is a $\Delta_n^{-1/2}$ -consistent estimator of the jump beta, with \mathcal{F} -conditional asymptotic covariance matrix given by

$$\Sigma(w) \equiv \Xi(w)^{-1} \left(\sum_{p \in \mathcal{P}} w_p^2 \mathbb{E} [\zeta_p^2 | \mathcal{F}] g(\Delta Z_{\tau_p}) g(\Delta Z_{\tau_p})^\top \right) \Xi(w)^{-1},$$

where

$$\mathbb{E} [\zeta_p^2 | \mathcal{F}] = (1, -\beta^{*\top} \partial g(\Delta Z_{\tau_p})) \left(\frac{1}{2} c_{\tau_p^-} + \left(k - \frac{1}{2} \right) c_{\tau_p} \right) (1, -\beta^{*\top} \partial g(\Delta Z_{\tau_p}))^\top.$$

It is easy to see that $\Sigma(w)$ can be minimized using the weight function

$$w(c_-, c_+, z, \beta) \equiv \frac{1}{(1, -\beta^\top \partial g(z)) \left(\frac{1}{2} c_- + \left(k - \frac{1}{2} \right) c_+ \right) (1, -\beta^\top \partial g(z))^\top}.$$

We refer to the associated estimator as the optimally weighted estimator. By construction, it is more efficient than an unweighted estimator. Moreover, Li, Todorov, and Tauchen (2016) establish the semiparametric efficiency bound for estimating the jump beta in the case without volatility-price cojumps. In this case, the optimally weighted estimator defined above attains the efficiency bound computed for the coarsely sampled data. The reason for using the coarser frequency in the analysis of the semiparametric efficiency of the jump beta estimation is that the limiting distribution of the jump regression coefficient is determined only by the increments containing the jumps. However, these increments are aggregated to a coarser scale in order to guard against the gradual jump phenomenon. In this regard, we should stress that the frequency used for jump detection as well as for the estimation of volatility has no bearing on the efficiency statement. The reason is that the error coming from the jump detection as well as volatility measurement is of higher order in the jump regression.

2.3.2 Higher-order refinement and bootstrap inference

We now develop refined inference for the jump regression estimate of β^* . We first derive a high-order asymptotic result and then propose a bootstrap procedure which we show achieves the higher order asymptotic refinement.

To motivate, we observe that while the weighted estimator $\hat{\beta}_n(w)$ depends on the spot covariance estimates $(\hat{c}_{n,i-}, \hat{c}_{n,i+})$, the sampling variability of the latter is not reflected in the asymptotic distribution described by Theorem 1. The reason is that the local volatility estimates enter only the weights and their sampling errors are annihilated in the second-order asymptotics. In finite samples, the sampling variability of the spot covariance estimates may still have some effect, because the latter enjoy only a nonparametric convergence rate. To account for such effects, we need a refined characterization of the asymptotic behavior of the weighted estimator which we now provide. For the analysis here we need the following additional assumption for the volatility process.

Assumption 3. *The process σ_t is also an Itô semimartingale of the form*

$$\begin{aligned} \sigma_t = & \sigma_0 + \int_0^t \tilde{b}_s ds + \int_0^t \tilde{\sigma}_s dW_s + \int_0^t \int_{\mathbb{R}} \tilde{\delta}(s, u) 1_{\{\|\tilde{\delta}(s, u)\| > 1\}} \mu(ds, du) \\ & + \int_0^t \int_{\mathbb{R}} \tilde{\delta}(s, u) 1_{\{\|\tilde{\delta}(s, u)\| \leq 1\}} (\mu - \nu)(ds, du), \end{aligned}$$

where the processes \tilde{b}_t and $\tilde{\sigma}_t$ are locally bounded and for a sequence of stopping times $(T_m)_{m \geq 1}$ increasing to infinity and a sequence $(\tilde{J}_m)_{m \geq 1}$ of λ -integrable bounded functions, $\|\tilde{\delta}(t, u)\|^2 \wedge 1 \leq \tilde{J}_m(u)$ for all $t \leq T_m$ and $u \in \mathbb{R}$.

Assumption 3 is needed for characterizing the stable convergence of the spot covariance estimates. This assumption is fairly unrestrictive and is satisfied by many models in finance. In particular, it allows for “leverage effect,” that is, the Brownian

motions W and \widetilde{W} can be correlated. Moreover, Assumption 3 allows for volatility jumps, and it does not restrict their activity and dependence with the price jumps. However, this assumption does rule out certain long-memory volatility models driven by the fractional Brownian motion (see Comte and Renault (1996)).

We also need some additional notation. We consider $d \times d$ random matrices $(\zeta_{p-}, \zeta_{p+})_{p \geq 1}$ which, conditional on \mathcal{F} , are centered Gaussian, mutually independent and independent of $(\kappa_p, \xi_{p-}, \xi_{p+})_{p \geq 1}$, with conditional covariances given by

$$\begin{cases} \mathbb{E}[\zeta_{p-}^{jk} \zeta_{p-}^{lm} | \mathcal{F}] = c_{\tau_p}^{jl} c_{\tau_p}^{km} + c_{\tau_p}^{jm} c_{\tau_p}^{kl}, \\ \mathbb{E}[\zeta_{p+}^{jk} \zeta_{p+}^{lm} | \mathcal{F}] = c_{\tau_p}^{jl} c_{\tau_p}^{km} + c_{\tau_p}^{jm} c_{\tau_p}^{kl}, \end{cases} \quad 1 \leq j, k, l, m \leq d.$$

We denote the first differential of w by $dw(c_-, c_+, z, b) = \dot{w}(c_-, c_+, z, b; dc_-, dc_+, dz, db)$ and then set, for $p \geq 1$,

$$\tilde{w}_p \equiv \dot{w}(c_{\tau_p-}, c_{\tau_p}, \Delta Z_{\tau_p}, \beta^*; \zeta_{p-}, \zeta_{p+}, 0, 0).$$

Finally, for notational simplicity, we set

$$\begin{cases} \Xi(w) \equiv \sum_{p \in \mathcal{P}} w_p g(\Delta Z_{\tau_p}) g(\Delta Z_{\tau_p})^\top, & \Lambda(w) \equiv \sum_{p \in \mathcal{P}} w_p g(\Delta Z_{\tau_p}) \varsigma_p, \\ \tilde{\Xi}(w) \equiv \sum_{p \in \mathcal{P}} \tilde{w}_p g(\Delta Z_{\tau_p}) g(\Delta Z_{\tau_p})^\top, & \tilde{\Lambda}(w) \equiv \sum_{p \in \mathcal{P}} \tilde{w}_p g(\Delta Z_{\tau_p}) \varsigma_p. \end{cases}$$

The higher-order asymptotic expansion for $\Delta_n^{-1/2}(\hat{\beta}_n(w) - \beta^*)$ is given in the following theorem.

Theorem 2. *Suppose Assumptions 1, 2 and 3, and $k_n \asymp \Delta_n^{-a}$ for some $a \in (0, 1/2)$.*

(a) We can decompose

$$\Delta_n^{-1/2} \left(\hat{\beta}_n(w) - \beta^* \right) = \mathcal{L}_n(w) + k_n^{-1/2} \mathcal{H}_n(w) + o_p(k_n^{-1/2}), \quad (2.15)$$

such that

$$(\mathcal{L}_n(w), \mathcal{H}_n(w)) \xrightarrow{\mathcal{L}\text{-s}} (\mathcal{L}(w), \mathcal{H}(w)),$$

where

$$\begin{cases} \mathcal{L}(w) \equiv \Xi(w)^{-1} \Lambda(w), \\ \mathcal{H}(w) \equiv \Xi(w)^{-1} \tilde{\Lambda}(w) - \Xi(w)^{-1} \tilde{\Xi}(w) \Xi(w)^{-1} \Lambda(w). \end{cases}$$

(b) If, in addition, there is no price-volatility cojump and W is independent of (σ, J) , then $\sup_x |\mathbb{P}(\mathcal{L}_n(w) \leq x | \sigma, J) - \mathbb{P}(\mathcal{L}(w) \leq x | \sigma, J)| = O_p(\Delta_n^{1/2})$.

The leading term $\mathcal{L}_n(w)$ in (2.15) is what drives the convergence in Theorem 1. The higher-order term $k_n^{-1/2} \mathcal{H}_n(w)$ is $O_p(k_n^{-1/2})$ and hence is asymptotically dominated by $\mathcal{L}_n(w)$. The limiting variable $\mathcal{H}_n(w)$ involves not only $(\varsigma_p)_{p \geq 1}$ but also $(\tilde{w}_p)_{p \geq 1}$, where the latter captures the sampling variability in the weights due to the spot variance estimates. Part (b) of Theorem 2 further shows that the conditional law of the leading term converges at a (fast) parametric rate under the uniform metric.

Because of the higher-order asymptotic effect played by $\hat{c}_{n,i\pm}$ in the efficient beta estimation, the user has a lot of freedom in setting the block size k_n . Indeed, as seen from Theorem 2, we need only $k_n \asymp \Delta_n^{-a}$ with a in the wide range of $(0, 1/2)$. This is unlike the block-based volatility estimators, see e.g., Jacod and Rosenbaum (2013), where one has significantly less freedom in choosing k_n . Having the refined asymptotic result in Theorem 2 helps since if k_n is relatively small, the higher-order term $k_n^{-1/2} \mathcal{H}_n(w)$ might have nontrivial finite sample effect.

We now introduce a bootstrap algorithm and show that (see Theorem 3 below)

it provides the higher-order approximation described in Theorem 2. With a mild adjustment, the same bootstrap sample can also be used to compute critical values for the specification test developed in Section 2.3.3. The bootstrap was first introduced to the high-frequency setting by Gonçalves and Meddahi (2009) and Dovonon, Gonçalves, and Meddahi (2013) for making inference for integrated variance and covariance matrices; also see Hounyo (2013) and Dovonon, Hounyo, Gonçalves, and Meddahi (2014). We apply here the bootstrap to make inference for jumps, which is therefore very different from prior work that concerns volatility inference.⁹

Algorithm 1 (Bootstrapping $\hat{\beta}_n(w)$).

Step 1. In each bootstrap sample, we generate a d -dimensional standard Brownian motion W^* and random times $(\tau_i^*)_{i \in \mathcal{J}_n}$ which are mutually independent and independent of the data, such that each τ_i^* is drawn uniformly from $[(i-1)\Delta_n, i\Delta_n]$.¹⁰ Set the diffusive return for each $i \in \mathcal{J}_n$ as

$$\Delta_{i,k}^n X^{*c} \equiv \begin{pmatrix} \Delta_{i,k}^n Y^{*c} \\ \Delta_{i,k}^n Z^{*c} \end{pmatrix} = \hat{c}_{n,i-}^{1/2} (W_{\tau_i^*}^* - W_{(i-1)\Delta_n}^*) + \hat{c}_{n,i+}^{1/2} (W_{(i-1+k)\Delta_n}^* - W_{\tau_i^*}^*). \quad (2.16)$$

Step 2. Set $\Delta_{i,k}^n Z^* = \Delta_{i,k}^n Z + \Delta_{i,k}^n Z^{*c}$ and $\Delta_{i,k}^n Y^* = \hat{\beta}_n(w)^\top g(\Delta_{i,k}^n Z) + \Delta_{i,k}^n Y^{*c}$ for $i \in \mathcal{J}_n$. Compute $\hat{\beta}_n^*$ as the OLS estimator by regressing $\Delta_{i,k}^n Y^*$ on $g(\Delta_{i,k}^n Z^*)$ in the subsample $i \in \mathcal{J}_n$.

⁹Dovonon, Hounyo, Gonçalves, and Meddahi (2014) consider an application of the bootstrap for approximating the null asymptotic distribution of jump tests, which mainly concerns the jump-robust inference for the integrated variance, rather than the jump process itself.

¹⁰We note that the Gaussian increments of W^* are only needed within two-sided k_n -windows around the jump returns. This fact is useful for reducing the computational cost in practice.

Step 3. For each $i \in \mathcal{J}_n$, set

$$\Delta_{i-k_n+j}^n X^{*c} = \hat{c}_{n,i-}^{1/2} \Delta_{i-k_n+j}^n W^*, \quad \Delta_{i+k+j}^n X^{*c} = \hat{c}_{n,i+}^{1/2} \Delta_{i+k+j}^n W^*, \quad 0 \leq j \leq k_n - 1, \quad (2.17)$$

and compute $(\hat{c}_{n,i-}^*, \hat{c}_{n,i+}^*)$ as

$$\begin{cases} \hat{c}_{n,i-}^* \equiv \frac{1}{k_n \Delta_n} \sum_{j=0}^{k_n-1} (\Delta_{i-k_n+j}^n X^{*c}) (\Delta_{i-k_n+j}^n X^{*c})^\top, \\ \hat{c}_{n,i+}^* \equiv \frac{1}{k_n \Delta_n} \sum_{j=0}^{k_n-1} (\Delta_{i+k+j}^n X^{*c}) (\Delta_{i+k+j}^n X^{*c})^\top. \end{cases} \quad (2.18)$$

Step 4. Compute $\hat{\beta}_n^*(w)$ in the bootstrap sample using (2.13) with $(\Delta_{i,k}^n Y, \Delta_{i,k}^n Z, \hat{w}_{n,i})_{i \in \mathcal{J}_n}$ replaced by $(\Delta_{i,k}^n Y^*, \Delta_{i,k}^n Z^*, \hat{w}_{n,i}^*)_{i \in \mathcal{J}_n}$, where $\hat{w}_{n,i}^* \equiv w(\hat{c}_{n,i-}^*, \hat{c}_{n,i+}^*, \Delta_{i,k}^n Z^*, \hat{\beta}_n^*)$.

In summary, Algorithm 1 suggests computing $\hat{\beta}_n^*(w)$ in the same way as $\hat{\beta}_n(w)$ using the bootstrap sample. One exception is that the computation of the spot covariances (see (2.18)) does not require truncation, because we only use the diffusive returns in the bootstrap. It is important to observe that the spot covariance matrices and the weights are also resampled so as to capture their sampling variability in the higher-order asymptotics.

Theorem 3, below, describes the convergence in probability of the \mathcal{F} -conditional law of the bootstrap estimator $\hat{\beta}_n^*(w)$. For a sequence of random variables A_n , we write $A_n \xrightarrow{\mathcal{L}|\mathcal{F}} A$ if the \mathcal{F} -conditional law of A_n converges in probability to that of A under any metric for the weak convergence of probability measures.¹¹

Theorem 3. *Suppose the same conditions as in Theorem 2. Then we can decompose*

$$\Delta_n^{-1/2} \left(\hat{\beta}_n^*(w) - \hat{\beta}_n(w) \right) = \mathcal{L}_n^*(w) + k_n^{-1/2} \mathcal{H}_n^*(w) + o_p(k_n^{-1/2}), \quad (2.19)$$

¹¹We note that $A_n \xrightarrow{\mathcal{L}|\mathcal{F}} A$ amounts to the convergence of \mathcal{F} -conditional law in a weak sense, namely the convergence is in probability for measure-valued random elements. This convergence is weaker than the almost sure convergence of the \mathcal{F} -conditional law of A_n towards that of A , but is stronger than the stable convergence in law.

such that

$$(\mathcal{L}_n^*(w), \mathcal{H}_n^*(w)) \xrightarrow{\mathcal{L}|\mathcal{F}} (\mathcal{L}(w), \mathcal{H}(w)),$$

where $(\mathcal{L}(w), \mathcal{H}(w))$ are defined as in Theorem 2.

Theorem 3 justifies using the \mathcal{F} -conditional distribution of $\Delta_n^{-1/2}(\hat{\beta}_n^*(w) - \hat{\beta}_n(w))$ to approximate the \mathcal{F} -conditional limiting distribution of $\Delta_n^{-1/2}(\hat{\beta}_n(w) - \beta^*)$. Importantly, the approximation not only captures the leading term $\mathcal{L}(w)$, but also the higher-order term $k_n^{-1/2}\mathcal{H}(w)$.¹² We further note that both $\mathcal{L}(w)$ and $\mathcal{H}(w)$ are \mathcal{F} -conditionally symmetric. Therefore, the basic bootstrap and the percentile bootstrap (see, e.g., Davison and Hinkley (1997)) can both be used for constructing bootstrap confidence intervals.

Overall, refined inference for the jump regression coefficients can be done either by the use of the higher order asymptotic result in Theorem 2 or the bootstrap procedure based on the result of Theorem 3. We provide no asymptotic justification for preferring one over the other. In both methods the higher order asymptotic effect from the estimation of volatility around the jump times is accounted for. In addition, both methods ignore errors in the jump regression which are of even higher order (than the error due to the estimation of volatility), like the errors in detecting the locations of the jump times as well as the error due to the time variation in the volatility in the local blocks around the jump times. The difference between the inference based on the higher order asymptotic result and the bootstrap method is that the latter is based on mimicking the finite sample distribution of the regression estimator assuming jumps are located correctly and volatility does not vary over the local blocks. The inference based on the higher order asymptotic result, on the other

¹²It is useful to note that the spot volatility estimates $\hat{c}_{n,i\pm}$ in Algorithm 1 can be taken differently from those used in the estimation of $\hat{\beta}_n(w)$. In particular, if these spot volatility estimates attain the optimal $\Delta_n^{-1/4}$ rate, then it can be shown that the \mathcal{F} -conditional distribution $\mathcal{L}_n^*(w)$ converges to that of $\mathcal{L}(w)$ under the uniform metric with rate $\Delta_n^{-1/4}$.

hand, is based on asymptotic expansion of the regression estimator in the above simplified setting (i.e., when assuming jumps are located correctly and volatility is constant over the local windows around the jump times). In that sense, the difference between the two methods of refined inference for the jump regression is similar to the difference between inference based on asymptotic theory and bootstrap in classical settings, see e.g., Section 2 of Horowitz (2001).¹³ Finally, on the practical side the bootstrap is conceptually simple to grasp in the sense that the econometrician only needs to repeatedly compute the mixed-scale OLS or WLS estimator in the bootstrap samples.

2.3.3 Specification testing and its bootstrap implementation

We proceed a specification test for (2.4), which generalizes the test of Li, Todorov, and Tauchen (2016) into a multivariate mixed-scale setting. Since (2.4) is no longer assumed to be correct, we introduce the pseudo-true parameter

$$\bar{\beta} \equiv \left(\sum_{\tau \in \mathcal{T}} g(\Delta Z_{\tau}) g(\Delta Z_{\tau})^{\top} \right)^{-1} \left(\sum_{\tau \in \mathcal{T}} g(\Delta Z_{\tau}) \Delta Y_{\tau} \right).$$

Clearly, $\bar{\beta}$ coincides with β^* whenever (2.4) is correctly specified, but $\bar{\beta}$ remains well-defined even under misspecification. Formally, the testing problem is to decide in

¹³The type of refinement offered by the bootstrap is nevertheless nonstandard and theoretically interesting because our bootstrap is not applied to an asymptotically pivotal statistic, see Section 3.2 of Horowitz (2001) for a review of standard results on the asymptotic refinement of the bootstrap for asymptotically pivotal statistics. Instead, here, the refinement accounts for a higher-order sampling variability from the nonparametrically constructed weights (due to spot covariance estimation) that are used for efficient estimation.

which of the following two sets the observed sample path falls:¹⁴

$$\begin{cases} \Omega_0 \equiv \{\Delta Y_\tau = \bar{\beta}^\top g(\Delta Z_\tau) \text{ for all } \tau \in \mathcal{T}\} \cap \{|\mathcal{P}| > q\}, & \text{(Null Hypothesis)} \\ \Omega_a \equiv \{\Delta Y_\tau \neq \bar{\beta}^\top g(\Delta Z_\tau) \text{ some } \tau \in \mathcal{T}\} \cap \{|\mathcal{P}| > q\}, & \text{(Alternative Hypothesis)}. \end{cases} \quad (2.20)$$

We note that the event $\{|\mathcal{P}| > q\}$ rules out the degenerate situation where (2.4) holds trivially (recall that q is the dimension of $g(\cdot)$). Like in the classical setting, this condition says that β^* is overidentified, so that a specification test is possible.

We carry out the test by examining whether the sum of squared residuals (SSR) of a linear regression is “close enough” to zero. The SSR statistic is given by

$$SSR_n \equiv \sum_{i \in \mathcal{J}_n} \left(\Delta_{i,k}^n Y - g(\Delta_{i,k}^n Z)^\top \hat{\beta}_n \right)^2. \quad (2.21)$$

We reject the null hypothesis that (2.4) is correctly specified at significance level $\alpha \in (0, 1)$ if SSR_n is greater than a critical value cv_n^α that is described in Algorithm 2 below. In practice, it may be useful to report the test in terms of the R^2 of the regression (2.9), that is,

$$R_n^2 \equiv 1 - \frac{SSR_n}{\sum_{i \in \mathcal{J}_n} \Delta_{i,k}^n Y^2}.$$

We reject the null hypothesis when $1 - R_n^2$ is greater than $cv_n^\alpha / \sum_{i \in \mathcal{J}_n} (\Delta_{i,k}^n Y)^2$.

Algorithm 2 (Bootstrapping Critical Values for the Specification Test).

Step 1. Generate $(\Delta_{i,k}^n X^{*c})_{i \in \mathcal{J}_n}$ as in step 1 of Algorithm 1.

Step 2. Set $\Delta_{i,k}^n Z^* = \Delta_{i,k}^n Z + \Delta_{i,k}^n Z^{*c}$ and $\Delta_{i,k}^n Y^* = \hat{\beta}_n^\top g(\Delta_{i,k}^n Z) + \Delta_{i,k}^n Y^{*c}$ for $i \in \mathcal{J}_n$.

Step 3. Set cv_n^α to be the $(1 - \alpha)$ -quantile of SSR_n^* of the bootstrap sample, where SSR_n^* is the SSR obtained from regressing $\Delta_{i,k}^n Y^*$ on $g(\Delta_{i,k}^n Z^*)$. \square

¹⁴Specifying hypotheses in terms of random events is unlike the classical setting of hypothesis testing (e.g., Lehmann and Romano (2005)), but is standard in the study of high frequency data; see Jacod and Protter (2012), and references and discussions therein.

Theorem 4. *Under Assumptions 1 and 2, the following statements hold.*

(a) *In restriction to Ω_0 , $\Delta_n^{-1}SSR_n$ converges stably in law to*

$$\sum_{p \in \mathcal{P}} \varsigma_p^2 - \left(\sum_{p \in \mathcal{P}} g(\Delta Z_{\tau_p}) \varsigma_p \right)^\top \left(\sum_{p \in \mathcal{P}} g(\Delta Z_{\tau_p}) g(\Delta Z_{\tau_p})^\top \right)^{-1} \left(\sum_{p \in \mathcal{P}} g(\Delta Z_{\tau_p}) \varsigma_p \right).$$

In restriction to Ω_a , SSR_n converges in probability to

$$\sum_{p \in \mathcal{P}} \Delta Y_{\tau_p}^2 - \left(\sum_{p \in \mathcal{P}} g(\Delta Z_{\tau_p}) \Delta Y_{\tau_p} \right)^\top \left(\sum_{p \in \mathcal{P}} g(\Delta Z_{\tau_p}) g(\Delta Z_{\tau_p})^\top \right)^{-1} \left(\sum_{p \in \mathcal{P}} g(\Delta Z_{\tau_p}) \Delta Y_{\tau_p} \right).$$

(b) *The test associated with the critical region $\{SSR_n > cv_n^\alpha\}$ has asymptotic level α under the null hypothesis and asymptotic power one under the alternative hypothesis, that is,*

$$\mathbb{P}(SSR_n > cv_n^\alpha | \Omega_0) \rightarrow \alpha, \quad \mathbb{P}(SSR_n > cv_n^\alpha | \Omega_a) \rightarrow 1.$$

2.4 Simulation Results

We now examine the asymptotic theory above in simulations that mimic our empirical setting in Section 2.5. We set the sample span $T = 1$ year, or equivalently, 250 trading days. Each day contains $m = 400$ high-frequency returns, roughly corresponding to 1-minute sampling. Each Monte Carlo sample contains $n = 100,000$ returns, which are expressed in annualized percentage terms. We set the fine scale $\Delta_n = 1/n$ and implement the mixed-scale jump regression at the coarse scale $k\Delta_n$, for $k = 3, 5$ and 10. While our main focus is on results with mixed scales, we also report results for $k = 1$ as a benchmark. There are 2,000 Monte Carlo trials.

We consider a data generating process that allows for important features such

as leverage effect and price-volatility co-jumps. For independent Brownian motions $W_{1,t}$, $W_{2,t}$, $B_{1,t}$ and $B_{2,t}$, we set

$$\left\{ \begin{array}{l} d \log (V_{1,t}^*) = -\lambda_N \mu_F dt + 0.5 (dB_{1,t} + J_{V,t} dN_t), \quad V_{1,0}^* = \bar{V}_1, \\ \log (V_{2,t}^*) = \log (\bar{V}_2 - \beta_C^2 \bar{V}_1) + B_{2,t}, \\ V_{1,t} = TOD_t V_{1,t}^*, \quad V_{2,t} = TOD_t V_{2,t}^*, \\ dZ_t = \sqrt{V_{1,t}} (\rho dB_{1,t} + \sqrt{1 - \rho^2} dW_{1,t}) + \varphi_{Z,t} dN_t, \\ dY_t = \beta_C \sqrt{V_{1,t}} (\rho dB_{1,t} + \sqrt{1 - \rho^2} dW_{1,t}) + \sqrt{V_{2,t}} dW_{2,t} + \varphi_{Y,t} dN_t, \end{array} \right. \quad (2.22)$$

where TOD_t is a daily periodic function that captures the time-of-day effect in volatility.¹⁵ The jump regression relationship is given by

$$\varphi_{Y,t} = \beta_J \varphi_{Z,t}, \quad (2.23)$$

and the parameters are, in annualized terms,

$$\left\{ \begin{array}{l} \bar{V}_1 = 18^2, \quad \bar{V}_2 = 26^2, \quad \rho = -0.7, \quad \beta_C = 0.89, \quad \beta_J = 1, \\ J_{V,t} \stackrel{i.i.d.}{\sim} \text{Exponential}(\mu_F), \quad \mu_F = 0.1, \\ \varphi_{Z,t} | V_{1,t} \stackrel{i.i.d.}{\sim} \mathcal{N} \left(0, \frac{\phi^2 V_{1,t}}{n} \right), \quad \phi = 7.5, 10, \text{ or } 12.5, \\ N_t \text{ is a Poisson process with intensity } \lambda_N = 20. \end{array} \right. \quad (2.24)$$

These parameters are calibrated to match some key features of the data used in Section 2.5. In particular, the signal-to-noise ratio parameter ϕ controls the relative size of price jumps with respect to that of the (local) 1-minute diffusive returns, which is about 10.3 in our empirical sample. The jump intensity λ_N is calibrated so that the average number of detected jumps in the simulation is close to what we

¹⁵The time-of-day effect is calibrated using the data in our empirical study by averaging across days for each fixed sampling time within a day.

observe in the data, which is about 10.6 jumps per year. When $\phi = 7.5, 10$ and 12.5 , the average number of detected jumps in the simulation are 8.7, 11.9 and 14.2, respectively.

In order to examine the power of the specification test, we also implement the test under the following alternative model:

$$\varphi_{Y,t} = \beta_J \varphi_{Z,t} - \frac{\gamma}{\phi \sqrt{V_1/n}} \varphi_{Z,t}^2 1_{\{\varphi_{Z,t} < 0\}},$$

where the normalization via the average jump size $\phi \sqrt{V_1/n}$ makes the interpretation of the parameter γ comparable across simulations. We note that the correctly specified model (2.23) corresponds to $\gamma = 0$. We generate the misspecified model by setting $\gamma = 1$ or 2 .

Tuning parameters are chosen as follows. We set $k_n = 60$, corresponding to a one-hour local window for spot covariance estimation. For each trading day $t \in \{1, \dots, 250\}$, the truncation thresholds for Z are chosen adaptively as

$$u_{n,t} = 7m^{-0.49} \sqrt{BV_t}, \quad u'_{n,t} = 4m^{-0.49} \sqrt{BV_t}.$$

Here, BV_t is a slightly modified version of the Bipower Variation estimator of Barndorff-Nielsen and Shephard (2004c):

$$BV_t \equiv \frac{m}{m-4} \sum_i |\Delta_i^n Z| |\Delta_{i+1}^n Z|,$$

where the sum is over all returns in day t but with the largest 3 summand excluded.¹⁶

¹⁶In empirical applications, there may be large consecutive returns with similar magnitude but opposite signs (i.e., bouncebacks). The bipower variation estimator is sensitive to such issues. Removing the largest three summand is a simple but effective finite-sample robustification in this respect.

Table 2.1: Monte Carlo Rejection Rates of Specification Tests

		$\gamma = 0$			$\gamma = 1$			$\gamma = 2$		
		1%	5%	10%	1%	5%	10%	1%	5%	10%
$\phi = 7.5$	$k = 1$	1.2	4.3	9.8	90.1	93.5	95.2	96.6	97.7	98.5
	$k = 3$	1.5	5.3	9.7	75.9	83.5	87.0	92.3	94.5	95.8
	$k = 5$	1.3	5.0	10.2	64.3	74.6	79.8	87.9	91.5	93.9
	$k = 10$	1.3	5.1	9.7	46.1	59.9	67.3	75.3	82.8	87.2
$\phi = 10$	$k = 1$	0.9	5.1	10.0	96.0	97.9	98.6	99.1	99.4	99.4
	$k = 3$	1.5	5.5	9.8	86.9	91.8	93.7	97.5	98.4	98.8
	$k = 5$	1.3	5.9	10.9	80.1	87.0	89.8	94.8	96.9	97.7
	$k = 10$	1.5	6.3	10.4	64.1	75.3	80.3	87.1	92.3	94.3
$\phi = 12.5$	$k = 1$	1.3	5.5	10.4	98.2	99.1	99.3	99.2	99.4	99.6
	$k = 3$	1.4	5.6	10.4	93.3	95.5	96.8	97.8	98.3	98.8
	$k = 5$	1.6	5.7	10.1	87.6	92.1	94.3	96.4	97.5	98.0
	$k = 10$	1.4	5.8	10.6	74.2	83.6	86.7	91.6	94.4	96.1

Note: We report the Monte Carlo rejection rates of the specification test at significance level 1%, 5% and 10%. We report results under the null hypothesis ($\gamma = 0$) and the alternative hypothesis ($\gamma = 1, 2$) for various mixed scales ($k = 1, 3, 5$ and 10) as well as various relative jump sizes ($\phi = 7.5, 10$ and 12.5). The inference is based on 1000 bootstrap draws. Each experiment has 2000 Monte Carlo trials.

The truncation threshold for Y is computed similarly. Finally, we use the procedure detailed in the supplemental material of Todorov and Tauchen (2012) to adjust the time-of-day effect.

In Table 2.1, we report the finite-sample rejection rates of the specification test described in Theorem 4. Under the null hypothesis (i.e., $\gamma = 0$), we see that the rejection rates are fairly close to the nominal levels across various jump sizes and mixed scales. Under the alternative model (i.e., $\gamma = 1$ or 2), the rejection rates are well above the nominal level. Not surprisingly, the finite-sample power decreases as we use coarser scale (i.e. larger k), but it is interesting to note that the drop of power from $k = 1$ to $k = 3$ is not severe. As ϕ and γ increase, the rejection rates approach one.

Table 2.2: Summary of Estimation and Coverage Results

		Mixed-Scale OLS				Mixed-Scale WLS			
		RMSE	99% CI	95% CI	90% CI	RMSE	99% CI	95% CI	90% CI
$\phi = 7.5$	$k = 1$	0.063	98.5	93.7	88.1	0.057	98.9	94.4	88.7
	$k = 3$	0.111	98.2	92.9	88.4	0.101	98.3	92.7	88.8
	$k = 5$	0.143	98.6	94.6	88.7	0.129	98.5	94.0	88.7
	$k = 10$	0.194	98.5	94.9	89.3	0.182	98.6	93.9	88.9
$\phi = 10$	$k = 1$	0.045	98.7	94.1	88.3	0.039	98.5	94.6	89.2
	$k = 3$	0.075	98.7	94.2	88.8	0.065	98.5	94.6	89.5
	$k = 5$	0.097	98.0	93.5	88.6	0.084	99.0	95.0	88.4
	$k = 10$	0.131	98.9	95.1	91.0	0.116	98.9	95.6	90.4
$\phi = 12.5$	$k = 1$	0.034	98.8	94.6	88.6	0.029	98.9	94.8	89.3
	$k = 3$	0.062	98.7	93.0	87.9	0.051	98.7	94.8	89.2
	$k = 5$	0.080	98.2	94.0	88.2	0.067	98.6	94.0	89.0
	$k = 10$	0.110	98.5	93.0	88.2	0.095	98.8	93.6	88.3

Note: We report the root mean squared error (RMSE) and the Monte Carlo coverage rates of confidence intervals (CI) at levels 99%, 95% and 90%. We report results for various mixed scales ($k = 1, 3, 5$ and 10) and relative jump sizes ($\phi = 7.5, 10$ and 12.5) for both mixed-scale OLS and WLS estimators. The CIs are constructed using Algorithm 1 and the percentile bootstrap based on 1000 bootstrap draws. Each experiment has 2000 Monte Carlo trials.

In Table 2.2, we report some summary statistics for the mixed-scale OLS and WLS estimators of the jump beta. We see that the WLS estimator is always more accurate than the OLS estimator as measured by the root mean squared error (RMSE). Moreover, the coverage rates of confidence intervals (CI) constructed using Algorithm 1 and the percentile bootstrap are generally very close to the nominal levels, regardless of the sampling scale and the jump size. Coverage results based on the basic bootstrap are very similar to the percentile bootstrap and, hence, are omitted for brevity.

2.5 Empirical application

We use the developed tools to study the systematic jump risk in the stocks comprising the Dow Jones Industrial Average Index in December 2014, except Visa Inc. (V) is replaced by Bank of America (BAC) to make a balanced panel covering January 3, 2007 to December 12, 2014. The proxy for the market is the front-month E-mini S&P 500 index futures contract, which is among the most liquid instruments in the world.¹⁷ In some of our analysis we also make use of the ETFs on the nine industry portfolios comprising the S&P 500 index. We remove market holidays and half trading days. We also remove the two “Flash Crashes” (May 6, 2010 and April 23, 2013) because the dramatic market fluctuations in these days are known to be due to market malfunctioning. The resultant sample contains 1982 trading days. The intraday observations are sampled at 1-minute frequency from 9:35 to 15:55 EST; the prices at the first and the last 5 minutes are discarded to guard against possible adverse microstructure effects at market open and close. Finally, the truncation and the window size for the local volatility estimation are set as in the Monte Carlo, after adjusting for the intraday diurnal pattern of volatility.¹⁸ For this choice of the truncation (corresponding to a move slightly higher than 7 standard deviations), we detect a total of 85 market jumps in our sample.

To gauge the importance of microstructure noise that is weakly dependent in time, we compare the average value of realized volatility at 1-minute (our sampling frequency) and the coarser sampling frequency of 5-minutes. In presence of weakly dependent noise, the realized volatility should be higher for the higher sampling frequency due to this type of noise. For our data set, the median value of the ratio

¹⁷Hasbrouck (2003) estimates that 90% of U.S. equity price formation takes place in the E-mini market futures market.

¹⁸We use the procedure detailed in the supplemental material of Todorov and Tauchen (2012).

of 1-minute realized volatility over 5-minute realized volatility is 1.08. This indicates a relatively modest impact of the noise at the frequency we use here.

We start our empirical analysis with illustrating how stocks react to market jumps using four representative market jumps in our sample (two positive and two negative). On Figure 2.1, we plot the prices of the market and a set of selected stocks before and after the market jump event. The top left panel shows the behavior around the market jump on September 18, 2013 which was associated with the (positive) surprise by the Fed of not tapering its QE programs. In this case, both the market and the BA stock reacted within the same minute and fully adjusted to their new higher levels. A similar example, but in the opposite direction, is illustrated on the top right panel of the figure. This panel plots the market and AXP prices around the market jump on August 5, 2014. There were growing fears on this day associated with the impact of geopolitical risks on the economy along with concerns among investors that the Fed might raise interest rates sooner than expected in the wake of signs that the economy is gaining strength. In this example, like the previous one, both the market and the stock adjust to their lower level within the minute. Our third example of a market jump is of October 1, 2008 in the midst of the recent financial crisis. In this case, the CVX stock appeared to take more than one minute to fully incorporate the positive market jump, a seemingly delayed reaction which could be driven by market microstructure issues (e.g., stale limit orders). Another example of this type is the reaction of the WMT stock to the market jump on February 23, 2010 which is displayed on the bottom right panel of Figure 2.1. This market jump was associated with a surprisingly weak consumer confidence index reflecting the pessimism among investors for the strength of the economic recovery. While the market reacted within the minute of the release of the negative consumer confidence data, the WMT stock took at least 2 minutes to fully incorporate the bad news.

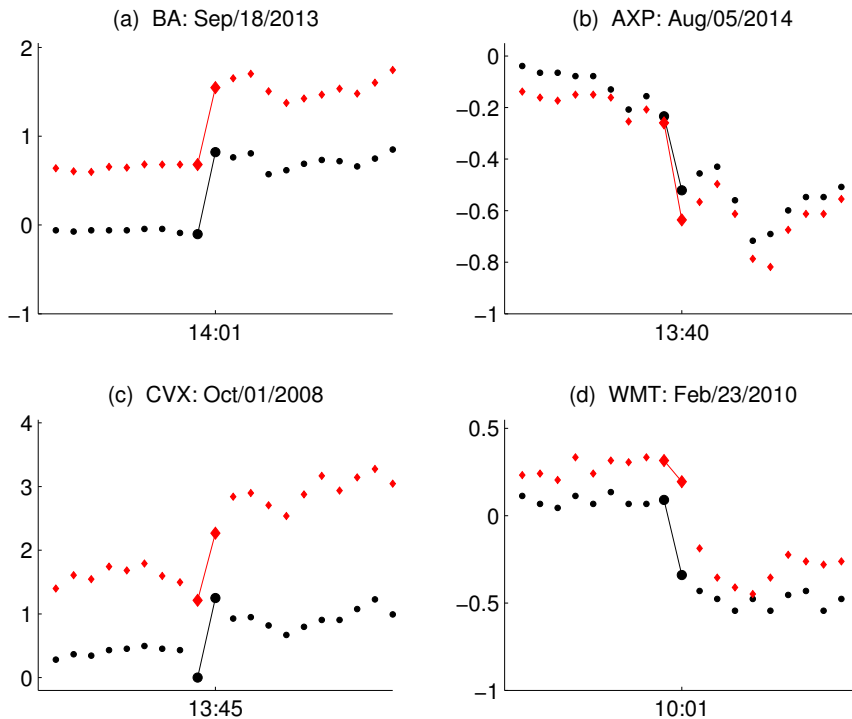


Figure 2.1: Market Jump Events and Stock Reaction.

Note: Circle and diamond dots correspond to cumulative (over the day) 1-minute log-returns on market and stock respectively. The connected dots correspond to the minute interval in which a market jump is detected.

Overall, the above four examples suggest that, in general, the stocks in our sample react quickly to the news triggering the market jumps. However, in some instances market microstructure related issues can confound the reaction of stocks to the market jumps. These issues, however, seem to be fairly short-lived. To verify that this is indeed the case, in Table 2.3, we report the jump beta estimates for all the stocks in the sample using aggregation of 3 and 5 minutes for the beta estimation (and using the whole sample). In the absence of confounding market microstructure effects, the two beta estimates should not be statistically different from each other. The results of the table show that this is largely the case. Indeed, the two beta estimates are fairly close with the median difference between the 5-minute and 3-minute estimates being only 0.01. The largest difference of 0.14 in our data set is for the DD stock,

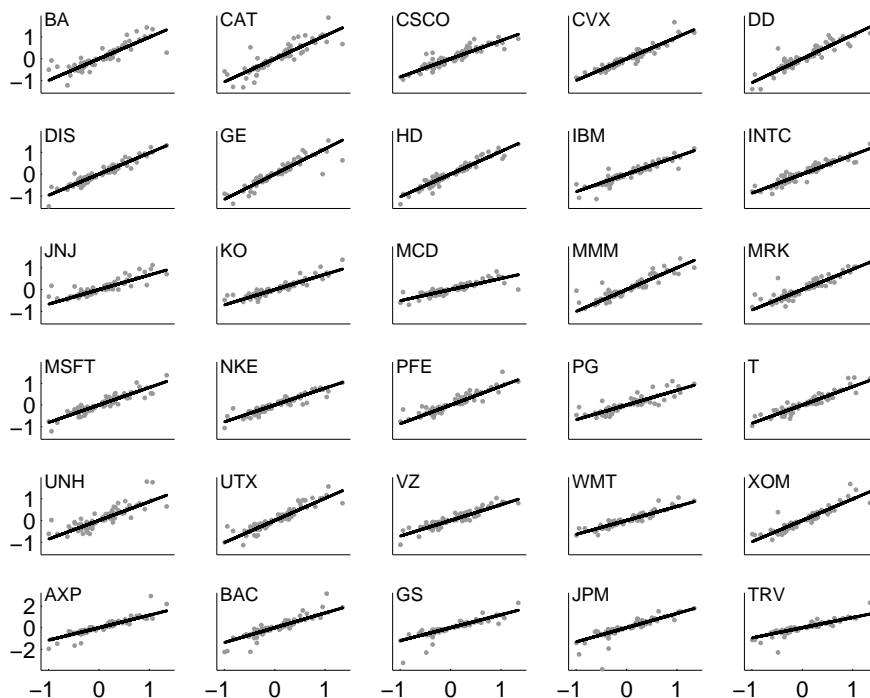


Figure 2.2: Scatter of Stock versus Market Returns at Market Jump Times of the Full Sample.

and this difference is only marginally statistically significant. Given this evidence, for the results that follow we will focus attention on the beta estimates based on three minute aggregation of returns following the market jump.

On Figure 2.2, we present scatter plots of stock jumps versus market jumps along with the fit implied by a constant market jump beta model for the whole sample. Overall, we see a very good fit. Most of the jump observations are fairly close to the fit implied by the constant jump beta model. Nevertheless, for some of the stocks, particularly those in the financial sector (bottom row), we see somewhat non-trivial deviations from the linear jump regression model. Of course, this can be merely due to the temporal variation in betas. In terms of levels, the jump betas of the stocks in the banking industry are systematically above one, while those of the consumer and healthcare sectors like MCD, WMT, JNJ and PG, are significantly below one.

Table 2.3: Full sample WLS beta estimates

Ticker	β	95% CI	β	95% CI
		$k = 3$		$k = 5$
AXP	1.15	[1.08; 1.20]	1.17	[1.09; 1.22]
BA	0.99	[0.92; 1.03]	1.02	[0.94; 1.07]
BAC	1.36	[1.27; 1.43]	1.36	[1.25; 1.44]
CAT	1.06	[0.99; 1.11]	1.08	[1.00; 1.13]
CSCO	0.84	[0.77; 0.90]	0.89	[0.82; 0.97]
CVX	0.99	[0.94; 1.03]	0.98	[0.92; 1.02]
DD	1.09	[1.03; 1.14]	1.23	[1.15; 1.27]
DIS	0.97	[0.92; 1.01]	0.98	[0.91; 1.02]
GE	1.16	[1.09; 1.21]	1.17	[1.08; 1.23]
GS	1.20	[1.12; 1.25]	1.21	[1.11; 1.27]
HD	1.05	[0.98; 1.09]	1.07	[0.99; 1.12]
IBM	0.81	[0.76; 0.84]	0.81	[0.75; 0.84]
INTC	0.88	[0.81; 0.94]	0.93	[0.85; 0.99]
JNJ	0.67	[0.62; 0.70]	0.67	[0.62; 0.70]
JPM	1.31	[1.24; 1.37]	1.29	[1.20; 1.34]
KO	0.70	[0.65; 0.74]	0.66	[0.60; 0.70]
MCD	0.51	[0.47; 0.54]	0.50	[0.46; 0.54]
MMM	1.00	[0.95; 1.03]	1.04	[0.97; 1.07]
MRK	0.94	[0.87; 0.97]	0.91	[0.84; 0.95]
MSFT	0.81	[0.75; 0.85]	0.81	[0.75; 0.87]
NKE	0.78	[0.72; 0.83]	0.82	[0.75; 0.87]
PFE	0.87	[0.80; 0.92]	0.88	[0.80; 0.94]
PG	0.68	[0.63; 0.71]	0.65	[0.59; 0.68]
T	0.84	[0.78; 0.88]	0.82	[0.75; 0.86]
TRV	0.94	[0.88; 0.97]	0.86	[0.79; 0.90]
UNH	0.86	[0.80; 0.91]	0.92	[0.84; 0.97]
UTX	1.02	[0.96; 1.05]	1.04	[0.97; 1.08]
VZ	0.72	[0.67; 0.76]	0.71	[0.65; 0.76]
WMT	0.64	[0.59; 0.67]	0.62	[0.57; 0.66]
XOM	0.98	[0.93; 1.02]	0.97	[0.91; 1.01]

Note: We report the efficient k -mixed-scale ($k = 3$ or 5) WLS estimates and their 95% confidence intervals (CI) of the 30 Dow stocks over the full sample. The CIs are from the percentile bootstrap using 1000 draws.

Given the overwhelming prior evidence on time variation in market betas, we next present results from testing for constancy of market jump betas over periods of years. The results are reported in the top panel of Table 2.4. We conduct the test over different aggregation frequencies ranging from one minute (no aggregation) to ten minutes. Naturally, more aggregation leads to diminishing power of detecting the variation in jump betas. This is consistent with our Monte Carlo results reported in the previous section. However, the drop of rejection rates going from one to three minutes evident from Table 2.4 is too big to be solely explained by the statistical effect of losing power when aggregating returns for the jump beta estimation. Instead, the relatively high rejection rates of the test for one minute aggregation are likely due to market microstructure effects like the ones illustrated on the bottom panels of Figure 2.1. At the three minute aggregation level, the rejection rates of the test are relatively low except for years 2007, 2008 and 2013. Some of these rejections can be still due to microstructure issues. However, some of the rejections probably reflect genuine variation of market jump betas, particularly during the period of the recent global financial crisis.

To further gauge the performance of the year-by-year linear jump regression model, the second panel of Table 2.4 reports the R^2 of the model fit at the market jump events. As seen from the table, the R^2 numbers are generally very high. For example, the time series average of R^2 at one- and three-minute aggregation are 0.93 and 0.89 respectively. As expected from theory, when increasing aggregation from one minute to ten minutes, the R^2 drops because the volatility of the diffusive aggregated increments around the jumps increases. Nevertheless, we see that with the exception of year 2008, the loss of R^2 going from one-minute to three-minute aggregation is quite moderate. Comparing the two panels of Table 2.4, we notice that there is no direct correspondence between the rejection rates and the magnitude

Table 2.4: Specification testing results for 30 DJIA stocks

	2007	2008	2009	2010	2011	2012	2013	2014
Cross-Sectional Rejection Rate								
$k = 1$	0.80	0.77	0.77	0.50	0.50	0.93	0.70	0.77
$k = 3$	0.50	0.43	0.07	0.17	0.10	0.40	0.33	0.17
$k = 5$	0.27	0.03	0.00	0.10	0.17	0.40	0.10	0.10
$k = 10$	0.17	0.10	0.10	0.00	0.00	0.17	0.07	0.03
Cross-Sectional Median of R^2								
$k = 1$	0.90	0.90	0.93	0.93	0.97	0.90	0.96	0.91
$k = 3$	0.86	0.80	0.87	0.92	0.95	0.88	0.95	0.86
$k = 5$	0.83	0.75	0.87	0.93	0.91	0.88	0.97	0.85
$k = 10$	0.82	0.81	0.93	0.86	0.87	0.82	0.97	0.81

Note: On the top panel, we report the cross-sectional rejection rate of the specification test at 1% significance level for the k -mixed samples year-by-year. On the bottom panel, we report the cross-sectional median of the R^2 s of the 30 stocks for the k -mixed samples.

of the R^2 of the linear jump regression. For example, focusing at the three-minute aggregation results, we can see that year 2007 is associated with the highest rejection rate of the linear jump regression model and yet has relatively high R^2 . On the other hand, year 2008 is associated with high rejection rate and is the lowest in the sample R^2 (using again the three-minute aggregation results). This difference can be explained with the different magnitude of the volatility around the jump event: it is relatively higher in 2008 than in 2007 and, as a result, the inference in the latter is sharper than the former.

To better assess the time variation in market jump betas, we plot next their time series (using yearly estimation intervals) on Figure 2.3. There are some clearly distinguishable time-series patterns evident from the figure. For example, the market jump betas of stocks in the financial sector, such as AXP, BAC and JPM, increase in the first two years in our sample and gradually decrease afterwards. On the other

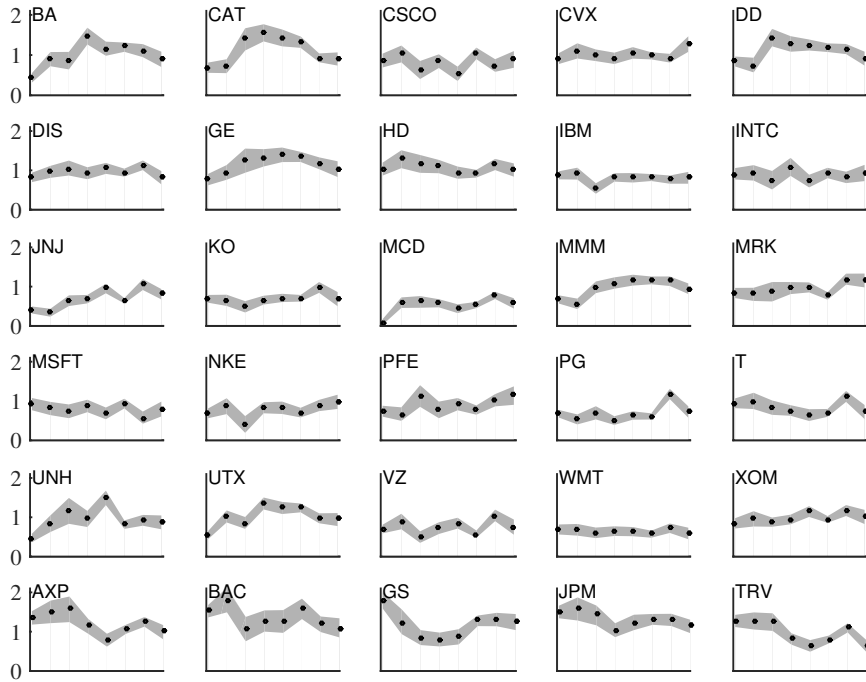


Figure 2.3: Time Series of Yearly Jump Betas, 2007–2014.

Note: The dots correspond to the yearly WLS beta estimates and the shaded areas correspond to the associated 95% confidence intervals.

hand, stocks such as INTC and WMT exhibit very little time variation.

The analysis so far has been based at the market jump times. We next investigate how stocks react to other systematic jump events. In particular, we focus attention on jump events in the nine industries comprising the S&P 500 index (our proxy for the market index) which are not detected as market jump events. In a market jump model in which the jumps in stocks are of two types, idiosyncratic and market, aggregate portfolios, such as the nine industry portfolios, should contain only jumps at the times when the market jump (as the idiosyncratic jumps get diversified away). However, some systematic jump events can have much bigger impact on a particular industry sector than on the market as a whole and, hence, the magnitude of an industry jump can be much bigger than that of the market co-jump. In such instances, given our discrete setting and high truncation level, we can fail to detect such jump events on

the market level but still find them in a particular industry sector portfolio. Hence, we label jump events in an industry sector, which are not detected as market jump times, as sector-specific jumps. These jumps have relatively much bigger importance for the particular sector than for the market.

Table 2.5: R^2 of the market factor for two types of jumps.

R^2 of Market-wide Jumps					
AXP	0.81	HD	0.92	NKE	0.84
BA	0.75	IBM	0.84	PFE	0.84
BAC	0.81	INTC	0.87	PG	0.74
CAT	0.77	JNJ	0.73	T	0.87
CSCO	0.82	JPM	0.70	TRV	0.78
CVX	0.90	KO	0.83	UNH	0.71
DD	0.84	MCD	0.72	UTX	0.85
DIS	0.89	MMM	0.81	VZ	0.85
GE	0.84	MRK	0.81	WMT	0.87
GS	0.72	MSFT	0.84	XOM	0.85

R^2 of Sector-specific Jumps					
AXP	0.45	HD	0.35	NKE	0.39
BA	0.33	IBM	0.22	PFE	0.46
BAC	0.75	INTC	0.48	PG	0.61
CAT	0.46	JNJ	0.50	T	0.53
CSCO	0.72	JPM	0.83	TRV	0.42
CVX	0.29	KO	0.31	UNH	0.21
DD	0.32	MCD	0.68	UTX	0.52
DIS	0.49	MMM	0.62	VZ	0.58
GE	0.46	MRK	0.32	WMT	0.29
GS	0.45	MSFT	0.78	XOM	0.33

To study the reaction of stocks to sector-specific jump events, we first associate with each of the stocks in our analysis the industry sector it belongs to.¹⁹ In Table 2.5 we report the R^2 for a linear jump regression model of the stock jump against the market jump at the sector-specific jump events for each of the stocks based on the

¹⁹The stocks in our study are all part of the S&P 500 index during the sample period. We, therefore, use the industry classification that is used to split the stocks in the S&P 500 index into nine industry portfolio ETFs.

whole sample. For comparison we also report the corresponding R^2 for the linear market jump model at the market jump times. The results present a rather mixed picture for the performance of the linear market jump model at the sector-specific jump events. For some stocks such as BAC, JPM, MCD and MSFT, the performance of the linear regression at the sector-specific jump events in terms of R^2 is comparable to its performance at the market jump events. However, for stocks like CVX, IBM, WMT and XOM, the R^2 of the regression at the sector-specific jump times is very low. Some of the loss of fit when comparing the performance of the linear jump model at market-wide jump events and sector-specific jump events can be due to the “signal” being smaller, that is, the market jump size at the sector-specific jump events being smaller in absolute value. This, however, cannot be the sole explanation, since as explained above, for some of the stocks in our sample the drop in R^2 is quite small. Another reason for the worsening fit at the sector-specific jump events can be due to larger “noise”, i.e., the diffusive volatility around the sector-specific jump events can be much bigger than around market-wide jump events for some of the stocks. Yet a third reason can be that the linear market jump model does not hold at the sector-specific jump events.

To get further insight in the performance of the linear market jump model at the sector-specific jump events, we display on Figure 2.4 scatter plots of the stock jumps against the market jumps at the sector-specific jump events for four representative (in terms of R^2) stocks. As seen from the figure, the performance of the model for IBM is very good with the observations being very close to the linear fit. On the other hand, for GE we notice that the jump observations are much more dispersed around the linear fit. This is suggestive of larger diffusive volatility around the sector-specific jump events for GE which consequently lowers the R^2 of the regression. Similar reasoning can explain the low R^2 for XOM. For this stock, however, we can also

notice a few outliers in the lower left corner of the plot which are indicative of model failure, i.e., that the market jumps cannot solely explain the XOM jumps at the sector-specific jump events. Finally, the fit for WMT is fairly poor with no strong association between the stock and market jumps at the sector-specific jump events. This is in sharp contrast with the performance of the linear jump market model for this stock at the market jump events.

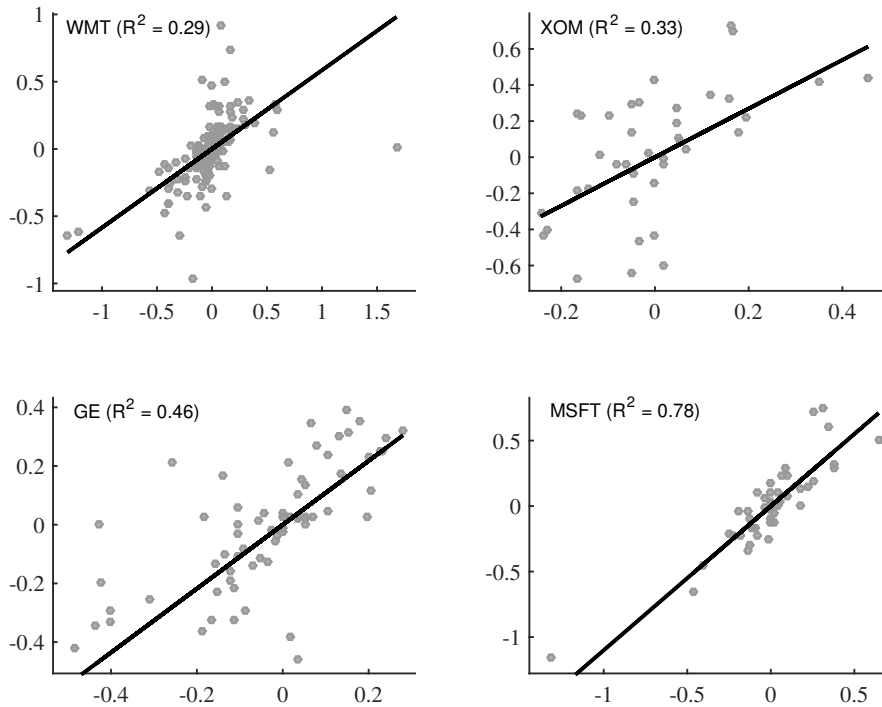


Figure 2.4: Scatter of Stock versus Market Returns at Sector-Specific Jump Times.

Overall, we can conclude that for some stocks the linear market jump model continues to work well at the sector-specific jump events. For many of the stocks, however, this is also associated with increased diffusive volatility around the sector-specific jump events which makes inference for the market jump beta at these events much noisier when compared with inference conducted at the market-wide jump events. Finally, for some of the stocks, the linear market jump model fails to account

for behavior of the stock market jumps at the sector-specific jump events and other factors are probably needed.

2.6 Conclusion

The dissertation touches upon two topics in volatility literature and makes contributions on each front. We propose a new mixed-scale jump regression framework for studying deterministic dependencies among jumps in a multivariate setting. A fine time scale is used to identify with high accuracy the times of large rare jumps in the explanatory variable(s). A coarser scale is then used to conduct the estimation in order to attenuate the effects of trading friction noise. We derive the asymptotic properties of an efficient estimator of the jump regression coefficients and a test for its specification. The limiting distributions of the estimator and the test statistic are non-standard, but a simple bootstrap method is shown to be valid for feasible inference. We further show that the bootstrap provides a higher-order refinement that accounts for the sampling variation in spot covariance estimates which are used to construct the efficient estimator. In a realistically calibrated Monte Carlo setting, which features leverage effects and price-volatility co-jumps, we report good size and power properties of the general specification test and good coverage properties of the confidence intervals.

The empirical application employs a 1-minute panel of Dow stock prices together with the front-month E-mini S&P 500 stock market index futures over the period 2007–2014. The 1-minute market index is used to locate jump times, and subsequent 3-minute sampling around the jump times is used to conduct the jump regression. We find a strong relationship between market jumps and stock price moves at market jump times. The market jump betas exhibit remarkable temporal stability and the

jump regressions have very high observed R^2 s. On the other hand, for many of the stocks in the sample, the relationship between stock and market jumps at sector-specific jump times is significantly noisier, and temporally more unstable, than the tight relationship seen at market jump times.

Chapter 3

Dynamic Semiparametric Models for ES (and VaR)

3.1 Introduction

The financial crisis of 2007-08 and its aftermath led to numerous changes in financial market regulation and banking supervision. One important change appears in the Third Basel Accord (Basel Committee, 2010), where new emphasis is placed on “Expected Shortfall” (ES) as a measure of risk, complementing, and in parts substituting, the more-familiar Value-at-Risk (VaR) measure. Expected Shortfall is the expected return on an asset conditional on the return being below a given quantile of its distribution, namely its VaR. That is, if Y_t is the return on some asset over some horizon (e.g., one day or one week) with conditional (on information set \mathcal{F}_{t-1}) distribution F_t , which we assume to be strictly increasing with finite mean, the α -level VaR and ES are:

$$\text{ES}_t = \mathbb{E}[Y_t | Y_t \leq \text{VaR}_t, \mathcal{F}_{t-1}] \quad (3.1)$$

$$\text{where } \text{VaR}_t = F_t^{-1}(\alpha), \text{ for } \alpha \in (0, 1) \quad (3.2)$$

$$\text{and } Y_t | \mathcal{F}_{t-1} \sim F_t \quad (3.3)$$

As Basel III is implemented worldwide (implementation is expected to occur in the period leading up to January 1st, 2019), ES will inevitably gain, and require, increasing attention from risk managers and banking supervisors and regulators. The

new “market discipline” aspects of Basel III mean that ES and VaR will be regularly disclosed by banks, and so a knowledge of these measures will also likely be of interest to these banks’ investors and counter-parties.

There is, however, a paucity of empirical models for expected shortfall. The large literature on volatility models (see Andersen et al. (2006) for a review) and VaR models (see Komunjer (2013) and McNeil et al. (2015)), have provided many useful models for these measures of risk. However, while ES has long been known to be a “coherent” measure of risk (Artzner et al. (1999)), in contrast with VaR, the literature contains relatively few models for ES; some exceptions are discussed below. This dearth is perhaps in part because regulatory interest in this risk measure is only recent, and may also be due to the fact that this measure is not “elicitable.” A risk measure (or statistical functional more generally) is said to be “elicitable” if there exists a loss function such that the risk measure is the solution to minimizing the expected loss. For example, the mean is elicitable using the quadratic loss function, and VaR is elicitable using the piecewise-linear or “tick” loss function. Having such a loss function is a stepping stone to building dynamic models for these quantities. We use recent results from Fissler and Ziegel (2016), who show that ES is *jointly elicitable* with VaR, to build new dynamic models for ES and VaR.

This chapter makes three main contributions. Firstly, we present some novel dynamic models for ES and VaR, drawing on the GAS framework of Creal et al. (2013), as well as successful models from the volatility literature, see Andersen et al. (2006). The models we propose are semiparametric in that they impose parametric structures for the dynamics of ES and VaR, but are completely agnostic about the conditional distribution of returns (aside from regularity conditions required for estimation and inference). The models proposed in this chapter are related to the class of “CAViaR” models proposed by Engle and Manganelli (2004), in that we directly parameterize

the measure(s) of risk that are of interest, and avoid the need to specify a conditional distribution for returns. The models we consider make estimation and prediction fast and simple to implement. Our semiparametric approach eliminates the need to specify and estimate a conditional density, thereby removing the possibility that such a model is misspecified, though at a cost of a loss of efficiency compared with a correctly specified density model.

Our second contribution is asymptotic theory for a general class of dynamic semiparametric models for ES and VaR. This theory is an extension of results for VaR presented in Weiss (1991) and Engle and Manganelli (2004), and draws on identification results in Fissler and Ziegel (2016) and results for M-estimators in Newey and McFadden (1994). We present conditions under which the estimated parameters of the VaR and ES models are consistent and asymptotically normal, and we present a consistent estimator of the asymptotic covariance matrix. We show via an extensive Monte Carlo study that the asymptotic results provide reasonable approximations in realistic simulation designs. In addition to being useful for the new models we propose, the asymptotic theory we present provides a general framework for other researchers to develop, estimate, and evaluate new models for VaR and ES.

Our third contribution is an extensive application of our new models and estimation methods in an out-of-sample analysis of forecasts of ES and VaR for four international equity indices over the period January 1990 to December 2016. We compare these new models with existing methods from the literature across a range of tail probability values (α) used in risk management. We use Diebold and Mariano (1995) tests to identify the best-performing models for ES and VaR, and we present simple regression-based methods, related to those of Engle and Manganelli (2004) and Nolde and Ziegel (2016), to “backtest” the ES forecasts.

Some work on expected shortfall estimation and prediction has appeared in the

literature, overcoming the problem of elicibility in different ways: Engle and Manganelli (2004b) discuss using extreme value theory, combined with GARCH or CAViaR dynamics, to obtain forecasts of ES. Cai and Wang (2008) propose estimating VaR and ES based on nonparametric conditional distributions, while Taylor (2008) and Gschöpf, Härdle, and Mihoci (2015) estimate models for “expectiles” (Newey and Powell (1987)) and map these to ES. Zhu and Galbraith (2011) propose using flexible parametric distributions for the standardized residuals from models for the conditional mean and variance. Drawing on Fissler and Ziegel (2016), we overcome the problem of elicibility more directly, and open up new directions for ES modeling and prediction.

In recent independent work, Taylor (2019) proposes using the asymmetric Laplace distribution to jointly estimate dynamic models for VaR and ES. He shows the intriguing result that the negative log-likelihood of this distribution corresponds to one of the loss functions presented in Fissler and Ziegel (2016), and thus can be used to estimate and evaluate such models. Unlike this chapter, Taylor (2019) provides no asymptotic theory for his proposed estimation method, nor any simulation studies of its reliability. However, given the link he presents, the theoretical results we present below can be used to justify *ex post* the methods of his paper.

The remainder of this chapter is structured as follows. In Section 3.2 we present new dynamic semiparametric models for ES and VaR and compare them with the main existing models for ES and VaR. In Section 3.3 we present asymptotic distribution theory for a generic dynamic semiparametric model for ES and VaR, and in Section 3.4 we study the finite-sample properties of the estimators in some realistic Monte Carlo designs. In Section 3.5 we apply the new models to daily data on four international equity indices, and compare these models both in-sample and out-of-sample with existing models. Section 3.6 concludes. Proofs and additional technical

details are presented in the appendix, and a supplemental appendix contains detailed proofs and additional analyses.

3.2 Dynamic models for ES and VaR

In this section we propose some new dynamic models for expected shortfall (ES) and Value-at-Risk (VaR). We do so by exploiting recent work in Fissler and Ziegel (2016) which shows that these variables are elicitable *jointly*, despite the fact that ES was known to be not elicitable on its own, see Gneiting (2011). The models we propose are based on the GAS framework of Creal et al. (2013) and Harvey (2013), which we briefly review in Section 3.2.2 below.

3.2.1 A consistent scoring rule for ES and VaR

Fissler and Ziegel (2016) show that the following class of loss functions (or “scoring rules”), indexed by the functions G_1 and G_2 , is consistent for VaR and ES. That is, minimizing the expected loss using any of these loss functions returns the true VaR and ES. In the functions below, we use the notation v and e for VaR and ES.

$$L_{FZ}(Y, v, e; \alpha, G_1, G_2) = (\mathbf{1}\{Y \leq v\} - \alpha) \left(G_1(v) - G_1(Y) + \frac{1}{\alpha} G_2(e) v \right) - G_2(e) \left(\frac{1}{\alpha} \mathbf{1}\{Y \leq v\} Y - e \right) - \mathcal{G}_2(e) \quad (3.4)$$

where G_1 is weakly increasing, G_2 is strictly increasing and strictly positive, and $\mathcal{G}'_2 = G_2$. We will refer to the above class as “FZ loss functions.”¹ Minimizing any

¹Consistency of the FZ loss function for VaR and ES also requires imposing that $e \leq v$, which follows naturally from the definitions of ES and VaR in equations (1) and (2). We discuss how we impose this restriction empirically in Sections 3.4 and 3.5 below.

member of this class yields VaR and ES:

$$(\text{VaR}_t, \text{ES}_t) = \arg \min_{(v,e)} \mathbb{E}_{t-1} [L_{FZ}(Y_t, v, e; \alpha, G_1, G_2)] \quad (3.5)$$

Using the FZ loss function for estimation and forecast evaluation requires choosing G_1 and G_2 . To do so, first define $\Delta L(Y_t, v_{1t}, e_{1t}, v_{2t}, e_{2t}) \equiv L(Y_t, v_{1t}, e_{1t}) - L(Y_t, v_{2t}, e_{2t})$ as the loss difference for two forecasts $(v_{j,t}, e_{j,t})$, $j \in \{1, 2\}$. We choose G_1 and G_2 so that the loss function generates ΔL that is homogeneous of degree zero, a property that has been shown in volatility forecasting applications to lead to higher power in Diebold-Mariano (1995) tests, see Patton and Sheppard (2009). Nolde and Ziegel (2016) show that there does *not* generally exist an FZ loss function that generates loss differences that are homogeneous of degree zero, however, we show in Proposition 1 below that zero-degree homogeneity may be attained by exploiting the fact that, for the values of α that are of interest in risk management applications (namely, values ranging from around 0.01 to 0.10), we may assume that $\text{ES}_t < 0$ a.s. $\forall t$. Proposition 1 shows that if we further impose that $\text{VaR}_t < 0$ a.s. $\forall t$, then, up to irrelevant location and scale factors, there is only *one* FZ loss function that generates loss differences that are homogeneous of degree zero.² The uniqueness of the loss function defined in Proposition 1 means, of course, that it also has the added benefit of there being no remaining shape or tuning parameters to be specified.

Proposition 1. *Define the FZ loss difference for two forecasts (v_{1t}, e_{1t}) and (v_{2t}, e_{2t}) as*

$$L_{FZ}(Y_t, v_{1t}, e_{1t}; \alpha, G_1, G_2) - L_{FZ}(Y_t, v_{2t}, e_{2t}; \alpha, G_1, G_2).$$

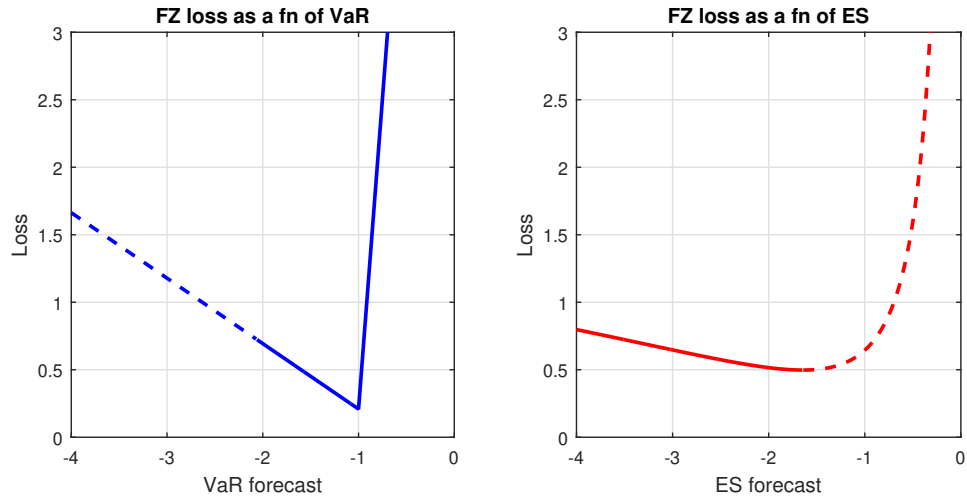
²If VaR can be positive, then there is one free shape parameter in the class of zero-homogeneous FZ loss functions (φ_1/φ_2 , in the notation of the proof of Proposition 1). In that case, our use of the loss function in equation (3.6) can be interpreted as setting that shape parameter to zero. This shape parameter does not affect the consistency of the loss function, as it is a member of the FZ class, but it may affect the ranking of misspecified models, see Patton (2016).

VaR and ES are both strictly negative, the loss differences generated by a FZ loss function are homogeneous of degree zero iff $G_1(x) = 0$ and $G_2(x) = -1/x$. The resulting “FZ0” loss function is:

$$L_{FZ0}(Y, v, e; \alpha) = -\frac{1}{\alpha e} \mathbf{1}\{Y \leq v\} (v - Y) + \frac{v}{e} + \log(-e) - 1 \quad (3.6)$$

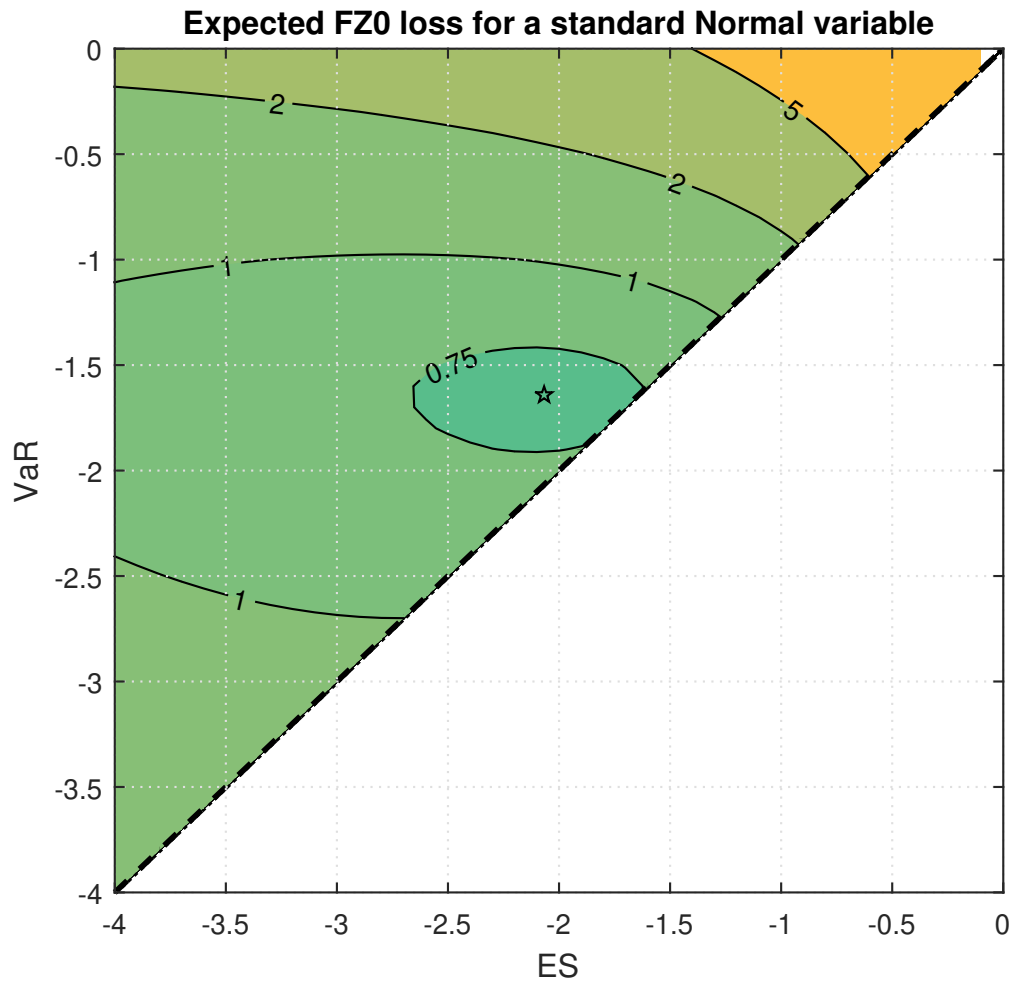
All proofs are presented in Appendix A.1. In Figure 3.1 we plot L_{FZ0} when $Y = -1$. In the left panel we fix $e = -2.06$ and vary v , and in the right panel we fix $v = -1.64$ and vary e . (These values for (v, e) are the $\alpha = 0.05$ VaR and ES from a standard Normal distribution.) The left panel shows that the implied VaR loss function resembles the “tick” loss function from quantile estimation, see Komunjer (2005) for example. In the right panel we see that the implied ES loss function resembles the “QLIKE” loss function from volatility forecasting, see Patton (2011) for example. In both panels, values of (v, e) where $v < e$ are presented with a dashed line, as by definition ES_t is below VaR_t , and so such values that would never be considered in practice. In Figure 3.2 we plot the contours of expected FZ0 loss for a standard Normal random variable. The minimum value, which is attained when $(v, e) = (-1.64, -2.06)$, is marked with a star, and we see that the “iso-expected loss” contours (that is, the level sets) of the expected loss function are boundaries of convex sets. Fissler (2017) shows that convexity of sublevel sets holds more generally for the FZ0 loss function under any distribution with finite first moments, unique α -quantiles, continuous densities, and negative ES.

Figure 3.1: Plot of FZ0 loss function



Note: This figure plots the FZ0 loss function when $Y = -1$ and $\alpha = 0.05$. In the left panel we fix $e = -2.06$ and vary v , in the right panel we fix $v = -1.64$ and vary e . Values where $v < e$ are indicated with a dashed line.

Figure 3.2: Contours of expected FZ0 loss when the target variable is standard Normal.



Note: Only values where $ES < VaR < 0$ are considered. The optimal value is marked with a star.

With the FZ0 loss function in hand, it is then possible to consider semiparametric dynamic models for ES and VaR:

$$(\text{VaR}_t, \text{ES}_t) = (v(\mathbf{Z}_{t-1}; \theta), e(\mathbf{Z}_{t-1}; \theta)) \quad (3.7)$$

that is, where the true VaR and ES are some specified parametric functions of elements of the information set, $\mathbf{Z}_{t-1} \in \mathcal{F}_{t-1}$. The parameters of this model are estimated via:

$$\hat{\theta}_T = \arg \min_{\theta} \frac{1}{T} \sum_{t=1}^T L_{FZ0}(Y_t, v(\mathbf{Z}_{t-1}; \theta), e(\mathbf{Z}_{t-1}; \theta); \alpha) \quad (3.8)$$

Such models impose a parametric structure on the dynamics of VaR and ES, through their relationship with lagged information, but require no assumptions, beyond regularity conditions, on the conditional distribution of returns. In this sense, these models are semiparametric. Using theory for M-estimators (see White (1994) and Newey and McFadden (1994) for example) we establish in Section 3.3 below the asymptotic properties of such estimators. Before doing so, we first consider some new dynamic specifications for ES and VaR.

3.2.2 A GAS model for ES and VaR

One of the challenges in specifying a dynamic model for a risk measure, or any other quantity of interest, is the mapping from lagged information to the current value of the variable. Our first proposed specification for ES and VaR draws on the work of Creal, *et al.* (2013) and Harvey (2013), who proposed a general class of models called “generalized autoregressive score” (GAS) models by the former authors, and “dynamic conditional score” models by the latter author. In both cases the models start from an assumption that the target variable has some parametric conditional

distribution, where the parameter (vector) of that distribution follows a GARCH-like equation. The forcing variable in the model is the lagged score of the log-likelihood, scaled by some positive definite matrix, a common choice for which is the inverse Hessian. This specification nests many well known models, including ARMA, GARCH (Bollerslev, 1986) and ACD (Engle and Russell, 1998) models. See Koopman *et al.* (2016) for an overview of GAS and related models.

We adopt this modeling approach and apply it to our M-estimation problem. In this application, the forcing variable is a function of the derivative and Hessian of the L_{FZ0} loss function rather than a log-likelihood. We will consider the following GAS(1,1) model for ES and VaR:

$$\begin{bmatrix} v_{t+1} \\ e_{t+1} \end{bmatrix} = \mathbf{w} + \mathbf{B} \begin{bmatrix} v_t \\ e_t \end{bmatrix} + \mathbf{A}\mathbf{H}_t^{-1}\nabla_t \quad (3.9)$$

where \mathbf{w} is a (2×1) vector and \mathbf{B} and \mathbf{A} are (2×2) matrices. The forcing variable in this specification is comprised of two components, \mathbf{H}_t and ∇_t . Using details provided in Appendix A.2.1, the latter can be shown to be:

$$\nabla_t \equiv \begin{bmatrix} \partial L_{FZ0}(Y_t, v_t, e_t; \alpha) / \partial v_t \\ \partial L_{FZ0}(Y_t, v_t, e_t; \alpha) / \partial e_t \end{bmatrix} = \begin{bmatrix} \frac{1}{\alpha v_t e_t} \lambda_{v,t} \\ \frac{-1}{\alpha e_t^2} (\lambda_{v,t} + \alpha \lambda_{e,t}) \end{bmatrix} \quad (3.10)$$

$$\text{where } \lambda_{v,t} \equiv -v_t (\mathbf{1}\{Y_t \leq v_t\} - \alpha) \quad (3.11)$$

$$\lambda_{e,t} \equiv \frac{1}{\alpha} \mathbf{1}\{Y_t \leq v_t\} Y_t - e_t \quad (3.12)$$

Note that the expression given for $\partial L_{FZ0} / \partial v_t$ only holds for $Y_t \neq v_t$. As we assume that Y_t is continuously distributed, this holds with probability one. The scaling

matrix, \mathbf{H}_t , is related to the Hessian:

$$\mathbf{I}_t \equiv \begin{bmatrix} \frac{\partial^2 \mathbb{E}_{t-1}[L_{FZ0}(Y_t, v_t, e_t; \alpha)]}{\partial v_t^2} & \frac{\partial^2 \mathbb{E}_{t-1}[L_{FZ0}(Y_t, v_t, e_t; \alpha)]}{\partial v_t \partial e_t} \\ \bullet & \frac{\partial^2 \mathbb{E}_{t-1}[L_{FZ0}(Y_t, v_t, e_t; \alpha)]}{\partial e_t^2} \end{bmatrix} = \begin{bmatrix} -\frac{f_t(v_t)}{\alpha e_t} & 0 \\ 0 & \frac{1}{e_t^2} \end{bmatrix} \quad (3.13)$$

The second equality above exploits the fact that $\partial^2 \mathbb{E}_{t-1}[L_{FZ0}(Y_t, v_t, e_t; \alpha)] / \partial v_t \partial e_t = 0$ under the assumption that the dynamics for VaR and ES are correctly specified. The first element of the matrix \mathbf{I}_t depends on the unknown conditional density of Y_t . We would like to avoid estimating this density, and we approximate the term $f_t(v_t)$ as being proportional to v_t^{-1} . This approximation holds exactly if Y_t is a zero-mean location-scale random variable, $Y_t = \sigma_t \eta_t$, where $\eta_t \sim iid F_\eta(0, 1)$, as in that case we have:

$$f_t(v_t) = f_t(\sigma_t v_\alpha) = \frac{1}{\sigma_t} f_\eta(v_\alpha) \equiv k_\alpha \frac{1}{v_t} \quad (3.14)$$

where $k_\alpha \equiv v_\alpha f_\eta(v_\alpha)$ is a constant with the same sign as v_t . We define \mathbf{H}_t to equal \mathbf{I}_t with the first element replaced using the approximation in the above equation.³ The forcing variable in our GAS model for VaR and ES then becomes:

$$\mathbf{H}_t^{-1} \nabla_t = \begin{bmatrix} \frac{-1}{k_\alpha} \lambda_{v,t} \\ \frac{-1}{\alpha} (\lambda_{v,t} + \alpha \lambda_{e,t}) \end{bmatrix} \quad (3.15)$$

Notice that the second term in the model is a linear combination of the two elements of the forcing variable, and since the forcing variable is premultiplied by a coefficient

³Note that we do *not* use the fact that the scaling matrix is exactly the inverse Hessian (e.g., by invoking the information matrix equality) in our empirical application or our theoretical analysis. Also, note that if we considered a value of α for which $v_t = 0$, then $v_\alpha = 0$ and we cannot justify our approximation using this approach. However, we focus on cases where $\alpha \ll 1/2$, and so we are comfortable assuming $v_t \neq 0$, making k_α invertible.

matrix, say $\tilde{\mathbf{A}}$, we can equivalently use

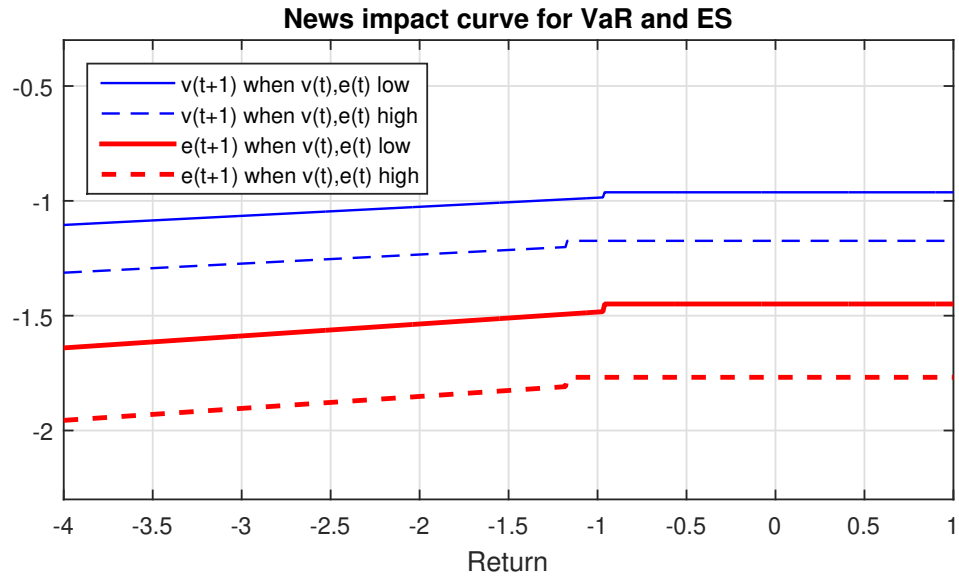
$$\tilde{\mathbf{A}}\mathbf{H}_t^{-1}\nabla_t = \mathbf{A}\lambda_t \quad (3.16)$$

$$\text{where } \lambda_t \equiv [\lambda_{v,t}, \lambda_{e,t}]'$$

We choose to work with the $\mathbf{A}\lambda_t$ parameterization, as the two elements of this forcing variable $(\lambda_{v,t}, \lambda_{e,t})$ are not directly correlated, while the elements of $\mathbf{H}_t^{-1}\nabla_t$ are correlated due to the overlapping term $(\lambda_{v,t})$ appearing in both elements. This aids the interpretation of the results of the model without changing its fit.

To gain some intuition for how past returns affect current forecasts of ES and VaR in this model, consider the “news impact curve” of this model, which presents (v_{t+1}, e_{t+1}) as a function of Y_t through its impact on $\lambda_t \equiv [\lambda_{v,t}, \lambda_{e,t}]'$, holding all other variables constant. Figure 3.3 shows these two curves for $\alpha = 0.05$, using the estimated parameters for this model when applied to daily returns on the S&P 500 index (details are presented in Section 3.5 below). We consider two values for the “current” value of (v, e) : 10% above and below the long-run average for these variables. We see that for values where $Y_t > v_t$, the news impact curves are flat, reflecting the fact that on those days the value of the realized return does not enter the forcing variable. When $Y_t \leq v_t$, we see that ES and VaR react linearly to Y and this reaction is through the $\lambda_{e,t}$ forcing variable; the reaction through the $\lambda_{v,t}$ forcing variable is a simple step (down) in both of these risk measures.

Figure 3.3: Plot of values of VaR and ES as a function of the lagged return.



Note: This figure shows the values of VaR and ES as a function of the lagged return, when the lagged values of VaR and ES are either low (10% below average) or high (10% above average). The function is based on the estimated parameters for daily S&P 500 returns.

3.2.3 A one-factor GAS model for ES and VaR

The specification in Section 3.2.2 allows ES and VaR to evolve as two separate, correlated, processes. In many risk forecasting applications, a useful simpler model is one based on a structure with only one time-varying risk measure, e.g. volatility. We will consider a one-factor model in this section, and will name the model in Section 3.2.2 a “two-factor” GAS model.

Consider the following one-factor GAS model for ES and VaR, where both risk measures are driven by a single variable, κ_t .⁴

$$v_t = a \exp \{ \kappa_t \} \quad (3.17)$$

$$e_t = b \exp \{ \kappa_t \}, \text{ where } b < a < 0$$

$$\text{and } \kappa_t = \omega + \beta \kappa_{t-1} + \gamma H_{t-1}^{-1} s_{t-1}$$

The forcing variable, $H_{t-1}^{-1} s_{t-1}$, in the evolution equation for κ_t is obtained from the FZ0 loss function, plugging in $(a \exp \{ \kappa_t \}, b \exp \{ \kappa_t \})$ for (v_t, e_t) . Using details provided in Appendix A.2.2, we find that the score and Hessian are:

$$\begin{aligned} s_t &\equiv \frac{\partial L_{FZ0}(Y_t, a \exp \{ \kappa_t \}, b \exp \{ \kappa_t \}; \alpha)}{\partial \kappa} = -\frac{1}{e_t} \left(\frac{1}{\alpha} \mathbf{1} \{ Y_t \leq v_t \} Y_t - \frac{1}{\alpha} \right) \\ \text{and } I_t &\equiv \frac{\partial^2 \mathbb{E}_{t-1} [L_{FZ0}(Y_t, a \exp \{ \kappa_t \}, b \exp \{ \kappa_t \}; \alpha)]}{\partial \kappa_t^2} = \frac{\alpha - k_\alpha a_\alpha}{\alpha} \end{aligned} \quad (3.19)$$

where k_α is a negative constant and a_α lies between zero and one. The Hessian, I_t , turns out to be a constant in this case, and since we estimate a free coefficient on our

⁴We use the structure in equation (3.17) to emphasize its similarity to conditional volatility models, which we include as competitor models in the next section. The one-factor model for ES and VaR can also be obtained by considering a zero-mean volatility model for Y_t , with *iid* standardized residuals, say denoted η_t . In this case, κ_t is the log conditional standard deviation of Y_t , and $a = F_\eta^{-1}(\alpha)$ and $b = \mathbb{E}[\eta | \eta \leq a]$. (We exploit this interpretation when linking these models to GARCH models in Section 3.2.5 below.)

forcing variable, we can set the scaling matrix, H_t , to any positive constant; we set H_t to one. Note that the VaR score, $\lambda_{v,t} = \partial L / \partial v$, turns out to drop out from the forcing variable. Thus the one-factor GAS model for ES and VaR becomes:

$$\kappa_t = \omega + \beta \kappa_{t-1} + \gamma \frac{1}{b \exp \{\kappa_{t-1}\}} \left(\frac{1}{\alpha} \mathbf{1} \{Y_{t-1} \leq a \exp \{\kappa_{t-1}\}\} Y_{t-1} - b \exp \{\kappa_{t-1}\} \right) \quad (3.20)$$

We drop the negative sign in s_t that its coefficient, γ , is positive rather than negative. This change, of course, does not affect the fit of the model. The FZ loss function only identifies (v_t, e_t) , and in the specification in equation (3.17) this implies that ω , a , and b are not separably identifiable: for any constant c , the parameter vectors $(\omega, a, b, \beta, \gamma)$ and $(\omega + c(1 - \beta), a \exp \{-c\}, b \exp \{-c\}, \beta, \gamma)$ yield identical sequences of (v_t, e_t) , and thus identical values of the objective function. Fixing any one of ω , a , or b resolves this problem; we set $\omega = 0$ for simplicity.

Foreshadowing the empirical results in Section 3.5, we find that this one-factor GAS model outperforms the two-factor GAS model in out-of-sample forecasts for most of the asset return series that we study.

3.2.4 Existing dynamic models for ES and VaR

As noted in the introduction, there is a relative paucity of dynamic models for ES and VaR, but there is not a complete absence of such models. The simplest existing model is based on a rolling window estimate of these quantities:

$$\begin{aligned} \widehat{\text{VaR}}_t &= \widehat{\text{Quantile}} \{Y_s\}_{s=t-m}^{t-1} \\ \widehat{\text{ES}}_t &= \frac{1}{\alpha m} \sum_{s=t-m}^{t-1} Y_s \mathbf{1} \{Y_s \leq \widehat{\text{VaR}}_s\} \end{aligned} \quad (3.21)$$

where $\widehat{Quantile} \{Y_s\}_{s=t-m}^{t-1}$ denotes the sample quantile of Y_s over the period $s \in [t-m, t-1]$. Common choices for the window size, m , include 125, 250 and 500, corresponding to six months, one year and two years of daily return observations respectively.

A more challenging competitor for the new ES and VaR models proposed in this chapter are those based on ARMA-GARCH dynamics for the conditional mean and variance, accompanied by some assumption for the distribution of the standardized residuals. These models all take the form:

$$\begin{aligned} Y_t &= \mu_t + \sigma_t \eta_t \\ \eta_t &\sim iid F_\eta(0, 1) \end{aligned} \tag{3.22}$$

where μ_t and σ_t^2 are specified to follow some ARMA and GARCH model, and $F_\eta(0, 1)$ is some arbitrary, strictly increasing, distribution with mean zero and variance one. What remains is to specify a distribution for the standardized residual, η_t . Given a choice for F_η , VaR and ES forecasts are obtained as:

$$\begin{aligned} v_t &= \mu_t + a\sigma_t, \quad \text{where } a = F_\eta^{-1}(\alpha) \\ e_t &= \mu_t + b\sigma_t, \quad \text{where } b = \mathbb{E}[\eta_t | \eta_t \leq a] \end{aligned} \tag{3.23}$$

Two parametric choices for F_η are common in the literature:

$$\begin{aligned} \eta_t &\sim iid N(0, 1) \\ \eta_t &\sim iid Skew t(0, 1, \nu, \lambda) \end{aligned} \tag{3.24}$$

There are various skew t distributions used in the literature; in the empirical analysis

below we use that of Hansen (1994). A nonparametric alternative is to estimate the distribution of η_t using the empirical distribution function (EDF), an approach that is also known as “filtered historical simulation,” and one that is perhaps the best existing model for ES, see the survey by Engle and Manganelli (2004b).⁵ We consider all of these models in our empirical analysis in Section 3.5.

3.2.5 GARCH and ES/VaR estimation

In this section we consider two extensions of the models presented above, in an attempt to combine the success and parsimony of GARCH models with this chapter’s focus on ES and VaR forecasting.

Estimating a GARCH model via FZ minimization

If an ARMA-GARCH model, including the specification for the distribution of standardized residuals, is correctly specified for the conditional distribution of an asset return, then maximum likelihood is the most efficient estimation method, and should naturally be adopted. If, on the other hand, we consider an ARMA-GARCH model only as a useful approximation to the true conditional distribution, then it is no longer clear that MLE is optimal. In particular, if the application of the model is to ES and VaR forecasting, then we might be able to improve the fitted ARMA-GARCH model by estimating the parameters of that model via FZ loss minimization, as discussed in Section 3.2.1. This estimation method is related to one discussed in Remark 1 of Francq and Zakoïan (2015).

⁵Some authors have also considered modeling the tail of F_η using extreme value theory, however for the relatively non-extreme values of α we consider here, past work (e.g., Engle and Manganelli (2004b), Nolde and Ziegel (2016) and Taylor (2017)) has found EVT to perform no better than the EDF, and so we do not include it in our analysis.

Consider the following model for asset returns:

$$\begin{aligned} Y_t &= \sigma_t \eta_t, \quad \eta_t \sim iid F_\eta(0, 1) \\ \sigma_t^2 &= \omega + \beta \sigma_{t-1}^2 + \gamma Y_{t-1}^2 \end{aligned} \tag{3.25}$$

The variable σ_t^2 is the conditional variance and is assumed to follow a GARCH(1,1) process. This model implies a structure analogous to the one-factor GAS model presented in Section 3.2.3, as we find:

$$\begin{aligned} v_t &= a \cdot \sigma_t, \quad \text{where } a = F_\eta^{-1}(\alpha) \\ e_t &= b \cdot \sigma_t, \quad \text{where } b = \mathbb{E}[\eta | \eta \leq a] \end{aligned} \tag{3.26}$$

Some further results on VaR and ES in dynamic location-scale models are presented in Appendix A.2.3. To apply this model to VaR and ES forecasting, we also have to estimate the VaR and ES of the standardized residual, denoted (a, b) . Rather than estimating the parameters of this model using (Q)MLE, we consider here estimating via FZ loss minimization. As in the one-factor GAS model, ω is unidentified and we set it to one,⁶ so the parameter vector to be estimated is (β, γ, a, b) . This estimation approach leads to a fitted GARCH model that is tailored to provide the best-fitting ES and VaR forecasts, rather than the best-fitting volatility forecasts.

⁶Similar to the one-factor GAS model, in this case we find that for any strictly positive constant c , the parameter vectors $(\omega, a, b, \beta, \gamma)$ and $(c\omega, a/\sqrt{c}, b/\sqrt{c}, \beta, c\gamma)$ yield identical sequences of (v_t, e_t) , and thus identical values of the objective function. Fixing any one of ω , a , or b resolves this problem. As ω must be strictly positive in a GARCH model, we cannot set it to zero as we did for the one-factor GAS model; instead we set it to one.

A hybrid GAS/GARCH model

Finally, we consider a direct combination of the forcing variable suggested by a GAS structure for a one-factor model of returns, described in equation (3.20), with the successful GARCH model for volatility. We specify:

$$\begin{aligned} Y_t &= \exp\{\kappa_t\} \eta_t, \quad \eta_t \sim iid F_\eta(0, 1) \\ \kappa_t &= \omega + \beta\kappa_{t-1} + \gamma \frac{1}{e_{t-1}} \left(\frac{1}{\alpha} \mathbf{1}\{Y_{t-1} \leq v_{t-1}\} Y_{t-1} - e_{t-1} \right) + \delta \log |Y_{t-1}| \end{aligned} \quad (3.27)$$

The variable κ_t is the log-volatility, identified up to scale. As the latent variable in this model is log-volatility, we use the lagged log absolute return rather than the lagged squared return, so that the units remain in line for the evolution equation for κ_t . There are five parameters in this model $(\beta, \gamma, \delta, a, b)$, and we estimate them using FZ loss minimization.

3.3 Estimation of dynamic models for ES and VaR

This section presents asymptotic theory for the estimation of dynamic ES and VaR models by minimizing FZ loss. Given a sample of observations (Y_1, \dots, Y_T) and a constant $\alpha \in (0, 0.5)$, we are interested in estimating and forecasting the conditional α quantile (VaR) and corresponding expected shortfall (ES) of Y_t . Suppose Y_t is a real-valued random variable that has, conditional on information set \mathcal{F}_{t-1} , distribution function $F_t(\cdot | \mathcal{F}_{t-1})$ and corresponding density function $f_t(\cdot | \mathcal{F}_{t-1})$. Let $v_1(\theta^0)$ and $e_1(\theta^0)$ be some initial conditions for VaR and ES and let $\mathcal{F}_{t-1} = \sigma\{Y_{t-1}, \mathbf{X}_{t-1}, \dots, Y_1, \mathbf{X}_1\}$, where \mathbf{X}_t is a vector of exogenous variables or predetermined variables, be the information set available for forecasting Y_t . The vector of unknown parameters to be estimated is $\theta^0 \in \Theta \subset \mathbb{R}^p$.

The conditional VaR and ES of Y_t at probability level α , that is $\text{VaR}_\alpha(Y_t|\mathcal{F}_{t-1})$ and $\text{ES}_\alpha(Y_t|\mathcal{F}_{t-1})$, are assumed to follow some dynamic model:

$$\begin{bmatrix} \text{VaR}_\alpha(Y_t|\mathcal{F}_{t-1}) \\ \text{ES}_\alpha(Y_t|\mathcal{F}_{t-1}) \end{bmatrix} = \begin{bmatrix} v(Y_{t-1}, \mathbf{X}_{t-1}, \dots, Y_1, \mathbf{X}_1; \theta^0) \\ e(Y_{t-1}, \mathbf{X}_{t-1}, \dots, Y_1, \mathbf{X}_1; \theta^0) \end{bmatrix} \equiv \begin{bmatrix} v_t(\theta^0) \\ e_t(\theta^0) \end{bmatrix}, \quad t = 1, \dots, T. \quad (3.28)$$

The unknown parameters are estimated as:

$$\hat{\theta}_T \equiv \arg \min_{\theta \in \Theta} L_T(\theta) \quad (3.29)$$

where $L_T(\theta) = \frac{1}{T} \sum_{t=1}^T L_{FZ0}(Y_t, v_t(\theta), e_t(\theta); \alpha)$

and the FZ loss function L_{FZ0} is defined in equation (3.6). Below we provide conditions under which estimation of these parameters via FZ loss minimization leads to a consistent and asymptotically normal estimator, with standard errors that can be consistently estimated. In Supplemental Appendix, we show that all of these conditions are satisfied for the widely-used GARCH(1,1) model, drawing on Lumsdaine (1996) and Carrasco and Chen (2002) among others. See Francq and Zakoian (2010) for a review of asymptotic theory for GARCH processes.

Assumption 4. (A) $L(Y_t, v_t(\theta), e_t(\theta); \alpha)$ obeys the uniform law of large numbers.

(B)(i) Θ is a compact subset of \mathbb{R}^p for $p < \infty$. (ii) $\{Y_t\}_{t=1}^\infty$ is a strictly stationary process. Conditional on all the past information \mathcal{F}_{t-1} , the distribution of Y_t is $F_t(\cdot|\mathcal{F}_{t-1})$ which, for all t , belongs to a class of distribution functions on \mathbb{R} with finite first moments and unique α -quantiles. (iii) $\forall t$, both $v_t(\theta)$ and $e_t(\theta)$ are \mathcal{F}_{t-1} -measurable and a.s. continuous in θ . (iv) If $\Pr[v_t(\theta) = v_t(\theta^0) \cap e_t(\theta) = e_t(\theta^0)] = 1 \quad \forall t$, then $\theta = \theta^0$.

Theorem 5 (Consistency). Under Assumption 4, $\hat{\theta}_T \xrightarrow{p} \theta^0$ as $T \rightarrow \infty$.

The proof of Theorem 5, provided in Appendix, is straightforward given Theorem 2.1 of Newey and McFadden (1994) and Corollary 5.5 of Fissler and Ziegel (2016). Assumption 4(A) can be satisfied by one of a variety of uniform laws of large numbers for the time series applications we consider here, see Andrews (1987) and Pötscher and Prucha (1989) for example. Assumption 4(B) is standard for parameter time series inference. Zwingmann and Holzmann (2016) show that if the α -quantile is not unique (violating part of our Assumption 4(B)(ii)), then the convergence rate and asymptotic distribution of (\hat{v}_T, \hat{e}_T) are non-standard, even in a setting with *iid* data. We do not consider such problematic cases here.

We next turn to the asymptotic distribution of our parameter estimator. In the assumptions below, K denotes a finite constant that can change from line to line, and we use $\|\mathbf{x}\|$ to denote the Euclidean norm of if \mathbf{x} is a vector, and the Frobenius norm if \mathbf{x} is a matrix.

Assumption 5. (A) For all t , we have (i) $v_t(\theta)$ and $e_t(\theta)$ are a.s. twice continuously differentiable in θ , (ii) $e_t(\theta^0) < v_t(\theta^0) \leq 0$.

(B) For all t , we have (i) conditional on all the past information \mathcal{F}_{t-1} , Y_t has a continuous density $f_t(\cdot|\mathcal{F}_{t-1})$ that satisfies $f_t(y|\mathcal{F}_{t-1}) \leq K < \infty$ and $|f_t(y'|\mathcal{F}_{t-1}) - f_t(y''|\mathcal{F}_{t-1})| \leq K|y' - y''|$, (ii) $\mathbb{E} \left[|Y_t|^{4+\delta} \right] \leq K < \infty$, for some $0 < \delta < 1$.

(C) There exists a neighborhood of θ^0 , $\mathcal{N}(\theta^0)$, such that for all t we have (i) $|1/e_t(\theta)| \leq K < \infty$, $\forall \theta \in \mathcal{N}(\theta^0)$, (ii) there exist some (possibly stochastic) \mathcal{F}_{t-1} -measurable functions $V(\mathcal{F}_{t-1})$, $V_1(\mathcal{F}_{t-1})$, $H_1(\mathcal{F}_{t-1})$, $V_2(\mathcal{F}_{t-1})$, $H_2(\mathcal{F}_{t-1})$ that satisfy $\forall \theta \in \mathcal{N}(\theta^0)$: $|v_t(\theta)| \leq V(\mathcal{F}_{t-1})$, $\|\nabla v_t(\theta)\| \leq V_1(\mathcal{F}_{t-1})$, $\|\nabla e_t(\theta)\| \leq H_1(\mathcal{F}_{t-1})$, $\|\nabla^2 v_t(\theta)\| \leq V_2(\mathcal{F}_{t-1})$, and $\|\nabla^2 e_t(\theta)\| \leq H_2(\mathcal{F}_{t-1})$.

(D) For some $0 < \delta < 1$ and for all t we have (i) $\mathbb{E} [V_1(\mathcal{F}_{t-1})^{3+\delta}]$, $\mathbb{E} [H_1(\mathcal{F}_{t-1})^{3+\delta}]$,

$$\mathbb{E} \left[V_2(\mathcal{F}_{t-1})^{\frac{3+\delta}{2}} \right], \mathbb{E} \left[H_2(\mathcal{F}_{t-1})^{\frac{3+\delta}{2}} \right] \leq K, \quad (ii) \mathbb{E} \left[V(\mathcal{F}_{t-1})^{2+\delta} V_1(\mathcal{F}_{t-1}) H_1(\mathcal{F}_{t-1})^{2+\delta} \right] \leq K,$$

$$(iii) \mathbb{E} \left[H_1(\mathcal{F}_{t-1})^{1+\delta} H_2(\mathcal{F}_{t-1}) |Y_t|^{2+\delta} \right], \mathbb{E} \left[H_1(\mathcal{F}_{t-1})^{3+\delta} |Y_t|^{2+\delta} \right] \leq K.$$

(E) The matrix \mathbf{D}_0 defined in Theorem 6 is (strictly) positive definite for T sufficiently large.

(F) $\{[Y_t, v_t(\theta^0), e_t(\theta^0), \nabla' v_t(\theta^0), \nabla' e_t(\theta^0)]\}$ is α -mixing with $\sum_{m=1}^{\infty} \alpha(m)^{(q-2)/q} < \infty$ for some $q > 2$.

(G) For any T , $\sup_{\theta \in \Theta} \sum_{t=1}^T \mathbf{1}\{Y_t = v_t(\theta)\} \leq K$ a.s.

Most of the above assumptions are standard. Assumption 5(A)(ii) imposes that the VaR is negative, but given our focus on the left-tail ($\alpha < 0.5$) of asset returns, this is not likely a binding constraint. Assumptions 5(B)–(E) are similar to those in Engle and Manganelli (2004a). Assumption 5(B)(ii) requires at least $4 + \delta$ moments of returns to exist, however 5(D) may actually increase the number of required moments, depending on the VaR-ES model employed. Our requirement of at least $4 + \delta$ moments of returns allows returns to be fat tailed, but not without limit: it rules out applications where kurtosis is not defined, for example Student's t distributions with degrees of freedom of four or less. (In our simulation study below, we show that the theory here has good finite sample properties when using a Skew t with five degrees of freedom.) Assumptions 5(C)–(D) are conditions on the magnitude of the VaR and ES paths, as well as first and second derivatives of these, making them somewhat hard to interpret. In the Supplemental Appendix we show that for a GARCH process these reduce to moment conditions on the observed returns. Assumption 5(F) is a standard condition on the amount of time series dependence, and allows us to invoke a CLT of Hall and Heyde (1980). Assumption 5(G) limits the number of exact equalities of realized returns and fitted VaR values; given assumption 5(B), in linear

models $K = \dim(\theta)$, while in nonlinear models it may be that $K < \dim(\theta)$.

Theorem 6 (Asymptotic Normality). *Under Assumptions 4 and 5, we have*

$$\sqrt{T}\mathbf{A}_0^{-1/2}\mathbf{D}_0(\hat{\theta}_T - \theta^0) \xrightarrow{d} N(0, I) \text{ as } T \rightarrow \infty \quad (3.30)$$

where

$$\mathbf{D}_0 = \mathbb{E} \left[\frac{f_t(v_t(\theta^0)|\mathcal{F}_{t-1})}{-e_t(\theta^0)\alpha} \nabla' v_t(\theta^0) \nabla v_t(\theta^0) + \frac{1}{e_t(\theta^0)^2} \nabla' e_t(\theta^0) \nabla e_t(\theta^0) \right] \quad (3.31)$$

$$\mathbf{A}_0 = \mathbb{E} [g_t(\theta^0)g_t(\theta^0)'] \quad (3.32)$$

$$\begin{aligned} g_t(\theta) &= \frac{\partial L(Y_t, v_t(\theta), e_t(\theta); \alpha)}{\partial \theta} \quad (3.33) \\ &= \nabla' v_t(\theta) \frac{1}{-e_t(\theta)} \left(\frac{1}{\alpha} \mathbf{1}\{Y_t \leq v_t(\theta)\} - 1 \right) \\ &\quad + \nabla' e_t(\theta) \frac{1}{e_t(\theta)^2} \left(\frac{1}{\alpha} \mathbf{1}\{Y_t \leq v_t(\theta)\} (v_t(\theta) - Y_t) - v_t(\theta) + e_t(\theta) \right) \end{aligned}$$

An outline of the proof of this theorem is given in Appendix, and the detailed lemmas underlying it are provided in the supplemental appendix. The proof of Theorem 6 builds on Huber (1967), Weiss (1991) and Engle and Manganelli (2004a), who focused on the estimation of quantiles.

Finally, we present a result for estimating the asymptotic covariance matrix of $\hat{\theta}_T$, thereby enabling the reporting of standard errors and confidence intervals.

Assumption 6. (A) *The deterministic positive sequence c_T satisfies $c_T = o(1)$ and $c_T^{-1} = o(T^{1/2})$.*

(B)(i) $T^{-1} \sum_{t=1}^T g_t(\theta^0)g_t(\theta^0)' - \mathbf{A}_0 \xrightarrow{p} \mathbf{0}$, where \mathbf{A}_0 is defined in Theorem 6 .

(ii) $T^{-1} \sum_{t=1}^T \frac{1}{e_t(\theta^0)^2} \nabla' e_t(\theta^0) \nabla e_t(\theta^0) - \mathbb{E} \left[\frac{1}{e_t(\theta^0)^2} \nabla' e_t(\theta^0) \nabla e_t(\theta^0) \right] \xrightarrow{p} \mathbf{0}$.

(iii) $T^{-1} \sum_{t=1}^T \frac{f_t(v_t(\theta^0)|\mathcal{F}_{t-1})}{-e_t(\theta^0)\alpha} \nabla' v_t(\theta^0) \nabla v_t(\theta^0) - \mathbb{E} \left[\frac{f_t(v_t(\theta^0)|\mathcal{F}_{t-1})}{-e_t(\theta^0)\alpha} \nabla' v_t(\theta^0) \nabla v_t(\theta^0) \right] \xrightarrow{p} \mathbf{0}$.

0.

Theorem 7. Under Assumptions 4-6, $\hat{\mathbf{A}}_T - \mathbf{A}_0 \xrightarrow{p} \mathbf{0}$ and $\hat{\mathbf{D}}_T - \mathbf{D}_0 \xrightarrow{p} \mathbf{0}$, where

$$\hat{\mathbf{A}}_T = T^{-1} \sum_{t=1}^T g_t(\hat{\theta}_T) g_t(\hat{\theta}_T)'$$

$$\hat{\mathbf{D}}_T = T^{-1} \sum_{t=1}^T \left\{ \frac{1}{2c_T} \mathbf{1} \left\{ |Y_t - v_t(\hat{\theta}_T)| < c_T \right\} \frac{\nabla' v_t(\hat{\theta}_T) \nabla v_t(\hat{\theta}_T)}{-\alpha e_t(\hat{\theta}_T)} + \frac{\nabla' e_t(\hat{\theta}_T) \nabla e_t(\hat{\theta}_T)}{e_t^2(\hat{\theta}_T)} \right\}$$

This result extends Theorem 3 in Engle and Manganelli (2004a) from dynamic VaR models to dynamic joint models for VaR and ES. The key choice in estimating the asymptotic covariance matrix is the bandwidth parameter in Assumption 6(A). In our simulation study below we set this to $T^{-1/3}$ and we find that this leads to satisfactory finite-sample properties.

The results here extend some very recent work in the literature: Dimitriadis and Bayer (2017) consider VaR-ES regression, but focus on *iid* data and linear specifications. These authors also consider a variety of FZ loss functions, in contrast with our focus on the FZ0 loss function, and they consider both M and GMM estimation, while we focus only on M estimation. Barendse (2017) considers “interquantile expectation regression,” which nests VaR-ES regression as a special case. He allows for time series data, but imposes that the models are linear. Our framework allows for time series data and nonlinear models.

3.4 Simulation study

In this section we investigate the finite-sample accuracy of the asymptotic theory for dynamic ES and VaR models presented in the previous section. For ease of comparison with existing studies of related models, such as volatility and VaR models, we consider a GARCH(1,1) for the DGP, and estimate the parameters by FZ loss

minimization. Specifically, the DGP is

$$Y_t = \sigma_t \eta_t \tag{3.34}$$

$$\sigma_t^2 = \omega + \beta \sigma_{t-1}^2 + \gamma Y_{t-1}^2$$

$$\eta_t \sim iid F_\eta(0, 1) \tag{3.35}$$

We set the parameters of this DGP to $(\omega, \beta, \gamma) = (0.05, 0.9, 0.05)$. We consider two choices for the distribution of η_t : a standard Normal, and the standardized skew t distribution of Hansen (1994), with degrees of freedom (ϑ) and skewness (λ) parameters in the latter set to $(5, -0.5)$. Under this DGP, the ES and VaR are proportional to σ_t , with

$$(\text{VaR}_t^\alpha, \text{ES}_t^\alpha) = (a_\alpha, b_\alpha) \sigma_t \tag{3.36}$$

We make the dependence of the coefficients of proportionality (a_α, b_α) on α explicit here, as we consider a variety of values of α in this simulation study: $\alpha \in \{0.01, 0.025, 0.05, 0.10, 0.20\}$. Interest in VaR and ES from regulators focuses on the smaller of these values of α , but we also consider the larger values to better understand the properties of the asymptotic approximations at various points in the tail of the distribution.

For a standard Normal distribution, with CDF and PDF denoted Φ and ϕ , we have:

$$a_\alpha = \Phi^{-1}(\alpha) \tag{3.37}$$

$$b_\alpha = -\phi(\Phi^{-1}(\alpha)) / \alpha$$

For Hansen's skew t distribution we can obtain VaR (a_α) from the inverse CDF, which is available in closed form. We obtain a closed-form expression for ES (b_α) by

extending results in Dobrev, et al. (2017) which provides analytical expressions for ES for (symmetric) Student’s t random variables. Details are presented in Appendix A.2.4. As noted in Section 3.2.5, FZ loss minimization does not allow us to identify ω in the GARCH model, and in our empirical work we set this parameter to one, however to facilitate comparisons of the accuracy of estimates of (a_α, b_α) in our simulation study we instead set ω at its true value. This is done without loss of generality and merely eases the presentation of the results. To match our empirical application, we replace the parameter a_α with $c_\alpha = a_\alpha/b_\alpha$, and so our parameter vector becomes $(\beta, \gamma, b_\alpha, c_\alpha)$.

We consider two sample sizes, $T \in \{2500, 5000\}$ corresponding to 10 and 20 years of daily returns respectively. These large sample sizes enable us to consider estimating models for quantiles as low as 1%, which are often used in risk management. We repeat all simulations 1000 times. To mitigate sensitivity to starting values, we initially estimate all models using a “smoothed” version of the FZ0 loss function, and use the resulting estimate as the starting value for the estimation problem using the original, “unsmoothed,” FZ0 loss function. Details are in Appendix A.3.

Table 3.1 presents results for the estimation of this model on standard Normal innovations, and Table 3.2 presents corresponding results for skew t innovations. The top row of each panel present the true parameter values, with the latter two parameters changing across α . The second row presents the median estimated parameter across simulations, and the third row presents the average bias in the estimated parameter. Both of these measures indicate that the parameter estimates are nicely centered on the true parameter values. The penultimate row presents the cross-simulation standard deviations of the estimated parameters, and we observe that these decrease with the sample size and increase as we move further into the tails (i.e., as α decreases), both as expected. Comparing the standard deviations across

Tables 3.1 and 3.2, we also note that they are higher for skew t innovations than Normal innovations, again as expected.

The last row in each panel presents the coverage probabilities for 95% confidence intervals for each parameter, constructed using the estimated standard errors, with bandwidth parameter $c_T = \lfloor T^{-1/3} \rfloor$. For $\alpha \geq 0.05$ we see that the coverage is reasonable, ranging from around 0.88 to 0.96. For $\alpha = 0.025$ or $\alpha = 0.01$ the coverage tends to be too low, particularly for the smaller sample size. Thus some caution is required when interpreting the standard errors for the models with the smallest values of α . In Table C.1 of the Supplemental Appendix we present results for (Q)MLE for the GARCH model corresponding to the results in Tables 3.1 and 3.2, using the theory of Bollerslev and Wooldridge (1992), and in Tables C.2 and C.3 we present results for CAViaR estimation of this model, using the “tick” loss function and the theory of Engle and Manganelli (2004a).⁷ We find that (Q)MLE has better finite sample properties than FZ minimization, but CAViaR estimation has slightly worse properties than FZ minimization.

Table 3.3 presents results for $T = 500$, which is relatively short given our interest in tail events, but may be of interest when only limited data are available or when structural breaks are suspected. We see here that the estimator remains approximately unbiased, however inference (e.g., through confidence intervals) is less reliable with this short sample.

⁷In (Q)MLE, the parameters to be estimated are (ω, β, γ) , and they are obtained by maximizing the sample average of the Normal log-likelihood. In “CAViaR” estimation, the parameters are $(\omega, \beta, \gamma, a_\alpha)$ and they are obtained by minimizing the sample average of the “tick” loss function, defined as $L(y, v; \alpha) = (\mathbf{1}\{y \leq v\} - \alpha)(v - y)$. Like FZ estimation, in the CAViaR approach we find that a_α and ω are not separately identified. As for the study of FZ estimation, we set ω to its true value to facilitate interpretation of the results, and estimate the remaining three parameters.

Table 3.1: Simulation results for Normal innovations

	$T = 2500$				$T = 5000$			
	β	γ	b_α	c_α	β	γ	b_α	c_α
$\alpha = 0.01$								
True	0.900	0.050	-2.665	0.873	0.900	0.050	-2.665	0.873
Median	0.901	0.049	-2.615	0.882	0.899	0.049	-2.671	0.877
Avg bias	-0.017	0.015	-0.108	0.008	-0.011	0.006	-0.089	0.004
St dev	0.077	0.076	1.095	0.022	0.049	0.033	0.805	0.015
Coverage	0.868	0.827	0.875	0.919	0.884	0.876	0.888	0.937
$\alpha = 0.025$								
True	0.900	0.050	-2.338	0.838	0.900	0.050	-2.338	0.838
Median	0.899	0.047	-2.329	0.842	0.897	0.048	-2.392	0.841
Avg bias	-0.017	0.007	-0.137	0.004	-0.011	0.002	-0.111	0.002
St dev	0.066	0.044	0.852	0.017	0.050	0.024	0.656	0.012
Coverage	0.898	0.870	0.911	0.931	0.912	0.888	0.925	0.923
$\alpha = 0.05$								
True	0.900	0.050	-2.063	0.797	0.900	0.050	-2.063	0.797
Median	0.901	0.048	-2.051	0.800	0.899	0.049	-2.094	0.799
Avg bias	-0.013	0.005	-0.097	0.002	-0.008	0.002	-0.081	0.001
St dev	0.062	0.046	0.707	0.015	0.041	0.021	0.511	0.010
Coverage	0.913	0.874	0.916	0.947	0.923	0.907	0.927	0.948
$\alpha = 0.10$								
True	0.900	0.050	-1.755	0.730	0.900	0.050	-1.755	0.730
Median	0.900	0.048	-1.769	0.730	0.898	0.048	-1.778	0.730
Avg bias	-0.015	0.006	-0.103	0.000	-0.009	0.001	-0.072	0.000
St dev	0.065	0.052	0.623	0.013	0.040	0.020	0.435	0.009
Coverage	0.917	0.883	0.925	0.954	0.922	0.902	0.934	0.960
$\alpha = 0.20$								
True	0.900	0.050	-1.400	0.601	0.900	0.050	-1.400	0.601
Median	0.898	0.048	-1.391	0.602	0.899	0.048	-1.417	0.602
Avg bias	-0.017	0.008	-0.091	0.000	-0.010	0.002	-0.064	0.000
St dev	0.078	0.072	0.547	0.014	0.044	0.022	0.374	0.010
Coverage	0.925	0.881	0.934	0.948	0.941	0.923	0.945	0.954

Note: This table presents results from 1000 replications of the estimation of VaR and ES from a GARCH(1,1) DGP with standard Normal innovations. Details are described in Section 3.4. The top row of each panel presents the true values of the parameters. The second, third, and fourth rows present the median estimated parameters, the average bias, and the standard deviation (across simulations) of the estimated parameters. The last row of each panel presents the coverage rates for 95% confidence intervals constructed using estimated standard errors.

Table 3.2: Simulation results for skew t innovations

	$T = 2500$				$T = 5000$			
	β	γ	b_α	c_α	β	γ	b_α	c_α
$\alpha = 0.01$								
True	0.900	0.050	-4.506	0.730	0.900	0.050	-4.506	0.730
Median	0.893	0.049	-4.376	0.750	0.895	0.048	-4.562	0.741
Avg bias	-0.047	0.038	-0.399	0.018	-0.028	0.014	-0.340	0.009
St dev	0.150	0.134	2.687	0.048	0.094	0.065	1.983	0.034
Coverage	0.797	0.797	0.809	0.894	0.837	0.853	0.839	0.936
$\alpha = 0.025$								
True	0.900	0.050	-3.465	0.695	0.900	0.050	-3.465	0.695
Median	0.895	0.047	-3.448	0.705	0.896	0.048	-3.520	0.701
Avg bias	-0.028	0.014	-0.254	0.008	-0.017	0.005	-0.198	0.004
St dev	0.101	0.069	1.591	0.034	0.068	0.033	1.192	0.023
Coverage	0.855	0.835	0.877	0.921	0.874	0.893	0.887	0.939
$\alpha = 0.05$								
True	0.900	0.050	-2.767	0.651	0.900	0.050	-2.767	0.651
Median	0.896	0.048	-2.760	0.656	0.898	0.048	-2.795	0.654
Avg bias	-0.021	0.007	-0.187	0.005	-0.011	0.003	-0.114	0.003
St dev	0.081	0.049	1.085	0.025	0.053	0.025	0.782	0.017
Coverage	0.906	0.883	0.921	0.937	0.916	0.904	0.922	0.951
$\alpha = 0.10$								
True	0.900	0.050	-2.122	0.577	0.900	0.050	-2.122	0.577
Median	0.897	0.048	-2.121	0.579	0.898	0.048	-2.140	0.578
Avg bias	-0.017	0.006	-0.125	0.003	-0.008	0.002	-0.069	0.002
St dev	0.066	0.045	0.745	0.020	0.040	0.022	0.510	0.014
Coverage	0.931	0.900	0.937	0.949	0.926	0.925	0.927	0.947
$\alpha = 0.20$								
True	0.900	0.050	-1.514	0.431	0.900	0.050	-1.514	0.431
Median	0.899	0.050	-1.485	0.432	0.899	0.049	-1.503	0.432
Avg bias	-0.019	0.006	-0.089	0.001	-0.008	0.002	-0.049	0.001
St dev	0.089	0.047	0.618	0.018	0.042	0.022	0.380	0.012
Coverage	0.916	0.888	0.922	0.938	0.929	0.916	0.940	0.944

Note: This table presents results from 1000 replications of the estimation of VaR and ES from a GARCH(1,1) DGP with skew t innovations. Details are described in Section 3.4. The top row of each panel presents the true values of the parameters. The second, third, and fourth rows present the median estimated parameters, the average bias, and the standard deviation (across simulations) of the estimated parameters. The last row of each panel presents the coverage rates for 95% confidence intervals constructed using estimated standard errors.

Table 3.3: Simulation results for T=500

	<i>Normal innovations</i>				<i>Skewed t innovations</i>			
	β	γ	b_α	c_α	β	γ	b_α	c_α
$\alpha = 0.01$								
True	0.900	0.050	-2.665	0.873	0.900	0.050	-4.506	0.730
Median	0.915	0.048	-2.165	0.917	0.906	0.033	-3.694	0.813
Avg bias	-0.041	0.063	0.056	0.042	-0.086	0.096	-0.250	0.071
St dev	0.161	0.189	1.550	0.049	0.233	0.264	3.552	0.095
Coverage	0.781	0.709	0.779	0.730	0.704	0.666	0.747	0.762
$\alpha = 0.025$								
True	0.900	0.050	-2.338	0.838	0.900	0.050	-3.465	0.695
Median	0.909	0.044	-2.136	0.860	0.906	0.030	-3.170	0.736
Avg bias	-0.028	0.031	-0.048	0.020	-0.053	0.048	-0.262	0.037
St dev	0.134	0.134	1.205	0.039	0.176	0.192	2.240	0.070
Coverage	0.862	0.765	0.868	0.899	0.817	0.717	0.835	0.875
$\alpha = 0.05$								
True	0.900	0.050	-2.063	0.797	0.900	0.050	-2.767	0.651
Median	0.905	0.040	-1.976	0.808	0.899	0.028	-2.671	0.672
Avg bias	-0.030	0.027	-0.133	0.011	-0.053	0.028	-0.366	0.021
St dev	0.134	0.142	1.098	0.033	0.174	0.165	1.817	0.053
Coverage	0.870	0.749	0.870	0.922	0.829	0.712	0.862	0.920
$\alpha = 0.10$								
True	0.900	0.050	-1.755	0.730	0.900	0.050	-2.122	0.577
Median	0.902	0.038	-1.694	0.736	0.897	0.031	-2.195	0.588
Avg bias	-0.032	0.024	-0.156	0.006	-0.058	0.025	-0.373	0.012
St dev	0.132	0.137	0.970	0.029	0.175	0.164	1.413	0.045
Coverage	0.883	0.756	0.880	0.950	0.845	0.746	0.876	0.922
$\alpha = 0.20$								
True	0.900	0.050	-1.400	0.601	0.900	0.050	-1.514	0.431
Median	0.899	0.037	-1.394	0.602	0.894	0.033	-1.564	0.435
Avg bias	-0.042	0.036	-0.177	0.001	-0.063	0.027	-0.290	0.004
St dev	0.147	0.174	0.854	0.031	0.187	0.167	1.070	0.038
Coverage	0.894	0.757	0.888	0.944	0.853	0.741	0.866	0.955

Note: This table presents results from 1000 replications of the estimation of VaR and ES from a GARCH(1,1) DGP with standard Normal innovations (left panel) or skew t innovations (right panel). Details are described in Section 3.4. The top row of each panel presents the true values of the parameters. The second, third, and fourth rows present the median estimated parameters, the average bias, and the standard deviation (across simulations) of the estimated parameters. The last row of each panel presents the coverage rates for 95% confidence intervals constructed using estimated standard errors.

Table 3.4: Sampling variation of FZ estimation relative to (Q)MLE and CAViaR

α	<i>Normal innovations</i>				<i>Skew t innovations</i>			
	$T = 2500$		$T = 5000$		$T = 2500$		$T = 5000$	
	β	γ	β	γ	β	γ	β	γ
Panel A: FZ/(Q)ML								
0.01	1.209	5.940	1.701	3.731	1.577	4.830	2.533	3.723
0.025	1.034	3.394	1.764	2.694	1.055	2.485	1.853	1.905
0.05	0.980	3.576	1.431	2.377	0.850	1.784	1.426	1.458
0.10	1.021	4.074	1.406	2.302	0.698	1.627	1.095	1.250
0.20	1.224	5.558	1.543	2.497	0.939	1.710	1.145	1.242
Panel B: FZ/CAViaR								
0.01	0.982	1.162	0.951	0.975	1.062	1.384	0.912	1.465
0.025	0.965	1.139	0.971	1.042	0.976	1.030	0.974	0.997
0.05	0.925	1.238	0.910	0.930	0.885	0.819	0.920	0.903
0.10	0.940	1.283	0.847	0.827	0.831	0.903	0.816	0.819
0.20	0.855	0.671	0.703	0.510	0.736	0.437	0.503	0.515

Note: This table presents the ratio of cross-simulation standard deviations of parameter estimates obtained by FZ loss minimization and (Q)MLE (Panel A), and CAViaR (Panel B). We consider only the parameters that are common to these three estimation methods, namely the GARCH(1,1) parameters β and γ . Ratios greater than one indicate the FZ estimator is more variable than the alternative estimation method; ratios less than one indicate the opposite.

In Table 3.4 we compare the efficiency of FZ estimation relative to (Q)MLE and to CAViaR estimation, for the parameters that all three estimation methods have in common, namely (β, γ) . As expected, when the innovations are standard Normal, FZ estimation is substantially less efficient than MLE, however when the innovations are skew t the loss in efficiency drops and for some values of α FZ estimation is actually more efficient than QMLE. This switch in the ranking of the competing estimators is qualitatively in line with results in Francq and Zakoïan (2015). In Panel B of Table 3.4, we see that FZ estimation is generally, though not uniformly, more efficient than CAViaR estimation.

In many applications, interest is focused on the forecasted values of VaR and ES rather than the estimated parameters of the models generating these forecasts. To study this, Table 3.5 presents results on the accuracy of the fitted VaR and ES estimates for the three estimation methods: (Q)MLE, CAViaR and FZ estimation. We consider the same two DGPs as above, and two others that represent more challenging environments for QMLE. In the two additional DGPs, we assume the same mean and volatility dynamics as before, and we additionally allow the degrees of freedom (ν) and skewness (λ) parameters in the skew t distribution to vary in such a way as to either “offset” or “amplify” the dynamics in volatility, resulting in VaR and ES series that are either approximately constant, or proportional to the conditional variance rather than the conditional standard deviation. These two simulation designs represent simple ways to obtain dynamics in VaR and ES that are “far” from the dynamics in volatility, and is an environment where QMLE would be expected to perform poorly. Details are provided in Appendix A.4.

To obtain estimates of VaR and ES from the (Q)ML estimates, we follow common empirical practice and compute the sample VaR and ES of the estimated standardized residuals. The columns labeled *MAE* present the mean absolute error from (Q)MLE, and in the next two columns of each panel we present the *relative MAE* of CAViaR and FZ to (Q)MLE.

For Normal innovations, reported in Panel A, MLE is the most accurate estimation method, as expected. Averaging across values of α , CAViaR is about 40% worse, while FZ is about 30% worse. For skew t innovations, reported in Panel B, the gap in performance closes somewhat, with CAViaR and FZ performing about 24% and 16% worse than QMLE. In Panels C and D we consider challenging environments for QMLE, where the dynamics in volatility, which is the focus in QMLE, are very different from those in VaR and ES, which are the focus in FZ estimation. Unsurpris-

ingly, QMLE does poorly in this case compared with FZ estimation, with MAE ratios (averaging across α) of 0.41 and 0.61 in these two panels, indicating that FZ does between 1.5 and 2.5 times better than QMLE in these simulation designs. CAViaR also outperforms QMLE in these designs, with average MAE ratios of 0.50 and 0.65.

Overall, these simulation results show that the asymptotic results of the previous section provide reasonable approximations in finite samples, with the approximations improving for larger sample sizes and less extreme values of α . Compared with MLE, estimation by FZ loss minimization is less accurate when the innovations are Normal or skew t , but when the dynamics in VaR and ES are different from those in volatility, the benefits of FZ estimation becomes apparent. Across all simulation designs, we find that FZ estimation is generally more accurate than estimation using the CAViaR approach of Engle and Manganelli (2004a), likely attributable to the fact that FZ estimation draws on information from two tail measures, VaR and ES, while CAViaR was designed to only model VaR.

Table 3.5: Mean absolute errors for VaR and ES estimates

α	$T = 2500$						$T = 5000$					
	VaR			ES			VaR			ES		
	MAE		MAE ratio	MAE		MAE ratio	MAE		MAE ratio	MAE		MAE ratio
	QML	CAViaR	FZ	QML	CAViaR	FZ	QML	CAViaR	FZ	QML	CAViaR	FZ
Panel A: <i>Normal innovations</i>												
0.01	0.069	1.368	1.369	0.084	1.487	1.345	0.049	1.404	1.387	0.060	1.443	1.344
0.025	0.055	1.305	1.288	0.064	1.341	1.290	0.038	1.306	1.291	0.044	1.348	1.313
0.05	0.043	1.302	1.271	0.051	1.332	1.289	0.031	1.314	1.264	0.036	1.350	1.290
0.10	0.034	1.322	1.253	0.042	1.394	1.302	0.024	1.365	1.265	0.029	1.449	1.320
0.20	0.026	1.443	1.257	0.033	1.652	1.377	0.018	1.458	1.241	0.023	1.706	1.377
Panel B: <i>Skew t innovations</i>												
0.01	0.196	1.327	1.381	0.342	1.249	1.252	0.138	1.369	1.375	0.245	1.256	1.248
0.025	0.120	1.228	1.244	0.205	1.166	1.166	0.087	1.245	1.234	0.145	1.197	1.185
0.05	0.084	1.193	1.166	0.141	1.154	1.129	0.061	1.184	1.143	0.101	1.164	1.119
0.10	0.056	1.168	1.089	0.098	1.160	1.083	0.041	1.155	1.067	0.071	1.158	1.069
0.20	0.034	1.301	1.087	0.066	1.404	1.121	0.024	1.316	1.066	0.048	1.409	1.089
Panel C: <i>Offsetting dynamics in VaR and ES</i>												
0.01	0.111	0.395	0.406	0.264	1.048	0.420	0.075	0.278	0.281	0.260	0.310	0.287
0.025	0.074	0.339	0.343	0.222	0.679	0.331	0.055	0.251	0.255	0.221	0.253	0.249
0.05	0.058	0.369	0.348	0.196	0.522	0.298	0.046	0.262	0.267	0.196	0.264	0.233
0.10	0.058	0.476	0.465	0.166	0.436	0.348	0.048	0.382	0.388	0.166	0.341	0.292
0.20	0.054	0.977	1.030	0.106	0.778	0.508	0.049	0.902	1.019	0.106	0.744	0.459
Panel D: <i>Amplifying dynamics in VaR and ES</i>												
0.01	0.141	0.533	0.457	0.340	0.700	0.415	0.150	0.543	0.457	0.363	0.397	0.412
0.025	0.144	0.650	0.551	0.288	0.705	0.501	0.148	0.658	0.616	0.268	0.476	0.552
0.05	0.127	0.698	0.658	0.191	0.737	0.666	0.124	0.777	0.724	0.176	0.652	0.703
0.10	0.084	0.581	0.637	0.112	0.781	0.745	0.082	0.689	0.726	0.105	0.734	0.780
0.20	0.047	0.518	0.559	0.066	0.844	0.709	0.046	0.508	0.553	0.063	0.827	0.731

Note: This table presents results on the accuracy of the fitted VaR and ES estimates for the three estimation methods: QML, CAViaR and FZ estimation. In the first column of each panel we present the mean absolute error (MAE) from QML, computed across all dates in a given sample and all 1000 simulation replications. The next two columns present the *relative* MAE of CAViaR and FZ to QML. Values greater than one indicate QML is more accurate (has lower MAE); values less than one indicate the opposite.

3.5 Forecasting equity index ES and VaR

We now apply the models discussed in Section 3.2 to the forecasting of ES and VaR for daily returns on four international equity indices. We consider the S&P 500 index, the Dow Jones Industrial Average, the NIKKEI 225 index of Japanese stocks, and the FTSE 100 index of UK stocks. Our sample period is 1 January 1990 to 31 December 2016, yielding between 6,630 and 6,805 observations per series (the exact numbers vary due to differences in holidays and market closures). In our out-of-sample analysis, we use the first ten years for estimation, and reserve the remaining 17 years for evaluation and model comparison.

Table 3.6 presents full-sample summary statistics on these four return series. Average annualized returns range from -2.7% for the NIKKEI to 7.2% for the DJIA, and annualized standard deviations range from 17.0% to 24.7%. All return series exhibit mild negative skewness (around -0.15) and substantial kurtosis (around 10). The lower two panels of Table 3.6 present the sample VaR and ES for four choices of α .

Table 3.7 presents results from standard time series models estimated on these return series over the in-sample period (Jan 1990 to Dec 1999). In the first panel we present the estimated parameters of the optimal ARMA(p, q) models, where the choice of (p, q) is made using the BIC. We note that for three of the four series the optimal model includes just a constant, consistent with the well-known lack of predictability of daily equity returns. The second panel presents the parameters of the GARCH(1,1) model for conditional variance, and the lower panel presents the estimated parameters the skew t distribution applied to the standardized residuals. All of these parameters are broadly in line with values obtained by other authors for these or similar series.

Table 3.6: Summary statistics on the four daily equity return series.

	S&P 500	DJIA	NIKKEI	FTSE
Mean (Annualized)	6.776	7.238	-2.682	3.987
Std dev (Annualized)	17.879	17.042	24.667	17.730
Skewness	-0.244	-0.163	-0.114	-0.126
Kurtosis	11.673	11.116	8.580	8.912
VaR-0.01	-3.118	-3.034	-4.110	-3.098
VaR-0.025	-2.324	-2.188	-3.151	-2.346
VaR-0.05	-1.731	-1.640	-2.451	-1.709
VaR-0.10	-1.183	-1.126	-1.780	-1.193
ES-0.01	-4.528	-4.280	-5.783	-4.230
ES-0.025	-3.405	-3.215	-4.449	-3.295
ES-0.05	-2.697	-2.553	-3.603	-2.643
ES-0.10	-2.065	-1.955	-2.850	-2.031

Note: This table presents summary statistics on the four daily equity return series studied in Section 3.5, over the full sample period from January 1990 to December 2016. The first two rows report the annualized mean and standard deviation of these returns in percent. The second panel presents sample Value-at-Risk for four choices of α , and the third panel presents corresponding sample Expected Shortfall estimates.

Table 3.7: ARMA, GARCH, and Skew t results

	SP500	DJIA	NIKKEI	FTSE
ϕ_0	0.056	0.056	-0.029	0.042
θ_1	–	–	–	0.075
R^2	0.000	0.000	0.000	0.006
ω	0.005	0.004	0.072	0.009
β	0.942	0.922	0.865	0.936
γ	0.052	0.077	0.105	0.053
ν	6.358	6.766	6.677	13.663
λ	-0.035	-0.059	-0.016	-0.024

Note: This table presents parameter estimates for the four daily equity return series studied in Section 3.5, over the in-sample period from January 1990 to December 1999. The first panel presents the optimal ARMA model according to the BIC, along with the R^2 of that model. The second panel presents the estimated GARCH(1,1) parameters, and the third panel presents the estimated parameters of the skewed t distribution applied to the estimated standardized residuals.

3.5.1 In-sample estimation

We now present estimates of the parameters of the models presented in Section 3.2, along with standard errors computed using the theory from Section 3.3. In the interests of space, we only report the parameter estimates for the S&P 500 index for $\alpha = 0.05$. The two-factor GAS model based on the FZ0 loss function is presented in the left panel of Table 3.8 . This model allows for separate dynamics in VaR and ES, and we present the parameters for each of these risk measures in separate columns. We impose that the \mathbf{B} matrix is diagonal for parsimony. We observe that the persistence of these processes is high, with the estimated b parameters equal to 0.993 and 0.994, similar to the persistence found in GARCH models (e.g., see Table 3.7). The model-implied average values of VaR and ES are -1.589 and -2.313, similar to the sample values of these measures reported in Table 3.6 . We observe that the coefficients on λ_e for both VaR and ES are small in magnitude and far from being statistically significant. The coefficients on λ_v are larger and more significant (the t -statistics are -2.95 and -2.58). The overall imprecision from the coefficients on the four forcing variables suggests that this model is over-parameterized. For example, proportionality of v_t and e_t would suggest that a one-factor model is sufficient. We can formally test for this in the context of the two-factor model by testing that $w_e/w_v = a_{ev}/a_{vv} = a_{ee}/a_{ve} \cap b_v = b_e$. We obtain a p -value of 0.77 for this restriction, indicating no evidence against proportionality.

The right panel of Table 3.8 shows three one-factor models for ES and VaR. The first is the one-factor GAS model, which is nested in the two-factor model presented in the left panel. We see a slight loss in fit (the average loss is slightly greater) but the parameters of this model are estimated with greater precision. The one-factor GAS model fits better than the GARCH model estimated via FZ loss minimization

(reported in the penultimate column).⁸ The “hybrid” model, augmenting the one-factor GAS model with a GARCH-type forcing variable, fits better than the other one-factor models, and also slightly better than the larger two-factor GAS model, and we observe that the coefficient on the GARCH forcing variable (δ) is significantly different from zero (with a t -statistic of 9.55). The computation times for these models is reported in the bottom row of Table 3.8; for comparison, the computation time for QML estimation of the GARCH model is 0.39 seconds.

⁸Recall that in all of the one-factor models, the intercept (ω) in the GAS equation is unidentified. We fix it at zero for the GAS-1F and Hybrid models, and at one for the GARCH-FZ model. This has no impact on the fit of these models for VaR and ES, but it means that we cannot interpret the estimated (a, b) parameters as the VaR and ES of the standardized residuals, and we no longer expect the estimated values to match the sample estimates in Table 3.6.

Table 3.8: Estimated parameters of GAS models for VaR and ES

	<i>GAS-2F</i>			<i>GAS-1F</i>	<i>GARCH-FZ</i>	<i>Hybrid</i>
	VaR	ES				
w	-0.009	-0.010	β	0.995	0.944	0.974
(s.e.)	(0.003)	(0.004)	(s.e.)	(0.002)	(0.058)	(0.006)
b	0.993	0.994	γ	0.007	0.031	0.003
(s.e.)	(0.002)	(0.003)	(s.e.)	(0.0001)	(0.010)	(0.003)
a_v	-0.358	-0.351	δ	–	–	0.017
(s.e.)	(0.109)	(0.129)	(s.e.)			(0.002)
a_e	-0.003	-0.003	a	-1.164	-1.955	-2.320
(s.e.)	(0.002)	(0.003)	(s.e.)	(0.420)	(0.256)	(4.671)
			b	-1.757	-2.829	-3.434
			(s.e.)	(0.634)	(0.522)	(6.874)
Avg loss	0.592			0.603	0.637	0.590
Time (secs)	44.773			0.594	0.767	1.343

Note: This table presents parameter estimates and standard errors for four GAS models of VaR and ES for the S&P 500 index over the in-sample period from January 1990 to December 1999. The left panel presents the results for the two-factor GAS model in Section 3.2.2. The right panel presents the results for the three one-factor models: a one-factor GAS model (from Section 3.2.3), and a GARCH model estimated by FZ loss minimization, and “hybrid” one-factor GAS model that includes an additional GARCH-type forcing variable (both from Section 3.2.5). The penultimate row of this table presents the average (in-sample) losses from each of these four models, and the bottom row presents the estimation time for each model (using Matlab R2018b on a 3.4GHz machine).

3.5.2 Out-of-sample forecasting

We now turn to the out-of-sample (OOS) forecast performance of the models discussed above, as well as some competitor models from the existing literature. We will focus initially on the results for $\alpha = 0.05$, given the focus on that percentile in the extant VaR literature. (Results for other values of α are considered below, with details provided in the supplemental appendix.) We will consider a total of ten models for forecasting ES and VaR. Firstly, we consider three rolling window methods, using window lengths of 125, 250 and 500 days. We next consider ARMA-GARCH models, with the ARMA model orders selected using the BIC, and assuming that the distribution of the innovations is standard Normal or skew t , or estimating it nonparametrically using the sample ES and VaR of the estimated standardized residuals. Finally we consider four new semiparametric dynamic models for ES and VaR: the two-factor GAS model presented in Section 3.2.2, the one-factor GAS model presented in Section 3.2.3, a GARCH model estimated using FZ loss minimization, and the “hybrid” GAS/GARCH model presented in Section 3.2.5. We estimate these models using the first ten years as our in-sample period, and retain those parameter estimates throughout the OOS period.

In Figure 3.4 below we plot the fitted 5% ES and VaR for the S&P 500 return series, using three models: the rolling window model using a window of 125 days, the GARCH-EDF model, and the one-factor GAS model. This figure covers both the in-sample and out-of-sample periods. The figure shows that the average ES was estimated at around -2%, rising as high as around -1% in the mid 90s and mid 00s, and falling to its most extreme values of around -10% during the financial crisis in late 2008. Thus, like volatility, ES fluctuates substantially over time.

Figure 3.5 zooms in on the last two years of our sample period, to better reveal

the differences in the estimates from these models. We observe the usual step-like movements in the rolling window estimate of VaR and ES, as the more extreme observations enter and leave the estimation window. Comparing the GARCH and GAS estimates, we see how they differ in reacting to returns: the GARCH estimates are driven by lagged squared returns, and thus move stochastically each day. The GAS estimates, on the other hand, only use information from returns when the VaR is violated, and on other days the estimates revert deterministically to the long-run mean. This generates a smoother time series of VaR and ES estimates. We investigate below which of these estimates provides a better fit to the data.

The left panel of Table 3.9 presents the average OOS losses, using the FZ0 loss function from equation (3.6), for each of the ten models, for the four equity return series. The lowest values in each column are highlighted in bold, and the second-lowest are in italics. We observe that the one-factor GAS model, labelled FZ1F, is the preferred model for the two US equity indices, while the Hybrid model is the preferred model for the NIKKEI and FTSE indices. The worst model is the rolling window with a window length of 500 days.

While average losses are useful for an initial look at OOS forecast performance, they do not reveal whether the gains are statistically significant. Table 3.10 presents Diebold-Mariano t -statistics on the loss differences, for the S&P 500 index. Corresponding tables for the other three equity return series are presented in Table C.4 of the supplemental appendix. The tests are conducted as “row model minus column model” and so a positive number indicates that the column model outperforms the row model. The column “FZ1F” corresponding to the one-factor GAS model contains all positive entries, revealing that this model out-performed all competing models. This outperformance is strongly significant for the comparisons to the rolling window forecasts, as well as the GARCH model with Normal innovations. The gains relative to the GARCH model with skew t or nonparametric innovations are not significant, with DM t -statistics of 1.79 and 1.53 respectively. Similar results are found for the best models for each of the other three equity return series. Thus the worst models are easily separated from the better models, but the best few models are generally not significantly different. The supplemental appendix presents results analogous to Table 3.9, but with $\alpha=0.025$, which is the value for ES that is the focus of the Basel III accord. The rankings and results are qualitatively similar to those for $\alpha=0.05$ discussed here.

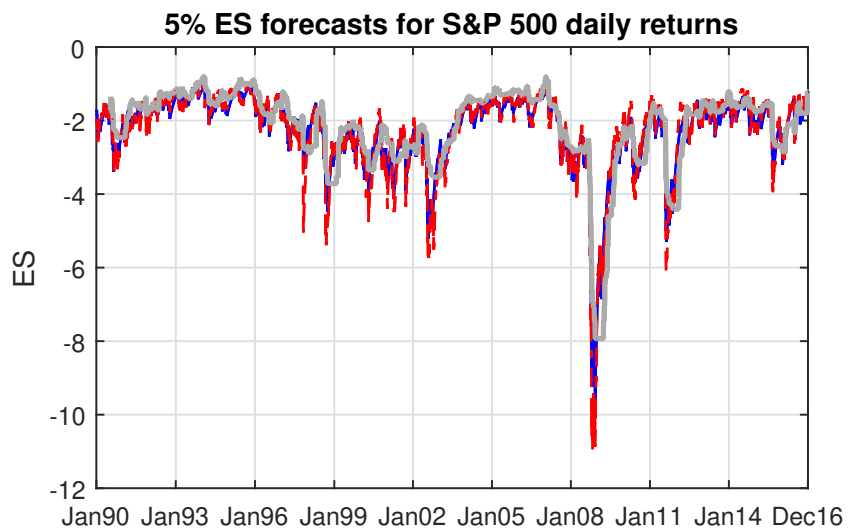
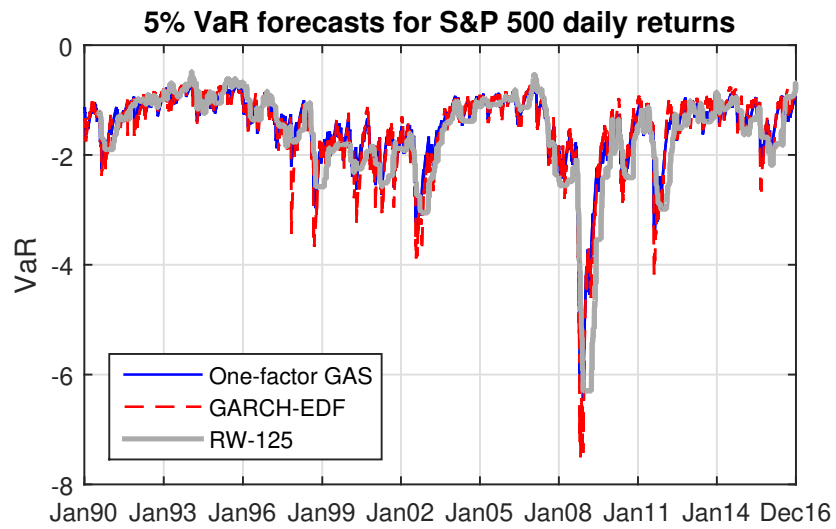


Figure 3.4: Plot of the estimated 5% VaR and ES or daily returns on the S&P 500 index, over the period January 1990 to December 2016.

Note: *The estimates are based on a one-factor GAS model, a GARCH model, and a rolling window using 125 observations.*

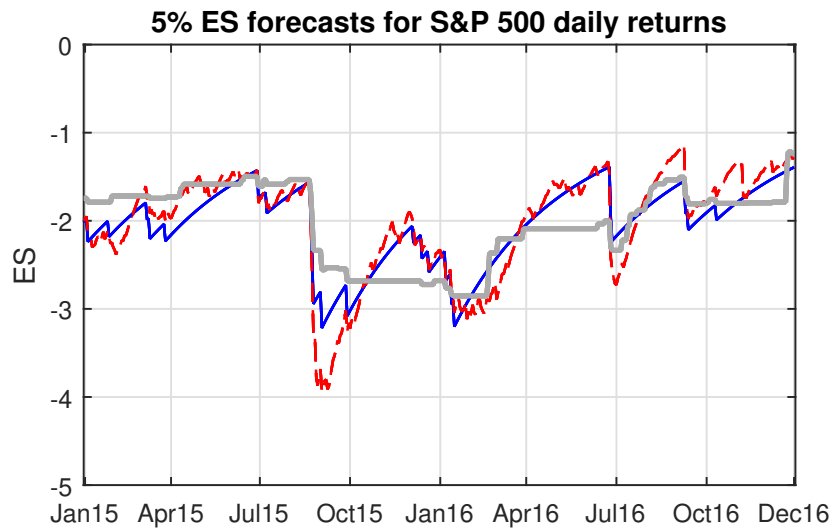
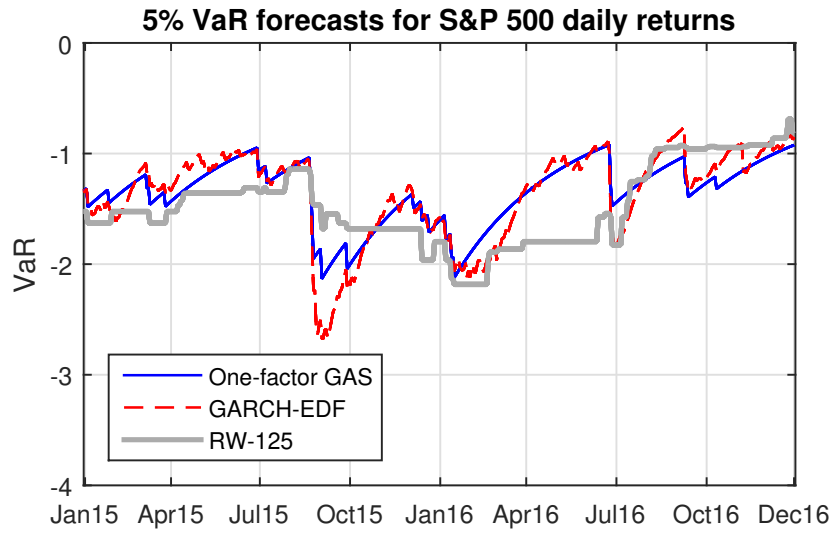


Figure 3.5: Plot of the estimated 5% VaR and ES for daily returns on the S&P 500 index, over the period January 2015 to December 2016.

Note: The estimates are based on a one-factor GAS model, a GARCH model, and a rolling window using 125 observations.

Table 3.9: Out-of-sample average losses and goodness-of-fit tests ($\alpha=0.05$)

	<i>Average loss</i>				<i>GoF p-values: VaR</i>				<i>GoF p-values: ES</i>			
	S&P	DJIA	NIK	FTSE	S&P	DJIA	NIK	FTSE	S&P	DJIA	NIK	FTSE
RW-125	0.914	0.864	1.290	0.959	0.039	0.021	0.002	0.000	0.046	0.028	0.011	0.000
RW-250	0.959	0.909	1.294	1.002	0.003	0.002	0.026	0.000	<i>0.062</i>	0.020	0.034	0.002
RW-500	1.023	0.975	1.318	1.056	0.002	0.003	0.001	0.000	0.024	0.021	0.002	0.000
GCH-N	0.876	0.811	1.170	0.871	0.043	0.010	0.536	0.001	0.001	0.000	0.195	0.000
GCH-Skt	0.865	0.799	1.168	<i>0.864</i>	0.006	0.006	0.109	0.001	0.005	0.004	0.273	0.000
GCH-EDF	0.862	<i>0.796</i>	<i>1.166</i>	0.865	0.005	0.006	0.580	0.001	0.019	0.018	0.519	0.000
FZ-2F	<i>0.859</i>	0.799	1.206	0.874	0.004	0.001	0.279	0.001	0.170	0.319	0.313	0.004
FZ-1F	0.850	0.791	1.190	0.871	0.007	0.000	0.218	0.000	<i>0.073</i>	0.002	0.550	0.001
GCH-FZ	0.862	0.797	1.166	0.870	0.009	0.008	0.519	0.000	0.027	0.035	0.459	0.000
Hybrid	0.870	0.796	1.165	0.859	0.000	0.007	0.464	0.000	0.002	0.038	0.453	0.000

Note: The left panel of this table presents the average losses, using the FZ0 loss function, for four daily equity return series, over the out-of-sample period from January 2000 to December 2016, for ten different forecasting models. The lowest average loss in each column is highlighted in bold, the second-lowest is highlighted in italics. The first three rows correspond to rolling window forecasts, the next three rows correspond to GARCH forecasts based on different models for the standardized residuals, and the last four rows correspond to models introduced in Section 3.2. The middle and right panels of this table present p -values from goodness-of-fit tests of the VaR and ES forecasts respectively. Values that are greater than 0.10 (indicating no evidence against optimality at the 0.10 level) are in bold, and values between 0.05 and 0.10 are in italics.

Table 3.10: Diebold-Mariano t -statistics on average out-of-sample loss differences

	RW125	RW250	RW500	G-N	G-Skt	G-EDF	FZ-2F	FZ-1F	G-FZ	Hybrid
RW125		-2.257	-3.527	1.952	2.478	2.625	2.790	3.600	2.736	2.684
RW250	2.257		-3.215	2.752	3.129	3.246	3.364	4.039	3.399	3.538
RW500	3.527	3.215		3.706	3.997	4.087	4.334	4.818	4.223	4.454
G-N	-1.952	-2.752	-3.706		3.526	2.965	1.418	2.483	2.847	0.634
G-Skt	-2.478	-3.129	-3.997	-3.526		1.954	0.626	1.791	1.179	-0.436
G-EDF	-2.625	-3.246	-4.087	-2.965	-1.954		0.335	1.529	-0.023	-0.756
FZ-2F	-2.790	-3.364	-4.334	-1.418	-0.626	-0.335		1.000	-0.329	-0.904
FZ-1F	-3.600	-4.039	-4.818	-2.483	-1.791	-1.529	-1.000		-1.624	-2.049
G-FZ	-2.736	-3.399	-4.223	-2.847	-1.179	0.023	0.329	1.624		-0.895
Hybrid	-2.684	-3.538	-4.454	-0.634	0.436	0.756	0.904	2.049	0.895	

Note: This table presents t -statistics from Diebold-Mariano tests comparing the average losses, using the FZ0 loss function, over the out-of-sample period from January 2000 to December 2016, for ten different forecasting models. A positive value indicates that the row model has higher average loss than the column model. Values greater than 1.96 in absolute value indicate that the average loss difference is significantly different from zero at the 95% confidence level. Values along the main diagonal are all identically zero and are omitted for interpretability. The first three rows correspond to rolling window forecasts, the next three rows correspond to GARCH forecasts based on different models for the standardized residuals, and the last four rows correspond to models introduced in Section 3.2.

To complement the study of the relative performance of these models for ES and VaR, we now consider goodness-of-fit tests for the OOS forecasts of VaR and ES. Under correct specification of the model for VaR and ES, we know that

$$\mathbb{E}_{t-1} \begin{bmatrix} \partial L_{FZ0}(Y_t, v_t, e_t; \alpha) / \partial v_t \\ \partial L_{FZ0}(Y_t, v_t, e_t; \alpha) / \partial e_t \end{bmatrix} = 0 \quad (3.38)$$

and we note that this implies that $\mathbb{E}_{t-1}[\lambda_{v,t}] = \mathbb{E}_{t-1}[\lambda_{e,t}] = 0$, where $(\lambda_{v,t}, \lambda_{e,t})$ are defined in equations (3.11)-(3.12). Thus the variables $\lambda_{v,t}$ and $\lambda_{e,t}$ can be considered as a form of “generalized residual” for this model. To mitigate the impact of serial correlation in these measures (which comes through the persistence of v_t and e_t) we use standardized versions of these residuals:

$$\begin{aligned} \lambda_{v,t}^s &\equiv \frac{\lambda_{v,t}}{v_t} = \mathbf{1}\{Y_t \leq v_t\} - \alpha \\ \lambda_{e,t}^s &\equiv \frac{\lambda_{e,t}}{e_t} = \frac{1}{\alpha} \mathbf{1}\{Y_t \leq v_t\} \frac{Y_t}{e_t} - 1 \end{aligned} \quad (3.39)$$

These standardized generalized residuals are also conditionally mean zero under correct specification, and we note that the standardized residual for VaR is simply the demeaned “hit” variable, which is the focus of well-known tests from the VaR literature, see Christoffersen (1998) and Engle and Manganelli (2004a). We adopt the “dynamic quantile (DQ)” testing approach of Engle and Manganelli (2004a), which is based on simple regressions of these generalized residuals on elements of the information set available at the time the forecast was made. Consider, then the following “DQ” and “DES” regressions:

$$\begin{aligned} \lambda_{v,t}^s &= a_0 + a_1 \lambda_{v,t-1}^s + a_2 v_t + u_{v,t} \\ \lambda_{e,t}^s &= b_0 + b_1 \lambda_{e,t-1}^s + b_2 e_t + u_{e,t} \end{aligned} \quad (3.40)$$

where $\mathbf{a} = [a_0, a_1, a_2]'$ and $\mathbf{b} = [b_0, b_1, b_2]'$ are the parameters of the regression and $u_{v,t}$ and $u_{e,t}$ are the regression residuals. We test forecast optimality by testing that all parameters in these regressions are zero, against the usual two-sided alternative. Similar “conditional calibration” tests are presented in Nolde and Ziegel (2017). One could also consider a joint test of both of the above null hypotheses, however we will focus on these separately so that we can determine which variable is well/poorly specified.

The right two panels of Table 3.9 present the p -values from the tests of the goodness-of-fit of the VaR and ES forecasts. Entries greater than 0.10 (indicating no evidence against optimality at the 0.10 level) are in bold, and entries between 0.05 and 0.10 are in italics. For the S&P 500 index and the DJIA, we see that only one model passes the ES tests: the two-factor GAS model, while no model passes the VaR tests. For the NIKKEI we see that all of the dynamic models pass these two tests, while all three of the rolling window models fail. For the FTSE index, on the other hand, we see that all ten models considered here fail both the goodness-of-fit tests. The outcomes for the NIKKEI and the FTSE each, in different ways, present good examples of the problem highlighted in Nolde and Ziegel (2017), that many different models may pass a goodness-of-fit test, or all models may fail, which makes discussing their relative performance difficult. To do so, one can look at Diebold-Mariano tests of differences in average loss, as we do in Table 3.10.

Finally, in Table 3.11 we look at the performance of these models across four values of α , to see whether the best-performing models change with how deep in the tails we are. We find that this is indeed the case: for $\alpha = 0.01$, the best-performing model across the four return series is the GARCH model estimated by FZ loss minimization, followed by the GARCH model with nonparametric residuals. These rankings also hold for $\alpha = 0.025$. For $\alpha = 0.05$ the two best models are the GARCH model

with nonparametric residuals and the Hybrid model, while for $\alpha = 0.10$ the two best models are the Hybrid model and the one-factor GAS model. These rankings are perhaps related to the fact that the forcing variable in the GAS model depends on observing a violation of the VaR, and for very small values of α these violations occur only infrequently. In contrast, the GARCH model uses the information from the squared residual, and so information from the data moves the risk measures whether a VaR violation was observed or not. When α is not too small, the forcing variable suggested by the GAS model applied to the FZ loss function starts to out-perform.

Table 3.11: Out-of-sample performance rankings for various alpha

	$\alpha = 0.01$					$\alpha = 0.025$				
	S&P	DJIA	NIK	FTSE	Avg	S&P	DJIA	NIK	FTSE	Avg
RW-125	8	8	10	8	8.50	8	8	9	7	8.00
RW-250	9	9	8	9	8.75	9	9	8	8	8.50
RW-500	10	10	9	10	9.75	10	10	10	10	10.00
G-N	7	7	5	4	5.75	7	6	4	3	5.00
G-Skt	6	3	1	2	3.00	5	3	1	1	2.50
G-EDF	5	2	2	1	2.50	2	2	3	2	2.25
FZ-2F	4	4	6	7	5.25	4	5	7	9	6.25
FZ-1F	3	6	7	6	5.50	3	4	6	5	4.50
G-FZ	2	1	3	3	2.25	1	1	2	4	2.00
Hybrid	1	5	4	5	3.75	6	7	5	6	6.00

	$\alpha = 0.05$					$\alpha = 0.10$				
	S&P	DJIA	NIK	FTSE	Avg	S&P	DJIA	NIK	FTSE	Avg
RW-125	8	8	8	8	8.00	8	8	8	8	8.00
RW-250	9	9	9	9	9.00	9	9	9	9	9.00
RW-500	10	10	10	10	10.00	10	10	10	10	10.00
G-N	7	7	5	6	6.25	3	4	7	4	4.50
G-Skt	5	6	4	2	4.25	7	6	6	3	5.50
G-EDF	3	3	2	3	2.75	4	2	3	5	3.50
FZ-2F	2	2	7	4	3.75	2	3	5	7	4.25
FZ-1F	1	1	6	7	3.75	1	7	2	2	3.00
G-FZ	4	5	3	5	4.25	6	5	4	6	5.25
Hybrid	6	4	1	1	3.00	5	1	1	1	2.00

Note: This table presents the rankings (with the best performing model ranked 1 and the worst ranked 10) based on average losses using the FZ0 loss function, for four daily equity return series, over the out-of-sample period from January 2000 to December 2016, for ten different forecasting models. The first three rows in each panel correspond to rolling window forecasts, the next three rows correspond to GARCH forecasts based on different models for the standardized residuals, and the last four rows correspond to models introduced in Section 3.2. The last column in each panel represents the average rank across the four equity return series.

3.6 Conclusion

With the implementation of the Third Basel Accord in the next few years, risk managers and regulators will place greater focus on expected shortfall (ES) as a measure of risk, complementing and partly substituting previous emphasis on Value-at-Risk (VaR). We draw on recent results from statistical decision theory (Fissler and Ziegel, 2016) to propose new dynamic models for ES and VaR. The models proposed are semiparametric, in that they impose parametric structures for the dynamics of ES and VaR, but are agnostic about the conditional distribution of returns. We also present asymptotic distribution theory for the estimation of these models, and we verify that the theory provides a good approximation in finite samples. We apply the new models and methods to daily returns on four international equity indices, over the period 1990 to 2016, and find the proposed new ES-VaR models outperform forecasts based on GARCH or rolling window models.

The asymptotic theory presented in this chapter facilitates considering a large number of extensions of the models presented here. Our models all focus on a single value for the tail probability (α), and extending these to consider multiple values simultaneously could prove fruitful. For example, one could consider the values 0.01, 0.025 and 0.05, to capture various points in the left tail, or one could consider 0.05 and 0.95 to capture both the left and right tails simultaneously. Another natural extension is to make use of exogenous information in the model; the models proposed here are all univariate, and one might expect that information from options markets, high frequency data, or news announcements to also help predict VaR and ES. We leave these interesting extensions to future research.

Chapter 4

A Consistent Joint Test of Dynamic Models for VaR and ES

4.1 Introduction

The Basel Accord III (Basel Committee, 2010) prompts the substitution of expected shortfall (ES) for the previously well-known Value-at-Risk (VaR) as measure of risk. The main reason for this change is the long-standing criticism over VaR about its limitation in capturing the tail risk. VaR, simply the quantile of return on an asset, only tells us the minimum loss in an extreme event, without any indication of the actual loss. Differently, ES defined as the expected return on an asset conditional on the return being below the VaR, gives more information about the tail behaviour of asset returns.

This transition generates a great demand for models to estimate and forecast ES, as well as procedures of backtesting these models. However, there are not many available models and evaluation tools for ES in the literature. The paucity of models and backtesting procedures for ES is perhaps attributable by the "non-elicitability" problem of ES, see Gneiting (2011). A risk measure (or statistical functional more generally) is said to be "elicitable" if there exists a loss function (or log-likelihood) such that the measure is the solution to minimizing the expected loss. For example, the mean is elicitable using the quadratic loss function, and VaR is elicitable using the piecewise-linear or "tick" loss function. The "non-elicitability" problem of ES

restricts not only the modeling of ES, but also the testing of it.

A recent result from Fissler and Ziegel (2016), which shows that ES is jointly "elicitable" with VaR, opened a new channel to build dynamic models for ES and VaR. Patton, Ziegel, and Chen (2019) used this new result and proposed modeling directly VaR and ES without making assumption on the conditional distribution of returns. With more and more models of ES becoming available, it is important to have powerful test of these ES models.

However, there are few papers about backtesting ES. Du and Escanciano (2016) gave a review on backtesting procedures for unconditional ES. They considered tests for conditional ES instead. However, their test relies on assumption of the conditional distribution of returns, which could be avoided if the task is only to forecast and backtest ES at a single probability level. Detailed discussion about this problem will be made later.

Contrary to the dearth of tests on conditional ES, tests on conditional VaR have been well studied in the literature. Kupiec (1995) considered the unconditional test of VaR. For the conditional test, Christoffersen (1998) proposed a likelihood ratio test to check a sequence of indicator functions $\{\mathbf{1}(Y_t \leq v_t(\theta_0))\}$ is i.i.d., which is equivalent to the correct specification of the conditional VaR, if the information set only consists of past realization of indicator functions. Berkowitz, Christoffersen, and Pelletier (2011) used the Portmanteau or Ljung-Box statistics to test zero autocorrelations of the indicator functions. Escanciano and Olmo (2010) showed that these unconditional and conditional tests can be misleading due to estimation errors, and proposed a correction of these tests to make them free of estimation risk. Different from these conditional tests that only checked the informativeness of the past indicator functions, the conditional test proposed by Engle and Manganelli (2004) could incorporate a variety of functions of the past information set. However, the power of the Engle and

Manganelli (2004) test still depends on the choice of the conditioning function. This problem was solved by Escanciano and Velasco (2010), who developed consistent test of parametric dynamic conditional quantiles.

We extended their framework to jointly test conditional VaR and ES. Although they consider test of a continuum of conditional quantiles, we only focus on joint test of VaR and ES at a single probability level, $\alpha \in (0, 0.5)$, meaning the left tail. There are two reasons for this. First, testing a continuum of conditional quantiles is essentially test of conditional distribution, which relies on assumption on the entire distribution. The distributional assumption is too strong and unnecessary if the focus is only on VaR and ES at a specific probability level, for example, 1% VaR and 2.5% ES as required by the Basel Accord. Instead of using a fully parametric model, one could use the CAViaR model proposed by Engle and Manganelli (2004) for forecasting VaR at a single probability level, and the extension of the CAViaR model by Patton, Ziegel, and Chen (2019) for jointly forecasting VaR and ES at a single probability level. These models are semiparametric in the sense that they impose parametric structures for the dynamics of ES and VaR, but are completely agnostic about the conditional distribution of returns. Second, test of a continuum of conditional joint VaR and ES is not a well-defined problem on its own because the ES at any given probability level α equals the integral of all VaRs from 0 to α . Finally, it is worth mentioning that although the method in Escanciano and Velasco (2010) also applies to a unique conditional quantile case, their simulation studies did not include it. They considered left tails, but only studied the performance of test for a continuum of conditional quantiles from 0.05 to 0.2, which could not provide full information about the performance of test for 5% VaR, let alone 1% VaR, which is required by the Basel Accord. To check the performance of our test in situations where it will be used in practice, our simulation studies focus on joint test of VaR

and ES at 2.5% probability level as required by the Basel Accord, and also 1%, 5%, 10% for the sake of comparison.

Throughout the chapter, $\|\cdot\|$ refers to the Euclidean norm, unless otherwise stated.

4.2 The Test Statistics and Its Asymptotic Theory

4.2.1 The Test Statistics

Consider a sample of observations Y_1, Y_2, \dots, Y_T , and let X_t be a vector of exogenous or predetermined variables, we assume the time series process $\{(Y_t, X_t)'\} : t = 1, 2, \dots\}$, defined on the probability space (Ω, \mathcal{A}, P) , is strictly stationary and ergodic. Fix a probability level $\alpha \in (0, 0.5)$, suppose the conditional VaR and ES of Y_t at the α probability level, are $v_{t-1}(\theta_0)$ and $e_{t-1}(\theta_0)$ respectively, given the information set available at time t, which is $\mathcal{F}_{t-1} = \{Y_{t-1}, X_{t-1}, \dots, Y_1, X_1, v_0(\theta_0), e_0(\theta_0)\}$, where $v_0(\theta_0)$ and $e_0(\theta_0)$ are the initial conditions for VaR and ES. If the conditional VaR and ES are specified correctly, there exists some finite-dimensional parameter $\theta_0 \in \Theta$, with Θ a compact subset of \mathbb{R}^p , such that:

$$\mathbb{E} \left[\begin{pmatrix} \mathbf{1}\{Y_t \leq v_{t-1}(\theta_0)\} - \alpha \\ \frac{1}{\alpha} Y_t \mathbf{1}\{Y_t \leq v_{t-1}(\theta_0)\} - e_{t-1}(\theta_0) \end{pmatrix} \middle| \mathcal{F}_{t-1} \right] = \mathbf{0} \text{ a.s.} \quad (4.1)$$

Clearly, the number of random variables in the information set \mathcal{F}_{t-1} grows to infinity with the sample size. Consistent test of time series models with infinite-dimensional information set has been studied in De Jong (1996). However, their test is difficult to implement. For the practicality of our test, similar to Escanciano and Velasco

(2010), we consider instead the following null hypothesis

$$H_0 : \mathbb{E} \left[\begin{pmatrix} \mathbf{1}\{Y_t \leq v_{t-1}(\theta_0)\} - \alpha \\ \frac{1}{\alpha} Y_t \mathbf{1}\{Y_t \leq v_{t-1}(\theta_0)\} - e_{t-1}(\theta_0) \end{pmatrix} \middle| I_{t-1} \right] = \mathbf{0} \text{ a.s. for some } \theta_0 \in \Theta \subset \mathbb{R}^p \quad (4.2)$$

against the alternative

$$H_A : P \left\{ \mathbb{E} \left[\begin{pmatrix} \mathbf{1}\{Y_t \leq v_{t-1}(\theta)\} - \alpha \\ \frac{1}{\alpha} Y_t \mathbf{1}\{Y_t \leq v_{t-1}(\theta)\} - e_{t-1}(\theta) \end{pmatrix} \middle| I_{t-1} \right] \neq \mathbf{0} \right\} > 0, \text{ for all } \theta \in \Theta \subset \mathbb{R}^p \quad (4.3)$$

where $I_{t-1} \in \mathcal{F}_{t-1}$ and the dimension of I_{t-1} remains fixed (at d) as the sample size goes to infinity. To simplify the notation, denote

$$\begin{cases} \Psi_{v,\alpha,t}(\theta) = \mathbf{1}\{Y_t \leq v_{t-1}(\theta)\} - \alpha \\ \Psi_{e,\alpha,t}(\theta) = \frac{1}{\alpha} Y_t \mathbf{1}\{Y_t \leq v_{t-1}(\theta)\} - e_{t-1}(\theta) \end{cases}$$

The key idea of consistent conditional moment test originating from Bierens (1982) is that the conditional moment condition can be characterized by an infinite number of unconditional moment conditions with a proper choice of the weight function. Although Bierens (1982) used the exponential weight function, subsequent work showed that other weight functions could also generate the consistent test, see White (1989), Bierens (1990), Lee, White, and Granger (1993) etc. Bierens and Ploberger (1997) gave the conditions that a general weight function needs to satisfy in order to make the test consistent. Escanciano and Velasco (2010) used the same weight function as Bierens (1982), for its out-performance in simulation over other choices such as indicator functions. Another advantage of using this exponential weight function is

that it could deliver a simple closed-form expression for the CvM test statistic when combined with the multivariate standard normal distribution as the measure for the index of the weight function. For this sake, we follow their choice. In this case, the null hypothesis could be characterized by

$$\mathbb{E} \left[\left(\Psi_{v,\alpha,t}(\theta_0), \Psi_{e,\alpha,t}(\theta_0) \right)' \cdot \exp(ix' I_{t-1}) \right] = \mathbf{0}, \text{ for some } \theta_0 \in \Theta \subset \mathbb{R}^p \quad (4.4)$$

Now define the empirical process indexed by $x \in \mathbb{R}^d$ and $\theta \in \Theta$,

$$S_{\alpha,n}(x, \theta) := (S_{v,\alpha,n}(x, \theta), S_{e,\alpha,n}(x, \theta))' := n^{-1/2} \sum_{t=1}^n \left(\Psi_{v,\alpha,t}(\theta), \Psi_{e,\alpha,t}(\theta) \right)' \cdot \exp(ix' I_{t-1})$$

Given a \sqrt{n} -consistent estimator $\hat{\theta}_n$ of θ , we are going to form test statistics based on the empirical process $\hat{R}_{\alpha,n}(x) \equiv (\hat{R}_{v,\alpha,n}(x), \hat{R}_{e,\alpha,n}(x))' \equiv S_{\alpha,n}(x, \hat{\theta}_n)$. To aid the proof of the convergence of this empirical process, we will also consider the infeasible process $R_{\alpha,n}(x) \equiv (R_{v,\alpha,n}(x), R_{e,\alpha,n}(x))' \equiv S_{\alpha,n}(x, \theta_0)$.

The process $\hat{R}_{\alpha,n}$ is a mapping from (Ω, \mathcal{A}, P) to $\ell^\infty(\Pi)$, where $\ell^\infty(\Pi)$ is the space of all complex-valued functions that are uniformly bounded on Π , with Π a compact subset of \mathbb{R}^d having positive Lebesgue measure and containing the origin.

If the hull hypothesis is true, the process $\hat{R}_{\alpha,n}$ should be close to zero for almost all $x \in \mathbb{R}^d$. To measure the distance between $\hat{R}_{\alpha,n}$ and zero, we need a norm, with Cramer-von Mises (CvM) functional and Kolmogorov-Smirnov (KS) functional as the two most popular choices, which are defined below:

$$CvM_{\alpha,n} := \int_{\Pi} \|\hat{R}_{\alpha,n}(x)\|^2 d\mu(x) \quad (4.5)$$

where μ is some integrating measure on Π .

$$KS_{\alpha,n} := \sup_{x \in \Pi} \|\widehat{R}_{\alpha,n}(x)\|^2 \quad (4.6)$$

Follow Escanciano and Velasco (2010), we use the CvM functional to form test statistics.

4.2.2 The Limiting Distribution of the Test Statistics under the Null

This section focuses on the limiting distribution of the empirical process $\widehat{R}_{\alpha,n}(x)$ under the null hypothesis H_0 , which is obtained by combining weak convergence of $R_{\alpha,n}(x)$ and the asymptotic extension of $\widehat{R}_{\alpha,n}(x)$ around $R_{\alpha,n}(x)$. The null limiting distribution of the test statistics then follows by the continuous mapping theorem (CMT). To derive the asymptotic results we need to make some assumptions. In the assumptions below K denotes a finite constant that can change from line to line.

Assumption A1. (a): $\{(Y_t, X_t')' : t = 1, 2, \dots\}$ is a strictly stationary and ergodic process.

Under H_0 , $\left\{ \left(\Psi_{v,\alpha,t}(\theta_0), \Psi_{e,\alpha,t}(\theta_0) \right)', \mathcal{F}_t \right\}$ is a martingale difference sequence.

(b): $\mathbb{E}\|I_0\| \leq K$, $\mathbb{E}[\|I_0\|Y_1^2] \leq K$, $\mathbb{E}Y_1^2 \leq K$, $\mathbb{E}e_0^2(\theta_0) \leq K$.

(c): Conditional on all the past information \mathcal{F}_{t-1} , Y_t has continuous density $f_{t-1}(\cdot)$ with derivative $f'_{t-1}(\cdot)$, that satisfies $\sup_{y \in \mathbb{R}} f_{t-1}(y) \leq K$ and $\sup_{y \in \mathbb{R}} f'_{t-1}(y) \leq K$.

(d): The true parameter θ_0 belongs to the interior of Θ , which is a compact subset of \mathbb{R}^p .

(e): $\forall t, v_t(\theta)$ and $e_t(\theta)$ are twice continuously differentiable in θ and satisfy $e_t(\theta^0) < v_t(\theta^0) \leq 0$. (f): $\mathbb{E}[\sup_{\theta \in \Theta} \|\nabla v_{t-1}(\theta)\|] \leq K$, $\mathbb{E}[\sup_{\theta \in \Theta} \|v_{t-1}^2(\theta) \nabla v_{t-1}(\theta)\|] \leq K$.

(g): $\mathbb{E}[\sup_{\theta \in \Theta} \|\nabla^2 v_{t-1}(\theta)\|] \leq K$, $\mathbb{E}[\sup_{\theta \in \Theta} \|\nabla v_{t-1}(\theta)\|^2] \leq K$, $\mathbb{E}[\|I_{t-1}\| \sup_{\theta \in \Theta} \|\nabla v_{t-1}(\theta)\|] \leq K$, $\mathbb{E}[\sup_{\theta \in \Theta} (|v_{t-1}(\theta)| \cdot \|\nabla v_{t-1}(\theta)\|^2)] \leq K$, $\mathbb{E}[\sup_{\theta \in \Theta} (|v_{t-1}(\theta)| \cdot \|\nabla^2 v_{t-1}(\theta)\|)] \leq K$, $\mathbb{E}[\sup_{\theta \in \Theta} \|\nabla^2 e_{t-1}(\theta)\|] \leq K$, $\mathbb{E}[\|I_{t-1}\| \sup_{\theta \in \Theta} (|v_{t-1}(\theta)| \cdot \|\nabla v_{t-1}(\theta)\|)] \leq K$, and $\mathbb{E}[\|I_{t-1}\| \sup_{\theta \in \Theta} \|\nabla e_{t-1}(\theta)\|] \leq K$.

Assumption A2. *Uniformly in $x \in \Xi$,*

$$\frac{1}{n} \sum_{t=1}^n f_{t-1}(v_{t-1}(\theta_0)) \nabla v_{t-1}(\theta_0) \exp(ix' I_{t-1}) = \mathbb{E}[f_{t-1}(v_{t-1}(\theta_0)) \nabla v_{t-1}(\theta_0) \exp(ix' I_{t-1})] + o_p(1).$$

and

$$\begin{aligned} & \frac{1}{n} \sum_{t=1}^n [v_{t-1}(\theta_0) f_{t-1}(v_{t-1}(\theta_0)) \nabla v_{t-1}(\theta_0) - \nabla e_{t-1}(\theta_0)] \exp(ix' I_{t-1}) \\ &= \mathbb{E} \left[\left(v_{t-1}(\theta_0) f_{t-1}(v_{t-1}(\theta_0)) \nabla v_{t-1}(\theta_0) - \nabla e_{t-1}(\theta_0) \right) \exp(ix' I_{t-1}) \right] + o_p(1). \end{aligned}$$

Assumption A3. *Under H_0 , the estimator θ_n is \sqrt{n} -consistent and satisfies*

$$Q_n = \sqrt{n}(\hat{\theta}_n - \theta_0) = n^{-1/2} \sum_{t=1}^n l_\alpha(Y_t, I_{t-1}, \theta_0) + o_p(1) \xrightarrow{d} Q$$

where $l_\alpha(\cdot)$ is such that $\mathbb{E}[l_\alpha(Y_1, I_0, \theta_0)] = \mathbf{0}$, $L_\alpha(\theta_0) = \mathbb{E}[l_\alpha(Y_1, I_0, \theta_0) l'_\alpha(Y_t, I_0, \theta_0)]$ exists and is positive definite and $\mathbb{E}[l_\alpha(Y_1, I_0, \theta_0)(\mathbf{1}\{Y_t \leq v_{t-1}(\theta_0)\} - \alpha)] = 0$ if $t \neq s$. Furthermore, as a process in $\ell^\infty(\Pi)$, $Q_n(\alpha)$ converges weakly to a Gaussian process $Q(\cdot)$ with zero mean and covariance function

$$K_Q = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{t=1}^n \sum_{s=1}^n \mathbb{E}[l_\alpha(Y_t, I_{t-1}, \theta_0) l'_\alpha(Y_s, I_{s-1}, \theta_0)].$$

Relying on a recent result by Fissler and Ziegel (2016) that VaR and ES jointly is the minimizer of a class of loss functions, Patton, Ziegel, and Chen (2019) pro-

posed directly modeling the dynamics of conditional VaR and ES, with parameters estimated by minimizing a specific choice of the loss function, whose difference is homogeneous of degree 0. Their estimator $\hat{\theta}$ satisfies Assumption A3, with

$$l_\alpha(Y_t, v_{t-1}(\theta_0), e_{t-1}(\theta_0)) = D_n^{-1}(-g_t(\theta_0))$$

where

$$D_n = n^{-1} \sum_{t=1}^n \mathbb{E} \left\{ \nabla' v_{t-1}(\theta_0) \frac{f_{t-1}(v_{t-1}(\theta_0))}{-e_{t-1}(\theta_0)\alpha} \nabla v_{t-1}(\theta_0) + \nabla' e_{t-1}(\theta_0) \frac{1}{e_{t-1}(\theta_0)^2} \nabla e_{t-1}(\theta_0) \right\}$$

and

$$g_{t-1}(\theta) = \nabla' v_{t-1}(\theta) \frac{1}{-e_{t-1}(\theta)} \left(\frac{1}{\alpha} \mathbf{1} \{Y_t \leq v_{t-1}(\theta)\} - 1 \right) + \quad (4.7)$$

$$\nabla' e_{t-1}(\theta) \frac{1}{e_{t-1}(\theta)^2} \left(\frac{1}{\alpha} \mathbf{1} \{Y_t \leq v_{t-1}(\theta)\} (v_{t-1}(\theta) - Y_t) - v_{t-1}(\theta) + e_{t-1}(\theta) \right) \quad (4.8)$$

To derive the limiting distribution of $\widehat{R}_{\alpha,n}$, we establish the limiting distribution of the infeasible empirical process $R_{\alpha,n}$ first and then bridge them. Under Assumption A1(a) and H_0 , $R_{\alpha,n}(x)$ is a zero-mean MDS for each $x \in \Pi$, using a stable CLT for stationary ergodic MDS, we have that the finite-dimensional distribution of $R_{\alpha,n}(x)$ converges to those of a multivariate normal distribution with a zero mean vector and variance-covariance matrix, that could be characterized by the following three types

of covariances:

$$E[\Psi_{v,\alpha,t}(\theta_0) \exp(ix'_1 I_{t-1}) \cdot \Psi_{v,\alpha,t}(\theta_0) \exp(ix'_2 I_{t-1})] = (\alpha - \alpha^2) \mathbb{E}[\exp(i(x_1 + x_2)' I_0)] \quad (4.9)$$

$$E[\Psi_{v,\alpha,t}(\theta_0) \exp(ix'_1 I_{t-1}) \cdot \Psi_{e,\alpha,t}(\theta_0) \exp(ix'_2 I_{t-1})] = (1 - \alpha) E[e_0(\theta_0) \exp(i(x_1 + x_2)' I_0)] \quad (4.10)$$

$$\begin{aligned} & E[\Psi_{e,\alpha,t}(\theta_0) \exp(ix'_1 I_{t-1}) \cdot \Psi_{e,\alpha,t}(\theta_0) \exp(ix'_2 I_{t-1})] \\ &= E\left([\alpha^{-2} Y_1^2 \mathbf{1}\{Y_1 \leq v_0(\theta_0)\} - e_0^2(\theta_0)] \exp(i(x_1 + x_2)' I_0)\right) \end{aligned} \quad (4.11)$$

Under Assumption A1(b), all of the three covariances are finite.

Theorem 8. *Under the null hypothesis H_0 and Assumptions A1 (a)-(b)*

$$R_{\alpha,n} \Rightarrow R_{\alpha,\infty}$$

where $R_{\alpha,\infty}$ is a two-dimensional empirical process with zero mean and covariance function that consists of the three types of covariances given by (4.9)-(4.10).

The proof for this theorem is given in the appendix.

As emphasized before, $R_{\alpha,n}(x)$ is unobserved and we could only form the test statistic using $\widehat{R}_{\alpha,n}(x)$. The following theorem quantifies the difference between these two empirical processes. Define the function

$$G_\alpha(x) := \mathbb{E} \left[\left(\begin{array}{c} f_{t-1}(v_{t-1}(\theta_0)) \nabla v_{t-1}(\theta_0) \exp(ix' I_{t-1}) \\ (v_{t-1}(\theta_0) f_{t-1}(v_{t-1}(\theta_0)) \nabla v_{t-1}(\theta_0) - \nabla e_{t-1}(\theta_0)) \exp(ix' I_{t-1}) \end{array} \right) \right], \quad x \in \Pi$$

Theorem 9. *Under the null hypothesis H_0 and Assumptions A1-A3,*

$$\sup_{x \in \Pi} \left| \widehat{R}_{\alpha,n}(x) - R_{\alpha,n}(x) - G_{\alpha}(x) n^{-1/2} \sum_{t=1}^n l_{\alpha}(Y_t, v_{t-1}(\theta_0), e_{t-1}(\theta_0)) \right| = o_p(1).$$

Combining the results of Theorem 8 and Theorem 9, we are ready to get the limiting distribution of $\widehat{R}_{\alpha,n}(x)$.

Corollary 1. *Under the assumptions of Theorem 9,*

$$\widehat{R}_{\alpha,n} \Rightarrow \widehat{R}_{\alpha,\infty},$$

where $\widehat{R}_{\alpha,\infty}(\cdot) = R_{\alpha,\infty}(\cdot) + G_{\alpha}(\cdot)Q$ (in distribution).

With the last corollary, the asymptotic distribution of continuous functionals of $\widehat{R}_{\alpha,n}(x)$, such as CvM_n and K_n defined in (4.5) and (4.6) respectively, follows by invoking the CMT.

Corollary 2. *Under the assumptions of Theorem 9, for any continuous functional $\Gamma(\cdot)$ from $\ell^{\infty}(\Pi)$ to \mathbb{R} ,*

$$\Gamma(\widehat{R}_{\alpha,n}) \xrightarrow{d} \Gamma(\widehat{R}_{\alpha,\infty}).$$

4.2.3 Consistency against all Fixed Alternatives

In this section, we show that the CvM test considered in this chapter is consistent against all fixed alternatives.

Assumption A4. *Under H_A , (i) there exists a $\theta_1 \in \Theta$ such that $\|\theta_n - \theta_1\| = o_p(1)$; (ii) $\mathbb{E} \left[\left(\Psi_{v,\alpha,t}(\theta_1), \Psi_{e,\alpha,t}(\theta_1) \right)' \cdot \exp(ix' I_{t-1}) \right] \neq \mathbf{0}, \forall x$ in a subset with positive Lebesgue measure on Π .*

If $\hat{\theta}_n$ is chosen as the estimator (we shall call it FZ estimator because it is obtained by minimizing the loss function in Fissler and Ziegel (2016)) proposed by Patton, Ziegel, and Chen (2019), conditions on the FZ estimator to satisfy Assumption A4(i) remain an open question. Noticing that the results in Theorem 1 of Bierens and Ploberger (1997) also hold when u is a multivariate random variable, without change of anything else, we could see a sufficient condition for A4 (ii) is that I_{t-1} is bounded. If I_{t-1} is not bounded, we can without loss of generality replace I_{t-1} by $\phi(I_{t-1})$, with $\phi(I_{t-1})$ a bounded one-to-one mapping. An example of such function is $\phi(\mathbf{x}) = [\arctan(x_1), \dots, \arctan(x_d)]'$, see for example Bierens (1982).

Theorem 10. *Under the alternative hypothesis H_A and Assumptions A1, A2 and A4,*

$$n^{-1/2} \widehat{R}_{\alpha,n}(\cdot) \xrightarrow{a.s.} \mathbb{E} \left[\left(\Psi_{v,\alpha,t}(\theta_1), \Psi_{e,\alpha,t}(\theta_1) \right)' \cdot \exp(i \cdot' I_{t-1}) \right].$$

By this theorem and the CMT,

$$\int_{\Pi} \|n^{-1/2} \widehat{R}_{\alpha,n}(x)\|^2 d\Phi(x) \xrightarrow{p} \int_{\Pi} \left\| \mathbb{E} \left[\left(\Psi_{v,\alpha,t}(\theta_1), \Psi_{e,\alpha,t}(\theta_1) \right)' \cdot \exp(ix' I_{t-1}) \right] \right\|^2 d\Phi(x) > 0,$$

provided that Φ is chosen absolutely continuous with respect to Lebesgue measure. In this case, the test statistic CvM_n will go to infinity under any fixed alternative, that is to say the test considered in this chapter is consistent against all alternatives.

4.3 Subsampling Approximation

As we could see from the previous section that the limiting distribution of the test statistic under the null hypothesis is case-dependent and can therefore not be tabulated. This is a common limitation of consistent conditional moment tests. In this article, following Escanciano and Velasco (2010), we get around this problem by ap-

plying the subsampling methodology to approximate the critical values of continuous functionals of $\widehat{R}_{\alpha,n}$. The subsampling procedure is described below.

The test statistic $\Gamma(\widehat{R}_{\alpha,n})$ can be viewed as a function of the data $\{Z_t = (Y_t, X_t)'\} : t = 1, 2, \dots\}$, $\Gamma(\widehat{R}_{\alpha,n}) = \Gamma(\widehat{R}_{\alpha,n}(Z_1, \dots, Z_n))$. Let $G_{\alpha,n}^\Gamma$ be the test statistic's cdf,

$$G_{\alpha,n}^\Gamma(w) = P(\Gamma(\widehat{R}_{\alpha,n}) \leq w)$$

The key of subsampling in time series is that the order of the original sequence should be preserved, see Politis, Romano, and Wolf (1999). Suppose the full sample size is n and the sample size is set as b (we are going to discuss the choice of b later in this section), then we will use "moving blocks" of size b of consecutive observations as legitimate subsamples, the first one being (Z_1, \dots, Z_b) , and the last being (Z_{n-b+1}, \dots, Z_n) . So each subsample of size b is indeed a sample of size b from the true DGP. Let $\Gamma(\widehat{R}_{\alpha,b,i}) = \Gamma(\widehat{R}_{\alpha,b}(Z_i, \dots, Z_{i+b-1}))$ be the test statistic computed with the i -th subsample, $i = 1, 2, \dots, n-b+1$. It is quite intuitive that the sampling distribution $G_{\alpha,n}^\Gamma(w)$ can be approximated by the empirical distribution of $\Gamma(\widehat{R}_{\alpha,b,i})$. That is, we approximate $G_{\alpha,n}^\Gamma$ by

$$G_{\alpha,n,b}^\Gamma(w) = \frac{1}{n-b+1} \sum_{i=1}^{n-b+1} \mathbf{1}\{\Gamma(\widehat{R}_{\alpha,b,i}) \leq w\}, \quad w \in [0, \infty)$$

Let $c_{\alpha,n,1-\tau,b}^\Gamma$ be the $(1-\tau)$ -th sample quantile of $G_{\alpha,n,b}^\Gamma(w)$, i.e.,

$$c_{\alpha,n,1-\tau,b}^\Gamma = \inf\{w : G_{\alpha,n,b}^\Gamma(w) \geq 1 - \tau\}.$$

The null hypothesis is rejected if $\Gamma(\widehat{R}_{\alpha,n}) > c_{\alpha,n,1-\tau,b}^\Gamma$. Let $c_{\alpha,1-\tau}^\Gamma$ be the $(1-\tau)$ -th quantile of $G_{\alpha,\infty}^\Gamma(w) = P(\Gamma(\widehat{R}_{\alpha,\infty}) \leq w)$. For the theoretical justification of subsampling approximation, we need an additional assumption on the serial dependence of

the DGP. Define the α -mixing coefficient as

$$\alpha(m) = \sup_{n \in \mathbb{Z}} \sup_{B \in \mathcal{F}_n, A \in \mathcal{P}_{n+m}} |P(A \cap B) - P(A)P(B)|, \quad m \geq 1,$$

where \mathcal{F}_n and \mathcal{P}_{n+m} are σ -fields defined as $\mathcal{F}_n := \sigma(Z_t, t \leq n)$ and $\mathcal{P}_{n+m} := \sigma(Z_t, t \geq n)$, respectively, with $\{Z_t = (Y_t, X_t)'\}$.

Assumption A5. $\{Z_t = (Y_t, X_t)'\} : t = 0, \pm 1, \pm 2, \dots\}$ is a strictly stationary strong mixing process with α -mixing coefficients satisfying

$$\sum_{m=1}^n \alpha(m) = o(n)$$

The mixing condition is the same as Escanciano and Velasco (2010). The following theorem guarantees the theoretical validity of the subsampling approximation.

Theorem 11. *Assume Assumptions A1-A5 and that $b/n \rightarrow 0$ and $b \rightarrow \infty$ as $n \rightarrow \infty$. Then,*

(i) *Under the null hypothesis H_0 ,*

$$c_{\alpha, n, 1-\tau, b}^\Gamma \xrightarrow{P} c_{\alpha, 1-\tau}^\Gamma.$$

and

$$P(\Gamma(\widehat{R}_{\alpha, n}) > c_{\alpha, n, 1-\tau, b}^\Gamma) \rightarrow \tau$$

(ii) *Under any fixed alternative hypothesis,*

$$P(\Gamma(\widehat{R}_{\alpha, n}) > c_{\alpha, n, 1-\tau, b}^\Gamma) \rightarrow 1$$

Theorem 11 implies that the proposed subsampling tests have a correct asymptotic

size, are consistent and are able to detect alternatives tending to the null at the parametric rate $n^{-1/2}$.

In practice, the choice of the tuning parameter b could affect the empirical size and power of the tests. Sakov and Bickel (2000) derived the optimal choice of b as $b = \lfloor kn^{2/5} \rfloor$, where $\lfloor \cdot \rfloor$ denotes the integer part. We follow their suggestion, similar to Escanciano and Velasco (2010). Our simulation shows that this subsampling procedure provides good approximation in finite sample for a wide range of the values for k .

Before discussing the simulation results, we want to mention about a recentering method that was introduced by Chernozhukov and Fernández-Val (2005) to achieve better power for the resulting subsampling test. Following Escanciano and Velasco (2010), aside from the original subsampling critical value described above, we computed another two subsampling critical values with recentered subsampling statistics $\Gamma(\widehat{R}_{\alpha,b,i}) - \Gamma(b^{1/2}n^{-1/2}\widehat{R}_{\alpha,n})$ and $\Gamma(\widehat{R}_{\alpha,b,i} - b^{1/2}n^{-1/2}\widehat{R}_{\alpha,n})$. However, unreported simulation results suggest that in our case both recentering methods could lead to substantial size distortions. On the contrary, the subsampling test using the critical values of the original uncentered subsampling statistics $\Gamma(\widehat{R}_{\alpha,b,i})$ have good size and power properties.

4.4 Finite Sample Performance

In this section, we shall show the Monte Carlo simulation results of the proposed test $CvM_{\alpha,n}$ and compare its finite-sample performance with other related tests.

4.4.1 Comparative Test Statistics

Following Escanciano and Velasco (2010), we use the distribution function of the d -variate standard normal random vector as the integrating measure μ of the CvM functional defined in (4.5). This choice could lead to a simple closed form for the CvM test statistic $CvM_{\alpha,n}$ which equals

$$CvM_{\alpha,n} = \frac{1}{n} \Psi'_{1,\alpha} W_{\text{exp}} \Psi'_{1,\alpha} + \frac{1}{n} \Psi'_{2,\alpha} W_{\text{exp}} \Psi'_{2,\alpha},$$

where W_{exp} is a $n \times n$ matrix with elements $w_{\text{exp},t,s} = \exp(-\frac{1}{2} \|I_{t-1} - I_{s-1}\|^2)$ and $\Psi_{1,\alpha}$ and $\Psi_{2,\alpha}$ are two $n \times 1$ vectors of the sample moment condition with respect to conditional VaR and conditional ES, respectively. That is,

$$\Psi_{1,\alpha} = (\Psi_{1,\alpha,1}(\hat{\theta}_n), \dots, \Psi_{v,\alpha,n}(\hat{\theta}_n))'$$

$$\Psi_{2,\alpha} = (\Psi_{2,\alpha,1}(\hat{\theta}_n), \dots, \Psi_{e,\alpha,n}(\hat{\theta}_n))'$$

We consider three choices of I_{t-1} : $I_{t-1} = Y_{t-1}$, $I_{t-1} = (Y_{t-1}, Y_{t-2})$ and $I_{t-1} = (Y_{t-1}, \dots, Y_{t-5})$ and denote the corresponding $CvM_{\alpha,n}$ test by $CvM_{\alpha,n}(1)$, $CvM_{\alpha,n}(2)$ and $CvM_{\alpha,n}(5)$ respectively. Motivated by Escanciano and Velasco (2010), we compare this consistent test with the extension of other well-known tests to our case of jointly testing conditional VaR and ES. The first comparative test statistic is the unconditional test proposed by Kupiec (1995), which in our case can be defined as:

$$K_{\alpha,n} := \left(\frac{1}{\sqrt{n}} \sum_{t=1}^n \Psi_{v,\alpha,t}(\hat{\theta}_n) \right)^2 + \left(\frac{1}{\sqrt{n}} \sum_{t=1}^n \Psi_{e,\alpha,t}(\hat{\theta}_n) \right)^2$$

This unconditional test is equal to the CvM functional of $\widehat{R}_{\alpha,n}(\cdot)$, with the integrating measure μ in (4.5) chosen as δ_0 , which is the delta-Dirac measure at zero. To make

it clear,

$$K_{\alpha,n} \equiv \int_{\Pi} \|\widehat{R}_{\alpha,n}(x)\|^2 d\delta_0(x)$$

Applying the CMT, we could get the asymptotic distribution of K_n easily.

Another test we are going to consider is the extension of the Christoffersen (1998)'s conditional or independence test, which is defined as

$$C_{\alpha,n} := \left(\frac{1}{\sqrt{n-1}} \sum_{t=2}^n \Psi_{v,\alpha,t}(\hat{\theta}_n) \Psi_{v,\alpha,t-1}(\hat{\theta}_n) \right)^2 + \left(\frac{1}{\sqrt{n-1}} \sum_{t=1}^n \Psi_{e,\alpha,t}(\hat{\theta}_n) \Psi_{e,\alpha,t-1}(\hat{\theta}_n) \right)^2$$

The same as we do with the CvM test statistic, these two tests K_n and C_n are also implemented by the subsampling approximation because subsampling approximation could take care of the estimation risk that affects the asymptotic distribution of K_n and C_n (cf. Escanciano and Olmo (2010)).

4.4.2 DGP

For comparison, we use the same true DGP with the same true parameters as Escanciano and Velasco (2010), which is a AR(1)-GARCH(1,1) model for Y_t . This model implies the following dynamics for the conditional VaR and ES:

$$v_{t-1}(\theta_0) = \mu_0 + \mu_1 Y_{t-1} + \sigma_t \Phi^{-1}(\alpha)$$

$$e_{t-1}(\theta_0) = \mu_0 + \mu_1 Y_{t-1} + \sigma_t \cdot (-\phi(\Phi^{-1}(\alpha))/\alpha)$$

$$\text{with } \sigma_t^2 = \omega + \gamma(Y_t - \mu_0 - \mu_1 Y_{t-1})^2 + \beta \sigma_{t-1}^2 \quad (4.12)$$

where $\Phi^{-1}(\alpha)$ and $\phi(\cdot)$ are the α -quantile and density of the standard Gaussian error distribution, respectively, and the parameters $\theta_0 = (\mu_0, \mu_1, \omega, \gamma, \beta)'$ are estimated by Quasi-Maximum Likelihood (QML). Please note that we assume that the innovation's quantile is known and given by $\Phi_c^{-1}(\alpha)$. For the alternative data generating process, we consider the following:

NULL: AR(1)-GARCH(1,1) model with i.i.d $N(0,1)$ innovations: $Y_t = 0.053 - 0.092Y_{t-1} + \epsilon_t$

ALT1: ARMA(1)-GARCH(1,1) model: $Y_t = 0.053 - 0.092Y_{t-1} + \epsilon_t + 0.7\epsilon_{t-1}$

ALT2: TAR model: $Y_t = 0.6Y_{t-1} + \epsilon_t$ if $Y_{t-1} \leq 1$ and $Y_t = -0.5Y_{t-1} + \epsilon_t$ if $Y_{t-1} > 1$

ALT3: Bilinear model (BIL): $Y_t = 0.6Y_{t-1} + 0.7u_{t-1}Y_{t-2} + u_t$

ALT4: Non-linear Moving Average model (NLMA): $Y_t = 0.8u_{t-1}^2 + u_t$

ALT5: AR(1)-GARCH(1,1) model with i.i.d $t(5,-0.5)$ innovations: $Y_t = 0.053 - 0.092Y_{t-1} + v_t$

ALT6: AR(1)-GARCH(1,1) model with i.i.d $t(5,0)$ innovations: $Y_t = 0.053 - 0.092Y_{t-1} + \tilde{v}_t$

ALT7: AR(1)-EGARCH(1,1) model: $Y_t = 0.053 - 0.092Y_{t-1} + \epsilon_t$,

$$\ln \sigma_t^2 = 0.01 + 0.9 \ln \sigma_{t-1}^2 + 0.3(|u_{t-1}| - \sqrt{2/\pi}) - 0.8u_{t-1}$$

ALT1-ALT5 are the same as in Escanciano and Velasco (2010), while ALT7 was used by Du and Escanciano (2016). In models NULL and ALT1-ALT2, $\epsilon_t = \sigma_t u_t$, with σ_t as in (4.12) with $(\omega, \gamma, \beta) = (0.013, 0.104, 0.880)$ and $u_t \sim$ i.i.d $N(0, 1)$. In ALT6, $v_t = \sigma_t w_t$, with σ_t as before and $w_t \sim$ i.i.d Hansen's skewed $t(5,-0.5)$. In ALT7, $\tilde{v}_t = \sigma_t \tilde{w}_t$, with σ_t as before and $\tilde{w}_t \sim$ i.i.d Hansen's skewed $t(5,0)$.

We consider two sample sizes $T = 1000$ and $T = 2500$ ¹ and four probability levels $\alpha = 1\%$, 2.5% , 5% , 10% . These probability levels are picked because 2.5% ES is required by the Basel Accord III and 1% VaR was used previously. As for

¹To eliminate the possible effect of the starting values, we simulate in each replication a time series doubling the sample size and drop the first half of the sample.

the number of subsamples, we follow the suggestion of Sakov and Bickel (2000) and set it to $b = \lfloor kn^{2/5} \rfloor$, with a wide range for the values of k , which is $k=6$ and 16 . These values correspond to $b = 95$ and 253 for $T = 1000$ and $b = 137$ and 365 for $T = 2500$, respectively. The simulation results are based on 1,000 replications throughout the chapter. Therefore, the maximal simulation standard error is $\max_{0 \leq p \leq 1} \sqrt{p(1-p/1000)} \approx 0.016$. In all experiments, the significant level is 0.05. Other choices of the significant level, such as 1% and 10%, is left for robustness check. All the simulation results are replicable and the code is available upon request.

4.4.3 Discussion on the Results

As emphasized before, different from Escanciano and Velasco (2010) which considers joint test of the conditional VaR in a continuum of probability levels, our focus is on the joint test for the conditional VaR and ES at a single probability level. In unreported simulation results, we observed that both the empirical size and power for joint test of the conditional VaR in a continuum of probability levels (we consider the average over the four probability levels) is mainly driven by that for the highest probability level. This result strengthens the superiority of test at a single probability level over in a continuum of probability levels in this case.

In Table 4.1-4.4 we report the rejection rates of the tests based on CvM_n, K_n, C_n for models NULL and ALT1-ALT7, for each of the four probability levels. The empirical size performance is satisfactory for CvM_n and K_n , especially with larger sample size and higher probability levels. However, for C_n , there are substantial upward size distortions, and the bias even increases with the sample size at low probability levels 1% and 2.5%. Considering its over-rejection, we drop C_n from the comparison pool at probability levels 1%, 2.5% and 5%, only keeping it at 10%

probability level when the size distortion disappears. Also, we noticed that the CvM test using 5 lags of Y_t in the information set has under rejection under the null DGP. This is perhaps due to the fact that the AR(1) model only depends on the first lag of Y_t . Since we have shown in the previous section that the CvM test proposed in this chapter is consistent against any fixed alternatives, we should expect that the CvM test has higher power than other tests under all, or at least some alternatives. For the convenience of comparison, we highlighted the test with the highest rejection rates under each of the 7 alternatives. The CvM_n test wins in ALT1 with more serial dependence and in ALT2-ALT3 of nonlinear time series. The empirical power of both the CvM_n test and the K_n test are equally good under ALT4. The K_n test out-performs the CvM_n test under ALT5 and ALT6. This is non-surprising because ALT5 and ALT6 only differs from the null DGP in the distribution of the innovations, resulting in a constant derivation from zero of the moment condition w.r.t both conditonal VaR and ES and the unconditional test usually captures well the constant difference. The rejection rates increase with the sample size, especially for the CvM_n test, confirming its consistency against these fixed alternatives.

Table 4.1: Empirical rejection rates for jointly testing 1% VaR and ES at 5% significant level based on 1000 simulations

n	k	Test	NULL	ALT1	ALT2	ALT3	ALT4	ALT5	ALT6	ALT7
1000	6	K_n	8.3	9.3	16.3	6.6	79.5	86	29.9	62.3
		C_n	28.6							
		$CvM_n(1)$	9.8	8.2	50.4	5.5	62.6	81.4	19.4	44.4
		$CvM_n(2)$	4.9	9.4	37.1	3.3	25.9	75.6	11.7	23
		$CvM_n(5)$	0	0.7	5.8	1.6	0.3	47.6	2.3	1.7
	16	K_n	5.4	6	11.1	7.3	60.8	77.8	21.9	55.8
		C_n	19.2							
		$CvM_n(1)$	6.1	7.1	42.6	10.8	69.7	77.6	18.9	53.2
		$CvM_n(2)$	6.4	12.3	34.9	8.6	45.1	76.6	14	37.2
		$CvM_n(5)$	1.1	3.1	9.8	2.9	1.9	61.4	6.6	5.5
2500	6	K_n	7	5.6	27.9	7.3	98.8	97	65	88.8
		C_n	51.4							
		$CvM_n(1)$	5.4	4.5	85.5	10.4	99.2	95	48.3	73
		$CvM_n(2)$	1.8	13.3	71.5	6	77	91.4	31.9	41.1
		$CvM_n(5)$	0	0.6	10.7	0.5	0	73.5	8.2	1.6
	16	K_n	6.2	5.5	17.8	7.5	94.2	90.4	49.6	86.3
		C_n	30.8							9.2
		$CvM_n(1)$	5.9	4.1	77	16.3	97.5	92.7	44.1	84.7
		$CvM_n(2)$	5	18.2	70	16.5	83.7	92	36.5	66.5
		$CvM_n(5)$	0.7	2.9	19.1	1.8	3.9	83.1	18	8.2

Table 4.2: Empirical rejection rates for jointly testing 2.5% VaR and ES at 5% significant level based on 1000 simulations

n	k	Test	NULL	ALT1	ALT2	ALT3	ALT4	ALT5	ALT6	ALT7
1000	6	K_n	6.1	6.2	13.7	4.6	95.7	85.8	12.9	54.4
		C_n	26							
		CvM_n (1)	5	5.3	60.9	12.1	98.7	78.2	7.2	46.8
		CvM_n (2)	2.6	15.5	52	10.2	80.1	71.5	4.6	23.4
		CvM_n (5)	0.4	1	6.7	1.3	1.2	41.6	1.2	0.9
	16	K_n	6.2	6.6	11.1	4.5	86.8	75.9	12.5	51
		C_n	16.6							
		CvM_n (1)	5.7	5.4	53.1	18.6	96.2	74.4	9.7	56.8
		CvM_n (2)	4.9	20.3	51.3	21.9	84	73.3	7.9	42.3
		CvM_n (5)	1.8	5.2	15.2	3.5	10.6	56.1	3.9	6.2
2500	6	K_n	5.8	5.8	22.6	4.4	100	97.1	26.2	88
		C_n	23.9							
		CvM_n (1)	3.5	4.3	95.3	27.9	100	93.5	13.1	76.7
		CvM_n (2)	1.6	41.3	87.8	27.9	100	89.3	7.4	48.2
		CvM_n (5)	0	2.7	22	0.7	19.4	67	1.4	1.4
	16	K_n	4.9	5.5	15.5	4.3	99.9	89.3	17.3	85.3
		C_n	17.7							
		CvM_n (1)	3.1	4.5	91.4	35.7	100	91.5	13.5	87.5
		CvM_n (2)	2.5	47.9	86.8	52	100	91	10.3	74.2
		CvM_n (5)	0.6	13.4	38.3	4.3	64.3	79.5	3.8	13.1

Table 4.3: Empirical rejection rates for jointly testing 5% VaR and ES at 5% significant level based on 1000 simulations

n	k	Test	NULL	ALT1	ALT2	ALT3	ALT4	ALT5	ALT6	ALT7
1000	6	K_n	6.4	6.1	11.9	8.8	100	79.7	6	43.6
		C_n	12.3							
		CvM_n (1)	4.5	4.5	72.4	24	100	67.9	3.8	44.1
		CvM_n (2)	2.2	27.3	60.8	26.9	98.9	57.1	1	24.9
		CvM_n (5)	0.1	3.2	6.3	0.9	11	24.5	0.2	1.2
	16	K_n	6.1	5.3	9.7	8.3	98.1	65	6	44.2
		C_n	10.6							
		CvM_n (1)	5.1	5.8	65.8	30.4	100	61.9	5.1	57.4
		CvM_n (2)	4.3	33.7	63.5	45.7	98.8	59.3	2.7	45.7
		CvM_n (5)	1.6	10.5	18.7	4.8	44.9	42	1.6	6.2
2500	6	K_n	6.1	4.9	21	11.3	100	95.6	6.1	82.8
		C_n	9.9							
		CvM_n (1)	2.8	3.3	96.5	49.1	100	87.9	2.8	76.1
		CvM_n (2)	0.9	68.7	92	55.5	100	80.6	1	50.2
		CvM_n (5)	0	14.7	28.3	1.2	97.9	48.8	0	2
	16	K_n	4.6	4.6	12.6	10	100	85.6	6.1	79.4
		C_n	9.1							
		CvM_n (1)	3.6	3.9	94.1	60.6	100	84.6	4	89.1
		CvM_n (2)	2.1	74.2	92.5	76.3	100	82.8	1.7	76.6
		CvM_n (5)	0.5	33.4	46.7	10.3	98.9	66	0.4	14.7

Table 4.4: Empirical rejection rates for jointly testing 10% VaR and ES at 5% significant level based on 1000 simulations

n	k	Test	NULL	ALT1	ALT2	ALT3	ALT4	ALT5	ALT6	ALT7
1000	6	K_n	7.1	7.4	13.5	19.7	100	46.1	25.6	23.2
		C_n	7.5	7.8	10.7	33.1	71.5	6	5.6	14.7
		CvM_n (1)	3.4	2.5	77.3	39.7	100	30.3	18.6	40.2
		CvM_n (2)	0.8	46.8	65.4	50.9	100	18.3	9.5	25
		CvM_n (5)	0.1	8.1	7.1	1.3	66.7	3.3	1	1.1
	16	K_n	6	5.4	8.9	16.1	99.8	35.1	17.7	25
		C_n	7.4	8.6	10.9	36.8	62.2	8	5.5	16.1
		CvM_n (1)	4.9	5.2	71.7	47.3	100	29.7	17.8	56.9
		CvM_n (2)	3	50.2	68.4	72.5	100	24.8	13	46.8
		CvM_n (5)	1.1	20.1	22.2	9.5	88.9	13.1	4.5	7.2
2500	6	K_n	5.4	5.1	20.6	39	100	75.1	52.4	58.8
		C_n	4.6	11.2	8.8	48.1	96.6	6.8	5.5	24.7
		CvM_n (1)	2.4	2.4	97.8	69.3	100	51.7	39.1	74.2
		CvM_n (2)	0.4	87.6	93.9	74.6	100	35.6	22.7	53
		CvM_n (5)	0	34.4	31.6	6.8	100	10.9	2.5	1.6
	16	K_n	4.3	5.1	9.7	33.9	100	54.9	42.6	56.7
		C_n	4.7	7.9	6.7	50.5	93.4	5.1	5.4	24.6
		CvM_n (1)	3.3	2.2	95.6	81.9	100	47.4	39.1	88.2
		CvM_n (2)	1.6	90.1	94	88.2	100	41.5	30.6	78.5
		CvM_n (5)	0.7	58.5	55.6	27.8	100	24.1	8.2	14.8

4.4.4 Regression-based Test

Patton, Ziegel, and Chen (2019) run the following regression (they call it "DQ" and "DES" regression, respectively):

$$\Psi_{v,\alpha,t}(\hat{\theta}_n) = a_0 + a_1\Psi_{v,\alpha,t-1}(\hat{\theta}_n) + a_2v_{t-1}(\hat{\theta}_n) + u_{v,t} \quad (4.13)$$

$$\Psi_{e,\alpha,t}^s(\hat{\theta}_n) = b_0^s + b_1^s\Psi_{e,\alpha,t-1}^s(\hat{\theta}_n) + b_2^se_{t-1}(\hat{\theta}_n) + u_{e,t}^s \quad (4.14)$$

where $\Psi_{e,\alpha,t}^s(\theta) = \frac{\Psi_{e,\alpha,t}(\theta)}{e_t(\theta)}$. This standardization is performed to reduce its possible serial correlation (which comes through the persistence of v_t and e_t). To analyze the impact of this standardization, we consider both their regression and the regression with the original $\Psi_{e,\alpha,t}(\theta)$. Based on regression 4.13 and 4.14, Patton, Ziegel, and Chen (2019) tested $a_0 = a_1 = a_2 = 0$ and $b_0 = b_1 = b_2 = 0$ (against the usual two-sided alternatives) separately to determine whether the dynamics of conditional VaR or conditional ES is well/poorly specified. To compare with the CvM test proposed by this chapter, we also perform joint test of $a_0 = a_1 = a_2 = 0$ and $b_0 = b_1 = b_2 = 0$ (still, against the usual two-sided alternatives). Inspired by the unconditional test and the Christoffersen test, we also consider the following two sets of regressions:

$$\Psi_{v,\alpha,t}(\hat{\theta}_n) = a_0 + u_{v,t} \quad (4.15)$$

$$\Psi_{e,\alpha,t}^s(\hat{\theta}_n) = b_0^s + u_{e,t}^s \quad (4.16)$$

and

$$\Psi_{v,\alpha,t}(\hat{\theta}_n) = a_0 + a_1\Psi_{v,\alpha,t-1}(\hat{\theta}_n) + u_{v,t} \quad (4.17)$$

$$\Psi_{e,\alpha,t}^s(\hat{\theta}_n) = b_0^s + b_1^s\Psi_{e,\alpha,t-1}^s(\hat{\theta}_n) + u_{e,t}^s \quad (4.18)$$

The same as regression (4.14), we also run regression (4.16) and (4.18) with both the original $\Psi_{e,\alpha,t}(\theta)$ and $\Psi_{e,\alpha,t}^s(\theta)$. In all the tests, the covariance matrix is calculated by the heteroskedasticity and autocorrelation consistent (HAC) estimator.

As we have discussed in the previous subsection, under our setting, the Christoffersen (1998) type test has severe size distortion, which disappears only when the probability level raises up to 10%. So it is important to check whether the regression-based test we are considering here also has the size issue. Table 5 reports the empirical size of both the individual and joint test of 1%, 2.5%, 5%, 10% VaR/ES in the 6 regressions at 5% significant level based on 1,000 simulations.

For regression 1 and 1^s, the test is regression-based test of unconditional moment condition w.r.t. VaR/ ES (standardized and unstandardized). As the results shows, standardization of ES does not create significant difference to the testing result. The size of individual test of unconditional moment of VaR/ES only is within the confidence interval for probability level $\alpha = 1\%$. However, there is more and more under-rejection as the probability level goes up and the downward bias in the empirical size could not go away when the sample size increase from 1,000 to 2,500. As for the joint test, there is over-rejection instead at 1% probability level, and also under-rejection at 5% and 10% probability level. The overall performance of regression-based unconditional test is not as satisfactory as the previous unconditional test, where both the over-rejection at 1% probability level and under-rejection at 5% and 10% level are not obvious. This finding is consistent with the result of Escanciano and Olmo (2010), which shows that the performance of widely used unconditional and conditional test of models of VaR could be greatly affected by the estimation error. This also indicates that the good performance of the previous unconditional test should be contributed to subsampling, which effectively takes care of the estimation error.

The test associated with regression 2 and 2^s is the regression-based Christoffersen's correlation test of moments of VaR/ES. As we have discussed before, the Christoffersen's correlation test has severe size distortion at low probability levels, which could not even be remedied by subsampling. Not surprisingly, similar over-rejection also exists in the regression-based test. If we focus on the joint test of VaR and ES so that the result from the regression-based test can be directly compared with our previous result, we could see that the size distortion is more severe in the regression-based test than in the previous correlation test. This is another evidence of how estimation error could impart the performance of conditional test of VaR/ES models.

As for regression 3 and 3^s which is extended from Engle and Manganelli (2004)'s dynamic quantile regression and conducted in Patton, Ziegel, and Chen (2019), we observe similar or even more severe size distortion problem than in regression 2 and 2^s.

In summary, all the regression-based test considered here have size distortion that are brought by estimation error so researchers and empiricists should use them with caution. The failure of these regression-based test as well as the previous subsampling-based Christoffersen test further highlights the comparative advantage of our test in detecting violations of the conditional moment condition of joint models of VaR and ES.

Table 4.5: Empirical size for regression-based test of $\alpha = 1\%, 2.5\%, 5\%, 10\%$ VaR and ES at 5% significant level based on 1000 simulations

Null hypothesis	T=1000				T=2500			
	$\alpha = 0.01$	0.025	0.050	0.100	$\alpha = 0.01$	0.025	0.050	0.100
Regression 1:	$\Psi_{v,\alpha,t}(\hat{\theta}_n) = a_0 + u_{v,t}$ $\Psi_{e,\alpha,t}(\hat{\theta}_n) = b_0 + u_{e,t}$							
$a_0 = 0$	4.0	2.4	1.2	0.2	3.8	2.4	0.9	0.6
$b_0 = 0$	5.0	3.4	1.4	0.6	4.9	3.0	1.8	0.6
$a_0 = b_0 = 0$	10.7	5.5	3.6	1.7	7.6	4.2	2.8	1.8
Regression 1 ^s :	$\Psi_{v,\alpha,t}(\hat{\theta}_n) = a_0 + u_{v,t}$ $\Psi_{e,\alpha,t}^s(\hat{\theta}_n) = b_0^s + u_{e,t}^s$							
$b_0^s = 0$	4.5	2.0	1.2	0.3	3.4	1.9	0.8	0.1
$a_0 = b_0^s = 0$	13.7	6.4	3.5	1.6	8.0	4.3	2.8	1.9
Regression 2:	$\Psi_{v,\alpha,t}(\hat{\theta}_n) = a_0 + a_1\Psi_{v,\alpha,t-1}(\hat{\theta}_n) + u_{v,t}$ $\Psi_{e,\alpha,t}(\hat{\theta}_n) = b_0 + b_1\Psi_{e,\alpha,t-1}(\hat{\theta}_n) + u_{e,t}$							
$a_0 = a_1 = 0$	44.7	29.4	5.9	3.9	28.7	11.6	6.4	1.5
$b_0 = b_1 = 0$	90	55.9	21.6	8.3	78.3	30.6	12.7	5.6
$a_0 = a_1 = b_0 = b_1 = 0$	46.8	51.6	38.9	12.5	37.5	47.4	23.2	8.4
Regression 2 ^s :	$\Psi_{v,\alpha,t}(\hat{\theta}_n) = a_0 + a_1\Psi_{v,\alpha,t-1}(\hat{\theta}_n) + u_{v,t}$ $\Psi_{e,\alpha,t}^s(\hat{\theta}_n) = b_0^s + b_1^s\Psi_{e,\alpha,t-1}^s(\hat{\theta}_n) + u_{e,t}^s$							
$b_0^s = b_1^s = 0$	90.6	55.1	12.5	4.2	78.1	21.9	6.4	1.9
$a_0 = a_1 = b_0^s = b_1^s = 0$	49.1	55.2	41.3	13.8	39.1	52.8	19.1	5.6
Regression 3:	$\Psi_{v,\alpha,t}(\hat{\theta}_n) = a_0 + a_1\Psi_{v,\alpha,t-1}(\hat{\theta}_n) + a_2v_{t-1}(\hat{\theta}_n) + u_{v,t}$ $\Psi_{e,\alpha,t}(\hat{\theta}_n) = b_0 + b_1\Psi_{e,\alpha,t-1}(\hat{\theta}_n) + b_2e_{t-1}(\hat{\theta}_n) + u_{e,t}$							
$a_0 = a_1 = a_2 = 0$	90.7	56.7	12.1	4.5	79.2	23.5	6.7	1.9
$b_0 = b_1 = b_2 = 0$	91.7	59.6	26.4	11.8	81.2	35	17	7.9
$a_0 = a_1 = a_2 = b_0 = b_1 = b_2 = 0$	95.1	84.2	50.9	21.6	90.5	66	32.7	14.6
Regression 3 ^s :	$\Psi_{v,\alpha,t}(\hat{\theta}_n) = a_0 + a_1\Psi_{v,\alpha,t-1}(\hat{\theta}_n) + a_2v_{t-1}(\hat{\theta}_n) + u_{v,t}$ $\Psi_{e,\alpha,t}^s(\hat{\theta}_n) = b_0^s + b_1^s\Psi_{e,\alpha,t-1}^s(\hat{\theta}_n) + b_2^se_{t-1}(\hat{\theta}_n) + u_{e,t}^s$							
$b_0^s = b_1^s = b_2^s = 0$	90.9	56.7	14.3	4.5	79.1	23.4	7.2	2.5
$a_0 = a_1 = a_2 = b_0^s = b_1^s = b_2^s = 0$	94.1	81.6	48.1	15.6	88.4	63.8	19.3	6.7

4.5 Proofs

In the proof, K denotes a positive constant, θ_* and x_* denote median values of the parameter θ , and the index of the related empirical process, respectively. All of them could change from line to line. Before giving the proof, we introduce some shorthand notation. $\mathbb{E}_{t-1}[\cdot] := \mathbb{E}[\cdot | \mathcal{F}_{t-1}]$.

The proof is based on Delgado and Escanciano (2007), which established weak convergence of a large class of empirical process that allows to be non-differentiable under general serial dependence, taking advantage of martingale difference conditions. By the fact that the asymptotic tightness of a multivariate empirical process follows by the asymptotic tightness in each dimension, we could extend their results from an univariate empirical process to a multivariate case.

We are going to state the extended weak convergence theorem first. Let for each $n \geq 1$, $I'_{n,0}, \dots, I'_{n,n-1}$, be an array of random vectors in \mathbb{R}^p , $p \in \mathbb{N}$, and $Y_{n,1}, \dots, Y_{n,n}$, be an array of real random variables (r.v.'s). Denote by $(\Omega_n, \mathcal{A}_n, P_n)$, $n \geq 1$, the probability space in which all the r.v.'s $\{Y_{n,t}, I'_{n,t-1}\}_{t=1}^n$ are defined. Let $\mathcal{F}_{n,t}$, $0 \leq t \leq n$, be a double array of sub σ -fields of \mathcal{A}_n such that $\mathcal{F}_{n,t-1} \subset \mathcal{F}_{n,t}$, $t = 1, \dots, n$. $\left\{ \left(w_1(Y_{n,t}, I_{n,t-1}, \gamma), w_2(Y_{n,t}, I_{n,t-1}, \gamma) \right)' \right\}_{t=0}^n$ is a bivariate square-integrable martingale difference sequence for each $\gamma \in \mathcal{H}$, that is,

$$\mathbb{E} \left[\left(w_1(Y_{n,t}, I_{n,t-1}, \gamma), w_2(Y_{n,t}, I_{n,t-1}, \gamma) \right)' \middle| \mathcal{F}_{n,t-1} \right] = \mathbf{0} \text{ a.s.}, 1 \leq t \leq n, \forall n \geq 1,$$

$\mathbb{E}w_1^2(Y_{n,t}, I_{n,t-1}, \gamma) < \infty$, $\mathbb{E}w_2^2(Y_{n,t}, I_{n,t-1}, \gamma) < \infty$, $w_1(Y_{n,t}, I_{n,t-1}, \gamma)$ and $w_2(Y_{n,t}, I_{n,t-1}, \gamma)$ are $\mathcal{F}_{n,t}$ -measurable for each $\gamma \in \mathcal{H}$, and $\forall t, 1 \leq t \leq n, \forall n \in \mathbb{N}$. The following result

gives sufficient conditions for the weak convergence of the empirical process

$$\alpha_{n,w}(\gamma) = \begin{pmatrix} n^{-1/2} \sum_{t=1}^n w_1(Y_{n,t}, I_{n,t-1}, \gamma) \\ n^{-1/2} \sum_{t=1}^n w_2(Y_{n,t}, I_{n,t-1}, \gamma) \end{pmatrix}, \gamma \in \mathcal{H}.$$

An important role in the weak convergence theorem is played by the conditional quadratic variation (CV) of the empirical process $\alpha_{n,w}$ on a finite partition $\mathcal{B} = \{H_k, 1 \leq k \leq N\}$ of \mathcal{H} , which in each dimension, is defined as

$$CV_{n,w_i}(\mathcal{B}) = \max_{1 \leq k \leq N} n^{-1} \sum_{t=1}^n E \left[\sup_{\gamma_1, \gamma_2 \in H_k} |w(Y_{n,t}, I_{n,t-1}, \gamma_1) - w(Y_{n,t}, I_{n,t-1}, \gamma_2)|^2 | \mathcal{F}_{n,t-1} \right], i = 1, 2. \quad (4.19)$$

For the weak convergence theorem we need the following two assumptions.

W1: For each $n \geq 1$, $\{(Y_{n,t}, I'_{n,t-1})\}_{t=1}^n$ is a strictly stationary and ergodic process. The sequence $\left\{ \left(w_1(Y_{n,t}, I_{n,t-1}, \gamma), w_2(Y_{n,t}, I_{n,t-1}, \gamma) \right)', \mathcal{F}_{n,t} \right\}_{t=0}^n$ is a bivariate square-integrable martingale difference sequence for each $\gamma \in \mathcal{H}$. Also, there exists three functions $C_{w_1}(\gamma_1, \gamma_2)$, $C_{w_2}(\gamma_1, \gamma_2)$, and $C_{w_1, w_2}(\gamma_1, \gamma_2)$ on $\mathcal{H} \times \mathcal{H}$ to \mathbb{R} such that uniformly in $(\gamma_1, \gamma_2) \in \mathcal{H} \times \mathcal{H}$

$$n^{-1} \sum_{t=1}^n w_1(Y_{n,t}, I_{n,t-1}, \gamma_1) w_1(Y_{n,t}, I_{n,t-1}, \gamma_2) = C_{w_1}(\gamma_1, \gamma_2) + o_{P_n}(1).$$

$$n^{-1} \sum_{t=1}^n w_1(Y_{n,t}, I_{n,t-1}, \gamma_1) w_2(Y_{n,t}, I_{n,t-1}, \gamma_2) = C_{w_1, w_2}(\gamma_1, \gamma_2) + o_{P_n}(1).$$

$$n^{-1} \sum_{t=1}^n w_2(Y_{n,t}, I_{n,t-1}, \gamma_1) w_2(Y_{n,t}, I_{n,t-1}, \gamma_2) = C_{w_2}(\gamma_1, \gamma_2) + o_{P_n}(1).$$

W2: $\alpha_{n,w}$ is a mapping from Ω_n to $\ell^\infty(\mathcal{H})$ and for every $\delta > 0$ there exists a finite partition $\mathcal{B}_\delta = \{H_k; 1 \leq k \leq N_\delta\}$ of \mathcal{H} , with N_δ being the number of elements

of such partitions, such that

$$\int_0^\infty \sqrt{\log(N_\delta)} d\delta < \infty \quad (4.20)$$

and

$$\sup_{\delta \in (0,1) \cap \mathbb{Q}} \frac{CV_{n,w_i}(\mathcal{B}_\delta)}{\delta^2} = O_{P_n}(1), \quad i = 1, 2. \quad (4.21)$$

Let $\alpha_{\infty,w}(\cdot)$ be a Gaussian process with zero mean and covariance function that could be characterized by the three functions $C_{w_1}(\gamma_1, \gamma_2)$, $C_{w_2}(\gamma_1, \gamma_2)$, and $C_{w_1, w_2}(\gamma_1, \gamma_2)$. Now we could state the weak convergence of $\alpha_{n,w}(\cdot)$.

Theorem A1. *If Assumptions W1 and W2 hold, then it follows that*

$$\alpha_{n,w} \Rightarrow \alpha_{\infty,w} \text{ in } \ell^\infty(\mathcal{H})$$

proof of Theorem A1. *This theorem is simply an extension of Theorem A1 in Delgado and Escanciano (2007) to the multivariate case.*

proof of Theorem 1.: *We shall apply Theorem A1 to $R_{\alpha,n}(x)$. Assumption W1 is easy to verify. The three covariance functions are given by (4.9)-(4.10) and they are finite by Assumption A1(c). The rest of the proof will focus on Assumption W2. Let $N_{\square}(\delta, \Pi, \|\cdot\|)$ be the minimum number of δ -brackets needed to cover Π . Then, $\mathcal{B}_\delta = \{H_k : 1 \leq k \leq N_{\square}(\delta, \Pi, \|\cdot\|)\}$ is a finite partition of Π . Since Π is assumed to be compact subset of \mathbb{R}^d , it is easy to show (see example 19.7 of Van der Vaart (1998)) that*

$$\int_0^\infty \sqrt{\log N_{\square}(\delta, \Pi, \|\cdot\|)} d\delta < \infty$$

Given this partition, we are going to bound $CV_{n,w_1}(\mathcal{B}_\delta)$ and $CV_{n,w_2}(\mathcal{B}_\delta)$ separately.

$$\begin{aligned}
& \sup_{x_1, x_2 \in H_k} |w_1(Y_t, I_{t-1}, x_1) - w_1(Y_t, I_{t-1}, x_2)|^2 \\
& \leq \left| \mathbf{1}\{Y_t \leq v_{t-1}(\theta_0)\} - \alpha \right|^2 \cdot \sup_{x_1, x_2 \in H_k} \left| \exp(ix'_1 I_{t-1}) - \exp(ix'_2 I_{t-1}) \right|^2 \\
& \leq \sup_{x_1, x_2 \in H_k} \left| \exp(ix'_1 I_{t-1}) - \exp(ix'_2 I_{t-1}) \right|^2 \\
& \leq \sup_{x_1, x_2 \in H_k} |x'_1 I_{t-1} - x'_2 I_{t-1}|^2 \\
& \leq |I_{t-1}|^2 \sup_{x_1, x_2 \in H_k} \|x_1 - x_2\|^2 \\
& \leq \|I_{t-1}\|^2 \delta^2
\end{aligned}$$

Thus, $CV_{n,w_1}(\mathcal{B}_\delta)$ defined in (4.19) is bounded by $\delta^2 n^{-1} \sum_{t=1}^n \|I_{t-1}\|^2$.

$$\begin{aligned}
& \mathbb{E}_{t-1} \left[\sup_{x_1, x_2 \in H_k} |w_2(Y_t, I_{t-1}, x_1) - w_2(Y_t, I_{t-1}, x_2)|^2 \right] \\
& \leq \sup_{x_1, x_2 \in H_k} \left| \exp(ix'_1 I_{t-1}) - \exp(ix'_2 I_{t-1}) \right|^2 \cdot \mathbb{E}_{t-1} \left| \alpha^{-1} Y_t \mathbf{1}\{Y_t \leq v_{t-1}(\theta_0)\} - e_{t-1}(\theta_0) \right|^2 \\
& \leq \delta^2 \|I_{t-1}\|^2 \cdot \alpha^{-2} \mathbb{E}_{t-1} Y_t^2
\end{aligned}$$

where the last inequality follows because

$$\begin{aligned}
& \mathbb{E}_{t-1} \left| \alpha^{-1} Y_t \mathbf{1}\{Y_t \leq v_{t-1}(\theta_0)\} - e_{t-1}(\theta_0) \right|^2 \\
& \leq \mathbb{E}_{t-1} \left| \alpha^{-1} Y_t \mathbf{1}\{Y_t \leq v_{t-1}(\theta_0)\} \right|^2 - e_{t-1}^2(\theta_0) \\
& \leq \alpha^{-2} \mathbb{E}_{t-1} Y_t^2
\end{aligned}$$

Thus, $CV_{n,w_2}(\mathcal{B}_\delta)$ defined in (4.19) is bounded by $\delta^2 \alpha^{-2} n^{-1} \sum_{t=1}^n \|I_{t-1}\|^2 \mathbb{E}_{t-1} Y_t^2$. Therefore, W2 of Theorem A1 holds under Assumption A1 (a)-(b) and the theorem is proved.

Theorem A2. *Assume Assumptions A1(a)-(f) and that there exist a $\theta_1 \in \Theta$ such that $\|\theta_n - \theta_1\| = o_p(1)$. Then, uniformly in $x \in \Pi$,*

$$\begin{aligned}
\widehat{R}_{v,\alpha,n}(x) &:= n^{-1/2} \sum_{t=1}^n \Psi_{v,\alpha,t}(\widehat{\theta}_n) \exp(ix' I_{t-1}) \\
&= n^{-1/2} \sum_{t=1}^n \{\Psi_{v,\alpha,t}(\theta_1) - \mathbb{E}_{t-1} \Psi_{v,\alpha,t}(\theta_1)\} \exp(ix' I_{t-1}) \\
&\quad + n^{-1/2} \sum_{t=1}^n \mathbb{E}_{t-1} \Psi_{v,\alpha,t}(\widehat{\theta}_n) \exp(ix' I_{t-1}) + o_p(1). \tag{4.22}
\end{aligned}$$

and

$$\begin{aligned}
\widehat{R}_{e,\alpha,n}(x) &:= n^{-1/2} \sum_{t=1}^n \Psi_{e,\alpha,t}(\widehat{\theta}_n) \exp(ix' I_{t-1}) \\
&= n^{-1/2} \sum_{t=1}^n \{\Psi_{e,\alpha,t}(\theta_1) - \mathbb{E}_{t-1} \Psi_{e,\alpha,t}(\theta_1)\} \exp(ix' I_{t-1}) \\
&\quad + n^{-1/2} \sum_{t=1}^n \mathbb{E}_{t-1} \Psi_{e,\alpha,t}(\widehat{\theta}_n) \exp(ix' I_{t-1}) + o_p(1). \tag{4.23}
\end{aligned}$$

proof of Theorem A2. *Define the metric on $\mathcal{W} = \Pi \times \Theta$ as*

$$d(p_1, p_2) := \sqrt{\|\theta_1 - \theta_2\| + \|x_1 - x_2\|^2}, \quad (p_1, p_2) \in \mathcal{W}.$$

Let $N_{\square}(\delta, \mathcal{W}, d)$ be the minimum number of δ -brackets needed to cover \mathcal{W} . Then, $\mathcal{B}_\delta = \{H_k : 1 \leq k \leq N_{\square}(\delta, \Pi, \|\cdot\|)\}$ is a finite partition of \mathcal{W} . Since Π is assumed to be compact subset of \mathbb{R}^d , and Θ a compact subset of \mathbb{R}^p , it is also easy to show that

$$\int_0^\infty \sqrt{\log N_{\square}(\delta, \mathcal{W}, d)} \, d\delta < \infty$$

Write $w_{1,t-1}(x, \theta) := \{\Psi_{v,\alpha,t}(\theta) - \mathbb{E}_{t-1}\Psi_{v,\alpha,t}(\theta_1)\} \exp(ix'I_{t-1})$.

$$w_{2,t-1}(x, \theta) := \{\Psi_{e,\alpha,t}(\theta) - \mathbb{E}_{t-1}\Psi_{e,\alpha,t}(\theta_1)\} \exp(ix'I_{t-1})$$

First, we show that the process

$$S_{1,n}(x, \theta) = n^{-1/2} \sum_{t=1}^n w_{1,t-1}(x, \theta)$$

is asymptotically tight w.r.t $(x, \theta) \in \mathcal{W} = \Pi \times \Theta$.

$$\begin{aligned} |x_1y_1 - x_2y_2| &= |x_1y_1 - x_2y_1 + x_2y_1 - x_2y_2| \\ &= |(x_1 - x_2)y_1 + x_2(y_1 - y_2)| \\ &\leq |(x_1 - x_2)y_1| + |x_2(y_1 - y_2)| \end{aligned}$$

$$\begin{aligned} &|w_{1,t-1}(x_1, \theta_1) - w_{1,t-1}(x_2, \theta_2)| \\ &\leq \left| \left(\Psi_{v,\alpha,t}(\theta_1) - \mathbb{E}_{t-1}\Psi_{v,\alpha,t}(\theta_1) - \Psi_{v,\alpha,t}(\theta_2) + \mathbb{E}_{t-1}\Psi_{v,\alpha,t}(\theta_2) \right) \exp(ix'_1 I_{t-1}) \right| \\ &\quad + \left| \left(\Psi_{v,\alpha,t}(\theta_2) - \mathbb{E}_{t-1}\Psi_{v,\alpha,t}(\theta_2) \right) \left(\exp(ix'_1 I_{t-1}) - \exp(ix'_2 I_{t-1}) \right) \right| \\ &\leq \left| \Psi_{v,\alpha,t}(\theta_1) - \mathbb{E}_{t-1}\Psi_{v,\alpha,t}(\theta_1) - \Psi_{v,\alpha,t}(\theta_2) + \mathbb{E}_{t-1}\Psi_{v,\alpha,t}(\theta_2) \right| \\ &\quad + 2 \left| \exp(ix'_1 I_{t-1}) - \exp(ix'_2 I_{t-1}) \right| \end{aligned}$$

Since $\forall x_1$ and α , $\left| \exp(ix'_1 I_{t-1}) \right| = 1$ and

$$\begin{aligned} \left| \Psi_{v,\alpha,t}(\theta_2) - \mathbb{E}_{t-1} \Psi_{v,\alpha,t}(\theta_2) \right| &\leq \left| \Psi_{v,\alpha,t}(\theta_2) \right| + \left| \mathbb{E}_{t-1} \Psi_{v,\alpha,t}(\theta_2) \right| \\ &\leq 1 + \mathbb{E}_{t-1} |\Psi_{v,\alpha,t}(\theta_2)| \\ &\leq 1 + 1 \end{aligned}$$

$$\begin{aligned} &|w_{1,t-1}(x_1, \theta_1) - w_{1,t-1}(x_2, \theta_2)|^2 \\ &\leq 2 \left| \Psi_{v,\alpha,t}(\theta_1) - \Psi_{v,\alpha,t}(\theta_2) - \mathbb{E}_{t-1} \Psi_{v,\alpha,t}(\theta_1) + \mathbb{E}_{t-1} \Psi_{v,\alpha,t}(\theta_2) \right|^2 \\ &\quad + 8 \left| \exp(ix'_1 I_{t-1}) - \exp(ix'_2 I_{t-1}) \right|^2 \\ &\leq 4 |\Psi_{v,\alpha,t}(\theta_1) - \Psi_{v,\alpha,t}(\theta_2)|^2 + 4 \mathbb{E}_{t-1} |\Psi_{v,\alpha,t}(\theta_1) - \Psi_{v,\alpha,t}(\theta_2)|^2 \\ &\quad + 8 \left| \exp(ix'_1 I_{t-1}) - \exp(ix'_2 I_{t-1}) \right|^2 \end{aligned}$$

$$\mathbb{E}_{t-1} \left[\sup_{p_1, p_2 \in H_k} |w_{1,t-1}(x_1, \theta_1) - w_{1,t-1}(x_2, \theta_2)|^2 \right] \leq K \cdot \sum_{i=1}^3 \mu^{(i)},$$

where each of $\mu^{(i)}$, $i = 1, 2, 3$ is defined and bounded below:

$$\begin{aligned} \mu^{(1)} &= \mathbb{E}_{t-1} \left[\sup_{p_1, p_2 \in H_k} |\Psi_{v,\alpha,t}(\theta_1) - \Psi_{v,\alpha,t}(\theta_2)|^2 \right] \\ &= \mathbb{E}_{t-1} \left[\mathbf{1} \left\{ \min_{p \in H_k} v_{t-1}(\theta) \leq Y_t \leq \max_{p \in H_k} v_{t-1}(\theta) \right\} \right] \\ &= \int_{\min_{p \in H_k} v_{t-1}(\theta)}^{\max_{p \in H_k} v_{t-1}(\theta)} f_{t-1}(x) dx \\ &\leq K \sup_{\theta \in H_k} \|\nabla v_{t-1}(\theta)\| \cdot \sup_{p_1, p_2 \in H_k} \|\theta_1 - \theta_2\| \\ &\leq K \delta^2 \sup_{\theta \in H_k} \|\nabla v_{t-1}(\theta)\| \end{aligned}$$

$$\begin{aligned}
\mu^{(2)} &= \sup_{p_1, p_2 \in H_k} \mathbb{E}_{t-1} |\Psi_{v, \alpha, t}(\theta_1) - \Psi_{v, \alpha, t}(\theta_2)|^2 \\
&\leq \mathbb{E}_{t-1} \left[\sup_{p_1, p_2 \in H_k} |\Psi_{v, \alpha, t}(\theta_1) - \Psi_{v, \alpha, t}(\theta_2)|^2 \right] \equiv \mu^{(1)}
\end{aligned}$$

$$\begin{aligned}
\mu^{(3)} &= \mathbb{E}_{t-1} \left[\sup_{p_1, p_2 \in H_k} |\exp(ix'_1 I_{t-1}) - \exp(ix'_2 I_{t-1})|^2 \right] \\
&\leq \delta^2 \|I_{t-1}\|^2
\end{aligned}$$

$CV_{n, w_1}(\mathcal{B}_\delta)$ in (4.19) is bounded by

$$K\delta^2 \left(n^{-1} \sum_{t=1}^n \sup_{\theta \in \Theta} \|\nabla v_{t-1}(\theta)\| + n^{-1} \sum_{t=1}^n \|I_{t-1}\|^2 \right)$$

Second, we show that the process

$$S_{2, n}(x, \theta) = n^{-1/2} \sum_{t=1}^n w_{2, t-1}(x, \theta)$$

is asymptotically tight w.r.t $(x, \theta) \in \mathcal{W} = \Pi \times \Theta$.

$$\begin{aligned}
&|w_{2, t-1}(x_1, \theta_1) - w_{2, t-1}(x_2, \theta_2)| \\
&\leq \left| \left(\Psi_{e, \alpha, t}(\theta_1) - \mathbb{E}_{t-1} \Psi_{e, \alpha, t}(\theta_1) - \Psi_{e, \alpha, t}(\theta_2) + \mathbb{E}_{t-1} \Psi_{e, \alpha, t}(\theta_2) \right) \exp(ix'_1 I_{t-1}) \right| \\
&\quad + \left| \left(\Psi_{e, \alpha, t}(\theta_2) - \mathbb{E}_{t-1} \Psi_{e, \alpha, t}(\theta_2) \right) \left(\exp(ix'_1 I_{t-1}) - \exp(ix'_2 I_{t-1}) \right) \right| \\
&\leq \left| \Psi_{e, \alpha, t}(\theta_1) - \mathbb{E}_{t-1} \Psi_{e, \alpha, t}(\theta_1) - \Psi_{e, \alpha, t}(\theta_2) + \mathbb{E}_{t-1} \Psi_{e, \alpha, t}(\theta_2) \right| \\
&\quad + \left| \Psi_{e, \alpha, t}(\theta_2) - \mathbb{E}_{t-1} \Psi_{e, \alpha, t}(\theta_2) \right| \cdot \left| \exp(ix'_1 I_{t-1}) - \exp(ix'_2 I_{t-1}) \right| \\
&= \alpha^{-1} \left| Y_t \left(\mathbf{1}\{Y_t \leq v_{t-1}(\theta_1)\} - \mathbf{1}\{Y_t \leq v_{t-1}(\theta_2)\} \right) \right. \\
&\quad \left. - \mathbb{E}_{t-1} \left[Y_t \left(\mathbf{1}\{Y_t \leq v_{t-1}(\theta_1)\} - \mathbf{1}\{Y_t \leq v_{t-1}(\theta_2)\} \right) \right] \right| \\
&\quad + \alpha^{-1} \left| Y_t \mathbf{1}\{Y_t \leq v_{t-1}(\theta_2)\} - \mathbb{E}_{t-1} [Y_t \mathbf{1}\{Y_t \leq v_{t-1}(\theta_2)\}] \right| \cdot \left| \exp(ix'_1 I_{t-1}) - \exp(ix'_2 I_{t-1}) \right|
\end{aligned}$$

$$\begin{aligned}
& |w_{2,t-1}(x_1, \theta_1) - w_{2,t-1}(x_2, \theta_2)|^2 \\
& \leq 2\alpha^{-2} \left| Y_t \left(\mathbf{1}\{Y_t \leq v_{t-1}(\theta_1)\} - \mathbf{1}\{Y_t \leq v_{t-1}(\theta_2)\} \right) \right. \\
& \quad \left. - \mathbb{E}_{t-1} \left[Y_t \left(\mathbf{1}\{Y_t \leq v_{t-1}(\theta_1)\} - \mathbf{1}\{Y_t \leq v_{t-1}(\theta_2)\} \right) \right] \right|^2 \\
& \quad + 2\alpha^{-2} \left| Y_t \mathbf{1}\{Y_t \leq v_{t-1}(\theta_2)\} - \mathbb{E}_{t-1}[Y_t \mathbf{1}\{Y_t \leq v_{t-1}(\theta_2)\}] \right|^2 \cdot \left| \exp(ix'_1 I_{t-1}) - \exp(ix'_2 I_{t-1}) \right|^2 \\
& \leq 4\alpha^{-2} \left| Y_t \left(\mathbf{1}\{Y_t \leq v_{t-1}(\theta_1)\} - \mathbf{1}\{Y_t \leq v_{t-1}(\theta_2)\} \right) \right|^2 \\
& \quad + 4\alpha^{-2} \mathbb{E}_{t-1} \left| Y_t \left(\mathbf{1}\{Y_t \leq v_{t-1}(\theta_1)\} - \mathbf{1}\{Y_t \leq v_{t-1}(\theta_2)\} \right) \right|^2 \\
& \quad + 2\alpha^{-2} \left| Y_t \mathbf{1}\{Y_t \leq v_{t-1}(\theta_2)\} - \mathbb{E}_{t-1}[Y_t \mathbf{1}\{Y_t \leq v_{t-1}(\theta_2)\}] \right|^2 \cdot \left| \exp(ix'_1 I_{t-1}) - \exp(ix'_2 I_{t-1}) \right|^2
\end{aligned}$$

$$\mathbb{E}_{t-1} \left[\sup_{p_1, p_2 \in H_k} |w_{2,t-1}(x_1, \theta_1) - w_{2,t-1}(x_2, \theta_2)|^2 \right] \leq K \cdot \sum_{i=1}^3 \mu^{(i)},$$

where each of $\mu^{(i)}$, $i = 1, 2, 3$ is defined and bounded below:

$$\begin{aligned}
\mu^{(1)} &= \mathbb{E}_{t-1} \left[\sup_{p_1, p_2 \in H_k} \left| Y_t \left(\mathbf{1}\{Y_t \leq v_{t-1}(\theta_1)\} - \mathbf{1}\{Y_t \leq v_{t-1}(\theta_2)\} \right) \right|^2 \right] \\
&= \mathbb{E}_{t-1} [Y_t^2 \cdot \mathbf{1}\{\min_{p \in H_k} v_{t-1}(\theta) \leq Y_t \leq \max_{p \in H_k} v_{t-1}(\theta)\}] \\
&= \int_{\min_{p \in H_k} v_{t-1}(\theta)}^{\max_{p \in H_k} v_{t-1}(\theta)} x^2 f_{t-1}(x) dx \\
&\leq K \sup_{\theta \in \Theta} \|v_{t-1}^2(\theta) \nabla v_{t-1}(\theta)\| \cdot \sup_{p_1, p_2 \in H_k} \|\theta_1 - \theta_2\| \\
&\leq K \delta^2 \sup_{\theta \in \Theta} \|v_{t-1}^2(\theta) \nabla v_{t-1}(\theta)\|
\end{aligned}$$

$$\begin{aligned}
\mu^{(2)} &= \sup_{p_1, p_2 \in H_k} \mathbb{E}_{t-1} \left| Y_t \left(\mathbf{1}\{Y_t \leq v_{t-1}(\theta_1)\} - \mathbf{1}\{Y_t \leq v_{t-1}(\theta_2)\} \right) \right|^2 \\
&\leq \mathbb{E}_{t-1} \left[\sup_{p_1, p_2 \in H_k} \left| Y_t \left(\mathbf{1}\{Y_t \leq v_{t-1}(\theta_1)\} - \mathbf{1}\{Y_t \leq v_{t-1}(\theta_2)\} \right) \right|^2 \right] \equiv \mu^{(1)}
\end{aligned}$$

$\mu^{(3)} = \mathbb{E}_{t-1}[(*)]$, where

$$\begin{aligned}
(*) &= \sup_{p_1, p_2 \in H_k} \left(\left| Y_t \mathbf{1}\{Y_t \leq v_{t-1}(\theta_2)\} - \mathbb{E}_{t-1}[Y_t \mathbf{1}\{Y_t \leq v_{t-1}(\theta_2)\}] \right|^2 \left| \exp(ix'_1 I_{t-1}) - \exp(ix'_2 I_{t-1}) \right|^2 \right) \\
&\leq \sup_{p_1, p_2 \in H_k} \left| Y_t \mathbf{1}\{Y_t \leq v_{t-1}(\theta_2)\} - \mathbb{E}_{t-1}[Y_t \mathbf{1}\{Y_t \leq v_{t-1}(\theta_2)\}] \right|^2 \\
&\quad \cdot \sup_{p_1, p_2 \in H_k} \left| \exp(ix'_1 I_{t-1}) - \exp(ix'_2 I_{t-1}) \right|^2
\end{aligned}$$

Since

$$\begin{aligned}
&\sup_{p_1, p_2 \in H_k} \left| Y_t \mathbf{1}\{Y_t \leq v_{t-1}(\theta_2)\} - \mathbb{E}_{t-1}[Y_t \mathbf{1}\{Y_t \leq v_{t-1}(\theta_2)\}] \right|^2 \\
&\leq 2 \sup_{p_1, p_2 \in H_k} \left| Y_t \mathbf{1}\{Y_t \leq v_{t-1}(\theta_2)\} \right|^2 + 2 \sup_{p_1, p_2 \in H_k} \mathbb{E}_{t-1} |Y_t \mathbf{1}\{Y_t \leq v_{t-1}(\theta_2)\}|^2 \\
&\leq 2 \sup_{p_1, p_2 \in H_k} \left| Y_t \mathbf{1}\{Y_t \leq v_{t-1}(\theta_2)\} \right|^2 + 2 \mathbb{E}_{t-1} \left[\sup_{p_1, p_2 \in H_k} |Y_t \mathbf{1}\{Y_t \leq v_{t-1}(\theta_2)\}|^2 \right]
\end{aligned}$$

and $\sup_{p_1, p_2 \in H_k} \left| \exp(ix'_1 I_{t-1}) - \exp(ix'_2 I_{t-1}) \right|^2 \leq \delta^2 \|I_{t-1}\|^2$ as we have shown before, it follows that

$$\begin{aligned}
\mu^{(3)} &\leq K \delta^2 \|I_{t-1}\|^2 \cdot \mathbb{E}_{t-1} \left[\sup_{p_1, p_2 \in H_k} |Y_t \mathbf{1}\{Y_t \leq v_{t-1}(\theta_2)\}|^2 \right] \\
&\leq K \delta^2 \|I_{t-1}\|^2 \cdot \mathbb{E}_{t-1} Y_t^2
\end{aligned}$$

$CV_{n, w_2}(\mathcal{B}_\delta)$ in (4.19) is bounded by

$$K \delta^2 \left(n^{-1} \sum_{t=1}^n \sup_{\theta \in \Theta} \|v_{t-1}^2(\theta) \nabla v_{t-1}(\theta)\| + n^{-1} \sum_{t=1}^n \|I_{t-1}\|^2 \cdot \mathbb{E}_{t-1} Y_t^2 \right)$$

The asymptotic tightness of $S_{1,n}(x, \theta)$ and $S_{2,n}(x, \theta)$ are then proved. Since $\|\theta_n - \theta_1\| =$

$o_p(1)$, it follows that

$$\sup_{x \in \Pi} |S_{1,n}(x, \hat{\theta}_n) - S_{1,n}(x, \theta_1)| = o_P(1),$$

and

$$\sup_{x \in \Pi} |S_{2,n}(x, \hat{\theta}_n) - S_{2,n}(x, \theta_1)| = o_P(1).$$

proof of Theorem 2.: Under H_0 , $\theta_1 = \theta_0$ and $\mathbb{E}_{t-1}[(\Psi_{v,\alpha,t}(\theta_0), \Psi_{e,\alpha,t}(\theta_0))]' = \mathbf{0}$ a.s. Using the asymptotic expansion of $\hat{R}_{v,\alpha,n}(x)$ and $\hat{R}_{e,\alpha,n}(x)$ in (4.22) and (4.23) respectively, we could get that, uniformly in $x \in \Pi$,

$$\begin{aligned} \hat{R}_{v,\alpha,n}(x) &= n^{-1/2} \sum_{t=1}^n \Psi_{v,\alpha,t}(\theta_0) \exp(ix' I_{t-1}) \\ &\quad + n^{-1/2} \sum_{t=1}^n \{\mathbb{E}_{t-1} \Psi_{v,\alpha,t}(\hat{\theta}_n) - \mathbb{E}_{t-1} \Psi_{v,\alpha,t}(\theta_0)\} \exp(ix' I_{t-1}) + o_P(1) \\ &= R_{v,\alpha,n}(x) + n^{-1/2} \sum_{t=1}^n \{F_{t-1}(v_{t-1}(\hat{\theta}_n)) - F_{t-1}(v_{t-1}(\theta_0))\} \exp(ix' I_{t-1}) + o_P(1) \end{aligned}$$

and

$$\begin{aligned} \hat{R}_{e,\alpha,n}(x) &= n^{-1/2} \sum_{t=1}^n \Psi_{e,\alpha,t}(\theta_0) \exp(ix' I_{t-1}) \\ &\quad + n^{-1/2} \sum_{t=1}^n \{\mathbb{E}_{t-1} \Psi_{e,\alpha,t}(\hat{\theta}_n) - \mathbb{E}_{t-1} \Psi_{e,\alpha,t}(\theta_0)\} \exp(ix' I_{t-1}) + o_p(1) \\ &= R_{e,\alpha,n}(x) + n^{-1/2} \sum_{t=1}^n \{\mathbb{E}_{t-1} \Psi_{e,\alpha,t}(\hat{\theta}_n) - \mathbb{E}_{t-1} \Psi_{e,\alpha,t}(\theta_0)\} \exp(ix' I_{t-1}) + o_p(1) \end{aligned}$$

$\forall x \in \Pi$,

$$\begin{aligned} & n^{-1/2} \sum_{t=1}^n \{F_{t-1}(v_{t-1}(\hat{\theta}_n)) - F_{t-1}(v_{t-1}(\theta_0))\} \exp(ix' I_{t-1}) \\ &= \left(\frac{1}{n} \sum_{t=1}^n f_{t-1}(v_{t-1}(\theta_*)) \nabla v_{t-1}(\theta_*) \exp(ix' I_{t-1}) \right) \cdot \sqrt{n}(\hat{\theta}_n - \theta_0) \end{aligned}$$

and

$$\begin{aligned} & n^{-1/2} \sum_{t=1}^n \{\mathbb{E}_{t-1} \Psi_{e,\alpha,t}(\hat{\theta}_n) - \mathbb{E}_{t-1} \Psi_{e,\alpha,t}(\theta_0)\} \exp(ix' I_{t-1}) \\ &= \left(\frac{1}{n} \sum_{t=1}^n [v_{t-1}(\theta_*) f_{t-1}(v_{t-1}(\theta_*)) \nabla v_{t-1}(\theta_*) - \nabla e_{t-1}(\theta_*)] \exp(ix' I_{t-1}) \right) \cdot \sqrt{n}(\hat{\theta}_n - \theta_0) \end{aligned}$$

Define

$$Q_{v,\alpha,n}(x, \theta) = \frac{1}{n} \sum_{t=1}^n f_{t-1}(v_{t-1}(\theta)) \nabla v_{t-1}(\theta) \exp(ix' I_{t-1})$$

and

$$Q_{e,\alpha,n}(x, \theta) = \frac{1}{n} \sum_{t=1}^n [v_{t-1}(\theta) f_{t-1}(v_{t-1}(\theta)) \nabla v_{t-1}(\theta) - \nabla e_{t-1}(\theta)] \exp(ix' I_{t-1})$$

If we could show that both $Q_{v,\alpha,n}(x, \theta)$ and $Q_{e,\alpha,n}(x, \theta)$ are stochastic equicontinuous, since $\|\theta_* - \theta_0\| = o_p(1)$, then uniformly in $x \in \Pi$,

$$\frac{1}{n} \sum_{t=1}^n f_{t-1}(v_{t-1}(\theta_*)) \nabla v_{t-1}(\theta_*) \exp(ix' I_{t-1}) = \frac{1}{n} \sum_{t=1}^n f_{t-1}(v_{t-1}(\theta_0)) \nabla v_{t-1}(\theta_0) \exp(ix' I_{t-1}) + o_p(1)$$

and

$$\begin{aligned} & \frac{1}{n} \sum_{t=1}^n [v_{t-1}(\theta_*) f_{t-1}(v_{t-1}(\theta_*)) \nabla v_{t-1}(\theta_*) - \nabla e_{t-1}(\theta_*)] \exp(ix' I_{t-1}) \\ &= \frac{1}{n} \sum_{t=1}^n [v_{t-1}(\theta_0) f_{t-1}(v_{t-1}(\theta_0)) \nabla v_{t-1}(\theta_0) - \nabla e_{t-1}(\theta_0)] \exp(ix' I_{t-1}) + o_p(1) \end{aligned}$$

Furthermore, uniformly in $x \in \Pi$,

$$\begin{aligned} & n^{-1/2} \sum_{t=1}^n \{F_{t-1}(v_{t-1}(\hat{\theta}_n)) - F_{t-1}(v_{t-1}(\theta_0))\} \exp(ix' I_{t-1}) \\ &= \left(\frac{1}{n} \sum_{t=1}^n f_{t-1}(v_{t-1}(\theta_0)) \nabla v_{t-1}(\theta_0) \exp(ix' I_{t-1}) \right) \cdot \sqrt{n}(\hat{\theta}_n - \theta_0) + o_p(1). \end{aligned}$$

and

$$\begin{aligned} & n^{-1/2} \sum_{t=1}^n \{\mathbb{E}_{t-1} \Psi_{e,\alpha,t}(\hat{\theta}_n) - \mathbb{E}_{t-1} \Psi_{e,\alpha,t}(\theta_0)\} \exp(ix' I_{t-1}) \\ &= \left(\frac{1}{n} \sum_{t=1}^n [v_{t-1}(\theta_0) f_{t-1}(v_{t-1}(\theta_0)) \nabla v_{t-1}(\theta_0) - \nabla e_{t-1}(\theta_0)] \exp(ix' I_{t-1}) \right) \cdot \sqrt{n}(\hat{\theta}_n - \theta_0) + o_p(1). \end{aligned}$$

This together with Theorem 1, Assumption A2 and A3 proves the theorem. Below, we establish the stochastic equicontinuity of $Q_{v,\alpha,n}(x, \theta)$ and $Q_{e,\alpha,n}(x, \theta)$ based on Theorem 21.10 of Andrews (1992).

$$\forall p_1 = (x'_1, \theta'_1)', p_2 = (x'_2, \theta'_2)' \in \mathcal{W} = \Pi \times \Theta,$$

$$\|Q_{v,\alpha,n}(x_1, \theta_1) - Q_{v,\alpha,n}(x_2, \theta_2)\| \leq \|p_1 - p_2\| \cdot \left\| \left(\frac{\partial Q_{v,\alpha,n}}{\partial \theta}(x_*, \theta_*), \frac{\partial Q_{v,\alpha,n}}{\partial x}(x_*, \theta_*) \right)' \right\|$$

where

$$\begin{aligned}
& \left\| \frac{\partial Q_{v,\alpha,n}}{\partial \theta}(x_*, \theta_*) \right\| \\
&= \left\| \frac{1}{n} \sum_{t=1}^n [f_{t-1}(v_{t-1}(\theta_*)) \nabla^2 v_{t-1}(\theta_*) + f'_{t-1}(v_{t-1}(\theta_*)) \nabla' v_{t-1}(\theta_*) \nabla v_{t-1}(\theta_*)] \exp(ix'_* I_{t-1}) \right\| \\
&\leq \frac{1}{n} \sum_{t=1}^n \left\| f_{t-1}(v_{t-1}(\theta_*)) \nabla^2 v_{t-1}(\theta_*) + f'_{t-1}(v_{t-1}(\theta_*)) \nabla' v_{t-1}(\theta_*) \nabla v_{t-1}(\theta_*) \right\| \\
&\leq K \cdot \frac{1}{n} \sum_{t=1}^n \left(\|\nabla^2 v_{t-1}(\theta_*)\| + \|\nabla v_{t-1}(\theta_*)\|^2 \right)
\end{aligned}$$

and

$$\begin{aligned}
\left\| \frac{\partial Q_{v,\alpha,n}}{\partial x}(x_*, \theta_*) \right\| &= \left\| \frac{1}{n} \sum_{t=1}^n [\exp(ix'_* I_{t-1}) i I'_{t-1}] f_{t-1}(v_{t-1}(\theta_*)) \nabla v_{t-1}(\theta_*) \right\| \\
&\leq K \cdot \frac{1}{n} \sum_{t=1}^n \|I_{t-1}\| \cdot \|\nabla v_{t-1}(\theta_*)\|
\end{aligned}$$

Under Assumption A1(g), $\frac{1}{n} \sum_{t=1}^n \sup_{\theta \in \Theta} \|\nabla^2 v_{t-1}(\theta)\| = O_P(1)$, $\frac{1}{n} \sum_{t=1}^n \sup_{\theta \in \Theta} \|\nabla v_{t-1}(\theta)\|^2 = O_P(1)$ and $\frac{1}{n} \sum_{t=1}^n \|I_{t-1}\| \cdot \sup_{\theta \in \Theta} \|\nabla v_{t-1}(\theta)\| = O_P(1)$, then by Theorem 21.10 of Andrews (1992), $Q_{v,\alpha,n}(x, \theta)$ is stochastic equicontinuous.

$$\|Q_{e,\alpha,n}(x_1, \theta_1) - Q_{e,\alpha,n}(x_2, \theta_2)\| \leq \|p_1 - p_2\| \cdot \left\| \left(\frac{\partial Q_{e,\alpha,n}}{\partial \theta}(x_*, \theta_*), \frac{\partial Q_{e,\alpha,n}}{\partial x}(x_*, \theta_*) \right)' \right\|$$

where

$$\begin{aligned}
& \left\| \frac{\partial Q_{e,\alpha,n}}{\partial \theta}(x_*, \theta_*) \right\| \\
&= \left\| \frac{1}{n} \sum_{t=1}^n \left[\left(f_{t-1}(\ast) + v_{t-1}(\theta_*) f'_{t-1}(\ast) \right) \nabla' v_{t-1}(\theta_*) \nabla v_{t-1}(\theta_*) \right. \right. \\
&\quad \left. \left. + v_{t-1}(\theta_*) f_{t-1}(\ast) \nabla^2 v_{t-1}(\theta_*) - \nabla^2 e_{t-1}(\theta_*) \right] \exp(ix'_* I_{t-1}) \right\| \\
&\leq K \cdot \frac{1}{n} \sum_{t=1}^n \left(\|\nabla^2 v_{t-1}(\theta_*)\| + |v_{t-1}(\theta_*)| \cdot \|\nabla v_{t-1}(\theta_*)\|^2 \right. \\
&\quad \left. + |v_{t-1}(\theta_*)| \cdot \|\nabla^2 v_{t-1}(\theta_*)\| + \|\nabla^2 e_{t-1}(\theta_*)\| \right)
\end{aligned}$$

and

$$\begin{aligned}
\left\| \frac{\partial Q_{2\alpha,n}}{\partial x}(x_*, \theta_*) \right\| &= \left\| \frac{1}{n} \sum_{t=1}^n [\exp(ix'_* I_{t-1}) i I'_{t-1}] [v_{t-1}(\theta_*) f_{t-1}(v_{t-1}(\theta_*)) \nabla v_{t-1}(\theta_*) - \nabla e_{t-1}(\theta_*)] \right\| \\
&\leq K \cdot \frac{1}{n} \sum_{t=1}^n \left(\|I_{t-1}\| \cdot |v_{t-1}(\theta_*)| \cdot \|\nabla v_{t-1}(\theta_*)\| + \|I_{t-1}\| \cdot \|\nabla e_{t-1}(\theta_*)\| \right)
\end{aligned}$$

Under Assumption A1(g), $\frac{1}{n} \sum_{t=1}^n \sup_{\theta \in \Theta} \|\nabla^2 v_{t-1}(\theta)\| = O_P(1)$, $\frac{1}{n} \sum_{t=1}^n \sup_{\theta \in \Theta} (|v_{t-1}(\theta)| \cdot \|\nabla v_{t-1}(\theta)\|^2) = O_P(1)$, $\frac{1}{n} \sum_{t=1}^n \sup_{\theta \in \Theta} (|v_{t-1}(\theta)| \cdot \|\nabla^2 v_{t-1}(\theta)\|) = O_P(1)$, $\frac{1}{n} \sum_{t=1}^n \sup_{\theta \in \Theta} \|\nabla^2 e_{t-1}(\theta)\| = O_P(1)$, $\frac{1}{n} \sum_{t=1}^n \|I_{t-1}\| \cdot \sup_{\theta \in \Theta} (|v_{t-1}(\theta)| \cdot \|\nabla v_{t-1}(\theta)\|) = O_P(1)$, and $\frac{1}{n} \sum_{t=1}^n \|I_{t-1}\| \cdot \sup_{\theta \in \Theta} \|\nabla e_{t-1}(\theta)\| = O_P(1)$, then by Theorem 21.10 of Andrews (1992), $Q_{e,\alpha,n}(x, \theta)$ is stochastic equicontinuous. So the proof is complete.

4.6 Conclusion

I propose a consistent specification test of dynamic joint models for VaR and ES and derive the limiting distribution of the test statistics under the null hypothesis. The asymptotic null distribution is non-standard, so I use subsampling approximation to

get the asymptotic critical values. The proposed test is confirmed via Monte Carlo studies to have better empirical size and power performance in finite samples than other existing tests. In particular, the severe size distortion of subsampling-based Christoffersen test and regression-based test highlights the comparative advantage of my test in detecting violations of the conditional moment condition of joint models for VaR and ES.

Chapter 5

Conclusions

This dissertation touches upon two topics in financial econometrics literature and makes contributions on each front. First, in the joint work with Jia Li, Viktor Todorov and George Tauchen, we propose a new mixed-scale jump regression framework for studying deterministic dependencies among jumps in a multivariate setting. We derive the asymptotic properties of an efficient estimator of the jump regression coefficients and a test for its specification. The limiting distributions of the estimator and the test statistic are non-standard, but a simple bootstrap method is shown to be valid for feasible inference. In the empirical application, we find a strong relationship between market jumps and stock price moves at market jump times. Second, in my joint work with Andrew J. Patton and Johanna F. Ziegel, we draw on recent results from statistical decision theory (Fissler and Ziegel, 2016) to propose new dynamic models for ES and VaR. We also present asymptotic distribution theory for the estimation of these models, and we verify that the theory provides a good approximation in finite samples. We apply the new models and methods to daily returns on four international equity indices, over the period 1990 to 2016, and find the proposed new ES-VaR models outperform forecasts based on GARCH or rolling window models. Finally, in my independent work, I propose a consistent specification test of dynamic joint models for VaR and ES and derive the limiting distribution of the test statistics under the null hypothesis. The asymptotic null distribution is non-standard, so I use subsampling approximation to get the asymptotic critical values. The proposed test is confirmed via Monte Carlo studies to have better empirical size and power

performance in finite samples than other existing tests. In particular, the severe size distortion of subsampling-based Christoffersen test and regression-based test highlights the comparative advantage of my test in detecting violations of the conditional moment condition of joint models for VaR and ES.

Appendix A

Appendix to Chapter 3

A.1 Proofs

Proof of Proposition 1. Theorem C.3 of Nolde and Ziegel (2017) shows that under the assumption that ES is strictly negative, the loss differences generated by a FZ loss function are homogeneous of degree zero iff $G_1(x) = \varphi_1 \mathbf{1}\{x \geq 0\}$ and $G_2(x) = -\varphi_2/x$ with $\varphi_1 \geq 0$ and $\varphi_2 > 0$. Denote the resulting loss function as $L_{FZ0}^*(Y, v, e; \alpha, \varphi_1, \varphi_2)$, and notice that:

$$\begin{aligned} & L_{FZ0}^*(Y, v, e; \alpha, \varphi_1, \varphi_2) \\ = & \varphi_1 (\mathbf{1}\{Y \leq v\} - \alpha) (\mathbf{1}\{v \geq 0\} - \mathbf{1}\{Y \geq 0\}) \\ & + \varphi_2 \left\{ -(\mathbf{1}\{Y \leq v\} - \alpha) \frac{1}{\alpha} \frac{v}{e} + \frac{1}{e} \left(\frac{1}{\alpha} \mathbf{1}\{Y \leq v\} Y - e \right) + \log(-e) \right\} \\ = & \varphi_1 (\mathbf{1}\{Y \leq v\} - \alpha) (\mathbf{1}\{v \geq 0\} - \mathbf{1}\{Y \geq 0\}) + \varphi_2 L_{FZ0}(Y, v, e; \alpha) \\ = & \varphi_2 L_{FZ0}(Y, v, e; \alpha) + \varphi_1 \alpha \mathbf{1}\{Y \geq 0\} \\ & + \varphi_1 \mathbf{1}\{v \geq 0\} (\mathbf{1}\{Y \leq v\} - \alpha - \mathbf{1}\{0 \leq Y \leq v\}) \end{aligned}$$

Under the assumption that $v < 0$, the third term vanishes. The second term is purely a function of Y and so can be disregarded; we can set $\varphi_1 = 0$ without loss of generality. The first term is affected by a scaling parameter $\varphi_2 > 0$, and we can set $\varphi_2 = 1$ without loss of generality. Thus we obtain the L_{FZ0} given in equation (3.6). If v can be positive, then setting $\varphi_1 = 0$ is interpretable as fixing this shape parameter value at a particular value. \square

Proof of Theorem 5. The proof is based on Theorem 2.1 of Newey and McFadden (1994). We only need to show that $E[L_T(\cdot)]$ is uniquely minimized at θ^0 , because the other assumptions of Newey and McFadden's theorem are clearly satisfied. By Corollary (5.5) of Fissler and Ziegel (2016), given Assumption 4(B)(iii) and the fact that our choice of the objective function L_{FZ0} satisfies the condition as in Corollary (5.5) of Fissler and Ziegel (2016), we know that $\mathbb{E}[L(Y_t, v_t(\theta), e_t(\theta); \alpha) | \mathcal{F}_{t-1}]$ is uniquely minimized at $(\text{VaR}_\alpha(Y_t | \mathcal{F}_{t-1}), \text{ES}_\alpha(Y_t | \mathcal{F}_{t-1}))$, which equals $(v_t(\theta^0), e_t(\theta^0))$ under correct specification. Combining this assumption and Assumption 1(B)(iv), we know that θ^0 is a unique minimizer of $\mathbb{E}[L_T(\cdot)]$, completing the proof. \square

Outline of proof of Theorem 6. We consider the population function $\lambda(\theta) = \mathbb{E}[g_t(\theta)]$, and take a mean-value expansion of $\lambda(\hat{\theta})$ around θ^0 . We show in Lemma 1 that:

$$\sqrt{T}(\hat{\theta} - \theta^0) = -\Lambda^{-1}(\theta^0) \frac{1}{\sqrt{T}} \sum_{t=1}^T g_t(\theta^0) + o_p(1)$$

where $\Lambda(\theta^*) = \left. \frac{\partial \mathbb{E}[g_t(\theta)]}{\partial \theta} \right|_{\theta=\theta^*}$

In the supplemental appendix we prove Lemma 1 by building on and extending Weiss (1991), who extends Huber (1967) to non-*iid* data. We draw on Weiss' Lemma A.1, and we verify that all five assumptions (N1-N5 in his notation) for that lemma are satisfied: N1, N2 and N5 are obviously satisfied given our Assumptions 4 and 5, and we show in Lemmas 3 - 6 that assumptions N3 and N4 are satisfied. Lemma 7 shows that a CLT applies for the sequence $\left\{ T^{-1/2} \sum_{t=1}^T g_t(\theta^0) \right\}$, with asymptotic covariance matrix $\mathbf{A}_0 = \mathbb{E}[g_t(\theta^0)g_t(\theta^0)']$. We denote $\Lambda(\theta^0)$ as \mathbf{D}_0 , leading to the stated result. \square

Proof of Theorem 7. Given Assumption 6B(i) and the result in Theorem 5, the proof that $\hat{\mathbf{A}}_T - \mathbf{A}_0 \xrightarrow{p} \mathbf{0}$ is standard and omitted. Next, define

$$\tilde{\mathbf{D}}_T = T^{-1} \sum_{t=1}^T \left\{ (2c_T)^{-1} \mathbf{1}\{|Y_t - v_t(\theta^0)| < c_T\} \frac{1}{-e_t(\theta^0)\alpha} \nabla v_t(\theta^0)' \nabla v_t(\theta^0) + \frac{1}{e_t(\theta^0)^2} \nabla e_t(\theta^0)' \nabla e_t(\theta^0) \right\}$$

To prove the result we will show that $\hat{\mathbf{D}}_T - \tilde{\mathbf{D}}_T = o_p(1)$ and $\tilde{\mathbf{D}}_T - \mathbf{D}_0 = o_p(1)$. Firstly, consider

$$\begin{aligned} \|\hat{\mathbf{D}}_T - \tilde{\mathbf{D}}_T\| &\leq \|(2Tc_T)^{-1} \\ &\times \sum_{t=1}^T \left\{ (\mathbf{1}\{|Y_t - v_t(\hat{\theta}_T)| < c_T\} - \mathbf{1}\{|Y_t - v_t(\theta^0)| < c_T\}) \frac{1}{-e_t(\hat{\theta}_T)\alpha} \nabla v_t(\hat{\theta}_T)' \nabla v_t(\hat{\theta}_T) \right. \\ &\quad + \mathbf{1}\{|Y_t - v_t(\theta^0)| < c_T\} \frac{1}{-e_t(\hat{\theta}_T)\alpha} \left(\nabla v_t(\hat{\theta}_T) - \nabla v_t(\theta^0) \right)' \nabla v_t(\hat{\theta}_T) \\ &\quad + \mathbf{1}\{|Y_t - v_t(\theta^0)| < c_T\} \left(\frac{1}{-\alpha e_t(\hat{\theta}_T)} - \frac{1}{-\alpha e_t(\theta^0)} \right) \nabla v_t(\theta^0)' \nabla v_t(\hat{\theta}_T) \\ &\quad \left. + \mathbf{1}\{|Y_t - v_t(\theta^0)| < c_T\} \frac{1}{-\alpha e_t(\theta^0)} \nabla v_t(\theta^0)' (\nabla v_t(\hat{\theta}_T) - \nabla v_t(\theta^0)) \right\} \\ &\quad + T^{-1} \sum_{t=1}^T \left\| \frac{1}{e_t(\hat{\theta}_T)^2} \nabla e_t(\hat{\theta}_T)' \nabla e_t(\hat{\theta}_T) - \frac{1}{e_t(\theta^0)^2} \nabla e_t(\theta^0)' \nabla e_t(\theta^0) \right\| \end{aligned}$$

The last line above was shown to be $o_p(1)$ in the proof of Theorem 6. The difficult quantity in the first term (over the first six lines above) is the indicator, and following the same steps as in Engle and Manganelli (2004a), that term is also $o_p(1)$. Next,

consider $\tilde{\mathbf{D}}_T - \mathbf{D}_0$:

$$\begin{aligned} \tilde{\mathbf{D}}_T - \mathbf{D}_0 &= \frac{1}{2Tc_T} \sum_{t=1}^T (\mathbf{1}\{|Y_t - v_t(\theta^0)| < c_T\} - \mathbb{E}[\mathbf{1}\{|Y_t - v_t(\theta^0)| < c_T\} | \mathcal{F}_{t-1}]) \\ &\quad \times \frac{\nabla' v_t(\theta^0) \nabla v_t(\theta^0)}{-e_t(\theta^0)\alpha} \\ &\quad + \frac{1}{T} \sum_{t=1}^T \left\{ \frac{1}{2c_T} \mathbb{E}[\mathbf{1}\{|Y_t - v_t(\theta^0)| < c_T\} | \mathcal{F}_{t-1}] \frac{1}{-e_t(\theta^0)\alpha} \nabla' v_t(\theta^0) \nabla v_t(\theta^0) \right. \\ &\quad \left. - \mathbb{E} \left[\frac{f_t(v_t(\theta^0))}{-e_t(\theta^0)\alpha} \nabla' v_t(\theta^0) \nabla v_t(\theta^0) \right] \right\} + o_p(1) \end{aligned}$$

Following Engle and Manganelli (2004a), assumptions 4-6 are sufficient to show $\tilde{\mathbf{D}}_T - \mathbf{D}_0 = o_p(1)$ and the result follows. \square

A.2 Derivations

A.2.1 Generic calculations for the FZ0 loss function

The FZ0 loss function is:

$$L_{FZ0}(Y, v, e; \alpha) = -\frac{1}{\alpha e} \mathbf{1}\{Y \leq v\} (v - Y) + \frac{v}{e} + \log(-e) - 1 \quad (\text{A.1})$$

Note that this is *not* homogeneous, as for any $k > 0$, $L_{FZ0}(kY, kv, ke; \alpha) = L_{FZ0}(Y, v, e; \alpha) + \log(k)$, but this loss function generates loss *differences* that are homogenous of degree zero, as the additive additional term above drops out.

We will frequently use the first derivatives of this loss function, and the second derivatives of the expected loss for an absolutely continuous random variable with

density f and CDF F . These are (for $v \neq Y$):

$$\nabla_v \equiv \frac{\partial L_{FZ0}(Y, v, e; \alpha)}{\partial v} = -\frac{1}{\alpha e} (\mathbf{1}\{Y \leq v\} - \alpha) \equiv \frac{1}{\alpha v e} \lambda_v \quad (\text{A.2})$$

$$\begin{aligned} \nabla_e &\equiv \frac{\partial L_{FZ0}(Y, v, e; \alpha)}{\partial e} && (\text{A.3}) \\ &= \frac{1}{\alpha e^2} \mathbf{1}\{Y \leq v\} (v - Y) - \frac{v}{e^2} + \frac{1}{e} \\ &= \frac{v}{\alpha e^2} (\mathbf{1}\{Y \leq v\} - \alpha) - \frac{1}{e^2} \left(\frac{1}{\alpha} \mathbf{1}\{Y \leq v\} Y - e \right) \\ &\equiv \frac{-1}{\alpha e^2} (\lambda_v + \alpha \lambda_e) \end{aligned}$$

where

$$\lambda_v \equiv -v (\mathbf{1}\{Y \leq v\} - \alpha) \quad (\text{A.4})$$

$$\lambda_e \equiv \frac{1}{\alpha} \mathbf{1}\{Y \leq v\} Y - e \quad (\text{A.5})$$

and

$$\frac{\partial^2 \mathbb{E} [L_{FZ0}(Y, v, e; \alpha)]}{\partial v^2} = -\frac{1}{\alpha e} f(v) \quad (\text{A.6})$$

$$\frac{\partial^2 \mathbb{E} [L_{FZ0}(Y, v, e; \alpha)]}{\partial v \partial e} = \frac{1}{\alpha e^2} (F(v) - \alpha) \quad (\text{A.7})$$

$$= 0, \text{ at the true value of } (v, e)$$

$$\frac{\partial^2 \mathbb{E} [L_{FZ0}(Y, v, e; \alpha)]}{\partial e^2} = \frac{1}{e^2} - \frac{2}{\alpha e^3} \{ (F(v) - \alpha) v - (\mathbb{E} [\mathbf{1}\{Y \leq v\} Y] - \alpha e) \} \quad (\text{A.8})$$

$$= \frac{1}{e^2}, \text{ at the true value of } (v, e)$$

A.2.2 Derivations for the one-factor GAS model for ES and VaR

Here we present the calculations to compute s_t and I_t for this model. Below we use:

$$\frac{\partial v}{\partial \kappa} = \frac{\partial^2 v}{\partial \kappa^2} = a \exp \{ \kappa \} = v \quad (\text{A.9})$$

$$\frac{\partial e}{\partial \kappa} = \frac{\partial^2 e}{\partial \kappa^2} = b \exp \{ \kappa \} = e \quad (\text{A.10})$$

And so we find (for $v_t \neq Y_t$)

$$s_t \equiv \frac{\partial L_{FZ0}(Y_t, v_t, e_t; \alpha)}{\partial \kappa_t} \quad (\text{A.11})$$

$$\begin{aligned} &= \frac{\partial L_{FZ0}(Y_t, v_t, e_t; \alpha)}{\partial v_t} \frac{\partial v_t}{\partial \kappa_t} + \frac{\partial L_{FZ0}(Y_t, v_t, e_t; \alpha)}{\partial e_t} \frac{\partial e_t}{\partial \kappa_t} \\ &= \left\{ -\frac{1}{\alpha e_t} (\mathbf{1} \{ Y_t \leq v_t \} - \alpha) \right\} v_t \\ &\quad + \left\{ -\frac{1}{e_t^2} \left(\frac{1}{\alpha} \mathbf{1} \{ Y_t \leq v_t \} Y_t - e_t \right) + \frac{v_t}{e_t^2} \frac{1}{\alpha} (\mathbf{1} \{ Y_t \leq v_t \} - \alpha) \right\} e_t \\ &= -\frac{1}{e_t} \left(\frac{1}{\alpha} \mathbf{1} \{ Y_t \leq v_t \} Y_t - e_t \right) \quad (\text{A.12}) \end{aligned}$$

$$\equiv -\lambda_{et}/e_t \quad (\text{A.13})$$

Thus, the λ_{vt} term drops out of s_t and we are left with $-\lambda_{et}/e_t$.

Next we calculate I_t :

$$\begin{aligned} I_t &\equiv \frac{\partial^2 \mathbb{E}_{t-1} [L_{FZ0}(Y_t, v_t, e_t; \alpha)]}{\partial \kappa_t^2} \quad (\text{A.14}) \\ &= \frac{\partial^2 \mathbb{E}_{t-1} [L_{FZ0}(Y_t, v_t, e_t; \alpha)]}{\partial v_t^2} \left(\frac{\partial v_t}{\partial \kappa_t} \right)^2 + \frac{\partial^2 \mathbb{E}_{t-1} [L_{FZ0}(Y_t, v_t, e_t; \alpha)]}{\partial v_t \partial e_t} \frac{\partial v_t}{\partial \kappa_t} \\ &\quad + \frac{\partial^2 \mathbb{E}_{t-1} [L_{FZ0}(Y_t, v_t, e_t; \alpha)]}{\partial e_t^2} \left(\frac{\partial e_t}{\partial \kappa_t} \right)^2 + \frac{\partial^2 \mathbb{E}_{t-1} [L_{FZ0}(Y_t, v_t, e_t; \alpha)]}{\partial v_t \partial e_t} \frac{\partial e_t}{\partial \kappa_t} \\ &\quad + \frac{\partial \mathbb{E}_{t-1} [L_{FZ0}(Y_t, v_t, e_t; \alpha)]}{\partial v_t} \frac{\partial^2 v_t}{\partial \kappa_t^2} + \frac{\partial \mathbb{E}_{t-1} [L_{FZ0}(Y_t, v_t, e_t; \alpha)]}{\partial e_t} \frac{\partial^2 e_t}{\partial \kappa_t^2} \end{aligned}$$

But note that under correct specification,

$$\frac{\partial^2 \mathbb{E}_{t-1} [L(Y_t, v_t, e_t; \alpha)]}{\partial v_t \partial e_t} = \frac{\partial \mathbb{E}_{t-1} [L(Y_t, v_t, e_t; \alpha)]}{\partial v_t} = \frac{\partial \mathbb{E}_{t-1} [L(Y_t, v_t, e_t; \alpha)]}{\partial e_t} = 0 \quad (\text{A.15})$$

and so the Hessian simplifies to:

$$I_t = \frac{\partial^2 \mathbb{E}_{t-1} [L_{FZ0}(Y_t, v_t, e_t; \alpha)]}{\partial v_t^2} \left(\frac{\partial v_t}{\partial \kappa_t} \right)^2 + \frac{\partial^2 \mathbb{E}_{t-1} [L_{FZ0}(Y_t, v_t, e_t; \alpha)]}{\partial e_t^2} \left(\frac{\partial e_t}{\partial \kappa_t} \right)^2 \quad (\text{A.16})$$

$$= -\frac{1}{\alpha e_t} f_t(v_t) v_t^2 + 1 \quad (\text{A.17})$$

$$= 1 - \frac{k_\alpha a_\alpha}{\alpha b_\alpha}, \text{ since } f_t(v_t) = \frac{k_\alpha}{v_t} \text{ and } \frac{v_t}{e_t} = \frac{a_\alpha}{b_\alpha}, \text{ for this DGP.} \quad (\text{A.18})$$

Thus although the Hessian could vary with time, as it is a derivative of the conditional expected loss, in this specification it simplifies to a (positive) constant.

A.2.3 ES and VaR in location-scale models

Dynamic location-scale models are widely used for asset returns and in this section we consider what such a specification implies for the dynamics of ES and VaR. Consider the following:

$$Y_t = \mu_t + \sigma_t \eta_t, \quad \eta_t \sim iid F_\eta(0, 1) \quad (\text{A.19})$$

where, for example, μ_t is some ARMA model and σ_t^2 is some GARCH model. For asset returns that follow equation (A.19) we have:

$$v_t = \mu_t + a\sigma_t, \quad \text{where } a = F_\eta^{-1}(\alpha) \quad (\text{A.20})$$

$$e_t = \mu_t + b\sigma_t, \quad \text{where } b = \mathbb{E}[\eta_t | \eta_t \leq a]$$

and we can recover (μ_t, σ_t) from (v_t, e_t) :

$$\begin{bmatrix} \mu_t \\ \sigma_t \end{bmatrix} = \frac{1}{b-a} \begin{bmatrix} b & -a \\ -1 & 1 \end{bmatrix} \begin{bmatrix} v_t \\ e_t \end{bmatrix} \quad (\text{A.21})$$

Thus under the conditional location-scale assumption, we can back out the conditional mean and variance from the VaR and ES. Next note that if $\mu_t = 0 \forall t$, then $v_t = c \cdot e_t$, where $c = a/b \in (0, 1)$. Daily asset returns often have means that are close to zero, and so this restriction is one that may be plausible in the data. A related, though less plausible, restriction is that $\sigma_t = \bar{\sigma} \forall t$, and in that case we have the simplification that $v_t = d + e_t$, where $d = (a - b) \bar{\sigma} > 0$.

A.2.4 VaR and ES for Hansen's skew t random variables

The VaR for Hansen's (1994) skew t variable can be obtained using an expression for the inverse CDF of the skew t distribution presented in Jondeau and Rockinger (2003). In a recent paper, Dobrev et al. (2017) present an analytical expression for the expected shortfall for a Student's t random variable, X , with degrees of freedom $\nu > 1$:

$$ES_x(\alpha; \nu) = \frac{\nu^{\nu/2}}{2\alpha\sqrt{\pi}} \frac{\Gamma(\frac{\nu-1}{2})}{\Gamma(\frac{\nu}{2})} \left((VaR_x(\alpha; \nu))^2 + \nu \right)^{-(\nu-1)/2} \quad (\text{A.22})$$

where $VaR_x(\alpha; \nu) = F_x^{-1}(\alpha; \nu)$ is the α -quantile of a Student's $t(\nu)$ variable. The VaR for a standardized (unit variance) Student's t variable, Y , is simply:

$$ES_y(\alpha; \nu) = \sqrt{\frac{\nu-2}{\nu}} ES_x(\alpha; \nu) \quad (\text{A.23})$$

We use the connection between the PDF of a standardized Student's t random variable and Hansen's (1994) skew t variable, Z , to obtain an analogous expression for

the expected shortfall of a skew t random variable. As in Hansen (1994), define

$$a \equiv 4\lambda c \left(\frac{\nu - 2}{\nu - 1} \right), \quad b \equiv \sqrt{1 + 3\lambda^2 - a^2}, \quad c \equiv \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right) \sqrt{\pi(\nu-2)}} \quad (\text{A.24})$$

Define

$$ES_z^*(\alpha; \nu, \lambda) = \frac{\tilde{\alpha}}{\alpha} (1 - \lambda) \left(-\frac{a}{b} + \frac{1 - \lambda}{b} ES_y(\alpha; \nu) \right) \quad (\text{A.25})$$

where $\tilde{\alpha} \equiv F_z \left(\frac{b}{1 - \lambda} \left(VaR_y(\alpha; \nu) + \frac{a}{b} \right); \nu, 0 \right)$

where $F_z(\cdot; \nu, \lambda)$ is the CDF of the skew t distribution with parameters (ν, λ) . Let $VaR_z(\alpha; \nu, \lambda) = F_z^{-1}(\alpha; \nu, \lambda)$. It can then be shown that

$$ES_z(\alpha; \nu, \lambda) = \begin{cases} ES_z^*(\alpha; \nu, \lambda), & VaR_z(\alpha; \nu, \lambda) \leq -a/b \\ \frac{1-\alpha}{\alpha} ES_z^*(1-\alpha; \nu, -\lambda), & VaR_z(\alpha; \nu, \lambda) > -a/b \end{cases} \quad (\text{A.26})$$

Note that when $\lambda = 0$ this simplifies to a symmetric unit-variance Student's t variable, and we recover the expression above, i.e., $ES_z(\alpha; \nu, 0) = ES_y(\alpha; \nu)$. A Matlab function for the VaR and ES of a skew t variable is available at the link given in the first footnote of this chapter.

A.3 Estimation using the FZ0 loss function

The FZ0 loss function, equation (3.6), involves the indicator function $\mathbf{1}\{Y_t \leq v_t\}$ and so necessitates the use of a numerical search algorithm that does not rely on differentiability of the objective function; we use the function `fminsearch` in Matlab. However, in preliminary simulation analyses we found that this algorithm was sensitive to the starting values used in the search. To overcome this, we initially consider a

“smoothed” version of the FZ0 loss function, where we replace the indicator variable with a Logistic function:

$$\tilde{L}_{FZ0}(Y, v, e; \alpha, \tau) = -\frac{1}{\alpha e} \Gamma(Y_t, v_t; \tau) (v - Y) + \frac{v}{e} + \log(-e) - 1 \quad (\text{A.27})$$

$$\text{where } \Gamma(Y_t, v_t; \tau) \equiv \frac{1}{1 + \exp\{\tau(Y_t - v_t)\}}, \text{ for } \tau > 0 \quad (\text{A.28})$$

where τ is the smoothing parameter, and the smoothing function Γ converges to the indicator function as $\tau \rightarrow \infty$. In GAS models that involve an indicator function in the forcing variable, we alter the forcing variable in the same way, to ensure that the objective function as a function of θ is differentiable. In these cases the loss function *and* the model itself are slightly altered through this smoothing.

In our empirical implementation, we obtain “smart” (or “warm”) starting values by first estimating the model using the “smoothed FZ0” loss function with $\tau = 5$. This choice of τ gives some smoothing for values of Y_t that are roughly within ± 1 of v_t . Call the resulting parameter estimate $\tilde{\theta}_T^{(5)}$. Since this objective function is differentiable, we can use more familiar gradient-based numerical search algorithms, such as `fminunc` or `fmincon` in Matlab, which are often less sensitive to starting values. We then re-estimate the model, using $\tilde{\theta}_T^{(5)}$ as the starting value, setting $\tau = 20$ and obtain $\tilde{\theta}_T^{(20)}$. This value of τ smoothes values of Y_t within roughly ± 0.25 of v_t , and so this objective function is closer to the true objective function. Finally, we use $\tilde{\theta}_T^{(20)}$ as the starting value in the optimization of the actual FZ0 objective function, with no artificial smoothing, using the function `fminsearch`, and obtain $\hat{\theta}_T$. We found that this approach largely eliminated the sensitivity to starting values.

A.4 Dynamics in the Skew t distribution

For each parameter vector (ϑ, λ) of the Skew t distribution, we define $h(\vartheta, \lambda) \equiv [h_v(\vartheta, \lambda), h_e(\vartheta, \lambda)] = (v, e)$ to be the VaR and ES at a given value of α . Given the functional form of the Skew t distribution, not all pairs of VaR and ES are attainable from a set of parameters (ϑ, λ) and so the function h is not invertible everywhere. We define a pseudo-inverse of this mapping as

$$h^{(-1)}(v, e) \equiv \arg \min_{(\vartheta, \lambda)} (h_v(\vartheta, \lambda) - v)^2 + (h_e(\vartheta, \lambda) - e)^2 \quad (\text{A.29})$$

In words, the pseudo-inverse returns the Skew t parameters (ϑ, λ) that lead to VaR and ES that are as close as possible, in a squared-error distance metric, to the target values (v, e) .

To obtain dynamics in (ϑ, λ) that “offset” those in the GARCH volatility process, we set $(\vartheta_t, \lambda_t) = h^{(-1)}(\bar{v}/\sigma_t, \bar{e}/\sigma_t)$, and we fix $(\bar{v}, \bar{e}) = (\Phi^{-1}(\alpha), -\phi(\Phi^{-1}(\alpha))/\alpha)$. If the pseudo-inverse was actually the proper inverse, we would find:

$$(v_t, e_t) = \sigma_t h(\vartheta_t, \lambda_t) = \sigma_t h(h^{(-1)}(\bar{v}/\sigma_t, \bar{e}/\sigma_t)) = (\bar{v}, \bar{e}) \quad \forall t \quad (\text{A.30})$$

In our simulation study, the time series of (v_t, e_t) in the “offsetting” case were approximately but not perfectly flat, e.g., for $\alpha = 0.05$, the min-max spreads for v_t and e_t were around 0.05, when $(\bar{v}, \bar{e}) = (-1.65, -2.06)$. As the dynamics in volatility are not completely offset, this is helpful for QMLE and the ratios reported in Panel C of Table 5 are better than if the dynamics could be offset completely.

To obtain “amplifying” dynamics, we set $(\vartheta_t, \lambda_t) = h^{(-1)}(\bar{a}\sigma_t, \bar{b}\sigma_t)$, and we fix

$(\bar{a}, \bar{b}) = (\bar{v}, \bar{e}) / (2\bar{\sigma})$. If $h^{(-1)}$ were a proper inverse this would lead to:

$$(v_t, e_t) = \sigma_t h(\vartheta_t, \lambda_t) = \sigma_t h(h^{(-1)}(\bar{a}\sigma_t, \bar{b}\sigma_t)) = (\bar{a}, \bar{b}) \sigma_t^2 = \left(\frac{\bar{v}\sigma_t}{2\bar{\sigma}}, \frac{\bar{e}\sigma_t}{2\bar{\sigma}}\right) \sigma_t \equiv (a_t, b_t) \sigma_t \quad (\text{A.31})$$

and so the coefficients linking VaR and ES to conditional standard deviation are also increasing in the standard deviation, yielding the “amplifying” feature of this model. In our simulation study, the time series of (v_t, e_t) were almost perfectly linear in σ_t^2 , with R^2 values from regressions of v_t and e_t on σ_t^2 of over 0.998 in all cases.

Appendix B

Supplemental Appendix to Chapter 3

This appendix contains three parts. Part 1 presents lemmas that provide further details on the proof of Theorem 6 presented in Chapter 3. Part 2 presents a detailed verification that the high level assumptions made in the theorems of Chapter 3 hold for the widely-used GARCH(1,1) process. Part 3 contains additional tables of analysis.

B.1 Detailed proofs

Throughout this appendix, we suppress the subscript on $\hat{\theta}_T$ for simplicity of presentation, and we denote the conditional distribution and density functions as F_t and f_t rather than $F_t(\cdot|\mathcal{F}_{t-1})$ and $f_t(\cdot|\mathcal{F}_{t-1})$.

In Lemmas 1 and 3 below, we will refer to the expected score, defined as:

$$\begin{aligned}\lambda(\theta) &= \mathbb{E}[g_t(\theta)] \\ &= \mathbb{E}\left[\frac{1}{-e_t(\theta)}\left(\frac{F_t(v_t(\theta))}{\alpha} - 1\right)\nabla v_t(\theta)' + \right. \\ &\quad \left. \frac{1}{e_t(\theta)^2}\left(\frac{F_t(v_t(\theta))}{\alpha}v_t(\theta) - \frac{1}{\alpha}\mathbb{E}_{t-1}[Y_t|1\{Y_t \leq v_t(\theta)\}] - v_t(\theta) + e_t(\theta)\right)\nabla e_t(\theta)'\right]\end{aligned}$$

Lemma 1. *Let*

$$\Lambda(\theta^*) = \left. \frac{\partial \mathbb{E}[g_t(\theta)]}{\partial \theta} \right|_{\theta=\theta^*}$$

Then under Assumptions 4 and 5,

$$\sqrt{T}(\hat{\theta} - \theta^0) = (\Lambda^{-1}(\theta^0) + o_p(1)) \left(-\frac{1}{\sqrt{T}} \sum_{t=1}^T g_t(\theta^0) + o_p(1) \right)$$

Proof of Lemma 1. Consider a mean-value expansion of $\lambda(\hat{\theta})$ around θ^0 :

$$\begin{aligned} \lambda(\hat{\theta}) &= \lambda(\theta^0) + \left. \frac{\partial \mathbb{E}[g_t(\theta)]}{\partial \theta} \right|_{\theta=\theta^*} (\hat{\theta} - \theta^0) \\ &= \Lambda(\theta^*)(\hat{\theta} - \theta^0) \end{aligned}$$

where θ^* lies between $\hat{\theta}$ and θ^0 , and noting that $\lambda(\theta^0) = 0$ and the definition of $\Lambda(\theta^*)$ given in the statement of the lemma. Proving the claim involves two results: (I) $\Lambda^{-1}(\theta^*) = \Lambda^{-1}(\theta^0) + o_p(1)$, and (II) $\sqrt{T}\lambda(\hat{\theta}) = -\frac{1}{\sqrt{T}} \sum_{t=1}^T g_t(\theta^0) + o_p(1)$. Part (I) is easy to verify: Since $v_t(\theta)$ and $e_t(\theta)$ are twice continuously differentiable, and $e_t(\theta^0) < 0$, $\Lambda(\theta)$ is continuous in θ and $\Lambda(\theta)$ is non-singular in a neighborhood of θ^0 . Then by the continuous mapping theorem, $\theta^* \xrightarrow{p} \theta^0 \Rightarrow \Lambda(\theta^*)^{-1} \xrightarrow{p} \Lambda^{-1}(\theta^0)$. Establishing (II) builds on Theorem 3 of Huber (1967) and Lemma A.1 of Weiss (1991), which extends Huber's conclusion to the case of non-*iid* dependent random variables. We are going to verify the conditions of Weiss's Lemma A.1. Since the other conditions are easily checked, we only need to show that $T^{-1/2} \sum_{t=1}^T g_t(\hat{\theta}) = o_p(1)$, which we show in Lemma 2, and that his assumptions N3 and N4 hold, which we show in Lemmas 3-6. \square

Lemma 2. Under Assumptions 4 and 5, $T^{-1/2} \sum_{t=1}^T g_t(\hat{\theta}) = o_p(1)$.

Proof of Lemma 2. Let $\{e_j\}_{j=1}^p$ be the standard basis of \mathbb{R}^p and define

$$L_T^j(a) = T^{-1/2} \sum_{t=1}^T L_{FZ0} \left(Y_t, v_t(\hat{\theta} + ae_j), e_t(\hat{\theta} + ae_j); \alpha \right)$$

where a is a scalar. Let $G_T^j(a)$ (a scalar) be the right partial derivative of $L_T^j(a)$, that is

$$G_T^j(a) = T^{-1/2} \sum_{t=1}^T \left(\frac{\nabla_j v_t(\hat{\theta} + ae_j)}{-e_t(\hat{\theta} + ae_j)} \left(\frac{1}{\alpha} \mathbf{1} \left\{ Y_t \leq v_t(\hat{\theta} + ae_j) \right\} - 1 \right) + \frac{\nabla_j e_t(\hat{\theta} + ae_j)}{e_t(\hat{\theta} + ae_j)^2} \left(\frac{1}{\alpha} \mathbf{1} \left\{ Y_t \leq v_t(\hat{\theta} + ae_j) \right\} (v_t(\hat{\theta} + ae_j) - Y_t) - v_t(\hat{\theta} + ae_j) + e_t(\hat{\theta} + ae_j) \right) \right)$$

$G_T^j(0) = \lim_{\xi_1 \rightarrow 0^+} G_T^j(\xi_1)$ is the right partial derivative of $L_T(\theta)$ at $\hat{\theta}$ in the direction θ_j , while $\lim_{\xi_2 \rightarrow 0^+} G_T^j(-\xi_2)$ is the left partial derivative of $L_T(\theta)$ at $\hat{\theta}$ in the direction θ_j . Because $L_T(\theta)$ achieves its minimum at $\hat{\theta}$, and its left and right partial derivatives exist, its left derivative must be non-positive and its right derivative must be non-negative. Thus,

$$\begin{aligned} |G_T^j(0)| &\leq \lim_{\xi_1 \rightarrow 0^+} G_T^j(\xi_1) - \lim_{\xi_2 \rightarrow 0^+} G_T^j(-\xi_2) \\ &= T^{-1/2} \sum_{t=1}^T \left(\frac{\nabla_j v_t(\hat{\theta})}{-e_t(\hat{\theta})} \frac{1}{\alpha} \mathbf{1} \left\{ Y_t = v_t(\hat{\theta}) \right\} + \frac{\nabla_j e_t(\hat{\theta})}{e_t(\hat{\theta})^2} \frac{1}{\alpha} (v_t(\hat{\theta}) - Y_t) \mathbf{1} \left\{ Y_t = v_t(\hat{\theta}) \right\} \right) \\ &= T^{-1/2} \sum_{t=1}^T \frac{|\nabla_j v_t(\hat{\theta})|}{-e_t(\hat{\theta})} \frac{1}{\alpha} \mathbf{1} \left\{ Y_t = v_t(\hat{\theta}) \right\} \end{aligned}$$

The second term in the penultimate line vanishes as $\mathbf{1}\{Y_t = v_t(\hat{\theta})\}(v_t(\hat{\theta}) - Y_t)$ is always zero.

By Assumption 5(C), for all t , $|\nabla_j v_t(\hat{\theta})| \leq \|\nabla v_t(\hat{\theta})\| \leq V_1(\mathcal{F}_{t-1})$ and $\left| 1/e_t(\hat{\theta}) \right| \leq$

H , thus:

$$|G_T^j(0)| \leq \frac{H}{\alpha} \left[T^{-1/2} \max_{1 \leq t \leq T} V_1(\mathcal{F}_{t-1}) \right] \left[\sum_{t=1}^T \mathbf{1} \{Y_t = v_t(\hat{\theta})\} \right]$$

H is finite by Assumption 5(C). Next note that for all $\epsilon > 0$,

$$\Pr \left[T^{-1/2} \max_{1 \leq t \leq T} V_1(\mathcal{F}_{t-1}) > \epsilon \right] \leq \sum_{t=1}^T \Pr [V_1(\mathcal{F}_{t-1}) > \epsilon T^{1/2}] \leq \sum_{t=1}^T \frac{\mathbb{E}[V_1(\mathcal{F}_{t-1})^3]}{\epsilon^3 T^{3/2}} \rightarrow 0$$

with the latter inequality following from Markov's inequality. Since $\mathbb{E}[V_1(\mathcal{F}_{t-1})^3]$ is finite by assumption 5(D), we then have that $T^{-1/2} \max_{1 \leq t \leq T} V_1(\mathcal{F}_{t-1}) = o_p(1)$. Finally, by Assumption 5(G) we have $\sum_{t=1}^T \mathbf{1} \{Y_t = v_t(\hat{\theta})\} = \mathcal{O}_{a.s.}(1)$. We therefore have $G_T^j(0) \xrightarrow{p} 0$. Since this holds for every j , we have $T^{-1/2} \sum_{t=1}^T g_t(\hat{\theta}) = o_p(1)$.

□

The following three lemmas show each of the three parts of Assumption N3 of Weiss (1991) holds. In the proofs below we make repeated use of mean-value expansions, and we use θ^* to denote a point on the line connecting $\hat{\theta}$ and θ^0 , and θ^{**} to denote a point on the line connecting θ^* and θ^0 . The particular point on the line can vary from expansion to expansion.

Lemma 3. *Under Assumption 4 and 5, Assumption N3(i) of Weiss (1991) holds:*

$$\|\lambda_T(\theta)\| \geq a \|\theta - \theta^0\|, \text{ for } \|\theta - \theta^0\| \leq d_0.$$

for T sufficiently large, where a and d_0 are strictly positive numbers.

Proof of Lemma 3. A mean-value expansion yields:

$$\lambda_T(\hat{\theta}) = \lambda_T(\theta^0) + \Lambda(\theta^*)(\hat{\theta} - \theta^0) = \Lambda_T(\theta^*)(\hat{\theta} - \theta^0)$$

since $\lambda_T(\theta^0) = 0$, where $\Lambda(\theta) = \partial \mathbb{E}[g_t(\theta)] / \partial \theta$. Using the fact that

$$\frac{\partial \mathbb{E}[Y_t \mathbf{1}\{Y_t \leq v_t(\theta)\} | \mathcal{F}_{t-1}]}{\partial \theta} = \frac{\partial}{\partial \theta} \left\{ \int_{-\infty}^{v_t(\theta)} y f_t(y) dy \right\} = v_t(\theta) f_t(v_t(\theta)) \nabla v_t(\theta)$$

we can write:

$$\begin{aligned} \Lambda(\theta) = & \mathbb{E} \left[\left(\frac{\nabla^2 v_t(\theta)}{-e_t(\theta)} + \frac{\nabla v_t(\theta)' \nabla e_t(\theta)}{e_t(\theta)^2} + \frac{\nabla e_t(\theta)' \nabla v_t(\theta)}{e_t(\theta)^2} \right) \left(\frac{F_t(v_t(\theta))}{\alpha} - 1 \right) \right. \\ & + \left(\nabla^2 e_t(\theta) \frac{1}{e_t(\theta)^2} + \frac{-2}{e_t(\theta)^3} \nabla e_t(\theta)' \nabla e_t(\theta) \right) \\ & \cdot \left(\left(\frac{F_t(v_t(\theta))}{\alpha} - 1 \right) v_t(\theta) - \frac{1}{\alpha} \mathbb{E}[Y_t \mathbf{1}\{Y_t \leq v_t(\theta)\} | \mathcal{F}_{t-1}] + e_t(\theta) \right) \\ & + \frac{f_t(v_t(\theta))}{-\alpha e_t(\theta)} \nabla' v_t(\theta) \nabla v_t(\theta) \\ & \left. + \frac{1}{e_t(\theta)^2} \nabla' e_t(\theta) \nabla e_t(\theta) \right] \Big| \mathcal{F}_{t-1} \end{aligned}$$

Evaluated at θ^0 , the first two terms of Λ drop out because $F_t(v_t(\theta^0)) = \alpha$ and $\frac{1}{\alpha} \mathbb{E}[Y_t \mathbf{1}\{Y_t \leq v_t(\theta^0)\} | \mathcal{F}_{t-1}] = e_t(\theta^0)$. Define D as

$$D \equiv \Lambda(\theta^0) = T^{-1} \sum_{t=1}^T \mathbb{E} \left[\frac{f_t(v_t(\theta^0))}{-\alpha e_t(\theta^0)} \nabla v_t(\theta^0)' \nabla v_t(\theta^0) + \frac{1}{e_t(\theta^0)^2} \nabla e_t(\theta^0)' \nabla e_t(\theta^0) \right]$$

Below we show that $\Lambda(\theta^*) = D + O(\|\hat{\theta} - \theta^0\|)$ by decomposing $\|\Lambda_T(\theta^*) - D\|$ into four terms and showing that each is bounded by a $O(\|\hat{\theta} - \theta^0\|)$ term.

First term: Using a mean-value expansion around θ^0 and Assumptions 5(C)-(D)

Third term:

$$\begin{aligned}
& \left\| \mathbb{E} \left[\frac{f_t(v_t(\theta^*))}{-e_t(\theta^*)\alpha} \nabla v_t(\theta^*)' \nabla v_t(\theta^*) - \frac{f_t(v_t(\theta^0))}{-e_t(\theta^0)\alpha} \nabla v_t(\theta^0)' \nabla v_t(\theta^0) \right] \right\| \\
&= \frac{1}{\alpha} \left\| T^{-1} \sum_{t=1}^T \mathbb{E} \left\{ \frac{f_t(v_t(\theta^*))}{-e_t(\theta^*)} \nabla v_t(\theta^*)' \nabla v_t(\theta^*) - \frac{f_t(v_t(\theta^*))}{-e_t(\theta^*)} \nabla v_t(\theta^0)' \nabla v_t(\theta^*) \right. \right. \\
&+ \frac{f_t(v_t(\theta^*))}{-e_t(\theta^*)} \nabla v_t(\theta^0)' \nabla v_t(\theta^*) - \frac{f_t(v_t(\theta^0))}{-e_t(\theta^*)} \nabla v_t(\theta^0)' \nabla v_t(\theta^*) \\
&+ \frac{f_t(v_t(\theta^0))}{-e_t(\theta^*)} \nabla v_t(\theta^0)' \nabla v_t(\theta^*) - \frac{f_t(v_t(\theta^0))}{-e_t(\theta^0)} \nabla v_t(\theta^0)' \nabla v_t(\theta^*) \\
&+ \left. \left. \frac{f_t(v_t(\theta^0))}{-e_t(\theta^0)} \nabla v_t(\theta^0)' \nabla v_t(\theta^*) - \frac{f_t(v_t(\theta^0))}{-e_t(\theta^0)} \nabla v_t(\theta^0)' \nabla v_t(\theta^0) \right\} \right\| \\
&= \frac{1}{\alpha} \left\| T^{-1} \sum_{t=1}^T \mathbb{E} \left\{ \frac{f_t(v_t(\theta^*))}{-e_t(\theta^*)} [\nabla^2 v_t(\theta^{**}) (\theta^* - \theta^0)] \nabla v_t(\theta^*) \right. \right. \\
&+ \frac{f_t(v_t(\theta^*)) - f_t(v_t(\theta^0))}{-e_t(\theta^*)} \nabla v_t(\theta^0)' \nabla v_t(\theta^*) \\
&+ \frac{f_t(v_t(\theta^0))}{e_t(\theta^{**})^2} (\theta^* - \theta^0) \nabla v_t(\theta^0)' \nabla v_t(\theta^*) \\
&+ \left. \left. \frac{f_t(v_t(\theta^0))}{-e_t(\theta^0)} \nabla v_t(\theta^0)' (\theta^* - \theta^0)^2 v_t(\theta^{**}) \right\} \right\| \\
&\leq \frac{1}{\alpha} T^{-1} \sum_{t=1}^T \mathbb{E} \{ V_2(\mathcal{F}_{t-1}) (KH \cdot V_1(\mathcal{F}_{t-1})) + KH \cdot V_1(\mathcal{F}_{t-1})^3 \\
&+ KH^2 H_1(\mathcal{F}_{t-1}) V_1(\mathcal{F}_{t-1})^2 + KH V_1(\mathcal{F}_{t-1}) V_2(\mathcal{F}_{t-1}) \} \cdot \|\theta^* - \theta^0\|
\end{aligned}$$

Fourth term: The bound on this term follows similar steps to that of the third

term:

$$\begin{aligned}
& \left\| T^{-1} \sum_{t=1}^T \mathbb{E} \left\{ \frac{1}{e_t(\theta^*)^2} \nabla e_t(\theta^*)' \nabla e_t(\theta^*) - \frac{1}{e_t(\theta^0)^2} \nabla e_t(\theta^0)' \nabla e_t(\theta^0) \right\} \right\| \\
= & \left\| T^{-1} \sum_{t=1}^T \mathbb{E} \left\{ \frac{1}{e_t(\theta^*)^2} \nabla e_t(\theta^*)' \nabla e_t(\theta^*) - \frac{1}{e_t(\theta^*)^2} \nabla e_t(\theta^0)' \nabla e_t(\theta^*) \right. \right. \\
& \quad \left. \left. + \frac{1}{e_t(\theta^*)^2} \nabla e_t(\theta^0)' \nabla e_t(\theta^*) - \frac{1}{e_t(\theta^0)^2} \nabla e_t(\theta^0)' \nabla e_t(\theta^*) \right. \right. \\
& \quad \left. \left. + \frac{1}{e_t(\theta^0)^2} \nabla e_t(\theta^0)' \nabla e_t(\theta^*) - \frac{1}{e_t(\theta^0)^2} \nabla e_t(\theta^0)' \nabla e_t(\theta^0) \right\} \right\| \\
\leq & T^{-1} \sum_{t=1}^T \{ H^2 \cdot \mathbb{E}[H_1(\mathcal{F}_{t-1})H_2(\mathcal{F}_{t-1})] + 2H^3 \mathbb{E}[H_1(\mathcal{F}_{t-1})^3] \} \\
+ & H^2 \mathbb{E}[H_1(\mathcal{F}_{t-1})H_2(\mathcal{F}_{t-1})] \|\theta^* - \theta^0\|
\end{aligned}$$

Therefore, $\Lambda_T(\theta^*) = D_T + O(\|\hat{\theta} - \theta^0\|) \Rightarrow \|\Lambda_T(\theta^*) - D_T\| \leq K\|\hat{\theta} - \theta^0\|$, where K is some constant $< \infty$, for T sufficiently large. By Assumption 5(E), D_T has eigenvalues bounded below by a positive constant, denoted as a , for T sufficiently large. Thus,

$$\begin{aligned}
\|\lambda_T(\hat{\theta})\| &= \|\Lambda_T(\theta^*) (\hat{\theta} - \theta^0)\| \\
&= \|D_T(\hat{\theta} - \theta^0) - (D_T - \Lambda_T(\theta^*))(\hat{\theta} - \theta^0)\| \\
&\geq \|D_T(\hat{\theta} - \theta^0)\| - \|(D_T - \Lambda_T(\theta^*))(\hat{\theta} - \theta^0)\| \\
&\geq (a - K\|\hat{\theta} - \theta^0\|) \cdot \|\hat{\theta} - \theta^0\|
\end{aligned}$$

The penultimate inequality holds by the triangle inequality, and the final inequality follows from Assumption 5(E) on the minimum eigenvalue of D_T . Thus, for T sufficiently large so that $a - K\|\hat{\theta} - \theta^0\| > 0$, the result follows. \square

Lemma 4. *Define*

$$\mu_t(\theta, d) = \sup_{\|\tau - \theta\| \leq d} \|g_t(\tau) - g_t(\theta)\| \quad (\text{B.1})$$

Then under Assumption 4 and 5, Assumption N3(ii) of Weiss (1991) holds

$$\mathbb{E}[\mu_t(\theta, d)] \leq bd, \text{ for } \|\theta - \theta^0\| + d \leq d_0, d \geq 0 \quad (\text{B.2})$$

for T sufficiently large, where b , d , and d_0 are strictly positive numbers.

Proof of Lemma 4. In this proof, the strictly positive constant c and the mean-value expansion term, τ^* , can change from line to line. Pick d_0 such that for any θ that satisfies $\|\theta - \theta^0\| \leq d_0$, all the conditions in Assumption 5(C) and 5(D) hold as well as $e_t(\theta) \leq v_t(\theta) \leq 0$. Let us expand $g_t(\theta)$ into six terms:

$$\begin{aligned} g_t(\theta) = & \frac{1}{\alpha} \frac{\nabla' v_t(\theta)}{-e_t(\theta)} \mathbf{1}\{Y_t \leq v_t(\theta)\} - \frac{\nabla' v_t(\theta)}{-e_t(\theta)} + \frac{1}{\alpha} \frac{v_t(\theta) \nabla' e_t(\theta)}{e_t(\theta)^2} \mathbf{1}\{Y_t \leq v_t(\theta)\} \\ & - \frac{v_t(\theta) \nabla' e_t(\theta)}{e_t(\theta)^2} - \frac{1}{\alpha} \frac{\nabla' e_t(\theta)}{e_t(\theta)^2} \mathbf{1}\{Y_t \leq v_t(\theta)\} Y_t + \frac{\nabla' e_t(\theta)}{e_t(\theta)} \end{aligned} \quad (\text{B.3})$$

We will bound $\mu_t(\theta, d)$ by considering six terms, $\mu_t(\theta, d)^{(i)}$, $i = 1, 2, \dots, 6$, defined below. Each term is shown to be bounded by a constant times d .

First term:

$$\mu_t(\theta, d)^{(1)} = \frac{1}{\alpha} \sup_{\|\tau - \theta\| \leq d} \left\| \frac{\nabla' v_t(\tau)}{-e_t(\tau)} \mathbf{1}\{Y_t \leq v_t(\tau)\} - \frac{\nabla' v_t(\theta)}{-e_t(\theta)} \mathbf{1}\{Y_t \leq v_t(\theta)\} \right\| \quad (\text{B.4})$$

Set $\tau_1 = \arg \min_{\|\tau - \theta\| \leq d} v_t(\tau)$ and $\tau_2 = \arg \max_{\|\tau - \theta\| \leq d} v_t(\tau)$. Since $v_t(\theta)$ and $e_t(\theta)$ are assumed to be twice continuously differentiable, τ_1 and τ_2 exist. We want to take the indicator function out from the ‘sup’ operator. To this end, let us discuss what $\alpha \cdot \mu_t(\theta, d)^{(1)}$ equals in two cases.

Case 1: $Y_t \leq v_t(\theta)$. (a) If $Y_t > v_t(\tau_2)$, $\alpha \cdot \mu_t(\theta, d)^{(1)} = \left\| \frac{\nabla' v_t(\theta)}{-e_t(\theta)} \right\|$. (b) If $Y_t < v_t(\tau_1)$, $\alpha \cdot \mu_t(\theta, d)^{(1)} = \sup_{\|\tau - \theta\| \leq d} \left\| \frac{\nabla' v_t(\tau)}{-e_t(\tau)} - \frac{\nabla' v_t(\theta)}{-e_t(\theta)} \right\|$. (c) If $v_t(\tau_1) \leq Y_t \leq v_t(\tau_2)$,

$$\begin{aligned} \alpha \cdot \mu_t(\theta, d)^{(1)} &= \max \left\{ \sup_{\|\tau - \theta\| \leq d, Y_t \leq v(\tau)} \left\| \frac{\nabla' v_t(\tau)}{-e_t(\tau)} - \frac{\nabla' v_t(\theta)}{-e_t(\theta)} \right\|, \left\| \frac{\nabla' v_t(\theta)}{-e_t(\theta)} \right\| \right\} \\ &\leq \sup_{\|\tau - \theta\| \leq d} \left\| \frac{\nabla' v_t(\tau)}{-e_t(\tau)} - \frac{\nabla' v_t(\theta)}{-e_t(\theta)} \right\| + \left\| \frac{\nabla' v_t(\theta)}{-e_t(\theta)} \right\| \end{aligned} \quad (\text{B.5})$$

Case 2: $Y_t > v_t(\theta)$,

$$\begin{aligned} \alpha \cdot \mu_t(\theta, d)^{(1)} &= \mathbf{1}\{Y_t \leq v(\tau_2)\} \cdot \sup_{\|\tau - \theta\| \leq d, Y_t \leq v(\tau)} \left\| \frac{\nabla' v_t(\tau)}{-e_t(\tau)} \right\| \\ &\leq \mathbf{1}\{Y_t \leq v(\tau_2)\} \cdot \sup_{\|\tau - \theta\| \leq d} \left\| \frac{\nabla' v_t(\tau)}{-e_t(\tau)} \right\| \end{aligned} \quad (\text{B.6})$$

$\|\theta - \theta^0\| + d \leq d_0$ implies that both θ and τ (which are in a d -neighborhood of θ) are in a d_0 -neighborhood of θ_0 , and so

$$\left\| \frac{\nabla' v_t(\theta)}{-e_t(\theta)} \right\| \leq \sup_{\|\tau - \theta\| \leq d} \left\| \frac{\nabla' v_t(\tau)}{-e_t(\tau)} \right\| \leq \sup_{\|\theta - \theta^0\| \leq d_0} \left\| \frac{\nabla' v_t(\theta)}{-e_t(\theta)} \right\| \quad (\text{B.7})$$

Thus,

$$\begin{aligned} &\alpha \cdot \mu_t(\theta, d)^{(1)} \\ &\leq (\mathbf{1}\{v_t(\tau_2) < Y_t \leq v_t(\theta)\} + \mathbf{1}\{v_t(\tau_1) \leq Y_t \leq v_t(\theta)\} + \mathbf{1}\{v_t(\theta) < Y_t \leq v_t(\tau_2)\}) \\ &\quad \cdot \left(\sup_{\|\theta - \theta^0\| \leq d_0} \left\| \frac{\nabla' v_t(\theta)}{-e_t(\theta)} \right\| + \sup_{\|\tau - \theta\| \leq d} \left\| \frac{\nabla' v_t(\tau)}{-e_t(\tau)} - \frac{\nabla' v_t(\theta)}{-e_t(\theta)} \right\| \right), \end{aligned}$$

where

$$\begin{aligned}\mathbb{E}_{t-1}[\mathbf{1}\{v_t(\tau_2) < Y_t \leq v_t(\theta)\}] &= \int_{v_t(\tau_2)}^{v_t(\theta)} f_t(y) dy \\ &\leq K|v_t(\tau_2) - v_t(\theta)| \leq KV_1(\mathcal{F}_{t-1})\|\tau_2 - \theta\| \leq KV_1(\mathcal{F}_{t-1})d\end{aligned}$$

and similarly,

$$\begin{aligned}\mathbb{E}[\mathbf{1}\{v_t(\theta) < Y_t \leq v_t(\tau_2)\}|\mathcal{F}_{t-1}] &\leq KV_1(\mathcal{F}_{t-1})d \quad (\text{B.8}) \\ \text{and } \mathbb{E}[\mathbf{1}\{v_t(\tau_1) < Y_t \leq v_t(\theta)\}|\mathcal{F}_{t-1}] &\leq KV_1(\mathcal{F}_{t-1})d\end{aligned}$$

Further

$$\sup_{\|\theta - \theta^0\| \leq d} \left\| \frac{\nabla' v_t(\theta)}{-e_t(\theta)} \right\| \leq HV_1(\mathcal{F}_{t-1}) \quad (\text{B.9})$$

and by the mean-value theorem,

$$\frac{\nabla' v_t(\tau)}{-e_t(\tau)} - \frac{\nabla' v_t(\theta)}{-e_t(\theta)} = \left\| \frac{\nabla^2 v_t(\tau^*)}{-e_t(\tau^*)} + \frac{\nabla' v_t(\tau^*) \nabla e_t(\tau^*)}{e_t(\tau^*)^2} \right\| \cdot (\tau - \theta) \quad (\text{B.10})$$

$$\Rightarrow \sup_{\|\tau - \theta\| \leq d} \left\| \frac{\nabla' v_t(\tau)}{-e_t(\tau)} - \frac{\nabla' v_t(\theta)}{-e_t(\theta)} \right\| \leq (HV_2(\mathcal{F}_{t-1}) + H^2V_1(\mathcal{F}_{t-1})H_1(\mathcal{F}_{t-1})) \cdot d. \quad (\text{B.11})$$

By Assumption 5(D), $\mathbb{E}[V_2(\mathcal{F}_{t-1})]$ and $\mathbb{E}[V_1(\mathcal{F}_{t-1})H_1(\mathcal{F}_{t-1})]$ are finite, so $\mathbb{E}[\mu_t(\theta, d)^{(1)}] \leq cd$, where c is a strictly positive constant.

Second term: $\mu_t(\theta, d)^{(2)} = \sup_{\|\tau - \theta\| \leq d} \left\| \frac{\nabla' v_t(\tau)}{-e_t(\tau)} - \frac{\nabla' v_t(\theta)}{-e_t(\theta)} \right\|$. It was shown in the derivations for the first term that $\mathbb{E}[\mu_t(\theta, d)^{(2)}] \leq cd$, where c is a strictly positive constant.

Third term:

$$\mu_t(\theta, d)^{(3)} = \frac{1}{\alpha} \sup_{\|\tau - \theta\| \leq d} \left\| \frac{v_t(\tau) \nabla' e_t(\tau)}{e_t(\tau)^2} \mathbf{1}\{Y_t \leq v_t(\tau)\} - \frac{v_t(\theta) \nabla' e_t(\theta)}{e_t(\theta)^2} \mathbf{1}\{Y_t \leq v_t(\theta)\} \right\|$$

Similar to the first term, $\alpha \cdot \mu_t(\theta, d)^{(3)}$ can be bounded by

$$\begin{aligned} & (\mathbf{1}\{v_t(\tau_2) < Y_t \leq v_t(\theta)\} + \mathbf{1}\{v_t(\tau_1) \leq Y_t \leq v_t(\theta)\} + \mathbf{1}\{v_t(\theta) < Y_t \leq v_t(\tau_2)\}) \\ & \cdot \sup_{\|\theta - \theta^0\| \leq d_0} \left\| \frac{v_t(\theta) \nabla' e_t(\theta)}{e_t(\theta)^2} \right\| + \sup_{\|\tau - \theta\| \leq d} \left\| \frac{v_t(\tau) \nabla' e_t(\tau)}{e_t(\tau)^2} - \frac{v_t(\theta) \nabla' e_t(\theta)}{e_t(\theta)^2} \right\| \end{aligned}$$

where

$$\begin{aligned} & \mathbb{E}[\mathbf{1}\{v_t(\tau_2) < Y_t \leq v_t(\theta)\} + \mathbf{1}\{v_t(\tau_1) \leq Y_t \leq v_t(\theta)\} + \mathbf{1}\{v_t(\theta) < Y_t \leq v_t(\tau_2)\} | \mathcal{F}_{t-1}] \\ & \leq 3KV_1 \mathcal{F}_{t-1} d \end{aligned}$$

and

$$\sup_{\|\theta - \theta^0\| \leq d} \left\| \frac{v_t(\theta) \nabla' e_t(\theta)}{e_t(\theta)^2} \right\| \leq H \cdot H_1(\mathcal{F}_{t-1})$$

where $e_t(\theta) \leq v_t(\theta) \leq 0$ is used, and by the mean-value theorem,

$$\begin{aligned} & \frac{v_t(\tau) \nabla' e_t(\tau)}{e_t(\tau)^2} - \frac{v_t(\theta) \nabla' e_t(\theta)}{e_t(\theta)^2} \\ & = \left\| \frac{\nabla' e_t(\tau^*) \nabla v_t(\tau^*)}{e_t(\tau^*)^2} - \frac{2v_t(\tau^*) \nabla' e_t(\tau^*) \nabla e_t(\tau^*)}{e_t(\tau^*)^3} + \frac{v_t(\tau^*) \nabla^2 e_t(\tau^*)}{e_t(\tau^*)^2} \right\| \cdot (\tau - \theta) \\ & \Rightarrow \sup_{\|\tau - \theta\| \leq d} \left\| \frac{v_t(\tau) \nabla' e_t(\tau)}{e_t(\tau)^2} - \frac{v_t(\theta) \nabla' e_t(\theta)}{e_t(\theta)^2} \right\| \\ & \leq (H^2 V_1(\mathcal{F}_{t-1}) H_1(\mathcal{F}_{t-1}) + 2H^2 H_1(\mathcal{F}_{t-1})^2 + H \cdot H_2(\mathcal{F}_{t-1})) \cdot d \end{aligned}$$

By Assumption 5(D), $\mathbb{E}[V_1(\mathcal{F}_{t-1}) H_1(\mathcal{F}_{t-1})]$, $\mathbb{E}[H_1(\mathcal{F}_{t-1})^2]$, $\mathbb{E}[H_2(\mathcal{F}_{t-1})] < \infty$. There-

fore, $\mathbb{E}[\mu_t(\theta, d)^{(3)}] \leq cd$, where c is a strictly positive constant.

Fourth term: $\mu_t(\theta, d)^{(4)} = \sup_{\|\tau-\theta\| \leq d} \left\| \frac{v_t(\tau)\nabla' e_t(\tau)}{e_t(\tau)^2} - \frac{v_t(\theta)\nabla' e_t(\theta)}{e_t(\theta)^2} \right\|$. In the derivations for the third term we showed that $\mathbb{E}[\mu_t(\theta, d)^{(4)}] \leq cd$, where c is a strictly positive constant.

Fifth term:

$$\mu_t(\theta, d)^{(5)} = \frac{1}{\alpha} \sup_{\|\tau-\theta\| \leq d} \left\| \frac{\nabla' e_t(\tau)}{e_t(\tau)^2} \mathbf{1}\{Y_t \leq v_t(\tau)\} Y_t - \frac{\nabla' e_t(\theta)}{e_t(\theta)^2} \mathbf{1}\{Y_t \leq v_t(\theta)\} Y_t \right\|$$

Similar to the first term, $\alpha \cdot \mu_t(\theta, d)^{(5)}$ can be bounded by

$$\begin{aligned} & (\mathbf{1}\{v_t(\tau_2) < Y_t \leq v_t(\theta)\} + \mathbf{1}\{v_t(\tau_1) \leq Y_t \leq v_t(\theta)\} + \mathbf{1}\{v_t(\theta) < Y_t \leq v_t(\tau_2)\}) \\ & \cdot |Y_t| \sup_{\|\theta-\theta^0\| \leq d_0} \left\| \frac{\nabla' e_t(\theta)}{e_t(\theta)^2} \right\| + |Y_t| \sup_{\|\tau-\theta\| \leq d} \left\| \frac{\nabla' e_t(\tau)}{e_t(\tau)^2} - \frac{\nabla' e_t(\theta)}{e_t(\theta)^2} \right\| \end{aligned}$$

where

$$\begin{aligned} & \mathbb{E} [\mathbf{1}\{v_t(\tau_2) < Y_t \leq v_t(\theta)\} | Y_t | | \mathcal{F}_{t-1}] \\ & = \int_{v_t(\tau_2)}^{v_t(\theta)} |y| f_t(y) dy \leq K |v_t(\tau_2)| \cdot |v_t(\tau_2) - v_t(\theta)| \\ & \leq KV(\mathcal{F}_{t-1})V_1(\mathcal{F}_{t-1})\|\tau_2 - \theta\| \leq KV(\mathcal{F}_{t-1})V_1(\mathcal{F}_{t-1})d \end{aligned}$$

and similarly,

$$\begin{aligned} & \mathbb{E} [\mathbf{1}\{v_t(\tau_1) < Y_t \leq v_t(\theta)\} | Y_t | | \mathcal{F}_{t-1}] \leq KV(\mathcal{F}_{t-1})V_1(\mathcal{F}_{t-1})d \\ \text{and } & \mathbb{E} [\mathbf{1}\{v_t(\theta) < Y_t \leq v_t(\tau_2)\} | Y_t | | \mathcal{F}_{t-1}] \leq KV(\mathcal{F}_{t-1})V_1(\mathcal{F}_{t-1})d \end{aligned}$$

Further

$$\sup_{\|\theta - \theta^0\| \leq d} \left\| \frac{\nabla' e_t(\theta)}{e_t(\theta)^2} \right\| \leq H^2 H_1(\mathcal{F}_{t-1})$$

and by the mean-value theorem,

$$\begin{aligned} \frac{\nabla' e_t(\tau)}{e_t(\tau)^2} - \frac{\nabla' e_t(\theta)}{e_t(\theta)^2} &= \left\| -\frac{2\nabla' e_t(\tau^*)\nabla e_t(\tau^*)}{e_t(\tau^*)^3} + \frac{\nabla^2 e_t(\tau^*)}{e_t(\tau^*)^2} \right\| \cdot (\tau - \theta) \\ \Rightarrow \sup_{\|\tau - \theta\| \leq d} \left\| \frac{\nabla' e_t(\tau)}{e_t(\tau)^2} - \frac{\nabla' e_t(\theta)}{e_t(\theta)^2} \right\| &\leq (2H^3 H_1(\mathcal{F}_{t-1})^2 + H^2 H_2(\mathcal{F}_{t-1})) \cdot d \end{aligned}$$

By Assumption 5(D), $\mathbb{E}[V(\mathcal{F}_{t-1})V_1(\mathcal{F}_{t-1})H_1(\mathcal{F}_{t-1})]$, $\mathbb{E}[H_1(\mathcal{F}_{t-1})^2|Y_t]$, $\mathbb{E}[H_2(\mathcal{F}_{t-1})|Y_t]$ $< \infty$. Therefore, $\mathbb{E}[\mu_t(\theta, d)^{(5)}] \leq cd$, where c is a strictly positive constant.

Sixth term:

$$\mu_t^{(6)}(\theta, d) = \sup_{\|\tau - \theta\| \leq d} \left\| \frac{\nabla' e_t(\tau)}{-e_t(\tau)} - \frac{\nabla' e_t(\theta)}{-e_t(\theta)} \right\|$$

By the mean-value theorem,

$$\begin{aligned} \frac{\nabla' e_t(\tau)}{-e_t(\tau)} - \frac{\nabla' e_t(\theta)}{-e_t(\theta)} &= \left\| \frac{\nabla' e_t(\tau^*)\nabla e_t(\tau^*)}{e_t(\tau^*)^2} + \frac{\nabla^2 e_t(\tau^*)}{-e_t(\tau^*)} \right\| \cdot (\tau - \theta) \\ \Rightarrow \sup_{\|\tau - \theta\| \leq d} \left\| \frac{\nabla' e_t(\tau)}{-e_t(\tau)} - \frac{\nabla' e_t(\theta)}{-e_t(\theta)} \right\| &\leq (H^2 H_1(\mathcal{F}_{t-1})^2 + H \cdot H_2(\mathcal{F}_{t-1})) \cdot d. \end{aligned}$$

By Assumption 5(D), $\mathbb{E}[H_1(\mathcal{F}_{t-1})^2]$, $\mathbb{E}[H_2(\mathcal{F}_{t-1})] < \infty$. Therefore, $\mathbb{E}[\mu_t(\theta, d)^{(6)}] \leq cd$, where c is a strictly positive constant.

Thus we have shown that $\mu_t(\theta, d) \leq \sum_{i=1}^6 \mu_t(\theta, d)^{(i)}$ with $\mathbb{E}[\mu_t(\theta, d)^{(i)}] \leq cd$, $\forall i = 1, 2, \dots, 6$, where c is a strictly positive constant, proving the lemma. \square

Lemma 5. *Under Assumption 4 and 5, Assumption N3(iii) of Weiss (1991) holds:*

$$\mathbb{E}[\mu_t(\theta, d)^q] \leq cd, \text{ for } \|\theta - \theta^0\| + d \leq d_0, \text{ and some } q > 2$$

for T sufficiently large, and where $c > 0$, $d \geq 0$ and $d_0 > 0$.

Proof of Lemma 5. In this proof, the strictly positive constant c and the mean-value expansion term, τ^* , can change from line to line. Pick d_0 such that for any θ that satisfies $\|\theta - \theta^0\| \leq d_0$, all the conditions in Assumption 5(C) and 5(D) hold as well as $e_t(\theta) \leq v_t(\theta) \leq 0$. Similar to Lemma 4, we will decompose $\mu_t(\theta, d)$ into six terms, $\mu_t(\theta, d)^{(i)}$, for $i = 1, 2, \dots, 6$. By Jensen's inequality, $\mathbb{E}[\mu_t(\theta, d)^q] \leq 6^{q-1} \sum_{i=1}^6 \mathbb{E}[(\mu_t(\theta, d)^{(i)})^q]$, $q > 2$. We will show that for some $0 < \delta < 1$, $\mathbb{E}[(\mu_t(\theta, d)^{(i)})^{2+\delta}] \leq cd$, $\forall i = 1, 2, \dots, 6$, where c is a strictly positive constant.

First term:

$$\mu_t(\theta, d)^{(1)} = \frac{1}{\alpha} \sup_{\|\tau - \theta\| \leq d} \left\| \frac{\nabla' v_t(\tau)}{-e_t(\tau)} \mathbf{1}\{Y_t \leq v_t(\tau)\} - \frac{\nabla' v_t(\theta)}{-e_t(\theta)} \mathbf{1}\{Y_t \leq v_t(\theta)\} \right\|$$

Set $\tau_1 = \arg \min_{\|\tau - \theta\| \leq d} v_t(\tau)$ and $\tau_2 = \arg \max_{\|\tau - \theta\| \leq d} v_t(\tau)$. Following the same argument as in the proof of Lemma 4, we obtain

$$\begin{aligned} & [\alpha \cdot \mu_t(\theta, d)^{(1)}]^{2+\delta} \\ & \leq (\mathbf{1}\{v_t(\tau_2) < Y_t \leq v_t(\theta)\} + \mathbf{1}\{v_t(\tau_1) \leq Y_t \leq v_t(\theta)\} + \mathbf{1}\{v_t(\theta) < Y_t \leq v_t(\tau_2)\}) \\ & \quad \cdot \left(\sup_{\|\theta - \theta^0\| \leq d_0} \left\| \frac{\nabla' v_t(\theta)}{-e_t(\theta)} \right\| \right)^{2+\delta} + \left(\sup_{\|\tau - \theta\| \leq d} \left\| \frac{\nabla' v_t(\tau)}{-e_t(\tau)} - \frac{\nabla' v_t(\theta)}{-e_t(\theta)} \right\| \right)^{2+\delta} \end{aligned}$$

where

$$\begin{aligned} & \mathbb{E}_{t-1} [\mathbf{1}\{v_t(\tau_2) < Y_t \leq v_t(\theta)\} + \mathbf{1}\{v_t(\tau_1) \leq Y_t \leq v_t(\theta)\} + \mathbf{1}\{v_t(\theta) < Y_t \leq v_t(\tau_2)\}] \\ & \leq 3KV_1(\mathcal{F}_{t-1})d \end{aligned}$$

and

$$\left(\sup_{\|\theta - \theta^0\| \leq d} \left\| \frac{\nabla' v_t(\theta)}{-e_t(\theta)} \right\| \right)^{2+\delta} \leq (HV_1(\mathcal{F}_{t-1}))^{2+\delta}$$

For $\left(\sup_{\|\tau - \theta\| \leq d} \left\| \frac{\nabla' v_t(\tau)}{-e_t(\tau)} - \frac{\nabla' v_t(\theta)}{-e_t(\theta)} \right\| \right)^{2+\delta}$, we need to combine the two following two results:

$$\begin{aligned} \sup_{\|\tau - \theta\| \leq d} \left\| \frac{\nabla' v_t(\tau)}{-e_t(\tau)} - \frac{\nabla' v_t(\theta)}{-e_t(\theta)} \right\| &\leq (HV_2(\mathcal{F}_{t-1}) + H^2V_1(\mathcal{F}_{t-1})H_1(\mathcal{F}_{t-1})) d \\ \left(\sup_{\|\tau - \theta\| \leq d} \left\| \frac{\nabla' v_t(\tau)}{-e_t(\tau)} - \frac{\nabla' v_t(\theta)}{-e_t(\theta)} \right\| \right)^{1+\delta} &\leq (2HV_1(\mathcal{F}_{t-1}))^{1+\delta} \end{aligned}$$

Combining with Assumption 5(D), we thus have $\mathbb{E}[(\mu_t(\theta, d)^{(1)})^{2+\delta}] \leq cd$.

Second term: $\mu_t(\theta, d)^{(2)} = \sup_{\|\tau - \theta\| \leq d} \left\| \frac{\nabla' v_t(\tau)}{-e_t(\tau)} - \frac{\nabla' v_t(\theta)}{-e_t(\theta)} \right\|$. It was shown in the derivations for the first term that $\mathbb{E}[(\mu_t(\theta, d)^{(2)})^{2+\delta}] \leq cd$.

Third term:

$$\mu_t(\theta, d)^{(3)} = \frac{1}{\alpha} \sup_{\|\tau - \theta\| \leq d} \left\| \frac{v_t(\tau) \nabla' e_t(\tau)}{e_t(\tau)^2} \mathbf{1}\{Y_t \leq v_t(\tau)\} - \frac{v_t(\theta) \nabla' e_t(\theta)}{e_t(\theta)^2} \mathbf{1}\{Y_t \leq v_t(\theta)\} \right\|$$

Similar to the first term, $(\alpha \cdot \mu_t(\theta, d)^{(3)})^{2+\delta}$ can be bounded by

$$\begin{aligned} &(\mathbf{1}\{v_t(\tau_2) < Y_t \leq v_t(\theta)\} + \mathbf{1}\{v_t(\tau_1) \leq Y_t \leq v_t(\theta)\} + \mathbf{1}\{v_t(\theta) < Y_t \leq v_t(\tau_2)\}) \\ &\cdot \left(\sup_{\|\theta - \theta^0\| \leq d_0} \left\| \frac{v_t(\theta) \nabla' e_t(\theta)}{e_t(\theta)^2} \right\| \right)^{2+\delta} + \left(\sup_{\|\tau - \theta\| \leq d} \left\| \frac{v_t(\tau) \nabla' e_t(\tau)}{e_t(\tau)^2} - \frac{v_t(\theta) \nabla' e_t(\theta)}{e_t(\theta)^2} \right\| \right)^{2+\delta} \end{aligned}$$

where

$$\begin{aligned} & \mathbb{E}_{t-1}(\mathbf{1}\{v_t(\tau_2) < Y_t \leq v_t(\theta)\} + \mathbf{1}\{v_t(\tau_1) \leq Y_t \leq v_t(\theta)\} + \mathbf{1}\{v_t(\theta) < Y_t \leq v_t(\tau_2)\}) \\ & \leq 3KV_1(\mathcal{F}_{t-1})d \end{aligned}$$

and

$$\left(\sup_{\|\theta - \theta^0\| \leq d} \left\| \frac{v_t(\theta) \nabla' e_t(\theta)}{e_t(\theta)^2} \right\| \right)^{2+\delta} \leq (H \cdot H_1(\mathcal{F}_{t-1}))^{2+\delta}$$

For $\left(\sup_{\|\tau - \theta\| \leq d} \left\| \frac{v_t(\tau) \nabla' e_t(\tau)}{e_t(\tau)^2} - \frac{v_t(\theta) \nabla' e_t(\theta)}{e_t(\theta)^2} \right\| \right)^{2+\delta}$, we need to combine the following two results:

$$\begin{aligned} & \sup_{\|\tau - \theta\| \leq d} \left\| \frac{v_t(\tau) \nabla' e_t(\tau)}{e_t(\tau)^2} - \frac{v_t(\theta) \nabla' e_t(\theta)}{e_t(\theta)^2} \right\| \\ & \leq (H^2 V_1(\mathcal{F}_{t-1}) H_1(\mathcal{F}_{t-1}) + 2H^2 H_1(\mathcal{F}_{t-1})^2 + H \cdot H_2(\mathcal{F}_{t-1})) d \\ & \quad \left(\sup_{\|\tau - \theta\| \leq d} \left\| \frac{v_t(\tau) \nabla' e_t(\tau)}{e_t(\tau)^2} - \frac{v_t(\theta) \nabla' e_t(\theta)}{e_t(\theta)^2} \right\| \right)^{1+\delta} \\ & \leq (2H \cdot H_1(\mathcal{F}_{t-1}))^{1+\delta} \end{aligned}$$

Combining with Assumption 5(D), we thus have $\mathbb{E}[(\mu_t(\theta, d)^{(3)})^{2+\delta}] \leq cd$.

Fourth term: $\mu_t(\theta, d)^{(4)} = \sup_{\|\tau - \theta\| \leq d} \left\| \frac{v_t(\tau) \nabla' e_t(\tau)}{e_t(\tau)^2} - \frac{v_t(\theta) \nabla' e_t(\theta)}{e_t(\theta)^2} \right\|$. It was shown in the derivations for the third term that $\mathbb{E}[(\mu_t(\theta, d)^{(4)})^{2+\delta}] \leq cd$.

Fifth term:

$$\mu_t(\theta, d)^{(5)} = \frac{1}{\alpha} \sup_{\|\tau - \theta\| \leq d} \left\| \frac{\nabla' e_t(\tau)}{e_t(\tau)^2} \mathbf{1}\{Y_t \leq v_t(\tau)\} Y_t - \frac{\nabla' e_t(\theta)}{e_t(\theta)^2} \mathbf{1}\{Y_t \leq v_t(\theta)\} Y_t \right\|$$

Similar to the first and third terms, $(\alpha \cdot \mu_t(\theta, d)^{(5)})^{2+\delta}$ can be bounded by

$$\begin{aligned} & (\mathbf{1}\{v_t(\tau_2) < Y_t \leq v_t(\theta)\} + \mathbf{1}\{v_t(\tau_1) \leq Y_t \leq v_t(\theta)\} + \mathbf{1}\{v_t(\theta) < Y_t \leq v_t(\tau_2)\}) \\ & \cdot |Y_t|^{2+\delta} \left(\sup_{\|\theta - \theta^0\| \leq d_0} \left\| \frac{\nabla' e_t(\theta)}{e_t(\theta)^2} \right\| \right)^{2+\delta} + |Y_t|^{2+\delta} \left(\sup_{\|\tau - \theta\| \leq d} \left\| \frac{\nabla' e_t(\tau)}{e_t(\tau)^2} - \frac{\nabla' e_t(\theta)}{e_t(\theta)^2} \right\| \right)^{2+\delta} \end{aligned}$$

where

$$\begin{aligned} & \mathbb{E}_{t-1}[\mathbf{1}\{v_t(\tau_2) < Y_t \leq v_t(\theta)\} |Y_t|^{2+\delta}] \\ & = \int_{v_t(\tau_2)}^{v_t(\theta)} |y|^{2+\delta} f_t(y) dy \leq K |v_t(\tau_2)|^{2+\delta} \cdot |v_t(\tau_2) - v_t(\theta)| \\ & \leq KV(\mathcal{F}_{t-1})^{2+\delta} V_1(\mathcal{F}_{t-1}) \|\tau_2 - \theta\| \leq KV(\mathcal{F}_{t-1})^{2+\delta} V_1(\mathcal{F}_{t-1}) d \end{aligned}$$

and similarly,

$$\begin{aligned} & \mathbb{E}[\mathbf{1}\{v_t(\tau_1) < Y_t \leq v_t(\theta)\} |Y_t|^{2+\delta} | \mathcal{F}_{t-1}] \leq KV(\mathcal{F}_{t-1})^{2+\delta} V_1(\mathcal{F}_{t-1}) d \\ & \text{and } \mathbb{E}[\mathbf{1}\{v_t(\theta) < Y_t \leq v_t(\tau_2)\} |Y_t|^{2+\delta} | \mathcal{F}_{t-1}] \leq KV(\mathcal{F}_{t-1})^{2+\delta} V_1(\mathcal{F}_{t-1}) d \end{aligned}$$

Further

$$\sup_{\|\theta - \theta^0\| \leq d} \left\| \frac{\nabla' e_t(\theta)}{e_t(\theta)^2} \right\| \leq H^2 H_1(\mathcal{F}_{t-1})$$

For $\left(\sup_{\|\tau - \theta\| \leq d} \left\| \frac{\nabla' e_t(\tau)}{e_t(\tau)^2} - \frac{\nabla' e_t(\theta)}{e_t(\theta)^2} \right\| \right)^{2+\delta}$, we need to combine the following two results:

$$\begin{aligned} & \sup_{\|\tau - \theta\| \leq d} \left\| \frac{\nabla' e_t(\tau)}{e_t(\tau)^2} - \frac{\nabla' e_t(\theta)}{e_t(\theta)^2} \right\| \leq (2H^3 H_1(\mathcal{F}_{t-1})^2 + H^2 H_2(\mathcal{F}_{t-1})) d \\ & \left(\sup_{\|\theta - \theta^0\| \leq d} \left\| \frac{\nabla' e_t(\theta)}{e_t(\theta)^2} \right\| \right)^{1+\delta} \leq (2H^2 H_1(\mathcal{F}_{t-1}))^{1+\delta} \end{aligned}$$

Combining with Assumption 5(D), we thus have $\mathbb{E}[(\mu_t(\theta d)^{(5)})^{2+\delta}] \leq cd$.

Sixth term:

$$\mu_t^{(6)}(\theta, d) = \sup_{\|\tau - \theta\| \leq d} \left\| \frac{\nabla' e_t(\tau)}{-e_t(\tau)} - \frac{\nabla' e_t(\theta)}{-e_t(\theta)} \right\|$$

We have

$$\begin{aligned} \sup_{\|\tau - \theta\| \leq d} \left\| \frac{\nabla' e_t(\tau)}{-e_t(\tau)} - \frac{\nabla' e_t(\theta)}{-e_t(\theta)} \right\| &\leq (H^2 H_1(\mathcal{F}_{t-1})^2 + H H_2(\mathcal{F}_{t-1})) d \\ \left(\sup_{\|\tau - \theta\| \leq d} \left\| \frac{\nabla' e_t(\tau)}{-e_t(\tau)} - \frac{\nabla' e_t(\theta)}{-e_t(\theta)} \right\| \right)^{1+\delta} &\leq (2H H_1(\mathcal{F}_{t-1}))^{1+\delta} \end{aligned}$$

Combining with Assumption 5(D), we thus have $\mathbb{E}[(\mu_t(\theta, d)^{(6)})^{2+\delta}] \leq cd$. Thus $\mathbb{E}[\mu_t(\theta, d)^{(i)2+\delta}] \leq cd, \forall i = 1, 2, \dots, 6$, proving the lemma. \square

Lemma 6. *Under Assumption 4 and 5, $E\|g_t(\theta^0)\|^{2+\delta} \leq M$, for all t and some $M > 0$.*

Proof of Lemma 6.

$$\begin{aligned}
\mathbb{E}\|g_t(\theta^0)\|^{2+\delta} &\leq 4^{1+\delta} \left\{ \mathbb{E} \left\| \frac{\nabla' v_t(\theta^0)}{-e_t(\theta^0)} \left(\frac{1}{\alpha} \mathbf{1}\{Y_t \leq v_t(\theta^0)\} - 1 \right) \right\|^{2+\delta} \right. \\
&\quad + \mathbb{E} \left\| \frac{v_t(\theta^0) \nabla' e_t(\theta^0)}{e_t(\theta^0)^2} \left(\frac{1}{\alpha} \mathbf{1}\{Y_t \leq v_t(\theta^0)\} - 1 \right) \right\|^{2+\delta} \\
&\quad + \mathbb{E} \left\| \frac{\nabla' e_t(\theta^0)}{e_t(\theta^0)^2} \frac{1}{\alpha} \mathbf{1}\{Y_t \leq v_t(\theta^0)\} Y_t \right\|^{2+\delta} \\
&\quad \left. + \mathbb{E} \left\| \frac{\nabla' e_t(\theta^0)}{e_t(\theta^0)} \right\|^{2+\delta} \right\} \\
&\leq 4^{1+\delta} \left\{ \left(\frac{1}{\alpha} + 1 \right)^{2+\delta} H^{2+\delta} \mathbb{E} [V_1(\mathcal{F}_{t-1})^{2+\delta}] \right. \\
&\quad + \left(\frac{1}{\alpha} + 1 \right)^{2+\delta} H^{2+\delta} \mathbb{E} [H_1(\mathcal{F}_{t-1})^{2+\delta}] \\
&\quad + \frac{1}{\alpha^{2+\delta}} H^{4+2\delta} \mathbb{E} [H_1(\mathcal{F}_{t-1})^{2+\delta} |Y_t|^{2+\delta}] \\
&\quad \left. + H^{2+\delta} \mathbb{E} [H_1(\mathcal{F}_{t-1})^{2+\delta}] \right\} \\
&\leq M
\end{aligned}$$

since all the four expectations in the penultimate inequality are finite by assumption 5(D). Assumption N4 of Weiss (1991) only requires $E\|g_t(\theta^0)\|^2 \leq M$, which is implied by the above. \square

Lemma 7. *Under Assumption 4 and 5, we have $T^{-1/2} \sum_{t=1}^T g_t(\theta^0) \xrightarrow{d} N(0, \mathbf{A}_0)$ as $T \rightarrow \infty$, where $\mathbf{A}_0 \equiv \mathbb{E}[g_t(\theta^0)g_t(\theta^0)']$.*

Proof of Lemma 7. First note that the sequence $\{g_t(\theta^0)\}$ is stationary by Assumption 1(B)(ii), and has zero mean. Under Assumption 5(F) and Lemma 6, we can use Corollary 5.1 of Hall and Heyde (1980) and the Cramer-Wold device to obtain the result. \square

B.2 Estimating a GARCH(1,1) model by FZ loss minimization

In this appendix we show that we can estimate the popular GARCH(1,1) model via FZ loss minimization. We then verify that the assumptions required to show this are implied by the Assumption 4 and 5 in Chapter 3. Throughout, $\|\mathbf{x}\|$ refers to the Euclidean norm if \mathbf{x} is a vector and to the Frobenius norm if \mathbf{x} is a matrix.

Appendix B.2.1: Model specification

Assume that the data generating process for Y_t is:

$$\begin{aligned} Y_t &= \sigma_t \eta_t, \quad \eta_t \perp\!\!\!\perp \sigma_t, \quad \eta_t \sim iid F_\eta(0, 1) \\ \sigma_t^2 &= \omega_0 + \beta_0 \sigma_{t-1}^2 + \gamma_0 Y_{t-1}^2 \end{aligned}$$

Under this model, the conditional VaR and ES of Y_t at a probability level $\alpha \in (0, 1)$, that is $VaR_\alpha(Y_t|\mathcal{F}_{t-1})$ and $ES_\alpha(Y_t|\mathcal{F}_{t-1})$, follow the dynamics:

$$\begin{aligned} \begin{bmatrix} VaR_\alpha(Y_t|\mathcal{F}_{t-1}) \\ ES_\alpha(Y_t|\mathcal{F}_{t-1}) \end{bmatrix} &= \begin{bmatrix} c_0 \cdot ES_\alpha(Y_t|\mathcal{F}_{t-1}) \\ b_0 \cdot \sigma_t \end{bmatrix} \tag{B.12} \\ \text{where } c_0 &\equiv F_\eta^{-1}(\alpha)/\mathbb{E}[\eta_t|\eta_t \leq F_\eta^{-1}(\alpha)] \\ b_0 &\equiv \mathbb{E}[\eta_t|\eta_t \leq F_\eta^{-1}(\alpha)] \end{aligned}$$

We fix the level $\alpha \in (0, 1)$ throughout this appendix. Our goal is to estimate the parameter vector $\theta^0 = [\beta_0, \gamma_0, b_0, c_0]$ by minimizing the FZ loss function. Note that the parameters do not include ω_0 because only two of the three parameters ω_0, b_0, γ_0 are identifiable under this model. A detailed discussion about the identification of the GARCH model via FZ loss minimization is provided in Section SA.2.3 of this

appendix.

In the simulation study (Section 4 of Chapter 3), for estimating the GARCH model via FZ loss minimization, we fix ω at its true value ω_0 . Put $\theta = [\beta, \gamma, b, c]$ and its true value is $\theta^0 = [\beta_0, \gamma_0, b_0, c_0]$. We will estimate θ^0 by

$$\begin{aligned}\theta_T &\equiv \arg \min_{\theta \in \Theta} L_T(\theta) \\ \text{where } L_T(\theta) &= \frac{1}{T} \sum_{t=1}^T L_{FZ0}(Y_t, v_t(\theta), e_t(\theta); \alpha) \\ \sigma_t^2(\theta) &= \omega_0 + \beta \sigma_{t-1}^2(\theta) + \gamma Y_{t-1}^2 \\ v_t(\theta) &= c \cdot e_t(\theta) \\ e_t(\theta) &= b \cdot \sigma_t(\theta)\end{aligned}$$

and the FZ loss function L_{FZ0} is defined in equation (3.6)

Appendix B.2.2: Assumptions to estimate GARCH by FZ minimization

GARCH Assumption 1: F_η has zero mean, unit variance, finite fourth moment, and a unique α -quantile, which is non-positive. It has density $f_\eta(\cdot)$ that satisfies $f_\eta(\cdot) \leq K$ and $|f_\eta(\lambda_1) - f_\eta(\lambda_2)| \leq K|\lambda_1 - \lambda_2|$.

The distributions we often assume for the innovations of GARCH model, like the normal distribution or t-distribution with degrees of freedom greater than four, all satisfy this assumption.

GARCH Assumption 2: $0 < \omega_0 < \infty$. The true parameter vector $\theta^0 = [\beta_0, \gamma_0, b_0, c_0] \in \Theta \in \mathbb{R}^4$ is in the interior of Θ , a compact and convex parameter space. Specifically, for any vector $[\beta, \gamma, b, c] \in \Theta$, assume that $\delta_1 \leq \beta \leq (1 - \delta_1)$, $\delta_1 \leq \gamma \leq (1 - \delta_1)$ for some constant $\delta_1 > 0$, $\delta_2 \leq c \leq (1 - \delta_2)$, $-B_1 \leq b \leq -B_2$,

for some constants $\delta_2, B_1, B_2 > 0$, and $(\beta + \gamma)^2 + (\mathbb{E}[\eta_t^4] - 1)\gamma^2 \leq 1 - \delta_3$ for some constant $\delta_3 > 0$.

This assumption is similar to Assumption 1 of Lumsdaine (1996) with the exception of the third condition on the parameter vector, which is used to validate the mixing condition in Assumption 5(F), which we now discuss. It is not hard to show that

$$\sigma \{(Y_t, v_t(\theta), e_t(\theta), \nabla' v_t(\theta), \nabla' e_t(\theta))\} \subset \sigma \{(Y_t, \sigma_t^2(\theta), \partial \sigma_t^2(\theta) / \beta)\}$$

and thus we need to consider the mixing properties of $(Y_t, \sigma_t^2(\theta), \partial \sigma_t^2(\theta) / \beta)$. Using Definition 3 of Carrasco and Chen (2002), $\{(Y_t, \sigma_t^2(\theta), \partial \sigma_t^2(\theta) / \beta), t \geq 0\}$ is a generalized hidden Markov model with a hidden chain $\{(\sigma_t^2(\theta), \partial \sigma_t^2(\theta) / \beta), t \geq 0\}$. By their Proposition 4, if $(\sigma_t^2(\theta), \partial \sigma_t^2(\theta) / \beta)$ is stationary and β -mixing then $(Y_t, \sigma_t^2(\theta), \partial \sigma_t^2(\theta) / \beta)$ is stationary and β -mixing with a decay rate at least as fast as that of $\{(\sigma_t^2(\theta), \partial \sigma_t^2(\theta) / \beta), t \geq 0\}$.

We use Proposition 3 of Carrasco and Chen (2002). First, we express $\{(\sigma_t^2(\theta), \partial \sigma_t^2(\theta) / \beta), t \geq 0\}$ in the polynomial random coefficient form:

$$\begin{pmatrix} \sigma_t^2(\theta) \\ \partial \sigma_t^2(\theta) / \beta \end{pmatrix} = \begin{pmatrix} \omega_0 \\ 0 \end{pmatrix} + \begin{pmatrix} \gamma \eta_{t-1}^2 + \beta & 0 \\ 1 & \beta \end{pmatrix} \begin{pmatrix} \sigma_{t-1}^2(\theta) \\ \partial \sigma_{t-1}^2(\theta) / \beta \end{pmatrix} \quad (\text{B.13})$$

First, note that by GARCH Assumption 1 $\{\eta_t^2\}$ satisfies their condition (e) and that by GARCH Assumption 2, their Assumption A_0 is obviously satisfied. For Assump-

tion A_1 , the spectral radius of $\begin{pmatrix} \beta & 0 \\ 1 & \beta \end{pmatrix}$ is $\beta < 1$. For Assumption A'_2 , the spectral radius of $\begin{pmatrix} \gamma \eta_{t-1}^2 + \beta & 0 \\ 1 & \beta \end{pmatrix}$ is $\mathbb{E}[(\gamma \eta_{t-1}^2 + \beta)^2] = (\beta + \gamma)^2 + (\mathbb{E}[\eta_t^4] - 1)\gamma^2 <$

1. Then, if $(\sigma_t^2(\theta), \partial\sigma_t^2(\theta)/\beta)$ is initialized from the invariant distribution (which we did in our simulations) then $\{(\sigma_t^2(\theta), \partial\sigma_t^2(\theta)/\beta), t \geq 0\}$ is strictly stationary and β -mixing with exponential decay. It is well known that β -mixing implies α -mixing and so Assumption 2(F) of Chapter 3 is satisfied.

GARCH Assumptions 1–2 imply that the distribution and density of Y_t conditional on \mathcal{F}_{t-1} satisfy Assumption 5(B)(i). Since $Y_t = \sigma_t \eta_t$ and $\sigma_t \in \mathcal{F}_{t-1}$,

$$\begin{aligned} F_t(x|\mathcal{F}_{t-1}) &= F_\eta\left(\frac{x}{\sigma_t}\right) \\ f_t(x|\mathcal{F}_{t-1}) &= \frac{1}{\sigma_t} f_\eta\left(\frac{x}{\sigma_t}\right) \end{aligned}$$

Thus,

$$|f_t(x|\mathcal{F}_{t-1})| \leq \frac{K}{\sqrt{\omega_0}}, \text{ since } \sigma_t^2 = \omega_0 + \beta\sigma_{t-1}^2 + \gamma y_{t-1}^2 \geq \omega_0 > 0$$

$$|f_t(\lambda_1|\mathcal{F}_{t-1}) - f_t(\lambda_2|\mathcal{F}_{t-1})| = \frac{1}{\sigma_t} |f_\eta\left(\frac{\lambda_1}{\sigma_t}\right) - f_\eta\left(\frac{\lambda_2}{\sigma_t}\right)| \leq \frac{K}{\sigma_t^2} |\lambda_1 - \lambda_2| \leq \frac{K}{\omega_0} |\lambda_1 - \lambda_2|$$

GARCH Assumption 3: $\mathbb{E}|Y_t|^{5+\delta} < \infty$, for some $\delta > 0$.

GARCH Assumption 3 is needed to show the uniform LLN Assumption 1(A) of Chapter 3 and also to ensure the moment conditions in Assumptions 5 (C) and (D).

For the GARCH model it is possible to obtain the results of Chapter 3 under a weaker version of Assumption 5(D). An inspection of the proofs shows that it is sufficient to replace Assumption 5(D) by the following.

Assumption 5(D’): For some $0 < \delta < 1$ and $\forall t$:

- (i) $\mathbb{E} \left[V_1(\mathcal{F}_{t-1})^{3+\delta} \right], \mathbb{E} \left[H_1(\mathcal{F}_{t-1})^{3+\delta} \right], \mathbb{E} \left[V_2(\mathcal{F}_{t-1})^{\frac{3+\delta}{2}} \right], \mathbb{E} \left[H_2(\mathcal{F}_{t-1})^{\frac{3+\delta}{2}} \right] \leq K,$
- (ii) $\mathbb{E} \left[V(\mathcal{F}_{t-1})^{2+\delta} V_1(\mathcal{F}_{t-1})^{1+\delta} \right] \leq K,$
- (iii) $\mathbb{E} \left[H_1(\mathcal{F}_{t-1})^{2+\delta} |Y_t|^{2+\delta} \right], \mathbb{E} \left[H_2(\mathcal{F}_{t-1})^{1+\delta} |Y_t|^{2+\delta} \right] \leq K.$

Assumption 5(D’) is in turn fulfilled if $\mathbb{E}|Y_t|^{4+\delta} < \infty$, for some $\delta > 0$. For

reasons of brevity, we omit the arguments and work instead with the stronger GARCH Assumption 3.

Appendix B.2.3: Identification

In Theorem 1, we discussed the identification of a general dynamic model for ES and VaR model by minimizing the FZ loss, with the form of a general model given by equation (3.4). Under correct specification of the model, that is

$(VaR_\alpha(Y_t|\mathcal{F}_{t-1}), ES_\alpha(Y_t|\mathcal{F}_{t-1})) = (v_t(\theta^0), e_t(\theta^0)) \forall t$ a.s., the condition required for identification is given by Assumption 4(B) (iv): $\Pr[v_t(\theta) = v_t(\theta^0) \cap e_t(\theta) = e_t(\theta^0)] = 1, \forall t \Rightarrow \theta = \theta^0$. This assumption is equivalent to

$$\Pr[v_t(\theta) = v_t(\theta^0) \cap e_t(\theta) = e_t(\theta^0), \forall t] = 1$$

In the case of the GARCH model we have:

$$\begin{aligned} & \Pr[\{v_t(\theta) = v_t(\theta^0)\} \cap \{e_t(\theta) = e_t(\theta^0)\}, \forall t] = 1 \\ \Rightarrow & \Pr[\{c \cdot e_t(\theta) = c_0 \cdot e_t(\theta^0)\} \cap \{e_t(\theta) = e_t(\theta^0)\}, \forall t] = 1 \\ \Rightarrow & \Pr[\{c = c_0\} \cap \{b \cdot \sigma_t(\theta) = b_0 \cdot \sigma_t(\theta^0)\}, \forall t] = 1 \\ \Rightarrow & c = c_0, \Pr[b^2 \cdot \sigma_t^2(\theta) = b_0^2 \cdot \sigma_t^2(\theta^0), \forall t] = 1 \\ \Rightarrow & c = c_0, \Pr[b^2(\omega + \beta\sigma_{t-1}^2(\theta) + \gamma Y_{t-1}^2) = b_0^2(\omega_0 + \beta_0\sigma_{t-1}^2(\theta^0) + \gamma_0 Y_{t-1}^2), \forall t] = 1 \\ \Rightarrow & c = c_0, \Pr[b^2\omega + \beta b_0^2\sigma_{t-1}^2(\theta^0) + b^2\gamma Y_{t-1}^2 = b_0^2\omega_0 + \beta_0 b_0^2\sigma_{t-1}^2(\theta^0) + b_0^2\gamma_0 Y_{t-1}^2, \forall t] = 1 \end{aligned}$$

where the third line holds because $e_t(\theta^0) = b_0\sigma_t(\theta^0)$ and we assume that $b_0 < 0$, thus $e_t(\theta^0) < 0$, and in the last line, we replaced $b^2\sigma_{t-1}^2(\theta)$ by $b_0^2\sigma_{t-1}^2(\theta^0)$ because we started with $b^2\sigma_t^2(\theta) = b_0^2\sigma_t^2(\theta^0), \forall t$ almost surely.

Since the GARCH model assumes that $Y_{t-1}|\sigma_{t-1}(\theta^0) \sim F_\eta(0, \sigma_{t-1}^2(\theta^0))$,

$$\begin{aligned} & \Pr [b^2\omega + \beta b_0^2\sigma_{t-1}^2(\theta^0) + b^2\gamma Y_{t-1}^2 = b_0^2\omega_0 + \beta_0 b_0^2\sigma_{t-1}^2(\theta^0) + b_0^2\gamma_0 Y_{t-1}^2, \forall t] = 1 \\ \Rightarrow & \Pr [\{b^2\gamma = b_0^2\gamma_0\} \cap \{b^2\omega + \beta b_0^2\sigma_{t-1}^2(\theta^0) = b_0^2\omega_0 + \beta_0 b_0^2\sigma_{t-1}^2(\theta^0)\}, \forall t] = 1 \\ \Rightarrow & b^2\gamma = b_0^2\gamma_0, \quad \Pr [b^2\omega + \beta b_0^2\sigma_{t-1}^2(\theta^0) = b_0^2\omega_0 + \beta_0 b_0^2\sigma_{t-1}^2(\theta^0), \forall t] = 1 \end{aligned}$$

If $\beta b_0^2 \neq \beta_0 b_0^2$ and $\Pr [b^2\omega + \beta b_0^2\sigma_{t-1}^2(\theta^0) = b_0^2\omega_0 + \beta_0 b_0^2\sigma_{t-1}^2(\theta^0), \forall t] = 1$ hold at the same time, then we have

$$\begin{aligned} & \Pr \left[\sigma_{t-1}^2(\theta^0) = \frac{b_0^2\omega_0 - b^2\omega}{\beta b_0^2 - \beta_0 b_0^2}, \forall t \right] = 1 \\ \Rightarrow & \Pr \left[\omega_0 + \beta_0 \sigma_{t-2}^2(\theta^0) + \gamma_0 Y_{t-2}^2 = \frac{b_0^2\omega_0 - b^2\omega}{\beta b_0^2 - \beta_0 b_0^2}, \forall t \right] = 1 \end{aligned}$$

This contradicts the assumption of the GARCH model, that $Y_{t-2}|\sigma_{t-2}(\theta^0) \sim F_\eta(0, \sigma_{t-2}^2(\theta^0))$. Thus, $\beta b_0^2 \neq \beta_0 b_0^2$ and $\Pr [b^2\omega + \beta b_0^2\sigma_{t-1}^2(\theta^0) = b_0^2\omega_0 + \beta_0 b_0^2\sigma_{t-1}^2(\theta^0), \forall t] = 1$ cannot hold at the same time. This means that $\Pr [b^2\omega + \beta b_0^2\sigma_{t-1}^2(\theta^0) = b_0^2\omega_0 + \beta_0 b_0^2\sigma_{t-1}^2(\theta^0), \forall t] = 1$ implies $\beta b_0^2 = \beta_0 b_0^2$, which further implies that $\beta = \beta_0$ and $b^2\omega = b_0^2\omega_0$. In summary, we have shown that

$$\begin{aligned} & \Pr [v_t(\theta) = v_t(\theta^0) \cap e_t(\theta) = e_t(\theta^0)] = 1, \forall t \\ \Rightarrow & c = c_0, \quad b^2\gamma = b_0^2\gamma_0, \quad \beta b_0^2 = \beta_0 b_0^2, \quad b^2\omega = b_0^2\omega_0 \\ \Rightarrow & c = c_0, \quad \beta = \beta_0, \quad b^2\gamma = b_0^2\gamma_0, \quad b^2\omega = b_0^2\omega_0 \end{aligned}$$

Therefore, Assumption 4(B)(iv) holds if we normalize one of the three parameters b, γ, ω . We choose to normalize ω .

Appendix B.2.4: Uniform LLN

In this section, we show that under the GARCH assumptions we have made in

Section SA.2.2, Assumption 4(A) is satisfied: $L_{FZ0}(Y_t, v_t(\theta), e_t(\theta); \alpha)$ obeys the uniform law of large numbers.

Since the parameter space is assumed to be compact, we can establish the uniform LLN by combining the pointwise LLN with stochastic equicontinuity.

Appendix B.2.4.1: LLN

The LLN is based on Davidson (1994, Corollary 19.3) which we restate here as Theorem 12 for convenience.

Theorem 12 (Davidson). *Suppose that $(X_t)_{t \in \mathbb{N}}$ satisfies:*

$$\sup_{t \in \mathbb{N}} \mathbb{E}|X_t|^{2+\delta} < \infty \text{ for some } \delta > 0$$

and $(X_t)_{t \in \mathbb{N}}$ is α mixing with $\sum_{m=1}^{\infty} m^{-1} \alpha(m)^{\delta/(2+\delta)} < \infty$, then

$$\frac{1}{n} \sum_{t=1}^n (X_t - \mathbb{E}[X_t]) \xrightarrow{L_2} 0.$$

Under Assumption 5(F), which we discussed in the context of the GARCH model in Section SA.2.2 above, implies that $L_{FZ0}(Y_t, v_t(\theta), e_t(\theta))$ is α -mixing with a decay rate no slower than that required by Theorem 12. Also, $L_{FZ0}(Y_t, v_t(\theta), e_t(\theta))$ is strictly stationary as we have shown that $(Y_t, \sigma_t^2(\theta), \partial \sigma_t^2(\theta) / \partial \beta)$ is strictly stationary. We then need only show that

$$\mathbb{E}|L_{FZ0}(Y_t, v_t(\theta), e_t(\theta); \alpha)|^{2+\delta} < \infty$$

$$\begin{aligned}
& |L_{FZ0}(Y_t, v_t(\theta), e_t(\theta); \alpha)| \\
&= \left| -\frac{1}{\alpha e_t(\theta)} \mathbf{1}\{Y_t \leq v_t(\theta)\} (v_t(\theta) - Y_t) + \frac{v_t(\theta)}{e_t(\theta)} + \log(-e_t(\theta)) - 1 \right| \\
&= \left| \frac{v_t(\theta)}{e_t(\theta)} \cdot \left(1 - \frac{1}{\alpha} \mathbf{1}\{Y_t \leq v_t(\theta)\}\right) + \frac{Y_t}{\alpha e_t(\theta)} \mathbf{1}\{Y_t \leq v_t(\theta)\} + \log(-e_t(\theta)) - 1 \right| \\
&= \left| c \cdot \left(1 - \frac{1}{\alpha} \mathbf{1}\{Y_t \leq v_t(\theta)\}\right) - \frac{\eta_t}{\alpha b} \mathbf{1}\{Y_t \leq v_t(\theta)\} + \log(-b\sigma_t(\theta)) - 1 \right| \\
&\leq c \left(1 + \frac{1}{\alpha}\right) + |\log(-b)| + 1 + \frac{|\eta_t|}{\alpha|b|} + |\log \sigma_t|
\end{aligned}$$

By Cr-inequality, it is sufficient to show that

$$\mathbb{E}|\eta_t|^{2+\delta} < \infty \quad \text{and} \quad \mathbb{E}|\log \sigma_t|^{2+\delta} < \infty.$$

The moment condition on η_t is directly implied by the structure of the model and GARCH Assumption 3. Recall that $\sigma_t^2 = \omega_0 + \beta\sigma_{t-1}^2 + \gamma y_{t-1}^2 \geq \omega_0 > 0$. Therefore, if $\sigma_t^2 < 1$ then $|\log \sigma_t| \leq |\log \sqrt{\omega_0}|$, and if $\sigma_t^2 \geq 1$ then $|\log \sigma_t| \leq \sigma_t$. In summary, $|\log \sigma_t| \leq |\log \sqrt{\omega_0}| + \sigma_t$. Therefore, by Cr-inequality

$$\begin{aligned}
\mathbb{E}|\log \sigma_t|^{2+\delta} &\leq \mathbb{E}(|\log \sqrt{\omega_0}| + \sigma_t)^{2+\delta} \\
&\leq 2^{1+\delta} (|\log \sqrt{\omega_0}|^{2+\delta} + \mathbb{E}\sigma_t^{2+\delta}).
\end{aligned}$$

Thus, a sufficient condition for $\mathbb{E}|\log \sigma_t|^{2+\delta} < \infty$ is $\mathbb{E}[\sigma_t^{2+\delta}] < \infty$ which is implied by GARCH Assumption 3. Hence, $L_{FZ0}(Y_t, v_t(\theta), e_t(\theta))$ obeys the law of large numbers for any fixed θ by Theorem 12.

Appendix B.2.4.2: Stochastic equicontinuity

The stochastic equicontinuity condition is derived using Davidson (1994, Theorem 21.10) which we restate here as Theorem 13 for convenience.

Theorem 13 (Davidson). *Let $Q_n(\cdot)$ be the objective function for an M -estimator.*

Suppose there exists $N \in \mathbb{N}$ such that

$$|Q_n(\theta) - Q_n(\theta')| \leq a_n h(\|\theta - \theta'\|), \text{ a.s.}$$

holds for all $\theta, \theta' \in \Theta$ and $n \geq N$, where h is a deterministic function with $h(x) \downarrow 0$ as $x \downarrow 0$, and $a_n = \mathcal{O}_p(1)$. Then $(Q_n)_{n \in \mathbb{N}}$ is stochastically equicontinuous.

Observe that

$$\mathbf{1}\{Y_t \leq v_t(\theta)\}(v_t(\theta) - Y_t) = \frac{1}{2}(v_t(\theta) - Y_t + |v_t(\theta) - Y_t|). \quad (\text{B.14})$$

Let $\theta_1 = [\beta_1, \gamma_1, b_1, c_1], \theta_2 = [\beta_2, \gamma_2, b_2, c_2] \in \Theta$. Then, using (B.14), we obtain

$$\begin{aligned} & \left| \frac{\mathbf{1}\{Y_t \leq v_t(\theta_1)\}(v_t(\theta_1) - Y_t)}{e_t(\theta_1)} - \frac{\mathbf{1}\{Y_t \leq v_t(\theta_2)\}(v_t(\theta_2) - Y_t)}{e_t(\theta_2)} \right| \\ &= \frac{1}{2} \left| \frac{v_t(\theta_1) - Y_t + |v_t(\theta_1) - Y_t|}{e_t(\theta_1)} - \frac{v_t(\theta_2) - Y_t + |v_t(\theta_2) - Y_t|}{e_t(\theta_2)} \right| \\ &\leq \frac{1}{2} \left| \frac{v_t(\theta_1) - Y_t}{e_t(\theta_1)} - \frac{v_t(\theta_2) - Y_t}{e_t(\theta_2)} \right| + \frac{1}{2} \left| \frac{|v_t(\theta_1) - Y_t|}{e_t(\theta_1)} - \frac{|v_t(\theta_2) - Y_t|}{e_t(\theta_2)} \right| \end{aligned} \quad (\text{B.15})$$

$$\begin{aligned} &\leq \left| \frac{v_t(\theta_1) - Y_t}{e_t(\theta_1)} - \frac{v_t(\theta_2) - Y_t}{e_t(\theta_2)} \right| \quad (\text{B.16}) \\ &= \left| \frac{v_t(\theta_1)}{e_t(\theta_1)} - \frac{v_t(\theta_2)}{e_t(\theta_2)} - \left(\frac{Y_t}{e_t(\theta_1)} - \frac{Y_t}{e_t(\theta_2)} \right) \right| \\ &= \left| c_1 - c_2 - \left(\frac{\eta_t}{b_1} - \frac{\eta_t}{b_2} \right) \right| \\ &\leq |c_1 - c_2| + \frac{|b_1 - b_2|}{|b_1 b_2|} |\eta_t|. \end{aligned}$$

The inequality between (B.15) and (B.16) holds because

$$\begin{aligned} \left| \frac{|v_t(\theta_1) - Y_t|}{e_t(\theta_1)} - \frac{|v_t(\theta_2) - Y_t|}{e_t(\theta_2)} \right| &= \left| \frac{|v_t(\theta_1) - Y_t|}{-|e_t(\theta_1)|} - \frac{|v_t(\theta_2) - Y_t|}{-|e_t(\theta_2)|} \right| \\ &= \left| \frac{|v_t(\theta_2) - Y_t|}{|e_t(\theta_2)|} - \frac{|v_t(\theta_1) - Y_t|}{|e_t(\theta_1)|} \right| \\ &\leq \left| \frac{v_t(\theta_2) - Y_t}{e_t(\theta_2)} - \frac{v_t(\theta_1) - Y_t}{e_t(\theta_1)} \right|. \end{aligned}$$

By Taylor's theorem,

$$|\log(-e_t(\theta_1)) - \log(-e_t(\theta_2))| = \left| \frac{1}{-e_t(\theta_1^*)} \right| \cdot \|\nabla e_t(\theta_1^*)\| \cdot \|\theta_1 - \theta_2\|$$

for some θ_1^* between θ_1 and θ_2 . Since,

$$\begin{aligned} \|\nabla e_t(\theta)\| &= \|b \cdot \nabla \sigma_t(\theta) + \sigma_t(\theta) \cdot [0, 0, 1, 0]\| \\ &\leq |b| \cdot \|\nabla \sigma_t(\theta)\| + \sigma_t(\theta) \\ \Rightarrow \frac{\|\nabla e_t(\theta)\|}{|e_t(\theta)|} &\leq \frac{|b| \cdot \|\nabla \sigma_t(\theta)\| + \sigma_t(\theta)}{|b \cdot \sigma_t(\theta)|} \leq \frac{\|\nabla \sigma_t(\theta)\|}{\sigma_t(\theta)} + \frac{1}{|b|}, \end{aligned}$$

we obtain

$$|\log(-e_t(\theta_1)) - \log(-e_t(\theta_2))| \leq \left(\frac{\|\nabla \sigma_t(\theta_1^*)\|}{\sigma_t(\theta_1^*)} + \frac{1}{|b_1^*|} \right) \cdot \|\theta_1 - \theta_2\|.$$

Therefore,

$$\begin{aligned} &|L_{FZ0}(Y_t, v_t(\theta_1), e_t(\theta_1); \alpha) - L_{FZ0}(Y_t, v_t(\theta_2), e_t(\theta_2); \alpha)| \\ &\leq \frac{1}{\alpha} \left(|c_1 - c_2| + \frac{|b_1 - b_2|}{|b_1 b_2|} |\eta_t| \right) + |c_1 - c_2| + \left(\frac{\|\nabla \sigma_t(\theta_1^*)\|}{\sigma_t(\theta_1^*)} + \frac{1}{|b_1^*|} \right) \cdot \|\theta_1 - \theta_2\| \\ &\leq \left(1 + \frac{1}{\alpha} + \frac{1}{B_2} + \frac{|\eta_t|}{\alpha B_2^2} + \frac{\|\nabla \sigma_t(\theta_1^*)\|}{\sigma_t(\theta_1^*)} \right) \cdot \|\theta_1 - \theta_2\| \end{aligned}$$

The last inequality holds because by GARCH Assumption 2, $|b_1|, |b_2|, |b_1^*| \geq B_2 > 0$.

Using Lemma 8 in Section SA.2.5, we obtain

$$\begin{aligned} \frac{\|\nabla\sigma_t(\theta)\|}{\sigma_t(\theta)} &\leq \frac{1}{2} \cdot \left[\gamma^{1/2}\beta^{-1/2}\sigma_t(\theta)^{-1} \sum_{i=2}^{\infty} (i-1)\beta^{(i-2)/2}|Y_{t-i}| + \gamma^{-1} \right] \\ &\leq \frac{1}{2} \cdot \left[(1-\delta_1)^{1/2}\delta_1^{-1/2}\omega_0^{-1/2} \sum_{i=2}^{\infty} (i-1)(1-\delta_1)^{(i-2)/2}|Y_{t-i}| + \delta_1^{-1} \right] \end{aligned}$$

using the bounds on γ and β in GARCH Assumption 2. Define

$$Z_t = \frac{1}{2} \cdot \left[(1-\delta_1)^{1/2}\delta_1^{-1/2}\omega_0^{-1/2} \sum_{i=2}^{\infty} (i-1)(1-\delta_1)^{(i-2)/2}|Y_{t-i}| + \delta_1^{-1} \right]$$

Then,

$$|L_T(\theta_1) - L_T(\theta_2)| \leq \left(1 + \frac{1}{\alpha} + \frac{1}{B_2} + \frac{1}{\alpha B_2^2} \cdot \frac{1}{T} \sum_{t=1}^T |\eta_t| + \frac{1}{T} \sum_{t=1}^T Z_t \right) \cdot \|\theta_1 - \theta_2\|.$$

$\mathbb{E}|\eta_t| = \text{const} < \infty$ and $\mathbb{E}[Z_t] = \text{const} < \infty$ (because $\mathbb{E}|Y_t| = \text{const} < \infty$ by GARCH Assumptions 2 and 3). Then, $\frac{1}{T} \sum_{t=1}^T |\eta_t| = \mathcal{O}_p(1)$ and $\frac{1}{T} \sum_{t=1}^T Z_t = \mathcal{O}_p(1)$.

Therefore, by Theorem 13, L_T is stochastically equicontinuous.

Appendix B.2.5: Assumptions 5(C) and 5(D)

We define

$$\begin{aligned} X_1(t; \beta) &= \sum_{i=2}^{\infty} (i-1)\beta^{i-2}Y_{t-i}^2 \\ X_2(t; \beta) &= \sum_{i=2}^{\infty} (i-1)\beta^{(i-2)/2}|Y_{t-i}| \\ X_3(t; \beta) &= \sum_{i=3}^{\infty} \frac{(i-1)(i-2)}{2}\beta^{i-3}Y_{t-i}^2 \end{aligned}$$

Lemma 8. *Under GARCH Assumption 2, we have*

$$\begin{aligned}
\|\nabla\sigma_t^2(\theta)\| &\leq \gamma X_1(t; \beta) + \gamma^{-1}\sigma_t^2(\theta) \\
\|\nabla\sigma_t(\theta)\| &\leq \frac{1}{2} \cdot [\gamma^{1/2}\beta^{-1/2}X_2(t; \beta) + \gamma^{-1}\sigma_t(\theta)] \\
\|\nabla^2\sigma_t^2(\theta)\| &\leq 2\left[\frac{\omega_0}{(1-\beta)^3} + \gamma X_3(t; \beta) + X_1(t; \beta)\right] \\
\|\nabla^2\sigma_t(\theta)\| &\leq \frac{\omega_0^{-1/2}}{4} \cdot \gamma\beta^{-1}X_2(t; \beta)^2 + \gamma^{-1/2}\beta^{-1/2}X_2(t; \beta) + \frac{1}{4}\gamma^{-2}\sigma_t(\theta) \\
&\quad + \omega_0^{-1/2} \cdot \left[\frac{\omega_0}{(1-\beta)^3} + \gamma X_3(t; \beta) + X_1(t; \beta)\right]
\end{aligned}$$

Proof of Lemma 8.

$$\sigma_t^2(\theta) = \omega_0 + \beta\sigma_{t-1}^2(\theta) + \gamma Y_{t-1}^2 = \frac{\omega_0}{1-\beta} + \gamma \sum_{i=1}^{\infty} \beta^{i-1} Y_{t-i}^2. \quad (\text{B.17})$$

Therefore,

$$\begin{aligned}
\nabla\sigma_t^2(\theta) &= \left[\omega_0/(1-\beta)^2 + \gamma \sum_{i=2}^{\infty} (i-1)\beta^{i-2} Y_{t-i}^2, \sum_{i=1}^{\infty} \beta^{i-1} Y_{t-i}^2, 0, 0 \right] \\
&= [\omega_0/(1-\beta)^2 + \gamma X_1(t; \beta), \gamma^{-1}(\sigma_t^2 - \omega_0/(1-\beta)), 0, 0] \quad (\text{B.18}) \\
\|\nabla\sigma_t^2(\theta)\| &\leq \omega_0/(1-\beta)^2 + \gamma X_1(t; \beta) + \gamma^{-1}(\sigma_t^2 - \omega_0/(1-\beta)) \\
&= \gamma X_1(t; \beta) + \gamma^{-1}\sigma_t^2 + \frac{\omega_0}{1-\beta} \left(\frac{1}{1-\beta} - \frac{1}{\gamma} \right) \\
&\leq \gamma X_1(t; \beta) + \gamma^{-1}\sigma_t^2,
\end{aligned}$$

since $\beta + \gamma \leq 1$ implies $1/(1-\beta) - 1/\gamma \leq 0$. Furthermore,

$$\|\nabla\sigma_t(\theta)\| = \frac{\|\nabla\sigma_t^2(\theta)\|}{2\sigma_t(\theta)} \leq \frac{1}{2} \left[\frac{\gamma \sum_{i=2}^{\infty} (i-1)\beta^{i-2} Y_{t-i}^2}{\sqrt{\frac{\omega_0}{1-\beta} + \gamma \sum_{i=1}^{\infty} \beta^{i-1} Y_{t-i}^2}} + \frac{\gamma^{-1}\sigma_t^2}{\sigma_t} \right].$$

For all $j \geq 2$, we have

$$\frac{\gamma(j-1)\beta^{j-2}Y_{t-j}^2}{\sqrt{\frac{\omega_0}{1-\beta} + \gamma \sum_{i=1}^{\infty} \beta^{i-1}Y_{t-i}^2}} \leq \frac{\gamma(j-1)\beta^{j-2}Y_{t-j}^2}{\sqrt{\gamma\beta^{j-1}Y_{t-j}^2}} = (j-1)\gamma^{1/2}\beta^{(j-3)/2}|Y_{t-j}|$$

as all summands in the denominator are positive. This implies

$$\begin{aligned} \|\nabla\sigma_t(\theta)\| &\leq \frac{1}{2} \cdot \left[\gamma^{1/2}\beta^{-1/2} \sum_{i=2}^{\infty} (i-1)\beta^{(i-2)/2}|Y_{t-i}| + \gamma^{-1}\sigma_t(\theta) \right] \\ &= \frac{1}{2} \cdot [\gamma^{1/2}\beta^{-1/2}X_2(t; \beta) + \gamma^{-1}\sigma_t(\theta)]. \end{aligned}$$

Using equation (B.18), we obtain

$$\nabla^2\sigma_t^2(\theta) = \begin{bmatrix} a_{11} & a_{12} & 0 & 0 \\ a_{21} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

where

$$\begin{aligned} a_{11} &= \frac{2\omega_0}{(1-\beta)^3} + \gamma \sum_{i=3}^{\infty} (i-1)(i-2)\beta^{i-3}Y_{t-i}^2 \\ a_{12} &= a_{21} = \sum_{i=2}^{\infty} (i-1)\beta^{i-2}Y_{t-i}^2 \end{aligned}$$

Thus, since the Frobenius norm is always less than the sum of the absolute values of the matrix entries,

$$\|\nabla^2\sigma_t^2(\theta)\| \leq 2\left[\frac{\omega_0}{(1-\beta)^3} + \gamma X_3(t; \beta) + X_1(t; \beta)\right].$$

Using that $\nabla\sigma_t(\theta) = \nabla\sigma_t^2(\theta)/(2\sigma_t(\theta))$, we find that

$$\nabla^2\sigma_t(\theta) = \frac{\nabla_t'^2(\theta)\nabla\sigma_t(\theta)}{-2\sigma_t^2(\theta)} + \frac{\nabla^2\sigma_t^2(\theta)}{2\sigma_t(\theta)} = \frac{\nabla'\sigma_t(\theta)\nabla\sigma_t(\theta)}{-\sigma_t(\theta)} + \frac{\nabla^2\sigma_t^2(\theta)}{2\sigma_t(\theta)},$$

therefore,

$$\|\nabla^2\sigma_t(\theta)\| \leq \frac{\|\nabla\sigma_t(\theta)\|^2}{\sigma_t(\theta)} + \frac{\|\nabla^2\sigma_t^2(\theta)\|}{2\sigma_t(\theta)}$$

Since $\sigma_t^2 \geq \omega_0 > 0$ and using our previous results, we obtain the claimed bound on $\|\nabla^2\sigma_t(\theta)\|$. \square

Lemma 9. *Under GARCH Assumption 2, it holds that*

$$\begin{aligned} |v_t(\theta)| &\leq V(\mathcal{F}_{t-1}) = B_1 \cdot S_1(\mathcal{F}_{t-1}) \\ \|\nabla e_t(\theta)\| &\leq H_1(\mathcal{F}_{t-1}) = B_1 \cdot S_2(\mathcal{F}_{t-1}) + S_1(\mathcal{F}_{t-1}) \\ \|\nabla v_t(\theta)\| &\leq V_1(\mathcal{F}_{t-1}) = H_1(\mathcal{F}_{t-1}) + V(\mathcal{F}_{t-1}) \\ \|\nabla^2 e_t(\theta)\| &\leq H_2(\mathcal{F}_{t-1}) = B_1 \cdot S_3(\mathcal{F}_{t-1}) + 2S_2(\mathcal{F}_{t-1}) \\ \|\nabla^2 v_t(\theta)\| &\leq V_2(\mathcal{F}_{t-1}) = H_2(\mathcal{F}_{t-1}) + 2H_1(\mathcal{F}_{t-1}), \end{aligned}$$

where

$$\begin{aligned} S_1(\mathcal{F}_{t-1}) &= \sqrt{\omega_0\delta_1^{-1} + \gamma \sum_{i=1}^{\infty} (1-\delta_1)^{i-1} Y_{t-i}^2} \\ S_2(\mathcal{F}_{t-1}) &= \frac{1}{2} \cdot [(1-\delta_1)^{1/2}\delta_1^{-1/2} X_2(t; 1-\delta_1) + \delta_1^{-1} S_1(\mathcal{F}_{t-1})] \\ S_3(\mathcal{F}_{t-1}) &= \frac{\omega_0^{-1/2}}{4} \cdot (\delta_1^{-1} - 1) X_2(t; 1-\delta_1)^2 + \delta_1^{-1} X_2(t; 1-\delta_1) + \frac{1}{4} \delta_1^{-2} S_1(\mathcal{F}_{t-1}) \\ &\quad + \omega_0^{-1/2} \cdot [\omega_0\delta_1^{-3} + (1-\delta_1) X_3(t; 1-\delta_1) + X_1(t; 1-\delta_1)]. \end{aligned}$$

Proof of Lemma 9. As a function in β and γ , $\sigma_t(\theta)$ is increasing in both arguments,

see equation (B.17), and, in fact, it does not depend on the parameters b and c . Therefore, $\sigma_t(\theta) \leq S_1(\mathcal{F}_{t-1})$. The quantities $X_1(t, \beta)$, $X_2(t, \beta)$, $X_3(t, \beta)$ defined in the beginning of this section are all increasing in β , and thus, bounded by $X_1(t, 1 - \delta_1)$, $X_2(t, 1 - \delta_1)$, $X_3(t, 1 - \delta_1)$, respectively. Recall that $|b| \leq B_1$ under GARCH Assumption 2.

The first inequality holds because $|v_t(\theta)| \leq |e_t(\theta)| = |b| \cdot \sigma_t(\theta)$. The remaining ones are implied by Lemma 8 and

$$\begin{aligned} \|\nabla e_t(\theta)\| &= \|b \cdot \nabla \sigma_t(\theta) + \sigma_t(\theta) \cdot [0, 0, 1, 0]\| \leq |b| \cdot \|\nabla \sigma_t(\theta)\| + \sigma_t(\theta) \\ \|\nabla v_t(\theta)\| &= \|c \cdot \nabla e_t(\theta) + e_t(\theta) \cdot [0, 0, 0, 1]\| \leq \|\nabla e_t(\theta)\| + |b| \cdot \sigma_t(\theta) \\ \|\nabla^2 e_t(\theta)\| &= \|b \nabla^2 \sigma_t(\theta) + [0, 0, 1, 0]' \nabla \sigma_t(\theta) + \nabla' \sigma_t(\theta) [0, 0, 1, 0]\| \\ &\leq |b| \cdot \|\nabla^2 \sigma_t(\theta)\| + 2 \|\nabla \sigma_t(\theta)\| \\ \|\nabla^2 v_t(\theta)\| &= \|c \nabla^2 e_t(\theta) + [0, 0, 0, 1]' \nabla e_t(\theta) + \nabla' e_t(\theta) [0, 0, 0, 1]\| \\ &\leq \|\nabla^2 e_t(\theta)\| + 2 \|\nabla e_t(\theta)\|. \end{aligned}$$

□

Lemma 10. *Let $\{X_i\}_{i \in \mathbb{N}}$ be a sequence of random variables and define $X = \sum_{i=1}^{\infty} a_i |X_i|$, where $a_i > 0$ for all $i \in \mathbb{N}$ and $\sum_{i=1}^{\infty} a_i < \infty$. Let $p > 1$. If $\sup_{i \in \mathbb{N}} \mathbb{E}|X_i|^p \leq K < \infty$ for some constant K , then $\mathbb{E}|X|^p \leq (\sum_{i=1}^{\infty} a_i)^p K$.*

Proof of Lemma 10. By Jensen's inequality, $\mathbb{E}|Z|^p \geq |\mathbb{E}Z|^p$. We rewrite X as

$$X = \sum_{i=1}^{\infty} a_i |X_i| = \left(\sum_{i=1}^{\infty} a_i \right) \cdot \sum_{i=1}^{\infty} \left(\sum_{i=1}^{\infty} a_i \right)^{-1} a_i |X_i|.$$

Note that $\sum_{i=1}^{\infty} (\sum_{i=1}^{\infty} a_i)^{-1} a_i = 1$, namely $\{(\sum_{i=1}^{\infty} a_i)^{-1} a_i\}_{i=1}^{\infty}$ is a probability mea-

sure. Then, using Jensen's inequality,

$$X^p = \left(\sum_{j=1}^{\infty} a_j \right)^p \cdot \left(\sum_{i=1}^{\infty} \left(\sum_{j=1}^{\infty} a_j \right)^{-1} a_i |X_i| \right)^p \leq \left(\sum_{j=1}^{\infty} a_j \right)^p \cdot \sum_{i=1}^{\infty} \left(\sum_{j=1}^{\infty} a_j \right)^{-1} a_i |X_i|^p.$$

Thus,

$$\begin{aligned} \mathbb{E}[X^p] &\leq \left(\sum_{j=1}^{\infty} a_j \right)^p \cdot \sum_{i=1}^{\infty} \left(\sum_{j=1}^{\infty} a_j \right)^{-1} a_i \mathbb{E}|X_i|^p \\ &\leq \left(\sum_{j=1}^{\infty} a_j \right)^p \cdot \sum_{i=1}^{\infty} \left(\sum_{j=1}^{\infty} a_j \right)^{-1} a_i K \\ &= \left(\sum_{j=1}^{\infty} a_j \right)^p K \end{aligned}$$

□

Lemma 11. *Under GARCH Assumption 2 and for any $p > 1$, $p_1, \dots, p_6 > 0$, the following statements hold for all t :*

(1) *If $\mathbb{E}|Y_t|^p < \infty$, then the following quantities are all finite: $\mathbb{E}[V^p(\mathcal{F}_{t-1})]$, $\mathbb{E}[V_1^p(\mathcal{F}_{t-1})]$, $\mathbb{E}[H_1^p(\mathcal{F}_{t-1})]$, $\mathbb{E}[V_2^{p/2}(\mathcal{F}_{t-1})]$, $\mathbb{E}[H_2^{p/2}(\mathcal{F}_{t-1})]$.*

(2) *If $p = p_1 + p_2 + p_3 + 2p_4 + 2p_5 + p_6$ and $\mathbb{E}|Y_t|^p < \infty$, then*

$$\mathbb{E}[V^{p_1}(\mathcal{F}_{t-1})V_1^{p_2}(\mathcal{F}_{t-1})H_1^{p_3}(\mathcal{F}_{t-1})V_2^{p_4}(\mathcal{F}_{t-1})H_2^{p_5}(\mathcal{F}_{t-1})|Y_t|^{p_6}] < \infty.$$

(3) *If $\mathbb{E}|Y_t|^{4+\delta} < \infty$ for some $\delta > 0$, all the moment conditions in Assumption 2(D)' could be satisfied.*

Proof of Lemma 11. Part (1) follows by combining Lemma 9 with Lemma 10, and part (2) is a consequence of part (1) and Hölder's inequality. □

Lemma 11 implies that GARCH Assumption 3 implies Assumption 5(D) of Chapter 3.

Appendix B.2.6: Assumption 2(E)

\mathbf{D}_0 is the Hessian of the expected loss at θ^0 , so it is positive semi-definite. Let $x = (x_1, \dots, x_4)' \in \mathbb{R}^4$ such that $x' \mathbf{D}_0 x = 0$. This implies that $x' \nabla v_t(\theta^0) = 0$, $x' \nabla e_t(\theta^0) = 0$ almost surely. We have, $x' \nabla v_t(\theta^0) = cx' \nabla e_t(\theta^0) + x_4 e_t(\theta^0)$. Therefore, $x_4 = 0$. Furthermore,

$$\begin{aligned} 2\sigma_t(\theta^0)x' \nabla e_t(\theta^0) &= 2\sigma_t(\theta^0)bx' \nabla \sigma_t(\theta^0) + 2x_3\sigma_t^2(\theta^0) \\ &= bx' \nabla \sigma_t^2(\theta^0) + 2x_3\sigma_t^2(\theta^0) = 0, \text{ a.s.} \end{aligned} \quad (\text{B.19})$$

The stationarity of $\{Y_t\}$ implies that $\sigma_t^2(\theta^0)$ is stationary. Therefore it also holds that

$$bx' \nabla \sigma_{t-1}^2(\theta^0) + 2x_3\sigma_{t-1}^2(\theta^0) = 0, \text{ a.s.} \quad (\text{B.20})$$

Computing (B.19) $-\beta \cdot$ (B.20), we obtain that a.s.

$$0 = bx'[\sigma_{t-1}^2(\theta^0), Y_{t-1}^2, 0, 0]' + 2x_3(\omega_0 + \gamma Y_{t-1}^2) = (bx_2 + 2\gamma x_3)Y_{t-1}^2 + (2\omega_0 x_3 + bx_1\sigma_{t-1}^2(\theta^0)). \quad (\text{B.21})$$

By the assumption that $Y_{t-1} | \sigma_{t-1}^2 \sim F_\eta(0, \sigma_{t-1}^2(\theta^0))$ and that $\sigma_{t-1}^2(\theta^0) = \omega_0 + \beta_0 \sigma_{t-2}^2(\theta^0) + \gamma_0 Y_{t-2}^2$, we can conclude from the above equation that $x_1 = x_2 = x_3 = 0$. Thus \mathbf{D}_0 is positive definite.

Appendix B.2.7: Assumption 5(G)

We now verify this assumption for the GARCH(1,1) model. Set $a = bc$, so that $v_t = a\sigma_t$. Then for $T \geq 5$, a necessary condition for $Y_t = v_t(\theta)$, $t = 1, \dots, T$ is given

by the set of equations

$$Y_t^2 = a^2 \beta^t \sigma_0^2 + a^2 \beta^{t-1} (\omega_0 + \gamma Y_0^2) + a^2 \sum_{k=1}^{t-1} \beta^{t-1-k} (\omega_0 + \gamma Y_k^2), \quad t = 1, \dots, 4$$

or, equivalently,

$$Y_1^2 = a^2 \beta \sigma_0^2 + a^2 (\omega_0 + \gamma Y_0^2) \tag{B.22}$$

$$Y_2^2 = \beta Y_1^2 + a^2 (\omega_0 + \gamma Y_1^2) \tag{B.23}$$

$$Y_3^2 = \beta Y_2^2 + a^2 (\omega_0 + \gamma Y_2^2) \tag{B.24}$$

$$Y_4^2 = \beta Y_3^2 + a^2 (\omega_0 + \gamma Y_3^2). \tag{B.25}$$

Solving equations (B.23)-(B.25) for β and equating the results, we obtain

$$\frac{a^2}{\omega_0} = \frac{Y_2^4 - Y_1^2 Y_3^2}{Y_2^2 - Y_1^2} = \frac{Y_3^4 - Y_2^2 Y_4^2}{Y_3^2 - Y_2^2}. \tag{B.26}$$

Therefore, a necessary condition such that $Y_t = v_t(\theta)$, $t = 1, \dots, T$ for some parameter $\theta \in \Theta$ is that (Y_1, \dots, Y_T) lies in the set $p^{-1}(0) = \{(Y_1, \dots, Y_T) \in \mathbb{R}^T | p(Y_1, \dots, Y_T) = 0\}$, where p is the polynomial function

$$p(Y_1, \dots, Y_T) = (Y_2^4 - Y_1^2 Y_3^2)(Y_3^2 - Y_2^2) - (Y_3^4 - Y_2^2 Y_4^2)(Y_2^2 - Y_1^2).$$

The set $p^{-1}(0)$ has Hausdorff dimension less than T . Therefore, as the distribution of (Y_1, \dots, Y_T) is assumed to be absolutely continuous from GARCH Assumption 1, we obtain the claim with $K = 4$.

Appendix B.2.8: Summary

We summarize the arguments showing that Assumption 4 and 5 of Chapter 3 are

satisfied under GARCH Assumptions 1–3.

Assumption 4: Part (A) holds as it has been shown in Section SA.2.4.1 that the uniform law of large number holds under our GARCH Assumptions. Part (B)(i)-(ii) are satisfied under GARCH Assumptions 1-2. Part (B)(iii) is easy to check. Concerning Part (B)(iv), we have shown in Section SA.2.3 that the GARCH model is identifiable when ω is normalized.

Assumption 5: Part (A)(i) is easy to check, (ii) is satisfied by GARCH Assumption 1. Part (B)(i) is satisfied by GARCH Assumption 1, (ii) is clearly weaker than GARCH Assumption 3. Part (C)(i) follows easily from $\sigma_t(\theta)^2 \geq \omega_0 > 0$ and the bounds on the parameter $|b|$. Part (C)(ii) has been shown in Lemma 9. Part (D) is implied by Lemma 11. Part (E) is discussed in Section SA.2.6. Part (F) is satisfied under GARCH Assumptions 2–3 as discussed in Section SA.2.2, and Part (G) is satisfied by GARCH Assumption 1 as discussed in Section SA.2.7.

Appendix C

Additional Tables for Chapter 3

Table C.1: Finite-sample performance of (Q)MLE

	$T = 2500$			$T = 5000$		
	ω	β	γ	ω	β	γ
Panel A: N(0,1) innovations						
True	0.050	0.950	0.050	0.050	0.950	0.050
Median	0.053	0.897	0.050	0.051	0.899	0.050
Avg bias	0.011	(0.011)	0.000	0.005	(0.005)	0.000
St dev	0.056	0.064	0.013	0.023	0.029	0.009
Coverage	0.936	0.930	0.928	0.936	0.933	0.937
Panel B: Skew t (5,-0.5) innovations						
True	0.050	0.950	0.050	0.050	0.950	0.050
Median	0.052	0.895	0.049	0.052	0.897	0.050
Avg bias	0.017	(0.023)	0.005	0.006	(0.008)	0.002
St dev	0.077	0.095	0.028	0.026	0.037	0.017
Coverage	0.899	0.907	0.897	0.913	0.907	0.903

Note: This table presents results from 1000 replications of the estimation of the parameters of a GARCH(1,1) model, using the Normal likelihood. In Panel A the innovations are standard Normal, and so estimation is then ML. In Panel B the innovations are standardized skew t , and so estimation is QML. Details are described in Section 3.4 of Chapter 3. The top row of each panel presents the true values of the parameters. The second, third, and fourth rows present the median estimated parameters, the average bias, and the standard deviation (across simulations) of the estimated parameters. The last row of each panel presents the coverage rates for 95% confidence intervals constructed using estimated standard errors.

Table C.2: Simulation results for Normal innovations, estimation by CAViaR

	$T = 2500$			$T = 5000$		
	β	γ	a_α	β	γ	a_α
$\alpha = 0.01$						
True	0.900	0.050	-2.326	0.900	0.050	-2.326
Median	0.901	0.048	-2.275	0.899	0.048	-2.347
Avg bias	-0.017	0.012	-0.120	-0.011	0.006	-0.095
St dev	0.079	0.066	0.957	0.051	0.034	0.718
Coverage	0.881	0.874	0.907	0.892	0.886	0.905
$\alpha = 0.025$						
True	0.900	0.050	-1.960	0.900	0.050	-1.960
Median	0.898	0.047	-1.953	0.896	0.047	-2.009
Avg bias	-0.018	0.005	-0.136	-0.012	0.002	-0.110
St dev	0.068	0.038	0.728	0.052	0.023	0.566
Coverage	0.906	0.879	0.934	0.913	0.892	0.918
$\alpha = 0.05$						
True	0.900	0.050	-1.645	0.900	0.050	-1.645
Median	0.901	0.047	-1.639	0.899	0.049	-1.667
Avg bias	-0.014	0.005	-0.085	-0.009	0.002	-0.070
St dev	0.068	0.037	0.597	0.045	0.023	0.436
Coverage	0.909	0.884	0.930	0.918	0.900	0.935
$\alpha = 0.10$						
True	0.900	0.050	-1.282	0.900	0.050	-1.282
Median	0.898	0.047	-1.291	0.898	0.048	-1.289
Avg bias	-0.016	0.006	-0.076	-0.010	0.003	-0.055
St dev	0.069	0.041	0.482	0.047	0.025	0.364
Coverage	0.916	0.883	0.933	0.921	0.896	0.937
$\alpha = 0.20$						
True	0.900	0.050	-0.842	0.900	0.050	-0.842
Median	0.898	0.048	-0.848	0.899	0.048	-0.840
Avg bias	-0.023	0.022	-0.058	-0.016	0.007	-0.049
St dev	0.091	0.107	0.391	0.063	0.044	0.304
Coverage	0.914	0.876	0.931	0.929	0.901	0.940

Note: This table presents results from 1000 replications of the estimation of VaR from a GARCH(1,1) DGP with standard Normal innovations. Details are described in Section 3.4 of Chapter 3. The top row of each panel presents the true values of the parameters. The second, third, and fourth rows present the median estimated parameters, the average bias, and the standard deviation (across simulations) of the estimated parameters. The last row of each panel presents the coverage rates for 95% confidence intervals constructed using estimated standard errors.

Table C.3: Simulation results for skew t innovations, estimation by CAViaR

	$T = 2500$			$T = 5000$		
	β	γ	a_α	β	γ	a_α
$\alpha = 0.01$						
True	0.900	0.050	-3.290	0.900	0.050	-3.290
Median	0.898	0.045	-3.272	0.899	0.045	-3.306
Avg bias	-0.041	0.022	-0.355	-0.027	0.008	-0.306
St dev	0.142	0.097	1.928	0.103	0.044	1.546
Coverage	0.771	0.805	0.827	0.785	0.808	0.823
$\alpha = 0.025$						
True	0.900	0.050	-2.408	0.900	0.050	-2.408
Median	0.899	0.047	-2.371	0.898	0.049	-2.414
Avg bias	-0.026	0.012	-0.190	-0.016	0.004	-0.144
St dev	0.103	0.067	1.135	0.070	0.033	0.862
Coverage	0.832	0.841	0.877	0.830	0.862	0.859
$\alpha = 0.05$						
True	0.900	0.050	-1.800	0.900	0.050	-1.800
Median	0.899	0.047	-1.780	0.899	0.049	-1.792
Avg bias	-0.023	0.008	-0.146	-0.013	0.004	-0.087
St dev	0.092	0.060	0.782	0.057	0.028	0.563
Coverage	0.863	0.861	0.892	0.883	0.871	0.890
$\alpha = 0.10$						
True	0.900	0.050	-1.223	0.900	0.050	-1.223
Median	0.900	0.049	-1.205	0.900	0.049	-1.217
Avg bias	-0.019	0.008	-0.074	-0.010	0.004	-0.043
St dev	0.080	0.050	0.495	0.050	0.027	0.356
Coverage	0.895	0.892	0.919	0.892	0.905	0.910
$\alpha = 0.20$						
True	0.900	0.050	-0.652	0.900	0.050	-0.652
Median	0.903	0.051	-0.619	0.902	0.051	-0.636
Avg bias	-0.027	0.026	-0.035	-0.016	0.009	-0.028
St dev	0.122	0.109	0.353	0.084	0.042	0.271
Coverage	0.867	0.887	0.897	0.890	0.889	0.916

Note: This table presents results from 1000 replications of the estimation of VaR from a GARCH(1,1) DGP with skew t innovations. Details are described in Section 3.4 of Chapter 3. The top row of each panel presents the true values of the parameters. The second, third, and fourth rows present the median estimated parameters, the average bias, and the standard deviation (across simulations) of the estimated parameters. The last row of each panel presents the coverage rates for 95% confidence intervals constructed using estimated standard errors.

Table C.4: Diebold-Mariano t-statistics on average out-of-sample loss differences for the DJIA, NIKKEI and FTSE100 ($\alpha = 0.05$)

	RW125	RW250	RW500	G-N	G-Skt	G-EDF	FZ-2F	FZ-1F	G-FZ	Hybrid
Panel A: DJIA										
RW125		-2.200	-3.536	2.324	2.860	2.935	3.006	3.821	3.244	3.494
RW250	2.200		-3.349	2.983	3.411	3.502	3.989	4.522	3.926	3.957
RW500	3.536	3.349		3.979	4.336	4.417	4.805	5.321	4.829	4.860
G-N	-2.324	-2.983	-3.979		3.573	2.787	0.791	1.419	1.472	1.670
G-Skt	-2.860	-3.411	-4.336	-3.573		1.385	-0.034	0.625	0.195	0.302
G-EDF	-2.935	-3.502	-4.417	-2.787	-1.385		-0.266	0.432	-0.119	-0.031
FZ-2F	-3.006	-3.989	-4.805	-0.791	0.034	0.266		1.085	0.192	0.247
FZ-1F	-3.821	-4.522	-5.321	-1.419	-0.625	-0.432	-1.085		-0.796	-0.613
G-FZ	-3.244	-3.926	-4.829	-1.472	-0.195	0.119	-0.192	0.796		0.126
Hybrid	-3.494	-3.957	-4.86	-1.670	-0.302	0.031	-0.247	0.613	-0.126	
Panel B: NIKKEI										
RW125		-0.225	-1.047	3.703	3.687	3.719	3.733	3.219	3.692	3.868
RW250	0.225		-1.162	4.048	4.058	4.098	3.897	3.582	4.076	4.249
RW500	1.047	1.162		3.733	3.748	3.785	3.768	3.387	3.773	3.847
G-N	-3.703	-4.048	-3.733		1.165	2.110	-1.841	-1.261	1.861	0.457
G-Skt	-3.687	-4.058	-3.748	-1.165		1.797	-1.888	-1.378	1.468	0.295
G-EDF	-3.719	-4.098	-3.785	-2.110	-1.797		-1.984	-1.522	-0.797	0.100
FZ-2F	-3.733	-3.897	-3.768	1.841	1.888	1.984		1.209	1.958	2.489
FZ-1F	-3.219	-3.582	-3.387	1.261	1.378	1.522	-1.209		1.487	2.624
G-FZ	-3.692	-4.076	-3.773	-1.861	-1.468	0.797	-1.958	-1.487		0.134
Hybrid	-3.868	-4.249	-3.847	-0.457	-0.295	-0.100	-2.489	-2.624	-0.134	

Note: Table continued on next page.

Table C.5: (cont'd) Diebold-Mariano t-statistics on average out-of-sample loss differences for the DJIA, NIKKEI and FTSE100 ($\alpha = 0.05$)

	RW125	RW250	RW500	G-N	G-Skt	G-EDF	FZ-2F	FZ-1F	G-FZ	Hybrid
Panel C: FTSE										
RW125		-2.329	-3.439	3.275	3.485	3.450	2.732	3.279	3.300	3.141
RW250	2.329		-2.751	4.146	4.337	4.305	3.663	4.264	4.160	4.025
RW500	3.439	2.751		4.682	4.845	4.817	4.232	4.848	4.696	4.661
G-N	-3.275	-4.146	-4.682		4.327	4.446	-0.210	-0.070	0.581	1.048
G-Skt	-3.485	-4.337	-4.845	-4.327		-3.853	-0.746	-0.877	-4.066	0.428
G-EDF	-3.450	-4.305	-4.817	-4.446	3.853		-0.648	-0.731	-3.949	0.545
FZ-2F	-2.732	-3.663	-4.232	0.210	0.746	0.648		0.213	0.249	1.401
FZ-1F	-3.279	-4.264	-4.848	0.070	0.877	0.731	-0.213		0.128	1.321
G-FZ	-3.300	-4.160	-4.696	-0.581	4.066	3.949	-0.249	-0.128		1.006
Hybrid	-3.141	-4.025	-4.661	-1.048	-0.428	-0.545	-1.401	-1.321	-1.006	

Note: This table presents t -statistics from Diebold-Mariano tests comparing the average losses, using the FZ0 loss function, over the out-of-sample period from January 2000 to December 2016, for ten different forecasting models. A positive value indicates that the row model has higher average loss than the column model. Values greater than 1.96 in absolute value indicate that the average loss difference is significantly different from zero at the 95% confidence level. Values along the main diagonal are all identically zero and are omitted for interpretability. The first three rows correspond to rolling window forecasts, the next three rows correspond to GARCH forecasts based on different models for the standardized residuals, and the last four rows correspond to models introduced in Section 3.2 of Chapter 3.

Table C.6: Out-of-sample average losses and goodness-of-fit tests ($\alpha=0.025$)

	<i>Average loss</i>				<i>GoF p-values: VaR</i>				<i>GoF p-values: ES</i>			
	S&P	DJIA	NIK	FTSE	S&P	DJIA	NIK	FTSE	S&P	DJIA	NIK	FTSE
RW-125	1.119	1.088	1.525	1.166	0.036	0.004	0.001	0.000	0.017	0.006	0.001	0.001
RW-250	1.164	1.117	1.525	1.209	0.009	0.009	0.006	0.000	0.037	<i>0.056</i>	0.015	0.006
RW-500	1.245	1.187	1.561	1.294	0.003	0.001	0.011	0.000	0.032	0.025	0.014	0.000
GCH-N	1.089	1.021	1.341	1.052	0.000	0.001	0.177	0.000	0.000	0.000	<i>0.053</i>	0.000
GCH-Skt	1.044	0.978	1.328	1.026	0.008	0.009	0.796	0.001	0.011	0.006	0.725	0.001
GCH-EDF	<i>1.028</i>	<i>0.969</i>	1.329	<i>1.042</i>	0.188	0.031	0.796	0.000	0.258	0.017	0.593	0.000
FZ-2F	1.039	0.998	1.421	1.242	0.000	0.002	0.341	0.000	0.001	0.001	0.158	0.000
FZ-1F	1.030	0.985	1.390	1.056	<i>0.057</i>	0.007	0.773	0.000	0.130	<i>0.058</i>	0.415	0.000
GCH-FZ	1.020	0.951	1.328	1.055	0.125	0.364	0.688	0.000	<i>0.222</i>	<i>0.403</i>	0.521	0.000
Hybrid	1.053	1.030	1.345	1.079	0.001	0.114	0.558	0.000	0.002	<i>0.075</i>	0.464	0.000

Note: The left panel of this table presents the average losses, using the FZ0 loss function, for four daily equity return series, over the out-of-sample period from January 2000 to December 2016, for ten different forecasting models. The lowest average loss in each column is highlighted in bold, the second-lowest is highlighted in italics. The first three rows correspond to rolling window forecasts, the next three rows correspond to GARCH forecasts based on different models for the standardized residuals, and the last four rows correspond to models introduced in Section 3.2 of Chapter 3. The middle and right panels of this table present p -values from goodness-of-fit tests of the VaR and ES forecasts respectively. Values that are greater than 0.10 (indicating no evidence against optimality at the 0.10 level) are in bold, and values between 0.05 and 0.10 are in italics.

Table C.7: Diebold-Mariano t-statistics on average out-of-sample loss differences for the S&P 500, DJIA, NIKKEI and FTSE100 (alpha=0.025)

	RW125	RW250	RW500	G-N	G-Skt	G-EDF	FZ-2F	FZ-1F	G-FZ	Hybrid
Panel A: S&P 500										
RW125		-1.836	-2.988	1.025	2.479	2.788	2.146	3.371	2.891	2.419
RW250	1.836		-2.815	1.725	2.747	3.004	2.602	3.712	3.135	2.992
RW500	2.988	2.815		2.823	3.673	3.893	3.630	4.624	4.023	4.045
G-N	-1.025	-1.725	-2.823		4.019	3.368	2.083	2.429	3.698	1.928
G-Skt	-2.479	-2.747	-3.673	-4.019		2.275	0.270	0.815	2.742	-0.594
G-EDF	-2.788	-3.004	-3.893	-3.368	-2.275		-0.592	-0.074	1.393	-1.483
FZ-2F	-2.146	-2.602	-3.630	-2.083	-0.270	0.592		0.487	1.227	-0.729
FZ-1F	-3.371	-3.712	-4.624	-2.429	-0.815	0.074	-0.487		0.579	-1.605
G-FZ	-2.891	-3.135	-4.023	-3.698	-2.742	-1.393	-1.227	-0.579		-2.172
Hybrid	-2.419	-2.992	-4.045	-1.928	0.594	1.483	0.729	1.605	2.172	
Panel B: DJIA										
RW125		-0.971	-2.294	1.892	2.981	3.051	3.132	3.590	3.332	1.840
RW250	0.971		-2.527	1.954	2.844	2.968	3.640	3.732	3.311	2.043
RW500	2.294	2.527		2.891	3.717	3.852	4.680	4.679	4.195	3.093
G-N	-1.892	-1.954	-2.891		3.705	2.900	0.765	1.305	3.236	-0.459
G-Skt	-2.981	-2.844	-3.717	-3.705		1.421	-0.706	-0.291	2.335	-2.666
G-EDF	-3.051	-2.968	-3.852	-2.900	-1.421		-1.022	-0.705	2.213	-2.693
FZ-2F	-3.132	-3.640	-4.680	-0.765	0.706	1.022		1.344	1.740	-1.229
FZ-1F	-3.590	-3.732	-4.679	-1.305	0.291	0.705	-1.344		1.539	-1.943
G-FZ	-3.332	-3.311	-4.195	-3.236	-2.335	-2.213	-1.740	-1.539		-3.127
Hybrid	-1.840	-2.043	-3.093	0.459	2.666	2.693	1.229	1.943	3.127	

Note: Table continued on the next page.

Table C.8: (cont'd): Diebold-Mariano t -statistics on average out-of-sample loss differences for the S&P 500, DJIA, NIKKEI and FTSE100 ($\alpha=0.025$)

	RW125	RW250	RW500	G-N	G-Skt	G-EDF	FZ-2F	FZ-1F	G-FZ	Hybrid
Panel C: NIKKEI										
RW125		0.010	-0.901	3.956	3.896	3.944	3.703	3.093	3.895	3.829
RW250	-0.010		-1.486	4.105	4.149	4.177	3.544	3.340	4.136	4.102
RW500	0.901	1.486		3.935	3.999	4.012	3.886	3.441	3.980	3.996
G-N	-3.956	-4.105	-3.935		1.799	2.032	-2.541	-2.010	2.052	-0.226
G-Skt	-3.896	-4.149	-3.999	-1.799		-0.785	-2.726	-2.532	-0.310	-0.977
G-EDF	-3.944	-4.177	-4.012	-2.032	0.785		-2.741	-2.499	0.459	-0.903
FZ-2F	-3.703	-3.544	-3.886	2.541	2.726	2.741		1.481	2.687	2.739
FZ-1F	-3.093	-3.34	-3.441	2.010	2.532	2.499	-1.481		2.454	2.971
G-FZ	-3.895	-4.136	-3.98	-2.052	0.310	-0.459	-2.687	-2.454		-0.919
Hybrid	-3.829	-4.102	-3.996	0.226	0.977	0.903	-2.739	-2.971	0.919	
Panel D: FTSE										
RW125		-1.557	-3.197	2.938	3.467	3.157	-1.683	2.978	2.570	2.173
RW250	1.557		-2.864	3.646	4.172	3.863	-0.758	3.788	3.355	2.985
RW500	3.197	2.864		4.350	4.789	4.532	1.179	4.688	4.173	3.972
G-N	-2.938	-3.646	-4.350		4.520	3.634	-3.549	-0.239	-0.340	-2.352
G-Skt	-3.467	-4.172	-4.789	-4.520		-4.471	-3.863	-1.996	-3.05	-3.991
G-EDF	-3.157	-3.863	-4.532	-3.634	4.471		-3.686	-0.949	-1.612	-3.218
FZ-2F	1.683	0.758	-1.179	3.549	3.863	3.686		3.924	3.468	3.271
FZ-1F	-2.978	-3.788	-4.688	0.239	1.996	0.949	-3.924		0.046	-1.602
G-FZ	-2.570	-3.355	-4.173	0.340	3.050	1.612	-3.468	-0.046		-2.354
Hybrid	-2.173	-2.985	-3.972	2.352	3.991	3.218	-3.271	1.602	2.354	

Note: This table presents t -statistics from Diebold-Mariano tests comparing the average losses, using the FZ0 loss function, over the out-of-sample period from January 2000 to December 2016, for ten different forecasting models. A positive value indicates that the row model has higher average loss than the column model. Values greater than 1.96 in absolute value indicate that the average loss difference is significantly different from zero at the 95% confidence level. Values along the main diagonal are all identically zero and are omitted for interpretability. The first three rows correspond to rolling window forecasts, the next three rows correspond to GARCH forecasts based on different models for the standardized residuals, and the last four rows correspond to models introduced in Section 3.2 of Chapter 3.

Bibliography

- AÏT-SAHALIA, Y. (2002): “Telling from Discrete Data Whether the Underlying Continuous-Time Model is a Diffusion,” *The Journal of Finance*, 57, 2075–2112.
- AÏT-SAHALIA, Y., AND J. JACOD (2009): “Testing for Jumps in a Discretely Observed Process,” *Annals of Statistics*, 37, 184–222.
- (2014): *High-Frequency Financial Econometrics*. Princeton University Press.
- ANDERSEN, T., T. BOLLERSLEV, F. DIEBOLD, AND P. LABYS (2001): “The Distribution of Realized Exchange Rate Volatility,” *Journal of the American Statistical Association*, 96, 42–55.
- ANDERSEN, T. G., AND T. BOLLERSLEV (1997): “Intraday Periodicity and Volatility Persistence in Financial Markets,” *Journal of Empirical Finance*, 4, 115–158.
- ANDERSEN, T. G., T. BOLLERSLEV, F. X. DIEBOLD, AND P. LABYS (2003): “Modeling and Forecasting Realized Volatility,” *Econometrica*, 71(2), pp. 579–625.
- ANDERSEN, T. G., T. BOLLERSLEV, F. X. DIEBOLD, AND C. VEGA (2003): “Micro Effects of Macro Announcements: Real-Time Price Discovery in Foreign Exchange,” *The American Economic Review*, 93(1), 251 – 277.
- ANDERSEN, T. G., T. BOLLERSLEV, F. X. DIEBOLD, AND G. WU (2006): “Realized Beta: Persistence and Predictability,” in , vol. 20 (Part 2) of *Advances in Econometrics: Econometric Analysis of Financial and Economic Time Series*, pp. 1–39. Emerald Group Publishing Limited.
- ANDERSEN, T. G., ET AL. (2006): “Volatility and Correlation Forecasting,” in *Handbook of Economic Forecasting*, vol. 1, pp. 777 – 878. Elsevier.
- ANDREWS, D. (1991): “Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation,” *Econometrica*, 59, 817–858.
- ANG, A., J. CHEN, AND Y. XING (2006): “Downside Risk,” *Review of Financial Studies*, 19(4), 1191 – 1239.
- ANG, A., R. J. HODRICK, Y. XING, AND X. ZHANG (2009): “High Idiosyncratic Volatility and Low Returns: International and Further U.S. Evidence.,” *Journal of Financial Economics*, 91(1), 1 – 23.
- ARTZNER, P., ET AL. (1999): “Coherent Measures of Risk,” *Mathematical Finance*, 9(3), 203–228.
- BANDI, F., AND R. RENÓ (2015): “Price and Volatility Co-jumps,” *Journal of Financial Economics*, *Forthcoming*.

- BARNDORFF-NIELSEN, O., S. GRAVERSEN, J. JACOD, M. PODOLSKIJ, AND N. SHEPHARD (2005): “A Central Limit Theorem for Realised Power and Bipower Variations of Continuous Semimartingales,” in *From Stochastic Analysis to Mathematical Finance, Festschrift for Albert Shiryaev*. Springer.
- BARNDORFF-NIELSEN, O., P. HANSEN, A. LUNDE, AND N. SHEPHARD (2009): “Realized kernels in practice: trades and quotes,” *The Econometrics Journal*, 12(3), C1–C32.
- BARNDORFF-NIELSEN, O., AND N. SHEPHARD (2001): “Non-Gaussian Ornstein–Uhlenbeck-Based Models and some of Their Uses in Financial Economics,” *Journal of the Royal Statistical Society, Series B*, 63, 167–241.
- (2002): “Econometric Analysis of Realized Volatility and its Use in Estimating Stochastic Volatility Models,” *Journal of the Royal Statistical Society, Series B*, 64, 253–280.
- (2003): “Realized Power Variation and Stochastic Volatility Models,” *Bernoulli*, 9, 243–265.
- (2007): “Variation, Jumps, Market Frictions and High Frequency Data in Financial Econometrics,” in *Advances in Economics and Econometrics. Theory and Applications, Ninth World Congress*, ed. by R. Blundell, T. Persson, and W. Newey. Cambridge University Press.
- BARNDORFF-NIELSEN, O., N. SHEPHARD, AND M. WINKEL (2006): “Limit Theorems for Multipower Variation in the Presence of Jumps in Financial Econometrics,” *Stochastic Processes and Their Applications*, 116, 796–806.
- BARNDORFF-NIELSEN, O. E., P. R. HANSEN, A. LUNDE, AND N. SHEPHARD (2008): “Designing Realized Kernels to Measure the ex post Variation of Equity Prices in the Presence of Noise,” *Econometrica*, 76, 1481–1536.
- BARNDORFF-NIELSEN, O. E., AND N. SHEPHARD (2004a): “Econometric Analysis of Realized Covariation: High Frequency Based Covariance, Regression, and Correlation in Financial Economics.,” *Econometrica*, 72(3), 885 – 925.
- (2004b): “Econometric Analysis of Realized Covariation: High Frequency Based Covariance, Regression, and Correlation in Financial Economics,” *Econometrica*, 72(3), pp. 885–925.
- BARNDORFF-NIELSEN, O. E., AND N. SHEPHARD (2004c): “Power and Bipower Variation with Stochastic Volatility and Jumps,” *Journal of Financial Econometrics*, 2, 1–37.

- BARNDORFF-NIELSEN, O. E., AND N. SHEPHARD (2006): “Econometrics of Testing for Jumps in Financial Economics Using Bipower Variation,” *Journal of Financial Econometrics*, 4, 1–30.
- BERKOWITZ, J., P. CHRISTOFFERSEN, AND D. PELLETIER (2011): “Evaluating value-at-risk models with desk-level data,” *Management Science*, 57(12), 2213–2227.
- BICKEL, P. J., C. A. J. KLAASSEN, Y. RITOV, AND J. A. WELLNER (1998): *Efficient and Adaptive Estimation for Semiparametric Models*. New York: Springer-Verlag.
- BIERENS, H. J. (1982): “Consistent model specification tests,” *Journal of Econometrics*, 20(1), 105–134.
- (1990): “A consistent conditional moment test of functional form,” *Econometrica*, pp. 1443–1458.
- BIERENS, H. J., AND W. PLOBERGER (1997): “Asymptotic theory of integrated conditional moment tests,” *Econometrica*, pp. 1129–1151.
- BILLINGSLEY, P. (1968): *Convergence of Probability Measures*. Wiley, New York.
- BINGHAM, N., C. GOLDIE, AND J. TEUGELS (1987): *Regular Variation*. Cambridge University Press.
- BLACK, F. (1976): “Studies of Stock Price Volatility Changes,” *Proceedings of the Business and Economics Section of the American Statistical Association*, pp. 177–181.
- BOLLERSLEV, T. (1990): “Modelling the Coherence in Short-run Nominal Exchange Rates: A Multivariate Generalized ARCH Model,” *Review of Economics and Statistics*, 72(3), 498 – 505.
- BORODIN, A., AND I. IBRAGIMOV (1994): *Limit Theorems for Functionals of Random Walks*. Tr. Mat. Inst. Steklova.
- CAI, Z., AND X. WANG (2008): “Nonparametric estimation of conditional VaR and expected shortfall,” *Journal of Econometrics*, 147(1), 120 – 130.
- CARR, P., AND L. WU (2003): “What Type of Process Underlies Options? A Simple Robust Test,” *The Journal of Finance*, 58, 2581–2610.
- CHAMBERLAIN, G. (1992): “Efficiency Bounds for Semiparametric Regression,” *Econometrica*, 60(3), pp. 567–596.

- CHANG, B. Y., P. CHRISTOFFERSEN, AND K. JACOBS (2013): “Market Skewness Risk and the Cross Section of Stock Returns,” *Journal of Financial Economics*, 107(1), 46 – 68.
- CHANG, B.-Y., P. CHRISTOFFERSEN, K. JACOBS, AND G. VAINBERG (2012): “Option-Implied Measures of Equity Risk,” *Review of Finance*, 16(2), 385 – 428.
- CHERNOZHUKOV, V., AND I. FERNÁNDEZ-VAL (2005): “Subsampling inference on quantile regression processes,” *Sankhyā: The Indian Journal of Statistics*, pp. 253–276.
- CHRISTOFFERSEN, P. F. (1998): “Evaluating interval forecasts,” *International economic review*, pp. 841–862.
- CLÉMENT, E., S. DELATTRE, AND A. GLOTER (2013): “An Infinite Dimensional Convolution Theorem with Applications to the Efficient Estimation of the Integrated Volatility,” *Stochastic Processes and their Applications*, 123, 2500–2521.
- COMTE, F., AND E. RENAULT (1996): “Long memory continuous time models,” *Journal of Econometrics*, 73, 101–149.
- CONRAD, J., R. F. DITTMAR, AND E. GHYSELS (2013): “Ex Ante Skewness and Expected Stock Returns,” *Journal of Finance*, 68(1), 85 – 124.
- CREAL, D., ET AL. (2013): “Generalized Autoregressive Score Models with Applications,” *Journal of Applied Econometrics*, 28(5), 777–795.
- DAVISON, A. C., AND D. V. HINKLEY (1997): *Bootstrap Methods and their Application*, Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.
- DE JONG, R. M. (1996): “The Bierens test under data dependence,” *Journal of Econometrics*, 72(1), 1–32.
- DELGADO, M. A., AND J. C. ESCANCIANO (2007): “Nonparametric tests for conditional symmetry in dynamic models,” *Journal of Econometrics*, 141(2), 652–682.
- DIEBOLD, F. X., AND R. S. MARIANO (1995): “Comparing Predictive Accuracy,” *Journal of Business & Economic Statistics*, 13(3), 253–263.
- DIEBOLD, F. X., AND R. S. MARIANO (2002): “Comparing predictive accuracy,” *Journal of Business & economic statistics*, 20(1), 134–144.
- DITTMAR, R. F. (2002): “Nonlinear Pricing Kernels, Kurtosis Preference, and Evidence from the Cross Section of Equity Returns,” *Journal of Finance*, 57(1), 369 – 403.

- DOVONON, P., S. GONÇALVES, AND N. MEDDAHI (2013): “Bootstrapping Realized Multivariate Volatility Measures,” *Journal of Econometrics*, 172(1), 49 – 65.
- DOVONON, P., U. HOUNYO, S. GONÇALVES, AND N. MEDDAHI (2014): “Bootstrapping High Frequency Jump Tests,” Discussion paper, Toulouse School of Economics.
- DU, Z., AND J. C. ESCANCIANO (2016): “Backtesting expected shortfall: accounting for tail risk,” *Management Science*, 63(4), 940–958.
- DUFFIE, D., D. FILIPOVIĆ, AND W. SCHACHERMAYER (2003): “Affine Processes and Applications in Finance,” *Annals of Applied Probability*, 13(3), 984–1053.
- DUFFIE, D., J. PAN, AND K. SINGLETON (2000): “Transform Analysis and Asset Pricing for Affine Jump-Diffusions,” *Econometrica*, 68, 1343–1376.
- EISENBAUM, N., AND H. KASPI (2007): “On the Continuity of Local Times of Borel Right Markov Processes,” *Annals of Probability*, 35, 915–934.
- ENGLE, R., AND B. KELLY (2012): “Dynamic Equicorrelation,” *Journal of Business and Economic Statistics*, 30, 212–228.
- ENGLE, R. F., AND S. MANGANELLI (2004): “CAViaR: Conditional autoregressive value at risk by regression quantiles,” *Journal of Business & Economic Statistics*, 22(4), 367–381.
- EPPS, T. W. (1979): “Comovements in Stock Prices in the Very Short Run.,” *Journal of the American Statistical Association*, 74(366), 291 – 298.
- ERAKER, B., M. S. JOHANNES, AND N. POLSON (2003): “The Impact of Jumps in Equity Index Volatility and Returns,” *The Journal of Finance*, 58, 1269–1300.
- ESCANCIANO, J. C., AND J. OLMO (2010): “Backtesting parametric value-at-risk with estimation risk,” *Journal of Business & Economic Statistics*, 28(1), 36–51.
- ESCANCIANO, J. C., AND C. VELASCO (2010): “Specification tests of parametric dynamic conditional quantiles,” *Journal of Econometrics*, 159(1), 209–221.
- FELLER, W. (1971): *An Introduction to Probability Theory and Its Applications*, Volume II. John Wiley.
- FERSON, W. E., AND C. R. HARVEY (1999): “Conditioning Variables and the Cross Section of Stock Returns.,” *Journal of Finance*, 54(4), 1325 – 1360.
- FERSON, W. E., S. KANDEL, AND R. F. STAMBAUGH (1987): “Tests of Asset Pricing with Time-Varying Expected Risk Premiums and Market Betas.,” *Journal of Finance*, 42(2), 201 – 220.

- FISSLER, T., AND J. F. ZIEGEL (2016): “Higher order elicibility and Osband’s principle,” *The Annals of Statistics*, 44(4), 1680–1707.
- FLORESCU, I., AND C. G. PASARICA (2009): “A Study About The Existence Of Leverage Effect In Stochastic Volatility Models,” *Physica A*, 388, 419–432.
- FORBES, K. J., AND R. RIGOBON (2002): “No Contagion, Only Interdependence: Measuring Stock Market Comovements,” *The Journal of Finance*, 57, 2223–2261.
- GEMAN, D., AND J. HOROWITZ (1980): “Occupation Densities,” *The Annals of Probability*, 8, 1–67.
- GNEITING, T. (2011): “Making and evaluating point forecasts,” *Journal of the American Statistical Association*, 106(494), 746–762.
- GOBBI, F., AND C. MANCINI (2012): “Identifying the Brownian Covariation from the Co-Jumps given Discrete Observations,” *Econometric Theory*, 28, 249–273.
- GOBET, E. (2001): “Local Asymptotic Mixed Normality Property for Elliptic Diffusion: A Malliavin Calculus Approach,” *Bernoulli*, 7(6), pp. 899–912.
- GONÇALVES, S., AND N. MEDDAHI (2009): “Bootstrapping Realized Volatility,” *Econometrica*, 77, 283–306.
- GRAHAM, B. S., AND J. L. POWELL (2012): “Identification and Estimation of Average Partial Effects in “Irregular” Correlated Random Coefficient Panel Data Models,” *Econometrica*, 80(5), pp. 2105–2152.
- GSCHÖPF, P., W. HÄRDLE, AND A. MIHOCI (2015): “Tail Event Risk Expectile based Shortfall,” Sfb 649 discussion papers, Humboldt University, Collaborative Research Center 649.
- GUIDOLIN, M., AND A. TIMMERMANN (2008): “International Asset Allocation under Regime Switching, Skew, and Kurtosis Preferences,” *Review of Financial Studies*, 21(2), 889 – 935.
- HANSEN, L. P., AND S. F. RICHARD (1987): “The Role of Conditioning Information in Deducing Testable,” *Econometrica*, 55(3), 587 – 613.
- HANSEN, P. R., AND A. LUNDE (2006): “Realized Variance and Market Microstructure Noise,” *Journal of Business and Economic Statistics*, 24, 127–161.
- HARDY, G., J. LITTLEWOOD, AND G. POLYA (1952): *Inequalities*. Cambridge University Press.
- HARVEY, A. (2013): *Dynamic Models for Volatility and Heavy Tails*. Cambridge University Press.

- HARVEY, C. R., AND A. SIDDIQUE (2000): “Conditional Skewness in Asset Pricing Tests,” *Journal of Finance*, 55(3), 1263 – 1295.
- HASBROUCK, J. (2003): “Intraday Price Formation in U.S. Equity Index Markets,” *The Journal of Finance*, 58(6), 2375–2399.
- HASBROUCK, J. (2015): “Securities Trading: Procedures and Principles,” Discussion paper, New York University.
- HESTON, S. (1993): “A closed-form solution for options with stochastic volatility with applications to bonds and currency options,” *Review of Financial Studies*, 6, 327–343.
- HOROWITZ, J. L. (2001): “The Bootstrap,” in *Handbook of Econometrics*, vol. 5. Elsevier.
- HOUNYO, U. (2013): “Bootstrapping Realized Volatility and Realized Beta under a Local Gaussianity Assumption,” Discussion paper, University of Oxford.
- HUANG, X., AND G. TAUCHEN (2005): “The Relative Contributions of Jumps to Total Variance,” *Journal of Financial Econometrics*, 3, 456–499.
- IBRAGIMOV, I., AND R. HAS’MINSKII (1981): *Statistical Estimation: Asymptotic Theory*. Springer, Berlin.
- JACOD, J. (2008): “Asymptotic Properties of Power Variations and Associated Functionals of Semimartingales,” *Stochastic Processes and their Applications*, 118, 517–559.
- JACOD, J., Y. LI, P. A. MYKLAND, M. PODOLSKIJ, AND M. VETTER (2009): “Microstructure Noise in the Continuous Case: The Pre-Averaging Approach,” *Stochastic Processes and Their Applications*, 119, 2249–2276.
- JACOD, J., AND M. PODOLSKIJ (2013): “A Test for the Rank of the Volatility Process: the random Perturbation Approach,” *Annals of Statistics*, forthcoming.
- JACOD, J., AND P. PROTTER (2012): *Discretization of Processes*. Springer.
- JACOD, J., AND M. REISS (2012): “A Remark on the Rates of Convergence for Integrated Volatility Estimation in the Presence of Jumps,” Discussion paper, arXiv: 1209.4173v1.
- JACOD, J., AND M. ROSENBAUM (2013): “Quarticity and Other Functionals of Volatility: Efficient Estimation,” *Annals of Statistics*, 41, 1462–1484.
- JACOD, J., AND V. TODOROV (2009): “Testing for common arrivals of jumps for discretely observed multidimensional processes,” *Annals of Statistics*, 37, 1792–1838.

- (2010): “Do Price and Volatility Jump Together?,” *Annals of Applied Probability*, 20, 1425–1469.
- JEGANATHAN, P. (1982): “On the Asymptotic Theory of Estimation when the Limit of the Log-likelihood is Mixed Normal,” *Sankhya, Ser. A*, 44, 173–212.
- (1983): “Some Asymptotic Properties of Risk Functions When the Limit of the Experiment Is Mixed Normal,” *Sankhya: The Indian Journal of Statistics, Series A (1961-2002)*, 45(1), pp. 66–87.
- JIANG, G. J., AND R. C. OOMEN (2008): “Testing for Jumps when Asset Prices are Observed with Noise - A “Swap Variance” Approach,” *Journal of Econometrics*, 144, 352–370.
- KALNINA, I. (2013): “Nonparametric Tests of Time Variation in Betas,” Discussion paper, University of Montreal.
- KERKHOF, J., AND B. MELENBERG (2004): “Backtesting for risk-based regulatory capital,” *Journal of Banking & Finance*, 28(8), 1845–1865.
- KING, M., E. SENTANA, AND S. WADHWANI (1994): “Volatility and Links between National Stock Markets,” *Econometrica*, 62, 901–933.
- KOMUNJER, I. (2013): “Quantile Prediction,” in *Handbook of Economic Forecasting*, vol. 2, pp. 961–994. Elsevier.
- KUPIEC, P. H. (1995): “Techniques for verifying the accuracy of risk measurement models,” *The Journal of Derivatives*, 3(2), 73–84.
- KYLE, A. S. (1985): “Continuous Auctions and Insider Trading,” *Econometrica*, 53(6), pp. 1315–1335.
- LEE, S., AND P. MYKLAND (2008): “Jumps in Financial Markets: A New Nonparametric Test and Jump Dynamics,” *Review of Financial Studies*, 21(6), 2535–2563.
- LEE, T.-H., H. WHITE, AND C. W. GRANGER (1993): “Testing for neglected nonlinearity in time series models: A comparison of neural network methods and alternative tests,” *Journal of Econometrics*, 56(3), 269–290.
- LEHMANN, E. L., AND J. P. ROMANO (2005): *Testing Statistical Hypothesis*. Springer.
- LETTAU, M., M. MAGGIORI, AND M. WEBER (2014): “Conditional Risk Premia in Currency Markets and Other Asset Classes.” Forthcoming in the *Journal of Financial Economics*.
- LI, J., V. TODOROV, AND G. TAUCHEN (2013): “Volatility Occupation Times,” *Annals of Statistics*, 41, 1865–1891.

- (2014): “Adaptive Estimation of Continuous-Time Regression Models using High-Frequency Data,” Discussion paper, Duke University.
- (2016): “Jump Regressions,” *Econometrica*, *Forthcoming*.
- LI, J., AND D. XIU (2013): “Spot Variance Regressions,” Discussion paper, Duke University and University of Chicago.
- MANCINI, C. (2001): “Disentangling the Jumps of the Diffusion in a Geometric Jumping Brownian Motion,” *Giornale dell’Istituto Italiano degli Attuari*, LXIV, 19–47.
- MARCUS, M. B., AND J. ROSEN (2006): *Markov Processes, Gaussian Processes, and Local Times*. Cambridge University Press.
- MCNEIL, A. J., AND R. FREY (2000): “Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach,” *Journal of empirical finance*, 7(3), 271–300.
- MERTON, R. C. (1973): “An Intertemporal Capital Asset Pricing Model,” *Econometrica*, 41(5), 867 – 887.
- (1975): “Theory of Finance from the Perspective of Continuous Time,” *Journal of Financial and Quantitative Analysis*, 10(4), 659 – 674.
- (1976): “Option Pricing When Underlying Stock Returns Are Discontinuous,” *Journal of Financial Economics*, 3(1/2), 125 – 144.
- MYKLAND, P., AND L. ZHANG (2006): “ANOVA for Diffusions and Ito Processes,” *Annals of Statistics*, 34, 1931–1963.
- (2009): “Inference for Continuous Semimartingales Observed at High Frequency,” *Econometrica*, 77, 1403–1445.
- NELSON, D. B. (1990): “ARCH Models as Diffusion Approximations,” *Journal of Econometrics*, 45, 7–38.
- (1991): “Conditional Heteroskedasticity in Asset Returns: A New Approach,” *Econometrica*, 59(2), 347 – 370.
- NEWBY, W. K., AND D. MCFADDEN (1994): “Large sample estimation and hypothesis testing,” in *Handbook of Econometrics*, vol. 4, pp. 2111 – 2245. Elsevier.
- NEWBY, W. K., AND J. L. POWELL (1987): “Asymmetric Least Squares Estimation and Testing,” *Econometrica*, 55(4), 819–847.
- NOLDE, N., AND J. F. ZIEGEL (2016): “Elicitability and backtesting: Perspectives for banking regulation,” *The Annals of Applied Statistics*, 11, 1833–1874.

- PATTON, A. (2013): *Copula Methods for Forecasting Multivariate Time Series* vol. 2 of *Handbook of Economic Forecasting*, pp. 899 – 960. Elsevier.
- PATTON, A. J., AND K. SHEPPARD (2009): *Evaluating Volatility and Correlation Forecasts* pp. 801–838. Springer Berlin Heidelberg.
- PATTON, A. J., J. F. ZIEGEL, AND R. CHEN (2019): “Dynamic semiparametric models for expected shortfall (and Value-at-Risk),” *Journal of Econometrics*, 211(2), 388 – 413.
- POLITIS, D., J. ROMANO, AND M. WOLF (1999): “Subsampling Springer-Verla,” *New York*.
- PROTTER, P. (2004): *Stochastic Integration and Differential Equations*. Springer-Verlag, Berlin, 2nd edn.
- REISS, M. (2011): “Asymptotic Equivalence for Inference on the Volatility from Noisy Observations,” *The Annals of Statistics*, 39(2), 772–802.
- REISS, M., V. TODOROV, AND G. TAUCHEN (2015): “Nonparametric Test for a Constant Beta between Ito Semimartingales based on High-Frequency Data,” *Stochastic Processes and their Applications*, 125, 2955–2988.
- RENAULT, E., C. SARISOY, AND B. J. WERKER (2014): “Efficient Estimation of Integrated Volatility and Related Processes,” Discussion paper, Brown University.
- REVUZ, D., AND M. YOR (1999): *Continuous Martingales and Brownian Motion*. Springer-Verlag, Berlin, Germany, third edn.
- ROBIN, J.-M., AND R. J. SMITH (2000): “Tests of Rank,” *Econometric Theory*, 16, 151–175.
- ROLL, R. (1987): “ R^2 ,” *Journal of Finance*, 43, 541–566.
- SAKOV, A., AND P. J. BICKEL (2000): “An Edgeworth expansion for the m out of n bootstrapped median,” *Statistics & probability letters*, 49(3), 217–223.
- SATO, K. (1999): *Lévy Processes and Infinitely Divisible Distributions*. Cambridge University Press.
- SINGLETON, K. (2006): *Empirical Dynamic Asset Pricing*. Princeton University Press.
- STEIN, C. (1956): “Efficient Nonparametric Testing and Estimation,” in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pp. 187–195. University of California Press, Berkeley, Calif.

- STOCK, J. H. (1994): “Unit Roots, Structural Breaks and Trends,” in *Handbook of Econometrics, volume 4*, ed. by R. Engle, and D. McFadden, pp. 2739–2841. Elsevier, Amsterdam.
- STOCK, J. H., AND M. W. WATSON (2002): “Macroeconomic Forecasting Using Diffusion Indexes,” *Journal of Business & Economic Statistics*, 20(2), pp. 147–162.
- TAYLOR, J. W. (2008): “Estimating Value at Risk and Expected Shortfall Using Expectiles,” *Journal of Financial Econometrics*, 6(2), 231–252.
- (2019): “Forecasting Value at Risk and Expected Shortfall Using a Semi-parametric Approach Based on the Asymmetric Laplace Distribution,” *Journal of Business & Economic Statistics*, 37(1), 121–133.
- TODOROV, V. (2009): “Estimation of Continuous-time Stochastic Volatility Models with Jumps using High-Frequency Data,” *Journal of Econometrics*, 148, 131–148.
- TODOROV, V., AND T. BOLLERSLEV (2010): “Jumps and Betas: A New Framework for Disentangling and Estimating Systematic Risks,” *Journal of Econometrics*, 157, 220–235.
- TODOROV, V., AND G. TAUCHEN (2011): “Volatility Jumps,” *Journal of Business and Economic Statistics*, 29, 356–371.
- TODOROV, V., AND G. TAUCHEN (2012): “The Realized Laplace Transform of Volatility,” *Econometrica*, 80, 1105–1127.
- TSYBAKOV, A. B. (2009): *Introduction to Nonparametric Estimation*. Springer.
- VAN DER VAART, A., AND J. WELLNER (1996): *Weak Convergence and Empirical Processes*. Springer-Verlag.
- VAN DER VAART, A. W. (1998): *Asymptotic statistics*, vol. 3. Cambridge university press.
- WEISS, A. A. (1991): “Estimating Nonlinear Dynamic Models Using Least Absolute Error Estimation,” *Econometric Theory*, 7(1), 46–68.
- WHITE, H. (1989): “An additional hidden unit test for neglected nonlinearity in multilayer feedforward networks,” in *Proceedings of the international joint conference on neural networks*, vol. 2, pp. 451–455.
- ZHANG, L., P. A. MYKLAND, AND Y. AÏT-SAHALIA (2005): “A Tale of Two Time Scales: Determining Integrated Volatility with Noisy High-Frequency Data,” *Journal of the American Statistical Association*, 100, 1394–1411.
- ZHU, D., AND J. W. GALBRAITH (2011): “Modeling and forecasting expected shortfall with the generalized asymmetric Student-t and asymmetric exponential power distributions,” *Journal of Empirical Finance*, 18(4), 765 – 778.

Biography

Rui Chen is a PhD candidate in the economics department at Duke University. She specializes in econometrics, with a focus on time series forecasting. She is also interested in various topics of both theoretical and empirical econometrics.