Econometric Forecasting
and
Textual Analysis in Finance

Ye Zeng

January, 2019

A dissertation submitted to
School of Business and Social Sciences, Aarhus University,
in partial fulfilment of the requirements of
the PhD degree in Economics and Business Economics
Acknowledgment

The thesis summarizes part of my three-year study at CREATES, Aarhus University. The research project leading to these results received funding from the European Union Research and Innovation programme Horizon 2020 under the Marie Skłodowska-Curie grant agreement No. 675044 (BigDataFinance). I would like to thank the European Union for financial funding and CREATES for providing excellent research environment. It must be stressed that all errors in the thesis are solely mine, and they should not be attached to the funding agency or the host institute in any way.

I also feel obliged to acknowledge many tender-hearted people, who, despite my lack of gregariousness, always reach out when I am in need. I would like to thank Professor Asger Lunde and Professor Kim Christensen, who have maintained professionalism throughout their supervision and shown great tolerance towards my caprices. I am grateful to Ms Solveig N. Sørensen, who with her constant geniality and proficiency helps me set aside non-academic problems and concentrate on my study. I express my sincere gratitude to Professor Shin Kanaya, who offered many encouraging advices when I was matriculated as Ph.D. I have many thanks to Professor Leopoldo Catania, who never grudges his time when I stop by with naïve questions, and have been a congenial and inspirational friend. I also thank Alice Buccioli – who will also be awarded doctorate soon – for always bringing an air of affability to the office. My special appreciation goes to the benign and hospitable Steenbach couple, who has been great example of urbanity and rendered a cozy ambiance to our home in Aarhus. Finally, I am deeply indebted to my wife, for her constant understanding of all my whimsical thoughts and her tremendous efforts to ease my work. Needless to say, I would not have grown to be of basic intellectual quality without the precious help offered by these people.

Ye Zeng
January 2019
Updates

During the pre-defense meeting on March 28th, 2019, Professor Michael Wolf (University of Zurich), Professor Ingmar Nolte (Lancaster University) and Professor Eric Hillebrand (Aarhus University) astutely made many crucial corrections. They also offered valuable suggestions that could greatly improve the thesis. I would like to express my sincere gratitude for their being inspirational as well as encouraging. I have tried to make revisions that the committee deems necessary. Some other helpful suggestions are left for future study. Finally, I must stress that all remaining errors are solely mine.

Ye Zeng
April 2019
Contents

Summary 1

Resumé III

1 Data-Driven Sample Split and Out-of-Sample Tests 1
1.1 Introduction .............................................. 2
1.2 The General Setting and The Main Idea .................. 6
1.3 Comparison of nonnested forecast models ............... 15
1.4 A modified sup-type test ................................ 20
1.5 Power under local alternatives .......................... 23
1.6 Monte Carlo Experiments ............................... 32
1.7 Discussions on other novel out-of-sample tests ......... 40
1.8 Discussions on test for nested models and comparison of forecast methods 42
1.9 Conclusions .............................................. 45
Appendix A Estimation of nuisance parameters .............. 47
Appendix B Proofs ............................................. 49
References .................................................... 81

2 Forecasting Daytime US Market Volatility with Overnight Information 83
2.1 Introduction .............................................. 84
2.2 Forecasting Daytime Volatility of US Stocks ............ 86
2.3 Dissecting the forecast .................................. 110
2.4 Conclusion ................................................. 115
Appendix A Volatility Measures ............................. 118
Appendix B HAR Models Using Both SPY and Emini Futures Data 122
Appendix C The relation Between Volatilities of SPY and ES ... 124
References .................................................... 131

3 How Does Diversity of Media Coverage Influence Firms' Risk Exposure? 133
3.1 Introduction .............................................. 134
3.2 Review of Related Literature ............................. 136
3.3 Data, Textual Analysis Techniques and Econometrics Method 138
3.4 Main Empirical Findings ................................ 147
3.5 Robustness of the Findings and Further Analysis ...... 155
3.6 Conclusion ................................................. 161
Appendix A Data Pre-processing and Software Usage ....... 163
Appendix B Review of Word2Vec ............................ 166
Appendix C Estimation of the dynamic panel ................ 172
Appendix D Examples and Additional Details ................ 174
Appendix E Firm by Firm Regression Results ................ 180
References .................................................... 191
The thesis has three self-contained chapters. The first two chapters are on econometric forecasting, while the last chapter may be subsumed to textual analysis in finance.

Chapter 1 conducts theoretical analysis of data-driven sample split and the resulting out-of-sample tests. Notwithstanding the growing concern about p-hacking among scientists, the econometrics community has paid little heed in pseudo out-of-sample analysis to the delicate split point, a potential leverage of the test statistic. The first half of this chapter calls attention to the problem of data-driven sample split. I build on Rossi and Inoue (2012) and Hansen and Timmermann (2012) and extend their analysis by formalizing various modes of sample split that might be common in practice, and show how they distort size of the test if conventional critical values are used. Then I adapt the approach of Rossi and Inoue (2012) and illustrate how to develop ad hoc treatment for data-driven split. The second half of the chapter turns to another question inspired by Rossi and Inoue (2012). The combination of a data-driven split mode and its treatment makes a novel out-of-sample test, and I seek to know if the resulting test can be more powerful than conventional test using fixed split point. To this end, I develop a sup-type test with the following features: (1) It does not require Monte Carlo approximation. (2) It involves a test statistic that is not asymptotically pivotal, but I develop feasible procedures to make the test asymptotically sized at the nominal level. (3) The limit distribution of the test statistic is a driftless (resp. drifted) time-changed Brownian motion under the null (resp. local alternatives). The last feature enables me to invoke known formulae for boundary crossing probability to calculate power of the sup-type test. The finding is two-fold: the sup-type test is more powerful than the conventional test in the presence of structural break, but not so in the absence. As an eclectic suggestion, the sup-type test should be used only if the researcher holds the a priori belief in structural break of predictive accuracy.

Chapter 2 analyzes predictive ability of volatility forecasting models that feature overnight information. We find that HAR (Corsi (2009)), HAR-CJ (Andersen et al. (2007)), SHAR (Patton and Sheppard (2015)) and HARQ (Bollerslev et al. (2016)) models augmented with volatility of the overnight US futures market produce considerably more accurate forecast. In out-of-sample analysis, superior predictive ability of the augmented models is confirmed by reality check. The last section delves into the overnight information and relates the improvement in volatility forecasting to the interaction between the overnight US market and the daytime European market; it also briefly discusses if scheduled FOMC meeting has bearing on the forecast.

Chapter 3 studies how diversity of media coverage affects firms’ exposure to systematic risk. The content diversity of a pool of news stories is defined as their average textual dissim-
ilarity. To capture semantic as well as syntactic relations, we use Google's word2vec embeddings (Mikolov et al. (2013); Mikolov et al. (2013)) and gauge the textual dissimilarity by Word Mover’s Distance (Kusner et al. (2015)). The dynamic panel featuring multifactor error structure (Chudik and Pesaran (2015)) is employed in investigating the effect of news diversity on risk exposure. The main finding is that diversity of media coverage predicts higher systematic risk on the next trading day, after we control salience, sentiment, social media attention and multiple stock market variables. Robustness checks under various model specifications and with respect to different word embeddings as well as textual dissimilarity measures are also provided. Finally, we interpret the finding by linking diversity of media coverage to investors’ difference of opinion.

References


Afhandlingen består af tre selvstændige kapitler. De to første kapitler handler om økonometrisk prognose, mens det sidste kapitel behandler anvendelser af tekstanalyse i finansiering.


Kapitel 3 undersøger, hvordan mangfoldighed i mediedækning påvirker virksomhedernes eksponering for systematisk risiko. Indholdsmangfoldigheden af en mængde nyhedshistorier er defineret som deres gennemsnitlige tekstmæssige ulighed. For at indfange semantiske såvel som syntaktiske relationer bruger vi Googles tekst2vec (Mikolov et al. (2013); Mikolov et al. (2013)) og måler den tekstmæssige ulighed af Word Mover's Distance (Kusner et al. (2015)). Det dynamiske panel med multifaktor-fejlstruktur (Chudik and Pesaran (2015)) bliver brugt til at undersøge effekten af nyhedsdiversiten på risikoeksponering. Vi ser, at mangfoldigheden af mediedækning forudsiger højere systematisk risiko på den næste handelsdag efter, at vi har kontrolleret salience, sentiment, opmærksomhed på sociale medier og flere aktiemarkedsvariabler. Robusthedskontrol under forskellige modelspecifikationer og med hensyn til forskellige ordindlejringer samt tekstmæssige ulighedstiltag (dis-similarity measures) gives også. Endelig fortolker vi ved at forbinde mangfoldigheden af mediedækning med investorerernes meningsforskel (investors’ difference of opinion).

References


CHAPTER 1

Data-Driven Sample Split and Out-of-Sample Tests

Ye Zeng
Aarhus University and CREATES

Abstract

This paper scrutinizes the cause and effect between the behavioral mode of sample split and limit distribution of the out-of-sample test statistic. We address two problems concerning size and power respectively. In terms of size, we provide concrete examples for Diebold (2015)’s comment that pseudo out-of-sample forecast comparison expands the scope of strategic $p$-hacking. We delineate different modes of sample split and show how they distort size of the test if conventional critical values are used. We then build on Rossi and Inoue (2012) and illustrate how ad hoc treatments for the sample split modes can be developed. The treatments are ad hoc, as different modes of sample split can result in the same degree of size distortion but require very different critical values to be sized at the nominal level. We show that achieving robustness to all modes of sample split comes at the cost of considerable loss of power. As of power, we ask if it is worthwhile to combine data-driven sample split and its treatment to develop novel out-of-sample test. Specifically, we modify the sup-type test proposed by Rossi and Inoue (2012) and analyze its asymptotic power under local alternatives. We find that the sup-type test is no more powerful than conventional test using fixed and small split point under the local alternative with Pitman drift, but becomes more powerful in the presence of structural break.

Keywords: Forecast Comparison; Diebold-Mariano Test; $p$-Hacking; Boundary Crossing Probability.

JEL classification: C12, C53, G17
1.1 Introduction

The utility of pseudo out-of-sample test in comparison of forecast models was recently questioned by Diebold (2015). As a matter of fact, pseudo out-of-sample test has many inherent subtleties, and the literature does not share unanimous view on them. On the one hand, pseudo out-of-sample test tends to have low power in finite sample (Inoue and Kilian (2005)); it is sub-optimal compared to full-sample model comparison methods as it discards data (Diebold (2015)). On the other hand, out-of-sample test might be less sensitive to mining over model specifications (Hansen and Timmerman (2015)); it might also be an effective tool for flagging structural change (Giacomini and Rossi (2010); Giacomini and Rossi (2016)), as noted by Diebold (2015). Apart from the technical reasons, it seems more natural to conduct pseudo out-of-sample analysis to simulate “real-time” prediction than to conduct full-sample regression if one’s interest consists solely in comparing historical predictive performances. It is therefore important to have a sound understanding of how the many subtleties affect pseudo out-of-sample test. Such understanding will help us better appreciate the merits and demerits of the pseudo out-of-sample approach, and it will be crucial for making decision on use or abstention from use of this approach.

This paper focuses on the delicate sample split point, which belongs to one of the subtleties affecting pseudo out-of-sample test. Specifically, we study the effects of different behavioral modes of sample split on pseudo out-of-sample forecast comparison and their due treatments. This problem would not have emerged if there were an unanimous rule of sample split to which every researcher adheres. Conventional tests, such as the ones proposed by West (1996), McCracken (2000), Clark and McCracken (2001), McCracken (2007), are developed under the assumption that the sample split point is a fixed fraction of the total sample size\(^1\). But just as other choices in statistical analysis, in practice researcher’s choice of sample split may be contingent on the data, sometimes without being realized by himself or herself (Gelman and Loken (2013))\(^2\). Data-driven sample split may invalidate conventional out-of-sample tests; this issue is made clear by Rossi and Inoue (2012) and Hansen and Timmermann (2012)\(^3\), who exemplify gross size distortions resulting from mining over sample split point in attempt to maximize the test statistic.

\(^1\)More precisely, if \(R_T\) denotes the sample split point and \(T\) the total sample size, they require \(R_T\) to be nonrandom and \(R_T/T\) converge to a constant.

\(^2\)Munafò et al. (2017) provide a brief summary of both psychological motives, such as confirmation bias (i.e., the tendency to focus on evidence that is in consonance with predisposed expectations and beliefs) and hindsight bias (i.e., one’s propensity for overstating his or her predictive ability after observing an event), and social institutions (e.g., publication bias), that could induce one to conduct data-mining, either knowingly or not. For confirmation bias, see Nickerson (1998), who calls it an “ubiquitous phenomenon”.

\(^3\)To the best of our knowledge, Rossi and Inoue (2012) and Hansen and Timmermann (2012) are the only ones among the literature of econometrics that are dedicated to investigating the effect of sample split. In general, the problem of data-mining over sample split point may be subsumed to the broader “p-hacking” problem (Simonsohn et al. (2014)), which has aroused many concerns among scientists. See, e.g., Munafò et al. (2017).
Building on Rossi and Inoue (2012) and Hansen and Timmermann (2012), we aim to provide deeper insights into the effect of sample split. We extend their analysis by delineating different modes of sample split and their effects. In particular, we distinguish between data-independent and data-driven sample split. “Data-independent” means that the researcher’s choice of sample split is a random variable independent of the data – the decision can be arbitrary or out of sophisticated considerations, as long as it is not contingent on the data. We show that data-independent sample split is innocuous in the sense that it does not inflate size of the out-of-sample tests at all. On the contrary, “data-driven” means that the researcher’s choice of sample split does not observe a rule specified before embarking on the empirical analysis and is ultimately dependent on the data; the sample would have been split into different proportions if the data turned out to be different. The data-driven mode of splitting the sample in order to maximize the test statistic elevates size of the test to far above the nominal level, as Rossi and Inoue (2012) and Hansen and Timmermann (2012) have already pointed out. We call attention to some subtleties. We add that there is a variety of data-driven modes, and the subtle problem is that a specific mode of sample split creates specific uncertainty that requires ad hoc treatment, that is, ad hoc adjustment of the critical value to make the test correctly sized.

It is therefore important to provide guidelines for development of ad hoc treatments for different modes of sample split, assuming that we are informed about the mode. To this end, we adapt the approach of Rossi and Inoue (2012) by building explicit link between modes of sample split and asymptotic distributions of the test statistic. Rossi and Inoue (2012) astutely point out that convergence of many conventional out-of-sample test statistics may be generalized to functional convergence, i.e., weak convergence of the test statistic viewed as stochastic process. With this finding they show how novel out-of-sample tests can be developed by explicitly mining over sample split points and then adjusting the critical value to offset the effect of data-mining. We borrow from them the technical framework but give different interpretations. To be precise, we make use of the functional convergence to derive asymptotic distribution of the out-of-sample test statistic under different modes of sample split, and – from the perspective of those who seek to make the test correctly sized – the resulting tests are interpreted as ad hoc treatments for the corresponding modes. In particular, the sup-type test of Rossi and Inoue (2012) is interpreted as the ad hoc treatment for what we term the sup-mode of sample split, and their average-type test is interpreted as the ad hoc treatment if one evaluates the test statistic at all possible sample split points and reports the arithmetic average.

We call attention to the effect of and the treatment for the pass-mode of sample split, which is not covered by Rossi and Inoue (2012) and Hansen and Timmermann (2012). “Pass” means that one starts off the out-of-sample analysis by trying a sample split point and then evaluating the test statistic; another split point will be tried if the test statistic fails to pass a desired threshold, and the conduct of “split and test” stops only if the test statistic passes
the threshold or the sample has been exhausted. The pass-mode of sample split is indeed an (pseudo) out-of-sample analogue of optional stopping\(^4\) – also called repeated significance test – a sampling mode whose (detrimental) effect has long been noted by statisticians. Feller (1940) was perhaps the first to point out that optional stopping can eventually result in significant test statistic even if the null hypothesis is true. The problem has since then been recurrently studied by Anscombe (1954), Armitage et al. (1969), Berger and Berry (1988), Kadane et al. (1996), Wagenmakers (2007) and so forth. But to our knowledge there has been no discussion on the parallel problem in pseudo out-of-sample analysis. Knowing the effect of the pass-mode of sample split is of equal importance as knowing that of the sup-mode, since the pass-mode of sample split might be equally common as or even more than the sup-mode\(^5\). Our analysis reveals a subtle attribute of the pass-mode. If one chooses a threshold that is just above the conventional critical value, the pass-mode can cause size distortion to almost the same degree as the sup-mode, but on the other hand requires a much smaller critical value than the latter to be correctly sized. The finding entails that the sup-type test developed by Rossi and Inoue (2012) might cost too much power to alleviate size distortion, if what’s in operation is the pass-mode\(^6\). Our analysis thus furnishes concrete examples for Diebold (2015)’s comment that pseudo out-of-sample analysis is subject to an expanded scope of strategic data-mining and achieving robustness comes at the cost of additional power loss\(^7\).

After discussing size of out-of-sample test in the presence of data-driven sample split, we turn to power of the test. The flip side of the treatment for a data-driven mode of sample split is a type of out-of-sample test. To be specific, the combination of a data-driven mode of sample split and its treatment makes a (novel) out-of-sample test, which is indeed what Rossi and Inoue (2012) utilize to develop their sup-type and average-type tests. A question inspired by Rossi and Inoue (2012) concerns whether such combination renders a test more powerful than the conventional one using fixed split point. On the one hand, they argue that “traditional tests may also lack power to detect predictive ability when implemented for an ‘ad-hoc’ choice of the window size” (p. 450), and “it is possible that when there is some predictive ability only over a portion of the sample, he or she [the researcher] may lack empirical evidence in favor of predictive ability because the window size was either too small

\(^4\)Optional stopping refers to the conduct of analyzing data as they accumulate and stopping the experiment as soon as the test statistic passes some desired threshold.

\(^5\)In dealing with \(p\)-hacking, Simonsohn et al. (2014) remark: “if a researcher \(p\)-hacks, attempting additional analyses in order to turn a nonsignificant result into a significant one... [they] are unlikely to pursue the lowest possible \(p\) value; rather, we suspect that \(p\)-hacking frequently stops upon obtaining significance.”

\(^6\)To illustrate, suppose someone has conducted the pass-mode of sample split and presents only the test statistic. Uninformed about the mode of sample split, we make the conservative decision to use critical value of the sup-type test. The critical value of the sup-type test does keep the rate of false discovery of unequal predictive ability below the stipulated size, but the robustness comes at the cost of considerably lower power than the \textit{ad hoc} test that is suited to the pass-mode.

\(^7\)Diebold’s original comments is primarily on the cost for achieving robustness to mining over variables/models. But it is evident that the same comments are equally fair on achieving robustness to mining over the sample split point.
or too large to capture it” (p. 434). On the other hand, they also note that there are “two opposite forces” that could offset each other, and remark that “it is not clear a priori whether our test would find more or less empirical evidence in favor of predictive ability” (p. 446). Answer to this question is important for practitioners. The combination of data-mining and its correction is neither parsimonious nor straightforward, since it involves a searching process and derivation of the correct asymptotic distribution. Therefore, it is worthwhile to take the roundabout route of “data-mining first, self-correcting after” only if it leads to more powerful test.

To formally address this question, we analyze asymptotic power of the sup-type test under some local alternatives and contrast it with the power of conventional test using fixed split point. To this end, we modify the test statistic proposed by West (1996) and McCracken (2000) – which is the same test statistic considered by Rossi and Inoue (2012) – and show that it converges weakly to a time-changed Wiener process under the null hypothesis of equal predictive accuracy and to a time-changed Wiener process with drift under local alternatives. With this finding, we derive analytic formulae for obtaining critical values of the sup-type test, thereby circumventing the use of Monte Carlo approximation as in Rossi and Inoue (2012). We also show that computation of asymptotic power of the sup-type test under local alternatives is essentially the same as computation of (nonlinear) boundary crossing probability of Wiener process (see, e.g., Siegmund (1986)). When the impact of estimation uncertainty vanishes, that is, when the framework collapses to the one of Diebold and Mariano (1995), the boundary is linear and we invoke some well-known formulae (Anderson (1960); Siegmund (1986); Hall (1997)) to derive analytic formulae for the power under local alternatives.

We draw two contrasting findings from the analysis. First, under the local alternative with Pitman drift, the sup-type test is no more powerful than the conventional test using fixed and small split point. Second, the sup-type test becomes more powerful than the conventional test under the local alternative with structural break. In this regard, our analysis formally justifies the claim of Rossi and Inoue (2012) that the sup-type test can be more powerful when there is structural break rendering predictive abilities of two models unequal. Some intuitive explanations on the findings follow. In the presence of structural break, the pre-break and post-break observations are not equally informative. For example, if two models once possessed equal predictive abilities before the structural break but no longer do so, the post-break observations are more likely to reveal the prevailing unequal predictive abilities than the pre-break observations. With too large split point the conventional test tends to overlook many evidential post-break observations, whereas with too small split point it tends to include many uninformative pre-break observations. By contrast, the sup-

---

8 The original remark: “By considering a wide variety of window sizes, our tests might be more likely to find empirical evidence in favor of predictive ability, as our Monte Carlo results have shown. However, by correcting statistical inference to take into account the search process across multiple window sizes, our tests might at the same time be less likely to find empirical evidence in favor of predictive ability.”
type test makes more effective use of the post-break observations as the test statistic tends to be maximized around the break time, and hence it becomes more powerful than the conventional test\(^9\). On the other hand, all observations are equally evidential under the local alternative with Pitman drift. In this case, the conventional test using small split point makes equally or even more effective use of the observations than the sup-type test, so the latter cannot be more powerful.

It should be noted, however, that our analysis of power has two limitations. First, the analysis is confined to comparison of nonnested models. Second, while the modified sup-type test developed in this paper is correctly sized under the null hypothesis whether the impact of estimation uncertainty vanishes or not, the formulae for its asymptotic power under local alternatives are applicable only when the estimation uncertainty becomes immaterial. The second limitation is due to our lack of closed-form formulae for nonlinear boundary crossing probability, which is crucial for calculating the power. Nonetheless, if the estimation uncertainty is small, the formulae derived in this paper might still be a good approximation, so the second limitation might not be very restrictive. On the other hand, the first limitation does restrict applicability of results in this paper, since we often run into nested models in practice. What keeps us from deriving analytic results for the nested case are the technical pitfalls summarized in Section 1.8. Because of these pitfalls, we leave a more thorough analysis of the nested case to future research.

The remainder of the paper is structured as follows. Section 1.2 lays out the general setting and explains how to adapt the approach of Rossi and Inoue (2012) to develop *ad hoc* treatments for data-driven modes of sample split. Section 1.3 generalizes West (1996)'s result to weak convergence of stochastic process. Section 1.4 develops a modified sup-type test for comparison of nonnested models that circumvents the use of Monte Carlo. Section 1.5 analyzes power of the (modified) sup-type test and that of the conventional test under local alternatives. Section 1.6 contains some small Monte Carlo experiments. Section 1.7 briefly discusses other novel out-of-sample tests obtained by explicit data-mining. Section 1.8 explains the pitfalls that keep us from analyzing sup-type test for comparison of nested models; it also explains the pitfalls that we face in deriving analogous results for comparison of forecast methods (Giacomini and White (2006)). Section 1.9 provides concluding remarks. Appendix A details how nuisance parameters are estimated in the Monte Carlo experiments. Finally, all proofs are relegated to the Appendix B.

### 1.2 The General Setting and The Main Idea

This section is by and large a reinterpretation of the method of Rossi and Inoue (2012). We first lay out the framework and give concrete examples for data-driven modes of sample split.

---

\(^9\)It might be argued that the conventional test with fixed split point around the break time could be more powerful than the sup-type test, but this test is infeasible if the break time is unknown.
split. Then we illustrate how the approach of Rossi and Inoue (2012) can be adapted to develop ad hoc treatments. No distinction between nested and nonnested models is drawn for this section.

### 1.2.1 Notations

Throughout this paper we let $\lfloor r \rfloor$ denote the integer part of a real number $r$. $[l, u]$ means the interval of all real numbers between $l$ and $u$, and, as a moderate abuse of notation, $[\lfloor l \rfloor, \lfloor u \rfloor]$ means the set of all integers between $\lfloor l \rfloor$ and $\lfloor u \rfloor$. $D^k[l, u]$ denotes the Skorokhod space of all càdlàg $\mathbb{R}^k$-valued functions defined on the interval $[l, u]$; when $k = 1$ we simply write $D[l, u]$. $C^k[l, u]$ and $C[l, u]$ denote the space of all continuous functions whose topology is induced by the uniform norm, that is, $\|x - y\| = \sup_{l \leq t \leq u} |x(t) - y(t)|$ for all $x, y \in C^k[l, u]$. We recall that the Skorokhod topology relativized to $C^k[l, u]$ coincides with the uniform topology (see Chapter VI, Jacod and Shiryaev (2003)), and since the limit process of the test statistic we consider is continuous, we shall use the uniform norm unless otherwise stated. Almost sure convergence and convergence in probability are denoted by $\overset{a.s.}{\rightarrow}$ and $\overset{p}{\rightarrow}$ respectively. Convergence in law of a scalar is denoted by $\overset{L}{\rightarrow}$, and weak convergence of a stochastic process by $\overset{w}{\rightarrow}$. Finally, $1(\cdot)$ denotes the indicator function.

### 1.2.2 The setting

The framework is consistent with that of West (1996), McCracken (2000), Clark and McCracken (2001), McCracken (2007) and so on. Let $h$ denote the forecast horizon, $y_{t+h}$ the variable to be forecast at period $t$ and $x_t$ the variables used to produce the forecast. We suppose the availability of the historical sample $\{x_t, y_t : 1 \leq t \leq T + h\}$, and that one would like to contrast two competing models, both of which seek to forecast $y_{t+h}$ with the data $\{x_s : s \leq t\}$ at period $t$. Parameters of the two models are denoted by $\theta$ and $\gamma$ respectively; in particular, pseudo-true parameters are denoted by $\theta^*$ and $\gamma^*$. As in most statistical applications, $\theta^*$ and $\gamma^*$ are assumed to be unknown and must be estimated before being plugged back into the models. This dictates that the sample must be split into two segments, with the first segment used to produce parameter estimates for the first forecast, and the second to contrast the pseudo “real-time” forecasts. Specifically, we let $1 \leq R_T \leq T$ denote the sample split point. Then the first forecast of $y_{R_T+h}$ is made with data available at time $R_T$, the second of $y_{R_T+1+h}$ with data available at time $R_T + 1$, and so on. We let $\hat{\theta}_{t,R_T}$ and $\tilde{\gamma}_{t,R_T}$ denote estimates of the parameters $\theta$ and $\gamma$; the estimation can be done with recursive window, rolling window or fixed window. Forecasts produced by plugging in $\theta$ and $\gamma$ are denoted by

---

10 This is called by Diebold (2015) the (“old-school”) WCM framework.

11 Variables of the two models are stacked together to form the vector $x_t$.

12 More detailed descriptions of the three estimation schemes can be found in p. 819, West and McCracken (1998).
\( \hat{y}_{t+h|t}(\theta) \) and \( \hat{y}_{t+h|t}(\gamma) \) respectively. The forecasts produced with estimated parameters and used in practice are thus \( \hat{y}_{t+h|t}(\hat{\theta}_{t,R_T}) \) and \( \hat{y}_{t+h|t}(\hat{\gamma}_{t,R_T}) \). The forecasts corresponding to the pseudo-true parameter values \( \theta^* \) and \( \gamma^* \) are \( \hat{y}_{t+h|t}(\theta^*) \) and \( \hat{y}_{t+h|t}(\gamma^*) \), and when it causes no confusion we may suppress their dependence on parameters and write \( \hat{y}_{t+h|t} \equiv \hat{y}_{t+h|t}(\theta^*) \) and \( \hat{y}_{t+h|t} \equiv \hat{y}_{t+h|t}(\gamma^*) \) for short. There will be \( T - R_T + 1 \) pseudo real-time forecasts produced by each model, and one might gauge deviations of the forecasts from the true values by pre-defined loss functions. We let \( L_{t+h}^{(1)}(y_{t+h}, \hat{y}_{t+h|t}) \) and \( L_{t+h}^{(2)}(y_{t+h}, \hat{y}_{t+h|t}) \) denote losses of the two forecasts. Again we write \( L_{t+h}^{(1)} \) and \( L_{t+h}^{(2)} \) for short when it causes no confusion. The null hypothesis we would like to test is equal predictive ability of the two models:

\[
E \left[ L_{t+h}^{(1)} \right] = E \left[ L_{t+h}^{(2)} \right] \quad \text{for all} \ t
\]

Once the pseudo out-of-sample forecasts are available, we can construct test statistic based on the time series \( \{L_{t+h}^{(1)}, L_{t+h}^{(2)} : R_T \leq t \leq T\} \). The concrete forms of the loss functions and the test statistic hinge on whether the models are nested or not (see, e.g., Clark and McCracken (2001)), and will be introduced after we explain treatments for different modes of sample split. What we must reiterate is that the null hypothesis (1.1) concerns the forecast models evaluated at the pseudo-true parameters \( \theta^* \) and \( \gamma^* \) but not the estimated parameters \( \hat{\theta}_{t,R_T} \) and \( \hat{\gamma}_{t,R_T} \). It therefore means equal predictive ability of the two forecast models and should be distinguished from equal predictive ability of two forecast methods (Giacomini and White (2006); Giacomini and Rossi (2010)).

For now we simply let \( d_T(R_T) \) denote the out-of-sample test statistic of equal predictive ability. The conventional tests, such as the ones developed by West (1996), McCracken (2000) and McCracken (2007), all assume that \( R_T \) is a predetermined fraction of \( T \) but not chosen after observing the data. Intuitively, intentional choice of \( R_T \) will imbue the test statistic with additional uncertainty and invalidate the asymptotic distribution derived without accounting for such uncertainty. For example, if one “fishes” through a range of sample split points to attain the largest possible value of the test statistic, the conventional critical values not adjusted to offset the effect of such “fishing” behavior will lead the test to over-reject the null hypothesis of equal predictive ability. This mode is the focal point of Rossi and Inoue (2012) and Hansen and Timmermann (2012), which might be quite common but is not the only mode of sample split. As we shall exemplify below, there are a multitude of modes of sample split that might be conducted. Importantly, some modes of sample split are innocuous in the sense that no adjustment to the conventional tests is required, while the others require ad hoc treatments. For clarity, we distinguish two modes of sample split.

**Definition 1.** When \( R_T \) is a \( \tilde{F}_{T+h} \equiv \sigma(\{x_t, y_t : 1 \leq t \leq T + h\}) \)-measurable random variable, the sample split is said to be data-driven. When \( R_T \) is chosen as a random variable indepen-

\[13\]More precisely, they assume that \( R_T \) is nonrandom and \( R_T/T \) converges.
dent of $\mathcal{F}_T$, the sample split is said to be \textit{data-independent}.

By our definition, fixing the sample split point at a constant fraction of the total sample is a data-independent mode of sample split. Another example is drawing $R_T$ at random from $[0.1T, 0.9T]$. In general, data-independent sample split can be arbitrary or out of scrupulous considerations. For example, if the researcher has read Hansen and Timmermann (2015) and realizes that under local alternatives with Pitman drift the test is asymptotically more powerful with long hold-out sample, he might find it preferable to draw $R_T$ as a random variable with left-skewed distribution\textsuperscript{14}. On the other hand, data-driven sample split includes the mode of maximizing (the absolute value of) the test statistic by mining over the sample split points, namely the mode considered by Rossi and Inoue (2012) and Hansen and Timmermann (2012). Other modes of sample split will be illustrated later. As we shall see, data-independent sample split is always innocuous, while data-driven sample split begets uncertainty that must be appropriately dealt with.

1.2.3 Data-independent sample split

We consider first data-independent sample split. One assumption we maintain throughout the paper is

\begin{quote}
\begin{center}
\textbf{Assumption A.} The sample split point $R_T$ comes from the interval $[\lfloor l T \rfloor, \lfloor u T \rfloor]$, where $0 < l < u < 1$ are constants fixed \textit{ex ante}.\end{center}
\end{quote}

The same assumption is imposed by both Rossi and Inoue (2012) and Hansen and Timmermann (2012), and it means that one decides, \textit{ex ante}, to reserve a fixed fraction of the total sample for initial estimation and another fixed fraction for out-of-sample evaluation. Further mining over the “reserved” fractions is thus ruled out.

We define a càdlàg process $S_T$ by

$$S_T(\xi) \equiv d_T(\lfloor \xi T \rfloor), \quad l \leq \xi \leq 1$$  \hspace{1cm} (1.2)

Putting the upper bound as 1 rather than $u$ is intended to ease our exposition of the asymptotic results. The process $S_T$ plays a key role in establishing asymptotics under different modes of sample split. Clearly, choosing $R_T$ from $[\lfloor l T \rfloor, \lfloor u T \rfloor]$ for $d_T(R_T)$ corresponds to choosing $\xi$ from $[l, u]$ for $S_T(\xi)$.

Now we are ready to prove that data-independent sample split does not invalidate conventional critical values.

\begin{proposition}
Let the process $S_T$ be given by (1.2). We impose the following conditions:
\end{proposition}

\textsuperscript{14}What keeps the researcher from simply fixing the sample split point at a very small fraction might be the concern about inaccurate parameter estimate.
1.2. The General Setting and The Main Idea

(a). The sample is split in such way that \( R_T \equiv \lfloor \delta T \rfloor \), where \( \delta \) is a random variable supported on \([l, u]\) and independent of the data;
(b). For any \( \xi \in [l, u] \), \( S_T(\xi) \) converges in law to \( S(\xi) \), where \( S(\xi) \) has the same law \( \mathcal{L} \) that does not depend on \( \xi \).

Then \( \mathcal{L} \) is also the asymptotic law of \( d_T(R_T) \).

Condition (a) is fulfilled by a good many modes of sample split. A leading case is \( R_T \) drawn from the uniform distribution over \([\lfloor l T \rfloor, \lfloor u T \rfloor]\). To see this, note first that \( d_T(R_T) = S_T(R_T/T) \), where \( S_T \) is defined in (1.2). Because \( S_T \) is a càdlàg step process that is fully described by its value at the jump points and \( R_T \) is assumed to be independent of the data, \( S_T(R_T/T) \) has the same law as \( S_T(\delta) \), where \( \delta \) is uniformly distributed over the interval \([l, u]\). The case where \( R_T \) is a fixed fraction of \( T \) (e.g., \( R_T = \frac{1}{2} T \)) is also encompassed by (a).

We can also entertain with other distributions over \([l, u]\), such as an \( U \)-shaped distribution that renders the split point more likely to be either early or late, or a hump-shaped distribution that makes the split point more likely to fall in the middle of the sample. The pointwise convergence in law in condition (b) is quite weak. On the other hand, the “equi-distribution” assumption, i.e., the law of \( S(\xi) \) does not depend on \( \xi \), is obviously satisfied by West (1996)’s test statistic for comparison of nonnested models, whose limit distribution is standard normal. For comparison of nested models, Hansen and Timmermann (2015) show that the \( F \)-test statistic of McCracken (2007) can be transformed to be \( \xi \)-invariant as well. Proposition 1 indeed slightly generalizes the results of West (1996), McCracken (2000) and McCracken (2007). In these papers \( R_T \) is assumed to be nonrandom and \( R_T/T \) converges to a constant. Proposition 1 instead allows \( R_T \) to be a random number independent of the data, and apparently \( R_T/T \) needs not converge in this case.

1.2.4 Data-driven sample split: Examples

Unlike Data-independent sample split, data-driven modes exert effects on the test statistic and the effects must be appropriately dealt with. There are many data-driven modes, and we give several examples that might be common in practice.

Example 1 (Sup-mode). The mode considered by Rossi and Inoue (2012) and Hansen and Timmermann (2012) may be called the sup-mode. The two-sided sup-mode is described by

\[
R_T \equiv \arg\max_{|IT| \leq r \leq uT} |d_T(r)| \tag{1.3}
\]

One attains the largest absolute value of the test statistic by conducting the two-sided sup-mode, i.e., \( |d_T(R_T)| = \sup_{|IT| \leq r \leq uT} |d_T(r)| \). There is also the one-sided sup-mode, which is described by

\[
R_T \equiv \arg\max_{|IT| \leq r \leq uT} d_T(r) \quad \text{or} \quad R_T \equiv \arg\min_{|IT| \leq r \leq uT} d_T(r) \tag{1.4}
\]
The conducts of the two-sided and the one-sided sup-modes may be driven by different psychological motives. The two-sided sup-mode leaves room for outcome switching (p.3, Munafò et al. (2017)), that is, either of the competing models can be argued to have better predictive ability as long as the null hypothesis gets rejected. The one-sided sup-mode may be associated with the predisposition to favoring one particular candidate of the competing models.

Example 2 (Pass-mode). The two-sided pass-mode that we mentioned in the introduction can be described by

\[ R_T \equiv \inf\{ |l_T| \leq r \leq |u_T| : |d_T(r)| > c \} \land |u_T| \]  

(1.5)

where \( c > 0 \) is a pre-specified threshold. In plain words, it means that one starts off by evaluating the test statistic at \(|l_T|\). \( R_T \) is set to \(|l_T|\) if the test statistic passes the threshold \( c \); otherwise one tries \(|l_T| + 1\) and checks again if the threshold has been passed. The process is repeated until one finds a sample split point that lifts the test statistic above the threshold, or if all possible split points are exhausted; in the latter case \( R_T \) is set to \(|u_T|\).

As we have mentioned in introduction, the pass-mode (1.5) is an analogue of the repeated significance test or optional stopping that has received quite a few discussions in the literature of clinical trials, biostatistics and psychology (Wagenmakers (2007); García-Pérez (2012)). Compared to the sup-mode, the pass-mode involves more subtle problems. First, every choice of \( c \) makes a specific pass-mode and thus requires ad hoc treatment. Second, there are many variants of the pass-mode. It is more accurate to call the mode described by (1.5) the forward pass-mode to stress that it starts searching from small split point. Reversing the search route results in the backward pass-mode whose effect and treatment are likely to be different from the forward pass-mode. One may even take different routes from the forward and backward ones. For example, one can take a “binomial tree” search route: it starts by trying \( R_T = \frac{1}{2} T \); if the test statistic fails to pass the threshold, the split points \( R_T = \frac{1}{4} T \) and \( R_T = \frac{3}{4} T \) are tried; if neither of them renders the test statistic larger than the threshold, the split points \( R_T = \frac{1}{8} T \), \( R_T = \frac{3}{8} T \), \( R_T = \frac{5}{8} T \) and \( R_T = \frac{7}{8} T \) are tried. The procedures are repeated until the threshold is passed. The different examples of the pass-mode in fact echo Diebold (2015)’s comment that pseudo out-of-sample analysis is subject to an expanded scope of data-mining. More explicitly, in “genuine” out-of-sample analysis the pass-mode or optional stopping can only take the forward route, while in pseudo out-of-sample analysis a multitude of search routes become possible. Finally, the pass-mode can actually distort size to the same degree as the sup-mode but requires less mining time, as it

15There are psychological experiments (Rosenthal and Gaito (1963); Rosenthal and Gaito (1964)) in which the p-value 0.05 exhibits a “cliff effect”, namely that the human subjects’ confidence in a result drops abruptly when the p-value just rises above the 0.05 level. In this light, perhaps people have the propensity for choosing a threshold slightly larger than the critical value for the 0.05 significance level. But certainly we do not know what would be the choice of a particular researcher in a particular situation.
stops mining as soon as the test statistic goes beyond the threshold.

**Example 3 (Mixed Sup-mode).** There are many other data-driven modes that might be equally common in practice. For example, define

\[ \mathcal{R}_T \equiv \{ [l_T] \leq r \leq [u_T] : |d_T(r)| > c \} \]

which is the set of all sample split points that render the test statistic greater in absolute value than the threshold \( c \). A common mode of sample split that is prone to result in over-rejection of the null hypothesis of equal predictive ability is described by

\[
R_T \equiv \begin{cases} 
\text{A random element of } \mathcal{R}_T, & \mathcal{R}_T \neq \emptyset \\
\text{A random element of } [l_T, u_T], & \mathcal{R}_T = \emptyset 
\end{cases} \tag{1.6}
\]

The definition of \( R_T \) means that one is indifferent to the split points in \( \mathcal{R}_T \) as they all make the test statistic large enough. Whether the set \( \mathcal{R}_T \) is empty depends on whether the supremum \( \sup_{l_T \leq r \leq u_T} |d_T(r)| \) passes the threshold. This mode can be viewed as a mixture of the sup-mode and a “random” choice; for this reason we call it the *mixed sup*-mode. The conduct of this mode is no more complicated than the sup-mode and the pass-mode, but its treatment requires more careful development since the mode involves a random set and a further choice variable. We will not further discuss the mixed-sup mode in this paper.

The above examples are just the tip of the iceberg and there are many other data-driven modes that might be as common as the mentioned ones, but in the sequel we shall focus only on the sup-mode and the pass-mode. Our next step is to adapt the approach of Rossi and Inoue (2012) to provide general guidelines for development of *ad hoc* treatments for different modes of sample split.

### 1.2.5 Data-driven sample split: The treatments

In this section we give a reinterpretation of the approach of Rossi and Inoue (2012). We study the cause and effect between data-driven modes of sample split and asymptotic distributions of the test statistic under the null hypothesis of equal predictive ability. Critical values drawn from the correct asymptotic distributions are termed *due treatments* for the mode in effect. Apart from the concern about size, we have also suggested that combination of a data-driven mode of sample split and its treatment makes a type of out-of-sample test. Our interest in the due treatments is hence rooted in the consideration for both size and power of the test.

The core idea is to establish weak convergence of the stochastic process \( S_T \) defined in (1.2) to some process \( S \). Once this is done, it remains to represent \( d_T(R_T) \) (or \( |d_T(R_T)| \)) as a functional of \( S_T \) and invoke the continuous mapping theorem. For importance of this idea...
we state it in the following proposition, which is slightly more general than Proposition 1 of Rossi and Inoue (2012).

**Proposition 2.** Suppose that the test statistic can be represented as $|d_T(R_T)| = \varphi_T(S_T)$, where $S_T$ is defined in (1.2). Assume in addition the following conditions:

(a) $S_T \xrightarrow{u} S$ on $\mathbb{D}[l,1]$, where $S$ is a continuous process;

(b) The sequence of functions $\{\varphi_T : \mathbb{C}[l,1] \rightarrow \mathbb{R}\}$ is pointwise equicontinuous with respect to the topology induced by the uniform norm.

(c) There exists a Borel mapping $\varphi : \mathbb{C}[l,1] \rightarrow \mathbb{R}$, such that $\varphi_T(S) \xrightarrow{L} \varphi(S)$.

Then the test statistic $|d_T(R_T)|$ converges in law to $\varphi(S)$.

**Remark 1.** Condition (b) is stated for a family of functions $\{\varphi_T\}$ rather than a single function $\varphi$, since we would like to encompass the case where $R_T$ is nonrandom and $\lim T \frac{R_T}{T} = \xi$, i.e., the condition imposed by West (1996), McCracken (2000) and McCracken (2007). To see this, note that we can define $\varphi_T(x) \equiv x(R_T/T)$, where $x$ is any continuous function. The sequence $\{\varphi_T : T \geq 1\}$ is clearly pointwise equicontinuous. Defining $\varphi(x) \equiv x(\xi)$, we have $\varphi_T(x) = x \left( \frac{R_T}{T} \right) \rightarrow x(\xi) = \varphi(x)$ by continuity of $x$, whence $\varphi_T(S) \xrightarrow{L} \varphi(S)$. The results in the cited papers then follow from our proposition.

If one conducts the two-sided sup-mode of sample split (1.3), the test statistic can be represented as

$$|d_T(R_T)| = \sup_{l \leq \xi \leq u} |S_T(\xi)|$$

Thus we obtain the sup-type test of Rossi and Inoue (2012), which in our framework is interpreted as the *ad hoc* treatment for the two-sided sup-mode.

**Corollary 3 (Treatment for two-sided sup-mode).** Suppose that one conducts the two-sided sup-mode of sample split described by (1.3), and $S_T$ converges weakly to a continuous process $S$. Then

$$|d_T(R_T)| \xrightarrow{L} \sup_{l \leq \xi \leq u} |S(\xi)|$$

The pass-mode of sample split engenders very different uncertainty that must be dealt with in a different way. Again applying Proposition 2 we obtain the following corollary.

**Corollary 4 (Treatment for two-sided pass-mode).** Suppose that one conducts the two-sided pass-mode of sample split described by (1.5). In addition, $S_T$ converges weakly to a continuous process $S$ satisfying $\mathbb{P}\left( \sup_{l \leq \xi \leq u} |S(\xi)| = c \right) = 0$, where $c$ is the threshold of the pass-mode. Then

$$|d_T(R_T)| \xrightarrow{L} |S(\tau)|$$

where $\tau \equiv \inf\{l \leq \xi \leq u : |S(\xi)| > c\} \wedge u$.

---

16What we mean by "pointwise equicontinuous" here is that for any $x \in C[l,1]$ and $\varepsilon > 0$, there exists $\delta > 0$, such that $\limsup_T |\varphi_T(x) - \varphi_T(y)| < \varepsilon$ as long as $\|x - y\| < \delta$.  

---

13
Remark 2. Interpretation of this corollary requires some cautions. First of all, it means that the pass-mode of sample split makes the test statistic converge in law to the limit process stopped at the first passage time. Distribution of the stopped process is very different from that of the running supremum. The critical values as quantiles of the limit distribution thus differ drastically for the two modes of sample split. When the threshold $c$ governing the pass-mode is small, distribution of the running supremum has longer right tail, so critical values of the sup-type test would be too conservative and incurs severe loss of power to alleviate size distortion if the pass-mode is in effect. This problem will become manifest in the Monte Carlo experiments. The second problem concerns critical values for the pass-mode. Although the convergence in law seems to cast light on behaviour of the test statistic, it is of little avail in finite sample. The reason is that $|S(\tau)| \leq c$ holds with probability one. This entails that any critical value based on the distribution of $|S(\tau)|$ must not exceed $c$. The problem immediately ensues: in finite sample, the tail-quantiles of $|d_T(R_T)|$ are often greater than $c$. This becomes evident if we consider Monte Carlo approximation of $|S(\tau)|$. We might simulate a discretized version of $S$ and approximate its value at the first passage time, and this value is not prevented from going beyond $c$. Indeed, finding suitable critical values in finite sample is the main challenge for developing the pass-type test.

Both Corollary 3 and Corollary 4 represent the idea that if a mode of sample split is conducted, one should use the (asymptotic) population analogue for choosing critical values. The same idea has indeed broader application than treating just uncertainty arising from sample split. We give two examples to illustrate how this idea can be applied to deal with different behavioral transformations of the test statistic.

Example 4 (The average-type test). Suppose that one computes the test statistic $d_T(R_T)$ for $R_T = [l_T], [l_T] + 1, \ldots, [u_T]$ and takes the average:

$$
\frac{1}{[u_T] - [l_T] + 1} \sum_{R_T = [l_T]}^{[u_T]} d_T(R_T)
$$

The conventional critical values are invalid for this average-type test statistic. Instead, given the assumption $S_T \xrightarrow{w} S$, Proposition 2 implies

$$
\frac{1}{[u_T] - [l_T] + 1} \sum_{R_T = [l_T]}^{[u_T]} d_T(R_T) \xrightarrow{L} \int_l^u S(\xi) d\xi
$$

so asymptotically valid critical values should be derived from the distribution of $\int_l^u S(\xi) d\xi$. This leads to the average-type test of Rossi and Inoue (2012), but in our framework it is interpreted as the ad hoc treatment for averaging the conventional test statistic evaluated at a range of split points.

\[\text{If the process } S \text{ fails to pass the threshold on } [l, u], \text{ then we would have } |S(\tau)| = |S(u)| < c; \text{ otherwise by continuity we must have } |S(\tau)| = c.\]
Example 5. In reviewing some structural exchange rate forecast models (Engel et al. (2007); Gourinchas and Rey (2007); Molodtsova and Papell (2009)), Rogoff and Stavrakeva (2008) argue that failure to sufficiently check robustness to alternative time windows have led many studies to overstate the models’ short-horizon predictive abilities. By “robustness” Rogoff and Stavrakeva (2008) mean that the test statistic evaluated at different forecast windows should stay above a prescribed level\(^{18}\). However, such definition of robustness would make the test largely undersized if conventional critical values are used. To see this, let \(c_{0.9}\) denote the conventional critical value for the 0.1 significance level. Under the null hypothesis that the structural model does not produce better forecast than a driftless random walk, we have
\[
\lim_T P(|d_T(R_T)| > c_{0.9}) = 0.1 \quad \text{if } R_T \text{ is any fixed point in } [\lfloor l_T \rfloor, \lfloor u_T \rfloor].
\]
We note that Rogoff and Stavrakeva (2008)’s definition implicitly asks for \(\inf_{\lfloor l_T \rfloor \leq R_T \leq \lfloor u_T \rfloor} |d_T(R_T)| > c_{0.9}\). Under the assumptions of Proposition 2,
\[
\lim_T P\left(\inf_{\lfloor l_T \rfloor \leq R_T \leq \lfloor u_T \rfloor} |d_T(R_T)| > c_{0.9}\right) = P\left(\inf_{l \leq \xi \leq u} |S(\xi)| > c_{0.9}\right)
\]
The probability on the right hand side is in fact much smaller than 0.1, that is, the true size is much smaller than the nominal level 0.1, which in turn means higher probability of making type II error. Therefore, it is likely that nonrobustness of the results in the cited papers, as Rogoff and Stavrakeva (2008) concern about, is due to use of a critical value not adjusted for the definition of robustness. \(\square\)

1.3 Comparison of nonnested forecast models

Having put forth the idea for developing treatments for data-driven modes of sample split, we turn to study specifically the comparison of nonnested forecast models\(^{19}\), whose framework is laid out by West (1996) and McCracken (2000). Our exposition follows West (1996) very closely and we seek to generalize his Theorem 4.1. The asymptotic results we derive differ partly from those in Rossi and Inoue (2012); see Remark 7 for discussions on the differences and their important implications. To keep the analysis rigorous, we carry out proofs in details and provide them in the appendix.

1.3.1 Uniform mean value expansion

West (1996) demonstrates that in addition to model uncertainty, the uncertainty stemming from estimation of the parameters should be taken into account in comparison of

\(^{18}\)The original argument of Rogoff and Stavrakeva (2008) is that \(p\)-value of the test conducted with different forecast windows should stay below certain level. For example, on p.16 they write “In order for a result to be considered robust, we would expect that the TU \(p\)-value is below 0.1 for almost all of the plotted forecast windows.”

\(^{19}\)We explain why we do not consider comparison of nested model in Section 1.8.
1.3. Comparison of Nonnested Forecast Models

Forecast models. The original framework of West (1996) assumes recursive estimation window; it is later extended by McCracken (2000) to accommodate rolling and fixed estimation windows. The test statistic we consider is

\[ d_T(R_T) \equiv T^{-1/2} \sum_{i=R_T}^{T} \left[ L^{(1)}_{t+h}(y_{t+h}, \hat{y}_{t+h+i}) - L^{(2)}_{t+h}(y_{t+h}, \tilde{y}_{t+h+i}) \right] \tag{1.7} \]

The above definition differs from that considered by West (1996), McCracken (2000) and Rossi and Inoue (2012) in two nuances. First, the normalizing factor is \( T \) rather than \((T - R_T + 1)^{-1/2}\). This is intended for obtaining a time-changed Wiener process with drift. Second, we do not include square root of the inverse of the long-run covariance matrix. This is intended for developing a modified sup-type test in the next section.

We introduce some additional notations and assumptions. Let

\[ f_{t+h}(\beta) \equiv L^{(1)}_{t+h}(y_{t+h}, \hat{y}_{t+h+i}(\theta)) - L^{(2)}_{t+h}(y_{t+h}, \tilde{y}_{t+h+i}(\gamma)) \tag{1.8} \]

be the loss differential evaluated at \( \beta = (\theta', \gamma')' \), which we assume to be \( k \)-dimensional. \( f_{t+h}(\beta^*) \) will simply be denoted by \( f_{t+h} \). We let \( \mu_t \equiv E[f_{t+h}] \); the null hypothesis of equal predictive ability corresponds to \( \mu_t = 0 \) for all \( t \). We will assume that \( \beta \rightarrow f_{t+h}(\beta) \) is twice differentiable, and we define the \( k \times 1 \)-vector \( F \equiv E[\partial_{\beta} f_{t+h}(\beta^*)] \). On the other hand, the estimator \( \hat{\beta}_t \) is assumed to take the form \( \hat{\beta}_t = \beta^* + B_t G_t \), where \( \{B_t : t \geq 1\} \) is a sequence of \( k \times q \) matrices converging almost surely to the constant matrix \( B \). We recall that the matrices \( F \) and \( B \) play key roles in the framework of West (1996) and McCracken (2000).

**Assumption B.** We impose the following assumptions for comparison of nonnested models:

1. In some neighborhood \( N \) around \( \beta^* \), the following conditions hold with probability approaching one:

   (a) \( \beta \rightarrow f_{t+h}(\beta) \) is twice continuously differentiable with respect to \( \beta \).

   (b) There exists measurable functions \( m_t \), such that \( \sup_{\beta \in N} \partial^2_{\beta} f_{t+h}(\beta) \leq m_t \) and \( \sup_t E[m_t] < \infty \).

2. The estimate satisfies \( \hat{\beta}_t - \beta^* = B_t G_t \), where \( B_t \) is a \( k \times q \) matrix (\( q \leq k \)) and \( G_t \) is a \( q \times 1 \) vector. Moreover,

   (a) \( B_t \xrightarrow{a.s.} B \) as \( t \rightarrow \infty \), where \( B \) is a constant matrix with full column rank.

   (b) \( G_t \) is \( G_t \equiv \frac{1}{t} \sum_{i=1}^{t} g_i(\beta^*) \), \( G_t \equiv \frac{1}{R_T} \sum_{i=t-R_T+1}^{t} g_i(\beta^*) \) and \( G_t \equiv \frac{1}{R_T} \sum_{i=1}^{R_T} h_i(\beta^*) \) in the recursive, rolling and fixed schemes respectively.

   (c) \( E[g_i(\beta^*)] = 0 \) for all \( i \).
Remark 4. There lies a subtlety in the assumptions. We have highlighted the subscript of uniform convergence of the covariance estimator. We do not need the high level assumption in Appendix A of Rossi and Inoue (2012) that stipulates (4.1) to the quicker decay of the temporal dependence is intended for generalizing West’s Equation and requires slightly higher decaying rate of the strong mixing coefficients. Specifically, the quicker decay of the temporal dependence is intended for generalizing West’s Equation (4.1) to uniform mean-value expansion, as we prove in Proposition 5 below. Finally, we do not need the high level assumption in Appendix A of Rossi and Inoue (2012) that stipulates uniform convergence of the covariance estimator.

Remark 3. We have dropped West (1996)’s Assumption A4 that requires $R_T$ to be nonrandom and $R_T/T$ converge. The reason is that we can already establish weak convergence under other assumptions, and West’s original result can be obtained by application of the continuous mapping theorem. The weak convergence is achieved by imposing Assumption B3(c) that requires slightly higher decaying rate of the strong mixing coefficients. Specifically, the quicker decay of the temporal dependence is intended for generalizing West’s Equation (4.1) to uniform mean-value expansion, as we prove in Proposition 5 below. Finally, we do not need the high level assumption in Appendix A of Rossi and Inoue (2012) that stipulates uniform convergence of the covariance estimator.

Remark 4. There lies a subtlety in the assumptions. We have highlighted the subscript of $\mu_t = \mathbb{E}[f_{t+h}(\beta^*)]$, because later in analysis of asymptotic power under local alternatives we would consider the case where $\mu_t$ depends on $t$, say $\mu_t = \frac{1}{\sqrt{T}}1(t \geq |\lambda T|)$ in the presence of structural break. See Assumption C below. However, if $\mu_t$ depends on $t$ (or $T$), the expectation $\mathbb{E}[\partial_\beta f_{t+h}(\beta^*)]$ may depend on $t$ (or $T$) as well, and in this case $U_t$ will not be covariance stationary. Of course, we could redefine

$$ U_t \equiv (\text{vec}(\partial_\beta f_{t+h}(\beta^*))', f_{t+h}(\beta^*) - \mu_t, g_t(\beta^*)')', $$

where $F_t \equiv \mathbb{E}[\partial_\beta f_{t+h}(\beta^*)]$, and $U_t$ thus defined may be covariance stationary. But a subtle problem immediately emerges: in Proposition 5 below, the constant vector $F$ should be replaced with the possibly time-varying vector $F_t$. Because $F_t$ is time-varying, the asymptotics established in Proposition 6 may need revision as well, and so are the subsequent analyses. The problem gets even trickier if we note that $\mathbb{E}[g_t(\beta^*)]$ may not be zero, but, say, $O(T^{-a})$ for some $a > 0$, under the local alternatives.

A meticulous analysis under local alternatives require further assumptions. We may, for instance, impose the following assumptions:

(a) $F_t \equiv \mathbb{E}[\partial_\beta f_{t+h}(\beta^*)]$ can be decomposed as $F_t = F + c_{F,t}$, where $F$ is a constant matrix and $\frac{1}{\sqrt{T}}\sum_{t=|\xi T|}^T c_{F,t}$ converges in the Skorokhod topology on $\mathbb{D}[I,1]$ to a càdlàg function $C_{F}(\xi)$.

\[\text{West (1996)} \text{ assumes a strong mixing rate of } 3d/(d-1)\]
1.3. Comparison of nonnested forecast models

(b) Let \( \hat{g}_t \equiv \mathbb{E}[g_t(\beta^*)] \). \( \frac{1}{\sqrt{T}} \sum_{t=\lfloor T \xi \rfloor}^T \hat{g}_t \) converges in the Skorohod topology on \( \mathbb{D}[l, 1] \) to a càdlàg function \( C_G(\xi) \).

(c) \( \text{vec}(\partial_\beta f_{t+h}(\beta^*))' - F_t, f_{t+h}(\beta^*) - \mu_t, g_t(\beta^*)' - \hat{g}_t' \) is covariance stationary.

The propositions established below should then be revised by including \( C_F(\xi) \) and \( C_G(\xi) \). In practice, it is important to make sure compatibility of these assumptions and the underlying model. Typically, under the null hypothesis (i.e., \( \mu_t = 0 \)) we would have \( c_{F,t} \equiv 0 \) and \( c_{G,t} \equiv 0 \). We will not pursue such generality in this paper.

To derive the limit process of \( S_T(\xi) \equiv d_T([T \xi]) \) we introduce another process

\[
Z_T(\xi) = T^{-1/2} \sum_{t=\lfloor T \xi \rfloor}^T (f_{t+h}(\hat{\beta}_t) - \mu_t), \ l \leq \xi \leq 1 \tag{1.9}
\]

\( Z_T \) differs from \( S_T \) in that the summands are subtracted by \( \mu_t \), and the two processes coincide under the null hypothesis of equal predictive ability \( H_0 : \mu_t = 0, \forall t \). The following proposition shows an uniform mean-value expansion for \( Z_T \).

**Proposition 5.** Let \( Z_T \) be given by (1.9). Under Assumption B, we have the uniform mean-value expansion

\[
Z_T(\xi) = T^{-1/2} \sum_{t=\lfloor T \xi \rfloor}^T (f_{t+h}(\beta^*) - \mu_t) + T^{-1/2} \sum_{t=\lfloor T \xi \rfloor}^T FBG_t + M_T(\xi), \ l \leq \xi \leq 1 \tag{1.10}
\]

where \( F = \mathbb{E}[\partial_\beta f_{t+h}(\beta^*)] \) and \( \sup_{l \leq \xi \leq 1} |M_T(\xi)| = o_p(1) \).

**Remark 5.** Proposition 5 generalizes West (1996)’s Equation (4.1), which gives the mean-value expansion (1.10) for a single \( \xi \) and does not involve the supremum. The same uniform expansion is tacitly assumed in Rossi and Inoue (2012). Similar expansion is also implicitly assumed by Rossi and Sekhposyan (2016). In fact, their Assumption A2(i) is a uniform version of the conclusion of West’s Lemma A4(a). We furnish a proof of this proposition in the appendix since it does not seem to follow readily from the original conclusion of West (1996) and there appears to be no straightforward adaptation of his proof.

1.3.2 Weak convergence

Once we have established the uniform mean-value expansion, weak convergence of the process \( Z_T \) can be proved by invoking suitable functional central limit theorem and continuous mapping theorem.

**Proposition 6.** Let \( W \) be a \((q+1)\)-dimensional standard Wiener process, and let

\[
X(\xi) \equiv \begin{pmatrix} X^{(1)}(\xi) \\ X^{(2)}(\xi) \end{pmatrix} \equiv \Omega^{1/2}W(\xi), \ \Omega \equiv \begin{bmatrix} v_{uf} & v_{fg} \\ v'_f & v'_g \end{bmatrix} \tag{1.11}
\]
where \(X^{(1)}\) and \(X^{(2)}\) are one-dimensional and \(q\)-dimensional respectively. According to the estimation scheme, define

```
recursive: \[ Z(\xi) \equiv X^{(1)}(1) - X^{(1)}(\xi) - F' B \int_\xi^1 \rho^{-1} X^{(2)}(\rho) d\rho, \ l \leq \xi \leq 1 \]
```

```
rolling: \[ Z(\xi) \equiv X^{(1)}(1) - X^{(1)}(\xi) - F' B \int_\xi^1 \xi^{-1} (X^{(2)}(\rho) - X^{(2)}(\rho - \xi)) d\rho, \ l \leq \xi \leq 1 \]
```

```
fixed: \[ Z(\xi) \equiv X^{(1)}(1) - X^{(1)}(\xi) - F' B \frac{1 - \xi}{\xi} X^{(2)}(\xi), \ l \leq \xi \leq 1 \]
```

Under Assumption B, the following assertions are true:

(a) \(Z_T\) converges weakly to \(Z\) in the \(D[l, 1]\).

(b) \(Z\) is a zero-mean Gaussian process. Moreover, defining \(V(\xi) \equiv \text{var}(Z(\xi))\) as the variance, it holds

\[
V(\xi) = (1 - \xi) v_{ff} + \lambda_{fg}(\xi) (F'Bv_{fg} + v_{fg}'B'F) + \lambda_{gg}(\xi) F'Bv_{gg}B'F
\]

where under different schemes, the functions \(\lambda_{fg}\) and \(\lambda_{gg}\) are

<table>
<thead>
<tr>
<th>Scheme</th>
<th>(\lambda_{fg})</th>
<th>(\lambda_{gg})</th>
</tr>
</thead>
<tbody>
<tr>
<td>recursive</td>
<td>((1 - \xi) + \xi \log(\xi))</td>
<td>((1 - \xi) + \xi \log(\xi))</td>
</tr>
<tr>
<td>rolling, (\xi \geq \frac{1}{2})</td>
<td>((1 - \xi)^2)</td>
<td>((1 - \xi)^2)</td>
</tr>
<tr>
<td>rolling, (l \leq \xi &lt; \frac{1}{2})</td>
<td>(\frac{3\xi}{2})</td>
<td>(\frac{4\xi}{3})</td>
</tr>
<tr>
<td>fixed</td>
<td>0</td>
<td>(\frac{3 - \xi}{3})</td>
</tr>
</tbody>
</table>

**Remark 6.** This proposition establishes weak convergence of stochastic process, hence generalizing Theorem 4.1 of West (1996). When \(R_T/T = \xi\) is fixed, the asymptotic variance of \(Z_T(\xi)\) differ from that derived in West (1996) only by a scaling factor \((1 - \xi)^{-1}\).

**Remark 7.** Working on the original test statistic of West (1996), Rossi and Inoue (2012) derive the limit process

\[
\hat{Z}(\xi) = \frac{W(1) - W(\xi)}{\sqrt{1 - \xi}}, \ l \leq \xi \leq u
\]

where \(W\) is an one-dimensional standard Wiener process. We remark that their proof has some typos (see Remark 11 in Appendix B). In particular, their result is correct only if \(F = 0\), namely that the impact of parameter estimation uncertainty vanishes altogether. In this case, \(\hat{Z}\) and \(Z\) differ only by a scaling factor: \(\hat{Z}(\xi) \equiv [(1 - \xi) v_{ff}]^{-1/2} Z(\xi)\). If \(F \neq 0\), uncertainty resulting from estimation must be explicitly accounted for, and critical values derived from the distribution of \(\hat{Z}\) will not make the test correctly sized. Recall that the nuisance parameters \(F\) and \(B\) are in fact the key features differentiating the framework of West (1996) from the
1.4. A modified sup-type test

We can combine results of Section 1.2 and Section 1.3 to develop a sup-type test whose critical values can be calculated analytically, thereby circumventing the use of Monte Carlo as in Rossi and Inoue (2012). More importantly, the modified sup-type test will be asymptotically sized at the nominal level whether or not the impact of estimation uncertainty vanishes, while, as we mentioned in Remark 7, the Monte Carlo critical values provided by Rossi and Inoue (2012) are justified only when estimation uncertainty becomes immaterial.

Throughout this section we maintain the null hypothesis of equal predictive ability $H_0 : \mu_t = 0$, so $S_T$ (i.e., the process associated with the test statistic) coincides with $Z_T$ defined in (1.9).

Nonetheless, when the effect of parameter estimation uncertainty is small, these critical values may still provide good approximations, even if they have not been adjusted for the estimation uncertainty.

---

Remark 8. Inspired by Hansen and Timmermann (2015), there is a another angle to view the case where $F = 0$. Consider

$$\tilde{Z}_T(\xi) \equiv T^{-1/2} \sum_{t=1}^T (f_{t+h}(\hat{\beta}_T) - \mu_t) - T^{-1/2} \sum_{t=1}^{[T\xi]} (f_{t+h}(\hat{\beta}_{[T\xi]}) - \mu_t)$$

which is the difference between two “in-sample” test statistics. More explicitly, $\hat{\beta}_T$ is the estimator computed with the full sample, while $\hat{\beta}_{[T\xi]}$ is the estimator computed with the smaller sample. By arguments similar to those advanced for Proposition 6, it can be shown that $\tilde{Z}_T$ converges weakly to a limit process $\tilde{Z}$, where $\tilde{Z}(\xi) \equiv X^{(1)}(1) - X^{(1)}(\xi)$ with $X^{(1)}$ defined in Proposition 6. The weak convergence holds irrespective of $F$, and it holds in particular when $F \neq 0$. On the other hand, the weak limit of $\tilde{Z}$ coincides with that of $Z$ when $F = 0$; see Equation (1.12). In other words, the case where estimation uncertainty vanishes (i.e., $F = 0$) is asymptotically equivalent to the case where one calculates the difference of two in-sample test statistics.

---

one of Diebold and Mariano (1995). Their presence has two important implications for our analysis. First, because the law of the process $Z$ hinges on the nuisance matrices $F$ and $B$, it is not possible to provide a single table of critical values by directly approximating the process via Monte Carlo; only ad hoc (approximate) critical values under specific $F$ and $B$ can be given. Therefore, the approximate critical values listed in Rossi and Inoue (2012) are valid only for the case $F = 0^{21}$. Second, we shall show that computation of the asymptotic power of the sup-type test is essentially the same as computation of boundary crossing probability (BCP) of Wiener process. When $F = 0$, the boundary is linear and we can resort to some well-known formulae for BCP. When $F \neq 0$, the boundary is nonlinear and, as far as we know, exact formula for BCP is still an open question.

---

21 Nonetheless, when the effect of parameter estimation uncertainty is small, these critical values may still provide good approximations, even if they have not been adjusted for the estimation uncertainty.
1.4.1 The pitfall of developing a pivotal statistic

Before delving into the modified sup-type test, let us explain why we do not follow the usual route and develop an asymptotically pivotal sup-type test statistic. Under the null hypothesis of equal predictive ability, the test statistic \( S_T \) coincides with \( Z_T \), and by Proposition 6 we can apply a time-change to \( S_T \) to make the resulting limit process a standard Brownian motion, hence free from the nuisance parameters \( F \) and \( B \). Let \( \nu \) denote the time-change, so that \( S_T \circ \nu \) converges to a process that involves neither \( F \) nor \( B \). But now the nuisance parameters \( F \) and \( B \) are passed from the limit distribution to the time-change \( \nu \); we may write it more explicitly as \( \nu(F, B) \) to highlight the fact. To obtain a feasible test, we first need to plug in estimates of \( F \) and \( B \), say \( \hat{F}_T \) and \( \hat{B}_T \), to obtain an estimate \( \hat{\nu}_T \equiv \nu(\hat{F}_T, \hat{B}_T) \) of the time-change \( \nu \), and then work on \( S_T \circ \hat{\nu}_T \). This is the usual approach to constructing an asymptotically pivotal test statistic.

The main pitfall of this approach lies in proof of the convergence of \( S_T \circ \hat{\nu}_T \). We might need to show that \( \hat{\nu}_T \) converges uniformly to \( \nu \), which turns out to be the same as uniform convergence of the covariance estimator. This is indeed the high-level assumption imposed by Rossi and Inoue (2012) (see their Appendix A), and in general this assumption is hard to verify.

The modified sup-type test circumvents the use of time-change and hence gets rid of the assumption of (or the proof for) uniform convergence of the covariance estimator. As we shall demonstrate soon, only pointwise convergence of the covariance estimator is required. The trick is to use a result due to Rao (1962), which states that weak convergence of probability measures on Euclidean space entails uniform convergence of the measures on certain class of sets.

1.4.2 Analytic formulae for computing the critical values

Under the null hypothesis of equal predictive accuracy, the two-sided sup-mode of sample split leads to the convergence

\[
d_T(R_T) = \sup_{l \leq \xi \leq u} |S_T(\xi)| \overset{d}{=} \sup_{l \leq \xi \leq u} |Z(\xi)|
\]

where \( Z \) is defined in Proposition 6. Convergence of the one-sided sup-type test statistic can also be obtained by the continuous mapping theorem. Cumulative distribution function of \( \sup_{l \leq \xi \leq u} |Z(\xi)| \) can be derived as soon as we observe that time-reserval

\[
L(\xi) \equiv Z(1 - \xi), \quad 0 \leq \xi \leq 1 - l
\]  

\[22\text{The notation } \circ \text{ means composition, i.e., } S_T \circ \nu(t) = S_T(\nu(t)).\]
1.4. A modified sup-type test

is a zero-mean time-changed Wiener process, a fact we prove in Lemma B13 in the appendix. Based on this fact, we can apply time-change to $L$ to obtain a standard Wiener process, and then we just need to invoke the well-known formulae for distribution of the running supremum. The following proposition gives formulae for computing critical values of the one-sided and the two-sided sup-type tests.

**Proposition 7.** Let $\Phi(\cdot; \mu, \sigma^2)$ denote the cumulative distribution function of the normal distribution with mean $\mu$ and variance $\sigma^2$. When $\mu = 0$ and $\sigma^2 = 1$, we write $\Phi(\cdot)$ for short. Denote by $p(\cdot; \sigma^2)$ the probability density function of the normal distribution with zero mean and variance $\sigma^2$. Under Assumptions A and B, we have the following formulae:

\[
\mathbb{P}\left(\sup_{l \leq \xi \leq u} Z(\xi) \leq z\right) = \int_{-\infty}^{z} \Phi\left(z - w; V(l) - V(u)\right) p(w; V(u)) dw
\]

(1.16)

\[
\mathbb{P}\left(\sup_{l \leq \xi \leq u} |Z(\xi)| \leq z\right) = \sum_{n=-\infty}^{\infty} \int_{-z}^{z} q_n(z, w) p(w; V(u)) dw
\]

(1.17)

where

\[
q_n(z, w) = \Phi(z - w; -4nz, V(l) - V(u)) - \Phi(-z - w; -4nz, V(l) - V(u))
\]

\[
- \Phi(z - w; 4nz - 2z - 2w, V(l) - V(u)) + \Phi(-z - w; 4nz - 2z - 2w, V(l) - V(u))
\]

(1.18)

and the variance process $\xi \mapsto V(\xi)$ is defined in (1.13).

### 1.4.3 The modified sup-type test

The formulae provided in Proposition 7 are not immediately usable, as the variances $V(l)$ and $V(u)$ involve the nuisance parameters $F$ and $B$ (see (1.13)). The quantity $d_T(R_T)$ is thus not asymptotically pivotal. But it is not difficult to develop test procedures that are asymptotically sized at the claimed level.

To save space we consider only the one-sided sup-type test, but the same idea readily carries over to the two-sided test. Throughout we fix the significance level at $\alpha$. The test procedures are summarized in Table 1.1.

To justify the test procedures means to prove that the test is asymptotically sized at the level $\alpha$, that is,

\[
\lim_{T} \mathbb{P}\left(\sup_{[lT] \leq r \leq [uT]} d_T(r) > \zeta(\hat{V}_T(l), \hat{V}_T(u))\right) = \alpha
\]

(1.19)

To this end, define two (random) sets

\[
A_T \equiv \left(\zeta(\hat{V}_T(l), \hat{V}_T(u)), \infty\right), \quad A \equiv \left(\zeta(V(l), V(u)), \infty\right)
\]

(1.20)
Table 1.1: Procedures of the one-sided sup-type test

Step 1: Obtain consistent estimators \( \hat{V}_T(l) \) and \( \hat{V}_T(u) \) for \( V(l) \) and \( V(u) \) respectively by plugging consistent estimates\(^a\) of \( F \) and \( B \) into (1.13).

Step 2: Define

\[
h(z; \hat{V}_T(l), \hat{V}_T(u)) = \int_{-\infty}^{z} \Phi \left( \frac{z - w}{\hat{V}_T(l) - \hat{V}_T(u)} \right) p(w; \hat{V}_T(u)) \, dw - (1 - \alpha)
\]

Find the zero\(^b\) of \( z \mapsto h(z; \hat{V}_T(l), \hat{V}_T(u)) \), which we denote by \( \zeta(\hat{V}_T(l), \hat{V}_T(u)) \).

Step 3: Reject the null hypothesis if and only if \( \sup_{l \leq r \leq u} d_T(r) > \zeta(\hat{V}_T(l), \hat{V}_T(u)) \).

\(^a\) McCracken (2000) provides guidance on how to obtain the consistent estimates.

\(^b\) The zero is a singleton as the integrand is strictly positive.

Equation (1.19) is equivalent to

\[
\lim_T \mathbb{P} \left( \sup_{l \leq \xi \leq u} Z_T(\xi) \in A_T \right) = \mathbb{P} \left( \sup_{l \leq \xi \leq u} Z(\xi) \in A \right)
\]  

(1.21)

**Proposition 8.** Given that \( \hat{V}_T(l) \) and \( \hat{V}_T(u) \) consistently estimate \( V(l) \) and \( V(u) \) respectively, the one-sided sup-type test described in Table 1.1 is asymptotically sized at the \( \alpha \) level, that is, equality (1.21) holds.

The proof utilizes a nice property of weakly convergent sequence of measures on Euclidean space: if the limit measure is absolutely continuous with respect to the Lebesgue measure, then the sequence converges uniformly on the class of measurable convex sets. This result was first established by Rao (1962). See also Theorem 8.2.18 in Bogachev (2007). Because of this property, Proposition 8 requires only pointwise consistency of the estimators \( \hat{V}_T(l) \) and \( \hat{V}_T(u) \) (cf. Appendix A of Rossi and Inoue (2012)).

### 1.5 Power under local alternatives

This section aims to cast light on the question inspired by Rossi and Inoue (2012): is it worthwhile to develop novel out-of-sample tests by combining data-mining split and its treatment? This question is related to power of the test: such combination is appealing only if it renders the test more powerful than the conventional one using fixed split point. The study of Rossi and Inoue (2012) seems to favour a positive answer. They argue that the conventional test using fixed split point \( R_T \) might lack empirical evidence in favor of unequal predictive ability – especially when there prevail unequal predictive abilities only
over a portion of the sample – because the window size is either too small or too large to capture it (p. 434). But in the meanwhile they also acknowledge that the novel tests are subject to counteracting forces and so it is unclear whether these tests will “find more or less empirical evidence in favour of predictive ability” (p. 446).

To address this question we inspect power of West’s test using fixed split point and the sup-type test proposed by Rossi and Inoue (2012) under local alternatives. Rossi and Inoue (2012) also propose an average-type test, but we will not consider this test here as it tends to have low power (see the discussion in Section 1.7). The local alternatives we consider involve either Pitman drift or one-time structural break. Finally, closed-form results are obtained only for the case where the impact of estimation uncertainty vanishes.

1.5.1 Limit process under local alternatives

To derive the limit process of \( S_T(\xi) \equiv d_T(\lfloor \xi T \rfloor) \) we impose an additional assumption.

**Assumption C.** As \( T \to \infty, \frac{1}{\sqrt{T}} \sum_{t=\lfloor \xi T \rfloor}^{T} \mu_t \) converges in the Skorohod topology on \( D[l, 1] \) to a càdlàg function \( C(\xi) \). In particular, we consider two types of local alternatives:

1. Local alternative with Pitman drift: \( \mu_t \equiv \frac{c}{\sqrt{T}}, \) where \( c \) is a nonnegative constant. In this case \( C(\xi) \equiv c(1 - \xi) \).

2. Local alternative with single structural break: \( \mu_t \equiv \frac{c}{\sqrt{T}} I(t \geq \lfloor \lambda T \rfloor), \) where \( c \) and \( \lambda \) are constant scalars with \( c \) being nonnegative. We further assume that \( l < \lambda < u \), that is, the structural break occurs on the interval which is mined over. In this case \( C(\xi) \equiv c(1 - \xi \lor \lambda) \).

**Remark 9.** Assumption C2 is similar to the local alternative analyzed by Clark and McCracken (2005). Under this assumption the competing models exhibit unequal predictive abilities after the structural break if \( c \neq 0 \). The case where the models once possessed unequal predictive abilities before the structural break can be studied in the same vein and will not be considered here. Nor will we consider local alternatives with multiple structural break, but the analysis is quite similar to the one with single structural break.

**Proposition 9.** Let Assumptions A, B and C be satisfied. The limit process is \( S(\xi) \equiv Z(\xi) + C(\xi) \) under the local alternatives, with \( Z \) defined by (1.12) and the function \( C \) defined in assump-

---

23 We remark that the formulae derived for the case where the impact of estimation uncertainty vanishes are applicable in framework of Diebold and Mariano (1995). As noted by Diebold (2015), the primitives of the Diebold-Mariano framework are the loss differentials, and the Diebold-Mariano test is essentially imposing assumptions on the loss differentials that would entail validity of some central limit theorem. Certainly, we could impose assumptions that makes applicable suitable functional central limit theorem, so that the Diebold-Mariano test statistic converges weakly to a standard Wiener process. The formulae we derive in this section then apply readily.
CHAPTER 1. DATA-DRIVEN SAMPLE SPLIT AND OUT-OF-SAMPLE TESTS

Moreover,

\[
P\left(\sup_{l \leq \xi \leq u} S(\xi) > x\right) = P\left(\sup_{v(1-u) \leq \xi \leq v(1-l)} \left(W(\xi) + C(1 - v^{-1}(\xi))\right) > x\right)
\]

\[
P\left(\sup_{l \leq \xi \leq u} |S(\xi)| > x\right) = P\left(\sup_{v(1-u) \leq \xi \leq v(1-l)} \left|W(\xi) + C(1 - v^{-1}(\xi))\right| > x\right)
\]

where \(W\) is an one-dimensional standard Wiener process, \(v(\xi) \equiv V(1 - \xi)\) with \(V\) defined by (1.13), and \(v^{-1}\) denotes the inverse of \(v\).

Proposition 9 shows that computation of asymptotic power of the sup-type test under local alternatives is essentially the same as computation of the probability that a Wiener process crosses the boundary \(\xi \rightarrow C(1 - v^{-1}(\xi))\). The equations (B30) and (B31) are used to calculate asymptotic power for one-sided and two-sided sup-type tests respectively. Shape of the boundary depends on both \(v\) and \(C\). When \(F = 0\), \(v\) is a linear function and so is \(v^{-1}\).

In this case, the boundary is linear under local alternative with Pitman drift and piecewise linear under local alternative with structural break. When \(F \neq 0\), \(v\) is nonlinear, and \(C\) will be nonlinear as well. Crossing probability of linear or piecewise linear boundary has explicit formula; see, e.g., Anderson (1960) (p. 180, Theorem 4.3) and Scheike (1992) (p. 450, Equation (2)). On the other hand, we are not aware of explicit formula for crossing probability of general nonlinear boundary.

1.5.2 The case where the impact of estimation uncertainty vanishes

When \(F = 0\), variance of the process \(Z\) becomes (see (1.13))

\[V(\xi) = (1 - \xi) v_{ff}\]

which is linear function of \(\xi\). Such linearity turns out to be crucial for obtaining closed form formula for the asymptotic power.

**Proposition 10.** Let Assumptions A, B and C be satisfied. Denote by \(p(w; \sigma^2)\) the density of the normal distribution with zero-mean and variance \(\sigma^2\), and by \(\Phi(\cdot)\) the c.d.f. of standard normal distribution. Consider the local alternative with Pitman drift \(\mu_t \equiv \frac{c}{\sqrt{T}}\). Asymptotic power of the conventional test and sup-type test can be calculated with the following formulae:

---

24 Boundary crossing probability (BCP) has long been applied in statistics and econometrics. It originates perhaps from analysis of the Kolmogorov-Smirnov statistic. Interest in such problems is further stimulated by development of sequential analysis, especially in design of efficient clinical trials. For statistical applications of BCP, see, e.g., Siegmund (1986) and the monograph of Siegmund (2013). In econometrics, BCP is widely applied in analysis of the cusum test. See, e.g., Brown et al. (1975) (p. 153), Krämer et al. (1988) (p. 1357), Ploberger and Krämer (1990), Ploberger and Krämer (1992) (p. 276) and Chu et al. (1996).

25 Computation of nonlinear boundary crossing probability is usually done by numerical approximations and Monte Carlo; see, e.g., Wang and Pötzelberger (1997) and Pötzelberger and Wang (2001).
1.5. Power under local alternatives

(a) One-sided: Denote by $P_1(a, b, t)$ the probability that $W$ crosses the boundary $a + bt$ before or at $t$, where $a > 0$. For the conventional test with fixed split point $R_T \equiv \lfloor lT \rfloor$,

$$P(S(l) > x) = \Phi \left( \frac{-x + c(1 - l)}{\sqrt{v_{ff}(1 - l)}} \right)$$

(1.24)

For the sup-type test that mines over $[\lfloor lT \rfloor, \lfloor uT \rfloor]$,

$$P\left( \sup_{l \leq \zeta \leq u} S(\zeta) > x \right)$$

$$= 1 - \int_{-\infty}^{x - c(1 - u)} p(w; v_{ff}(1 - u)) \left[ 1 - P_1\left( a(w), -\frac{c}{v_{ff}}, v_{ff}(u - l) \right) \right] \, dw$$

where $a(w) \equiv x - c(1 - u) - w$. Formula for $P_1(a, b, t)$ is given in Lemma B16 of the appendix.

(b) Two-sided: For two linear boundaries $a_1 + bt$ and $a_2 + bt$, where $a_1 > 0 > a_2$, denote by $P_2(a_1, a_2, b, t)$ the probability that $W$ crosses either of the boundaries before or at $t$. For the conventional test,

$$P(|S(l)| > x) = \Phi \left( \frac{-x + c(1 - l)}{\sqrt{v_{ff}(1 - l)}} \right) + \Phi \left( \frac{-x - c(1 - l)}{\sqrt{v_{ff}(1 - l)}} \right)$$

(1.26)

while for the sup-type test,

$$P\left( \sup_{l \leq \zeta \leq u} |S(\zeta)| > x \right)$$

$$= 1 - \int_{-x - c(1 - u)}^{x - c(1 - u)} p(w; v_{ff}(1 - u)) \left[ 1 - P_2\left( a_1(w), a_2(w), -\frac{c}{v_{ff}}, v_{ff}(u - l) \right) \right]$$

(1.27)

where $a_1(w) \equiv x - c(1 - u) - w$ and $a_2(w) \equiv -x - c(1 - u) - w$. Formula for $P_2(a_1, a_2, b, t)$ is given in Lemma B17 of the appendix.

Proposition 10 enables us to compare asymptotic power of the conventional test using fixed split point and the sup-type test under the local alternative with Pitman drift. For the conventional test, the noncentrality parameter $c(1 - l)$ grows when $l$ decreases (since we assume $c \geq 0$), which suggests that, ceteris paribus, the test becomes asymptotically more powerful as the hold-out sample size extends; equivalently, the test loses power as the hold-out sample size gets shorter\(^{26}\). We consider the conventional test with fixed split point $R_T \equiv 26$

\(^{26}\)The finding is indeed in consonance with the one by Hansen and Timmermann (2015). They derive asymptotic distribution of the test statistic for nested models under the local alternative with Pitman drift as well, and find that the non-centrality parameter, which is crucial for detecting unequal predictive abilities, decreases as the hold-out sample gets shorter.
Figure 1.1: Asymptotic power under local alternative with Pitman drift \((F = 0)\)

Note: This figure plots asymptotic power of the conventional test with fixed sample split point \(R_T \equiv \lfloor IT \rfloor\) (the blue surface) and the sup-type test that mines over \([\lfloor IT \rfloor, \lceil uT \rceil]\) (the yellow surface), where \(u = 1 - l\). The power is calculated under the condition \(F = 0\) and the local alternative with Pitman drift \(\mu_t \equiv \frac{c}{v_{ff}}, c\) being a constant scalar. Panels on the left: one-sided tests; Panels on the right: the two-sided tests. Analytic formulae for the local power are given in Proposition 10.

\([IT]\). On the other hand, for the sup-type test we suppose that the mining is conducted over \([l, u]\), where \(u = 1 - l\). Figure 1.1 contrasts asymptotic power of the conventional test and the sup-type test, where we let \(c, l\) and \(v_{ff}\) vary. The three panels on the first column correspond to the one-sided tests and the three panels on the right to the two-sided tests. We observe that the conventional test (i.e., the blue surface) is more powerful than the sup-type test (the yellow surface) in all cases, although gap between the two surfaces is in general very small.

The above finding suggests that the conventional test using fixed small split point is actually more powerful than, and hence preferable to, the sup-type test under the local alternative with Pitman drift. But would the sup-type test become more powerful in the presence of structural break? To compute the power we establish the following proposition by means of the same technique used in proving Proposition 10.
Proposition 11. Let Assumptions A, B and C be satisfied. Denote by \( p(w; \sigma^2) \) the density of the normal distribution with zero-mean and variance \( \sigma^2 \), and by \( \Phi(\cdot) \) the c.d.f. of standard normal distribution. Consider the local alternative with one-time structural break \( \mu_t \equiv \frac{c}{\sqrt{T}} 1(t \geq [\lambda T]) \). Asymptotic power of the conventional test and the sup-type test can be calculated with the following formulae:

(a) One-sided: For the conventional test with fixed split point \( R = [IT] \),

\[
\Pr (S(l) > x) = \Phi \left( \frac{-x + c(1 - \lambda)}{\sqrt{v_{ff}(1 - l)}} \right)
\]

while for the sup-type test that mines over \([IT], [uT]\),

\[
\Pr \left( \sup_{l \leq \xi \leq u} S(\xi) > x \right) = 1 - \int_{-\infty}^{x - c(1 - \lambda) - \tilde{w}} \int_{-\infty}^{x - c(1 - \lambda) - \tilde{w}} p(\tilde{w}; v_{ff}(1 - u)) p(w; v_{ff}(u - \lambda)) \times \left[ 1 - \exp \left( -\frac{2(x - c(1 - \lambda) - \tilde{w} - w)(x - c(1 - \lambda) - \tilde{w} - w)}{v_{ff}(u - \lambda)} \right) \right] \frac{2\Phi \left( \frac{x - c(1 - \lambda) - \tilde{w} - w}{\sqrt{v_{ff}(1 - l)}} \right)}{dwd\tilde{w}} \tag{1.29}
\]

(b) Two-sided: For two linear boundaries \( a_1 + bt \) and \( a_2 + bt \), where \( a_1 > 0 > a_2 \), denote by \( P^2(a_1, a_2, b, t) \) the probability that \( W \) crosses either of the boundaries before or at \( t \). Furthermore, let \( P^{(x)}(a_1, a_2, b, t) \) be the conditional probability that \( W \) crosses either of the boundaries before or at \( t \), conditional on the terminal value \( W(t) = x \). For the conventional test with fixed split point \( R = [IT] \),

\[
\Pr (|S(l)| > x) = \Phi \left( \frac{-x + c(1 - \lambda)}{\sqrt{v_{ff}(1 - l)}} \right) + \Phi \left( \frac{-x - c(1 - \lambda)}{\sqrt{v_{ff}(1 - l)}} \right)
\]

while for the sup-type test that mines over \([IT], [uT]\),

\[
\Pr \left( \sup_{l \leq \xi \leq u} |S(\xi)| > x \right)
= 1 - \int_{x - c(1 - u)}^{x - c(1 - \lambda) - \tilde{w}} \int_{x - c(1 - \lambda) - \tilde{w}}^{x - c(1 - \lambda) - \tilde{w}} p(\tilde{w}; v_{ff}(1 - u)) p(w; v_{ff}(u - \lambda)) \times P_{2}^{(w)}(\tilde{a}_1(\tilde{w}), \tilde{a}_2(\tilde{w}), -\frac{c}{v_{ff}}, v_{ff}(u - \lambda)) \times \left[ 1 - P_2(a_1(w + \tilde{w}), a_2(w + \tilde{w}), 0, v_{ff}(\lambda - l)) \right] d\tilde{w}d\tilde{w}
\]

where \( \tilde{a}_1(\tilde{w}) = x - c(1 - u) - \tilde{w}, \tilde{a}_2(\tilde{w}) = -x - c(1 - u) - \tilde{w}, a_1(w) = x - c(1 - \lambda) - w \) and \( a_2(w) = -x - c(1 - u) - w \). Formulae for \( P^{(w)}(a_1, a_2, b, t) \) and \( P(a_1, a_2, b, t) \) are given in
Lemma B18 and Lemma B19 respectively.

Remark 10. The idea for establishing these lengthy formulae is just to apply the Markov property repeatedly and the (conditional) boundary crossing probabilities. At the cost of higher numerical complexity, we can also obtain formulae for asymptotic power in the presence of multiple structural break, which is not pursued in this paper. ■

Again we compare asymptotic power of the conventional test using fixed split point \( R_T \equiv \lfloor lT \rfloor \) and the sup-type test mining over \( \lfloor lT \rfloor, \lfloor uT \rfloor \). We consider two settings: in the first setting, we fix the break time \( \lambda \equiv 0.75l + 0.25u \) and let \( c, l \) and \( v_{ff} \) vary; in the second setting, we fix \( v_{ff} \equiv 1 \) and let \( c, l \) and \( \delta \equiv \frac{u-\lambda}{u-l} \) vary. Figure 1.2 visualizes power of the two tests in the first setting. In contrast to the case under the local alternative with Pitman drift, the sup-type test (the yellow surface) becomes more powerful than the conventional test (the blue surface) in the presence of structural break. Gap between the two surfaces gets wider as the constant \( c \) gets larger. Take the one-sided tests (the left column) for example. When \( c = 1 \) the biggest difference in power between the sup-type test and the conventional test is about 1%. When \( c = 3 \) or \( c = 5 \), the biggest difference becomes about 5.6%. The gain in power achieved by data-mining becomes more pronounced in the second setting, which is visualized in Figure 1.3. As in the first setting, the sup-type test is more powerful than the conventional test. The difference in power is most conspicuous in the two panels at the bottom, where the constant \( c = 5 \). For example, when \( l = 0.15 \) and \( \frac{u-\lambda}{u-l} = 0.2 \), the two-sided sup-type test rejects the null of equal predictive accuracy about 15.57% more often than the conventional test. These findings suggest that the sup-type test becomes remarkably more powerful than the conventional test using fixed small split point when there is a pronounced structural change towards unequal predictive abilities occurring late in the sample. Finally, it is worthwhile to point out that in all cases both tests tend to be more powerful when \( \lambda \) approaches \( l \) (i.e., when \( \frac{u-\lambda}{u-l} \) approaches 1). This is expected, since the closer is \( \lambda \) to \( l \), the more post-break observations evidencing unequal predictive accuracy, and hence the more powerful are the tests.

To sum, we draw two findings from the analysis under local alternatives. When the local alternative is "smooth", i.e., with Pitman drift, the conventional test with fixed small split point is asymptotically more powerful, so there is nothing to gain from conducting the sup-type test. On the other hand, when the local alternative involves structural break, our analysis corroborates the claim of Rossi and Inoue (2012) that the sup-type test can be more powerful than the conventional test.
1.5. Power under local alternatives

Figure 1.2: Asymptotic power under local alternative with one-time structural break ($F = 0$)

Note: This figure plots asymptotic power of the conventional test with fixed sample split point $R_T \equiv \lfloor lT \rfloor$ (the blue surface) and the sup-type test that mines over $[\lfloor lT \rfloor, \lfloor uT \rfloor]$ (the yellow surface), where $u = 1 - l$. The power is calculated under the condition $F = 0$ and the local alternative with break $\mu_t \equiv \frac{c}{2T}1(t \geq \lfloor \lambda T \rfloor)$, $c$ being a constant scalar and $\lambda \equiv 0.75l + 0.25u$. Panels on the left: one-sided tests; Panels on the right: the two-sided tests. Analytic formulae for computing the power are given in Proposition 11.
Figure 1.3: Asymptotic power under local alternative with one-time structural break ($F = 0$)

Note: This figure plots asymptotic power of the conventional test with fixed sample split point $R_T \equiv \lfloor IT \rfloor$ (the blue surface) and the sup-type test that mines over $[\lfloor IT \rfloor, \lfloor uT \rfloor]$ (the yellow surface), where $u = 1 - l$. The power is calculated under the condition $F = 0$ and the local alternative with break $\mu_t \equiv \frac{c}{\sqrt{T}}1(t \geq \lfloor AT \rfloor)$, $c$ being a constant scalar. $v_{ff} \equiv 1$ is fixed and $\lambda$ and $l$ are allowed to vary. In particular, $\delta \equiv \frac{u - \lambda u - l}{u - l}$ varies from 0.1 to 0.9. Panels on the left: one-sided tests; Panels on the right: the two-sided tests. Analytic formulae for computing the power are given in Proposition 11.
1.5.3 The case where the impact of estimation uncertainty does not vanish

All analytic formulae derived in Section 1.5.2 hinge on the critical assumption $F = 0$, which means that the impact of estimation uncertainty vanishes. Generalization to the case where $F \neq 0$ is not as trivial as it might appear. The main pitfall is nonlinearity of the variance process (see (1.13)), which causes computation of power of the sup-type test to involve crossing probability of a nonlinear boundary. Moreover, Proposition 9 shows that the crossing probability involves the inverse function of $v(\xi) \equiv V(1 - \xi)$, which does not seem to admit a closed-form expression. We are not aware of any general formula for nonlinear boundary crossing probability, and more thorough analysis of the general case is thus left for future study.

1.6 Monte Carlo Experiments

This section presents two small Monte Carlo experiments. The first aims to illustrate the problem of size distortion caused by different modes of sample split and the cost of power in achieving robustness. The second aims to cast light on whether combining data-driven sample split and its treatment renders the test more powerful than the conventional test using fixed split point.

1.6.1 Size results

As of size, it would be more interesting to ape a researcher’s mining behaviour. Conventionally, for comparison of nonnested forecast models researchers use the Diebold-Mariano-West test statistic (cf. (1.9))

$$\tilde{d}_T(R_T) = T^{-1/2} \tilde{V}(\xi)^{-1/2} \sum_{t=[\xi T]}^T f_{t+h}(\hat{\beta}_t), \text{ where } R_T \equiv [\xi T]$$

If the researcher’s choice of sample split point $R_T$ is fixed ex ante, $\tilde{d}_T(R_T)$ is asymptotically standard normal. By Proposition 1, convergence to the standard normal distribution still holds if the choice of $R_T$ is data-independent in the sense of definition 1.

We use a plain vanilla example to illustrate the problems of size distortion and loss of power in achieving robustness. The DGP is:

$$y_{t+1} = v_t + w_t + \epsilon_{t+1}, \quad v_t \overset{i.i.d.}{\sim} N(0, \sigma_v^2), \quad w_t \overset{i.i.d.}{\sim} N(0, \sigma_w^2), \quad \epsilon_t \overset{i.i.d.}{\sim} N(0, \sigma_\epsilon^2)$$

(1.33)
The forecast models being compared are

\begin{align*}
\text{Model 1: } y_{t+1} &= \alpha v_t + \eta_{1,t} \\
\text{Model 2: } y_{t+1} &= \beta w_t + \eta_{2,t}
\end{align*}

The coefficients are estimated by OLS with increasing window. We consider five split modes:

M1. Halve the initial sample, that is, set \( R_T = \frac{1}{2} T \);

M2. \( R_T \) is “arbitrarily” drawn as a random number uniformly distributed over \([\lfloor l T \rfloor, \lfloor u T \rfloor]\);

M3. Start from \( R_T = \lfloor l T \rfloor \) and stop searching once \(| \tilde{d}_T(R_T) | \) passes \( c = 2.5 \); otherwise set \( R_T = \lfloor l T \rfloor \);

M4. Start from \( R_T = \lfloor l T \rfloor \) and stop searching once \(| \tilde{d}_T(R_T) | \) passes \( c = 2.25 \); otherwise set \( R_T = \lfloor u T \rfloor \);

M5. Start from \( R_T = \lfloor l T \rfloor \) and stop searching once \(| \tilde{d}_T(R_T) | \) passes \( c = 2 \); otherwise set \( R_T = \lfloor u T \rfloor \);

M6. \( R_T \equiv \arg\max_{\lfloor l T \rfloor \leq r \leq \lfloor u T \rfloor} | \tilde{d}_T(r) | \).

For the experiment, we fix \( T = 1000 \) and \( \sigma_\epsilon = \frac{1}{4} \sigma_v \). We let \( \sigma_v = \sigma_w \) change from 0.1 to 1 with step size 1. Due to symmetry of \( v_t \) and \( w_t \) in the DGP (1.33), the two models possess equal predictive accuracy precisely when \( \sigma_v = \sigma_w \). In total we simulate 5,000 samples.

Now what if we fail to adjust for the uncertainty resulting from sample split and wrongly use the conventional critical values, i.e., the quantiles of the standard normal distribution? Figure 1.4 plots the rejection rates under the six modes of sample split, with nominal size being 0.05. The first observation is that data-independent sample split, either having \( R_T = \frac{1}{2} T \) or drawing it randomly from \([\lfloor l T \rfloor, \lfloor u T \rfloor]\), does not have any adverse effect on the size. On the contrary, the behavior of maximizing the test statistic results in severe size distortion. When \( l = 0.15 \) and \( u = 0.85 \), for example, the rejection rate is elevated above 25%. We can also see how the size can be distorted artificially by the pass-mode split. Again take the case where \( l = 1 - u = 0.15 \) for example. Setting the threshold to \( c = 2.5 \) renders the true size about 5% larger than the nominal level. The true size can be further lifted to 15% if we put \( c = 2.25 \).

The most subtle problem emerges when we set \( c = 2 \), which is slightly above the conventional critical value 1.96. On the one hand, Figure 1.4 shows that the associated pass-mode elevates the true size to almost the same level as the sup-mode. On the other hand, the treatments for the pass-mode and the sup-mode are very different. To illustrate, in Figure 1.5 we plot the tail probabilities of the limit distributions of \( \tilde{d}_T(R_T) \) under different modes of sample split, that is,

\[
\lim_{T \to \infty} \mathbb{P}(|\tilde{d}_T(R_T)| > x)
\]

The probabilities are approximated by Monte Carlo. Tail probability of the sup-mode envelopes tail probabilities of other modes. In this light, the sup-mode is like an “upper bound”

\footnote{Note that \( \tilde{d}_T(R_T) \) defined in (1.32) differs from \( d_T(R_T) \) defined in (1.7), because of the presence of \( \tilde{V}(\xi) \).}
Figure 1.4: Plain vanilla DGP with uncorrected critical values

$l = 0.05, u = 0.95$

$l = 0.10, u = 0.90$

$l = 0.15, u = 0.85$

$l = 0.20, u = 0.80$

$l = 0.25, u = 0.75$

$l = 0.30, u = 0.70$

Note: This figure plots the rejection rates with the plain vanilla DGP (1.33) and the conventional critical value 1.96. We set $T = 1000$ and $\sigma_e = \frac{1}{4}\sigma_w$, and let $\sigma_v = \sigma_w$ increase from 0.1 to 1 with step size 0.1. The data-mining is conducted on $[[IT], [uT]]$, where $u \equiv 1 - l$. In total we have simulated 5,000 samples. The modes of sample split are: M1: halve the total sample (i.e., $R_T \equiv \frac{1}{2}T$); M2: draw $R_T$ as a random variable uniformly distributed over $[[IT], [uT]]$; M3 to M5: pass-mode with thresholds 2.5, 2.25 and 2 respectively; M6: sup-mode.
for size distortion in out-of-sample analysis. Now, the subtlety is that the pass-mode with threshold $c = 2$ results in almost the same degree of size distortion as the sup-mode – see crosses of the curves with the vertical reference line – but to make the test correctly sized we need to pick a critical value which is much smaller than that required to make the sup-mode correctly sized. When $l = 0.15$ and $u = 0.85$, for example, to make the test sized at the 0.05 level in the presence of the sup-mode split, the critical value should be about 2.75. But in the presence of the pass-mode with threshold $c = 2$, the critical value should be about 2.05.

### 1.6.2 Loss of power in achieving robustness

Continuing with the plain vanilla example, we illustrate the potential loss of power in achieving “robustness” to data-driven sample split. Suppose that we conservatively use treatments for the sup-mode split regardless of the true underlying modes of sample split. This could incur considerable loss of power if what’s in operation is the pass-mode. Figure 1.6 plots the rejection rates of the test when the split point $R_T$ is determined by the pass-mode with threshold $c = 2$. We try two critical values 2.75, the one that makes the test correctly sized in the presence of the sup-mode split, and 2.2, the one that makes the test correctly sized in the presence of pass-mode split. In every panel the largest difference in power is about 24%, which is the price paid for achieving robustness.

### 1.6.3 Power results

In this Monte Carlo experiment, we study if it is worthwhile to combine data-mining and its treatment to make a novel out-of-sample test. Specifically, we compare power of the (two-sided) modified sup-type test developed in Section 1.4 and the conventional test using fixed split point. To be clear, by conventional test statistic we mean

$$
\hat{S}_T \equiv T^{-1/2} \tilde{V}(I)^{-1/2} \sum_{t=\lfloor lT \rfloor}^{T} f_{t+h}(\hat{\beta}_t),
$$

namely that the fixed split point $R_T \equiv \lfloor lT \rfloor$ is adopted; the modified sup-type test is assumed to mine over $[\lfloor lT \rfloor, \lfloor uT \rfloor]$, where $u = 1 - l$.

We revisit a data generating process, labeled DGP2, in Rossi and Inoue (2012). The DGP is

$$
\begin{bmatrix}
    y_{t+1} \\
    v_{t+1} \\
    w_{t+1}
\end{bmatrix} =
\begin{bmatrix}
    0.3 & d_{t,T} & 0.5 \\
    0 & 0.5 & 0 \\
    0 & 0 & 0.5
\end{bmatrix} \begin{bmatrix}
    y_t \\
    v_t \\
    w_t
\end{bmatrix} +
\begin{bmatrix}
    \varepsilon_{y,t+1} \\
    \varepsilon_{v,t+1} \\
    \varepsilon_{w,t+1}
\end{bmatrix}
$$

where $y_0 = v_0 = w_0 = 0$, and $(\varepsilon_{y,t+1}, \varepsilon_{v,t+1}, \varepsilon_{w,t+1}) \overset{i.i.d.}{\sim} N(0, I_3)$. The variable to be forecast is

We have not derived analytic formulae for the limit distribution of $\tilde{d}_T(R_T)$, so we just use Monte Carlo to approximate it here.
1.6. Monte Carlo Experiments

**Figure 1.5:** Tail probabilities of nonnested forecast comparison

\[ l = 0.05, \; u = 0.95 \]

![Graph showing tail probabilities for different modes of sample split with legend indicating various modes including Step mode, Pass mode with different parameters, and Std Normal critical value (α=0.05).]

\[ l = 0.1, \; u = 0.9 \]

![Graph showing tail probabilities for different modes of sample split with legend indicating various modes including Step mode, Pass mode with different parameters, and Std Normal critical value (α=0.05).]

\[ l = 0.15, \; u = 0.85 \]

![Graph showing tail probabilities for different modes of sample split with legend indicating various modes including Step mode, Pass mode with different parameters, and Std Normal critical value (α=0.05).]

\[ l = 0.2, \; u = 0.8 \]

![Graph showing tail probabilities for different modes of sample split with legend indicating various modes including Step mode, Pass mode with different parameters, and Std Normal critical value (α=0.05).]

\[ l = 0.25, \; u = 0.75 \]

![Graph showing tail probabilities for different modes of sample split with legend indicating various modes including Step mode, Pass mode with different parameters, and Std Normal critical value (α=0.05).]

\[ l = 0.3, \; u = 0.7 \]

![Graph showing tail probabilities for different modes of sample split with legend indicating various modes including Step mode, Pass mode with different parameters, and Std Normal critical value (α=0.05).]

Note: This figure plots the limit tail probabilities of the test statistic (1.32), i.e., \( \lim_{T} P \left( |d_T(R_T)| > x \right) \), under various modes of sample split. If the sample split point \( R_T \) is independent of the data, the test statistic is asymptotically standard normal.
Figure 1.6: Potential loss of power

Note: This figure shows the rejection rates of the plain vanilla example when the sample split point $R_T$ is determined by the pass-mode with threshold 2.
1.6. Monte Carlo Experiments

\( y_{t+1} \). The two nonnested models producing the forecasts at time \( t \) are

Model 1: 
\[
y_{t+1} = y_t \theta_1 + w_t \theta_2 + \eta_{1,t+1}
\]
Model 2: 
\[
y_{t+1} = y_t \gamma_1 + v_t \gamma_2 + \eta_{2,t+1}
\]

The models cannot be estimated by OLS due to endogeneity of \( y_t \), and we resort to 2SLS estimator. For model 1 the instruments are \((w_{t-1}, w_t)\)'r, and for model 2 the instruments are \((v_{t-1}, v_t)'\). Forecast accuracy is gauged by squared error, namely (setting \( \beta \equiv (\theta_1, \theta_2, \gamma_1, \gamma_2)'\))

\[
f_{t+1}(\beta) = (y_{t+1} - y_t \theta_1 - w_t \theta_2)^2 - (y_{t+1} - y_t \gamma_1 - v_t \gamma_2)^2
\]

We make two important notes before proceeding. First, it is easy to check that

\[
F = \mathbb{E}\left[ \frac{\partial f(\beta^*)}{\partial \beta} \right] = \mathbb{E}\left[ \begin{bmatrix} -2d_{t,T} y_t v_t \\ -2d_{t,T} w_t v_t \\ y_t w_t \\ v_t w_t \end{bmatrix} \right] \neq 0
\]

Thus in view of Proposition 6, the estimation uncertainty does not vanish, and \( F \) and \( B \) must be estimated. Details about estimations of \( F \) and \( B \) (and in addition \( \Omega \)) can be found in appendix A. Second, whether the two models have equal predictive abilities depends on \( d_{t,T} \). When \( d_{t,T} = 0.5 \), the two models produce equally accurate forecasts, since \( v_t \) and \( w_t \) play symmetric roles. When \( d_{t,T} > 0.5 \), forecast produced by model 1 becomes more accurate.

In this experiment we set

\[
d_{t,T} \equiv 0.5 + \delta 1(t > \lceil \tau T \rceil)
\]

where \( \tau \) goes from 0.3 to 0.7 with step size 0.1 and \( \delta \) ranges from 0 to 1 with the same step size. The sample size \( T = 600 \), and in total we draw 1,000 samples. \( \delta = 0 \) corresponds to the null hypothesis of equal predictive accuracy. On the other hand, when \( \delta > 0 \), the second model becomes more accurate after the break. We emphasize that this is not the local alternative stated in Assumption C, but fixed alternative. Obviously, under the fixed alternative both the sup-type test and the conventional test are consistent, that is, the probability that the false null gets rejected converges to 1 as \( T \to \infty \).

Rejection rates of the conventional test with fixed split point \( R_T \equiv [0.15T] \) and the modified sup-type test that mines over \([0.15T], [0.85T] \) are listed in Table 1.2. We have colored the cells with greater rejection rates. The nominal significance level is 0.05. At first glance, the conventional test appears more powerful, but we note that it tends to be slightly oversized, whereas the sup-type test is undersized. Taking into account the difference in sizes, it would be fairer to conclude that the match ends in a tie. It is somewhat surprising that the sup-type test does not dominate even though structural break is present. Nonetheless,
Table 1.2: Comparison of power of the conventional test and the two-sided sup-type test

<table>
<thead>
<tr>
<th>$\delta$</th>
<th>$\tau = 0.3$</th>
<th>$\tau = 0.4$</th>
<th>$\tau = 0.5$</th>
<th>$\tau = 0.6$</th>
<th>$\tau = 0.7$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Conventional</td>
<td>Sup-type</td>
<td>Conventional</td>
<td>Sup-type</td>
<td>Conventional</td>
</tr>
<tr>
<td>0</td>
<td>0.058</td>
<td>0.047</td>
<td>0.050</td>
<td>0.046</td>
<td>0.045</td>
</tr>
<tr>
<td>0.1</td>
<td>0.183</td>
<td>0.173</td>
<td>0.147</td>
<td>0.138</td>
<td>0.114</td>
</tr>
<tr>
<td>0.2</td>
<td>0.517</td>
<td>0.498</td>
<td>0.382</td>
<td>0.392</td>
<td>0.265</td>
</tr>
<tr>
<td>0.3</td>
<td>0.768</td>
<td>0.771</td>
<td>0.659</td>
<td>0.673</td>
<td>0.523</td>
</tr>
<tr>
<td>0.4</td>
<td>0.944</td>
<td>0.943</td>
<td>0.859</td>
<td>0.872</td>
<td>0.762</td>
</tr>
<tr>
<td>0.5</td>
<td>0.984</td>
<td>0.989</td>
<td>0.959</td>
<td>0.972</td>
<td>0.882</td>
</tr>
<tr>
<td>0.6</td>
<td>0.997</td>
<td>0.997</td>
<td>0.989</td>
<td>0.994</td>
<td>0.969</td>
</tr>
<tr>
<td>0.7</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.999</td>
<td>0.983</td>
</tr>
<tr>
<td>0.8</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>0.9</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>1</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Note: This table lists rejection rates of the conventional test with fixed split point $R_T = [0.15T]$ and the modified sup-type test mining over $[\lfloor 0.15T \rfloor, \lfloor 0.85T \rfloor]$. The DGP is given by (1.35), where $d_{i,T} = 0.5 + \delta 1(i > \lfloor \tau T \rfloor)$. The color indicates the maximum between the rejection rates of the conventional test and the sup-type test.
if we put together the Monte Carlo and the analytic results derived in Section 1.5, we may conclude that the sup-type test is at least equally powerful and has the potential to be more in the presence of structural break. Thus, so long as one does not mind the computational cost, the sup-type test might be more preferable than the conventional test.

1.7 Discussions on other novel out-of-sample tests

1.7.1 The average-type test

We have analyzed asymptotic power of the modified sup-type test, whose original version is proposed by Rossi and Inoue (2012). They have also proposed an average-type test, which in our framework is interpreted as the ad hoc treatment for averaging the conventional test statistic evaluated at different split points (see Example 4). It can be shown under the same conditions of Proposition 9 that

\[
P\left(\int_{\tau}^{\infty} S(\xi) d\xi > x\right) = P\left(\int_{\tau}^{\infty} \left(W(\xi) + C(1 - \nu^{-1}(\xi))\right) d\xi > x\right)
\]

\[
P\left(\int_{\tau}^{\infty} |S(\xi)| d\xi > x\right) = P\left(\int_{\tau}^{\infty} |W(\xi) + C(1 - \nu^{-1}(\xi))| d\xi > x\right)
\]

Thus, computation of power of the one-sided average-type test boils down to computation of cumulative distribution of the integral of Wiener process, which is elementary and can be done by, for example, invoking the stochastic Fubini theorem. Development of the two-sided average-type test requires different techniques as the handy stochastic Fubini theorem does not apply.

We will not analyze power of the average-type test, since the test tends to have low power. In fact, the reader might check Table 8 to Table 10 in Rossi and Inoue (2012) (pp. 445-446) and will find that the average-type test rejects the null hypothesis of equal predictive ability much less often than both the conventional test and the sup-type test in the presence of structural break.\(^{28}\) The cause of the relatively low power of the average-type test has not been discussed in the original paper, so here we provide some heuristic explanations. Consider the case in which two models exhibit unequal predictive abilities after a structural break that occurs at time \(\tau\). For a test to detect the unequal predictive abilities, it must make effective use of the post-break observations; intuitively, the more post-break observations used, the more evidence for unequal predictive abilities, and the more likely for the test to reject the null. By this reasoning, the conventional test statistic \(d_{T}(R_{1})\) evaluated at \(R_{1} < \tau\) is more likely to suggest unequal predictive abilities than the test statistic \(d_{T}(R_{2})\) evaluated

\(^{28}\)In addition, in their empirical study the authors contrast either of the UIRP and the Taylor rule model with a random walk model. They find dramatically less evidence on unequal predictive accuracy with the average-type test than with the sup-type test. See Table 12 therein. The dramatic difference might be explained by the same reason we raise here.
at \( R_2 > \tau \). Now suppose that \( [l T] < \tau < [u T] \), and compare the conventional test statistic \( |d_T([l T])| \) using fixed split point \( R_T \equiv [l T] \) and the (two-sided) average-type test computed as \( A_T \equiv \frac{1}{[u T]-[l T]+1} \sum_{R_T=[l T]}^{[u T]} |d_T(R_T)| \). As said above, when \( R_T > \tau \) the test statistic \( |d_T(R_T)| \) becomes less likely to reject the null hypothesis since it ignores observations in \( [\tau, R_T] \) that would have provided evidence for unequal predictive abilities. By assigning weights to those \( |d_T(R_T)| \) with \( R_T > \tau \), the average-type test statistic essentially downplays the informative post-break observations and hence possesses lower power than the conventional test using split point earlier than the break time \( \tau \).

### 1.7.2 The pass-type test

We might wonder if a *pass-type* test can be developed by conducting the pass-mode split described in Example 2. There is, however, a major pitfall in the development: determining the critical values. The asymptotic distribution of the pass-type test statistic throws little light on the development. According to Corollary 4, asymptotic distribution of the test statistic is bounded from above by the threshold, but in finite sample a critical value larger than the threshold might be needed. One possible way to facilitate the asymptotic analysis is to revise the pass-mode and make the boundary time-varying. To be clear, the revised pass-mode of split may be described by

\[
R_T \equiv \inf\{ [l T] \leq r \leq [u T] : |d_T(r)| > c(r) \} \wedge [u T] \tag{1.36}
\]

where the threshold \( c \) is no longer a constant, but a boundary function \( r \mapsto c(r) \). By choosing the boundary, it is possible to have size of the associated pass-type test controlled at the nominal level. In fact, we are essentially drawing an analogy between the pass-type test and the sequential test widely used in clinical trials and structural stability assessment, and the suggested solution is a common approach to designing feasible sequential test. For more discussion on the approach, see, e.g., Anatolyev and Kosenok (2018).

It should be noted that the suggested solution touches only one side of the development, namely how to develop a pass-type test that is correctly sized. An even more important problem is whether by choosing certain boundary we can render the pass-type test more powerful than the conventional test using fixed split point – it is only in this case that we find the pass-type test appealing. This problem is not addressed here, but left for future study.
1.8 Discussions on test for nested models and comparison of forecast methods

1.8.1 Nested models

The analysis from Section 1.3 onwards is confined to comparison of nonnested models. We do not deal with nested models due to some technical pitfalls, which are briefly explained below.

Consider two predictive regression models for an $h$-period forecast horizon:

$$y_{t+h} = u_t' \beta^* + \varepsilon_{t+h}$$
$$y_{t+h} = u_t' \gamma^* + v_t' \delta^* + \eta_{t+h}$$

where $u_t$ is $k$-dimensional vector and $v_t$ is $q$-dimensional vector. For notational simplicity, let $\theta \equiv (\gamma, \delta)$, and correspondingly $\theta^* \equiv (\gamma^*, \delta^*)$. At each period $t$, we let $\hat{y}_{t+h|t}(\beta) \equiv u_t' \beta$ denote forecast of $y_{t+h}$ produced by the first model (with $\beta^*$ replaced by $\beta$), and let $\tilde{y}_{t+h|t}(\theta) \equiv u_t' \gamma + v_t' \delta$ denote forecast of $y_{t+h}$ produced by the second model. The test of equal predictive ability, as gauged by squared error, has the null hypothesis

$$H_0 : \mathbb{E}[(y_{t+h} - \hat{y}_{t+h|t}(\beta^*))^2] = \mathbb{E}[(y_{t+h} - \tilde{y}_{t+h|t}(\theta^*))^2]$$

which is equivalent to $\delta^* = 0$. The widely used test for this hypothesis developed by McCracken (2007) uses the statistic

$$d_T(R_T) \equiv \frac{\sum_{t=R_T}^{T} (y_{t+h} - \hat{y}_{t+h|t}(\hat{\beta}_t))^2 - (y_{t+h} - \tilde{y}_{t+h|t}(\hat{\theta}_t))^2}{\hat{\sigma}_\varepsilon^2}$$

where $\hat{\beta}_t$ and $\hat{\theta}_t$ are recursively formed OLS estimators, $R_T$ is the sample split point, and $\hat{\sigma}_\varepsilon^2$ is a consistent estimator of $\text{var}(\varepsilon_{t+h})$. Asymptotics of this test statistic is derived in McCracken (2007) and further simplified by Hansen and Timmermann (2015). Letting $S_T(\xi) \equiv d_T(\lfloor \xi T \rfloor)$, Theorem 3 of Hansen and Timmermann (2015) suggests that

$$S_T(\xi) \overset{w}{\to} \sum_{i=1}^{q} \lambda_i \left[ W_i^2(1) - \xi^{-1} W_i^2(\xi) + \log(\xi) + (1 - \xi)c^2 + 2ca_i(W_i(1) - W_i(\xi)) \right]$$

(1.37)

under the local alternative $\delta = cT^{-1/2}b$, where $b$ is a suitably normalized vector, $c$ is a real-number, $\lambda_i$ and $a_i$ are nuisance parameters, and $(W_1, \cdots, W_q)'$ is a $q$-dimensional standard Wiener process.

We have not analyzed the sup-type test for the nested case since the distribution functions of the sup-type test statistics $\sup_{t \leq \xi \leq u} S_T(\xi)$ and $\sup_{t \leq \xi \leq u} |S_T(\xi)|$ seem to warrant a
separate study. Even for the one-sided test, we need to calculate under the null hypothesis (i.e., $c = 0$) the distribution function of $\sup_{\xi \leq u} \left( W^2_i (1) - \xi^{-1} W^2_i (\xi) + \log(\xi) \right)$ to obtain critical values, and we are not aware of a closed-form formula.

Nonetheless, we have some intuitions for the potential of the sup-type test in the nested case. The weak convergence in (1.37) is established under the local alternative with Pitman drift. As noted by Hansen and Timmermann (2015) (p. 2495), the non-centrality parameter $(1 - \xi)c^2$ is crucial for detecting unequal predictive abilities. The larger is $\xi$, namely the shorter is the hold-out sample, the smaller the non-centrality parameter, and hence the lower the power of the test. This observation is actually in line with the observation that we had for comparison of nonnested models under the local alternative with Pitman drift. We naturally conjecture that, under the local alternative $\delta = cT^{-1/2}b$, the sup-type test is not more powerful than the conventional test using fixed small split point in comparison of nested models. On the other hand, we conjecture that a different conclusion might be drawn in the presence of structural break, as in the case for nonnested models. Consider the local alternative $\delta_t \equiv cT^{-1/2}b1(t \geq \lfloor \lambda T \rfloor)$, where $\lambda \in (0, 1)$ is the relative break time. It seems that the proof of Theorem 3 in Hansen and Timmermann (2015) can be adapted to derive the weak limit of the test statistic $S_T$ under such local alternative. In particular, the non-centrality parameter $(1 - \xi)c^2$ in (1.37) would be replaced by $(1 - \xi \vee \lambda)c^2$. The fact that $(1 - \xi \vee \lambda)c^2 = (1 - \lambda)c^2$ for all $\xi \leq \lambda$ suggests that there might not be any gain in using a small split point. Intuitively, the pre-break observations are not as evidential as the post-break ones. Furthermore, it is likely that, as in the case for nonnested models, inclusion of the pre-break observations "dilutes" the information in the test statistic, rendering the conventional test using fixed small split point less powerful than the sup-type test.

Certainly, the discussions in the preceding paragraph are only conjectures, and we leave thorough analysis of the nested case to future study.

### 1.8.2 Comparison of Forecast Methods

The paper has so far been focusing on comparison of forecast models. Giacomini and White (2006) developed an alternative framework for comparison of forecast methods. To recap, let $z_t \equiv (y_t, x_t')'$, where $y_t$ is the variable to be forecast. At each period $t$, we suppose the availability of two forecasts of $y_{t+h}$, which are denoted by $\hat{y}_{t+h|t}(\hat{\beta}_{t,\hat{m}_t})$ and $\tilde{y}_{t+h|t}(\tilde{\gamma}_{t,\tilde{m}_t})$ respectively. Here $\hat{\beta}_{t,\hat{m}_t}$ is an estimator, either parametric or non-parametric, that is dependent on the latest $\hat{m}_t$ observations $z_t, z_{t-1}, \ldots, z_{t-\hat{m}_t+1}$, and $\tilde{y}_{t+h|t}(\hat{\beta}_{t,\hat{m}_t})$ is assumed to have the functional form

$$\hat{y}_{t+h|t}(\hat{\beta}_{t,\hat{m}_t}) \equiv \hat{f}(z_t, z_{t-1}, \ldots, z_{t-\hat{m}_t+1}; \hat{\beta}_{t,\hat{m}_t})$$
1.8. DISCUSSIONS ON TEST FOR NESTED MODELS AND COMPARISON OF FORECAST METHODS

Similar assumptions are imposed on \( \tilde{y}_{t, \tilde{m}_t} \) and \( \tilde{y}_{t+h|t}(\tilde{\gamma}_{t, \tilde{m}_t}) \). Letting \( L_{t+h} \) denote the loss function, the null hypothesis to be tested is

\[
\mathbb{E} \left[ L_{t+h}(y_{t+h}, \tilde{y}_{t+h|t}(\tilde{\gamma}_{t, \tilde{m}_t})) \right] = \mathbb{E} \left[ L_{t+h}(y_{t+h}, \tilde{y}_{t+h|t}(\tilde{\gamma}_{t, \tilde{m}_t})) \right] \quad \text{for all } t \tag{1.38}
\]

In contrast to the null hypothesis (1.1), the loss functions in (1.38) are not evaluated at the probability limits of the estimators \( \hat{\gamma}_{t, \hat{m}_t} \) and \( \tilde{\gamma}_{t, \tilde{m}_t} \), but the estimators themselves. This is a distinguishing feature of the Giacomini-White framework: what are being compared are not predictive abilities of two forecast models, but predictive abilities of two forecast methods, which include in addition to the underlying models all practicalities in generating the forecasts. For example, one might compare two methods having the same forecast model but different estimation windows; this is noted in the original paper of Giacomini and White (2006) and reiterated in the recent review by Elliott and Timmermann (2016).

Here we would like to discuss briefly is the role of the sample split point in the Giacomini-White framework. In this framework, the sample split point in the pseudo out-of-sample analysis is tied to the estimation window sizes \( \hat{m}_t \) and \( \tilde{m}_t \). The asymptotic distribution of the Giacomini-White test statistic is derived under the assumption that (1) the window sizes \( \hat{m}_t \) and \( \tilde{m}_t \) are random integers adapted to the \( \sigma \)-filtration generated by \( \{z_t\} \); (2) the sequences \( \{\hat{m}_t\} \) and \( \{\tilde{m}_t\} \) are uniformly bounded by a constant \( M \). This constant \( M \) typically serves as the sample split point in the pseudo out-of-sample analysis: we begin by computing the forecasts \( \hat{y}_{M+h|M}(\hat{\gamma}_{M, \hat{m}_M}) \) and \( \hat{y}_{M+h|M}(\tilde{\gamma}_{M, \tilde{m}_M}) \) with \( \{z_M, z_{M-1}, \ldots, z_{M-\tilde{m}_M+1}\} \) and \( \{z_M, z_{M-1}, \ldots, z_{M-\hat{m}_M+1}\} \) respectively; then we go on to compute \( \hat{y}_{M+1+h|M+1}(\hat{\gamma}_{M+1, \hat{m}_{M+1}}) \) and \( \hat{y}_{M+1+h|M+1}(\tilde{\gamma}_{M+1, \tilde{m}_{M+1}}) \) with \( \{z_{M+1}, z_M, \ldots, z_{M-\tilde{m}_{M+1}+2}\} \) and \( \{z_{M+1}, z_M, \ldots, z_{M-\hat{m}_{M+1}+2}\} \) respectively; the process is repeated until the sample is exhausted.

In sharp contrast to the framework for comparison of forecast models, the Giacomini-White framework explicitly allows for data-driven sample split, as long as the window sizes are determined in a “look-back” manner.\(^{29}\) Thus, in the Giacomini-White framework the problem of \( p \)-hacking arises only if one chooses the window sizes in a “look-ahead” way, thus violating the adaptedness assumption on \( \hat{m}_t \) and \( \tilde{m}_t \). This is the case, for example, if one decides to set \( \hat{m}_t \) and \( \tilde{m}_t \) to the same “constant” \( m \), but \( m \) is determined by the sup-mode or the pass-mode mining described in this paper.

Despite many appealing features of the Giacomini-White framework and its great importance to practitioners, we have not analyzed it in this paper, as we believe that extension of the asymptotic results established in Giacomini and White (2006) is non-trivial and warrants independent study. The main non-trivial technicality lies in adaptedness of the estimation window sizes. When the window sizes \( \hat{m}_t \) and \( \tilde{m}_t \) are adapted to the \( \sigma \)-filtration of the observations – in plain words, when they are only contingent on historical data – Giacomini

\(^{29}\) As remarked in their original paper, one can follow Pesaran and Timmermann (2007) in determining the estimation windows.
and White show that the partial sum of the loss differential

$$\Delta L_{t+h} \equiv L_{t+h}\left(y_{t+h}, \hat{y}_{t+h|t}(\hat{\beta}_t, \hat{m}_t)\right) - L_{t+h}\left(y_{t+h}, \tilde{y}_{t+h|t}(\tilde{\gamma}_t, \tilde{m}_t)\right)$$

satisfies certain mixing and moment conditions. In this case, we can invoke suitable central limit theorem for near-epoch dependent process to construct a Wald-type statistic that converges weakly to a chi-square limit. But in our study, we are primarily interested in the case where $\hat{m}_t$ and $\tilde{m}_t$ can depend on the whole sample, and hence need not be adapted. Lifting the adaptedness assumption, the loss differential process $\{\Delta L_{t+h}\}$ is no longer mixing, and it is unclear whether it will converge to a Gaussian limit. Given this pitfall, we decide to leave more thorough analysis of Giacomini-White test with data-driven sample split to future research.

### 1.9 Conclusions

The first half of the paper analyzes the size inflation caused by data-driven sample split. The problem was recently discussed in Rossi and Inoue (2012) and Hansen and Timmermann (2012); we add that there are a multitude of data-driven split modes, each requiring an ad hoc treatment, and achieving robustness often comes at the expense of considerable loss of power. This problem is certainly not new and was already known in the 1940s (Feller (1940)). It catches researchers’ attention again as the scientific community has become increasingly concerned about $p$-hacking in the recent years (Simonsohn et al. (2014); Munafò et al. (2017)).

Inspired by Rossi and Inoue (2012), the second half of the paper turns to study if data-driven sample split makes more powerful test than the conventional test using fixed split point. To facilitate asymptotic analysis, we modify the original sup-type test proposed by Rossi and Inoue (2012) to circumvent Monte Carlo and make it applicable whether impact of the estimation uncertainty vanishes or not. Analytic results under local alternatives show that the sup-type test is less powerful than the conventional test using fixed and small split point in the absence of structural break, but becomes more powerful in the presence.

There are many problems to be further studied. Developing a feasible pass-type test and assessing its power is one of the problems, as discussed in Section 1.7.2. Another problem concerns out-of-sample test for nested models. Most analytic results derived in this paper are predicated on the fact that the limit process of the test statistic is a time-changed Wiener process (with drift). This is usually not case when we consider test for nested models. For example, limit process of the out-of-sample $F$ test statistic – which is originally developed in McCracken (2007) and further analyzed by Hansen and Timmermann (2015) – involves stochastic integral of Wiener process and might require different techniques in analyzing the sup-type test. We leave these problems for future research.
Acknowledgment

I have benefited from many helpful suggestions kindly offered by Michael Wolf, Eric Hillebrand, Ingmar Nolte, Asger Lunde and participants at the CREATES lunch seminar. I would like to thank CREATES for research support. The research project leading to these results receives funding from the European Union Research and Innovation programme Horizon 2020 under grant agreement No. 675044 (BigDataFinance). All errors are solely mine.
Appendix A  Estimation of nuisance parameters

The sup-type test makes use of the formulae in Proposition 7, which involve the unknown variances $V(l)$ and $V(u)$ that must be estimated. We let $\hat{V}(l)$ and $\hat{V}(u)$ denote their estimators below. These two estimators in turn rely on estimators of the nuisance parameters $F$, $B$ and $\Omega$. In this appendix, we give details about their estimation.

The vector $F$: By definition, $F \equiv \mathbb{E}[\frac{\partial f_{t+1}(\beta^*)}{\partial \beta}]$, where

$$\frac{\partial f_{t+1}(\beta^*)}{\partial \beta} = 2 \begin{bmatrix} -y_t \eta_{1,t+1} \\ -v_t \eta_{1,t+1} \\ y_t \eta_{2,t+1} \\ v_t \eta_{2,t+1} \end{bmatrix}$$

$F$ can be estimated by

$$\hat{F}_T \equiv \frac{2}{T-[IT]+1} \sum_{t=[IT]}^T \begin{bmatrix} -y_t(y_{t+1} - y_t \hat{\theta}_1 - w_t \hat{\theta}_2) \\ -w_t(y_{t+1} - y_t \hat{\theta}_1 - w_t \hat{\theta}_2) \\ y_t(y_{t+1} - y_t \hat{\gamma}_1 - v_t \hat{\gamma}_2) \\ v_t(y_{t+1} - y_t \hat{\gamma}_1 - v_t \hat{\gamma}_2) \end{bmatrix}$$

The same estimator is used for both $\hat{V}(l)$ and $\hat{V}(u)$.

The matrix $B$: By definition, $B$ is the (almost sure) limit of $B_t^*$ (see Assumption B). In our Monte Carlo experiments, we use

$$\hat{B}_T \equiv \begin{bmatrix} \frac{1}{[uT]} \sum_{s=\max(T-[uT],0)}^{T-1} w_{s-1} y_s \\ \frac{1}{[uT]} \sum_{s=\max(T-[uT],0)}^{T-1} w_{s-1} w_s \\ \frac{1}{[uT]} \sum_{s=\max(T-[uT],0)}^{T-1} y s w_s \\ \frac{1}{[uT]} \sum_{s=\max(T-[uT],0)}^{T-1} y^2 s \end{bmatrix}$$

for both $\hat{V}(l)$ and $\hat{V}(u)$.

The matrix $\Omega$: $\Omega$ is the long-run covariance of $(f_{t+1}(\beta^*) - \mu_t, g_t(\beta^*))'$ (see Assumption B), where for the DGP in our Monte Carlo experiment

$$g_t(\beta) = \begin{bmatrix} w_{s-1}(y_{s+1} - y_s \theta_1 - w_s \theta_2) \\ w_s y_{s+1} - y_s \theta_1 - w_s \theta_2 \\ v_{s-1}(y_{s+1} - y_s \gamma_1 - v_s \gamma_2) \\ v_s y_{s+1} - y_s \gamma_1 - v_s \gamma_2 \end{bmatrix}$$

We let $\hat{\Omega}_T$ be the HAC estimator calculated with $(f_{t+1}(\hat{\beta}_t) - \hat{f}, g_t(\hat{\beta}_t))'$, where $\hat{\beta}_t$ is the “insample” 2SLS estimator using data available at $t$, and $\hat{f} \equiv \frac{1}{T-[IT]+1} \sum_{t=[IT]}^T f_{t+1}(\hat{\beta}_t)$. The
A. Estimation of nuisance parameters

Bartlett kernel and bandwidth \([0.75T^{1/3}]\) are used. Consistency of such HAC estimator is proved in McCracken (2000). Again, the same \(\hat{\Omega}_T\) is used to construct both \(\hat{V}(l)\) and \(\hat{V}(u)\).
Appendix B  Proofs

B.1 Proofs for Section 1.2.3

Proof of Proposition 1. Given the assumptions, the characteristic function of \( d_T(R_T) \) coincides with that of \( S_T(\delta) \). Write \( \mu \) as the probability measure on \([l, u]\) induced by \( \delta \). By independence of \( \delta \) and the data,

\[
\mathbb{E}\left[\exp(ixS_T(\delta))\right] = \int_l^u \mathbb{E}\left[\exp(ixS_T(\xi))\right] \mu(d\xi)
\]

Next, since \( S(\xi) \) has the same law \( \mathcal{L} \) for any \( \xi \in [l, u] \), its characteristic function does not depend on \( \xi \), and we may write it as \( \varphi(x) \equiv \mathbb{E}\left[\exp(ixS(\xi))\right] \). In light of Lévy’s continuity theorem, \( \lim_T \mathbb{E}\left[\exp(ixS_T(\xi))\right] = \varphi(x) \). Hence, invoking the dominated convergence theorem, we arrive at the convergence

\[
\lim_T \mathbb{E}\left[\exp(ixS_T(\delta))\right] = \int_l^u \varphi(x) \mu(d\xi) = \varphi(x)
\]

The proof is concluded by applying Lévy’s continuity theorem again. \( \square \)

B.2 Proofs for Section 1.2.5

Proof of Proposition 2. When \( \{\varphi_T\} \) all coincide with \( \varphi \), it is simply the continuous mapping theorem. The more general case follows from Theorem 8.4.1. of Bogachev (2007). \( \square \)

Proof of Corollary 4. Proposition 2 is still applicable if we deem the passage time as a functional on the processes, but here we invoke Skorokhod representation theorem (see, e.g., Section 8.5 of Bogachev (2007)) to aid in the proof. Let \( \tau_T \equiv \inf\{l \leq \xi \leq u : |S_T(\xi)| > c\} \land u \), so that \( |d_T(R_T)| \) can be represented as \( |S_T(\tau_T)| \). By virtue of Skorokhod’s representation theorem we can construct a probability space \( (\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}}) \) on which define the \( \mathbb{D}[l, 1] \)-valued random elements \( \tilde{S}_T \) and \( \tilde{S} \), such that \( \tilde{S}_T \overset{d}{=} S_T, \tilde{S} \overset{d}{=} S \), and \( \tilde{S}_T \overset{a.s.}{\rightarrow} \tilde{S} \). The corresponding passage times are \( \tilde{\tau}_T \equiv \inf\{l \leq \xi \leq u : |\tilde{S}_T(\xi)| > c\} \land u \) and \( \tilde{\tau} \equiv \inf\{l \leq \xi \leq u : |\tilde{S}(\xi)| > c\} \land u \) respectively.

Clearly \( |\tilde{S}_T(\tilde{\tau}_T)| \overset{d}{=} |S_T(\tau_T)| \) and \( |\tilde{S}(\tilde{\tau})| \overset{d}{=} |S(\tau)| \). Hence it suffices to prove \( |\tilde{S}(\tilde{\tau}_T)| \overset{a.s.}{\rightarrow} |\tilde{S}(\tilde{\tau})| \), and the problem essentially boils down to the case in which all terms are non-random.

Continuity of the limit process \( S \) validates the use of uniform metric in lieu of the Skorokhod metric. Also the assumption has ruled out the case \( \sup_{l \leq \xi \leq u} |S(\xi)| = c \). Hence we just need to consider two distinct cases \( \sup_{l \leq \xi \leq u} |S(\xi)| < c \) and \( \sup_{l \leq \xi \leq u} |S(\xi)| > c \). In the former case, convergence in the uniform metric dictates \( \tilde{\tau}_T = \tilde{\tau} = u \) and \( |\tilde{S}_T(\tilde{\tau}_T)| \rightarrow |\tilde{S}(\tilde{\tau})| = |\tilde{S}(u)| \) as \( T \rightarrow \infty \). In the latter case, \( \tilde{\tau} < u \) and again, convergence in the uniform metric entails \( \tilde{\tau} \rightarrow \tilde{\tau} \) and \( \tilde{S}_T(\tilde{\tau}_T) \rightarrow \tilde{S}(\tilde{\tau}) \) as \( T \rightarrow \infty \). \( \square \)
B.3 Proofs for Section 1.3.1

The key to proof of (1.10) is generalization of Lemma A4 of West (1996). In particular, his Lemma A4(b) and A4(c) can be easily generalized to convergence under the uniform metric, so we focus on generalization of his Lemma A4(a). To make it clear, in the proofs below we shall repeatedly use the following two inequalities without explicit citation or reference.

**Lemma B1.** Let $X \in \mathcal{A}$ and $Y \in \mathcal{B}$ be two random variables, where $\mathcal{A}$ and $\mathcal{B}$ are $\sigma$-algebras. If $\alpha(\mathcal{A}, \mathcal{B})$ denotes the $\alpha$-mixing coefficient, then for any $p, q > 1$ such that $p^{-1} + q^{-1} < 1$, we have

$$|\text{cov}(X, Y)| \leq 8\alpha(\mathcal{A}, \mathcal{B})^{1-p^{-1}-q^{-1}} \|X\|_p \|Y\|_q$$  \hspace{1cm} (B1)

**Lemma B2.** Let $\mathcal{A}$ and $\mathcal{B}$ be two $\sigma$-algebras. Then for any $\mathcal{A}$-measurable random variable $X$ and $1 \leq p \leq r \leq \infty$,

$$\|\mathbb{E}[X|\mathcal{A}] - \mathbb{E}[X]\|_p \leq 6\alpha(\mathcal{A}, \mathcal{B})^{1/p-1/r} \|X\|_r$$  \hspace{1cm} (B2)

Inequality (B1) can be found on p.278 of Hall and Heyde (1980) and will be referred to as covariance inequality in the sequel. Inequality (B2) is (2.3) of McLeish (1975) and will be referred to as McLeish’s inequality.

**Uniform mean-value expansion: The recursive scheme**

**Lemma B3.** Let $v_t \equiv \partial_{\beta} f_{t+h}(\beta^*) - F$. Under the recursive scheme, $\max_{1 \leq R \leq T} |\sum_{t=1}^R v_t BG_t| = o_p(T^{1/2})$, where $G_t \equiv \frac{1}{t} \sum_{i=1}^t g_i$.

**Proof.** In this proof, we use $g_i$ and $G_t$ to mean $Bg_i$ and $BG_t$ respectively, just to ease the already heavy use of notations. This is innocuous since $B$ is a constant matrix. Without loss of generality, we assume that $v_t$ and $g_i$ are scalars. We shall write $a \ll b$ if there exists a constant $\kappa$ that is independent of $t$ and $T$ such that $a \leq \kappa b$. $\|\cdot\|_p$ denotes the $L^p$-norm. Finally, because of Assumption B3(c) we can pick a constant $\delta > 1$ that satisfies $a_k^{1-1/d} \ll k^{-4\delta}$, where $a_k$ is the $\alpha$-mixing coefficient.

We begin with

$$\max_{1 \leq R \leq T} \left| \sum_{t=1}^R v_t G_t \right| \leq \max_{1 \leq R \leq T} \left| \sum_{t=1}^R (v_t G_t - \mathbb{E}[v_t G_t]) \right| + \max_{1 \leq R \leq T} \left| \sum_{t=1}^R \mathbb{E}[v_t G_t] \right|$$

The term $\max_{1 \leq R \leq T} |\sum_{t=1}^R \mathbb{E}[v_t G_t]|$ is $o(T^{1/2})$. In fact, by the covariance inequality and assumption B3(c),

$$|\mathbb{E}[v_t G_t]| = \left| t^{-1} \sum_{i=1}^t \mathbb{E}[v_t g_i] \right| \leq t^{-1} \sum_{i=1}^t |\mathbb{E}[v_t g_i]|$$

50
The proof proceeds by means of the maximal inequality of McLeish (1975) (see his Lemma 1.5). For ease of reference we reproduce this maximal inequality below. We break down the remaining proof into several steps.

**Step 1:** Define $\mathcal{F}_t \equiv \sigma(U_s : 1 \leq s \leq t)$ for $t \geq 1$, where $U_t$ is defined in Assumption B3, and set $\mathcal{F}_t \equiv \{ \emptyset, \Omega \}$ for $-\infty \leq t \leq 0$. Let $S_R \equiv \sum_{t=1}^{R} (v_t G_t - \mathbb{E}[v_t G_t])$, $1 \leq R \leq T$. First, observe that

$$v_t G_t - \mathbb{E}[v_t G_t] = \mathbb{E}[v_t G_t - \mathbb{E}[v_t G_t] | \mathcal{F}_t] - \mathbb{E}[v_t G_t - \mathbb{E}[v_t G_t] | \mathcal{F}_{-\infty}]$$

$$= \sum_{k=0}^{\infty} \left( \mathbb{E}[v_t G_t - \mathbb{E}[v_t G_t] | \mathcal{F}_{t-k}] - \mathbb{E}[v_t G_t - \mathbb{E}[v_t G_t] | \mathcal{F}_{t-k-1}] \right)$$

Defining $Y_{R,k} \equiv \sum_{t=1}^{R} \left( \mathbb{E}[v_t G_t - \mathbb{E}[v_t G_t] | \mathcal{F}_{t-k}] - \mathbb{E}[v_t G_t - \mathbb{E}[v_t G_t] | \mathcal{F}_{t-k-1}] \right)$, we have

$$S_R = \sum_{t=1}^{R} (v_t G_t - \mathbb{E}[v_t G_t])$$

$$= \sum_{t=1}^{\infty} \sum_{k=1}^{R} \left( \mathbb{E}[v_t G_t - \mathbb{E}[v_t G_t] | \mathcal{F}_{t-k}] - \mathbb{E}[v_t G_t - \mathbb{E}[v_t G_t] | \mathcal{F}_{t-k-1}] \right)$$

$$= \sum_{k=0}^{\infty} \sum_{t=1}^{R} \left( \mathbb{E}[v_t G_t - \mathbb{E}[v_t G_t] | \mathcal{F}_{t-k}] - \mathbb{E}[v_t G_t - \mathbb{E}[v_t G_t] | \mathcal{F}_{t-k-1}] \right)$$

$$= \sum_{k=0}^{\infty} Y_{R,k}$$

As in McLeish (1975), let $\{a_k : k \geq 0\}$ be a nonincreasing sequence of positive numbers to be
determined later, which satisfies $b \equiv \sum_{k=0}^{\infty} a_k < \infty$. By Jensen's inequality,

$$|S_R|^2 = \left| \sum_{k=0}^{\infty} Y_{R,k} \right|^2 = b^2 \left| \sum_{k=0}^{\infty} \frac{Y_{R,k}}{a_k} \right|^2 \leq b^2 \sum_{k=0}^{\infty} \left| \frac{Y_{R,k}}{a_k} \right|^2 \leq b \sum_{k=0}^{\infty} \frac{|Y_{R,k}|^2}{a_k}$$

Then

$$\max_{1 \leq R \leq T} |S_R|^2 \leq \max_{1 \leq R \leq T} \left( \sum_{k=0}^{\infty} a_k \left( \sum_{k=0}^{\infty} a_k^{-1} |Y_{R,k}|^2 \right) \right) \leq \left( \sum_{k=0}^{\infty} a_k \right) \left( \sum_{k=0}^{\infty} a_k^{-1} \max_{1 \leq R \leq T} |Y_{R,k}|^2 \right)$$

For each $k$, $\{Y_{R,k} : R \geq 1\}$ is a martingale with respect to $\{\mathcal{F}_{t-k} : R \geq 1\}$. Hence by Doob's $L^2$-inequality,

$$\mathbb{E}[\max_{1 \leq R \leq T} |S_R|^2] \leq \left( \sum_{k=0}^{\infty} a_k \right) \left( \sum_{k=0}^{\infty} a_k^{-1} \mathbb{E}[\max_{1 \leq R \leq T} |Y_{R,k}|^2] \right) \leq 4 \left( \sum_{k=0}^{\infty} a_k \right) \left( \sum_{k=0}^{\infty} a_k^{-1} \mathbb{E}[|Y_{T,k}|^2] \right) \tag{B3}$$

**Step 2:** In step 1 we have reproduced the maximal inequality of McLeish (1975). Next, by the martingale property we obtain

$$\mathbb{E}[Y_{T,k}^2] = \sum_{t=1}^{T} \mathbb{E}[v_t G_t - \mathbb{E}[v_t G_t] | \mathcal{F}_{t-k}] - \mathbb{E}[v_t G_t - \mathbb{E}[v_t G_t] | \mathcal{F}_{t-k-1}] \|_2^2$$

Hence

$$\sum_{k=0}^{\infty} a_k^{-1} \mathbb{E}[Y_{T,k}^2]$$

$$= \sum_{k=0}^{\infty} \sum_{t=1}^{T} a_k^{-1} \mathbb{E}[v_t G_t - \mathbb{E}[v_t G_t] | \mathcal{F}_{t-k}] - \mathbb{E}[v_t G_t - \mathbb{E}[v_t G_t] | \mathcal{F}_{t-k-1}] \|_2^2$$

$$= \sum_{t=1}^{T} \sum_{k=0}^{\infty} a_k^{-1} \mathbb{E}[v_t G_t - \mathbb{E}[v_t G_t] | \mathcal{F}_{t-k}] - \mathbb{E}[v_t G_t - \mathbb{E}[v_t G_t] | \mathcal{F}_{t-k-1}] \|_2^2$$

$$= \sum_{t=1}^{T} \sum_{k=0}^{\infty} a_k^{-1} \|v_t G_t - \mathbb{E}[v_t G_t] | \mathcal{F}_{t-k} \|_2^2 - \|v_t G_t - \mathbb{E}[v_t G_t] | \mathcal{F}_{t-k-1} \|_2^2 \tag{B4}$$

where the last equality is due to the fact that $\mathbb{E}[v_t G_t - \mathbb{E}[v_t G_t] | \mathcal{F}_s] = \mathbb{E}[v_t G_t] - \mathbb{E}[v_t G_t] = 0$ for $s \leq 0$.

Next, observe

$$\sum_{k=0}^{\infty} a_k^{-1} \mathbb{E}[v_t G_t - \mathbb{E}[v_t G_t] | \mathcal{F}_{t-k}] - \mathbb{E}[v_t G_t - \mathbb{E}[v_t G_t] | \mathcal{F}_{t-k-1}] \|_2^2$$

$$= \sum_{k=0}^{t} a_k^{-1} \left( \|v_t G_t - \mathbb{E}[v_t G_t] | \mathcal{F}_{t-k} \|_2^2 - \|v_t G_t - \mathbb{E}[v_t G_t] | \mathcal{F}_{t-k-1} \|_2^2 \right)$$

$$= \sum_{k=0}^{t} a_k^{-1} \|v_t G_t - \mathbb{E}[v_t G_t] | \mathcal{F}_{t-k} \|_2^2 - \sum_{k=0}^{t} a_k^{-1} \|v_t G_t - \mathbb{E}[v_t G_t] | \mathcal{F}_{t-k-1} \|_2^2$$

$$= t (a_k^{-1} - a_{k-1}^{-1}) \|v_t G_t - \mathbb{E}[v_t G_t] | \mathcal{F}_{t-k} \|_2^2 + a_0^{-1} \|v_t G_t - \mathbb{E}[v_t G_t] \|_2^2 \tag{B5}$$
For $1 \leq k \leq t$, we have
\[
\|E[v_t G_t - E[v_t G_t]|\mathcal{F}_{t-k}]\|_2^2 \\
= \left\| \sum_{i=1}^{t} t^{-1} E[v_t g_i - E[v_t g_i]|\mathcal{F}_{t-k}] \right\|_2^2 \\
= \left\| \sum_{i=1}^{t-k} t^{-1} g_i E[v_t - E[v_t g_i]|\mathcal{F}_{t-k}] + \sum_{i=t-k+1}^{t} t^{-1} E[v_t g_i - E[v_t g_i]|\mathcal{F}_{t-k}] \right\|_2^2 \\
\leq \left\| \sum_{i=1}^{t-k} t^{-1} g_i E[v_t|\mathcal{F}_{t-k}] \right\|_2^2 + \left\| \sum_{i=1}^{t-k} t^{-1} g_i E[v_t g_i] \right\|_2^2 + \left\| \sum_{i=t-k+1}^{t} t^{-1} E[v_t g_i - E[v_t g_i]|\mathcal{F}_{t-k}] \right\|_2^2 \\
\leq \left( \sum_{i=1}^{t-k} g_i \right)^2 \left( \sum_{i=1}^{t-k} g_i^2 \right)^2 + 2 \sum_{i=1}^{t-k-1} \sum_{j=2}^{t-k} g_i g_j \left( \sum_{i=1}^{t-k} g_i^2 \right)^2
\]
\[(B6)\]

In the next step, we give proper upper bounds to the three terms appearing on the right hand side of the above inequality.

**Step 3-1:** Let us give an upper bound to the first term on the right hand side of (B6). Observe
\[
\left( \sum_{i=1}^{t-k} g_i \right)^2 \left( \sum_{i=1}^{t-k} g_i^2 \right)^2 = \sum_{i=1}^{t-k} g_i^2 \mathbb{E}[v_t|\mathcal{F}_{t-k}]^2 + 2 \sum_{i=1}^{t-k-1} \sum_{j=2}^{t-k} g_i g_j \mathbb{E}[v_t|\mathcal{F}_{t-k}]
\]

Firstly,
\[
E \left[ \sum_{i=1}^{t-k} g_i^2 \mathbb{E}[v_t|\mathcal{F}_{t-k}] \right] \\
= \sum_{i=1}^{t-k} \mathbb{E}\left[ g_i^2 \mathbb{E}[v_t|\mathcal{F}_{t-k}] \right] \\
\leq \sum_{i=1}^{t-k} \| g_i \|_4^2 \| \mathbb{E}[v_t|\mathcal{F}_{t-k}] \|_4^2 \quad \text{(Hölder's inequality)} \\
\leq 6 \sum_{i=1}^{t-k} \| g_i \|_4^2 \| v_t \|_{4d}^2 \alpha_k^{2(1/4-1/4d)} \quad \text{(covariance inequality)} \\
\leq 6 \sum_{i=1}^{t-k} \| g_i \|_4^2 \| v_t \|_{4d}^2 k^{-2\delta} \quad \text{(Assumption B3(c))} \\
\ll (t - k) k^{-2\delta} \quad \text{(Assumption B3(b))}
\]

Secondly,
\[
\left| E \left[ \sum_{i=1}^{t-k-1} \sum_{j=2}^{t-k} g_i g_j \mathbb{E}[v_t|\mathcal{F}_{t-k}] \right] \right| \\
= \left| \sum_{i=1}^{t-k-1} \sum_{j=2}^{t-k} \mathbb{E}[g_i g_j \mathbb{E}[v_t|\mathcal{F}_{t-k}]] \right| \\
\leq \sum_{i=1}^{t-k-1} \sum_{j=2}^{t-k} |\mathbb{E}[g_i g_j \mathbb{E}[v_t|\mathcal{F}_{t-k}]]| \\
\leq \sum_{i=1}^{t-k-1} \sum_{j=2}^{t-k} |\text{cov}(g_i, g_j \mathbb{E}[v_t|\mathcal{F}_{t-k}])|
\]

53
Combining (B7) and (B8), we obtain the estimate

\[
\leq 8 \sum_{i=1}^{t-k-1} \sum_{j=2}^{t-k} \|g_i\|_{4d} \|g_j\|_4 \|E^2[v_t|\mathcal{F}_{t-k}]\|_{4/3} \alpha_{j-i}^{1/14d-3/4} \quad \text{(covariance inequality)}
\]

\[
\leq 8 \sum_{i=1}^{t-k-1} \sum_{j=2}^{t-k} \|g_i\|_{4d} \|g_j\|_4 \|E^2[v_t|\mathcal{F}_{t-k}]\|_2 \alpha_{j-i}^{14-4d} \quad \text{(Hölder's inequality)}
\]

\[
= 8 \sum_{i=1}^{t-k-1} \sum_{j=2}^{t-k} \|g_i\|_{4d} \|g_j\|_4 \|E[v_t|\mathcal{F}_{t-k}]\|_2^2 \alpha_{j-i}^{14-4d} \quad \text{(McLeish's inequality)}
\]

\[
\ll k^{-26} \sum_{i=1}^{t-k-1} \sum_{j=2}^{t-k} (j-i)^{-\delta} \quad \text{(Assumption B3(c))}
\]

\[
= k^{-26} \sum_{i=1}^{t-k-1} (t-k-l)^{-\delta}
\]

\[
= (t-k)k^{-26} \sum_{l=1}^{t-k-1} \frac{t-k-l}{t-k} l^{-\delta}
\]

\[
\leq (t-k)k^{-26} \sum_{l=1}^{\infty} l^{-\delta} \quad \ll (t-k)k^{-26}
\]

(B8)

Combining (B7) and (B8), we obtain the estimate

\[
\left\| \sum_{i=1}^{t-k} t^{-1} g_i \mathbb{E}[v_t|\mathcal{F}_{t-k}] \right\|_2 = t^{-1} \left( \mathbb{E} \left( \left( \sum_{i=1}^{t-k} g_i \mathbb{E}[v_t|\mathcal{F}_{t-k}] \right)^2 \right) \right)^{1/2} \ll t^{-1} (t-k)^{1/2} k^{-\delta}
\]

(B9)

**Step 3-2:** Next we bound the second term on the right hand side of (B6). We have

\[
\left\| \sum_{i=1}^{t-k} t^{-1} g_i \mathbb{E}[v_t|\mathcal{F}_{t-k}] \right\|_2 \leq t^{-1} \sum_{i=1}^{t-k} \|g_i\|_2 \|\mathbb{E}[v_t g_i]\|
\]

\[
\leq 8 t^{-1} \sum_{i=1}^{t-k} \|g_i\|_2 \|v_t\|_{2d} \|g_i\|_{2d} \alpha_{t-i}^{-1/d} \quad \text{(covariance inequality)} \quad \text{(B10)}
\]

\[
\ll t^{-1} \sum_{i=1}^{t-k} (t-i)^{-4\delta} \quad \text{(Assumption B3)}
\]

\[
= t^{-1} \sum_{i=k}^{t-k} i^{-4\delta}
\]

**Step 3-3:** For the third term, when \(i \geq t-k+1\), by McLeish’s inequality and Assumption B3 we have

\[
\left\| \mathbb{E}[v_t g_i - \mathbb{E}[v_t g_i|\mathcal{F}_{t-k}]] \right\|_2 \leq 6 \|v_t g_i\|_{2d} \alpha_{t-k}^{1/2 - 1/2d} \ll (t-k)^{-2\delta}
\]

54
Thus
\[
\left\| \sum_{i=t-k+1}^{t} t^{-1} \mathbb{E}[v_t g_i - \mathbb{E}[v_t g_i | \mathcal{F}_{t-k}]] \right\|_2 \\
\leq \sum_{i=t-k+1}^{t} t^{-1} \left\| \mathbb{E}[v_t g_i - \mathbb{E}[v_t g_i | \mathcal{F}_{t-k}]] \right\|_2 \\
\ll t^{-1} \sum_{i=t-k+1}^{t} (i - (t - k))^{-2\delta} \\
\ll t^{-1} \sum_{i=1}^{k} i^{-2\delta}
\]

Amalgamating (B9), (B10) and (B11), we obtain
\[
\| \mathbb{E}[v_t G_t - \mathbb{E}[v_t G_t] | \mathcal{F}_{t-k}] \|_2 \ll t^{-1} \left\{ (t-k)^{1/2} k^{-\delta} + \sum_{i=k}^{t-1} i^{-4\delta} + \sum_{i=1}^{k} i^{-2\delta} \right\} \ll t^{-1} (t-k)^{1/2} k^{-\delta} \quad \text{(B12)}
\]

and it follows
\[
\sum_{k=1}^{t} (a_k^{-1} - a_{k-1}^{-1}) \| \mathbb{E}[v_t G_t - \mathbb{E}[v_t G_t] | \mathcal{F}_{t-k}] \|_2^2 \ll t^{-2} \sum_{k=1}^{t} (a_k^{-1} - a_{k-1}^{-1})(t-k)^{-2\delta} \quad \text{(B13)}
\]

Step 4: We are left to bound the last term \( a_0^{-1} \| v_t G_t - \mathbb{E}[v_t G_t] \|_2 \) in (B5). This can be done following the same argument as in step 3. We have
\[
\| v_t G_t - \mathbb{E}[v_t G_t] \|_2 \leq \| v_t G_t \|_2 + \| \mathbb{E}[v_t G_t] \|
\]

Here \( v_t G_t = t^{-1} \sum_{i=1}^{t} v_t g_i \), and
\[
\left( \sum_{i=1}^{t} v_t g_i \right)^2 = \sum_{i=1}^{t} g_i^2 v_i^2 + 2 \sum_{i=1}^{t} \sum_{j=2}^{t} g_i g_j v_i^2
\]

By Hölder’s inequality and Assumption B3(2)\(^{30}\),
\[
\left| \mathbb{E} \left[ \sum_{i=1}^{t-1} \sum_{j=2}^{t} g_i g_j v_i^2 \right] \right| \\
\leq \sum_{i=1}^{t-1} \sum_{j=2}^{t} |\mathbb{E}[g_i g_j v_i^2]| \\
= \sum_{i=1}^{t-1} \sum_{j=2}^{t} |\text{cov}(g_i, g_j v_i^2)| \\
\leq 8 \sum_{i=1}^{t-1} \sum_{j=2}^{t} \|g_i\|_4 d \|g_j v_i^2\|_{4/3} a_{j-i}^{-1/4d - 3/4} \quad \text{(covariance inequality)}
\]

\(^{30}\)It is for this inequality that we assume slightly faster decay of the dependence, i.e., \( \alpha \)-mixing coefficients of size \(-4d/(d - 1)\) as opposed to \(-3d/(d - 1)\) in West (1996).
Combining (B13) and (B14), we obtain

$$\leq 8 \sum_{i=1}^{t-1} \sum_{j=2}^{t} \|g_i\|_{4d} \|g_j\|_4 \|v_i^2\|_2^{a_{i-j}^{1/4-1/4d}}$$

(Hölder’s inequality)

$$\ll \sum_{i=1}^{t-1} \sum_{j=2}^{t} (j-i)^{-\delta}$$

(Assumption B3)

$$= \sum_{l=1}^{t-1} (t-l)^{-\delta}$$

$$\ll t \sum_{l=1}^{\infty} l^{-\delta}$$

$$\ll t$$

Hence $E[\sum_{i=1}^{t} v_t g_i] \ll t$ and $\|v_t G_t\|_2 \ll t^{-1/2}$. On the other hand, for $i < t$ by covariance inequality and Assumption B3 we have

$$\|E[v_t g_i]\| \leq 8a_{i-i}^{-1/d} \|v_t\|_{2d} \|g_i\|_{2d} \ll (t-i)^{-4\delta}$$

and so

$$\|v_t G_t - E[v_t G_t]\|_2$$

$$\leq \|v_t G_t\|_2 + |E[v_t G_t]|$$

$$\ll t^{-1/2} + t^{-1} \sum_{i=1}^{t} |E[v_t g_i]|$$

$$\ll t^{-1/2} + t^{-1} \sum_{i=1}^{t} i^{-4\delta}$$

$$\ll t^{-1/2}$$

**Step 5:** Combining (B13) and (B14), we obtain

$$\sum_{k=0}^{t} a_k^{-1} \|E[v_t G_t - E[v_t G_t]|\mathcal{F}_{t-k}] - E[v_t G_t - E[v_t G_t]|\mathcal{F}_{t-k-1}]\|^2_2$$

$$\leq \sum_{k=1}^{t} (a_k^{-1} - a_{k-1}^{-1}) \|E[v_t G_t - E[v_t G_t]|\mathcal{F}_{t-k}]\| + a_0^{-1} \|v_t G_t - E[v_t G_t]\|^2_2$$

$$\ll t^{-2} \sum_{k=1}^{t-1} (a_k^{-1} - a_{k-1}^{-1})(t-k)k^{-\delta} + t^{-1} a_0^{-1}$$

$$\ll t^{-1} \left\{ \sum_{k=1}^{t-1} (a_k^{-1} - a_{k-1}^{-1})k^{-2\delta} + a_0^{-1} \right\}$$

and it follows

$$\sum_{k=0}^{\infty} a_k^{-1} E[Y_{1,k}^2] \ll \sum_{t=1}^{T} t^{-1} \left\{ \sum_{k=1}^{t-1} (a_k^{-1} - a_{k-1}^{-1})k^{-2\delta} + a_0^{-1} \right\}$$

Now choose $a_k \equiv k^{-\delta}$ for $k \geq 1$ and $a_0 \equiv 1$. Then $\sum_{k=0}^{\infty} a_k < \infty$ and $a_k^{-1} - a_{k-1}^{-1} \ll k^{\delta-1}$. It
follows
\[
\sum_{k=1}^{t-1} (a_k^{-1} - a_{k-1}^{-1}) k^{-2\delta} + a_0^{-1} \ll 1 + \sum_{k=1}^{t-1} k^{-\delta-1} < \infty
\]
and
\[
\sum_{k=0}^{\infty} a_k^{-1} E[Y_{t,k}^2] \ll \sum_{t=1}^{T} t^{-1} \ll \log(T)
\]
We have thus reached
\[
E \left[ \sum_{1 \leq R \leq T} |S_R|^2 \right] \leq 4 \left( \sum_{k=0}^{\infty} a_k \right) \left( \sum_{k=0}^{\infty} a_k^{-1} E[Y_{t,k}^2] \right) \ll \log(T) = o(T)
\]
whence we conclude \( \max_{1 \leq R \leq T} |S_R|^2 = o_p(T^{1/2}) \).

Lemma B4. Under Assumption B, the following quantities are \( o_p(1) \):

(a) \( \sup_{|IT| \leq R_T \leq T} \left| T^{-1/2} \sum_{t=R_T}^{T} (\partial_\beta f_{t+h}(\beta^*) - F) BG_t \right| = o_p(1) \);

(b) \( \sup_{|IT| \leq R_T \leq T} \left| T^{-1/2} \sum_{t=R_T}^{T} (\partial_\beta f_{t+h}(\beta^*) - F) (B_t - B) G_t \right| = o_p(1) \);

(c) \( \sup_{|IT| \leq R_T \leq T} \left| T^{-1/2} \sum_{t=R_T}^{T} F(B_t - B) G_t \right| = o_p(1) \);

Proof. (a) This follows immediately from Lemma B3.

(b) We have
\[
\sup_{|IT| \leq R_T \leq T} \left| T^{-1/2} \sum_{t=R_T}^{T} (\partial_\beta f_{t+h}(\beta^*) - F) (B_t - B) G_t \right| \leq T^{-1/2} \sum_{t=|IT|}^{T} |(\partial_\beta f_{t+h}(\beta^*) - F) (B_t - B) G_t| = o_p(1)
\]

 Lemma A4(b) of West (1996) means that \( T^{-1/2} \sum_{t=R_T}^{T} |(\partial_\beta f_{t+h}(\beta^*) - F) (B_t - B) G_t| = o_p(1) \) for any \( R_T \) such that \( R_T / T \) converges. In particular, it applies to \( R_T \equiv |IT| \), and so we conclude (b).

(c) The conclusion is also implicitly contained in Lemma A4(c) of West (1996). We have
\[
\sup_{|IT| \leq R_T \leq T} \left| T^{-1/2} \sum_{t=R_T}^{T} F(B_t - B) G_t \right| \leq \sup_{t\geq|IT|} \left| B_t - B \right| ||F|| T^{-1/2} \sum_{t=|IT|}^{T} |G_t|
\]
West (1996) has shown that \( T^{-1/2} \sum_{t=|IT|}^{T} |G_t| = O_p(1) \), and because of \( \sup_{|IT| \leq T} |B_t - B| = o_p(1) \) (by the almost sure convergence) we conclude the proof.

Proof of Proposition 5 (recursive scheme). Given Assumption B1, with probability approaching one we have
\[
f_{t+h}(\hat{\beta}_t) = f_{t+h}(\beta^*) + \partial_\beta f_{t+h}(\beta^*) (\hat{\beta}_t - \beta^*) + r_{t+h}
\]
where the remainder \( r_{t+h} \) has the form
\[
r_{t+h} \equiv \frac{1}{2} (\hat{\beta}_t - \beta^*) \partial_\beta^2 f_{t+h}(\hat{\beta}_t) (\hat{\beta}_t - \beta^*)
\]

57
With such expansion,

$$T^{-1/2} \sum_{t=[\xi]}^{T} (f_{t+h}(\hat{\beta}_t) - \mu_t)$$

$$= T^{-1/2} \sum_{t=[\xi]}^{T} (f_{t+h}(\beta^*) - \mu_t) + T^{-1/2} \sum_{t=[\xi]}^{T} \partial f_{t+h}(\beta^*)(\hat{\beta}_t - \beta^*) + T^{-1/2} \sum_{t=[\xi]}^{T} r_{t+h}$$

We have

$$\sup_{\lfloor l T \rfloor \leq \xi \leq \lfloor u T \rfloor} \left| T^{-1/2} \sum_{t=[\xi]}^{T} r_{t+h} \right| \leq T^{-1/2} \sum_{t=[lT]}^{T} \left| r_{t+h} \right|$$

where West (1996) has implicitly proved that the right hand side is $o_p(1)$ (see proof of his Equation (4.1)). The second term on the right hand side of (B15) can be further transformed to

$$T^{-1/2} \sum_{t=[\xi]}^{T} \partial f_{t+h}(\beta^*)(\hat{\beta}_t - \beta^*)$$

Define

$$M_T(\xi) \equiv T^{-1/2} \sum_{t=[\xi]}^{T} (\partial f_{t+h}(\beta^*) - F)(B_t - B_t)G_t + T^{-1/2} \sum_{t=[\xi]}^{T} (\partial f_{t+h}(\beta^*) - F)BG_t$$

In light of Lemma B4, $\sup_{\lfloor l T \rfloor \leq \xi \leq \lfloor u T \rfloor} |M_T(\xi)| = o_p(1)$, and so we conclude the proof. □

Uniform mean-value expansion: The rolling scheme and the fixed scheme

The proof for the rolling scheme and the fixed scheme proceeds in almost the same way. In particular, we only need to establish the conclusion of Lemma B3 under these schemes, and the proof is still based on maximal inequality. For its similarity, we omit the proof here.

B.4 Proofs for Section 1.3.2

In this proof, we let $f_{t+h} \equiv f_{t+h}(\beta^*)$ for short. The core result for development of treatments for different modes of sample split is the weak convergence established in Proposition
6. We take a different route from West (1996) and McCracken (2000), who calculate first the long-run covariance of the partial sum

\[ T^{-1/2} \sum_{t=\lfloor \xi T \rfloor}^{T} \begin{pmatrix} f_{t+h} - \mu_t \\ G_t \end{pmatrix} \]

for fixed \( \xi \) and then apply a central limit theorem to show that the asymptotic law is a normal distribution with the calculated long-run covariance. In our proof, we establish first weak convergence of the above partial sum process. The limit turns out to be a Gaussian process involving some stochastic integrals, which can be handled by the stochastic Fubini theorem.

The proof of Proposition 6 is divided according to the estimation scheme.

Weak convergence: The recursive scheme

The asymptotics established in this section are under the recursive estimation scheme unless otherwise stated.

Lemma B5. Under Assumption B, we have the weak convergence

\[ T^{-1/2} \sum_{t=\lfloor \xi T \rfloor}^{T} \begin{pmatrix} f_{t+h} - \mu_t \\ G_t \end{pmatrix} \xrightarrow{w} \begin{pmatrix} X_T^{(1)}(1) - X_T^{(1)}(\cdot) \\ f, \rho^{-1} X_T^{(2)}(\cdot) d\rho \end{pmatrix} \]  

(B16)

Proof. Define

\[ X_T(\xi) \equiv \begin{pmatrix} X_T^{(1)}(\xi) \\ X_T^{(2)}(\xi) \end{pmatrix} \equiv \sum_{t=1}^{\lfloor \xi T \rfloor} \begin{pmatrix} T^{-1/2} f_{t+h} - \mu_t \\ T^{-1/2} g_t \end{pmatrix} \]

In light of Corollary 29.14 of Davidson (1994), the sequence \( \{X_T : T \geq 1\} \) converges weakly to \( \Omega^{1/2}W \), where \( W \) is a standard Wiener process. Obviously, \( \Omega^{1/2}W \) has the same law as the process \( X \). Observe that

\[ T^{-1/2} \sum_{t=\lfloor \xi T \rfloor}^{T} \begin{pmatrix} f_{t+h} \\ G_t \end{pmatrix} = \begin{pmatrix} \sum_{t=1}^{T} T^{-1/2} f_{t+h} - \sum_{t=1}^{\lfloor T/\xi \rfloor} T^{-1/2} f_{t+h} \\ \int f, T^{-1/2} G_{[\rho T]} \times T d\rho \end{pmatrix} \]

\[ = \begin{pmatrix} \sum_{t=1}^{T} T^{-1/2} f_{t+h} - \sum_{t=1}^{\lfloor T/\xi \rfloor} T^{-1/2} f_{t+h} \\ \int f, \sum_{t=1}^{\lfloor \rho T \rfloor} T^{-1/2} g_t \times \frac{T}{T} d\rho \end{pmatrix} + \begin{pmatrix} 0 \\ T^{-1/2} G_T \end{pmatrix} \]

\[ = \begin{pmatrix} X_T^{(1)}(1) - X_T^{(1)}(\lfloor T/\xi \rfloor - 1) \\ \int f, X_T^{(2)}(\cdot) \times \frac{T}{T} d\rho \end{pmatrix} + \begin{pmatrix} 0 \\ T^{-1/2} G_T \end{pmatrix} \]

31More precisely, they consider the slightly more general case where \( R_T / T \) converges.
Because $G_T = o_p(1)$ by Assumptions B2 and B3 and the weak law of large number for mixing-gale (see, e.g., Andrews (1988)), it suffices to establish weak convergence of the first term in the above equation. To this end, define a sequence of mapping $\varphi_T : D^{k+1}[I, 1] \rightarrow D^{k+1}[I, 1]$ by

$$\varphi_T(x)(\xi) \equiv \begin{pmatrix} x^{(1)}(1) - x^{(1)}(\frac{\lfloor \xi T \rfloor - 1}{T}) \\ \int_1^{\xi} x^{(2)}(\rho) \times \frac{T}{[\rho T]} d\rho \end{pmatrix}$$

It is easy to see that $\{\varphi_T : T \geq 1\}$ is equicontinuous with respect to the uniform topology. Moreover,

$$\varphi_T(X) \overset{a.s.}{\rightarrow} \begin{pmatrix} X^{(1)}(1) - X^{(1)}(\cdot) \\ \int_1^{1} \rho^{-1} X^{(2)}(\rho) d\rho \end{pmatrix}$$

Applying the extended continuous mapping theorem (see, e.g., Theorem 8.4.1., Bogachev (2007)), we obtain

$$\begin{pmatrix} X^{(1)}_T(1) - X^{(1)}(\frac{\lfloor \xi T \rfloor - 1}{T}) \\ \int_1^{X^{(2)}_T} \times \frac{T}{[\rho T]} d\rho \end{pmatrix} = \varphi_T(X_T) \overset{w}{\rightarrow} \begin{pmatrix} X^{(1)}(1) - X^{(1)}(\cdot) \\ \int_1^{1} \rho^{-1} X^{(2)}(\rho) d\rho \end{pmatrix}$$

The proof is thus completed.

**Remark 11.** The asymptotic limit of $\sum_{t=\lfloor \xi T \rfloor}^{T} G_t$ derived in Rossi and Inoue (2012) is incorrect. In our notation, their proof implicitly claims that the partial sum process $T^{-1/2} \sum_{t=1}^{\lfloor \xi T \rfloor} G_t$ converges to a Wiener process; see the first paragraph in p.451 of Rossi and Inoue (2012). This is incorrect since $\{G_t\}$ itself is not a near-epoch dependent sequence, so the FCLT of Wooldridge and White (1988), which they invoke, cannot be applied directly to $\{G_t\}$.

**Lemma B6.** The following equality holds up to indistinguishability of the processes:

$$\int_1^{\xi} \rho^{-1} X^{(2)}(\rho) d\rho = - \log(\xi) X^{(2)}(\xi) - \int_1^{\xi} \log(\rho) dX^{(2)}(\rho)$$

(B17)

**Proof.** The proof is a simple application of the stochastic Fubini theorem (see, e.g., Veraar (2012)):
Because the above equation holds for each \( \xi \) and the processes are continuous, we deduce that the equality is up to indistinguishability.

**Lemma B7.** Define

\[
Y(\xi) \equiv \begin{pmatrix} X^{(1)}(1) - X^{(1)}(\xi) \\ \int_{\xi}^{1} \rho^{-1} X^{(2)}(\rho) d\rho \end{pmatrix}, \quad l \leq \xi \leq 1
\]  

(B18)

Then \( Y \) is a zero-mean Gaussian process and for each \( \xi \),

\[
\text{var}(Y(\xi)) = \begin{pmatrix} (1 - \xi) v_{ff} + (1 - \xi) + \xi \log(\xi) \right) v_{fg} \\ (1 - \xi) + \xi \log(\xi) v'_{fg} + 2(1 - \xi) + \xi \log(\xi) v_{gg} \end{pmatrix}
\]  

(B19)

**Proof.** In view of Lemma B6,

\[
Y(\xi) = \begin{pmatrix} \int_{\xi}^{1} dX^{(1)}(r) \\ \int_{\xi}^{1} \log(r) dX^{(2)}(r) \end{pmatrix} - \begin{pmatrix} 0 \\ \log(\xi) X^{(2)}(1) \end{pmatrix} = \int_{\xi}^{1} \varphi(r) dX(r) - \int_{0}^{\xi} \eta(\xi) dX(r)
\]

where

\[
\varphi(r) \equiv \begin{pmatrix} 1 & 0 \\ 0 & -\log(r) I_k \end{pmatrix}, \quad \eta(\xi) \equiv \begin{pmatrix} 0 & 0 \\ 0 & -\log(\xi) I_k \end{pmatrix}
\]

To establish that \( Y \) is Gaussian, we need to show that its finite dimensional distributions are Gaussian. For notational ease, we show that \( (Y(\xi_1), Y(\xi_2)) \) is Gaussian, with \( \xi_1 < \xi_2 \). Let \( \lambda_1 \) and \( \lambda_2 \) be two \((k + 1)\)-dimensional vector. In light of the Crámer-Wold device, it suffices to show that \( \lambda'_1 Y(\xi_1) + \lambda'_2 Y(\xi_2) \) is a Gaussian random variable. We have

\[
\lambda'_1 Y(\xi_1) + \lambda'_2 Y(\xi_2) = \lambda'_1 \int_{\xi_1}^{1} \varphi(r) dX(r) + \int_{0}^{\xi_1} \lambda'_1 \eta(\xi_1) dX(r) + \int_{\xi_1}^{1} \lambda'_2 \varphi(r) dX(r) + \int_{0}^{\xi_2} \lambda'_2 \eta(\xi_2) dX(r)
\]

\[
= \int_{0}^{\xi_1} (\lambda'_1 \eta(\xi_1) + \lambda'_2 \eta(\xi_2)) dX(r) + \int_{\xi_1}^{\xi_2} (\lambda'_1 \varphi(r) + \lambda'_2 \varphi(r)) dX(r)
\]

Because the integrands are deterministic, the three terms on the right hand side of the above equation are all Gaussian random variable. Moreover, they are independent, whence \( \lambda'_1 Y(\xi_1) + \lambda'_2 Y(\xi_2) \) is a Gaussian random variable. This implies that \( Y \) is a Gaussian process.

That \( Y(\xi) \) has zero mean is evident. We are left to derive the variance of \( Y(\xi) \). First,

\[
\mathbb{E}[ (X^{(1)}(1) - X^{(1)}(\xi))(X^{(1)}(1) - X^{(1)}(\xi))' ] = (1 - \xi) v_{ff}
\]

61
For the covariance, we have
\[
E \left[ \left( X^{(1)}(1) - X^{(1)}(\xi) \right) \left( \int_\xi^1 \rho^{-1} X^{(2)}(\rho) d\rho \right) \right]'' = E \left[ \left( X^{(1)}(1) - X^{(1)}(\xi) \right) \left( -\log(\xi) X^{(2)}(\xi) - \int_\xi^1 \log(r) dX^{(2)}(r) \right) \right]''
\]
\[
= -E \left[ \left( X^{(1)}(1) - X^{(1)}(\xi) \right) \log(\xi) X^{(2)}(\xi) \right] - E \left[ \left( X^{(1)}(1) - X^{(1)}(\xi) \right) \int_\xi^1 \log(r) dX^{(2)}(r) \right]
\]
By independence, the first term on the right hand side is equal to zero. Denoting by \langle \cdot, \cdot \rangle the predictable quadratic covariation of two processes, the second term can be calculated by
\[
E \left[ \left( X^{(1)}(1) - X^{(1)}(\xi) \right) \int_\xi^1 \log(r) dX^{(2)}(r) \right] = E \left[ \int_\xi^1 dX^{(1)}(r) \int_\xi^1 \log(r) dX^{(2)}(r) \right]
\]
\[
= E \left[ \int_\xi^1 \log(r) d\langle X^{(1)}, X^{(2)} \rangle \right]
\]
\[
= \int_\xi^1 \log(r) dr \times v_{fg}
\]
Hence
\[
E \left[ \left( X^{(1)}(1) - X^{(1)}(\xi) \right) \left( \int_\xi^1 \rho^{-1} X^{(2)}(\rho) d\rho \right) \right] = (\xi \log(\xi) + (1 - \xi)) v_{fg}
\]
We are left to compute the variance of \( \int_\xi^1 \rho^{-1} X^{(2)}(\rho) d\rho \). Again we resort to the equality established in Lemma B6:
\[
\int_\xi^1 \rho^{-1} X^{(2)}(\rho) d\rho = -\log(\xi) X^{(2)}(\xi) - \int_\xi^1 \log(r) dX^{(2)}(r)
\]
The first term is Gaussian random variable with covariance matrix \( \xi \log^2(\xi) v_{gg} \); the second term is also Gaussian random variable, since the integrand is deterministic, and its covariance matrix is \( \int_\xi^1 \log^2(r) dr v_{gg} \). Applying integration by part,
\[
\int_\xi^1 \log^2(r) dr = -\xi \log^2(\xi) + 2\xi \log(\xi) + 2(1 - \xi)
\]
Finally, the two terms are independent. Summing up the covariance matrices, we obtain
\[
\int_\xi^1 \rho^{-1} X(\rho) d\rho \sim N(0, 2((1 - \xi) + \xi \log(\xi)) v_{gg})
\]
The proof is thus complete.
Proof of Proposition 6 (recursive scheme). By Proposition 5,

\[ T^{-1/2} \sum_{t=1}^{T} f_{t+h}(\hat{\beta}_t) = (1, FB) T^{-1/2} \sum_{t=\lfloor \xi T \rfloor}^{T} \left( f_{t+h} \right) G_t + M_T(\xi) \]

In view of Lemma B5 and continuous mapping theorem,

\[ (1, FB) T^{-1/2} \sum_{t=\lfloor \xi T \rfloor}^{T} \left( f_{t+h} \right) \xrightarrow{w} X^{(1)}(1) - X^{(1)}(\cdot) - \int_{\rho}^{1} \rho^{-1} X^{(2)}(\rho) d\rho \]

On the other hand, \( \sup_{\xi \leq T} M_T(\xi) = o_P(1) \) established in Proposition 5. Thus

\[ T^{-1/2} \sum_{t=\lfloor \xi T \rfloor}^{T} (f_{t+h}(\hat{\beta}_t) - \mu_t) \xrightarrow{w} X^{(1)}(1) - X^{(1)}(\cdot) - \int_{\rho}^{1} \rho^{-1} X^{(2)}(\rho) d\rho \]

The proof is then completed. \( \square \)

Weak convergence: The rolling scheme

Next we turn to the proof for the rolling scheme. The proof is almost the same as for the recursive scheme.

Lemma B8. Under Assumptions A1-A3, we have the weak convergence

\[ T^{-1/2} \sum_{t=\lfloor \xi T \rfloor}^{T} \left( f_{t+h} - \mu_t \right) G_t \xrightarrow{w} \begin{pmatrix} X^{(1)}(1) - X^{(1)}(\cdot) \\ X^{(2)}(\rho) - X^{(2)}(\rho - \xi) \end{pmatrix} \]

(B20)

Proof. As in the proof of Lemma B5, we define

\[ X_T(\xi) \equiv \begin{pmatrix} X^{(1)}_T(\xi) \\ X^{(2)}_T(\xi) \end{pmatrix} \equiv \sum_{t=1}^{\lfloor T/\xi \rfloor} \begin{pmatrix} T^{-1/2}(f_{t+h} - \mu_t) \\ T^{-1/2} G_t \end{pmatrix} \]

which converges weakly to the process \( X \equiv \Omega^{1/2} W \). Observe the representation

\[ T^{-1/2} \sum_{t=\lfloor \xi T \rfloor}^{T} \left( f_{t+h} - \mu_t \right) G_t = \begin{pmatrix} \sum_{t=1}^{T} T^{-1/2}(f_{t+h} - \mu_t) - \sum_{t=1}^{\lfloor T/\xi \rfloor} T^{-1/2}(f_{t+h} - \mu_t) \\ \int_{\rho}^{1} T^{-1/2} G_{[\rho T]} \times T d\rho \end{pmatrix} + \begin{pmatrix} 0 \\ T^{-1/2} G_T \end{pmatrix} \]

which converges to

\[ \begin{pmatrix} T^{-1/2} G_{[\rho T]} \times T d\rho \\ T^{-1/2} G_T \end{pmatrix} \]

The proof is concluded by invoking continuous mapping theorem and Slutsky’s lemma. \( \square \)
Lemma B9. The following equalities hold up to indistinguishability of the processes:

\[
\int_{\xi}^{1} X^{(2)}(\rho) d\rho = \int_{0}^{1} (1 - \xi \lor r) dX^{(2)}(r) = X^{(2)}(1) - \xi X^{(2)}(\xi) - \int_{\xi}^{1} r dX^{(2)}(r)
\]

\[
\int_{\xi}^{1} X^{(2)}(\rho - \xi) d\rho = \int_{0}^{1-\xi} (1 - \xi - r) dX^{(2)}(r) = (1 - \xi) X^{(2)}(1 - \xi) - \int_{0}^{1-\xi} r dX^{(2)}(r)
\]

Proof. Again we simply need to employ the stochastic Fubini theorem. For the first integral,

\[
\int_{\xi}^{1} X^{(2)}(\rho) d\rho = \int_{\xi}^{1} \int_{0}^{\rho} dX^{(2)}(r) d\rho = \int_{0}^{1} \int_{\xi \lor r}^{1} d\rho dX^{(2)}(r) = \int_{0}^{1} (1 - \xi \lor r) dX^{(2)}(r) = X^{(2)}(1) - \xi X^{(2)}(\xi) - \int_{\xi}^{1} r dX^{(2)}(r)
\]

For the second integral, we first apply the change of time (see, e.g., Proposition (1.4), Chapter V of Revuz and Yor (2013)):

\[
\int_{\xi}^{1} X^{(2)}(\rho - \xi) d\rho = \int_{0}^{1-\xi} X^{(2)}(\rho) d\rho
\]

The stochastic integral on the right hand side can then be handled by stochastic Fubini theorem:

\[
\int_{0}^{1-\xi} X^{(\rho)} d\rho = \int_{0}^{1-\xi} \int_{0}^{\rho} dX^{(2)}(r) d\rho = \int_{0}^{1-\xi} \int_{r}^{1-\xi} d\rho dX^{(2)}(r) = \int_{0}^{1-\xi} (1 - \xi - r) dX^{(2)}(r) = (1 - \xi) X^{(2)}(1 - \xi) - \int_{0}^{1-\xi} r dX^{(2)}(r)
\]

The proof is thus complete.

Lemma B10. Define

\[
Y(\xi) = \begin{pmatrix}
X^{(1)}(1) - X^{(1)}(\xi) \\
\int_{\xi}^{1} \xi^{-1} (X^{(2)}(\rho) - X^{(2)}(\rho - \xi)) d\rho
\end{pmatrix}
\]
Then $Y$ is a zero-mean Gaussian process. Moreover, for each $\xi \geq 1/2$

$$\text{var}(Y(\xi)) = \begin{pmatrix} (1 - \xi) v_{ff} & \frac{(1-\xi)^2}{2\xi} v_{fg} \\ \frac{(1-\xi)^2}{2\xi} v_{fg}' & \left( \frac{1-\xi^2}{\xi} - \frac{(1-\xi)^2}{3\xi^2} \right) v_{gg} \end{pmatrix}$$  \hspace{1cm} (B23)

while for $l \leq \xi < 1/2$,

$$\text{var}(Y(\xi)) = \begin{pmatrix} (1 - \xi) v_{ff} & \left(1 - \frac{3\xi}{2}\right) v_{fg} \\ \left(1 - \frac{3\xi}{2}\right) v_{fg}' & \left(1 - \frac{4\xi}{3}\right) v_{gg} \end{pmatrix}$$  \hspace{1cm} (B24)

**Proof.** That $Y$ is a zero-mean Gaussian process can be established as under the recursive estimation scheme. Consider the covariance matrix. Using the equalities proved in Lemma B9, we have

$$\text{cov}(X^{(1)}(1) - X^{(1)}(\xi), \int_0^1 \xi^{-1}(X^{(2)}(\rho) - X^{(2)}(\rho - \xi)) d\rho)$$

$$= \mathbb{E} \left[ \left( X^{(1)}(1) - X^{(1)}(\xi) \right) \int_0^1 \xi^{-1}(X^{(2)}(\rho) - X^{(2)}(\rho - \xi)) d\rho \right]$$

$$= \mathbb{E} \left[ \int_0^1 1(\rho \geq \xi) \xi^{-1}(1 - \xi \vee r) - (1 - \xi - r) 1(r \leq 1 - \xi) dX^{(2)}(r) \right]$$

$$= \mathbb{E} \left[ \int_0^1 1(\rho \geq \xi) \xi^{-1}(1 - \xi \vee r) - (1 - \xi - r) 1(r \leq 1 - \xi) d\xi \right]$$

$$= \mathbb{E} \left[ \int_\xi^1 \xi^{-1}(1 - \rho - \xi) 1(\xi \leq r \leq 1 - \xi) d\rho \times v_{fg} \right]$$

$$= \left( \left( \xi^{-1}(1 - \rho - \xi) \right) - \int_\xi^1 \xi^{-1}(1 - \xi - r) 1(\xi \leq r \leq 1 - \xi) d\rho \right) \times v_{fg}$$

$$= \left( \int_\xi^1 \xi^{-1}(1 - \rho - \xi) - \frac{(1 - \xi^2)}{2\xi} - 1(\xi < 1/2) \frac{(2\xi - 1)^2}{2\xi} \right) \times v_{fg}$$

It is easy to check that the covariance coincides with the stated one. The proof for the variances is similar and hence omitted. \hfill \Box

**Proof of Proposition 6 (rolling scheme).** Similar to the proof for the recursive scheme. \hfill \Box

**Weak convergence: The fixed scheme**

We are left to deal with the fixed scheme, which is simpler as no stochastic integral gets involved.

**Lemma B11.** Under Assumptions A1-A3, we have the weak convergence

$$T^{-1/2} \sum_{t=\lceil T \rceil}^{T} \left( \frac{f_{t+h}}{G_t} \right) \xrightarrow{w} \left( X^{(1)}(1) - X^{(1)}(\cdot) \right)$$  \hspace{1cm} (B25)
where \( \iota(\xi) \equiv \xi \) is the identity mapping.

**Proof.** Again define
\[
X_T(\xi) \equiv \left(\begin{array}{c}
X^{(1)}_T(\xi) \\
X^{(2)}_T(\xi)
\end{array}\right) = \sum_{t=1}^{[\xi T]} \left( T^{-1/2} f_{t+h} \right) \left( T^{-1/2} g_t \right)
\]
which converges weakly to \( X \equiv \Omega^{1/2} W \). Observe that
\[
T^{-1/2} \sum_{t=[\cdot T]}^{T} f_{t+h} G_t = \left( \sum_{t=1}^{T} T^{-1/2} f_{t+h} - \sum_{t=1}^{[\cdot T]} T^{-1/2} f_{t+h} \right) = \left( X^{(1)}_T(1) - X^{(1)}_T(\lfloor \cdot T \rfloor - 1) \right)
\]
The proof is concluded by application of the continuous mapping theorem. \(\square\)

**Lemma B12.** Define
\[
Y(\xi) \equiv \left(\begin{array}{c}
X^{(1)}(1) - X^{(1)}(\xi) \\
\frac{1-\xi}{\xi} X^{(2)}(\xi)
\end{array}\right), \quad l \leq \xi \leq 1 \quad (B26)
\]
Then \( Y \) is a zero-mean Gaussian process and for each \( \xi \),
\[
\text{var}(Y(\xi)) = \left(\begin{array}{cc}
(1-\xi) v_{ff} & 0 \\
0 & \frac{(1-\xi)^2}{\xi} v_{gg}
\end{array}\right) \quad (B27)
\]
**Proof.** That \( Y \) is a zero-mean Gaussian process can be proved by the same argument in Lemma B7. The variances of \( X^{(1)}(1) - X^{(1)}(\xi) \) and \( \frac{1-\xi}{\xi} X^{(2)}(\xi) \) can be computed by definition of the Gaussian process \( X \). The zero-covariance is due to independence of \( X^{(1)}(1) - X^{(1)}(\xi) \) and \( X^{(2)}(\xi) \). \(\square\)

**Proof of Proposition 6 (fixed scheme).** Similar to the proof for the recursive scheme. \(\square\)

### B.5 Proofs for Section 1.4.2

**Lemma B13.** Let \( Z \) be given by (1.12). Then the process \( L(\xi) \equiv Z(1-\xi), \quad 0 \leq \xi \leq (1-l) \) is a time-changed Wiener process starting at 0. Moreover, the quadratic variation of \( L \) is \( \langle L, L \rangle(\xi) \equiv V(1-\xi) \), where \( V(\xi) \) is defined by (1.13).

**Proof.** The claim is equivalent to that \( Z \) is a zero-mean Lévy Gaussian martingale starting at zero, and we will prove the latter. We only consider the recursive scheme, and proof under the other two schemes are almost the same. Obviously \( L(0) = Z(1) = 0 \). \( Z \) has been shown to be a zero-mean Gaussian process; then \( L \) is a zero-mean Gaussian process as well, since whether or not a process is Gaussian depends solely on its finite dimensional distributions. We are simply left to show that \( L \) has independent increments, which together with the zero-mean property implies the martingale property. Consider \( 0 \leq \xi < 1-l \) and arbitrary \( 0 < \delta < \ldots \)
1 − l − ξ. Then

\[ L(\xi + \delta) - L(\xi) = X^{(1)}(1 - \xi) - X^{(1)}(1 - \xi - \delta) - FB + \int_{1-\xi-\delta}^{1-\xi} \rho^{-1} X^{(2)}(\rho) d\rho \]

which is independent of any \( L(\eta) \) with \( \eta < \xi \). Thus we deduce that \( L \) is a zero-mean Lévy Gaussian martingale starting at 0.

Because \( L \) is a zero-mean Lévy Gaussian martingale, its predictable quadratic variation coincides with its variance process, that is,

\[ \langle L, L \rangle (\xi) = \text{var}(L(\xi)) = \text{var}(Z(1 - \xi)) = V(1 - \xi) \]

whence we conclude the proof. \( \square \)

**Proof of Proposition 7.** Let \( v(\xi) \equiv \langle L, L \rangle (\xi) = V(1 - \xi) \). We shall use \( v^{-1} \) to denote the inverse of \( v \). Following Lemma B13, the time-changed process \( W \equiv L \circ v^{-1} \) is standard Wiener process. Hence

\[
\mathbb{P}\left( \sup_{l \leq \xi \leq u} Z(\xi) \leq z \right) = \mathbb{P}\left( \sup_{1-u \leq \xi \leq 1-l} L(\xi) \leq z \right) = \mathbb{P}\left( \sup_{v(1-u) \leq \xi \leq v(1-l)} L \circ v^{-1}(\xi) \leq z \right) = \mathbb{P}\left( \sup_{v(1-u) \leq \xi \leq v(1-l)} W(\xi) \leq z \right)
\]

\[
= \int_{-\infty}^{z} \mathbb{P}\left( \sup_{0 \leq \xi \leq v(1-l) - v(1-u)} W(\xi) \leq z - w \right) p(w; v(1-u)) d\omega
= \int_{-\infty}^{z} \Phi\left( \frac{z - w}{\sqrt{v(1-l) - v(1-u)}} \right) p(w; v(1-u)) d\omega
= \int_{-\infty}^{z} \Phi\left( \frac{z - w}{\sqrt{V(l) - V(u)}} \right) p(w; V(u)) d\omega
\]

where the fourth equality is due to Markov property of the Wiener process and the fifth equality to the coincidence in distribution of the running supremum of Wiener process and its absolute terminal value (see, e.g., Proposition 3.7, Revuz and Yor (2013)).

\( (b) \) The same trick can be applied to establish (1.17). First,

\[
\mathbb{P}\left( \sup_{l \leq \xi \leq u} |Z(\xi)| \leq z \right) = \mathbb{P}\left( \sup_{1-u \leq \xi \leq 1-l} |L(\xi)| \leq z \right)
\]
B. Proofs

\begin{equation}
\mathbb{P} \left( \sup_{v(1-u) \leq v(1-l)} |L \circ v^{-1}(\xi)| \leq z \right)
\end{equation}

\begin{equation}
= \mathbb{P} \left( \sup_{v(1-u) \leq v(1-l)} |W(\xi)| \leq z \right)
\end{equation}

\begin{equation}
= \int_{-z}^{z} \mathbb{P} \left( \sup_{0 \leq \xi \leq v(1-l) - v(1-u)} W(\xi) \leq z - w, \inf_{0 \leq \xi \leq v(1-l) - v(1-u)} W(\xi) \geq -z - w \right) p(w; v(1-u)) dw
\end{equation}

Next, we invoke Lévy’s triple law (see, e.g., Theorem 6.18, Schilling and Partzsch (2014)) to compute the probability appearing in the integral:

\begin{equation}
\mathbb{P} \left( \sup_{0 \leq \xi \leq v(1-l) - v(1-u)} W(\xi) \leq z - w, \inf_{0 \leq \xi \leq v(1-l) - v(1-u)} W(\xi) \geq -z - w \right)
\end{equation}

\begin{equation}
= \int_{-z-w}^{z-w} \frac{1}{\sqrt{2\pi(v(1-l) - v(1-u))}} \sum_{n=-\infty}^{\infty} \left[ \exp \left( -\frac{(x+4nz)^2}{2(v(1-l) - v(1-u))} \right) - \exp \left( -\frac{(x+2z+2w-4nz)^2}{2(v(1-l) - v(1-u))} \right) \right] dx
\end{equation}

\begin{equation}
= \sum_{n=-\infty}^{\infty} q_n(z, w)
\end{equation}

Substituting the above formula back to the integral leads to the desired result. \qed

B.6 Proofs for Section 1.4.3

To establish (1.21) we introduce two lemmas. The first lemma is due to Rao (1962).

**Lemma B14.** Suppose \( \mu \) is a Borel measure on the \( k \)-dimensional Euclidean space \( \mathbb{R}^k \) such that every convex set of \( \mathbb{R}^k \) has \( \mu \)-null boundary. Then a sequence of measures \( \{\mu_n\} \) converges weakly to \( \mu \) if and only if

\begin{equation}
\sup_{C \in \mathcal{C}} |\mu_n(C) - \mu(C)| \to 0
\end{equation}

where \( \mathcal{C} \) denotes the class of all measurable convex sets. In particular, the uniform convergence in (B28) holds if \( \mu \) is absolutely continuous with respect to the Lebesgue measure.

**Proof.** See Theorem 4.2, Rao (1962). \qed

The second lemma concerns “continuity” of the zero \( \zeta \).

**Lemma B15.** Suppose that \( \hat{V}_T(l) \xrightarrow{p} V(l) \) and \( \hat{V}_T(u) \xrightarrow{p} V(u) \). Then \( \zeta(\hat{V}_T(l), \hat{V}_T(u)) \xrightarrow{p} \zeta(V(l), V(u)) \) as \( T \to \infty \).

**Proof.** We use the usual subsequence argument to reduce the proof to pointwise convergence. We simply need to show that every subsequence of \( \{\zeta(\hat{V}_T(l), \hat{V}_T(u)) : T \geq 1\} \) has a further subsequence converging almost surely to \( \zeta(V(l), V(u)) \). For notational ease, we still use
\{\zeta(\hat{V}_T(l), \hat{V}_T(u))\} to denote a subsequence. Because \(\hat{V}_T(l)\) and \(\hat{V}_T(u)\) consistently estimate \(V(l)\) and \(V(u)\) respectively, passing to subsequences we assume without loss of generality that \(\hat{V}_T(l) \xrightarrow{a.s.} V(l)\) and \(\hat{V}_T(u) \xrightarrow{a.s.} V(u)\). With almost sure convergence, we may simply treat all quantities as nonrandom.

We claim that \(\liminf_T \zeta(V_T(l), V_T(u)) \geq \zeta(V(l), V(u))\) holds. Otherwise for some \(\varepsilon_0 > 0\), \(\zeta(\hat{V}_T(l), \hat{V}_T(u)) < \zeta(V(l), V(u)) - \varepsilon_0\) along a subsequence \(\{\hat{V}_T_k(l), \hat{V}_T_k(u) : k \geq 1\}\). If this were the case, we would have

\[
1 - \alpha = \limsup_k \int_{-\infty}^{\zeta(\hat{V}_T_k(l), \hat{V}_T_k(u))} \Phi\left(\frac{\zeta(\hat{V}_T_k(l), \hat{V}_T_k(u)) - w}{\sqrt{\hat{V}_T_k(l) - \hat{V}_T_k(u)}}\right)p(w; \hat{V}_T_k(u)) \, dw
\leq \limsup_k \int_{-\infty}^{\zeta(V(l), V(u)) - \varepsilon_0} \Phi\left(\frac{\zeta(V(l), V(u)) - \varepsilon_0 - w}{\sqrt{V(l) - V(u)}}\right)p(w; V(u)) \, dw
= \int_{-\infty}^{\zeta(V(l), V(u)) - \varepsilon_0} \Phi\left(\frac{\zeta(V(l), V(u)) - \varepsilon_0 - w}{\sqrt{V(l) - V(u)}}\right)p(w; V(u)) \, dw
< \int_{-\infty}^{\zeta(V(l), V(u))} \Phi\left(\frac{\zeta(V(l), V(u)) - w}{\sqrt{V(l) - V(u)}}\right)p(w; V(u)) \, dw
= 1 - \alpha
\]

where the first line is by definition of \(\zeta(\hat{V}_T_k(l), \hat{V}_T_k(u))\), the second line by monotonicity and positivity of the integrand, the third line by dominated convergence theorem, the fourth line by strict positivity of the integrand, and the fifth line by definition of \(\zeta(V(l), V(u))\). This is clearly a contradiction, so it must be that \(\liminf_T \zeta(V_T(l), V_T(u)) \geq \zeta(V(l), V(u))\). By the same argument we can prove \(\limsup_T \zeta(V_T(l), V_T(u)) \leq \zeta(V(l), V(u))\), whence we conclude the proof.

Now we are ready to justify the testing procedures.

**Proof of Proposition 8.** The sets defined in (1.20) are clearly convex. The sequence \(\sup_{t \leq \xi \leq u} Z_T(\xi) : T \geq 1\) converges in distribution to \(\sup_{t \leq \xi \leq u} Z(\xi)\), and by Proposition 7 the distribution of \(\sup_{t \leq \xi \leq u} Z(\xi)\) is absolutely continuous with respect to the Lebesgue measure. With the class \(\mathcal{C}\) defined in Lemma B14,

\[
\left| P\left( \sup_{t \leq \xi \leq u} Z_T(\xi) \in A_T \right) - P\left( \sup_{t \leq \xi \leq u} Z(\xi) \in A \right) \right|
\leq \left| P\left( \sup_{t \leq \xi \leq u} Z_T(\xi) \in A_T \right) - P\left( \sup_{t \leq \xi \leq u} Z(\xi) \in A_T \right) \right| + \left| P\left( \sup_{t \leq \xi \leq u} Z(\xi) \in A_T \right) - P\left( \sup_{t \leq \xi \leq u} Z(\xi) \in A \right) \right|
\leq \sup_{C \in \mathcal{C}} \left| P\left( \sup_{t \leq \xi \leq u} Z_T(\xi) \in C \right) - P\left( \sup_{t \leq \xi \leq u} Z(\xi) \in C \right) \right| + \left| P\left( \sup_{t \leq \xi \leq u} Z(\xi) \in A_T \right) - P\left( \sup_{t \leq \xi \leq u} Z(\xi) \in A \right) \right|
\leq \sup_{C \in \mathcal{C}} \left| P\left( \sup_{t \leq \xi \leq u} Z_T(\xi) \in C \right) - P\left( \sup_{t \leq \xi \leq u} Z(\xi) \in C \right) \right| + \left| P\left( \sup_{t \leq \xi \leq u} Z(\xi) \in A_T \Delta A \right) \right|
\]

(B29)
where \( A_T \Delta A \) is the symmetric difference of \( A_T \) and \( A \). Lemma B14 entails convergence to zero of the first term on the right hand side of (B29). For the second term, let \( B_\epsilon \) denote the open ball centered at \( \xi(V(l), V(u)) \) with radius \( \epsilon \). Then

\[
\mathbb{P}\left( \sup_{l \leq \xi \leq u} Z(\xi) \in A_T \Delta A \right) = \mathbb{P}\left( \xi(\hat{V}_T(l), \hat{V}_T(u)) \wedge \xi(V(l), V(u)) \leq \sup_{l \leq \xi \leq u} Z(\xi) \leq \xi(\hat{V}_T(l), \hat{V}_T(u)) \vee \xi(V(l), V(u)) \right) \leq \mathbb{P}\left( \sup_{l \leq \xi \leq u} Z(\xi) \in B_\epsilon \right) + \mathbb{P}\left( \left| \xi(\hat{V}_T(l), \hat{V}_T(u)) - \xi(V(l), V(u)) \right| > \epsilon \right)
\]

Applying Lemma B15, as \( T \) tends to infinity we have

\[
\limsup_T \mathbb{P}\left( \sup_{l \leq \xi \leq u} Z(\xi) \in A_T \Delta A \right) \leq \mathbb{P}\left( \sup_{l \leq \xi \leq u} Z(\xi) \in B_\epsilon \right)
\]

Further letting \( \epsilon \to 0 \), we obtain \( \lim_T \mathbb{P}\left( \sup_{l \leq \xi \leq u} Z(\xi) \in A_T \Delta A \right) = 0 \). Now from (B29) we deduce

\[
\lim_T \mathbb{P}\left( \sup_{l \leq \xi \leq u} Z_T(\xi) \in A_T \right) = \mathbb{P}\left( \sup_{l \leq \xi \leq u} Z(\xi) \in A \right) = \alpha
\]

concluding the proof.

B.7 Proofs for Section 1.5

Proofs for the case where \( F = 0 \) and \( \mu_t = \frac{c}{\sqrt{t}} \) are based on the following two lemmas.

**Lemma B16** (One-sided BCP). Let \( W \) be an one-dimensional standard Wiener process and denote by \( \Phi \) the standard normal c.d.f. Consider the linear boundary \( a + bt \), where \( a > 0 \). Denote by \( P_1(a, b, t) \) the probability that \( W \) crosses the boundary before or at \( t \), that is,

\[
P_1(a, b, t) \equiv \mathbb{P}\left( T^* \leq t \right), \text{ where } T^* \equiv \inf\{u : W(u) = a + bu\}
\]

The crossing probability can be computed by the following formula:

\[
P_1(a, b, t) = \Phi\left( \frac{-a - bt}{\sqrt{t}} \right) + \exp(-2ba)\Phi\left( \frac{-a + bt}{\sqrt{t}} \right) \tag{B30}
\]

**Proof.** Equation (B30) is the well-known Bachelier-Lévy formula, whose proof can be found in standard textbooks (see, e.g., p. 147, Jeanblanc et al. (2009)).

**Lemma B17** (Two-sided BCP). Let \( W \) be an one-dimensional standard Wiener process and denote by \( \Phi \) the standard normal c.d.f. Consider two linear boundaries \( a_1 + bt \) and \( a_2 + bt \), where \( a_1 \) > \( a_2 \). Denote by \( P_2(a_1, a_2, b, t) \) the probability that \( W \) crosses either of the boundaries

\[
P_2(a_1, a_2, b, t) \equiv \mathbb{P}\left( \min\{T_1, T_2\} \leq t \right), \text{ where } T_1 \equiv \inf\{u : W(u) = a_1 + bu\}, T_2 \equiv \inf\{u : W(u) = a_2 + bu\}
\]

The crossing probability can be computed by the following formula:

\[
P_2(a_1, a_2, b, t) = \exp(-2ba)\Phi\left( \frac{-a + bt}{\sqrt{t}} \right) + \Phi\left( \frac{-a - bt}{\sqrt{t}} \right) - \exp(-2ba)\Phi\left( \frac{-a + bt}{\sqrt{t}} \right) \tag{B31}
\]

**Proof.** Equation (B31) is the well-known Bachelier-Lévy formula, whose proof can be found in standard textbooks (see, e.g., p. 147, Jeanblanc et al. (2009)).
Chapter 1. Data-Driven Sample Split and Out-of-Sample Tests

before or at \(t\), that is,

\[
P_2(a_1, a_2, b, t) \equiv P(T^* \leq t), \text{ where } T^* \equiv \inf\{u : W(u) = a_1 + bu\} \wedge \inf\{u : W(u) = a_2 + bu\}
\]

The crossing probability can be computed by the formula below:

\[
P_2(a_1, a_2, b, t) = \sum_{j=0}^{\infty} (-1)^j \exp(-a_1 b) \left[ \exp(-r_j b) \Phi \left( \frac{-r_j + bt}{\sqrt{t}} \right) + \exp(r_j b) \Phi \left( \frac{-r_j + bt}{\sqrt{t}} \right) \right]
\]

\[
+ \sum_{j=0}^{\infty} (-1)^j \exp(-a_2 b) \left[ \exp(s_j b) \Phi \left( \frac{s_j + bt}{\sqrt{t}} \right) + \exp(-s_j b) \Phi \left( \frac{s_j + bt}{\sqrt{t}} \right) \right]
\]

where

\[
r_j \equiv \begin{cases} j(a_1 - a_2) + a_1, & j \text{ even} \\ j(a_1 - a_2) - a_2, & j \text{ odd} \end{cases}, \quad s_j \equiv \begin{cases} j(a_1 - a_2) - a_2, & j \text{ even} \\ j(a_1 - a_2) + a_1, & j \text{ odd} \end{cases}
\]

\[
(B31)
\]

**Proof.** Formula for the two-sided BCP is less known; it is perhaps first established by Anderson (1960). Here (B31) is a simplification due to Hall (1997).

**Proof of Proposition 10.** (a) Proof for the conventional test is straightforward and omitted. For the sup-type test, by Proposition 9 and Markov property of Wiener process we have

\[
P\left( \sup_{t \leq \xi \leq u} S(\xi) \leq x \right)
\]

\[
= P\left( \sup_{t(1-u) \leq \xi \leq t(1-u)} \left( W(\xi) + cv^{-1}(\xi) \right) \leq x \right)
\]

\[
= P\left( \sup_{t(1-u) \leq \xi \leq t(1-u)} \left( W(\xi) - W(t(1-u)) + c(v^{-1}(\xi) - (1-u)) \right) \leq x - W(t(1-u)) - c(1-u) \right)
\]

\[
= \int_{-\infty}^{x-c(1-u)} p(w; t(1-u))
\]

\[
\times P\left( \sup_{t(1-u) \leq \xi \leq t(1-u)} \left( W(\xi) - W(t(1-u)) + c(v^{-1}(\xi) - (1-u)) \right) \leq x - w - c(1-u) \right) \, dw
\]

\[
= \int_{-\infty}^{x-c(1-u)} p(w; t(1-u)) P\left( \sup_{0 \leq \xi \leq v_f(t(1-u))} \left( W(\xi) + \frac{c}{v_f} \xi \right) \leq a(w) \right) \, dw
\]

\[
= \int_{-\infty}^{x-c(1-u)} p(w; v_f(t(1-u))) \left[ 1 - P\left( a(w), -\frac{c}{v_f}, v_f(t(1-u)) \right) \right] \, dw
\]

\[
(B32)
\]

(b) Proof for the conventional test is simple and hence omitted. Proof for the sup-type test is similar to that of (a); by Proposition 9 and Markov property of Wiener process,

\[
P\left( \sup_{t \leq \xi \leq u} |S(\xi)| \leq x \right)
\]
\[ \mathbb{P} \left( \sup_{t \leq \xi \leq u} S(\xi) \leq x, \inf_{t \leq \xi \leq u} S(\xi) \geq -x \right) \]

\[ = \mathbb{P} \left( \sup_{v(1-u) \leq \xi \leq v(1-1)} (W(\xi) + cv^{-1}(\xi)) \leq x, \inf_{v(1-u) \leq \xi \leq v(1-1)} (W(\xi) + cv^{-1}(\xi)) \geq -x \right) \]

\[ = \int_{-x-c(1-u)}^{x-c(1-u)} p(w; v_{ff}(1-u)) \]

\[ \times \mathbb{P} \left[ \sup_{0 \leq \xi \leq v_{ff}(u-1)} (W(\xi) + \frac{c}{v_{ff}}\xi) \leq a_1(w), \inf_{0 \leq \xi \leq v_{ff}(u-1)} (W(\xi) + \frac{c}{v_{ff}}\xi) \geq a_2(w) \right] d\nu \]

\[ = \int_{-x-c(1-u)}^{x-c(1-u)} p(w; v_{ff}(1-u)) \left[ 1 - P_2\left( a_1(w), a_2(w), -\frac{c}{v_{ff}}, v_{ff}(u-1) \right) \right] d\nu \]

The proof is thus complete.

To prove Proposition 11, we resort to formulae for boundary crossing probability conditional on the terminal value.

**Lemma B18** (One-sided Conditional BCP). Let \( W \) be an one-dimensional standard Wiener process. Suppose \( a \) and \( b \) are constant scalars, where \( a > 0 \). Then

\[ \mathbb{P} \{ W(u) \geq a + bu \text{ for some } u \leq t \mid W(t) = x \} = \begin{cases} \exp \left( -\frac{2a(bt+a-x)}{t} \right), & \text{if } x \leq a + bt \\ 1, & \text{if } x > a + bt \end{cases} \quad \text{(B33)} \]

**Proof.** The proof can be found on p. 375, Siegmund (1986).

**Lemma B19** (Two-sided Conditional BCP). Let \( W \) be an one-dimensional standard Wiener process. Consider two linear boundaries \( a_1 + bt \) and \( a_2 + bt \), where \( a_1 > 0 > a_2 \). Let \( P_{2}^{(x)}(a_1, a_2, b, t) \) denote the probability that \( W \) crosses either of the boundaries before or at \( t \) conditional on \( W(t) = x \), that is,

\[ P_{2}^{(x)}(a_1, a_2, b, t) \equiv \mathbb{P} \{ T^* \leq t \mid W(t) = x \}, \text{ where } T^* \equiv \inf u : W(u) = a_1 + bu \wedge \inf u : W(u) = a_2 + bu \]

Then for \( a_2 + bt \leq x \leq a_1 + bt \),

\[ P_{2}^{(x)}(a_1, a_2, b, t) \]

\[ = \sum_{j=1}^{\infty} \exp \left( \frac{2}{t} (j(a_1 - a_2) + a_2)(x - bt - j(a_1 - a_2) - a_2) \right) - \sum_{j=1}^{\infty} \exp \left( \frac{2}{t} j(a_1 - a_2)(x - bt - j(a_1 - a_2)) \right) \]

\[ + \sum_{j=1}^{\infty} \exp \left( \frac{2}{t} (j(a_1 - a_2) - a_1)(-x + bt - j(a_1 - a_2) + a_1) \right) - \sum_{j=1}^{\infty} \exp \left( \frac{2}{t} j(a_1 - a_2)(-x + bt - j(a_1 - a_2)) \right) \quad \text{(B34)} \]

**Proof.** The proof can be found in Section 3 of Hall (1997).
Proof of Proposition 11. In this proof, we let $P^{(w)}$ denote the probability measure under which the Wiener process $W$ starts from $w$. The corresponding expectation is denoted by $E^{(w)}$. When $w = 0$, we simply write $P$ and $E$ as usual.

(a) For the conventional test, we have

$$P(S(l) > x) = P(\frac{Z(l)}{\sqrt{v_{ff}(1-l)}} > \frac{x - c(1-\lambda)}{\sqrt{v_{ff}(1-l)}})$$

$$= 1 - \Phi \left( \frac{x - c(1-\lambda)}{\sqrt{v_{ff}(1-l)}} \right)$$

For the sup-type test, by Proposition 9 and the Markov property,

$$P \left( \sup_{l \leq \xi \leq u} S(\xi) \leq x \right)$$

$$= P \left( \sup_{v_{ff}(1-l) \leq \xi \leq v_{ff}(1-l)} \left( W(\xi) + c(1-\lambda) \right) \leq x \right)$$

$$= P \left( \sup_{v_{ff}(1-l) \leq \xi \leq v_{ff}(1-l)} W(\xi) + c \frac{\xi}{v_{ff}} \leq x \right) P \left( \sup_{v_{ff}(1-l) \leq \xi \leq v_{ff}(1-l)} W(\xi) \leq x - c(1-\lambda) \right)$$

$$= E \left[ \mathbf{1} \left( \sup_{v_{ff}(1-l) \leq \xi \leq v_{ff}(1-l)} W(\xi) + \frac{c}{v_{ff}} \xi \leq x \right) P \left( \sup_{v_{ff}(1-l) \leq \xi \leq v_{ff}(1-l)} W(\xi) \leq x - c(1-\lambda) \right) \right]$$

$$= E \left[ \mathbf{1} \left( \sup_{v_{ff}(1-l) \leq \xi \leq v_{ff}(1-l)} W(\xi) + \frac{c}{v_{ff}} \xi \leq x \right) \right] \Phi \left( W(v_{ff}(1-\lambda)) \right)$$

$$= E \left[ \mathbf{1} \left( \sup_{v_{ff}(1-l) \leq \xi \leq v_{ff}(1-l)} W(\xi) + \frac{c}{v_{ff}} \xi \leq x \right) \right] \Phi \left( W(v_{ff}(1-\lambda)) \right)$$
where

\[ \varphi(w) \equiv \mathbb{P}\left( \sup_{v_{ff}(1 - \lambda) \leq \xi \leq v_{ff}(1 - \lambda)} W(\xi) \leq x - c(1 - \lambda) \mid W(v_{ff}(1 - \lambda)) = w \right) \]

\[ = \mathbb{P}(w) \left( \sup_{0 \leq \xi \leq v_{ff}(1 - \lambda)} W(\xi) \leq x - c(1 - \lambda) \right) \]

\[ = \mathbb{P}\left( \sup_{0 \leq \xi \leq v_{ff}(1 - \lambda)} W(\xi) \leq x - c(1 - \lambda) - w \right) \]

\[ = \mathbb{P}\left( \left| W(v_{ff}(1 - \lambda)) \right| \leq x - c(1 - \lambda) - w \right) \]

\[ = 2\Phi\left( \frac{x - c(1 - \lambda) - w}{\sqrt{v_{ff}(1 - \lambda)}} \right) - 1 \]

Still by Markov property and Lemma B18,

\[ \mathbb{P}\left( \sup_{t \leq \xi \leq u} S(\xi) \leq x \right) \]

\[ = \mathbb{E}\left[ \mathbb{E}\left[ 1 \left( \sup_{v_{ff}(1 - u) \leq \xi \leq v_{ff}(1 - l)} \left( W(\xi) + \frac{c}{v_{ff}} \xi \right) \leq x \right) \varphi\left( W(v_{ff}(1 - \lambda)) \right) \mid \mathcal{F}_{v_{ff}(1 - u)} \right] \right] \]

\[ = \mathbb{E}\left[ \mathbb{E}\left[ 1 \left( \sup_{v_{ff}(1 - u) \leq \xi \leq v_{ff}(1 - l)} \left( W(\xi) + \frac{c}{v_{ff}} \xi \right) \leq x \right) \varphi\left( W(v_{ff}(1 - \lambda)) \right) \mid W_{v_{ff}(1 - u)} \right] \right] \]

\[ = \int_{-\infty}^{x-c(1-u)} p(\tilde{w} \mid v_{ff}(1 - u)) \]

\[ \times \mathbb{E}\left[ \mathbb{E}\left[ 1 \left( \sup_{v_{ff}(1 - u) \leq \xi \leq v_{ff}(1 - l)} \left( W(\xi) + \frac{c}{v_{ff}} \xi \right) \leq x \right) \varphi\left( W(v_{ff}(1 - \lambda)) \right) \mid W_{v_{ff}(1 - u)} = \tilde{w} \right] \right] d\tilde{w} \]

\[ = \int_{-\infty}^{x-c(1-u)} p(\tilde{w} \mid v_{ff}(1 - u)) \mathbb{E}^{(\tilde{w})}\left[ 1 \left( \sup_{v_{ff}(1 - u) \leq \xi \leq v_{ff}(1 - l)} \left( W(\xi) + \frac{c}{v_{ff}} \xi \right) \leq x \right) \varphi\left( W(v_{ff}(1 - \lambda)) \right) \right] d\tilde{w} \]

\[ = \int_{-\infty}^{x-c(1-u)} p(\tilde{w} \mid v_{ff}(1 - u)) \mathbb{E}\left[ \varphi\left( W(v_{ff}(1 - \lambda)) \right) \right] d\tilde{w} \]

\[ = \int_{-\infty}^{x-c(1-u)} \int_{-\infty}^{x-c(1-\lambda)-\tilde{w}} p(\tilde{w} \mid v_{ff}(1 - u)) p(w \mid v_{ff}(u - \lambda)) \]

\[ \times \mathbb{P}\left( \sup_{0 \leq \xi \leq v_{ff}(u - \lambda)} \left( W(\xi) + \frac{c}{v_{ff}} \xi \right) \leq x - c(1 - u) - \tilde{w} \mid W(v_{ff}(u - \lambda)) = w \right) \varphi(w + \tilde{w}) dwd\tilde{w} \]

\[ = \int_{-\infty}^{x-c(1-u)} \int_{-\infty}^{x-c(1-\lambda)-\tilde{w}} p(\tilde{w} \mid v_{ff}(1 - u)) p(w \mid v_{ff}(u - \lambda)) \]

\[ \times \left[ 1 - \exp\left( -\frac{2(x - c(1-u) - \tilde{w})(x - c(1-\lambda) - \tilde{w} - w)}{v_{ff}(u - \lambda)} \right) \right] \left[ 2\Phi\left( \frac{x - c(1-\lambda) - \tilde{w} - w}{\sqrt{v_{ff}(1 - \lambda)}} \right) - 1 \right] dwd\tilde{w} \]

where the rightmost term is the claimed formula.

\( (b) \) Proof for the conventional test is similar to the one-sided case. To establish the for-
mula for the two-sided sup-type test, we define the sets

\[
A_1 \equiv \left( \sup_{\nu_f(1-u) \leq \xi \leq \nu_f(1-\lambda)} \left[ W(\xi) + \frac{c}{v_f} \xi \right] \leq x \right)
\]

\[
A_2 \equiv \left( \sup_{\nu_f(1-\lambda) \leq \xi \leq \nu_f(1-\lambda)} \left[ W(\xi) + c(1-\lambda) \right] \leq x \right)
\]

\[
B_1 \equiv \left( \sup_{\nu_f(1-u) \leq \xi \leq \nu_f(1-\lambda)} \left[ W(\xi) + \frac{c}{v_f} \xi \right] \geq -x \right)
\]

\[
B_2 \equiv \left( \sup_{\nu_f(1-u) \leq \xi \leq \nu_f(1-\lambda)} \left[ W(\xi) + c(1-\lambda) \xi \right] \geq -x \right)
\]

With the sets thus defined,

\[
\mathbb{P} \left( \sup_{t \leq \xi \leq u} |S(\xi)| \leq x \right)
\]

\[
= \mathbb{P} \left( \sup_{\nu_f(1-u) \leq \xi \leq \nu_f(1-\lambda)} \left| W(\xi) + c(1-\lambda) \lor v_f^{-1} \xi \right| \leq x \right)
\]

\[
= \mathbb{P} (A_1 \cap A_2 \cap B_1 \cap B_2)
\]

\[
= \mathbb{E} \left[ \mathbb{I}(A_1 \cap B_1) \mathbb{P}(A_2 \cap B_2 | \mathcal{F}(\nu_f(1-\lambda))) \right]
\]

\[
= \mathbb{E} \left[ \mathbb{I}(A_1 \cap B_1) \mathbb{P}(A_2 \cap B_2 | W(\nu_f(1-\lambda))) \right]
\]

\[
= \mathbb{E} \left[ \mathbb{I}(A_1 \cap B_1) \varphi(W(\nu_f(1-\lambda))) \right]
\]

where

\[
\varphi(w) \equiv \mathbb{P} \left( A_2 \cap B_2 | W(\nu_f(1-\lambda)) = w \right)
\]

\[
= \mathbb{P}(w) \left( \sup_{0 \leq \xi \leq \nu_f(\lambda-\lambda)} W(\xi) \leq x - c(1-\lambda), \inf_{0 \leq \xi \leq \nu_f(\lambda-\lambda)} W(\xi) \geq -x - c(1-\lambda) \right)
\]

\[
= \mathbb{P} \left( \sup_{0 \leq \xi \leq \nu_f(\lambda-\lambda)} W(\xi) \leq x - c(1-\lambda) - w, \inf_{0 \leq \xi \leq \nu_f(\lambda-\lambda)} W(\xi) \geq -x - c(1-\lambda) - w \right)
\]

\[
= 1 - P_2(a_1(w), a_2(w), 0, \nu_f(\lambda-\lambda))
\]

with \(a_1(w)\) and \(a_2(w)\) defined in the proposition.

Next, we further define two sets

\[
A'_1(\tilde{w}) \equiv \left( \sup_{0 \leq \xi \leq \nu_f(u-\lambda)} \left[ W(\xi) + \frac{c}{v_f} \xi \right] \leq x - c(1-u) - \tilde{w} \right)
\]

\[
B'_1(\tilde{w}) \equiv \left( \sup_{0 \leq \xi \leq \nu_f(u-\lambda)} \left[ W(\xi) + \frac{c}{v_f} \xi \right] \geq -x - c(1-u) - \tilde{w} \right)
\]
Continuing from (B35), we have

\[
P \left( \sup_{t \leq \xi \leq u} |S(\xi)| \leq x \right)
= E \left[ 1(A_1 \cap B_1) \varphi(W(v_{ff}(1 - \lambda))) \right]
= E \left[ E \left[ 1(A_1 \cap B_1) \varphi(W(v_{ff}(1 - \lambda))) \big| F(v_{ff}(1 - u)) \right] \right]
= E \left[ E \left[ 1(A_1 \cap B_1) \varphi(W(v_{ff}(1 - \lambda))) \big| W(v_{ff}(1 - u)) \right] \right]
= \int_{x-c(1-u)}^{x-c(1-u)} p(\bar{w}; v_{ff}(1-u)) E \left[ 1(A_1 \cap B_1) \varphi(W(v_{ff}(1 - \lambda))) \big| W(v_{ff}(1 - u)) = \bar{w} \right] d\bar{w}
= \int_{x-c(1-u)}^{x-c(1-u)} p(\bar{w}; v_{ff}(1-u)) E \left[ 1(A_1' \cap B_1') \varphi(W(v_{ff}(u - \lambda) + \bar{w})) \big| W(v_{ff}(u - \lambda)) = w \right] \varphi(w + \bar{w}) dwd\bar{w}
= \int_{-\infty}^{x-c(1-u)} \int_{-\infty}^{x-c(1-u)} p(\bar{w}; v_{ff}(1-u)) p(w; v_{ff}(u - \lambda))
\times \left[ 1 - \exp \left( - \frac{2(x - c(1-u) - \bar{w})(x - c(1-\lambda) - \bar{w} - w)}{v_{ff}(u - \lambda)} \right) \right]
\times \left[ 2 \varphi \left( \frac{x - c(1-\lambda) - \bar{w} - w}{\sqrt{v_{ff}(\lambda - l)}} \right) \right] dwd\bar{w}
\]

which is as claimed. \( \square \)
References


Gelman, A. and Loken, E. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no fishing expedition or p-hacking and the research hypothesis was posited ahead of time. *Working Paper*, 2013.


Forecasting Daytime US Market Volatility with Overnight Information

Giorgio Mirone
Aarhus University and CREATEs
Ye Zeng
Aarhus University and CREATEs

Abstract

This paper highlights the sizeable benefit derived from inclusion of information on the overnight US futures market in forecasting volatility of the daytime US stock market. We find that augmenting the HAR, HAR-CJ, SHAR and HARQ models with volatility of the overnight US futures market leads to considerably more accurate forecast. In out-of-sample analysis, we confirm superior predictive ability of the models featuring overnight information by White's reality check. Finally, we dissect the forecasts by relating them to the interaction between the overnight US market and the daytime European market and to the FOMC meeting.

Keywords: High Frequency Data; Overnight Information; Volatility Forecasting; HAR; Futures Market.

JEL classification: C53, C58.
2.1 Introduction

To prosper in the financial market, one must keep an eye on the unceasing information flow to stay abreast of the prevailing state of affairs. Being informed is a prerequisite for taking judicious action. This is particularly true for volatility forecasting, a crucial task on which hinge the profit-oriented activities such as asset pricing and portfolio management. If the model overlooks a market-quivering event occurring at nightfall, it is liable to underforecast volatility on the coming day and miss prompt adjustment of holdings. The implication is that overnight information must be built explicitly into volatility forecasting model, yet few of the extant literature are attentive to such a crucial matter.

The present paper seeks to enrich the literature with an analysis of the effect on forecasting daytime volatility of the US stock market exerted by inclusion of the information on the overnight \footnote{In this paper, the term “overnight” means the breaktime between two adjacent opening periods of the US stock market. It should be underlined that “overnight” refers not only to the period from 16:00 to 09:30 (New York time) on the next day during the week, but also to 16:00 Friday to 09:30 the next Monday. It might also refer to a recess covering holidays, if any. All overnight periods are deemed homogeneous, in the sense that we do not distinguish between overnight periods covering and not covering weekend or holiday.} Emini S&P 500 futures market. The US stock market closes from 16:00 to 09:30 (New York time) on the next business day, during which investors around the world do not cease acting and the media do not stop disseminating information. It poses the problem of quantifying overnight volatility in the absence of transaction data, a component forecasters would like to integrate into their model. In this paper, we propose – not as a perfect solution, but, frankly speaking, as an expedient – to exploit the overnight volatility of the Emini S&P 500 futures (ticker: ES), which trades almost around the clock (see data description below), and show its expediency of forecasting daytime volatility of some US stocks.

We divide our empirical analysis into several parts, each assuming a specific task. We first consider forecasting daytime volatility of the US stock market as a whole, which can be gauged by volatility of the ETF SPDR (also known as SPY) that tracks closely the S&P 500 index. We take the state-of-the-art HAR model \citep{Corsi2009} as benchmark, which features no overnight information, and augment it with overnight volatility of ES. We document remarkable improvement in accuracy of one-day ahead forecast: the augmented model featuring overnight information has an in-sample $R^2$ that is about 15% higher than the benchmark, and its superior predictive ability is confirmed by the reality check \citep{White2000} in out-of-sample analysis. We go on to study if we can still derive benefit from overnight information with longer forecast horizons, say up to one month. We find that the improve-
ment brought by overnight information diminishes rapidly as the forecast horizon extends, though it doesn’t vanish altogether.

Next, in lieu of HAR, we take as benchmark three variants of it – the HAR-CJ model (Andersen et al. (2007)), the SHAR model (Chen and Ghysels (2010); Patton and Sheppard (2015)) and the HARQ model (Bollerslev et al. (2016)) – and pose the question: to maximize the benefit derived from the overnight information, i.e., to make the most accurate forecast with overnight information, should one simply augment the benchmark model with overnight realized volatility, or should one further extend the augmented model in the spirit of the benchmark? Our findings suggest that the answer hinges on the loss function used to evaluate the model: further extension of the augmented model is preferable if symmetric loss function such as MSE is used, but it is not desirable if asymmetric loss function such as QLIKE is adopted for evaluation.

The aforementioned findings apply chiefly to SPY, whose volatility is intimately related to the volatility of ES. We could expect to derive equally sizeable benefits in forecasting daytime volatility of individual stocks if the single stock futures (SSF) were traded overnight. Yet unlike the Emini futures, SSFs do not trade around the clock; it is not even actively traded during the limited opening hours of the exchange. We do not tackle the problem of measuring volatility of individual stock in the absence of overnight transaction. Instead, we augment the HAR model with overnight volatility of ES, which arguably gauges systematic risk as opposed to idiosyncratic risk, and make comparison of the forecasts produced with and without inclusion of overnight information. Out of about 50% of the stocks that we consider, inclusion of overnight volatility of ES improves forecast accuracy, sometimes considerably. Interestingly, we find that the higher the correlation between volatility of the stock in concern and volatility of ES, the larger the cut in MSE attained by inclusion of overnight volatility of ES, but at the same time the lower the reduction in QLIKE loss.

In the last part of our analysis we dissect the volatility forecasts by relating them to the interaction between the overnight US market and the daytime European market and to the FOMC meeting. The close interaction between the overnight US market and the daytime European market suggests that inclusion of overnight information enables the model to catch up swiftly with escalated risk engendered by impactful events that originate from outside the US. Neglecting the overnight information, by contrast, causes the model to understate the intensified market stress. On the other hand, stratifying losses according to whether there is FOMC meeting, we find that inclusion of overnight volatility of ES typically leads to more improvement in forecast accuracy in the absence of FOMC meeting.

Literature that are closely related to the present paper include Tsiakas (2008), Martens (2002), Taylor (2007), and Chen et al. (2012). Devising a stochastic volatility model that is conditioned on lagged overnight information, Tsiakas (2008) highlights the predictive power of information garnered during non-trading hours. Our research differs from his in several aspects. First, as opposed to the open-to-close and close-to-open returns used in his paper,
2.2. Forecasting Daytime Volatility of US Stocks

2.2.1 Data

We make use of the following datasets in our empirical analysis:

D1. High frequency transaction data of 24 highly liquid stocks and the ETF SPDR 500 (also known as SPY);
D2. High frequency transaction data of the Emini S&P 500 index-futures (also known as ES);
D3. High frequency transaction data of the FTSE 100 index-futures.

D1 and D2 span from 29th January 2001 through 31st December 2013, covering 3,226
trading days of the US stock market in total. D3 is shorter: it ranges from 4th January 2010 to 31st December 2013. In this section we capitalize on the information of overnight segment of D2 to forecast volatility measures built on D1; the interaction between the overnight segment of D2 and the daytime segment of D3 is investigated in Section 2.3. Before proceeding it is important to make clear the overlapped and non-overlapped segments of the three markets’ trading hours. Figure 2.1 depicts the trading and non-trading hours of the three markets (all in New York local time). The US stock market opens from 09:30 to 16:00 on business days. The ES contract trades from Sunday to Friday. On each of these days its transaction begins at 18:00 and halts at 16:15 the next day for 15 minutes. After the halt the trades last for 45 minutes, then pause for another 45 minutes due to daily maintenance, and finally start anew after the maintenance. The FTSE 100 futures contract is transacted from 08:00 to 17:30 in London local time, which corresponds to the period 03:00 to 12:30 in New York.

On most days multiple ES contracts are available, each having a peculiar maturity date. Therefore rules must be specified to determine which contract is used on each day to create a consecutive time series. We would like to adopt the most liquid contract, which for most of the time is the front month contract, i.e., the one with the closest maturity. But when the front month contract is about to expire, its transaction diminishes rapidly, as traders switch to the contract in a further-out month. Given these facts, we define a roll-over window as two weeks (10 business days) prior to expiration of the front month contract. Outside the roll-over window the front month contract is used; upon moving into the roll-over window, we start to monitor daily number of transactions of both the front month and the further-out contracts and roll to the latter as soon as it has more transactions.

### 2.2.2 Notations

The volatility measures employed in our empirical analysis are reviewed in Appendix A. Basically, we let $S$ and $F$ denote respectively the stock whose daytime volatility is to be forecast – either SPY or a single stock – and the Emini S&P 500 futures contract whose overnight volatility is to be included in forecasting. $RV(\cdot)$ denotes realized volatility, $TBV(\cdot)$ threshold bipower variation, $JV(\cdot)$ jump variation, $RSV^+(\cdot)$ realized positive semivariance and $RSV^-(\cdot)$ realized negative semivariance, and $TRQ(\cdot)$ the tri-power quarticity. Figure 2.1 depicts the time windows over which the key variables, $RV(S)_{t+1}^D$ and $RV(F)_{t}^N$, are computed respectively.

---

2. We do not have access to data with longer history.
3. We do not take into account the pre-open and after hours.
4. It should be underlined that the overnight volatility measures before opening of the US stock exchange on every Monday is computed with data of ES from Sunday 18:00 to Monday 09:30 (New York local time). From Friday 16:00 to Sunday 18:00 (New York local time) the ES is closed as well.
5. This is slightly different from the conventional roll-over of trading of Emini futures that occurs eight calendar days before the expiration date.
2.2. Forecasting Daytime Volatility of US Stocks

Note: This figure summarizes trading (blue) as well as non-trading hours (black) of the US stocks (D1), the Emini S&P 500 index-futures (D2) and the FTSE 100 index-futures (D3). The blue tick labels are New York time and the red tick labels are local time of the exchange on which the asset is traded. The US stock market opens from 09:30 to 16:00 New York time on business days (excluding pre-open and after-hours trades). The ES contract trades from Sunday to Friday. On each of these days its transaction begins at 18:00 (New York time) and halts at 16:15 (New York time) the next day for 15 minutes. After the halt the trades last for 45 minutes, then pause for another 45 minutes due to daily maintenance, and finally start anew after the maintenance. The FTSE 100 futures contract is transacted from 08:00 to 17:30 London time, which corresponds to 03:00 to 12:30 in New York.


### 2.2.3 Summary statistics of the ES data

**Figure 2.2:** Average number of transactions of Emini S&P 500 futures

Note: This graph presents average number of transactions of the Emini S&P 500 futures in each 15-minute interval of a day. Since the Emini futures contracts are traded on CME, the x-axis is set to Chicago local time, which is an hour earlier than New York local time.

Figure 2.2 presents the average number of transactions of ES in every 15-minute interval. The x-axis is set to Chicago local time since the Emini futures are traded on CME. As mentioned above, transaction of the ES futures is halted from 15:15 to 15:30 and 16:15 to 17:00 (Chicago time) on each trading day, so the number of transactions is zero during these periods. An immediate observation is that transaction of the futures is much more active during the day, especially from 08:30 to 15:00 (Chicago time) when NYSE is open. There are still a number of transactions within the fifteen minutes right after closing of NYSE. The distribution of transactions from 17:00 to 02:00 of the next day (Chicago time) is relatively uniform and stays at a low level of less than 200 transactions. With trading of the FTSE 100 futures beginning around 02:00 (Chicago time), transaction of ES doubles as well, and its distribution before opening of NYSE (i.e., 08:30 in Chicago) is roughly U-shaped.

Table 2.1 summarizes minimum, median, mean and maximum of daytime and overnight volatility measures of ES. Almost all statistics of the overnight measures are smaller than their daytime counterparts. Maximums of the overnight volatility measures are only 50% to 60% of maximums of the corresponding daytime measures. To better appreciate the distributional differences, in Figure 2.3 we provide kernel density plots for daytime and overnight logarithmic transform of the volatility measures, where the densities are underlaid with his-
2.2. Forecasting Daytime Volatility of US Stocks

| Table 2.1: Summary statistics of daytime and overnight volatility measures of ES |
|-------------------------------|-------|-------|-------|-------|-------|
| **Panel A: Daytime and overnight volatility measures** |
| | Min | Median | Mean | Max |
| **Day** | **Night** | **Day** | **Night** | **Day** | **Night** | **Day** | **Night** |
| RV | 0.0277 | 0.0122<sup>p</sup> | 0.5200 | 0.2226 | 1.0922 | 0.5106 | 62.4236 | 34.5409 |
| TBV | 0.0033<sup>p</sup> | 0.0069<sup>p</sup> | 0.4023 | 0.1448 | 0.9076 | 0.3518 | 38.1025 | 22.3949 |
| JV<sup>+</sup> | <0.0001<sup>p</sup> | 0.0002<sup>p</sup> | 0.0759 | 0.0609 | 0.1978 | 0.1592 | 24.3211 | 12.1460 |
| RSV<sup>+</sup> | 0.0179 | 0.0059<sup>p</sup> | 0.2497 | 0.1086 | 0.5518 | 0.2552 | 42.1047 | 20.5266 |
| RSV<sup>−</sup> | 0.0056 | 0.0005<sup>p</sup> | 0.2526 | 0.1085 | 0.5404 | 0.2554 | 20.3188 | 14.0143 |
| **Panel B: Summary statistics of RV<sup>F</sup><sub>Y</sub>/RV<sup>F</sup><sub>D</sub>** |
| | Min | Median | Mean | Max |
| | 0.0317 | 0.4510 | 0.5735 | 12.4836 |

Note: When marked with *p*, the statistic is computed on a subset of the corresponding measures which are strictly positive.

...tograms. Densities of overnight volatility measures are seen to be positioned to the left and have longer left tails, confirming that the overnight ES market is less volatile than daytime ES market in general.

2.2.4 Forecasting daytime volatility of SPY

**Benchmark: HAR (one-day ahead)**

We begin with forecast of daytime volatility of the ETF SPY; results of individual stocks are presented in the next subsection. The reason for separating the analyses of SPY and individual stocks is two-fold: first, volatility of SPY reflects market-wide risk as opposed to idiosyncratic risk; second, there is an intimate relation between volatilities of SPY and ES, whereupon are the considerable gains we present below predicated. To keep the paper concise we relegate discussion on such relation to Appendix C.

We adopt first the HAR model introduced by Corsi (2009) as the benchmark and forecast one-day ahead volatility; longer horizons and other variants of HAR will be discussed later. The HAR model is specified as

\[
RV(S)_{t+1}^{D} = \alpha_0 + \alpha_D RV(S)^D_t + \alpha_W RV(S)^W_t + \alpha_M RV(S)^M_t + \varepsilon_{t+1}
\]

where \(RV(S)^W_t = \frac{1}{5} \sum_{i=0}^{4} RV(S)^D_{t-i}\) and \(RV(S)^M_t = \frac{1}{22} \sum_{i=0}^{21} RV(S)^D_{t-i}\) are weekly and monthly averages of daytime realized volatilities respectively. Model (2.1) does not feature information during the overnight period 16:00 to 09:30 (New York time), and one simple extension is to augment it with the latest overnight volatility measure, which we denote by \(V_N^t\). The
What we seek to know is the effect of putting $V_t^N$ as $RV(F)_t^N$, i.e., the overnight realized volatility of ES. The corresponding extension is termed HAR-F (F for futures). In addition we consider the squared overnight return of the SPY ETF itself as an alternative overnight volatility measure. Defining overnight return of SPY is a subtle problem. SPY trades roughly at one tenth of the S&P 500 index. As pointed out by Ahoniemi and Lanne (2013), the S&P 500 index values are computed from last trades of all constituents, making the first published index value of the day dominated by numerous prices of the previous trading day. The calculation method of S&P 500 index hence tends to engender stale opening value, i.e., value equal to closing value of the previous trading day. Ahoniemi and Lanne (2013) contend that use of stale opening value leads to misinterpretation of the overnight return. They propose to calculate overnight return of the S&P 500 index as return from closing of the previous day up until 09:35 (New York time), namely five minutes past NYSE’s opening. We follow their suggestion in our empirical analysis.$^6$ The overnight return of SPY from day-$(t-1)$ to day-$t$ will thus be denoted by $r(S)_t^{N,09:35}$. When $r(S)_t^{N,09:35}$ is plugged into (2.2), the augmented model is

$$RV(S)_{t+1}^D = \alpha_0 + \alpha_D RV(S)_t^D + \alpha_W RV(S)_t^W + \alpha_M RV(S)_t^M + \theta N V_t^N + \epsilon_{t+1}$$

$^6$We have considered the usual close-to-open overnight return of SPY as well. The results are indeed almost the same as those obtained using close-to-09:35 returns, with the latter yielding slightly better, if not the same, forecasts. This is expected, since the suggestion made by Ahoniemi and Lanne (2013) is primarily for volatility estimation rather than volatility forecasting.
resulting model is termed HAR-$r(S)_N^{09:35}$.

The HAR-F model simply augments HAR with the latest overnight volatility of ES. Naturally, we wonder the effect of further lagged overnight volatilities, so we also consider a “complete” extension of HAR, named HAR-F-C (C for complete), which further includes weekly and monthly averages of overnight realized volatilities:

\[
RV(S)_{t+1}^D = \alpha_0 + \alpha_D RV(S)_t^D + \alpha_W RV(S)_t^W + \alpha_M RV(S)_t^M \\
+ \theta_N RV(F)_t^N + \theta_W RV(F)_t^{N,W} + \theta_M RV(F)_t^{N,M} + \varepsilon_{t+1}
\]

(2.3)

where $RV(F)_t^{N,W} \equiv \frac{1}{5} \sum_{i=0}^{4} RV(F)_{t-i}^N$ and $RV(F)_t^{N,M} \equiv \frac{1}{22} \sum_{i=0}^{21} RV(F)_{t-i}^N$.

To appraise the models we consider two loss functions: the squared error (SE) and the Quasi-Likelihood (QLIKE), the first being symmetric and the second asymmetric. The loss functions are given by

\[
SE(RV_t, \hat{RV}_t) \equiv (RV_t - \hat{RV}_t)^2 \\
QLIKE(RV_t, \hat{RV}_t) \equiv \frac{RV_t}{\hat{RV}_t} - \log \left( \frac{RV_t}{\hat{RV}_t} \right) - 1
\]

where $RV_t$ refers to (the object) realized volatility and $\hat{RV}_t$ the fit or forecast from any model. According to Patton (2011), both loss functions are homogeneous and robust. They are homogeneous in the sense that ranking of forecast is invariant to re-scaling of data. Hence, unit of price or return does not affect ranking of forecast. In particular, it suggests that the choice between percentage and decimal returns has no bearing on the ranking. On the other hand, the loss functions are robust in the sense that substituting imperfect measure for the latent volatility in the loss function, i.e., substituting realized volatility for quadratic variation (see Appendix A), does not affect ranking of the models. For detailed explanation we refer to Patton (2011) and references therein.

Panel A of Table 2.2 summarizes in-sample results. Inclusion of the squared close-to-09:35 returns only leads to marginal improvement in terms of (adjusted) $R^2$. By contrast, HAR-F has better in-sample fitting, with (adjusted) $R^2$ being almost 15% higher. Another notable fact is that when $RV(F)_t^N$ is included, weights are “shifted” from past daily, weekly and monthly daytime volatilities of SPY to the latest overnight volatility of ES. In the HAR model, $\alpha_D$, $\alpha_W$ and $\alpha_M$ are about 0.26, 0.47 and 0.18 respectively, while in the HAR-F the estimates become 0.12, 0.23 and 0.12, which are 30% to 50% smaller.

The coefficient estimate of $RV(F)_t^N$ is significant and slightly above 1. The sharp drop in coefficients of the daytime volatility components of SPY and the much larger coefficient of the overnight volatility of ES motivate us to consider a distributed lag model which uses only $RV(F)_t^N$ to forecast daytime volatility of SPY. The row labeled “DL-F” in panel A of Table 2.2 shows in-sample results of the distributed lag model. The adjusted $R^2$ is 0.62, which is larger than that of HAR. Nonetheless, past daytime volatilities of SPY still help to improve forecast
Table 2.2: In-sample and out-of-sample results of the HAR group (one-day ahead forecasting).

**Panel A: In-sample**

<table>
<thead>
<tr>
<th>Model</th>
<th>$\alpha_0$</th>
<th>$\alpha_D$</th>
<th>$\alpha_W$</th>
<th>$\alpha_M$</th>
<th>$\theta_N$</th>
<th>$\theta_W$</th>
<th>$\theta_M$</th>
<th>Adj. $R^2$</th>
<th>MSE</th>
<th>QLIKE</th>
</tr>
</thead>
<tbody>
<tr>
<td>HAR</td>
<td>0.0901</td>
<td>0.2637</td>
<td>0.4679</td>
<td>0.1826</td>
<td></td>
<td></td>
<td></td>
<td>0.5685</td>
<td>2.4317</td>
<td>0.1387</td>
</tr>
<tr>
<td></td>
<td>(0.0563)</td>
<td>(0.1367)</td>
<td>(0.1743)</td>
<td>(0.1024)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HAR-$r(S)^{N,09:35}$</td>
<td>0.0648</td>
<td>0.1941</td>
<td>0.4249</td>
<td>0.1927</td>
<td>0.2674</td>
<td></td>
<td></td>
<td>0.6047</td>
<td>2.2268</td>
<td>0.1363</td>
</tr>
<tr>
<td></td>
<td>(0.0582)</td>
<td>(0.1532)</td>
<td>(0.1409)</td>
<td>(0.0955)</td>
<td>(0.1079)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HAR-F</td>
<td>0.0542</td>
<td>0.1174</td>
<td>0.2314</td>
<td>0.1228</td>
<td>1.0034</td>
<td></td>
<td></td>
<td>0.6967</td>
<td>1.7087</td>
<td>0.1345</td>
</tr>
<tr>
<td></td>
<td>(0.0470)</td>
<td>(0.1068)</td>
<td>(0.1102)</td>
<td>(0.0883)</td>
<td>(0.2147)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HAR-F-C</td>
<td>0.0484</td>
<td>0.1016</td>
<td>0.1169</td>
<td>0.5428</td>
<td>1.0241</td>
<td>0.3117</td>
<td>-0.9339</td>
<td>0.7032</td>
<td>1.6711</td>
<td>0.1328</td>
</tr>
<tr>
<td></td>
<td>(0.0449)</td>
<td>(0.1123)</td>
<td>(0.1685)</td>
<td>(0.1990)</td>
<td>(0.2419)</td>
<td>(0.6053)</td>
<td>(0.5066)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DL-F</td>
<td>0.2903</td>
<td></td>
<td></td>
<td></td>
<td>1.5381</td>
<td></td>
<td></td>
<td>0.6259</td>
<td>2.1096</td>
<td>0.2441</td>
</tr>
<tr>
<td></td>
<td>(0.0891)</td>
<td></td>
<td></td>
<td></td>
<td>(0.1909)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Panel B: Out-of-sample**

<table>
<thead>
<tr>
<th>Model</th>
<th>Rolling $R^2$</th>
<th>Rolling MSE</th>
<th>Rolling QLIKE</th>
<th>Increasing $R^2$</th>
<th>Increasing MSE</th>
<th>Increasing QLIKE</th>
</tr>
</thead>
<tbody>
<tr>
<td>HAR</td>
<td>0.5032</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.5186</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>HAR-$r(S)^{N,09:35}$</td>
<td>0.5161</td>
<td>0.9303</td>
<td>1.0222</td>
<td>0.5398</td>
<td>0.9403</td>
<td>1.0136</td>
</tr>
<tr>
<td>HAR-F</td>
<td>0.6487</td>
<td>0.6469*†</td>
<td>0.8414*†</td>
<td>0.6640</td>
<td>0.7001</td>
<td>0.9867</td>
</tr>
<tr>
<td>HAR-F-C</td>
<td>0.6184</td>
<td>0.7023</td>
<td>0.9011</td>
<td>0.6278</td>
<td>0.7736</td>
<td>0.9234</td>
</tr>
<tr>
<td>DL-F</td>
<td>0.6077</td>
<td>0.7026</td>
<td>1.2711</td>
<td>0.5817</td>
<td>0.8612</td>
<td>1.7332</td>
</tr>
</tbody>
</table>

Note: This table lists in-sample and out-of-sample results of the HAR model and its extensions. The dependent variable is the daytime (realized) volatility of SPY and the forecast is made one day ahead. The benchmark HAR model is specified in (2.1). The HAR-$r(S)^{N,09:35}$ is HAR augmented with squared close-to-09:35 return of SPY itself. The HAR-F model is HAR augmented with overnight volatility of ES. The HAR-F-C model is specified in (2.3). DL-F is the distributed lag model below:

$RV(S)_{t+1}^{D} = \alpha_0 + \theta_N RV(F)_{t}^{N} + \varepsilon_{t+1}$

Panel A: The regressions are based on percentage returns. In parentheses are Newey-West estimates of standard errors computed with Bartlett kernel and bandwidth 5. $\theta_N$ is always the coefficient of overnight volatility measures (squared close-to-09:35 return of SPY itself or volatility of ES).

Panel B: In the rolling scheme, the window size is fixed at 800. In the increasing scheme, the estimation window starts with the first 800 observations and keeps expanding. The entries are average ratios of the losses of the models to the corresponding ones of the benchmark HAR model. Entries marked with a symbol means that null hypothesis of the (modified) reality check (see (2.4) and (2.5)) is rejected at the significance level 5%. *: test is conducted on the full sample; †: test is conducted after removing top 1% least accurate forecasts produced by each model.
accuracy if in concern is the QLIKE loss, which is seriously enlarged in the distributed lag model.

We also perform pseudo out-of-sample one-day ahead forecast with both rolling and increasing schemes. In the rolling scheme, the window size is fixed at 800, that is, on each day estimation of coefficients are done with data of the most recent 800 days. In the increasing scheme, the estimation window starts with the first 800 observations and keeps expanding as the day on which forecast is made moves on. Panel B of Table 2.2 summarizes the Mincer-Zarnowitz $R^2$ and average ratios of loss functions of the models to the corresponding ones of the benchmark HAR model.

For the rolling scheme, inclusion of close-to-09:35 returns of SPY itself merely brings about 1% increment of the Mincer-Zarnowitz $R^2$. The models featuring overnight volatility of ES produce much more accurate forecasts. In the HAR-F model, for example, not only does the Mincer-Zarnowitz $R^2$ increase about 15% (relative to the benchmark HAR), but all losses are cut down remarkably. Interestingly, the simple distributed lag model does have larger $R^2$ and smaller squared loss when compared with HAR, though it shows lower forecast accuracy when QLIKE loss is used for evaluation.

Results of the increasing scheme is somewhat mixed. For HAR-$r(S)^N,09:35$ in the increasing scheme, the results are roughly the same as those reported in the rolling scheme. Forecast of the simple distributed lag model has smaller MSE, but, as in the rolling scheme, is much worse than those of the benchmark HAR model in terms of QLIKE loss. As for the two models exploiting overnight information, both of them have reduced MSE as well as QLIKE loss. Neither of them seems uniformly better than the other: HAR-F has lower MSE while HAR-F-C becomes preferable when QLIKE loss is in concern.

To formally infer whether including overnight volatility of ES leads to significantly improved forecast accuracy, we apply a modified version of the Reality Check proposed by White (2000). The same modification is made by Bollerslev et al. (2016) to test whether their HARQ model outperforms other variants of HAR. To illustrate, let $L$ denote the loss function, $RV_t$ the realized volatility and $RV_{t}^{(j)}$, $j = 0, 1, \ldots, k$ forecasts of $RV_t$ produced by model $j$, where model 0 plays the special role of baseline and the other models are competitors. Hypotheses of the (modified) reality check are

$$H_0 : \min_{j=1, \ldots, k} \mathbb{E}[L(RV_t, RV_{t}^{(j)}) - L(RV_t, RV_{t}^{(0)})] \leq 0 \quad (2.4)$$

and

$$H_1 : \min_{j=1, \ldots, k} \mathbb{E}[L(RV_t, RV_{t}^{(j)}) - L(RV_t, RV_{t}^{(0)})] > 0 \quad (2.5)$$

The null hypothesis means that at least one of the competing models has expected loss no greater than the baseline model. Rejection of the null hypothesis thus evidences superior predictive ability of the baseline model. In our analysis HAR-F is taken as the baseline. Fol-
lowing White (2000), we implement reality check with the stationary bootstrap (Politis and Romano (1994)) which accommodates weak dependence of time series data. The average block length is set to 12 and 999 bootstrap samples are drawn for each test. Some untabulated results show that our conclusions are robust to choice of the block length and the number of bootstrap samples. For each loss function we consider reality check on two samples. The first one is the full sample, i.e., the sample of all forecasts. The second sample is obtained after removing the top 1% least accurate forecasts of each model\(^7\).

Results of reality check are listed in panel B of Table 2.2. For the rolling scheme, test on the full-sample is in favor of superior predictive ability of HAR-F, whichever loss function we consider. Removing the least accurate forecasts does not affect the conclusion. It is noteworthy that HAR-F-C, the model further including past weekly and monthly averages of overnight volatility, is among the competitors, but the null hypothesis is still rejected. In other words, HAR-F shows superior predictive ability over HAR-F-C, though the latter features complete past information\(^8\). For the increasing scheme, the reality check does not suggest superior predictive ability of HAR-F: This is expected when QLIKE is taken as yardstick, as the competitor HAR-F-C has substantially lower QLIKE loss. On the other hand, with the squared loss and the full sample of forecasts, \(p\)-value of the test statistic actually falls randomly around 5%\(^9\). Nonetheless, we draw the conservative conclusion that HAR-F does not exhibit superior predictive ability if increasing scheme is adopted.

**Benchmark: HAR (longer horizons)**

We naturally wonder if the latest overnight information of the futures market still provides incremental information as the forecast horizon extends. Thus we turn to longer forecast horizons, say up to one-month (22 days) ahead. Let \(h\) denote the number of days ahead. We would like to forecast the average daytime volatility of SPY on the upcoming \(h\) trading days\(^10\), i.e., \(\bar{RV}(S)_{t+1}^{D} = \frac{1}{h} \sum_{j=1}^{h} RV(S)_{t+j}^{D}\). The benchmark model and the corresponding extension is obtained by substituting \(\frac{1}{h} \sum_{j=1}^{h} RV(S)_{t+j}^{D}\) for \(RV(S)_{t+1}^{D}\) in (2.1) and (2.2) respectively. To be specific, the benchmark HAR model is

\[
\frac{1}{h} \sum_{j=1}^{h} RV(S)_{t+j}^{D} = \alpha_0 + \alpha_D RV(S)^D_t + \alpha_W RV(S)^W_t + \alpha_M RV(S)^M_t + \epsilon_{t+1}
\] (2.6)

---

\(^7\)Removal of these forecasts may create technical problems in application of the reality check, but circumscribed by scope of the present paper, we do not delve into this issue here.

\(^8\)Interestingly, when developing the HARQ model, Bollerslev et al. (2016) also consider the “full” Q-model which includes past weekly and monthly averages of realized quarticity, and they find that the “full” model is inferior to the simpler HARQ model in terms of predictive ability.

\(^9\)The randomness results from approximating the bootstrap test statistic by Monte Carlo. In most of our experimentations we find the test statistic to be between 5% and 6%.

\(^10\)Forecasting average of future volatility is often called *direct* approach in the literature, as opposed to the iterative approach. The direct approach is applied by Bollerslev et al. (2018) as well. Some helpful discussion and comparison of the approaches can be found in Ghysels et al. (2009).
and the corresponding HAR-F model is
\[
\frac{1}{h} \sum_{j=1}^{h} RV(S)_{t+j}^D = \alpha_0 + \alpha_D RV(S)_t^D + \alpha_W RV(S)_t^W + \alpha_M RV(S)_t^M + \theta_N RV(F)_t^N + \varepsilon_{t+1}
\] (2.7)

The other models are defined analogously. We still consider a distributed lag model which uses a single regressor \( RV(F)_t^N \) (together with intercept) to forecast \( \frac{1}{h} \sum_{j=1}^{h} RV(S)_{t+j}^D \).

Constrained by space we only tabulate results of one-week (5 days) ahead forecast in Table 2.3 and of one-month (22 days) ahead forecast in Table 2.4. On the other hand, Figure 2.4 plots in-sample adjusted \( R^2 \)'s of HAR, HAR-F and the distributed lag model over multiple horizons (the left panel) and coefficient estimates and 95% confidence intervals of the overnight volatility of ES (the right panel). From the in-sample results we observe, not surprisingly, that improvement in forecast accuracy brought by inclusion of \( RV(F)_t^N \) diminishes as the forecast horizon extends. As shown in Figure 2.4, for one-day ahead forecast, adjusted \( R^2 \) of HAR-F is about 15% larger than that of HAR, and even the DL-F model has better in-sample fitting. The difference between \( R^2 \) of HAR-F and \( R^2 \) of HAR keeps shrinking as the forecast horizon extends and drops to about 2% for one-month ahead forecast.

Contrariwise, the difference between \( R^2 \) of HAR and \( R^2 \) of the distributed lag model widens as forecast horizon extends. Another immediate observation from Figure 2.4 is that coefficient of the overnight volatility of ES, though still significant, keeps getting smaller. Indeed, it drops sharply from 1.00 to 0.72 as the forecast horizon extends from one-day to two-day ahead. These observations all suggest diminishing impact of overnight volatility of ES when we look further into the future.

Turning to the out-of-sample results listed in panel B of Table 2.3, we observe that although HAR-F still has on average smaller loss than HAR, except for QLIKE loss in one-week ahead forecast with increasing estimation window. Null hypothesis of the reality check is rejected occasionally. With the rolling scheme and one-week ahead forecast, HAR-F exhibits superior predictive ability if MSE is taken as yardstick and the least accurate forecasts get removed, or if QLIKE loss is used for evaluation. With increasing scheme, no combination of the loss function and the estimation scheme renders HAR-F superior to its competitors. If we look further into the one-month ahead forecast (panel B of Table 2.4), HAR-F shows superior ability in reducing QLIKE loss for both estimation schemes. It also surpasses the competitors in terms of MSE if rolling scheme is adopted. Overall, a fair conclusion drawn from the results of reality check would be that inclusion of overnight volatility of ES still endows the HAR-F model with superior predictive ability on occasions.

**Benchmark: HAR-CJ, SHAR and HARQ**

Next we take three other popular variants of HAR as benchmark and augment them with overnight volatility of ES. We do not consider squared overnight return of SPY, since it has
Figure 2.4: Adjusted $R^2$ and $\theta_N$

Note: The left panel compares adjusted $R^2$'s of HAR, HAR-F and the distributed lag model over multiple forecast horizons. The right panel shows how estimate of $\theta_N$ (i.e., $\hat{\theta}_N$) in the HAR-F model, namely the coefficient of overnight volatility of ES (see (2.7)), varies as the forecast horizon $h$ (number of trading days) extends. $\hat{\theta}_N$ is marked by * and the 95% confidence interval of $\theta_N$ based on robust standard errors is represented by a line.
2.2. Forecasting Daytime Volatility of US Stocks

Table 2.3: In-sample and out-of-sample results of the HAR group (one-week ahead forecasting).

Panel A: In-sample

<table>
<thead>
<tr>
<th>Model</th>
<th>$\alpha_0$</th>
<th>$\alpha_D$</th>
<th>$\alpha_W$</th>
<th>$\alpha_M$</th>
<th>$\theta_N$</th>
<th>$\theta_W$</th>
<th>$\theta_M$</th>
<th>Adj. $R^2$</th>
<th>MSE</th>
<th>QLIKE</th>
</tr>
</thead>
<tbody>
<tr>
<td>HAR</td>
<td>0.1443</td>
<td>0.2245</td>
<td>0.3553</td>
<td>0.2828</td>
<td>0.6622</td>
<td>1.3805</td>
<td>0.1143</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HAR-r(S)</td>
<td>0.1358</td>
<td>0.2011</td>
<td>0.3408</td>
<td>0.2862</td>
<td>0.0901</td>
<td>0.6678</td>
<td>1.3573</td>
<td>0.1116</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HAR-F</td>
<td>0.1239</td>
<td>0.1416</td>
<td>0.2212</td>
<td>0.2489</td>
<td>0.5690</td>
<td>0.7191</td>
<td>1.1478</td>
<td>0.1116</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HAR-F-C</td>
<td>0.1196</td>
<td>0.1302</td>
<td>0.1002</td>
<td>0.5632</td>
<td>0.5556</td>
<td>0.3334</td>
<td>-0.6971</td>
<td>0.7238</td>
<td>1.1278</td>
<td>0.1038</td>
</tr>
<tr>
<td>DL-F</td>
<td>0.4430</td>
<td>1.2385</td>
<td>0.5602</td>
<td>1.7986</td>
<td>0.2445</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: This table lists in-sample and out-of-sample results of the HAR model and its extensions. The dependent variable is the 5-day average of daytime (realized) volatility of SPY and the forecast is made one week (5 days) ahead. The benchmark HAR model is specified in (2.6) (with $h = 5$). The HAR-r(S) model is HAR augmented with squared close-to-09:35 return of SPY itself. The HAR-F model is HAR augmented with overnight volatility of ES (see (2.7)). The HAR-F-C model is obtained by substituting $\sum_{j=1}^{5} RV(S)_t + j$ for the dependent variable in (2.3). DL-F is the distributed lag model below:

$$1 + z \sum_{j=1}^{5} RV(S)_t + j$$

Panel B: Out-of-sample

<table>
<thead>
<tr>
<th>Model</th>
<th>$R^2$</th>
<th>MSE</th>
<th>QLIKE</th>
</tr>
</thead>
<tbody>
<tr>
<td>HAR</td>
<td>0.6679</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>HAR-r(S)</td>
<td>0.6612</td>
<td>1.0080</td>
<td>0.9956</td>
</tr>
<tr>
<td>HAR-F</td>
<td>0.7263</td>
<td>0.7774</td>
<td>0.9195</td>
</tr>
<tr>
<td>HAR-F-C</td>
<td>0.7067</td>
<td>0.8494</td>
<td>0.9482</td>
</tr>
<tr>
<td>DL-F</td>
<td>0.5367</td>
<td>1.1358</td>
<td>1.7336</td>
</tr>
</tbody>
</table>

Note: In the rolling scheme, the window size is fixed at 800. In the increasing scheme, the estimation window starts with the first 800 observations and keeps expanding. The entries are average ratios of the losses of the models to the corresponding ones of the benchmark HAR model. In parentheses are Newey-West estimates of standard errors computed with Bartlett kernel and bandwidth 5.
Table 2.4: In-sample and out-of-sample results of the HAR group (one-month ahead forecasting).

<table>
<thead>
<tr>
<th>Model</th>
<th>$\alpha_0$</th>
<th>$\alpha_D$</th>
<th>$\alpha_W$</th>
<th>$\alpha_M$</th>
<th>$\theta_N$</th>
<th>$\theta_W$</th>
<th>$\theta_M$</th>
<th>Adj. $R^2$</th>
<th>MSE</th>
<th>QLIKE</th>
</tr>
</thead>
<tbody>
<tr>
<td>HAR</td>
<td>0.2944</td>
<td>0.1239</td>
<td>0.3477</td>
<td>0.2475</td>
<td></td>
<td></td>
<td></td>
<td>0.5736</td>
<td>1.3771</td>
<td>0.1714</td>
</tr>
<tr>
<td>(0.0670)</td>
<td>(0.0282)</td>
<td>(0.1228)</td>
<td>(0.0943)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HAR-$r(S)^{N,09:35}$</td>
<td>0.2889</td>
<td>0.1087</td>
<td>0.3382</td>
<td>0.2498</td>
<td>0.0586</td>
<td></td>
<td></td>
<td>0.5766</td>
<td>1.3672</td>
<td>0.1698</td>
</tr>
<tr>
<td>(0.0651)</td>
<td>(0.0322)</td>
<td>(0.1197)</td>
<td>(0.0927)</td>
<td>(0.0304)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HAR-F</td>
<td>0.2832</td>
<td>0.0786</td>
<td>0.2743</td>
<td>0.2290</td>
<td>0.3111</td>
<td></td>
<td></td>
<td>0.5952</td>
<td>1.3071</td>
<td>0.1691</td>
</tr>
<tr>
<td>(0.0663)</td>
<td>(0.0246)</td>
<td>(0.1055)</td>
<td>(0.0946)</td>
<td>(0.0691)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HAR-F-C</td>
<td>0.2835</td>
<td>0.0791</td>
<td>0.2993</td>
<td>0.2097</td>
<td>0.3263</td>
<td>-0.0705</td>
<td>0.0420</td>
<td>0.5950</td>
<td>1.3068</td>
<td>0.1689</td>
</tr>
<tr>
<td>(0.0529)</td>
<td>(0.0406)</td>
<td>(0.1584)</td>
<td>(0.1670)</td>
<td>(0.2638)</td>
<td>(0.2379)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DL-F</td>
<td>0.5915</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.9417</td>
<td></td>
<td>0.4117</td>
<td>1.9014</td>
<td>0.2815</td>
</tr>
<tr>
<td>(0.0680)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.1266)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Panel A: In-sample

<table>
<thead>
<tr>
<th>Model</th>
<th>$R^2$</th>
<th>MSE</th>
<th>QLIKE</th>
</tr>
</thead>
<tbody>
<tr>
<td>HAR</td>
<td>0.6164</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>HAR-$r(S)^{N,09:35}$</td>
<td>0.6082</td>
<td>1.0359</td>
<td>0.9976</td>
</tr>
<tr>
<td>HAR-F</td>
<td>0.6237</td>
<td>0.9611*†</td>
<td>0.9822*</td>
</tr>
<tr>
<td>HAR-F-C</td>
<td>0.6239</td>
<td>1.0029</td>
<td>1.0260</td>
</tr>
<tr>
<td>DL-F</td>
<td>0.3487</td>
<td>1.3765</td>
<td>1.5258</td>
</tr>
</tbody>
</table>

Panel B: Out-of-sample

Note: This table lists in-sample and out-of-sample results of the HAR model and its extensions. The dependent variable is the 22-day average of daytime (realized) volatility of SPY and the forecast is made one month (22 days) ahead. The benchmark HAR model is specified in (2.6) (with $h = 22$). The HAR-$r(S)^{N,09:35}$ is HAR augmented with squared close-to-09:35 return of SPY itself. The HAR-F model is HAR augmented with overnight volatility of ES (see (2.7)). The HAR-F-C model is obtained by substituting $\sum_{j=1}^{22} RV(D)_{t+j}$ for the dependent variable in (2.3). DL-F is the distributed lag model below:

$$\frac{1}{22} \sum_{j=1}^{22} RV(S)_{t+j} = \alpha_0 + \theta_N RV(F)^N_t + \epsilon_{t+1}$$

Panel A: The regressions are based on percentage returns. In parentheses are Newey-West estimates of standard errors computed with Bartlett kernel and bandwidth 5. $\theta_N$ is always the coefficient of overnight volatility measures (squared close-to-09:35 return of SPY itself or volatility of ES).

Panel B: In the rolling scheme, the window size is fixed at 800. In the increasing scheme, the estimation window starts with the first 800 observations and keeps expanding. The entries are average ratios of the losses of the models to the corresponding ones of the benchmark HAR model. Entries marked with a symbol means that null hypothesis of the (modified) reality check (see (2.4) and (2.5)) is rejected at the significance level 5%. *: test is conducted on the full sample; †: test is conducted after removing top 1% least accurate forecasts produced by each model.
2.2. Forecasting Daytime Volatility of US Stocks

been found to bring about very marginal improvement; nor do we consider “complete” extension of the benchmark model, i.e., the extension featuring overnight volatilities of ES over the past month. By contrast, our interest lies primarily in the difference between simply augmenting the benchmark model with overnight volatility of ES and further extending the augmented model in the spirit of the benchmark. Only results for one-day ahead forecast are reported and discussed, since for the untabulated results for forecasts with longer horizons, comments on the group with HAR being benchmark model apply as well.

The first benchmark we consider is the HAR-CJ model proposed by Andersen et al. (2007), which extends HAR by decomposing realized volatility into the empirical counterparts of diffusive and jump variations (see Appendix A). Andersen et al. (2007) use the Bipower variation (BV) proposed by Barndorff-Nielsen and Shephard (2004) to measure diffusive variation, but we follow Corsi et al. (2010) and employ the threshold bipower variation (TBV) instead. We also employ the C-Tz statistic for jump detection. Review of these statistics can be found in Appendix A. Given daytime RV and TBV of SPY, we define

\[ J(S)_t^D \equiv \max(RV(S)_t^D - TBV(S)_t^D)1_{|C-Tz_t|>1.96} = JV(S)_t^D 1_{|C-Tz_t|>1.96}, \]

that is, \( J(S)_t^D \) is equal to \( JV(S)_t^D \) only if the jump detection statistic is significant at the 5% level. The corresponding measure for diffusive variation is given by \( C(S)_t^D \equiv RV(S)_t^D - J(S)_t^D \). The HAR-CJ model\(^{11}\) is specified as

\[ RV(S)_{t+1} = \alpha_0 + \alpha_D C(S)_t^D + \alpha_W C(S)_t^W + \alpha_M C(S)_t^M + \beta_D J(S)_t^D + \theta N RV(F)_t^N + \epsilon_{t+1} \quad (2.8) \]

We consider two extensions featuring overnight information. The first extension augments HAR-CJ with \( RV(F)_t^N \):

\[ RV(S)_{t+1} = \alpha_0 + \alpha_D C(S)_t^D + \alpha_W C(S)_t^W + \alpha_M C(S)_t^M + \beta_D J(S)_t^D + \theta N RV(F)_t^N + \epsilon_{t+1} \quad (2.9) \]

This model will be named HAR-CJ-F. We may further “rough up” the model by decomposing \( RV(F)_t^N \) into \( C(F)_t^N \) and \( J(F)_t^N \), which are defined in a way analogous to \( C(S)_t^D \) and \( J(S)_t^D \) respectively. The resulting model, which we call HAR-CJ-FCJ, takes the form

\[ RV(S)_{t+1} = \alpha_0 + \alpha_D C(S)_t^D + \alpha_W C(S)_t^W + \alpha_M C(S)_t^M + \beta_D J(S)_t^D + \kappa_N C(F)_t^N + \lambda_N J(F)_t^N + \epsilon_{t+1} \quad (2.10) \]

Panel A of Table 2.5 summarizes in-sample results of HAR-CJ, HAR-CJ-F and HAR-CJ-FCJ. Similar to results of the HAR group, past daytime volatilities of SPY are downweighted as soon as overnight volatility measures of ES are augmented. A notable fact is that coefficient

\(^{11}\)In Corsi et al. (2010), the model is called HAR-TCJ to stress the fact that threshold bipower variation rather than bipower variation is employed.
of SPY’s daytime jump (i.e., $\beta_D$) is attenuated as well – it drops from 0.20 in HAR-CJ to 0.05 in HAR-CJ-F. As of in-sample fitting, inclusion of overnight information renders (adjusted) $R^2$ of HAR-CJ-F about 12% higher than that of the benchmark. Comparing HAR-CJ-F and HAR-CJ-FCJ, further decomposition of overnight realized volatility into diffusive and jump variations brings about only 3% increment in $R^2$, but in the meanwhile incurs considerable QLIKE loss.

Out-of-sample results are listed in panel B of Table 2.5, where the same window size 800 is again adopted for both estimation schemes. An immediate observation is that HAR-CJ-F outperforms the benchmark HAR-CJ under any combination of estimation scheme and loss function. For the rolling scheme, the improvement achieved by including overnight volatility of ES is especially sizeable. On the other hand, HAR-CJ-FCJ has poorer forecast than HAR-CJ-F if evaluated by QLIKE loss; under the increasing scheme its QLIKE loss is even higher than that of the benchmark HAR-CJ. Here again we apply the modified reality check with HAR-CJ-F being the baseline. With QLIKE loss the null hypothesis is rejected at the 5% significance level for both estimation schemes. With squared loss the result is a bit different. If the least accurate forecasts are not removed, we obtain $p$-value higher than 30%, whichever estimation scheme is adopted. This is expected, since MSE of HAR-CJ-FCJ is substantially lower than that of HAR-CJ-F. If we implement reality check after removal of the least accurate forecasts in the rolling scheme, $p$-value of the test immediately drops to about 0.01. Again, change of the test results from insignificant to significant may be ascribed to sensibility of squared loss to the least accurate forecasts.

Next we turn to the SHAR model proposed by Chen and Ghysels (2010) and Patton and Sheppard (2015), in which past daily realized volatility of a stock is decomposed into realized semivariances. To be specific, the SHAR model is specified as

$$RV(S)^D_{t+1} = \alpha_0 + \alpha_D^+ RS^+(S)^D_t + \alpha_D^- RS^-(S)^D_t + \alpha W RV(S)^W_t + \alpha M RV(S)^M_t + \epsilon_{t+1}$$

Again we scrutinize two extensions: the first extension, named SHAR-F, simply augments SHAR with overnight realized volatility of ES:

$$RV(S)^D_{t+1} = \alpha_0 + \alpha_D^+ RS^+(S)^D_t + \alpha_D^- RS^-(S)^D_t + \alpha W RV(S)^W_t + \alpha M RV(S)^M_t + \theta_N RV(F)^N_t + \epsilon_{t+1}$$

The second extension, termed SHAR-FS, further decomposes $RV(F)^N_t$ into realized semivariances:

$$RV(S)^D_{t+1} = \alpha_0 + \alpha_D^+ RS^+(S)^D_t + \alpha_D^- RS^-(S)^D_t + \alpha W RV(S)^W_t + \alpha M RV(S)^M_t + \theta_N RS^+(F)^N_t + \theta_N RS^-(F)^N_t + \epsilon_{t+1}$$

In-sample results are tabulated in panel A of Table 2.6. Similar to results presented above,
2.2. **Forecasting Daytime Volatility of U.S. Stocks**

Table 2.5: In-sample and out-of-sample results of the HAR-CJ group (one-day ahead forecasting).

**Panel A: In-sample**

| Model | $\alpha_0$ | $\alpha_D$ | $\alpha_W$ | $\alpha_M$ | $\beta_D$ | $\theta$ | $\kappa$ | $\lambda$ | Adj. $R^2$ | MSE | QLIKE
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>HAR-CJ</td>
<td>0.1201</td>
<td>0.2704</td>
<td>0.5254</td>
<td>0.1510</td>
<td>0.1987</td>
<td>0.5706</td>
<td>2.4193</td>
<td>0.1410</td>
<td>(0.0520)</td>
<td>(0.1646)</td>
<td>(0.2002)</td>
</tr>
<tr>
<td>HAR-CJ-F</td>
<td>0.0707</td>
<td>0.1301</td>
<td>0.2490</td>
<td>0.1146</td>
<td>0.0489</td>
<td>1.0000</td>
<td>0.6971</td>
<td>1.7058</td>
<td>0.1355</td>
<td>(0.0441)</td>
<td>(0.1241)</td>
</tr>
<tr>
<td>HAR-CJ-FCJ</td>
<td>0.1059</td>
<td>0.0521</td>
<td>0.2351</td>
<td>0.1056</td>
<td>0.0892</td>
<td>1.4561</td>
<td>-0.0421</td>
<td>0.7282</td>
<td>1.5300</td>
<td>0.1519</td>
<td>(0.0413)</td>
</tr>
</tbody>
</table>

**Panel B: Out-of-sample**

| Model | $R^2$ | MSE | QLIKE
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>HAR-CJ</td>
<td>0.4736</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>HAR-CJ-F</td>
<td>0.6228</td>
<td>0.6241†</td>
<td>0.8387 ∗</td>
</tr>
<tr>
<td>HAR-CJ-FCJ</td>
<td>0.6765</td>
<td>0.5419</td>
<td>0.9851</td>
</tr>
</tbody>
</table>

Note: The in-sample results are based on percentage returns. In parentheses are Newey-West estimates of standard errors computed with Bartlett kernel and bandwidth 5.

Increasing: In the rolling scheme, the window size is fixed at 800. In the increasing scheme, the estimation window starts with the first 800 observations and expands.

Entries marked with a symbol means that null hypothesis of the (modified) reality check (see (2.4) and (2.5)) is rejected at the significance level 5%.

*: test is conducted on the full sample; †: test is conducted after removing top 1% least accurate forecasts produced by each model.

By each model.
when the benchmark SHAR is augmented with overnight realized volatility of ES, all coefficients of past daytime volatilities of SPY are downweighted. Further disintegrating overnight realized volatility into realized semivariances, the leverage effect of overnight realized negative semivariance immediately manifests, and it is much more impactful than daytime realized negative semivariance of SPY itself, with \( \hat{\theta}_N \) being four times larger than \( \hat{\alpha}_D \). However, similar to results of the HAR-CJ group, further decomposition of the realized volatility incurs higher QLIKE loss for the SHAR-FS model. Turning to panel B of Table 2.6, we observe once again the superior predictive ability of SHAR-F acquired by inclusion of overnight realized volatility of ES. When SHAR-F is taken as baseline, the reality check formally confirms its superior predictive ability if estimated with the rolling scheme and evaluated by QLIKE loss. Similar to in-sample results, further disintegrating \( RV(F)_t^N \) into realized semivariances leads to lower squared loss but much higher QLIKE loss, whichever estimation scheme we adopt.

The last benchmark we consider is the HARQ model introduced by Bollerslev et al. (2016). The HARQ model is specified as

\[
RV(S)_{t+1}^D = \alpha_0 + \left( \alpha_D + \beta_D \sqrt{TPQ(S)_t^D} \right) RV(S)_t^D + \alpha_W RV(S)_t^W + \alpha_M RV(S)_t^M + \epsilon_{t+1} \tag{2.14}
\]

where \( TPQ(S)_t^D \) is the Tri-Power Quarticity of Barndorff-Nielsen and Shephard (2006). Bollerslev et al. (2016) argue that the dynamic modeling of realized volatility for the purpose of forecasting integrated variance, namely the probability limit of realized volatility, suffers from the classical error-in-variables problem. Thus, they suggest to include the quarticity term to account explicitly for the temporal variation in the errors when modeling realized volatility. Roughly speaking, the larger the quarticity, the noisier the realized volatility, and the more should it be attenuated in forecasting. The attenuation can be understood by treating \( \alpha_D + \beta_D \sqrt{TPQ(S)_t^D} \) as the coefficient of \( RV(S)_t^D \): Bollerslev et al. (2016) find that \( \beta_D \) is often negative on empirical data; therefore, the higher is \( TPQ(S)_t^D \), the more is \( RV(S)_t^D \) downweighted.

Like the analysis above, we consider two extensions of HARQ. The first extension, named HARQ-F\(^{12} \), augments HARQ with overnight volatility of ES:

\[
RV(S)_{t+1}^D = \alpha_0 + \left( \alpha_D + \beta_D \sqrt{TPQ(S)_t^D} \right) RV(S)_t^D + \alpha_W RV(S)_t^W + \alpha_M RV(S)_t^M + \theta_N RV(F)_t^N + \epsilon_{t+1} \tag{2.15}
\]

The second extension, named HARQ-FQ, further includes a quarticity term to attenuate the

\(^{12}\)It should be noted that in Bollerslev et al. (2016), HARQ-F refers to the “full HARQ” model which use two more quarticity terms to further attenuate past weekly and monthly realized volatilities; see Equation (12) in their paper. Here we do not consider this “full HARQ” model, since the results in Bollerslev et al. (2016) show that the full model is at best on par with the simpler HARQ model. Therefore, in our paper HARQ-F refers to the HARQ model augmented with overnight volatility of Emini futures.
### 2.2 Forecasting Daytime Volatility of U.S. Stocks

Table 2.6: In-sample and out-of-sample results of the SHAR model and its extensions in forecasting the daytime volatility of SPY and the forecast is made one day ahead. The dependent variable is the percentage return of the security. The forecast is made one day ahead. The benchmark SHAR model is specified in (2.11). The SHAR-F model is defined in (2.12). The SHAR-FS model is defined in (2.13).

#### Panel A: In-sample

<table>
<thead>
<tr>
<th></th>
<th>α0</th>
<th>α+</th>
<th>α−</th>
<th>αM</th>
<th>θN</th>
<th>θ−</th>
<th>Adj. R²</th>
<th>MSE</th>
<th>QLIKE</th>
</tr>
</thead>
<tbody>
<tr>
<td>SHAR</td>
<td>0.0514</td>
<td>-0.3860</td>
<td>1.2250</td>
<td>0.3973</td>
<td>0.1440</td>
<td>0.6120</td>
<td>2.1987</td>
<td>0.1328</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0685)</td>
<td>(0.3977)</td>
<td>(0.4963)</td>
<td>(0.1330)</td>
<td>(0.1082)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SHAR-F</td>
<td>0.0372</td>
<td>-0.2113</td>
<td>0.6378</td>
<td>0.2164</td>
<td>0.1080</td>
<td>0.9092</td>
<td>0.7074</td>
<td>1.6375</td>
<td>0.1278</td>
</tr>
<tr>
<td></td>
<td>(0.0526)</td>
<td>(0.2595)</td>
<td>(0.3447)</td>
<td>(0.1068)</td>
<td>(0.0913)</td>
<td>(0.1998)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SHAR-FS</td>
<td>0.0387</td>
<td>-0.2063</td>
<td>0.5269</td>
<td>0.2391</td>
<td>0.0525</td>
<td>-0.6487</td>
<td>2.8206</td>
<td>0.7489</td>
<td>1.4233</td>
</tr>
<tr>
<td></td>
<td>(0.0485)</td>
<td>(0.1815)</td>
<td>(0.2585)</td>
<td>(0.1030)</td>
<td>(0.0928)</td>
<td>(0.5498)</td>
<td>(0.7405)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Panel B: Out-of-sample

<table>
<thead>
<tr>
<th></th>
<th>R²</th>
<th>MSE</th>
<th>QLIKE</th>
</tr>
</thead>
<tbody>
<tr>
<td>SHAR</td>
<td>0.5479</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SHAR-F</td>
<td>0.6576</td>
<td>0.7262</td>
<td>0.8715</td>
</tr>
<tr>
<td></td>
<td>*</td>
<td>†</td>
<td></td>
</tr>
<tr>
<td>SHAR-FS</td>
<td>0.7127</td>
<td>0.6097</td>
<td>1.0541</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: This table lists in-sample and out-of-sample results of the SHAR model and its extensions. The dependent variable is the realized volatility of SPY and the forecast is made one day ahead. The benchmark SHAR model is specified in (2.11). The SHAR-F model is defined in (2.12). The SHAR-FS model is defined in (2.13). The regressions are based on percentage returns. In parentheses are Newey-West estimates of standard errors. Panel B: In the rolling scheme, the window size is fixed at 800. In the increasing scheme, the estimation window starts with the first 800 observations and keeps expanding. The entries are average ratios of the losses of the models to the corresponding ones with the benchmark model. *: test is conducted on the full sample; †: test is conducted after removing top 1% least accurate forecasts produced by each model.
Table 2.7: In-sample and out-of-sample results of the HARQ group (one-day ahead forecasting).

### Panel A: In-sample

<table>
<thead>
<tr>
<th></th>
<th>$\alpha_0$</th>
<th>$\alpha_D$</th>
<th>$\alpha_W$</th>
<th>$\alpha_M$</th>
<th>$\beta_D$</th>
<th>$\theta_M$</th>
<th>$\kappa_N$</th>
<th>Adj. $R^2$</th>
<th>MSE</th>
<th>QLIKE</th>
</tr>
</thead>
<tbody>
<tr>
<td>HARQ</td>
<td>-0.0362</td>
<td>0.6670</td>
<td>0.3446</td>
<td>0.0822</td>
<td>-0.0080</td>
<td></td>
<td>0.5955</td>
<td></td>
<td>2.2784</td>
<td>0.1464</td>
</tr>
<tr>
<td></td>
<td>(0.0635)</td>
<td>(0.1174)</td>
<td>(0.1383)</td>
<td>(0.1267)</td>
<td>(0.0024)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HARQ-F</td>
<td>-0.0157</td>
<td>0.3549</td>
<td>0.1743</td>
<td>0.0690</td>
<td>-0.0051</td>
<td>0.9482</td>
<td></td>
<td>0.7049</td>
<td>1.6613</td>
<td>0.1309</td>
</tr>
<tr>
<td></td>
<td>(0.0459)</td>
<td>(0.1000)</td>
<td>(0.0987)</td>
<td>(0.1019)</td>
<td>(0.0018)</td>
<td>(0.2028)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HARQ-FQ</td>
<td>-0.0045</td>
<td>0.3656</td>
<td>0.1830</td>
<td>0.0790</td>
<td>-0.0052</td>
<td>0.8464</td>
<td>0.0025</td>
<td>0.7053</td>
<td>1.6589</td>
<td>0.1266</td>
</tr>
<tr>
<td></td>
<td>(0.0510)</td>
<td>(0.1231)</td>
<td>(0.1231)</td>
<td>(0.0791)</td>
<td>(0.0022)</td>
<td>(0.4410)</td>
<td>(0.0114)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Panel B: Out-of-sample

<table>
<thead>
<tr>
<th></th>
<th>$R^2$</th>
<th>MSE</th>
<th>QLIKE</th>
</tr>
</thead>
<tbody>
<tr>
<td>HARQ</td>
<td>0.1949</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>HARQ-F</td>
<td>0.2612</td>
<td>0.7570</td>
<td>0.8396 *</td>
</tr>
<tr>
<td>HARQ-FQ</td>
<td>0.3077</td>
<td>0.5286</td>
<td>0.9094</td>
</tr>
</tbody>
</table>

Note: This table lists in-sample and out-of-sample results of the HARQ model and its extensions. The dependent variable is the daytime (realized) volatility of SPY and the forecast is made one day ahead. The benchmark HARQ model is specified in (2.14). The HARQ-F model is defined in (2.15). The HARQ-FQ model is defined in (2.16).

Panel A: The regressions are based on percentage returns. In parentheses are Newey-West estimates of standard errors computed with Bartlett kernel and bandwidth 5.

Panel B: In the rolling scheme, the window size is fixed at 800. In the increasing scheme, the estimation window starts with the first 800 observations and keeps expanding. The entries are average ratios of the losses of the models to the corresponding ones of the benchmark HARQ model. Entries marked with a symbol means that null hypothesis of the (modified) reality check (see (2.4) and (2.5)) is rejected at the significance level 5%. *: test is conducted on the full sample; †: test is conducted after removing top 1% least accurate forecasts produced by each model.
2.2. Forecasting Daytime Volatility of US Stocks

Overnight volatility:

\[
RV(S)^D_t = \alpha_0 + \left( \alpha_D + \beta_D \sqrt{TPQ(S)^D_t} \right) RV(S)^D_t + \alpha_W RV(S)^W_t + \alpha_M RV(S)^M_t + \left( \theta_N + \kappa_N \sqrt{TPQ(F)^N_t} \right) RV(F)^N_t + \epsilon_{t+1}
\] (2.16)

Both in-sample and out-of-sample results are listed in Table 2.7. We draw three notable findings from the in-sample results. First, inclusion of overnight volatility of the Emini futures once again brings about pronounced improvement in fitting. For example, the adjusted \(R^2\) of the HARQ-F model is about 10% larger than that of the benchmark HARQ. Second, further introducing the tri-power quarticity of overnight Emini futures itself only renders the fitting marginally better. Third, the parameter \(\beta_D\) is significantly negative in all three models, which agrees with the rationale of Bollerslev et al. (2016) that daytime realized volatility is attenuated on days of high measurement error; however, the parameter \(\kappa_N\) in the HARQ-FQ model is positive and insignificant, which suggests that the overnight realized volatility of ES is not attenuated.

The out-of-sample results are somewhat striking. The Mincer-Zarnowitz \(R^2\)’s are much smaller than those for the HAR-CJ and SHAR groups, suggesting quite poor forecast of the HARQ models. As of comparison of the three models, we find that both extensions produce substantially more accurate forecasts than the benchmark HARQ. Which extension is better depends again on the loss function used for evaluation: the HARQ-FQ model makes more accurate forecast when squared error is taken as yardstick, while the HARQ-F model performs better if QLIKE loss is in concern.

In summary, with any of HAR-CJ, SHAR and HARQ being benchmark, including overnight volatility of ES renders forecast of daytime volatility of SPY more accurate. Whether or not further extending the augmented model in the spirit of the benchmark model depends ultimately on the loss function used for evaluation. If squared loss is in concern, our results are in favor of further extensions (HAR-CJ-FCJ, SHAR-FS and HARQ-FQ). If, on the other hand, the models are appraised by asymmetric loss function such as QLIKE, further extensions become sub-optimal, and augmenting the benchmark model simply with overnight realized volatility of ES suffices to induce sizeable improvement in forecast accuracy.

2.2.5 Individual Stocks

Thus far we have been concentrating on forecasting daytime volatility of the ETF SPY. The remarkable improvements induced by inclusion of overnight volatility of ES is arguably due to the intimate relation between volatility of SPY and volatility of ES. We discuss such relation in Appendix C. We would expect to gain equally sizeable improvements in forecasting daytime volatility of individual stocks if single stocks futures (henceforth SSF) were traded.
overnight. However, SSF of the US market is traded on OneChicago Exchange only from 08:15 to 15:00 (Chicago time) on business days. Moreover, on average there are only a few transactions for most SSFs in a day, rendering construction of high-frequency-data-based volatility measures more difficult. Given these facts, we find it more feasible as well as more plausible to include overnight volatility of the Emini S&P 500 futures, which gauges market-wide risk, in forecasting daytime volatility of individual stocks. In the sequel we forecast volatility of 24 individual stocks, taking the HAR model (2.1) as benchmark and augmenting it with overnight volatility of ES. The augmented model is still termed HAR-F. The comparison is mainly made between HAR and HAR-F13.

In-sample results are tabulated in Table 2.8. Mixed conclusions may be drawn from the results: whether inclusion of overnight volatility of ES leads to improvement hinges on the loss function used to evaluate the forecasts. If the forecasts are evaluated by squared loss, which is also the objective function being minimized in estimation, the HAR-F model is overall preferable to the benchmark HAR, as for most stocks the MSE reduces and the adjusted $R^2$ increases, sometimes notably (e.g., AA, AXP and XOM). On the other hand, if the models are evaluated by QLIKE loss, then the results are by and large neutral or even stand against inclusion of overnight volatility of ES. Interestingly, many stocks having sizeable cut in MSE have, contrariwise, considerably larger QLIKE loss. Examples include AA and XOM, whose MSEs drop drastically after inclusion of overnight volatility of ES, but meanwhile their QLIKE losses surge by 10% and 16% respectively. For a few stocks, such as AIG, BAC and C, we do observe reduced QLIKE loss. Perhaps not by chance, MSEs of forecasts of the three firms are just marginally reduced after inclusion of overnight volatility of ES. Such results is likely to be caused by the suffering of these firms during the financial crisis in 2008, during which many unusually high volatilities and many inaccurate forecasts cluster.

Out-of-sample results are listed in Table 2.9. Again we take HAR-F as baseline and apply the (modified) reality check to contrast its performance with HAR. Like the in-sample results, the out-of-sample results are not uniformly in favor of the HAR-F model either. When QLIKE loss is in concern, for 11 out of 24 stocks the reality check confirms superior predictive ability of the HAR-F model (without removing the least accurate forecasts). But when squared loss is concerned, for all stocks predictive ability of HAR and HAR-F are indistinguishable unless we remove some least accurate forecasts, though the latter shows on average lower numeric loss. There is one interesting fact about how loss of HAR-F relative to that of HAR varies with the correlation between the stock’s daytime volatility and ES’s daytime volatility. In Figure 2.5 we plot the loss ratios against correlation of the daytime volatilities.

---

13We initially conducted analysis of three models, the third one being HAR augmented with squared overnight return (i.e., close-to-open return) of the stock itself. Unlike the case for SPY, a comparison between this model and HAR-F seems a priori a fair match, as overnight volatility of ES no longer approximates overnight volatility of the stock in concern, and conclusions in Appendix C no longer hold. The results turned out to counter our intuition, however. For almost all stocks, the third model performs even much worse than HAR, let alone HAR-F. Only for the stock VZ did we observe slightly more accurate forecast produced by this model. Therefore, we choose not to include this model in the discussion below.
3.2. Forecasting

In-sample results of individual stocks (one-day ahead forecasting)

Table 2.8:
<table>
<thead>
<tr>
<th>Ticker</th>
<th>Rollng</th>
<th></th>
<th></th>
<th>Increasing</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Adj.$R^2$</td>
<td>MSE</td>
<td>QLIKE</td>
<td>Adj.$R^2$</td>
<td>MSE</td>
<td>QLIKE</td>
</tr>
<tr>
<td>AA</td>
<td>0.5144</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.5236</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td>0.6407</td>
<td>0.6196</td>
<td>0.1069</td>
<td>0.6660</td>
<td>0.6265</td>
<td>1.1078</td>
</tr>
<tr>
<td>AIG</td>
<td>0.4909</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.5189</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td>0.4958</td>
<td>0.1090</td>
<td>0.9508$^+$</td>
<td>0.5232</td>
<td>0.1018</td>
<td>0.9406$^+$</td>
</tr>
<tr>
<td>AXP</td>
<td>0.4486</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.4682</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td>0.5597</td>
<td>0.7501</td>
<td>1.1099</td>
<td>0.5778</td>
<td>0.7869</td>
<td>1.0729</td>
</tr>
<tr>
<td>BA</td>
<td>0.6000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.6078</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td>0.6751</td>
<td>0.7854$^+$</td>
<td>0.9253$^+$</td>
<td>0.6773</td>
<td>0.8673</td>
<td>0.9669</td>
</tr>
<tr>
<td>BAC</td>
<td>0.5700</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.5729</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td>0.5638</td>
<td>0.1006</td>
<td>0.9210$^+$</td>
<td>0.5722</td>
<td>0.9921</td>
<td>0.9290$^+$</td>
</tr>
<tr>
<td>C</td>
<td>0.2625</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.2716</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td>0.2649</td>
<td>0.9366</td>
<td>0.9825$^+$</td>
<td>0.2763</td>
<td>0.9355</td>
<td>0.9025$^+$</td>
</tr>
<tr>
<td>CAT</td>
<td>0.6524</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.6582</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td>0.7213</td>
<td>0.7779$^+$</td>
<td>0.9560$^+$</td>
<td>0.7280</td>
<td>0.7985</td>
<td>1.0070</td>
</tr>
<tr>
<td>DD</td>
<td>0.5684</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.5840</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td>0.7175</td>
<td>0.6426</td>
<td>0.9375$^+$</td>
<td>0.7317</td>
<td>0.6791</td>
<td>1.0406</td>
</tr>
<tr>
<td>DIS</td>
<td>0.4207</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.5330</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td>0.5360</td>
<td>0.6771</td>
<td>0.9087$^+$</td>
<td>0.6223</td>
<td>0.9265</td>
<td>1.0145</td>
</tr>
<tr>
<td>GE</td>
<td>0.5139</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.5425</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td>0.5276</td>
<td>0.9436$^+$</td>
<td>0.9122$^+$</td>
<td>0.5560</td>
<td>0.9765</td>
<td>0.9645</td>
</tr>
<tr>
<td>HD</td>
<td>0.5278</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.5441</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td>0.6618</td>
<td>0.7258</td>
<td>1.0153</td>
<td>0.6686</td>
<td>0.8388</td>
<td>1.1485</td>
</tr>
<tr>
<td>IBM</td>
<td>0.5445</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.5601</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td>0.6275</td>
<td>0.8058</td>
<td>0.9628</td>
<td>0.6383</td>
<td>0.8865</td>
<td>1.0825</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ticker</th>
<th>Rollng</th>
<th></th>
<th></th>
<th>Increasing</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Adj.$R^2$</td>
<td>MSE</td>
<td>QLIKE</td>
<td>Adj.$R^2$</td>
<td>MSE</td>
<td>QLIKE</td>
</tr>
<tr>
<td>JNJ</td>
<td>0.3476</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.4909</td>
<td>0.6756</td>
<td>0.9089$^+$</td>
</tr>
<tr>
<td></td>
<td>0.4889</td>
<td>0.5232</td>
<td>0.1018</td>
<td>0.5232</td>
<td>0.1018</td>
<td>0.9406$^+$</td>
</tr>
<tr>
<td>JPM</td>
<td>0.5138</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.5189</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td>0.5676</td>
<td>0.8884</td>
<td>3.4673</td>
<td>0.5880</td>
<td>0.9152</td>
<td>1.1925</td>
</tr>
<tr>
<td>KO</td>
<td>0.4963</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.5186</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td>0.5987</td>
<td>0.7531</td>
<td>0.9453$^+$</td>
<td>0.6066</td>
<td>0.9254</td>
<td>1.1079</td>
</tr>
<tr>
<td>MCD</td>
<td>0.2807</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.3390</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td>0.4008</td>
<td>0.7497</td>
<td>0.9981</td>
<td>0.4179</td>
<td>1.0615</td>
<td>1.2776</td>
</tr>
<tr>
<td>MMM</td>
<td>0.2970</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.3290</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td>0.4356</td>
<td>0.7093</td>
<td>0.7978$^+$</td>
<td>0.4561</td>
<td>0.8104</td>
<td>0.9515</td>
</tr>
<tr>
<td>MRK</td>
<td>0.0394</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.0876</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td>0.0709</td>
<td>0.9516</td>
<td>0.9368$^+$</td>
<td>0.1399</td>
<td>0.9593</td>
<td>1.0626</td>
</tr>
<tr>
<td>PG</td>
<td>0.3441</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.3938</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td>0.4432</td>
<td>0.7511</td>
<td>0.9580</td>
<td>0.5114</td>
<td>0.8214</td>
<td>1.1051</td>
</tr>
<tr>
<td>UTX</td>
<td>0.5093</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.5501</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td>0.6572</td>
<td>0.6421</td>
<td>0.8759$^+$</td>
<td>0.6734</td>
<td>0.7929</td>
<td>1.0203</td>
</tr>
<tr>
<td>VZ</td>
<td>0.4259</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.4874</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td>0.5191</td>
<td>0.7778</td>
<td>1.0093</td>
<td>0.5689</td>
<td>0.9230</td>
<td>1.1405</td>
</tr>
<tr>
<td>WHR</td>
<td>0.4259</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.4874</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td>0.5191</td>
<td>0.7778</td>
<td>1.0093</td>
<td>0.5689</td>
<td>0.9230</td>
<td>1.1405</td>
</tr>
<tr>
<td>WMT</td>
<td>0.4068</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.4523</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td>0.6122</td>
<td>0.6339</td>
<td>1.0717</td>
<td>0.6331</td>
<td>0.8027</td>
<td>1.1984</td>
</tr>
<tr>
<td>XOM</td>
<td>0.4141</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.4491</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td>0.5352</td>
<td>0.6969</td>
<td>0.9979</td>
<td>0.5808</td>
<td>0.7623</td>
<td>1.1878</td>
</tr>
</tbody>
</table>

Note: This table lists out-of-sample results of HAR (the shaded) and HAR-F (the unshaded) applied to forecasting daytime volatility of individual stocks. The forecasts are made one day ahead. In the rolling scheme, the window size is fixed at 800. In the increasing scheme, the estimation window starts with the first 800 observations and keeps expanding. The entries are average ratios of the losses of the models to the corresponding ones of the benchmark HAR model. Entries marked with a symbol means that null hypothesis of the (modified) reality check (see (2.4) and (2.5)) is rejected at the significance level 5%. $^*$: test is conducted on the full sample; $^+$: test is conducted after removing the top 1% least accurate forecasts.
2.3. DISSECTING THE FORECAST

Figure 2.5: Loss ratio against correlation of daytime volatilities

Note: These graphs plot out-of-sample (average) loss ratio of the HAR-F model to the HAR model of each individual stock against the correlation coefficients of the stock’s daytime realized volatility and that of ES.

The upper panels suggest that the higher the correlation between the stock's and ES's volatilities, the more the squared loss gets reduced by inclusion of overnight volatility of ES. The lower panels suggest the contrary. Higher correlation is accompanied by higher QLIKE loss, especially with the increasing estimation scheme. Summarizing the findings, it may be fair – though indecisive – to say that the effect of including overnight volatility of ES seemingly depends on peculiar characteristics of the firm, and which of HAR and HAR-F should be adopted in practice depends heavily on the loss function.

2.3 Dissecting the forecast

2.3.1 The interaction with the daytime European market

In the process of gleaning valuable information from the overnight ES market, its interaction with the daytime European market catches our attention. Circumscribed by data availability, we can only scrutinize interplay of the two markets during 2010 to 2013, which happens to cover the European Sovereign Debt Crisis. Below we focus primarily on econometric properties of the price and volatility series of the two markets. For studies dedicated to the debt crisis, see, e.g., Lane (2012) and Argyrou and Kontonikas (2012).

The principal characteristic reflected by the data is the highly positive correlation of the overnight ES market and the daytime FTSE market. On each calendar day, prices of ES and FTSE 100 are sampled every minute during 02:00 to 08:30 Chicago Time, which corresponds...
Figure 2.6: Histogram of correlation coefficients

Note: The figure shows the histogram of correlation coefficients of intra-day price series of the Emini S&P 500 futures and the FTSE 100 futures. On each calendar day, we compute the correlation coefficient of the one-minute price series of ES (sampled during 02:00 to 08:30 Chicago Time) and the one-minute price series of FTSE 100 (sampled during 08:00 to 14:30 Chicago Time).

to 08:00 to 14:30 London Time. We then compute their correlation coefficient after obtaining the two price series. Figure 2.6 presents histogram of the correlation coefficients of the two markets’ price series. On over 20% of the days, the intra-day price series have a correlation coefficient higher than 0.95; the proportion further goes up to 40% if we consider correlation coefficients higher than 0.9. High correlation of the price series translates into high correlation of the volatilities, as can be observed from Figure 2.7, where we plot realized volatilities of the two markets with five-minute data sampled during the aforementioned hours. The two volatility series almost mirror each other; their correlation coefficient during the sample period is up to 0.95.

We may exemplify the close comovement of the overnight ES market and the daytime FTSE market through the lens of high-frequency data. Figure 2.8 presents one-minute price series of the two markets on 9th August 2011, on which FTSE 100 has the highest volatility during our sample period. After climbing for a short while right after opening, the FTSE 100 sheds stiffly, losing more than 300 points, which is its biggest intra-day fall since July 2010. Soon after touching the daily low, it rallies 6% and almost recovers all losses by 14:30 (London time). The overnight ES market is seen to keep a close pace with the daytime FTSE.

---

14 The sampling window is chosen mainly to align the price series: trades of the FTSE 100 futures start from 08:00 in London, namely 02:00 in Chicago; on the other hand, the US stock market opens at 09:30 in New York, which corresponds to 08:30 in Chicago and 14:30 in London.

15 Here and below, the prices are denominated in Euro and US dollar respectively. We do not use a common denominator, say US dollar, since this is not the primary focus of our analysis.
2.3. DISSECTING THE FORECAST

Figure 2.7: Realized Volatilities of FTSE 100 and ES during 4th January 2010 and 31st December 2013

Note: This graph shows realized volatilities of the FTSE 100 futures and the Emini S&P 500 futures during 4th January 2010 and 31st December 2013. On each calendar day, realized volatility of ES is computed with 5-minute price series sampled during 02:00 and 08:30 Chicago time; realized volatility of FTSE 100 is computed with data sampled at the same frequency during the same period, that is, 08:00 to 14:30 London Time.

100: humps and troughs of the volatilities are by and large overlapped.

How does the interaction between the two markets translate into more accurate volatility forecast achieved by inclusion of the overnight information of ES? From the above observations, it is natural to infer that the overnight US market would resonate and its ensuing daytime volatility would surge if the daytime European market is vibrated by an impactful event. Overlooking such event, if any, in forecasting daytime volatility of the US market on the coming day could lead to great loss in accuracy. Panel A of Table 2.10 corroborates our argument. We first find out the top 5% most volatile days of the daytime FTSE 100 market\(^{16}\) and compare forecasts of daytime volatility of the US market on the coming day calculated with and without inclusion of overnight information. For all cases inclusion of overnight information brings about sizeable improvement in the forecast accuracy; the losses of HAR-F are about 20% to 35% smaller than those of HAR. On 9th August 2011, the day we just discussed, realized volatility of the daytime US market rises up to 2.43. The forecasts of HAR-F are 1.77 and 1.86 under the rolling and the increasing schemes respectively; the forecasts of HAR are 1.04 and 0.98, which largely understate the escalated risk. Without including the overnight information, the HAR model misses the roller-coaster ride of both the daytime European and the overnight US market that Figure 2.8 shows\(^{17}\). In the same panel of Table

---

\(^{16}\) As above, we calculate realized volatilities of the FTSE 100 futures from 08:00 to 14:30 London local time.

\(^{17}\) Indeed, on this day several stock markets that open earlier than the US market, including Japan, Hong
we also list loss ratios conditional on the bottom 95% volatilities of the daytime FTSE 100 market. There is large cut in improvement in forecast accuracy attained by inclusion of overnight information of ES. For both of the in-sample forecast and the out-of-sample forecast with increasing scheme, HAR-F incurs higher QLIKE loss than HAR. These observations lead us to the conclusion that part of the improvement we find in Section 2.2 can be ascribed to the interaction between overnight ES market and the daytime European market. Of course, we do not claim that the European market necessarily “leads” the US market. That an event originates from Europe (like the debt crisis we discuss here) impacts the US market is but one possibility of the many; the transatlantic effect could take the converse direction or could be exerted on the two markets simultaneously.

2.3.2 The effect of FOMC meeting

It is well known that stock market volatility evolves abnormally around the FOMC meeting. Bomfim (2003) finds evidence of the “calm-before-the-storm” effect in volatility before FOMC announcements. He also finds that surprise in actual interest rate (i.e., difference between actual and expected target federal funds rates) decisions causes stock market volatility to rise in the short run. Lucca and Moench (2015) find that on scheduled FOMC days, volatility of the stock market is relatively low before the announcement, but it spikes steeply at 2:15pm when the FOMC statement is released.

These observations inspire us to examine volatility forecasts made on FOMC meeting days. Our sample period covers 107 FOMC meeting days in total. Again, for both HAR and HAR-F we divide the volatility forecasts into two groups according to whether there is FOMC meeting on the day when we make the forecast. For each group, we calculate both in-sample and out-of-sample loss ratios. The results are tabulated in panel B of Table 2.10. There are several notable facts. First, when there is no FOMC meeting, inclusion of overnight volatility of the Emini futures market renders the forecast more accurate, as evidenced by both in-sample and out-of-sample results. Second, when there is FOMC meeting, inclusion of overnight volatility occasionally incurs higher QLIKE loss. For example, when there is FOMC meeting, the in-sample result shows that HAR-F has on average 2.7% higher QLIKE loss. Third, inclusion of overnight volatility of ES typically leads to more improvement in forecast accuracy when there is no FOMC meeting. For example, when there is no FOMC meeting, inclusion of the overnight volatility of ES reduces the squared error by 37%, which is substantially higher than the reduction in the presence of FOMC meeting (= 22%). The only exception is the QLIKE loss under the increasing estimation scheme. In this case, QLIKE loss of the forecast made by HAR-F is 13% smaller than forecast by HAR in the presence of FOMC meeting, while the QLIKE losses of HAR-F and HAR are almost the same when there is no FOMC meeting.

Kong and Australia, all crash heavily.
2.3. Dissecting the Forecast

**Figure 2.8:** One-minute Price Series of FTSE 100 and Emini S&P 500 on 9th August 2011

*Note:* This graph shows one-minute transaction price series of the FTSE 100 futures (the upper panel) and of the Emini S&P 500 futures (the lower panel) during 02:00 and 08:30 (GMT-5) on 9th August 2011. The prices are sampled by previous tick interpolation, viz. at every minute the most recent price is used. The shaded rectangles in the background are realized volatilities of every 30-minute window, which are adjusted by a factor so that height of the tallest rectangle is exactly 80% of range of the y-axis.
Table 2.10: Stratified loss ratios

<table>
<thead>
<tr>
<th>Panel A: Stratified according to daytime volatility of FTSE 100</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MSE</strong></td>
</tr>
<tr>
<td>In-Sample</td>
</tr>
<tr>
<td>Top 5% RV_{FTSE}</td>
</tr>
<tr>
<td>HAR</td>
</tr>
<tr>
<td>HAR-F</td>
</tr>
<tr>
<td>Bottom 95% RV_{FTSE}</td>
</tr>
<tr>
<td>HAR</td>
</tr>
<tr>
<td>HAR-F</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Panel B: Stratified according to whether there is FOMC meeting</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MSE</strong></td>
</tr>
<tr>
<td>In-Sample</td>
</tr>
<tr>
<td>FOMC meeting</td>
</tr>
<tr>
<td>HAR</td>
</tr>
<tr>
<td>HAR-F</td>
</tr>
<tr>
<td>No FOMC meeting</td>
</tr>
<tr>
<td>HAR</td>
</tr>
<tr>
<td>HAR-F</td>
</tr>
</tbody>
</table>

Note: This table presents stratified loss ratios, defined as the ratio of the average loss of HAR-F to the average loss of HAR, according to the level of daytime volatility of the FTSE 100 futures market (panel A), or to whether there is FOMC meeting on the day when we make the forecast (panel B). The columns “Rolling” and “Increasing” are out-of-sample results.

2.4 Conclusion

We carry out empirical analysis of the effect brought by inclusion of overnight volatility of the Emini S&P 500 futures (a.k.a. ES) in forecasting daytime volatility of the ETF SPDR (a.k.a. SPY) and 24 stocks of the US market. The analysis and conclusions are summarized as follows.

For the ETF SPY, we start off by taking the HAR model as benchmark. We augment it by the latest overnight volatility of ES. We have also considered two other extensions: the first one includes squared close-to-09:35 return of SPY itself and the second one includes past overnight volatilities of ES up to one month. For one-day ahead forecast, we find that inclusion of the latest overnight volatility of ES brings about sizeable improvements in terms of all measures – the (adjusted) $R^2$ is about 15% higher, and both MSE and QLIKE are greatly reduced. Superior predictive ability of the model featuring overnight information is confirmed by Reality Check in out-of-sample analysis. We also find that further including averages of past weekly and monthly overnight volatilities of ES only marginally reduces squared loss and meanwhile incurs higher QLIKE loss. On the other hand, the improvement in forecast
2.4. Conclusion

Accuracy brought by inclusion of overnight information diminishes rapidly as the forecast horizon extends.

We then take three variants of HAR – the HAR-CJ model, the SHAR model and the HARQ model – as benchmark. For each benchmark model we consider two extensions: one simply augments the benchmark with overnight volatility of ES, and the other further extends the augmented model in the spirit of the benchmark. For all models, we find that inclusion of overnight volatility of ES renders the forecast much more accurate. On the other hand, whether the augmented model should be further extended in the spirit of the benchmark hinges on the loss function used to evaluate the models. Our findings are in favor of further extension of the overnight volatility if squared loss is concerned, but stand against such decomposition if the models are appraised by QLIKE loss.

The remarkable improvement achieved by inclusion of overnight volatility of ES is predicated on the intimate relation between volatility of an underlying stock and volatility of futures contracts written on it, the discussion on which is provided in Appendix C. In the absence of actively traded single stock futures, we consider forecasting daytime volatility of individual stocks with and without inclusion of volatility of the overnight Emini futures market, still taking HAR as benchmark. For some stocks we observe more accurate forecasts resulting from inclusion of overnight volatility of ES, while for the others we can not draw the same conclusion. An interesting finding concerns how the loss varies with correlation between the stock’s volatility and ES’s volatility: the higher the correlation coefficient, the larger the reduction in squared loss brought by inclusion of overnight volatility of ES, but in the meanwhile the lower the cut in QLIKE loss.

Finally, we dissect the forecasts by relating them to the daytime European market and the FOMC meeting. Stratifying losses of the forecasts according to volatility of the daytime European market, we find that inclusion of overnight information of the Emini futures enables the model to timely take into account impactful events that possibly originate from outside the US and hence swiftly catch up with the escalated risk of the daytime US market on the coming day. On the other hand, stratifying losses according to whether there is FOMC meeting, we find that inclusion of overnight volatility of ES typically leads to more improvement in forecast accuracy in the absence of FOMC meeting.

Acknowledgment

We thank Asger Lunde, Kim Christensen, Anders B. Kock, Leopoldo Catania, Kevin Sheppard, Fulvio Corsi, Eric Hillebrand, Ingmar Nolte, and seminar participants at CREATES, the University of Oxford, and the IAAE 2018 annual conference in Montréal for helpful comments. We are grateful to Dr. Sandra Nolte for providing the European FTSE 100 data. Part of this research was conducted while the first author was visiting the Department of Economics, University of Oxford, and the generosity and hospitality of the department are
gratefully acknowledged. The research is supported by CREATEs, funded by the Danish Na-
tional Research Foundation. The research project leading to these results received funding
from the European Union Research and Innovation programme Horizon 2020 under grant
agreement No. 675044 (BigDataFinance).
Appendix A  Volatility Measures

A.1  The setting

In this appendix we review the volatility measures employed in our empirical analysis. We let \( X \) denote the logarithmic price of an asset, which is either a stock or a futures contract written on it. When we refer to the stock (resp. the futures), we use \( S \) (resp. \( F \)) to denote the logarithmic price instead. We assume that \( X \) is an Itô semimartingale:

\[
X_t = X_0 + \int_0^t b_{X,s} \, ds + \int_0^t \sigma_{X,s} \, dW_s + J_t
\]

Here \( b_X \) and \( \sigma_X \) are, respectively, scalar and \( d \)-dimensional predictable processes satisfying the usual measurability and integrability conditions; \( W \) is a \( d \)-dimensional standard Brownian motion; and \( J \) is a jump process. The spot volatility process \( \sigma_X \) in (A1) and the Brownian motion \( W \) are purposefully defined to be multi-dimensional as opposed to the oft-used one-dimensional assumption in the literature. The reason will become apparent in Appendix C, where we study the relation between volatilities of SPY and ES.

The quadratic variation (see, e.g., Andersen and Bollerslev (1998), Andersen et al. (2001a) and Andersen et al. (2001b)) of \( X \) is defined by

\[
[X]_t \equiv \int_0^t ||\sigma_{X,s}||^2 \, ds + \sum_{0<s\leq t} (\Delta X_s)^2
\]

The two summands on the right hand side of Equation (A2) are often called diffusive variation and jump variation respectively. These measures quantify risk of different categories. The quadratic variation and the jump variation measure the asset's gross risk and tail risk respectively. The three variations in Equation (A2) are in essence latent and need be estimated in practice.

A.2  Realized volatility

For quadratic variation we resort to its empirical counterpart, the realized volatility. Assume the availability of an equidistant grid of observations \( X_{s+i\Delta_n}, i = 0, 1, \cdots, n \) over the period \([s,t]\), where \( \Delta_n \) is the time elapsed between two adjacent observations and \( n = \left\lfloor \frac{t-s}{\Delta_n} \right\rfloor \).

We denote by \( r(X)_{s+i\Delta_n} \equiv X_{s+i\Delta_n} - X_{s+(i-1)\Delta_n} \) the \( i \)-th intra-grid return. Realized volatility of \( X \) over the grid is defined below and, as is well known, consistently estimate \([X]_t - [X]_s\) when \( \Delta_n \) tends to zero: (see, e.g., Chapter VI.4, He et al. (1992))

\[
RV(X)_{s,t} \equiv \sum_{i=1}^{n} r(X)_{s+i\Delta_n}^2 \xrightarrow{p} [X]_t - [X]_s
\]
Throughout this paper, all realized volatilities, be it daytime or overnight, are estimated with ∆ₙ being five minutes. The 5-min realized volatility is found to mitigate the effect of market microstructure noise, whose presence may cause the semimartingale assumption (A1) to be violated (Hansen and Lunde (2006); Bandi and Russell (2008)). Many noise-robust estimators have been developed over the years (see, e.g., Zhang (2006), Barndorff-Nielsen et al. (2008) and Jacod et al. (2009)). Nonetheless, the extensive study by Liu et al. (2015) suggests that the 5-min realized volatility is hardly outperformed in terms of estimation accuracy, so we adopt it for our empirical study.

A.3 Threshold bipower variation and jump variation

The HAR-CJ model requires measures of the diffusive and jump variations respectively. We employ the threshold bipower variation (TBV) proposed by Corsi et al. (2010), which derives from the bipower variation (Barndorff-Nielsen and Shephard (2004)) and the threshold realized volatility (Mancini (2004) and Mancini (2009)). Over the same grid Xₛ+iΔₙ, i = 0, 1, ⋯, n, the threshold bipower variation is given by

\[
TBV(X)_{s,t} \equiv \frac{\pi}{2} \sum_{i=2}^{n} |r(X)_{s+i\Delta_n}|^2 |r(X)_{s+(i-1)\Delta_n}|^2 \mathbb{1}_{|r(X)_{s+i\Delta_n}|^2 \leq \theta_s+i\Delta_n} \mathbb{1}_{|r(X)_{s+(i-1)\Delta_n}|^2 \leq \theta_s+(i-1)\Delta_n} \tag{A4}
\]

where \(\mathbb{1}_{A}\) is an indicator function of a measurable set \(A\), and \(\theta_s+i\Delta_n \equiv c_\theta^2 \hat{V}_{s+i\Delta_n}\) is a sequence of (local) thresholds, with \(c_\theta\) being a constant threshold scale and \(\hat{V}_{s+i\Delta_n}\) a sequence of local volatility estimates. Threshold bipower variation serves as a weakly consistent estimator for the diffusive variation, i.e.,

\[
TBV(X)_{s,t} \overset{p}{\longrightarrow} \int_s^t ||\sigma_{X,u}||^2 \, du
\]

Although the jump variation can be trivially estimated by the difference between realized volatility and threshold bipower variation, such trivial estimator is deficient as it might attain negative value on finite sample. Instead, we truncate negative values and make use of the following estimator of the jump variation:

\[
JV(X)_{s,t} \equiv \max(RV(X)_{s,t} - TBV(X)_{s,t}, 0) \tag{A5}
\]

In our empirical analysis, we let \(\Delta_n = 5\) minutes, namely that the prices are sampled every five minutes. The threshold scale \(c_\theta\) is set to 3 as in Corsi et al. (2010). On the other hand, the local volatility estimates \(\hat{V}_{s+i\Delta_n}\) are obtained following exactly the procedures described in Appendix B of Corsi et al. (2010).
A. Volatility Measures

A.4 Jump detection statistic

Another component of the HAR-CJ model is the jump detection statistic. Whereas the statistic can be built on the threshold bipower variation, Corsi et al. (2010) argue that such statistic would be problematic since returns exceeding the threshold are completely annihilated, thereby inducing a negative bias in finite sample. They propose to attenuate the bias by replacing the summand in (A4) with

\[ Z_1(r(X)_{s+i\Delta n}, \theta_{s+i\Delta n}) Z_1(r(X)_{s+(i-1)\Delta n}, \theta_{s+(i-1)\Delta n}), \]

where for any \( \gamma \in \mathbb{R} \),

\[ Z_\gamma(x, y) \equiv \begin{cases} |x|, & \text{if } x^2 \leq y \\ \frac{1}{2n(-c_0)} \left( \frac{2}{c_0} \right)^{\gamma/2} \Gamma\left( \frac{\gamma+1}{2}, \frac{c_0^2}{2} \right), & \text{if } x^2 > y \end{cases} \]  

(A6)

In other words, returns which would have been truncated in the threshold bipower variation are replaced with the expected values as if they were normally distributed. The resulting quantity, termed corrected threshold bipower variation (CTBV), is

\[ CTV(X)_{s,t} \equiv \mu_4^{-3} \Delta_n^{-1} \sum_{i=3}^{n} \prod_{j=1}^{3} Z_4^i(r(X)_{s+(i-j+1)\Delta n}, \theta_{s+(i-j+1)\Delta n}) \]

where \( \mu_4 \equiv \mathbb{E}[|N|^{4/3}] \approx 0.8309 \) with \( N \) being a standard normal random variable and \( Z_4^i(\cdot, \cdot) \) is obtained by setting \( \gamma = 4/3 \) in Equation (A6). The resulting jump detection statistic, termed C-Tz statistic, is defined by:

\[ C-Tz(X)_{s,t} \equiv \left( \frac{n}{t-s} \right)^{-1/2} \frac{(RV(X)_{s,t} - CTV(X)_{s,t}) \cdot RV(X)_{s,t}^{-1}}{\sqrt{\frac{\pi^2}{4} + \pi - 5} \max \{ 1, \frac{CTTV(X)_{s,t}}{CTBV(X)_{s,t}} \} } \sim N(0, 1) \]  

(A7)

In our empirical analysis, all tuning parameters of the jump detection statistic are the same as those for the threshold bipower variation. In particular, the sampling frequency \( \Delta_n \) is five minutes.
A.5 Realized semivariances

For the SHAR model we decompose realized volatility into realized positive and negative semivariances, which are devised to measure upside and downside risk respectively. Over the same grid $X_{s+i\Delta_n}, i = 0, 1, \cdots, n$, realized positive and negative semivariances are defined as

$$RSV^+(X)_{s,t} \equiv \sum_{i=1}^{n} r(X)_{s+i\Delta} 1_{(r(X)_{s+i\Delta}>0)}$$

$$RSV^-(X)_{s,t} \equiv \sum_{i=1}^{n} r(X)_{s+i\Delta} 1_{(r(X)_{s+i\Delta}<0)}$$

(A8)

According to Barndorff-Nielsen et al. (2010) the following convergences in probability hold:

$$RSV^+(X)_{s,t} \xrightarrow{p} \frac{1}{2} \int_s^t \|\sigma_{X,u}\|^2 \, du + \sum_{s<u\leq t} (\Delta X_u)^2 1_{(\Delta X_u>0)}$$

$$RSV^-(X)_{s,t} \xrightarrow{p} \frac{1}{2} \int_s^t \|\sigma_{X,u}\|^2 \, du + \sum_{s<u\leq t} (\Delta X_u)^2 1_{(\Delta X_u<0)}$$

We see that the asymptotic limits of realized semivariances differ only in the signed jump variations. The sampling frequency $\Delta_n$ is set to five minutes throughout our empirical analysis.

A.6 Tri-Power quarticity

Finally, the HARQ model requires the tri-power quarticity\(^{18}\) as input. This estimator, provided by Barndorff-Nielsen and Shephard (2006), is given by

$$TPQ(X)_{s,t} = n \mu_{4/3}^{-3} \sum_{i=1}^{n-2} |r(X)_{s+i\Delta} |^{4/3} \sum_{i=1}^{n-2} |r(X)_{s+(i+1)\Delta} |^{4/3} |r(X)_{s+(i+2)\Delta} |^{4/3}$$

\[(A9)\]

where $n \equiv \lfloor \frac{t-s}{\Delta_n} \rfloor$, $\mu_{4/3} \equiv 2^{3/2} \Gamma(7/6)/\Gamma(1/2)$ and $\Gamma(\cdot)$ is the Gamma function. The reason for using the quarticity is that under certain conditions, $RV(X)_{s,t}$ may be regarded as the quadratic variation $|X|_t - |X|_s$ plus $\eta \sim MN \left(0, 2 \int_s^t \|\sigma_{X,r}\|^2 \, dr \right)$, where $MN$ stands for mixed normal distribution. Thus, Bollerslev et al. (2016) argue that quarticity term gauges the measurement error in realized volatility, so it should be included in modelling to attenuate realized volatility. In our empirical analysis, we let $\Delta_n = 5$ minutes when calculating the tri-power quarticity.

\(^{18}\)Bollerslev et al. (2016) in fact consider mainly realized quarticity, but we use tri-power quarticity instead as it remains consistent for integrated quarticity in the presence of jumps. The robustness checks in Bollerslev et al. (2016) show that performance of the HARQ models based on different quarticity estimators are generally close.
Appendix B  HAR Models Using Both SPY and Emini Futures Data

In this appendix we present additional regression results. We consider three more HAR specifications, all aiming to forecast one-day ahead daytime volatility of SPY. The first specification replaces past daily, weekly and monthly realized volatilities of SPY by the corresponding ones of the Emini futures:

\[
RV(S)_{t+1}^D = \alpha_0 + \beta_D RV(F)_t^D + \beta_W RV(F)_t^W + \beta_M RV(F)_t^M + \epsilon_{t+1}
\]  
(B10)

The second specification includes past realized volatilities of both SPY and the Emini futures:

\[
RV(S)_{t+1}^D = \alpha_0 + \alpha_D RV(S)_t^D + \alpha_W RV(S)_t^W + \alpha_M RV(S)_t^M + \beta_D RV(F)_t^D + \beta_W RV(F)_t^W + \beta_M RV(F)_t^M + \epsilon_{t+1}
\]  
(B11)

The third specification further augments the second specification with overnight volatility of the Emini futures.

\[
RV(S)_{t+1}^D = \alpha_0 + \alpha_D RV(S)_t^D + \alpha_W RV(S)_t^W + \alpha_M RV(S)_t^M + \beta_D RV(F)_t^D + \beta_W RV(F)_t^W + \beta_M RV(F)_t^M + \theta_N RV(F)_t^N + \epsilon_{t+1}
\]  
(B12)

In-sample results are listed in panel A of Table 2.11. It is not surprising to find that the specification using past daytime volatilities of Emini futures (HAR (ES only)) and the specification using past daytime volatilities of both SPY and ES of the Emini futures produce almost the same results as the benchmark HAR specification using past daytime volatilities of SPY itself. In fact, this is arguably due to the high correlation of volatilities of a futures contract and its underlying asset, which is explained in Appendix C. On the other hand, the specification using not only past daytime volatilities of SPY and ES, but also the latest overnight volatility of ES, does not produce more accurate forecast than the more parsimonious HAR-F specification that augments the benchmark HAR. This is corroborated by the out-of-sample results listed in panel B of the table. We see that for both estimation schemes, the HAR-F specification outperforms all competitors in terms of all measures, with the only exception that the most complicated specification “HAR-F (SPY+ES)” has the smallest QLIKE loss when the models are estimated with increasing window.

Therefore, for forecasting one-day ahead volatility of SPY, the most critical component is the overnight volatility of the Emini futures. The HAR-F specification which simply augments the benchmark HAR with \(RV(F)_t^N\) produces the most accurate forecast; we do not gain improvement by further including past daytime volatilities of the Emini futures.
### Table 2.11: In-sample and out-of-sample results (one-day ahead forecasting)

<table>
<thead>
<tr>
<th></th>
<th>HAR (ES only)</th>
<th>HAR (SPY+ES)</th>
<th>HAR-F (SPY+ES)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \alpha )</td>
<td>0.0901</td>
<td>0.0563</td>
<td>0.0360</td>
</tr>
<tr>
<td>( \beta_D )</td>
<td>0.2637</td>
<td>0.1692</td>
<td>0.1124</td>
</tr>
<tr>
<td>( \beta_W )</td>
<td>0.4679</td>
<td>0.4131</td>
<td>0.4121</td>
</tr>
<tr>
<td>( \alpha_M )</td>
<td>0.1826</td>
<td>(0.1024)</td>
<td>(0.1022)</td>
</tr>
<tr>
<td>( \beta_M )</td>
<td>0.5685</td>
<td>(0.1367)</td>
<td>(0.1743)</td>
</tr>
<tr>
<td>( \theta_N )</td>
<td>0.1387</td>
<td>0.1414</td>
<td>0.1414</td>
</tr>
<tr>
<td>( \sigma^2 )</td>
<td>0.5680</td>
<td>0.5680</td>
<td>0.5680</td>
</tr>
<tr>
<td>( \text{MSE} )</td>
<td>2.4317</td>
<td>2.4317</td>
<td>2.4317</td>
</tr>
<tr>
<td>( \text{R}^2 )</td>
<td>0.5687</td>
<td>0.5687</td>
<td>0.5687</td>
</tr>
</tbody>
</table>

**Panel A: In-sample**

#### Rolling

<table>
<thead>
<tr>
<th>Model</th>
<th>( \text{R}^2 )</th>
<th>MSE</th>
<th>QLIKE</th>
</tr>
</thead>
<tbody>
<tr>
<td>HAR</td>
<td>0.5032</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>HAR (ES only)</td>
<td>0.5013</td>
<td>1.0096</td>
<td>1.0161</td>
</tr>
<tr>
<td>HAR (SPY+ES)</td>
<td>0.4752</td>
<td>1.0875</td>
<td>1.0013</td>
</tr>
<tr>
<td>HAR-F</td>
<td>0.6487</td>
<td>0.6952</td>
<td>0.7001</td>
</tr>
</tbody>
</table>

**Panel B: Out-of-sample**

<table>
<thead>
<tr>
<th>Model</th>
<th>( \text{R}^2 )</th>
<th>MSE</th>
<th>QLIKE</th>
</tr>
</thead>
<tbody>
<tr>
<td>HAR</td>
<td>0.5032</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>HAR (ES only)</td>
<td>0.5013</td>
<td>1.0096</td>
<td>1.0161</td>
</tr>
<tr>
<td>HAR (SPY+ES)</td>
<td>0.4752</td>
<td>1.0875</td>
<td>1.0013</td>
</tr>
<tr>
<td>HAR-F</td>
<td>0.6487</td>
<td>0.6952</td>
<td>0.7001</td>
</tr>
</tbody>
</table>

**Note:** This table lists in-sample and out-of-sample results of the HAR model and its extensions. The dependent variable is the daytime (realized) volatility of SPY and the forecast is made one day ahead. The benchmark HAR model is specified in (2.1). The models “HAR (ES only), “HAR (SPY+ES)” and “HAR-F (SPY+ES)” are specified by (B10), (B11) and (B12) respectively. The regressions are based on percentage returns. In parentheses are Newey-West estimates of standard errors computed with Bartlett kernel and bandwidth 5.

**Panel B:** In the rolling scheme, the window size is fixed at 800. In the increasing scheme, the estimation window starts with the first 800 observations and keeps expanding. The entries are average ratios of the losses of the models to the corresponding ones of the benchmark HAR model. Entry marked with * means that null hypothesis of the (modified) reality check (see (2.4) and (2.5)) is rejected at the significance level 5%.
Appendix C  The relation Between Volatilities of SPY and ES

The remarkable gain in forecasting daytime volatility of SPY by inclusion of overnight volatility of ES is predicated on the intimate relation between volatilities of SPY and ES, or more generally, between volatilities of a stock and futures contract written on it. In this appendix we briefly discuss such relation. Suppose the availability of a futures contract written on the stock of our interest, whose no-arbitrage logarithmic price and maturity are denoted by $F$ and $T$ respectively. Also suppose that there is a risk-free asset with short rate process denoted by $r$. Assuming existence of the equivalent local martingale measure $Q$, it is known that (see, e.g., Schroder (1999)):

$$\exp(F_t) = E^Q[\exp(S_T)|\mathcal{F}_t], \quad t \leq T \tag{C13}$$

In general, Equation (C13) admits no closed-form solution unless extra assumptions are imposed. The short rate process plays a fundamental role here. We discuss three nested cases below, where we always assume $t \leq T$, that is, we only consider the price processes before expiration of the futures contract.

1. Consider a plain vanilla model, in which $r$ is deterministic (not necessarily constant). It follows that

$$F_t = S_t + \int_t^T r_s \, ds.$$  

The integrated short rate process merely adds a deterministic drift to the stock price. There are several implications of the above equality:

(a) At the “local” level, we have $F^c = S^c$ and $\Delta F = \Delta S$. In other words, not only the continuous local martingale parts but the jump parts of the two assets are indistinguishable.

(b) At the “global” level, we have $\int_0^t ||\sigma_{F,s}||^2 \, ds = \int_0^t ||\sigma_{S,s}||^2 \, ds$ and $\sum_{0<s \leq t} (\Delta F_s)^2 = \sum_{0<s \leq t} (\Delta S_s)^2$ for any $t > 0$. That is, diffusive variations of the two assets are indistinguishable, and the same applies to the jump variations.

In a nutshell, volatilities of the two assets are identical when the short rate process is non-random.

2. What if $r$ is stochastic? We consider a parametric setting, in which

$$S_t = S_0 + \int_0^t b_{S,s} \, ds + \int_0^t \sigma_{S,s} dW^{(1)}_s + \int_0^t \int \delta S(s,z) \mu(ds,dz)$$

This is equivalent to the “no free lunch with vanishing risk” condition in Delbaen and Schachermayer (1994). Under the $Q$ measure, the discounted stock price process $S_t \exp(\int_0^t r_s \, ds)$ is a local martingale. Imposing additional conditions such as boundedness, the discounted stock price process becomes a martingale.
and \( r \) is a CIR process (Cox et al. (1985):

\[
  r_t = r_0 + \int_0^t \kappa (\theta - r_s) ds + \int_0^t \sqrt{r_s} dW_s^{(2)}
\]

where \( W^{(1)} \) and \( W^{(2)} \) are two one-dimensional Wiener processes. To derive a closed-form solution to (C13), we further assume that \( \exp \left\{ \int_0^T r_s ds \right\} \) and \( \exp \left\{ \int_0^T \sigma_{S,s} dW_s^{(1)} \right\} \) are independent conditional on \( \mathcal{F}_t \). Then the following equation can be shown as in Ramaswamy and Sundaresan (1985):

\[
  F_t = S_t + \alpha_t + \beta_t r_t
\]

where \( \alpha \) and \( \beta \) are continuously differentiable deterministic functions. Applying Itô’s formula to \( \beta_t r_t \), we further have

\[
  F_t = (S_0 + \beta_0 r_0) + \left( \alpha_t + \int_0^t (\kappa \beta_s (\theta - r_s) + \beta_s' r_s) ds \right) + \left( \int_0^t \sigma_{S,s} dW_s^{(1)} + \int_0^t \sqrt{r_s} dW_s^{(2)} \right)
\]

\[
  + \int_0^t \int_{\mathbb{R}} \delta_S(s,z) \mu(ds,dz)
\]

where \( \beta_s' \) is the first order derivative of \( \beta_s \). We now see why in Equation (A1) the driving Wiener process \( W \) is assumed to be multi-dimensional. Again, from the above equality we know:

(a) At the “local” level, the jump parts are still indistinguishable, i.e., \( \Delta F = \Delta S \), but the continuous local martingale parts now differ. In particular,

\[
  F^c_t = \int_0^t \sigma_{S,s} dW_s^{(1)} + \int_0^t \sqrt{r_s} dW_s^{(2)} = S_t^c + \int_0^t \sqrt{r_s} dW_s^{(2)}
\]

The difference \( \int_0^t \sqrt{r_s} dW_s^{(2)} \) is seen to be induced by volatility of the short rate process.

(b) At the “global” level, therefore, the jump variations are indistinguishable, i.e.,

\[
  \sum_{0 < s \leq t} (\Delta F_s)^2 = \sum_{0 < s \leq t} (\Delta S_s)^2,
\]

whereas the diffusive variations and the quadratic variations might be different. For instance, if we further assume independence between \( W^{(1)} \) and \( W^{(2)} \), it holds

---

20Conditional independence of the two random variables requires additional assumptions on the instantaneous volatility process \( \sigma_S \) and the correlation between the driving Wiener processes \( W^{(1)} \) and \( W^{(2)} \). Conditional independence might not hold if the dynamics of \( \sigma_S \) is assumed to be a function of the short rate. It is outside the scope of this paper to formulate assumptions that imply desired conditional independence. However, once conditional independence is taken for granted, the futures price can be expressed as \( \exp(F_t) = S_t \mathbb{E}[\exp(\int_0^T r_s ds) | \mathcal{F}_t] \), where the conditional expectation could be evaluated, upon changing sign of the short rate, in the same way as deducting the no-arbitrage price of a zero-coupon bond that has par value 1 and matures at time \( T \).
C. The relation between volatilities of SPY and ES

\[ [F^c]_t = [S^c]_t + \nu^2 \int_0^t \beta_s^2 r_s \, ds \]
\[ [F]_t = [F^c]_t + \sum_{0 < s \leq t} (\Delta F_s)^2 = [S]_t + \nu^2 \int_0^t \beta_s^2 r_s \, ds \]

Thus the stochastic short rate \( r \) might beget a nonnegative difference \( \nu^2 \int_0^t \beta_s^2 r_s \, ds \), which means that the futures exhibit more volatility than its underlying stock.

3. In the most general nonparametric setting, Corollary 3 of Schroder (1999) provides an alternative formula to (C13):

\[ F_t = S_t + \log \left( E^\tilde{Q} \left[ \exp \left( \int_t^T r_s \, ds \right) \bigg| \mathcal{F}_t \right] \right), \quad (C14) \]

where \( \tilde{Q} \) is another equivalent local martingale measure. The second summand in (C14) is a continuous \( P \)-semimartingale. To see this, set \( Z_t \equiv E^\tilde{Q} \left[ \exp \left( \int_t^T r_s \, ds \right) \bigg| \mathcal{F}_t \right] \) and \( V_t \equiv \log(Z_t) \). Observe that

\[ Z_t = \exp \left( - \int_0^t r_s \, ds \right) E^\tilde{Q} \left[ \exp \left( \int_0^T r_s \, ds \right) \bigg| \mathcal{F}_t \right] \]

The continuity of \( Z_t \) follows from continuity of \( \exp(-\int_0^t r_s \, ds) \). By Itô’s formula, \( V_t \) is a continuous \( \tilde{Q} \)-semimartingale, and by Girsanov-Meyer theorem it is a continuous \( P \)-semimartingale as well, whence we draw the following conclusions

(a) At the “local” level, the jump parts are indistinguishable, that is, \( \Delta F = \Delta S \), while the continuous local martingale parts are not. In particular, \( F^c_t = S^c_t + V^c_t \).

(b) At the “global” level, the jump variations are indistinguishable, that is, \( \sum_{0 < s \leq t} (\Delta F_s)^2 = \sum_{0 < s \leq t} (\Delta S_s)^2 \), while diffusive variations and quadratic variations might not be identical. Indeed the quadratic covariation between \( S \) and \( V \) plays a critical role here. The continuity of \( V \) implies that \( [S, V]_t = [S^c, V^c]_t \). We have

\[ [F]_t - [S]_t = [F^c]_t - [S^c]_t = [S, V]_t + [V]_t \]

When, in particular, the quadratic covariation \( [S, V] \) is identically zero, the futures price has higher quadratic variation as well as diffusive variation than the stock price.

We have thus seen that except in the most trivial case in which short rate is nonrandom, volatilities of futures contract and its underlying stock as measured by quadratic variation need not coincide, while the two markets still have equal “tail risk” as measured by jump variation.
C.1 Some empirical results

Given that theoretically quadratic variations of a stock and a futures written on it need not be indistinguishable, it is natural to wonder how different are the quadratic variations. Here we present some results for the SPY-ES pair. The difference $D_t \equiv [F]_t - [S]_t$ is consistently estimated by difference of realized volatilities, $RV(F)_t - RV(S)_t$. On each trading day we compute $RV(F)_t - RV(S)_t$ and further construct two-sided 95% confidence interval for $D_t$, which takes the form

$$ (RV(F)_t - RV(S)_t) \pm 1.96 \sqrt{ \frac{1}{n} \sum_{i=1}^{[t/\Delta_n]} \left( (\Delta_i^n F)^2 - (\Delta_i^n S)^2 \right)^2 - \frac{1}{n-1} (RV(F)_t - RV(S)_t)^2 } \quad (C15) $$

The realized volatilities are all computed with prices sampled every five minutes, a rule-of-thumb frequency that has been documented to mitigate market microstructure noise effectively (see, e.g., Liu et al. (2015)).

Figure 2.9 visualizes the difference $D_t$ and its (daily) confidence intervals year by year. Constrained by space the plot is only from 2005 onwards. The subplot of 2010 seems exotic as on 6th May there occurred the “flash crash” and the difference of realized volatilities on this date ($=-4.5154$) is much higher than on the rest of days in this year. Table 2.12 lists
Table 2.12: Summary statistics of $RV(F)^D_t - RV(S)^D_t$ and 95% confidence interval of $[F]_t - [S]_t$

<table>
<thead>
<tr>
<th></th>
<th>Panel A: $RV(F)^D_t - RV(S)^D_t$</th>
<th>Panel B: 95% confidence interval of $[F]_t - [S]_t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>0.0106</td>
<td>Widest: [-17.7862, 23.0961]</td>
</tr>
<tr>
<td>min</td>
<td>-6.0174</td>
<td>Tightest: [-0.0075, 0.0042]</td>
</tr>
<tr>
<td>Q25</td>
<td>-0.0089</td>
<td>0 $\notin$ CI</td>
</tr>
<tr>
<td>Q50</td>
<td>0.0133</td>
<td></td>
</tr>
<tr>
<td>Q75</td>
<td>0.0383</td>
<td></td>
</tr>
<tr>
<td>max</td>
<td>2.6549</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Note: Panel A reports summary statistics of $RV(F)^D_t - RV(S)^D_t$. “Q” stands for quantile. The first two columns of panel B report the widest and tightest confidence intervals of $</td>
<td>F</td>
</tr>
</tbody>
</table>

We observe that the difference falls by and large around zero, although the extremums deviate a lot. The average difference is 0.01, accounting for merely 0.98% and 0.97% of the average realized volatility of SPY (= 1.08) and that of ES (= 1.09) respectively. The tightest confidence interval is obtained on 26th December 2013. On this date the difference of realized volatilities is -0.0017. The widest confidence interval is obtained on 18th April 2001, which is also the date with largest difference (= 2.65) of realized volatilities. The number of confidence intervals that do not contain 0 is about 11.90%. In other words, if we are to conduct the hypothesis test $H_0: |F|_t = |S|_t$ (with two-sided alternative), then for close to 90% of the dates we cannot reject the null hypothesis of equal quadratic variations. The empirical data thus provides mixed evidence: on most of dates volatilities of SPY and ES as measured by quadratic variations are quite close to each other, while occasionally they deviate far from each other.
References


CHAPTER 3

How Does Diversity of Media Coverage Influence Firms’ Risk Exposure?

Miha Torkar
Jožef Stefan Institute
Ye Zeng
Aarhus University and CREATEs

Abstract

This article studies how diversity of media coverage affects firms’ exposure to systematic risk, as gauged by realized beta. The content diversity of a pool of news stories is defined as their average semantic dissimilarity, where dissimilarity is in turn gauged by the Word Mover’s Distance (Kusner et al. (2015)). To investigate the effect of the diversity measure, we employ the multifactor error structure panel proposed by Chudik and Pesaran (2015). The main finding is that diversity of media coverage positively predicts systematic risk on the next trading day. The effect still prevails after we further control salience, sentiment and social media attention. Robustness checks under various model specifications and with respect to different word embeddings as well as textual dissimilarity measures are also provided.

JEL classification: G12, G14.

Keywords: Beta; Media coverage; Word Mover’s Distance; Word2Vec; Textual similarity.
3.1 Introduction

The media industry would not have been so diverse if news were merely intended to be a mirror image of reality. The media does not mechanically take in the information and parrot them. Rather, they produce news stories that are highly differentiated and vary along many dimensions, such as focus, accuracy, depth and style. A group of stories may focus on the same firm but pay attention to different ongoing events; or the stories may cover the same event in differing depths. The journalists often lack consensus on the consequence of an event and express divided opinions; or they might be unanimous in implication of the event but differ in the styles of conveying their judgements. Insofar as news are deemed economic goods, diversity of their content can be explained by consumer demands and the media’s attempt to cater to the consumers. For discussions on the market forces that shape diversity of media coverage, see Chapter 1 of Hamilton (2004).

By supplying news the media does not merely gratify their readers’ wants; they also exert reverse influence on the readers’ judgements of prevailing affairs. The influence of news diversity on politics has long been acknowledged, and this article calls attention to the influence of news diversity on business performance. Specifically, we study how news diversity affects a firm’s exposure to systematic risk, as gauged by beta. Our research aims to enrich the literature by answering the following questions: (1) does diversity of media coverage exert influence on beta; (2) if yes, what’s the direction of the influence?

To this end, on every trading day we retrieve a sequence of news stories and extract sentences that explicitly mention the firm of interest. For each pair of the news stories we calculate the Word Mover’s Distance (Kusner et al. (2015)), a semantic dissimilarity measure built on the prevalent word2vec embedding (Mikolov et al. (2013); Mikolov et al. (2013a)). Diversity of media coverage is defined as the average dissimilarity of the news stories. On the other hand, to capture the time-varying feature of beta (Ferson et al. (1987); Shanken (1990); Lewellen and Nagel (2006); Andersen et al. (2006); Patton and Verardo (2012)), we draw on the literature of high frequency econometrics and calculate realized beta (Barndorff-Nielsen and Shephard (2004); Andersen et al. (2006); Patton and Verardo (2012)) as the ratio of firm-market covariation to market volatility on a daily basis. Realized betas and diversity mea-

---

1For example, heterogeneous beliefs of consumers can cause the media to slant towards diverse coverage, as by consumption the readers expect the news not only to inform but also to confirm, interpret and persuade (Mullainathan and Shleifer (2005)). Another factor is competition. The common belief is the positive correlation between diversity of news and competition. The Federal Communications Commission once stated that it “has traditionally assumed that there is a positive correlation between viewpoints expressed and ownership of the outlet” (Federal Communications Commission (2003)). Interestingly, in some situations competition is found to diminish diversity, as separate ownership of outlets leads to duplication of news offerings (Steiner (1952); Hamilton (2004)).

2The U.S. Supreme Court claims “One of the most vital of all general interests is the dissemination of news from as many different sources and with as many different facets and colors as is possible. That interest... presupposes that right conclusions are more likely to be gathered out of a multitude of tongues, than through any kind of authoritative selection” (U.S.Supreme Court (1945)).
asures of 75 constituents of S&P 500 sampled on more than 1,000 trading days are cast into the multifactor error structure dynamic panel proposed by Chudik and Pesaran (2015) and we draw inference based on the dynamic common correlated effects estimator. Use of this model is motivated by three considerations. Firstly, it operates in the “large N large T” setting and renders possible the specification of heterogeneous effect of news diversity across firms. Secondly, the model features lagged beta, which is important as the literature of high frequency econometrics have documented persistence of variation and covariation (Andersen et al. (2001a); Andersen et al. (2001b); Andersen et al. (2003)), and realized beta as their ratio is likely to be persistent as well. Thirdly, the model features latent common factors that are present in the error as well as the regressors. In our context, these factors capture unmodeled macroeconomic events that affect risk exposure and media coverage of the firms at the same time. All three considerations suggest adequacy of the model raised by Chudik and Pesaran (2015).

Our empirical analysis proceeds along a specific-to-general route. We begin with a parsimonious specification that involves only first order lags of the diversity measure and variables related to stock market activities. The mean group estimate of diversity turns out to be significant and positive. We then further control characteristics abstracted from textual news data, including sentiment, salience of media coverage and social media attention. News sentiment is measured by the average fraction of positive and negative words per news, a popular measure widely employed in the literature (Tetlock (2007); Tetlock et al. (2008); García (2013); Chen et al. (2014)). Salience of media coverage is defined as the average number of headlines that mention the firm explicitly, and social media attention is gauged by Facebook shares of the news story. Our main conclusion still holds after controlling for these measures. Robustness of the finding is further checked under various sub-sampling methods, model specifications and combinations of word embedding and textual dissimilarity measures. It turns out that robustness of our main conclusion that diversity of media coverage positively predicts firms’ exposure to systematic risk hinges critically on the significance level. At the 10% level, the main conclusion holds irrespective of the sub-sampling method, homogeneity/heterogeneity of the slopes and inclusion of further lags of all variables but the diversity measure itself. But if one feels like running less risk of type I error and adopt the more conservative 5% level, the main conclusion may not be claimed robust, especially when we sub-sample in various ways. Finally, it seems important to build the diversity measure on a word embedding that encodes semantic information and the word mover’s distance. Diversity measures calculated with cosine dissimilarity or Jaccard index (Jaccard (1901)) do not lead to the same conclusion.

Overall, our empirical analysis suggests that diversity of media coverage positively predicts a firm’s risk exposure. We have two interrelated interpretations of the finding. The first interpretation is that diverse media coverage may result from the media’s differing opinion on the firm’s value. The media’s differing opinion might simply reflect the more broad phe-
nomenon that the investors have disagreements on the firm, which in turn cause the firm’s stock to become riskier. The second interpretation leans towards the causal role of the media. For investors, attention is a scarce factor and can be paid to only a few news stories, by which their views and opinions are influenced and shaped. When the media’s coverage of the firm is highly diverse, the investors are likely to be informed about different events and or different facets of the same event, hence forming different judgements of the firm’s value. Again, the differing judgements are the ultimate cause for increased risk exposure. Both interpretations are grounded on investors’ opinions, and thus can be related to previous work that study how differences of opinion affect stock performance. Assuming investors differ in interpretations of public news, Banerjee and Kremer (2010) show that investors’ disagreement tends to heighten volatility. Carlin et al. (2014) also find that increased disagreement among investors is associated with higher volatility. Our research may be viewed as providing new empirical evidence on the positive interaction between differences of opinion and risk exposure.

The remainder is structured as follows. Section 3.2 briefly reviews related literature and summarizes what differentiate our work from them. Section 3.3 gives a concise overview of data, textual analysis techniques and econometric methods employed in this paper. Section 3.4 presents main empirical findings. Section 3.5 performs various robustness checks. Section 3.6 concludes. Being an interdisciplinary study, the paper features long appendices that provide details about the dataset, the language processing techniques and the econometrics method. Appendix A describes data preprocessing and software usage. Appendix B gives a concise review of the word2vec embedding, the cornerstone of our diversity measure. Appendix C explains estimation of the dynamic panel with multifactor error structure. Finally, Appendix D collects various examples and additional details.

3.2 Review of Related Literature

This paper can be subsumed to study of the role of media in finance. Tetlock (2014) and Tetlock (2015) provide excellent reviews on this subject. The paradigm of such study consists of application of textual analysis techniques to media data, such as newspaper articles, social media posts and search queries, to quantify certain characteristics of the media and investigate how they influence the financial market. The emphasis may be put on textual analysis, data source, news characteristic, and media coverage. Study focusing on textual analysis may be represented by Das and Chen (2007), who develop methods for extracting investor sentiment from stock message boards, as well as Loughran and McDonald (2011), who develop a dictionary for sentiment classification. Study stressing data source may be represented by Chen et al. (2014), who find that views expressed on Seeking Alpha predict stock returns and earnings surprises. Notable characteristics of news stories that have been investigated include tone and unusualness. Tonal measures quantify sentiment expressed
in the news stories. In his influential paper, Tetlock (2007) finds that pessimism expressed in The Wall Street Journal predicts downward pressure on market prices followed by a reversion to fundamentals. Tetlock et al. (2008) further find that the fraction of negative words in firm-specific news stories signals low earnings. Other notable studies employing tonal measures include García (2013), Chen et al. (2014), Zhang et al. (2016), Heston and Sinha (2017) and so forth. Recently, Mamaysky and Glasserman (2016) call our attention to another facet of news stories. They construct a measure quantifying unusualness of phrases used by journalists and show that the unusualness measure interacted with negative tonal measure predicts higher market volatility. Predictive power of the unusualness measure is further inspected in Calomiris and Mamaysky (2018).

Our research differs from the abovementioned studies by focusing on the fourth dimension— the media coverage. To be specific, we are interested in content diversity, namely the number of unique perspectives or story selections offered in a marketplace (Hamilton (2004), p. 22). On the other hand, most related studies on the effect of media coverage concentrate on other facets, such as geography, salience, and readership. Engelberg and Parsons (2011) focus on exposure of investors to local press and find that coverage of earnings announcement by local media stimulates trading volume of local markets. Klibanoff et al. (1998) focus on salience as measured by front-page appearance on The New York Times and find that country-specific news affect pricing of closed-end country funds. Barber and Odean (2008) study another facet of salience. They find that individual investors are more likely to buy than sell attention-grabbing stocks, such as the ones mentioned in news. Da et al. (2011) find that increase in Google Search Volume Index predicts higher stock prices in the next 2 weeks and reversion within the year. Fang and Peress (2009) define “media coverage” as the number of news mentioning a firm and find that stocks with no media coverage earn higher returns than stocks with high coverage. Our work differs from the cited studies in one important point: we delve into the news stories and quantify the content diversity by analyzing semantics of the texts.

This paper is also closely related to Tetlock (2011). Tetlock studies whether investors react to stale information. He defines staleness of a news story as its (average) textual similarity to the ten previous news, where similarity between two news is defined as their Jaccard index (Jaccard (1901)). Conceptually, the flip side of Tetlock’s staleness measure (e.g., 1−staleness) resembles our diversity measure, but we point out some important differences between his and our work. First, the dependent variable of our interest is a firm’s risk exposure, while Tetlock focuses on return and volume. Second, we do not distinguish between “old” and “new” stories after forming the pool of news stories, as Tetlock does. Instead, we compare each pair of the stories in the pool to determine the (average) difference in their contents, and our diversity measure aims to gauge the breadth as opposed to novelty of information in the news stories. Third, the technicalities underpinning his staleness measure and our diversity measure are very different. The Jaccard index is defined as the number
of unique words in the intersection of two texts divided by the number of unique words in their union. It is based on matching of word forms, i.e., orthographic appearances of the words, and does not take into account word senses, i.e., meanings of the words. Ignoring word senses, the Jaccard index is likely to understate\(^3\) textual similarity, and if we were to gauge diversity by, say, \(1 - \text{Jaccard}(\text{news}_1, \text{news}_2)\), the diversity measure would be upward biased. By contrast, we employ the Word Mover's Distance (WMD) introduced by Kusner et al. (2015) in the computing the news diversity measure. WMD builds on the prevailing word2vec embedding (Mikolov et al. (2013); Mikolov et al. (2013a)), which encodes semantic as well as syntactic relations between word senses. In this light, our diversity measure may better capture various narratives of the journalists. These differences should suffice to differentiate our and Tetlock's work from each other.

Finally, our work is also related to what Loughran and McDonald (2016) summarize under the rubrics “thematic structure in documents” and “measuring document similarity”. Hoberg and Phillips (2016) employ the bag-of-words representation and the cosine similarity measure to classify firms according to their 10-K product descriptions. Lang and Stice-Lawrence (2015) resort to similar approach in comparison of annual report of a firm with its peers. Huang et al. (2017) use the Latent Dirichlet Allocation (Blei et al. (2003)) to compare analyst reports issued right after earnings conference calls with the content of the calls themselves. The textual analysis techniques employed by the cited papers are very different from ours, and so are the focuses of their study.

3.3 Data, Textual Analysis Techniques and Econometrics Method

3.3.1 Stock and News data

We consider 75 constituents of S&P 500. Tickers along with references used to extract news stories about the firms are listed in Table A3. We download transaction prices and volumes (i.e., shares transacted) of these firms from NYSE's Transactions And Quotes (TAQ) database. We use in addition data of the ETF SPDR, also known as SPY, which closely tracks the S&P 500 index. The data spans from January 2nd 2014 through May 31st 2018, covering 1,612 calendar days and 1,110 trading days respectively. The raw data are cleaned following the procedures advocated by Barndorff-Nielsen et al. (2009), except that we do not implement their rule T4 for removing “outliers”. For ease of reference the cleaning procedures are

\(^3\)Consider the sentences \(S_0 = \text{“AIG informs shareholders about court decision”}\) and \(S_1 = \text{“AIG tells investors about the Judge’s ruling”}\) given in Appendix D.1. Excluding the function words “about” and the “the”, the two sentences have nine unique words in total, but the only common word is the subject “AIG”. The Jaccard index between \(S_0\) and \(S_1\) is \(J(S_0, S_1) = \frac{1}{9}\). Given that \(S_0\) and \(S_1\) are actually talking about the same event, the Jaccard index does not provide suggestive evidence on the semantic similarity between them.
described in Appendix A.1. The textual news data are provided by EventRegistry and cover the same period as the stock data. A detailed description of retrieval of the news articles is provided in Appendix A.2.

### 3.3.2 Measure of systematic risk exposure

The classic CAPM assumes an invariant beta. In the past a few decades, the invariance assumption were questioned by many researchers. Huang and Litzenberger (1988), for example, say that "it is unlikely that risk premiums and betas on individual assets are stationary over time" (p. 303). Lewellen and Nagel (2006) use short-window regressions to directly estimate betas and find them to vary considerably over time. Patton and Verardo (2012) provide evidence that betas increase on earnings announcement days and revert to their average levels afterwards.

**Definition of beta**

The time-varying feature of beta motivates us to draw on the literature of high-frequency econometrics (e.g., Barndorff-Nielsen and Shephard (2004) and Andersen et al. (2006)) and calculate beta with intraday prices. Let \((X_1, X_2, \ldots, X_N)\)' be a vector process of arbitrage-free log-prices of \(N + 1\) individual stocks, and let \(\tilde{X}\) denote the arbitrage-free log-price of the market portfolio. For our empirical analysis, the ETF SPDR is taken to be \(\tilde{X}\). We assume that dynamics of the price processes are stipulated by the following stochastic differential equations:

\[
X_j(t) = X_j(t-1) + \int_{t-1}^{t} \mu_j(s) ds + \int_{t-1}^{t} \beta_j(s) \tilde{\sigma}(s) d\tilde{W}(s) + \int_{t-1}^{t} \sigma_j(s) dW_j(s),
\]

\[
\tilde{X}(t) = \tilde{X}(t-1) + \int_{t-1}^{t} \tilde{\mu}(s) ds + \int_{t-1}^{t} \tilde{\sigma}(s) d\tilde{W}(s).
\]

where \(\mu_j\) and \(\tilde{\mu}\) are drifts of the returns, \(\tilde{\sigma}\) the instantaneous volatility of the market portfolio, \(\sigma_j\) the idiosyncratic instantaneous volatility of stock-\(j\), and \((W_1, W_2, \ldots, W_N, \tilde{W})\)' an \((N + 1)\)-dimensional standard Wiener process. The process \(\beta_j\) may be interpreted as the stock's instantaneous exposure to systematic risk. The quadratic variation of the market portfolio over the period \([t - 1, t]\), an informative measure of volatility (Andersen et al. (2001a); Andersen et al. (2001b)), is defined by

\[
[X, \tilde{X}]_{t-1, t} = \int_{t-1}^{t} \tilde{\sigma}^2(s) ds.
\]

On the other hand, the quadratic covariation between the stock and the market portfolio is

\[
[X_{j-1}, \tilde{X}]_{t-1, t} = \int_{t-1}^{t} \beta_j(s) \tilde{\sigma}^2(s) ds.
\]
3.3. Data, Textual Analysis Techniques and Econometrics Method

A measure of the stock’s exposure to systematic risk, which we call integrated beta, is defined as the ratio of quadratic covariation between the stock and the market portfolio to the quadratic variation of the market portfolio itself:

$$I\beta_{j,t} \equiv \frac{[X_{j-1}, \tilde{X}]_{t-1,t}}{[X, \tilde{X}]_{t-1,t}} = \frac{\int_{t-1}^{t} \beta_j(s) \tilde{\sigma}^2(s) ds}{\int_{t-1}^{t} \tilde{\sigma}^2(s) ds}.$$  (3.4)

Obviously, when the instantaneous risk exposure $\beta_j$ is a constant, i.e., $\beta_j(s) \equiv \beta$, the integrated beta becomes constant as well.

**Estimation of beta**

We could have used $I\beta_{j,t}$ as the measure of systematic risk exposure if it were observable. But it is not even ex post, and we have to substitute it by a reliable approximate. The approximate can be obtained by replacing both $[\tilde{X}, \tilde{X}]_{t-1,t}$ and $[X_{j-1}, \tilde{X}]_{t-1,t}$ with their respective estimates built on intra-day transaction prices. Trivial estimators such as realized variance and realized covariance (see, e.g., Andersen et al. (2003) and Barndorff-Nielsen and Shephard (2004)) could be used if the transaction prices were the same as the no-arbitrage prices. But the literature (Andersen et al. (2000); Bandi and Russell (2008)) suggest that the transaction price may not be modeled directly as the price in an arbitrage-free framework, but more plausibly as no-arbitrage price plus a noise component induced by the market microstructure\(^4\). Denoting by $X_j^*$ and $\tilde{X}^*$ the transaction prices of the stock and the market portfolio respectively, we assume that they comply with the following data generating processes:

$$X_j^*(t) \equiv X_j(t) + \epsilon_j(t),$$

$$\tilde{X}^*(t) \equiv \tilde{X}(t) + \tilde{\epsilon}(t).$$  (3.5)

where $\epsilon_j(t)$ and $\tilde{\epsilon}(t)$ capture the market microstructure noises. Realized (co)variance no longer consistently estimates the quadratic (co)variation in the presence of market microstructure noise, so we must use more sophisticated noise-robust estimators. There are several noise-robust estimators, and in the empirical analysis we use the multivariate realized kernel proposed by Barndorff-Nielsen et al. (2011)\(^5\). The estimator first uses the so-called refresh-time sampling to synchronize the transaction prices; the kernel estimator is then computed with a formula resembling the one for the celebrated HAC estimator (Newey and West (1987); Andrews (1991)). See pp. 150-151 in Barndorff-Nielsen et al. (2011) for details.

We let $RK(X_j^*, \tilde{X}^*)_{t-1,t}$ denote the bivariate realized kernel, a 2-by-2 matrix-valued estimator computed with transaction prices $X_j^*$ and $\tilde{X}^*$ sampled during $[t - 1, t]$. The asymp-

---

\(^4\)See the extensive study by Hansen and Lunde (2006).

\(^5\)Other noise-robust estimators include the pre-averaging estimator proposed by Christensen et al. (2010) and the quasi maximum likelihood estimator by Shephard and Xiu (2017).
totic theory in Barndorff-Nielsen et al. (2011) suggests the convergence in probability

$$\text{RK}(X^*_j, \tilde{X}^*_j)_{t-1,t} \overset{p}{\to} \left[ \int_{t-1}^t \beta^2_j(s) \sigma^2(s) ds + \int_{t-1}^t \sigma^2_j(s) ds \right].$$

(3.6)

Define the realized beta of stock- \( j \) on day \( t \) by

$$R\beta_j, t \equiv \frac{\text{RK}(X^*_j, \tilde{X}^*_j)_{1,2}^{1,2}}{\text{RK}(X^*_j, \tilde{X}^*_j)_{2,2}^{1,2}}.$$

(3.7)

where \( \text{RK}(X^*_j, \tilde{X}^*_j)_{1,2}^{1,2} \) and \( \text{RK}(X^*_j, \tilde{X}^*_j)_{2,2}^{2,2} \) are the (1,2)-entry and the (2,2)-entry of the multivariate realized kernel respectively. Then \( R\beta_j, t \) consistently estimates the integrated beta defined in (3.4), and we use it throughout the paper as a measure of systematic risk exposure.

**Interpretation of beta**

Our framework is consistent with the ones adopted by Bollerslev and Zhang (2003), Andersen et al. (2005), Andersen et al. (2006), Todorov and Bollerslev (2010) and Patton and Verardo (2012). In the cited papers, the ratio of covariation between a stock and the market portfolio to the variation of market portfolio itself is interpreted as exposure to systematic risk. There are, however, some nuances. First, the market portfolio used in these papers are different. Bollerslev and Zhang (2003) consider about 6400 stocks per year, and the market portfolio is constructed as the logarithmic transform of the value-weighted percentage returns across all stocks. Similar construction of portfolio is done by Andersen et al. (2005). On the other hand, Todorov and Bollerslev (2010) and Patton and Verardo (2012) use the S&P 500 index as the measure of the aggregate market, with the corresponding high-frequency returns constructed from the prices for the ETF SPDR. In our work, we follow Todorov and Bollerslev (2010) and Patton and Verardo (2012) in treating SPDR as the market portfolio. Second, in (3.1) we have abstracted away potential jumps. Jump risk is certainly of interest to practitioners, given the evidence that option traders price expected variation in equity returns associated with jumps differently from the expected variation associated with diffusive price moves (Pan (2002); Todorov (2009)). The setting of Todorov and Bollerslev (2010) allows explicitly for jumps, and they design an estimator which consistently estimates the jump beta\(^6\) in the absence of market microstructure noise. Here we make two remarks. First, the multivariate realized kernel still consistently estimates the quadratic covariation in the simultaneous presence of jump as well as market microstructure noise. But when there is jump, the quadratic covariation can be decomposed into diffusive covariation and jump covariation. If one’s interest lies primarily in the jump risk, a different estimator that enables

\(^6\)The notion “jump beta” is also featured in Li et al. (2017a) and Li et al. (2017b), but in a slightly different technical setting.
us to isolate jump covariation from diffusive covariation, such as the one proposed by Jacod et al. (2017), should be used. Second, in this paper we do not seek to isolate jump risk, but focus on systematic risk on the whole. The reason is that the approaches to estimating jump beta, such as the one proposed by Todorov and Bollerslev (2010), only permits estimation on days when there are actually jump in the market portfolio, while the econometrics model we adopt predicts one-day ahead risk exposure on a daily basis. We shall therefore proceed without discussing potential presence of jumps in the sequel.

### 3.3.3 Diversity of media coverage

Our measure of the diversity of media coverage aims to gauge content diversity, namely breadth of the information in the news stories. The definition is motivated by Hamilton (2004) (p. 22), who defines “diversity” as the number of unique perspectives or story selections offered in a marketplace. To automate the calculation of diversity measure we borrow some tools from natural language processing. Ideally, the tools should be able to capture semantic relations\(^7\), i.e., relations between word senses, because what a journalist chooses in crafting a story is ultimately word sense rather than word form. Methods based on matching of word forms are suboptimal, since they neglect potential synonymy and are prone to overstate dissimilarity of two texts.

Our diversity measure builds on the Word Mover’s Distance (WMD), a textual dissimilarity measure proposed by Kusner et al. (2015). The cornerstone of WMD is word2vec (Mikolov et al. (2013); Mikolov et al. (2013a)), which has two features. First, the studies by Mikolov et al. (2013) and Mikolov et al. (2013a) suggest that word embeddings produced by word2vec preserve semantic and syntactic relations to a certain extent. However, as Jurafsky and Martin (2018) remark, the reason as to why the word embeddings encode these relations remain largely unknown. Second, word2vec represents words as dense low-dimensional vectors. The raw input of word2vec are one-hot encoded vectors (i.e., vectors with a single one and zeros elsewhere) representing words in the given vocabulary. The length of each one-hot encoded word is the same as the size of the vocabulary, which is often very large. On the contrary, the word2vec embeddings are typically vectors in low-dimensional space without many zeros in its coordinates. For example, Google’s word2vec embeddings are trained with a vocabulary of 3 million words, where the embeddings lie in the 300 × 1 Euclidean space. Representing words as dense low-dimensional vectors, word2vec substantially reduces the computational complexity of many NLP tasks.

Taking advantage of word2vec embeddings, Kusner et al. (2015) show that WMD leads to higher accuracy in many document classification experiments. Though word2vec is an important ingredient, to introduce WMD we may simply assume that each word from a vo-

---

\(^7\)Major semantic relations include synonymy (i.e., having the same or nearly the same meaning but different spellings), antonymy (i.e., having opposite meanings), hypernymy and hyponymy (i.e., the relation between a generic term and its specific instance).
cabulary has already been embedded into a vector space without knowing details of the embedding algorithm. Hence, to keep it straightforward we shall only review WMD here. The review of word2vec is relegated to Appendix B.

**Word Mover's Distance**

Our presentation draws on Section 4 of Kusner et al. (2015). Assume the availability of a vocabulary $V$, which ideally should include most commonly used words. The vocabulary delimits the words that are machine-readable; words not in the vocabulary (e.g., function words) will not enter the calculation of WMD. We take for granted an embedding $e : V \rightarrow \mathbb{R}^d$. Each word $w \in V$ is represented as a $d$-dimensional real vector $e(w)$. The semantic dissimilarity between two words $w_1$ and $w_2$ is gauged by the Euclidean distance between their embeddings:

$$c(w_1, w_2) \equiv \|e(w_1) - e(w_2)\|_2.$$  

We use the notation $c$ in the above equation, as the semantic dissimilarity $c(w_1, w_2)$ may be viewed as the “cost” of transporting $w_1$ to $w_2$, and vice versa. This helps to formulate the computation of semantic dissimilarity between two texts as an optimal transport problem.

Consider two normalized texts $t \equiv (w_1, w_2, \ldots, w_m)$ and $\tilde{t} \equiv (\tilde{w}_1, \tilde{w}_2, \ldots, \tilde{w}_n)$. Here normalization means retaining only unique words that are contained in the vocabulary. For word $w_j$ we let $f_j$ denote its frequency in the raw text and define

$$p_j \equiv \frac{f_j}{\sum_{i=1}^{m} f_i}, \quad 1 \leq j \leq m,$$

which is the probability mass of $w_j$ in the first text. For the second text we define $\tilde{p}_k, 1 \leq k \leq n$ in the same vein. A transport plan from $t$ to $\tilde{t}$ is a mapping $T : \{w_1, w_2, \ldots, w_m\} \times \{\tilde{w}_1, \tilde{w}_2, \ldots, \tilde{w}_n\} \rightarrow [0, \infty)$ that satisfies the constraints

$$\sum_{k=1}^{n} T(w_j, \tilde{w}_k) = p_j \quad \text{and} \quad \sum_{j=1}^{m} T(w_j, \tilde{w}_k) = \tilde{p}_k. \quad (3.8)$$

$T(w_j, \tilde{w}_k)$ may be thought of as the fraction of the word $w_j$ transported to vacancies of the word $\tilde{w}_k$. Intuitively, the first constraint may be interpreted as that distribution of $w_j$ among words in $\tilde{t}$ sums up to the total “shares” of $w_j$ in $t$; the second constraint may be understood as that all vacancies of $\tilde{w}_k$ in $\tilde{t}$ are occupied by words from $t$. The total cost of the transport plan $T$ is

$$\sum_{j=1}^{m} \sum_{k=1}^{n} T(w_j, \tilde{w}_k) c(w_j, \tilde{w}_k).$$

The word mover’s distance (henceforth WMD) is defined as the minimal cost for transporting
words from \( t \) to \( \tilde{t} \), i.e.,

\[
WMD(t, \tilde{t}) \equiv \min \left\{ \sum_{j=1}^{m} \sum_{k=1}^{n} T(w_j, \tilde{w}_k) c(w_j, \tilde{w}_k) : T \text{ satisfies (3.8)} \right\}, \tag{3.9}
\]

The WMD is a special case of the Earth Mover’s Distance (EMD) that is widely used in image retrieval (Rubner et al. (2000)). It is metric, as it is the same as first Mallows distance between the multinomial distributions \( \{p_j\} \) and \( \{\tilde{p}_k\} \) (Levina and Bickel (2001)). A concrete example of WMD is given in Appendix D.1.

It should be noted that by construction the WMD \textit{per se} can not capture semantic dissimilarity between the texts. In fact, the novelty of WMD consists in its formulating the computation of textual dissimilarity as an optimal transport problem. The remarkable accuracy of document classification reported in Kusner et al. (2015) is arguably due to the word2vec embedding which WMD capitalizes on. The word2vec is a group of algorithms employing shallow neural networks to produce word embeddings that encode semantic and syntactic information; it is such information that render WMD adequate for gauging news diversity. For its importance, we provide a concise review of word2vec in Appendix B.

The diversity measure

We illustrate computation of the diversity measure for a specific firm \( i \) unless otherwise stated. A news story is considered relevant if it mentions firm \( i \) at least once in the main body. On the trading day \( t \) we retrieve the most recent 10 relevant news stories published before 09:30 New York local time, the moment NYSE opens for trading. For each pair of the news stories, say News\(_i\) and News\(_j\) (\( i \neq j \)), we compute their semantic distance as follows. We extract all sentences that have explicit reference to the firm. References of the firms are listed in Table A3 in the appendix. Sentences extracted from News\(_i\) (resp. News\(_j\)) are collected in the set \( S(\text{News}_i) \) (resp. \( S(\text{News}_j) \)). The semantic distance between News\(_i\) and News\(_j\) is defined as

\[
SD(\text{News}_i, \text{News}_j) \equiv \frac{1}{|S(\text{News}_i)||S(\text{News}_j)|} \sum_{s \in S(\text{News}_i), \tilde{s} \in S(\text{News}_j)} WMD(s, \tilde{s}),
\]

namely the average semantic distance between sentences from the two news. Table A1 and Table A2 in the appendix give examples of news with short and long semantic distance respectively. The diversity measure is defined as

\[
diversity_{i,t} \equiv \frac{1}{45} \sum_{i=1}^{9} \sum_{j=i+1}^{10} SD(\text{News}_i, \text{News}_j), \tag{3.10}
\]

that is, \( \text{diversity}_{i,t} \) is the average semantic distance between the news stories.

Construction of the running pool of news stories merits more discussions. A common
practice in the literature is to fix a time interval and compute the textual measure with all news released during this interval. For example, on each day one can sample all news published around the clock and compute the diversity measure with them. This method is not ideal for our analysis. On the one hand, we have limited the number of news sources to 28; on the other hand, some firms catch much less attention than the others. For both reasons, some less attention-catching firms have no or few news on many trading days, and it does not seem plausible to gauge news diversity if there is no or only one news. The method we adopt ensures that the diversity measure is computed with sufficient news (= 10 in our analysis) on every trading day. If on day $t$ there is no news about the firm, news published on previous days are used to calculate the diversity measure. A consequence of our approach is that the diversity measure might be stale on some days, hence inducing a positive serial correlation. The “staleness” of our news diversity measure may be gauged by the autocorrelation coefficients; for several firms, the first autocorrelation coefficient is about 0.8. The high-degree of staleness may adversely enlarge standard errors and hence lower the probability of detecting the effect of news diversity, if we include multiple lags of the diversity measure. This problem will be analyzed when we conduct various robustness checks later.

We also note that working at the sentence level has obvious pros and cons. On the positive side, by confining ourselves to sentences that explicitly mention the firm, the extracted information are sure to be relevant. The diversity measure would be heavily contaminated if we consider the whole document as long as it mentions the firm. On the negative side, our approach may incur considerable loss of information. Very often the surrounding context of the extracted sentence does not mention the firm explicitly, but still contains relevant information. But it is an elusive task to make the computer capable of determining informativeness of the surrounding context, and we do not want to make the language processing procedures too complicated. As Loughran and McDonald (2016) advocate (p. 1192), applications of textual analysis to finance should be based on straightforward characteristics of the data so that little “econometric exorcism” gets involved. Adhering to this principle, we decide to work at the sentence level to keep the textual analysis transparent and the diversity measure least noisy.

---

8To illustrate, consider the case in which firm $i$ is covered by the media on day $t-1$ but not on day $t$. Clearly, our algorithm entails $\text{diversity}_{i,t-1} = \text{diversity}_{i,t}$, as the pool of news remains the same on day $t$.

9For example, the daily briefing by Bloomberg on January 5, 2018 mentions that “Amazon’s digital assistant is coming with you”. But the briefing does not focus on Amazon and covers other stories about Apple, Los Angeles, employment data and so forth. Working on the document level will inevitably involve these irrelevant stories and render the diversity measure very noisy.

10In the aforementioned briefing, what succeeds the extracted sentence is “The first Alexa-enabled digital glasses are here. The augmented-reality eyewear, to be unveiled at CES, the consumer electronics show, were made by tiny Vuzix Corp”. These unextracted sentences certainly contain valuable information to a human reader.
3.3.4 Multifactor Error Structure Model

We present specification of the econometric model here and illustrate its estimation in Appendix C. We employ the dynamic panel featuring heterogeneous slopes and multifactor error structure proposed by Chudik and Pesaran (2015) to study the relation between news diversity and beta. The model is specified by

\[ R\beta_{i,t} = c_i + \phi_i R\beta_{i,t-1} + \theta_i \text{diversity}_{i,t-1} + \delta'_i x_{i,t-1} + u_{i,t}, \]  

(3.11)

\[ u_{i,t} = \gamma'_i f_t + \epsilon_{i,t}, \]  

(3.12)

\[ \omega_{i,t} = \begin{bmatrix} \text{diversity}_{i,t} \\ x_{i,t} \end{bmatrix} = c_{w,i} + \alpha_i R\beta_{i,t-1} + \Gamma'_i f_t + v_{i,t}, \]  

(3.13)

for firms \( i = 1, 2, \ldots, N \) and periods \( t = 1, 2, \ldots, T \). Here \( c_i \) and \( c_{w,i} \) are fixed effects of firm \( i \), \( x_{i,t} \) a \( k_x \times 1 \) vector of firm-specific control variables, \( \epsilon_{i,t} \) the idiosyncratic errors, \( f_t \) a \( m \times 1 \) vector of unobserved factors common to all firms (\( m \) unknown as well), \( \Gamma_i \) a \( m \times k \) matrix of factor loadings, \( \alpha_i \) a \((k_x + 1) \times 1\) vector of unknown coefficients, and \( v_{i,t} \) idiosyncratic errors for diversity \( i_t \) and \( x_{i,t} \). The coefficients are all assumed to randomly fluctuate around the same constant mean, and the idiosyncratic errors \( \epsilon_{i,t} \) are allowed to be spatially dependent; see Assumption 1 and Assumption 4 in Chudik and Pesaran (2015). Our interest lies mainly in the coefficient \( \theta_i \) and its group mean \( \theta \equiv E[\theta_i] \).

There are several reasons for choosing this model. First of all, our dataset covers 75 firms and more than 1,000 trading days, so it is reasonable to operate in a “large \( N \) large \( T \)” setting where both \( N \) and \( T \) grow to infinity in asymptotic analysis, instead of the classic “large \( N \) small \( T \)” setting\textsuperscript{11}. The “large \( N \) large \( T \)” setting renders possible modeling with and estimation of heterogeneous slopes, i.e., \( \pi_i \equiv (\phi_i, \theta_i, \delta'_i)' \), which would otherwise be incidental and non-estimatable in the “small \( T \)” setting. The assumption of heterogeneous slopes seems more sensible than the assumption of homogeneous slopes, as the effect of media coverage is more likely to be firm-specific rather than uniform across all firms. The second reason is the inclusion of the lagged dependent variable. The heterogeneous panel studied by Pesaran (2006) has similar multifactor error structure, but it does not feature lagged dependent variable. Inclusion of lagged beta seems crucial to our analysis. By definition, realized beta is the ratio of the covariation of the firm and the market to the market’s own variation. Quite a few studies have documented persistence in (co)variation (Andersen et al. (2001b); Andersen et al. (2001a); Andersen et al. (2003)), and it is natural to expect that such persistence will carry over to the realized beta\textsuperscript{12}. This is why we consider (3.11) a plausible specification.

\textsuperscript{11}The recent textbook by Pesaran (2015) provides introduction to analysis under both settings.

\textsuperscript{12}Andersen et al. (2006) argue that realized beta is less persistent than realized covariance and realized volatility. Their argument is based on quarterly beta computed with daily returns, while our beta is calculated on each day with intra-day high-frequency returns. In the high-frequency setting, asynchronous trading and market microstructure noise become worrisome and render the theoretical treatment of beta very different from theirs. Therefore, in our framework persistence of beta cannot be ruled out \textit{a priori}.
Also note that the feedback effect of beta on news diversity and other control variables, such as salience and social media attention, is explicitly into the model through (3.13). Explicit modeling of the feedback effect is important. The journalists do not fabricate stories with no reference to reality, but craft news to communicate with the reader what they have experienced and what they expect to happen in the market. In this light, their writings are subject to the influence of firms’ risk exposure. Another appealing feature of the model is spatial dependence of the idiosyncratic errors $\varepsilon_{i,t}$ in (3.12). Allowing spatial dependence renders the model more realistic, since investors do not value firms as if they were independent of each other and the media does not cover a firm as if its businesses and prospects had no bearing on other firms. The last and most important reason is the presence of the common latent effect $f_t$. $f_t$ captures non-firm-specific factors that affect performance and media coverage of all firms simultaneously, which include the macroeconomic environment, political climate, social aggregate time preference and risk aversion etc. All the mentioned factors are not directly modeled and hence may be thought of being built into $f_t$.\[13\]

### 3.4 Main Empirical Findings

#### 3.4.1 Control variables and summary statistics

The first set of regression controls include variables related to firm-\(i\)'s stock market activities, such as open-to-close percentage return $r_{i,t}$; one-sided percentage return $r_{i,t}^-$; log-arithmetic trading volumes $\text{vol}_{i,t-1}$; the log of Amihud (2002)'s illiquidity measure, which is defined as

$$\text{illiq}_{i,t} = \log\left(10^6 \times \frac{|r_{i,t}|}{\text{dollar volume}_{i,t}}\right).$$

The second set of controls include firm-specific news variables, such as word count $\text{word}_{i,t}$, salience $\text{sal}_{i,t}$, relevance $\text{rel}_{i,t}$, positive sentiment $\text{pos}_{i,t}$, negative sentiment $\text{neg}_{i,t}$ and social media attention $\text{atten}_{i,t}$. $\text{word}_{i,t}$ is the log of average words per news story. $\text{sal}_{i,t}$ is the average number of news that mention the firm explicitly in the headline. $\text{rel}_{i,t}$ is the average relevance of the latest 10 news stories to firm-\(i\), where relevance of each news story is defined as

$$\text{rel}_{i,t} = \frac{\text{Frequency of mention}}{\text{Word count of the news}}.$$
The tonal measures build on the Loughran-McDonald sentiment word lists (Loughran and McDonald (2011)), which categorize words frequently used by financial press into Negative, Positive, Uncertainty, Litigious, Modal and Constraining. For each news story, positive sentiment measure is defined as the fraction of words with positive connotation, and \( \text{pos}_{i,t} \) is the average positive sentiment per news story. \( \text{neg}_{i,t} \) is defined in the same way using the list of negative words. Finally, for many news stories EventRegistry keeps track of its number of shares on Facebook. The social media attention \( \text{atten}_{i,t} \) is defined as
\[
\text{atten}_{i,t} = \log(1 + \text{facebook shares}_{i,t}).
\] (3.14)

It must be emphasized that the dataset is incomplete. For some news stories EventRegistry provides no data for facebook shares, which creates missing observations and hence renders the panel data unbalanced. EventRegistry also keeps track of shares on other social media, such as twitter, linkedin, google plus and so forth. These alternatives, especially twitter, seem to be more prominent in the literature (Bartov et al. (2017); Gu and Kurov (2018)). However, data for the alternative media are even more incomplete. For a few firms we even have no news whose shares on twitter is tracked by EventRegistry. Limited by the data availability, we only use Facebook shares in empirical analysis. The use of Facebook may nonetheless cast some new light on the role of social media in financial market.

Table 3.1: Summary statistics of all variables

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>sd</th>
<th>min</th>
<th>p25</th>
<th>p75</th>
<th>max</th>
<th>skew</th>
<th>kurt</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{diversity}_{i,t} )</td>
<td>1.003</td>
<td>0.096</td>
<td>0.000</td>
<td>0.964</td>
<td>1.064</td>
<td>1.230</td>
<td>−1.837</td>
<td>10.307</td>
</tr>
<tr>
<td>( R\beta_{i,t} )</td>
<td>0.966</td>
<td>0.426</td>
<td>−2.406</td>
<td>0.710</td>
<td>1.190</td>
<td>6.497</td>
<td>0.688</td>
<td>6.381</td>
</tr>
<tr>
<td>( \text{vol}_{i,t} )</td>
<td>13.911</td>
<td>1.010</td>
<td>9.028</td>
<td>13.260</td>
<td>14.558</td>
<td>17.920</td>
<td>−0.092</td>
<td>3.769</td>
</tr>
<tr>
<td>( \text{illiq}_{i,t} )</td>
<td>−5.103</td>
<td>1.202</td>
<td>−12.970</td>
<td>−5.789</td>
<td>−4.249</td>
<td>−1.289</td>
<td>−0.743</td>
<td>3.792</td>
</tr>
<tr>
<td>( r_{i,t} )</td>
<td>0.011</td>
<td>1.126</td>
<td>−10.995</td>
<td>−0.558</td>
<td>0.611</td>
<td>12.111</td>
<td>−0.207</td>
<td>7.641</td>
</tr>
<tr>
<td>( r_{i,t}^- )</td>
<td>−0.397</td>
<td>0.702</td>
<td>−10.995</td>
<td>−0.558</td>
<td>0.000</td>
<td>0.000</td>
<td>−3.159</td>
<td>19.296</td>
</tr>
<tr>
<td>( \text{word}_{i,t} )</td>
<td>0.060</td>
<td>0.003</td>
<td>0.045</td>
<td>0.058</td>
<td>0.062</td>
<td>0.084</td>
<td>0.390</td>
<td>4.355</td>
</tr>
<tr>
<td>( \text{rel}_{i,t} )</td>
<td>0.004</td>
<td>0.005</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.055</td>
<td>1.613</td>
<td>6.913</td>
</tr>
<tr>
<td>( \text{pos}_{i,t} )</td>
<td>0.015</td>
<td>0.005</td>
<td>0.000</td>
<td>0.012</td>
<td>0.018</td>
<td>0.050</td>
<td>0.510</td>
<td>3.598</td>
</tr>
<tr>
<td>( \text{neg}_{i,t} )</td>
<td>0.025</td>
<td>0.010</td>
<td>0.001</td>
<td>0.018</td>
<td>0.030</td>
<td>0.141</td>
<td>1.137</td>
<td>6.523</td>
</tr>
<tr>
<td>( \text{sal}_{i,t} )</td>
<td>0.107</td>
<td>0.143</td>
<td>0.000</td>
<td>0.000</td>
<td>0.200</td>
<td>1.000</td>
<td>1.730</td>
<td>6.349</td>
</tr>
<tr>
<td>( \text{atten}_{i,t} )</td>
<td>0.066</td>
<td>0.017</td>
<td>0.007</td>
<td>0.054</td>
<td>0.077</td>
<td>0.137</td>
<td>0.205</td>
<td>3.080</td>
</tr>
</tbody>
</table>

Note: The table presents summary statistics of the variables, including diversity of media coverage (\( \text{diversity}_{i,t} \)), realized beta (\( R\beta_{i,t} \)), logarithmic trading volume (\( \text{vol}_{i,t} \)), the log of Amihud (2002)’s illiquidity measure (\( \text{illiq}_{i,t} \)), percentage open-to-close return (\( r_{i,t} \)), percentage one-sided return (\( r_{i,t}^- \)), the log of average words per news story (\( \text{word}_{i,t} \)), relevance (\( \text{rel}_{i,t} \)), the average positive sentiment (\( \text{pos}_{i,t} \)), the average negative sentiment (\( \text{neg}_{i,t} \)), salience (\( \text{sal}_{i,t} \)), and social media attention (\( \text{atten}_{i,t} \)).

\(^{14}\text{A word of caution: after capturing a news story, EventRegistry keeps track of its social media shares for up to five days. Very often the vast majority of social media shares are created within the first 24 hours after the news released. Because we use lagged social media attention \( \text{atten}_{i,t−1} \) as regressor, we might to a certain extent be free of look-ahead bias. Certainly, we cannot completely rule out the endogeneity problem induced by the social media attention measure. Hence the regression results with social media attention involved should be interpreted with extra caution.}\)
Chapter 3. How Does Diversity of Media Coverage Influence Firms’ Risk Exposure?

Table 3.1 presents summary statistics of all variables used in our analysis. The measure of diversity of media coverage ranges from 0 to 1.23. The mean level is about 1.003, and the 25th and 75th percentiles are 0.964 and 1.064 respectively. Its distribution is left-skewed and strikingly leptokurtic, with the kurtosis being greater than 10. Distribution of beta is also characterized by excessive kurtosis (> 6). The average beta is slightly below one. The return \( r_{i,t} \) has the stylistic left-skewed and leptokurtic distribution. As of the tonal measures, the news stories in our sample seem to lean towards pessimism, as mean level and skewness of \( \text{neg}_{i,t} \) are both larger than those of \( \text{pos}_{i,t} \). Like the negative tonal measure, the salience measure is right-skewed and has larger-than-normal kurtosis. Its mean level is about 10.7%, which means that on average one out of ten news stories mentions the firm in headline. Finally, the social media attention measure is slightly skewed to the right and has the smallest kurtosis.

Table 3.2: Refresh rates of news

<table>
<thead>
<tr>
<th>Ticker</th>
<th>Ref.%</th>
<th>Ticker</th>
<th>Ref.%</th>
<th>Ticker</th>
<th>Ref.%</th>
<th>Ticker</th>
<th>Ref.%</th>
<th>Ticker</th>
<th>Ref.%</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAPL</td>
<td>98.83</td>
<td>CELG</td>
<td>55.79</td>
<td>GE</td>
<td>96.28</td>
<td>MCD</td>
<td>88.34</td>
<td>PM</td>
<td>40.07</td>
</tr>
<tr>
<td>ABBV</td>
<td>64.00</td>
<td>CL</td>
<td>36.18</td>
<td>GILD</td>
<td>64.95</td>
<td>MDLZ</td>
<td>62.02</td>
<td>QCOM</td>
<td>46.98</td>
</tr>
<tr>
<td>ABT</td>
<td>58.14</td>
<td>COP</td>
<td>64.68</td>
<td>GM</td>
<td>97.92</td>
<td>MDT</td>
<td>57.79</td>
<td>SBUX</td>
<td>96.75</td>
</tr>
<tr>
<td>ACN</td>
<td>81.69</td>
<td>COST</td>
<td>50.23</td>
<td>GS</td>
<td>97.65</td>
<td>MET</td>
<td>60.89</td>
<td>SLB</td>
<td>57.69</td>
</tr>
<tr>
<td>AIG</td>
<td>49.77</td>
<td>CSCO</td>
<td>93.65</td>
<td>HD</td>
<td>89.26</td>
<td>MO</td>
<td>45.35</td>
<td>T</td>
<td>97.02</td>
</tr>
<tr>
<td>AMGN</td>
<td>68.17</td>
<td>CVX</td>
<td>90.09</td>
<td>HON</td>
<td>69.41</td>
<td>MON</td>
<td>77.80</td>
<td>TWX</td>
<td>93.73</td>
</tr>
<tr>
<td>AMZN</td>
<td>97.56</td>
<td>DHR</td>
<td>19.53</td>
<td>HPQ</td>
<td>90.18</td>
<td>MRK</td>
<td>63.42</td>
<td>UNH</td>
<td>61.78</td>
</tr>
<tr>
<td>AXP</td>
<td>90.36</td>
<td>DIS</td>
<td>72.52</td>
<td>IBM</td>
<td>97.73</td>
<td>MS</td>
<td>98.19</td>
<td>UNP</td>
<td>40.59</td>
</tr>
<tr>
<td>BA</td>
<td>97.92</td>
<td>DUK</td>
<td>48.94</td>
<td>INTC</td>
<td>95.36</td>
<td>MSFT</td>
<td>98.56</td>
<td>UPS</td>
<td>89.42</td>
</tr>
<tr>
<td>BAC</td>
<td>96.11</td>
<td>EBAY</td>
<td>97.01</td>
<td>JNJ</td>
<td>88.54</td>
<td>NKE</td>
<td>96.92</td>
<td>USB</td>
<td>28.05</td>
</tr>
<tr>
<td>BIIB</td>
<td>50.69</td>
<td>EMR</td>
<td>22.78</td>
<td>JPM</td>
<td>96.75</td>
<td>ORCL</td>
<td>89.79</td>
<td>UTX</td>
<td>61.13</td>
</tr>
<tr>
<td>BLK</td>
<td>88.26</td>
<td>ESRX</td>
<td>44.98</td>
<td>KO</td>
<td>95.47</td>
<td>OXY</td>
<td>32.76</td>
<td>VZ</td>
<td>96.11</td>
</tr>
<tr>
<td>BMY</td>
<td>65.39</td>
<td>F</td>
<td>98.10</td>
<td>LLY</td>
<td>66.09</td>
<td>PEP</td>
<td>87.63</td>
<td>WFC</td>
<td>97.10</td>
</tr>
<tr>
<td>C</td>
<td>95.56</td>
<td>FB</td>
<td>98.83</td>
<td>LMT</td>
<td>90.82</td>
<td>PFE</td>
<td>93.17</td>
<td>WMT</td>
<td>92.57</td>
</tr>
<tr>
<td>CAT</td>
<td>63.59</td>
<td>FOXA</td>
<td>84.41</td>
<td>MA</td>
<td>88.08</td>
<td>PG</td>
<td>87.52</td>
<td>XOM</td>
<td>80.80</td>
</tr>
</tbody>
</table>

Note: The table presents refresh rates (Ref.%). For each firm, the refresh rate is defined as the percentage of trading days with news from the sources listed in Table A4 mentioning the firm.

As mentioned earlier, the running pool of news used to calculate the diversity measure is updated on a day only if there are news mentioning the firm published on this day. Dated news would be used in calculation if there are no news on the day, causing the diversity measure to be stale. Staleness of the news diversity measure is the opposite pole of the news refresh rate, which we define as the percentage of trading days with news. Table 3.2 lists news refresh rates for all firms. Not surprisingly, the media’s attention to the firms is quite heterogeneous: the refresh rate goes from as high as 97% (e.g., AAPL, AMZN, MSFT) to as low as 22% (e.g., Emerson Electric co. (EMR)). Low refresh rates tend to induce positive serial correlation. In Figure 3.1 we constrast time series and autocorrelation coefficients of Apple Inc. (ticker: AAPL) and Emerson Electric co. (ticker: EMR). Grabbing very little attention of the media, the news diversity measure of Emerson Electric exhibits much more
staleness. Its autocorrelation function decays linearly from a level very close to one. By contrast, Apple's diversity measure exhibits more fluctuations, and its first order autocorrelation is less than 0.15. One potential consequence of the staleness is enlarged standard error and hence widened confidence interval when multiple lags of realized beta gets included in the regression. This problem will be discussed later when we reach the robustness checks.

**Figure 3.1:** Time series of the news diversity

3.4.2 Predicting firms’ risk exposure with diversity of media coverage

We test the hypothesis that diversity of media coverage predicts firms’ systematic risk exposure, as gauged by realized beta. In the first step, we include only lagged stock market control variables, such as the log-volume $\text{vol}_{i,t-1}$, the illiquidity measure $\text{illiq}_{i,t-1}$, open-to-close percentage return $r_{i,t-1}$ and one-sided percentage return $r^+_{i,t-1}$. Inclusion of the one-side return is inspired by the **leverage effect**, i.e., the phenomenon that beta responds asymmetrically to positive and negative returns (Braun et al. (1995)).

Table 3.3 displays the regression results, where we report the mean group estimates and their standard errors. The first column corresponds to the regression that drops the diversity measure from (3.11). The mean squared errors (MSE) of this model will be used as benchmark and we calculate the ratios of other specifications’ to it. The second column corresponds to the specification that drops lagged realized beta. This specification is the same as the panel distributed lag model with multifactor error structure proposed by Pesaran (2006).
The MSE becomes substantially (about 26.3%) higher if we exclude lagged beta, which reflects the importance of including it in the model. The third column corresponds to the specification (3.11) with the controls being stock market variables. The mean group estimate of the diversity measure is found to be positive and significant at the 5% level. It suggests that more diverse media coverage predicts higher exposure to systematic risk on the following day. Other variables, except log-volume, are found to predict beta as well. The lagged beta is significant at the 1% level and its mean group estimate is about 0.287, suggesting that a unit increase in beta predicts 28.7% higher beta the day after, ceteris paribus. The controls illiq, r, t−1, and r−, t−1 are all significant at the 1% level. In particular, mean group estimate of the one-sided negative return r−, t is significantly negative, which echoes the finding of Braun et al. (1995), namely that beta responds more to negative returns than to positive returns.

Figure 3.2: Firm-specific effects of the diversity of media coverage

Note: The bar plots show \( \hat{\theta}_i \), namely estimate of the slope of news diversity (see (3.11)), of each firm in Table 3.2. The x-axes correspond to firms and the y-axes to the estimates. The bars are colored according to the sectors to which the firms belong. The lower-right panel summarizes the sectors and their colors. The sectorization is based on the Global Industry Classification Standard (GICS), an industry taxonomy developed by MCSI and Standard & Poor.

We can inspect the effect of diversity of media coverage at the firm level, since the augmented OLS \( \hat{\theta}_i \) delivers consistent estimate of \( \theta_i \) in the “large N, large T” setting (see Appendix C). Figure 3.2 shows \( \hat{\theta}_i \) of each firm. We sort the estimates in ascending order and divide them into three sub-figures. The x-axes correspond to firms and the y-axes to \( \hat{\theta}_i \). We also classify\(^\text{15}\) the firms and color the bars accordingly; the five classes can be found

\(^{15}\)The classification is adapted from the Global Industry Classification Standard (GICS), an industry taxonomy
### Table 3.3: The influence of diversity of media coverage

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>diversity$_{i,t-1}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R\beta_{i,t-1}$</td>
<td>0.241***</td>
<td></td>
<td>0.239***</td>
<td>0.234***</td>
<td>0.234***</td>
<td>0.232***</td>
</tr>
<tr>
<td></td>
<td>(0.012)</td>
<td></td>
<td>(0.012)</td>
<td>(0.012)</td>
<td>(0.012)</td>
<td>(0.012)</td>
</tr>
<tr>
<td>vol$_{i,t-1}$</td>
<td></td>
<td>0.010</td>
<td>0.001</td>
<td>0.002</td>
<td>0.002</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td>(0.007)</td>
<td></td>
<td>(0.007)</td>
<td>(0.007)</td>
<td>(0.007)</td>
<td>(0.007)</td>
</tr>
<tr>
<td>illiq$_{i,t-1}$</td>
<td>-0.005***</td>
<td>-0.002</td>
<td>-0.004***</td>
<td>-0.004**</td>
<td>-0.004**</td>
<td>-0.004**</td>
</tr>
<tr>
<td></td>
<td>(0.002)</td>
<td></td>
<td>(0.002)</td>
<td>(0.002)</td>
<td>(0.002)</td>
<td>(0.002)</td>
</tr>
<tr>
<td>$r_{i,t-1}$</td>
<td>0.027***</td>
<td>0.030***</td>
<td>0.027***</td>
<td>0.027***</td>
<td>0.027***</td>
<td>0.027***</td>
</tr>
<tr>
<td></td>
<td>(0.004)</td>
<td></td>
<td>(0.004)</td>
<td>(0.004)</td>
<td>(0.004)</td>
<td>(0.004)</td>
</tr>
<tr>
<td>$r_{i,t-1}^{-}$</td>
<td>-0.069***</td>
<td>-0.071***</td>
<td>-0.069***</td>
<td>-0.068***</td>
<td>-0.068***</td>
<td>-0.068****</td>
</tr>
<tr>
<td></td>
<td>(0.007)</td>
<td></td>
<td>(0.007)</td>
<td>(0.007)</td>
<td>(0.007)</td>
<td>(0.007)</td>
</tr>
<tr>
<td>word$_{i,t-1}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.322</td>
<td></td>
<td>0.154</td>
<td>0.076</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rel$_{i,t-1}$</td>
<td></td>
<td>-0.508</td>
<td>-0.225</td>
<td>-0.225</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.415)</td>
<td></td>
<td>(0.525)</td>
<td>(0.514)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>neg$_{i,t-1}$</td>
<td></td>
<td>-0.159</td>
<td>-0.128</td>
<td>-0.038</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.341)</td>
<td></td>
<td>(0.339)</td>
<td>(0.342)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pos$_{i,t-1}$</td>
<td></td>
<td>0.015</td>
<td>0.048</td>
<td>-0.023</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.223)</td>
<td></td>
<td>(0.230)</td>
<td>(0.216)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sal$_{i,t-1}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.034**</td>
<td>-0.033**</td>
<td></td>
<td>(0.014)</td>
<td>(0.015)</td>
<td></td>
</tr>
<tr>
<td>atten$_{i,t-1}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.204*</td>
<td></td>
<td>(0.104)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSE Ratio</td>
<td>1.000</td>
<td>1.263</td>
<td>0.998</td>
<td>0.992</td>
<td>0.991</td>
<td>0.989</td>
</tr>
<tr>
<td>Obs.</td>
<td>72448</td>
<td>72448</td>
<td>72448</td>
<td>72448</td>
<td>72448</td>
<td>72383</td>
</tr>
</tbody>
</table>

Note: This table displays results of the dynamic panel regression of $R\beta_{i,t}$ on lagged diversity of media coverage (diversity$_{i,t-1}$) and control variables. The model is specified by (3.11), (3.12) and (3.13). The table reports mean group estimates, and in parentheses are their standard errors. Asterisks indicate significance levels (* $p < 0.1$, ** $p < 0.05$, *** $p < 0.001$). The second last row lists ratio of mean squared errors (MSE) of each specification to that of the first specification. The last row gives number of observations used in the regression. The second to the fifth specifications have fewer observations as some firms start to have enough ($\geq 10$) news stories for calculating the diversity measure later than the first day of the sample. The last specification involves even fewer observations, since for some news stories EventRegistry does not have record of Facebook shares.

in the lower-right panel. For 29 out of 75 firms news diversity is found to predict lower risk exposure, whereas for the rest it predicts rise of beta. Overall, the distribution of $\hat{\theta}_i$ is right skewed, so the mean group estimate is positive. The lowest two $\hat{\theta}_i$’s are those of Mondeléze International (ticker: MDLZ), a company manufactures confectionery and beverages, and EBAY (ticker: EBAY), an E-commerce company that facilitates consumer-to-consumer (C2C) and business-to-consumer (B2C) sales. For EBAY, 1% increase of the diversity measure predicts -0.35% lower beta on the coming day. The highest two $\hat{\theta}_i$’s are those of U.S. Bankcorp (ticker: USB), a bank holding company, and Biogen Inc. (ticker: BIIB), a biotech-developed by MSCI and Standard & Poor for use by the global financial community. GICS has 11 sectors in total; in addition to the ones shown in Figure 3.2 there is the sector “Real Estate”. In Figure 3.2 we further group the sectors lest the colors gets too overwhelming.
nology company specializing in development and delivery of therapies for the treatment of neurodegenerative, hematologic, and autoimmune diseases. For Biogen, an 1% increase in diversity of media coverage predicts 0.58% higher beta on the coming day. Turning to sectorization of the firms, we do not recognize any pattern in the figures, and in each of the panels we find multiple firms from all sectors. Perhaps a firm’s principal business activities have no bearing on the media’s causal role in influencing its risk exposure.

3.4.3 Controlling tonal measures

Study of tonal measures is a burgeoning field in application of textual analysis to finance, but so far most work focus on how tonal measures are related to abnormal return or turnover (Tetlock (2007); Tetlock et al. (2008); Chen et al. (2014)). We naturally wonder if the tonal measures also influence firm’s risk exposure and, if yes, whether the diversity measure still contains incremental information for predicting it.

Regression results after further controlling the (log of) average word count (word\(_{i,t}\)) per news, relevance (rel\(_{i,t}\)) and two tonal measures pos\(_{i,t}\) and neg\(_{i,t}\) are listed in the fourth column of Table 3.3. Inclusion of the tonal measures makes little difference to our main conclusion. The mean group estimate of diversity\(_{i,t−1}\) remains significant at the 5% level, and it is the same as the mean group estimate when we include only stock market variables. Mean group estimates of the stock market variables are also by and large the same. On the other hand, none of the additional four textual measures is found to predict beta. It should be reiterated that the tonal measures, unlike the diversity measure, are calculated by counting sentiment words in the whole news story rather than the sentences explicitly mentioning the firm. This is a common practice in the literature (see, e.g., Chen et al. (2014)) and we simply follow it. But built in this way the tonal measures are arguably cruder than the diversity measure, and we should not rush to the conclusion that tonal measures carry no information for predicting risk exposure. Since the paper’s focus is on news diversity, we leave a refined analysis of the relation between tonal measures and risk exposure to future study.

3.4.4 Controlling salience

Two intuitive reasons arouse our interest in salience of media coverage\(^{16}\), as gauged by the average number of headlines mentioning the firm. First, the headline is a scarce means employed by the media to disperse information that they believe to be most attention-grabbing. Promoting a specific firm to the headline hence reflects the media’s judgement

\(^{16}\) The effect of salient media coverage has been recurrently studied in the literature. See, e.g., Klibanoff et al. (1998), Barber and Odean (2008), Fang and Peress (2009) and Lawrence et al. (2018). It is also worth mentioning the study of Huberman and Regev (2001), which analyzes how salient media coverage of a nonevent about EntreMed, a biotechnology company that once announced breakthrough in cancer-curing medicine, causes its price to skyrocket several months after the announcement.
of its informational value, and the judgement per se might contain information on the firm’s risk exposure. Second, scarcity of the investors’ attention entails their different processing of information about and hence different treatments of firms with and without headline coverage, and the different treatments might in turn translate into significant influence of salient media coverage on risk exposure.

If salient media coverage does influence firms’ exposure to systematic risk, we naturally wonder if the news diversity measure provides additional information after controlling salience. Regression results are displayed in the fifth column of Table 3.3. Our main conclusion is not affected by inclusion of the salience measure. The mean group estimate of diversity$_i,t−1$ is about 0.04 and still significant at the 5% level. Estimates of the other four crude textual measures have changed drastically, though they are still insignificant. By contrast, estimates of the stock market variables remain the same. It is interesting that the mean group estimate of sal$_i,t−1$ is significant at the 5% level. It suggests that salient media coverage negatively influences a firm’s risk exposure: 1% more headline coverage renders beta 0.039% lower on the next day. A possible explanation, as we indicate above, is that investors discriminate between firms with and without salient coverage. If the investors get informed through headlines about positive change in a firm’s fundamental value, they might sell shares of some firms with no headline coverage and divert the capital to the firm being headlined. The supply and demand force price of the firm with headline coverage to rise and prices of the firms’ without headline coverage to fall, and the opposite movements translate into the decrease in beta.

### 3.4.5 Controlling social media attention

Investors nowadays rely increasingly on social media in gleaning and disseminating information. Compared to traditional media, social media is almost barrier-free, highly interactive and highly participatory. These advantages lead to the immense popularity of social media among investors. Social media has also aroused academic interest. A growing body of literature is harnessing social media data to predict the stock market. Chen et al. (2014) find that investors’ opinions expressed in both articles and commentaries on Seeking Alpha predict future stock returns and earnings surprises. Bartov et al. (2017) find that aggregate opinion from individual tweets help to predict firms’ quarterly earnings and announcement returns. Gu and Kurov (2018) find that twitter sentiment is useful for forecasting stock returns at the firm-level. Interestingly, all cited studies ascribe the predictive information gleaned from social media to the “wisdom of crowds”. Another strand of the literature studies how firms capitalize on social media to facilitate information dissemination among investors. Blankespoor et al. (2013), for example, argue that firms reduce information asymmetry, as gauged by the depth of bid-ask spreads, by sending links to press releases via twitter. The mentioned studies all focus on abnormal returns or bid-ask spreads, and many of them re-
late stock market variables to sentiment measures. We pose a different question: we ask if social media attention influences firms’ exposure to systematic risk, and, if yes, whether our main conclusion is robust to inclusion of the attention measure.

We take advantage of EventRegistry’s tracking system and compute the social media attention measure $\text{atten}_{i,t}$ (see (3.14)). The regression results are shown in the last column of Table 3.3. Again, our main conclusion remains valid at the 5% significance level: the mean-group estimate of the diversity measure ($=0.038$) is still positive and does not differ from estimates under other specifications. Estimates of stock market variables are also qualitatively the same. As of textual measures, the two crude measures word count and relevance and the two tonal measures remain insignificant, while mean group estimate of the salience measure is still significant. The attention measure is significant only at the 10% level. It is interesting that the mean group estimates of $\text{sal}_{i,t-1}$ and $\text{atten}_{i,t-1}$ have opposite signs, which suggests that the traditional media and the social media users play non-overlapped roles in influencing firms’ risk exposure.

### 3.5 Robustness of the Findings and Further Analysis

#### 3.5.1 Subsampling the firms

We begin the robustness check by sub-sampling the firms. We consider four sub-samplings in total:

1. Randomly sub-sample 50 firms.
2. Randomly sub-sample 50 firms with different seed.
3. Use data from March 18, 2016 to May 31, 2018, namely second half of the dataset.
4. Sub-sample the top 50 firms with most frequent news refresh rates (see Table 3.2).

The regression results are listed in Table 3.4, where the columns are ordered according to the sub-sampling methods. It turns out that for the sub-sampling methods (1), (2) and (4), we cannot claim robustness of the main conclusion unless we take on the more aggressive significance level 10%. On the other hand, mean group estimates of $R\beta_{i,t-1}$, $r_{i,t-1}$ and $r_{i,t-1}^-$ remain highly significant. When we sub-sample the second half of the data, mean group estimate ($=0.103$) of the diversity measure is significant at the 5% level, but it is substantially larger than the estimates in other cases. In the meanwhile, mean group estimate of $R\beta_{i,t-1}$ is substantially smaller than the estimates in other cases.

#### 3.5.2 Other model specifications

Specification of the model (3.11) can be modified in at least two dimensions. First, we could constrain the slopes $\pi_i \equiv (\phi_i, \theta_i, \delta_i')'$ to be homogeneous. Models with homogeneous
3.5. **ROBUSTNESS OF THE FINDINGS AND FURTHER ANALYSIS**

Table 3.4: The influence of diversity of media coverage: sub-samples

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>diversity&lt;sub&gt;i,t−1&lt;/sub&gt;</td>
<td>0.048*</td>
<td>0.042*</td>
<td>0.103**</td>
<td>0.045*</td>
</tr>
<tr>
<td></td>
<td>(0.027)</td>
<td>(0.024)</td>
<td>(0.042)</td>
<td>(0.023)</td>
</tr>
<tr>
<td>Rβ&lt;sub&gt;i,t−1&lt;/sub&gt;</td>
<td>0.230***</td>
<td>0.241***</td>
<td>0.156***</td>
<td>0.266***</td>
</tr>
<tr>
<td></td>
<td>(0.016)</td>
<td>(0.015)</td>
<td>(0.014)</td>
<td>(0.017)</td>
</tr>
<tr>
<td>vol&lt;sub&gt;i,t−1&lt;/sub&gt;</td>
<td>0.011</td>
<td>0.014*</td>
<td>0.009</td>
<td>0.019**</td>
</tr>
<tr>
<td></td>
<td>(0.008)</td>
<td>(0.008)</td>
<td>(0.009)</td>
<td>(0.008)</td>
</tr>
<tr>
<td>illid&lt;sub&gt;i,t−1&lt;/sub&gt;</td>
<td>−0.004*</td>
<td>−0.003</td>
<td>−0.010***</td>
<td>−0.004**</td>
</tr>
<tr>
<td></td>
<td>(0.002)</td>
<td>(0.002)</td>
<td>(0.003)</td>
<td>(0.002)</td>
</tr>
<tr>
<td>R&lt;sub&gt;i,t−1&lt;/sub&gt;</td>
<td>0.021***</td>
<td>0.025***</td>
<td>0.035***</td>
<td>0.026***</td>
</tr>
<tr>
<td></td>
<td>(0.005)</td>
<td>(0.005)</td>
<td>(0.007)</td>
<td>(0.005)</td>
</tr>
<tr>
<td>vol&lt;sub&gt;−1&lt;/sub&gt;</td>
<td>−0.059***</td>
<td>−0.063***</td>
<td>−0.084***</td>
<td>−0.067***</td>
</tr>
<tr>
<td></td>
<td>(0.008)</td>
<td>(0.008)</td>
<td>(0.013)</td>
<td>(0.009)</td>
</tr>
<tr>
<td>word&lt;sub&gt;i,t−1&lt;/sub&gt;</td>
<td>0.684</td>
<td>0.223</td>
<td>1.147</td>
<td>−0.019</td>
</tr>
<tr>
<td></td>
<td>(0.727)</td>
<td>(0.782)</td>
<td>(1.001)</td>
<td>(0.563)</td>
</tr>
<tr>
<td>rel&lt;sub&gt;i,t−1&lt;/sub&gt;</td>
<td>−0.463</td>
<td>−0.604</td>
<td>−0.748</td>
<td>−0.551</td>
</tr>
<tr>
<td></td>
<td>(0.639)</td>
<td>(0.613)</td>
<td>(0.787)</td>
<td>(0.527)</td>
</tr>
<tr>
<td>neg&lt;sub&gt;i,t−1&lt;/sub&gt;</td>
<td>−0.274</td>
<td>0.041</td>
<td>−0.194</td>
<td>−0.266</td>
</tr>
<tr>
<td></td>
<td>(0.495)</td>
<td>(0.514)</td>
<td>(0.721)</td>
<td>(0.466)</td>
</tr>
<tr>
<td>pos&lt;sub&gt;i,t−1&lt;/sub&gt;</td>
<td>0.005</td>
<td>−0.257</td>
<td>0.224</td>
<td>0.102</td>
</tr>
<tr>
<td></td>
<td>(0.215)</td>
<td>(0.249)</td>
<td>(0.548)</td>
<td>(0.230)</td>
</tr>
<tr>
<td>sal&lt;sub&gt;i,t−1&lt;/sub&gt;</td>
<td>−0.002</td>
<td>−0.031*</td>
<td>−0.045</td>
<td>−0.011</td>
</tr>
<tr>
<td></td>
<td>(0.015)</td>
<td>(0.016)</td>
<td>(0.033)</td>
<td>(0.015)</td>
</tr>
<tr>
<td>atten&lt;sub&gt;i,t−1&lt;/sub&gt;</td>
<td>0.356**</td>
<td>0.313**</td>
<td>0.324**</td>
<td>0.245*</td>
</tr>
<tr>
<td></td>
<td>(0.152)</td>
<td>(0.140)</td>
<td>(0.161)</td>
<td>(0.138)</td>
</tr>
<tr>
<td>Obs.</td>
<td>48662</td>
<td>48565</td>
<td>35267</td>
<td>48919</td>
</tr>
</tbody>
</table>

**Note:** The model is still specified by (3.11), (3.12) and (3.13), but the regressions are done on sub-samples: (1) Randomly sub-sample 50 firms. (2) Randomly sub-sample 50 firms with a different seed. (3) Use data from March 18, 2016 to May 31, 2018. (4) Sub-sample the top 50 firms with most frequent news refresh rates (see Table 3.2). The table reports mean group estimates, and in parentheses are their standard errors. Asterisks indicate significance levels (* p < 0.1, ** p < 0.05, *** p < 0.001).

Slopes are quite popular in the literature of textual analysis in finance. To be aligned with the literature, we consider specifications with homogeneous slopes and draw inference with the pooled mean group estimator. Second, Equation (3.11) includes only first order lags as in the original paper of Chudik and Pesaran (2015), but we could add more lags without invalidating asymptotic properties of the mean group estimator reviewed in Appendix C. We consider below specifications which either impose the homogeneity constraint or add lags, and specifications with homogeneous slopes as well as more lagged regressors.

The regression results are listed in Table 3.5. The note below the table summarizes the corresponding specifications of all columns. The first column displays pooled mean group estimates of the specification that imposes the constraint of homogeneous slopes, i.e., $\boldsymbol{\pi}_t \equiv \boldsymbol{\pi}$. The pooled mean group estimate of diversity<sub>i,t−1</sub> does not differ much from the mean group estimate (cf. Table 3.3), but it is significant only at the 10% level. The second

---

Chapter 3. How Does Diversity of Media Coverage Influence Firms’ Risk Exposure?

Table 3.5: The influence of diversity of media coverage: various model specifications

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>diversity(i, t-1)</td>
<td>0.039*</td>
<td>0.021</td>
<td>0.038**</td>
<td>0.037**</td>
<td>0.033*</td>
</tr>
<tr>
<td></td>
<td>(0.023)</td>
<td>(0.031)</td>
<td>(0.018)</td>
<td>(0.018)</td>
<td>(0.017)</td>
</tr>
<tr>
<td>(R\beta_{i, t-1})</td>
<td>0.261***</td>
<td>0.159***</td>
<td>0.159***</td>
<td>0.175***</td>
<td>0.175***</td>
</tr>
<tr>
<td></td>
<td>(0.016)</td>
<td>(0.008)</td>
<td>(0.008)</td>
<td>(0.010)</td>
<td>(0.010)</td>
</tr>
<tr>
<td>vol(i, t-1)</td>
<td>0.004</td>
<td>0.012*</td>
<td>0.012**</td>
<td>0.004</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td>(0.006)</td>
<td>(0.006)</td>
<td>(0.006)</td>
<td>(0.006)</td>
<td>(0.006)</td>
</tr>
<tr>
<td>illiq(i, t-1)</td>
<td>-0.005**</td>
<td>-0.005***</td>
<td>-0.005**</td>
<td>-0.006***</td>
<td>-0.006***</td>
</tr>
<tr>
<td></td>
<td>(0.002)</td>
<td>(0.002)</td>
<td>(0.002)</td>
<td>(0.002)</td>
<td>(0.002)</td>
</tr>
<tr>
<td>(r_{i, t-1})</td>
<td>0.026***</td>
<td>0.025***</td>
<td>0.024***</td>
<td>0.025***</td>
<td>0.025***</td>
</tr>
<tr>
<td></td>
<td>(0.005)</td>
<td>(0.004)</td>
<td>(0.004)</td>
<td>(0.005)</td>
<td>(0.005)</td>
</tr>
<tr>
<td>(r_{i, t-1})</td>
<td>-0.068***</td>
<td>-0.066***</td>
<td>-0.065***</td>
<td>-0.067***</td>
<td>-0.067***</td>
</tr>
<tr>
<td></td>
<td>(0.009)</td>
<td>(0.007)</td>
<td>(0.007)</td>
<td>(0.009)</td>
<td>(0.009)</td>
</tr>
<tr>
<td>word(i, t-1)</td>
<td>0.090</td>
<td>-0.651</td>
<td>-0.560</td>
<td>-0.015</td>
<td>-0.012</td>
</tr>
<tr>
<td></td>
<td>(0.668)</td>
<td>(0.855)</td>
<td>(0.843)</td>
<td>(0.490)</td>
<td>(0.489)</td>
</tr>
<tr>
<td>rel(i, t-1)</td>
<td>0.100</td>
<td>-0.502</td>
<td>-0.423</td>
<td>0.065</td>
<td>0.054</td>
</tr>
<tr>
<td></td>
<td>(0.666)</td>
<td>(0.758)</td>
<td>(0.709)</td>
<td>(0.593)</td>
<td>(0.591)</td>
</tr>
<tr>
<td>neg(i, t-1)</td>
<td>-0.141</td>
<td>-0.691</td>
<td>-0.634</td>
<td>-0.359</td>
<td>-0.354</td>
</tr>
<tr>
<td></td>
<td>(0.342)</td>
<td>(0.533)</td>
<td>(0.536)</td>
<td>(0.352)</td>
<td>(0.352)</td>
</tr>
<tr>
<td>pos(i, t-1)</td>
<td>0.072</td>
<td>-0.228</td>
<td>-0.253</td>
<td>-0.125</td>
<td>-0.126</td>
</tr>
<tr>
<td></td>
<td>(0.185)</td>
<td>(0.361)</td>
<td>(0.365)</td>
<td>(0.192)</td>
<td>(0.193)</td>
</tr>
<tr>
<td>sal(i, t-1)</td>
<td>-0.026*</td>
<td>-0.003</td>
<td>-0.006</td>
<td>-0.006</td>
<td>-0.006</td>
</tr>
<tr>
<td></td>
<td>(0.014)</td>
<td>(0.024)</td>
<td>(0.023)</td>
<td>(0.015)</td>
<td>(0.015)</td>
</tr>
<tr>
<td>atten(i, t-1)</td>
<td>0.183</td>
<td>0.643***</td>
<td>0.653***</td>
<td>0.278*</td>
<td>0.278*</td>
</tr>
<tr>
<td></td>
<td>(0.119)</td>
<td>(0.171)</td>
<td>(0.173)</td>
<td>(0.146)</td>
<td>(0.147)</td>
</tr>
<tr>
<td>Obs.</td>
<td>72383</td>
<td>69776</td>
<td>69776</td>
<td>69776</td>
<td>69776</td>
</tr>
<tr>
<td>(R^2)</td>
<td>0.08</td>
<td>0.21</td>
<td>0.20</td>
<td>0.13</td>
<td>0.13</td>
</tr>
</tbody>
</table>

Note: The table displays regression results of the following specifications:

1. The slopes \(\pi_i \equiv (\phi_i, \theta_i, \delta_i')^{'}\) in (3.11) are assumed to be homogeneous, that is, \(\pi_i \equiv \pi\) for a constant \(\pi\).
2. We include weekly lags (i.e., \(t-1\) to \(t-5\) lags) of all regressors.
3. We include weekly lags of all regressors but the diversity measure.
4. We include weekly lags of the regressors, as in (2), and constrain the slopes to be homogeneous, as in (1).
5. We include further lags of the regressors but the diversity measure, as in (4), and constrain the slopes to be homogeneous, as in (1).

The table reports mean group estimates, and in parentheses are their standard errors. To conserve space we do not list estimates of lags from \(t-2\) to \(t-5\). Asterisks indicate significance levels (* \(p < 0.1\), ** \(p < 0.05\), *** \(p < 0.001\)).

The column corresponds to the specification that involves \(t-1\) up to \(t-5\) lags of all regressors. To conserve space we do not report estimates of the lags from \(t-2\) to \(t-5\). Noticeably, the mean group estimate (= 0.021) is no longer significant at all conventional levels, and it is substantially lower than the estimate when only first order lags of the regressors are included. In untabulated results, the mean group estimates of diversity\(i, t-j\), \(2 \leq j \leq 5\), are insignificant as well. The insignificance seems to be induced by the substantially larger standard error and hence much wider confidence interval: when we include only first order lag of the diversity measure as in Table 3.3, the standard error of diversity\(i, t-1\) is about 0.018, but the standard
error becomes 0.031 in this specification. We suspect that enlargement of the standard error is due to the colinearity induced by the high autocorrelation of the diversity measure, especially for the firms with low frequency of news coverage. Therefore we estimate a specification that includes weekly lags of all regressors but the diversity measure. Regression results of this specification are listed in the third column of the table. With only first order lag of the diversity measure included, the standard error falls back to 0.018 again, and the mean-group estimate of diversity $\gamma_{i,t-1}$ is significant at the 5% level. This seems to confirm our suspicion. The fourth column of Table 3.5 corresponds to the specification where we include weekly lags of all regressors and further impose the homogeneous constraint on the slopes. The specification underlying the fifth column is the same, except that we include only first order lag of the diversity measure. In either case the pooled mean group estimate of diversity $\gamma_{i,t-1}$ is significant at the 10% level. It should be noted that pooling the slopes tend to diminish the goodness of fit: the $R^2$ is no larger than 13% when we pool the slopes, but gets larger than 20% if we allow the slopes to be heterogeneous.

3.5.3 Alternative methods for building the diversity measure

For the last set of robustness check, we turn attention to different ways of building the diversity measure. To recapitulate, the diversity measure we have been using thus far is obtained in two steps: We first use the word2vec embedding to represent words as numeric vectors, and then we apply the word mover’s distance (WMD) to calculate the textual dissimilarity. The word2vec is not the only embedding we could use; the GloVe model developed by Pennington et al. (2014) is another popular unsupervised learning algorithm for distributed representation of words. Nor is WMD the only textual dissimilarity measure we could use; it may be replaced by the conventional cosine dissimilarity.

We check if our main conclusion is affected by the textual analysis techniques used to calculate the diversity measure. We consider the following alternative combinations of textual analysis techniques:

- **GloVe + WMD**: Instead of word2vec, we use the GloVe representation of words. The textual dissimilarity is still gauged by WMD. The resulting diversity measure is denoted by diversity $^{(G,W)}_{i,t}$.

- **Word2Vec + Cosine dissimilarity**: We still use the word2vec embedding, but the textual dissimilarity is gauged by the cosine dissimilarity. The resulting diversity measure is denoted by diversity $^{(W,C)}_{i,t}$. To illustrate the calculation of cosine dissimilarity, let $t = (w_1, w_2, \cdots, w_m)$ and $\tilde{t} = (\tilde{w}_1, \tilde{w}_2, \cdots, \tilde{w}_n)$ be two texts. If $e$ denote the word2vec embedding, the Bag-Of-Words (BOW) representation of $t$ is defined as the arithmetic average of the embeddings of individual words, that is,

$$BOW(t) \equiv \frac{1}{m} \sum_{i=1}^{m} e(w_i).$$
Similarly, the BOW representation of \( \tilde{t} \) is defined as
\[
BOW(\tilde{t}) \equiv \frac{1}{n} \sum_{j=1}^{n} e(\tilde{w}_j).
\]

The cosine similarity between the two texts is
\[
\text{cosine}(t, \tilde{t}) = \frac{\langle BOW(t), BOW(\tilde{t}) \rangle}{\|BOW(t)\| \|BOW(\tilde{t})\|},
\]
where \( \langle BOW(t), BOW(\tilde{t}) \rangle \) is the inner product between the Bag-Of-Words representations and \( \|\cdot\| \) is the Euclidean norm. The cosine similarity takes value between \(-1\) and 1; it is equal to 1 only if the two texts have exactly the same BOW representation. The cosine dissimilarity is simply defined as \(1 - \text{cosine}(t, \tilde{t})\).

- **GloVe + Cosine dissimilarity**: This is the same as the forgoing combination, except that we use the GloVe embedding instead. The resulting diversity measure is denoted by diversity\(^{(G,C)}_{i,t}\).
- **Jaccard Index**: We also build the diversity measure using Jaccard index. The Jaccard index of two texts \( t \) and \( \tilde{t} \) is defined as
\[
\text{Jaccard}(t, \tilde{t}) = \frac{\# \text{ unique words in } t \cap \tilde{t}}{\# \text{ unique words in } t \cup \tilde{t}},
\]
which is bounded between 0 and 1. The Jaccard index may be deemed as a similarity measure. It takes the maximum 1 if and only if the two texts are exactly the same. The dissimilarity of \( t \) and \( \tilde{t} \) may be gauged by \(1 - \text{Jaccard}(t, \tilde{t})\). We denote the resulting diversity measure by diversity\(^{(J)}_{i,t}\).

All alternative diversity measures except diversity\(^{(J)}_{i,t}\) involve an embedding – either word2vec or GloVe – that encodes semantic information, and hence implicitly take into account word senses. The two diversity measures diversity\(_{i,t}\) and diversity\(^{(G,W)}_{i,t}\) are highly correlated, as GloVe embeddings are very similar to word2vec embeddings. On our sample, the correlation coefficient between diversity\(_{i,t}\) and diversity\(^{(G,W)}_{i,t}\) is up to 0.95. The Jaccard index is based on matching of word forms and ignore semantic relations, which could render diversity\(^{(J)}_{i,t}\) upward biased.

Results of the regressions involving different diversity measures are listed in Table 3.6. Following discussion in the previous subsection, in all regressions we include weekly lags of all variables except the diversity measures. To save space only estimates of first order lags are reported. There are several notable results. First, as long as we use WMD to gauge textual dissimilarity, substituting GloVe for word2vec does not make any difference to our conclusion. This can be told by comparing the first two columns: the mean group estimates of diversity\(_{i,t}\) and diversity\(^{(G,W)}_{i,t}\) are almost the same, and both of them are significant at the 5% level. Mean group estimates of other variables are also qualitatively the same. Because of
The influence of diversity of media coverage: alternative diversity measures

Note: The table presents regression results using alternative diversity measures. All diversity measures but diversity are obtained in two steps: first choose a word embedding, then calculate the diversity measure with either word mover’s distance (WMD) or cosine dissimilarity.

<table>
<thead>
<tr>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
<th>(8)</th>
<th>(9)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obs</td>
<td>76862</td>
<td>76862</td>
<td>76862</td>
<td>76862</td>
<td>76862</td>
<td>76862</td>
<td>76862</td>
<td>76862</td>
</tr>
<tr>
<td>R²</td>
<td>(0.166)</td>
<td>(0.166)</td>
<td>(0.173)</td>
<td>(0.173)</td>
<td>(0.169)</td>
<td>(0.174)</td>
<td>(0.167)</td>
<td>(0.166)</td>
</tr>
<tr>
<td>β₁</td>
<td>-0.093</td>
<td>0.033</td>
<td>0.040*</td>
<td>0.040**</td>
<td>0.036**</td>
<td>0.037**</td>
<td>0.032*</td>
<td>0.009</td>
</tr>
<tr>
<td>β₂</td>
<td>0.025</td>
<td>0.025</td>
<td>0.025</td>
<td>0.025</td>
<td>0.025</td>
<td>0.025</td>
<td>0.025</td>
<td>0.025</td>
</tr>
<tr>
<td>β₃</td>
<td>0.009*</td>
<td>0.009</td>
<td>0.008</td>
<td>0.008</td>
<td>0.009</td>
<td>0.009</td>
<td>0.009</td>
<td>0.009</td>
</tr>
<tr>
<td>β₄</td>
<td>0.155***</td>
<td>0.157***</td>
<td>0.155***</td>
<td>0.158***</td>
<td>0.155***</td>
<td>0.158***</td>
<td>0.156***</td>
<td>0.158***</td>
</tr>
<tr>
<td>β₅</td>
<td>-0.006</td>
<td>0.002</td>
<td>-0.004*</td>
<td>-0.004**</td>
<td>-0.005***</td>
<td>-0.005***</td>
<td>-0.005***</td>
<td>-0.005***</td>
</tr>
<tr>
<td>β₆</td>
<td>0.009</td>
<td>0.009</td>
<td>0.009</td>
<td>0.009</td>
<td>0.009</td>
<td>0.009</td>
<td>0.009</td>
<td>0.009</td>
</tr>
<tr>
<td>β₇</td>
<td>0.008</td>
<td>0.008</td>
<td>0.008</td>
<td>0.008</td>
<td>0.008</td>
<td>0.008</td>
<td>0.008</td>
<td>0.008</td>
</tr>
<tr>
<td>β₈</td>
<td>0.008</td>
<td>0.008</td>
<td>0.008</td>
<td>0.008</td>
<td>0.008</td>
<td>0.008</td>
<td>0.008</td>
<td>0.008</td>
</tr>
<tr>
<td>β₉</td>
<td>0.008</td>
<td>0.008</td>
<td>0.008</td>
<td>0.008</td>
<td>0.008</td>
<td>0.008</td>
<td>0.008</td>
<td>0.008</td>
</tr>
</tbody>
</table>

Table 3.6: Robustness of the Findings and Further Analysis
Chapter 3. How Does Diversity of Media Coverage Influence Firms’ Risk Exposure?

the high correlation, neither diversity\textsubscript{\textit{i},\textit{t−1}} nor diversity\textsubscript{(\textit{G,W})\textit{i},\textit{t−1}} is significant if both are present in the regression; see column (3). Second, if the diversity measure is built on cosine dissimilarity, it does not predict risk exposure at any conventional significance level, whichever word embedding we use. This can be seen from the results in columns (4) to (7). More importantly, the results suggest that the diversity measure built on WMD contain information that is not encoded in the diversity measures built on cosine dissimilarity. In column (5), for example, the mean group estimate of diversity\textsubscript{\textit{i},\textit{t}} remains significant at the 5% level, though the diversity measure diversity\textsubscript{(\textit{W,C})\textit{i},\textit{t}} built on cosine dissimilarity is also included in regression. Third, results in column (8) suggest that the diversity measure diversity\textsubscript{(\textit{J})\textit{i},\textit{t}} predicts neither rise nor drop of the firm's exposure to systematic risk. Finally, results in column (9) show that inclusion of diversity\textsubscript{(\textit{J})\textit{i},\textit{t}} renders the mean group estimate of diversity\textsubscript{\textit{i},\textit{t}} substantially higher.

3.5.4 Brief summary of the robustness checks

Robustness of our main conclusion that diversity of media coverage positively predicts firms’ exposure to systematic risk is tied to the significance level governing the test. At the 10% level, the main conclusion holds regardless of the sub-sampling method, pooling of the slopes and inclusion of further lags of all variables but the diversity measure itself. But if one feels like running less risk of type I error and adopt the more conservative 5% level, the main conclusion may not be claimed robust, especially when we sub-sample in various ways. Finally, it seems important to build the diversity measure on a word embedding that encodes semantic information and the word mover’s distance. Diversity measures calculated with cosine dissimilarity or Jaccard index do not lead to the same conclusion.

3.6 Conclusion

This paper explores the interactions between diversity of media coverage and firms’ exposure to systematic risk, as measured by beta. We employ the recently developed word2vec embedding (Mikolov et al. (2013); Mikolov et al. (2013a)) and the word mover’s distance (Kusner et al. (2015)) to gauge the breadth of information in the news stories. We find that news diversity positively predicts risk exposure on the coming day. The conclusion remains valid after we control for stock market variables that are known to predict beta and various textual measures that reflect sentiment, salience of media coverage and social media attention. All in all, our study suggests another way of applying natural language processing techniques to analyze the role of media in financial market.
Acknowledgment

We thank Asger Lunde, Dunja Mladenic and participants of the weekly seminar at Artificial Intelligence Laboratory, Jožef Stefan Institute for helpful comments. We are grateful to Gregor Leban for granting access to EventRegistry. The second author thanks CREATES for research support and Jožef Stefan Institute for hosting him in summer 2018, during which the paper was drafted. The research project leading to these results receives funding from the European Union Research and Innovation programme Horizon 2020 under grant agreement No. 675044 (BigDataFinance).
Appendix A Data Pre-processing and Software Usage

A.1 Cleaning of TAQ data

We follow the rules described in Barndorff-Nielsen et al. (2009), except that we do not implement their rule T4 for dropping “outliers” (i.e., transactions with price distant from highest bid or lowest ask).

P1. Remove transactions that occur outside the interval 9:30am to 4pm EST.

P2. Remove transactions with zero price.

P3. Keep only transactions that occur in the stock’s primary listing.

T1. Remove transactions with correction record, that is, trades with a nonzero correction indicator \( \text{CORR} \).

T2. Remove transactions with abnormal sale condition, that is, trades where \( \text{COND} \) has a code other than '@', 'E' and 'F'.

T3. If multiple transactions have identical timestamp, use the median price and sum up the volumes.

A.2 Retrieval of news articles

Here we elaborate retrieval of the news articles from EventRegistry\textsuperscript{18}. To download news articles from EventRegistry we make use of the python package \texttt{eventregistry}\textsuperscript{19}. Key options for customizing download include concepts, news sources and language. Concepts of an article comprehends both concrete entities and abstract ideas mentioned in the article that have Wikipedia entry. For instance, in the article titled "Amazon’s Biggest Business Will Have Nothing To Do With Retail (AMZN)" and published by Business Insider, identified concepts include but not limited to “Amazon”, “Business”, “Retail” and so forth, since all these terms have stand-alone pages on Wikipedia. A firm might have multiple related concepts. Usually, the firm’s name is the concept that is linked to the most news, so we use it to retrieve news stories. We do not consider other concepts, such as the ones related to the firm’s prod-

\textsuperscript{18}EventRegistry is a repository of events: the events are automatically identified by analysing news articles collected from numerous sources around the world. The collection is done by virtue of the Newsfeed service which monitors RSS feeds of about 100,000 news outlets globally. Whenever a new article is released in the RSS feed, the Newsfeed service downloads all available information about the article and passes it to EventRegistry’s internal system, which then groups articles according to the events they refer to. The system also tags each article with all concepts mentioned therein. On average EventRegistry captures about 200,000 news articles a day, which are written in various languages, with English, Spanish and German being the most common ones. EventRegistry is indeed featured by its cross-lingual event-classification system. However, for our research we mainly download news articles stored by EventRegistry and exploit no further functionality. For more detailed introduction to EventRegistry, see Leban et al. (2014) and Rupnik et al. (2016).

\textsuperscript{19}Available on \\url{https://github.com/EventRegistry/event-registry-python}. Download requires user-specific API key whose acquisition is elaborated on the website.
The reason is two-fold: first, it is rare that a news story mentions product of a firm but not its name, thus using only the firm’s name omits just a few relevant stories. Second, even if there are such news stories, it is hard to keep track of the related concepts, since a firm’s products and services vary from time to time.

EventRegistry has a broad spectrum of news sources, ranging from newspapers with worldwide influence such as The Wall Street Journal and The New York Times to the now burgeoning web-based financial advisors such as Seeking Alpha and The Motley Fool. The comprehensive list of news sources might be beneficial in some studies, but it also arouses the concern about credibility of the sources. In fact, some channels that are known to disseminate unsubstantiated hearsays are also on EventRegistry’s track list. News stories dispersed by these channels can certainly influence their reader’s valuation of the firms and it would be interesting to inspect the role played by these channels, but in this paper we focus on those credible sources whose stories are likely to be evidence-based. We select 28 news sources that top the ranking of news websites by Alexa.com in terms of their traffic data. The news sources are listed in Table A4. The list encompasses many media with global reach, such as Yahoo, Bloomberg and Reuters. It also includes the sources widely used in textual analysis in finance, such as The New York Times and The Wall Street Journal (Tetlock (2007); Tetlock et al. (2008)).

In terms of language, only English stories are kept. Finally, to each article EventRegistry attaches a Boolean value indicating whether it duplicates another article. We remove all articles labeled as duplicated. The remaining ones are used for our empirical analysis.

The python package eventregistry is used to retrieve articles. The procedures are summarized below:

R1. Set ConceptUri to full name of the firm.
R2. Set dateStart to 2014/1/1 and dateEnd to 2018/5/31.
R3. Set SourceGroupUri to those in Table A4.
R4. Set location and category and set by default.
R5. Remove articles not written in English.
R6. Remove articles marked as duplicated.

---

20 Take Microsoft for example. Many products of Microsoft, such as windows (the operating system), office (the software) and surface (the touchscreen PC), have their own concepts, namely “Microsoft Windows”, “Microsoft Office” and “Microsoft Surface”.

21 Posts on these web-based financial advisors have also caught attention of researchers. Chen et al. (2014), for example, show how sentiment expressed in posts on Seeking Alpha helps to predict returns and earnings surprise.

22 Alexa.com is a subsidiary of Amazon that provides web traffic data and analytics. The original ranking is not confined to financial press and includes magazines such as Nationalgeographic.com, which we do not include in the study.

23 Duplication does not mean that the article is conceptually the same as another article, i.e., the two articles are classified by EventRegistry into the same event. It means that the articles captured at different instants have exactly the same content. This can happen since some sources in the list (e.g., Yahoo News) not only publish their own stories but serve as aggregator, that is, they aggregate links to stories published by other sources.
A.3 Software usage

Our research has benefited from many publicly available packages. We use the python suite NLTK (Bird and Loper (2004)) for text preprocessing. For word embedding, we use Google's pretrained word2vec model (Mikolov et al. (2013); Mikolov et al. (2013a); Mikolov et al. (2013b)). The model is trained on about 100 billion words extracted Google News and contains a vocabulary of 3 million words and phrases, each embedded into the 300 × 1 real Euclidean space. The model is published on July 30, 2013 before the first date of our sample, so its usage induces no look-ahead bias. The word mover's distance is computed with the python toolkit gensim\textsuperscript{24}. Finally, the dynamic panel is estimated by the Stata package xtdcce2, which is contributed by Ditzen (2018).

\textsuperscript{24}See Řehůřek and Sojka (2010). Implementation of WMD is based on Pele and Werman (2008) and Pele and Werman (2009).
Appendix B  Review of Word2Vec

This section briefly reviews word2vec, the building block of our diversity measure. Word2vec has two models for word embedding: continuous bag-of-words (CBOW) and skip-gram. We only review the skip-gram model as it is what we utilize in the empirical analysis. Though neural network is the workhorse of word2vec, we shall abstract it away in the first half of the review and focus on conveying the main idea. Only in the second half of the review do we need (basic) notions of neural network.

B.1 Skip-gram: The main idea

 Throughout the sequel we suppose availability of a vocabulary \( \mathcal{V} \), whose size is denoted by \( V \). It helps to treat the vocabulary as ordered, that is, \( \mathcal{V} \equiv \{ v_j : 1 \leq j \leq V \} \). Each word in the vocabulary may be viewed as an element of \( \mathbb{R}^V \). In particular, the \( i \)th word \( v_i \) in the vocabulary may be represented as a \( V \times 1 \) vector with a single 1 on the \( i \)th coordinate and zeros elsewhere, that is,

\[
v_i = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \\ \end{pmatrix}_{V \times 1} \quad \leftarrow \text{\( i \)th.} \tag{B1}
\]

The one-hot encoded representation is too sparse for most computational purposes, and it carries no semantic or syntactic information. Our goal is to embed words in the vocabulary into a denser space, and ideally the embedding should cluster words with closely related meanings or the same lemma\(^{25} \). Formally, we look for an embedding \( e : \mathbb{R}^V \rightarrow \mathbb{R}^d \), such that \( d \) is low enough for efficient computation, and the embeddings \( e(w_1) \) and \( e(w_2) \) have short Euclidean distance if the words \( w_1 \) and \( w_2 \) are near synonyms, or are hypernym and hyponym, or share the same lemma.

The skip-gram model obtains the embedding by training a neural network for the (pseudo) task of context prediction. We describe the task first and then give explanation. Suppose the availability of a training corpus, which we represent as a sequence of ordered words \( (w_1, w_2, \ldots, w_N) \). Choose a hyperparameter \( C \) and a parametric model \( \{ p(w' | w; \theta) : \theta \in \Theta \} \), where \( \Theta \) is a finite dimensional space and \( p(w' | w; \theta) \) is the posterior probability of the word \( w' \) given \( w \). The exact meaning of the posterior probability will become clear below, and its exact form will be given in the next subsection. The task is to find the maximum

\(^{25}\)Lemma is the citation form of a word in the dictionary. For example, the words “studies” and “studied” have the same lemma “study”.
of the following objective function:

$$J(\theta) \equiv \prod_{j=1}^{N} \prod_{j-C \leq k \leq j+C, k \neq j} p(w_k|w_j; \theta), \theta \in \Theta.$$  \hfill (B2)

Some explanations are in order. First of all, formulation of the task draws on distributional semantics. The tenet of distributional semantics is that words occur in the same contexts tend to purport similar meanings, an idea dating back to Firth (1957). For example, the word “risk” tends to co-occur with the words “insurance”, “insurer”, “hedge” etc., as in the following sentences, and the co-occurrence characterizes proximity of these words’ senses to the sense of “risk”:

1. The insurer may hedge its own risk by taking out reinsurance.
2. Insurance is used to hedge against the risk of a contingent or uncertain loss.

The idea suggests that co-occurrence of words can be exploited to produce embedding that preserves semantic as well as syntactic relations. We start off by building a parametric model that predicts the context of given a word $w$, say the preceding $C$ words as well as the succeeding $C$ words of $w$. The set of the $2C$ surrounding words is called context of $w$ and the hyperparameter $C$ is termed context window size. The prediction accuracy of the model is gauged by the posterior probability $p(w_k|w_j; \theta)$, which may be understood as the probability that $w'$ occurs in the context of $w$. If $w'$ appears frequently in the context of $w$, an accurate predictive model should have high $p(w'|w; \theta)$ to capture the frequent co-occurrence.

Bearing the idea in mind, we turn attention to the training corpus. In the corpus, the word $w_j$ has the context $\{w_{j-C}, \ldots, w_{j-1}, w_{j+1}, \ldots, w_{j+C}\}$. The knowledge of co-occurrence of $w_j$ and its context words can be used to train the model. Intuitively, a well-trained model should reflect the co-occurrence by assigning high value to the posterior probabilities $p(w_k|w_j; \theta)$, $j-C \leq k \leq j+C$ and hence to their product

$$\prod_{j-C \leq k \leq j+C, k \neq j} p(w_k|w_j; \theta).$$

Of course, each word in the corpus and its associated context can be used to train the model, so we further take product of the posterior probabilities and obtain the objective function $J(\theta)$ in (B2).

Figure A1 illustrates the training with a corpus having only one sentence. The context window size is set to $C \equiv 2$. In the first step, the given word is “insurance”, and accuracy of the prediction gauged by

$$p(\text{company}|\text{insurance}; \theta) \times p(\text{uses}|\text{insurance}; \theta) \times p(\text{to}|\text{insurance}; \theta) \times p(\text{hedge}|\text{insurance}; \theta).$$

In the second step, the context window rolls on, and we predict the context given the word


B. Review of Word2Vec

Figure A1: An example of skip-gram training

The company uses insurance to hedge the risk of loss.
The company uses insurance to hedge the risk of loss.
The company uses insurance to hedge the risk of loss.

“hedge”. Accuracy of the prediction is gauged by

\[ p(\text{insurance}|\text{hedge}; \theta) \times p(\text{to}|\text{hedge}; \theta) \times p(\text{the}|\text{hedge}; \theta) \times p(\text{risk}|\text{hedge}; \theta). \]

In the last step, the given word is “risk” and accuracy of the prediction is gauged by

\[ p(\text{hedge}|\text{risk}; \theta) \times p(\text{the}|\text{risk}; \theta) \times p(\text{of}|\text{risk}; \theta) \times p(\text{loss}|\text{risk}; \theta). \]

The task is thus to find the maximum of

\[
J(\theta) \equiv p(\text{company}|\text{insurance}; \theta) \times p(\text{uses}|\text{insurance}; \theta) \times p(\text{to}|\text{insurance}; \theta) \times p(\text{hedge}|\text{insurance}; \theta) \\
\times p(\text{insurance}|\text{hedge}; \theta) \times p(\text{to}|\text{hedge}; \theta) \times p(\text{the}|\text{hedge}; \theta) \times p(\text{risk}|\text{hedge}; \theta) \\
\times p(\text{hedge}|\text{risk}; \theta) \times p(\text{the}|\text{risk}; \theta) \times p(\text{of}|\text{risk}; \theta) \times p(\text{loss}|\text{risk}; \theta).
\]

We said at the beginning that context prediction is just a pseudo task. We are not interested in context prediction per se; rather, we are interested in the parameter

\[
\hat{\theta} = \arg \max_{\theta \in \Theta} J(\theta).
\]

Intuitively, finding \( \hat{\theta} \) may be deemed as finding the model that “best” captures co-occurrences of the words in the corpus. Since co-occurrence reflects proximity of word senses, the semantic and syntactic relations should be encoded in the maximum \( \hat{\theta} \), which could serve as the embedding. This gives an inkling of why the skip-gram model is able to produce word embeddings that preserve semantic and syntactic relations.

B.2 Skip-gram: The underlying neural network

A shallow neural network with single hidden layer is introduced when it comes to the parameterization of the posterior probabilities \( \{p(w'|w; \theta) : \theta \in \Theta\} \).

We give the formula for \( p(v_j|v_i; \theta) \), where \( v_i \) and \( v_j \) are, respectively, the \( i \)th and the \( j \)th words in the vocabulary. The neural network is depicted in Figure A2. Between the input

\[26\text{We stress that this is only an intuitive explanation. Jurafsky and Martin (2018) provides more intuitions in the section on word2vec. But the exact reason as to why word2vec leads to unprecedented performance in many language processing tasks is, to our knowledge, still unknown.} \]

168
The neural network underlying the skip-gram model

$$v_i \xrightarrow{E} e_i \xrightarrow{F} f'_1 e_i \xrightarrow{\text{softmax}} e' \xrightarrow{\text{softmax}} p(v_1 | v_i; \theta) \xrightarrow{\text{softmax}} p(v_2 | v_i; \theta) \xrightarrow{\text{softmax}} p(v_V | v_i; \theta)$$

and the hidden layers is a $d \times V$ matrix

$$E \equiv \begin{bmatrix} e_1 & e_2 & \cdots & e_V \end{bmatrix}, \quad e_j \in \mathbb{R}^d.$$

By the one-hot encoded representation (B1), $Ev_i$ is equal to $e_i$, the $i$th column of $E$. On the other hand, between the hidden- and the output-layer is another $d \times V$ matrix

$$F \equiv \begin{bmatrix} f_1 & f_2 & \cdots & f_V \end{bmatrix}, \quad f_j \in \mathbb{R}^d.$$

The posterior probabilities are parameterized by the $\theta \equiv (E, F)$. In particular, $p(v_j | v_i; \theta)$ is defined by

$$p(v_j | v_i; \theta) \equiv \frac{\exp(f'_j e_i)}{\sum_{k=1}^V \exp(f'_k e_i)}.$$

The formula for $p(w' | w; \theta)$, where $w$ and $w'$ are arbitrary words in the vocabulary, is simple: we just need to find the indexes $i$ and $j$ such that $v_i = w$ and $v_j = w'$, and set $p(w' | w; \theta) = p(v_j | v_i; \theta)$.

With $p(w' | w; \theta)$ parameterized, we could work on a corpus and train the model for context prediction to obtain $\hat{\theta} = (\hat{E}, \hat{F})$. The skip-gram word embedding \footnote{The skip-gram model actually learns two embeddings: $\hat{E}$ and $\hat{F}$. $\hat{E}$ is often termed target embedding while $\hat{F}$ is called context embedding. We use only the target embedding in the empirical analysis.} lies in the columns of the matrix $\hat{E}$. More explicitly, the embedding $e : \mathbb{R}^V \to \mathbb{R}^d$ is defined as

$$e(w) \equiv \hat{E}w.$$
B.3 Skip-gram: Some computational concerns

The discussions thus far have conveyed the main idea behind the skip-gram model. To train the model in practice, we still need to address several computational problems. Here we just give a brief review of the problems and their practical solutions. The reader may refer to the original paper by Mikolov et al. (2013a) for details.

The first problem is to identify phrase such as “New York Times” and “United Kingdom” that forms a meaningful unit. The meaning of such phrase is not simple composition of the meanings of its individual component words, so it is preferable to have embedding for the phrase as a whole. In our analysis, phrase identification and embedding is important for the firms whose names consist of multiple words (e.g., “American International Group”). Besides, identifying common phrases also helps to cut down size of the corpus and accelerate the training process. Mikolov et al. (2013a) use a simple data-driven approach to capture the phrases (p. 6). It turns out that their method is able to capture some but not all phrases appearing in our sample. For example, “American International Group” is successfully identified, but “Procter & Gamble” is not; in fact, “Procter &" is incorrectly identified as a meaningful phrase.

The second problem is related to words with too frequent occurrence in the training corpus. In a large training corpus, the definite article “the” co-occurs with almost every word, and, intuitively, it provides less information than the rare words. To make efficient use of the words, Mikolov et al. (2013a) apply a simple subsampling approach: each word \( w \) in the training set is discarded with probability given by the formula

\[
P(w) \equiv \max \left( 1 - \sqrt{\frac{\tau}{f(w)}}, 0 \right), \tag{B6}
\]

where \( f(w) \) is the frequency of \( w \) in the corpus and \( \tau \) is a user-chosen threshold, typically around \( 10^{-5} \). The more frequent is a word, the more likely it gets discarded. Though the formula is chosen heuristically, Mikolov et al. (2013a) find it to accelerate training and even significantly improve the accuracy of embeddings of rare words.

The third problem concerns efficient computation of \( \hat{\theta} \equiv (\hat{E}, \hat{F}) \), the maximum of \( J(\theta) \) defined in (B2). Obviously, \( \hat{\theta} \) is also the maximum of the log-linear transform \( L(\theta) \equiv \frac{1}{N} \log(J(\theta)) \). There is a computational pitfall in finding \( \hat{\theta} = \text{argmax} \ L(\theta) \). To illustrate, let \( \iota \) denote the “index” mapping that locates a word \( w \) in the vocabulary, i.e., \( w = v_{\iota(w)} \). As a common practice in training neural network model, \( \hat{\theta} \) is approximated using the (stochastic) gradient descent method. This method requires calculation of the derivative of \( L(\theta) \). Consider the partial derivative of \( L(\theta) \) over \( e_{\iota(w_j)} \). With the parameterization (B4), applica-
The computational pitfall is created by the sum $\sum_{l=1}^{V} p(v_l|v_\iota(w_j); \theta)f_i(w_l)$: the summation costs too many computational power when $V$ is large. To render the computation more tractable, Mikolov et al. (2013a) resort to the negative sampling method. Interested readers may refer to Goldberg and Levy (2014) for a detailed explanation on negative sampling.
Appendix C  
Estimation of the dynamic panel

Specification of the model is highly sophisticated, but the core coefficients \( \pi_i \equiv (\phi_i, \theta_i, \delta'_i)' \) in (3.11) can be easily estimated by applying OLS to an augmented regression. The key observation is that the common latent factor \( f_t \) can be proxied by detrended cross-section averages of the regressand and the regressors, which is defined by

\[
\tilde{z}_t \equiv \bar{z}_t - \bar{c}, \quad \text{where} \quad \bar{z}_t \equiv \frac{1}{N} \sum_{i=1}^{N} \begin{bmatrix} R\beta_{i,t} \\ \text{diversity}_{i,t} \\ x_{i,t} \end{bmatrix},
\]

where the number of lags \( p_T \) grows to infinity as \( T \to \infty \) and \( p_T^3/T \) converges to some positive constant. Let \( \hat{\pi}_i \equiv (\hat{\phi}_i, \hat{\theta}_i, \hat{\delta}'_i)' \) denote the OLS estimator of (C9). \( \hat{\pi}_i \) turns out to be consistent as both \( N \) and \( T \) grow to infinity, that is,

\[
\hat{\pi}_i - \pi_i \xrightarrow{p} 0 \quad \text{as} \quad (N, T) \xrightarrow{j} \infty,
\]

where \( \xrightarrow{p} \) means convergence in probability and \( \xrightarrow{j} \) the joint convergence. This remarkable result does not hinge on any relative convergence rate of \( N \) and \( T \); see Theorem 1 of Chudik and Pesaran (2015). As said above, our interest lies primarily in the mean levels of the coefficients \( \pi \equiv E[\pi_i] \). Consistency of firm-specific estimates readily carries over to consistency of the mean group estimate \( \hat{\pi}_{MG} \equiv \frac{1}{N} \sum_{i=1}^{N} \hat{\pi}_i \), namely

\[
\hat{\pi}_{MG} \xrightarrow{p} \pi, \quad \text{as} \quad (N, T) \xrightarrow{j} \infty.
\]
To draw inference on the coefficients we need further convergence in distribution of the mean group estimator. Under the additional assumption that $N/T$ converges to some positive constant, Chudik and Pesaran (2015) establish (see Theorem 3 therein)

$$\sqrt{N}(\hat{\pi}_{MG} - \pi) \xrightarrow{d} N(0, \Omega).$$

(C10)

The limit of $N/T$ does not enter the asymptotic covariance matrix $\Omega$, so the joint convergence only adds minor restriction. Consistent estimation of the $\Omega$ is achieved by

$$\hat{\Omega}_{MG} \equiv \frac{1}{N-1} \sum_{i=1}^{N} (\hat{\pi}_i - \hat{\pi}_{MG})(\hat{\pi}_i - \hat{\pi}_{MG})',$$

with which inference on the coefficients can be drawn as usual.

The only tuning parameter that needs our subjective choice is the truncation lag $p_T$. Following Chudik and Pesaran (2015), we simply set $p_T \equiv \lfloor T^{1/3} \rfloor = 10$ in the empirical analysis. All inferences are based on the mean group estimator $\hat{\pi}_{MG}$. Chudik and Pesaran (2015) have also proposed a half-panel jackknife method that corrects the $O(T^{-1})$ bias, but since we have sufficiently large $T$, we do not use the jackknife estimator in this paper.
Appendix D  Examples and Additional Details

D.1  An example of WMD

Figure A3: An example of Word Mover’s Distance

\[
\begin{align*}
S_1 & \quad \text{AIG tells investors about the judge’s ruling.} \\
& \quad 0.74 = 0.00 + 0.15 + 0.19 + 0.21 + 0.19 \\
S_0 & \quad \text{AIG informs shareholders about court decision.} \\
& \quad 0.99 = 0.00 + 0.25 + 0.22 + 0.25 + 0.27 \\
S_2 & \quad \text{AIG falls after disappointing earnings released.}
\end{align*}
\]

Figure A3 gives two examples of WMD. Both examples involve the sentence \(S_0 = \text{“AIG informs shareholders about court decision”}\). The first example shows calculation of the WMD between \(S_0\) and the sentence \(S_1 = \text{“AIG tells investors about the judge’s ruling”}\). The function words “about” and “the” are not involved in the calculation. The optimal transport plan from \(S_1\) to \(S_0\) turns out to move all shares of a word, say \(w \in S_1\), to the word \(\tilde{w}^* \in S_0\) to which \(w\) has the closest semantic distance\(^{29}\), i.e., \(\tilde{w}^* = \arg\min_{\tilde{w} \in S_0} \|e(w) - e(\tilde{w})\|\). Arrows in the figure indicate the transportations. We see that “AIG” is moved to “AIG”, “tells” to “informs”, “investors” to “shareholders”, “judge’s” to “court”\(^{30}\) and “ruling” to “decision”. The word mover’s distance between \(S_1\) and \(S_0\) is \(WMD(S_1, S_0) = 0.74\), which is obtained by summing transportation costs of individual words. The second example shows calculation of the WMD between \(S_0\) and the sentence \(S_2 = \text{“AIG falls after disappointing earnings released”}\). As in the first example, the optimal transport plan turns out to move all shares of each \(w \in S_2\) to the word in \(S_0\) to which \(w\) has the shortest semantic distance. Interestingly, the adjective “disappointing” in \(S_2\) is transported to the noun “decision” in \(S_0\); this is possibly due to frequent co-occurrence of the two words (e.g., “disappointing decision”) in the

\(^{29}\)We emphasize that this is not always the case, even if words in the two texts have the same uniform distribution. For example, it may be that two words \(w_1\) and \(w_2\) in the first text both have shortest semantic distance with the word \(\hat{w}\) in the second text, but since \(c(w_1, \hat{w}) < c(w_2, \hat{w})\), all vacancies of \(\hat{w}\) will be occupied by \(w_1\), and \(w_2\) must be transported to another word to which it does not have the shortest semantic distance.

\(^{30}\)Before calculating the WMD, we let the program automatically remove all punctuation marks, including the single quote. This causes “judge’s” to become “judges” in the actual calculation. The embeddings of “judge” and “judges” are different, as the word2vec does not apply any lemmatization or stemming to reduce words in the raw dataset to common base forms. The semantic distance between “judge” and “court” is 0.16, which is shorter than the semantic distance between “judges” and “court”.

174
training dataset of the underlying word2vec embedding. $WMD(S_2, S_0) = 0.99$ is larger than $WMD(S_1, S_0) = 0.74$. The difference between the two WMDs faithfully reflects the fact that $S_1$ shares the same topic with $S_0$ while $S_2$ does not.\(^{31}\)

\(^{31}\)The difference between $WMD(S_1, S_0)$ and $WMD(S_2, S_0)$ might not appear strikingly large to eyeball, but we stress that distribution of the WMD is highly leptokurtic. For AIG, $WMD(S_1, S_0) = 0.74$ is below the 5% quantile, whereas $WMD(S_2, S_0) = 0.99$ is at the 90% quantile.
D.2 Two examples of the diversity measure

Table A1: An example of short semantic distance between two news stories

<table>
<thead>
<tr>
<th>Bloomberg</th>
<th>Reuters</th>
</tr>
</thead>
<tbody>
<tr>
<td>American International Group Inc. won court approval of a $970.5 million settlement with investors who accused the insurer of mis-stating its exposure to the subprime mortgage market, leading to a liquidity crisis and $182.3 billion in federal bailouts.</td>
<td>American International Group Inc shareholders won approval on Friday of a $970.5 million settlement resolving claims they were misled about its subprime mortgage exposure, leading to a liquidity crisis and $182.3 billion in federal bailouts.</td>
</tr>
<tr>
<td>The investors, urging a federal judge to approve the accord, called it one of the largest securities class action recoveries stemming from the 2008 financial crisis. AIG was sued that year by investors who said the New York-based company misstated its exposure to the subprime mortgage market through its securities lending program and its credit-default swap portfolio.</td>
<td>U.S. District Judge Laura Taylor Swain in Manhattan granted final approval at a hearing to what lawyers for the investors call one of the largest class action settlements to come out of the 2008 financial crisis.</td>
</tr>
<tr>
<td>“It’s a great settlement and we are extremely gratified to have been able to achieve this result,” Jeffrey Golan, a lawyer for one of the lead plaintiffs in the case, the State of Michigan Retirement Systems, said Friday in a phone interview.</td>
<td>It marks the largest shareholder class action settlement in a case where no criminal or regulatory enforcement actions were ever pursued, the plaintiffs’ lawyers have said.</td>
</tr>
<tr>
<td>AIG, bailed out by the U.S. government during the crisis, will pay $960 million while the accounting firm PricewaterhouseCoopers LLP will pay $10.5 million, according to a Manhattan federal court filing. Lawyers for the lead plaintiffs on Friday also won approval of almost $116.5 million in legal fees, plus other litigation expenses, according to court filings.</td>
<td>AIG said it was pleased with the judge’s order.</td>
</tr>
<tr>
<td>U.S. District Judge Laura Taylor Swain set a May 5 deadline for investors to submit claims under the agreement. More than 40,000 already have been filed with the claims administrator in the case. The judge said at least 77 funds have opted out of the accord and may now sue independently, according to the court.</td>
<td>Swain noted on Friday that no potential class member had objected to the terms of the deal, which she said was “fair, reasonable and adequate” and should be approved. She added that the amount was “very substantial” and that shareholders would face significant risk if they continued to litigate instead of settling.</td>
</tr>
<tr>
<td>“We are pleased that the settlement has been approved by the court and look forward to putting this litigation behind us,” Jon Diat, an AIG spokesman, said in an e-mailed statement.</td>
<td>The settlement covers investors who bought AIG securities between March 16, 2006, and Sept. 16, 2008, when the company received its first bailout.</td>
</tr>
<tr>
<td>A separate related suit was filed by former employees who claimed AIG invested their retirement funds too heavily in the company’s own stock. That case is still pending. In 2012, AIG repaid the bailout, which reached $182.3 billion.</td>
<td>Swain overruled an objection by two people who bought AIG shares before the beginning of that period and said they should be included in the class.</td>
</tr>
<tr>
<td>The case is In re American International Group Inc. 08-cv-05722, U.S. District Court, Southern District of New York (Manhattan).</td>
<td>For the lawyers’ work, Swain on Friday awarded plaintiffs law firms Barrack, Rodos &amp; Bacine and The Miller Law Firm $116.46 million in fees plus more than $4 million in expenses.</td>
</tr>
<tr>
<td>Investors led by the State of Michigan Retirement Systems, which oversees several state pension plans, accused AIG of failing to disclose the risks it took on through its portfolio of credit default swaps and a securities lending program.</td>
<td>Investors also led by the State of Michigan Retirement Systems, which oversees several state pension plans, accused AIG of failing to disclose the risks it took on through its portfolio of credit default swaps and a securities lending program.</td>
</tr>
<tr>
<td>They said the failures led investors to buy stock and debt they otherwise would not have bought, resulting in billions of dollars in losses.</td>
<td>They said the failures led investors to buy stock and debt they otherwise would not have bought, resulting in billions of dollars in losses.</td>
</tr>
<tr>
<td>The government has since sold off its stake in AIG, resulting in a positive return of $22.7 million to the U.S. Treasury Department and Federal Reserve.</td>
<td>The government has since sold off its stake in AIG, resulting in a positive return of $22.7 million to the U.S. Treasury Department and Federal Reserve.</td>
</tr>
</tbody>
</table>

**NOTE:** This table gives an example of short semantic distance (= 0.2065) between two news stories. The two news, published by Bloomberg and Reuters respectively, are used in calculation of the news diversity measure of American International Group on March 23, 2015. Only the bold-faced sentences are used in computing the semantic distance, as only these sentences contain the reference “American International Group” (the ticker AIG is not used as reference). Six out of eight other news (not listed here) used to compute the diversity measure focus on the same event. The remaining two cover different topics; see Table A2. The diversity measure for AIG on this date is 0.3062.
**Table A2: An example of long semantic distance between two news stories**

**Bloomberg** - U.S. bank regulators are scheduled to vote next week on whether to extend MetLife Inc.’s deadline for filing a so-called living will as the insurer suits the government for subjecting it to stricter oversight.

The Federal Reserve and the Federal Deposit Insurance Corp. are “actively considering the possibility” of pushing back MetLife’s July 1, 2016, date for submitting a plan explaining how it would unwind itself if it were to enter bankruptcy, according to a court document released Friday night.

The request stems from MetLife’s suit against the Financial Stability Oversight Council, a group of regulators led by Treasury Secretary Jacob J. Lew. The panel in December deemed the insurer a systemically important non-bank, subjecting it to Fed oversight. Companies designated by the council must draw up a resolution plan, known as a living will.

The U.S. District Court in Washington has asked the FSOC to push back the deadline by about six months so that it can hear arguments on the case before MetLife “begins to expend time and money to prepare its resolution plan.” The Fed and FDIC are expected to vote in the next week on whether to grant an extension, according to the court filing.

MetLife’s lawsuit against the council, whose members include Fed Chair Janet Yellen and FDIC Chairman Martin Gruenberg, is the first challenging an FSOC decision. The council has labeled three other non-banks systemically important: insurers American International Group Inc. and Prudential Financial Inc., and General Electric Co.’s finance arm.

New York-based MetLife prefers that the court rule on the case in the first few months of 2016 because of the “substantial preparation and investment of resources” needed to comply with the FSOC designation. The systemically important label means that regulators think the company’s failure could pose risks to the financial system, though it doesn’t imply that the firm is currently facing difficulty.

Bank-holding companies with more than $50 billion in assets, such as Citigroup Inc. and Bank of America Corp., are automatically overseen by the Fed under the Dodd-Frank law, enacted in response to the 2008 crisis credit.

FDIC spokeswoman Barbara Hagenbaugh and Fed spokesman Eric Kollig declined to comment.

**Reuters** - A growing number of U.S. companies, including MillerCoors and AIG, are stepping up the battle against online ad fraud by demanding proof that their ads have been seen by real people instead of computers hijacked by cybercriminals.

Spurred by a warning in December by the Association of National Advertisers (ANA) that businesses are losing $6.3 billion a year to so-called "click fraud," these companies now stipulate in advertising contracts that they will only pay for online ads when given proof that humans clicked on them.

“We don’t want to be paying for non-human traffic,” said Mark Clowes, global head of advertising at American International Group Inc (AIG.N), the largest commercial insurer in the country.

In a typical click fraud scheme, a crook infects many computers with malicious software, and directs the machines - called bots - to visit a webpage, click on an ad or watch a video. The fake traffic fools a marketer into thinking the site is popular, so it pays to place ads and links on the site.

A study conducted by cybersecurity firm White Ops for the ANA found that bots viewed almost one-fourth of online video ads and 11 percent of display ads. The study, published in December, involved 36 participants including AIG and MillerCoors (SAB.L) (TAP.N).

While click fraud has been going on for years, the ANA report has galvanized advertisers to fight back, ad executives said. What had been a trickle of references in contracts is becoming a flood.

“We’ve written into all of our contracts that our clients insist on at least 95 percent human traffic and anything less requires a make good or credit,” said Barry Lowenthal, president of The Media Kitchen, an ad buying agency whose clients include Victoria’s Secret Pink.

“By the end of 2015 we expect that every major agency and every major advertiser will include these kinds of clauses in their terms and conditions.”

U.S. digital advertising revenue is expected to reach $59 billion this year, according to eMarketer.

Online ads are sold through a complex, opaque system that can involve many parties including advertising agencies, website publishers and automated ad exchanges - the middlemen who aggregate and resell ad space across many websites.

Website publishers often pay third parties to direct visitors to their sites, a practice known as “source traffic.” According to the 57-page ANA study, about half of all click fraud came from publishers paying for third-party traffic providers.

*(omitted to conserve space...)*

**Note:** This table gives an example of long semantic distance (= 0.9156) between two news stories. The two news, published by Bloomberg and Reuters respectively, are used in calculation of the news diversity measure of American International Group on March 23, 2015. Only the bold-faced sentences are used in computing the semantic distance, as only these sentences contain the reference “American International Group” (the ticker AIG is not used as reference). All other eight news used to compute the diversity measure focus on another event discussed in Table A1. The diversity measure for AIG on this date is 0.3062.
### D.3 Stock tickers and references for news retrieval

**Table A3: Ticker symbols and EventRegistry concepts for news retrieval**

<table>
<thead>
<tr>
<th>Ticker</th>
<th>Reference</th>
<th>Ticker</th>
<th>Reference</th>
<th>Ticker</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAPL</td>
<td>Apple</td>
<td>ABBV</td>
<td>Abbvie</td>
<td>ABT</td>
<td>Abbott</td>
</tr>
<tr>
<td>ACN</td>
<td>Accenture</td>
<td>AIG</td>
<td>American International Group</td>
<td>AMGN</td>
<td>Amgen</td>
</tr>
<tr>
<td>AMZN</td>
<td>Amazon</td>
<td>AXP</td>
<td>American Express</td>
<td>BA</td>
<td>Boeing</td>
</tr>
<tr>
<td>BAC</td>
<td>Bank of America</td>
<td>BIIB</td>
<td>Biogen</td>
<td>BLK</td>
<td>Blackrock</td>
</tr>
<tr>
<td>BMY</td>
<td>Bristol-Myers Squibb</td>
<td>C</td>
<td>Citigroup</td>
<td>CAT</td>
<td>Caterpillar</td>
</tr>
<tr>
<td>CELG</td>
<td>Celgene</td>
<td>CL</td>
<td>Colgate-Palmolive</td>
<td>COP</td>
<td>ConocoPhillips</td>
</tr>
<tr>
<td>COST</td>
<td>Costco Wholesale</td>
<td>CSCO</td>
<td>Cisco</td>
<td>CVS</td>
<td>CVS Health</td>
</tr>
<tr>
<td>DHR</td>
<td>Danaher</td>
<td>DIS</td>
<td>Walt Disney</td>
<td>DUK</td>
<td>Duke Energy</td>
</tr>
<tr>
<td>EBay</td>
<td>Ebay</td>
<td>EMR</td>
<td>Emerson Electric</td>
<td>ESRX</td>
<td>Express Scripts</td>
</tr>
<tr>
<td>F</td>
<td>Ford</td>
<td>FB</td>
<td>Facebook</td>
<td>FOXA</td>
<td>Century Fox</td>
</tr>
<tr>
<td>GE</td>
<td>General Electric</td>
<td>GILD</td>
<td>Gilead Sciences</td>
<td>GM</td>
<td>General Motors</td>
</tr>
<tr>
<td>GS</td>
<td>Goldman Sachs</td>
<td>HD</td>
<td>Home Depot</td>
<td>HON</td>
<td>Honeywell</td>
</tr>
<tr>
<td>HPQ</td>
<td>HP</td>
<td>IBM</td>
<td>IBM</td>
<td>INTC</td>
<td>Intel</td>
</tr>
<tr>
<td>INJ</td>
<td>Johnson &amp; Johnson</td>
<td>JPM</td>
<td>JPMorgan</td>
<td>KO</td>
<td>Coca-Cola</td>
</tr>
<tr>
<td>LLY</td>
<td>Eli Lilly</td>
<td>LMT</td>
<td>Lockheed Martin</td>
<td>MA</td>
<td>Mastercard</td>
</tr>
<tr>
<td>MCD</td>
<td>McDonald's</td>
<td>MDLZ</td>
<td>Mondelēz</td>
<td>MDT</td>
<td>Medtronic</td>
</tr>
<tr>
<td>MET</td>
<td>MetLife</td>
<td>MO</td>
<td>Altria</td>
<td>MON</td>
<td>Monsanto</td>
</tr>
<tr>
<td>MRK</td>
<td>Merck &amp; Co</td>
<td>MS</td>
<td>Morgan Stanley</td>
<td>MSFT</td>
<td>Microsoft</td>
</tr>
<tr>
<td>NKE</td>
<td>Nike</td>
<td>ORCL</td>
<td>Oracle</td>
<td>OXY</td>
<td>Occidental Petroleum</td>
</tr>
<tr>
<td>PEP</td>
<td>PepsiCo</td>
<td>PFE</td>
<td>Pfizer</td>
<td>PG</td>
<td>Procter &amp; Gamble</td>
</tr>
<tr>
<td>PM</td>
<td>Philip Morris</td>
<td>QCOM</td>
<td>Qualcomm</td>
<td>SBUX</td>
<td>Starbucks</td>
</tr>
<tr>
<td>SLB</td>
<td>Schlumberger</td>
<td>T</td>
<td>AT&amp;T</td>
<td>TWX</td>
<td>Time Warner</td>
</tr>
<tr>
<td>UNH</td>
<td>UnitedHealth</td>
<td>UNP</td>
<td>Union Pacific</td>
<td>UPS</td>
<td>United Parcel Service</td>
</tr>
<tr>
<td>USB</td>
<td>U.S. Bancorp</td>
<td>UTX</td>
<td>United Technologies</td>
<td>VZ</td>
<td>Verizon</td>
</tr>
<tr>
<td>WFC</td>
<td>Wells Fargo</td>
<td>WMT</td>
<td>Walmart</td>
<td>XOM</td>
<td>ExxonMobil</td>
</tr>
</tbody>
</table>
D.4 News sources

<table>
<thead>
<tr>
<th>Site</th>
<th>News Source</th>
<th>Site</th>
<th>News Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>cnn.com</td>
<td>CNN</td>
<td>money.cnn.com</td>
<td>CNN Money</td>
</tr>
<tr>
<td>theguardian.com</td>
<td>The Guardian</td>
<td>nypost.com</td>
<td>New York Post</td>
</tr>
<tr>
<td>indiatimes.com</td>
<td>indiatimes.com</td>
<td>cbsnews.com</td>
<td>CBS News</td>
</tr>
<tr>
<td>news.yahoo.com</td>
<td>Yahoo News</td>
<td>nbcnews.com</td>
<td>NBC News</td>
</tr>
<tr>
<td>foxnews.com</td>
<td>Fox News</td>
<td>usnews.com</td>
<td>US News &amp; World Report</td>
</tr>
<tr>
<td>washingtonpost.com</td>
<td>The Washington Post</td>
<td>economist.com</td>
<td>The economist</td>
</tr>
<tr>
<td>forbes.com</td>
<td>Forbes</td>
<td>fortune.com</td>
<td>Fortune</td>
</tr>
<tr>
<td>huffingtonpost.com</td>
<td>The Huffington Post</td>
<td>foxbusiness.com</td>
<td>Fox Business</td>
</tr>
<tr>
<td>usatoday.com</td>
<td>USA Today</td>
<td>washingtontimes.com</td>
<td>Washington Times</td>
</tr>
<tr>
<td>bloomberg.com</td>
<td>Bloomberg Business</td>
<td>businessinsider.com</td>
<td>Business Insider</td>
</tr>
<tr>
<td>cnbc.com</td>
<td>CNBC</td>
<td>marketwatch.com</td>
<td>MarketWatch</td>
</tr>
<tr>
<td>wsj.com</td>
<td>The Wall Street Journal</td>
<td>latimes.com</td>
<td>Los Angeles Times</td>
</tr>
<tr>
<td>reuters.com</td>
<td>Reuters</td>
<td>euronews.com</td>
<td>Euronews English</td>
</tr>
</tbody>
</table>
Appendix E  Firm by Firm Regression Results

In response to a referee’s request, this appendix presents additional results for firm-by-firm regressions. Specifically, for firm-\(i\) we estimate the linear time series model

\[ R\beta_{i,t} = c_{i} + \phi_{i}R\beta_{i,t-1} + \theta_{i}\text{diversity}_{i,t-1} + \delta^\prime_{i}x_{i,t-1} + u_{i,t} \]

This is exactly the regression equation for the dynamic heterogeneous model, but we drop the two equations (3.12) and (3.13) specifying multifactor error structure. The same set of controls is used, and we apply OLS estimator to the time series of each firm. The regression results are listed on the next a few pages, where we present only coefficient estimates of the diversity measure and the stock market controls to save space. Interpretation of the results is challenging. Perhaps the only discernible pattern is the positive autocorrelation in beta: for all firms the coefficient of lagged beta is positive and highly significant. For all other variables the results are mixed. The coefficient of the diversity measure is significant for a few firms and insignificant for the rest, and the significant ones are mostly negative. Setting aside the diversity measure, for many firms the coefficients for \(r_{i,t-1}\) and \(r_{i,t-1}^-\) are insignificant, which is in sharp contrast to the panel regression results presented in this paper. Same comments apply to other stock market variables.
### CHAPTER 3. HOW DOES DIVERSITY OF MEDIA COVERAGE INFLUENCE FIRMS’ RISK EXPOSURE?

<table>
<thead>
<tr>
<th></th>
<th>AAPL</th>
<th>ABBV</th>
<th>ABT</th>
<th>ACN</th>
<th>AIG</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{diversity}_{i,t-1} )</td>
<td>0.334</td>
<td>-0.109</td>
<td>-0.096</td>
<td>0.015</td>
<td>0.255***</td>
</tr>
<tr>
<td>( R\beta_{i,t-1} )</td>
<td>0.512***</td>
<td>0.262***</td>
<td>0.265***</td>
<td>0.146***</td>
<td>0.374***</td>
</tr>
<tr>
<td>( \text{vol}_{i,t-1} )</td>
<td>0.037*</td>
<td>-0.070**</td>
<td>0.022</td>
<td>0.011</td>
<td>0.087***</td>
</tr>
<tr>
<td>( \text{illiq}_{i,t-1} )</td>
<td>0.022*</td>
<td>0.002</td>
<td>0.006</td>
<td>0.007</td>
<td>0.012</td>
</tr>
<tr>
<td>( r_{i,t-1} )</td>
<td>0.003</td>
<td>0.055***</td>
<td>-0.026</td>
<td>-0.022</td>
<td>-0.047**</td>
</tr>
<tr>
<td>( r_{i,t-1}^- )</td>
<td>-0.006</td>
<td>-0.116***</td>
<td>0.029</td>
<td>0.004</td>
<td>0.081*</td>
</tr>
<tr>
<td>( R^2 )</td>
<td>0.302</td>
<td>0.142</td>
<td>0.080</td>
<td>0.046</td>
<td>0.188</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>AMGN</th>
<th>AMZN</th>
<th>AXP</th>
<th>BA</th>
<th>BAC</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{diversity}_{i,t-1} )</td>
<td>-0.305*</td>
<td>0.091</td>
<td>0.364***</td>
<td>0.046</td>
<td>0.181</td>
</tr>
<tr>
<td>( R\beta_{i,t-1} )</td>
<td>0.233***</td>
<td>0.476***</td>
<td>0.417***</td>
<td>0.239***</td>
<td>0.577***</td>
</tr>
<tr>
<td>( \text{vol}_{i,t-1} )</td>
<td>-0.036</td>
<td>0.091***</td>
<td>-0.049***</td>
<td>-0.024</td>
<td>0.019</td>
</tr>
<tr>
<td>( \text{illiq}_{i,t-1} )</td>
<td>-0.030*</td>
<td>-0.008</td>
<td>-0.008</td>
<td>-0.036***</td>
<td>0.001</td>
</tr>
<tr>
<td>( r_{i,t-1} )</td>
<td>0.047*</td>
<td>0.040*</td>
<td>0.012</td>
<td>0.056***</td>
<td>0.042*</td>
</tr>
<tr>
<td>( r_{i,t-1}^- )</td>
<td>-0.101**</td>
<td>-0.095**</td>
<td>-0.024</td>
<td>-0.143***</td>
<td>-0.087**</td>
</tr>
<tr>
<td>( R^2 )</td>
<td>0.084</td>
<td>0.319</td>
<td>0.213</td>
<td>0.092</td>
<td>0.384</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>BIIB</th>
<th>BLK</th>
<th>BMY</th>
<th>C</th>
<th>CAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{diversity}_{i,t-1} )</td>
<td>0.390**</td>
<td>0.089</td>
<td>-0.316**</td>
<td>0.364***</td>
<td>-0.063</td>
</tr>
<tr>
<td>( R\beta_{i,t-1} )</td>
<td>0.390***</td>
<td>0.142***</td>
<td>0.168***</td>
<td>0.555***</td>
<td>0.394***</td>
</tr>
<tr>
<td>( \text{vol}_{i,t-1} )</td>
<td>0.052</td>
<td>0.013</td>
<td>-0.080***</td>
<td>0.111***</td>
<td>0.023</td>
</tr>
<tr>
<td>( \text{illiq}_{i,t-1} )</td>
<td>0.008</td>
<td>-0.001</td>
<td>-0.002</td>
<td>0.022</td>
<td>-0.002</td>
</tr>
<tr>
<td>( r_{i,t-1} )</td>
<td>-0.014</td>
<td>0.026</td>
<td>0.040*</td>
<td>0.015</td>
<td>0.017</td>
</tr>
<tr>
<td>( r_{i,t-1}^- )</td>
<td>-0.013</td>
<td>-0.031</td>
<td>-0.074*</td>
<td>-0.016</td>
<td>-0.034</td>
</tr>
<tr>
<td>( R^2 )</td>
<td>0.214</td>
<td>0.030</td>
<td>0.057</td>
<td>0.414</td>
<td>0.190</td>
</tr>
</tbody>
</table>
### E. Firm by Firm Regression Results

<table>
<thead>
<tr>
<th></th>
<th>CELG</th>
<th>CL</th>
<th>COP</th>
<th>COST</th>
<th>CSCO</th>
</tr>
</thead>
<tbody>
<tr>
<td>diversity&lt;sub&gt;<em>i,t</em>,1&lt;/sub&gt;</td>
<td>−0.031</td>
<td>−0.187**</td>
<td>0.174</td>
<td>0.143*</td>
<td>0.013</td>
</tr>
<tr>
<td></td>
<td>(0.161)</td>
<td>(0.094)</td>
<td>(0.174)</td>
<td>(0.074)</td>
<td>(0.105)</td>
</tr>
<tr>
<td>Rβ&lt;sub&gt;<em>i,t</em>,1&lt;/sub&gt;</td>
<td>0.279***</td>
<td>0.264***</td>
<td>0.518***</td>
<td>0.164***</td>
<td>0.327***</td>
</tr>
<tr>
<td></td>
<td>(0.029)</td>
<td>(0.030)</td>
<td>(0.026)</td>
<td>(0.030)</td>
<td>(0.029)</td>
</tr>
<tr>
<td>vol&lt;sub&gt;<em>i,t</em>,1&lt;/sub&gt;</td>
<td>−0.004</td>
<td>−0.052**</td>
<td>0.249***</td>
<td>−0.101***</td>
<td>−0.083***</td>
</tr>
<tr>
<td></td>
<td>(0.039)</td>
<td>(0.026)</td>
<td>(0.040)</td>
<td>(0.025)</td>
<td>(0.021)</td>
</tr>
<tr>
<td>illiq&lt;sub&gt;<em>i,t</em>,1&lt;/sub&gt;</td>
<td>0.027</td>
<td>−0.035***</td>
<td>0.044**</td>
<td>−0.025**</td>
<td>−0.024*</td>
</tr>
<tr>
<td></td>
<td>(0.021)</td>
<td>(0.013)</td>
<td>(0.020)</td>
<td>(0.012)</td>
<td>(0.012)</td>
</tr>
<tr>
<td>r&lt;sub&gt;<em>i,t</em>,1&lt;/sub&gt;</td>
<td>0.046*</td>
<td>0.063***</td>
<td>−0.011</td>
<td>0.104***</td>
<td>0.080***</td>
</tr>
<tr>
<td></td>
<td>(0.025)</td>
<td>(0.029)</td>
<td>(0.022)</td>
<td>(0.025)</td>
<td>(0.022)</td>
</tr>
<tr>
<td>r&lt;sub&gt;−i,t_,1&lt;/sub&gt;</td>
<td>−0.094**</td>
<td>−0.139***</td>
<td>0.029</td>
<td>−0.196***</td>
<td>−0.146***</td>
</tr>
<tr>
<td></td>
<td>(0.045)</td>
<td>(0.053)</td>
<td>(0.041)</td>
<td>(0.047)</td>
<td>(0.040)</td>
</tr>
<tr>
<td>R&lt;sup&gt;2&lt;/sup&gt;</td>
<td>0.141</td>
<td>0.125</td>
<td>0.429</td>
<td>0.060</td>
<td>0.159</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>CVX</th>
<th>DHR</th>
<th>DIS</th>
<th>DUK</th>
<th>EBAY</th>
</tr>
</thead>
<tbody>
<tr>
<td>diversity&lt;sub&gt;<em>i,t</em>,1&lt;/sub&gt;</td>
<td>0.041</td>
<td>−0.008</td>
<td>0.233***</td>
<td>0.165</td>
<td>−0.242</td>
</tr>
<tr>
<td></td>
<td>(0.134)</td>
<td>(0.124)</td>
<td>(0.065)</td>
<td>(0.106)</td>
<td>(0.201)</td>
</tr>
<tr>
<td>Rβ&lt;sub&gt;<em>i,t</em>,1&lt;/sub&gt;</td>
<td>0.484***</td>
<td>0.136***</td>
<td>0.356***</td>
<td>0.447***</td>
<td>0.299***</td>
</tr>
<tr>
<td></td>
<td>(0.027)</td>
<td>(0.036)</td>
<td>(0.028)</td>
<td>(0.027)</td>
<td>(0.030)</td>
</tr>
<tr>
<td>vol&lt;sub&gt;<em>i,t</em>,1&lt;/sub&gt;</td>
<td>0.096***</td>
<td>−0.006</td>
<td>0.057***</td>
<td>0.018</td>
<td>−0.028</td>
</tr>
<tr>
<td></td>
<td>(0.024)</td>
<td>(0.022)</td>
<td>(0.022)</td>
<td>(0.033)</td>
<td>(0.032)</td>
</tr>
<tr>
<td>illiq&lt;sub&gt;<em>i,t</em>,1&lt;/sub&gt;</td>
<td>0.004</td>
<td>0.013</td>
<td>0.006</td>
<td>0.003</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>(0.013)</td>
<td>(0.011)</td>
<td>(0.011)</td>
<td>(0.016)</td>
<td>(0.017)</td>
</tr>
<tr>
<td>r&lt;sub&gt;<em>i,t</em>,1&lt;/sub&gt;</td>
<td>0.013</td>
<td>−0.019</td>
<td>−0.020</td>
<td>−0.018</td>
<td>−0.013</td>
</tr>
<tr>
<td></td>
<td>(0.020)</td>
<td>(0.021)</td>
<td>(0.022)</td>
<td>(0.035)</td>
<td>(0.026)</td>
</tr>
<tr>
<td>r&lt;sub&gt;−i,t_,1&lt;/sub&gt;</td>
<td>−0.021</td>
<td>0.037</td>
<td>0.035</td>
<td>−0.001</td>
<td>0.024</td>
</tr>
<tr>
<td></td>
<td>(0.038)</td>
<td>(0.040)</td>
<td>(0.039)</td>
<td>(0.063)</td>
<td>(0.051)</td>
</tr>
<tr>
<td>R&lt;sup&gt;2&lt;/sup&gt;</td>
<td>0.361</td>
<td>0.157</td>
<td>0.277</td>
<td>0.234</td>
<td>0.106</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>EMR</th>
<th>ESRX</th>
<th>F</th>
<th>FB</th>
<th>FOXA</th>
</tr>
</thead>
<tbody>
<tr>
<td>diversity&lt;sub&gt;<em>i,t</em>,1&lt;/sub&gt;</td>
<td>−0.155</td>
<td>−0.409***</td>
<td>−0.074</td>
<td>0.355**</td>
<td>0.288</td>
</tr>
<tr>
<td></td>
<td>(0.131)</td>
<td>(0.145)</td>
<td>(0.140)</td>
<td>(0.148)</td>
<td>(0.178)</td>
</tr>
<tr>
<td>Rβ&lt;sub&gt;<em>i,t</em>,1&lt;/sub&gt;</td>
<td>0.337***</td>
<td>0.172***</td>
<td>0.278***</td>
<td>0.596***</td>
<td>0.096***</td>
</tr>
<tr>
<td></td>
<td>(0.031)</td>
<td>(0.030)</td>
<td>(0.029)</td>
<td>(0.024)</td>
<td>(0.033)</td>
</tr>
<tr>
<td>vol&lt;sub&gt;<em>i,t</em>,1&lt;/sub&gt;</td>
<td>0.052*</td>
<td>−0.029</td>
<td>0.025</td>
<td>0.053*</td>
<td>−0.020</td>
</tr>
<tr>
<td></td>
<td>(0.029)</td>
<td>(0.034)</td>
<td>(0.031)</td>
<td>(0.028)</td>
<td>(0.035)</td>
</tr>
<tr>
<td>illiq&lt;sub&gt;<em>i,t</em>,1&lt;/sub&gt;</td>
<td>0.034**</td>
<td>0.002</td>
<td>−0.002</td>
<td>−0.013</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td>(0.017)</td>
<td>(0.016)</td>
<td>(0.019)</td>
<td>(0.016)</td>
<td>(0.020)</td>
</tr>
<tr>
<td>r&lt;sub&gt;<em>i,t</em>,1&lt;/sub&gt;</td>
<td>−0.028</td>
<td>0.039</td>
<td>0.038</td>
<td>0.090***</td>
<td>−0.028</td>
</tr>
<tr>
<td></td>
<td>(0.024)</td>
<td>(0.026)</td>
<td>(0.025)</td>
<td>(0.023)</td>
<td>(0.023)</td>
</tr>
<tr>
<td>r&lt;sub&gt;−i,t_,1&lt;/sub&gt;</td>
<td>0.042</td>
<td>−0.054</td>
<td>−0.070</td>
<td>−0.187***</td>
<td>0.011</td>
</tr>
<tr>
<td></td>
<td>(0.045)</td>
<td>(0.045)</td>
<td>(0.045)</td>
<td>(0.041)</td>
<td>(0.045)</td>
</tr>
<tr>
<td>R&lt;sup&gt;2&lt;/sup&gt;</td>
<td>0.165</td>
<td>0.055</td>
<td>0.118</td>
<td>0.505</td>
<td>0.035</td>
</tr>
</tbody>
</table>
## Chapter 3. How Does Diversity of Media Coverage Influence Firms’ Risk Exposure?

<table>
<thead>
<tr>
<th></th>
<th>GE</th>
<th>GILD</th>
<th>GM</th>
<th>GS</th>
<th>HD</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{diversity}_{i,t-1}$</td>
<td>0.226**</td>
<td>0.087</td>
<td>-0.159</td>
<td>0.042</td>
<td>-0.072</td>
</tr>
<tr>
<td></td>
<td>(0.110)</td>
<td>(0.134)</td>
<td>(0.142)</td>
<td>(0.191)</td>
<td>(0.102)</td>
</tr>
<tr>
<td>$R\beta_{i,t-1}$</td>
<td>0.210***</td>
<td>0.339***</td>
<td>0.326***</td>
<td>0.339***</td>
<td>0.242***</td>
</tr>
<tr>
<td></td>
<td>(0.030)</td>
<td>(0.029)</td>
<td>(0.029)</td>
<td>(0.029)</td>
<td>(0.029)</td>
</tr>
<tr>
<td>$\text{vol}_{i,t-1}$</td>
<td>0.036*</td>
<td>0.148***</td>
<td>0.062**</td>
<td>0.052</td>
<td>-0.021</td>
</tr>
<tr>
<td></td>
<td>(0.022)</td>
<td>(0.038)</td>
<td>(0.029)</td>
<td>(0.033)</td>
<td>(0.022)</td>
</tr>
<tr>
<td>$\text{illiq}_{i,t-1}$</td>
<td>0.003</td>
<td>-0.007</td>
<td>0.027*</td>
<td>-0.002</td>
<td>-0.015</td>
</tr>
<tr>
<td></td>
<td>(0.012)</td>
<td>(0.018)</td>
<td>(0.015)</td>
<td>(0.016)</td>
<td>(0.010)</td>
</tr>
<tr>
<td>$r_{i,t-1}$</td>
<td>-0.008</td>
<td>0.028</td>
<td>-0.017</td>
<td>0.012</td>
<td>0.063***</td>
</tr>
<tr>
<td></td>
<td>(0.020)</td>
<td>(0.024)</td>
<td>(0.021)</td>
<td>(0.027)</td>
<td>(0.019)</td>
</tr>
<tr>
<td>$r^-_{i,t-1}$</td>
<td>0.015</td>
<td>-0.064</td>
<td>0.022</td>
<td>-0.028</td>
<td>-0.100***</td>
</tr>
<tr>
<td></td>
<td>(0.036)</td>
<td>(0.044)</td>
<td>(0.038)</td>
<td>(0.048)</td>
<td>(0.033)</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.063</td>
<td>0.245</td>
<td>0.135</td>
<td>0.154</td>
<td>0.096</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>HON</th>
<th>HPQ</th>
<th>IBM</th>
<th>INTC</th>
<th>JNJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{diversity}_{i,t-1}$</td>
<td>-0.128</td>
<td>-0.101</td>
<td>0.104</td>
<td>0.038</td>
<td>0.028</td>
</tr>
<tr>
<td></td>
<td>(0.089)</td>
<td>(0.186)</td>
<td>(0.115)</td>
<td>(0.172)</td>
<td>(0.103)</td>
</tr>
<tr>
<td>$R\beta_{i,t-1}$</td>
<td>0.264***</td>
<td>0.154***</td>
<td>0.355***</td>
<td>0.379***</td>
<td>0.360***</td>
</tr>
<tr>
<td></td>
<td>(0.029)</td>
<td>(0.030)</td>
<td>(0.029)</td>
<td>(0.029)</td>
<td>(0.029)</td>
</tr>
<tr>
<td>$\text{vol}_{i,t-1}$</td>
<td>0.029</td>
<td>0.058*</td>
<td>0.011</td>
<td>0.001</td>
<td>0.028</td>
</tr>
<tr>
<td></td>
<td>(0.019)</td>
<td>(0.033)</td>
<td>(0.021)</td>
<td>(0.025)</td>
<td>(0.020)</td>
</tr>
<tr>
<td>$\text{illiq}_{i,t-1}$</td>
<td>0.002</td>
<td>0.046**</td>
<td>0.011</td>
<td>-0.037*</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>(0.011)</td>
<td>(0.018)</td>
<td>(0.010)</td>
<td>(0.015)</td>
<td>(0.010)</td>
</tr>
<tr>
<td>$r_{i,t-1}$</td>
<td>-0.004</td>
<td>-0.022</td>
<td>-0.026</td>
<td>0.066***</td>
<td>0.019</td>
</tr>
<tr>
<td></td>
<td>(0.023)</td>
<td>(0.021)</td>
<td>(0.022)</td>
<td>(0.021)</td>
<td>(0.024)</td>
</tr>
<tr>
<td>$r^-_{i,t-1}$</td>
<td>0.002</td>
<td>0.030</td>
<td>0.034</td>
<td>-0.128***</td>
<td>-0.034</td>
</tr>
<tr>
<td></td>
<td>(0.042)</td>
<td>(0.040)</td>
<td>(0.040)</td>
<td>(0.040)</td>
<td>(0.042)</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.129</td>
<td>0.064</td>
<td>0.132</td>
<td>0.199</td>
<td>0.168</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>JPM</th>
<th>KO</th>
<th>LLY</th>
<th>LMT</th>
<th>MA</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{diversity}_{i,t-1}$</td>
<td>-0.151</td>
<td>0.043</td>
<td>-0.147</td>
<td>0.198</td>
<td>0.024</td>
</tr>
<tr>
<td></td>
<td>(0.115)</td>
<td>(0.089)</td>
<td>(0.158)</td>
<td>(0.151)</td>
<td>(0.133)</td>
</tr>
<tr>
<td>$R\beta_{i,t-1}$</td>
<td>0.409***</td>
<td>0.385***</td>
<td>0.145***</td>
<td>0.294***</td>
<td>0.182***</td>
</tr>
<tr>
<td></td>
<td>(0.028)</td>
<td>(0.028)</td>
<td>(0.030)</td>
<td>(0.030)</td>
<td>(0.030)</td>
</tr>
<tr>
<td>$\text{vol}_{i,t-1}$</td>
<td>-0.026</td>
<td>0.009</td>
<td>0.050*</td>
<td>0.060**</td>
<td>0.018</td>
</tr>
<tr>
<td></td>
<td>(0.023)</td>
<td>(0.021)</td>
<td>(0.029)</td>
<td>(0.024)</td>
<td>(0.024)</td>
</tr>
<tr>
<td>$\text{illiq}_{i,t-1}$</td>
<td>-0.032***</td>
<td>-0.003</td>
<td>-0.002</td>
<td>0.023*</td>
<td>0.024*</td>
</tr>
<tr>
<td></td>
<td>(0.011)</td>
<td>(0.012)</td>
<td>(0.015)</td>
<td>(0.014)</td>
<td>(0.012)</td>
</tr>
<tr>
<td>$r_{i,t-1}$</td>
<td>0.073***</td>
<td>-0.017</td>
<td>0.017</td>
<td>-0.025</td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td>(0.020)</td>
<td>(0.029)</td>
<td>(0.023)</td>
<td>(0.027)</td>
<td>(0.020)</td>
</tr>
<tr>
<td>$r^-_{i,t-1}$</td>
<td>-0.141***</td>
<td>-0.021</td>
<td>-0.036</td>
<td>0.026</td>
<td>-0.014</td>
</tr>
<tr>
<td></td>
<td>(0.035)</td>
<td>(0.053)</td>
<td>(0.043)</td>
<td>(0.048)</td>
<td>(0.035)</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.224</td>
<td>0.189</td>
<td>0.040</td>
<td>0.148</td>
<td>0.058</td>
</tr>
</tbody>
</table>
### E. Firm by Firm Regression Results

<table>
<thead>
<tr>
<th></th>
<th>MCD</th>
<th>MDLZ</th>
<th>MDT</th>
<th>MET</th>
<th>MO</th>
</tr>
</thead>
<tbody>
<tr>
<td>diversity_{i,t-1}</td>
<td>0.182</td>
<td>-0.306**</td>
<td>-0.392***</td>
<td>-0.038</td>
<td>-0.004</td>
</tr>
<tr>
<td></td>
<td>(0.113)</td>
<td>(0.125)</td>
<td>(0.123)</td>
<td>(0.095)</td>
<td>(0.096)</td>
</tr>
<tr>
<td>Rβ_{i,t-1}</td>
<td>0.259***</td>
<td>0.164***</td>
<td>0.179***</td>
<td>0.490***</td>
<td>0.302***</td>
</tr>
<tr>
<td></td>
<td>(0.030)</td>
<td>(0.033)</td>
<td>(0.031)</td>
<td>(0.029)</td>
<td>(0.029)</td>
</tr>
<tr>
<td>vol_{i,t-1}</td>
<td>0.020</td>
<td>-0.055*</td>
<td>0.017</td>
<td>0.062*</td>
<td>-0.042</td>
</tr>
<tr>
<td></td>
<td>(0.017)</td>
<td>(0.030)</td>
<td>(0.024)</td>
<td>(0.034)</td>
<td>(0.030)</td>
</tr>
<tr>
<td>illiq_{i,t-1}</td>
<td>0.002</td>
<td>-0.019</td>
<td>0.004</td>
<td>-0.007</td>
<td>0.011</td>
</tr>
<tr>
<td></td>
<td>(0.010)</td>
<td>(0.017)</td>
<td>(0.012)</td>
<td>(0.017)</td>
<td>(0.013)</td>
</tr>
<tr>
<td>r_{i,t-1}</td>
<td>0.005</td>
<td>0.031</td>
<td>-0.024</td>
<td>-0.011</td>
<td>0.027</td>
</tr>
<tr>
<td></td>
<td>(0.022)</td>
<td>(0.026)</td>
<td>(0.021)</td>
<td>(0.026)</td>
<td>(0.023)</td>
</tr>
<tr>
<td>r_{i,t-1}</td>
<td>-0.025</td>
<td>-0.081*</td>
<td>0.028</td>
<td>0.023</td>
<td>-0.049</td>
</tr>
<tr>
<td></td>
<td>(0.041)</td>
<td>(0.048)</td>
<td>(0.039)</td>
<td>(0.046)</td>
<td>(0.040)</td>
</tr>
<tr>
<td>R²</td>
<td>0.085</td>
<td>0.069</td>
<td>0.072</td>
<td>0.262</td>
<td>0.158</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>MON</th>
<th>MRK</th>
<th>MS</th>
<th>MSFT</th>
<th>NKE</th>
</tr>
</thead>
<tbody>
<tr>
<td>diversity_{i,t-1}</td>
<td>0.041</td>
<td>-0.213**</td>
<td>0.222</td>
<td>0.142</td>
<td>0.088</td>
</tr>
<tr>
<td></td>
<td>(0.117)</td>
<td>(0.098)</td>
<td>(0.168)</td>
<td>(0.147)</td>
<td>(0.132)</td>
</tr>
<tr>
<td>Rβ_{i,t-1}</td>
<td>0.358***</td>
<td>0.257***</td>
<td>0.426***</td>
<td>0.473***</td>
<td>0.311***</td>
</tr>
<tr>
<td></td>
<td>(0.028)</td>
<td>(0.030)</td>
<td>(0.028)</td>
<td>(0.027)</td>
<td>(0.029)</td>
</tr>
<tr>
<td>vol_{i,t-1}</td>
<td>0.169***</td>
<td>0.048*</td>
<td>0.145***</td>
<td>-0.058***</td>
<td>-0.064***</td>
</tr>
<tr>
<td></td>
<td>(0.021)</td>
<td>(0.025)</td>
<td>(0.031)</td>
<td>(0.021)</td>
<td>(0.018)</td>
</tr>
<tr>
<td>illiq_{i,t-1}</td>
<td>0.042***</td>
<td>0.006</td>
<td>0.013</td>
<td>-0.017</td>
<td>-0.028**</td>
</tr>
<tr>
<td></td>
<td>(0.011)</td>
<td>(0.012)</td>
<td>(0.015)</td>
<td>(0.012)</td>
<td>(0.012)</td>
</tr>
<tr>
<td>r_{i,t-1}</td>
<td>-0.043**</td>
<td>0.031</td>
<td>0.003</td>
<td>0.036*</td>
<td>0.036*</td>
</tr>
<tr>
<td></td>
<td>(0.022)</td>
<td>(0.022)</td>
<td>(0.022)</td>
<td>(0.020)</td>
<td>(0.021)</td>
</tr>
<tr>
<td>r_{i,t-1}</td>
<td>0.066*</td>
<td>-0.042</td>
<td>-0.005</td>
<td>-0.067*</td>
<td>-0.099***</td>
</tr>
<tr>
<td></td>
<td>(0.038)</td>
<td>(0.039)</td>
<td>(0.040)</td>
<td>(0.035)</td>
<td>(0.038)</td>
</tr>
<tr>
<td>R²</td>
<td>0.332</td>
<td>0.100</td>
<td>0.270</td>
<td>0.251</td>
<td>0.134</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>ORCL</th>
<th>OXY</th>
<th>PEP</th>
<th>PFE</th>
<th>PG</th>
</tr>
</thead>
<tbody>
<tr>
<td>diversity_{i,t-1}</td>
<td>0.114</td>
<td>-0.083</td>
<td>-0.225**</td>
<td>-0.168*</td>
<td>-0.375***</td>
</tr>
<tr>
<td></td>
<td>(0.104)</td>
<td>(0.167)</td>
<td>(0.111)</td>
<td>(0.099)</td>
<td>(0.103)</td>
</tr>
<tr>
<td>Rβ_{i,t-1}</td>
<td>0.317***</td>
<td>0.267***</td>
<td>0.305***</td>
<td>0.266***</td>
<td>0.435***</td>
</tr>
<tr>
<td></td>
<td>(0.029)</td>
<td>(0.031)</td>
<td>(0.029)</td>
<td>(0.030)</td>
<td>(0.027)</td>
</tr>
<tr>
<td>vol_{i,t-1}</td>
<td>-0.011</td>
<td>0.113***</td>
<td>0.060***</td>
<td>-0.007</td>
<td>0.023</td>
</tr>
<tr>
<td></td>
<td>(0.022)</td>
<td>(0.040)</td>
<td>(0.019)</td>
<td>(0.022)</td>
<td>(0.019)</td>
</tr>
<tr>
<td>illiq_{i,t-1}</td>
<td>0.003</td>
<td>-0.003</td>
<td>0.013</td>
<td>-0.015</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>(0.012)</td>
<td>(0.019)</td>
<td>(0.012)</td>
<td>(0.013)</td>
<td>(0.012)</td>
</tr>
<tr>
<td>r_{i,t-1}</td>
<td>0.018</td>
<td>0.034</td>
<td>-0.073**</td>
<td>0.062***</td>
<td>-0.013</td>
</tr>
<tr>
<td></td>
<td>(0.020)</td>
<td>(0.027)</td>
<td>(0.030)</td>
<td>(0.023)</td>
<td>(0.028)</td>
</tr>
<tr>
<td>r_{i,t-1}</td>
<td>-0.027</td>
<td>-0.048</td>
<td>0.104*</td>
<td>-0.108**</td>
<td>-0.013</td>
</tr>
<tr>
<td></td>
<td>(0.037)</td>
<td>(0.052)</td>
<td>(0.053)</td>
<td>(0.042)</td>
<td>(0.050)</td>
</tr>
<tr>
<td>R²</td>
<td>0.122</td>
<td>0.210</td>
<td>0.118</td>
<td>0.099</td>
<td>0.237</td>
</tr>
</tbody>
</table>
# Chapter 3. How Does Diversity of Media Coverage Influence Firms’ Risk Exposure?

<table>
<thead>
<tr>
<th></th>
<th>PM</th>
<th>QCOM</th>
<th>SBUX</th>
<th>SLB</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>diversity&lt;sub&gt;i,t-1&lt;/sub&gt;</td>
<td>0.036</td>
<td>0.094</td>
<td>-0.202*</td>
<td>0.031</td>
<td>0.033</td>
</tr>
<tr>
<td></td>
<td>(0.097)</td>
<td>(0.122)</td>
<td>(0.122)</td>
<td>(0.101)</td>
<td>(0.081)</td>
</tr>
<tr>
<td>( R\beta_{i,t-1} )</td>
<td>0.313***</td>
<td>0.191***</td>
<td>0.346***</td>
<td>0.388***</td>
<td>0.236***</td>
</tr>
<tr>
<td></td>
<td>(0.029)</td>
<td>(0.031)</td>
<td>(0.028)</td>
<td>(0.028)</td>
<td>(0.030)</td>
</tr>
<tr>
<td>vol&lt;sub&gt;i,t-1&lt;/sub&gt;</td>
<td>-0.041</td>
<td>-0.070**</td>
<td>-0.049**</td>
<td>0.085**</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>(0.027)</td>
<td>(0.027)</td>
<td>(0.021)</td>
<td>(0.039)</td>
<td>(0.025)</td>
</tr>
<tr>
<td>illiq&lt;sub&gt;i,t-1&lt;/sub&gt;</td>
<td>0.021</td>
<td>0.013</td>
<td>-0.019</td>
<td>-0.002</td>
<td>0.007</td>
</tr>
<tr>
<td></td>
<td>(0.013)</td>
<td>(0.013)</td>
<td>(0.012)</td>
<td>(0.019)</td>
<td>(0.014)</td>
</tr>
<tr>
<td>( r_{i,t-1} )</td>
<td>-0.022</td>
<td>0.009</td>
<td>0.062***</td>
<td>-0.001</td>
<td>0.030</td>
</tr>
<tr>
<td></td>
<td>(0.026)</td>
<td>(0.017)</td>
<td>(0.021)</td>
<td>(0.027)</td>
<td>(0.027)</td>
</tr>
<tr>
<td>( r_{i,t-1}^* )</td>
<td>0.033</td>
<td>-0.003</td>
<td>-0.127***</td>
<td>-0.028</td>
<td>-0.089*</td>
</tr>
<tr>
<td></td>
<td>(0.046)</td>
<td>(0.030)</td>
<td>(0.039)</td>
<td>(0.053)</td>
<td>(0.049)</td>
</tr>
<tr>
<td>( R^2 )</td>
<td>0.147</td>
<td>0.098</td>
<td>0.183</td>
<td>0.176</td>
<td>0.095</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>TWX</th>
<th>UNH</th>
<th>UNP</th>
<th>UPS</th>
<th>USB</th>
</tr>
</thead>
<tbody>
<tr>
<td>diversity&lt;sub&gt;i,t-1&lt;/sub&gt;</td>
<td>-0.057</td>
<td>-0.095</td>
<td>0.094</td>
<td>-0.041</td>
<td>0.478***</td>
</tr>
<tr>
<td></td>
<td>(0.132)</td>
<td>(0.131)</td>
<td>(0.088)</td>
<td>(0.099)</td>
<td>(0.133)</td>
</tr>
<tr>
<td>( R\beta_{i,t-1} )</td>
<td>0.470***</td>
<td>0.207***</td>
<td>0.174***</td>
<td>0.235***</td>
<td>0.346***</td>
</tr>
<tr>
<td></td>
<td>(0.028)</td>
<td>(0.030)</td>
<td>(0.030)</td>
<td>(0.030)</td>
<td>(0.029)</td>
</tr>
<tr>
<td>vol&lt;sub&gt;i,t-1&lt;/sub&gt;</td>
<td>0.061**</td>
<td>0.062**</td>
<td>0.010</td>
<td>-0.021</td>
<td>0.019</td>
</tr>
<tr>
<td></td>
<td>(0.025)</td>
<td>(0.029)</td>
<td>(0.021)</td>
<td>(0.021)</td>
<td>(0.025)</td>
</tr>
<tr>
<td>illiq&lt;sub&gt;i,t-1&lt;/sub&gt;</td>
<td>0.032**</td>
<td>-0.001</td>
<td>-0.013</td>
<td>0.007</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>(0.013)</td>
<td>(0.014)</td>
<td>(0.013)</td>
<td>(0.012)</td>
<td>(0.013)</td>
</tr>
<tr>
<td>( r_{i,t-1} )</td>
<td>-0.009</td>
<td>0.016</td>
<td>0.030</td>
<td>0.004</td>
<td>-0.024</td>
</tr>
<tr>
<td></td>
<td>(0.021)</td>
<td>(0.024)</td>
<td>(0.020)</td>
<td>(0.025)</td>
<td>(0.026)</td>
</tr>
<tr>
<td>( r_{i,t-1}^* )</td>
<td>0.016</td>
<td>-0.060</td>
<td>-0.055</td>
<td>-0.035</td>
<td>0.031</td>
</tr>
<tr>
<td></td>
<td>(0.040)</td>
<td>(0.046)</td>
<td>(0.037)</td>
<td>(0.047)</td>
<td>(0.046)</td>
</tr>
<tr>
<td>( R^2 )</td>
<td>0.357</td>
<td>0.082</td>
<td>0.044</td>
<td>0.073</td>
<td>0.175</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>UTX</th>
<th>VZ</th>
<th>WFC</th>
<th>WMT</th>
<th>XOM</th>
</tr>
</thead>
<tbody>
<tr>
<td>diversity&lt;sub&gt;i,t-1&lt;/sub&gt;</td>
<td>0.090</td>
<td>0.054</td>
<td>0.006</td>
<td>-0.152</td>
<td>-0.172*</td>
</tr>
<tr>
<td></td>
<td>(0.103)</td>
<td>(0.095)</td>
<td>(0.111)</td>
<td>(0.124)</td>
<td>(0.097)</td>
</tr>
<tr>
<td>( R\beta_{i,t-1} )</td>
<td>0.297***</td>
<td>0.272***</td>
<td>0.499***</td>
<td>0.281***</td>
<td>0.528***</td>
</tr>
<tr>
<td></td>
<td>(0.029)</td>
<td>(0.029)</td>
<td>(0.027)</td>
<td>(0.030)</td>
<td>(0.027)</td>
</tr>
<tr>
<td>vol&lt;sub&gt;i,t-1&lt;/sub&gt;</td>
<td>0.008</td>
<td>-0.040*</td>
<td>0.007</td>
<td>-0.005</td>
<td>0.049**</td>
</tr>
<tr>
<td></td>
<td>(0.020)</td>
<td>(0.024)</td>
<td>(0.026)</td>
<td>(0.020)</td>
<td>(0.023)</td>
</tr>
<tr>
<td>illiq&lt;sub&gt;i,t-1&lt;/sub&gt;</td>
<td>-0.002</td>
<td>-0.006</td>
<td>0.000</td>
<td>0.007</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td>(0.012)</td>
<td>(0.014)</td>
<td>(0.013)</td>
<td>(0.011)</td>
<td>(0.011)</td>
</tr>
<tr>
<td>( r_{i,t-1} )</td>
<td>-0.011</td>
<td>0.040</td>
<td>-0.020</td>
<td>0.013</td>
<td>0.008</td>
</tr>
<tr>
<td></td>
<td>(0.023)</td>
<td>(0.027)</td>
<td>(0.023)</td>
<td>(0.022)</td>
<td>(0.021)</td>
</tr>
<tr>
<td>( r_{i,t-1}^* )</td>
<td>-0.010</td>
<td>-0.101**</td>
<td>0.032</td>
<td>-0.025</td>
<td>-0.013</td>
</tr>
<tr>
<td></td>
<td>(0.041)</td>
<td>(0.050)</td>
<td>(0.042)</td>
<td>(0.038)</td>
<td>(0.040)</td>
</tr>
<tr>
<td>( R^2 )</td>
<td>0.113</td>
<td>0.107</td>
<td>0.290</td>
<td>0.103</td>
<td>0.360</td>
</tr>
</tbody>
</table>
References


Declaration of co-authorship

Full name of the PhD student: Ye Zeng

This declaration concerns the following article/manuscript:

<table>
<thead>
<tr>
<th>Title:</th>
<th>Forecasting Daytime US Market Volatility with Overnight Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authors:</td>
<td>Giorgio Mirone and Ye Zeng</td>
</tr>
</tbody>
</table>

The article/manuscript is: Published ☐ Accepted ☐ Submitted ☐ In preparation ☒

If published, state full reference:

If accepted or submitted, state journal:

Has the article/manuscript previously been used in other PhD or doctoral dissertations?

No ☐ Yes ☒ If yes, give details: An earlier version of this paper is used in the first author’s PhD thesis (chapter 3), under the name “The HAR-F model: Incorporating overnight futures data in daytime stock volatility forecasting”.

The PhD student has contributed to the elements of this article/manuscript as follows:

A. Has essentially done all the work
B. Major contribution
C. Equal contribution
D. Minor contribution
E. Not relevant

<table>
<thead>
<tr>
<th>Element</th>
<th>Extent (A-E)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Formulation/identification of the scientific problem</td>
<td>C</td>
</tr>
<tr>
<td>2. Planning of the experiments/methodology design and development</td>
<td>C</td>
</tr>
<tr>
<td>3. Involvement in the experimental work/clinical studies/data collection</td>
<td>C</td>
</tr>
<tr>
<td>4. Interpretation of the results</td>
<td>C</td>
</tr>
<tr>
<td>5. Writing of the first draft of the manuscript</td>
<td>C</td>
</tr>
<tr>
<td>6. Finalization of the manuscript and submission</td>
<td>C</td>
</tr>
</tbody>
</table>

Signatures of the co-authors

<table>
<thead>
<tr>
<th>Date</th>
<th>Name</th>
<th>Signature</th>
</tr>
</thead>
<tbody>
<tr>
<td>January 17, 2019</td>
<td>Giorgio Mirone</td>
<td></td>
</tr>
</tbody>
</table>

In case of further co-authors please attach appendix

Date: Jan 17, 2019

Signature of the PhD student

Ye Zeng

1 of 1
**Declaration of co-authorship**

Full name of the PhD student: Ye Zeng

This declaration concerns the following article/manuscript:

<table>
<thead>
<tr>
<th>Title:</th>
<th>How Does Diversity of Media Coverage Influence Firms' Risk Exposure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authors:</td>
<td>Miha Torkar and Ye Zeng</td>
</tr>
</tbody>
</table>

The article/manuscript is: Published □ Accepted □ Submitted □ In preparation ✗

If published, state full reference:

If accepted or submitted, state journal:

Has the article/manuscript previously been used in other PhD or doctoral dissertations?

No □ Yes ✗ If yes, give details:

The PhD student has contributed to the elements of this article/manuscript as follows:

A. Has essentially done all the work
B. Major contribution
C. Equal contribution
D. Minor contribution
E. Not relevant

<table>
<thead>
<tr>
<th>Element</th>
<th>Extent (A-E)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Formulation/identification of the scientific problem</td>
<td>C</td>
</tr>
<tr>
<td>2. Planning of the experiments/methodology design and development</td>
<td>C</td>
</tr>
<tr>
<td>3. Involvement in the experimental work/clinical studies/data collection</td>
<td>C</td>
</tr>
<tr>
<td>4. Interpretation of the results</td>
<td>C</td>
</tr>
<tr>
<td>5. Writing of the first draft of the manuscript</td>
<td>C</td>
</tr>
<tr>
<td>6. Finalization of the manuscript and submission</td>
<td>C</td>
</tr>
</tbody>
</table>

**Signatures of the co-authors**

<table>
<thead>
<tr>
<th>Date</th>
<th>Name</th>
<th>Signature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan 9, 2019</td>
<td>Miha Torkar</td>
<td>[Signature]</td>
</tr>
</tbody>
</table>

Date: Jan 9, 2019

In case of further co-authors please attach appendix

Signature of the PhD student

Ye Zeng

1 of 1