

Journal of Financial Econometrics

Volume 18, Number 3, Summer 2020

Special Issue Articles on Predictive Modeling, Volatility,
and Risk Management in Financial Markets: In Memory
of Peter F. Christoffersen (Part I)

Introduction

- 471 Predictive Modeling, Volatility, and Risk
Management in Financial Markets: In Memory
of Peter F. Christoffersen (Part I)
*Francis X. Diebold, René Garcia and
Kris Jacobs*
-

Articles

- 473 The Term Structures of Expected Loss and Gain
Uncertainty
*Bruno Feunou, Ricardo Lopez Aliouchkin,
Roméo Tédongap and Lai Xu*
-
- 502 Realized Volatility Forecasting with Neural
Networks
Andrea Bucci
-
- 532 Realized Variance Modeling: Decoupling
Forecasting from Estimation
*Fabrizio Cipollini, Giampiero M. Gallo and
Alessandro Palandri*
-
- 556 Using the Extremal Index for Value-at-Risk
Backtesting
*Axel Bücher, Peter N. Posch and
Philipp Schmidtke*
-

Downloaded from <https://academic.oup.com/jfec/issue/18/3> by Oxford University Press USA user on 29 December 2021

<https://academic.oup.com/jfec>

ISSN (Print) 1479-8409
ISSN (Online) 1479-8417

Continued on back cover

Official journal of the Society for Financial Econometrics

OXFORD
UNIVERSITY PRESS

Editors

- Allan Timmermann
*University of California,
San Diego*
- Fabio Trojani
*University of Geneva and
Swiss Finance Institute*

Co-Editors

- Silvia Goncalves
McGill University
- Bryan Kelly
Yale University
- Dacheng Xiu
University of Chicago

Editorial Assistant

- Sarah King

Associate Editors

- Caio Almeida
Princeton University
- Drew Creal
University of Chicago
- Frank De Jong
Tilburg University
- Yanqin Fan
Vanderbilt University
- Andras Fulop
ESSEC Business School
- Nicola Fusari
Johns Hopkins University
- Patrick Gagliardini
University of Lugano
- Kay Gieseke
Stanford University
- Massimo Guidolin
University Bocconi
- Joel Hasbrouck
New York University
- Peter Hansen
University of North Carolina
- Nikolaus Hautsch
University of Vienna

Advisory Board

- Torben Gustav Andersen
Northwestern University
- John Y. Campbell
Harvard University
- Francis X. Diebold
University of Pennsylvania
- Robert F. Engle
*New York University and
University of California,
San Diego*
- A. Ronald Gallant
Pennsylvania State University
- John F. Geweke
*University of Technology,
Sydney*
- Eric Ghysels
*University of North Carolina
at Chapel Hill*
- Christian Gouriéroux
*University of Toronto and
INSEE-CREST*

- Christopher S. Jones
University of Southern California
- Frank Kleiberger
University of Amsterdam
- Markus Leippold
University of Zurich
- Haitao Li
*Cheung Kong Graduate
School of Business*
- Jia Li
Duke University
- Yingying Li
*Hong Kong University of
Science and Technology*
- André Lucas
Timbergen Institute
- Robin L. Lumsdaine
American University
- Loriano Mancini
University of Lugano
- Marcelo Medeiros
*Pontifical Catholic University of
Rio de Janeiro*
- Per Mykland
University of Chicago

- Lars Hansen
University of Chicago
- Wolfgang Härdle
*Humboldt-Universität
zu Berlin*
- Ravi Jagannathan
Northwestern University
- Adrian Pagan
*Australian National
University*
- George Tauchen
Duke University

Founding Editors

- René Garcia
*Finance, Law and
Accounting Department,
Edhec Business School*
- Eric Renault
*Department of Economics,
University of North Carolina
at Chapel Hill,
CIRANO and CIREQ*

- Roberto Renò
University of Verona
- Jeroen Rombouts
ESSEC Business School
- Alberto Rossi
University of Maryland
- Olivier Scaillet
*University of Geneva and
Swiss Finance Institute*
- Andrea Vedolin
Boston University
- Jules van Binsbergen
Wharton University
- Bas J. M. Werker
Tilburg University
- Michael Wolf
University of Zurich
- Jun Yu
*Singapore Management
University*
- Paolo Zaffaroni
Imperial College London
- Harold H. Zhang
*University of Texas
at Dallas*

Aims and Scope

The goal of *Journal of Financial Econometrics* is to reflect and advance the relationship between econometrics and finance, both at the methodological and at the empirical levels. Estimation, testing, learning, prediction and calibration in the framework of asset pricing or risk management represent the core focus. The scope includes topics relating to volatility processes, return modelling, dynamic conditional moments, machine learning, big data, fintech, extreme values, long memory, dynamic mixture models, endogenous sampling transaction data, and microstructure of financial markets.

Contents

Special Issue Articles on Predictive Modeling, Volatility, and Risk Management in Financial Markets: In Memory of Peter F. Christoffersen (Part I)

Introduction

Predictive Modeling, Volatility, and Risk Management in Financial Markets: In Memory of Peter F. Christoffersen (Part I) <i>Francis X. Diebold, René Garcia and Kris Jacobs</i>	471
--	-----

Articles

The Term Structures of Expected Loss and Gain Uncertainty <i>Bruno Feunou, Ricardo Lopez Aliouchkin, Roméo Tédongap and Lai Xu</i>	473
Realized Volatility Forecasting with Neural Networks <i>Andrea Bucci</i>	502
Realized Variance Modeling: Decoupling Forecasting from Estimation <i>Fabrizio Cipollini, Giampiero M. Gallo and Alessandro Palandri</i>	532
Using the Extremal Index for Value-at-Risk Backtesting <i>Axel Bücher, Peter N. Posch and Philipp Schmidtke</i>	556
Mixed-Frequency Macro–Finance Factor Models: Theory and Applications <i>Elena Andreou, Patrick Gagliardini, Eric Ghysels and Mirco Rubin</i>	585
Implied Default Probabilities and Losses Given Default from Option Prices <i>Jennifer Conrad, Robert F. Dittmar and Allaudeen Hameed</i>	629

Predictive Modeling, Volatility, and Risk Management in Financial Markets: In Memory of Peter F. Christoffersen (Part I)

Francis X. Diebold¹, René Garcia² and Kris Jacobs³

¹University of Pennsylvania, ²University of Montreal and ³University of Houston

Peter F. Christoffersen left us in 2018, much too soon, at the age of fifty-one years. He was a world-renowned financial econometrics researcher, educator, lecturer, administrator (including hosting the 2014 SoFiE conference at the University of Toronto), and public servant (including the U.S. Federal Reserve System's Model Validation Committee, charged with reviewing the models used for bank supervision and regulation). If Peter was an esteemed colleague, he was equally a dear friend. His unbridled optimism, relaxed personality, and remarkable humility endeared him to all who knew him.

We honor Peter's path-breaking research in this special issue. Peter's style is marked by a masterful blend of intuition, theoretical rigor, and always, empirical relevance. It influenced and inspired countless others in academics and industry, worldwide. It has four basic, and highly intertwined, organizational themes:

1. Predictive models and their evaluation (e.g., his classic early work on evaluating the conditional calibration of interval forecasts, [Christoffersen 1998](#)—one of the *International Economic Review*'s ten most-cited papers since its founding in 1960)
2. Financial market risk measurement and management (e.g., his celebrated text, [Christoffersen 2003](#))
3. Asset return volatility modeling and forecasting (e.g., his survey, [Andersen et al. 2013](#))
4. Financial derivative markets with emphasis on options (e.g., [Christoffersen et al. 2009](#), one of his many widely cited papers).

We humbly offer this two-part special issue as a tribute to Peter. The included papers reflect his style and interests, not only methodologically as characterized above, but also in their wide variety of substantive applications, clearly testifying to the depth and breadth of the Christoffersen legacy.

References

- Andersen, T. G., T. Bollerslev, P. F. Christoffersen, and F. X. Diebold. 2013. "Financial Risk Measurement for Financial Risk Management." In G. Constantinides, M. Harris, and R. Stulz (eds.), *Handbook of the Economics of Finance*, Volume 2, Part B. Elsevier, 1127–1220.

- Christoffersen, P. 1998. Evaluating Interval Forecasts. *International Economic Review* 39: 841–862.
- Christoffersen, P. F. 2003. *Elements of Financial Risk Management*, Academic Press.
- Christoffersen, P. F., S. Heston, and K. Jacobs. 2009. The Shape and Term Structure of the Index Option Smirk: Why Multifactor Stochastic Volatility Models Work so Well. *Management Science* 55: 1914–1932.

The Term Structures of Expected Loss and Gain Uncertainty*

Bruno Feunou¹, Ricardo Lopez Aliouchkin², Roméo Tédongap³ and Lai Xu⁴

¹Bank of Canada, ²Syracuse University, ³ESSEC Business School and ⁴Syracuse University

Address correspondence to Bruno Feunou, Bank of Canada, 234 Wellington Street, Ottawa, Ontario, Canada K1A 0G9, or e-mail: feun@bankofcanada.ca

Received August 23, 2019; revised August 23, 2019; accepted April 16, 2020

Abstract

We document that the term structures of risk-neutral expected loss and gain uncertainty on S&P 500 returns are upward sloping on average. These shapes mainly reflect the higher premium required by investors to hedge downside risk and the belief that potential gains will increase in the long run. The term structures exhibit substantial time-series variation with large negative slopes during crisis periods. Through the lens of a flexible Jump-Diffusion framework, we evaluate the ability of existing reduced-form option pricing models to replicate these term structures. We stress that three ingredients are particularly important: (i) the inclusion of jumps; (ii) disentangling the price of negative jump risk from its positive analog in the stochastic discount factor specification; and (iii) specifying three latent factors.

Key words: Quadratic payoff, quadratic loss, quadratic gain, quadratic risk premium, options

JEL classification: G12

Financial economists have long agreed that to better understand asset returns, and also uncertainty about these returns, it is necessary to break them down into several components, each reflecting a different aspect through which an investment opportunity can be perceived, analyzed, and evaluated. Since a return (r) can be classified as either a loss ($-l$) if nonpositive or a gain (g) if nonnegative, it is natural to break it down into these two components, formally $r = g - l$, where $l = \max(0, -r)$ and $g = \max(0, r)$. This decomposition

* We would like to thank the editor Francis X. Diebold and two anonymous referees for their helpful comments, which greatly improved the article. Feunou gratefully acknowledges financial support from the the Canadian Derivatives Institute (CDI). Lopez Aliouchkin and Tédongap acknowledge the research grant from the Thule Foundation's Skandia research programs on "Long-Term Savings." The views expressed in this article are those of the authors and do not necessarily reflect those of the Bank of Canada.

of returns also leads to a similar decomposition of return uncertainty into loss and gain components, namely loss uncertainty and gain uncertainty (also referred to as downside and upside, respectively, or in the most recent literature as bad and good, respectively. See, for example, [Barndorff-Nielsen, Kinnebrock, and Shephard, 2010](#), [Patton and Sheppard, 2015](#), [Bekaert, Engstrom, and Ermolov, 2015](#), and [Kilic and Shaliastovich, 2019](#), just to name a few). Likewise, investment returns are assessed over a given horizon, which, together with the maturity of the payoff, is among the key elements that guide investment choices.

We argue that expectations of (per-period) asset returns uncertainty and its loss and gain components across different investment horizons (i.e., their respective term structures) are critical for understanding market views about short- and long-term loss and gain potential. When dealing with expectations of asset returns uncertainty, it is also important to distinguish between physical and risk-neutral expectations. On the one hand, physical expectations of uncertainty measure the degree to which investors anticipate that they could be wrong about their returns forecast. On the other hand, risk-neutral expectations of uncertainty additionally indicate how much investors are willing to pay for risk hedging or would require for risk compensation. The shape of the term structure of risk-neutral expectations of loss (gain) uncertainty reflects both the expected path of future loss (gain) volatility and different risk premia associated with downside (upside) risk at different maturities.

The primary goal of this article is to provide an empirical investigation of physical and risk-neutral expectations of loss uncertainty and gain uncertainty across different investment horizons. The main challenge resides in estimating or measuring these expectations using available financial data. Since current period uncertainty on a future period return is not observed, a large body of the literature relies on model-free measures that can readily be computed using realized returns. A popular measure of uncertainty is the realized variance that cumulates higher frequency squared returns over the investment horizon. We assume we observe returns at regular intra-month time intervals of length δ . The monthly realized return $r_{t-1,t}$ and the monthly realized variance $RV_{t-1,t}$ are defined by aggregating $r_{t-1+j\delta}$ and $r_{t-1+j\delta}^2$, respectively:

$$r_{t-1,t} = \sum_{j=1}^{1/\delta} r_{t-1+j\delta} \quad \text{and} \quad RV_{t-1,t} = \sum_{j=1}^{1/\delta} r_{t-1+j\delta}^2, \quad (1)$$

where $1/\delta$ is the number of higher-frequency returns in a monthly period (e.g., $\delta = 1/21$ for daily returns) and $r_{t-1+j/21}$ denotes the j th daily return of the monthly period starting from day $t - 1$ and ending on day t . The loss component of realized variance cumulates higher-frequency squared losses, $l_{t-1+j\delta}$, while the gain component sums up higher-frequency squared gains, $g_{t-1+j\delta}$, that is,

$$RV_{t-1,t}^l = \sum_{j=1}^{1/\delta} l_{t-1+j\delta}^2 \quad \text{and} \quad RV_{t-1,t}^g = \sum_{j=1}^{1/\delta} g_{t-1+j\delta}^2. \quad (2)$$

Thus, the realized variance is the sum of its loss and gain components.

As thoroughly discussed in [Feunou et al. \(2019\)](#), estimating or measuring risk-neutral expectations of the loss and gain realized variance is not feasible, nor are loss and gain variance swaps traded such that their strikes could then be observed measures for these risk-

neutral expectations. To illustrate this difficulty, consider the risk-neutral expectation of the monthly realized variance based on daily returns. It is the sum of risk-neutral expectations of daily squared returns, $\mathbb{E}^Q[r_{t-1+j\delta}^2]$, each of which would be computable in theory, for example, using the formula of [Bakshi, Kapadia, and Madan \(2003\)](#) for the price of the quadratic contract. Empirically, this would require data on one-day-to-maturity out-of-the-money options, which are not currently available. The same applies to risk-neutral expectations of the monthly loss and gain realized variance. It is therefore important to rely on a measure of asset returns uncertainty for which the unobserved expectations of loss and gain components can be consistently estimated or measured from the data. The quadratic payoff is one of such measures and is the focus in this article.

The quadratic payoff is the square of the realized return over the investment horizon; that is, the monthly quadratic payoff is simply defined as the squared monthly realized return, $r_{t-1,t}^2$. The quadratic payoff and the realized variance are therefore related as follows:

$$r_{t-1,t}^2 = RV_{t-1,t} + 2RA_{t-1,t}, \quad \text{where} \quad RA_{t-1,t} = \sum_{i=1}^{1/\delta-11/\delta-i} \sum_{j=1}^{1/\delta-i} r_{t-1+j\delta} r_{t-1+i\delta}, \quad (3)$$

and $RA_{t-1,t}$ can be defined as the realized autocovariance. Like the realized variance, the quadratic payoff is also a model-free measure of the asset returns uncertainty. The loss component of the quadratic payoff (or quadratic loss) is the squared loss, while the gain component (or quadratic gain) is the squared gain over the investment horizon. Formally and in monthly terms, this means $l_{t-1,t}^2$ and $g_{t-1,t}^2$, respectively. Also similar to the realized variance, the quadratic payoff is the sum of its loss and gain components. Contrary to the realized variance, both physical and risk-neutral expectations of quadratic loss and gain for various horizons can be consistently estimated or measured from the data. We provide full details in [Online Appendix](#) Section A.1. We therefore rely on the quadratic payoff as the measure of asset returns uncertainty when analyzing the term structure of expected loss uncertainty and gain uncertainty.

Using a large panel of S&P 500 Index options data with time to maturity ranging from 1 month to 12 months, we build model-free risk-neutral expected quadratic loss and gain term structures. Our methodology follows from [Bakshi, Kapadia, and Madan \(2003\)](#) and is similar to that used to compute the VIX index. Likewise, using high-frequency S&P 500 Index return data and relying on a state-of-the-art variance forecasting model considered by [Bekaert and Hoerova \(2014\)](#), we build physical expected quadratic loss and gain term structures. We ask to what extent variations in these term structures reflect changes in the anticipated path of future loss and gain uncertainty and, therefore, the extent to which they reflect changes in the risk premia.

Our results reveal new important findings. First, the average term structure of the physical expected quadratic loss is downward sloping (with a slope of -4.73 percentage square units), while the average term structure of the risk-neutral expected quadratic loss is upward sloping (with a slope of 3.63 percentage square units). This means that, on average, investors anticipate that the (per-period) loss potential decreases with the investment horizon; yet, at the same time on the market, hedging the long-term loss potential of stocks is more expensive than hedging the short-term loss potential. Second, the average term structure of the physical expected quadratic gain is upward sloping (with a slope of 7.09 percentage square units), while the average term structure of the risk-neutral expected quadratic

gain is slightly upward sloping and almost flat (with a slope of 1.01 percentage square units). Likewise, this means that, on average, investors foresee that the (per-period) gain potential increases with the investment horizon; yet, at the same time on the market, speculating on the short-term gain potential of stocks is almost as costly as speculating on the long-term gain potential.

Our estimates of physical and risk-neutral expectations of quadratic loss and quadratic gain allow us to compute the associated risk premia by taking the appropriate difference between the physical and risk-neutral expectations. We follow [Feunou et al. \(2019\)](#) and measure the loss quadratic risk premium (QRP) as the risk-neutral minus the physical expected quadratic loss. It is the premium paid for downside risk hedging and thus a measure of downside risk. Likewise, we measure the gain QRP as the physical minus the risk-neutral expected quadratic gain. It is the premium received for upside risk compensation and thus a measure of upside risk. We subsequently analyze the term structures of loss and gain QRPs, and we find that both term structures are upward sloping (with slopes of 8.36 and 6.09 percentage square units, respectively). Therefore, on average, the (per-period) downside and upside risks are both higher for long-term investments relative to short-term investments in stocks, and since the equity premium is a remuneration of both types of risk, this confirms the upward-sloping average term structure of the equity premium found elsewhere in the literature.

The secondary goal of this article is to evaluate whether leading option pricing models that predominantly appear to be special cases of the model of [Andersen, Fusari, and Todorov \(2015\)](#) (henceforth AFT) are able to replicate the actual term structures of the risk-neutral expected quadratic loss and gain. Key features of the three-factor AFT model are its flexibility and its ability to completely disentangle the negative from the positive jump dynamics. To enhance our understanding of the model ingredients underlying the statistical properties of the quadratic loss and gain, we also estimate several restricted variants of the AFT model. These include, among others, the two-factor diffusion model of [Christoffersen, Heston, and Jacobs \(2009\)](#) (denoted as the baseline model AFT0) and a version of the AFT model where the negative and positive jump dynamics are equal (denoted by AFT3). The AFT3 model essentially represents the vast majority of option and variance swaps models studied in the literature so far (see, e.g., [Bates, 2012](#), [Christoffersen, Jacobs, and Ornathanalai, 2012](#), [Eraker, 2004](#), [Chernov et al., 2003](#), [Huang and Wu, 2004](#), [Amengual and Xiu, 2018](#), and [Ait-Sahalia, Karaman, and Mancini, 2015](#)).¹ We find that accounting for jumps in asset prices is essential for the model to fit the term structure of the risk-neutral expected quadratic loss and gain. The AFT0 model overestimates the risk-neutral expected quadratic gain and underestimates the risk-neutral expected quadratic loss, but is able to fit the term structure of the risk-neutral expected quadratic payoff. We also find that a jump process rather than a diffusion process is the most important in fitting the term structure of the risk-neutral expected quadratic loss, while it appears to be the opposite for the term structure of the risk-neutral expected quadratic gain.

1 Some authors consider asymmetry in the jump size distribution (see, e.g., [Amengual and Xiu, 2018](#)). However, the jump size distribution is assumed to be constant and the time variation in jumps comes through the jump intensity, which is assumed to be the same regardless of the sign of the jump.

The AFT model is primarily used for the risk-neutral dynamics of asset prices, and we further couple it with a pricing kernel specification that maps the risk-neutral dynamics into the physical dynamics. All parameters are estimated to maximize the joint likelihood of risk-neutral expected quadratic loss and gain across the term structure together with the second and third risk-neutral cumulants of asset returns. We examine the ability of various pricing kernel specifications in matching the actual dynamics of the term structures of physical expected quadratic loss and gain. This is equivalent to matching the actual term structures of loss and gain QRP. Our results unequivocally point to the importance of disentangling the price of negative jumps from the price of positive jumps. In other words, a restricted version of the pricing kernel imposing the same price for the negative and positive jump risk is unable to match the dynamics of the loss and gain QRP together. This restricted version represents the vast majority of pricing kernels studied in the literature (see, e.g., Eraker, 2004, Santa-Clara and Yan, 2010, Christoffersen, Jacobs, and Ornthanalai, 2012, and Bates, 2012) and highlights its inability to account for the actual joint dynamics of the loss and gain QRP.

Our article is related to the recent literature that analyzes the term structure of variance swaps. Ait-Sahalia, Karaman, and Mancini (2015) and Amengual and Xiu (2018) specify reduced-form models for the term structure of total variance. Dew-Becker et al. (2017) investigate the ability of existing structural models to fit the observed term structure of variance swaps. We contribute to this literature by investigating the term structures of the two variance components. Our article also relates to another strand of the literature documenting the importance of analyzing loss and gain components of variance (risk-neutral or physical) and VRP. Barndorff-Nielsen, Kinnebrock, and Shephard (2010) provide theoretical arguments supporting the splitting of the total realized variance into loss and gain components.

The remainder of the article is organized as follows. Section 1 introduces definitions and notations of all quantities, the data, and the methodology for constructing the risk-neutral and physical term structure of expected quadratic loss and gain, and presents key empirical facts that any economically sound model should be able to replicate. Section 2 introduces the AFT model and provides some details on its properties, including the implied closed form for both the risk-neutral and physical expectation of the quadratic loss and gain. Section 3 provides details on the estimation of the AFT model. Section 4 evaluates the ability of the AFT model and its variants to fit the empirical facts. Section 5 concludes.

1 Methodology, Data, and Preliminary Analysis

In this section, we start by introducing the quadratic payoff and its loss and gain components, namely, the quadratic loss and the quadratic gain. Next, we introduce a heuristic theoretical framework to understand the difference between the quadratic loss and the quadratic gain. We discuss the methodology to measure the risk-neutral and physical expectations of the quadratic payoff, the quadratic loss and the quadratic gain, over a given investment horizon. For the purpose of computing these term structures, we present the data and provide descriptive statistics. Finally, we provide a preliminary analysis based on principal components extracted from these term structures.

1.1 Definitions

Let S_t denote the S&P 500 Index price at the end of day t , and for any investment horizon τ , let $r_{t,t+\tau}$ denote its (log) return from end of day t to end of day $t + \tau$, given by $r_{t,t+\tau} = \ln(S_{t+\tau}/S_t)$. Both the log return $r_{t,t+\tau}$ and the quadratic payoff $r_{t,t+\tau}^2$ are subject to a gain–loss decomposition as follows:

$$r_{t,t+\tau} = g_{t,t+\tau} - l_{t,t+\tau} \quad \text{and} \quad r_{t,t+\tau}^2 = g_{t,t+\tau}^2 + l_{t,t+\tau}^2, \tag{4}$$

where the gain $g_{t,t+\tau} = \max(0, r_{t,t+\tau})$ and the loss $l_{t,t+\tau} = \max(0, -r_{t,t+\tau})$ represent the positive and negative parts of the asset payoff, respectively. In other words, the gain and loss are nonnegative amounts flowing in and out of an average investor’s wealth, respectively. Since a positive gain and a positive loss cannot occur simultaneously, we observe that $g_{t,t+\tau} \cdot l_{t,t+\tau} = 0$. This gain–loss decomposition of an asset’s payoff is exploited in an asset pricing context by [Bernardo and Ledoit \(2000\)](#).

Our goal in this article is to study how the time series dynamics of risk-neutral expectations $\mathbb{E}_t^{\mathbb{Q}}[r_{t,t+\tau}^2]$ and $\mathbb{E}_t^{\mathbb{Q}}[g_{t,t+\tau}^2]$, and of the physical expectations $\mathbb{E}_t^{\mathbb{P}}[r_{t,t+\tau}^2]$ and $\mathbb{E}_t^{\mathbb{P}}[g_{t,t+\tau}^2]$, vary with the investment horizon τ , where the exponents \mathbb{Q} and \mathbb{P} indicate that the values are under the risk-neutral and the physical measures, respectively. Knowledge of these term structures can be relevant in various risk management contexts. Indeed, one can learn about investors’ anticipations of the degree of loss and gain uncertainty every day for each investment horizon, and also how much investors are willing to pay for hedging risk or would require for compensation of the associated risks over a given investment horizon.

Given the risk-neutral and physical expectations of the same random quantity, one can readily take their difference to measure the associated risk premium. Following [Feunou et al. \(2019\)](#), we define the difference between the risk-neutral and physical expectations of the quadratic payoff as the QRP, where the loss and gain components, called loss QRP and gain QRP and denoted by $QRP_t^l(\tau)$ and $QRP_t^g(\tau)$, respectively, are formally given by

$$QRP_t^l(\tau) \equiv \mathbb{E}_t^{\mathbb{Q}}[l_{t,t+\tau}^2] - \mathbb{E}_t^{\mathbb{P}}[l_{t,t+\tau}^2] \quad \text{and} \quad QRP_t^g(\tau) \equiv \mathbb{E}_t^{\mathbb{P}}[g_{t,t+\tau}^2] - \mathbb{E}_t^{\mathbb{Q}}[g_{t,t+\tau}^2]. \tag{5}$$

[Equation \(5\)](#) shows that the loss QRP (QRP^l) represents the premium paid for the insurance against fluctuations in loss uncertainty, while the gain QRP (QRP^g) is the premium earned to compensate for the fluctuations in gain uncertainty. Thus, the (net) QRP ($QRP \equiv QRP^l - QRP^g$) represents the net cost of insuring fluctuations in loss uncertainty, that is, the premium paid for the insurance against fluctuations in loss uncertainty net of the premium earned to compensate for the fluctuations in gain uncertainty. Our study of the term structures of the risk-neutral and physical expected quadratic loss and gain naturally leads to examining the term structures of the loss and gain QRPs.

1.2 Decomposing the Quadratic Payoff into Loss and Gain: A Theory

For simplicity, let us denote the risk-neutral and physical expectations as the following:

$$\mu_n^{\mathbb{Q}+}(t, \tau) \equiv \mathbb{E}_t^{\mathbb{Q}}[g_{t,t+\tau}^n], \quad \mu_n^{\mathbb{Q}-}(t, \tau) \equiv \mathbb{E}_t^{\mathbb{Q}}[l_{t,t+\tau}^n], \quad \text{and} \quad \mu_n^{\mathbb{Q}}(t, \tau) \equiv \mathbb{E}_t^{\mathbb{Q}}[r_{t,t+\tau}^n], \tag{6}$$

$$\mu_n^{\mathbb{P}+}(t, \tau) \equiv \mathbb{E}_t^{\mathbb{P}}[g_{t,t+\tau}^n], \quad \mu_n^{\mathbb{P}-}(t, \tau) \equiv \mathbb{E}_t^{\mathbb{P}}[l_{t,t+\tau}^n], \quad \text{and} \quad \mu_n^{\mathbb{P}}(t, \tau) \equiv \mathbb{E}_t^{\mathbb{P}}[r_{t,t+\tau}^n]. \tag{7}$$

To understand the difference between $\mu_2^{Q+}(t, \tau)$ and $\mu_2^{Q-}(t, \tau)$, we follow Proposition 2 of Duffie, Pan, and Singleton (2000),

$$\mu_2^{Q-}(t, \tau) = \frac{\mathbb{E}_t^Q[r_{t,t+\tau}^2] + \Lambda^Q(t, \tau)}{2}, \quad \mu_2^{Q+}(t, \tau) = \frac{\mathbb{E}_t^Q[r_{t,t+\tau}^2] - \Lambda^Q(t, \tau)}{2}, \tag{8}$$

where $\Lambda^Q(t, \tau)$, the wedge between the risk-neutral expected quadratic loss and gain, is given by

$$\Lambda^Q(t, \tau) = \frac{2}{\pi} \int_0^{+\infty} \frac{\text{Im}(\varphi_{t,\tau}^{(2)}(-iv))}{v} dv, \tag{9}$$

with $\varphi_{t,\tau}(\cdot)$ being the time- t conditional risk-neutral moment-generating function of $r_{t,t+\tau}$ and $\varphi_{t,\tau}^{(2)}(\cdot)$ its second-order derivative, and $\text{Im}(\cdot)$ refers to the imaginary coefficient of a complex number. From Equation (8), it is apparent that studying the term structure of $\mu_2^{Q-}(t, \tau)$ and $\mu_2^{Q+}(t, \tau)$ amounts to studying the term structure of the quadratic payoff $\mathbb{E}_t^Q[r_{t,t+\tau}^2]$ and the term structure of $\Lambda^Q(t, \tau)$. Several papers in the literature have already dealt successfully with $\mathbb{E}_t^Q[r_{t,t+\tau}^2]$, and the consensus seems to be that a two-factor diffusion model provides a good statistical representation (see Christoffersen, Heston, and Jacobs, 2009). We now try to understand conceptually the potential drivers of the wedge $\Lambda^Q(t, \tau)$.

We use the following power series expansion of the moment-generating function $\varphi_{t,\tau}(\cdot)$,

$$\varphi_{t,\tau}(v) = \sum_{n=0}^{\infty} \frac{v^n}{n!} \mu_n^Q(t, \tau),$$

to establish that

$$\Lambda^Q(t, \tau) = \lim_{v \rightarrow \infty} \left\{ \sum_{j=1}^{\infty} \frac{(-1)^j v^{2j-1}}{(2j-1)(2j-1)!} \mu_{2j+1}^Q(t, \tau) \right\}, \tag{10}$$

which is a weighted average of odd high-order noncentral moments. Since only the odd high moments are included, the wedge $\Lambda^Q(t, \tau)$ is closely related to the asymmetry in the distribution of $r_{t,t+\tau}$. In the summation, when focusing on $j = 1$, it is apparent that $\Lambda^Q(t, \tau)$ is the opposite of the third-order noncentral moment $\mu_3^Q(t, \tau)$ (up to a positive multiplicative constant). Recall that $\mu_3^Q(t, \tau)$ is related to the first three central moments as follows:

$$\mu_3^Q(t, \tau) = \kappa_3^Q(t, \tau) + 3\mu_1^Q(t, \tau)\kappa_2^Q(t, \tau) + [\mu_1^Q(t, \tau)]^3,$$

where $\kappa_n^Q(t, \tau) \equiv \mathbb{E}_t^Q[(r_{t,t+\tau} - \mu_1^Q(t, \tau))^n]$. Hence, we conclude that the wedge between the risk-neutral expected quadratic loss and gain increases with the asymmetry in the risk-neutral distribution. A negative skewness implies larger risk-neutral expected quadratic losses, while a positive skewness yields the opposite effect. The wedge between the risk-neutral expected quadratic loss and gain still exists and is always negative when the distribution is symmetric (all odd-order central moments for a symmetric distribution are zero). In that case, the wedge increases in absolute value as the volatility increases.

1.3 Constructing Expectations

1.3.1 Inferring the risk-neutral expectation from option prices

In practice, previous literature estimates the risk-neutral conditional expectation of quadratic payoff directly from a cross-section of option prices. [Bakshi, Kapadia, and Madan \(2003\)](#) provide model-free formulas linking the risk-neutral moments of stock returns to explicit portfolios of options. These formulas are based on the basic notion, first presented in [Bakshi and Madan \(2000\)](#), that any payoff over a time horizon can be spanned by a set of options with different strikes with the same maturity as the investment horizon.

We adopt the notation in [Bakshi, Kapadia, and Madan \(2003\)](#) and define $V_t(\tau)$ as the time- t price of the τ -maturity quadratic payoff on the underlying stock. [Bakshi, Kapadia, and Madan \(2003\)](#) show that $V_t(\tau)$ can be recovered from the market prices of out-of-the-money (OTM) call and put options as follows:

$$V_t(\tau) = \int_{S_t}^{\infty} \frac{1 - \ln(K/S_t)}{K^2/2} C_t(\tau; K) dK + \int_0^{S_t} \frac{1 + \ln(S_t/K)}{K^2/2} P_t(\tau; K) dK, \quad (11)$$

where S_t is the time- t price of the underlying stock, and $C_t(\tau; K)$ and $P_t(\tau; K)$ are time- t option prices with maturity τ and strike K , respectively. The risk-neutral expected quadratic payoff is then

$$\mathbb{E}_t^{\mathbb{Q}} \left[r_{t,t+\tau}^2 \right] = e^{r_f \tau} V_t(\tau), \quad (12)$$

where r_f is the continuously compounded interest rate.

We compute $V_t(\tau)$ on each day and maturity. In theory, computing $V_t(\tau)$ requires a continuum of strike prices, while in practice we only observe a discrete and finite set of them. Following [Jiang and Tian \(2005\)](#) and others, we discretize the integrals in [Equation \(11\)](#) by setting up a total of 1001 grid points in the moneyness (K/S_t) range from 1/3 to 3. First, we use cubic splines to interpolate the implied volatility inside the available moneyness range. Second, we extrapolate the implied volatility using the boundary values to fill the rest of the grid points. Third, we calculate option prices from these 1001 implied volatilities using the Black-Scholes formula proposed by [Black and Scholes \(1973\)](#).² Next, we compute $V_t(\tau)$ if there are four or more OTM option implied volatilities (see, e.g., [Conrad, Dittmar, and Ghysels, 2013](#), and others). Finally, to obtain $V_t(30)$ on a given day, we interpolate and extrapolate $V_t(\tau)$ with different τ . This process yields a daily time series of the risk-neutral expected quadratic payoff for each maturity $\tau = 30, 60, \dots, 360$ days.

Note that the price of the quadratic payoff $V_t(\tau)$ in [Equation \(11\)](#) is the sum of a portfolio of OTM call options and a portfolio of OTM put options:

$$V_t(\tau) = V_t^g(\tau) + V_t^l(\tau), \quad (13)$$

where

$$V_t^l(\tau) = \int_0^{S_t} \frac{1 + \ln(S_t/K)}{K^2/2} P_t(\tau; K) dK \quad \text{and} \quad V_t^g(\tau) = \int_{S_t}^{\infty} \frac{1 - \ln(K/S_t)}{K^2/2} C_t(\tau; K) dK. \quad (14)$$

[Feunou et al. \(2019\)](#) analytically prove that $V_t^l(\tau)$ is the price of the quadratic loss and $V_t^g(\tau)$ is the price of the quadratic gain. We present that proof in Section A.6 of the [Online](#)

2 Since S&P 500 options are European, we do not have issues with the early exercise premium.

Appendix accompanying this article. Hence, the risk-neutral expectations of quadratic loss and gain are as follows:

$$\mathbb{E}_t^Q \left[l_{t,t+\tau}^2 \right] = e^{r_t \tau} V_t^l(\tau) \quad \text{and} \quad \mathbb{E}_t^Q \left[g_{t,t+\tau}^2 \right] = e^{r_t \tau} V_t^g(\tau). \tag{15}$$

1.3.2 Estimating the physical conditional expected quadratic payoff

A regression model can be used to estimate the expectations of the quadratic payoff and truncated returns over different periods using actual returns data. To compute these expectations, we assume that, conditional on time- t information, log returns $r_{t,t+\tau}$ have a normal distribution with mean $\mu_{t,\tau} = \mathbb{E}_t[r_{t,t+\tau}] = Z_t^\top \beta_\mu$ and variance $\sigma_{t,\tau}^2 = \mathbb{E}_t[RV_{t,t+\tau}]$, where $\mathbb{E}_t[RV_{t,t+\tau}] = Z_t^\top \beta_\sigma$ and $RV_{t,t+\tau}$ is the realized variance between end of day t and end of day $t + \tau$. We then have

$$\mathbb{E}_t \left[r_{t,t+\tau}^2 \right] = \mu_{t,\tau}^2 + \sigma_{t,\tau}^2 \quad \text{and} \quad \begin{cases} \mathbb{E}_t \left[l_{t,t+\tau}^2 \right] &= \left(\mu_{t,\tau}^2 + \sigma_{t,\tau}^2 \right) \Phi \left(-\frac{\mu_{t,\tau}}{\sigma_{t,\tau}} \right) - \mu_{t,\tau} \sigma_{t,\tau} \phi \left(\frac{\mu_{t,\tau}}{\sigma_{t,\tau}} \right) \\ \mathbb{E}_t \left[g_{t,t+\tau}^2 \right] &= \left(\mu_{t,\tau}^2 + \sigma_{t,\tau}^2 \right) \Phi \left(\frac{\mu_{t,\tau}}{\sigma_{t,\tau}} \right) + \mu_{t,\tau} \sigma_{t,\tau} \phi \left(\frac{\mu_{t,\tau}}{\sigma_{t,\tau}} \right), \end{cases} \tag{16}$$

under the log-normality assumption, where $\Phi(\cdot)$ and $\phi(\cdot)$ are the standard normal cumulative distribution functions and density, respectively.

The first part of (16) implies that the difference between the physical expected quadratic payoff $\mathbb{E}_t[r_{t,t+\tau}^2]$ and the physical expected realized variance $\sigma_{t,\tau}^2$ is due to the nonzero drift ($\mu_{t,\tau} \neq 0$). If the drift equals zero, the expected quadratic payoff is exactly the same as the expected realized variance. In Figure A7 of the [Online Appendix](#), we plot the average term structure of the drift ($\mu_{t,\tau}$) and the squared drift ($\mu_{t,\tau}^2$). As expected, the term structure of the drift is essentially flat. However, the term structure of the squared drift is increasing in the investment horizon, which explains the increasing discrepancy between the expected quadratic payoff and the expected realized variance at longer horizons. We also report descriptive statistics of both quantities in [Table A4](#) of the [Online Appendix](#). We find that both the drift and the squared drift of any investment horizon exhibit significant variations over time. Overall, these results highlight the importance of studying the term structure of the quadratic payoff.

The second part of [Equation \(16\)](#) implies that the wedge between the physical expected quadratic gain and loss is a function of the drift $\mu_{t,\tau}$. If the drift equals zero, then estimates of the physical expected quadratic gain and loss are equal. This is clearly a consequence of the normality assumption. In general, the wedge between the physical expected quadratic gain and loss could also be driven by other factors, for example, skewness. In the [Online Appendix](#) (Section A.10), we relax the normality assumption and instead assume that log returns $r_{t,t+\tau}$ follow a binormal distribution. This is an analytically tractable distribution that accommodates empirically plausible values of skewness and kurtosis, and nests the familiar Gaussian distribution (for details, see [Feunou, Jahan-Parvar, and Tédongap, 2013](#)). In [Table A5](#) of the [Online Appendix](#), we report the correlations between the physical expected quadratic loss (gain) estimated under the assumption of normal and binormal distributed log returns for a 1- to 12-month investment horizon. We find that correlations between these two physical expected quadratic loss (gain) quantities range between 0.984

(0.980) and 0.996 (0.993). Further, in Table A6 of the Online Appendix, we find that the average term structures of these quantities have consistent patterns and similar values. Taken together, this evidence suggests that our main results are robust to the distributional assumption of returns.

An estimate of $\mu_{t,\tau}$ is obtained as the fitted value from a linear regression of returns onto the vector of predictors, while an estimate of $\sigma_{t,\tau}^2$ is obtained as the fitted value from a linear regression of the total realized variance onto the same predictors. Those estimates are further plugged into the formulas in Equation (16) to obtain estimates of the physical expectations of the squared returns and truncated returns.

The specification of predictors in Z has been documented in a long list of previous literature. It is now widely accepted that models based on high-frequency realized variance dominate standard GARCH-type models (e.g., Chen and Ghysels, 2011), and thus, we follow this literature. Bekaert and Hoerova (2014) examine state-of-the-art models in the literature and consider the most general specification, where Z is a combination of a forward-looking volatility measure, the continuous variations, and the jump variations and negative returns in the last day, last week, or last month:

$$\begin{aligned} RV_{t,t+\tau} = & c + \alpha VIX_t^2 + \beta^m C_{t-21,t} + \beta^w C_{t-5,t} + \beta^d C_t \\ & + \gamma^m J_{t-21,t} + \gamma^w J_{t-5,t} + \gamma^d J_t \\ & + \delta^m I_{t-21,t} + \delta^w I_{t-5,t} + \delta^d I_{t-1,t} + \varepsilon_{t+\tau}^{(\tau)}, \end{aligned} \quad (17)$$

where C_t and J_t are respectively continuous and discontinuous components of the daily realized variance RV_t , $C_{t-b,t}$ and $J_{t-b,t}$ respectively aggregate C_{t-j} and J_{t-j} for $j = 0, 1, \dots, b-1$ (i.e., over a horizon b), and $I_{t-b,t}$ is the loss component of the return from day $t-b$ to day t . The conditional variance $\sigma_{t,\tau}^2$ is the fitted time series from the regression (17), for values of $\tau = 21, 42, \dots, 252$ days. Likewise, the conditional mean $\mu_{t,\tau}$ is the fitted time series from the regression (17) where the left-hand side is replaced by the τ -period log returns $r_{t,t+\tau}$.

Unlike the log returns and the realized variance, which are closed to temporal aggregation, the quadratic payoff and its loss and gain components are not. This suggests that the term structure of physical expectations of the quadratic payoff and its components are unlikely to be a flat line unless the mean $\mu_{t,\tau}$ is negligible for all considered horizons.

1.4 Data

1.4.1 Option data

For the estimation of the S&P 500 risk-neutral quadratic payoff, we rely on S&P 500 option prices obtained from the IvyDB OptionMetrics database for the January 1996 to December 2015 period. We exclude options with missing bid-ask prices, missing implied volatility, zero bids, negative bid-ask spreads, and options with zero open interest (see, e.g., Carr and Wu, 2009). Following Bakshi, Kapadia, and Madan (2003), we restrict the sample to out-of-the-money options. We further remove options with moneyness lower than 0.2 or higher than 1.8. To ensure that our results are not driven by misleading prices, we follow Conrad, Dittmar, and Ghysels (2013) and exclude options that do not satisfy the usual option price bounds, for example, call options with a price higher than the underlying price and options with less than seven days to maturity.

1.4.2 Return data

To construct the physical realized variance and perform volatility forecasts, we obtain intradaily S&P 500 cash index data spanning the period from January 1990 to December 2015 from TickData.com, for a total of 6542 trading days. On a given day, we use the last record in each five-minute interval to build a grid of five-minute equity index log returns. Following Andersen et al. (2001, 2003) and Barndorff-Nielsen, Kinnebrock, and Shephard (2010), we construct the realized variance on any given trading day t , where $r_{j,t}$ is the j th five-minute log return and n_t is the number of (five-minute) intradaily returns recorded on that day.³ We add the squared overnight log return to the realized variance. The realized variances between day t and $t + \tau$ are computed by accumulating the daily realized variances.

1.5 Preliminary Analysis

1.5.1 The term structure of the risk-neutral expected quadratic payoff

To estimate $\mathbb{E}_t^{\mathbb{Q}}[I_{t,t+\tau}^2]$ or $\mathbb{E}_t^{\mathbb{Q}}[g_{t,t+\tau}^2]$ for each maturity τ , we use options with maturity close to τ and do interpolations.⁴ In the top left panel of Figure 1, we plot the time-series average of the risk-neutral expected quadratic payoff and its loss and gain components for maturities of 1, 3, 6, 9, and 12 months. We find that the average term structures of the risk-neutral expected quadratic payoff and its loss component are, in general, upward sloping. On the other hand, we find that the term structure of the risk-neutral expected gain quadratic payoff is flat.

To investigate time variations in these term structures, in the two middle panels of Figure 1, we plot the 6-month maturity (the level) and the 12-month minus 2-month maturity (the slope) for the risk-neutral expected quadratic payoff and its components, respectively. We find that both the level and the slope display important time variations and have spikes and troughs during crises. We also notice that, although the slopes are mostly positive, they are negative during crises. These observed patterns are in line with the fact that during crises, investors expect a recovery in the long run rather than in the short run.

Relationship to Dew-Becker et al. (2017). Compute the term structure of forward variance prices as $F_t^{r\nu;\tau} \equiv \mathbb{E}_t^{\mathbb{Q}}[RV_{t+\tau-1,t+\tau}]$. The forward variance price $F_t^{r\nu;\tau}$ is essentially the month- t risk-neutral expectation of the realized variance from end of month $t + \tau - 1$ to end of month $t + \tau$. The authors compute these forward prices using traded variance swaps. Since traded loss and gain variance swaps do not exist, one cannot estimate risk-neutral expectations of the loss and gain components using their approach. Previous literature (see, e.g., Kilic and Shaliastovich, 2019) has shown that loss and gain components are important for asset prices. Our methodology allows us to compute the term structure of forward prices not only for the quadratic payoff but also for its loss and gain components as $F_t^{r;\tau} \equiv \mathbb{E}_t^{\mathbb{Q}}[I_{t,t+\tau-1,t+\tau}^2]$, $F_t^{l;\tau} \equiv \mathbb{E}_t^{\mathbb{Q}}[I_{t,t+\tau-1,t+\tau}^l]$, and $F_t^{g;\tau} \equiv \mathbb{E}_t^{\mathbb{Q}}[I_{t,t+\tau-1,t+\tau}^g]$, respectively.

3 On a typical trading day, we observe $n_t = 78$ five-minute returns.

4 In the data, we do not always observe options with the exact maturity τ . In order to find $\mathbb{E}_t^{\mathbb{Q}}[I_{t,t+\tau}^2]$ or $\mathbb{E}_t^{\mathbb{Q}}[g_{t,t+\tau}^2]$ at the exact maturity τ , we either interpolate or extrapolate to find the exact value. For example, if we wish to find $\mathbb{E}_t^{\mathbb{Q}}[I_{t,t+30}^2]$ (i.e., with maturity $\tau = 30$ days), we interpolate between $\mathbb{E}_t^{\mathbb{Q}}[I_{t,t+\tau_1}^2]$ and $\mathbb{E}_t^{\mathbb{Q}}[I_{t,t+\tau_2}^2]$ to obtain $\mathbb{E}_t^{\mathbb{Q}}[I_{t,t+30}^2]$, where τ_1 is the closest observed maturity below 30 days and τ_2 is the closest observed maturity over 30 days. In cases where we do not observe τ_2 in the data, we extrapolate τ_1 to obtain the exact maturity.

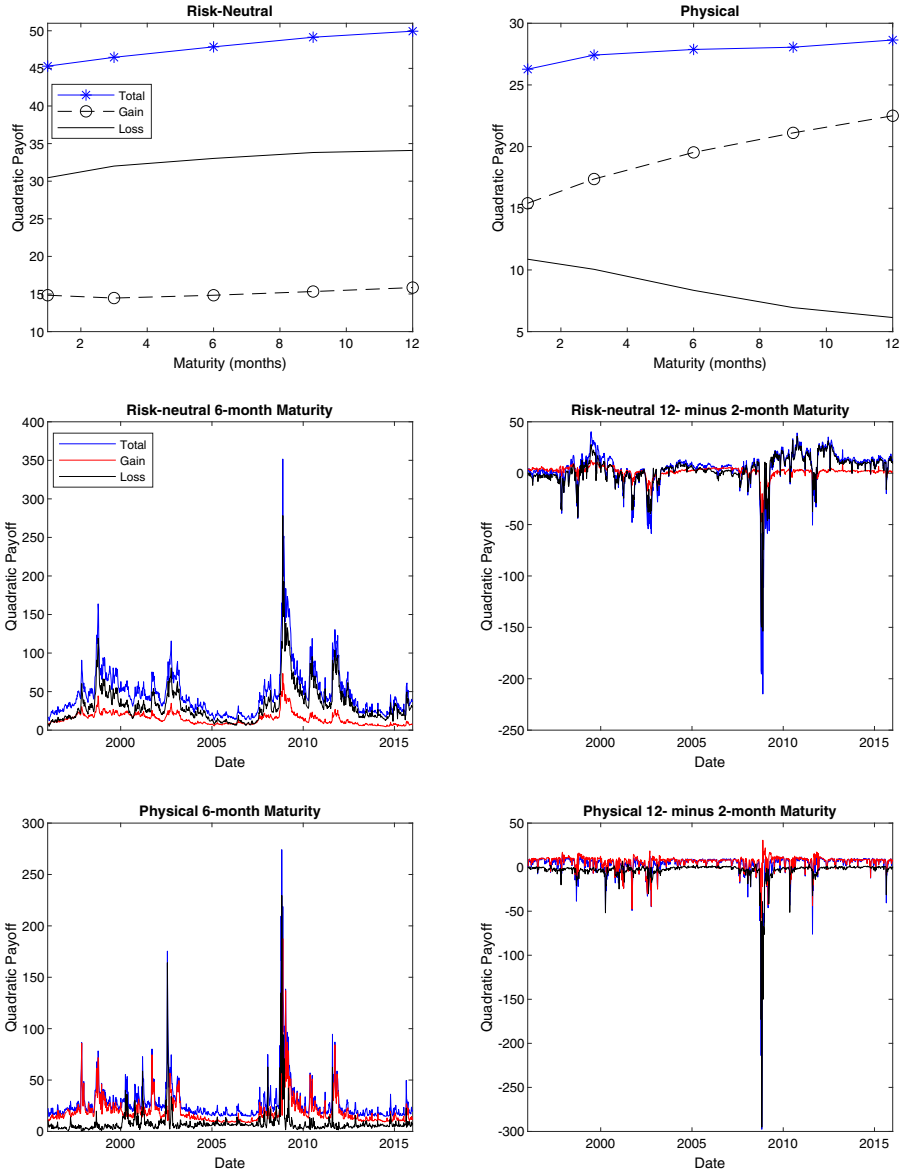


Figure 1 Term Structure of Expected Quadratic Payoff. In this figure, in the top two panels we plot the average S&P 500 expected quadratic payoff and its loss and gain components for maturities of 1, 3, 6, 9, and 12 months. The top left is for the risk-neutral quantities, while the top right is for the physical quantities. In the middle panels, we plot the level (6-month maturity) and the slope (12-month minus the 2-month maturity) of the risk-neutral expected quadratic payoff and its loss and gain components. In the bottom panels, we plot the physical level and slope. All reported values are monthly and in squared percentage units. The sample period is from January 1996 to December 2015.

In Panel A of [Figure A1](#) in the [Online Appendix](#), we plot the term structure of average forward prices for the risk-neutral expected quadratic payoff and its loss and gain components. We find that the estimated term structure of forward prices for the quadratic payoff is concave, with both the level and slope similar to the forward prices in [Dew-Becker et al. \(2017\)](#). This finding is despite the differences in sample periods and the fact that our forward prices are computed for the quadratic payoff while their forward prices are computed from traded variance swaps. Most importantly, as mentioned above, in contrast to [Dew-Becker et al. \(2017\)](#) we are able to separately estimate the forward prices for the quadratic loss and gain, while loss and gain variance swaps do not exist. This approach allows us to investigate not only the term structure of the quadratic payoff but also its components. We see that the term structures of the average forward prices for the quadratic loss and gain are in general upward sloping. We also find that, across all horizons, the quadratic loss forward prices are higher than the quadratic gain forward prices.

1.5.2 The term structure of the physical expected quadratic payoff

In the top right panel of [Figure 1](#), we plot the time-series average of the term structure of the physical expected quadratic payoff and its loss and gain components for maturities of 1, 3, 6, 9, and 12 months. We find that the term structure of the expected quadratic loss is downward sloping. Since the quadratic loss is a measure of loss uncertainty, this finding suggests that investors face more uncertainty about losses in the short run relative to the long run. On the other hand, we find that the term structure of the expected quadratic gain is upward sloping. Since the expected quadratic gain is a measure of the gain uncertainty, this finding suggests that investors face more uncertainty about gains in the long run relative to the short run. Comparing their term structures, we observe that the level of the expected quadratic gain dominates the level of the expected quadratic loss across all horizons, and even more so in the long run, leading to the upward-sloping pattern in the total expected quadratic payoff. The relatively larger values of the expected quadratic gain are consistent with the fact that the S&P 500 cash index has historically yielded a positive annual return of 7%.

To evaluate the time variation in these term structures, in the two bottom panels of [Figure 1](#), we plot the 6-month maturity (the level) and the 12-month minus 2-month maturity (the slope) for the expected quadratic payoff and its components, respectively. As for its risk-neutral counterpart, we find substantial variations in both the level and the slope. We find that the expected quadratic payoff and expected quadratic gain have in general positive and occasionally negative slopes. On the other hand, the expected quadratic loss slopes are almost always negative and very negative during the 2008 financial crisis. These observed patterns are in line with the fact that investors expect a growth opportunity in the long run rather than in the short run.

Further, in the top two panels of [Figure 1](#), we observe a common upward-sloping pattern for the term structures of the risk-neutral and physical expected quadratic payoff and its components. The only exception is the physical expected quadratic loss, which is downward sloping. [Bakshi, Kapadia, and Madan \(2003\)](#) show that under certain conditions, the risk-neutral distribution can be obtained by exponentially tilting the real-world density, with the tilt determined by the risk aversion of investors. This means that the observed upward-sloping risk-neutral expected quadratic loss relative to the downward-sloping term

structure of the physical quadratic loss may be explained by investors' increasing risk aversion as the investment horizon increases.

In [Table 1](#), we present the time-series means of the risk-neutral and physical expected quadratic payoff together with their loss and gain components. For each mean we also report, in parentheses, [Newey and West \(1987\)](#) adjusted standard errors. The mean values for the risk-neutral expected quadratic payoff increase as the maturity horizon increases from 45.30 at 1 month to 49.94 at 12 months in monthly percentage squared units. The mean values for the physical expected quadratic payoff are much lower but also increase as the maturity horizon increases from 26.28 at 1 month to 28.64 at 12 months. The mean values for the risk-neutral expected quadratic loss are much higher than the values for the risk-neutral expected quadratic gain for any given horizon, and the wedge is the same for different maturity horizons. For example, at 2 months, the risk-neutral expected quadratic loss is 31.16 and the risk-neutral expected quadratic gain is 14.39; the wedge is about 17, which is similar to the wedge at 4, 6, 8, and 12 months. However, the physical expected quadratic loss is much lower than the physical expected quadratic gain for any given horizon, and this wedge is increasing as the horizon increases. For example, at 2 months, the physical expected quadratic loss is 10.05 and the physical expected quadratic gain is 16.45; the wedge is about 6, and this wedge is strictly increasing to roughly 16–12 months. We also see that the standard errors of the means for all these quantities are decreasing in the maturity. Finally, we find that all the means are statistically different from zero.

Rolling-Window Parameter Estimation. We use the full sample to estimate the drifts in returns and the expectations of realized variances over different periods. These estimated quantities enter [Equation \(16\)](#) to compute the expectations of the truncated returns. Such an approach may incur forward-looking bias in the parameter estimation stage. To check the robustness of our results, we estimate drifts and expected realized variances using a set of different rolling windows: 60, 72, 84, 96, 108, and 120 months of daily data. Detailed comparisons to the full-sample expected quadratic gain and loss are discussed in the [Online Appendix](#). We plot the correlation between the rolling-window and full-sample expected quadratic gain in [Figure A9](#) of the [Online Appendix](#), which reveals that all the correlations are well above 0.7. Except for the expected quadratic gain using a 60-month window size, all other correlations are higher than 0.77. The correlation can be as high as 0.88 for the expected quadratic loss based on a 120-month rolling window.

To further compare the full-sample and rolling-window expected truncated returns, we fix the rolling window size to 120 months, and for each investment horizon τ , we estimate time-series regressions of the following form:

$$\begin{aligned}\mathbb{E}\left[g_{FS,\tau}^2\right] &= \alpha + \beta_{roll}^g \mathbb{E}\left[g_{roll,\tau}^2\right] + \varepsilon \\ \mathbb{E}\left[l_{FS,\tau}^2\right] &= \alpha + \beta_{roll}^l \mathbb{E}\left[l_{roll,\tau}^2\right] + \varepsilon\end{aligned}\quad (18)$$

If these expectations constructed from rolling-window parameters are exactly the same as the ones constructed from full-sample parameters, we should find that β_{roll}^g and β_{roll}^l are not statistically significantly different from one. In [Figure A10](#) of the [Online Appendix](#), we plot these coefficients for investment horizons from 1 to 12 months. Panel A shows that β_{roll}^g is not statistically different from one over horizons of 6, 7, and 8 months. On the other

Table 1 Descriptive statistics

Mean												
Maturity	1	2	3	4	5	6	7	8	9	10	11	12
$\mathbb{E}[r^2]$	26.28 (2.49)	26.49 (2.07)	27.42 (2.07)	28.20 (2.01)	27.97 (1.66)	27.88 (1.45)	27.97 (1.34)	28.06 (1.25)	28.06 (1.16)	28.16 (1.09)	28.38 (1.04)	28.64 (1.00)
$\mathbb{E}[l^2]$	10.87 (1.58)	10.05 (1.30)	10.05 (1.54)	9.95 (1.54)	9.09 (1.13)	8.35 (0.83)	7.90 (0.73)	7.43 (0.66)	6.95 (0.55)	6.56 (0.47)	6.32 (0.44)	6.14 (0.42)
$\mathbb{E}[g^2]$	15.41 (1.07)	16.45 (0.95)	17.37 (0.87)	18.25 (0.88)	18.88 (0.89)	19.53 (0.91)	20.07 (0.89)	20.63 (0.86)	21.11 (0.84)	21.59 (0.82)	22.06 (0.80)	22.50 (0.78)
$\mathbb{E}^{\mathbb{P}}[r^2]$	45.30 (3.53)	45.55 (3.03)	46.47 (2.92)	47.07 (2.66)	47.39 (2.43)	47.86 (2.38)	48.40 (2.34)	48.81 (2.32)	49.14 (2.29)	49.33 (2.26)	49.40 (2.19)	49.94 (2.18)
$\mathbb{E}^{\mathbb{P}}[l^2]$	30.46 (2.58)	31.16 (2.26)	32.01 (2.24)	32.50 (2.04)	32.71 (1.85)	33.03 (1.84)	33.39 (1.84)	33.65 (1.84)	33.81 (1.85)	33.83 (1.83)	33.74 (1.77)	34.09 (1.78)
$\mathbb{E}^{\mathbb{P}}[g^2]$	14.84 (0.98)	14.39 (0.80)	14.46 (0.73)	14.57 (0.68)	14.68 (0.64)	14.84 (0.61)	15.01 (0.59)	15.16 (0.58)	15.33 (0.57)	15.50 (0.56)	15.66 (0.56)	15.85 (0.55)
QRP	19.04 (1.35)	19.07 (1.43)	19.07 (1.43)	18.89 (1.36)	19.43 (1.41)	20.01 (1.44)	20.44 (1.41)	20.77 (1.43)	21.10 (1.46)	21.19 (1.45)	21.04 (1.43)	21.32 (1.44)
QRP^l	19.61 (1.42)	21.13 (1.54)	21.98 (1.63)	22.57 (1.65)	23.64 (1.69)	24.70 (1.71)	25.50 (1.69)	26.23 (1.71)	26.89 (1.74)	27.29 (1.73)	27.43 (1.69)	27.97 (1.70)
QRP^g	0.56 (0.33)	2.05 (0.37)	2.91 (0.40)	3.68 (0.46)	4.20 (0.44)	4.69 (0.46)	5.06 (0.47)	5.46 (0.47)	5.78 (0.47)	6.10 (0.47)	6.40 (0.47)	6.65 (0.48)

Notes: In this table, we report the time-series mean for all maturities ranging from 1 to 12 months for a set of variables that includes the risk-neutral expected quadratic payoff ($\mathbb{E}^{\mathbb{P}}[r^2]$, $\mathbb{E}^{\mathbb{P}}[l^2]$, $\mathbb{E}^{\mathbb{P}}[g^2]$), the expected quadratic payoff ($\mathbb{E}[r^2]$, $\mathbb{E}[l^2]$, $\mathbb{E}[g^2]$), and the QRP (QRP , QRP^l , QRP^g). Below each mean, in parentheses, we also report the Newey and West (1987) standard error. All reported statistics are monthly squared percentage values. The sample period is from January 1996 to December 2015.

hand, we find in Panel B that β_{roll}^l is not statistically different from one over a 3-month horizon. Overall, because of the difficulty predicting returns over short horizons, the quantities constructed from the full-sample parameters inevitably differ from the rolling-window counterparts.

1.5.3 The term structure of the QRP

Next, we turn to study the term structure of the QRP. In Table 1, we also present the time-series means of the QRP and its loss and gain components. On average, the QRP is positive, equal to 19.04, 20.01, and 21.32 at 1, 6, and 12 months, respectively. Both the loss QRP and the gain QRP are positive. However, the loss QRP is dominating the gain QRP at all horizons. For example, QRP^l is 21.98 while QRP^g is 2.91 at 3 months; QRP^l is 26.89 while QRP^g is 5.78 at 9 months. The average QRP^g is small at the 1-month horizon and not statistically different from zero. In general, we observe that the standard error of the average QRP (and its components), which represents the insurance cost (against downside risk, upside risk, or the net cost of hedging downside risk), increases with the horizon. Nevertheless, apart from the 1-month average QRP^g , we find that all means are significantly different from zero.

Table 2 Principal component analysis

Principal component	1	2	3	First 3
Information sets			Explanatory power (%)	
$\mathbb{E}[l^2]$ and $\mathbb{E}[g^2]$	58.01	37.10	3.13	98.24
$\mathbb{E}^{\mathbb{P}}[l^2]$ and $\mathbb{E}^{\mathbb{P}}[g^2]$	87.33	8.78	2.76	98.87
QRP^l and QRP^g	73.05	15.22	6.18	94.45
$\mathbb{E}[l^2]$, $\mathbb{E}[g^2]$, QRP^l and QRP^g	56.73	26.73	7.93	91.39

Notes: In this table, we report in percentage the explanatory power of each of the first three principal components, and their total explanatory power, for a number of different information sets. These include the term structure of the loss and gain components of the physical expected quadratic payoff, the risk-neutral expected quadratic payoff, and the QRP, each separately. We also report the explanatory power of the first three principal components of the term structure of components of the physical expected quadratic payoff together with the term structure of the components of the QRP. The sample period is from January 1996 to December 2015.

1.5.4 Principal component analysis

In general, structural and reduced-form asset pricing models have a very tight factor structure, implying that different expectations (whether risk-neutral or physical) are all driven by a very low number of factors (e.g., in reduced-form option pricing models, the largest number of factors considered in the literature so far is three). Nevertheless, our analysis deals with the joint term structures of two uncertainty components (loss and gain) under two different probability measures (\mathbb{Q} and \mathbb{P}). To pin down the number of factors observed in the data, we run a principal component analysis of the term structure of four quantities: the loss and gain components of the physical and risk-neutral expected quadratic payoff. Alternatively, one can choose to use the loss and gain components of the physical expected quadratic payoff and the QRP or the loss and gain components of the risk-neutral expected quadratic payoff and the QRP. There is no difference between these three choices.

Table 2 shows the explanatory powers of the first three principal components. We find that the first three principal components are enough to explain 91.39% of the variation in the term structure of the loss and gain physical expected quadratic payoff and the QRP (there are forty-eight variables because we include four quantities with twelve maturities). The first principal component explains 56.73%, the second explains 26.73%, and the third explains 7.93% of the variations. The immediate implication of these findings is that any model (whether reduced-form or structural) that aims to jointly fit these various term structures should include at least three factors.

2 A Model for the Joint Term Structure of Quadratic Loss and Gain

In search of a flexible reduced-form model to accommodate different kinds of distribution asymmetry and the term structure of $\mu_2^{\mathbb{Q}-}(t, \tau)$ and $\mu_2^{\mathbb{Q}+}(t, \tau)$, we study the recent model proposed by Andersen, Fusari, and Todorov (2015). This model is ideal for our analysis for three reasons. First, it is built to disentangle the dynamics of positive and negative jumps. Second, it is a three-factor framework, which would maximize the model's chances of fitting the term structure of the expected quadratic loss and gain and their risk premia since we find that three principal components are needed to fit the targeted term structures in Section 1.5.4. Third, since it is an affine model, it is tractable and enables us to compute all

the quantities of interest in closed form. In this section, we discuss the Andersen, Fusari, and Todorov (2015) model and some variants of this three-factor model. We use the two-factor diffusion model of Christoffersen, Heston, and Jacobs (2009) as the baseline model. Finally, we introduce a set of different specifications for the pricing kernel, including the baseline specification in which jumps are not priced.

2.1 Andersen, Fusari, and Todorov (2015)'s Risk-Neutral Specification

In the three-factor jump-diffusive stochastic volatility model of Andersen, Fusari, and Todorov (2015), the underlying asset price evolves according to the following general dynamics (under \mathbb{Q}):

$$\begin{aligned} \frac{dS_t}{S_{t-}} &= (r_{f,t} - \delta_t)dt + \sqrt{V_{1t}}dW_{1t}^{\mathbb{Q}} + \sqrt{V_{2t}}dW_{2t}^{\mathbb{Q}} + \eta\sqrt{V_{3t}}dW_{3t}^{\mathbb{Q}} + \int_{\mathbb{R}^2} (e^x - 1)\mu^{\mathbb{Q}}(dt, dx, dy) \\ dV_{1t} &= \kappa_1(\bar{v}_1 - V_{1t})dt + \sigma_1\sqrt{V_{1t}}dB_{1t}^{\mathbb{Q}} + \mu_1\int_{\mathbb{R}^2}x^2\mathbf{1}_{\{x < 0\}}\mu(dt, dx, dy) \\ dV_{2t} &= \kappa_2(\bar{v}_2 - V_{2t})dt + \sigma_2\sqrt{V_{2t}}dB_{2t}^{\mathbb{Q}} \\ dV_{3t} &= -\kappa_3V_{3t}dt + \mu_3\int_{\mathbb{R}^2}\left[(1 - \rho_3)x^2\mathbf{1}_{\{x < 0\}} + \rho_3y^2\right]\mu(dt, dx, dy), \end{aligned}$$

where $r_{f,t}$ and δ_t refer to the instantaneous risk-free rate and the dividend yield, respectively; $(W_{1t}^{\mathbb{Q}}, W_{2t}^{\mathbb{Q}}, W_{3t}^{\mathbb{Q}}, B_{1t}^{\mathbb{Q}}, B_{2t}^{\mathbb{Q}})$ is a five-dimensional Brownian motion with $corr(W_{1t}^{\mathbb{Q}}, B_{1t}^{\mathbb{Q}}) = \rho_1$ and $corr(W_{2t}^{\mathbb{Q}}, B_{2t}^{\mathbb{Q}}) = \rho_2$ while the remaining Brownian motions are mutually independent; and $\mu^{\mathbb{Q}}(dt, dx, dy) \equiv \mu(dt, dx, dy) - \nu_t^{\mathbb{Q}}(dx, dy)dt$, where $\nu_t^{\mathbb{Q}}(dx, dy)$ is the risk-neutral compensator for the jump measure μ , and is assumed to be

$$\nu_t^{\mathbb{Q}}(dx, dy) = \left\{ \left(c_t^- \mathbf{1}_{\{x < 0\}} \lambda_- e^{-\lambda_- |x|} + c_t^+ \mathbf{1}_{\{x > 0\}} \lambda_+ e^{-\lambda_+ |x|} \right) \mathbf{1}_{\{y=0\}} \right\} \tag{19}$$

where time-varying negative and positive jumps are governed by distinct coefficients: c_t^- and c_t^+ , respectively. These coefficients evolve as affine functions of the state vectors

$$c_t^- = c_0^- + c_1^- V_{1t-} + c_2^- V_{2t-} + c_3^- V_{3t-}, \quad c_t^+ = c_0^+ + c_1^+ V_{1t-} + c_2^+ V_{2t-} + c_3^+ V_{3t-}.$$

These three factors have distinctive features: V_{2t} is a pure-diffusion process, V_{3t} is a pure-jump process, and innovation in V_{1t} combines a diffusion and a jump component. Furthermore, one of the key features of the AFT model is its ability to break the tight link between expected negative and positive jump variation imposed by other traditional jump diffusion models.

To better understand the ability of the key features of this general model to match the observed term structures of risk-neutral expected quadratic loss and gain, we focus on two dimensions. The first is the number of factors. Compared to the three-factor framework, we ask whether two factors are enough and which two-factor alternatives generate the best fit. The second dimension is the model's ability to differentiate between the negative and positive jump distribution. We ask whether the symmetric jump distribution can still fit the term structures. We label the unrestricted general model AFT4 and consider the following nested specifications:

- AFT0: There are no jumps. This corresponds to the two-factor diffusion model studied extensively in [Christoffersen, Heston, and Jacobs \(2009\)](#). This is equivalent to suppressing all the jump related components ($\eta = 0$ and $\mu_1 = 0$) and the third factor V_{3t} .
- AFT1: There is no pure-jump process. This corresponds to suppressing V_{3t} .
- AFT2: There is no pure-diffusion process. This corresponds to suppressing V_{2t} . In this model, both variance factors V_{1t} and V_{3t} jump, implying that it can be used to judge the benefit of having jumps in volatility, a subject of much debate in the option pricing literature.
- AFT3: The expected negative jump variation equals the expected positive jump variation. This corresponds to a three-factor model that assumes the same distribution for positive and negative jumps. It is equivalent to imposing that $\lambda_- = \lambda_+$ and $c_t^- = c_t^+$. The AFT3 is representative of most of the existing option pricing and variance swap models as it does not differentiate between the intensity of positive and negative jumps (see, e.g., [Bates, 2012](#), [Christoffersen, Jacobs, and Ornathanalai, 2012](#), [Eraker, 2004](#), [Chernov et al., 2003](#), [Huang and Wu, 2004](#), [Amengual and Xiu, 2018](#), and [Ait-Sahalia, Karaman, and Mancini, 2015](#)).

One interesting model variation is the three-factor model in which $\eta = 0$. This makes V_{3t} a pure-jump process in the sense that it only drives the jump intensity while not entering in the diffusive volatility.⁵ In the [Online Appendix](#), we compare two three-factor models in which $\eta = 0$ and $\eta \neq 0$ and find that $\eta \neq 0$ is important for accurate pricing of truncated second moments.

2.2 The Radon–Nikodym Derivative

In this article, our goal is to understand the statistical properties of the stock returns distribution that are essential to reproduce the observed term structures of $\mu_2^{\mathbb{Q}^+}(t, \tau)$, $\mu_2^{\mathbb{Q}^-}(t, \tau)$ and the stochastic discount factor specifications that are able to replicate the observed spreads $\mu_2^{\mathbb{Q}^-}(t, \tau) - \mu_2^{\mathbb{P}^-}(t, \tau)$ and $\mu_2^{\mathbb{Q}^+}(t, \tau) - \mu_2^{\mathbb{P}^+}(t, \tau)$. To do that, we need to specify a Radon–Nikodym derivative (the law of change of measure). We specify the most flexible Radon–Nikodym derivative preserving the same model structure under the physical dynamic. Our Radon–Nikodym derivative is the product of the two derivatives separately governing the compensation of continuous variations and jump variations:

$$\left(\frac{d\mathbb{Q}}{d\mathbb{P}}\right)_t = \left(\frac{d\mathbb{Q}}{d\mathbb{P}}\right)_t^c \left(\frac{d\mathbb{Q}}{d\mathbb{P}}\right)_t^j,$$

where

$$\left(\frac{d\mathbb{Q}}{d\mathbb{P}}\right)_t^c = \exp \left\{ \int_0^t \theta_s^\top dW_s^\mathbb{P} + \int_0^t \bar{\theta}_s^{v\top} d\bar{B}_s^\mathbb{P} - \frac{1}{2} \int_0^t \left(\theta_s^\top \theta_s + \bar{\theta}_s^{v\top} \bar{\theta}_s^v \right) ds \right\},$$

and

5 Several papers, including [Santa-Clara and Yan \(2010\)](#), [Christoffersen, Jacobs, and Ornathanalai \(2012\)](#), and [Andersen et al. \(2015\)](#), find evidence for a pure-jump component in the pricing of S&P 500 options.

$$\left(\frac{dQ}{dP}\right)_t^j = \mathcal{E}\left(\int_0^t \int_{R^2} \Psi_s(x, y) \mu^{\mathbb{P}}(ds, dx, dy)\right),$$

with \mathcal{E} referring to the stochastic exponential, $dW_{jt}^{\mathbb{P}} \equiv dW_{jt}^{\mathbb{Q}} + \theta_t^j(j)dt$, $d\bar{B}_{jt}^{\mathbb{P}} \equiv d\bar{B}_{jt}^{\mathbb{Q}} + \bar{\theta}_t^j(j)dt$, $dB_{jt}^{\mathbb{Q}} = \rho_j dW_{jt}^{\mathbb{Q}} + \sqrt{1 - \rho_j^2} d\bar{B}_{jt}^{\mathbb{Q}}$, and $\mu^{\mathbb{P}}(dt, dx, dy) = \mu(dt, dx, dy) - \nu_t^{\mathbb{P}}(dx, dy)dt$.

With the appropriate choice for the price of risk parameters θ_t^j , $\bar{\theta}_t^j$ and the physical compensator $\nu_t^{\mathbb{P}}(dx, dy)$, we can show that the resulting physical dynamic preserves the exact structure as the risk-neutral dynamic. In particular, the price of jump risk, $\Psi_t(x, y)$, is given by

$$\Psi_t(x, y) \equiv \frac{\nu_t^{\mathbb{Q}}(dx, dy)}{\nu_t^{\mathbb{P}}(dx, dy)} - 1 \quad \text{where} \quad \frac{\nu_t^{\mathbb{Q}}(dx, dy)}{\nu_t^{\mathbb{P}}(dx, dy)} = \begin{cases} \frac{c_t^+ \lambda_+}{c_t^{\mathbb{P}+} \lambda_+^{\mathbb{P}}} \exp\left(-(\lambda_+ - \lambda_+^{\mathbb{P}})x\right) & x > 0 \quad y = 0 \\ \frac{c_t^- \lambda_-}{c_t^{\mathbb{P}-} \lambda_-^{\mathbb{P}}} \exp\left((\lambda_- - \lambda_-^{\mathbb{P}})x\right) & x < 0 \quad y = 0 \\ \frac{c_t^- \lambda_-}{c_t^{\mathbb{P}-} \lambda_-^{\mathbb{P}}} \exp\left((\lambda_- - \lambda_-^{\mathbb{P}})y\right) & x = 0 \quad y < 0 \end{cases}$$

Is the premium inherent in hedging bad shocks substantially different from the one required to be exposed to good shocks? The evidence presented in Section 2 overwhelmingly points to two very different premia. Another more challenging question is whether we need to specify a maximum flexible pricing kernel where all the parameters for jump intensities are shifted from the \mathbb{Q} - to the \mathbb{P} -measure by different amounts, or whether there is a parsimonious specification (one that imposes more restrictions between the \mathbb{Q} - and \mathbb{P} -dynamics) that is able to simultaneously replicate the observed dynamics of the term structures of the loss and gain QRP. To shed light on these issues, in the estimation investigation we distinguish between the following restrictions on the Radon–Nikodym derivative (where the unrestricted specification is labeled RND4):

1. RND0: Jumps are not priced. This is equivalent to imposing $c_j^{\mathbb{P}+} = c_j^+$, $c_j^{\mathbb{P}-} = c_j^-$, for $j = 0, 1, 2, 3$ $\lambda_-^{\mathbb{P}} = \lambda_-$ and $\lambda_+^{\mathbb{P}} = \lambda_+$. Note that this is the equivalent of setting $\Psi_t(x, y) = 0$, or equivalently, $(\frac{dQ}{dP})_t^j = 1$.
2. RND1: The price of positive jumps equals the price of negative jumps, or more formally $\Psi_t(x, y)$ is independent of the sign of x . Note that this is the implicit restriction imposed by traditional affine jump diffusion option pricing models, for example, Eraker (2004), Santa-Clara and Yan (2010), Christoffersen, Jacobs, and Ornthanalai (2012), and Bates (2012).
3. RND2: Negative jumps are not priced $\iff \lambda_-^{\mathbb{P}} = \lambda_-$, $c_j^{\mathbb{P}-} = c_j^-$, for $j = 0, 1, 2, 3$.
4. RND3: Positive jumps are not priced $\iff \lambda_+^{\mathbb{P}} = \lambda_+$, $c_j^{\mathbb{P}+} = c_j^+$, for $j = 0, 1, 2, 3$.

3 Estimation

We largely rely on the recent paper by Feunou and Okou (2018), which proposes to estimate affine option pricing models using risk-neutral moments instead of raw option prices. Unlike option prices, cumulants (central moments) are linear functions of unobserved

factors. Hence, using cumulants enables us to circumvent major challenges usually encountered in the estimation of latent factor option pricing models.

Given that the AFT model is affine, the linear Kalman filter appears as a natural estimation technique. The AFT model can easily be casted in a (linear) state-space form where the measurement equations relate the observed or model-free risk-neutral cumulants to the latent factors (state variables) and the transition equations describe the dynamics of these factors. However, unlike the setup in Feunou and Okou (2018), we are mainly interested in the term structures of expected quadratic loss and gain, which turn out to be nonlinear functions of the factors. Hence, we will have two sets of measurement equations: (i) linear equations, which relate the risk-neutral variances and third-order cumulants to the factors; and (ii) nonlinear equations, which relate the risk-neutral expected quadratic loss and gain to the factors. We will use only the first set of measurement equations in the linear Kalman filtering step, and conditional on the filtered factors, we will compute the likelihood of the risk-neutral expected quadratic loss and gain.

3.1 Risk-Neutral Cumulants Likelihood

On a given day t , we stack together the n^{th} -order risk-neutral cumulant observed at distinct maturities in a vector denoted by $CUM_t^{(n)Q} = (CUM_{t,\tau_1}^{(n)Q}, \dots, CUM_{t,\tau_j}^{(n)Q})^\top$, where $n \in \{2, 3\}$. We further stack the second and third cumulant vector in $CUM_t^Q = (CUM_t^{(2)Q\top}, CUM_t^{(3)Q\top})^\top$ to build a $2J \times 1$ vector. This implies the following linear measurement equation:

$$CUM_t^Q = \Gamma_0^{cum} + \Gamma_1^{cum} V_t + \Omega_{cum}^{1/2} \vartheta_t^{cum}, \tag{20}$$

where the dimension of the unobserved state vector (V_t) is 3. Notably, Γ_0^{cum} and Γ_1^{cum} are $2J \times 1$ and $2J \times 3$ matrices of coefficients whose analytical expressions depend explicitly on Q-parameters as shown in Feunou and Okou (2018). The last term in Equation (20) is a vector of observation errors, where Ω_{cum} is a $2J \times 2J$ diagonal covariance matrix and ϑ_t^{cum} denotes a $2J \times 1$ vector of independent and identically distributed (i.i.d.) standard Gaussian disturbances.

As shown in Feunou and Okou (2018), the transition equation for the three factors in the AFT model is

$$V_{t+1} = \Phi_0 + \Phi_1 V_t + \Sigma(V_t)^{1/2} \varepsilon_{t+1}, \tag{21}$$

where Φ_0 , Φ_1 , and $\Sigma(V_t)$ are functions of model parameters. They are given in the Online Appendix (Section A12) to save space. The system (20)–(21) gives the state-space representation of the AFT model. The marginal moments (mean and variance) of the latent vector are used to initialize the filter by setting $V_{0|0} = -(K_1^P)^{-1} K_0^P$ and $vec(P_{0|0}) = (I_9 - \Phi_1 \otimes \Phi_1)^{-1} vec(\Sigma(V_{0|0}))$, where I_9 is a 9×9 identity matrix and \otimes is the Kronecker product. Now, consider that $V_{t|t}$ and $P_{t|t}$ are available at a generic iteration t . Then, the filter proceeds recursively through the forecasting step:

$$\begin{cases} V_{t+1|t} = \Phi_0 + \Phi_1 V_{t|t} \\ P_{t+1|t} = \Phi_1 P_{t|t} \Phi_1^\top + \Sigma(V_{t|t}) \\ CUM_{t+1|t}^Q = \Gamma_0 + \Gamma_1 V_{t+1|t} \\ M_{t+1|t} = \Gamma_1 P_{t+1|t} \Gamma_1^\top + \Omega_{cum} \end{cases} \tag{22}$$

and the updating step:

$$\begin{cases} V_{t+1|t+1} = \left[V_{t+1|t} + P_{t+1|t} \Gamma_1^\top M_{t+1|t}^{-1} \left(CUM_{t+1}^Q - CUM_{t+1|t}^Q \right) \right]_+, \\ P_{t+1|t+1} = P_{t+1|t} - P_{t+1|t} \Gamma_1^\top M_{t+1|t}^{-1} \Gamma_1 P_{t+1|t}, \end{cases} \tag{23}$$

where $[V]_+$ returns a vector whose i^{th} element is $\max(V_i, 0)$. This additional condition ensures that latent factor estimates remain positive for all iterations—a crucial property for stochastic volatility factors that cannot assume negative values. Finally, we construct a Gaussian quasi log-likelihood for the cumulants:

$$Lik^{CUM} = -\frac{1}{2} \sum_{t=1}^T \left[\ln \left((2\pi)^{2J} \det(M_{t|t-1}) \right) + \xi_{t,cum}^\top M_{t|t-1}^{-1} \xi_{t,cum} \right], \tag{24}$$

where $\xi_{t,cum} \equiv CUM_t^Q - CUM_{t|t-1}^Q$.

3.2 Risk-Neutral Expected Quadratic Loss and Gain Likelihood

We use Equation (8) to compute the model-implied $\mu_2^{Q+}(t, \tau)$ and $\mu_2^{Q-}(t, \tau)$. Note that $\mathbb{E}_t^Q[r_{t,t+\tau}^2] = CUM_{t,\tau}^{(2)Q} + (CUM_{t,\tau}^{(1)Q})^2$ and both $CUM_{t,\tau}^{(2)Q}$ and $CUM_{t,\tau}^{(1)Q}$ are computed analytically within the AFT framework following Feunou and Okou (2018). We follow Fang and Oosterlee (2008) to approximate $\Lambda^Q(t, \tau)$ analytically. All the details are provided in Section A7 of the Online Appendix. Hence, both $\mu_2^{Q+}(t, \tau)$ and $\mu_2^{Q-}(t, \tau)$ are nonlinear functions of the factor V_t . We construct a Gaussian quasi log-likelihood for the truncated moments $Tmom_t$,

$$Lik^{Tmom} = -\frac{1}{2} \sum_{t=1}^T \left[\ln \left((2\pi)^{2J} \det(\Omega_{Tmom}) \right) + \xi_{t,Tmom}^\top \Omega_{Tmom}^{-1} \xi_{t,Tmom} \right], \tag{25}$$

where $\xi_{t,Tmom} = Tmom_t^{(Obs)} - Tmom_t(V_{t|t})$, Ω_{Tmom} denotes the measurement error variance, $V_{t|t}$ is obtained through the filtering procedure (see Equations (22) and (23)),

$$\begin{aligned} Tmom_t^+ &= \left(\mu_{2,\tau_1}^{Q+}, \dots, \mu_{2,\tau_j}^{Q+} \right)^\top; \quad Tmom_t^- = \left(\mu_{2,\tau_1}^{Q-}, \dots, \mu_{2,\tau_j}^{Q-} \right)^\top; \quad Tmom_t \\ &\equiv \left(Tmom_t^{+\top}, Tmom_t^{-\top} \right)^\top, \end{aligned}$$

and $Tmom_t^{(Obs)}$ is the time- t observed risk-neutral truncated moments (computed model-free using Equation (13)). Parameters for different models are estimated via a maximization of $Lik^{CUM} + Lik^{Tmom}$.

3.3 Discussions

Given the heteroscedasticity and non-normality of factors and some nonlinear measurement equations, the Kalman filter is not optimal in this case.⁶ Two other alternatives could have been considered regarding nonlinear measurement equations: (i) locally linearize the nonlinear measurement equation (this is known as the extended Kalman filter); or (ii) use a deterministic sampling technique (known as the unscented transformation) to accurately

6 Regarding the issue of heteroscedastic and non-Gaussian factors, we refer readers to Duan and Simonato (1999) for extensive discussions and Monte Carlo analyses suggesting that the loss of optimality is very minimal.

estimate the true mean and covariance (this is known as the unscented Kalman filter). We want to emphasize that the Kalman filter described in this article uses nonlinear equations for parameter estimation (see Equation (25)), which implies that nonlinear equations affect the filtering of unobserved factors in an indirect way. While the extended Kalman filter is appealing as it allows the nonlinear measurement equations to directly affect the filtered factors, in our case its implementation is difficult mainly because nonlinear measurements and their derivatives are approximate.

Nevertheless, of first importance is evaluating the impact of this estimation choice on our main results. To do so, we implement the extended Kalman filter for the most flexible model (AFT4) and compare its fit with the classic Kalman filter. Using the approach in Fang and Oosterlee (2008), we derive a simple analytical approximation of the Jacobian required for the effective implementation of the extended Kalman filter. Because of space constraints, we report these derivations and all empirical results in the Online Appendix (Table A3, Figures A5 and A6). We find that the two methods provide a very similar fit. Thus, our main conclusions are robust to the choice of the estimation method.

4 Results

In this section, we evaluate the ability of different models to fit the term structure of the expected quadratic payoff and its loss and gain components. We use Christoffersen, Heston, and Jacobs (2009) (AFT0) as our baseline model. We compare this baseline model with two other two-factor alternatives AFT1 and AFT2, and two more three-factor models, AFT3 with a symmetric jump distribution and AFT4 in Andersen, Fusari, and Todorov (2015). Finally, we evaluate the ability of different pricing kernels with various flexibility to fit the QRP and its loss and gain components.

4.1 Fitting the Risk-Neutral Expectations

We examine the performance of different models by relying on the root-mean-squared error (RMSE):

$$RMSE \equiv \sqrt{\frac{1}{T} \sum_{t=1}^T (Mom_t^{Mkt} - Mom_t^{Mod})^2},$$

where Mom_t^{Mkt} is the time t observed risk-neutral moment and Mom_t^{Mod} is the model-implied equivalent. Results are reported in Table 3.⁷ Overall, regarding the fitting of the term structure of the risk-neutral expected gain and loss, we find that the benchmark two-factor diffusion model (AFT0) is outperformed by all the other variants.

With respect to the risk-neutral quadratic loss fit, the AFT0 model's RMSE increases with the horizon and ranges from 1% at 2 months to 2.15% at 1 year. The average RMSE is 1.73%, which is far higher than other models' RMSEs. The best performer is the AFT4 model with an average error of 0.77%, which offers approximately a 56% improvement over the benchmark AFT0 model. This performance of the AFT4 model is robust across horizons, with an RMSE as low as 0.45% around horizons of 4–5 months, which is an

7 All risk-neutral parameter estimates, together with their standard deviations, are reported in Table A2 of the Online Appendix.

Table 3 Risk-neutral moments RMSEs

τ	Quadratic loss					Quadratic gain				
	AFT0	AFT1	AFT2	AFT3	AFT4	AFT0	AFT1	AFT2	AFT3	AFT4
2	1.08	0.97	1.14	0.77	0.68	2.56	0.69	1.50	1.65	0.82
3	1.47	0.60	0.81	0.62	0.50	2.12	0.72	1.13	1.37	0.63
4	1.73	0.40	0.57	0.52	0.44	1.99	0.78	0.95	1.14	0.69
5	1.82	0.45	0.46	0.49	0.46	1.93	0.85	0.91	1.08	0.80
6	1.77	0.66	0.55	0.64	0.54	1.98	0.92	0.98	1.06	1.00
7	1.74	0.88	0.74	0.80	0.66	2.07	0.99	1.03	1.07	1.19
8	1.71	1.06	0.89	0.97	0.83	2.15	1.06	1.07	1.08	1.37
9	1.70	1.23	1.05	1.12	0.95	2.21	1.12	1.10	1.03	1.51
10	1.86	1.36	1.20	1.23	1.03	2.28	1.19	1.13	1.00	1.63
11	1.99	1.49	1.29	1.32	1.11	2.38	1.21	1.12	1.00	1.72
12	2.15	1.60	1.39	1.36	1.22	2.45	1.29	1.06	1.07	1.80
Avg	1.73	0.97	0.92	0.89	0.77	2.19	0.98	1.09	1.14	1.20

τ	Volatility					Skewness				
	AFT0	AFT1	AFT2	AFT3	AFT4	AFT0	AFT1	AFT2	AFT3	AFT4
2	0.74	1.09	1.04	0.69	0.31	0.97	0.19	0.26	0.30	0.47
3	0.33	0.70	0.70	0.61	0.11	0.95	0.16	0.19	0.22	0.27
4	0.51	0.44	0.43	0.47	0.18	0.97	0.15	0.15	0.17	0.19
5	0.47	0.29	0.24	0.31	0.21	0.92	0.13	0.12	0.13	0.13
6	0.43	0.21	0.19	0.19	0.21	0.88	0.11	0.09	0.09	0.09
7	0.36	0.29	0.26	0.23	0.18	0.85	0.08	0.09	0.08	0.09
8	0.22	0.40	0.39	0.33	0.15	0.81	0.08	0.11	0.09	0.11
9	0.20	0.55	0.52	0.44	0.12	0.79	0.09	0.13	0.09	0.15
10	0.36	0.67	0.64	0.50	0.16	0.75	0.11	0.17	0.10	0.17
11	0.54	0.81	0.78	0.61	0.20	0.72	0.13	0.19	0.12	0.19
12	0.71	0.93	0.93	0.75	0.28	0.74	0.14	0.19	0.15	0.22
Avg	0.44	0.58	0.56	0.47	0.19	0.85	0.12	0.15	0.14	0.19

Notes: In this table we report the root-mean-squared error.

$$RMSE \equiv \sqrt{\frac{1}{T} \sum_{t=1}^T (Mom_t^{Mkt} - Mom_t^{Mod})^2}$$

where Mom_t^{Mkt} is the time t risk-neutral moment value observed on the market, and Mom_t^{Mod} is the corresponding model-implied equivalent. All variance RMSEs are in annual percentage units. The sample period is from January 1996 to December 2015.

improvement of nearly 75% over the benchmark AFT0 model. The three-factor models (AFT3 and AFT4) outperform the other two-factor models (AFT1 and AFT2). The AFT4 model offers an improvement of approximately 15% over the AFT3 model, which underscores the importance of accounting for asymmetry in the jump distribution.

Turning to the risk-neutral quadratic gain fit, the AFT0 model’s average RMSE is 2.19%, which is roughly 50% higher than other variant’s RMSEs. The best-performing

model on this front is the AFT1 model with an average RMSE of 0.98%, while the performance of the AFT2, AFT3, and AFT4 models is similar. However, the AFT0 model fits the term structure of the total risk-neutral quadratic payoff remarkably well with an average RMSE of 0.44%, which confirms the findings of [Christoffersen, Heston, and Jacobs \(2009\)](#). The best performer for the term structure of the total quadratic payoff is again the AFT4 model with an RMSE of about 0.19%, which offers an improvement of about 57% over the benchmarks AFT0 and AFT3. In accordance with our findings regarding the quadratic loss, this result highlights the importance of asymmetry in the jump distribution for fitting the term structure of the risk-neutral variance.

On the term structure of the risk-neutral skewness dimension, the benchmark AFT0 is the worst performer with an average RMSE of 0.85, whereas all the other variants have a similar fit, with an average RMSE of approximately 0.15, which is an almost 80% improvement over the benchmark AFT0. These results underscore the importance of jumps when fitting the term structure of risk-neutral skewness. To better understand our findings, in [Figure 2](#) we plot the observed and model implied average term structure of risk-neutral moments (the top and middle panels). The AFT0 model is clearly unable to fit the average term structure of risk-neutral expected quadratic gain or loss. It overestimates the risk-neutral expected quadratic gain and underestimates the risk-neutral expected quadratic loss, which explains why it is able to fit the term structure of the total risk-neutral expected quadratic payoff well. Not surprisingly, the AFT0 model is outperformed by all the other variants when it comes to fitting the term structure of skewness. The most likely explanation is that jumps are essential to generate skewness; accounting for only the leverage effect is not enough.

The ranking between two-factor models (AFT1 and AFT2) is mixed. The model without a pure-jump process (AFT1) dominates the one without a pure-diffusion process (AFT2) when fitting the term structure of the risk-neutral expected quadratic loss and gain in the short end. However, this result is reversed in the long end. [Figure 2](#) shows that, on average, the AFT1 model fits the term structure of the risk-neutral expected quadratic gain remarkably well, while the AFT2 model fits the term structure of the risk-neutral expected quadratic loss very well. These results suggest that incorporating a pure-jump process and having jumps in the volatility are essential for the distribution of the loss uncertainty, while a pure-diffusion process is a key ingredient for the distribution of the gain uncertainty.

Both of the two-factor variants are outperformed by the most general specification (AFT4), which overall is able to reproduce the term structure of the truncated and total risk-neutral moments remarkably well. Comparing the two three-factor models (AFT3 and AFT4), we evaluate the importance of introducing a wedge between the negative and positive jump distributions. The results are mixed on this front. For the term structure of the risk-neutral expected quadratic loss and the risk-neutral expected quadratic payoff, the AFT4 model clearly outperforms the AFT3 model (which has no asymmetry in the jump distribution). However, there is no clear winner for the gain uncertainty and skewness. The AFT4 model has a better fit on the short end of the risk-neutral expected quadratic gain, while the AFT3 is preferred on the long end.

4.2 Fitting the QRP

We focus on the most flexible specification of the AFT model (AFT4) and evaluate the fitting ability of different pricing kernel specifications discussed in Section 2.2. In the bottom

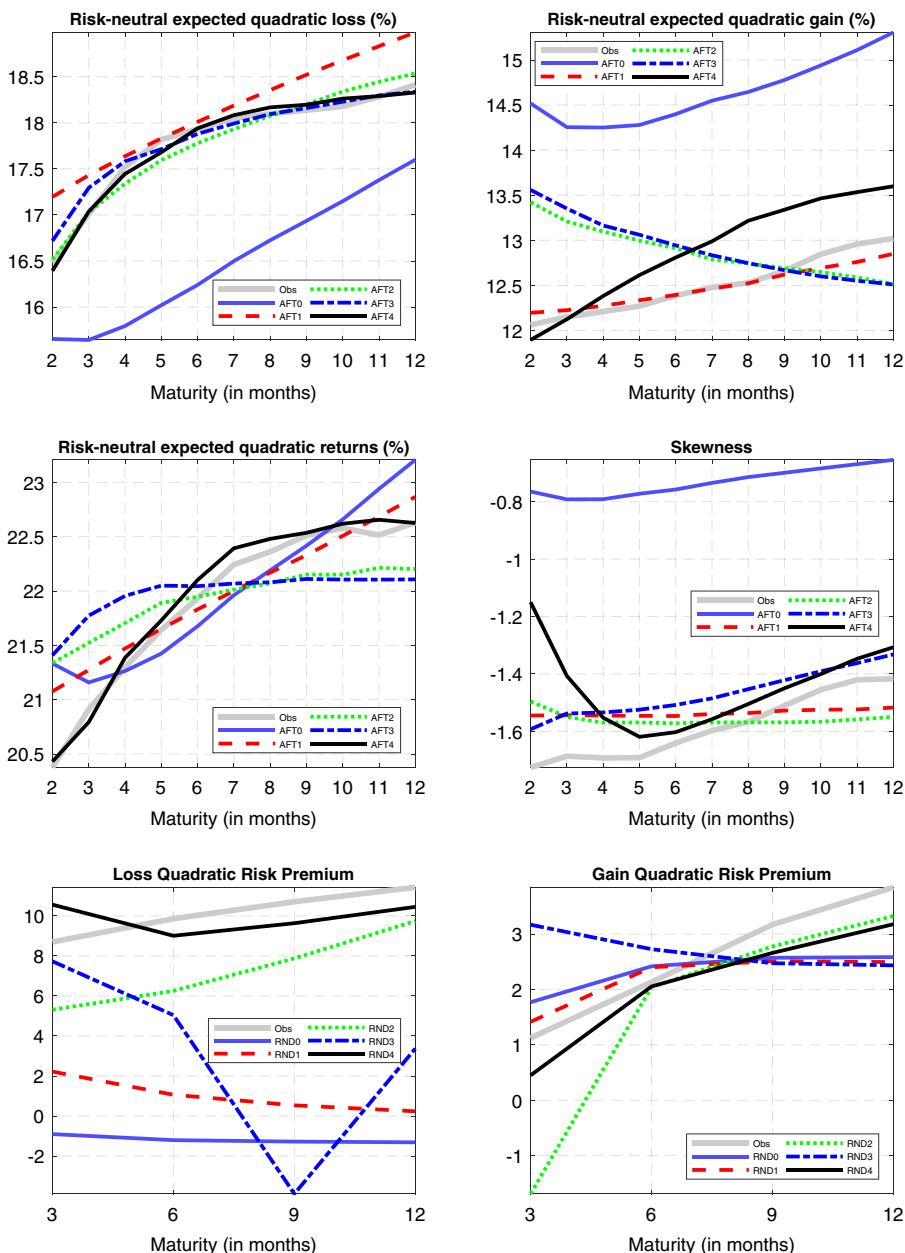


Figure 2 Observed and model-implied average term structure of risk-neutral moments and risk-premium. In this figure we plot the observed and model-implied average term structure of risk-neutral moments (the first two rows) and QRP (the third row). All values of the risk-neutral expected quadratic payoff and its components are reported in annualized volatility terms. For the risk premium, we use the most flexible specification of the AFT model (AFT4). The sample period is from January 1996 to December 2015.

panels of Figure 2, we plot the observed and model implied average term structure of the QRP. It is readily apparent that the most flexible Radon–Nikodym derivative (*RND4*) is the only one that is able to adequately fit the average term structure of the QRP and its loss and gain components. The worst performer is *RND0*, which assumes that jumps are not priced. Not pricing jumps generates a negative average term structure in the net and loss QRP. Not pricing either positive jumps (*RND2*) or negative jumps (*RND3*) is also strongly rejected. Finally, even though a symmetric Radon–Nikodym derivative (*RND1*), which gives the same price to both the positive and negative jumps, is able to replicate the positive sign for all the term structures, it falls short in capturing the level. Overall, we find that it is imperative to price jumps asymmetrically in the pricing kernel.

To confirm these visual findings, we report the RMSE in Table 4. In addition to the loss and gain QRP RMSEs reported in the top panels, we also report the total QRP (denoted Net QRP since it is the difference between the loss and gain QRP) in the bottom left panel and the skewness risk premium (SRP, which is defined as the sum of the loss and gain QRP) in the bottom right panel. The numbers are roughly in line with the visual findings of Figure 2. Except for the very short maturity (3 months), *RND4* yields the smallest RMSE across maturities and for different types of risk premia. *RND0* is the worst performer, which implies that pricing jumps is important for the dynamics of the QRP and its components. The average RMSE for the *RND4* model is 3.5%, 1.2%, 1.4%, and 3.2% for the loss, gain, net, and sum QRP, respectively. These numbers are substantial improvements over the benchmarks *RND0* (which assumes that jumps are not priced) and *RND1* (which gives the same price to both the positive and negative jumps). To be more specific, on the one hand, the *RND4* model offers approximately 70%, 30%, 80%, and 72% improvements over *RND0* for the fitting of the loss, gain, and net QRP, and the SRP, respectively. On the other hand, the *RND4* model offers approximately 62%, 33%, 75%, and 66% improvement over *RND1* for the fitting of the loss, gain, and net QRP, and the SRP, respectively.

We can further scrutinize the overall performance results by maturity. Table 4 reveals that the superiority of the *RND4* pricing kernel holds across the maturity spectrum. For the loss, net QRP, and the SRP, the relative improvement increases with the maturity and reaches 80% at the 1-year horizon.

5 Conclusion

In this article, we investigate how the amount of money paid by investors to hedge negative spikes in the stock market changes with the investment horizon. For this purpose, we estimate the quadratic payoff and its loss and gain components across time and horizon. We uncover new empirical facts that challenge most of the existing option and variance swaps pricing models. Among these facts, we find an average upward-sloping term structure for the risk-neutral expected quadratic payoff and its components. We also find upward-sloping term structures for the physical expected quadratic payoff and quadratic gain but a downward-sloping term structure for the physical expected quadratic loss. There is significant time variation in the slopes of these term structures, and we observe that they are negative and spike during financial downturns. Finally, we find that at least three principal components are required to explain the cross-section (across maturity or horizon) of the risk-neutral and physical expected quadratic payoff and its components.

Table 4 QRP RMSEs

τ	Loss QRP					Gain QRP				
	RND0	RND1	RND2	RND3	RND4	RND0	RND1	RND2	RND3	RND4
3	9.86	6.39	3.92	3.55	6.36	1.72	1.95	2.79	2.80	1.12
6	11.53	9.21	4.08	6.04	2.73	1.53	1.62	0.93	1.41	0.92
9	12.53	10.82	3.85	8.01	2.73	1.69	1.70	1.18	1.43	1.23
12	13.21	11.63	4.09	9.16	2.55	1.95	1.93	1.62	1.73	1.52
Avg	11.78	9.51	3.99	6.69	3.59	1.72	1.80	1.63	1.84	1.20

τ	Net QRP					Skewness RP				
	RND0	RND1	RND2	RND3	RND4	RND0	RND1	RND2	RND3	RND4
3	7.33	4.86	2.63	3.28	2.77	10.90	7.31	4.08	4.85	6.60
6	7.85	6.15	2.16	3.90	1.17	11.98	9.84	3.71	6.59	2.27
9	7.42	6.11	1.70	4.12	0.94	11.88	10.25	3.01	7.52	2.13
12	7.05	5.88	1.32	4.12	0.91	11.80	10.40	2.72	8.06	1.92
Avg	7.41	5.75	1.95	3.86	1.45	11.64	9.45	3.38	6.76	3.23

Notes: In this table we report the root-mean-squared error.

$$RMSE \equiv \sqrt{\frac{1}{T} \sum_{t=1}^T (QRP_t^{Mkt} - QRP_t^{Mod})^2},$$

where QRP_t^{Mkt} is the time t QRP value observed on the market, QRP_t^{Mod} is the corresponding model-implied equivalent. All variance RMSEs are in annual percentage units. Net QRP is the difference between loss and gain QRP, while skewness RP is the sum. The sample period is from January 1996 to December 2015.

To replicate these empirical facts, we focus on the Andersen, Fusari, and Todorov (2015) model and some of its restricted variants. This model is particularly appealing as it completely disentangles the dynamics of negative and positive jumps. In addition, the model has three factors, which is an essential ingredient as suggested by our principal component analysis. We find that models without an asymmetric treatment of positive and negative jumps are overall rejected as they are unable to fit the term structure of the risk-neutral expected quadratic loss. Notably, this category of models covers most of the existing option and variance swap pricing models found in the literature. We also evaluate different pricing kernel specifications and find that disentangling the price of negative jumps from its positive counterpart is essential for replicating the observed term structures of the loss and gain QRP.

Supplementary Data

Supplementary data are available at *Journal of Financial Econometrics* online.

References

Ait-Sahalia, Y., M. Karaman, and L. Mancini. 2020. “The term structure of equity and variance risk premia.” *Journal of Econometrics*. Forthcoming.

- Amengual, D., and D. Xiu. 2018. Resolution of Policy Uncertainty and Sudden Declines in Volatility. *Journal of Econometrics* 203: 297–315.
- Andersen, T., N. Fusari, and V. Todorov. 2015. The Risk Premia Embedded in Index Options. *Journal of Financial Economics* 117: 558–584.
- Andersen, T. G., T. Bollerslev, F. X. Diebold and H. Ebens 2001. The Distribution of Realized Stock Return Volatility. *Journal of Financial Economics* 61: 43–76.
- Andersen, T. G., T. Bollerslev, F. X. Diebold, and P. Labys. 2003. Modeling and Forecasting Realized Volatility. *Econometrica* 71: 579–625.
- Bakshi, G., and D. Madan. 2000. Spanning and Derivative-Security Valuation. *Journal of Financial Economics* 55: 205–238.
- Bakshi, G., N. Kapadia, and D. Madan. 2003. Stock Return Characteristics, Skew Laws and the Differential Pricing of Individual Equity Options. *Review of Financial Studies* 16: 101–143.
- Barndorff-Nielsen, O. E., S. Kinnebrock, and N. Shephard. 2010. “Measuring Downside Risk: Realised Semivariance.” In T. Bollerslev, and J. Russell, and M. Watson (eds.), *Volatility and Time Series Econometrics: Essays in Honor of Robert F. Engle*. Oxford: Oxford University Press, pp. 117–136.
- Bates, D. S. 2012. U.S. stock Market Crash Risk, 1926–2010. *Journal of Financial Economics* 105: 229–259.
- Bekaert, G., E. Engstrom, and A. Ermolov. 2015. Bad Environments, Good Environments: A Non-Gaussian Asymmetric Volatility Model. *Journal of Econometrics* 186: 258–275.
- Bekaert, G., and M. Hoerova. 2014. The Vix, the Variance Premium and Stock Market Volatility. *Journal of Econometrics* 183: 181–192. Analysis of Financial Data.
- Bernardo, A. E., and O. Ledoit. 2000. Gain, Loss, and Asset Pricing. *Journal of Political Economy* 108: 144–172.
- Black, F., and M. Scholes. 1973. The Pricing of Options and Corporate Liabilities. *Journal of Political Economy* 81: 637–654.
- Carr, P., and L. Wu. 2009. Variance Risk Premiums. *Review of Financial Studies* 22: 1311–1341.
- Chen, X., and E. Ghysels. 2011. News Good or Bad and Its Impact on Volatility Predictions over Multiple Horizons. *Review of Financial Studies* 24: 46–81.
- Chernov, M., R. Gallant, E. Ghysels, and G. Tauchen. 2003. Alternative Models for Stock Price Dynamics. *Journal of Econometrics* 116: 225–257.
- Christoffersen, P., S. Heston, and K. Jacobs. 2009. The Shape and Term Structure of the Index Option Smirk: Why Multifactor Stochastic Volatility Models Work so Well. *Management Science* 55: 1914–1932.
- Christoffersen, P., K. Jacobs, and C. Ornathanalai. 2012. Dynamic Jump Intensities and Risk Premiums: Evidence from S&P500 Returns and Options. *Journal of Financial Economics* 106: 447–472.
- Conrad, J., R. F. Dittmar, and E. Ghysels. 2013. Ex Ante Skewness and Expected Stock Returns. *The Journal of Finance* 68: 85–124.
- Dew-Becker, I., S. Giglio, A. Le, and M. Rodriguez. 2017. The Price of Variance Risk. *Journal of Financial Economics* 123: 225–250.
- Duan, J.-C., and J.-G. Simonato. 1999. Estimating Exponential-Affine Term Structure Models by Kalman Filter. *Review of Quantitative Finance and Accounting* 13: 111–135.
- Duffie, D., J. Pan, and K. Singleton. 2000. Transform Analysis and Option Pricing for Affine Jump-Diffusions. *Econometrica* 68: 1343–1377.
- Eraker, B. 2004. Do Stock Prices and Volatility Jump? Reconciling Evidence from Spot and Option Prices. *The Journal of Finance* 59: 1367–1404.
- Fang, F., and C. Oosterlee. 2008. A Novel Pricing Method for European Options Based on Fourier-Cosine Series Expansions. *SIAM Journal on Scientific Computing* 31: 826–848.

- Feunou, B., M. R. Jahan-Parvar, and R. Tédongap. 2013. Modeling Market Downside Volatility. *Review of Finance* 17: 443–481.
- Feunou, B., and C. Okou. 2018. Risk-Neutral Moment-Based Estimation of Affine Option Pricing Models. *Journal of Applied Econometrics* 33: 1007–1025.
- Feunou, B., R. Lopez Aliouchkin, R. Tédongap and L. Xu 2019. “Loss Uncertainty, Gain Uncertainty, and Expected Stock Returns.” Working Paper, Bank of Canada, Syracuse University and ESSEC Business School.
- Huang, J.-Z., and L. Wu. 2004. Specification Analysis of Option Pricing Models Based on Time-Changed Lévy Processes. *The Journal of Finance* 59: 1405–1439.
- Jiang, G. J., and Y. S. Tian. 2005. The Model-Free Implied Volatility and Its Information Content. *Review of Financial Studies* 18: 1305–1342.
- Kilic, M., and I. Shaliastovich. 2019. Good and Bad Variance Premia and Expected Returns. *Management Science* 65: 2522–2544.
- Newey, W. K., and K. D. West. 1987. A Simple, Positive Semi-Definite Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica* 55: 703–708.
- Patton, A., and K. Sheppard. 2015. Good Volatility, Bad Volatility: Signed Jumps and the Persistence of Volatility. *Review of Economics and Statistics* 97: 683–697.
- Santa-Clara, P., and S. Yan. 2010. Crashes, Volatility, and the Equity Premium: Lessons from s&cp 500 Options. *The Review of Economics and Statistics* 92: 435–451.

Realized Volatility Forecasting with Neural Networks

Andrea Bucci^{1,2}

¹Università Politecnica delle Marche and ²Università degli Studi G. d'Annunzio Chieti e Pescara

Address correspondence to Andrea Bucci, Università degli Studi "G. d'Annunzio" Chieti-Pescara, Viale Pindaro 42, Pescara, Italy, or e-mail: andrea.bucci@unich.it.

Received September 2, 2019; revised September 2, 2019; editorial decision April 14, 2020; accepted April 14, 2020

Abstract

In the last few decades, a broad strand of literature in finance has implemented artificial neural networks as a forecasting method. The major advantage of this approach is the possibility to approximate any linear and nonlinear behaviors without knowing the structure of the data generating process. This makes it suitable for forecasting time series which exhibit long-memory and nonlinear dependencies, like conditional volatility. In this article, the predictive performance of feed-forward and recurrent neural networks (RNNs) was compared, particularly focusing on the recently developed long short-term memory (LSTM) network and nonlinear autoregressive model process with exogenous input (NARX) network, with traditional econometric approaches. The results show that RNNs are able to outperform all the traditional econometric methods. Additionally, capturing long-range dependence through LSTM and NARX models seems to improve the forecasting accuracy also in a highly volatile period.

Key words: neural network, machine learning, stock market volatility, realized volatility

JEL classification: C22, C24, C58, G17

Measuring and predicting stock market volatility has received growing attention from both academics and practitioners over the last years. It is well known that stock return volatility varies over time (Engle, 1982; Bollerslev, 1986) and asymmetrically responds to unexpected news (Black, 1976; Nelson, 1990), which may cause distortions in the estimation of volatility and in the definition of its underlying process. For these reasons, some authors suggested to estimate stock market volatility through a smooth transition or a threshold model (De Pooter, Martens, and Van Dijk, 2008; McAleer and Medeiros, 2008). The nonlinearity makes the estimation of these models difficult, since the sample log-likelihood can exhibit local maxima and may be generally hard to solve with confidence. Furthermore, this class of models is in general greedy in requiring a substantial amount of data to identify the states

and presents poor out-of-sample forecasting performance (Clements and Krolzig, 1998; Pavlidis, Paya, and Peel, 2012).

In this framework, this article aims to capture the nonlinear relationships between aggregate stock market volatility, measured by realized volatility, and a set of financial and macroeconomic variables through artificial neural networks (ANNs). Realized volatility has been shown to be subject to structural breaks and regime-switching, hence the need to use a nonlinear adaptive modeling approach such neural networks. This method allows approximating arbitrarily well a wide class of linear and nonlinear functions without knowing the data generating process. Furthermore, ANNs are found to be particularly useful to forecast volatile financial variables exhibiting nonlinear dependence, such as stock prices, exchange rates, and realized volatility; see Donaldson and Kamstra (1996a,b). Even if neural networks have been around since 1950s, only in the last two decades they have been used in finance, showing that they can outperform linear models in capturing complex relationships in which linear models fail to perform well. In particular, they seem to be suitable for modeling the dynamics of realized volatility in relation to macroeconomic and financial determinants, that may drive the dynamics of realized volatility not linearly.

ANNs have been commonly implemented for predicting stock prices (White, 1988; Kamijo and Tanigawa, 1990; Khan, 2011), while there has been little effort on forecasting volatility through neural networks. Moreover, neural networks have been mostly employed in combination with GARCH models (Hajizadeh et al., 2012; Maciel, Gomide, and Ballini, 2016). For instance, Donaldson and Kamstra (1997) investigated the usefulness of a semi-nonparametric GARCH model to capture nonlinear relationships, proving that the ANN model performs better than all competing models. Hu and Tsoukalas (1999), instead, combined the forecasts from four conditional volatility models within a neural network's architecture, showing that the ANNs predict accurately well the targeted variable during crisis periods. Arnerić, Poklepovic, and Aljinović (2014) based their neural networks on the squared innovations deriving from a GARCH model. They relied on a Jordan neural network (JNN) and showed that an neural networks (NN) model provides superior forecasting accuracy in comparison with other linear and nonlinear models. An early contribution to this literature is Hamid and Iqbal (2004), which compared the forecasting performance of neural networks using implied volatility and realized volatility. In their manuscript, neural networks were able to outperform implied volatility forecasts and were in line with realized volatility. Kristjanpoller, Fadic, and Minutolo (2014) also made use of neural networks to forecast, as in this manuscript, monthly realized volatility and returns of three Latin-American stock market indexes.

Fernandes, Medeiros, and Scharth (2014) extended these studies by specifying a neural-network heterogeneous autoregressive (HAR) with exogenous variables to improve implied volatility forecasts. A recent paper by Vortelinos (2017) implemented a neural network to forecast a nonparametric volatility measure. The author concluded that the persistence in realized volatility is not well approximated by a feed-forward network. More recently, Rosa et al. (2014) and Miura, Pichl, and Kaizoji (2019) relied on the use of neural networks to provide out-of-sample forecasts of realized volatility, both showing that neural networks were able to outperform linear models.

This article contributes to this literature investigating whether a totally nonparametric model is able to outperform econometric methods in forecasting realized volatility. In particular, the analysis performed here compares the forecasting accuracy of time series models with

several neural networks architectures, as the feed-forward neural network (FNN), the Elman neural network (ENN), the JNN, a long short-term memory (LSTM) neural network, and the nonlinear autoregressive model process with exogenous input (NARX) neural network.

The latent volatility is estimated through the ex-post measurement of volatility based on high-frequency data, namely realized volatility; see [Andersen et al. \(2001\)](#) and [Barndorff-Nielsen and Shephard \(2002\)](#). Since macroeconomic and financial variables, which are sampled at lower frequencies, are included in the model, realized volatility is estimated on a monthly basis from daily squared returns.

The remainder of this article is organized as follows: Section 1 illustrates the data set, the estimation method of the volatility, and the set of macroeconomic and financial predictors. Section 2 introduces the neural network models. The choice of the architecture of the neural networks is presented in Section 3. In Section 4, the performance of the ANNs is assessed in terms of forecasting accuracy, while Section 5 concludes.

1 Data and Volatility Measurement

The data set employed in this study comprises monthly observations from February 1950 to December 2017 for a total of 815 observations. The realized variance for month t is computed as the sum of squared daily returns, $\sum_{i=1}^{N_t} r_{i,t}^2$, where $r_{i,t}$ is the i th daily continuously compounded return in month t and N_t denotes the number of trading days during month t . Given that the natural logarithm of realized volatility is approximately Gaussian ([Andersen et al., 2001](#)), the realized volatility is here defined as the log of the square root of the realized variance (RV):

$$RV_t = \ln \sqrt{\sum_{i=1}^{N_t} r_{i,t}^2}, \quad (1)$$

where $r_{i,t}$ is the daily return of the Standard & Poor's (S&P) index. The logarithm of the realized volatility is highly persistent, as indicated by the time series plot in [Figure 1](#) and by the autocorrelation function (ACF) in [Figure 2](#), suggesting that a long-memory detecting model should be implemented ([Rossi and Santucci de Magistris, 2014](#)). In order to understand whether a nonlinear model was truly necessary, the nonlinearity tests discussed in [Terasvirta \(1994\)](#) and [Keenan \(1985\)](#) were performed. The tests rejected the null of linearity with a p -value lower than 0.001.

Since volatility exhibits a highly variable behaviour, one may also suspect that its dynamics are partly driven by several economic variables. A strand of literature has focused on the identification of economic drivers of volatility. In a seminal work, [Schwert \(1989\)](#) found that volatility behaves in a countercyclical way respect to economic activity. Afterward, both [Engle, Ghysels, and Sohn \(2009\)](#) and [Diebold and Yilmaz \(2009\)](#) showed a strong link between macroeconomic fundamentals and stock return volatility. More recently, [Paye \(2012\)](#) and [Christiansen, Schmeling, and Schrimpf \(2012\)](#) examined the role of a large set of macroeconomic and financial variables on the dynamics of realized volatility. They proved that the presence of exogenous variables helps increasing forecasting accuracy. The same findings are showed by [Bucci, Palomba, and Rossi \(2019\)](#), which analyzed the forecasting accuracy of realized covariance through a Vector Logistic Smooth Transition Autoregressive model.

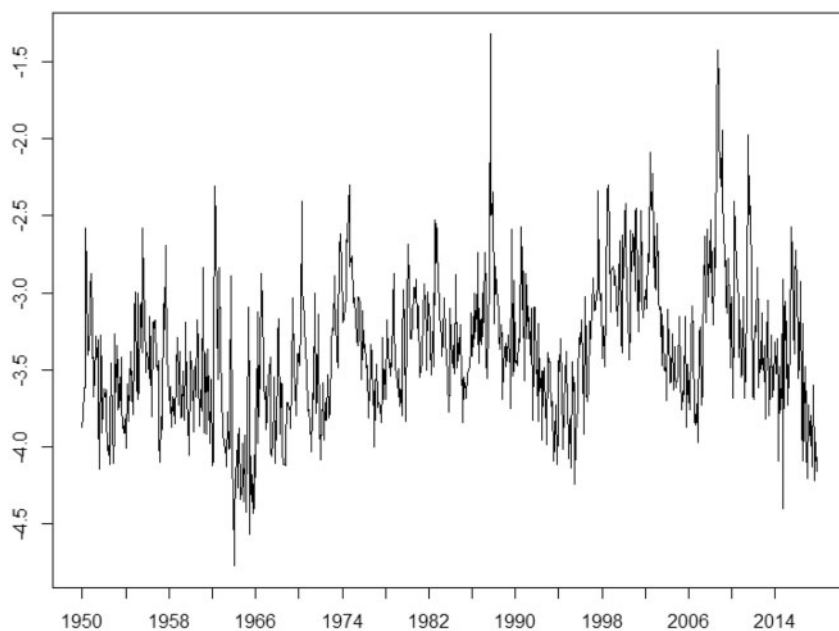


Figure 1 log RV from February 1950 through December 2017.

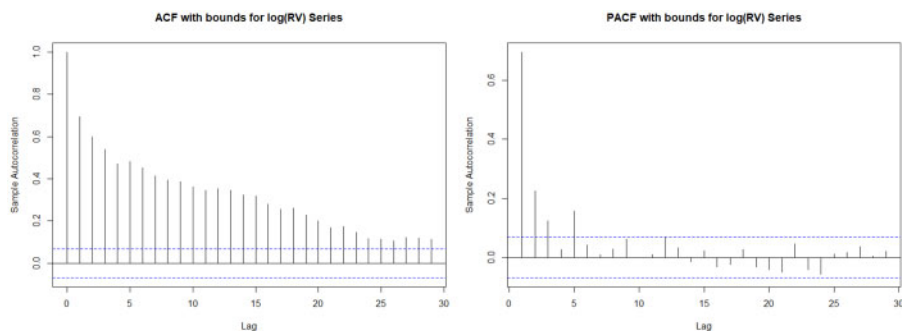


Figure 2 ACF and partial autocorrelation function of RV.

Understanding which are the volatility predictors can be crucial for investment decisions, and for policy makers and monetary authorities. Thus, this analysis relies on a comprehensive set of macroeconomic and financial variables as volatility predictors.

As in [Paye \(2012\)](#) and [Christiansen, Schmeling, and Schrimpf \(2012\)](#), I include in the analysis many predictive variables from return predictability literature ([Mele, 2007, 2008](#)).

First, the set of determinants comprehends the dividend-price (DP) and the earnings-price ratio (EP), commonly included in the set of the excess returns predictors, see also [Welch and Goyal \(2008\)](#). The well-known leverage effect (i.e., negative returns reflect higher volatility) is gathered through the equity market excess return (MKT). As a measure

of risk factors, the [Fama and French's \(1993\)](#) factors (High Minus Low, HML, and Small Minus Big, SMB) are considered in the analysis. The short-term reversal factor (STR) is included to capture the component of stock returns unexplained by “fundamentals.”

A set of bond market variables enriches the set of determinants, as the Treasury bill (T-bill) rate, the rate of return on long-term government bond and the term spread difference (TS) of long-term bond yield three-month T-bill rate. The default spread (DEF) completes the set of financial determinants to approximate credit risk.

The inclusion of macroeconomic variables, as inflation rate and industrial production growth, follows [Schwert \(1989\)](#) and [Engle, Ghysels, and Sohn \(2009\)](#). Including these variables permits to assess whether volatility is countercyclical or not. A description of the variables in the data is shown in [Table 1](#).

2 Neural Networks

ANNs can be seen as nonparametric tools, inspired by the structure of the human brain, for modeling and predicting the unknown function generating the observed data ([Arnerić, Poklepovic, and Aljinović, 2014](#)). The structure of the network can be modified to approximate a wide range of statistical and econometric models. For this reason, ANNs have been widely employed to forecast time series in different areas, like finance, medicine, biology, engineering, and physics. Empirical research indicates that ANNs are particularly suitable for forecasting volatile financial variables that exhibit nonlinear behaviors, like stock market returns or stock market volatility ([Maheu and McCurdy, 2002](#)), since they are capable of detecting nonlinear structure that linear models cannot detect. In this way, the researcher can implement neural networks without any a priori knowledge of the data generating process.

The neural network is specified as a collection of neurons (or nodes), grouped in layers, that connect to each other. The nodes of a layer are connected to the nodes of the following layer through weights and an activation function.¹ There exists a wide variety of learning algorithms to obtain these weights, the most popular being the backpropagation (BP). This algorithm is based on the gradient descent rule and allows to update the weights at each iteration, until there is no improvement in the error function, which is typically defined as the *mean squared error* (MSE).²

When the size of the network is too large, because of the number of hidden layers and hidden nodes, the training algorithm can be very slow. Although some rules have been suggested in the literature to find the optimal number of hidden layers and neurons ([Gnana Sheela and Deepa, 2013](#)), there is no commonly agreed solution to this issue. [Stinchcombe and White \(1992\)](#) proved that a single hidden neural network is a *universal approximator*, meaning that the network can approximate a wide range of linear and nonlinear functions,

- 1 An activation function is implemented in order to introduce nonlinearity to the network. Many activation functions, like sigmoid, hyperbolic tangent, and exponential, can be used in this framework, provided that they satisfy the condition of differentiability to apply the chain rule in the BP algorithm.
- 2 Other loss functions can be also implemented, such as mean absolute error and mean absolute percentage error.

Table 1 Variables description

Symbol	Variable	Data source	
		Description	Source
DP	Dividend yield ratio S&P	Dividends over the past year relative to current market prices; S&P 500 index	Robert Shiller’s website
EP	Earning price ratio S&P 500	Earnings over the past year relative to current market prices; S&P 500 index	Robert Shiller’s website
MKT	Market excess return	Fama–French’s market factor: return of U.S. stock market minus one-month T-bill rate	Kenneth French’s website
HML	Value factor	Fama–French’s HML factor: average return on value stocks minus average return on growth stock	Kenneth French’s website
SMB	Size premium factor	Fama–French’s SMB factor: average return on small stocks minus average return on big stocks	Kenneth French’s website
STR	Short-term reversal factor	Fama–French’s STR: average return on stocks with low prior return minus average return on stock with high prior return	Kenneth French’s website
TB	T-bill rate	Three-month T-bill rate	Datastream
TS	Term spread	Difference of long-term bond yield and three-month T-bill	Datastream
DEF	Default spread	Measure of default risk of corporate bonds: difference of BAA and AAA bond yields	Datastream
INF	Monthly Inflation	US inflation rate	Datastream
IP	Monthly industrial production growth rate	US industrial production growth	OECD database

if a sufficient number of hidden nodes is included. For this reason, a single hidden layer network is assumed throughout the present article. Assuming then a three-layer neural network and a single output variable, the output function is of the form:

$$f_t(x_t, \theta) = F\left(\beta_0 + \sum_{j=1}^q G(x_t \gamma'_j) \beta_j\right), \tag{2}$$

where F is the output activation function, G is the hidden units’ activation function, β_j , with $j = 1, \dots, q$, are the weights from hidden unit j to the output unit, $x_t = \{1, x_{1,t}, \dots, x_{s,t}\}$ is the $1 \times m$ vector of input variables at time t (with $m = s + 1$), β_0 is the bias of the final output, $\gamma_j = \{\gamma_{1,j}, \dots, \gamma_{m,j}\}$ is the $1 \times m$ vector of weights for the connections between the inputs and the hidden neuron j , q is the number of hidden units, and $\theta = \{\beta_0, \dots, \beta_q, \gamma'_1, \dots, \gamma'_q\}$ is the vector of all network weights. This version, with three input variables including the bias and two hidden nodes (i.e., $m = 3$ and $j = 2$), is depicted in [Figure 3](#) and assumes that information moves forward from the input layer to the output layer. Accordingly, it is also called *FNN*.

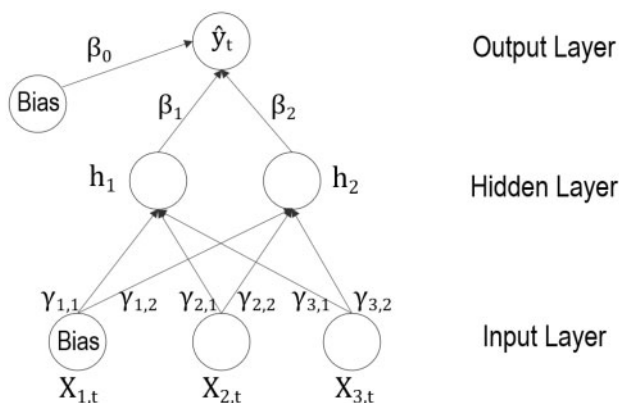


Figure 3 FNN with a single hidden layer.

Modern practice allows choosing F and G among a variety of functions. In the most used form of FNN, the output activation function is an identity function, that is, $F(a) = a$. In this case, Equation (2) can be written as follows

$$f_t(x_t, \theta) = \beta_0 + \sum_{j=1}^q G(x_t \gamma_j') \beta_j. \quad (3)$$

A common choice for G is the logistic function, that is, $G(a) = \frac{1}{1+e^{-a}}$, although any continuous, differentiable, and monotonic function may be implemented. This function, bounded between 0 and 1, permits the network to reproduce any nonlinear pattern and replicate the way a real neuron becomes active. In particular, the neuron shows a high level of activation for G close to 1, while it exhibits a poor response when G is close to 0.

Researchers usually refer to FNN as a *static network*, since a given set of input variables is used to forecast the target output variable at time t . Hence, feed-forward networks show no memory, even when sample information exhibits temporal dependence. The so-called *recurrent neural networks* (RNNs) overcome this shortcoming by allowing internal feed-backs. This type of networks allows propagating data from input to output, but also from later layers to earlier layers. Such models have many potential applications in economic and finance, when nonlinear time dependence and long-memory exist. For this reason, the use of RNN in forecasting volatility has attracted a large number of researchers (Schittenkopf, Dorffner, and Dockner, 2000; Tino, Schittenkopf, and Dorffner, 2001). This article focuses on four recurrent architectures: Elman and Jordan recurrent networks, LSTM networks, and NARX neural networks.

In the ENN, proposed by Elman (1990), the input layer has additional neurons which are fed back from the hidden layer (see Figure 4). The output of the ENN, with an identity function as output activation function, can be represented as

$$f_t(x_t, \theta) = \beta_0 + \sum_{j=1}^q h_{tj} \beta_j$$

$$h_{tj} = G(x_t \gamma_j' + h_{t-1} \delta_j') \quad j = 1, \dots, q \quad (4)$$

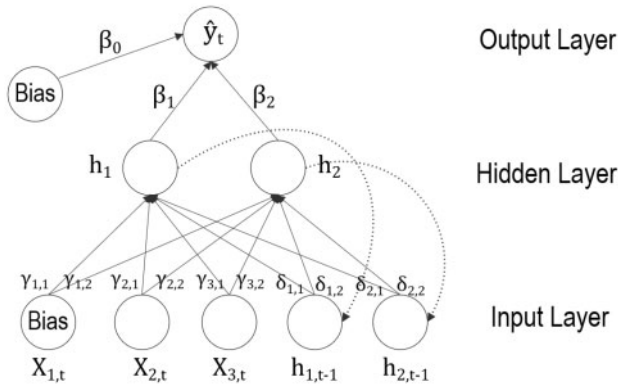


Figure 4 ENN with a single hidden layer.

where $h_{t-1} = (h_{t-1,1}, \dots, h_{t-1,q})$ is the vector of lagged hidden-unit activations, and $\delta_j = \{\delta_{1,j}, \dots, \delta_{q,j}\}$ is the vector of connection weights between the j th hidden unit and the lagged hidden units.

Jordan (1986), instead, introduced an RNN with a feedback from the output layer, as in Figure 5. Thus, the network output at time $t - 1$ is used as additional input for the network at time t . Specifically, the output of the JNN can be specified as follows

$$f_t(x_t, \theta) = \beta_0 + \sum_{j=1}^q G(x_t \gamma'_j + \hat{y}_{t-1} \psi_j) \beta_j \tag{5}$$

where \hat{y}_{t-1} is equal to $f_{t-1}(x_{t-1}, \theta)$ and ψ is the weight between the lagged output and the j th hidden unit.

Equations (4) and (5) indicate that the outputs of these RNNs can be expressed in terms of current and past inputs. This makes them similar to the distributed lag model or Autoregressive (AR) representation of an Autoregressive Moving Average (ARMA) model. Furthermore, differently from FNNs, RNNs are able to incorporate information of past observations without including them in the network.

Although extremely appealing, ENN and JNN suffer from the so-called “vanishing gradient problem.” In such methods, the network weights are updated through a training algorithm based on the gradient descent rule. When this kind of algorithm is implemented, the magnitude of the gradients gets exponentially smaller (vanishes) at each iteration, making the steps very small and resulting in an extremely slow learning process. In such cases, a local minimum might be reached.

One of the cause of this shortcoming is the choice of the activation function. For example, a logistic activation function maps all the input values in a relatively small range, that is $[0,1]$. As a result, even a large change in the input will produce a small change in the output, vanishing the gradient very fast.

LSTM was introduced by Hochreiter and Schmidhuber (1997) to alleviate the vanishing gradient problem through a mechanism based on memory cells. LSTM extends the RNN architecture by replacing each hidden unit with a memory block. Each block contains one or more self-connected memory cells and is equipped with three multiplicative units called

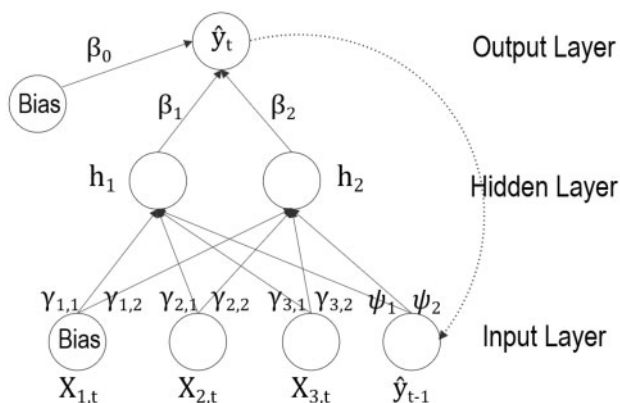


Figure 5 JNN with a single hidden layer.

input, forget, and output gates. These gates allow the memory cells to store and access information, in order to determine which information should be persisted. In this way, LSTMs are capable of retaining relevant information of input signals, overlooking the unnecessary parts.

Figure 6 illustrates the structure of a simple LSTM memory block with a one cell architecture. In the figure, x_t denotes the vector of input variables at time t , c_t and c_{t-1} correspond to the cell state at time t and $t - 1$, respectively, while h_t and h_{t-1} denote the hidden state or output of the cell at time step t and $t - 1$, respectively. The input gate is identified by i_t , f_t indicates the forget gate, while o_t is the output gate. Both input and output gates have the same role as in the RNNs. The new instance, that is, the forget gate, is responsible for removing the unnecessary information from the cell state. The information at time t , given by x_t and h_{t-1} , is passed through the forget gate f_t , which determines if the information should be retained or not using a sigmoid function. Basically, a zero response of the sigmoid function means that the information should be discarded, while a value close to 1 implies that the information should be stored. Meanwhile, the same information is processed by the input gate to add information to the cell state c_t . Additionally, a nonlinear layer, $\phi = \tanh$, is introduced to generate a vector of candidate values, \tilde{c}_t , to update the state of c_t . The output gate is used to regulate the output values of an LSTM cell, using a logistic function to filter the output. The final output of the memory cell, h_t , is then computed by feeding the cell state, c_t , into a \tanh layer and multiplying it by the value of the output gate. The entire process can be synthesized by the following equations:

$$f_t = \sigma(W_f h_{t-1} + U_f x_t + b_f) \tag{6}$$

$$i_t = \sigma(W_i h_{t-1} + U_i x_t + b_i) \tag{7}$$

$$\tilde{c}_t = \tanh(W_c h_{t-1} + U_c x_t + b_c) \tag{8}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \tag{9}$$

$$o_t = \sigma(W_o h_{t-1} + U_o x_t + V_o c_t + b_o) \tag{10}$$

$$h_t = o_t \odot \tanh(c_t) \tag{11}$$

$$\hat{y}_t = h_t \tag{12}$$

where W_f , W_i , W_c , W_o , U_f , U_i , U_c , and U_o are the weight matrices of forget, input, memory

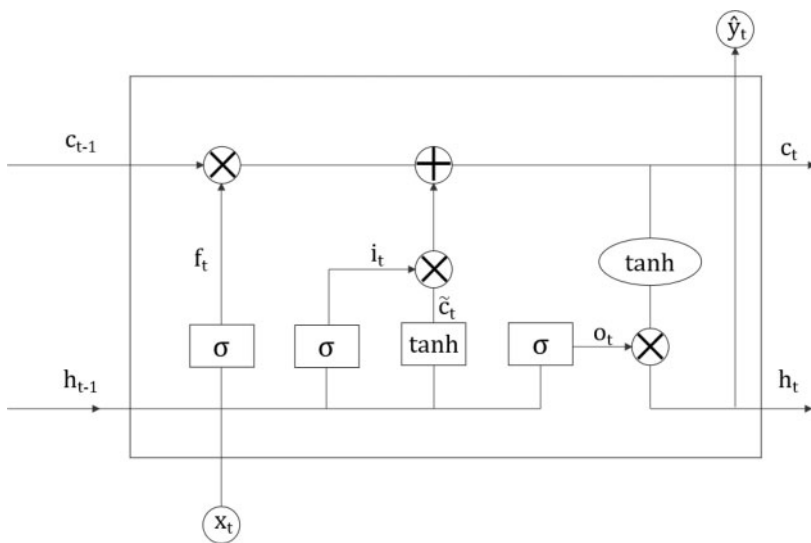


Figure 6 Basic LSTM memory cell.

Notes: The memory cell has four key components: an input gate, a neuron with self-current connection, a forget gate, and an output gate. The inputs (the predictors at time t and the outputs of the previous steps) are passed through the memory cell with some non-linear and linear interactions. Linear interactions of the cell state are point-wise addition \oplus and point-wise multiplication \otimes . Non-linear interactions are logistic functions, σ .

cell state, and output gates, respectively, V_c is the weight matrix of the cell state, \hat{y}_t is the output of the neural network, $b_f, b_i, b_c,$ and b_o are the biases of the related gates, σ is a sigmoid or logistic function and \odot is the Hadamard product function.

LSTM can operate where long-memory effects are present in the underlying structure of the times series, similarly to HAR or Autoregressive Fractional Integral Moving Average (ARFIMA) models. Accordingly, there are numerous applications of LSTM models in finance, see, for example, Heaton, Polson, and Witte (2016), Bao, Yue, and Rao (2017), Pichl and Kaizoji (2017), Kim and Won (2018), Di Persio and Honchar (2017), and Xiong, Nichols, and Shen (2016).

A further way to deal with long-term dependencies and mitigate the effect of the vanishing gradient problem is the NARX neural network. This network, introduced by Lin et al. (1996), addresses the vanishing gradient problem by using an orthogonal mechanism with direct connections or delays from the past. Some authors (Bianchi et al., 2017) showed that NARX networks accurately predict time series with long-term dependencies, while others (Menezes and Barreto, 2006) demonstrated that this method accurately forecasts nonlinear time series.

NARX networks can be specified in a two-fold way. The first mode is called *parallel (P) architecture*, in which the output is fed back to the input of the FNN. The NARX-P architecture behaves like a JNN where, at each training epoch, the output is trained and used in the subsequent time steps (differently from JNN, this architecture relies on a greater number of lags). The second mode is called *series-parallel (SP) architecture*, here the observed output is used as additional input instead of feeding back the estimated output. The

structure is that of a regular FNN with d additional inputs equal to d delays of the real target variable.

In this article, I consider only NARX-SP networks with zero input order and a one-dimensional output. Thus, the output function of the NARX networks with zero input order is defined by

$$\hat{y}_t = \Psi[x_t, y_{t-1}, \dots, y_{t-d}] \quad (13)$$

where x_t and y_t are, respectively, the input and the output of the network at time t , d is the output order and Ψ is a multilayer perceptron as in Figure 7. This architecture can be represented by the following equation:

$$f_t(x_t, \theta) = \beta_0 + \sum_{j=1}^q G \left(x_t \gamma_j' + \sum_{d=1}^{n_d} y_{t-d} \psi_{d,j} \right) \beta_j \quad (14)$$

where $\psi_{d,j}$ is the weight associated to the d th delay of the output.

In the following section, I specify the architecture for the above models, selecting the final set of inputs, the number of hidden nodes, and the training algorithm.

3 Neural Networks Architecture

The overall task of constructing a neural network passes through a process of trial and error. Some authors, [Anders and Korn \(1996\)](#) and [Panchal et al. \(2010\)](#) among others, suggested various ways to define information criteria that could help driving the choice of the neural network architecture. However, the most reliable approach remains the training of different architectures and the choice of the network producing the lowest forecasting error.

First, the researcher should choose a set of inputs. Variable selection represents a crucial phase for the identification of the neural networks' architecture. While the initial set of determinants can be guided by the economic theory (see Section 1), a subset of these predictors should be used to reduce the number of weights to be trained (equal to $(1 + m)q + 1$) and enable algorithms to work properly. In the related literature, there are several methods to optimally detect the relevant explanatory variables. Here, I selected the variables through a Least Absolute Shrinkage and Selection Operator (LASSO) regression, introduced by [Tibshirani \(1996\)](#). This method performs estimation and model selection in the same step by penalizing the absolute size of the regression coefficients, based on a penalty coefficient, λ ; see [Zou \(2006\)](#) for the mathematical details. To assess the results of the analysis, I examined which variables really affected realized volatility for two samples: the entire sample of observations, from January 1950 to December 2017, and a subsample,³ from February 1973 to June 2009. All independent and control variables were lagged by one year to mitigate the possibility of simultaneity or reverse causality bias, while the number of lags of the dependent variables was assessed through information criteria. As further explained in the

3 This subsample was used to validate the approach in a more volatile period. The starting month of this sample has been determined through a breakpoint analysis, while the final monthly observation coincided with the end of the *Great Recession* according to the National Bureau of Economic Research.

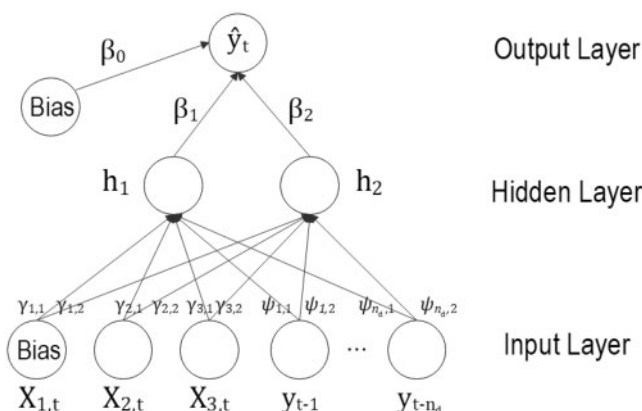


Figure 7 Architecture of an NARX network.

next paragraph, the final set of variables was selected on the training sample, and the resulting sets of variables were $X_a = \{RV_{t-1}, RV_{t-2}, RV_{t-3}, DP_{t-1}, MKT_{t-1}, STR_{t-1}, DEF_{t-1}\}$ for the entire sample, and $X_b = \{RV_{t-1}, RV_{t-2}, RV_{t-3}, MKT_{t-1}, STR_{t-1}\}$ for the subsample. The lack of significance of pure macroeconomic variables, that is, inflation rate and industrial production growth, is in line with the findings of Schwert (1989) and Christiansen, Schmeling, and Schrimpf (2012), once again underlying the relevance of premium risk’s determinants.

Choosing the set of explanatory variables entails a two-fold risk. On the one side, the so-called *look ahead bias*⁴ may occur. On the other side, the variables selected through this method, that is, LASSO, may not be relevant in a neural network framework. The choice of two samples and two different sets of explanatory variables may help alleviating these drawbacks. Moreover, the former issue was circumvented by selecting the relevant variables on the training sample (see the following section for details), where the number of observations was approximately equal to two-thirds of the entire number of observations. Furthermore, the neural networks have been implemented without macroeconomic and financial determinants, in order to understand if the lags of the dependent variable, alone, were sufficient to provide accurate forecasts.

Once a set of determinants has been identified, the researcher can proceed to select the number of hidden layers and hidden neurons. To assess the performance of an architecture, the researcher must modify the number of hidden units or by adding or removing certain network connections, and then evaluate them by comparing the MSE attained in compared architectures.

As previously mentioned, a single hidden layer was assumed throughout the article, while the selection of the optimal number of hidden neurons was trickier. Since a standard and accepted method for determining the number of hidden nodes does not exist, I evaluated the performance of the networks⁵ for each sample by the lowest training MSE for an

4 Look ahead bias involves using information not available during the period analyzed.

5 LSTM hidden units follow a different setting in comparison with other neural networks; thus, the number of hidden units is set to fifty, comparably to similar studies.

increasing number of hidden nodes, where the maximum number of hidden nodes was equal to the total number of inputs (i.e., 7 and 5, respectively), as suggested by [Tang and Fishwick \(1993\)](#). To avoid the optimization algorithm being trapped in a local minimum, the network weights were re-estimated using 300 sets of random starting values. [Table 2](#) provides the MSE for each architecture in the entire sample, while the results for the subsample are showed in [Table 3](#). Therefore, the number of hidden nodes was selected according to the lowest MSE. As in the case of explanatory variables selection, the choice of the architecture was made on the training samples.

A gradient descent with momentum and adaptive learning rate (*gdx*) BP has been used to train the feed-forward, Elman, Jordan, and LSTM architectures. Despite it converges more slowly in comparison to other algorithms, the trained weights iteratively adapt to the shape of the error surface at each iteration, reducing the risk of a local minimum. The NARX network has been trained using a Bayesian regularization (BR) algorithm, since the predictive performance of the BR algorithm is more robust when an NARX architecture is implemented, see [Guzman, Paz, and Tagert \(2017\)](#).

4 Assessing Forecast Accuracy

The forecasting ability of the ANNs was compared to an autoregressive fractionally integrated moving average with the same set of explanatory variables selected in the previous section (ARFIMAX) and without determinants (ARFIMA). The set of competing models also included a logistic smooth transition autoregressive model (LSTAR), also entailing exogenous variables (LSTARX), where the number of lags was set relying on the Akaike and the Bayesian information criteria. The analysis was performed over a period from January 1950 to December 2017 and a period from February 1973 to June 2009.

LAG selection of the ARFIMA was assessed on the training sample through information criteria. ARFIMA(0, d , 1) and ARFIMAX(0, d , 0) were selected for the larger sample, while ARFIMA(0, d , 0) and ARFIMAX(2, d , 2) were used for the subsample. In order to determine the number of regimes of the smooth transition models, the presence of structural breaks was evaluated through the method introduced by [Bai and Perron \(2003\)](#). A single structural break (and two regimes) was identified, while lagged realized volatility was used as transition variable.

The forecasting accuracy of the neural networks model was further compared with the forecasts from an HAR model for realized volatility, discussed in [Andersen, Bollerslev, and Diebold \(2007\)](#) and [Corsi \(2009\)](#). Since the realized volatility used in this article was observed monthly, we made use of quarterly and annual aggregation periods.

Realized volatility forecasts were produced on 245 out-of-sample observations (from August 1997 to December 2017) for the entire sample, while 22 out-of-sample forecasts (from September 2007 to June 2009) were produced for the subsample. The number of out-of-sample observations was equal to one-third of the entire sample in the former case, while it started from the beginning of the *Great Recession* for the latter. This helped understanding whether NNs were able to outperform econometric models in a highly volatile context and in the presence of greater persistence.

Table 2 MSE for increasing number of hidden nodes—entire sample

Model	No. of hidden	Performance	No. of weights	Model	No. of hidden	Performance	No. of weights
FNNX	1*	0.1029	10	FNN	1	0.1168	6
	2	0.1245	19		2	0.1260	11
	3	0.1062	28		3	0.1177	16
	4	0.1074	37		4	0.1171	21
	5	0.1035	46		5*	0.1136	26
	6	0.1069	58		6	0.1168	31
	7	0.1063	71		7	0.1175	36
ENNX	1	0.1027	11	ENN	1	0.1177	7
	2	0.1111	23		2	0.1479	15
	3	0.1031	37		3	0.1179	25
	4*	0.1006	53		4	0.1230	37
	5	0.1131	71		5	0.1210	51
	6	0.1070	91		6	0.1223	67
	7	0.1115	113		7*	0.1170	85
JNNX	1	0.1006	11	JNN	1	0.1165	7
	2	0.1236	21		2	0.1315	13
	3*	0.1004	31		3	0.1148	19
	4	0.1044	41		4	0.1158	25
	5	0.1040	51		5*	0.1147	31
	6	0.1062	61		6	0.1152	37
	7	0.1085	71		7	0.1154	43
NARX	1	0.0990	10	NAR	1	0.1141	6
	2	0.0981	19		2	0.1123	11
	3	0.0978	28		3	0.1160	16
	4	0.0968	37		4	0.1123	21
	5*	0.0953	46		5	0.1110	26
	6	0.0967	55		6	0.1113	31
	7	0.0966	64		7*	0.1100	36

Notes: The table includes the number of hidden nodes, the performance in terms of MSE, and the number of weights trained for each architecture. Each architecture has a maximum of iterations equal to 1000. The presence of the *X* in the name of the model indicates the use of exogenous determinants other than lagged realized variance. The asterisk denotes the selected number of hidden nodes. Values in boldface represent the lowest MSE.

The one-step-ahead ($k = 1$) out-of-sample forecasts were generated from a rolling window scheme, re-estimating the parameters at each step. In addition, multistep-ahead forecasts have been considered. The five-step-ahead ($k = 5$) forecasts were iteratively produced from a rolling window estimation. At each step ahead, the information was updated with the prediction of the previous step. The resulting set of variables used to make the forecasts five-step ahead is the following:

Table 3 MSE for increasing number of hidden nodes—subsample

Model	No. of hidden	Performance	No. of weights	Model	No. of hidden	Performance	No. of weights
FNNX	1	0.1450	8	FNN	1	0.1534	6
	2	0.1745	15		2	0.1751	11
	3*	0.1196	22		3*	0.1474	16
	4	0.1467	29		4	0.1491	21
	5	0.1534	36		5	0.1484	26
ENNX	1	0.1192	9	ENN	1	0.1298	7
	2	0.1693	19		2	0.1667	15
	3	0.1158	31		3	0.1425	25
	4*	0.1126	45		4*	0.1420	37
	5	0.1205	61		5	0.1471	51
JNNX	1	0.1201	9	JNN	1*	0.1425	7
	2	0.1289	17		2	0.1541	13
	3*	0.1115	25		3	0.1494	19
	4	0.1176	33		4	0.1498	25
	5	0.1169	41		5	0.1478	31
NARX	1	0.1111	8	NAR	1	0.1145	6
	2	0.1009	15		2	0.1291	11
	3	0.1119	22		3*	0.1056	16
	4*	0.0998	29		4	0.1177	21
	5	0.1020	36		5	0.1158	26

Notes: The table includes the number of hidden nodes, the performance in terms of MSE, and the number of weights trained for each architecture. Each architecture has a maximum of iterations equal to 1000. The presence of the X in the name of the model indicates the use of exogenous determinants other than a lagged variance. The asterisk denotes the selected number of hidden nodes. Values in boldface represent the lowest MSE.

$$\hat{y}_{t+1} = \{y_t, y_{t-1}, y_{t-2}, z_t\}$$

$$\hat{y}_{t+2} = \{\hat{y}_{t+1}, y_t, y_{t-1}, z_t\}$$

$$\hat{y}_{t+3} = \{\hat{y}_{t+2}, \hat{y}_{t+1}, y_t, z_t\}$$

$$\hat{y}_{t+4} = \{\hat{y}_{t+3}, \hat{y}_{t+2}, \hat{y}_{t+1}, z_t\}$$

$$\hat{y}_{t+5} = \{\hat{y}_{t+4}, \hat{y}_{t+3}, \hat{y}_{t+2}, z_t\},$$

where $z_t = \{DP_{t-1}, MKT_{t-1}, STR_{t-1}, DEF_{t-1}\}$ in the entire sample, and $z_t = \{MKT_{t-1}, STR_{t-1}\}$ in the subsample. The set of input variables used in models ARFIMA, LSTAR, HAR, FNN, ENN, JNN, LSTM, and NAR did not include z_t .

The relative performance of the out-of-sample forecasting accuracy was assessed using MSE and the quasi-likelihood (QLIKE), which belong to the family of loss functions robust to a noisy volatility proxy; see Patton (2011). The predictive performance of the competing models was also simultaneously compared via *model confidence set* (MCS), introduced by Hansen, Lunde, and Nason (2011). The MCS procedure consists in a sequence of equal predictive accuracy tests through which a set of superior models (SSM) is defined, given a certain confidence level. For a set of forecasts from M models, MCS tests, through a pairwise

comparison of loss difference $d_{l,j,t}$ from model l and model j , whether all models provide equal predictive accuracy. Assuming $d_{l,j,t}$ stationary, the null hypothesis assumes the following form:

$$H_0 : E[d_{l,j,t}] = 0, \quad \forall l, j \in M. \quad (15)$$

Given a confidence level α , a model is discarded when the null hypothesis of equal forecasting ability is rejected. The SSM is then defined as the set of models not-rejecting the null hypothesis. The p -values were computed using a stationary bootstrap. The lower the p -value of an object, the lower the probability of being included in the SSM, see Hansen, Lunde, and Nason (2011) for further details.

As shown by the average of the loss functions in Table 4, the majority of the neural networks were able to outperform the traditional long-memory detecting models from Panel A in the larger sample, when analyzing forecasts for $k = 1$. The unique model exhibiting an MSE and QLIKE in line with NNs, or better in some comparisons with Panel B, was the HAR model. In most cases, the exclusion of the explanatory variables worsened the forecasting accuracy, highlighting that the dynamics of realized volatility are somehow linked to macroeconomic and financial conditions. The lowest loss functions were provided by the models in Panel C, where the long-memory neural networks were stored. The best overall performance in terms of forecasting accuracy measures was exhibited by LSTMX and NARX in Panel C, which exhibited the lowest MSE and QLIKE. NARX was the unique model belonging to the 75% MCS ($\hat{M}_{75\%}^*$), regardless of the loss function considered. Even though, the average loss functions varied differently according to the kind of loss considered. For example, in Panel C the QLIKE function was almost equal for all the models, while MSE was severely smaller in the models entailing the use of macroeconomic and financial variables. This may be driven by the definition of the loss functions. In fact, MSE is a symmetric measure, while QLIKE penalizes more negative biases. In this context, all the models seemed to underestimate the observed RV (see also Figures 8 and 9), meaning that the differences between real RV and the forecasts tended to be positive and the QLIKE more in line among compared models.

The analysis of multistep-ahead forecasts, in the entire sample, further highlighted the predictive ability of long-term memory detecting RNNs. In fact, NARX and LSTMX neural networks exhibited the lowest MSE and QLIKE, excluding the overall best performance of the HAR model. Compared to one-step-ahead forecasts, the differences in terms of average losses in multistep-ahead forecasts were much less pronounced between the models compared. This emerged also from the p -values, which assumed values not close to zero for almost all the models, and from the fact that the majority of the models was included in the 75% MCS ($\hat{M}_{75\%}^*$). An explanation of these results derives from the nature of the forecasts, since in multistep-ahead forecasts, the information used is the same for repeated steps, flattening the results from models highly different. The average losses of the models in Panel A were higher than the most part of the models in Panels B and C, with the only exception of HAR model's losses and the QLIKE loss for the ARFIMAX model. Finally, among the models belonging to 75% MCS ($\hat{M}_{75\%}^*$), the p -values were higher for HAR model, JNNs, LSTMX, NAR, and NARX neural networks, both for MSE and QLIKE.

In the more volatile period of time, September 2007–June 2009, in Table 5, classical long-memory detecting models in Panel A seemed not accurate in terms of one-step-ahead

Table 4 MCS with 10,000 bootstraps (entire sample: 1997.08–2017.12)

Model	$k = 1$				$k = 5$			
	MSE		QLIKE		MSE		QLIKE	
	Loss	P_{MCS}	Loss	P_{MCS}	Loss	P_{MCS}	Loss	P_{MCS}
Panel A: time series models								
ARFIMAX	0.167	0.000	3.317	0.000	0.185	0.683**	3.323	0.867**
ARFIMA	0.205	0.000	3.325	0.000	0.219	0.201*	3.333	0.147*
LSTARX	0.145	0.000	3.316	0.000	0.323	0.000	3.352	0.000
LSTAR	0.130	0.000	3.311	0.000	0.245	0.205*	3.332	0.267**
HAR	0.115	0.000	3.308	0.000	0.102	1.000**	3.306	1.000**
Panel B: FFN, JNN, and ENN								
FNNX	0.132	0.000	3.313	0.000	0.178	0.680**	3.325	0.212*
FNN	0.130	0.000	3.311	0.000	0.176	0.757**	3.325	0.308**
ENNX	0.133	0.000	3.313	0.000	0.164	1.000**	3.321	1.000**
ENN	0.138	0.000	3.312	0.000	0.172	0.992**	3.322	0.872**
JNNX	0.136	0.000	3.314	0.000	0.158	1.000**	3.315	1.000**
JNN	0.134	0.000	3.312	0.000	0.161	1.000**	3.314	1.000**
Panel C: long-term dependence detecting neural networks								
LSTMX	0.042	0.005	3.293	0.004	0.152	1.000**	3.313	1.000**
LSTM	0.110	0.000	3.307	0.000	0.185	0.639*	3.327	0.025
NARX	0.018	1.000**	3.288	1.000**	0.146	1.000**	3.316	1.000**
NAR	0.075	0.000	3.301	0.003	0.164	1.000**	3.317	1.000**

Notes: This table reports the average loss over the evaluation sample and the MCS p -values calculated on the basis of the range statistics. The realized volatility forecasts in $\hat{M}_{90\%}^*$ and $\hat{M}_{75\%}^*$ are identified by one and two asterisks, respectively. Values in boldface represent the lowest average losses. The set of input variables used in models ARFIMA, LSTAR, HAR, FNN, ENN, JNN, LSTM, and NAR did not include exogenous variables other than the lags of the dependent variable.

forecasts. Nevertheless, ARFIMAX model exhibited an average MSE and QLIKE in line with the compared neural networks. Once again, the lowest loss functions were exhibited by two models from Panel C, that is, LSTM and NARX. From a theoretical point of view, this result is not surprising, given that the LSTM has shown stronger performance in similar works in presence of long dependencies (Heaton, Polson, and Witte, 2016; Pichl and Kaizoji, 2017). Furthermore, the prediction differences between neural networks and linear models may indicate a nonlinear behavior of the log-realized variance during financially stressed periods; see also Choudhry, Papadimitriou, and Shabi (2016). In this volatile framework, the five-step-ahead forecasts provided mixed results. All the models were included in the 90% MCS when the MSE was used as loss function, while only two models, LSTARX and LSTM, were not included in the 75% MCS ($\hat{M}_{75\%}^*$). A similar result was obtained with QLIKE as a loss function. Among models without exogenous variables, FNN exhibited the lowest average losses, while the most part of the simple neural networks from Panel B showed average losses lower than the more complex networks from Panel C. This may imply that a simpler model should be implemented when few multistep-ahead forecasts need to be produced.

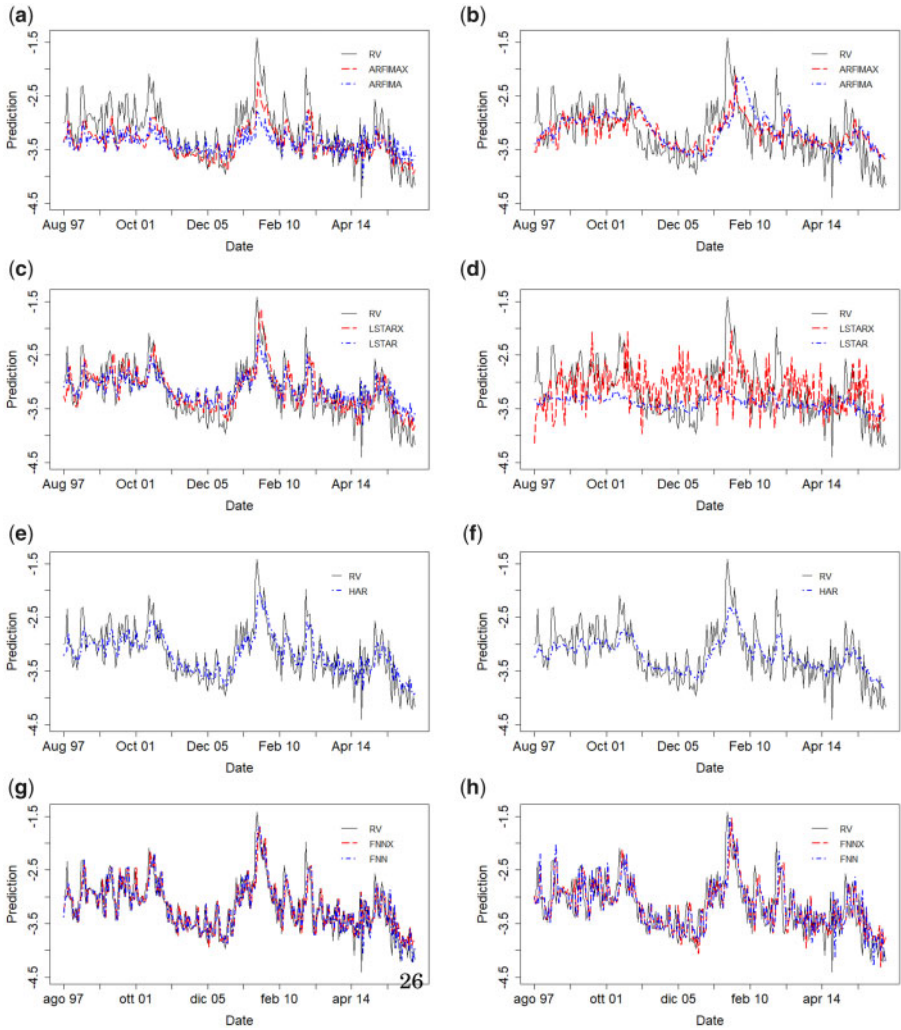


Figure 8 Entire sample forecasts comparison. The black line represents the realized volatility of S&P 500, the red dashed line describes the out-of-sample forecasts with exogenous variables, while the blue dashed line depicts the forecasts from a model without financial and macroeconomic variables. The left panel entails the one-step-ahead forecasts, while five-step-ahead forecasts are showed in the right column. (a) ARFIMA one-step-ahead forecasts. (b) ARFIMA five-step-ahead forecasts. (c) LSTAR one-step-ahead forecasts. (d) LSTAR five-step-ahead forecasts. (e) HAR one-step-ahead forecasts. (f) HAR five-step-ahead forecasts. (g) FNN one-step-ahead forecasts. (h) FNN five-step-ahead forecasts. (i) ENN one-step-ahead forecasts. (j) ENN five-step-ahead forecasts. (k) JNN one-step-ahead forecasts. (l) JNN five-step-ahead forecasts. (m) LSTM one-step-ahead forecasts. (n) LSTM five-step-ahead forecasts. (o) NARX one-step-ahead forecasts. (p) NARX five-step-ahead forecasts.

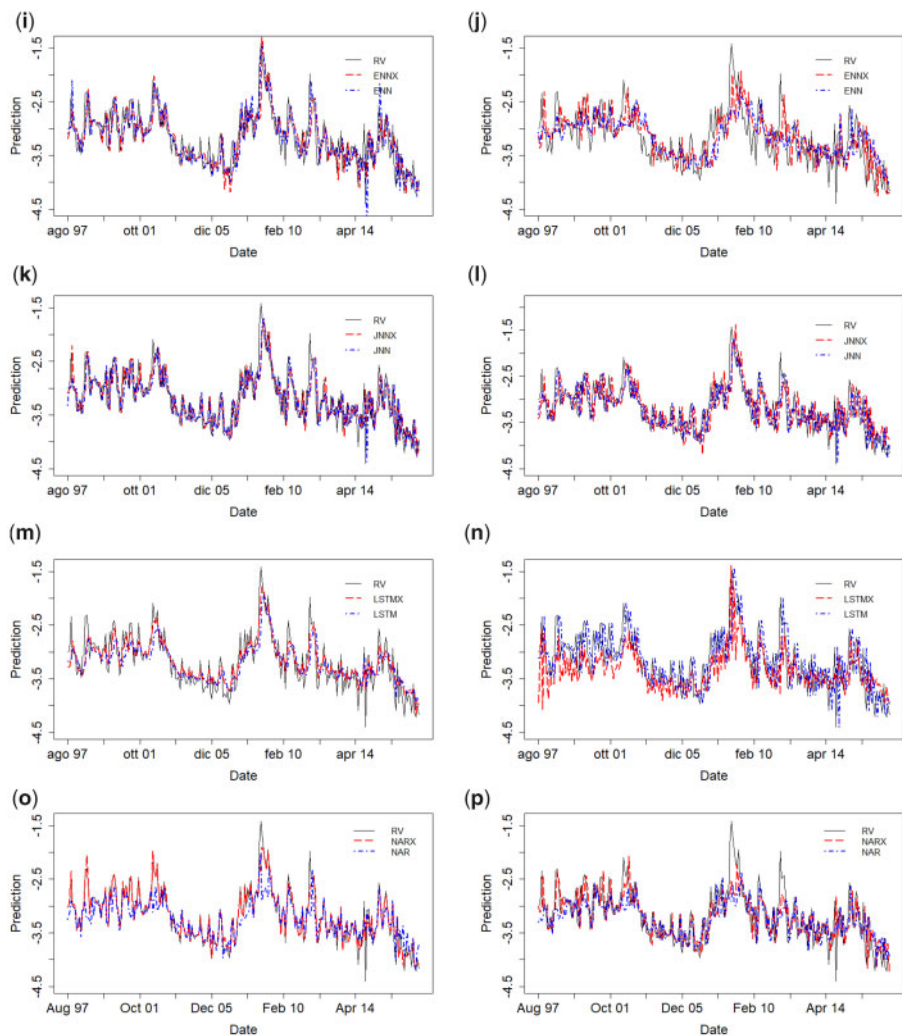


Figure 8 Continued

Additionally, the test of equal predictive accuracy of Diebold–Mariano (DM) (Diebold and Mariano, 1995) was used as a robustness check. The pairwise comparison test in Table 6 supported our previous findings when $k=1$ in both the samples, since the null of equal forecast accuracy was rejected only for long-memory detecting models in Panel C and for the HAR model. Thus, HAR, LSTM, and NARX models seemed to be the unique models predicting realized volatility one-step-ahead forecast better than the simple random walk model. A similar result was observed in the subsample when $k=1$, where the forecast accuracy of the NARX and LSTM models was the unique significantly different from a random walk. When $k=5$, neural networks model was able to significantly outperform the predictive accuracy of the benchmark method (i.e., $\hat{y}_{t+5} = y_t$), but the test statistics were

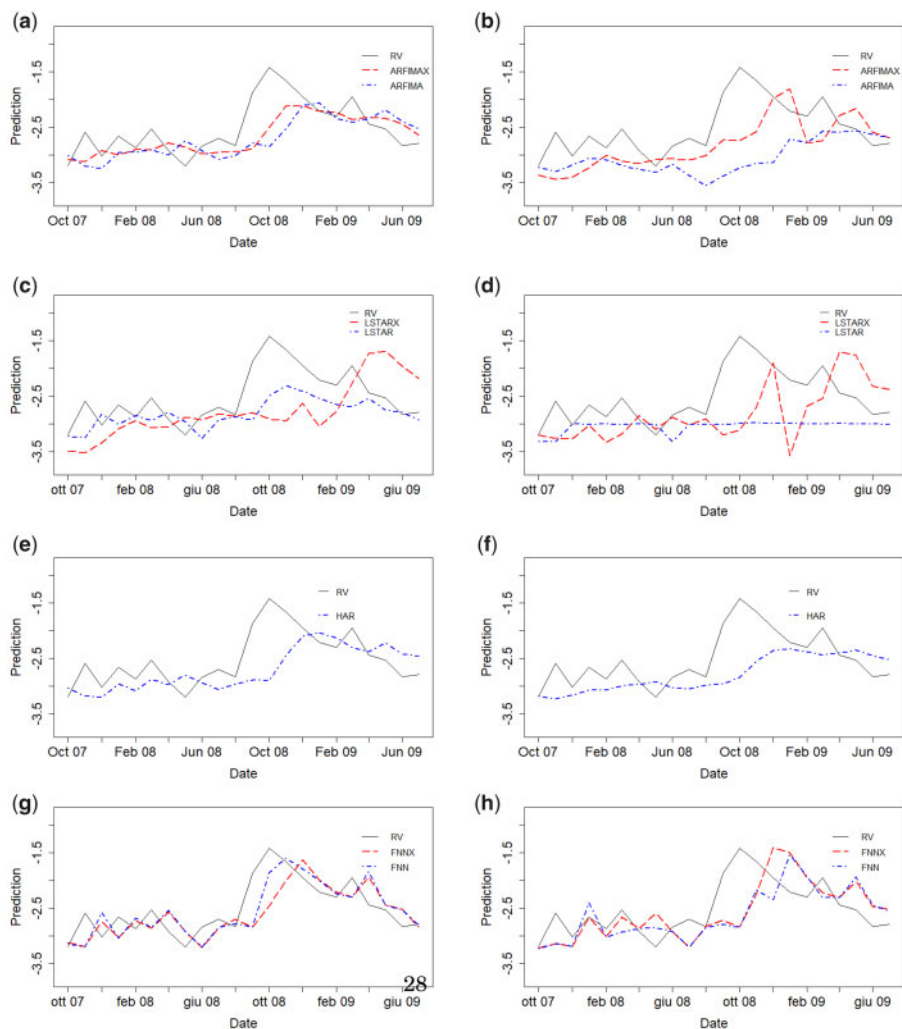


Figure 9 Subsample sample forecasts comparison. The black line represents the realized volatility of S&P 500, the red dashed line describes the out-of-sample forecasts with exogenous variables, while the blue dashed line depicts the forecasts from a model without financial and macroeconomic variables. The left panel entails the one-step-ahead forecasts, while five-step-ahead forecasts are shown in the right column. (a) ARFIMA one-step-ahead forecasts. (b) ARFIMA five-step-ahead forecasts. (c) LSTAR one-step-ahead forecasts. (d) LSTAR five-step-ahead forecasts. (e) HAR one-step-ahead forecasts. (f) HAR five-step-ahead forecasts. (g) FNN one-step-ahead forecasts. (h) FNN five-step-ahead forecasts. (i) ENN one-step-ahead forecasts. (j) ENN five-step-ahead forecasts. (k) JNN one-step-ahead forecasts. (l) JNN five-step-ahead forecasts. (m) LSTM one-step-ahead forecasts. (n) LSTM five-step-ahead forecasts. (o) NARX one-step-ahead forecasts. (p) NARX five-step-ahead forecasts.

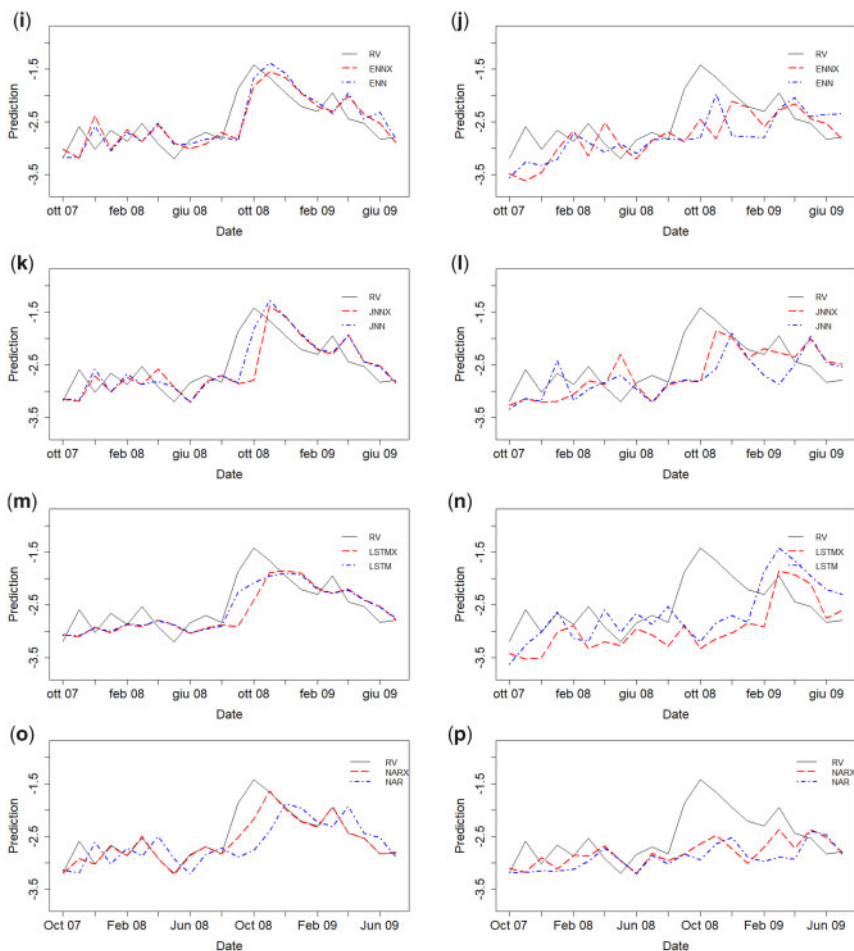


Figure 9 Continued

comparable among the neural networks. Still, in the entire sample, the greatest significant difference between the forecasts of the benchmark and the models compared was found with the HAR (with a DM statistics of 7.428) and the NARX models, which provided a test statistic of 4.772 and 4.758. In the subsample, as already assessed by the MCS in Table 5, the test statistics were mostly comparable among the models.

Finally, the forecasts were compared pairwise via encompassing tests, discussed in [Fair and Shiller \(1989\)](#) and [Chong and Hendry \(1986\)](#). The idea behind encompassing tests is that, given the realized variable, y_{t+k} , and two sets of forecasts of the variable, $\hat{y}_{1,t+k}$ and $\hat{y}_{2,t+k}$, y_{t+k} can be explained by pooling the forecasts as follows:

$$y_{t+k} = \alpha_0 + \alpha_1 \hat{y}_{1,t+k} + \alpha_2 \hat{y}_{2,t+k} + u_t.$$

Table 5 MCS with 10,000 bootstraps (subsample: 2007.09–2009.06)

Model	$k = 1$				$k = 5$			
	MSE		QLIKE		MSE		QLIKE	
	Loss	P_{MCS}	Loss	P_{MCS}	Loss	P_{MCS}	Loss	P_{MCS}
Panel A: time series models								
ARFIMAX	0.166	0.998**	2.857	1.000**	0.295	1.000**	2.890	1.000**
ARFIMA	0.244	0.306**	2.878	0.546**	0.571	0.593**	2.939	0.890**
LSTARX	0.467	0.000	2.957	0.000	0.503	0.121*	2.947	0.134*
LSTAR	0.221	0.000	2.872	0.449**	0.488	0.440**	2.930	0.813**
HAR	0.247	0.453**	2.878	0.653**	0.263	1.000**	2.885	1.000**
Panel B: FFN, JNN, and ENN								
FNNX	0.176	0.103	2.864	0.147*	0.259	1.000**	2.890	1.000**
FNN	0.138	1.000**	2.849	1.000**	0.234	1.000**	2.874	1.000**
ENNX	0.142	1.000**	2.850	1.000**	0.286	1.000**	2.883	1.000**
ENN	0.143	1.000**	2.853	1.000**	0.283	1.000**	2.887	1.000**
JNNX	0.216	0.309**	2.875	0.295**	0.244	1.000**	2.879	1.000**
JNN	0.137	1.000**	2.855	1.000**	0.290	1.000**	2.892	1.000**
Panel C: long-term dependence detecting neural networks								
LSTMX	0.151	1.000**	2.853	1.000**	0.533	0.426**	2.936	0.688**
LSTM	0.080	1.000**	2.833	1.000**	0.460	0.148*	2.952	0.137*
NARX	0.052	1.000**	2.824	1.000**	0.266	1.000**	2.884	1.000**
NAR	0.237	0.170*	2.878	0.260**	0.358	1.000**	2.905	0.740**

Notes: This table reports the average loss over the evaluation sample and the MCS p -values calculated on the basis of the range statistics. The realized volatility forecasts in $\hat{M}_{90\%}^*$ and $\hat{M}_{75\%}^*$ are identified by one and two asterisks, respectively. Values in boldface represent the lowest average losses.

Forecasts encompassing can be tested by analyzing the significance of the coefficients. For example, if the hypothesis $\alpha_1 = 0$ is rejected, while $\alpha_2 = 0$ not, the first model encompasses the second one. Moreover, if the null hypothesis $\alpha_2 = 0$ and $\alpha_1 = 1$ cannot be rejected, the forecasts from model 1 are also unbiased. Thus, ideally a model should have a coefficient close to or greater than 1 to significantly encompass the compared model.

Since the number of out-of-sample forecasts is limited in the subsample, only the forecasts from the entire sample were compared via encompassing tests in Tables 7 and 8.

The comparison of one-step-ahead forecasts highlighted that the coefficient of the HAR in encompassing regressions is close to 1 when included in a regression with forecasts from models in Panels A and B; thus, the HAR forecast encompasses those forecasts. For instance, if we regress RV on the HAR and the FNN forecasts, we get a (statistically significant) coefficient on HAR of 0.845 (from the FNN column and HAR row), while we get a coefficient on the FNN forecast of 0.146 (from the HAR column and FNN row), suggesting that the HAR forecast encompasses the FNN forecast. In contrast, the HAR forecast is encompassed by the LSTM, LSTMX, and NARX forecasts. From Table 7, NARX emerges as the unique method able to encompass the other forecasts and with a (statistically significant) coefficient always near to 1. All the models from Panel C encompassed the forecasts from models in Panels A and B, HAR excluded.

Table 6 DM test of equal predictive accuracy

Model	Entire sample		Subsample	
	$k = 1$	$k = 5$	$k = 1$	$k = 5$
Panel A: time series models				
ARFIMAX	-1.714*	4.284**	1.648	2.527**
ARFIMA	-3.490***	2.660***	0.130	0.025
LSTARX	-0.182	-1.953*	-2.518**	0.462
LSTAR	1.308	0.812	0.470	0.755
HAR	3.100***	7.428***	0.064	2.731**
Panel B: FFN, JNN, and ENN				
FNNX	1.454	3.552***	1.343	3.602***
FNN	2.457**	3.523***	1.241	3.663***
ENNX	0.778	4.258***	1.191	2.542**
ENN	-0.232	4.015***	1.129	2.557**
JNNX	0.203	4.009***	0.904	2.466**
JNN	0.496	4.044***	1.251	2.480**
Panel C: long-term dependence detecting neural networks				
LSTMX	7.647***	4.457***	1.761*	0.271
LSTM	2.713***	3.075***	2.264**	0.867
NARX	9.179***	4.772***	3.057***	3.010***
NAR	7.707***	4.758***	0.276	1.788*

Notes: This table reports the t -statistics for the DM test where the null hypothesis is the equivalence of the predictive accuracy of the compared models with the information available at time t (i.e., $\hat{y}_{t+h} = y_t$). *, **, and *** indicate a significant difference between the forecasting abilities at 1%, 5%, and 10% level, respectively. A positive and statistically significant difference means that the model in the line predicts better than simply using y_t .

When $k = 5$, the results were mixed. There was less clear discrimination between methods, with more evidence of some value to combining forecasts. If we had to pick a single model from Table 8, it would be the HAR model.

The graphical representation of the out-of-sample forecasts in the two samples was provided in Figures 8 and 9.

In Figure 8, the ARFIMA models both underestimated the RV during the sub-prime crises in sub-figure (a), while the five-step ahead forecasts were in line with the trend of RV. LSTAR models were more precise than ARFIMA when $k = 1$, especially during the financial crisis. This was not true for the five-step-ahead forecasts that were inaccurate both with and without exogenous variables. HAR model seemed to slightly underestimate RV both for $k = 1$ and $k = 5$. The graphical representations of the one-step-ahead forecasts from FNN, ENN, and JNN almost replicated the observed RV, both in the model with and without determinants. Furthermore, FNNs seemed to be precise also in five-step-ahead forecasts, while just JNNX's representation was in line with the observed RV. LSTM one-step-ahead forecasts seemed to be smoother than the observed RV, while the same forecasts from the NARX model (i.e., the red line in Figure 8, letter o) were almost overlapped in the period 1997–2010. In the five-step-ahead forecasts, FNN and JNN models provided precise forecasts, while ENNs were less accurate. Finally, LSTM multistep-ahead forecasts were in

Table 7 Encompassing tests—entire sample ($k = 1$)

Model	Panel A: time series models										Panel B: FNN, JNN, and ENN										Panel C: long-term dependence detecting neural networks									
	ARFIMAX	ARFIMA	LSTARX	LSTAR	HAR	FNNX	FNN	ENN	JNNX	JNN	LSTMX	LSTM	NARX	NAR	ARFIMAX	ARFIMA	LSTARX	LSTAR	HAR	FNNX	FNN	ENN	JNNX	JNN	LSTMX	LSTM	NARX	NAR		
Panel A: time series models																														
ARFIMAX	—	0.921***	0.911***	0.480**	-0.137	0.555***	0.396*	0.453**	0.378*	0.600***	0.547***	-0.786***	-0.162	0.403***	0.471***	0.760***	0.592***	0.303	0.039	—	0.053	0.334**	0.312**	0.326*	0.398*	-0.434***	0.108	-0.097***	0.253***	
ARFIMA	0.587***	—	0.988***	-0.603*	0.073	0.115	-0.137	0.409**	0.097	0.102	0.069	-0.394***	0.317*	0.484***	0.556***	0.840***	0.603***	0.561**	0.141	0.738***	—	0.419***	0.371**	0.759***	-0.455***	0.168	-0.084***	0.254**		
LSTARX	0.275**	0.576***	—	0.333***	0.017	0.297***	0.270***	0.192	0.244*	0.324**	0.304**	-0.543***	-0.332*	0.241***	0.510***	0.632***	0.632***	0.426***	0.199	0.472***	0.384***	—	0.502**	0.468***	-0.400***	0.132	-0.083***	0.270***		
LSTAR	0.808***	1.638***	0.908***	—	0.154	0.801**	0.372	0.604***	0.454	0.940***	0.805**	-0.641***	0.292	0.426***	0.522***	0.691***	0.569***	0.476***	0.208	0.468**	0.402**	0.412**	—	0.495***	0.479***	-0.385***	0.184	-0.051*	0.247***	
HAR	1.028***	0.984***	1.001***	0.901***	—	0.971***	0.845**	0.785***	0.743***	0.976***	0.959***	-0.775***	0.138	0.426***	0.428***	0.739**	0.556***	0.205	0.035	0.275	0.030	0.291**	0.271**	—	0.389***	0.099	-0.108***	0.241***		
Panel B: FNN, JNN, and ENN																														
FNNX	0.471***	0.760***	0.592***	0.303	0.039	—	0.053	0.334**	0.312**	0.326*	0.398*	-0.434***	0.108	0.253***	0.461***	0.754***	0.571**	0.291	0.048	0.404*	0.051	0.328**	0.286**	0.495**	-0.401***	0.102	-0.092***	0.255***		
FNN	0.556***	0.840***	0.603***	0.561**	0.141	0.738***	—	0.419***	0.371**	0.759***	0.739**	-0.455***	0.168	0.254**	0.510***	0.632***	0.632***	0.426***	0.199	0.472***	0.384***	—	0.502**	0.468***	-0.400***	0.132	-0.083***	0.270***		
ENN	0.510***	0.632***	0.632***	0.426***	0.199	0.472***	0.384***	—	0.352*	0.502**	0.468***	-0.400***	0.132	0.270***	0.522***	0.691***	0.569***	0.476***	0.208	0.468**	0.402**	0.412**	—	0.495***	0.479***	-0.385***	0.184	-0.051*	0.247***	
JNNX	0.428***	0.739**	0.556***	0.205	0.035	0.275	0.030	0.291**	0.271**	—	0.293	-0.389***	0.099	0.241***	0.428***	0.739**	0.556***	0.205	0.035	0.275	0.030	0.291**	0.271**	—	0.389***	0.099	-0.108***	0.241***		
JNN	0.461***	0.754***	0.571**	0.291	0.048	0.404*	0.051	0.328**	0.286**	0.495**	—	0.401***	0.102	0.255***	0.461***	0.754***	0.571**	0.291	0.048	0.404*	0.051	0.328**	0.286**	0.495**	-0.401***	0.102	-0.092***	0.255***		
Panel C: long-term dependence detecting neural networks																														
LSTMX	1.734***	1.369***	1.673***	1.633***	1.885***	1.654**	1.697**	1.648***	1.662***	1.618***	1.632***	—	-1.973**	0.293**	1.734***	1.369***	1.673***	1.633***	1.885***	1.654**	1.697**	1.648***	1.662***	1.618***	1.632***	—	-1.973**	0.293**	0.899***	
LSTM	1.219***	0.955***	1.431**	0.873**	0.951**	0.965**	0.886**	0.924**	0.845**	0.971**	0.967**	-0.908**	—	0.399***	1.219***	0.955***	1.431**	0.873**	0.951**	0.965**	0.886**	0.924**	0.845**	0.971**	0.967**	-0.908**	-0.070*	0.399***		
NARX	1.097***	1.089***	1.098***	1.111***	1.121***	1.124***	1.117**	1.117**	1.095**	1.134***	1.122***	0.826***	1.094**	0.918***	1.097***	1.089***	1.098***	1.111***	1.121***	1.124***	1.117**	1.117**	1.095**	1.134***	1.122***	0.826***	—	0.918***		
NAR	1.034***	1.091***	1.086***	1.1019***	1.010***	1.035**	1.022**	1.010***	1.009***	1.039***	1.028**	0.424**	0.970**	—	1.034***	1.091***	1.086***	1.1019***	1.010***	1.035**	1.022**	1.010***	1.009***	1.039***	1.028**	0.424**	0.210***	—		

Notes: The table presents the test statistics on the null hypothesis that the coefficient of the model in the row is equal to zero in the regression $\hat{y}_{i,t+1} = \alpha_0 + \alpha_1 \hat{y}_{i,t+1} + \alpha_2 \hat{y}_{i,t+1} + u_{i,t}$, where $\hat{y}_{i,t+1}$ is the forecast from the i th model in the row and $\hat{y}_{i,t+1}$ is the forecast from the model in the j th column. The forecast evaluation period covers August 1997–December 2017 ($N = 245$). *, **, and *** indicate the significance at 1%, 5%, and 10% level, respectively, respectively.

Table 8 Encompassing tests—entire sample ($k = 5$)

Model	Panel A: time series models					Panel B: FNN, JNN, and ENN					Panel C: long-term dependence detecting neural networks				
	ARFIMAX	ARFIMA	LSTARX	LSTAR	HAR	FNNX	FNN	ENNX	ENN	JNNX	JNN	LSTMX	LSTM	NARX	NAR
Panel A: time series models															
ARFIMAX	-	1.036***	0.917***	0.649***	-0.405***	0.453***	0.435***	0.302**	0.324**	0.271**	0.232**	0.325***	0.430***	0.361***	0.377***
ARFIMA	-0.086	-	0.604***	0.400***	-0.385***	0.248***	0.238***	0.028	-0.012	0.164**	0.132*	0.201***	0.242***	0.201**	0.167*
LSTARX	0.069	0.100	-	0.142**	0.009	0.132*	0.118*	0.109	0.117*	0.122**	0.131**	0.048	0.116*	0.148**	0.115*
LSTAR	1.746***	1.944***	-	0.532**	0.009	1.371***	1.331***	1.316***	1.494***	1.139***	1.028***	1.351***	1.313***	1.143***	1.279***
HAR	1.654***	1.678***	1.371***	1.256***	-	1.584***	1.528***	0.152	1.688***	1.640***	1.449***	1.115***	1.537***	1.546***	1.603***
Panel B: FNN, JNN, and ENN															
FNNX	0.524***	0.587***	0.662***	0.504***	-0.168**	-	0.071	0.307***	0.438***	0.255***	0.263***	0.337***	0.015	0.332***	0.399***
FNN	0.529***	0.586***	0.653***	0.508***	-0.118	0.603***	-	0.358***	0.450***	0.285***	0.276***	0.351***	0.354	0.357***	0.413***
ENNX	0.613***	0.728***	0.726***	0.564***	-0.124	0.465***	0.401***	-	0.497***	0.345***	0.309***	0.413***	0.403***	0.390***	0.443***
ENN	0.596***	0.766***	0.739***	0.558***	-0.289***	0.438***	0.414**	0.378***	-	0.121	0.084	0.326***	0.413***	0.263**	0.346***
JNNX	0.705***	0.745***	0.793***	0.657***	-0.204*	0.615***	0.584**	0.577***	0.720***	-	0.287**	0.440***	0.587***	0.687***	0.681***
JNN	0.736***	0.772***	0.810***	0.683***	-0.057	0.638***	0.616***	0.615***	0.764***	0.560***	-	0.476***	0.620***	0.708***	0.754***
Panel C: long-term dependence detecting neural networks															
LSTMX	0.832***	0.880***	0.952***	0.797***	0.264***	0.727***	0.707***	0.710***	0.753***	0.592***	0.549***	-	0.708***	0.656***	0.721***
LSTM	0.498**	0.551**	0.615***	0.477***	-0.117*	0.617***	0.304	0.334**	0.423***	0.263***	0.252**	0.333**	-	0.331***	0.386***
NARX	0.695***	0.754***	0.821***	0.660***	-0.150	0.581***	0.554**	0.552***	0.639***	0.141	0.142	0.409***	0.554***	-	0.603***
NAR	0.769***	0.866***	0.940***	0.726***	-0.251*	0.616***	0.589***	0.566***	0.645***	0.197	0.109	0.391***	0.587***	0.326**	-

Notes: The table presents the coefficient of the first model in the regression $y_{i,t+s} = \alpha_0 + \alpha_1 \hat{y}_{i,t+s} + \alpha_2 \hat{y}_{i,t+s} + u_t$, where $\hat{y}_{i,t+s}$ is the multistep ahead forecast from the i th model in the row and $\hat{y}_{i,t+s}$ is the forecast from the model in the j th column. The forecast evaluation period covers August 1997–December 2017 ($N = 245$). *, **, and *** indicate the significance at 1%, 5%, and 10% level, respectively.

line with the forecasts from other neural networks, while NARX models forecasts, despite a little underestimation during volatility peaks, almost replicated the real trend of RV.

In Figure 9, the representations of the out-of-sample forecasts in the subsample were heterogeneous. One-step-ahead forecasts from ARFIMA, LSTAR, and HAR models appeared to be smoother than the observed RV in the subsample, while the five-step-ahead forecasts from LSTAR models were highly inaccurate. When $k = 1$, the forecasts from FNN, ENN, JNN, and NARX seemed to almost overlap with the observed RV. Excluding FNNs and JNNX, the neural networks seemed instead to be less precise when $k = 5$, especially in the volatility peak in October 2008.

The analysis here presented has a two-fold implication. On the one side, the neural networks outperformed the compared classical methods in both the sample analyzed. In particular, the forecasts from long-term detecting networks proved to be the most accurate among the compared methods. This is not surprising given the well-known long memory of the realized volatility and the need for a model which accounts for that feature. On the other side, the use of macroeconomic and financial variables to make predictions increased the forecasting accuracy, even in the more complex models, although a causal effect cannot be assessed with neural networks.

5 Conclusions

In this article, a flexible nonlinear tool for forecasting volatility has been applied. The purpose of the article was to understand whether ANNs were able to capture linear and nonlinear relations and provide more accurate forecasts than traditional econometric methods. The target variable to be forecast was the logarithm of realized volatility, while the models included also macroeconomic and financial variables as determinants.

The most attractive feature of ANNs is that, by modifying the structure of the network, any linear and nonlinear function can be approximated. Moreover, in comparison with traditionally employed nonlinear time series model, such as smooth transition autoregressive model and threshold autoregressive model, they do not necessitate the knowledge of the number of regimes to be trained and require a minor computational effort in the estimation of the parameters.

Out-of-sample comparisons indicated that neural networks provide significant benefits in predicting relations expected to be nonlinear, such as between realized volatility and its determinants.

In a comparison of feed-forward and RNNs with traditional econometric methods, the best performing models appeared to be LSTM and NARX neural networks. The results further showed that these long-term dependence detecting models consistently outperformed competing for neural networks, like FNN, ENN, and JNN. The superior forecasting ability of LSTM and NARX was also assessed in a period where the stock market volatility was particularly high, like the recent financial crisis.

Since a researcher is often interested in producing forecasts for a horizon greater than one, multistep-ahead recursive forecasts were further compared. Among the main results, it emerged that long-term memory detecting neural networks had good performance when a large sample is analyzed, and provided comparable performance with other methods when a smaller and more volatile sample was evaluated.

Interestingly, in this article, realized volatility was predicted accurately well in a nonlinear framework. Several papers (McAleer and Medeiros, 2008; Hillebrand and Medeiros, 2010) showed that nonlinear models were not able to improve forecasting accuracy of realized volatility when compared with linear models. There could be several reasons that explain the findings of this manuscript. First, the analysis of out-of-sample forecasts proved that the models significantly able to outperform the linear models were the long-term dependence detecting models, that is, LSTM and NARX. So far, this seems to be the first attempt to implement such approaches to predict realized volatility. Moreover, the results highlighted the usefulness of the macroeconomic and financial variables in forecasting the realized volatility of S&P 500 index. Although there are many papers on forecasting realized volatility through a linear model, the literature on analyzing the role of volatility determinants in a nonlinear framework is still scarce.

Although appealing, there are still some issues concerning neural networks. The number of parameters to be trained can be extremely high even with a limited number of input variables. This is a stark contrast to the number of parameters of an ARFIMA or an HAR model. However, the number of trained weights does not differ excessively from the number of parameters in a smooth transition autoregressive model with multiple regimes. Furthermore, the network models do not lend themselves to the easy interpretation of explanatory variables due to the structure of the layers. On this purpose, the author acknowledges that this article was mainly focused on providing superior forecasting accuracy rather than interpreting causal relationships.

In future works, the performance of RNNs should be tested with different architectures, for example, by modifying the activation function or enlarging the number of hidden layers.

Moreover, in this article, I have exclusively focused on univariate time series while, in practice, multivariate forecasting problems require to forecast a set of possibly dependent time series. An important future direction is to extend the strategies developed in this article to the multivariate setting.

References

- Anders, U., and O. Korn. 1996. "Model Selection in Neural Networks." ZEW Discussion Paper No. 96–21.
- Andersen, T. G., T. Bollerslev, and F. X. Diebold. 2007. "Roughing It Up: Including Jump Components in the Measurement, Modeling and Forecasting of Return Volatility." CREATES Research Paper No. 2007–18.
- Andersen, T. G., T. Bollerslev, F. X. Diebold, and H. Ebens. 2001. The Distribution of Realized Stock Return Volatility. *Journal of Financial Economics* 61: 43–76.
- Arnerić, J., T. Poklepovic, and Z. Aljinović. 2014. GARCH Based Artificial Neural Networks in Forecasting Conditional Variance of Stock Returns. *Croatian Operational Research Review* 5: 329–343.
- Bai, J., and P. Perron. 2003. Computation and Analysis of Multiple Structural Change Models. *Journal of Applied Econometrics* 18: 1–22.
- Bao, W., H. Yue, and Y. Rao. 2017. A Deep Learning Framework for Financial Time Series Using Stacked Autoencoders and Long-Short Term Memory. *PLoS One* 12: e0180944.
- Barndorff-Nielsen, O. E., and N. Shephard. 2002. Econometric Analysis of Realised Volatility and Its Use in Estimating Stochastic Volatility Models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64: 253–280.

- Bianchi, F. M., M. C. Kampffmeyer, A. Rizzi, and R. Jenssen. 2017. "An Overview and Comparative Analysis of Recurrent Neural Networks for Short Term Load Forecasting." *CoRR*, abs/1705.04378.
- Black, F. 1976. Noise. *Journal of Finance* 41: 529–543.
- Bollerslev, T. 1986. Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics* 31: 307–327.
- Bucci, A., G. Palomba, and E. Rossi. 2019. "Does macroeconomics help in predicting stock markets volatility comovements? A nonlinear approach." Working Paper No. 440, Dipartimento di Scienze Economiche e Sociali, Università Politecnica delle Marche.
- Chong, Y. Y., and D. F. Hendry. 1986. Econometric Evaluation of Linear Macro-Economic Models. *The Review of Economic Studies* 53: 671–690.
- Choudhry, T., F. I. Papadimitriou, and S. Shabi. 2016. Stock Market Volatility and Business Cycle: Evidence from Linear and Nonlinear Causality Tests. *Journal of Banking & Finance* 66: 89–101.
- Christiansen, C., M. Schmeling, and A. Schrimpf. 2012. A Comprehensive Look at Financial Volatility Prediction by Economic Variables. *Journal of Applied Econometrics* 27: 956–977.
- Clements, M. P., and H.-M. Krolzig. 1998. A Comparison of the Forecast Performance of Markov-Switching and Threshold Autoregressive Models of US GNP. *The Econometrics Journal* 1: 47–75.
- Corsi, F. 2009. A Simple Approximate Long-Memory Model of Realized Volatility. *Journal of Financial Econometrics* 7: 174–196.
- De Pooter, M., M. Martens, and D. Van Dijk. 2008. Predicting the Daily Covariance Matrix for S&P 100 Stocks Using Intraday Data - But Which Frequency to Use?. *Econometric Reviews* 27: 199–229.
- Di Persio, L., and O. Honchar. 2017. Recurrent Neural Networks Approach to the Financial Forecast of Google Assets. *International Journal of Mathematics and Computers in Simulation* 11: 7–13.
- Diebold, F. X., and R. S. Mariano. 1995. Comparing Predictive Accuracy. *Journal of Business & Economic Statistics* 13: 253–263.
- Diebold, F. X., and K. Yilmaz. 2009. Measuring Financial Asset Return and Volatility Spillovers, with Application to Global Equity Markets. *The Economic Journal* 119: 158–171.
- Donaldson, G. R., and M. Kamstra. 1996a. A New Dividend Forecasting Procedure That Rejects Bubbles in Asset Prices. *Review of Financial Studies* 8: 333–383.
- Donaldson, G. R., and M. Kamstra. 1996b. Forecast Combining with Neural Networks. *Journal of Forecasting* 15: 49–61.
- Donaldson, G. R., and M. Kamstra. 1997. An Artificial Neural network-GARCH Model for International Stock Return Volatility. *Journal of Empirical Finance* 4: 17–46.
- Elman, J. L. 1990. Finding Structure in Time. *Cognitive Science* 14: 179–211.
- Engle, R. F. 1982. Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica* 50: 987–1007.
- Engle, R. F., E. Ghysels, and B. Sohn. 2009. "On the Economic Sources of Stock Market Volatility." NYU Working Paper No. FIN-08-043
- Fair, R., and R. J. Shiller. 1989. The Informational Context of Ex Ante Forecasts. *The Review of Economics and Statistics* 71: 325–331.
- Fama, E., and K. French. 1993. Common Risk Factors in the Returns on Stocks and Bonds. *Journal of Financial Economics* 33: 3–56.
- Fernandes, M., M. C. Medeiros, and M. Scharth. 2014. Modeling and Predicting the CBOE Market Volatility Index. *Journal of Banking & Finance* 40: 1–10.
- Gnana Sheela, K., and S. Deepa. 2013. Review on Methods to Fix Number of Hidden Neurons in Neural Networks. *Mathematical Problems in Engineering* 2013: 11.

- Guzman, S. M., J. O. Paz, and M. L. M. Tagert. 2017. The Use of NARX Neural Networks to Forecast Daily Groundwater Levels. *Water Resources Management* 31: 1591–1603.
- Hajizadeh, E., A. Seifi, M. Fazel Zarandi, and I. Turksen. 2012. A Hybrid Modeling Approach for Forecasting the Volatility of S&P 500 Index Return. *Expert Systems with Applications* 39: 431–436.
- Hamid, S. A., and Z. Iqbal. 2004. Using Neural Networks for Forecasting Volatility of S&P 500 Index Futures Prices. *Journal of Business Research* 57: 1116–1125.
- Hansen, P. R., A. Lunde, and J. M. Nason. 2011. The Model Confidence Set. *Econometrica* 79: 435–497.
- Heaton, J., N. Polson, and J. Witte. 2016. “Deep Learning in Finance.” CoRR, abs/1602.06561.
- Hillebrand, E., and M. C. Medeiros. 2010. The Benefits of Bagging for Forecast Models of Realized Volatility. *Econometric Reviews* 29: 571–593.
- Hochreiter, S., and J. Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9: 1735–1780.
- Hu, Y. M., and C. Tsoukalas. 1999. “Combining Conditional Volatility Forecasts Using Neural Networks: An Application to the EMS Exchange Rates. *Journal of International Financial Markets, Institutions & Money* 9: 407–422.
- Jordan, M. I. 1986. “Serial Order: A Parallel Distributed Processing Approach.” Discussion Paper, Institute for Cognitive Science Report 8604, University of California San Diego.
- Kamijo, K., and T. Tanigawa. 1990. “Stock Price Pattern Recognition - A Recurrent Neural Network Approach.” 1990 IJCNN International Joint Conference on Neural Networks. San Diego, CA, USA: IEEE, pp. 215–221.
- Keenan, D. M. 1985. A Tukey Nonadditivity-Type Test for Time Series Nonlinearity. *Biometrika* 72: 39–44.
- Khan, A. I. 2011. Financial Volatility Forecasting by Nonlinear Support Vector Machine Heterogeneous Autoregressive Model: Evidence from Nikkei 225 Stock Index. *International Journal of Economics and Finance* 3: 138–150.
- Kim, H. Y., and C. H. Won. 2018. Forecasting the Volatility of Stock Price Index: A Hybrid Model Integrating LSTM with Multiple GARCH-Type Models. *Expert Systems with Applications* 103: 25–37.
- Kristjanpoller, W., A. Fadic, and M. C. Minutolo. 2014. Volatility Forecast Using Hybrid Neural Network Models. *Expert Systems with Applications* 41: 2437–2442.
- Lin, T., B. G. Horne, P. Tino, and C. L. Giles. 1996. Learning Long-Term Dependencies in NARX Recurrent Neural Networks. *IEEE Transactions on Neural Networks* 7: 1329–1338.
- Maciel, L., F. Gomide, and R. Ballini. 2016. Evolving Fuzzy-GARCH Approach for Financial Volatility Modeling and Forecasting. *Computational Economics* 48: 379–398.
- Maheu, J. M., and T. H. McCurdy. 2002. Nonlinear Features of Realized Volatility. *Review of Economics and Statistics* 84: 668–681.
- McAleer, M., and M. Medeiros. 2008. A Multiple Regime Smooth Transition Heterogeneous Autoregressive Model for Long Memory and Asymmetries. *Journal of Econometrics* 147: 104–119.
- Mele, A. 2007. Asymmetric Stock Market Volatility and the Cyclical Behavior of Expected Returns. *Journal of Financial Economics* 86: 446–478.
- Mele, A. 2008. “Understanding Stock Market Volatility - A Business Cycle Perspective.” Working Paper.
- Menezes, J., and G. Barreto. 2006. “A New Look at Nonlinear Time Series Prediction with NARX Recurrent Neural Network.” 2006 Ninth Brazilian Symposium on Neural Networks (SBRN’06). Ribeiro Preto, Brazil: IEEE, pp. 160–165.

- Miura, R., L. Pichl, and T. Kaizoji. 2019. "Artificial Neural Networks for Realized Volatility Prediction in Cryptocurrency Time Series." In H. Lu, H. Tang, and Z. Wang (eds.), *Advances in Neural Networks – ISNN 2019*. Cham: Springer International Publishing, pp. 165–172.
- Nelson, D. B. 1990. Stationarity and Persistence in the GARCH (1,1) Model. *Econometric Theory* 6: 318–334.
- Panchal, G., A. Ganatra, Y. Kosta, and D. Panchal. 2010. Searching Most Efficient Neural Network Architecture Using Akaike's Information Criterion (AIC). *International Journal of Computer Applications* 5: 41–44.
- Patton, A. J. 2011. Volatility Forecast Comparison Using Imperfect Volatility Proxies. *Journal of Econometrics* 160: 246–256.
- Pavlidis, E. G., I. Paya, and D. A. Peel. 2012. Forecast Evaluation of Nonlinear Models: The Case of Long-Span Real Exchange Rates. *Journal of Forecasting* 31: 580–595.
- Paye, B. S. 2012. D  ja Vol: Predictive Regressions for Aggregate Stock Market Volatility Using Macroeconomic Variables. *Journal of Financial Economics* 106: 527–546.
- Pichl, L., and T. Kaizoji. 2017. Volatility Analysis of Bitcoin Price Time Series. *Quantitative Finance and Economics* 1: 474–485.
- Rosa R., L. Maciel, F. Gomide, and R. Ballini. 2014. "Evolving Hybrid Neural Fuzzy Network for Realized Volatility Forecasting with Jumps." *2014 IEEE Conference on Computational Intelligence for Financial Engineering Economics (CIFER)*. London: IEEE, pp. 481–488.
- Rossi, E., and P. Santucci de Magistris. 2014. Estimation of Long Memory in Integrated Variance. *Econometric Reviews* 33: 785–814.
- Schittenkopf, C., G. Dorffner, and E. J. Dockner. 2000. Forecasting Time-Dependent Conditional Densities: A Semi Non-parametric Neural Network Approach. *Journal of Forecasting* 19: 355–374.
- Schwert, W. G. 1989. Why Does Stock Market Volatility Change over Time. *The Journal of Finance* 44: 1115–1153.
- Stinchcombe, M., and H. White. 1992. "Using Feedforward Networks to Distinguish Multivariate Populations." *Proceedings of the International Joint Conference on Neural Networks*. Baltimore, MD, USA: IEEE, pp. 788–793.
- Tang, Z., and P. A. Fishwick. 1993. Feed-Forward Neural Nets as Models for Time Series Forecasting. *ORSA Journal on Computing* 5: 374–385.
- Terasvirta, T. 1994. Specification, Estimation, and Evaluation of Smooth Transition Autoregressive Models. *Journal of the American Statistical Association* 89: 208–218.
- Tibshirani, R. 1996. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58: 267–288.
- Tino, P., C. Schittenkopf, and G. Dorffner. 2001. Financial Volatility Trading Using Recurrent Neural Networks. *IEEE Transactions on Neural Networks* 12: 865–874.
- Vortelinos, D. I. 2017. Forecasting Realized Volatility: HAR against Principal Components Combining, Neural Networks and GARCH. *Research in International Business and Finance* 39: 824–839.
- Welch, I., and A. Goyal. 2008. A Comprehensive Look at the Empirical Performance of Equity Premium Prediction. *Review of Financial Studies* 21: 1455–1508.
- White, H. 1988. "Economic Prediction Using Neural Networks: The Case of IBM Daily Stock Returns." *IEEE 1988 International Conference on Neural Networks*. San Diego, CA, USA: IEEE, pp. 451–458.
- Xiong, R., E. P. Nichols, and Y. Shen. 2016. "Deep Learning Stock Volatility with Google Domestic Trends." *CoRR*, abs/1512.04916.
- Zou, H. 2006. The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association* 101: 1418–1429.

Realized Variance Modeling: Decoupling Forecasting from Estimation*

Fabrizio Cipollini¹, Giampiero M. Gallo ^{2,3}, and Alessandro Palandri^{1,4}

¹DiSIA “G.Parenti”, Università di Firenze, ²Italian Court of Audits, ³New York University in Florence and ⁴DCU Business School, Dublin City University

Address correspondence to Giampiero M. Gallo, NYU in Florence, Italy, or e-mail: giampiero.gallo@nyu.edu.

Received July 25, 2019; revised July 25, 2019; editorial decision April 16, 2020; accepted April 20, 2020

Abstract

This article evaluates the in-sample fit and out-of-sample forecasts of various combinations of realized variance models and functions delivering estimates (estimation criteria). Our empirical findings highlight that: independently of the econometrician’s forecasting loss (FL) function, certain estimation criteria perform significantly better than others; the simple ARMA modeling of the log realized variance generates superior forecasts than the Heterogeneous Autoregressive (HAR) family, for any of the FL functions considered; the (2, 1) parameterizations with negative lag-2 coefficient emerge as the benchmark specifications generating the best forecasts and approximating long-range dependence as does the HAR family.

Key words: forecast evaluation, Heterogeneous Autoregressive (HAR) model, realized variance, variance forecasting, variance modeling

JEL classification: C22, C52, C53, C58, G17

* The views expressed in the article are those of the authors and do not involve the responsibility of the Corte dei conti. We thank two anonymous referees for a very careful reading of our paper and for providing us with very useful comments. We are grateful to Frank Diebold for his encouragement and comments on the first draft. Thanks are due to Sébastien Laurent, Alessandra Luati, Monia Luparelli and Roberto Renó, to IAF 2018, QFFE 2019, and SIS 2019 conference participants and to seminar participants at the Department of Economics of the University of Verona. In our interactions with Peter Christoffersen since the early days at Penn, we always appreciated Peter’s deep knowledge, passion, and availability to share his work. The intellectual curiosity, passion for research, and a talent for mentorship that permeated his successful career have been an example and a point of reference. This is yet another occasion to renew the acknowledgment of his brilliant comments on our [Engle and Gallo \(2006\)](#) paper that helped us pass the final refereeing hurdle. In our last exchange dating back to May 2018, Peter had found the time to dig in his folders and retrieve replication files for his 2010 RFS paper with K. Mimouni. His untimely departure leaves us with even more questions to ask, but with his open and friendly smile to remember.

The technical capability of recording and storing intradaily data has given a tremendous boost to the literature both on measurement issues (various forms of aggregation of ultra-high frequency data into a daily value) and on modeling their dynamics (extending the autoregressive flavor behind the GARCH class of models). There exist recent refinements within a somewhat consolidated menu of realized variance choices [plain vanilla, robustness to market microstructure noise, jumps, etc.; cf. [Park and Linton \(2012\)](#), for a survey]; in this work, we focus on the realized kernel variance ([Barndorff-Nielsen et al., 2008](#)) to investigate a series of aspects related to modeling and forecasting. We will take it either as such, or as its square root, or yet as its logarithm, but for the sake of simplicity we generically talk here about realized variance modeling.

In this article, we consider that evaluating forecasts out-of-sample (OOS) is a matter of subjective taste about how to judge the distance between a predicted outcome and the actual value. When the latter is observable (say realized variance), the choice of a *forecasting loss* (FL) is a consequence of individual preferences, and hence it is not subject to qualitative assessment. However, when the actual value entering the loss is not observable and forecasts are evaluated using a proxy (say realized variance proxying for the underlying conditional variance), not all loss functionals are robust,¹ as shown by [Patton \(2011\)](#). In this respect, we consider several examples of FL functions and discuss their robustness when the actual value is not observable. In order to produce forecasts, two other elements are important: the *model specification* (the equation reproducing conditional variance dynamics), and the in-sample (IS) *estimation criterion* (EC), that is, the distance between observed and fitted values delivering parameter estimates. As for the adoption of such elements, we take the view that there are no natural a priori choices, but they must be geared toward obtaining the best results in terms of the given FL. In this respect, we argue that the choice of the latter should not force the same function to be repeated as the EC.² When the estimated model and the data generating process coincide, [Hansen and Dumitrescu \(2018\)](#) show that the asymptotically preferred estimation criteria are those that deliver consistent (paramount) and relatively more efficient parameter estimates, regardless of the FL; hence, maximum likelihood is the best criterion. The critical requirement of consistency implies that in the presence of *influential observations* (e.g., occasional spikes in volatility), the preferred EC is the one that is more robust, first, and efficient, second. Furthermore, in the common forecasting setting where model parameters are estimated over rolling windows of finite samples, convergence of the parameter estimates to the corresponding true values is not attained and an EC that replicates the FL may generate greater losses than an alternative criterion if the latter delivers more efficient estimates.

A first contribution of the article is to explore whether, for a given FL, the IS EC with the same functional form produces the best OOS results by specification or, rather, other estimation criteria are to be preferred. A second contribution is to investigate the capability of a popular model selection tool such as the Bayes (Schwarz) Information Criterion (BIC) to identify the specification which performs the best OOS. Third, we compare the

- 1 [Patton \(2011\)](#) defines a loss function as 'robust' if its ranking of any two forecasts is the same whether the ranking is done using the actual value or some conditionally unbiased proxy.
- 2 [Christoffersen and Jacobs \(2004\)](#) address the issue of the loss functions used in parameter estimation and in model evaluation in reference to option valuation, advocating an alignment in what we call estimation criterion and forecasting loss.

benchmark Heterogeneous Autoregressive (HAR) specifications of Corsi (2009) and Andersen, Bollerslev, and Diebold (2007) to several realized variance models reminiscent of well-known GARCH parameterizations which, for the most part, correspond to the simple ARMA modeling of either realized variance, its square root, or its logarithm. Across model classes, here we focus on the core specifications that could be extended to accommodate several refinements (asymmetry, see Engle and Gallo (2006); jumps, see Andersen, Bollerslev, and Diebold (2007); measurement errors, see Bollerslev, Patton, and Quaevdliog (2016), etc.).

We explore these issues on a panel of twenty-eight constituents of the Dow Jones 30 Index using daily realized kernel variance observations from January 2005 to December 2015. The sample is split into six five-year IS periods; we estimate all combinations of specifications and estimation criteria on the first IS period, generating OOS static forecasts for the ensuing one-year period, then we move the IS window by one year and repeat the procedure. Our results may be summarized as follows: there are FL functions not particularly apt to be repeated as estimation criteria; we cast some doubts about the BIC ability to identify *ex ante* the best OOS specification; the ARMA modeling of the log realized variance provides the best IS and OOS results. In general, we identify the (2,1) structure with a negative lag-2 coefficient to be a good parameterization. As a reading key to these results, we find it informative to relate the goodness of the OOS forecasts to the structure's capability to mimic long memory features: as a matter of fact, these specifications deliver a long memory approximation which is equivalent to that of the HAR family, but overall superior OOS forecasts across assets.

The article is structured as follows. Section 1 discusses the FLs employed to compare model forecasts and the estimation criteria used to estimate parameters. Section 2 presents the specifications for the variance dynamics and Section 3 the information criterion used for model selection. The empirical results are presented in Section 4. In Section 5 we provide a general discussion on the estimation criteria and the (2,1) parameterizations. Section 6 concludes.

1. FL and EC

A loss function maps events onto real numbers according to the preference orderings of an individual. When estimating, values of the unknown coefficients of a parametric model are obtained from the minimization of a loss function IS. When forecasting, the quality of a specification is assessed through the calculation of a loss function based on the distance between the predicted outcomes and the actual values, OOS. In what follows, we argue that the priority given to forecasting requires to define separately the *FL* and the *EC*.

With the expression *FL* we refer to the loss function used to evaluate the model performance OOS. With the term *EC* we refer to the objective function used to obtain the parameter estimates. In general, the choice of this function responds to some features: theoretical properties of the resulting estimator, tractability, etc. Distinguishing between forecast evaluation and model estimation allows us to investigate, empirically and in the context of variance modeling, whether the EC that coincides with the FL produces the best OOS results. For FL and EC we consider some of the most common functionals adopted in the literature to measure the distance between observed and predicted values. Prevailing measures are quadratic, which corresponds to least squares estimation, and Kullback–Leibler,

which corresponds to quasi-maximum-likelihood estimation when the chosen density function is linear–exponential.

1.1 Quadratic Distance

We characterize the general quadratic FL by:

$$\sum_{t=T_1+1}^{T_2} [f(RV_t) - f(\sigma_t^2)]^2,$$

where $t = T_1 + 1, \dots, T_2$ is the OOS period and $f(\sigma_t^2) \equiv \mathbb{E}_{t-1}[f(RV_t)]$. For the monotonic function $f(\cdot)$ we choose the identity, the square root (prefix SD) and the logarithm (prefix LN) which translate, respectively, into the specific quadratic LS, SDLS³ and LNLS FLs. As an example, for the square root function we have that the FL synthesizes the distance between $RV_t^{1/2}$ and $\sigma_t \equiv E_{t-1}[RV_t^{1/2}]$, $t = T_1 + 1, \dots, T_2$. When RV_t enters the loss as a proxy of the true conditional variance, SDLS and LNLS are not robust FLs. However, in the case of RV_t computed from five-minute returns, the optimal forecasts of SDLS and LNLS are only 1% and 2% smaller than the true conditional variance, respectively (cf. Table 2 in Patton (2011)).

By the same token, we define the quadratic EC:

$$\sum_{t=1}^{T_1} [g(RV_t) - g(\sigma_t^2)]^2,$$

where $t = 1, \dots, T_1$ is the IS period and $g(\cdot)$ is a monotonic function which may not coincide with the choice of $f(\cdot)$ in the FL. As before, choosing $g(\cdot)$ between the identity, square root and logarithmic functions, leads to LS, SDLS and LNLS estimation criteria, respectively.

To be clear, hence, we can have a SDLS FL (a matter of individual preference), but an LS EC (a matter of estimation convenience). For example, since they allow for the simple OLS estimation of the model parameters, LS is the EC of choice for the HAR of Corsi (2009) and LNLS for the LOG – HAR of Andersen, Bollerslev and Diebold, (2007), but we may want to measure the FL in terms of the realized volatility. We point out that σ_t^2 in the EC is that defined by the FL: $\sigma_t^2 = f^{-1}(\mathbb{E}_{t-1}[f(RV_t)])$. Therefore, unless $g=f$, what the model parameters are trained to predict, $\mathbb{E}_{t-1}[g(RV_t)]$, may differ from the object of interest in forecasting, $\mathbb{E}_{t-1}[f(RV_t)]$. And furthermore, notice how despite the symmetry of the quadratic FLs, when EC differs from FL, the generalized errors defined by the former do not enter symmetrically in the latter.³ For insights on optimal predictions under asymmetric losses, see Christoffersen and Diebold (1997), among others.

1.2 Kullback–Leibler Distance

Of the many density functions that may characterize a Kullback–Leibler distance, we focus on the Gaussian density to generate the QML FL:

3 Consider the case of SDLS EC and LS FL: substituting the generalized residual $\hat{\varepsilon}_t \equiv RV_t^{1/2} - \sigma_t$ in the FL yields $\sum_t \hat{\varepsilon}_t^4 + 4\sigma_t \hat{\varepsilon}_t^3 + 4\sigma_t^2 \hat{\varepsilon}_t^2$ which, because of the term $\hat{\varepsilon}_t^3$, is not symmetric.

$$\sum_{t=T_1+1}^{T_2} \left[\ln \sigma_t^2 + \frac{RV_t}{\sigma_t^2} \right],$$

where $t = T_1 + 1, \dots, T_2$ is the OOS period and σ_t^2 is the conditional expectation $\mathbb{E}_{t-1}[RV_t]$ of realized variance. Similarly, the QML EC is given by:

$$\sum_{t=1}^{T_1} \left[\ln \sigma_t^2 + \frac{RV_t}{\sigma_t^2} \right] \quad (1)$$

where $t = 1, \dots, T_1$ is the IS period and $\sigma_t^2 = f^{-1}(\mathbb{E}_{t-1}[f(RV_t)])$, as defined by the FL. Since the Gaussian belongs to the family of *linear-exponential* distributions, minimization of the QML EC is in fact a quasi-maximum-likelihood estimation with associated properties (Gouriéroux, Monfort and Trognon, 1984). If intradaily returns are normally distributed, Gaussian QML is maximum likelihood. Although differently motivated, the EC in Equation (1) is the estimator of choice for the univariate MEM of Brownlees, Cipollini and Gallo (2012).

1.3 Mismatch Bias

In general, an EC that differs from the FL induces a bias in the forecasts. While the magnitude of the bias depends on the (unknown) data generating process, for the cases considered, its sign may be derived from Jensen's inequality (Table 1). Since for LS, SDLS, and QML FLs, the quantity to forecast is positive, an EC that induces a negative bias and has a relatively smaller variance than an unbiased criterion may be interpreted as performing some degree of shrinkage toward zero (providing a well-known tradeoff between bias and variance). For the LNLS FL, the same reasoning applies to every $RV_t > 1$. Therefore, should the nonpositive bias of the LNLS EC, for every FL, be accompanied by a small variance, it could result in a well-performing EC (the opposite would be true for a nonnegative bias of the LS and a large variance).

2 Variance Modeling

In this section, we present some specifications present in the literature on realized variance modeling: two HAR specifications and several other models reminiscent of popular GARCH parameterizations. In line with established results in this field, we focus on the (1, 1) and (2, 1) parameterizations, where the former is found to be well suited to generate good forecasts, as highlighted by Hansen and Lunde (2005), while the latter occasionally provides better fit and forecasts. We present all models as parameterizing σ_t^2 . Although nonstandard, this is consistent with our setup in which the model parameters are estimated IS to provide the best predictions $\mathbb{E}_{t-1}[g(\sigma_t^2)]$, but are ultimately used to produce OOS forecasts $\mathbb{E}_{t-1}[f(\sigma_t^2)]$. While our empirical analysis focuses on one-step ahead forecasts, k -step-ahead forecasts $\mathbb{E}_{t-1}[f(\sigma_{t+k-1}^2)]$ may be generated for any of the following parameterizations, granting that in general their calculation requires numerical integration for $k > 1$.

2.1 HAR Specifications

The HAR, introduced by Corsi (2009), has rapidly achieved the benchmark status for models of realized variances. Features contributing to this role are the simplicity with which its

Table 1. Sign of the bias for the transformation of the quantity minimizing the EC (to the left of the minus sign) and the quantity minimizing the FL (to the right of the minus sign)

	FL : LS/QML	FL : SDLS	FL : LNLS
EC : LS/QML	$\mathbb{E}_{t-1}(RV_t) - \mathbb{E}_{t-1}(RV_t) = 0$	$\mathbb{E}_{t-1}(RV_t)^{1/2} - \mathbb{E}_{t-1}(RV_t^{1/2}) > 0$	$\ln \mathbb{E}_{t-1}(RV_t) - \mathbb{E}_{t-1}(\ln RV_t) > 0$
EC : SDLS	$\mathbb{E}_{t-1}(RV_t^{1/2})^2 - \mathbb{E}_{t-1}(RV_t) < 0$	$\mathbb{E}_{t-1}(RV_t^{1/2}) - \mathbb{E}_{t-1}(RV_t^{1/2}) = 0$	$2 \ln \mathbb{E}_{t-1}(RV_t^{1/2}) - \mathbb{E}_{t-1}(\ln RV_t) > 0$
EC : LNLS	$\exp \{ \mathbb{E}_{t-1}(\ln RV_t) \} - \mathbb{E}_{t-1}(RV_t) < 0$	$\exp \{ \frac{1}{2} \mathbb{E}_{t-1} \{ \ln RV_t \} \} - \mathbb{E}_{t-1}(RV_t^{1/2}) < 0$	$\mathbb{E}_{t-1}(\ln RV_t) - \mathbb{E}_{t-1}(\ln RV_t) = 0$

parameters may be estimated and its ability to reproduce long memory features: "... the mixing of relatively few volatility components is capable of reproducing a remarkably slow volatility autocorrelation decay that is almost indistinguishable from that of a hyperbolic pattern over most empirically relevant forecast horizons" (Andersen, Bollerslev and Diebold, 2007). The HAR models σ_t^2 as a function of past realizations over daily, weekly, and monthly time intervals:

$$\sigma_t^2 = \omega + \alpha_1 \cdot RV_{t-1} + \alpha_2 \cdot \frac{1}{5} \sum_{i=1}^5 RV_{t-i} + \alpha_3 \cdot \frac{1}{22} \sum_{i=1}^{22} RV_{t-i} \quad (2)$$

corresponding to an AR(22) process for RV_t with parameter constraints. Its parameters may be estimated by ordinary least squares when the EC is LS, whereas for different choices of the EC the estimates are not available in closed form. Necessary and sufficient conditions for the positivity of σ_t^2 are $\omega > 0$, $\alpha_3 \geq 0$, $\alpha_2/5 + \alpha_3/22 \geq 0$ and $\alpha_1 + \alpha_2/5 + \alpha_3/22 \geq 0$.

The LOG – HAR, introduced by Andersen, Bollerslev and Diebold, (2007), is an alternative specification linear in the logarithms:

$$\sigma_t^2 = \exp \left\{ \omega + \alpha_1 \cdot \ln RV_{t-1} + \alpha_2 \cdot \ln \left(\frac{1}{5} \sum_{i=1}^5 RV_{t-i} \right) + \alpha_3 \cdot \ln \left(\frac{1}{22} \sum_{i=1}^{22} RV_{t-i} \right) \right\} \quad (3)$$

Notice that the presence of the logarithms of averages places the LOG – HAR outside the class of AR processes. Its parameters may be estimated by ordinary least squares for LNLS EC, while for different choices of the EC the estimates are not available in closed form.

2.2 MVAR Specification

With MVAR we indicate the parameterization of σ_t^2 in terms of its lags and lags of RV_t or, equivalently, the ARMA modeling of realized variance RV_t . Without stretching this and subsequent parallels, an MVAR could be seen as an open-to-close GARCH (Bollerslev, 1986), in the limiting case of one intradaily observation (i.e., the realized variance collapses to the squared open-to-close return). The MVAR(2, 1) specification is given by:

$$\sigma_t^2 = \omega + \alpha_1 \cdot RV_{t-1} + \alpha_2 \cdot RV_{t-2} + \beta_1 \cdot \sigma_{t-1}^2. \quad (4)$$

Necessary and sufficient conditions for the positivity of σ_t^2 are $\omega > 0$, $\alpha_1, \beta_1 \geq 0$ and $\alpha_2 \geq -\alpha_1 \beta_1$. MVAR estimated by QML coincides with the MEM of Cipollini, Engle and Gallo (2013) while associated to the LS criterion it reduces to standard ARMA modeling and estimation.

2.3 MVOL Specification

With MVOL we denote the parameterization of σ_t in terms of its lags and lags of $RV_t^{1/2}$ or, equivalently, the ARMA modeling of realized volatility $RV_t^{1/2}$. It is reminiscent of a TGARCH(p, q) of Zakoian (1994) without the asymmetric term, which would result when realized volatilities are replaced by the absolute value of the open-to-close returns. The symmetric MVOL(2, 1) specification is given by:

$$\sigma_t^2 = \{\omega + \alpha_1 \cdot \text{RV}_{t-1}^{1/2} + \alpha_2 \cdot \text{RV}_{t-2}^{1/2} + \beta_1 \cdot \sigma_{t-1}\}^2 \quad (5)$$

Although σ_t^2 is positive by construction, the marginal effects of its determinants do not exhibit abrupt sign changes if and only if σ_t is also positive. Necessary and sufficient conditions for the positivity of σ_t are $\omega > 0$, $\alpha_1, \beta_1 \geq 0$ and $\alpha_2 \geq -\alpha_1\beta_1$. QML estimation of MVOL coincides with the MEM in [Brownlees, Cipollini and Gallo \(2012\)](#) while adopting the SDLS criterion reduces to standard ARMA modeling and estimation.

2.4 MLOG Specification

MLOG is the ARMA modeling of the log realized variances $\ln \text{RV}_t$ or, equivalently, the parameterization of $\ln \sigma_t^2$ in terms of its lags and lags of $\ln \text{RV}_t$. In the way of analogues, it is related to the log-GARCH(p, q) of [Geweke \(1986\)](#) as the limiting case of MLOG when the log realized variance reduces to log-squared residual for a single intraday observation. The MLOG(2, 1) specification is given by:

$$\sigma_t^2 = \exp \{ \omega + \alpha_1 \cdot \ln \text{RV}_{t-1} + \alpha_2 \cdot \ln \text{RV}_{t-2} + \beta_1 \cdot \ln \sigma_{t-1}^2 \} \quad (6)$$

When associated with the LNLS criterion, it reduces to standard ARMA modeling and estimation. QML, on the other hand, would be the natural estimator within the MEM framework. Dynamic specifications analogous to the MLOG are not uncommon in the context of Autoregressive Conditional Durations, among which [Bauwens, Galli and Giot \(2008\)](#) and [Taylor and Xu \(2017\)](#) are examples of QML⁴ and LNLS estimates, respectively.

2.5 MEXP Specification

The MEXP(2, 1) specification is obtained by substituting realized volatilities for the absolute value of the returns, a specification which evokes the EGARCH(p, q) parameterization ([Nelson, 1991](#)), but without the asymmetric term:

$$\sigma_t^2 = \exp \left\{ \omega + \alpha_1 \cdot \frac{\text{RV}_{t-1}^{1/2}}{\sigma_{t-1}} + \alpha_2 \cdot \frac{\text{RV}_{t-2}^{1/2}}{\sigma_{t-2}} + \beta_1 \cdot \ln \sigma_{t-1}^2 \right\}; \quad (7)$$

the MEXP reproduces the symmetric EGARCH in the limiting case of a single intraday observation. Due to poor performance of both IS (never providing the best description of the data) and OOS (always generating the largest losses), we omit presenting and discussing results pertaining to the MEXP(2, 1) specification.⁵

4 To be precise: [Bauwens et al. \(2008\)](#) estimate an ACD analogous to MLOG by minimizing the Kullback–Leibler distance based on the exponential distribution. Since the exponential belongs to the family of *linear–exponential* distributions, the resulting estimator is also quasi-maximum-likelihood.

5 Table entries for the MEXP(2, 1) are available upon request.

3 Model Evaluation

We perform IS model evaluation and selection by means of the BIC. For the quadratic estimation criteria of Section 1.1, we construct the BIC by treating the generalized residuals $g(RV_t) - g(\hat{\sigma}_t^2)$ as Gaussian:

$$\text{BIC} = T_1 \ln \left(\frac{1}{T_1} \sum_{t=1}^{T_1} [g(RV_t) - g(\hat{\sigma}_t^2)]^2 \right) + k \ln T_1$$

where k is the number of parameters, T_1 the sample size, and $\hat{\sigma}_t^2$ the model's prediction. For the Kullback–Leibler EC of Section 1.2, the BIC is immediately obtained from the log-likelihood function:

$$\text{BIC} = \sum_{t=1}^{T_1} \ln \hat{\sigma}_t^2 + k \ln T_1,$$

where $\hat{\sigma}_t^2$ is the prediction from the QML-estimated model. Notice how the BIC is calculated on the condensed EC in Equation (1) from which the average of $RV_t/\hat{\sigma}_t^2$ is dropped on the ground that deviations from its limiting value of one are neither data-driven nor model-driven but only reflect initial value choices.⁶ Since, for equally parameterized specifications, every information criterion produces identical model rankings, Akaike's and Hannan–Quinn's may produce results that differ from those we present only in the comparisons of differently parameterized specifications. Furthermore, given that BIC is the most conservative of the three when it selects a richer parameterization, so do Akaike's and Hannan–Quinn's. Since OOS measures of fit do not depend explicitly on the number of parameters k , we evaluate OOS forecasts directly from the FL functions of Sections 1.1 and 1.2.

4 Empirical Results

The data used in this study pertains to twenty-eight of the thirty constituents of the Dow Jones 30 Index. The sample has 11 years of high-frequency daily observations from March 1, 2005 to December 31, 2015 for a total of 2768 days. Two series (TRV and V) are not included in the study because they are not available for the full sample period.⁷ Tickers of the twenty-eight included stocks are: AAPL, AXP, BA, CAT, CSCO, CVX, DD, DIS, GE, GS, HD, IBM, INTC, JNJ, JPM, KO, MCD, MMM, MRK, MSFT, NKE, PFE, PG, UNH, UTX, VZ, WMT, XOM. The raw tick-by-tick TAQ data is cleaned using the procedure of [Brownlees and Gallo \(2006\)](#) and the series of realized variances calculated following [Barndorff-Nielsen et al. \(2011\)](#) with Parzen kernel. The sample is split into six five-year IS periods: 2005–2009 (1259 obs.), 2006–2010 (1259 obs.), 2007–2011 (1260 obs.), 2008–2012 (1259 obs.), 2009–2013 (1258 obs.), and 2010–2014 (1258 obs.). All model combinations are estimated on each of the six IS periods, and for each of them, OOS forecasts are generated for the following one-year period: 2010 (252 obs.), 2011 (252 obs.), 2012 (250 obs.), 2013 (252 obs.), 2014 (252 obs.), and 2015 (251 obs.).

6 In fact, for all the specifications considered, when the initial value σ_0^2 is treated as an unknown parameter and estimated, the average ratio RV_t/σ_t^2 is equal to 1.

7 TRV data are available only from February 26, 2007 while V data are missing from April 8, 2006 to February 26, 2007.

In what follows, we summarize the message behind the application of our strategy, given the variety of elements to be considered: we have eight different specifications (MVAR, MVOL, MLOG—each with (1, 1) and (2, 1) variants, HAR and LOG – HAR), four FLs (LS, SDLS, LNLS, and QML), four estimation criteria (same list with different meaning) and twenty-eight tickers across six partitions into IS/OOS periods (a total of 168 instances). The results are grouped in three sets of tables: in the first set (2–5), we consider in turn each of the four FLs, and, by specification, we report the percentage of instances in which each EC provides forecasts in the 75% Model Confidence Set (MCS) of Hansen, Lunde and Nason (2011) and the percentage it delivers the best OOS performance (total by row is 168). In the second group of Tables 6–9, for the same FLs, we report the percentage of instances each specification lies in the 75% MCS for a given EC together with the average loss (across 168 instances), marking the lowest value of the FL by specification. In the third group of Tables 10–13, we compare specifications by fixing the FL and the EC to be the same, and reporting the frequency by which each specification is best IS (based on BIC, total across rows is 100), best OOS (based on FL, total across rows is 100), best OOS among those instances where that specification had the best IS, the average and the median FL by specification across all 168 instances. Four fundamental questions can be addressed on the basis of this evidence:

1. **Is the FL the best EC?** The short answer is: not always. For a given FL, in Tables 2–9 we evaluate the conditional variance specifications (rows) when estimated by LS, SDLS, LNLS and QML estimation criteria (columns). From Tables 2–4 it emerges that when the FL is quadratic the EC that produces the best OOS results in most instances is LNLS, with SDLS a second. On the other hand, when the FL is QML, Table 5 shows that overall QML is the preferred EC. Similarly, the EC that is most present in the 75% MCS for every FL is LNLS followed by SDLS.

These results are confirmed in Tables 6–9 which, for every conditional variance specification (rows) and EC (columns), report the average value of the given FL over the 168 instances. The lowest average OOS FL measured by LS, SDLS, and LNLS is obtained, for every variance specification, when estimated using the LNLS EC. Similarly, the

Table 2. LS FL for OOS evaluations

Model		LS		SDLS		LNLS		QML	
MVAR (1,1)		43.5%	[16.1%]	79.2%	[19.6%]	93.5%	[56.6%]	45.8%	[7.7%]
	(2,1)	45.8%	[13.7%]	87.5%	[25.6%]	89.9%	[47.0%]	58.9%	[13.7%]
MVOL (1,1)		29.8%	[8.3%]	75.0%	[17.3%]	98.2%	[61.9%]	50.0%	[12.5%]
	(2,1)	37.5%	[18.3%]	79.2%	[24.4%]	97.0%	[49.4%]	60.7%	17.9%
MLOG (1,1)		30.4%	[7.1%]	69.0%	[20.8%]	98.8%	[56.6%]	55.4%	[15.5%]
	(2,1)	37.5%	[8.3%]	77.4%	[19.6%]	97.6%	[47.0%]	63.1%	[25.0%]
HAR		42.9%	[4.9%]	88.1%	[23.2%]	89.9%	[45.8%]	56.0%	[16.1%]
LNHAR		34.5%	[7.1%]	76.2%	[19.1%]	97.6%	[54.2%]	63.1%	[19.6%]

Notes: For every model specification, we report the percentages (twenty-eight tickers times six periods = 168 possible instances) with which the estimation criteria (LS, SDLS, LNLS, and QML) are in the 75% Model Confidence Set and, in square brackets, the percentages when those criteria provide the best performance (these add to 100 by row).

Table 3. SDLS FL for OOS evaluations

Model	LS		SDLS		LNLS		QML	
MVAR (1,1)	43.5%	[8.9%]	78.0%	[12.5%]	92.9%	[78.0%]	46.4%	[0.6%]
(2,1)	45.8%	[3.6%]	86.3%	[22.0%]	89.9%	[74.4%]	56.5%	[0.0%]
MVOL (1,1)	29.2%	[3.6%]	73.2%	[8.9%]	98.2%	[85.1%]	49.4%	[2.4%]
(2,1)	39.3%	[1.8%]	79.8%	[14.9%]	97.6%	[81.6%]	61.3%	[1.8%]
MLOG (1,1)	30.4%	[1.2%]	70.8%	[16.7%]	98.8%	[79.2%]	54.2%	[3.0%]
(2,1)	40.5%	[2.4%]	77.4%	[25.0%]	97.0%	[67.3%]	62.5%	[5.4%]
HAR	44.0%	[3.6%]	86.9%	[20.8%]	90.5%	[75.0%]	56.5%	[0.6%]
LNHAR	32.7%	[1.8%]	76.2%	[13.7%]	97.6%	[79.2%]	61.9%	[5.4%]

Notes: For every model specification, we report the percentages (twenty-eight tickers times six periods = 168 possible instances) with which the estimation criteria (LS, SDLS, LNLS, and QML) are in the 75% Model Confidence Set and, in square brackets, the percentages when those criteria provide the best performance (these add to 100 by row).

Table 4. LNLS FL for OOS evaluations

Model	LS		SDLS		LNLS		QML	
MVAR (1,1)	42.3%	[3.0%]	79.2%	[10.1%]	93.5%	[86.9%]	47.6%	[0.0%]
(2,1)	43.5%	[0.6%]	86.9%	[7.1%]	89.3%	[91.7%]	56.5%	[0.6%]
MVOL (1,1)	32.1%	[0.6%]	74.4%	[9.5%]	98.2%	[89.9%]	48.2%	[0.0%]
(2,1)	39.3%	[0.0%]	79.2%	[4.8%]	97.6%	[95.2%]	58.9%	[0.0%]
MLOG (1,1)	31.5%	[0.0%]	71.4%	[8.3%]	99.4%	[91.7%]	56.5%	[0.0%]
(2,1)	39.3%	[0.0%]	79.2%	[8.9%]	97.6%	[91.1%]	64.9%	[0.0%]
HAR	44.0%	[0.0%]	88.1%	[8.9%]	90.5%	[91.1%]	58.3%	[0.0%]
LNHAR	33.9%	[0.0%]	76.2%	[7.1%]	97.6%	[92.9%]	61.9%	[0.0%]

Notes: For every model specification, we report the percentages (twenty-eight tickers times six periods = 168 possible instances) with which the estimation criteria (LS, SDLS, LNLS, and QML) are in the 75% Model Confidence Set and, in square brackets, the percentages when those criteria provide the best performance (these add to 100 by row).

QML EC produces the lowest OOS QML FLs. Hence, the answer is affirmative for LNLS and QML, but not for LS and SDLS.

2. Does the best IS BIC deliver the best OOS specification? The short answer is: no. The details for this issue can be retrieved from [Tables 10–13](#): out of the 168 instances, when a specification is selected as the best IS BIC, it maintains the role of best OOS specification only 25%, 22%, 41%, and 4% of the times for LS, SDLS, LNLS, and QML, respectively. The average OOS FL calculated across all specifications selected by the best IS BIC is never the smallest, when compared to the values associated with either specification, no matter what the FL is. Nevertheless, the average for the quadratic FL functions resulting from a best IS BIC selection is preceded only by both MLOG's.

The results pertaining to LNLS, the best quadratic EC, are robust to the choice of the information criterion, with Akaike's and Hannan–Quinn's at most redistributing

Table 5. QML FL for OOS evaluations

Model	LS		SDLS		LNLS		QML	
MVAR (1,1)	43.5%	[7.7%]	78.6%	[19.6%]	93.5%	[29.2%]	48.8%	[43.5%]
(2,1)	44.0%	[10.1%]	86.9%	[16.7%]	90.5%	[23.2%]	57.1%	[50.0%]
MVOL (1,1)	29.2%	[13.1%]	73.8%	[24.4%]	98.2%	[23.2%]	50.0%	[39.3%]
(2,1)	38.1%	[15.5%]	78.6%	[19.6%]	97.6%	[17.9%]	59.5%	[47.0%]
MLOG (1,1)	31.5%	[16.1%]	71.4%	[31.0%]	99.4%	[14.3%]	56.5%	[38.7%]
(2,1)	38.1%	[18.5%]	78.6%	[26.2%]	97.6%	[13.1%]	64.9%	[42.3%]
HAR	43.5%	[7.7%]	87.5%	[16.7%]	91.1%	[23.2%]	58.9%	[52.4%]
LNHAR	32.7%	[16.7%]	75.6%	[29.8%]	97.6%	[10.7%]	61.3%	[42.9%]

Notes: For every model specification, we report the percentages (twenty-eight tickers times six periods = 168 possible instances) with which the estimation criteria (LS, SDLS, LNLS, and QML) are in the 75% Model Confidence Set and, in square brackets, the percentages when those criteria provide the best performance (these add to 100 by row).

Table 6. LS FL for OOS evaluations

Model	LS		SDLS		LNLS		QML	
MVAR (1,1)	48.8%	13.293	56.5%	13.514	64.3%	13.897	51.2%	15.160
(2,1)	62.5%	13.225	73.2%	13.533	78.0%	13.899	57.1%	15.169
MVOL (1,1)	53.6%	12.444	62.5%	12.189	70.8%	12.161	55.4%	12.669
(2,1)	63.1%	12.405	82.1%	12.237	87.5%	12.226	69.0%	12.762
MLOG (1,1)	70.8%	12.020	69.6%	11.742	76.8%	11.720	62.5%	11.868
(2,1)	69.0%	12.086	91.7%	11.731	95.8%	11.706	94.6%	11.799
HAR	63.7%	13.315	72.0%	13.612	78.6%	13.952	60.7%	15.296
LNHAR	61.3%	12.727	76.8%	12.212	84.5%	12.118	69.6%	12.892

Notes: For each of the four estimation criteria (LS, SDLS, LNLS, and QML), we report the percentages (twenty-eight tickers times six periods = 168 possible instances) with which the various specifications are in the 75% Model Confidence Set and the average loss. For every specification, the lowest loss (across estimation criteria) is reported in bold.

the 1.79% of MLOG(1,1) across the (2, 1) and HAR specifications. In contrast, model selection is particularly sensitive to the choice of information criterion in the QML case: using Akaike's, the least conservative of the three, the IS column of Table 13 from top to bottom would read: 0.00%, 7.14%, 0.00%, 19.05%, 0.00%, 22.02%, 13.69%, 38.10%. Nevertheless, the instances in which Akaike's delivers the best OOS specification is still a mere 11.31%. As such, the ability of information criteria to identify what will be the best OOS specification is very disappointing in the case of the QML EC, all the more so given that for the latter the information criteria are a natural consequence of the form of the QML EC itself.

To put it differently, our results show that to go through an intermediate step of model selection based on the best IS BIC hardly ensures to be coupled with the best OOS performance. For example, at least for our big caps tickers, a better strategy in terms of all OOS FL values would be to adopt the MLOG specification for every asset.

Table 7. SDLS FL for OOS evaluations

Model	LS		SDLS		LNLS		QML	
MVAR (1,1)	50.0%	0.170	55.4%	0.143	63.1%	0.141	50.6%	0.151
(2,1)	61.9%	0.161	72.6%	0.138	78.0%	0.136	58.9%	0.147
MVOL (1,1)	55.4%	0.158	63.1%	0.128	69.6%	0.126	54.2%	0.135
(2,1)	63.1%	0.150	82.7%	0.126	86.9%	0.124	69.0%	0.132
MLOG (1,1)	69.6%	0.146	69.6%	0.121	76.2%	0.120	61.3%	0.126
(2,1)	69.6%	0.147	91.7%	0.120	95.2%	0.118	94.6%	0.123
HAR	62.5%	0.162	72.0%	0.139	78.0%	0.137	61.9%	0.147
LNHAR	59.5%	0.164	76.8%	0.126	83.9%	0.124	70.2%	0.134

Notes: For each of the four estimation criteria (LS, SDLS, LNLS, and QML), we report the percentages (twenty-eight tickers times six periods = 168 possible instances) with which the various specifications are in the 75% Model Confidence Set and the average loss. For every specification, the lowest loss (across estimation criteria) is reported in bold.

Table 8. LNLS FL for OOS evaluations

Model	LS		SDLS		LNLS		QML	
MVAR (1,1)	48.8%	0.317	55.4%	0.209	64.3%	0.193	51.8%	0.216
(2,1)	62.5%	0.296	72.6%	0.201	77.4%	0.188	58.3%	0.211
MVOL (1,1)	54.2%	0.288	62.5%	0.194	71.4%	0.186	55.4%	0.207
(2,1)	63.1%	0.270	82.7%	0.188	88.1%	0.181	69.0%	0.202
MLOG (1,1)	69.0%	0.258	69.0%	0.187	77.4%	0.182	62.5%	0.201
(2,1)	71.4%	0.258	91.7%	0.184	95.8%	0.178	94.6%	0.197
HAR	63.1%	0.299	70.8%	0.203	78.6%	0.189	61.3%	0.211
LNHAR	58.9%	0.293	76.8%	0.189	84.5%	0.182	70.8%	0.203

Notes: For each of the four estimation criteria (LS, SDLS, LNLS, and QML), we report the percentages (twenty-eight tickers times six periods = 168 possible instances) with which the various specifications are in the 75% Model Confidence Set and the average loss. For every specification, the lowest loss (across estimation criteria) is reported in bold.

Clearly, it would be a hasty generalization to extend such recommendation to medium and small caps or other asset classes.⁸

3. **Is there an overall best OOS specification?** The short answer is: yes, but not overwhelmingly so, and there are some disappointing performances. Tables 10–13 provide further IS and OOS performance measures for when the FL functional is used as EC. The striking result is that MLOG produces the best OOS specifications 60.12%, 62.50%, 66.67%, and 53.57% of the times for LS, SDLS, LNLS, and QML, respectively. In fact, what emerges is that the specification producing the smallest average OOS values of the FL considered is MLOG(2, 1) followed very closely by MLOG(1, 1) and MVOL(2, 1). In contrast, HAR scores the best OOS specification just 7.74%, 10.71%, 8.33%, and

8 A judicious approach would suggest to rely on model selection procedures unless the superiority of a specific model for a given asset class is substantiated by empirical evidence. Similarly, the performance of a maintained model should be monitored over time and asset classes to detect significant changes and intervene when necessary.

Table 9. QML FL for OOS evaluations

Model	LS		SDLS		LNLS		QML	
MVAR (1,1)	48.2%	1.171	55.4%	1.142	63.7%	1.140	50.6%	1.135
(2,1)	62.5%	1.166	71.4%	1.141	76.8%	1.140	57.7%	1.135
MVOL (1,1)	53.6%	1.158	61.3%	1.136	70.2%	1.139	54.8%	1.134
(2,1)	62.5%	1.154	82.7%	1.136	87.5%	1.139	69.0%	1.133
MLOG (1,1)	70.2%	1.146	70.2%	1.134	76.8%	1.139	61.3%	1.133
(2,1)	69.6%	1.147	91.7%	1.134	95.8%	1.139	94.0%	1.132
HAR	63.1%	1.167	71.4%	1.141	79.2%	1.140	61.3%	1.135
LNHAR	58.9%	1.162	76.2%	1.136	84.5%	1.140	69.6%	1.134

Notes: For each of the four estimation criteria (LS, SDLS, LNLS, and QML), we report the percentages (twenty-eight tickers times six periods = 168 possible instances) with which the various specifications are in the 75% Model Confidence Set and the average loss. For every specification, the lowest loss (across estimation criteria) is reported in bold.

Table 10. LS FL and EC

Model	IS	OOS	OOS IS	A-Loss	M-Loss
MVAR (1,1)	0.00%	1.79%	0.00%	13.2932	1.2859
(2,1)	0.60%	7.74%	0.00%	13.2254	1.1525
MVOL (1,1)	0.00%	1.19%	0.00%	12.4442	1.1717
(2,1)	0.00%	10.12%	0.00%	12.4045	1.1243
MLOG (1,1)	44.05%	33.33%	32.43%	12.0195**	1.1647
(2,1)	39.29%	26.79%	16.67%	12.0863	1.1457
HAR	4.76%	7.74%	0.00%	13.3150	1.1433
LNHAR	11.31%	11.31%	42.11%	12.7265**	1.1624
Best IS BIC		25.00%		12.2505	1.1554

Notes: The first three columns report the percentage a specification is best: IS, IS (based on BIC); OOS, OOS (LS loss); OOS|IS, best OOS, limiting to the best IS. A-Loss and M-Loss are the average and median losses. The Best IS BIC row contains OOS results for the best specifications selected IS by the information criterion. Diebold–Mariano tests are conducted on the differences between individual A-Losses and corresponding value on the Best IS BIC row (*, **, *** signifies a better performance—10%, 5%, 1% significance—the reverse for the corresponding *, **, ***).

23.21% of the instances and generates average OOS FLs that are in between those of MVAR(1, 1) and MVAR(2, 1), with all other specifications exhibiting lower OOS FLs. LOG – HAR produces the best OOS fit 11.31%, 14.29%, 11.31%, and 7.14% of the instances and averages OOS FLs smaller than HAR and MVAR only.

4. **Is there an overall best IS specification?** The short answer is: same as with OOS. Even though the main focus of this study is on OOS forecasts, we recognize that there are instances where IS fit is of primary importance, for example, a reduced-form model capturing most relevant data features to be used as an *auxiliary model* in the indirect estimation of a *maintained model*. Tables 10–13 reproduce results similar to the OOS: neither HAR nor LOG – HAR emerges as best IS specification as they provide the best IS fit to the customary sequence of estimation criteria (4.76%, 4.17%, 1.19%, 13.10%) and (11.31%, 16.67%, 14.29%, 29.17%) of the instances, respectively. On the other

Table 11. SDLS FL and EC

Model	IS	OOS	OOS IS	A-Loss	M-Loss
MVAR (1,1)	0.00%	0.00%	0.00%	0.1429***	0.0817
(2,1)	0.00%	5.36%	0.00%	0.1383***	0.0776
MVOL (1,1)	0.00%	0.00%	0.00%	0.1283***	0.0775
(2,1)	5.95%	7.14%	0.00%	0.1259**	0.0759
MLOG (1,1)	42.26%	12.50%	9.86%	0.1213	0.0757
(2,1)	30.95%	50.00%	53.85%	0.1197***	0.0752
HAR	4.17%	10.71%	0.00%	0.1392***	0.0775
LNHAR	16.67%	14.29%	7.14%	0.1262**	0.0752
Best IS BIC		22.00%		0.1217	0.0757

Notes: The first three columns report the percentage a specification is best: IS, IS (based on *BIC*); OOS, OOS (SDLS loss); OOS|IS, best OOS, limiting to the best IS. A-Loss and M-Loss are the average and median losses. The Best IS BIC row contains OOS results for the best specifications selected IS by the information criterion. Diebold–Mariano tests are conducted on the differences between individual A-Losses and corresponding value on the Best IS BIC row (*, **, *** signifies a better performance—10%, 5%, 1% significance—the reverse for the corresponding *, **, ***).

hand, MLOG is selected as best IS specification in 20% more instances than LOG – HAR for QML EC and 400% more for the quadratic criteria.

Jointly, HAR and LOG – HAR are the preferred specifications 16.07%, 20.84%, 15.48%, and 42.27% of the instances for LS, SDLS, LNLS, and QML estimation criteria, respectively. However, the corresponding performance of MLOG of 83.34%, 73.21%, 76.79%, and 35.12% is substantially better (except for QML). Furthermore, if the HAR and LOG – HAR pair is compared with the MLOG and MVOL pair, the latter is found to provide the best IS for any EC considered.

To these, we add another question aimed at providing a robustness check:

5. **Is the good performance of the log-specification driven by a few large jumps?** The short answer is no. The question is whether the robustness afforded by the log-specification proves to be beneficial when we consider a time series of continuous variation only. Paralleling [Tables 12](#) and [13](#) (reporting the results for given LNLS, respectively QML EC and FL, and providing the main support in favor of the log-specification), we concentrate our robustness check on obtaining results from a jump robust variance estimator. For each asset, we identify a day t as a *jump*-day, testing whether the plain vanilla realized variance is significantly higher ($\alpha = 0.01$) than the realized bipower variation (cf. [Huang and Tauchen, 2005](#)); the *realized continuous variation* time-series, C_t , is equal to the plain vanilla realized variance in *normal* days, and to the bipower variation in *jump* days ([Andersen, Bollerslev and Huang, 2011](#), Eq. 12). Using C_t for each asset and subperiod, [Tables 14](#) and [15](#) qualitatively confirm all findings in [Tables 12](#) and [13](#), among which the relatively low capability of the BIC to identify what turns out to be the best OOS specification. In [Table 14](#), the only difference that arises in the switch from the RV_t to the C_t data is the improved performance of the MVOL(2, 1) specification, while the best OOS forecast performance of MLOG(2, 1) stands for both types of series. All other specifications produce statistically significant higher average losses. The

Table 12. LNLS FL and EC

Model	IS	OOS	OOS IS	A-Loss	M-Loss
MVAR (1,1)	0.00%	0.00%	0.00%	0.1934 ^{***}	0.1821
(2,1)	0.00%	6.55%	0.00%	0.1884 ^{***}	0.1767
MVOL (1,1)	0.00%	0.60%	0.00%	0.1861 ^{***}	0.1771
(2,1)	7.74%	6.55%	7.69%	0.1809 ^{***}	0.1724
MLOG (1,1)	1.79%	10.12%	0.00%	0.1821 ^{***}	0.1744
(2,1)	75.00%	56.55%	53.17%	0.1779 ^{***}	0.1704
HAR	1.19%	8.33%	0.00%	0.1887 ^{***}	0.1763
LNHAR	14.29%	11.31%	8.33%	0.1821 ^{***}	0.1739
Best IS BIC		41.00%		0.1787	0.1712

Notes: The first three columns report the percentage a specification is best: IS, IS (based on *BIC*); OOS, OOS (LNLS loss); OOS|IS, best OOS, limiting to the best IS. A-Loss and M-Loss are the average and median losses. The Best IS BIC row contains OOS results for the best specifications selected IS by the information criterion. Diebold–Mariano tests are conducted on the differences between individual A-Losses and corresponding value on the Best IS BIC row (*, **, *** signifies a better performance—10%, 5%, 1% significance—the reverse for the corresponding **, ***, ***).

Table 13. QML FL and EC

Model	IS	OOS	OOS IS	A-Loss	M-Loss
MVAR (1,1)	6.55%	2.38%	0.00%	1.1354 ^{**}	1.1108
(2,1)	0.00%	6.55%	0.00%	1.1346	1.1057
MVOL (1,1)	16.07%	2.38%	0.00%	1.1338 [*]	1.1086
(2,1)	0.00%	4.76%	0.00%	1.1329 ^{***}	1.1053
MLOG (1,1)	32.74%	19.05%	7.27%	1.1328 ^{***}	1.1070
(2,1)	2.38%	34.52%	0.00%	1.1323 ^{***}	1.1053
HAR	13.10%	23.21%	9.09%	1.1346	1.1074
LNHAR	29.17%	7.14%	4.08%	1.1343	1.1073
Best IS BIC		4.00%		1.1344	1.1070

Notes: The first three columns report the percentage a specification is best: IS, IS (based on *BIC*); OOS, OOS (QML loss); OOS|IS, best OOS, limiting to the best IS. A-Loss and M-Loss are the average and median losses. The Best IS BIC row contains OOS results for the best specifications selected IS by the information criterion. Diebold–Mariano tests are conducted on the differences between individual A-Losses and corresponding value on the Best IS BIC row (*, **, *** signifies a better performance—10%, 5%, 1% significance—the reverse for the corresponding **, ***, ***).

only minor difference in Table 15 is that the ranks of best (MLOG(2, 1)) and second-best (MVOL(2, 1)) specifications are reversed.

5 General Discussion

5.1 On the Estimation Criteria

A possible reading key to the empirical findings promoting LNLS as the preferred EC for LS and SDLS FLs is the balance it strikes between bias and variance of the estimated parameters, and consequently of the forecasts. In an ideal setting in which estimated model and

Table 14. Analysis from Table 11 repeated on continuous variation data

Model	IS	OOS	OOS IS	A-Loss	M-Loss
MVAR (1,1)	0.00%	1.79%	0.00%	0.2218 _{***}	0.2101
(2,1)	1.79%	9.52%	0.00%	0.2164 _{***}	0.2064
MVOL (1,1)	1.79%	2.98%	0.00%	0.2165 _{***}	0.2070
(2,1)	50.00%	19.64%	15.48%	0.2115 _{***}	0.2034
MLOG (1,1)	0.69%	7.14%	0.00%	0.2149 _{***}	0.2070
(2,1)	26.19%	36.90%	25.00%	0.2107 _{***}	0.2035
HAR	8.33%	15.48%	7.14%	0.2167 _{***}	0.2068
LNHAR	11.31%	11.31%	10.53%	0.2137 _{***}	0.2052
Best IS BIC		16.00%		0.2124	0.2049

Notes: LNLS FL and EC. The first three columns report the percentage a specification is best: IS, IS (based on BIC); OOS, OOS (LNLS loss); OOS|IS, best OOS, limiting to the best IS. A-Loss and M-Loss are the average and median losses. The Best IS BIC row contains OOS results for the best specifications selected IS by the information criterion. Diebold–Mariano tests are conducted on the differences between individual A-Losses and corresponding value on the Best IS BIC row (*, **, *** signifies a better performance—10%, 5%, 1% significance—the reverse for the corresponding *, **, ***).

Table 15. Analysis from Table 12 repeated on continuous variation data

Model	IS	OOS	OOS IS	A-Loss	M-Loss
MVAR (1,1)	18.45%	4.17%	0.00%	1.0685 _{**}	1.0412
(2,1)	0.00%	13.69%	0.00%	1.0670 _{**}	1.0395
MVOL (1,1)	28.57%	2.98%	0.00%	1.0675	1.0413
(2,1)	0.00%	14.88%	0.00%	1.0660 _{***}	1.0398
MLOG (1,1)	5.95%	12.50%	10.00%	1.0677	1.0427
(2,1)	1.19%	19.64%	0.00%	1.0666 _{**}	1.0413
HAR	35.12%	26.79%	16.95%	1.0671	1.0374
LNHAR	10.71%	5.36%	0.00%	1.0676	1.0399
Best IS BIC		6.00%		1.0678	1.0436

Notes: QML FL and EC. The first three columns report the percentage a specification is best: IS, IS (based on BIC); OOS, OOS (QML loss); OOS|IS, best OOS, limiting to the best IS. A-Loss and M-Loss are the average and median losses. The Best IS BIC row contains OOS results for the best specifications selected IS by the information criterion. Diebold–Mariano tests are conducted on the differences between individual A-Losses and corresponding value on the Best IS BIC row (*, **, *** signifies a better performance—10%, 5%, 1% significance—the reverse for the corresponding *, **, ***).

data generating process coincide, the results of Hansen and Dumitrescu (2018) apply and the LNLS EC generates forecasts that are inferior⁹ to those of LS and SDLS EC when the latter are the FLs. On the other hand, in the case of misalignment between model and data generating process, the standard bias–variance interpretation of the LS and SDLS forecasting mean squared error applies. It follows that bias and variance of parameter estimates

9 LNLS delivers unconditionally unbiased predictions of $\ln RV_t$ but biased predictions of RV_t (LS FL) and $RV_t^{1/2}$ (SDLS FL) in a context where unbiasedness alone is paramount to the minimization of the forecasting loss.

and forecasts are smallest for LNLS (log transformation), and smaller for SDLS (square root transformation¹⁰) than LS (no transformation).

One example of misalignment between model and data generating process is when the former is not able to accommodate occasional spikes in volatility (e.g., a flash crash). In such a case we are dealing with influential observations, with two possible effects: one relates to estimation and the possibility of biased estimates,¹¹ the other to forecasting, since such a (lagged) observation will exert its impact on the subsequent predictions in view of the estimated persistence. As far as the first effect is concerned, we note that, given a prediction $\hat{\sigma}_t^2$ the impact of an influential observation on the generalized residual, $\hat{\varepsilon}_t = g(RV_t) - g(\hat{\sigma}_t^2)$ for the quadratic criteria (Section 1.1) and on $\hat{\varepsilon}_t = RV_t - \hat{\sigma}_t^2$ for the QML criterion (Section 1.2), is smallest for LNLS followed by SDLS and then by LS and QML equally. Moreover, the impact of such an observation on the first-order conditions is inversely proportional to $\hat{\sigma}_t^n$ with $n = 0$ for LS, $n = 1$ for SDLS, $n = 2$ for LNLS and $n = 4$ for QML, respectively.

5.2 The Effect of (2, 1) Parameterizations on Autocorrelations

While the relevance of the moving-average coefficient β_1 is in line with expectations, the finding that adding an α_2 term produces better forecasts (Tables 6–9) is somewhat surprising in view of the long-established result that a (1, 1) specification for conditional variance suffices (e.g., Hansen and Lunde, 2005). Figure 1 reports the distributions of the (2, 1) parameters estimated by LNLS and QML. The striking empirical regularity is the negative sign of the $\hat{\alpha}_2$ for every specification, asset, and subsample period. Empirically, such a negative lag-2 coefficient has the remarkable effect of dynamically mitigating the impact of the observed lagged volatility, which is particularly useful in the case of an influential observation: at lag 1, it would have the customary impact of increasing predicted volatility; at lag 2, the negative coefficient results in a quicker dampening effect given that such a spike in volatility is isolated and not accompanied by subsequent high values.

Tables 16–19 report IS diagnostics in terms of percentage of instances that the residuals' autocorrelations at lags 1–5, 10, 15, and 20 test significant at 5%. For the quadratic estimation criteria, the (2, 1) and HAR specifications (except LOG – HAR estimated by LS) exhibit statistically insignificant lag-1 autocorrelations, as opposed to the (1, 1) parameterizations. However, HAR and (1, 1) specifications display higher percentages of significant lag-2 autocorrelations than the (2, 1) with $\alpha_2 < 0$. The QML EC favors a more balanced cleaning of the autocorrelations at lags 1 (higher than other EC) and 2 (lower than other EC). Incidentally, MLOG(2, 1) estimated by LNLS, which produces the best OOS forecasts, presents the lowest percentage of significant autocorrelations.

5.3 Long Memory Approximation

In what follows, we show that our (2, 1) specifications replicate the ability to approximate the long memory pattern observed in the autocorrelation of realized variances, a feature which has made the HAR model popular. In Figure 2, we show the cross-sectional average (over twenty-eight assets) of the empirical autocorrelations of realized variances (labeled

10 Closer similarity between log and square root than log and linear helps explain the empirical finding that SDLS comes in a close second in the rankings of EC not matching the FL (excluding QML).

11 For a detailed treatment of the robustness of M-estimators, see Hampel et al. (2005).

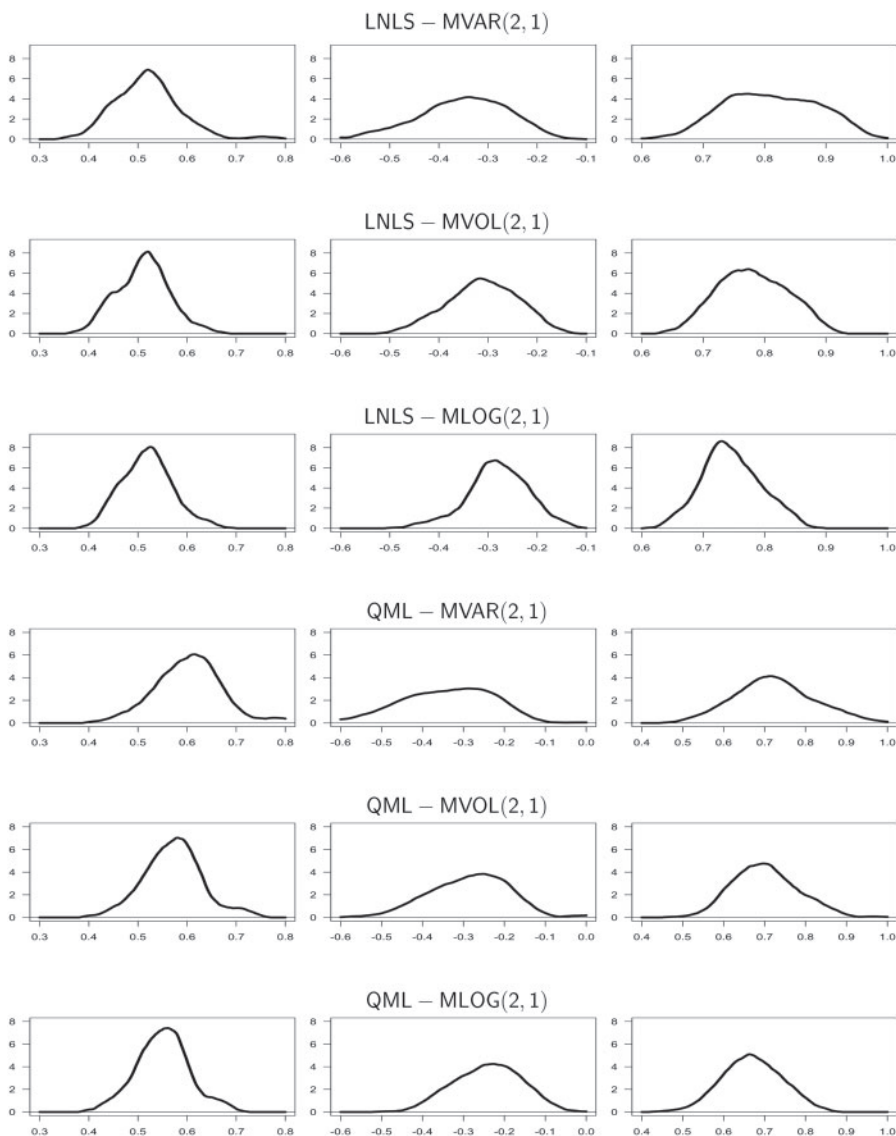


Figure 1. Parameter density estimates over the six IS periods and all the assets considered. From left to right, the columns contain the densities of the parameters α_1 , α_2 , and β_1 for the three specifications reported.

DATA) and the autocorrelation functions from several models.¹² Concentrating on the tail of the correlogram, we notice that the (2, 1) specifications and the HAR match the DATA from lag 3, respectively, 8/9, and that, from lag 10 on, they are substantially on top of one

12 Using the cross-sectional average of the parameter estimates for the (2010–2014) subsample and 10^6 Monte Carlo simulations, the results for MVAR(2, 1), MVOL(2, 1), MLOG(2, 1), HAR and LOG – HAR are used to plot the respective autocorrelation functions.

Table 16. LS EC, IS residual diagnostics

Model	1	2	3	4	5	10	15	20
MVAR (1,1)	17.26%	33.33%	65.48%	35.12%	11.31%	34.52%	34.52%	6.55%
(2,1)	0.00%	32.14%	49.40%	38.69%	11.31%	27.98%	35.71%	9.52%
MVOL (1,1)	13.10%	37.50%	68.45%	42.26%	9.52%	35.12%	32.74%	5.36%
(2,1)	0.60%	18.45%	50.00%	44.64%	15.48%	39.29%	33.93%	11.90%
MLOG (1,1)	28.57%	42.86%	67.86%	51.19%	8.33%	34.52%	33.33%	9.52%
(2,1)	0.00%	14.88%	54.76%	50.60%	18.45%	42.26%	38.69%	13.10%
HAR	0.00%	55.95%	44.64%	34.52%	33.33%	28.57%	35.12%	0.00%
LNHAR	29.76%	51.79%	46.43%	54.17%	16.67%	29.76%	32.74%	4.76%

Notes: For each specification we report the percentage (twenty-eight tickers times six periods = 168 possible instances) of 5% significant residual autocorrelations at lags 1–5, 10, 15, and 20.

Table 17. SDLS EC, IS residual diagnostics

Model	1	2	3	4	5	10	15	20
MVAR (1,1)	20.83%	16.07%	80.95%	16.67%	7.74%	27.38%	43.45%	3.57%
(2,1)	0.00%	56.55%	45.24%	25.00%	7.74%	14.29%	38.69%	0.00%
MVOL (1,1)	7.14%	16.07%	85.12%	16.07%	5.36%	22.02%	36.31%	2.98%
(2,1)	0.00%	35.71%	47.02%	18.45%	4.76%	10.12%	30.95%	0.00%
MLOG (1,1)	1.79%	22.62%	86.31%	17.26%	2.38%	22.02%	30.95%	2.98%
(2,1)	0.60%	20.83%	50.00%	19.05%	2.38%	14.29%	28.57%	0.60%
HAR	5.36%	69.05%	36.90%	17.86%	20.24%	18.45%	36.31%	0.60%
LNHAR	7.14%	60.12%	33.33%	23.21%	7.14%	16.07%	32.74%	0.60%

Notes: For each specification we report the percentage (twenty-eight tickers times six periods = 168 possible instances) of 5% significant residual autocorrelations at lags 1–5, 10, 15, and 20.

Table 18. LNLS EC, IS residual diagnostics

Model	1	2	3	4	5	10	15	20
MVAR (1,1)	11.90%	19.64%	63.10%	5.36%	2.98%	16.07%	7.14%	1.79%
(2,1)	7.74%	26.19%	2.38%	10.12%	6.55%	1.79%	2.38%	0.00%
MVOL (1,1)	11.31%	20.24%	70.83%	6.55%	2.38%	11.90%	5.95%	1.79%
(2,1)	1.79%	17.86%	1.19%	0.60%	2.38%	0.60%	2.38%	0.00%
MLOG (1,1)	10.71%	23.81%	76.19%	7.74%	1.19%	9.52%	5.36%	1.19%
(2,1)	0.00%	4.17%	1.19%	0.00%	0.60%	1.19%	2.38%	0.60%
HAR	7.74%	40.48%	1.79%	1.19%	4.76%	3.57%	3.57%	0.00%
LNHAR	3.57%	35.12%	0.00%	0.00%	0.60%	2.98%	2.38%	0.00%

Notes: For each specification we report the percentage (twenty-eight tickers times six periods = 168 possible instances) of 5% significant residual autocorrelations at lags 1–5, 10, 15, and 20.

Table 19. QML EC, IS residual diagnostics

Model	1	2	3	4	5	10	15	20
MVAR (1,1)	30.95%	11.31%	34.52%	5.95%	2.38%	12.50%	23.21%	2.38%
(2,1)	13.69%	16.07%	5.36%	9.52%	6.55%	5.95%	14.88%	0.00%
MVOL (1,1)	16.07%	19.05%	45.24%	8.33%	1.79%	10.71%	17.26%	3.57%
(2,1)	13.10%	8.93%	7.74%	6.55%	4.17%	4.17%	10.71%	0.00%
MLOG (1,1)	4.76%	39.29%	63.10%	10.71%	1.79%	9.52%	11.31%	3.57%
(2,1)	17.26%	4.17%	17.26%	5.36%	1.79%	4.17%	5.95%	0.60%
HAR	20.24%	26.79%	5.36%	10.12%	5.95%	6.55%	14.29%	0.00%
LNHAR	32.74%	14.88%	2.38%	2.38%	7.74%	4.76%	5.95%	0.00%

Notes: For each specification we report the percentage (twenty-eight tickers times six periods = 168 possible instances) of 5% significant residual autocorrelations at lags 1–5, 10, 15, and 20.

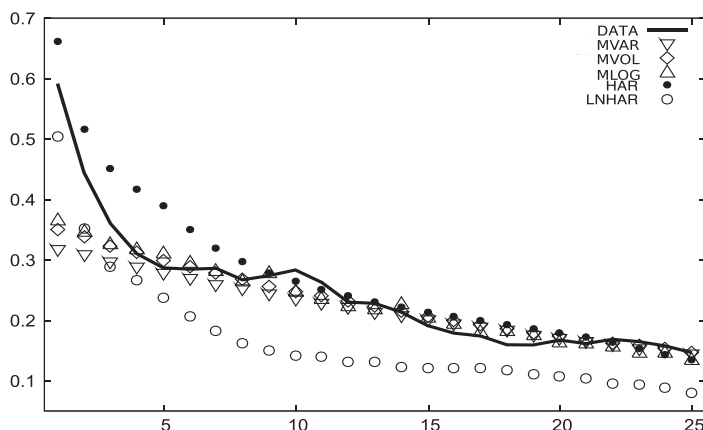


Figure 2. Autocorrelation functions for the variance processes with parameters equal to the cross-sectional averages (over twenty-eight assets) of the LNLS estimates in the most recent subsample (2010–2014). MVAR, MVOL, and MLOG specifications refer to the (2, 1) parameterizations. Autocorrelations are computed over 10^6 simulations with additive standard Gaussian innovations. DATA is the cross-sectional average of the autocorrelation functions of realized variances.

another, both replicating long-range dependence. One may note that LOG – HAR follows the profile of the decay, but underestimates the magnitude of the autocorrelations at all lags. At lag-1, we find autocorrelations that are approximately 0.6 for the DATA, 0.67 and 0.5 for the HAR and LOG – HAR, respectively, but only 0.3–0.4 for the (2, 1) specifications.

The lack of ability by the (2, 1) models in replicating the autocorrelation profile of the DATA at lag 1 is not an impediment to their better performance in forecasting relative to HAR. In fact matching the DATA autocorrelation pattern is not necessarily conducive to superior forecasts. Consider the following example in which RV_t and two competing forecasts $\hat{\sigma}_{1,t}^2$ and $\hat{\sigma}_{2,t}^2$ are given by:

$$RV_t = \sigma_t^2 + \nu_t, \quad \hat{\sigma}_{1,t}^2 = \sigma_t^2 + S_1 \eta_{1,t}, \quad \hat{\sigma}_{2,t}^2 = \sigma_t^2 + S_2 \eta_{2,t},$$

with $1 < S_2 < S_1$; ν_t and $\eta_{2,t}$ i.i.d. standard random variables; σ_t^2 a process with lag- j autocovariance γ_j , and $\gamma_0 = 1$; $\eta_{1,t}$ a standard random variable with lag- j autocovariance

$\psi_j = 0.5S_1^{-2}(S_1^2 - 1)\gamma_j$. Then, the autocorrelations of the three processes are $\rho_j(\text{RV}) = 0.5\gamma_j$, $\rho_j(\hat{\sigma}_1^2) = 0.5\gamma_j$ and $\rho_j(\hat{\sigma}_2^2) = (1 + S_2^2)^{-1}\gamma_j$, respectively. The forecasting mean squared errors of the competing specifications are: $\mathbb{E}[(\text{RV}_t - \hat{\sigma}_{1,t}^2)^2] = \mathbb{E}[(\text{RV}_t - \sigma_t^2)^2] + S_1^2$ and $\mathbb{E}[(\text{RV}_t - \hat{\sigma}_{2,t}^2)^2] = \mathbb{E}[(\text{RV}_t - \sigma_t^2)^2] + S_2^2$. Therefore, for $1 < S_2 < S_1$, the autocorrelation structure of $\hat{\sigma}_{1,t}^2$ is an exact match of that of RV_t while $\hat{\sigma}_{2,t}^2$ exhibits lower autocorrelations. This notwithstanding, the forecasting mean squared error of $\hat{\sigma}_{2,t}^2$ is lower than that of $\hat{\sigma}_{1,t}^2$.

Alternative specifications may thus produce more precise forecasts than HAR even though the lag 1 autocorrelation of HAR is closer to the one from the DATA. A possible explanation is that autoregressive models with (homoskedastic) measurement error follow an ARMA process (e.g., Staudenmayer and Buonaccorsi, 2005; Bollerslev, Patton and Quaadvlieg, 2016) and this may account for a good performance of the (2, 1) specifications. Moreover, Bollerslev, Patton and Quaadvlieg (2016) highlighted the possible misspecification of the HAR in the presence of measurement errors. This argument is supported by the contemporaneous correlations between realized variances and the model forecasts: 0.6231 for MLOG(2, 1), 0.6219 for MVOL(2, 1), 0.6174 for LOG – HAR, 0.6153 for MVAR(2, 1), and 0.5026 for HAR.

6 Conclusions

In this article, we start from the idea that OOS forecast evaluation relies on the choice of a distance between predicted outcomes and actual values of realized variance that reflects subjective preferences. Conditional on a specific OOS FL, we show that the same functional form may not be the most appropriate choice as an IS EC in which distance between actual and fitted values is used to deliver parameter estimates. To this end, we have examined several models reminiscent of well-known GARCH parameterizations alongside HAR specifications and we have handled several combinations of IS estimation criteria and OOS FLs.

We find the (2, 1) parameterizations to be best suited to model the dynamics and forecast realized variances, volatilities, and log-variances. In particular, the specification delivering the best OOS forecasts is MLOG together with MVOL, a result qualitatively confirmed by a robustness check performed on continuous variation time series. Interestingly, our empirical findings point to the presence of a negative lag-2 coefficient in all estimated (2, 1) specifications. We interpret it as a dampening agent which, by limiting to one day most of the impact of a shock to variance, induces a mimicking effect of long memory properties that are indistinguishable from those of the HAR and more sustained than those of the LOG – HAR.

With respect to the estimators considered, we find that for all quadratic FLs, models estimated using the LNLS EC provide the best OOS forecasts. On the other hand, for the QML FL, the QML EC does better than LNLS, although marginally so. Our findings further suggest a judicious approach to model selection which should rely on information criteria unless the superiority of a specific model is substantiated by empirical evidence.


References

Andersen, T. G., T. Bollerslev, and F. X. Diebold. 2007. Roughing It Up: Including Jump Components in the Measurement, Modeling and Forecasting of Return Volatility. *Review of Economics and Statistics* 89: 701–720.

- Andersen, T. G., T. Bollerslev, and X. Huang. 2011. A Reduced Form Framework for Modeling Volatility of Speculative Prices Based on Realized Variation Measures. *Journal of Econometrics* 160: 176–189.
- Barndorff-Nielsen, O. E., P. R. Hansen, A. Lunde, and N. Shephard. 2008. Designing Realised Kernels to Measure the Ex-Post Variation of Equity Prices in the Presence of Noise. *Econometrica* 76: 1481–1536.
- Barndorff-Nielsen, O. E., P. R. Hansen, A. Lunde, and N. Shephard. 2011. Multivariate Realised Kernels: Consistent Positive Semi-Definite Estimators of the Covariation of Equity Prices with Noise and Non-Synchronous Trading. *Journal of Econometrics* 162: 149–169.
- Bauwens, L., F. Galli, and P. Giot. 2008. The Moments of log-ACD Models. *Quantitative and Qualitative Analysis in Social Sciences* 2: 1–28.
- Bollerslev, T. 1986. Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics* 31: 307–327.
- Bollerslev, T., A. J. Patton, and R. Quaedvlieg. 2016. Exploiting the Errors: A Simple Approach for Improved Volatility Forecasting. *Journal of Econometrics* 192: 1–18.
- Brownlees, C. T., F. Cipollini, and G. M. Gallo. 2012. “Multiplicative Error Models.” In L. Bauwens, C. Hafner, and S. Laurent (eds.), *Volatility Models and Their Applications*, Hoboken, NJ: Wiley, pp. 223–47.
- Brownlees, C. T. and G. M. Gallo. 2006. Financial Econometric Analysis at Ultra-High Frequency: Data Handling Concerns. *Computational Statistics & Data Analysis* 51: 2232–2245.
- Christoffersen, P. and F. Diebold. 1997. Optimal Prediction under Asymmetric Loss. *Econometric Theory* 13: 808–817.
- Christoffersen, P. and K. Jacobs. 2004. The Importance of the Loss Function in Option Valuation. *Journal of Financial Economics* 72: 291–318.
- Cipollini, F., R. F. Engle, and G. M. Gallo. 2013. Semiparametric Vector MEM. *Journal of Applied Econometrics* 28: 1067–1086.
- Corsi, F. 2009. A Simple Approximate Long-Memory Model of Realized Volatility. *Journal of Financial Econometrics* 7: 174–196.
- Engle, R. F. and G. M. Gallo. 2006. A Multiple Indicators Model for Volatility Using Intra-Daily Data. *Journal of Econometrics* 131: 3–27.
- Geweke, J. 1986. Modeling the Persistence in Conditional Variances: A Comment. *Econometric Reviews* 5: 57–61.
- Gouriéroux, C., A. Monfort, and A. Trognon. 1984. Pseudo Maximum Likelihood Methods: Theory. *Econometrica* 52: 681–700.
- Hampel, F., E. Ronchetti, P. Rousseeuw, and W. Stahel. 2005. *Robust Statistics: The Approach Based on Influence Functions*, 2nd edn. New York: Wiley.
- Hansen, P. R. and E. Dumitrescu. 2018. *Parameter Estimation with out-of-Sample Objective*, WP SSRN 3178896. Chapel Hill, NC: University of North Carolina.
- Hansen, P. R. and A. Lunde. 2005. A Forecast Comparison of Volatility Models: Does Anything Beat a GARCH(1,1)? *Journal of Applied Econometrics* 20: 873–889.
- Hansen, P. R., A. Lunde, and J. M. Nason. 2011. The Model Confidence Set. *Econometrica* 79: 453–497.
- Huang, X. and G. Tauchen. 2005. The Relative Contribution of Jumps to Total Price Variance. *Journal of Financial Econometrics* 3: 456–499.
- Nelson, D. B. 1991. Conditional Heteroskedasticity in Asset Returns: A New Approach. *Econometrica* 59: 347–370.
- Park, S. and O. Linton. 2012. “Realized Volatility: Theory and Applications.” In L. Bauwens, C. Hafner and S. Laurent (eds.), *Volatility Models and Their Applications*, Wiley, pp. 319–45.

- Patton, A. 2011. Volatility Forecast Comparison Using Imperfect Volatility Proxies. *Journal of Econometrics* 160: 246–256.
- Staudenmayer, J. and J. Buonaccorsi. 2005. Measurement Error in Linear Autoregressive Models. *Journal of the American Statistical Association* 100: 841–852.
- Taylor, N. and Y. Xu. 2017. The Logarithmic Vector Multiplicative Error Model: An Application to High Frequency NYSE Stock Data. *Quantitative Finance* 17: 1021–1035.
- Zakoian, J.-M. 1994. Threshold Heteroskedastic Models. *Journal of Economic Dynamics and Control* 18: 931–955.

Using the Extremal Index for Value-at-Risk Backtesting*

Axel Bücher¹, Peter N. Posch² and Philipp Schmidtke ²

¹Heinrich Heine University Düsseldorf and ²TU Dortmund University

Address correspondence to Philipp Schmidtke, TU Dortmund University, Chair of Finance, Otto-Hahn-Str. 6, 44227 Dortmund, Germany, e-mail: philipp.schmidtke@udo.edu.

Received 23 August 2019; revised 23 August 2019; editorial decision 16 April 2020; accepted 22 April 2020

Abstract

We introduce a set of new Value-at-Risk independence backtests by establishing a connection between the independence property of Value-at-Risk forecasts and the extremal index, a general measure of extremal clustering of stationary sequences. For this purpose, we introduce a sequence of relative excess returns whose extremal index is to be estimated. We compare our backtest to both popular and recent competitors using Monte Carlo simulations and find considerable power in many scenarios. In an applied section, we perform realistic out-of-sample forecasts with common forecasting models and discuss advantages and pitfalls of our approach.

Key words: VaR backtesting, extremal index, independence, risk measures

JEL classification: C52, C53, C58

In spite of its usage as a risk measure for more than 20 years, researchers are still engaged in exploring new forecasting and backtesting procedures for the Value-at-Risk (VaR). The latter procedures are typically based on a statistical test which tries to assess whether a certain desirable property is met for the observed sequence of VaR violations: first, the concept of *correct unconditional coverage* aims at checking whether the number of overall violations is justifiable. From an academic perspective, we typically seek for a forecasting procedure which yields neither too many nor too few violations. On the other hand, regulators are usually interested in situations where the risk is not underestimated, resulting in a focus on

* The authors are grateful to the editor Prof. Diebold, two anonymous referees, whose comments on an earlier version of this manuscript lead to a significant improvement, participants at the IFABS 2018 Conference in Porto, and the Annual Meeting of the German Finance Association (DGF) 2018. This work has been supported in part by the Collaborative Research Center “Statistical Modeling of Nonlinear Dynamic Processes” (SFB 823, Subproject A7) of the German Research Foundation (DFG).

not too many violations. Second, the *correct independence* aspect focuses on possible serial dependence of violations and aims at checking whether the sequence of violations behaves like an independent sequence. This concept becomes most important if unconditional coverage is statistically satisfied, that is, an unconditional test cannot be rejected. In that case, a test using information about the way how violations occur has still potential to reject the forecasts. Available independence backtests may have power only with respect to a lack of independence, or with respect to both the lack of independence and of correct unconditional coverage. The latter ones are called *conditional coverage tests*.

In general, tackling the independence property is challenging. This is mainly due to the fact that risk forecasts deal with low-probability events and an often short testing sample. As a consequence, observing many violations is unlikely, which naturally results in small effective sample sizes and, therefore, bad power properties. In addition, some of the classical tests explicitly assume an alternative model incorporating a special kind of dependence, which may also result in a loss of power if, in fact, a more general form of dependence is present.

Despite these natural difficulties, the independence hypothesis itself is relevant. As a matter of fact, most financial time series exhibit large degrees of heteroskedasticity and, therefore, require time-changing risk forecasts. Renouncement would lead to a probably threatening violation clustering, something a sound risk management should always aim to prevent.

We contribute to the backtesting literature by introducing a new test for the independence hypothesis, which is particularly sensitive to deviations from independence among the most extreme observations. Unlike standard methods, the new test does not use solely the 0-1-violation sequence. Instead, we assess whether a series of VaR-adjusted returns, coined *relative excess returns*, exhibits a significant tendency for that its most extreme observations occur in clusters. As a measure for that tendency, we employ the *extremal index*, a natural measure of clustering of extreme observations stemming from extreme value theory. We implement the approach with two different extremal index estimators, the first one (Süveges and Davison, 2010) leading to a more classic 0-1-test, while the second one (Northrop, 2015; Berghaus and Bücher, 2018) enables the processing of more detailed information. We find considerable power improvements in many cases in comparison to common competing tests, with the second test often showing the most convincing results. Finally, it is important to note that our approach is not designed to have power against *incorrect unconditional coverage*, and that we do not present a unified test tackling *correct conditional coverage*. Thus, our approach can rather be regarded as a potentially powerful supplement to other tests.

As is well known, VaR lacks some important features of risk measures. The most common alternative measure is provided by the Expected Shortfall (ES), which will soon replace the VaR as the standard regulatory measure of risk for banks (BCBS, 2016). However, since VaR and ES are closely related, it does not come as a surprise that VaR and its backtests also play a prominent role in some ES backtests. For example, Kratz, Lok, and McNeil (2018) propose a joint backtest for several VaR levels as an intuitive way to implicitly backtest ES. A second example is BCBS (2016) itself, where out-of-sample backtesting is based on VaR as well. However, both in general and in the aforementioned examples, the issue of a possible lack of independence is rarely addressed. Since the implementation of our idea is relatively independent of the specific VaR level, we see this as a promising approach in this respect.

The remainder of this article is structured as follows. Section 1 provides preliminaries about the notation, a more detailed description of the backtesting problem, and mathematical details on the extremal index. Section 2 introduces our new approach of independence backtesting based on the extremal index. In Section 3, we perform a detailed analysis of the small-sample properties, while Section 4 focuses on some empirical implications. Finally, Section 5 concludes, while less important aspects are deferred to an [Online Appendix](#).

1 Preliminaries on Backtesting and the Extremal Index

In this section, we introduce our notation, review the essentials of VaR backtesting, and provide a brief introduction to the extremal index.

1.1 Backtesting the VaR

Consider a random return r_t of a financial asset in a period t , usually a day.¹ Suppose this return is continuously distributed with c.d.f. F_t , conditional on the information set \mathcal{F}_{t-1} which embodies all information up to period $t - 1$. We define the VaR at level p as $\text{VaR}_p^{(t)} := -F_t^{-1}(p)$, where F_t^{-1} denotes the generalized inverse of F_t . Throughout the article, we will refer to p as VaR level, usual values are 5% and 1%, whereas $q = 1 - p$ will be called the VaR confidence level. Note that, with this definition, we report large losses and hence VaRs as positive numbers.

A violation at time t occurs if $r_t < -\widehat{\text{VaR}}_p^{(t)}$, where $\widehat{\text{VaR}}_p^{(t)}$ denotes a forecast of the true VaR at period t , calculated based on information from \mathcal{F}_{t-1} . Using a series of VaR forecasts corresponding to observed returns r_1, \dots, r_n , we define the violation sequence $(I_t)_{t=1}^n$, by

$$I_t = \begin{cases} 1 \text{ (violation),} & \text{if } r_t < -\widehat{\text{VaR}}_p^{(t)} \\ 0 \text{ (compliance),} & \text{if } r_t \geq -\widehat{\text{VaR}}_p^{(t)} \end{cases}. \quad (1.1)$$

The time points t where violations occur, that is $I_t = 1$, are called violation times or violations indices. Suppose there are N_1 violations, that is, $N_1 = \{I_t = 1\}$, and order the violation times increasingly $t_1 < \dots < t_{N_1}$. We define the inter-violation durations D_i as $D_i := t_{i+1} - t_i$, where $i = 1, \dots, N_1 - 1$. If the VaR forecasts happen to be completely correct, that is $\widehat{\text{VaR}}_p^{(t)} = \text{VaR}_p^{(t)}$ for all t , then the violation sequence forms an i.i.d. Bernoulli sequence with success probability p , that is $I_t \stackrel{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$. This implies $N_1 = \sum_{t=1}^n I_t \sim \text{Binom}(n, p)$ and $D_i \stackrel{\text{i.i.d.}}{\sim} \text{Geom}(p)$.

The goal of backtesting is to assess whether a sequence of n *ex ante* VaR forecasts are appropriate in relation to the realized returns. This is usually done by stressing one of the above-mentioned properties of the violation sequence or the durations.

1 The methods introduced in this article are applicable for arbitrary period lengths such as hourly, daily, or monthly periods if the returns are non-overlapping. We focus on daily returns in subsequent analyses because this is the most common case in the academic and regulatory literature. For the latter, see, for instance, [BCBS \(1996b\)](#) or [BCBS \(2016\)](#).

Since [Christoffersen \(1998\)](#) backtests are classified according to their focus, see also the discussion in the introduction. The property of forecasts being completely unsuspecting is called correct *conditional coverage* (cc) and may be written as

$$\text{cc} : I_t \stackrel{\text{i.i.d.}}{\sim} \text{Bernoulli}(p), \quad t = 1, \dots, n. \quad (1.2)$$

Before this term was introduced, assessing VaR forecasts was solely concerned with the aspect of *unconditional coverage* (uc), which is defined by

$$\text{uc} : E(I_t) = p, \quad t = 1, \dots, n. \quad (1.3)$$

In other words, uc is concerned about whether the frequency of violations is reasonable in the sense that, for all time points t , the probability of observing a violation equals p , which is the probability had the true VaR been used for the calculation of I_t . A simple way to get a first impression about the latter property is to calculate the actual number of violations N_1 and compare the result to its expectation np under the assumption $\widehat{\text{VaR}}_p^{(t)} = \text{VaR}_p^{(t)}$. See, for example, [Kupiec \(1995\)](#) for an early test or [BCBS \(1996b\)](#) for the Traffic Light Approach used by the Basel Committee.

Unconditional coverage is complemented by the *independence property* (ind), given by

$$\text{ind} : I_1, \dots, I_n \text{ are stochastically independent.} \quad (1.4)$$

Rather than on how many violations occur, the focus is on how the violations occur over time. A simple graphical way to check this property is to look at a plot of VaR violations and assess visually whether there are any patterns. However, detecting a failure of the independence property can be fairly hard due to the natural scarcity of violations if the VaR level is sufficiently small or the backtesting sample is not large enough. Still, possible dependence among violations can be extremely important for risk managers, as subsequent violations can sum up and result in an overall loss of threatening magnitude.

From a more technical perspective, plain independence of violations does not necessarily imply the absence of violation clustering. This can be seen by an example in [Ziggel et al. \(2014\)](#), where I_t and I_{t-k} are, in fact, independent but clustering can still happen.²

1.2 The Extremal Index

Loosely spoken, the extremal index θ , a parameter in the interval $[0, 1]$, measures the tendency of a (strictly) stationary time series to form temporal clusters of extreme values. The formal definition is as follows, see, for example, [Embrechts, Klüppelberg, and Mikosch \(1997, p. 416\)](#).

Definition 1.1. Let (e_t) be a strictly stationary sequence with stationary c.d.f. $F(x) = \Pr(e_1 \leq x)$ and let θ be a non-negative number. Assume that, for every $\tau > 0$, there exists a sequence $(u_n) = (u_n(\tau))$ such that

- 2 See equation 28 of [Ziggel et al. \(2014\)](#). The authors argue that the independence property should be replaced by an i.i.d. property. Although the prevention and detection of violation clustering is also our aim, we continue to speak of independence when we mean the absence of violation clustering in the remainder of the article.

$$\lim_{n \rightarrow \infty} n\Pr(e_1 > u_n) = \tau,$$

and

$$\lim_{n \rightarrow \infty} \Pr(M_n \leq u_n) = \exp(-\theta\tau),$$

where $M_n = \max\{e_1, \dots, e_n\}$. Then θ is called the *extremal index* of the sequence (e_t) , and it can be shown to lie necessarily in $[0, 1]$.

The definition is fairly abstract and certainly needs some explanation. Consider an i.i.d. sequence first, and assume that the c.d.f. F of e_1 is continuous and, for simplicity, invertible. For given $\tau > 0$, we may then choose $u_n = F^{-1}(1 - \tau/n)$ to guarantee that $n\Pr(e_1 > u_n) = n\{1 - F(F^{-1}(1 - \tau/n))\} = \tau$, that is, the first limiting relationship in the above definition is satisfied. Since $n\Pr(e_1 > u_n) = E[\sum_{i=1}^n 1(e_i > u_n)]$ by linearity of expectation, we obtain that we can expect, on average, to observe τ exceedances of the threshold u_n in a sequence of length n . At the same time, it may well happen that we do not observe a single exceedance of the threshold, and this event is exactly $\{M_n \leq u_n\}$. For the i.i.d. case, we obtain

$$\Pr(M_n \leq u_n) = \Pr(e_1 \leq u_n) \cdots \Pr(e_n \leq u_n) = F(u_n)^n = \left(1 - \frac{n\{1 - F(u_n)\}}{n}\right)^n,$$

which converges to $\exp(-\tau)$. As a consequence, we obtain that the extremal index of an i.i.d. sequence is $\theta = 1$. Note that $\theta = 1$ does not imply independence.

For more general time series, a similar calculation is typically much more difficult. It has, however, been shown that, under weak conditions on the serial dependence and if $\Pr(M_n \leq u_n)$ does converge, then the limit is always of the form $\exp(-\theta\tau)$ with θ being independent of the level τ , as requested in the above definition (Leadbetter, 1983). The extremal index has been shown to exist for many common time series models, including, for example, GARCH models (Mikosch and Starica, 2000), and is often smaller than 1 as in the i.i.d. case.

A common interpretation of the extremal index is as follows: the reciprocal of the extremal index, that is, $1/\theta$, represents, in a suitable elaborated asymptotic framework, the expected size of a temporal cluster of extreme observations, see Embrechts, Klüppelberg, and Mikosch (1997, p. 421). As a consequence, $\theta = 1$ means that extreme observations typically occur by oneself, while values below 1 means that extreme observations tend to occur in temporal clusters, that is, close by in time, with the expected number of ‘close-by-extreme-observations’ being equal to $1/\theta$. It is exactly this interpretation which leads us to consider backtests based on the extremal index in Section 2.

2 Backtesting Based on the Extremal Index

Given some arbitrary forecasts $\widehat{\text{VaR}}_p^{(t)}$, consider the following negative return-VaR ratio

$$e_t := e_t(\widehat{\text{VaR}}_p^{(t)}) := -\frac{r_t}{\widehat{\text{VaR}}_p^{(t)}}, \tag{2.1}$$

which we call *relative excess returns*. Similar to the fact that the use of correct VaR forecasts leads to an i.i.d. Bernoulli violation sequence, the use of correct forecasts should result in no (or only low) serial dependence of the sequence (e_t) .

Note, the negative sign in front of the ratio, which implies that by looking at the right tail of (e_t) , we essentially look at the left tail of (r_t) . There is an obvious relationship with the violation sequence (I_t) defined in Equation (1.1): we have $\{e_t > 1 \iff I_t = 1\}$ and $\{e_t \leq 1 \iff I_t = 0\}$, but (e_t) obviously contains much more information.

We propose to check for extremal clustering in the right tail of (e_t) by using the extremal index, whence the resulting tests can, in fact, be expected to be particularly sensitive to deviations from independence in the far right tail of e_t , that is, in the most important part for risk management needs. More precisely, recalling that an independent sequence has extremal index 1, we aim at checking for what we coin *no cluster property* (noc):

$$\text{noc} : \theta_e := \theta((e_t)_t) = 1. \tag{2.2}$$

In Section 2.1, we argue that this approach is sensible, at least for mean-scale models. In Section 2.2, we introduce estimators for the extremal index that will be used in Section 2.3 to formally define the new backtests. Sections 2.4 and 2.5 provide extensions to more general risk measures and distributional backtests, respectively.

2.1 Relative Excess Returns of Mean-Scale Models

It is instructive to consider the relative excess return series (e_t) , with $\widehat{\text{VaR}}_p^{(t)} = \text{VaR}_p^{(t)}$, in a general mean-scale model defined by

$$r_t = \mu_t + \sigma_t z_t, \quad \text{where } z_t \stackrel{\text{i.i.d.}}{\sim} F_z \tag{2.3}$$

and where $E(z_t) = 0$, $\text{Var}(z_t) = 1$, and μ_t, σ_t are \mathcal{F}_{t-1} -measurable. As a consequence, the conditional VaR using information up to $t - 1$ can be written as

$$\text{VaR}_p^{(t)} = -\mu_t - \sigma_t F_z^{-1}(p), \tag{2.4}$$

which implies that

$$e_t = \frac{\mu_t + \sigma_t z_t}{\mu_t + \sigma_t F_z^{-1}(p)}. \tag{2.5}$$

We are next going to argue that the sequence (e_t) is either an i.i.d. sequence (zero mean case) or at least approximately serially independent (non-zero mean case), in particular when looking only at the right tail. This suggests to backtest the VaR-forecasts by checking for serial independence or the absence of extremal clustering of the relative excesses (e_t) .

2.1.1 The zero mean case

If $\mu_t \equiv 0$, then Formula (2.5) simplifies to $e_t = z_t / F_z^{-1}(p)$. As a consequence, (e_t) is an i.i.d. sequence due to the i.i.d. property of the innovations (z_t) .

In practice, the possibility of a non-zero mean cannot be ruled out. However, it is often argued that financial returns show only insignificant means, see, for example, [Hansen and Lunde \(2005\)](#). In that paper, a large number of mean-scale models are examined with respect to their volatility forecasting performance, relative to simple specifications such as the classical GARCH(1,1) model. Three different specifications for the conditional mean are used and it is concluded that the performance is almost identical across the three versions. In other words, for financial returns, the mean is often negligible, especially in the short term. This stylized fact is also assured by the popularity of methods and models which explicitly use the assumption of zero conditional means. A prime example is provided by the famous square-root-of-time rule for time-scaling of the volatility and VaR. The rule is well appreciated among academics and practitioners and is even implemented in regulatory standards for VaR scaling from daily returns to 10-day returns ([BCBS, 1996a](#)).

2.1.2 The general case

Next, consider the general case with $\mu_t \neq 0$ being allowed. The event $\{e_t > y\}$ can then be rewritten as

$$S_t(y) = \left\{ z_t < y \left(\frac{\mu_t}{\sigma_t} + F_z^{-1}(p) \right) - \frac{\mu_t}{\sigma_t} \right\},$$

and this representation suggests that the relative excess returns are in general not serially independent: the events $\{e_t > y\}$ and $\{e_{t+1} > x\}$ are connected through the conditional mean and volatility. However, we argue that the serial dependence is actually either vanishing or low in certain typical cases.

The first case is $x = y = 1$, in which case $S_t(1) = \{z_t < F_z^{-1}(p)\} = \{I_t = 1\}$, which is obviously independent over time. In fact, we are left with the classical violation sequence (I_t) .

Next, in case of either $\mu_{t+1} \approx 0$ or $\sigma_{t+1} \rightarrow \infty$, we get at least approximate equality of $S_t(y)$ and $\{z_t < F_z^{-1}(p)\}$ and hence approximate serial independence of (e_t) . Note that large volatilities σ_t are typically present in periods of financial turmoil, which are in turn associated with our phenomenon of interest, that is, violation clustering.

More generally, the serial dependence vanishes for $x, y \geq 1$ whenever $-F_z^{-1}(p) \gg \mu_t/\sigma_t$ for all t with high probability, which is reasonable for large values of q . In that case, $S_t(y)$ implies $z_t \ll F_z^{-1}(p)$, so that only very small values of z_t may trigger the event $S_t(y)$. Since z_t is an i.i.d. sequence, the events $S_t(y)$ are approximately serially independent too, with high probability.

2.2 Estimators for the Extremal Index

Perhaps not surprisingly, a huge variety of estimators for the extremal index has been described in the literature. Early estimators include the blocks and the runs method, see [Smith and Weissman \(1994\)](#) or [Beirlant et al. \(2004\)](#) for an overview. In this section, we describe the classical blocks estimator and two more recent methods which will be applied in the subsequent parts of this article. In what follows, let e_1, \dots, e_n be an observed stretch from a strictly stationary time series whose extremal index exists and is larger than 0.

2.2.1 The classical blocks estimator

One of the most intuitive estimators is the classical blocks estimator, see [Smith and Weissman \(1994\)](#). This estimator is closely related to the definition of clusters and its relationship to the extremal index and relies on a block size $b = b_n$ and a threshold $u = u_n$ to be chosen by the statistician.

Divide the sample e_1, \dots, e_n into n/b disjoint blocks of size b .³ Let $M_i^{dj} = \max\{e_{(i-1)b+1}, \dots, e_{ib}\}$ denotes the maximum of the observations in the i th disjoint block. The set of exceedances within a block containing at least one exceedance (i.e., $M_i^{dj} > u$) is regarded as a cluster. Since $1/\theta$ is the expected cluster size, this suggests to set

$$\hat{\theta}_n^{CB} = \frac{\sum_{i=1}^{n/b} 1(M_i^{dj} > u)}{\sum_{i=1}^n 1(e_i > u)},$$

which equals the number of clusters over the number of exceedances and yields the classical blocks estimator.

2.2.2 The K -gap estimator

The K -gap estimator by [Süveges and Davison \(2010\)](#) is based on inter-exceedance times between the extreme observations (the latter bears some similarities with the duration times introduced in a backtesting context in Section 1.1). The foundations of the estimator are laid in [Ferro and Segers \(2003\)](#) where it is shown that the inter-exceedance times, appropriately standardized, weakly converge to a limiting mixture model. This remains true after truncation by the so-called gap parameter $K \in \mathbb{N}$, as shown for $K = 1$ in [Süveges \(2007\)](#) and in the general K -gap case in [Süveges and Davison \(2010\)](#).

The K -gap estimator does depend on a high threshold $u = u_n$ to be chosen by the statistician and is constructed as follows. Let $N_1 = \sum_{t=1}^n 1(e_t > u)$ denotes the number of exceedances of the threshold u . Let $1 \leq j_1 < \dots < j_{N_1} \leq n$ denotes the time points at which an exceedance has occurred, and let $T_i = j_{i+1} - j_i$ denotes the inter-exceedance time, for $i = 1, \dots, N_1 - 1$. The K -gaps are introduced by truncating with $K > 0$, that is

$$S_i^{(K)} = \max\{T_i - K, 0\}.$$

The mentioned limiting mixture model means that a transformed inter-exceedance time (K -gaps also) follows either an exponential distribution with mean θ (with probability θ , inter-exceedance time positive) or equals to zero (with probability $1 - \theta$). In general, the log-likelihood of $(S_i^{(K)})$ is intractable, but a pseudo-log-likelihood may be derived under the assumption of independence of the inter-exceedance times:

3 We assume that the number of blocks n/b is an integer. If this is not the case, a possible remainder block of smaller size than b must be discarded (typically at the beginning or the end of the observation period).

$$\log L_K(\theta; S_i^{(K)}) = (N_1 - 1 - N_C) \log(1 - \theta) + 2N_C \log \theta - \theta \sum_{i=1}^{N_1-1} \bar{F}(u) S_i^{(K)},$$

where $N_C = \sum_{i=1}^{N_1-1} 1(S_i^{(K)} \neq 0)$ and where $\bar{F}(x) = \Pr(e_1 > x)$. The maximization of this log-likelihood yields a closed-form estimator of the extremal index given by

$$\hat{\theta}_n^G = \hat{\theta}_n^G(u, K) = \frac{\Sigma_2 - (\Sigma_2^2 - 8N_C \Sigma_1)^{1/2}}{2\Sigma_1} \quad (2.6)$$

where $\Sigma_1 = \sum_{i=1}^{N_1-1} \bar{F}(u) S_i^{(K)}$ and $\Sigma_2 = \Sigma_1 + N_1 - 1 + N_C$. In practice, one must replace the unknown function F by the empirical c.d.f. \hat{F}_n and $u = u_n$ by $\hat{F}_n^{-1}(q)$, for some value $q = q_n$ near 1. Note that the asymptotic behavior of the estimator has only been derived under additional (unrealistic) assumptions such as knowledge of the c.d.f. F and independence of the inter-exceedance times. Finally, it is important to note that we used a minor modification of the above estimator throughout our Monte Carlo experiments. The modification aims at a proper handling of the possibly censored inter-exceedance times at the start and the end of the observation period, see Section A.2 in the [Online Appendix](#) for details.

2.2.3 A block-based maximum likelihood estimator

A sliding block version of a maximum likelihood estimator for the extremal index has been proposed and theoretically analyzed in [Northrop \(2015\)](#) and [Berghaus and Bücher \(2018\)](#), respectively. Unlike other blocks estimators for the extremal index, it is only depending on one parameter to be chosen by the statistician, namely a block length parameter $b = b_n$. The estimator has a simple closed-form expression and is defined as follows: first, given a block length b , let $M_t^{\text{sl}} = \max\{e_t, \dots, e_{t+b-1}\}$ and $Z_t^{\text{sl}} = b\{1 - F(M_t^{\text{sl}})\}$, where F denotes the c.d.f. of e_1 and where $t = 1, \dots, n - b + 1$. It can be shown that the transformed block maxima Z_t^{sl} are asymptotically b -dependent and exponentially distributed with mean θ^{-1} . Hence, after replacing F by its empirical counterpart \hat{F}_n , the reciprocal of the sample mean of $\hat{Z}_t^{\text{sl}} = b\{1 - \hat{F}_n(M_t^{\text{sl}})\}$ can be used to estimate the extremal index⁴:

$$\hat{\theta}_n^B = \hat{\theta}_n^B(b) = \left(\frac{1}{n - b + 1} \sum_{t=1}^{n-b+1} \hat{Z}_t^{\text{sl}} \right)^{-1}. \quad (2.7)$$

Under regularity conditions on the time series and if $b = b_n \rightarrow \infty$ with $b = o(n)$, it follows from theorem 3.1 in [Berghaus and Bücher \(2018\)](#) that

$$\sqrt{n/b}(\hat{\theta}_n^B - \theta) \rightarrow N(0, \theta^4 \sigma_{\text{sl}}^2),$$

where θ denotes the true extremal index and where $\sigma_{\text{sl}}^2 = \sigma_{\text{sl}}^2(\theta) > 0$ denotes the asymptotic variance of the sliding blocks estimator. It is worthwhile to mention that $\sigma_{\text{sl}}^2 = 0.2726$ in

4 [Berghaus and Bücher \(2018\)](#) propose an additional bias correction, which we do not describe here in detail, but which we employ throughout the simulation studies and the empirical applications.

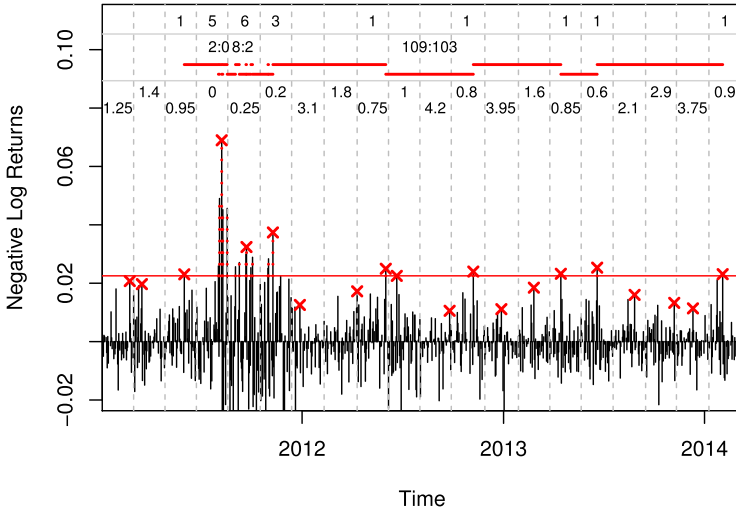


Figure 1. This plot illustrates the three extremal index estimators described in Section 2.2. The data set consists of about 800 daily returns on the S&P 500 index from January 3, 2011 till March 10, 2014. The first line at the top documents cluster sizes, the second line shows three examples for durations and six-gaps, and the last line reports transformed block maxima.

case the extremal index is equal to 1, that is, the limiting distribution is pivotal; see example 3.3 in [Berghaus and Bücher \(2018\)](#).

2.2.4 An example

In [Figure 1](#), we illustrate the classical blocks estimator from Section 2.2.1, the K -gap estimator from Section 2.2.2, and the disjoint variant of the block-based Maximum Likelihood estimator from Section 2.2.3 (obtained by using $\hat{Z}_t^{dj} = b\{1 - \hat{F}_n(M_t^{dj})\}$ with $M_t^{dj} = \max\{e_{bt-1+1}, \dots, e_{bt}\}$ for $t = 1, \dots, n/b$).

The data consist of about three years of negative daily log returns on the S&P 500 index. The solid red horizontal line corresponds to the *ex post* 97.5% empirical quantile of the negative return data, that is, $u = 0.0225$. Hence, we have exactly twenty values above this threshold. All exceedances of the threshold are labeled with a vertical dotted red line. The gray dashed lines mark the edges of the disjoint blocks. We choose a block length of $b = 40$ returns, resulting in $n/b = 20$ disjoint blocks.

The first line at the top of the [Figure 1](#) shows the empirical cluster sizes used for the disjoint classical blocks estimator from Section 2.2.1. The inverse of the average cluster size provides the classical blocks estimator for the extremal index, with a value of $\hat{\theta}^{CB} = 0.45$ for the particular example.

The second line corresponds to the K -gaps estimator from Section 2.2.2 with $K = 6$, which is depending on the threshold u and the gap parameter K , but not the block size b . The partly displaced horizontal red lines serve as an illustration for the durations between exceedances. Three numerical examples are provided above those lines. For instance, 109:103 means that the inter-exceedance time was 109 days. This value, truncated with $K = 6$, leads to a K -gap of 103. Since this inter-exceedance time is quite high, the truncating

alters little. However, in the first example, the duration is 2, resulting in a K -gap of 0. The final estimated value is $\hat{\theta}_n^G = 0.51$.

The blocks estimator $\hat{\theta}_n^B$ from Section 2.2.3 is only depending on the block size b and is based on computing maxima within each block. In particular, such a block maxima (red crosses) can also be below the threshold, as, for example, for blocks 1–2 in the picture. The block maxima are then transformed to the pseudo-observations \hat{Z}_t^{dj} , which are reported in the third line of the plot. Here, the inverse of the mean yields an estimate of $\hat{\theta}_n^B = 0.62$.

In this example, all estimates are quite similar and show that the negative S&P 500 returns exhibit a large degree of extremal dependence in their right tail.⁵ However, it is well known that such estimates can deviate largely depending on the estimator and parameter choice.⁶

2.3 The Backtesting Procedure for VaR

The backtesting procedure we propose is as follows: first, given a sequence of VaR forecasts $\hat{\text{VaR}}_p^{(t)}$ and observed returns r_t , for $t = 1, \dots, n$, calculate (e_t) as defined in Equation (2.1). Second, calculate $\hat{\theta}_n = \hat{\theta}_n(e_1, \dots, e_n)$ with $\hat{\theta}_n$ denoting any of the extremal index estimators from Section 2.2. Finally, reject the VaR-forecasts if $\hat{\theta}_n$ is significantly smaller than 1. Regarding the extremal index estimators, we only consider the sliding blocks estimator from Section 2.2.3 and the K -gap estimator from Section 2.2.2; the resulting tests will subsequently be denoted by $\Theta_{\text{noc}}^B = \Theta_{\text{noc}}^B(b)$ and $\Theta_{\text{noc}}^G = \Theta_{\text{noc}}^G(u, K)$, respectively.

2.3.1 Block maxima-based test

Regarding test Θ_{noc}^B , we first need to choose a block length parameter b . A preliminary Monte Carlo simulation study to compare several values of b , details can be found in Table A.3 of the Online Appendix; guides us to choose $b = 40$ across all further analyses. Although more suitable choices may be possible depending on the data generating process (DGP), we set a general data-dependent strategy for the choice of b aside, thus possibly reducing power in some cases.

Critical values for test Θ_{noc}^B could, in principle, be calculated based on the normal approximation described in Section 2.2.3: if the extremal index is 1, then $\hat{\theta}_b - 1$ is approximately centered normal with variance $0.2726 \cdot b/n$, no matter the stationary distribution F_e or the serial dependence of the time series outside the upper tail. However, the fact that the limiting distribution is pivotal also allows to approximate it by simulating from an arbitrary model for which the extremal index is 1. We hence opt for calculating critical values by first simulating $\tilde{e}_1, \dots, \tilde{e}_n$ from the model

5 This implies extremal dependence in the left tail of the S&P 500 returns.

6 See, for example, Tables 8.1.8 and 8.1.9 in Embrechts, Klüppelberg, and Mikosch (1997).

$$\tilde{e}_t = \frac{\tilde{r}_t}{-\widehat{\text{VaR}}_p^{(t)}}, \quad \tilde{r}_t \stackrel{\text{i.i.d.}}{\sim} N(0, 1), \quad \widehat{\text{VaR}}_p^{(t)} = -\Phi^{-1}(p), \tag{2.8}$$

where Φ denotes the c.d.f. of the standard normal distribution, and by then calculating $\hat{\theta}_n^B = \hat{\theta}_n^B(\tilde{e}_1, \dots, \tilde{e}_n)$ with the same block length parameter as chosen above, that is, $b = 40$. Note that such a simulation-based approach is also common for other classical backtesting procedures where asymptotic distributions are known.

2.3.2 K-gap test

Let us next describe details on test Θ_{noc}^G , which depends on the choice of both K and $u = u_n$. Regarding the choice of u , we simply set $u = 1$, which essentially means that we leave the extreme value context and are back to the 0-1-violation sequence from Section 1.1 (note that $e_t > 1$ if and only if $I_t = 1$). In particular, this viewpoint suggests to obtain critical values of the test simply by generating i.i.d. Bernoulli(p)-sequences \tilde{I}_t , and to calculate $\hat{\theta}_n^G = \hat{\theta}_n^G(K)$ by considering each time points where $\tilde{I}_t = 1$ as an exceedance (note that $\hat{\theta}_n^G$ only depends on those time points). Regarding the choice of the K -gap parameter, a further preliminary simulation study (details are presented in Table A.2 in the Online Appendix) prompts us to choose $K = 6$ for all further analyses. Since the K -gap test can be implemented with and without taking care of censored durations at the start and the end of the observation period, we distinguish three versions of this test via $\Theta_{\text{noc,u}}^G(K)$, $\Theta_{\text{noc,c,n}}^G(K)$, and $\Theta_{\text{noc,c,l}}^G(K)$. Details on their (minor) differences can be found in Section A.2 of the Online Appendix. Throughout this article, the two above-described simulation-based approaches, as well as all other similar approaches, are based on $N = 10,000$ replications, and corresponding p -values are computed as described in the Online Appendix, see also Dufour (2006).

2.3.3 Relation to the independence hypothesis and tests

By definition, the no cluster property, noc, in Equation (2.2) is quite different from the classical independence hypothesis, ind, in Equation (1.4), and a comparison between the hypotheses and tests is of interest.

As argued in Section 2.1, at least for (mean)-scale models, the use of correct forecasts implies that both ind and noc are met, both of which are properties related to the serial dependence of a respective forecast adjusted time series. This prompts us to regard corresponding tests for those hypotheses as valid backtests for the forecasts, despite the difference between the hypotheses.

When comparing respective tests, it is important to notice that the tests based on the extremal index follow quite different estimation approaches, with Θ_{noc}^G being based on inter-exceedance times and Θ_{noc}^B being based on block maxima. As described in the previous paragraph, Θ_{noc}^G in fact only depends on the 0-1-violation sequence and is hence closely related to classical independence backtests and also to the independence hypothesis, ind, in Equation (1.4).⁷

7 A more appropriate notation would hence be Θ_{ind}^G instead of Θ_{noc}^G .

Table 1. Rejection rates for several DGPs, backtesting samples sizes, VaR levels, significance levels, and backtesting procedures

λ	n	Significance level: 5%						Significance level: 10%					
		LR _{ind} ^{Mar}	LR _{ind} ^{Geo}	GMM _{ind,c} ^(VD)	MCS _{ind}	$\Theta_{noc,c,l}^G(6)$	$\Theta_{noc}^B(40)$	LR _{ind} ^{Mar}	LR _{ind} ^{Geo}	GMM _{ind,c} ^(VD)	MCS _{ind}	$\Theta_{noc,c,l}^G(6)$	$\Theta_{noc}^B(40)$
Panel A: 5% VaR													
0.8706	250	0.146	0.145	0.228	0.228	0.282	0.279	0.206	0.235	0.309	0.349	0.388	0.398
	1000	0.228	0.594	0.581	0.523	0.628	0.747	0.307	0.712	0.705	0.682	0.733	0.849
	2500	0.537	0.947	0.908	0.832	0.893	0.977	0.652	0.969	0.952	0.920	0.934	0.990
0.9828	250	0.063	0.058	0.068	0.084	0.089	0.103	0.124	0.106	0.131	0.166	0.154	0.169
	1000	0.052	0.083	0.141	0.172	0.135	0.239	0.097	0.154	0.212	0.277	0.225	0.360
	2500	0.070	0.204	0.275	0.335	0.246	0.465	0.122	0.288	0.368	0.466	0.352	0.595
0.9914	250	0.056	0.047	0.051	0.052	0.061	0.065	0.106	0.096	0.097	0.112	0.113	0.119
	1000	0.046	0.054	0.081	0.099	0.093	0.126	0.088	0.121	0.142	0.170	0.150	0.214
	2500	0.056	0.082	0.132	0.167	0.086	0.221	0.114	0.147	0.198	0.259	0.151	0.338
1 (H_0)	250	0.052	0.048	0.051	0.047	0.051	0.052	0.103	0.102	0.099	0.096	0.104	0.096
	1000	0.047	0.048	0.060	0.056	0.048	0.050	0.088	0.099	0.105	0.106	0.107	0.101
	2500	0.052	0.064	0.049	0.051	0.045	0.045	0.105	0.102	0.103	0.102	0.090	0.093
Panel B: 1% VaR													
0.8706	250	0.189	0.136	0.121	0.117	0.248	0.307	0.245	0.223	0.177	0.216	0.387	0.439
	1000	0.220	0.188	0.168	0.155	0.377	0.729	0.346	0.279	0.220	0.268	0.477	0.838
	2500	0.382	0.456	0.297	0.244	0.636	0.973	0.483	0.558	0.371	0.392	0.730	0.989
0.9828	250	0.115	0.074	0.081	0.063	0.070	0.088	0.189	0.136	0.152	0.126	0.151	0.161
	1000	0.091	0.052	0.100	0.120	0.096	0.212	0.136	0.101	0.141	0.211	0.169	0.350

(continued)

Table 1. (continued)

λ	n	Significance level: 5%						Significance level: 10%					
		LR _{ind} ^{Mar}	LR _{ind,c} ^{Geo}	GMM _{ind,c} ^(VD)	MCS _{ind}	$\Theta_{noc,c,l}^C(6)$	$\Theta_{noc}^B(40)$	LR _{ind} ^{Mar}	LR _{ind,c} ^{Geo}	GMM _{ind,c} ^(VD)	MCS _{ind}	$\Theta_{noc,c,l}^C(6)$	$\Theta_{noc}^B(40)$
0.9914	2500	0.098	0.115	0.178	0.190	0.134	0.454	0.162	0.194	0.242	0.295	0.214	0.578
	250	0.092	0.073	0.080	0.055	0.069	0.071	0.169	0.132	0.142	0.109	0.145	0.124
	1000	0.048	0.048	0.084	0.091	0.053	0.130	0.121	0.096	0.123	0.159	0.115	0.222
	2500	0.081	0.070	0.126	0.142	0.093	0.207	0.135	0.125	0.176	0.234	0.159	0.331
	250	0.063	0.067	0.068	0.044	0.072	0.052	0.124	0.139	0.130	0.092	0.137	0.097
1 (H_0)	10001	0.040	0.052	0.046	0.044	0.048	0.043	0.090	0.108	0.092	0.094	0.093	0.094
	2500	0.056	0.057	0.056	0.050	0.045	0.054	0.107	0.108	0.105	0.097	0.093	0.102

Notes: The setting is borrowed from table 4 in Ziggel et al. (2014). The DGPs produce clustered violations by the usage of a constant VaR forecast obtained as the unconditional empirical VaR of a simulated path of length 100,000. The additional DGP with $\lambda = 1$ allows to check for the sizes of the tests. Furthermore, the DGPs are simulated subject to at least two violations. The rejection rates are based on 5000 Monte-Carlo replications. The tests use Monte-Carlo p -values with simulated distributions of the statistics. Here, 10,000 replications subject to H_0 are used.

Table 2. Rejection rates of HS VaR forecasts with two estimation window sizes $T_e \in \{250, 500\}$ across several backtesting samples sizes, VaR levels, and backtesting procedures

T_e	n	Significance level: 5%						Significance level: 10%					
		LR _{ind} ^{Mar}	LR _{ind,c} ^{Geo}	GMM _{ind,c} ^(VD)	MCS _{ind}	$\Theta_{noc,c,l}^G(6)$	$\Theta_{noc}^B(40)$	LR _{ind} ^{Mar}	LR _{ind,c} ^{Geo}	GMM _{ind,c} ^(VD)	MCS _{ind}	$\Theta_{noc,c,l}^G(6)$	$\Theta_{noc}^B(40)$
Panel A: 5% VaR													
250	250	0.204	0.334	0.440	0.481	0.452	0.591	0.270	0.433	0.512	0.597	0.539	0.689
	500	0.287	0.701	0.740	0.758	0.700	0.890	0.351	0.777	0.810	0.839	0.776	0.928
	750	0.331	0.844	0.852	0.860	0.814	0.963	0.404	0.889	0.902	0.922	0.879	0.980
	1000	0.402	0.924	0.924	0.922	0.901	0.988	0.490	0.955	0.955	0.964	0.937	0.995
	1500	0.591	0.986	0.982	0.985	0.972	1.000	0.698	0.992	0.991	0.994	0.986	1.000
500	250	0.196	0.337	0.414	0.459	0.446	0.575	0.253	0.422	0.489	0.568	0.525	0.670
	500	0.282	0.679	0.726	0.740	0.680	0.870	0.347	0.754	0.791	0.821	0.750	0.913
	750	0.387	0.877	0.880	0.875	0.832	0.967	0.448	0.915	0.922	0.931	0.879	0.983
	1000	0.442	0.944	0.943	0.943	0.902	0.993	0.523	0.972	0.967	0.973	0.937	0.997
	1500	0.637	0.988	0.983	0.983	0.971	0.999	0.736	0.994	0.993	0.992	0.984	0.999
Panel B: 1% VaR													
250	250	0.219	0.156	0.186	0.198	0.243	0.617	0.262	0.226	0.218	0.283	0.351	0.703
	500	0.250	0.277	0.341	0.367	0.395	0.899	0.339	0.379	0.389	0.489	0.510	0.936
	750	0.251	0.398	0.436	0.415	0.513	0.973	0.391	0.501	0.489	0.559	0.612	0.984
	1000	0.252	0.515	0.515	0.463	0.584	0.992	0.452	0.612	0.577	0.607	0.689	0.996
	1500	0.319	0.696	0.648	0.563	0.722	1.000	0.461	0.773	0.716	0.702	0.812	1.000
500	250	0.202	0.153	0.191	0.209	0.234	0.573	0.260	0.228	0.222	0.285	0.321	0.677
	500	0.254	0.325	0.372	0.401	0.419	0.885	0.353	0.407	0.408	0.496	0.503	0.926
	750	0.304	0.504	0.518	0.525	0.575	0.970	0.448	0.584	0.568	0.631	0.654	0.985
	1000	0.328	0.642	0.650	0.623	0.661	0.992	0.510	0.723	0.692	0.734	0.752	0.996
	1500	0.431	0.815	0.773	0.734	0.793	0.999	0.544	0.864	0.822	0.827	0.853	0.999

Notes: The setting is borrowed from table 3 in Candelson et al. (2011). Compared to this, we use the independence version of all tests and add the MCS independence test. The rejection rates are based on 5000 Monte-Carlo replications. The tests use Monte-Carlo p -values with simulated distributions of the statistics. Here, 10,000 replications subject to H_0 are used.

Table 3. Rejection rates for several forecasting models using stochastic volatility are shown

FC	n	Significance level: 5%					Significance level: 10%						
		LR ^{Mar} _{Ind}	LR ^{Geo} _{Ind,c}	GMM ^(VD) _{Ind,c}	MCS _{Ind}	$\Theta^{G_{\text{noc},c-1}}(6)$	$\Theta^B_{\text{noc}}(40)$	LR ^{Mar} _{Ind}	LR ^{Geo} _{Ind,c}	GMM ^(VD) _{Ind,c}	MCS _{Ind}	$\Theta^{G_{\text{noc},c-1}}(6)$	$\Theta^B_{\text{noc}}(40)$
Panel A: 5% VaR True VaRs	250	0.053	0.056	0.067	0.057	0.056	0.062	0.112	0.103	0.112	0.110	0.107	0.111
	1000	0.045	0.041	0.050	0.043	0.048	0.051	0.092	0.094	0.091	0.089	0.107	0.108
	2500	0.052	0.051	0.057	0.046	0.042	0.062	0.093	0.098	0.104	0.104	0.093	0.115
	250	0.046	0.053	0.060	0.052	0.063	0.055	0.093	0.110	0.115	0.098	0.110	0.097
	1000	0.048	0.051	0.058	0.060	0.084	0.072	0.100	0.113	0.112	0.116	0.144	0.138
GARCH(1,1)	2500	0.079	0.103	0.094	0.078	0.127	0.105	0.133	0.154	0.149	0.140	0.193	0.168
	250	0.118	0.076	0.118	0.126	0.182	0.143	0.173	0.129	0.191	0.205	0.284	0.223
	1000	0.128	0.265	0.266	0.238	0.443	0.313	0.191	0.387	0.389	0.376	0.567	0.446
	2500	0.319	0.562	0.469	0.411	0.681	0.571	0.432	0.652	0.623	0.553	0.771	0.669
	250	0.100	0.192	0.306	0.315	0.404	0.387	0.149	0.282	0.390	0.430	0.512	0.516
ARCH(1)	1000	0.095	0.733	0.761	0.716	0.801	0.889	0.146	0.817	0.846	0.825	0.861	0.930
	2500	0.200	0.978	0.973	0.948	0.959	0.993	0.289	0.991	0.989	0.981	0.978	0.998
	250	0.050	0.046	0.044	0.038	0.046	0.049	0.094	0.097	0.092	0.083	0.092	0.109
	1000	0.049	0.043	0.050	0.045	0.043	0.058	0.101	0.092	0.098	0.098	0.085	0.120
	2500	0.055	0.051	0.036	0.047	0.047	0.065	0.110	0.106	0.090	0.096	0.093	0.136
GJR-GARCH(1,1)	250	0.060	0.062	0.056	0.038	0.051	0.060	0.108	0.101	0.116	0.082	0.104	0.112
	1000	0.031	0.050	0.054	0.046	0.054	0.076	0.113	0.106	0.104	0.110	0.100	0.150
	2500	0.069	0.068	0.059	0.057	0.054	0.133	0.122	0.116	0.111	0.102	0.112	0.200
	250	0.100	0.068	0.061	0.067	0.102	0.142	0.160	0.120	0.120	0.127	0.201	0.229
	1000	0.107	0.085	0.099	0.099	0.184	0.346	0.255	0.155	0.151	0.174	0.290	0.482
ARCH(1)	2500	0.218	0.158	0.127	0.116	0.310	0.605	0.296	0.240	0.190	0.219	0.424	0.705
	250	0.118	0.122	0.138	0.142	0.225	0.379	0.177	0.178	0.168	0.210	0.298	0.487
	1000	0.053	0.400	0.352	0.340	0.573	0.892	0.179	0.501	0.405	0.457	0.654	0.937
	2500	0.121	0.744	0.615	0.536	0.827	0.999	0.182	0.804	0.673	0.664	0.871	0.999

Notes: We use simulated data from an estimated GJR-GARCH(1,1) model with Student- t innovations, see Section A.4 of the Online Appendix. For each iteration, the corresponding model is fitted using the first 1000 simulated returns, remaining returns are used for forecasting and backtesting. The rejection rates are based on 5000 Monte-Carlo replications. The tests use Monte-Carlo p -values with simulated distributions of the statistics. Here, 10,000 replications subject to H_0 are used.

On the other hand, the block maxima-based test $\Theta_{\text{noc}}^{\text{B}}$ is really a test for noc in Equation (2.2), and as such quite different to classical backtests. Two important drawbacks are to be kept in mind when applying $\Theta_{\text{noc}}^{\text{B}}$: first, since an extremal index of one does not imply independence, the test cannot have power against certain incorrect forecasts. Second, our motivation from Section 2.1 only concerns (mean-)scale models, whence the test may in fact wrongly reject the null hypothesis for certain alternative models not of the mean-scale type. On the other hand, an application of $\Theta_{\text{noc}}^{\text{B}}$ also comes along with the advantage that it may use the data in a more informative way, given that the test is depending on (e_t) instead of just the violation sequence (I_t) . This may result in more power against certain types of forecasts that are not able to capture the tail in such a way that noc holds.

2.4 Extensions to More General Risk Measures Including ES

Backtesting the ES recently received increased attention due to its upcoming implementation as a standard risk measure for regulatory purposes in banking (BCBS, 2016). Most available backtests of ES focus on unconditional coverage, see Kerkhof and Melenberg (2004), Wong (2008), Costanzino and Curran (2015), and Kratz, Lok, and McNeil (2018). Only recently, Du and Escanciano (2017) propose to additionally use a Box–Pierce test to test for autocorrelation in a certain sequence of cumulative violations. This test can hence be regarded as the first ES backtest for independence (rather: serial uncorrelation). In this section, we extend the basic idea from Section 2.3 to obtain a further backtest for ES that is particularly sensitive to certain deviations from independence in the tails.

Recall that the main idea of the VaR method from Section 2.3 consists of checking whether the relative excess returns in Equation (2.1) do not show any sign of extremal clustering. The sensibility of such an approach was explained in Section 2.1 for mean-scale models, and the arguments can, in fact, be generalized to any risk measure which is translation invariant and positively homogeneous. Indeed, recall that a risk measure $\rho : M \rightarrow \mathbb{R}$, M a set of random variables, satisfies translation invariance if, for all $R \in M$ and every $c \in \mathbb{R}$, we have $\rho(R + c) = \rho(R) - c$ (the change of the sign stems from interpreting R as a return and not a loss). Positive homogeneity is satisfied if $\rho(c; R) = c; \rho(R)$ for all $R \in M$ and $c > 0$ (McNeil, Frey, and Embrechts, 2005). By the same arguments as in Section 2.1, it is sensible to backtest a sequence of forecasts $\hat{\rho}_t$ by checking whether the sequence

$$e_t = -\frac{r_t}{\hat{\rho}_t} \quad (2.9)$$

does not show any sign of extremal clustering. Indeed, for location-scale models as defined in Equation (2.3) and for $\hat{\rho}_t = \rho_t = \rho(r_t | \mathcal{F}_{t-1})$, we obtain that

$$e_t = -\frac{r_t}{\rho_t} = \frac{\mu_t + \sigma_t z_t}{\mu_t - \sigma_t \rho(z_t)},$$

which simplifies to Equation (2.5) if we use $\rho(z_t) = -F_z^{-1}(p)$, that is, VaR. As a consequence, it is sensible to apply the methodology described in Section 2.3 to the sequence $(e_t)_t$ defined in Equation (2.9), for any translation invariant and homogeneous risk measure. We do not pursue this any further in this document.

Table 4. Rejection rates for VaRs forecasts computed with an estimated correct model are shown

n_{Est}	n	Significance level: 5%						Significance level: 10%					
		LR _{ind,c} ^{Mar}	LR _{ind,c} ^{Geo}	GMM _{ind,c} ^(VD)	MCS _{ind}	$\Theta_{noc,c,l}^G(6)$	$\Theta_{noc}^B(40)$	LR _{ind,c} ^{Mar}	LR _{ind,c} ^{Geo}	GMM _{ind,c} ^(VD)	MCS _{ind}	$\Theta_{noc,c,l}^G(6)$	$\Theta_{noc}^B(40)$
Panel A: 5% VaR													
∞	250	0.053	0.056	0.067	0.057	0.056	0.062	0.112	0.103	0.112	0.110	0.107	0.111
	1000	0.053	0.051	0.054	0.054	0.055	0.054	0.096	0.105	0.108	0.103	0.106	0.102
5000	2500	0.048	0.050	0.053	0.052	0.043	0.056	0.095	0.102	0.100	0.098	0.094	0.106
	250	0.052	0.050	0.062	0.049	0.061	0.058	0.106	0.103	0.111	0.098	0.103	0.107
1000	1000	0.056	0.050	0.054	0.052	0.063	0.052	0.095	0.101	0.101	0.100	0.119	0.108
	2500	0.052	0.059	0.057	0.061	0.060	0.064	0.106	0.117	0.105	0.112	0.109	0.118
500	250	0.057	0.057	0.058	0.058	0.063	0.056	0.113	0.110	0.120	0.109	0.121	0.115
	1000	0.060	0.070	0.070	0.073	0.080	0.075	0.110	0.133	0.127	0.119	0.137	0.137
5000	2500	0.081	0.096	0.082	0.077	0.114	0.099	0.136	0.157	0.150	0.140	0.174	0.160
	250	0.070	0.064	0.071	0.062	0.082	0.068	0.131	0.116	0.139	0.119	0.134	0.121
1000	1000	0.076	0.099	0.093	0.087	0.123	0.108	0.124	0.176	0.158	0.148	0.193	0.173
	2500	0.101	0.143	0.116	0.101	0.166	0.156	0.172	0.208	0.191	0.170	0.232	0.223
Panel B: 1% VaR													
∞	250	0.050	0.046	0.044	0.038	0.046	0.049	0.094	0.097	0.092	0.083	0.092	0.109
	1000	0.043	0.047	0.057	0.049	0.046	0.059	0.095	0.103	0.105	0.093	0.095	0.112
5000	2500	0.058	0.055	0.055	0.050	0.042	0.075	0.104	0.102	0.099	0.100	0.089	0.142
	250	0.046	0.056	0.053	0.042	0.039	0.054	0.091	0.099	0.103	0.087	0.086	0.104
1000	1000	0.046	0.056	0.050	0.041	0.047	0.052	0.109	0.118	0.095	0.094	0.098	0.115
	2500	0.049	0.049	0.049	0.047	0.040	0.074	0.093	0.097	0.096	0.094	0.089	0.144
1000	250	0.046	0.053	0.059	0.049	0.042	0.064	0.100	0.101	0.109	0.095	0.088	0.109
	1000	0.047	0.050	0.063	0.054	0.054	0.087	0.127	0.112	0.112	0.106	0.107	0.158
500	2500	0.057	0.063	0.058	0.051	0.066	0.130	0.106	0.125	0.109	0.108	0.129	0.207
	250	0.083	0.060	0.074	0.054	0.065	0.085	0.147	0.115	0.123	0.112	0.130	0.158
1000	1000	0.045	0.057	0.082	0.070	0.070	0.117	0.146	0.109	0.135	0.124	0.129	0.193
	2500	0.083	0.080	0.081	0.067	0.095	0.179	0.143	0.134	0.132	0.126	0.162	0.259

Notes: We use simulated data from an estimated GJR-GARCH(1,1) model with student- t innovations, see Section A.4 of the Online Appendix. For each iteration, the corresponding model is fitted using the first n_{Est} simulated returns, remaining returns are used for forecasting and backtesting. The rejection rates are based on 5000 Monte-Carlo replications. The tests use Monte-Carlo p -values with simulated distributions of the statistics. Here, 10,000 replications subject to H_0 are used.

2.5 An Extension to Distributional Backtests

The general idea from Section 2.3 may also be applied to backtesting forecasts of the entire conditional distribution (or density), see also Berkowitz (2001). More precisely, suppose that \hat{F}_t is a distributional forecast of the conditional c.d.f. of r_t given \mathcal{F}_{t-1} , the latter being denoted by F_t . The role of the VaR-adjusted return series (e_t) may then be played by the probability integral transform sequence $u_t = 1 - \hat{F}_t(r_t)$, $t = 1, \dots, n$. In case $\hat{F}_t = F_t$, the sequence is known to constitute an i.i.d. sequence of uniformly distributed random variables on the interval $[0, 1]$, see Rosenblatt (1952). As in the previous section, a distributional backtest that is particularly sensitive to deviations from independence in the upper right tail of u_t is obtained by comparing the estimated extremal index of u_1, \dots, u_n with 1.

3 Size and Power Analysis

In this section, we compare our new approach to several classical independence backtesting procedures in terms of their empirical size and power properties by means of a large-scale Monte Carlo simulation study.

The following competitors to our tests are considered: the Markov chain-based likelihood-ratio test $LR_{\text{ind}}^{\text{Mar}}$ from Christoffersen (1998), the Geometric likelihood-ratio test $LR_{\text{ind,c}}^{\text{Geo}}$ from Berkowitz, Christoffersen, and Pelletier (2011), the test $GMM_{\text{ind,c}}^{(\text{VD})}$ based on the generalized method of moments from Candelon et al. (2011),⁸ and the test based on squared durations MCS_{ind} by Ziggel et al. (2014). More details on these tests can be found in Section A.1 of the Online Appendix. Regarding our tests and as described in Section 2.3, the parameter b is set to $b = 40$ for test $\Theta_{\text{noc}}^{\text{B}}$ and to $K = 6$ for test $\Theta_{\text{noc,c,l}}^{\text{G}}$.

Throughout, the general procedure to obtain empirical rejection rates is as follows: for each combination of DGP, VaR level, and sample size, we generate 5000 random samples. We then perform the mentioned tests, based on p -values that are computed as described in Section 2.3 and in Sections A.1 and A.3 in the Online Appendix.

3.1 Power Properties When True Unconditional VaRs Are Used

The first simulation experiment is guided by table 4 in Ziggel et al. (2014), the purpose being to compare (independence) backtests in situations where clustered violations are likely due to the use of unconditional instead of conditional VaRs. The DGP is as follows:

$$r_t = \sigma_t z_t, \quad t = 1, \dots, n,$$

with $z_t \stackrel{\text{i.i.d.}}{\sim} N(0, 1)$, $\sigma_1 = 1$ and

$$\sigma_t^2 = \lambda \sigma_{t-1}^2 + (1 - \lambda) z_{t-1}^2, \quad t = 2, \dots, n.$$

As in Ziggel et al. (2014), the parameter λ is chosen from the set $\{0.8706, 0.9829, 0.9914, 1\}$ (case $\lambda = 1$ will correspond to the null hypothesis) and the sample size n is chosen from the set $\{250, 1000, 2500\}$. Note that results for sample sizes as small as 250 should be regarded with a little caution. For the block maxima-based test

8 With (VD), we indicate that we choose the parameter k for the GMM test to be 3 for the 1% VaR level and 5 for the 5% VaR level.

Table 5. This table presents backtesting results for several one-step-ahead forecasts adopting a rolling window scheme assessed by different backtests

FC	N_t	p -value	$LR_{\text{ind}}^{\text{Mar}}$	$LR_{\text{ind},c}^{\text{Geo}}$	$GMM_{\text{ind},c}^{(\text{VD})}$	MCS_{ind}	$\Theta_{\text{noc},c,1}^G(6)$	$\Theta_{\text{noc}}^B(40)$
Turbulent period from January 15, 2008 till December 31, 2011 (1000 Obs)								
Panel A: 1% VaR								
HS (250)	22	0.0010***	0.0931	0.0000***	0.0000***	0.0000***	0.0000***	0.0000***
HS (500)	26	0.0000***	0.0729	0.0000***	0.0000***	0.0000***	0.0000***	0.0000***
Panel B: 0.05% VaR								
skew- t (250)	2	0.1103	0.0516	0.0022**	0.3884	0.2532	0.1994	0.0000***
skew- t (500)	0	0.3175	0.6975	0.3706	-	-	-	0.0000***
Panel C: 1% VaR and normal conditional distribution								
GJR-GARCH(1,1)	31	0.0000***	0.0728	0.1657	0.0510	0.9629	0.1906	0.7029
GARCH(1,1)	30	0.0000***	0.0729	0.3656	0.2057	0.2526	0.1013	0.2034
ARCH(1)	45	0.0000***	0.1648	0.0002***	0.0000***	0.0001***	0.0002***	0.0000***
Panel D: 1% VaR and t conditional distribution								
GJR-GARCH(1,1)	18	0.0221*	0.1043	0.9076	0.8708	0.4345	0.4328	0.6975
GARCH(1,1)	20	0.0051**	0.0956	0.0498*	0.0991	0.0673	0.0203*	0.2573
ARCH(1)	31	0.0000***	0.9994	0.0000***	0.0000***	0.0000***	0.0000***	0.0000***
Calm Period from January 1, 2012 till December 22, 2015 (1000 Obs)								
Panel A: 1% VaR								
HS (250)	1	0.3604	0.0034***	0.0811	0.6869	0.3444	0.0033**	0.0003***
HS (500)	8	0.5121	0.0002***	0.0003***	0.0025**	0.0050**	0.0001**	0.0002***
Panel B: 0.05% VaR								
skew- t (250)	0	0.3175	0.6975	0.3706	-	-	-	0.0000***
skew- t (500)	1	0.5336	0.2614	-	0.0813	0.1698	0.1994	0.0000***
Panel C: 1% VaR and normal conditional distribution								
GJR-GARCH(1,1)	16	0.0788	0.1283	0.0154*	0.0224*	0.9888	0.7150	0.9411
GARCH(1,1)	22	0.0010***	0.0390*	0.2466	0.0554	0.9562	0.1509	0.7815

(continued)

Table 5. (continued)

FC	N_1	p -value	LP_{ind}^{Mar}	$LP_{ind,c}^{Geo}$	$GMM_{ind,c}^{(VD)}$	MCS _{ind}	$\Theta_{noc,c,1}^G(6)$	$\Theta_{noc}^B(40)$
ARCH(1)	46	0.0000***	0.0015**	0.0000***	0.0000***	0.0000***	0.0001***	0.0000***
Panel D: 1% VaR and t conditional distribution								
GJR-GARCH(1,1)	11	0.7520	0.3960	0.1028	0.1455	0.9014	0.7150	0.9529
GARCH(1,1)	13	0.3604	0.0677	0.1979	0.0788	0.9349	0.3716	0.7952
ARCH(1)	3	0.0091**	0.9937	0.0950	0.0289*	0.0015**	0.7150	0.0000***

Notes: In Panel A, 1% VaR forecasts are performed with an unconditional non-parametric method. In Panel B, 0.05% VaR forecasts are made using a skew- t distribution. In each case, the numbers in brackets report the size of the rolling window. GARCH model refits are done every five days, unconditional methods are refitted on a daily basis. Panel C belongs to 1% VaR forecasts using three different GARCH models with rolling window size 1000 and a conditional normal distribution. Panel D equals to Panels C but uses a conditional t -distribution instead. The out-of-sample periods are a troubled and a calm market period of 1000 returns each. Hence, in Panel A 10 violations, in Panels B 0.5 violations, and in Panels C again 10 violations have to be expected. Column FC on the left side reports FCs. The second column reports both the number of violations N_1 and the corresponding asymptotic p -value of the unconditional backtest by Christoffersen (1998). The remaining columns show results of several independence backtests. The numbers are Monte-Carlo p -values. Asterisks mark levels of significance: *** at 0.1%, ** at 1%, and * at 5%.

$\Theta_{\text{noc}}^{\text{B}}$, taking blocks of length $b = 40$ results in only six disjoint block maxima (and slightly more distinct values for the sliding block maxima), to which we then fit an exponential distribution. However, since we do not rely on asymptotics but rather on simulations to calculate critical values, we still think that an application to such small sample sizes is sensible. The classical tests suffer from similar drawbacks in the absence of violations. To make this issue less important and in line with Ziggel et al. (2014), the simulation study is performed conditional on the restriction of at least two violations per backtesting sample.⁹

Recall from Section 2.1 that the true conditional VaR of the above-described model is given by $\text{VaR}_p^{(t)} = -\sigma_t \Phi^{-1}(p)$ and that the use of $\widehat{\text{VaR}}_p^{(t)} = \text{VaR}_p^{(t)}$ in Formula (2.1) would result in an i.i.d. sequence of relative excess returns. Serial dependence (and in particular extremal clustering) is now introduced by instead setting $\widehat{\text{VaR}}_p^{(t)} = \widehat{\text{VaR}}_p$ (independent of t) to the empirical VaR computed from a preliminary simulation with 100,000 returns.

The estimated rejection rates are reported in Table 1 (an extended version can be found in Table A.5 in the Online Appendix, where we also consider the confidence level $\alpha = 1\%$ and additional modifications of the tests). All tests exhibit a reasonable approximation of the intended level (case $\lambda = 1$). Only for $n = 250$ and the 1% VaR level, the nominal significance levels are consistently slightly surpassed, which is possibly due to above-mentioned restriction on the number of violations. In terms of power, the proposed extremal index test $\Theta_{\text{noc}}^{\text{B}}$ typically yields the largest numbers, which on top are often much larger than for the classical competitors. In the few cases, a non-extremal index test yields larger power, the improvement over the extremal index versions is rather small.

With some exceptions (in particular test $\Theta_{\text{noc}}^{\text{B}}$), the rejection rates are higher in the 5% VaR Panel. A possible explanation is that a small number of violations cannot yield the same evidence for serial dependence like a large number of violations are capable of. For $\Theta_{\text{noc}}^{\text{B}}$ instead, the rejection rates change barely for different levels, a likely explanation being that the input data (relative excess returns) are approximately the same up to a scaling factor. Moreover, by construction, $\Theta_{\text{noc}}^{\text{B}}$ is also able to use information of events where violations did occur almost, see also Figure 1. This questions whether it is meaningful to assess the independence property of 1% VaR forecasts in small samples based solely on violation sequences.

3.2 How Often Can We Reject Historical Simulation?

The second simulation is inspired by Candelon et al. (2011) and Christoffersen and Pelletier (2004). Again, returns are simulated using a mean-scale model with $\mu_t \equiv 0$ and innovations $z_t \stackrel{\text{i.i.d.}}{\sim} \sqrt{\frac{d-2}{d}} \varepsilon_t$, where ε_t follows a Student's t -distribution with d degrees of freedom. The conditional variance involves an asymmetric leverage effect and is given by

$$\sigma_t^2 = \omega + \gamma \sigma_{t-1}^2 (z_{t-1} - \theta)^2 + \beta \sigma_{t-1}^2, \quad t \in \mathbb{N}_{\geq 2},$$

where $\gamma = 0.1, \theta = 0.5, \beta = 0.85, \omega = 3.9683 \cdot 10^{-6}$ and $d = 8$. We set $\sigma_1^2 = \omega$ and use a burn-in period of length $N_{\text{burn-in}} = 200$ before forecasting is started.

9 The probability that a generated sample of size n violates this condition is negligible in all cases except for the 1% VaR level and $n = 250$ case, where approximately 28.6% of randomly drawn samples exhibit at most one violation.

Time-varying forecasts are obtained by applying the popular and realistic VaR forecasting technique of (unconditional) “Historical Simulation” (HS): given an integer $T_e \in \{250, 500\}$, we estimate the conditional VaR at time t by the respective empirical quantile (multiplied with -1) of the T_e observations prior to time point t . The experiment is hence in contrast to the scenario from Section 3.1, where one fixed VaR forecast was used for all t . Still, since HS is not able to capture the dynamics of the time-varying volatility adequately either, the forecasting method (FC) should be rejected. We hence allow for an assessment of the methods’ power in a more realistic environment. Note that we have to simulate $N_{\text{burn-in}} + T_e + n$ returns in total per replication to obtain the results of the simulation experiment.

Results of the simulation experiment are reported in Table 2 (an extended version can be found in Table A.6 in the Online Appendix). Focusing only on the classical methods first, we find that typically one of $\text{LR}_{\text{ind},e}^{\text{Geo}}$, $\text{GMM}_{\text{ind}}^{(\text{VD})}$, or MCS_{ind} backtests yields the largest power. The 0-1-Extremal Index approach $\Theta_{\text{noc}}^{\text{G}}$ shows overall comparable rejection rates to these—sometimes the power is larger, sometimes smaller, and sometimes there is barely any difference. Finally, the second extremal index test $\Theta_{\text{ind}}^{\text{B}}$ is able to improve the power in every case under consideration, sometimes by a considerable amount.

3.3 Rejection Rates of (Misspecified) Stochastic Volatility Models

In this section, we study backtest rejection rates for forecasts based on estimated, but possibly misspecified stochastic volatility models. The underlying DGP is fixed as a certain GJR-GARCH(1,1) model with Student- t innovations and a non-zero mean parameter μ , with the model parameters being chosen as the estimated values obtained by fitting the model to daily S&P 500 log returns from January 1, 2012 to January 1, 2015 (754 observations), see Appendix A.4 in the Online Appendix for details. Note that all parameters of the model were found to be highly significant in the fit, including the mean parameter $\mu = 6.91 \times 10^{-4}$. Hence, in light of the discussion in Section 2.1, the present setting also serves as a robustness check of the extremal index tests against a non-zero mean.

For each Monte Carlo repetition, a time series of length $n + 1000$, with $n \in \{250, 1000, 2500\}$, is simulated from the above-described model. Three FCs are then investigated, based on either a GJR-GARCH(1,1), a GARCH(1,1), or an ARCH(1) model fit to the first 1000 observations of the time series, and a subsequent VaR forecast based on the respective estimated model and the realized returns up to time $t - 1$. Note that the three models are included in each other, and that the latter two models are, by construction, misspecified. We hence expect increasing rejection rates in this order.

The results are presented in Table 3 (an extended version can be found in Table A.7 in the Online Appendix). For comparison, the first FC corresponds to the usage of the true VaRs of the simulated data, which is not available in practice but for which the null hypothesis is met. The remaining methods correspond to the three mentioned forecasting models. The main findings are summarized in the next three paragraphs.

3.3.1 Size

The FC “True VaRs” serves as a size benchmark. Overall, all methods exhibit a reasonable approximation of the nominal size. Deviations may in most cases be explained by

simulation variance of the simulated distributions of the test statistics, as well as the Monte Carlo simulations itself. However, we also observe a larger deviation for $\Theta_{\text{noc}}^{\text{B}}$ at the 1% VaR level and a backtesting sample size of $n=2500$. For example, the rejection rate is 13.6% at the 10% significance level. A possible explanation is the non-zero mean in the DGP, an issue that is further investigated in Section A.6 in the [Online Appendix](#). This kind of oversizedness does not occur for the other extremal index test $\Theta_{\text{noc}}^{\text{G}}$.

3.3.2 Estimation risk

The forecast based on estimating the (true) GJR-GARCH(1,1) is slightly more likely to be rejected than the true VaRs. We further check the sensitivity of the backtests to estimation uncertainty in the next Section 3.4.

3.3.3 Rejection rates of misspecified models

As expected, ARCH(1) is most likely to be rejected, followed by the standard GARCH model. Interestingly, $\Theta_{\text{noc}}^{\text{G}}$ performs often better than $\Theta_{\text{noc}}^{\text{B}}$ in the 5% VaR Panel. For the 1% VaR Panel, the decrease in the number of violations leads to a better performance of $\Theta_{\text{noc}}^{\text{B}}$. In most cases, both tests are able to improve the power substantially compared to the classical competitors.

3.4 Estimation Risk

The results in [Table 3](#) have shown that estimating the correct model is not sufficient to get correct forecasts. Hence, an additional aspect of the forecasting task in general is the ability to estimate the potentially correct model sufficiently accurately: which sample size is necessary to get (almost) true forecasts if the correct model is used? To shed light on this issue, we report in [Table 4](#) (an extended version can be found in [Table A.8](#) in the [Online Appendix](#)) results of a similar task as in the previous section. We estimate the true model using varying lengths of sample sizes n_{Est} ranging from 500 to 5000 (recall that we used a fixed value of $n_{\text{Est}} = 1000$ in the previous section). The table reveals that, across all tests, the extremal index approaches are most likely to reject the estimated model. A large amount of data is hence needed for the rejection rates to approach the nominal significance level.

The previous findings suggest to adapt the backtesting approaches in a way that is able to explicitly take care of the estimation uncertainty involved in the estimation of a correct model. Such a modification has for instance been worked out in [Escanciano and Olmo \(2010\)](#), among others, for tests that are based on the violation sequence. Corresponding adaptations for the extremal index-based tests constitute an interesting research problem, which, however, is beyond the scope of this article.

4 Empirical Applications

After we have investigated the power of the extremal index approach and competing methodologies in theoretical setups, we now shed light on the practical implications of our approach. Our focus is on four questions, which one might summarize under the title “Historical Simulation, Few Violations, and the Rejection of GARCH Models”.

The first question aims again at HS, which is not only widespread in the academic literature as is evident from the frequent use in simulation studies (as in this article and others) but also one of the most popular forecasting approaches used in practice. See, for example, Pérignon and Smith (2010) who report that HS was the most used procedure in 2005 with a percentage of 47.4% among their sample. Despite its prevalence, HS in its classical form should be rejected as a correct conditional approach due to its lack of a quick reaction to changing volatility. Therefore, we check backtesting results of HS in two different periods. First, we backtest HS for a 1% VaR on the S&P 500 index in a phase containing the last financial crisis (January 15, 2008 till December 31, 2011), and second the subsequent relatively calm phase (January 1, 2012 till December 22, 2015). Both backtesting periods consist of exactly 1000 returns which lead to an expectation of ten violations. We re-use these periods for Questions 2 and 3 below. The data were downloaded from Yahoo Finance.

The second question addresses a finding of Pérignon, Deng, and Wang (2008) and Pérignon and Smith (2010). In the first-named paper, the authors report that disclosed VaR numbers of Canadian banks were way too conservative in the past (seventy-four violations expected, only two violations happened) and suggest two explanations. First, it is conjectured that markets will severely punish banks who underestimate risk which makes them possibly very conservative. Second, a lack of correct accounting for diversification across departments of a bank could yield too large risk estimates, too. In the second paper, this conservatism is also found in another sample containing U.S., Canadian, and international banks. Interestingly, in the subsequent financial crisis, almost all banks suffered substantial losses which are surprising given that market risk measurements of banks were considered conservative before. Therefore, we analyze how the extremal index approach enables to assess independence even in the absence of many violations. We achieve this by calculating incorrect conditional VaR forecasts at an unconventional low level of 0.05% by simply fitting a skewed Student's t -distribution (Azzalini and Capitanio, 2003) to the $T_e \in \{250, 500\}$ observations prior to time t . Note that Eling (2014) found that this distribution can provide a good fit for asset returns. Due to the low VaR level, only 0.5 violations can be expected throughout the considered time periods.

The third question we consider is about distinguishing different specifications of the volatility process of GARCH-type models at the 1% VaR level. Note that GARCH model-based forecasts possibly provide the most common alternative to HS, aiming at more accurate forecasts due to their particular focus on time-changing aspects. However, there are many different models available and a modeler has to assess which of them captures the dynamic behavior the best. Hence, we adopt two very popular GARCH specifications (Bollerslev, 1986; Glosten, Jagannathan, and Runkle, 1993) as well as ARCH(1) and report how backtesting results differ. Of course, it is important to note that backtesting is not the appropriate method for comparing the accuracy of two or more forecasters. Nevertheless, it is interesting to see whether disparities can be made visible by backtesting.

Finally, the fourth question¹⁰ concerns the specification of the innovation distribution, which, next to the choice of the model for the volatility, is possibly most crucial for forecasting accuracy, see, for example, Kole et al. (2017). Although the innovation distribution might primarily influence the ability to obtain correct unconditional coverage, we aim at

10 We thank an anonymous referee for his proposal of investigating this issue.

checking whether we can also detect differences regarding the independence/no clustering property. For that purpose, we calculate forecasts using the volatility dynamics of Question 3 both with a normal distribution and a t -distribution for the innovation sequence.

For each forecasting exercise, we perform one-day ahead forecasts based on a rolling window scheme of previous returns. Questions 1 and 2 use estimation sample sizes T_e of 250 and 500. In the GARCH case, we chose windows of $T_e = 1000$ returns. Respective results for all four questions are presented in Table 5. Panel A corresponds to Question 1, Panel B to Question 2, and Panel C/D to Question 3/4.

First, we focus on Question 1. Panel A of the turbulent period shows that both HS approaches yield way too many violations. Most independence backtests are able to reject both methods. Here, the only failing backtest is LR_{ind}^{Mar} . Throughout the calm period, HS forecasts are more appropriate (see the smaller number of violations N_1) but can still be rejected by the use of independence backtests. Especially, both extremal index approaches reject the forecasts, but also LR_{ind}^{Mar} which failed in the turbulent phase. This somewhat surprising change can be explained by the fact that LR_{ind}^{Mar} can only detect violations that occurred on subsequent days. Moreover, we observe that the rejection of the longer estimation sample using $T_e = 500$ returns appears to be easier, as expected from the simulation results and their interpretation in Section 3.

Next, we turn to Question 2. We observe that the number of violations in each case of the Panel B is quite close to the expectation. Hence, we cannot reject unconditional coverage in any of these cases. However, due to the nature of the FC employed, we would still like to reject the independence hypothesis. As the main result from Question 1, we have seen that most independent backtests can reject unconditional forecasts quite satisfactorily. However, if violations are very rare as in the present setting, then a correct assessment can become either impossible or the assessment itself rather meaningless. Table 5 shows that binary tests are often not feasible or cannot reject the null.¹¹ Instead, the extremal index approach $\Theta_{noc}^B(40)$ decouples its result from the presence of violations and yields very similar p -values as for the HS scenarios in Question 1.

Finally, the results for Questions 3 and 4 are presented in Panels C and D. Regarding Question 3, we find that the ARCH(1) model is generally rejected in the turbulent phase (except by the Markov test) and also in the calm phase for normal innovations. The GARCH and GJR-GARCH models are harder to reject. In almost no case, an independence backtest is able to reject one of these two models, which is, to some extent, in line with the literature (Hansen and Lunde, 2005). Only in a few cases, the p -values are below 5%. The extremal index backtest Θ_{noc}^B shows barely a sign of misspecification, which is noteworthy due to its often large power in our simulations.

Regarding Question 4, we find that modeling the innovation sequence by a t -distribution yields typically more appropriate forecasts in terms of unconditional coverage compared to the normal distribution. In the turbulent period, we obtain consistently fewer violations with the t -distribution. Nevertheless, there are still too much of them. In the

11 Note that we report a p -value whenever our procedures return a finite value and “-” if not. The former happens also sometimes in cases with $N_1 \leq 1$ where actually no meaningful information is available. In these cases, the p -values should be interpreted with caution.

calm period, both distributions achieve fewer violations, while the t -distribution is again more suitable than the normal distribution. Regarding the violation of independence, we find some evidence that p -values tend to be slightly smaller in the normal case. This is in line with the impression obtained from the number of violations.

5 Conclusion

In this article, a new idea for the assessment of VaR forecasts with respect to violation clustering is worked out in detail. For that purpose, we implement two recently proposed estimators for the extremal index, derive corresponding new backtests, and compare them to existing ones. The results show that especially the sliding blocks estimator from Northrop (2015) and Berghaus and Bücher (2018) is suitable for this task. The corresponding backtest exhibits substantial power improvements in many theoretical scenarios and can easily reject unconditional forecasters even in the absence of violations, a feature lacked by many other backtests. The latter feature is possibly interesting to detect bad forecasts even if they are conservative, since conservative forecasts can fail to accurately adapt to changing markets, too. Furthermore, we briefly hint at possible extensions to backtesting ES, which may become more important in the future.

Supplementary Data

Supplementary data are available at *Journal of Financial Econometrics* online.

Acknowledgments

This work has been supported in part by the Collaborative Research Center “Statistical Modeling of Nonlinear Dynamic Processes” (SFB 823, Subproject A7) of the German Research Foundation (DFG). By looking at other papers published in the *Journal of Financial Econometrics* I see that funding is typically included at this place. Should we still include a separate funding section? Furthermore, note that we have added two conferences to the acknowledgements where earlier versions of the paper have been presented.

References

- Azzalini, A., and Capitanio A.. 2003. Distributions Generated by Perturbation of Symmetry with Emphasis on a Multivariate Skew t -Distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65: 367–389.
- BCBS 1996a. Overview of the Amendment to the Capital Accord to incorporate Market Risks. *Basel Committee on Banking Supervision*.
- BCBS 1996b. Supervisory Framework for the Use of Backtesting in Conjunction with the Internal Models Approach to Market Risk Capital Requirements. *Basel Committee on Banking Supervision*.
- BCBS 2016. Minimum Capital Requirements for Market Risk. *Basel Committee on Banking Supervision*.
- Beirlant, J., Goegebeur Y., Segers J., and Teugels J.. 2004. *Statistics of Extremes: Theory and Applications*. Wiley Series in Probability and Statistics. Chichester: John Wiley & Sons Ltd.
- Berghaus, B., and Bücher A.. 2018. Weak Convergence of a Pseudo Maximum Likelihood Estimator for the Extremal Index. *The Annals of Statistics* 46: 2307–2335.

- Berkowitz, J. 2001. Testing Density Forecasts, with Applications to Risk Management. *Journal of Business & Economic Statistics* 19: 465–474.
- Berkowitz, J., Christoffersen P., and Pelletier D.. 2011. Evaluating Value-at-Risk Models with Desk-Level Data. *Management Science* 57: 2213–2227.
- Bollerslev, T. 1986. Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics* 31: 307–327.
- Candelon, B., Colletaz G., Hurlin C., and Tokpavi S.. 2011. Backtesting Value-at-Risk: A GMM Duration-Based Test. *Journal of Financial Econometrics* 9: 314–343.
- Christoffersen, P., and Pelletier D.. 2004. Backtesting Value-at-Risk: A Duration-Based Approach. *Journal of Financial Econometrics* 2: 84–108.
- Christoffersen, P. F. 1998. Evaluating Interval Forecasts. *International Economic Review* 39: 841.
- Costanzino, N., and Curran M.. 2015. Backtesting General Spectral Risk Measures with Application to Expected Shortfall. *The Journal of Risk Model Validation* 9: 21–31.
- Du, Z., and Escanciano J. C.. 2017. Backtesting Expected Shortfall: Accounting for Tail Risk. *Management Science* 63: 940–958.
- Dufour, J.-M. 2006. Monte Carlo Tests with Nuisance Parameters: A General Approach to Finite-Sample Inference and Nonstandard Asymptotics. *Journal of Econometrics* 133: 443–477.
- Eling, M. 2014. Fitting Asset Returns to Skewed Distributions: Are the Skew-Normal and Skew-Student Good Models?. *Insurance: Mathematics and Economics* 59: 45–56.
- Embrechts, P., Klüppelberg C., and Mikosch T.. 1997. *Modelling Extremal Events for Insurance and Finance*, Volume 33 of Applications of Mathematics. Berlin u.a.: Springer.
- Escanciano, J. C., and Olmo J.. 2010. Backtesting Parametric Value-at-Risk with Estimation Risk. *Journal of Business & Economic Statistics* 28: 36–51.
- Ferro, C. A. T., and Segers J.. 2003. Inference for Clusters of Extreme Values. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65: 545–556.
- Glosten, L. R., Jagannathan R., and Runkle D. E.. 1993. On the Relation Between the Expected Value and the Volatility of the Nominal Excess Return on Stocks. *The Journal of Finance* 48: 1779.
- Hansen, P. R., and Lunde A.. 2005. A Forecast Comparison of Volatility Models: Does Anything Beat a GARCH(1,1)?. *Journal of Applied Econometrics* 20: 873–889.
- Kerkhof, J., and Melenberg B.. 2004. Backtesting for Risk-Based Regulatory Capital. *Journal of Banking & Finance* 28: 1845–1865.
- Kole, E., Markwat T., Opschoor A., and van Dijk D.. 2017. Forecasting Value-at-Risk under Temporal and Portfolio Aggregation. *Journal of Financial Econometrics* 15: 649–677.
- Kratz, M., Lok Y. H., and McNeil A. J.. 2018. Multinomial VaR Backtests: A Simple Implicit Approach to Backtesting Expected Shortfall. *Journal of Banking & Finance* 88: 393–407.
- Kupiec, P. H. 1995. Techniques for Verifying the Accuracy of Risk Measurement Models. *The Journal of Derivatives* 3: 73–84.
- Leadbetter, M. R. 1983. Extremes and Local Dependence in Stationary Sequences. *Zeitschrift für Wahrscheinlichkeitstheorie Und Verwandte Gebiete* 65: 291–306.
- McNeil, A. J., Frey R., and Embrechts P.. 2005. *Quantitative Risk Management. Princeton Series in Finance*. Princeton, NJ: Princeton University Press.
- Mikosch, T., and Starica C.. 2000. Limit Theory for the Sample Autocorrelations and Extremes of a GARCH (1, 1) Process. *The Annals of Statistics* 28: 1427–1451.
- Northrop, P. J. 2015. An Efficient Semiparametric Maxima Estimator of the Extremal Index. *Extremes* 18: 585–603.
- Pérignon, C., Deng Z. Y., and Wang Z. J.. 2008. Do Banks Overstate Their Value-at-Risk?. *Journal of Banking & Finance* 32: 783–794.
- Pérignon, C., and Smith D. R.. 2010. The Level and Quality of Value-at-Risk Disclosure by Commercial Banks. *Journal of Banking & Finance* 34: 362–377.

- Rosenblatt, M. 1952. Remarks on a Multivariate Transformation. *The Annals of Mathematical Statistics* 23: 470–472.
- Smith, R. L., and Weissman I.. 1994. Estimating the Extremal Index. *Journal of the Royal Statistical Society: Series B (Methodological)* 56: 515–528.
- Süveges, M. 2007. Likelihood Estimation of the Extremal Index. *Extremes* 10: 41–55.
- Süveges, M., and Davison A. C.. 2010. Model Misspecification in Peaks over Threshold Analysis. *The Annals of Applied Statistics* 4: 203–221.
- Wong, W. K. 2008. Backtesting Trading Risk of Commercial Banks Using Expected Shortfall. *Journal of Banking & Finance* 32: 1404–1415.
- Ziggel, D., Berens T., Weiß G. N., and Wied D.. 2014. A New Set of Improved Value-at-Risk Backtests. *Journal of Banking & Finance* 48: 29–41.

Mixed-Frequency Macro–Finance Factor Models: Theory and Applications*

Elena Andreou^{1,2}, Patrick Gagliardini^{3,4}, Eric Ghysels^{2,5} and Mirco Rubin⁶

¹University of Cyprus, ²CEPR, ³Università della Svizzera italiana, ⁴Swiss Finance Institute,

⁵University of North Carolina - Chapel Hill and ⁶EDHEC Business School

Address correspondence to Elena Andreou, University of Cyprus, University Avenue 1, P. O. Box 20537, 1678 Nicosia, Cyprus, or e-mail: elena.andreou@ucy.ac.cy.

Received April 20, 2020; revised April 20, 2020; accepted April 30, 2020

Abstract

This article presents tests for the existence of common factors spanning two large panels/groups of macroeconomic and financial variables, and the estimation of common and group-specific factors. New analytical results are derived regarding (i) the difference in the asymptotic distribution of the test statistics when aggregating the data first and then extracting the principal components (PCs), or vice versa, as well as (ii) the estimation of the common factor and its asymptotic distribution, extending the work of Andreou et al. (2019). We find that although there is no empirical evidence for one common factor, with constant loadings, in the United States during the period 1963–2017, there is evidence of one common macro–finance factor during the pre- and post-Great Moderation regimes. The aforementioned approaches of estimating PCs yield almost identical common and group-specific (financial and macro) factors which turn out to be significant in predicting key economic indicators, such as real Gross Domestic Product (GDP) growth and the CBOE Volatility Index, among others.

Key words: large panel, unobservable pervasive factors, mixed frequency, canonical correlations, forecasting models

JEL classification: C22, C38, C53, C55

This article contributes to our understanding of factor models, one of many areas which was of keen interest to Peter Christoffersen, in particular when it touched on research in

* We would like to thank the Editor, Frank Diebold, and two referees for insightful comments which helped improve our article. The first author would like to acknowledge that this work was funded by the Republic of Cyprus through the Research and Innovation Foundation (Project: INTERNATIONAL/USA/0118/0043). The authors would like to thank Mathias Drehmann for sharing his business and financial cycle indicators series.

both finance and econometrics.¹ Christoffersen was a real scholar and we would like to illustrate this by reporting on an exchange we had with him. At some point, we asked Christoffersen whether we could use his realized skewness series for our own research—obviously citing the original work (Amaya et al., 2015). Christoffersen graciously sent us his data series, and we are happy to put that series—among others—to good use in this article.

In this article, new analytical results are derived for the asymptotic distribution of the principal components (PCs) as well as the test for common factors between groups/panels of variables of mixed data frequencies, when either aggregating the data first and then extracting the PCs or when applying PC analysis (PCA) first and then aggregating the estimates. In addition, the asymptotic theory results are derived for the common factor estimation methods. New empirical results are also presented to test for the existence of common factors spanning two large panels/groups of macroeconomic and financial variables, as well as to estimate common and group-specific factors related to each of the aforementioned panels.

Hence this article contributes to the macro–finance literature in extracting the common factor (CF) also referred to as macro–finance factor between the financial sector and the U.S. real/nominal economy indicators. A macro–finance factor is the common part among the spaces of pervasive factors in the macro and finance panels. In other words, it is a common factor among the PCs extracted independently from the panels of macro and finance series. The role of the common as well as financial- and macro-specific factors for forecasting key macroeconomic and financial indicators is evaluated both in-sample (IS) and out-of-sample (OOS), uncovering some interesting results.

Macro panel data are often sampled at a low frequency (LF) (e.g., annual/quarterly), whereas higher frequency (HF) (e.g., daily/weekly) data are typically collected pertaining to financial indicators. Prime examples are, for example, for the macro panels, the Stock and Watson (2008) quarterly data as well as the McCracken and Ng (2016) FRED quarterly/monthly data dominated by macroeconomic indicators, versus higher frequency financial indicators such as the (intra)daily stock market or exchange rate indices, the Gilchrist and Zakrajsek (2012) monthly credit spreads panel and the Fama and French portfolios of sorted equity returns. Extracting the common factor (the evidence suggests there is only a single common factor—more on this later) between the two large panels/groups of macroeconomic and financial markets series provides a way to study how the U.S. common macro–finance component has evolved over time, its cyclical behavior, which variables drive this common factor, how the factor and/or its loadings might have changed over the last 55 years, as well as the role of the common factor in Granger causing and forecasting key economic indicators.

In extracting factors from mixed-frequency group panels, two approaches are pursued: the first approach is to aggregate the HF data and then perform PCA while the second approach refers to extracting the PCA first. Andreou et al. (2019) established the large sample distributional properties of the statistic for testing the number of common and group-specific factors in the first approach. While the two alternative approaches were also compared in Andreou et al. (2019) with accompanying simulation evidence, this article provides analytical results deriving the asymptotic expansion of PCs estimates following these two approaches and how these affect the distribution of the test for common factors

1 A partial list of his contributions in the area includes Christoffersen, Ghysels, and Swanson (2002), Christoffersen, Fournier, and Jacobs (2017), Christoffersen et al. (2014), Christoffersen and Langlois (2013), among others.

between the groups/panels. Moreover, conditions are presented under which the asymptotic distributions of the common factor test following the two approaches coincide.

Our empirical analysis presents evidence that the two approaches of aggregation first/PCA last or PCA first/aggregation last, yield almost identical PCs estimates as well as inferences regarding the number of factors and the common factor test. For our given groups of the monthly financial panel and the quarterly U.S. macro panel, we find that although there is no evidence for one common factor in the United States during the full sample period starting from the 1960s, there is, however, evidence of one CF during the pre- and post-Great Moderation (GM) regimes. We show that this is due to structural changes in the loadings of the common factor during the period 1963–2017, which is driven by almost all the categories of macro and financial variables considered in this study. In addition, group-specific factors, namely HF financial factors and LF macro factors are extracted. The forecasting role of the aforementioned factors (common and group specific) is further investigated in predicting key macroeconomic as well as financial indicators such as real GDP and consumption of services and nondurable goods growth, the Moody’s corporate bond default spread, the CBOE Volatility Index (VIX) and Variance Risk Premium (VRP) as well as the Exchange Traded Fund (ETF) iShares Core S&P500 returns. Our mixed-frequency group factors are also compared with other well-known factors in the literature extracted from different but related panels such as the mixed-frequency small panel factor measuring real business conditions of [Aruoba, Diebold, and Scotti \(2009\)](#), the large panel of corporate spreads factor of [Gilchrist and Zakrajsek \(2012\)](#), the large panel extracting an activity index/factor by [Brave and Butters \(2012\)](#), among others. Last but not least, the role of groups/panels of mixed sampling frequencies of data in estimating common and group-specific factors via PCs is also compared with the traditional approach that extracts factors from a single panel that stacks all variables (both macro and financial) together in common (low) frequency.

The article is organized as follows: Section 1 presents the mixed-frequency group factor model and the test for common factors. Section 2 provides the asymptotic results of the factors and the common factor test for the two approaches: PCA first or PCA last as well as the asymptotic distribution of estimators of the common factor. Section 3 presents a comprehensive empirical analysis, and Section 4 concludes the article. [Online Appendix](#), henceforth referred to as OA, provides proofs, supplementary theoretical results, an extensive description of the dataset used in the empirical application, and additional empirical results.

1 Group Factor Models

In this section, we revisit the class of group factor models studied by [Andreou et al. \(2019\)](#), henceforth AGGR.² We use the following notation for the group factor model setting, assuming two groups:

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} \Lambda_1^c & \Lambda_1^s & 0 \\ \Lambda_2^c & 0 & \Lambda_2^s \end{bmatrix} \begin{bmatrix} f_t^c \\ f_{1,t}^s \\ f_{2,t}^s \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix}, \tag{1.1}$$

2 See also [Kose, Otrok, and Whiteman \(2008\)](#), [Goyal, Pérignon, and Villa \(2008\)](#), [Chen \(2012\)](#), [Wang \(2012\)](#), [Ando and Bai \(2015\)](#), and [Breitung and Eickmeier \(2016\)](#) for recent contributions to the group factor model literature.

where $y_{j,t} = [y_{j,1t}, \dots, y_{j,N_j t}]'$ collects observations for N_j individuals in group j , $\Lambda_j^c = [\lambda_{j,1}^c, \dots, \lambda_{j,N_j}^c]'$ and $\Lambda_j^s = [\lambda_{j,1}^s, \dots, \lambda_{j,N_j}^s]'$ are the matrices of factor loadings, and $\varepsilon_{j,t} = [\varepsilon_{j,1t}, \dots, \varepsilon_{j,N_j t}]'$ is the vector of error terms, with $j = 1, 2$, and $t = 1, \dots, T$, related to our empirical analysis of the macro and financial groups/panels. The dimensions of the common factor f_t^c and the group-specific factors $f_{1,t}^s, f_{2,t}^s$ are, respectively, k^c, k_1^s , and k_2^s . The errors and factor processes are stationary, serially mixing, and satisfy the assumptions on weak cross-sectional dependence and existence of higher-order moments in AGGR, Appendix A. The group-specific factors $f_{1,t}^s$ and $f_{2,t}^s$ are orthogonal to the common factor f_t^c . Since the unobservable factors can be standardized, we assume (see Assumption A.2 in AGGR):

$$E \begin{bmatrix} f_t^c \\ f_{1,t}^s \\ f_{2,t}^s \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad \text{and} \quad V \begin{bmatrix} f_t^c \\ f_{1,t}^s \\ f_{2,t}^s \end{bmatrix} = \begin{bmatrix} I_{k^c} & 0 & 0 \\ 0 & I_{k_1^s} & \Phi \\ 0 & \Phi' & I_{k_2^s} \end{bmatrix}, \quad (1.2)$$

where I_k denotes the identity matrix of order k . We allow for a nonzero covariance Φ between group-specific factors. Under an identification condition implied by the set of assumptions in AGGR, the rotational invariance of Equations (1.1) and (1.2) allows only for separate rotations among the components of $f_{1,t}^s$, among those of $f_{2,t}^s$, and finally those of f_t^c , and therefore maintains the interpretation of common and group-specific factors.

We consider the generic setting of Equation (1.1) and let $k_j = k^c + k_j^s$, for $j = 1, 2$, be the dimensions of the pervasive factor spaces for the two groups, and define $\underline{k} = \min(k_1, k_2)$. We collect the factors of each group in the k_j -dimensional vectors $b_{j,t} = (f_t^c, f_{j,t}^s)'$ and define their variance and covariance matrices: $V_{j\ell} := E(b_{j,t} b_{\ell,t}')$, $j, \ell = 1, 2$. From Equation (1.2) we have $V_{jj} = I_{k_j}$ for $j = 1, 2$. AGGR show that the factor space dimensions k^c, k_1^s, k_2^s are identifiable using canonical correlation analysis applied to $b_{1,t}$ and $b_{2,t}$. In particular, according to their Proposition 1, it is shown that the number of common factors k^c , the common factor space spanned by f_t^c , and the spaces spanned by group-specific factors can be identified from the canonical correlations and canonical variables of $b_{1,t}$ and $b_{2,t}$. Therefore, the factor space dimensions k^c and k_j^s and factors f_t^c and $f_{j,t}^s$, $j = 1, 2$, are identifiable (up to a rotation) from information that can be inferred by disjoint PCA on the two subgroups. Indeed, disjoint PCA on the two subgroups allows us to identify the dimensions k_1 and k_2 , and vectors $b_{1,t}$ and $b_{2,t}$ up to linear one-to-one transformations. Therefore, our group factor model provides some key insights in estimating and testing for the existence of the CF between these two groups/panels, while at the same time estimating the group-specific financial and macro factors which are orthogonal to the common factor.

Assuming for a moment that the true number of factors $k_j > 0$ in each subgroup $j = 1, 2$, is known, and also that the true number of common factors $k^c > 0$, is known, then the following estimation procedure for the common factors can be implemented. Let $b_{1,t}$ and $b_{2,t}$ be estimated (up to a rotation) by extracting the first k_j PCs from each subpanel j , and denote by $\hat{b}_{j,t}$ these PC estimates of the factors, $j = 1, 2$. Let $\hat{H}_j = [\hat{b}_{j,1}, \dots, \hat{b}_{j,T}]'$ be the (T, k_j) matrix of estimated PCs extracted from panel $Y_j = [y_{j,1}, \dots, y_{j,T}]'$ associated with the largest k_j eigenvalues of matrix $\frac{1}{N_j T} Y_j Y_j'$, $j = 1, 2$. Let $\hat{V}_{j\ell}$ denote the empirical covariance matrix of

the estimated vectors $\hat{h}_{j,t}$ and $\hat{h}_{\ell,t}$, that is, $\hat{V}_{j\ell} = \frac{1}{T} \hat{H}'_j \hat{H}_\ell = \frac{1}{T} \sum_{t=1}^T \hat{h}_{j,t} \hat{h}'_{\ell,t}$, $j, \ell = 1, 2$, and let matrices \hat{R} and \hat{R}^* be defined as:

$$\hat{R} := \hat{V}_{11}^{-1} \hat{V}_{12} \hat{V}_{22}^{-1} \hat{V}_{21}, \quad \text{and} \quad \hat{R}^* := \hat{V}_{22}^{-1} \hat{V}_{21} \hat{V}_{11}^{-1} \hat{V}_{12}, \tag{1.3}$$

where \hat{R} and \hat{R}^* have the same nonzero eigenvalues. The k^c largest eigenvalues of \hat{R} (resp. \hat{R}^*), denoted by $\hat{\rho}_\ell^2$, $\ell = 1, \dots, k^c$, are the first k^c squared sample canonical correlation between $\hat{h}_{1,t}$ and $\hat{h}_{2,t}$. The associated k^c canonical directions, collected in the (k_1, k^c) matrix \hat{W}_1 (resp. (k_2, k^c) matrix \hat{W}_2), are the eigenvectors associated with the k^c largest eigenvalues of matrix \hat{R} (resp. \hat{R}^*), normalized to have length 1 with respect to \hat{V}_{11} (resp. \hat{V}_{22}). It also holds that $\hat{W}'_1 \hat{V}_{11} \hat{W}_1 = I_{k^c}$, and $\hat{W}'_2 \hat{V}_{22} \hat{W}_2 = I_{k^c}$.

AGGR consider two estimators of the common factors vector, that are $\hat{f}_t^c = \hat{W}'_1 \hat{h}_{1,t}$ and $\hat{f}_t^{c*} = \hat{W}'_2 \hat{h}_{2,t}$. Note that, $\frac{1}{T} \sum_{t=1}^T \hat{f}_t^c \hat{f}_t^{c'} = I_{k^c}$, and similarly for \hat{f}_t^{c*} , that is, the estimated common factor values match IS the normalization condition of identity variance–covariance matrix in Equation (1.2). In this article, we explore the idea that linear combinations of these two “basis” estimators also yield valid estimators. More specifically, let us consider the estimator

$$\hat{f}_t^{c*} = S(\omega) \left(\hat{f}_t^c + \omega \hat{f}_t^{c*} \right), \tag{1.4}$$

where scalar parameter ω is the weight. Transformation by matrix $S(\omega) = [(1 + \omega^2)I_{k^c} + 2\omega\hat{D}]^{-1/2}$, where $\hat{D} = \text{diag}(\hat{\rho}_1, \dots, \hat{\rho}_{k^c})$ ensures property $\frac{1}{T} \sum_{t=1}^T \hat{f}_t^{c*} (\hat{f}_t^{c*})' = I_{k^c}$ in sample.³

Note that the idea we explore is reminiscent of forecast combinations, originated by Bates and Granger (1969) who studied optimal mean square error (MSE) forecast combinations. The natural question which emerges is how to choose the weight. One possibility is suggested by revisiting the work in Goyal, Pérignon, and Villa (2008) and references therein. They consider the estimator of the common factors that is obtained by the rows of the (T, k^c) matrix of standardized eigenvectors of matrix $\frac{1}{T}(\hat{H}_1 \hat{H}'_1 + \hat{H}_2 \hat{H}'_2)$ associated with the k^c largest eigenvalues. The computations in Section D.2 of Online Appendix of AGGR show that the rows of the eigenvectors matrix are $(\hat{f}_t^c + \hat{f}_t^{c*})'$, $t = 1, \dots, T$, up to normalization. Hence, the Goyal, Pérignon, and Villa (2008) estimator corresponds to the linear combination in Equation (1.4) with weight $\omega = 1$, that is, the equally weighted linear combination of the two basis estimators \hat{f}_t^c and \hat{f}_t^{c*} .

An alternative choice for ω is provided by the optimal weight which minimizes the asymptotic MSE (AMSE) of the factor estimator which would be more in line with the approach put forward by Bates and Granger (1969) (see also Timmermann, 2006 for a survey). We consider the asymptotics with $N_1, N_2, T \rightarrow \infty$ such that

$$N_2/N_1 \rightarrow \mu > 0, \quad \sqrt{T}/N_2 = o(1), \quad N_2/T^{5/2} = o(1). \tag{1.5}$$

Further, to simplify the exposition, we focus here on the setting with $k^c = 1$, that is, a single common factor as we find in our empirical analysis, and conditionally homoskedastic errors that are uncorrelated across series and panels (see Online Appendix for a more

3 To see this, use $\frac{1}{T} \sum_{t=1}^T \hat{f}_t^c (\hat{f}_t^{c*})' = \hat{W}'_1 \hat{V}_{12} \hat{W}_2 = \hat{D}$. One could use a matrix-valued weight ω as well.

general analysis). From AGGR [Online Appendix](#) Section D.5, we have the joint asymptotic distribution

$$\begin{bmatrix} \sqrt{N_1} \left(\hat{\mathcal{H}}_c \hat{f}_t^c - f_t^c - \frac{1}{T} \beta_{1,t}^c \right) \\ \sqrt{N_2} \left(\hat{\mathcal{H}}_c^* \hat{f}_t^{c*} - f_t^c - \frac{1}{T} \beta_{2,t}^c \right) \end{bmatrix} \xrightarrow{d} N(0, \Sigma_u), \tag{1.6}$$

where the asymptotic variance is the (2,2) diagonal matrix $\Sigma_u = \text{diag}(\Sigma_{u,11}^{(cc)}, \Sigma_{u,22}^{(cc)})$, with $\Sigma_{u,jj} = \Sigma_{\Lambda,j}^{-1} \Omega_{\Lambda,j} \Sigma_{\Lambda,j}^{-1}$, $\Sigma_{\Lambda,j} = \lim_{N_j \rightarrow \infty} \frac{1}{N_j} \sum_{i=1}^{N_j} \lambda_{j,i} \lambda'_{j,i}$ and $\Omega_{\Lambda,j} = \lim_{N_j \rightarrow \infty} \frac{1}{N_j} \sum_{i=1}^{N_j} \gamma_{j,i} \lambda_{j,i} \lambda'_{j,i}$, for $\lambda_{j,i} = (\lambda_{j,i}^c, \lambda_{j,i}^s)'$ and $\gamma_{j,i} = E[\hat{e}_{j,i,t}^2]$, $j = 1, 2$, and (cc) denotes the upper-left element of a matrix. Random variables $\hat{\mathcal{H}}_c$ and $\hat{\mathcal{H}}_c^*$ converge to 1 in probability for the suitable sign fix of the latent factor. The bias terms are $\beta_{j,t}^c = \bar{\gamma}_j (\Sigma_{\Lambda,j}^{-1} b_{j,t})^{(c)}$, $j = 1, 2$, where $\bar{\gamma}_j = \lim_{N_j \rightarrow \infty} \frac{1}{N_j} \sum_{i=1}^{N_j} \gamma_{j,i}$. From [Equation \(1.6\)](#), we obtain the AMSE of the linear combination \hat{f}_t^{c*} in [Equation \(1.4\)](#), which depends on the factor realization f_t via the asymptotic bias. In [Online Appendix](#), we show that the average (across factor realizations) AMSE is minimized for

$$\omega = \frac{\frac{1}{N_1} \Sigma_{u,11}^{(cc)} + \frac{1}{T^2} (B_{11} - B_{12})}{\frac{1}{N_2} \Sigma_{u,22}^{(cc)} + \frac{1}{T^2} (B_{22} - B_{12})}, \tag{1.7}$$

where $B_{jj} = \bar{\gamma}_j^2 [\Sigma_{\Lambda,j}^{-2}]^{(cc)}$, $j = 1, 2$, and $B_{12} = \bar{\gamma}_1 \bar{\gamma}_2 [\Sigma_{\Lambda,1}^{-1} V_{12} \Sigma_{\Lambda,2}^{-1}]^{(cc)}$. When $N_1/T^2 = o(1)$, the bias terms do not matter, and the optimal weight ω depends positively on the ratio of the error variances $\Sigma_{u,11}^{(cc)}/\Sigma_{u,22}^{(cc)}$ and the ratio of the cross-sectional dimensions N_2/N_1 .⁴ If N_1/T^2 does not shrink to zero, there is an effect from the bias terms. Of course, the parametric family [Equation \(1.4\)](#) encompasses the AGGR estimators \hat{f}_t^c and \hat{f}_t^{c*} , which correspond to choices $\omega = 0$ and $\omega = +\infty$, respectively.

For a given choice of the weight ω , let $\hat{F}^{c*} = [\hat{f}_1^{c*}, \dots, \hat{f}_T^{c*}]'$ be the (T, k^c) matrix of estimated common factors, and $\hat{\Lambda}_j^c = [\hat{\lambda}_{j,1}^c, \dots, \hat{\lambda}_{j,N_j}^c]'$ the (N_j, k^c) matrix collecting the estimated loadings:

$$\hat{\Lambda}_j^c = Y_j' \hat{F}^{c*} (\hat{F}^{c*}{}' \hat{F}^{c*})^{-1} = \frac{1}{T} Y_j' \hat{F}^{c*}, \quad j = 1, 2. \tag{1.8}$$

Moreover, let $\zeta_{j,t}^c = y_{j,t} - \hat{\Lambda}_j^c \hat{f}_t^{c*}$ be the residuals of the regression of $y_{j,t}$ on the estimated common factor \hat{f}_t^{c*} , for $j = 1, 2$ and $\Xi_j = [\zeta_{j,1}, \dots, \zeta_{j,T}]'$ be the (T, N_j) matrix of the regression residuals, for $j = 1, 2$. Estimators of the specific factors $\hat{f}_{1,t}^s$ (resp. $\hat{f}_{2,t}^s$) are defined as the first k_1^s (resp. k_2^s) PCs of subpanel Ξ_1 (resp. Ξ_2), namely, the columns of the (T, k_j^s) matrix $\hat{F}_j^s = [\hat{f}_{j,1}^s, \dots, \hat{f}_{j,T}^s]'$ are the eigenvectors associated with the k_j^s largest eigenvalues of matrix $\frac{1}{N_j T} \Xi_j \Xi_j'$, normalized to have $\hat{F}_j^s \hat{F}_j^s / T = I_{k_j^s}$ for $j = 1, 2$. By construction, the estimated common factors in the columns of \hat{F}^{c*} are orthogonal in sample to the estimated

4 AGGR assumes that $N_2 = \min\{N_1, N_2\}$ without loss of generality. Note that depending on the application, N_2 may pertain to either the LF or HF data panel.

group-specific factors \hat{F}_j^s , for $j = 1, 2$. Finally, the loadings of the group-specific factors are estimated by

$$\hat{\Lambda}_j^s = Y_j' \hat{F}_j^s \left(\hat{F}_j^s ' \hat{F}_j^s \right)^{-1} = \frac{1}{T} \Xi_j' \hat{F}_j^s, \quad j = 1, 2, \tag{1.9}$$

where $\hat{\Lambda}_j^s = [\hat{\lambda}_{j,1}^s, \dots, \hat{\lambda}_{j,N_j}^s]'$ is the (N_j, k_j^s) matrix collecting the estimated loadings.

How does one determine the dimension k^c of the common factor space? To answer this question, we first consider the case where the number of pervasive factors k_1 and k_2 in each subpanel is known, hence $\underline{k} = \min(k_1, k_2)$ is also known, and we relax this assumption below. The dimension k^c is the number of unit canonical correlations between $b_{1,t}$ and $b_{2,t}$, see Proposition 1 in AGGR. We consider the hypotheses: $H(0) = \{1 > \rho_1 \geq \dots \geq \rho_{\underline{k}}\}$, $H(1) = \{\rho_1 = 1 > \rho_2 \geq \dots \geq \rho_{\underline{k}}\}$, \dots , $H(k^c) = \{\rho_1 = \dots = \rho_{k^c} = 1 > \rho_{k^c+1} \geq \dots \geq \rho_{\underline{k}}\}$, \dots , and finally, $H(\underline{k}) = \{\rho_1 = \dots = \rho_{\underline{k}} = 1\}$, where $\rho_1, \dots, \rho_{\underline{k}}$ are the ordered canonical correlations of $b_{1,t}$ and $b_{2,t}$. Hypothesis $H(0)$ corresponds to the absence of common factors. Generically, $H(k^c)$ corresponds to the case of k^c common factors and $k_1 - k^c$ and $k_2 - k^c$ group-specific factors in each group. The largest possible number of common factors is $\underline{k} = \min(k_1, k_2)$. In order to select the number of common factors, let us consider the following sequence of tests: $H_0 = H(k^c)$ against $H_1 = \cup_{0 \leq r < k^c} H(r)$, for each $k^c = \underline{k}, \underline{k} - 1, \dots, 1$. To test H_0 against H_1 , for any given $k^c = \underline{k}, \underline{k} - 1, \dots, 1$ we consider:

$$\hat{\xi}(k^c) = \sum_{\ell=1}^{k^c} \hat{\rho}_\ell. \tag{1.10}$$

The statistic $\hat{\xi}(k^c)$ corresponds to the sum of the k^c largest sample canonical correlations of $\hat{b}_{1,t}$ and $\hat{b}_{2,t}$. We reject the null $H_0 = H(k^c)$ when $\hat{\xi}(k^c) - k^c$ is negative and large.

The critical value is obtained from the large sample distribution of the statistic under the joint asymptotics $N_1, N_2, T \rightarrow \infty$, and the assumptions in Equation (1.5), as provided in AGGR. Then, let $\hat{\Sigma}_U = (N_2/N_1) \hat{\Sigma}_{u,11}^{(cc)} + \hat{\Sigma}_{u,22}^{(cc)}$, with $\hat{\Sigma}_{u,ij} = (\frac{1}{N_j} \hat{\Lambda}_j' \hat{\Lambda}_j)^{-1} (\frac{1}{N_j} \hat{\Lambda}_j' \hat{\Gamma}_j \hat{\Lambda}_j) (\frac{1}{N_j} \hat{\Lambda}_j' \hat{\Lambda}_j)^{-1}$ where $\hat{\Lambda}_j = [\hat{\Lambda}_j^c : \hat{\Lambda}_j^s]$, $\hat{\Lambda}_j^c$ and $\hat{\Lambda}_j^s$ are the loadings estimators, $\hat{\Gamma}_j = \text{diag}(\hat{\gamma}_{j,i}, i = 1, \dots, N_j)$ with $\hat{\gamma}_{j,i} = \frac{1}{T} \sum_{t=1}^T \hat{e}_{j,i,t}^2$, and $\hat{e}_{j,i,t} = y_{j,i,t} - \hat{\lambda}_{j,i}^c f_t^{c*} - \hat{\lambda}_{j,i}^s f_{j,t}^{s*}$, for $j = 1, 2$. Define the test statistic:

$$\tilde{\xi}(k^c) := N\sqrt{T} \left(\frac{1}{2} \text{tr} \left\{ \hat{\Sigma}_U^2 \right\} \right)^{-1/2} \left[\hat{\xi}(k^c) - k^c + \frac{1}{2N} \text{tr} \left\{ \hat{\Sigma}_U \right\} \right], \tag{1.11}$$

with $N = \min\{N_1, N_2\}$. Then Theorem 2 of AGGR, which holds under the assumptions that the errors are conditionally homoskedastic martingale difference sequences and are not cross-sectionally correlated, shows that: (i) under the null hypothesis $H_0 = H(k^c)$ of k^c common factors, we have: $\tilde{\xi}(k^c) \xrightarrow{d} N(0, 1)$ and (ii) under the alternative hypothesis $H_1 = \cup_{0 \leq r < k^c} H(r)$, we have: $\tilde{\xi}(k^c) \xrightarrow{p} -\infty$. Importantly, the asymptotic distribution and rate of convergence of the test statistic $\tilde{\xi}(k^c)$ in Theorem 2 of AGGR are unchanged when the true numbers of pervasive factors k_1 and k_2 are unknown, and is estimated by consistent selection methods as those provided, among others, by Bai and Ng (2002), Onatski (2010), and Ahn and Horenstein (2013). The above test statistics can be used to determine the number of common factors (and therefore by difference the number of group-specific factors) by means of a sequential testing procedure, see Proposition 2 in AGGR.

2 Mixed-Frequency Group Factor Model: PCA First or Last?

When we apply the theory of group factor models to mixed-frequency data, some additional issues emerge, not studied by AGGR. It is the purpose of this section to expand on such issues. First, we pose the case of mixed frequency as a group factor model in the first subsection. Then, we address specific issues hitherto unresolved—namely the interchange of aggregation and PCA. We are able to solve explicitly the comparison under some restrictive conditions and provide practical guidance to empirical work. The derivations in this section are complementary to the Monte Carlo simulations reported in AGGR. They provide analytic support instead of simulation-based evidence.

2.1 Model Structure

Let $t = 1, 2, \dots, T$ be the LF time units. Each time period $(t - 1, t]$ is divided into M subperiods with HF dates $t - 1 + m/M$, with $m = 1, \dots, M$ and the cross-section is of size N_H for the HF data and N_L for the LF data. We let $x_{m,t}^{H,i}$, for $i = 1, \dots, N_H$, be the HF data observation i during subperiod m of LF period t . Similarly, $x_t^{L,i}$, with $i = 1, \dots, N_L$, is the observation of the i^{th} LF series at t . These observations are gathered into the N_H -dimensional vectors $x_{m,t}^H$, for all m , and the N_L -dimensional vector x_t^L , respectively.

There are three types of latent pervasive factors, $g_{m,t}^C$, $g_{m,t}^H$, and $g_{m,t}^L$, respectively, of dimension k^C , k^H , and k^L . The former represents a vector of factors which affect both HF and LF data, and the other two types of factors affect exclusively high (superscript H) and low (marked by L) frequency data. The latent factor model with HF data sampling is:

$$\begin{aligned} x_{m,t}^H &= \Lambda_{HC} g_{m,t}^C + \Lambda_H g_{m,t}^H + e_{m,t}^H, \\ x_{m,t}^{L*} &= \Lambda_{LC} g_{m,t}^C + \Lambda_L g_{m,t}^L + e_{m,t}^L, \end{aligned} \quad (2.1)$$

where $m = 1, \dots, M$ and $t = 1, \dots, T$, and Λ_{HC} , Λ_H , Λ_{LC} , and Λ_L are matrices of factor loadings. The vector $x_{m,t}^{L*}$ is unobserved for each HF subperiod and the measurements, denoted by x_t^L , depend on the observation scheme, which can be either flow sampling or stock sampling (or some general linear scheme).

In the case of flow sampling, the LF observations are the sum (or average) of all $x_{m,t}^{L*}$ across all m , that is, $\bar{x}_t^L = \sum_{m=1}^M x_{m,t}^{L*}$.⁵ Then, model (2.1) implies:

$$\begin{aligned} \bar{x}_t^H &= \Lambda_{HC} g_{m,t}^C + \Lambda_H g_{m,t}^H + e_{m,t}^H, \quad m = 1, \dots, M, \\ \bar{x}_t^L &= \Lambda_{LC} \sum_{m=1}^M g_{m,t}^C + \Lambda_L \sum_{m=1}^M g_{m,t}^L + \sum_{m=1}^M e_{m,t}^L. \end{aligned} \quad (2.2)$$

Let us define the aggregated variables and innovations $\bar{x}_t^H := \sum_{m=1}^M x_{m,t}^H$, $\bar{e}_t^U := \sum_{m=1}^M e_{m,t}^U$, $U = H, L$, and the aggregated factors: $\bar{g}_t^U := \sum_{m=1}^M g_{m,t}^U$, $U = C, H, L$. Then we can stack the observations \bar{x}_t^H and \bar{x}_t^L and write:

$$\begin{bmatrix} \bar{x}_t^H \\ \bar{x}_t^L \end{bmatrix} = \begin{bmatrix} \Lambda_{HC} & \Lambda_H & 0 \\ \Lambda_{LC} & 0 & \Lambda_L \end{bmatrix} \begin{bmatrix} \bar{g}_t^C \\ \bar{g}_t^H \\ \bar{g}_t^L \end{bmatrix} + \begin{bmatrix} \bar{e}_t^H \\ \bar{e}_t^L \end{bmatrix}, \quad (2.3)$$

- 5 In the case of stock sampling, the LF observations are the end-of-period values $x_{m,t}^{L*}$ (or the values at some other given date m within a subperiod). The analysis proceeds analogously, replacing summation over subperiods with evaluation at the end-of-period. We cover the flow sampling here because it corresponds to the empirical analysis reported in later sections.

that is, the group factor model, with common factor \bar{g}_t^C and group-specific factors \bar{g}_t^H and \bar{g}_t^L . The normalized latent common and group-specific factors \bar{g}_t^U , $U = C, H, L$, satisfy the counterpart of Equation (1.2).

Finally, the results in AGGR can be applied for identification and inference in the mixed-frequency factor model—see their section 5 for details. In particular, AGGR show under mild assumptions on the factor loadings, that the HF values $g_{m,t}^C$ and $g_{m,t}^H$ of the common and high-frequency factors (HFFs) are identifiable. Not surprisingly, only the flow-sampled values \bar{g}_t^L of the low-frequency factor (LFF) are identifiable from the LF observations of the corresponding group (or panel).

2.2 Aggregation and PCA

Inference on the factor spaces and their dimensions can be conducted in two ways which are described below. We focus first on the inference on the number of common factors and leave the estimation of factor values for the next subsection.

(1) First flow sample the data and obtain a group factor model for observables $\bar{x}_t^{H,i}$ and $\bar{x}_t^{L,i}$, with pervasive factors, loadings matrices, and idiosyncratic errors given by

$$b_{1,t} = [\bar{g}_t^C', \bar{g}_t^H']', \quad \lambda_{1,i} = [\lambda'_{HC,i}, \lambda'_{H,i}]', \quad \varepsilon_{1,i,t} = \bar{e}_t^{H,i}, \tag{2.4}$$

$$b_{2,t} = [\bar{g}_t^C', \bar{g}_t^L']', \quad \lambda_{2,i} = [\lambda'_{LC,i}, \lambda'_{L,i}]', \quad \varepsilon_{2,i,t} = \bar{e}_t^{L,i}, \tag{2.5}$$

in the HF and LF panels, respectively. Next, apply PCA in each group to get PC estimates $\hat{h}_{1,t}$ and $\hat{h}_{2,t}$, compute the canonical correlations $\hat{\rho}_\ell$ and canonical directions \hat{W}_1 and \hat{W}_2 , and the test statistic $\hat{\xi}(k^C) = \sum_{\ell=1}^{k^C} \hat{\rho}_\ell$ for k^C common factors. Theorem 1 of AGGR implies that under the null $H(k^C)$ the test statistic (after recentering and standardization) is asymptotically standard Gaussian.

(2) First perform PCA on, respectively, the HF and LF panels to extract $b_{1,m,t} = [g_{m,t}^C, g_{m,t}^H]'$ at HF and $b_{2,t} = [\bar{g}_t^C', \bar{g}_t^L']'$ at LF, and then flow sample the HF factor estimates to get $\check{h}_{1,t} = \sum_{m=1}^M \check{h}_{1,m,t}$, compute the canonical correlations $\check{\rho}_\ell$ among LF PCs $\check{h}_{1,t}$ and $\hat{h}_{2,t}$, the canonical directions \check{W}_1 and \check{W}_2 , and the test statistic $\check{\xi}(k^C) = \sum_{\ell=1}^{k^C} \check{\rho}_\ell$ for k^C common factors. The “check” symbol notation highlights the difference with approach (1). If the HF panel data, PC estimates obey an asymptotic expansion of the same type as the one in Proposition 3 of AGGR, then upon aggregation:

$$\check{h}_{1,t} = \check{\mathcal{H}}_1 \left[b_{1,t} + \frac{1}{\sqrt{N_H}} u_{1,t} + \frac{1}{T} \check{b}_{1,t} + \frac{1}{\sqrt{N_H T M}} \check{d}_{1,t} + \check{v}_{1,t} \right],$$

where $b_{1,t}$ is as in Equation (2.4) and

$$\begin{aligned} u_{1,t} &= \frac{1}{N_H} \sum_{i=1}^{N_H} \lambda_{1,i} \lambda'_{1,i} \Big)^{-1} \frac{1}{\sqrt{N_H}} \sum_{i=1}^{N_H} \lambda_{1,i} \bar{e}_t^{H,i}, \quad \lambda_{1,i} = [\lambda'_{HC,i}, \lambda'_{H,i}]', \\ \check{b}_{1,t} &= \frac{1}{N_H} \sum_{i=1}^{N_H} \lambda_{1,i} \lambda'_{1,i} \Big)^{-1} \frac{1}{T M} \sum_{t=1}^T \sum_{m=1}^M b_{1,m,t} b'_{1,m,t} \Big)^{-1} \frac{1}{M} \sum_{m=1}^M \eta_{1,m,t}^2 b_{1,m,t}, \end{aligned} \tag{2.6}$$

$\eta_{1,m,t}^2 = plim_{N_H \rightarrow \infty} \frac{1}{N_H} \sum_{i=1}^{N_H} E[(e_{m,t}^{H,i})^2 | \mathcal{F}_t]$, $\check{d}_{1,t} = \sum_{m=1}^M d_{1,m,t}$, $\check{\vartheta}_{1,t} = \sum_{m=1}^M \vartheta_{1,m,t}$ is a remainder term and $\mathcal{F}_t = \sigma(F_s, s \leq t)$ is the sigma field generated by current and past factor values $F_t = (f_t^c, f_{1,t}^s, f_{2,t}^s)'$.⁶ The asymptotic distribution of the test statistic follows from Theorem 1 in AGGR by substituting the quantities in Equation (2.6) for $j=1$ and those in Equation (2.5) for $j=2$. Specifically, under the null hypothesis $H_0 = H(k^C)$ and the assumptions in Equation (1.5) that correspond to the regularity conditions in Theorem 1 of AGGR, the asymptotic distribution of the test statistics $\check{\xi}(k^c) = \sum_{\ell=1}^{k^c} \check{\vartheta}_{\ell,t}$ in approach (2) is such that:

$$N_L \sqrt{T} \left(\Omega_{U,1} + \frac{N_U}{T^2} \check{\Omega}_{U,2} \right)^{-1/2} \left[\check{\xi}(k^C) - k^C + \frac{1}{2N_L} tr \left\{ \check{\Sigma}_{cc}^{-1} \check{\Sigma}_U \right\} + \frac{1}{2T^2} tr \left\{ \check{\Sigma}_{cc}^{-1} \check{\Sigma}_B \right\} \right] \xrightarrow{d} N(0, 1), \tag{2.7}$$

where $\check{\Sigma}_U = \frac{1}{T} \sum_{t=1}^T \left(\mu_N^2 \check{\Sigma}_{u,11,t}^{(cc)} + \check{\Sigma}_{u,22,t}^{(cc)} - \mu_N \check{\Sigma}_{u,12,t}^{(cc)} - \mu_N \check{\Sigma}_{u,21,t}^{(cc)} \right)$, and

$$\check{\Delta} \tilde{b}_t = \tilde{b}_{1,t} - b_{2,t} - \left(\frac{1}{T} \sum_{s=1}^T (\tilde{b}_{1,s} - b_{2,s}) F_s' \right) \times \frac{1}{T} \sum_{s=1}^T F_s F_s' \Big)^{-1} F_t$$

$$\check{\Sigma}_{cc} = \frac{1}{T} \sum_{t=1}^T \tilde{g}_t^C \tilde{g}_t^{C'} \quad , \quad \check{\Sigma}_B = \frac{1}{T} \sum_{t=1}^T \widetilde{\Delta b}_t^{(c)} \widetilde{\Delta b}_t^{(c)'}$$

$$\Omega_{U,1} = \frac{1}{2} \sum_{h=-\infty}^{\infty} E \left[tr \left\{ \Sigma_{U,t}(h) \Sigma_{U,t}(h)' \right\} \right], \quad \check{\Omega}_{U,2} = \sum_{h=-\infty}^{\infty} E \left[tr \left\{ \Sigma_{U,t}(h) \Delta b_{t-b}^{(c)} \Delta b_t^{(c)'} \right\} \right],$$

$$\Delta b_t = \tilde{b}_{1,t} - \bar{b}_{2,t} - E \left[\left(\tilde{b}_{1,t} - \bar{b}_{2,t} \right) F_t' \right] V(F_t)^{-1} F_t,$$

$$\Sigma_{U,t}(h) = \mu^2 \Sigma_{u,11,t}^{(cc)}(h) + \Sigma_{u,22,t}^{(cc)}(h) - \mu \Sigma_{u,12,t}^{(cc)}(h) - \mu \Sigma_{u,21,t}^{(cc)}(h), \quad h = \dots, -1, 0, 1, \dots,$$

$$\check{\Sigma}_{u,11} = \frac{1}{N_H} \sum_{i=1}^{N_H} \lambda_{1,i} \lambda_{1,i}' \Big)^{-1} \frac{1}{N_H} \sum_{i=1}^{N_H} \sum_{\ell=1}^{N_H} \lambda_{1,i} \lambda_{1,\ell}' Cov \left(\check{e}_t^{H,i}, \check{e}_t^{H,\ell} | \mathcal{F}_t \right) \frac{1}{N_H} \sum_{i=1}^{N_H} \lambda_{1,i} \lambda_{1,i}' \Big)^{-1}$$

and similarly for $\check{\Sigma}_{u,22}$ and $\check{\Sigma}_{u,12}$ using the LF quantities.

The term $b_{2,t}$ is defined in AGGR as $b_{2,t} = \left(\frac{1}{N_L} \sum_{i=1}^{N_L} \lambda_{2,i} \lambda_{2,i}' \right)^{-1} \left(\frac{1}{T} \sum_{t=1}^T b_{2,t} b_{2,t}' \right)^{-1} \eta_{2,t}^2 b_{2,t}$ with $\eta_{2,t}^2 = plim_{N_L \rightarrow \infty} \frac{1}{N_L} \sum_{i=1}^{N_L} E[\check{e}_{2,i,t}^2 | \mathcal{F}_t]$ and $\bar{b}_{2,t} = \Sigma_{\Lambda,2}^{-1} \eta_{2,t}^2 b_{2,t}$ is its large cross-sectional limit, $\tilde{b}_{1,t} = \Sigma_{\Lambda,1}^{-1} E[b_{1,m,t} b_{1,m,t}']^{-1} \frac{1}{M} \sum_{m=1}^M \eta_{1,m,t}^2 b_{1,m,t}$ is the large cross-sectional limit of $\check{b}_{1,t}$ defined in Equation (2.6), and $\Sigma_{\Lambda,1} = \lim_{N_H \rightarrow \infty} \frac{1}{N_H} \sum_{i=1}^{N_H} \lambda_{1,i} \lambda_{1,i}'$ and similarly for $\Sigma_{\Lambda,2}$. The upper index (c) denotes the upper ($k^c, 1$) block of a vector, and the upper index (cc) denotes the upper-left (k^c, k^c) block of a matrix.

The zero-mean terms $u_{j,t}$, for $j=1, 2$, which drive asymptotic normality in the expansions of the PCs estimates are the same in both approaches (1) and (2). Hence, matrix $\Omega_{U,1}$ in the variance of the asymptotic distribution of the test statistic is also the same. Instead, the bias components $b_{1,t}$ and $\tilde{b}_{1,t}$ differ, which explain the different recentering term $\check{\Sigma}_B$ and variance contribution $\check{\Omega}_{U,2}$ when PCA is performed first compared to the result in

6 The remainder term $\check{\vartheta}_{1,t}$ is the flow-sampled value of higher-order terms $\vartheta_{1,m,t}$ in the asymptotic expansion of the PCs in Group 1. A probability bound on its magnitude follows from Proposition 3 in AGGR.

AGGR Theorem 1.⁷ Therefore, the test statistics generally differ depending on whether aggregation or PCA is performed first.

There is an important special case in which the asymptotic distributions of the test statistics in the two approaches coincide. Namely, let us assume that the HF error processes are uncorrelated across individual series and panels, at all leads and lags, and are conditionally homoskedastic martingale difference sequences given the unobservable factors:

$$\begin{aligned} \text{Cov}\left(e_{m,t}^{U,i}, e_{n,t-b}^{V,\ell} \mid \mathcal{F}_t\right) &= 0, \quad \text{if either } U \neq V, \text{ or } i \neq \ell, \\ E\left[e_{m,t}^{U,i} \mid \left\{e_{n,s}^{U,i}\right\}_{n < mVs < t}, \mathcal{F}_t\right] &= 0, \\ E\left[\left(e_{m,t}^{U,i}\right)^2 \mid \left\{e_{n,s}^{U,i}\right\}_{n < mVs < t}, \mathcal{F}_t\right] &= \gamma_{U,i} \quad (\text{say}), \end{aligned} \tag{2.8}$$

where $e_{t-1+m/M}^{U,i} \equiv e_{m,t}^{U,i}$, for $U = H, L$, where $\left\{e_{n,s}^{U,i}\right\}_{n < mVs < t}$ consists of all error terms previous to subperiod m of date t . Then, $\Omega_{U,2} = \tilde{\Omega}_{U,2} = 0$ and $\tilde{\Sigma}_B = \check{\Sigma}_B = 0$ in both approaches, and the test statistic under the null hypothesis $H(k^C)$ is such that:

$$N_L \sqrt{T} \left(\frac{1}{2} \text{tr} \left\{ \Sigma_U^2 \right\} \right)^{-1/2} \left[\tilde{\xi}(k^C) - k^C + \frac{1}{2N_L} \text{tr} \left\{ \tilde{\Sigma}_{cc}^{-1} \tilde{\Sigma}_U \right\} \right] \xrightarrow{d} N(0, 1), \tag{2.9}$$

where $\tilde{\Sigma}_{u,11} = M \left(\frac{1}{N_H} \sum_{i=1}^{N_H} \lambda_{1,i} \lambda'_{1,i} \right)^{-1} \left(\frac{1}{N_H} \sum_{i=1}^{N_H} \lambda_{1,i} \lambda'_{1,i} \gamma_{H,i} \right) \left(\frac{1}{N_H} \sum_{i=1}^{N_H} \lambda_{1,i} \lambda'_{1,i} \right)^{-1}$ and similarly for $\tilde{\Sigma}_{u,22}$ using the corresponding LF quantities, $\tilde{\Sigma}_U = (N_L/N_H) \tilde{\Sigma}_{u,11} + \tilde{\Sigma}_{u,22}$, $\tilde{\Sigma}_{cc} = \frac{1}{T} \sum_{t=1}^T \tilde{g}_t^C \tilde{g}_t^{C'}$, and Σ_U being the large sample limit of $\tilde{\Sigma}_U$ as $N_H, N_L \rightarrow \infty$. The same distributional result as Equation (2.9) holds for $\tilde{\xi}(k^C)$. Even if the recentring and rescaling terms in the asymptotic distribution are the same whenever aggregation or PCA is performed first, the test statistic values $\tilde{\xi}(k^C)$ and $\check{\xi}(k^C)$ in the two approaches differ, because the canonical correlations estimates differ.

2.3 Mixed-Frequency Factor Estimation

In this subsection, we consider the estimation of the factor values and focus in particular on the asymptotic distribution of the estimator of the common factor. Building on the previous subsection, there are two approaches depending on whether aggregation (flow sampling) is performed before or after PCA. To simplify the exposition, we focus on the case $k^C = 1$.

(i) When data are flow-sampled first, the estimator of the LF values of the common factor is

$$\hat{g}_t^{C*} = S(\omega) \left(\hat{g}_t^C + \omega \hat{g}_t^{C*} \right), \tag{2.10}$$

where $S(\omega) = (1 + \omega^2 + 2\omega\hat{\rho}_1)^{-1/2}$ and ω is the weight, with $\hat{g}_t^C = \hat{W}'_1 \hat{h}_{1,t}$ and $\hat{g}_t^{C*} = \hat{W}'_2 \hat{h}_{2,t}$. Then, we use the residuals from \hat{g}_t^{C*} to estimate the specific factors \hat{g}_t^H and \hat{g}_t^L , and the common and group-specific factor loadings to get $\hat{\Lambda}_1 = [\hat{\Lambda}_{HC} : \hat{\Lambda}_H]$, $\hat{\Lambda}_2 = [\hat{\Lambda}_{LC} : \hat{\Lambda}_L]$ in analogy with the procedure introduced in Section 1

7 Vectors $d_{1,t}$ and $\check{d}_{1,t}$ differ, but are both a (stochastic, asymptotically nonsingular) linear transformation of vector $h_{1,t}$, and therefore their contribution to the test statistic is asymptotically negligible. Further, matrices \mathcal{H}_1 and $\check{\mathcal{H}}_1$ that correspond to rotations of the factor estimates, also differ in approaches (1) and (2), but they are immaterial for the values of estimators and test statistics.

for the general group-factor framework. Finally, we estimate the HFF values $\hat{g}_{m,t}^C$ and $\hat{g}_{m,t}^H$ from the cross-sectional regression of $x_{m,t}^H$ onto the estimated loadings:

$$\begin{bmatrix} \hat{g}_{m,t}^C \\ \hat{g}_{m,t}^H \end{bmatrix} = \left(\hat{\Lambda}'_1 \hat{\Lambda}_1 \right)^{-1} \hat{\Lambda}'_1 x_{m,t}^H, \quad m = 1, \dots, M, \quad t = 1, \dots, T.$$

(ii) When PCA is performed first, we have HF estimates $\check{g}_{m,t}^C = \check{W}'_1 \check{b}_{1,m,t}$, for $m = 1, \dots, M$, and LF estimates $\check{g}_t^{C*} = \check{W}'_2 \check{b}_{2,t}$, of the common factor values. We follow the principle of linear combination to obtain another and possibly more efficient (in a sense to be defined below) HF estimator of the common factor. Consider the estimation of the common factor value for subperiod m . The linear combination of $\check{g}_{n,t}^C$, $n = 1, \dots, M$, and \check{g}_t^{C*} , which yields an asymptotically unbiased estimator of $g_{m,t}^C$, has the form

$$\check{g}_{m,t}^{C*} = \check{g}_{m,t}^C + \alpha_m \left(\check{g}_t^{C*} - \check{g}_t^C \right), \tag{2.11}$$

for a coefficient α_m , up to a standardization to impose unit sample variance, where $\check{g}_t^C = \sum_{n=1}^M \check{g}_{n,t}^C = \check{W}'_1 \check{b}_{1,t}$. When we aggregate across subperiods, we get $\check{g}_t^{C*} := \sum_{m=1}^M \check{g}_{m,t}^{C*} = \alpha \check{g}_t^C + (1 - \alpha) \check{g}_t^{C*}$, where $\alpha := 1 - \sum_{m=1}^M \alpha_m$. Therefore, for the flow-sampled estimators, we get a linear combination analogous to Equation (2.10) with relative weight:

$$\omega = (1 - \alpha) / \alpha. \tag{2.12}$$

Finally, we use the residuals from $\check{g}_{m,t}^{C*}$ on the HF panel, and the residuals of \check{g}_t^{C*} on the LF panel, to extract the group-specific factor estimates $\check{g}_{m,t}^H$ and \check{g}_t^L .

How to choose the weights? If we mimic the Goyal, Pérignon, and Villa (2008) estimator in our mixed-frequency setting, we chose $\omega = 1$ in (2.10) for the aggregation-first estimator. To have an analogous choice for the PCA-first estimator, Equation (2.12) suggests to have $\alpha = 1/2$. Imposing equal weights across subperiods as the simplest option, this yields $\alpha_m = (2M)^{-1}$ in Equation (2.11).

Another approach to the determination of the weights consists of minimizing the asymptotic MSE of the estimators. For this purpose, we might consider either the flow-sampled estimates \hat{g}_t^{C*} and \check{g}_t^{C*} or the HF estimates $\hat{g}_{m,t}^{C*}$ and $\check{g}_{m,t}^{C*}$. In this article, we consider the former option and assume that the HF weights $\alpha_m = \frac{1}{M}(1 - \alpha) = \frac{1}{M} \frac{\omega}{1 + \omega}$ for the PCA-first estimator are homogenous across subperiods. The latter option in a general framework is more challenging and is left for future research. Note that we impose Equation (2.8) to derive the results below. In Online Appendix, we derive the asymptotic Gaussian distributions of the flow-sampled estimators:

$$\sqrt{N} \left(\hat{\mathcal{H}}_C^* \hat{g}_t^{C*} - \bar{g}_t^C - \frac{1}{T} \bar{\beta}_t^{C*}(\omega) \right) \Rightarrow N(0, V^*(\omega)),$$

and

$$\sqrt{N} \left(\hat{\mathcal{H}}_C^* \check{g}_t^{C*} - \bar{g}_t^C - \frac{1}{T} \tilde{\beta}_t^{C*}(\omega) \right) \Rightarrow N(0, V^*(\omega)),$$

where $\tilde{\beta}_t^{C*}(\omega) = \frac{1}{1+\omega}(\tilde{\beta}_{1,t}^C + \omega\tilde{\beta}_{2,t}^C)$, $\tilde{\beta}_t^{C*}(\omega) = \frac{1}{1+\omega}(\tilde{\beta}_{1,t}^C + \omega\tilde{\beta}_{2,t}^C)$ and $V^*(\omega) = \frac{1}{(1+\omega)^2}(\omega^2\mu\Sigma_{u,11}^{(cc)} + \Sigma_{u,22}^{(cc)})$, with $\tilde{\beta}_{1,t}^C = M\tilde{\gamma}_H(\Sigma_{\Lambda,1}^{-1}b_{1,t})^{(c)}$, $\tilde{\beta}_{2,t}^C = M\tilde{\gamma}_L(\Sigma_{\Lambda,2}^{-1}b_{2,t})^{(c)}$, $\tilde{\beta}_{1,t}^C = \frac{1}{M}\tilde{\gamma}_H(\Sigma_{\Lambda,1}^{-1}E[b_{1,m,t}b'_{1,m,t}]^{-1}b_{1,t})^{(c)}$, and $\Sigma_{u,ji}^{(cc)} = M(\Sigma_{\Lambda,j}^{-1}\Omega_{\Lambda,j}\Sigma_{\Lambda,j}^{-1})^{(cc)}$, $j = 1, 2$, $\Sigma_{\Lambda,1} = \lim_{N_H \rightarrow \infty} \frac{1}{N_H} \sum_{i=1}^{N_H} \lambda_{1,i}\lambda'_{1,i}$ and $\Omega_{\Lambda,1} = \lim_{N_H \rightarrow \infty} \frac{1}{N_H} \sum_{i=1}^{N_H} \gamma_{H,i}\lambda_{1,i}\lambda'_{1,i}$, and similarly for $j = 2$ on the LF panel. The flow-sampled estimators with aggregation first or PCA first have the same asymptotic variance (for given ω) and different asymptotic bias at order $1/T$. The asymptotic bias is negligible if $N/T^2 = o(1)$, and the two approaches PCA first or last are asymptotically equivalent.⁸ The average AMSE of the aggregation-first estimator is

$$\frac{1}{T^2} E \left[\tilde{\beta}_t^{C*}(\omega)^2 \right] + \frac{1}{N} V^*(\omega) = \frac{1}{(1+\omega)^2} \left[\frac{1}{T^2} (B_{11} + \omega^2 B_{22} + 2\omega B_{12}) + \frac{1}{N} (\mu\Sigma_{u,11}^{(cc)} + \omega^2\Sigma_{u,22}^{(cc)}) \right] \tag{2.13}$$

where $B_{11} = M^2\tilde{\gamma}_H^2[\Sigma_{\Lambda,1}^{-2}]^{(cc)}$ and similarly for B_{22} , and $B_{12} = M^2\tilde{\gamma}_H\tilde{\gamma}_L[\Sigma_{\Lambda,1}^{-1}V_{12}\Sigma_{\Lambda,2}^{-1}]^{(cc)}$. It is minimized for

$$\omega = \frac{\frac{1}{N_H}\Sigma_{u,11}^{(cc)} + \frac{1}{T^2}(B_{11} - B_{12})}{\frac{1}{N_L}\Sigma_{u,22}^{(cc)} + \frac{1}{T^2}(B_{22} - B_{12})}. \tag{2.14}$$

For the PCA-first estimator, we get similar formulas with B_{11} and B_{12} replaced by $\tilde{B}_{11} = \frac{1}{M^2}\tilde{\gamma}_H^2[\Sigma_{\Lambda,1}^{-1}E(b_{1,m,t}b'_{1,m,t})^{-2}\Sigma_{\Lambda,1}^{-1}]^{(cc)}$ and $\tilde{B}_{12} = \tilde{\gamma}_H\tilde{\gamma}_L[\Sigma_{\Lambda,1}^{-1}E(b_{1,m,t}b'_{1,m,t})^{-1}V_{12}\Sigma_{\Lambda,2}^{-1}]^{(cc)}$, respectively.

We can compare the AMSE of the aggregation-first and PCA-first estimators for specific DGPs. As in the Monte Carlo experiments in AGGR, let us assume that the HF dynamics of the latent factors is given by the VAR(1) process $g_{m,t} = a_F g_{m-1,t} + \sqrt{\zeta}\eta_{m,t}$ with common AR coefficient a_F , where $g_{m,t} = (g_{m,t}^C, g_{m,t}^H, g_{m,t}^L)'$ is the stacked factor vector, $\eta_{m,t} = (\eta_{m,t}^C, \eta_{m,t}^H, \eta_{m,t}^L)'$ \sim IID(0, Σ_η), matrix Σ_η has identity matrices as diagonal blocks, $Cov(\eta_{m,t}^H, \eta_{m,t}^L) = \Phi$, and zero elements elsewhere. The scale of the innovation variance is $\zeta = \frac{1-a_F^2}{M^2\kappa}$ with $\kappa = 1 - \frac{2}{M^2} \sum_{m=1}^M m(1 - a_F^{M-m})$ to ensure the standardization $V(b_{j,t}) = I_{k_j}$. Then, we have $E[b_{1,m,t}b'_{1,m,t}] = \frac{1}{M^2\kappa}I_{k_1}$, which yields $\tilde{B}_{11} = \kappa B_{11}$ and $\tilde{B}_{12} = \kappa B_{12}$ with $\kappa < 1$. Therefore, from Equation (2.13), we can see that for this DGP the PCA-first estimator has smaller asymptotic bias and average AMSE for any given choice of weight $\omega > 0$ (the same in both approaches), such as $\omega = 1$ for the analogue of the Goyal, Pérignon, and Villa (2008) estimator, as well as for the optimal choices of the weights (as long as these are positive, and $B_{12} > 0$).

3 Empirical Analysis

3.1 Data Description

We employ two large panels/groups of variables available at two different sampling frequencies. The first panel comprises $N_L = 188$ LF, quarterly U.S. macroeconomic variables while the second panel comprises $N_H = 116$ HF, monthly financial variables.⁹ The choice

- 8 This corroborates the findings in our empirical analysis, where we have $T = 218$ and $N = 116$ and we find that the two approaches yield very similar estimates.
- 9 Note that therefore in the empirical analysis $N_H < N_L$, but this does not matter since the large sample results can be derived in this case by interchanging the roles of N_H and N_L .

of quarterly frequency for the macro data is based on maximizing N_L in this group to incorporate important indicators from the National Income and Product Accounts related to GDP, government expenditure, investment, among others. Similarly, the choice of the monthly frequency is constrained by the trade-off between increasing N_H aiming to enlarge the cross section of financial variables to include, for example, interest rates and credit spreads, and at the same time covering a long-span of time series, T . While we acknowledge that many financial series are available at a much higher frequency (e.g., daily and/or intra-daily), this choice would compromise both N and T as many of these higher frequencies series are not available since the early 1960s and this would challenge our inferential framework which is based on large N and T .

The time series period is 1963m7–2017m12, with $T=218$ quarterly and $TM = 654$ monthly observations. The macro panel is based on the quarterly macro indicators of FRED-QD (McCracken and Ng, 2016).¹⁰ The financial panel includes the following financial indicator categories: (i) Interest Rates, (ii) Stock Markets, (iii) Exchange Rates, (iv) Soft (a) and Hard (b) Commodities.¹¹ All variables are transformed to represent stationary variables and each series is demeaned and standardized in the panel following either FRED-QD or the corresponding transformations for stationarity in Stock and Watson (2002) and Brave and Butters (2014).

We also consider the following well-known factors in the literature extracted from different but related panels, such as the ADS factor of Aruoba, Diebold, and Scotti (2009) measuring real business conditions and based on a mixed-frequency small panel, the Chicago Federal National Activity Index (CFNAI) and the National Financial Conditions Indicator (NFCI; Brave and Butters, 2014) extracted from larger panels, as well as the credit spreads index of Gilchrist and Zakrajsek (2012) available from authors.

The role of the aforementioned factors from the literature, as well as our mixed-frequency group factors, namely the common and group-specific factors, are further investigated in predicting key macroeconomic as well as financial indicators such as real GDP and consumption of services and nondurable goods growth, the Moody's corporate bond default spread, the CBOE's VIX also referred to as the "fear index", the VRP (available from Zhou, 2018), as well as the ETF iShares Core S&P500 Index.

3.2 Extracting the Common and Group-Specific Factors

Within the mixed-frequency group factor model comprising U.S. quarterly (Low) frequency (LF) macroeconomic indicators and monthly (High) frequency (HF) financial variables, we investigate whether there is a CF spanning these two panels, as well as group-specific

- 10 The panel includes the following eleven categories of variables: (i) National Income and Product (NIPA), (ii) Industrial Production, (iii) Employment and Unemployment, (iv) Housing, (v) Inventories, Orders and Sales, (vi) Prices, (vii) Earnings and Productivity, (viii) Money and Credit, (ix) Household Balance Sheet, (x) Consumer Expectations, and (xi) Nonhousehold balance sheet. The macro variables in each category are listed in [Online Appendix Table OA.1](#). Our macro panel excludes the following FRED-QD categories: Exchange Rates, Interest Rates, and Stock Markets, since most of these variables are available at monthly frequency and belong to the financial panel.
- 11 The financial variables in each category and the corresponding data sources are listed in [Online Appendix Table OA.2](#).

Financial/HF and Macro/LFFs, HFF, and LFF, respectively. Employing the methods developed in AGGR and further expanded in Sections 1 and 2, we find that although there is no common factor in the United States during the full sample period from 1963 to 2017, there is however, evidence of a single CF during the pre- and post-GM periods. These results are presented in Tables 1 and 2 which report the estimated number of pervasive factors in the HF and LF panels, as well as the canonical correlations and test statistics for the common factors, respectively. Following the analysis in the previous section, we apply the CF test using the PCA approach first as well as PCA last (i.e., aggregation first) to examine how inferences related to the number of factors and the CF test is affected. It is worth mentioning at the outset that we find that these two approaches of estimating factors yield very similar results for this empirical application.

3.2.1 Group specific factors

We start by selecting the number of factors in each subpanel and each subperiod. In Table 1, we report the results for the IC_{p2} information criterion of Bai and Ng (2002); similar results apply for the IC_{p1} and we choose the maximum number of factors (k_{\max}) equal to ten in order to avoid excluding potentially important factors from the panels.¹² The IC_{p1} and IC_{p2} dominate the other criteria in Bai and Ng (2002). We focus the discussion on the number of factors in each subperiod (pre- and post-GM), given that we find one common factor in each of these regimes. For the panel/group of financial variables at monthly frequency (x^H) and quarterly frequency (\bar{x}^H), the IC_{p2} criterion, during the pre-GM period, selects six factors for x^H and eight factors for \bar{x}^H . In contrast, during the post-GM period, IC_{p2} selects nine factors for the x^H and ten factors for \bar{x}^H . On the contrary, for the quarterly macro variables panel/group, IC_{p2} selects five and six factors in the pre- and post-GM periods, respectively. The inference on the number of factors is robust to the PCA first or last approach, as shown in Table 1. Last but not least, we compare our inference on the number of group/frequency specific factors with the traditional approach of applying the IC_{p2} to a single panel with a common low (quarterly) frequency which stacks all the variables together, denoted by $[\bar{x}^H, \bar{x}^L]$ in Table 1. In the latter case, the IC_{p2} chooses seven factors in the pre-GM vis-à-vis nine factors in the post-GM. In our subsequent empirical applications, we proceed with the aggregation first approach (also followed in the empirical analysis of AGGR).

Most criteria for factor selection, including the IC_{p2} , choose factors in an unconditional setup, that is, without conditioning on the variable(s) of interest that the factors aim to explain or forecast. Moreover, from the total number of factors chosen from the panels and subperiods, it is expected that different factors will have varying explanatory power for different dependent variables of interest (e.g., macro or financial) and for alternative subperiods. Hence given that we are interested in the role of these factors in a conditional setup, our empirical analysis considers the above number of factors for each subpanel and regime in order to avoid any omitted factors/variables (hence the choice of $k_{\max} = 10$) in explaining key macro and financial variables. Subsequently, we reassess the conditional

12 Similar results apply to $k_{\max} = 8$ found in Online Appendix Tables OA.3 and OA.4. Note that for smaller values of $k_{\max} < 6$, we ignore some of the estimated factors in each subpanel and regime vis-à-vis $k_{\max} = 8$ or 10 which also turn out to be significant in the conditional setup for explaining key macro and financial variables, as discussed in the next subsection.

Table 1 Estimated number of pervasive factors in HF and LF panels

		Full Sample				Pre-GM				Post-GM			
		x^H	\bar{x}^H	\bar{x}^L	$[\bar{x}^H \bar{x}^L]$	x^H	\bar{x}^H	\bar{x}^L	$[\bar{x}^H \bar{x}^L]$	x^H	\bar{x}^H	\bar{x}^L	$[\bar{x}^H \bar{x}^L]$
IC_{p2}	Aggregation first	-	10	8	10	-	8	5	7	-	10	6	9
	PCA first	8	-			6	-			9	-		

Table 2 Canonical correlations and test statistics for common factors

Aggregation first/PCA last

Full sample			Pre-GM			Post-GM		
$(cv = -2.0003 \Rightarrow k^c = 0)$			$(cv = -1.9048 \Rightarrow k^c = 1)$			$(cv = -1.9477 \Rightarrow k^c = 1)$		
i	$\hat{\rho}_i$	$\tilde{\xi}(i)$	i	$\hat{\rho}_i$	$\tilde{\xi}(i)$	i	$\hat{\rho}_i$	$\tilde{\xi}(i)$
1	0.839	-5.139	1	0.913	0.376	1	0.911	-0.187
2	0.782	-4.027	2	0.800	-4.015	2	0.812	-4.842
3	0.715	-7.211	3	0.696	-8.325	3	0.701	-8.721
4	0.590	-9.171	4	0.25	-8.987	4	0.545	-8.303
5	0.378	-12.51	5	0.150	-9.436	5	0.376	-7.872
6	0.170	-14.04				6	0.144	-3.731
7	0.034	-8.384						
8	0.025	-8.300						

PCA first/aggregation last

FULL SAMPLE			Pre-GM			Post-GM		
$(cv = -2.0003 \Rightarrow k^c = 0)$			$(cv = -1.9048 \Rightarrow k^c = 1)$			$(cv = -1.9477 \Rightarrow k^c = 1)$		
i	$\hat{\rho}_i$	$\tilde{\xi}(i)$	i	$\hat{\rho}_i$	$\tilde{\xi}(i)$	i	$\hat{\rho}_i$	$\tilde{\xi}(i)$
1	0.834	-4.567	1	0.894	-1.098	1	0.906	-0.195
2	0.701	-10.640	2	0.722	-6.721	2	0.786	-5.852
3	0.603	-11.000	3	0.626	-11.000	3	0.670	-9.609
4	0.550	-12.850	4	0.241	-8.714	4	0.434	-13.370
5	0.258	-13.710	5	0.228	-7.521	5	0.356	-8.260
6	0.182	-14.180				6	0.200	-4.103
7	0.080	-9.204						
8	0.038	-6.405						

Notes: \bar{x}^H is the (T, N_H) panel of the quarterly data computed as the sum of the HF monthly (TM, N_H) panel data, x^H , and \bar{x}^L is the (T, N_L) panel of the LF quarterly data and $k_{max} = 10$. The number of observations is given by $N_H = 116$ for monthly financial variables, $N_L = 188$ for quarterly macroeconomic variables, $T_{post} = 128$ during the post-GM period (1986q1–2017q4), $T_{pre} = 82$ during the pre-GM period (1963q3–1983q4), $T_{full} = 218$ during the full sample period (1963q3–2017q4). $\hat{\rho}_i$ and $\tilde{\xi}(i)$ refer to the canonical correlation and test static of the i common factor, respectively. k^c is the estimated number of common factors defined as $k^c = \max\{i : 1 \leq i \leq k_{max}, \tilde{\xi}(i) \geq cv\}$, where cv refers to the critical value reported above which is defined in AGGR as $-c(N\sqrt{T})^\gamma$ with $c = 0.95$ and $\gamma = 0.1$.

significance of factors using both IS and OOS criteria such as the goodness-of-fit, significance/thresholding for targeted predictors (Bai and Ng, 2008) and testing based on mean squared forecasting error criteria, discussed in Subsection 3.3.

3.2.2 Common factor

The estimated canonical correlations in each of the two subpanels of LF and HF data and the test statistics, reported in Table 2, provide evidence that there is one common factor in the two subperiods, before and after the mid-1980s. The inference on a single common factor in these two regimes is also robust whether we apply the PCA first or last approach, as shown by the two panels in Table 2. Note also that while results reported in Tables 1 and 2 refer to 1984q1 being the change point, as reported in Stock and Watson (2008), the results on the existence of one common factor are robust to other break dates in the mid-1980s during the period 1984q1–1985q4, which is also consistent with other studies in the literature.¹³ Given that the inference on a single CF for these two regimes is robust following the two approaches, PCA first or last, we proceed to compare the actual PC estimates from these two approaches shown in Figure 1a and b, for the pre- and post-GM periods, respectively. The CFs estimated following the two approaches are very closely correlated as shown by the two PCs which are almost superimposed in Figure 1a and b, with the correlation of the factors from PCA first and last being 0.95 and 0.98, during the pre- and post-GM regimes, respectively. Moreover, the persistence of the CFs as measured by the simple AR(1) coefficient is estimated to be 0.79 (and 0.91) for PCA first and 0.88 (and 0.93) for the PCA last for the monthly CF in the first regime (and in the second regime). Further evidence on the factor estimates obtained from the two approaches is provided in Online Appendix Table OA.6 which reports the correlation matrices of all the factors showing that the corresponding PCs (from aggregation or PCA first) yield correlations of 0.94–1.00.

In Table 3, we provide additional evidence which shows the alternative CF estimation methods discussed in Sections 1 and 2, focusing on the post-GM period. The common factor estimators CF_1 , CF_2 , CF_3 , and CF_4 are based on Equation (2.10) with $\omega = 0$, $\omega = +\infty$, $\omega = 1$, and ω as specified in Equation (2.14), respectively. These results show that not only the correlations of the alternative CFs are very high across the different estimation types (reported in Table 3), but also the time series behavior of these CF estimates is almost identical as shown in Figure 2. Hence, in the subsequent analysis, we use the third estimation type, CF_3 , that is, Equation (2.10) with $\omega = 1$. Moreover, given the empirical evidence that the two approaches (PCA last or first) yield almost identical factor estimates we proceed with one of them namely aggregation first/PCA last, not only for conciseness in reporting results, but also because the aggregation first approach is more comparable to the common frequency single panel PCA approach (according to which all data are aggregated to a

13 Enlarging the panel to include other financial indicators such as the Fama–French portfolios related to the forty-nine industries and 100 portfolios sorted on size and book to market, we find no common factor during the full sample and the two subperiods. This evidence suggests that a financial panel dominated by these U.S. portfolio-type stock market variables may mask the existence of a common factor between the macro and financial panels and other financial indicators (including the stock market indices). Further evidence related to the role of specific stock market variables (e.g., the VXO) in driving the common factor is provided below, related to the changing structure of the CF during the pre- and post-GM periods.

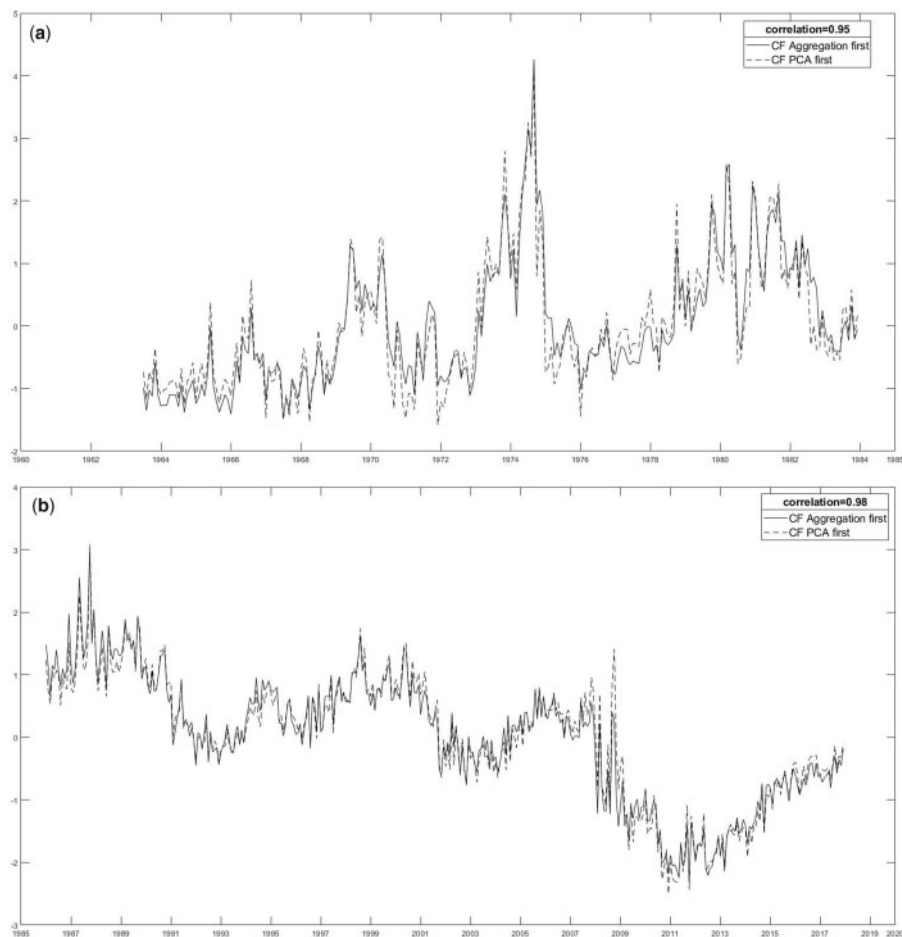


Figure 1 Common factor (CF) estimates using PCA first and last approaches during the (a) pre-GM and (b) post-GM.

common LF and all variables are collapsed to a common panel in order to estimate the PCA at the end). This latter approach would be denoted as the common frequency factors (CFFs).

The evidence of a CF during the pre- and post-GM is further investigated by testing for a structural change in the loadings of the CF (as well as the HFFs and the LFFs) in the mid-1980s, which seems to affect the inference on the existence of a common factor during the full sample period, 1963–2017. If factor loadings have a break which is not only small, but also the change point is sufficiently independent across the cross section of time series used to estimate the factors, then its effect is averaged out across the many series in the panel and the PCs estimates are not affected (e.g., [Stock and Watson, 2002](#)). Applying the LM (resp. supLM) test for a break in the loadings of factor models proposed by [Breitung and](#)

Table 3 Correlation coefficients between common factor (CF) estimation types during the post-GM

	CF ₁	CF ₂	CF ₃	CF ₄
Panel A: Correlations of HF CFs				
CF ₁	1	0.888	0.997	1
CF ₂	0.888	1	0.915	0.894
CF ₃	0.997	0.915	1	0.998
CF ₄	1	0.894	0.998	1
Panel B: Correlations of LF CFs:				
CF ₁	1	0.927	0.998	1
CF ₂	0.927	1	0.942	0.931
CF ₃	0.998	0.942	1	0.999
CF ₄	1	0.931	0.999	1

Notes: The common factor estimators CF₁, CF₂, CF₃, and CF₄ based on Equation (2.10) with $\omega = 0$, $\omega = +\infty$, $\omega = 1$, and ω as specified in Equation (2.14), respectively.

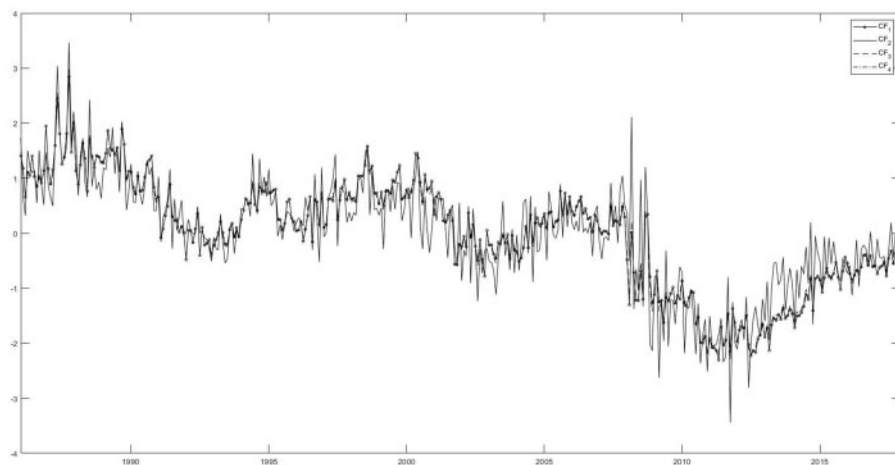


Figure 2 Common factor (CF) estimation types CF₁, CF₂, CF₃, and CF₄ during the post-GM.

Eickmeier (2011), we report empirical evidence in Table 4 that in the mid-1980s, 50% (resp. 73%) of the CF loadings associated with the LF/macro series change, whereas 55% (resp. 67%) of the HF/financial loadings in the CF change, as shown by the total percentage of rejections of the null (no break) hypothesis. Note that applying the Chow test for breaks in factor loadings proposed in Stock and Watson (2008), we also find evidence of breaks.¹⁴ Yet, given that the LM test for known break and supLM for unknown breaks are valid under more general assumptions and exhibit better finite sample properties (Breitung and Eickmeier, 2011), we focus our discussion on the LM-type tests. Within the cross section,

14 We present only the results for the Breitung and Eickmeier (2011) supLM test. The results of the Chow test proposed by Stock and Watson (2008) are found in Online Appendix Table OA.5.

Table 4 LM-type tests for the GM structural break in the loadings of dynamic factor models

	CF (Quarterly loadings) (%)	CF (Monthly loadings) (%)	CFF1 (%)	CFF2 (%)	CFF1, CFF2 (%)	CFFs (IC_{p2}) (%)	CF, HFFs, LFFs (%)
LM test (Breitung and Eickmeier, 2011)							
Total	50.0	55.2	46.1	60.5	68.4	89.5	99.0
Interest rates		67	93	74	96	100	100
U.S. stock market indices		69	62	69	77	100	100
Exchange rates		64	4	28	20	96	100
Commodities		41	22	14	22	49	94
NIPA	59		82	100	95	100	100
IP	60		93	80	93	100	100
Employment	70		66	89	98	100	100
Housing	100		0	100	100	100	100
Inventories, orders, and sales	50		83	100	100	100	100
Prices	9		24	33	46	91	100
Earnings and Productivity	40		50	60	90	100	100
Money and Credit	62		38	77	77	92	100
Household Balance Sheets	78		33	100	100	100	100
Consumer Expectations	100		0	100	100	100	100
Non-Household Balance Sheets	27		45	91	100	100	100
supLM test (Breitung and Eickmeier, 2011)							
Total	73.4	67.2	50.7	71.7	74.7	96.7	98.7
Interest rates		96	96	85	100	100	100
U.S. stock market indices		92	69	85	85	100	100
Exchange rates		60	8	36	36	100	100
Commodities		49	22	33	29	84	92
NIPA	91		82	100	100	100	100
IP	100		93	87	100	100	100
Employment	89		70	93	100	100	100
Housing	100		9	100	100	100	100
Inventories, orders, and sales	100		83	100	100	100	100
Prices	35		24	52	57	98	100
Earnings and Productivity	70		90	100	100	100	100
Money and Credit	69		46	77	77	92	100
Household Balance Sheets	89		44	100	100	100	100
Consumer Expectations	100		0	100	100	100	100
Non-Household Balance Sheets	55		64	100	100	100	100

Notes: The reported values refer to the percentage of rejections of the null hypothesis of no breaks in the loadings. The above results refer to the case of estimating mixed-frequency group factors (CF, HFFs, and LFFs) with aggregation first (i.e., PCA last) which is more comparable to the CFFs. Same results apply to the case of estimating the factors with the PCA first approach. The reported results for the LM test refer to the subsample 1986q1–2017q4. Results are robust to other known break dates, namely 1984q4 as well as 1985q4. CF refers to the common factor from the MFF model. CFF extracted from the stacked panel of all low/quarterly frequency variables. The details of the variable categories and the variable definitions are found in [Online Appendix](#). Bold values indicate the percentage of rejections for the total number of variables for the null hypothesis of no structural break of the LM test (for 5% significance level).

there is strong evidence that the GM is associated with changes in the CF loadings of all variable categories in the two groups, macro and financial. Moreover, there is evidence that a large percentage of CF loadings change within many variables categories. In contrast, the changes in the loadings appear to be relatively smaller (than the total change in the loadings) for the consumer/producer prices and non-household balance sheet indicators (in the macro panel), as well as for exchange rates and commodity prices (in the financial panel). The supLM test results show strong evidence that approximately more than 90% of the loadings of the CF related to the individual series in the following categories change due to the GM: interest rates and stock market return indices in monthly financial panel, as well as National Income and Product Accounts (NIPA), IP, Employment, Housing, Inventories, Orders and Sales, Household Balance Sheet, and Consumer Expectations in the quarterly macro panel. Similarly, the correlation of the loadings of the CF in these two regimes, marked by the GM, is quite small, for both the HF (0.09) and LF (0.12) series loadings. Last but not least, the results on the structural break analysis and the percentage changes in the loadings of different series categories are the same whether we apply the PCA first or last approach.

The time series behavior of the estimated CF during the pre- and post-GM as well as the corresponding full sample CF is presented in Figure 3. The dashed and dotted lines refer to the CF in the aforementioned regimes vis-à-vis the solid line which refers to the full sample CF. The CFs in Figure 3 present at least three interesting features of the macro–finance factor. First, there is a shift in the mean of the estimated PCs in the two regimes and ignoring the break seems to overestimate the mean of the full sample CF in the early 1960s and underestimate it in the mid-1980s, as shown by comparing the solid, dashed, and dotted lines representing the CFs in the three periods. Interpreting the PCs in Figure 3 as the CFs we find that the GM has caused an increase in the mean of this factor, conditional on the two regimes, as shown by the relatively higher mean during the mid-1980s until the early 2000 (compared to the mean of the CF shown by the dashed line). Second, the CF in the post-GM period has a strong cyclical behavior (vis-à-vis that in the first regime), suggesting that during the recent period the CF is dominated by the behavior of business cycle macro series as well as financial cycle-related series, as opposed to that of financial asset returns series (such as FX and stock market returns). Hence, our analysis provides additional and complementary evidence in the literature related to the GM structural break in the loadings of factor models, demonstrating how this has affected the loadings of the U.S. CF as well as the inference and behavior of this common component, while allowing us to study the behavior of group-specific (financial or macro) factors jointly. Third, we observe that during the U.S. NBER recession dates marked by the grey areas in Figure 3, the CF in most cases exhibits relative peaks associated, for example, with the recent global financial crisis in 2007–2008, the dotcom bubble in 2001, the banking strains in early 1990s, and the two oil crises in the mid-1970s and early 1980s, followed by downturns after each crisis/recession. In Figure 4, we relate our CF during the post-GM period with the U.S. business cycle and financial cycle of Drehmann, Borio, and Tsatsaronis (2012) and observe that our CF is dominated by long cycles similar to those of the financial cycle in the 1980s and during 2000–2017.¹⁵ The financial cycle has 0.45 correlation with the CF in the post-GM regime as

15 The Drehmann, Borio, and Tsatsaronis (2012) financial cycle is a frequency-based (band-pass) filter capturing medium-term cycles using five financial variables: credit to private and non-financial

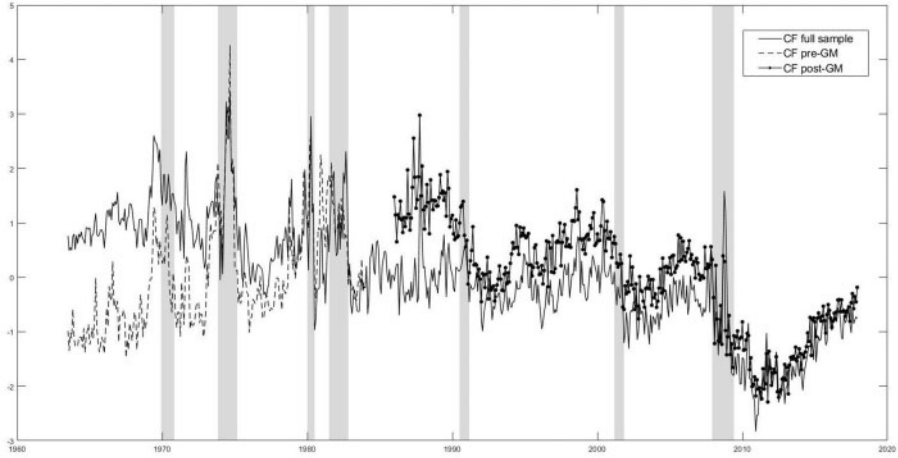


Figure 3 Common factor (CF) and NBER recessions during the full period (1963m07–2017m12) and during the pre- and post-GM (1963m07–1983m12 and 1986m01–2017m12, respectively) using the PCA last approach.

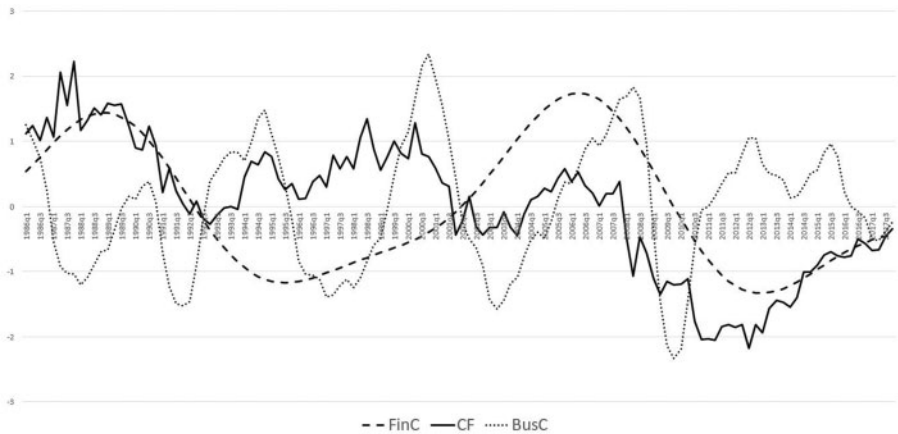


Figure 4 The Business Cycle (BusC), the Financial Cycle (FinC), and the CF during the post-GM period.

opposed to the business cycle with which correlation is very low, that is, 0.04 (reported later in the last two rows of Table 5).

The changing structure of the estimated macro–finance common factor in the two regimes is further investigated by examining the different categories that drive the CFs, as well as the HFFs, LFFs, and CFFs, during the two regimes. The categories of financial and macro variables and their R^2 are reported in Table 6. Further details of the specific

sector, the ratio of credit to GDP, equity prices, residential property prices, and an index of aggregate asset prices including residential and commercial property and equity prices. Similarly, the business cycle is a band-pass filter capturing fluctuations in real GDP over a period of one to eight years.

Table 5 Correlation Matrices of Factors and Variables during the pre- and post-GM period

Panel A: Correlation matrix of quarterly CF, HFFs, LFFs from mixed-frequencies group factor model and the CFF with other factors and variables in the pre-GM period

	Mixed-frequency group factors															CFFs						
	CF	HF1	HF2	HF3	HF4	HF5	LF1	LF2	LF3	LF4	LF5	CFF1	CFF2	CFF3	CFF4	CFF5	CFF6	CFF7				
GDP	0.52	0.20	-0.26	0.06	0.01	-0.11	0.55	-0.34	-0.18	0.35	-0.53	-0.27	0.46	-0.07	-0.15	0.15	-0.03					
RealCons	0.66	0.09	-0.09	0.05	0.11	0.03	0.13	-0.18	-0.17	-0.08	-0.65	0.01	0.32	0.06	-0.05	-0.03	-0.03					
Baa-Aaa	-0.35	-0.26	0.45	-0.08	-0.19	0.24	-0.30	0.43	-0.03	0.28	0.28	0.32	-0.51	-0.12	-0.07	-0.11	0.01					
ADS	0.61	0.17	-0.27	0.08	0.00	-0.23	0.39	-0.20	-0.07	-0.14	-0.56	-0.20	0.44	-0.09	-0.11	0.14	0.02					
CFNAI	0.56	0.18	-0.32	0.22	-0.01	-0.20	0.52	-0.26	0.06	-0.27	-0.54	-0.27	0.49	-0.22	-0.03	0.12	0.05					
NFCI	-0.90	0.19	0.25	-0.33	-0.04	0.21	-0.15	0.20	0.00	0.12	0.76	-0.15	-0.58	0.28	-0.05	-0.07	0.20					
GZ_SPRD	-0.49	-0.14	0.48	-0.10	0.44	0.20	-0.47	-0.03	0.04	0.17	0.37	0.30	-0.43	0.40	0.45	0.02	-0.14					
RM_RF	0.35	0.02	-0.36	-0.03	0.16	0.04	-0.03	-0.40	0.05	-0.03	-0.21	-0.01	0.45	0.20	0.00	-0.09	-0.01					
SMB	0.11	0.17	-0.33	0.11	0.19	-0.03	0.15	-0.25	0.11	0.04	-0.06	-0.22	0.32	0.10	0.15	-0.02	-0.03					
HML	-0.04	-0.07	0.09	0.11	-0.04	-0.28	-0.01	0.08	-0.15	-0.21	0.03	0.04	-0.06	-0.19	0.06	0.18	0.07					
UMD	-0.08	0.02	-0.05	-0.02	-0.22	0.05	0.07	0.12	0.19	0.01	0.10	-0.08	-0.05	-0.12	-0.15	-0.09	0.00					
FinC	0.13	0.34	0.23	0.33	-0.41	0.18	0.21	0.50	0.02	-0.12	-0.30	-0.23	-0.32	-0.40	-0.04	-0.37	-0.13					
BusC	-0.52	0.46	0.11	0.09	-0.56	-0.11	0.34	0.63	0.02	-0.13	0.37	-0.51	-0.48	-0.38	-0.25	-0.12	0.03					

(continued)

Table 5 Continued

Panel B: Correlation matrix of quarterly CF, HF, LF, LF5 from mixed-frequencies factor model and the CFFs with other factors and variables in the post-GM period

	Mixed-frequency group factors															CFFs							
	CF1	HF1	HF2	HF3	HF4	HF5	HF6	HF7	HF8	LF1	LF2	LF3	LF4	LF5	CFF1	CFF2	CFF3	CFF4	CFF5	CFF6	CFF7	CFF8	CFF9
GDP	0.30	0.16	-0.51	0.02	-0.10	-0.13	0.00	0.18	0.06	-0.70	-0.12	-0.38	-0.18	0.21	-0.74	0.02	0.01	0.05	0.10	-0.05	-0.12	0.20	0.03
RealCons	0.46	0.12	-0.39	-0.08	0.06	-0.12	-0.07	0.11	0.08	-0.51	0.00	-0.08	-0.25	-0.07	-0.66	-0.25	0.03	-0.03	-0.02	-0.09	-0.07	0.07	0.13
Baa-Aaa	0.05	0.11	0.45	0.07	0.24	0.06	-0.18	-0.23	0.12	0.24	0.41	0.34	-0.06	0.03	0.29	-0.20	0.25	-0.35	-0.15	-0.04	-0.12	-0.26	0.09
logVIX	0.02	-0.13	0.65	-0.06	0.19	0.10	0.27	0.13	0.22	0.49	0.00	-0.08	-0.03	-0.08	0.57	-0.26	0.08	-0.07	-0.08	-0.09	0.04	0.38	0.14
VRP	0.08	-0.20	-0.10	-0.11	-0.09	0.38	0.04	0.15	0.09	-0.05	0.00	-0.29	-0.03	-0.08	-0.05	-0.10	-0.22	0.20	-0.01	0.25	-0.27	0.19	0.04
ETF_iSHARES	-0.18	-0.16	-0.75	0.16	-0.33	0.05	0.18	0.14	-0.05	-0.46	-0.39	-0.22	-0.06	0.00	-0.54	0.45	-0.45	0.27	0.16	0.31	0.17	0.14	-0.02
ADS	0.18	0.10	-0.65	0.05	-0.17	-0.16	-0.04	0.24	-0.03	-0.80	-0.18	-0.20	-0.03	-0.04	-0.79	0.18	-0.11	0.07	0.15	-0.05	-0.12	0.18	0.03
CFNAI	0.20	0.08	-0.61	0.08	-0.20	-0.27	-0.06	0.20	-0.03	-0.81	-0.15	-0.12	-0.01	-0.06	-0.79	0.17	-0.11	0.00	0.21	-0.12	-0.09	0.12	-0.01
NFCI	-0.02	0.08	0.72	0.15	-0.03	0.42	0.21	-0.06	0.14	0.76	0.00	0.01	0.18	-0.03	0.75	-0.18	0.30	-0.09	0.06	0.35	0.02	0.20	0.06
GZ_SPDR	-0.35	0.01	0.68	-0.12	0.10	0.17	0.14	-0.21	0.15	0.70	0.09	0.08	-0.11	0.03	0.82	-0.02	0.19	0.06	-0.12	-0.01	0.13	-0.10	0.11
SKEW	-0.45	0.07	-0.17	0.18	0.07	-0.18	-0.07	-0.01	0.46	-0.16	0.07	0.21	-0.12	0.03	-0.02	0.45	-0.04	-0.10	-0.14	-0.10	0.09	-0.25	0.45
logSKEW	-0.45	0.07	-0.17	0.18	0.07	-0.18	-0.07	-0.01	0.46	-0.16	0.07	0.21	-0.12	0.02	-0.01	0.45	-0.04	-0.10	-0.13	-0.09	0.09	-0.24	0.45
RM_RF	0.01	0.03	-0.25	0.09	0.03	-0.20	0.03	0.19	-0.04	-0.37	-0.07	-0.14	0.11	-0.10	-0.30	0.16	-0.06	-0.04	0.02	-0.15	-0.03	0.19	-0.05
SMB	-0.10	0.21	0.03	-0.27	-0.05	-0.13	0.14	0.07	-0.06	-0.01	-0.16	-0.28	0.04	0.10	0.01	0.05	0.20	0.29	0.03	-0.23	0.05	0.12	-0.12
HML	-0.05	-0.20	-0.04	-0.19	-0.02	-0.02	-0.07	0.10	-0.01	-0.05	-0.02	-0.20	0.03	0.01	-0.02	-0.01	-0.11	0.18	0.00	-0.13	0.04	-0.13	-0.13
UMD	0.12	-0.14	0.10	0.03	0.12	0.03	0.02	-0.09	-0.11	0.09	0.11	0.21	-0.11	-0.07	0.06	-0.17	-0.10	-0.16	-0.08	0.02	0.04	-0.04	-0.01
FinC	0.45	-0.03	0.14	0.05	-0.17	0.06	0.17	-0.12	-0.04	0.34	-0.04	-0.03	0.27	0.01	0.06	-0.42	0.07	-0.05	0.24	0.23	0.16	0.04	-0.25
BusC	-0.04	-0.08	0.19	0.26	0.13	-0.20	-0.06	-0.26	-0.09	0.10	0.19	0.51	-0.13	0.08	0.17	0.01	-0.03	-0.42	-0.02	-0.08	0.19	-0.24	-0.02

Notes: The quarterly variables used in the correlation matrices are Real GDP growth (GDP), Real Consumption growth (RealCons), Moody's default spread (Baa-Aaa), VIX (logVIX), VRP, ETF iShares Core S&P500 (ETF_iSHARES), the *Aruoba, Diebold, and Scotti (2009)* index (ADS), the Chicago Fed National Activity Index (CFNAI), the NFCI, the *Gilchrist and Zakrajsek (2012)* spread (GZ_SPDR), the *Amaya et al. (2015)* realized skewness (SKEW and logSKEW), the Fama-French factors (Excess market returns [RM_RF], SMB, HML, and UMD) and the *Drehmann, Borio, and Tsatsaronis (2012)* U.S. Financial Cycle (FinC) and Business Cycle (BusC) indicators. Pairwise correlations are reported for all variables for the pre- and post-GM, except the following variables for the subperiods in the parenthesis: CFNAI(1967q1), NFCI (pre: 1971q1), and GZ_SPDR (pre: 1973q1–1983q4). Bold values indicate the maximum correlation of each factor with the corresponding variable in the first column.

Table 6 MFF (CF, HFF, and LFF) and CFF and their R^2 with alternative variable categories

Panel A: Pre-GM Factors			Panel B: Post-GM Factors		
Factors	R^2	Category of variables	Factors	R^2	Category of variables
CF	0.42–0.85	Interest rates	CF	0.52–0.64	Commodities spreads
	0.59	Stock markets: VXO		0.30–0.46	Interest rates
	0.28–0.49	Commodities spreads			
	0.61	Consumer Expectations		0.28–0.56	Employment
	0.40–0.56	Non-Household Balance Sheets		0.55	IP
	0.29–0.50	Money and Credit		0.37	Household Balance Sheets
	0.27–0.38	NIPA		0.28–0.37	NIPA
	0.32–0.37	Earnings and Productivity		0.26–0.36	Money and Credit
	0.29–0.32	Household Balance Sheets		0.36	Consumer Expectations
HFF1	0.29	Employment	HFF1	0.28–0.81	Interest rates
	0.35–0.52	Interest rates			
	0.36–0.45	Stock Markets			
	0.42	IP			
HFF2	0.34	Employment	HFF2	0.34–0.55	Stock Markets
	0.26	Prices			
	0.25–0.44	Commodities spreads, Exchange rates			
	0.26–0.29	Interest rates		0.32–0.42	Exchange rates
				0.26–0.34	Interest rates
HFF3	0.28	Employment	HFF3	0.27–0.58	Household Balance Sheets
	0.37–0.45	Interest rates		0.36	Inventories, orders, and sales
	0.30	Exchange rates		0.26–0.30	NIPA
HFF4	0.26–0.31	Interest rates, Exchange rates	HFF4	0.25	Housing
	0.44–0.46	Stock Markets		0.27–0.46	Commodities spreads
HFF5	0.30	Interest rates	HFF5	0.26–0.37	Interest rates
				0.38	Money and Credit
LFF1	0.25–0.36	Interest rates	LFF1	0.32	Inventories, orders, and sales
	0.26	Stock markets		0.26–0.62	Exchange rates
	0.53–0.76	Employment, IP		0.26–0.42	Stock Markets
	0.56	NIPA		0.29–0.34	Exchange rates
LFF2			LFF2	0.37–0.52	Interest rates
	0.29	Interest rates		0.26–0.28	Commodities spreads
	0.31–0.46	IP		0.34	Interest rates
	0.27–0.42	Employment		0.34	Stock Markets
	0.31–0.37	NIPA		0.66–0.83	Employment
	0.26–0.36	Inventories, orders, and sales		0.72	NIPA
	0.31	Non-Household Balance Sheets		0.62–0.71	IP
0.26–0.31	Housing	0.67	Inventories, orders, and sales		
LFF3	0.25–0.66	Prices	LFF3	0.48	Money and Credit
				0.45	Earnings and Productivity
LFF4			LFF4	0.31	Inventories, orders, and sales
	0.26–0.50	Prices, earnings, and productivity		0.26	Money and Credit
				0.25–0.26	NIPA
				0.61–0.71	Earnings and Productivity
		LFF5			

variables in each category can be found in Table OA.7 in [Online Appendix](#) (with the corresponding acronyms for each variable used to extract the factors). Focusing on the CFs results reported in the top of Panels A and B in [Table 6](#), the evidence suggests that the CF loads on three main financial categories during the pre-GM regime, namely interest rates spreads (both government and corporate default spreads) with $R^2 = 0.42\text{--}0.85$, commodities spreads (the difference between future and spot prices) with $R^2 = 0.28\text{--}0.49$ and the VXO with $R^2 = 0.59$.¹⁶ In contrast, in the post-GM, the CF is no longer driven by the VXO (and any other stock market indicators). In fact, the R^2 of the regression of the CF and VXO in the second regime drops to 0.01. Our group factor model reveals that in the post-GM, the VXO instead drives the second HF (monthly) financial factor, HFF2, with $R^2 = 0.55$, as shown in Table OA.7, Panel B. Moreover, interest rates spreads while being one of the main drivers of the CF in both regimes, their R^2 becomes weaker in the post-GM period with $R^2 = 0.30\text{--}0.46$, while commodities spreads are still highly correlated with the CF in the post-GM with $R^2 = 0.52\text{--}0.64$. These results further explain the changing role of the drivers of the CF, which in the last four decades, was mainly driven by commodities spreads and interest rates and not by the VXO or any other stock market indices. We find that the main drivers of the macro-finance CF are interest rates and credit spread factors (also found by [Gilchrist and Zakrajsek, 2012](#) for the U.S. economic activity), as well as commodities spreads and returns (also found by [Gospodinov and Ng, 2013](#) for explaining the U.S. inflation and by [Chaise, Ferrara, and Giannone, 2017](#) for global economic activity). These relationships of the CF with specific variables are further analyzed in the next subsection using dynamic partial correlations in a predictive context. Hence, we will investigate the role of our CF in Granger causing as well forecasting out of sample key financial and macro variables (including the VIX).

Turning to the LF quarterly macro variables, we find that many different categories drive the CF. In the pre-GM regime, the CF is driven by variables in the following categories ranked in terms of the higher R^2 first: Consumer Expectations ($R^2 = 0.61$), non-Household Balance Sheets ($R^2 = 0.40\text{--}0.56$), Money and Credit ($R^2 = 0.29\text{--}0.50$), National Income and Product Accounts ($R^2 = 0.27\text{--}0.38$), Earnings and Productivity ($R^2 = 0.32\text{--}0.37$), Household Balance Sheets ($R^2 = 0.29\text{--}0.32$), and Employment ($R^2 = 0.29$). While in the post-GM, the aforementioned variable categories are still important with similar R^2 , the non-Household Balance Sheets as well as Earnings and Productivity, are no longer correlated with the CF. Instead the Industrial Production (namely Capacity Utilization: Manufacturing with $R^2 = 0.55$) becomes an important driver of the CF in the post-GM and the role of the Employment category increases with $R^2 = 0.28\text{--}0.56$.

Last but not least, we obtain the correlation of the CF with some key economic and financial variables which we will evaluate in terms of forecasting, as well as with other well-known factors in the literature in the two regimes. [Table 5](#) shows that the correlation of the CF with real GDP and consumption growth is higher in the pre-GM period rather than in the post-GM period (in Panels A and B, respectively). The same applies for the correlation

16 For the full sample factor estimation, we consider the CBOE S&P100 Volatility Index (VXO) due to the longer historical sample since the 1960s. In the post-GM period, we examine the role of the factors in predicting the VIX which is not only a broader index (referring to the S&P500) but also a benchmark indicator in the VRP. Note that in the post-GM the correlation of the VXO and VIX is 0.97.

between the CF with the ADS index or CFNAI, for which their correlation drops below one-third compared to that in the first regime. More importantly, while in the first regime the CF is highly correlated with the financial factors and indicators, for example, the NFCI (0.90), Baa-Aaa (0.35), excess stock market returns (0.35), their correlation with the CF in the second regime drops to 0.02, 0.05, and 0.01, respectively.

Naturally, we wish to investigate how our Mixed-Frequency Group Factors (MFFs), that is, the CF, the HFFs, and the LFFs, compare to the traditional approach which employs aggregated (quarterly) data and pools the macro and financial panels to extract the CFFs from a single panel. First, we find evidence consistent with the literature above, that the loadings of the CFFs also exhibit a break associated with the GM. Evidence from the LM (resp. supLM) test in Table 4 reveals that using all the CFFs chosen by the IC_{p2} criterion, there is overwhelming evidence of a break in 90% (resp. 97%) of the loadings of all the series in the panel (both the macro and financial) in the mid-1980s. However, in order to compare the results derived from our CF, we focus on the first and second CFFs (CFF1 and CFF2 or both together), which have the highest correlation with the CF in pre- and post-break regimes. This result is reported in Table 7 (in the first column of Panels A and B), where the CF is highly correlated with the CFF1 (0.90) in the pre-GM regime and with the CFF2 (0.82) in the post-GM regime. The supLM, based on Table 4, yields very similar results whether using the CF or CFF2 or both CFF1 and CFF2 with regard to the total percentage of loadings changing, whereas the CFF2 also provides closer results to those of the CF when it comes to the various variable category loadings changes (as shown, for instance, by the consumer expectations category) as opposed to CFF1. Last, applying the LM-type tests to all CFFs or all the MFFs (CF, HFFs, and LFFs) selected by the IC_{p2} , we find that all loadings change in almost all categories. Our results suggest that focusing on the CF loading change point tests we are able to identify more heterogeneity in the change point of the loadings of different variable categories that drive the macro–finance factor.

The relationship of the CFFs and mixed-frequency group factors as well as key macro and financial variables is further analyzed in Tables 5 and 7. Table 7 shows that the IC_{p2} criterion selects a larger number of factors (HFF, LFF, and CFF) in the post-GM period rather than in the pre-GM period. As expected, the CF, LFFs, and HFFs are highly correlated with different CFFs in the two subperiods. For instance, in the pre-GM, the first CFF (CFF1) is highly correlated with the CF and the second CFF (CFF2) with both HFF1 and LFF1. In contrast, in the post-GM, CFF1 is correlated with HFF2 and LFF1 whereas CFF2 with just CF. These results suggest that while extracting factors from the mixed-frequency group factor model it is possible to identify and label common versus group-specific factors even in different regimes, whereas, this is less obvious in the CFF model. Hence, in many cases, it is difficult to isolate what is the driving group of the CFFs. Turning to Table 5, we find that in both subperiods while CFF1 is highly correlated with real GDP, Consumption growth, the CFNAI, and the ADS index, in the post-GM period CFF1 becomes highly correlated with NFCI and GZ_spread as opposed to the pre-GM period.

3.3 Predictive Evidence

In this subsection, we investigate the role of our estimated mixed-frequency group factors using IS and OOS predictive regression models in explaining and forecasting key macro and financial variables, namely the real GDP and Consumption growth, the Moody's Baa-

Table 7 Correlation matrices of factors during the pre- and post-GM period

Panel A: Correlation matrix of quarterly CF, HFFs, LFFs from mixed-frequencies group factor model and the CFFs in the pre-GM, 1963q3–1983q4.

	Mixed-frequency group factors									
	CF	HFF1	HFF2	HFF3	HFF4	HFF5	LFF1	LFF2	LFF3	LFF4
Common frequency factors										
CFF1	0.90	0.26	0.28	0.09	0.08	0.00	0.17	0.15	0.03	0.03
CFF2	0.19	0.88	0.30	0.09	0.09	0.17	0.79	0.06	0.05	0.05
CFF3	0.29	0.26	0.86	0.14	0.21	0.01	0.18	0.72	0.06	0.14
CFF4	0.00	0.22	0.06	0.67	0.60	0.29	0.32	0.35	0.03	0.19
CFF5	0.19	0.06	0.18	0.67	0.68	0.04	0.04	0.24	0.23	0.13
CFF6	0.04	0.14	0.13	0.24	0.30	0.84	0.31	0.02	0.03	0.14
CFF7	0.05	0.01	0.05	0.00	0.03	0.07	0.01	0.09	0.14	0.22

Panel B: Correlation matrix of quarterly CF, HFFs, LFFs from mixed-frequencies factor model and the CFFs in the post-GM, 1986q1–2017q4

	Mixed-frequency factors														
	CF	HFF1	HFF2	HFF3	HFF4	HFF5	HFF6	HFF7	HFF8	LFF1	LFF2	LFF3	LFF4	LFF5	
Common frequency factors															
CFF1	0.41	0.23	0.81	0.04	0.02	0.20	0.11	0.09	0.01	0.80	0.10	0.08	0.12	0.05	
CFF2	0.82	0.20	0.28	0.37	0.14	0.11	0.04	0.12	0.01	0.39	0.09	0.10	0.10	0.06	
CFF3	0.09	0.95	0.29	0.05	0.01	0.02	0.00	0.03	0.00	0.10	0.02	0.08	0.20	0.15	
CFF4	0.23	0.06	0.25	0.80	0.31	0.25	0.18	0.04	0.00	0.15	0.36	0.61	0.15	0.06	
CFF5	0.19	0.05	0.20	0.16	0.90	0.24	0.02	0.01	0.02	0.10	0.34	0.05	0.41	0.02	
CFF6	0.13	0.04	0.19	0.39	0.18	0.82	0.08	0.20	0.01	0.24	0.10	0.03	0.04	0.00	
CFF7	0.03	0.01	0.12	0.08	0.09	0.32	0.58	0.54	0.05	0.19	0.54	0.35	0.22	0.13	
CFF8	0.16	0.01	0.08	0.13	0.09	0.08	0.61	0.67	0.08	0.06	0.28	0.26	0.14	0.11	
CFF9	0.00	0.00	0.04	0.01	0.03	0.05	0.06	0.22	0.72	0.03	0.14	0.25	0.49	0.12	

(continued)

Table 7 Continued

Panel C: Correlation matrix of quarterly CF, HFFs, LFFs from mixed-frequencies factor model and the CFFs during the entire period 1963q3–2017q4

	Mixed-frequency factors															
	CF1	HFF1	HFF2	HFF3	HFF4	HFF5	HFF6	HFF7	LFF1	LFF2	LFF3	LFF4	LFF5	LFF6	LFF7	
Common frequency factors	CFF1	0.76	0.22	0.52	0.15	0.05	0.10	0.14	0.01	0.50	0.03	0.10	0.03	0.00	0.01	0.07
	CFF2	0.58	0.16	0.63	0.13	0.08	0.21	0.17	0.19	0.73	0.02	0.10	0.05	0.06	0.06	0.01
	CFF3	0.05	0.92	0.35	0.01	0.11	0.02	0.03	0.08	0.08	0.26	0.56	0.26	0.02	0.08	0.01
	CFF4	0.04	0.14	0.30	0.90	0.08	0.14	0.14	0.12	0.17	0.12	0.23	0.18	0.16	0.25	0.30
	CFF5	0.02	0.17	0.08	0.14	0.94	0.11	0.09	0.04	0.10	0.40	0.17	0.02	0.04	0.12	0.04
	CFF6	0.22	0.06	0.10	0.22	0.04	0.92	0.17	0.07	0.02	0.04	0.03	0.05	0.24	0.15	0.30
	CFF7	0.04	0.14	0.25	0.26	0.22	0.20	0.70	0.37	0.30	0.22	0.22	0.01	0.12	0.06	0.09
	CFF8	0.14	0.04	0.01	0.09	0.04	0.10	0.46	0.77	0.00	0.39	0.48	0.11	0.02	0.04	0.05
	CFF9	0.07	0.08	0.10	0.04	0.17	0.03	0.31	0.09	0.16	0.53	0.02	0.04	0.09	0.02	0.07
	CFF10	0.02	0.00	0.10	0.02	0.07	0.03	0.20	0.21	0.14	0.42	0.20	0.55	0.09	0.23	0.10

Bold values indicate the maximum correlation of each mixed-frequency factor with the corresponding common frequency factor (CFF).

Aaa default spread, the VIX, the VRP, and the ETF iShares Core S&P500 returns. We focus at forecasting these variables at quarterly frequency given that in many cases the quarterly frequency is the frequency of interest of policy makers as well as for comparison purposes across all variables. We consider traditional linear factor augmented distributed lag (FADL) models when all variables and factors are at the same, low (quarterly) frequency estimated by least squares (referred to as Linear-LS). Among these specifications, we include the models with the traditional CFFs considered in the literature. Additionally, we estimate the corresponding FADL-MIDAS models by nonlinear least squares (referred to as MIDAS-NLS), given that predictors/factors are available at higher frequency (monthly) than the dependent variable and some MIDAS weighting schemes can be estimated by NLS. Alternative HF weighting polynomials (the Almon and the Step) are used for estimating the FADL-MIDAS models. The predictive role and information content of our factors is also assessed relative to other related and established U.S. factors in the literature, such as the CFNAI, the NFICI, the ADS index, the GZ_spread, and the four Fama–French factors, excess market returns (RM-RF), small-minus-big (SMB), high-minus-low (HML), and momentum (UMD).

3.3.1 IS predictive evidence

The IS linear and MIDAS predictive models are presented in Tables 8 and 9 and include our MFFs extracted from the mixed-frequency group factor model and/or the aforementioned well-known factors in the literature as well as the traditional CFFs. More precisely in Table 8 we report the results for real GDP and Consumption growth as well as the corporate bonds default spread for the two subperiods marked by the GM, while in Table 9 we report the results for the VIX, VRP, and ETF returns for the more recent period, due to the shorter data sample available. Given the large model space involved in estimating the above models for all predictors, lag lengths, and HF weighting polynomials, we focus on reporting the results for those models where predictors turn out to be significant following the Bai and Ng (2008) targeted predictors approach with hard thresholding (based on the 10% significance level and the heteroskedastic and autocorrelation Newey–West standard errors). For the lag length p in these FADL type models, we consider $p = 1$ up to four quarters and select the number of lags using the BIC (which is a consistent information criterion and selects parsimonious models—a desirable property for forecasting models). For most models reported in Tables 8 and 9, the BIC selects one lag. For each dependent variable, the FADL and FADL-MIDAS models are compared in terms of the BIC. We highlight in bold the three models with the lowest BIC values and mark with a $+$ the model that yields the lowest BIC among these. Of special interest is the predictive or Granger causal role of the CF evaluated via the significance and estimated regression coefficient of the CF ($\hat{\beta}_{CF}$) which are also reported in Tables 8 and 9.¹⁷

In the top panel of Table 8, we present the results of real GDP growth in the two regimes. The first column reports the alternative model specifications in each row while the

17 Alternative approaches of dealing with the large model space such as alternative criteria for model selection, model averaging, shrinkage, among others, can also be pursued in this context. Although these are complementary approaches, our analysis aims at uncovering the predictive role of the common macro–finance as well as the group-specific factors in comparison to other factors and hence we consider the model selection approach.

Table 8 Real GDP growth (GDP), Real Consumption growth (RealCons) and Moody's default spread (Baa-Aaa) predictive models BIC results: FADL Linear-LS and FADL MIDAS models with alternative factors (CFFs and MFFs) during the pre- and post-GM period

Predictive model specifications	GDP 1963q3–1983q4			GDP 1986q1–2017q4		
	Targeted predictors	$\hat{\beta}_{CF}$	BIC	Targeted predictors	$\hat{\beta}_{CF}$	BIC
LD		-	-6.189		-	-7.557
LD, all CFFs [C_{p2}] in Linear-LS		-	-6.378		-	-7.558
LD, CFFs in Linear-LS	CFF:1,3,4,5	-	-6.518	CFF:1,2,3,5,7,8	-	-7.652
LD, MFFs in linear-LS	CF	0.238***	-6.505	CF, HFF4, LFF1	0.050***	-7.666
LD, MFFs in Almon MIDAS	CF, HFF6, LFF1	0.164***	-6.434	CF, LFF:1,5, HFF:1,5	0.027**	-7.792
LD, MFFs in step MIDAS	CF, LFF1	0.176**	-6.433	CF, LFF:1,5, HFF:1,5	0.012**	-7.773
LD, ADS, MFFs in Linear-LS	ADS	-	-7.216 ⁺	ADS, HFF4	-	-8.067
LD, ADS, MFFs in Almon MIDAS	ADS, LFF5, HFF1	-	-7.142	ADS	-	-8.195
LD, ADS, MFFs in step MIDAS	ADS, CF, LFF5, HFF:1,2,3	0.062***	-7.119	ADS	-	-8.223 ⁺
LD, CFNAI, MFFs in Linear-LS	CFNAI	-	-6.626	CF, CFNAI, HFF4, LFF1	-	-7.717
LD, CFNAI, MFFs in Almon MIDAS	CFNAI	-	-6.671	CFNAI, HFF5, LFF1	-	-8.101
LD, CFNAI, MFFs in step MIDAS	CFNAI, CF, LFF5, HFF:1,2,3	0.021***	-6.801	CFNAI, HFF5	-	-8.070
LD, NFCl, MFFs in Linear-LS	CF, HFF3	0.242***	-6.476	CF, LFF1	0.045**	-7.619
LD, NFCl, MFFs in Almon MIDAS	NFCl, CF, LFF:1,2, HFF:3,5,6	0.288***	-6.230	CF, LFF:1,5, HFF5	0.027**	-7.695
LD, NFCl, MFFs in step MIDAS	CF, LFF:1,5, HFF:1,5	0.232***	-6.322	CF, LFF:3,5, HFF5	0.035***	-7.547
LD, GZ_SPRD, MFFs in Linear-LS	HFF:3,6	-	-6.091	CF, HFF4, LFF1	0.050***	-7.666
LD, GZ_SPRD, MFFs in Almon MIDAS	CF, HFF3	0.154***	-6.496	CF, HFF:2,5, LFF1	0.033***	-7.766
LD, GZ_SPRD, MFFs in step MIDAS	HFF:1,6	-	-6.019	GZ_SPRD, LFF5, HFF6	-	-7.662
LD, FFs, MFFs in Linear-LS	CF, LFF5, HFF3	0.243***	-6.442	HML, UMD, RM_RF, CF, LFF1, HFF:2,3,4,5	0.058***	-7.578
LD, FFs, MFFs in Almon MIDAS	CF, LFF1	0.166***	-6.483	RM_RF, UMD, HML, CF, LFF1, HFF:2,3,5	0.030***	-7.739
LD, FFs, MFFs in step MIDAS	-	-	-6.189	RM_RF, HML, LFF1, HFF:1,2,5	-	-7.569

(continued)

Table 8 Continued

Predictive model specifications	RealCons 1963q3–1983q4			RealCons 1986q1–2017q4		
	Targeted predictors	$\hat{\beta}_{CF}$	BIC	Targeted predictors	$\hat{\beta}_{CF}$	BIC
LD		-	-7.840		-	-8.578
LD, all CFFs [C_{p2}] in Linear-LS		-	-7.922		-	-8.634
LD, CFFs in Linear-LS	CFF:1,2,3	-	-8.125	CFF:1,2,7,9	-	-8.775
LD, MFFs in Linear-LS	CF, HFF2	0.093***	-8.104	CF, HFF:4,5,7, LFF:1,3,4	0.057***	-8.768
LD, MFFs in Almon MIDAS	CF	0.073***	-8.138 ⁺	CF, HFF:4,5, LFF:1,3	0.030***	-8.796
LD, MFFs in step MIDAS	-	-	-	LFF1, HFF3	-	8.573
LD, ADS, MFFs in Linear-LS	ADS, CF, LFF4, HFF:2,4	0.078***	-8.015	ADS, CF, LFF:1,3,4, HFF:4,5,7	0.053***	-8.751
LD, ADS, MFFs in Almon MIDAS	ADS	-	-7.849	ADS, CF, LFF:2,3, HFF:4,5	0.019***	-8.879 ⁺
LD, ADS, MFFs in step MIDAS	ADS	-	-7.988	ADS, LFF1, HFF4	-	-8.725
LD, CFNAI, MFFs in Linear-LS	CFNAI, CF, LFF4, HFF:2,4,7	0.081***	-7.951	CF, LFF:1,3,4, HFF:4,5,7	0.057***	-8.768
LD, CFNAI, MFFs in Almon MIDAS	CFNAI, HFF4	-	-7.829	CFNAI, CF, LFF:2,3, HFF:4,5	0.017***	-8.875
LD, CFNAI, MFFs in step MIDAS	-	-	-	CFNAI, LFF:2,3, HFF4	-	-8.722
LD, NFCl, MFFs in Linear-LS	HFF2	-	-7.918	CF, LFF:1,3,4, HFF:4,5,7	0.057***	-8.768
LD, NFCl, MFFs in Almon MIDAS	NFCl, CF, HFF:5,7	0.098***	-7.817	CF, LFF1, HFF:4,5	0.031***	-8.795
LD, NFCl, MFFs in step MIDAS	CF, HFF:5,6	-	-7.972	LFF1	-	-
LD, GZ_SPRD, MFFs in Linear-LS	HFF7	-	-7.798	CF, LFF:1,3,4, HFF:4,5,7	0.057***	-8.768
LD, GZ_SPRD, MFFs in Almon MIDAS	-	-	-	CF, LFF:1,3, HFF:4,5	0.030***	-8.796
LD, GZ_SPRD, MFFs in step MIDAS	GZ_SPRD, LFF:4,5, HFF:3,5,6	-	-7.282	LFF1, HFF:3,4	-	-8.579
LD, FFs, MFFs in Linear-LS	UMD, CF, HFF7	0.093***	-7.946	CF, LFF:1,3,4, HFF:4,7	0.056***	-8.733
LD, FFs, MFFs in Almon MIDAS	HFF4	-	-7.810	CF, LFF:1,3, HFF4	0.031***	-8.752
LD, FFs, MFFs in step MIDAS	UMD	-	-7.876	LFF1	-	-

(continued)

Table 8 Continued

Predictive model specifications	Baa-Aaa 1963q3–1983q4		Baa-Aaa 1986q1–2017q4	
	Targeted predictors	$\hat{\beta}_{CF}$	Targeted predictors	$\hat{\beta}_{CF}$
LD		-		-
LD, all CFFs [IC_{CF}] in Linear-LS		-0.201		-0.328
LD, CFFs in Linear-LS	CFF:1,4,5,6	-0.773		-0.419
LD, MFFs in Linear-LS	CF, LFF:1,3, HFF:2,4,6	-0.896	CFF:1,2,4,5,6	-0.434
LD, MFFs in Almon MIDAS	CF, HFF:1,3,4,6	0.054***	LFF1, HFF:2,5	-0.534
LD, MFFs in step MIDAS	CF, LFF:2,5, HFF:1,2,3,5,6,7	0.042***	LFF1, HFF:1,2,5	-0.854
LD, ADS, MFFs in Linear-LS	ADS, CF, LFF:1,3, HFF:2,4,6,7	0.024***	HFF:1,2,3,5	-0.962
LD, ADS, MFFs in Almon MIDAS	ADS, CF, HFF:1,4,6	0.040***	HFF:2,5	-0.601
LD, ADS, MFFs in step MIDAS	ADS, CF, LFF:2,5, HFF:1,3,5,6,7	0.030***	ADS, HFF:1,2,5	-0.928
LD, CFNAI, MFFs in Linear-LS	CFNAI, CF, LFF3, HFF:2,4,6,7	0.000***	ADS, HFF:1,2,3,5,7	-1.062
LD, CFNAI, MFFs in Almon MIDAS	CFNAI, CF, LFF2, HFF:1,4,6	0.037***	HFF:2,5	-0.601
LD, CFNAI, MFFs in step MIDAS	CFNAI, CF, HFF:3,4,5,6,7	0.028***	HFF:1,2,5	-0.909
LD, NFCL, MFFs in Linear-LS	NFCL, LFF:2,3, HFF:2,3,4,5,7	0.006***	CFNAI, HFF:1,2,3,5,7	-1.028
LD, NFCL, MFFs in Almon MIDAS	HFF4	-	LFF1, HFF:2,5	-0.572
LD, NFCL, MFFs in step MIDAS	NFCL, CF, HFF:1,2,3,5,6,7	-0.172	HFF:1,2,5	-0.909
LD, GZ_SPRD, MFFs in Linear-LS	CF, LFF3, HFF:2,4,6	0.048***	NFCL, CF, HFF:1,5	0.004**
LD, GZ_SPRD, MFFs in Almon MIDAS	HFF:3,6,8	0.048***	GZ_SPRD, LFF1, HFF:2,4,5,7	-0.532
LD, GZ_SPRD, MFFs in step MIDAS	CF, HFF:3,4,5,6,7	-	GZ_SPRD, LFF1, HFF:1,2,5	-0.845
LD, FFs, MFFs in Linear-LS	CF, LFF:1,3, HFF:2,4,6	0.025***	GZ_SPRD, HFF:1,5	-1.188
LD, FFs, MFFs in Almon MIDAS	RM_RF, SMB, CF, LFF3, HFF:2,4,7,8	0.054***	LFF1, HFF7	-0.350
LD, FFs, MFFs in step MIDAS	RM_RF, UMD, CF, LFF:3,5, HFF:2,3,4,5,6,7,8	0.040***	LFF1, HFF:1,2,5	-0.891
		0.037***	UMD, HFF:1,2	-0.777

Notes: Bold BIC values refer to the models with the three lowest BIC for each dependent variable and the BIC values with a ⁺ denotes the model with minimum BIC for each variable. Targeted predictors are based on the Bai and Ng (2008) hard thresholding approach using 10% significance level, Newey–West standard errors with the Bartlett kernel and data-driven bandwidth selection. The corresponding significance of the regression coefficient of the CF refers to the following significance levels: ***1%, **5%, and *10%. The common factor (CF), the HFFs, and the LFFs are estimated using the quarterly (LF) and monthly (HF) frequency macro and financial series panels, respectively, using the MFF model. The MIDAS model step ($s = 3$) and Almon ($d = 1$) are reported given that yield parsimonious representations with low BIC. The CFFs are estimated from all the series at the quarterly LF stacked in a common panel. The predictive model involves the following predictors: the lagged dependent (LD) term, the MFFs, namely the Common Factor (CF), the HFFs and LFFs, the CFFs, the Aruoba, Diebold, and Scotti (2009) (ADS) index, the Chicago National Activity and Financial Conditions Indices (CFNAI and NFCL, respectively), the four Fama–French (FF) factors (RM-RE, SMB, HML, UMD) and Gilchrist and Zakrajsek (2012) spread (GZ_SPRD).

Table 9 VIX (logVIX), VRP, and ETF iShares Core S&P500 (ETF_iSHARES) predictive models BIC results: FADL Linear-LS and FADL MIDAS NLS models with alternative Factors (CFFs and MFFs)

Predictive model specifications	logVIX targeted predictors 1990q1–2017q4	$\hat{\beta}_{CF}$	BIC
LD		–	–1.542
LD, all CFFs [IC_{p2}] in Linear-LS		–	–1.261
LD, CFFs in Linear-LS	CFF:2,5,6	–	–1.466
LD, MFFs in Linear-LS	CF, LFF5, HFF:2,3,4,5,6,7,8	0.015***	–1.386
LD, MFFs in Almon MIDAS	CF, LFF3, HFF:2,3,4,6,7,8	0.013***	–1.671
LD, MFFs in step MIDAS	LFF:3,5, HFF:2,3,5,6,7,8	–	–1.536
LD, ADS, MFFs in Linear-LS	CF, LFF5, HFF:2,3,4,5,6,7,8	0.015***	–1.386
LD, ADS, MFFs in Almon MIDAS	CF, LFF3, HFF:2,3,4,6,7,8	0.013***	–1.671
LD, ADS, MFFs in step MIDAS	ADS, LFF:3,5, HFF:2,5,6,7,8	–	–1.504
LD, CFNAI, MFFs in Linear-LS	CF, LFF5, HFF:2,3,4,5,6,7,8	0.015***	–1.386
LD, CFNAI, MFFs in Almon	CF, LFF3, HFF:2,3,4,6,7,8	0.013***	–1.671
LD, CFNAI, MFFs in step MIDAS	CFNAI, LFF:3,5, HFF:2,3,5,6,7,8	–	–1.536
LD, NFCI, MFFs in Linear-LS	CF, LFF5, HFF:2,3,4,5,6,7,8	0.015***	–1.386
LD, NFCI, MFFs in Almon MIDAS	CF, LFF:3,4, HFF:2,3,4,5,6,7,8	0.012***	–1.628
LD, NFCI, MFFs in step MIDAS	LFF:3,5, HFF:2,5,6,8	–	–1.606
LD, GZ_SPRD, MFFs in Linear-LS	CF, LFF:3,5, HFF:2,3,4,5,6,7,8	0.014**	–1.358
LD, GZ_SPRD, MFFs in Almon MIDAS	GZ_SPRD, CF, LFF3, HFF:2,4,6,7,8	0.019***	–1.649
LD, GZ_SPRD, MFFs in step MIDAS	LFF3, LFF5, HFF:2,5,6,7,8	–	–1.567
LD, FFs, MFFs in Linear-LS	CF, HFF:1,5,7,8	–	–1.370
LD, FFs, MFFs in Almon MIDAS	RM_RF, CF, LFF:3,5, HFF:1,2,4,7,8	0.005**	–1.630
LD, FFs, MFFs in step MIDAS	RM_RF, LFF:3,5, HFF:2,7,8	–	–1.688 ⁺
LD, logSKEW, MFFs in Linear-LS	CF, HFF:2,3,4,6,7,8	0.011**	–1.373
LD, logSKEW, MFFs in Almon MIDAS	CF, LFF3, HFF:2,3,4,5,6,7,8	0.012***	–1.644
LD, logSKEW, MFFs in step MIDAS	LFF:3,5, HFF:2,5,6,7,8	–	–1.567
Predictive model specifications	VRP targeted predictors 1990q1–2017q4	$\hat{\beta}_{CF}$	BIC
LD		–	8.288
LD, all CFFs [IC_{p2}] in Linear-LS		–	8.423
LD, CFFs in Linear-LS	CFF:2,3,7,8,9	–	8.297
LD, MFFs in Linear-LS	CF, HFF:1,2,7,8	1.649**	8.343
LD, MFFs in Almon MIDAS	CF, HFF:1,7,8	1.103***	8.282
LD, MFFs in step MIDAS	HFF:1,2,3,5,7	–	8.438
LD, ADS, MFFs in Linear-LS	CF, HFF:1,2,3,6,7,8	2.121**	8.360
LD, ADS, MFFs in Almon MIDAS	ADS, CF, LFF1, HFF:1,2,3,5,7,8	0.679*	8.188
LD, ADS, MFFs in step MIDAS	ADS, LFF1, HFF:1,2,3,5,6	–	8.369
LD, CFNAI, MFFs in Linear-LS	CF, HFF:1,2,7,8	1.649**	8.343
LD, CFNAI, MFFs in Almon	CFNAI, CF, LFF1, HFF:1,2,5,7,8	–	8.249
LD, CFNAI, MFFs in step MIDAS	CFNAI, HFF:1,2,3,5,8	–	8.460
LD, NFCI, MFFs in Linear-LS	NFCI, CF, HFF:1,7,8	2.356***	8.260
LD, NFCI, MFFs in Almon MIDAS	NFCI, CF, HFF:1,5,7	1.372***	8.169 ⁺
LD, NFCI, MFFs in step MIDAS	HFF:1,3,5,7	–	8.436
LD, GZ_SPRD, MFFs in Linear-LS	CF, HFF:1,7,8	1.370*	8.343
LD, GZ_SPRD, MFFs in Almon MIDAS	CF, HFF:1,7,8	1.103***	8.282
LD, GZ_SPRD, MFFs in step MIDAS	CF, HFF:1,2,3,5,6,7	0.239**	8.439
LD, FFs, MFFs in Linear-LS	CF, LFF5, HFF:3,4,5,7,8	–	8.463
LD, FFs, MFFs in Almon MIDAS	RM_RF, HML, CF, HFF:1,7,8	1.069***	8.177
LD, FFs, MFFs in step MIDAS	RM_RF, HML, HFF:1,3,7	–	8.403
LD, SKEW, MFFs in Linear-LS	CF, HFF:1,2,3,7,8	1.795***	8.345
LD, SKEW, MFFs in Almon MIDAS	CF, HFF:1,7,8	1.103***	8.282
LD, SKEW, MFFs in step MIDAS	CF, HFF:1,2,3,5,7	–	8.427

(continued)

Table 9 VIX (logVIX), VRP, and ETF iShares Core S&P500 (ETF_iSHARES) predictive models BIC results: FADL Linear-LS and FADL MIDAS NLS models with alternative Factors (CFFs and MFFs)

Predictive model specifications	ETF_iSHARES targeted predictors 2000q3–2017q4	$\hat{\beta}_{CF}$	BIC
LD		–	–1.993
LD, all CFFs [IC_{p2}] in Linear-LS		–	–2.035
LD, CFFs in Linear-LS	CFF:2,5,8	–	–1.899
LD, MFFs in Linear-LS	CF, HFF7	0.010***	–1.944
LD, MFFs in Almon MIDAS	CF, HFF:1,2,3,5	0.006***	–2.712
LD, MFFs in step MIDAS	CF, LFF1, HFF:2,3,5,7,8	0.000**	–3.466
LD, ADS, MFFs in Linear-LS	ADS, CF, HFF7	0.014***	–2.100
LD, ADS, MFFs in Almon MIDAS	CF, HFF:2,3,5	0.006***	–2.724
LD, ADS, MFFs in step MIDAS	CF, HFF:2,3,5,7,8	0.002**	–3.372
LD, CFNAI, MFFs in Linear-LS	CF, HFF7	0.010***	–1.944
LD, CFNAI, MFFs in Almon	CF, HFF:1,2,3,5	0.006***	–2.712
LD, CFNAI, MFFs in step MIDAS	HFF:2,3,4,5,7	–	–3.523⁺
LD, NFICI, MFFs in Linear-LS	NFICI, HFF:1,2,3	–	–1.962
LD, NFICI, MFFs in Almon MIDAS	CF, HFF:1,2,3,5	0.006***	–2.712
LD, NFICI, MFFs in step MIDAS	CF, HFF:2,3,5,7,8	0.002**	–3.372
LD, GZ_SPRD, MFFs in Linear-LS	GZ_SPRD, CF, HFF7	0.014***	–1.888
LD, GZ_SPRD, MFFs in Almon MIDAS	CF, HFF:1,2,3,5	0.006***	–2.712
LD, GZ_SPRD, MFFs in step MIDAS	HFF:2,3,5,7,8	–	–3.290
LD, FFs, MFFs in Linear-LS	CF, HFF7	0.010***	–1.944
LD, FFs, MFFs in Almon MIDAS	HFF:2,3,5	–	–2.644
LD, FFs, MFFs in step MIDAS	RM_RF, SMB, HFF:2,4,7	–	–3.443
LD, logSKEW, MFFs in Linear-LS	CF, LFF2, HFF7	0.010***	–1.885
LD, logSKEW, MFFs in Almon MIDAS	logSKEW, HFF:1,2,3,5,7	–	–2.686
LD, logSKEW, MFFs in step MIDAS	HFF:2,3,5,7	–	–3.369

Notes: Bold BIC values refer to the models with the three lowest BIC for each dependent variable and the BIC values with a ⁺ denotes the model with minimum BIC for each variable. Targeted predictors are based on the Bai and Ng (2008) hard thresholding approach using 10% significance level, Newey–West standard errors with the Bartlett kernel and data-driven bandwidth selection. The corresponding significance of the regression coefficient of the CF refers to the following significance levels: ***1%, **5%, and *10%. The common factor (CF), the HFFs, and the LFFs are estimated using the quarterly (LF) and monthly (HF) frequency macro and financial series panels, respectively, using the MFF model. The MIDAS model step ($s = 3$) and Almon ($d = 1$) are reported given that yield parsimonious representations with low BIC. The CFFs are estimated from all the series at the quarterly LF stacked in a common panel. The predictive model involves the following predictors: the lagged dependent (LD) term, the MFFs, namely the Common Factor (CF), the HFFs and LFFs, the CFFs, the Aruoba, Diebold, and Scotti (2009) (ADS) index, the Chicago National Activity and Financial Conditions Indices (CFNAI and NFICI, respectively), the four Fama–French (FF) factors (RM-RF, SMB, HML, UMD), Gilchrist and Zakrajsek (2012) spread (GZ_SPRD) and the Amaya, Christoffersen, Jacobs, and Vasquez (2015) Realized Skewness (SKEW).

second and fifth columns list the significant predictors/factors in each model that correspond to each of the two regimes. The corresponding BIC values for each model in the first and second regime are reported in fourth and last columns, respectively. Last but by no means least, the estimated regression coefficient of the CF and its statistical significance in the pre- and post-GM periods are also reported. There are two interesting results to note

about the GDP growth predictive models: first, following the targeted predictors approach, the model with the lowest BIC is a simple linear model and a MIDAS model in the first and second regimes, respectively, with just the lagged ADS index, followed by the MIDAS models which also include some of the significant LFFs and HFFs and the CF, especially in the first regime. Second, comparing the estimated predictive coefficients of the common factor in all the reported models in the pre- and post-GM regimes, we find that while it is significant in both regimes, its estimated value has dropped by at least a third in the second regime. Hence, although still statistically significant, the actual value of the predictive coefficient of the CF in explaining GDP growth has decreased in the last three decades.

This finding also extends to the real Consumption growth models reported in the second panel, for which the estimated CF coefficient drops by a half in the recent regime. More interestingly, the results for the default spread, found in the last panel, show that the CF turns out to be significant in many models in the first regime, whereas in the second regime it turns out to lose its Granger causality role in almost all models, except in a single model with the lowest BIC. For real Consumption growth, the model with the minimum BIC is a MIDAS specification with just the lagged CF in the first regime and the CF along with the ADS, LFFs, and HFFs in the second regime. Similarly, for the default spread in the first regime the MIDAS model with the CF, HFFs as well as the ADS index and the LFFs is the model with the best fit. In the second regime, the MIDAS model with the CF, HFFs, and NFCI yields the lowest BIC for the U.S. corporate bond default rate. Summarizing [Table 8](#) shows that while the CF Granger causes the real GDP and consumption of services and nondurable goods growth in the two regimes, its predictive estimated effect is much lower in the post-GM regime. These findings not only extend to the case of the default spread but the results from the alternative models show that the Granger causality role of the CF is much weaker in the post-GM regime relative to the pre-GM period for this spread.

Turning to [Table 9](#), we report the results for predictive models for the $\log(\text{VIX})$, VRP, and the ETF iShares Core S&P500 returns during the second regime and based on data availability. Within these predictive models, we also consider the Realized Skewness (SKEW) proposed in [Amaya et al. \(2015\)](#) which turns out to be a significant predictor in the log specification of a model for ETF returns along with other HFFs. The reported results in [Table 9](#) provide two broad conclusions: first, for these three key financial indicators, the models that yield the lowest BIC are MIDAS specifications, which include a subset of our MFFs along with the excess market returns for the VIX and the NFCI for the VRP. Second, even if the CF is not driven by any stock market indices (as discussed in the previous subsection) and even though in the post-GM the VXO is no longer closely related with the CF, the evidence in [Table 9](#) shows that in the recent regime in many models the CF is a strongly significant predictor Granger causing the VIX, VRP, and ETF returns in the last three decades. Given the recent financial crisis, we add a simple dummy variable in the constant of all the AR, FADL, and FADL-MIDAS models (which takes the value one during 2008q4 and zero otherwise) and find that while this is significant for almost all models for all dependent variables (except for real consumption), it does not affect the significance of the predictive regression coefficient of the CF and the models selected by BIC. Hence, the results reported in [Tables 8](#) and [9](#) are robust to excluding the recent global financial crisis.

In comparing the CFFs in linear-LS models with our MFFs in either linear-LS or MIDAS-NLS models, we find that the MFFs perform relatively better in terms of BIC for all variables (except for real GDP growth in the pre-GM period). Comparing the model

results in rows 2–6 in [Tables 8](#) and [9](#) for each variable and regime, we find that the models with MFFs provide the best fit (in terms of BIC) when combined with MIDAS instead of linear-LS specifications especially in the second regime. Hence our empirical evidence suggests that it is the combination of the information content of the MFFs as well as their role in MIDAS predictive regressions vis-à-vis that of the CFFs (in linear-LS models) that provides goodness of fit improvements. Yet, what is not obvious to isolate and infer in predictive models with CFFs, as opposed to those with MFFs, is the role of the CF as shown in either linear or MIDAS models with the CF.

Summarizing, the large dimensional empirical analysis reduces the dimensionality of the data via mixed-frequency group factor models, yet we still face the large model space of alternative factors, predictors, and predictive model specifications. We find that model selection approaches favor MIDAS predictive regression models and mixed-frequency group factors as predictors of key macro and financial variables. The overall results in [Tables 8](#) and [9](#) show that the lowest BIC favors MIDAS specifications with some of our mixed-frequency group factors (i.e., CF, HFFs, and LFFs) mainly for the second regime, as opposed to the traditional CFFs. Additionally, the ADS, the NFCI, and excess market returns turn out to be additional significant predictors, among the aforementioned factors for models with the lowest BIC. For instance, the ADS factor turns out to be significant in the models with the best fit (based on BIC) for the real GDP growth and the default spread in the pre-GM period and for the real Consumption growth for the post-GM period. This is an interesting finding given that both the MFFs and ADS are based on the idea of deriving factors from mixed-frequency grouped data, albeit of different sizes and types of cross-sectional information. For the VIX, VRP, and ETF returns, we find that our HFFs (along with other factors such as the CF, LFF, NFCI, excess market returns) yield predictive models with the lowest BIC.

3.3.2 OOS predictive evidence

We analyze the OOS predictive ability of our factors (MFFs and CFFs) reporting in [Table 10](#) the root mean squared errors (RMSE) ratios of linear and MIDAS models vis-à-vis the random walk (RW) model, often considered as a simple benchmark model for both macro and financial indicators. Given the evidence of structural change, we focus on evaluating the forecasting performance of the models in the more recent post-GM and longer period. Panel A refers to the results for the three variables that have the longer sample period (Real GDP, Consumption, and Baa-Aaa), whereas Panel B refers to the financial variables with the shorter sample (VIX, VRP, and ETF returns). The IS period for the models in Panel A refer to 1986q1–2001q4, while for the variables in Panel B, namely for the VIX and VRP, the IS period is 1990q1–2003q4 and for the ETF returns it is 2000q3–2007q1, due to the shorter data availability. For the factor augmented predictive models (reported in each row of [Table 10](#)), we focus on evaluating the forecasting ability of the corresponding model (in each row) with the significant predictors during the IS period, given these are more parsimonious representations than the model with all factors. For the OOS forecasting evaluation, we pursue two approaches, the fixed and the recursive sample schemes. We report the results one-quarter ahead given the sample sizes. While the fixed OOS approach is pursued in many studies for evaluating macro forecasting models, the recursive OOS approach is more realistic especially for financial data as it is not subject to the look ahead bias criticism. The MSE-F test ([Gonçalves, McCracken, and Perron, 2017](#)) is performed to

Table 10 OOS forecasting ability of MFFs, CFFs, as well as other types of factors

Predictive model specifications	GDP		RealCons		Baa-Aaa	
	Fixed sample	Recursive sample	Fixed sample	Recursive sample	Fixed sample	Recursive sample
LD	1.00		1.00		1.03	
LD, all CFFs [C_{p2}]	0.88*	0.81*	0.79*	0.61*	1.15	0.57*
LD, CFFs in Linear-LS	0.77*	0.83*	0.55*	0.65*	1.00	0.59*
LD, MFFs in Linear-LS	0.82*	0.81*	0.59*	0.79*	0.89*	0.52*
LD, ADS, MFFs in Linear-LS	0.64*	0.69*	0.59*	0.72*	0.96*	0.54*
LD, CFNAI, MFFs in Linear-LS	0.81*	0.78*	0.59*	0.79*	0.96*	0.54*
LD, NFCl, MFFs in Linear-LS	0.86*	0.79*	0.59*	0.79*	0.89*	0.52*
LD, GZ_SPRD, MFFs in Linear-LS	0.82*	0.81*	0.59*	0.79*	0.77*	0.60*
LD, FFs, MFFs in Linear-LS	0.78*	0.77*	0.53*	0.82*	0.91*	0.57*

Predictive model specifications	GDP		RealCons		Baa-Aaa	
	Fixed sample	Recursive sample	Fixed sample	Recursive sample	Fixed sample	Recursive sample
LD, MFFs in Almon	0.74*	0.79*	0.47*	0.93*	0.91*	0.62*
LD, ADS, MFFs in Almon	0.58*	0.63*	0.49*	0.67*	0.94*	0.51*
LD, CFNAI, MFFs in Almon	0.59*	0.64*	0.44*	0.73*	0.96*	0.58*
LD, NFCl, MFFs in Almon	0.77*	0.79*	0.49*	0.88*	0.96*	0.58*
LD, GZ_SPRD, MFFs in Almon	0.77*	0.80*	0.47*	0.93*	0.91*	0.41*
LD, FFs, MFFs in Almon MIDAS	0.90*	0.80*	0.52*	0.93*	0.91*	0.58*

Panel A1: RMSE ratios for Linear-LS models vis-à-vis the RW model for Real GDP growth (GDP), Real Consumption growth (RealCons), and Moody's default spread (Baa-Aaa)

Panel A2: RMSE ratios for Almon MIDAS models vis-à-vis the RW model for Real GDP growth (GDP), Real Consumption growth (RealCons), and Moody's default spread (Baa-Aaa)

Table 10 Continued

Predictive model specifications	logVIX		VRP		ETF_iSHARES	
	Fixed sample	Recursive sample	Fixed sample	Recursive sample	Fixed sample	Recursive sample
LD	0.99		1.00		1.00	
LD, all CFFs [IC_{p2}]	1.76	0.70*	0.94*	0.97*	1.36	1.53
LD, CFFs in Linear-LS	0.87*	0.70*	0.93*	0.98	0.98	1.07
LD, MFFs in Linear-LS	1.20	0.76*	0.96*	0.93*	0.98	1.26
LD, ADS, MFFs in Linear-LS	1.20	0.76*	0.99	0.98	0.85*	1.19
LD, CFNAI, MFFs in Linear-LS	1.20	0.76*	0.96*	0.93*	0.98	1.26
LD, NFCL, MFFs in Linear-LS	1.20	0.76*	0.94*	1.04	2.07	1.46
LD, GZ_SPRD, MFFs in Linear-LS	1.20	0.77*	0.97	1.03	1.01	1.27
LD, FFs, MFFs in Linear-LS	1.16	0.72*	1.25	1.04	0.98	1.26
LD, SKEW, MFFs in Linear-LS	1.06	0.74*	0.94*	0.92*	1.00	1.27

Panel B2: RMSE ratios for Almon MIDAS models vis-à-vis the RW model for VIX (log VIX), VRP, and ETF iShares Core S&P500 (ETF_iSHARES)						
Predictive Model Specifications	Fixed sample		Recursive sample		Recursive sample	
	Fixed sample	Recursive sample	Fixed sample	Recursive sample	Fixed sample	Recursive sample
LD, MFFs in Almon	0.98	0.78*	0.91*	0.96*	1.26	1.84
LD, ADS, MFFs in Almon	0.98	0.78*	0.92*	0.98	0.86*	1.84
LD, CFNAI, MFFs in Almon	0.98	0.78*	0.93*	0.99	1.26	1.84
LD, NFCL, MFFs in Almon	1.08	0.80*	0.91*	1.08	1.26	1.84
LD, GZ_SPRD, MFFs in Almon	0.77*	0.75*	0.91*	0.96*	1.26	1.84
LD, FFs, MFFs in Almon MIDAS	0.78*	0.54*	0.99	0.97*	0.78*	1.08
LD, SKEW, MFFs in Almon	1.11	0.78*	0.91*	0.96*	1.31	2.00

Notes: The IS period refers to 1986q1–2001q4 for the Real GDP growth (GDP), Real Consumption growth (RealCons), and the Moody's default spread (Baa-Aaa), 1990q1–2003q4 for the VIX (logVIX) and VRP, and 2000q3–2007q1 for the ETF iShares Core S&P500 (ETF_iSHARES). The OOS period refers to 2002q1–2017q4 for the GDP, RealCons and Baa-Aaa, 2004q1–2017q4 for the logVIX and VRP, and 2007q2–2017q4 for the ETF_iSHARES. For the GZ_SPRD, the end date is 2016M08 hence IS period is 1986q1–2001q2 and OOS period 2001q3–2016q3. The OOS analysis is performed based on a fixed sample as well as the recursive estimation of the factors in pseudo real-time at each forecast origin using the recent vintage of data. * denotes the cases where the MSE-F statistic is greater than the critical values at 5% significance level which implies that the competing model as specified in the first column performs significantly better than the RW benchmark model. Bold RMSE ratios refer to the min RMSE across all models for a given dependent variable. The predictive model involves the following predictors: The lagged dependent (LD) term, the MFFs, namely the Common Factor (CF), the HFFs and LFFs, the CFFs, the *Arno*, *Diebold*, and *Scotti* (2009) (ADS) index, the Chicago National Activity and Financial Conditions Indices (CFNAI and NFCL, respectively), the four Fama–French (FF) factors (RMI-RF, SMB, HML, UMD), *Gilchrist and Zakrajsek* (2012) spread (GZ_SPRD) and the *Amaya, Christoffersen, Jacobs, and Vasquez* (2015) Realized Skewness (SKEW).

evaluate if the factor augmented predictive regression models yield statistically significant predictive gains vis-à-vis other benchmark models like the RW (or the AR reported in the first row of Panels A1 and B1), as marked in Table 10. Similarly, the model with the lowest and significant RMSE ratio for each dependent variable and each family of models and forecasting scheme is marked in bold in Table 10.

Four broad results can be highlighted from Table 10: first, the RMSE ratios of almost all factor augmented predictive models (linear or MIDAS) vis-à-vis the RW or the AR model (given it has the same RMSE as the RW), for forecasting the real GDP growth, the real Consumption growth and the default spread, are statistically significant and less than one (shown in Panels A1 and A2). This is also the case for most (but not all) the models for the VIX, VRP, and ETF returns, which seem harder to forecast OOS. This could be due to the nature of these financial series and/or the shorter sample available. Second, for the best performing model with the lowest RMSE ratio, the recursive OOS approach yields similar and in some cases improved forecasting gains vis-à-vis the fixed sample OOS scheme for all variables, except the ETF returns. Interestingly, although the recursive method is a more demanding OOS forecasting scheme, it not only improves over the fixed sample OOS but for at least two cases—namely the default spread and the VIX—it also substantially improves the forecasting gains, as shown by the corresponding lowest RMSE ratios. Third, the models with the lowest RMSE ratios refer to models with our MFFs rather than the traditional CFFs as predictors, in almost all cases. Additional factors turn out to improve the recursive OOS forecasts, namely the ADS index for the GDP, the CFNAI for the Consumption, the NFCI and GZ spread for the corporate default spread, the excess market returns for the VIX, along with some of our MFFs. Finally, in most cases, the best performing FADL-MIDAS models (with the lowest RMSE) provide forecasting gains over the corresponding best performing FADL linear-LS model (comparing the corresponding best RMSE models in Panels A1 and A2 or Panels B1 and B2).

To further investigate the OOS forecasting performance of the models in the recent post-GM period, we employ the Model Confidence Set (MCS) procedure developed by Hansen, Lunde, and Nason (2011). In line with Table 10, we use the significant predictors for each model and focus on the recursive scheme. In Table 11, we report the RMSE, the p -value and the ranking of each model. Following Hansen, Lunde, and Nason (2011), we use 75% and 90% confidence levels. The MCS results are consistent with those in Table 10 and show that for all variables at least one of the Linear-LS and PDL/Almon MIDAS models perform better than RW, which in most cases ranks among the worst predictive models.

Concluding, we provide evidence that for quarterly real GDP and Consumption growth, the Moody's corporate bond spread as well as the VIX, VRP, and ETF returns, during the pre- and post-GM periods, our CF, as well as the group-specific macro and financial factors have significant IS and OOS forecasting abilities. This evidence is especially strong in the context of MIDAS predictive regressions, vis-à-vis the traditional linear predictive regressions, for example, FADL type models, as well as the RW and AR benchmark models. Moreover, comparing the role of the CF, during the pre- and post-GM, in Granger causing GDP or Consumption growth as well as the default spread, we find that while the CF is significant in both subperiods, it has a relatively smaller estimated coefficient in the post-GM period. Hence, during the recent period, the role of the CF, while still significant, turns out to be relatively much weaker in Granger causing the aforementioned key economic variables. These predictive regression models results extend the evidence of a break in the

Table 11 MCS by Hansen, Lunde, and Nason (2011) for OOS forecasting ability of mixed-frequency factors, CFFs, as well as other types of factors

Panel A1: RMSE for Linear-LS models and the RW model for Real GDP growth (GDP), Real Consumption growth (RealCons), and Moody’s default spread (Baa-Aaa)

Specification	GDP			RealCons			Baa-Aaa		
	Recursive sample			Recursive sample			Recursive sample		
	RMSE	p-value	Rank	RMSE	p-value	Rank	RMSE	p-value	Rank
LD, all CFFs [ICP2]	0.51	0.73**	4	0.24	1.00**	1	0.27	0.82**	2
LD, CFFs in Linear-LS	0.52	0.62**	5	0.25	0.94**	2	0.28	0.57*	3
LD, MFFs in Linear-LS	0.51	0.49**	7	0.31	0.32*	3	0.25	1.00**	1
LD, ADS, MFFs in Linear-LS	0.44	1.00**	1	0.28	0.22*	6	0.26	0.27*	6
LD, CFNAI, MFFs in Linear-LS	0.49	0.54**	6	0.32	0.29*	4	0.26	0.27*	6
LD, NFCI, MFFs in Linear-LS	0.50	0.36**	8	0.28	0.22*	6	0.25	1.00**	1
LD, GZ_SPRD, MFFs in Linear-LS	0.48	0.78**	3	0.28	0.22*	6	0.38	0.45*	4
LD, FFs, MFFs in Linear-LS	0.51	0.49**	7	0.28	0.22*	6	0.27	0.18*	7
RW	0.63	0.86**	2	0.39	0.29*	5	0.48	0.36*	5

Panel A2: RMSE for PDL/Almon MIDAS models and the RW for Real GDP growth (GDP), Real Consumption growth (RealCons), and Moody’s default spread (Baa-Aaa)

Specification	Recursive sample			Recursive sample			Recursive sample		
	RMSE	p-value	Rank	RMSE	p-value	Rank	RMSE	p-value	Rank
LD, MFFs in PDL/Almon	0.50	0.29*	5	0.37	0.43*	3	0.31	0.15*	3
LD, ADS, MFFs in PDL/Almon	0.40	1.00**	1	0.26	1.00**	1	0.25	0.12*	5
LD, CFNAI, MFFs in PDL/Almon	0.41	0.86**	2	0.29	0.06	5	0.29	0.13*	4
LD, NFCI, MFFs in PDL/Almon	0.50	0.29*	5	0.35	0.77**	2	0.29	0.13*	4
LD, GZ_SPRD, MFFs in PDL/Almon	0.51	0.31*	4	0.37	0.43*	3	0.20	1.00**	1
LD, FFs, MFFs in PDL/Almon MIDAS	0.51	0.15*	6	0.37	0.43*	3	0.29	0.13*	4
RW	0.63	0.51*	3	0.39	0.21*	4	0.49	0.17*	2

Panel B1: RMSE for Linear-LS models and the RW for VIX (logVIX), Variance Risk Premium (VRP), and ETF iShares (ETF_iShares)

Specification	logVIX			VRP			ETF_iShares		
	Recursive sample			Recursive sample			Recursive sample		
	RMSE	p-value	Rank	RMSE	p-value	Rank	RMSE	p-value	Rank
LD, all CFFs [ICP2]	0.11	1.00**	2	15.57	0.85**	5	0.102	0.64**	5
LD, CFFs in Linear-LS	0.11	1.00**	1	15.67	0.94**	4	0.072	0.15*	8
LD, MFFs in Linear-LS	0.12	0.65**	5	14.88	1.00**	2	0.083	0.94**	3
LD, ADS, MFFs in Linear-LS	0.12	0.65**	5	15.74	0.24*	9	0.077	0.99**	2
LD, CFNAI, MFFs in Linear-LS	0.12	0.65**	5	14.88	1.00**	2	0.083	0.94**	3
LD, NFCI, MFFs in Linear-LS	0.12	0.65**	5	16.74	0.79**	7	0.096	0.59**	6
LD, GZ_SPRD, MFFs in Linear-LS	0.12	0.64**	6	16.50	0.72**	8	0.083	0.51**	7
LD, FFs, MFFs in Linear-LS	0.12	1.00**	3	16.67	0.86**	6	0.083	0.94**	3
LD, SKEW, MFFs in Linear-LS	0.12	0.95**	4	14.87	1.00**	1	0.084	0.91**	4
RW	0.16	0.08	7	16.02	0.98**	3	0.066	1.00**	1

Table 11 Continued

Panel B2: RMSE for PDL/Almon MIDAS models and the RW for VIX (logVIX), VRP, and ETF iShares (ETF_iShares)

Specification	Recursive sample			Recursive sample			Recursive sample		
	RMSE	<i>p</i> -value	Rank	RMSE	<i>p</i> -value	Rank	RMSE	<i>p</i> -value	Rank
LD, MFFs in PDL/Almon	0.13	0.00	3	15.34	1.00**	1	0.14	0.00	4
LD, ADS, MFFs in PDL/Almon	0.13	0.00	3	15.62	0.90**	3	0.14	0.00	4
LD, CFNAI, MFFs in PDL/Almon	0.13	0.00	3	15.83	0.90**	4	0.14	0.00	4
LD, NFICI, MFFs in PDL/Almon	0.13	0.00	4	17.31	0.40**	6	0.14	0.00	4
LD, GZ_SPRD, MFFs in PDL/Almon	0.12	0.03	2	15.34	1.00**	1	0.14	0.00	4
LD, FFs, MFFs in PDL/Almon MIDAS	0.09	1.00**	1	15.49	1.00**	2	0.08	0.56	2
LD, SKEW, MFFs in PDL/Almon	0.13	0.00	3	15.34	1.00**	1	0.15	0.17	3
RW	0.16	0.00	5	15.98	0.84**	5	0.07	1.00**	1

Notes: The IS period refers to 1986q1–2001q4 for the Real GDP growth (GDP), Real Consumption growth (RealCons) and the Moody's default spread (Baa-Aaa), 1990q1–2003q4 for the VIX (logVIX) and VRP, and 2000q3–2007q1 for the ETF iShares Core S&P500 (ETF_iSHARES). The OOS period refers to 2002q1–2017q4 for the GDP, RealCons, and Baa-Aaa, 2004q1–2017q4 for the logVIX and VRP, and 2007q2–2017q4 for the ETF_iSHARES. For the GZ_SPRD, the end date is 2016M08 hence IS period is 1986q1–2001q2 and OOS period is 2001q3–2016q3. The OOS analysis is performed based on the recursive estimation of the factors in pseudo real-time at each forecast origin using the recent vintage of data. Bold values refer to the models ranking first according to MCS. The results based on significance level 10% and 25% are identified by (*) and (**), respectively. RMSEs of GDP and RealCons are multiplied by 100.

loadings of the factor models (reported in the previous subsection), to the conditional setup related to the estimated impact of the CF as a predictor of key macro and financial variables.

4 Conclusions

This article contributes further to our understanding of group factor models for extracting PCs from large panels of mixed data frequencies, allowing us to estimate and test for the existence of the common factor among the groups as well as the group-specific factors. New analytical results are derived for the asymptotic distribution of the PCs and test statistics for the existence of the common factor especially regarding the alternative approaches of aggregating the data first and then extracting PCs, or applying PCA first and then aggregating the factor estimates. Our framework provides an interesting setup to study the common factors among two large groups/panels of quarterly macro and monthly financial indicators, in order to test for the existence of a CF as well as estimate the group-specific financial and macro factors. Interestingly, we find one CF in the United States during the pre- and post-GM, since the early 1960s. Structural break analysis reveals that the loadings of the CF have changed during the pre- and post-GM and that the loadings of certain financial

assets have become relatively weaker during the recent regime. Our empirical analysis shows that the estimated PCs are almost identical whether we pursue the PCA approach first and then aggregate the factors or whether we aggregate the data and then apply PCA. The forecasting role of our factors, as well as other established factors in the literature, is further investigated in predicting key macro and financial indicators, such as real GDP and Consumption, the VIX, the VRP, corporate bond default spreads, and ETF iShares Core S&P500 returns, via FADL and FADL-MIDAS type models. Our empirical results provide evidence of significant forecasting gains of our factors for these key economic indicators and show that the CF, while being significant in both regimes, has a weaker predictive effect in the recent period covering the last three decades.

Supplemental Data

Supplemental data is available at *Journal of Financial Econometrics* online.

References

- Ahn, S. C., and A. R. Horenstein. 2013. Eigenvalue Ratio Test for the Number of Factors. *Econometrica* 81: 1203–1227.
- Amaya, D., P. Christoffersen, K. Jacobs, and A. Vasquez. 2015. Does Realized Skewness Predict the Cross-Section of Equity Returns?. *Journal of Financial Economics* 118: 135–167.
- Ando, T., and J. Bai. 2015. Multifactor Asset Pricing with a Large Number of Observable Risk Factors and Unobservable Common and Group-Specific Factors. *Journal of Financial Econometrics* 13: 556–604.
- Andreou, E., P. Gagliardini, E. Ghysels, and M. Rubin. 2019. Inference in Group Factor Models with an Application to Mixed-Frequency Data. *Econometrica* 87: 1267–1305.
- Aruoba, S., F. Diebold, and C. Scotti. 2009. Real-Time Measurement of Business Conditions. *Journal of Business & Economic Statistics* 27: 417–427.
- Bai, J., and S. Ng. 2002. Determining the Number of Factors in Approximate Factor Models. *Econometrica* 70: 191–221.
- Bai, J., and S. Ng. 2008. Forecasting Economic Time Series Using Targeted Predictors. *Journal of Econometrics* 146: 304–317.
- Bates, J. M., and C. W. Granger. 1969. The Combination of Forecasts. *Journal of the Operational Research Society* 20: 451–468.
- Brave, S., and R. A. Butters. 2012. Diagnosing the Financial System: Financial Conditions and Financial Stress. *International Journal of Central Banking* 8: 191–239.
- Brave, S., and R. A. Butters. 2014. Nowcasting Using the Chicago Fed National Activity Index. *Economic Perspectives* 38: 19–37.
- Breitung, J., and S. Eickmeier. 2011. Testing for Structural Breaks in Dynamic Factor Models. *Journal of Econometrics* 163: 71–84.
- Breitung, J., and S. Eickmeier. 2016. Analyzing International Business and Financial Cycles Using Multi-Level Factor Models: A Comparison of Alternative Approaches. *Advances in Econometrics* 15: 177–214.
- Chaise, S. D., L. Ferrara, and D. Giannone. 2017. “Common Factors of Commodity Prices.” Working paper 645, Banque de France.
- Chen, P. 2012. “Common Factors and Specific Factors.” Working Paper.
- Christoffersen, P., C. Dorion, K. Jacobs, and L. Karoui. 2014. Nonlinear Filtering in Affine Term Structure Models. *Management Science* 60: 2248–2268.

- Christoffersen, P., M. Fournier, and K. Jacobs. 2017. The Factor Structure in Equity Options. *The Review of Financial Studies* 31: 595–637.
- Christoffersen, P., E. Ghysels, and N. R. Swanson. 2002. Let's Get "Real" about Using Economic Data. *Journal of Empirical Finance* 9: 343–360.
- Christoffersen, P., and H. Langlois. 2013. The Joint Dynamics of Equity Market Factors. *Journal of Financial and Quantitative Analysis* 48: 1371–1404.
- Drehmann, M., C. Borio, and K. Tsatsaronis. 2012. "Characterising the Financial Cycle: Don't Lose Sight of the Medium Term!." BIS Working Papers 380, Bank for International Settlements.
- Gilchrist, S., and E. Zakrajsek. 2012. Credit Spreads and Business Cycle Fluctuations. *American Economic Review* 102: 1692–1720.
- Gonçalves, S., M. W. McCracken, and B. Perron. 2017. Tests of Equal Accuracy for Nested Models with Estimated Factors. *Journal of Econometrics* 198: 231–252.
- Gospodinov, N., and S. Ng. 2013. Commodity Prices, Convenience Yields, and Inflation. *Review of Economics and Statistics* 95: 206–219.
- Goyal, A., C. Pérignon, and C. Villa. 2008. How Common Are Common Return Factors across the NYSE and Nasdaq?. *Journal of Financial Economics* 90: 252–271.
- Hansen, P., A. Lunde, and J. Nason. 2011. The Model Confidence Set. *Econometrica* 79: 453–497.
- Kose, A. M., C. Otrok, and C. H. Whiteman. 2008. Understanding the Evolution of World Business Cycles. *Journal of International Economics* 75: 110–130.
- McCracken, M., and S. Ng. 2016. FRED-MD: A Monthly Database for Macroeconomic Research. *Journal of Business & Economic Statistics* 36: 574–589.
- Onatski, A. 2010. Determining the Number of Factors from Empirical Distribution of Eigenvalues. *Review of Economics and Statistics* 92: 1004–1016.
- Stock, J., and M. Watson. 2002. Forecasting Using Principal Components from a Large Number of Predictors. *Journal of the American Statistical Association* 97: 1167–1179.
- Stock, J., and M. Watson. 2008. "Forecasting in Dynamic Factor Models Subject to Structural Instability." In N. Shephard and J. Castle (eds.), *The Methodology and Practice of Econometrics: Festschrift in Honor of D.F. Hendry*, pp. 1–57. Oxford University Press.
- Timmermann, A. 2006. "Forecast Combinations." In G. Elliott, C. Granger, and A. Timmermann (eds.), *Handbook of Economic Forecasting*, Volume 1, pp. 136–96. Elsevier.
- Wang, P. 2012. "Large Dimensional Factor Models with a Multi-Level Factor Structure: Identification, Estimation, and Inference." Working Paper, Hong Kong University of Science and Technology.
- Zhou, H. 2018. Variance Risk Premia, Asset Predictability Puzzles, and Macroeconomic Uncertainty. *Annual Review of Financial Economics* 10: 481–497.

Implied Default Probabilities and Losses Given Default from Option Prices*

Jennifer Conrad¹, Robert F. Dittmar², and Allaudeen Hameed³

¹Kenan-Flagler Business School, University of North Carolina, ²Ross School of Business, University of Michigan and ³NUS Business School, National University of Singapore

Address correspondence to Jennifer Conrad, Department of Finance, Kenan-Flagler Business School, University of North Carolina Chapel Hill, North Carolina, or e-mail: j_conrad@kenan-flagler.unc.edu.

Received May 21, 2019; revised May 21, 2019; editorial decision April 16, 2020; accepted April 16, 2020

Abstract

We propose a novel method of estimating default probabilities using equity options data. The resulting default probabilities are highly correlated with estimates of default probabilities extracted from CDS spreads, which assume constant losses given default. Additionally, the option-implied default probabilities are higher in bad economic times and for firms with poorer credit ratings and financial positions. A simple inferred measure of loss given default is related to underlying business conditions, and varies across sectors; the time series properties of this measure are similar after controlling for liquidity effects.

Key words: contingent pricing, default probabilities, recovery rates

JEL classification: G12, G13

From the perspective of academics, market professionals, and regulators, one of the attractive features of a credit default swap (CDS) contract is its window into market perceptions of credit risk. Based on no-arbitrage pricing formulations and assumptions about the recovery rate on the asset, one can use information in the quoted spread on a CDS contract to

* This paper has benefitted from the comments of Xudong An, Nina Boyarchenko, Davie Henn, Ken Singleton, Yin-Hua Yeh, and Adam Zawadowski, as well as seminar participants at the 2010 Financial Economics in Rio conference at FGV Rio de Janeiro, the 2011 FMA Asian conference, the 2012 American Finance Association conference, the 2012 European Finance Association conference, the 2013 NUS-RMI conference, the 2017 China International Review conference, Aalto, Arizona State, Georgetown, Goethe, Indiana, Purdue, and Rice Universities, Hong Kong University of Science and Technology, the Stockholm School of Economics, and the Universities of British Columbia, Mannheim, Toronto, Utah, and Western Ontario. Hameed gratefully acknowledges financial support from NUS Academic Research Grant. All errors are the responsibility of the authors.

infer the market's implied risk-neutral probability of default; see, for example, [Duffie and Singleton \(1999\)](#). This measure stands as a market-based nonparametric alternative to agency credit ratings and structural models of default. However, the market for CDS reached its peak in late 2007; since that time, this market has substantially decreased in size.¹ Recent market events suggest that the reduction in the size of the CDS market will not be reversed soon, as noted in [Augustin et al. \(2016\)](#). Thus, the decline in the single-name CDS market represents a loss for those interested in market-driven estimates regarding the credit risk of a corporate entity.

In this article, we propose a novel measure of the risk-neutral probability of default based on option prices. It is well known that option prices are informative about the risk-neutral distribution of equity payoffs; see, for example, [Breedon and Litzenberger \(1978\)](#). In addition, the equity payoff depends on default risk. In principle, if absolute priority holds, the value of equity will be zero in the case of a default as in [Merton \(1974\)](#). Alternatively, one can define a default region of the equity payoff distribution and use the cumulative probability of that region inferred from option prices to identify a probability of default. Importantly, an option-based approach has the advantage that, since options are traded on a large number of underlying securities, this method could provide estimates of default probability for a broader cross-section of firms.

Our results indicate that estimates of the levels of implied default probabilities extracted from equity options are strongly, but not perfectly, correlated with default probabilities estimated using CDS. If we assume constant loss given default (LGD) (as is frequently done in the CDS market), the median correlation between estimates of default probabilities for the cross-section of firms extracted from these two markets is 0.52. Aggregated across firms, the two estimates of default probabilities are highly correlated through time, with a correlation coefficient of 0.66. The default probabilities estimated from equity options prices increase monotonically with lower credit ratings; in addition, the relations between default probabilities estimated from equity options and firm characteristics are similar to the relations estimated between CDS default probabilities and firm characteristics. Overall, the evidence suggests that equity options can provide important information concerning the probability of default for a broad spectrum of underlying firms.

We also investigate whether the imperfect correlation between option- and CDS-implied default probabilities reflects the fact that estimates of LGD embedded in CDS rates vary across firms and through time, as argued in [Berndt et al. \(2018\)](#), [Doshi, Elkamhi, and Ornthanalai \(2018\)](#), [Schuermann \(2004\)](#), and [Altman et al. \(2005\)](#). We find that a simple proxy for LGD, defined as the ratio of the CDS spread to option-implied default probability, covaries positively with the frequency of default in the aggregate sample; the ratio is high during the early 2000s and the financial crisis, declines in the mid-2000s and after the financial crisis, and has been relatively low during 2013–2017. While some of this difference in the option-implied default probability and CDS spread is related to measures of illiquidity, the time series patterns in this variable remain after removing variation due to illiquidity.

Our article is related to the literature investigating the ability of option prices to provide information about default. In an earlier paper whose intuition is closely related to our work, [Le \(2015\)](#) develops models of option and CDS prices and uses those models

1 The Bank for International Settlements reported notional principal outstanding in the last half of 2007 of \$61 trillion; in the first half of 2019, this figure stood at \$7.8 trillion.

sequentially to estimate default probabilities and recovery rates. His method of estimating default probabilities differs from ours; for example, he assumes a specific process for the dynamics of equity prices and estimates the probability of default, where default is associated with equity prices diffusing or jumping to zero. In contrast, we estimate the risk-neutral density at a particular point in time from a range of option prices and consider various default thresholds. Moreover, our sample (which extends through February 2017) enables us to document the time series properties of LGD over a longer period, including the credit crisis. [Capuano \(2008\)](#) uses a cross-entropy functional to infer default probabilities, although [Vilsmeier \(2014\)](#) notes that the entropy approach has issues with accuracy and numerical stability (and provides some technical fixes for those problems); our approach is simpler computationally. [Carr and Wu \(2011\)](#) show that deep out-of-the-money (OTM) options can be used to synthesize a default insurance contract, and, as a consequence, infer the probability of default. Their approach is simple and intuitive, but necessitates the existence of options that are very deep out of the money, limiting the number of firms for which these probabilities can be calculated.

This article is also related to others that have inferred LGDs or other credit-related measures from the market prices of various securities; the papers in this literature differ with regard to the securities that are required for estimation of LGD, and, in some cases, the processes that rates of LGD are assumed to follow. [Duffie and Singleton \(1999\)](#) and [Das and Sundaram \(2007\)](#) provide examples of how LGD might be inferred from securities with the same probability of default but different payout structure or priority. Similarly, [Madan and Unal \(1998\)](#) empirically investigate the separation of default probability and LGD using junior and senior debt prices where those are available; [Madan and Güntay \(2003\)](#) also exploit differences in debt priority to infer LGDs. [Doshi, Elkamhi, and Ornthanalai \(2018\)](#), as noted above, use the term structure of CDS to estimate LGD. [Bakshi, Madan, and Zhang \(2006\)](#) use a risky debt model with stochastic rates of LGD to infer measures of LGD from risky bond prices. More recently, [Berndt et al. \(2018\)](#) combine information from CDS rates, firm-specific default estimates from Markit, and information regarding expected default from Moody's Analytics to estimate credit risk premia.

The remainder of the article is organized as follows. In Section 1, we discuss the methodology we employ for extracting risk-neutral default probabilities from options and from CDS spreads. We describe the data that we employ in this article in Section 2, and present estimates of default probabilities and their relation to various firm characteristics. In Section 3, we explore the differences between the two measures of default probability, including cross-sectional and time series variation in LGD and liquidity effects. We conclude in Section 4.

1 Risk-Neutral Probabilities Implied by CDS and Option Prices

1.1 Pricing CDSs

In order to infer risk-neutral default probabilities from the prices of CDSs, we follow a model widely used in practice for their valuation, detailed in [O'Kane and Turnbull \(2003\)](#). In the discussion that follows, we assume that the swap being valued is a one-year CDS contract with quarterly premium payments, and that there is no information on CDS from which to infer risk-neutral default probabilities for horizons of less than one year. Under

these assumptions, the practice is to assume a flat default probability term structure over the year.²

When a CDS contract is struck, the swap premium is set such that the value of the premium leg, received by the writer of the swap, is equal to the value of the protection leg, received by the swap purchaser. Assuming that premiums are accrued in case of default during a quarter, the value of the premium leg is given by

$$\frac{1}{2} s_t \sum_{j=1}^4 0.25 e^{-r(\frac{j}{4})\frac{j}{4}} (e^{-\lambda_t \frac{j-1}{4}} + e^{-\lambda_t \frac{j}{4}}), \quad (1)$$

where $r(\tau)$ is the continuously compounded zero coupon Treasury yield with maturity τ expressed as an annual rate and λ_t is the default intensity. Because of the assumption of a flat term structure of default probability over one year, this intensity is invariant to maturity for the one-year horizon, although it is indexed by t to indicate that default intensity may change over time. The quantity $e^{-\lambda_t \tau}$ represents the risk-neutral probability that the entity survives to time τ . Intuitively, expression (1) simply calculates the present value of the swap payments received by the swap writer, conditional on survival of the entity.

The value of the protection leg is the risk-neutral expected loss on the CDS,

$$(LGD) \sum_{j=1}^{12} e^{-r(\frac{j}{12})\frac{j}{12}} (e^{-\lambda_t \frac{j-1}{12}} - e^{-\lambda_t \frac{j}{12}}), \quad (2)$$

where *LGD* designates the loss given default as a fraction of the amount owed. Expression (2) is a discrete approximation to an integral that represents the expected risk-neutral loss on the underlying entity. O'Kane and Turnbull (2003) show that for a constant default intensity, the approximation error is given by $\frac{r(\tau)}{2M}$, where $r(\tau)$ is the continuously compounded risk free rate over the constant default intensity horizon and M is the number of summation periods. The authors suggest that for $M = 12$ as above and a risk-free rate of 3%, the absolute value of the error is 1 basis point on a spread of 800 basis points.

The above expressions allow for the determination of the break-even CDS spread if one has estimates of LGDs and risk-neutral survival probabilities. Alternatively, the expressions can be used to infer risk-neutral default probabilities given rates of LGD and constant maturity CDS spreads. For example, using the expressions above, the break-even CDS spread is given by

$$s_t = \frac{(LGD) \sum_{j=1}^{12} e^{-r(\frac{j}{12})\frac{j}{12}} (e^{-\lambda_t \frac{j-1}{12}} - e^{-\lambda_t \frac{j}{12}})}{\frac{1}{2} \sum_{j=1}^4 0.25 e^{-r(\frac{j}{4})\frac{j}{4}} (e^{-\lambda_t \frac{j-1}{4}} + e^{-\lambda_t \frac{j}{4}})}. \quad (3)$$

In the absence of any market frictions, Equation (3) is a nonlinear equation in the default intensity, λ_t .

For our initial calculations, we assume a constant rate of $LGD = 0.60$, consistent with common practice. Given the data on the risk-free term structure and one-year CDSs, we

2 Note that this valuation model is similar to the one described in Augustin and Saleh (2017), with the exception that we adjust for accrued interest.

then solve Equation (3) for λ_t for each reference entity and date in our sample. Given this default intensity, the probability of default of entity i is given by

$$Q_{i,t}^C = 1 - e^{-\lambda_{i,t}}, \tag{4}$$

where the superscript C indicates that CDS data are used to infer default probabilities.

1.2 Measuring Default Probabilities from Option Prices

An options-based approach to extracting default probabilities is based on the seminal work of Breeden and Litzenberger (1978), who show that one can recover the risk-neutral density of equity returns from option prices. Given the risk-neutral density, the risk-neutral probability of default can be thought of as the mass under the density up to the return that corresponds to a default event.

We construct the risk-neutral density using estimates of the risk-neutral moments computed as in Bakshi, Kapadia, and Madan (2003) and the Normal Inverse Gaussian (NIG) method developed in Eriksson, Ghysels, and Wang (2009). Specifically, Bakshi, Kapadia, and Madan (2003) (BKM) show that one can use traded option prices to compute estimates of the variance, skewness, and kurtosis of the risk-neutral distribution. These moments in turn serve as the inputs to the NIG method, which estimates the distribution. Eriksson, Ghysels, and Wang (2009) show that the NIG has several advantages to alternatives such as Gram–Charlier series expansions in pricing options. In particular, the distribution prevents negative probabilities, which the expansions can generate for the levels of skewness and kurtosis implied by option prices. The density is also known in closed form, avoiding the computational intensity of expansion approaches. Details of the estimation process are provided in Online Appendix A.

Once the risk-neutral distribution is estimated, we measure the probability of default for entity i at time t as the cumulative density of the NIG distribution at a default threshold α ,

$$Q_{i,t}^O(\tau) = \int_{-\infty}^{\alpha} f_{NIG}(x, \mathcal{E}_{i,t}(\tau), \mathcal{V}_{i,t}(\tau), \mathcal{S}_{i,t}(\tau), \mathcal{K}_{i,t}(\tau)) dx \tag{5}$$

where f_{NIG} is the NIG density function evaluated at a log return of x with parameters calculated as in BKM.³ The superscript O in Equation (5) indicates that the risk-neutral probability has been recovered from options data. The exact functional form of the density is provided in Online Appendix A.

A critical detail in this procedure is the definition of the default threshold, α . In the Merton (1974) model, equity has zero value in the case of default. However, the density at $x = \ln(0)$ cannot be calculated. Carr and Wu (2011) deal with this problem by assuming that there is a range of values for the stock price, $[A, B]$, in which default occurs. Specifically, prior to default, the equity value is assumed to be greater than B , and upon default the value is assumed to drop below the value $A \in [0, B)$. In their empirical implementation, the authors set $A = 0$, and choose the lowest priced put with positive bid price and positive open interest with strike price less than \$5 and option delta less than or equal to 15% in absolute value for their estimation.

3 In a few cases, our estimates of the kurtosis are too small given the calculated skewness. In order to calculate the cumulative density, it is necessary that $\mathcal{K}_{i,t} > 3 + \frac{5}{3} \mathcal{S}_{i,t}^2$. In cases in which this restriction is violated, we set the kurtosis to $\mathcal{K}_{i,t} = 3 + \frac{5}{3} \mathcal{S}_{i,t}^2 + 1e - 14$.

Table 1. Drop in equity prices over 12 months prior to bankruptcy

Rating	N	Mean	Std.	Min.	Max.
A	3	0.14	0.18	0.00	0.35
BBB	26	0.08	0.11	0.00	0.47
BB	67	0.13	0.20	0.00	0.98
B	197	0.12	0.16	0.00	0.95
CCC	59	0.20	0.22	0.00	0.84
CC	9	0.45	0.23	0.18	0.86
D	22	0.38	0.30	0.02	1.00

Notes: The table presents the magnitude of the drop in equity prices for firms that file bankruptcy over the previous 12 months. Bankruptcy filing dates are from the updated version of the data in [Chava and Jarrow \(2004\)](#) and are merged with CRSP data on equity prices and returns. We include only those firms that have CRSP data 12 months prior to the bankruptcy and in the month of bankruptcy filing and whose price declined over the 12 months prior to filing. The table presents summary statistics by credit rating for the decline in price by S&P credit rating for those firms for which ratings data are available. We present means, standard deviations, minima, and maxima of the ratio of the value of the stock price in the month of bankruptcy filing or delisting price to the price 12 months prior.

For our initial choice of the default threshold, we begin with an updated version of the data on bankruptcy filing dates from [Chava and Jarrow \(2004\)](#).⁴ We merge these data with data taken from the Center for Research in Security Prices (CRSP), and calculate the percentage decline in price ΔP from 12 months prior to the bankruptcy filing to either the delisting date or the CRSP price observed in the bankruptcy month.⁵

We examine the extent to which these declines are related to other measures of credit risk and to credit ratings. In this sample, there are 383 firms for which we have Standard and Poor's (S&P) long-term credit ratings for the borrowers 12 months prior to bankruptcy. The 12-month average declines in price for these firms are depicted by credit rating in [Table 1](#), where we group together all firms of a particular letter grade (i.e., "A", "A+", and "A-"). The table suggests that there is a clear relation between credit rating and price decline in the 12 months leading up to bankruptcy. Between "BBB"-rated and "CC"-rated firms, there is a near monotonic decline in losses. The prices of "BBB"-rated firms are on average 8% of their 12-month prior levels and the prices of "CC"-rated firms are 45% of their 12-month prior levels on average. The magnitude of price declines does not increase perfectly with credit rating; the average price decline of "A"-rated firms is higher than that of "BBB"-rated firms and the price decline of "BB"-rated firms is higher than that of "B"-rated firms.⁶ However, the overall pattern suggests that average price declines in the 12 months leading to bankruptcy filing decrease with credit rating.

4 Thanks to Sudheer Chava and Claus Schmitt for making these data available.

5 There are 1560 bankruptcy events in which the price declined over the previous 12 months, with an average decline in price of 79.40%. There are an additional 86 cases in which returns are positive over the 12-month period.

6 The average price declines of "A"-rated firms are driven by the very small number of "A"-rated firms (3) that have filed for bankruptcy. These firms are PG&E, with a stock price drop of 63.6%

Based on this evidence, for much of the analysis in the article, we let α , the default threshold, vary across credit ratings; we term this the Rating-derived α . According to S&P, from 1981 to 2015, no AAA-rated credit has defaulted, and defaults of “AA”-rated are rare. As a consequence, we assign $\alpha = 0.05$ to “AAA”-rated firms and $\alpha = 0.10$ to “AA”-rated firms, respectively. We assign $\alpha = 0.15$ to “A”-rated firms and increment α by 0.05 as we move to lower levels of credit ratings from “BBB” to “B”, consistent with the lower price declines associated with these firms. Finally, firms with ratings of “CCC” and below are assigned $\alpha = 0.35$. In our sample, these choices imply an average default threshold of $\alpha = 0.18$.

In robustness checks, we also compute default probabilities for each firm assuming a constant critical value across firms, and allow those critical values to vary from $\alpha = 0.01$ to $\alpha = 0.40$. In later sections of the article, we compare results obtained using Rating-derived alphas to those obtained using a constant threshold of 0.15, and all results with constant thresholds are available from the authors upon request.

2 Risk-Neutral Probabilities Implied by CDS and Option Prices

2.1 Data Description

Data on CDS are obtained from Markit. The initial sample consists of daily representative CDS quotes on all entities covered by Markit over the period January 2002 through November 2019. With the standardization of CDS contracts in 2009, new CDS contracts began trading with fixed coupon of 100 or 500 basis points, with upfront payment depending on the perceived credit risk of the underlying bond issuer. The CDS rate provided by Markit is “at market” composite CDS quotes, computed based on the bid and ask quotes obtained from two or more anonymous CDS dealers. We assume that the composite CDS rate is the rate at which the market value of the default swap is zero, without an upfront payment. While the five-year contract is generally thought to be the most liquid, our proposed measure of default probability relies on options data, of which few are struck for maturities in excess of one year. As a consequence, we restrict attention to entities that have quote data available on one-year CDS. We use these quoted prices, together with zero coupon discount rates, to solve for the default intensity, λ_t in Equation (3), assuming a constant rate of LGD at 60%. Discount rates are obtained by fitting the extended Nelson and Siegel (1987) model in Svensson (1994) using all non-callable Treasury securities from CRSP. Our initial sample consists of 364 entities for which we have at least one default intensity observation.

Options data are from OptionMetrics. The calculation of the risk-neutral moments requires the computation of integrals over a continuum of strikes. However, options are struck at discrete intervals. In addition, while the CDS in our sample has a constant one-year maturity, the maturity of options available in our sample varies and there are relatively few contracts available that are close to one year to maturity. We follow Hansis, Schlag, and Vilkov (2010) and Chang, Christoffersen, and Jacobs (2013) in constructing the volatility surface for options at 365 days to maturity using a cubic spline. We interpolate implied volatilities over the support of option deltas ranging from -99 to 99 at one-delta intervals, setting implied volatilities constant for deltas outside the span of observed option prices.

prior to filing in April 2001; Armstrong Cork, with a stock price drop of 93.8% prior to filing in December 2000; and Lehman Brothers, with a drop of 99.9% prior to delisting in September 2008.

Table 2. Summary statistics for probabilities of default

Panel A: Distribution					
Data Source	Mean	Std.	p5	p50	p95
Option	2.87	2.61	0.71	2.03	7.77
CDS	1.98	3.75	0.15	0.77	6.56

We then convert implied volatilities to option prices and integrate over OTM calls and puts using the rectangular approximation in [Dennis and Mayhew \(2002\)](#). In order to be included in the sample, we require that options have positive open interest, positive bid and offer prices, at least two OTM puts and OTM calls, offer prices greater than bid prices and offer prices greater than \$0.05. We also eliminate options where the offer price is greater than five times the bid price. Our sample of option-implied default probabilities yields 299 firms out of the 364 firms with default intensity information from CDS data.

We merge data from Markit, OptionMetrics, and CRSP data on the basis of the ticker and firm name to obtain the permno as a unique identifier for each firm. Next, we merge the matched sample of option- and CDS-implied default probabilities with credit ratings data from Compustat. We retain observations for which Compustat has a S&P ratings grade for the month of the observation. Since credit ratings data on Compustat end in February 2017, the combined sample extends only through that date. The final sample of firms, which have at least one time series observation with an option-implied default probability, a CDS-implied default probability, and a S&P credit rating, consists of 276 firms over the period January 2002 through February 2017.⁷

2.2 Descriptive Statistics

Summary statistics for the default probabilities implied by options and CDSs in our sample of firms are presented in [Table 2](#). For each firm, the default probability is calculated as the time series average of the weekly estimates obtained from CDS or options data using Rating-derived alphas. We report the mean, standard deviation, 5th, 50th, and 95th percentiles of the distribution of default probabilities in Panel A. Additionally, we report the 5th, 50th, and 95th percentile of the distribution of the correlation between CDS- and option-implied probabilities in Panel B.

The summary statistics indicate that across the distribution of firms, option-implied probabilities are on average higher than CDS-implied default probabilities. The mean and median option-implied default probability are 2.87% and 2.03%, compared to 1.98% and 0.77% for CDS. In addition, the option-implied probabilities exhibit less cross-sectional variation than CDS-implied default probabilities, with standard deviations of 2.61% and 3.75%, respectively. These results suggest that the distribution of CDS-implied default probabilities is more positively skewed than option-implied default probabilities; we explore various explanations for the skew later in the article.

7 These data are sourced from S&P Annual Global Corporate Default Study and Ratings Transitions 2015.

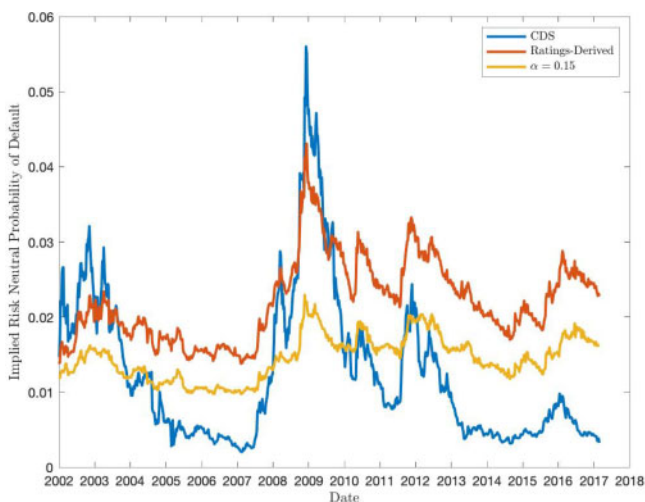


Figure 1. Implied default probabilities.

Notes: The figure plots the time series of aggregate default probabilities. Default probabilities are measured using one-year CDS spreads with an assumed rate of LGD of 60% and using the risk-neutral distribution implied by option prices with a ratings-dependent default threshold and a constant threshold $\alpha = 0.15$. CDS data are obtained from Markit and options data from OptionMetrics. Data are sampled at the weekly frequency and aggregated by taking the average across the firms in the sample. The data cover the period January 2002 through February 2017, and cover 276 firms.

We observe a substantial but imperfect correlation between CDS-implied and option-implied probabilities. The median correlation coefficient between default probabilities is 0.52; even at the 95th percentile, the correlation is less than 1.0.⁸

In [Figure 1](#), we plot the time series of cross-sectionally averaged weekly default probabilities implied by CDS and options. For the options-based default probabilities, we show the averages when the default threshold varies by credit rating, as well as when the default threshold is held constant at 0.15. The plot shows that the CDS measure of default probability, as well as both option-implied default probabilities, exhibit common features; default probabilities are uniformly low during the economic expansion and spike during times of economic turbulence. In particular, the default probabilities rise sharply during the recession in the early 2000s and the financial crisis of 2007–2009. Default probabilities also spike in late 2011, corresponding to the uncertainty surrounding the U.S. Congress' willingness to raise the federal debt ceiling and the subsequent downgrade of U.S. sovereign debt by S&P. Both measures of default probabilities are low during the economic expansions of mid-2000s and the later part of the sample in 2013–2017. The result that default probabilities are countercyclical is consistent with the evidence in ([Chen et al., 2009](#)), who find that they are better able to explain Baa-Aaa spreads using a model that is calibrated to match

8 At the 5th percentile, note that the correlation between the two default probabilities is negative. However, we find that this result is driven by the large gaps in the time series for some firms, with virtually all of the negative correlations concentrated in firms with relatively few time series observations.

Table 3. Summary statistics for probabilities of default by credit rating

Credit Rating		Option		CDS	
		Mean	Median	Mean	Median
AAA	9	0.45	0.35	0.23	0.19
AA	32	0.90	0.87	0.44	0.24
A	116	1.35	1.37	0.61	0.32
BBB	158	2.16	2.08	1.27	0.72
BB	88	3.71	3.64	2.87	2.34
B	49	6.73	6.09	4.72	3.59
CCC+ and below	9	14.39	14.51	15.38	11.48

Notes: The table presents summary statistics for risk-neutral probabilities of default implied by CDS spreads and prices of options on the equity of the same firm, grouped by credit rating. Firms with credit ratings augmented by ‘+’ or ‘-’ are grouped together; for example, rating ‘BBB’ refers to firms with a credit rating of ‘BBB+’, ‘BBB’, or ‘BBB-’. CDS-implied default probabilities are calculated using a Nelson–Siegel–Svensson zero coupon term structure and constant one-year maturity CDS, assuming a LGD of 60% on the underlying bond. Option-implied default probabilities are measured using the BKM procedure for computing risk-neutral moments, and then computing risk-neutral probabilities of the NIG distribution on the basis of these moments. Risk-neutral moments are computed using the implied volatility surface of options with 365 days to maturity. We calculate the average default probability for each firm, conditional on its ratings group, and report the mean and median of these averages by ratings group. Options and CDS data are sampled at the weekly frequency, and ratings at the monthly frequency for 276 firms over the period January 2002 through February 2017.

Sources: CDS data are from Markit, options data are from OptionMetrics, and ratings data are from Compustat.

default rates that are countercyclical. We find that all three estimates of default probabilities exhibit this tendency: the correlation between both of the two option-implied default probabilities and the CDS-implied default probability through time is relatively high, at 0.66 and 0.56, respectively.

The greater variability in CDS-implied default probabilities observed in Figure 1, with higher default probabilities compared to option-implied probabilities during market downturns and lower default probabilities compared to option-implied probabilities during market upturns, is consistent with the evidence in Table 2. It may also reflect variation in LGDs through the cycle, in contrast with the (assumed) constant LGDs embedded in the CDS-implied estimates of default probabilities shown here. That is, this pattern is consistent with expected LGD covarying positively with true default probabilities. We investigate this possibility later in the article.

2.3 Default Probabilities and Credit Ratings

To investigate further the cross-sectional variation in implied default probabilities, we report summary statistics for default probabilities by credit rating. In Table 3, we present mean and median default probabilities and the number of firms conditional on ratings class. Since firms may migrate across ratings, the total N reported in the table differs from the total number of firms reported in the sample. Thus, the $N = 9$ for “AAA”-rated firms indicates that there are nine firms that at some point in this time series have been rated “AAA.” As above, we group together firms with a “+,” “-,” or no modifier.

The summary statistics in Table 3 indicate that for both CDS- and option-implied default probabilities, the average risk-neutral probability of default increases monotonically across ratings classes. Using options (CDS) data, average default probabilities increase from 0.45% (0.23%) for “AAA”-rated firms to 14.39% (15.38%) for firms rated “CCC+” and below. There is a similar pattern of monotonic increase in median default probabilities across ratings classes, although CDS-implied median default probabilities are typically lower than the mean default probabilities. Across all ratings except “CCC,” the average and median option-implied probability is higher than that of the CDS-implied probability, consistent with the aggregate evidence reported earlier.

Of course, since the default thresholds in Table 3 vary by credit rating, some of these results may be mechanically induced. In untabulated results, we examine default probabilities across ratings classes, while keeping the default threshold constant across firms. Our results indicate that, as credit ratings deteriorate, the default probabilities implied by options for firms with poor credit ratings and relatively high default thresholds (high α) are more similar to those implied by CDS than those implied by relatively low thresholds (low α). Similarly, the default probabilities implied by options for firms with strong credit ratings are more similar to those implied by CDS when the threshold is low. This evidence is consistent with default thresholds that vary both cross-sectionally [as in Chava and Jarrow (2004)] and over time [as in Chen, Collin-Dufresne, and Goldstein (2008)] as credit ratings migrate.

2.4 Firm Characteristics and Probability of Default

The evidence presented above indicates that option-implied risk-neutral probabilities of default are substantially and positively correlated with CDS-implied risk-neutral probabilities of default, and are cross-sectionally correlated with S&P credit ratings. We analyze cross-sectional variation in default probabilities in more detail in this section, by examining the relation between both estimates of default probability and firm characteristics. In particular, we use a variant of Campbell, Hilscher, and Szilagyi (2008), who specify a pooled logit model for prediction of default. Instead of using a limited dependent variable based on an observation of default, we regress continuous estimates of default probabilities on firm-specific variables.

$$Q_{it}^k = a_{it} + \mathbf{b}'_{it}\mathbf{x}_{it} + u_{it},$$

where Q_{it}^k is the Week t observation of the risk-neutral probability, $k = \{C, O\}$ indexes CDS- and option-implied default probabilities and \mathbf{x}_{it} is a vector of firm-specific characteristics. In constructing these characteristics, we sample market variables at the weekly frequency contemporaneously with the default probabilities. Accounting data items are sampled at the quarterly frequency and lagged one-quarter relative to the default probabilities.⁹ The firm-specific variables that we use comprise the fundamental variables examined in Campbell, Hilscher, and Szilagyi (2008) and are described in Online Appendix B.

The results of these regressions are presented in Table 4. We split the sample into financial and nonfinancial firms, defined by the firm’s GICS sector from Compustat. Campbell, Hilscher, and Szilagyi (2008) consider only nonfinancial firms, and ratios such as leverage

9 Results are qualitatively unchanged if we sample the last weekly observation of the quarter for the market variables and default probabilities, rather than using all weekly observations.

Table 4. Relation between firm characteristics and default probabilities

Characteristic	CDS		Option	
	Financial	Nonfinancial	Financial	Nonfinancial
<i>NIMTA</i>	-17.32*** (2.18)	-11.20*** (0.34)	-24.37*** (1.30)	-5.99*** (0.22)
<i>TLMTA</i>	2.90*** (0.13)	1.54*** (0.04)	1.89*** (0.08)	2.33*** (0.02)
<i>EXRET</i>	-0.46*** (0.15)	-0.35*** (0.04)	1.56*** (0.09)	0.38*** (0.02)
<i>SIGMA</i>	7.45*** (0.09)	7.33*** (0.04)	1.99*** (0.05)	4.84*** (0.02)
<i>RSIZE</i>	-1.00*** (0.02)	-0.80*** (0.01)	-0.86*** (0.01)	-0.36*** (0.00)
<i>CASHMTA</i>	4.51*** (0.29)	1.62*** (0.09)	3.91*** (0.18)	2.21*** (0.06)
<i>BM</i>	0.04*** (0.01)	-0.47*** (0.01)	-0.06*** (0.01)	-0.32*** (0.01)
<i>R</i> ²	0.54	0.46	0.40	0.59

Notes: The table examines the relationship between default probabilities implied by either CDS spreads using a 60% LGD assumption or options with a critical threshold that varies with credit rating, and firm-specific characteristics. Default probabilities at the end of each month are regressed on a set of nine firm-specific variables: *NIMTA*, the ratio of net income to market value of total assets, *TLMTA*, the ratio of total liabilities to market value of assets, *EXRET*, the monthly log return on the firm's equity in excess of that of the S&P 500, *SIGMA*, the volatility of the firm's equity return over the past three months, *RSIZE*, the log ratio of the market capitalization of the firm's equity to that of the S&P 500, *CASHMTA*, and *BM*, the firm's ratio of book value of equity to market value of equity. Point estimates are the average of monthly regression coefficients, and standard errors in parentheses are corrected using the Newey-West procedure. We present results for financial firms and for firms excluding financial firms, defined as those in GICS sector 40.

Sources: Data for CDS are obtained from Markit, data for options is obtained from Option Metrics, return information is obtained from CRSP, financial statement and ownership information is obtained from Compustat.

*, **, *** denotes significance at the 10%, 5%, and 1% critical level, respectively.

and book-to-market ratio are likely to be very different for financial firms than nonfinancial firms.

The results for nonfinancial firms indicate that the relations between CDS-implied and option-implied default probabilities and firm characteristics are similar. The implied default probabilities for nonfinancial firms are statistically significantly decreasing in profitability, book-to-market, and relative size when default probabilities are measured using either options or CDS. Both the default probabilities are statistically significantly increasing in leverage and return volatility. All of these relations are consistent with the results in [Campbell, Hilscher, and Szilagyi \(2008\)](#). For example, the negative association between book-to-market ratios and default probabilities is consistent with the positive relation between market-to-book equity ratio and bankruptcy in [Campbell, Hilscher, and Szilagyi \(2008\)](#). This result is also consistent with the evidence in [Hovakimian, Kayhan, and Titman \(2012\)](#), who find that firms with higher proportions of tangible assets (or low book-to-market ratios) have lower default probabilities. Additionally, both default probabilities are

positively and statistically significantly associated with cash holdings, suggesting a greater precautionary saving motive for holding cash as in Acharya, Davydenko, and Strebulaev (2012). For only one variable, excess returns, does the relation to the two measures of default probability change signs in this subsample.

Results for financial firms are broadly similar to those for nonfinancial firms. When default probabilities are measured by options, the regression coefficient associated with firm characteristics is similar for financial and nonfinancial firms in terms of sign and statistical significance. We obtain similar findings for CDS-implied default probabilities, with the exception of its relation to book-to-market. Overall, the results of this analysis suggest that the relation of firm characteristics to both CDS and option-implied default probabilities are consistent with the underlying fundamentals of the firm.

3 Sources of Difference Between Option- and CDS-Implied Default Probabilities

The evidence presented so far suggests that default probabilities inferred from CDS and option prices contain broadly similar information regarding the time series properties of probabilities of default. Additionally, the evidence suggests that credit ratings and firm characteristics that are hypothesized to be related to the probability of default are related to both option- and CDS-implied default probabilities; that is, both estimates of default probabilities are capturing cross-sectional information.

However, as noted above, default probabilities estimated from the CDS and equity options market are not perfectly correlated. It is possible that these differences may simply arise from estimation error; in particular, the option-implied default probabilities are based on estimation of risk-neutral moments and the imposition of the NIG distributional assumption. In this section, we consider various reasons that the probabilities from the two markets might differ.

The first possibility we consider is that differences in option and CDS prices are due to variation in rates of LGD. A second possibility is that aggregate and security-level liquidity may impact option and CDS prices, and therefore the imputed default probabilities derived from these markets. For example, at the security level, our option-based probability estimates begin by calculating implied volatility surfaces using OTM puts and calls. These contracts, especially deep OTM contracts, are likely to have less liquidity than near-the-money puts and calls. Further, we are interpolating the volatility surface at a maturity of 365 days, where there are likely to be fewer available contracts and lower liquidity. In addition, CDS is also likely to suffer from liquidity issues. We are using one-year CDS contracts, which have lower liquidity than five-year contracts. Finally, CDS are relatively sparsely traded at the beginning of our sample period and some contracts also suffer from liquidity issues during the financial crisis.

A third possibility involves the default threshold, α , that is used when estimating default probabilities from the options market. That is, higher α 's may result in option-implied default probabilities that better match CDS-implied default probabilities in times when CDS-implied probabilities are high and for firms with poorer credit ratings. Thus, as in Chen, Collin-Dufresne, and Goldstein (2008), it may be that the market perceives the default threshold for equity as being different during times of financial market stress or when a firm is closer to its default boundary; although our estimation method lets α vary as the

credit rating varies, the market's perception of the default threshold may change at a higher frequency than a firm's credit rating. In addition, there may be other economic rationales for a varying default threshold, such as strategic bankruptcies; that is, in some circumstances, firms may find it beneficial to default even if they are solvent and able to make debt payments, as in [Davyydenko and Strebulaev \(2007\)](#).

3.1 A Simple Measure of LGD

To analyze differences in the default probability estimates in these two markets, and in particular to explore whether these differences may be related to LGDs, we consider a simplified version of the relation between the CDS spread, the default probability, and the LGD. If we consider a simple one-period CDS contract:

$$S_t = (Q_t^C) \cdot (LGD_t),$$

the CDS spread is the product of the default probability and the LGD. While the relationship between spreads, default probabilities, and LGD is more precisely given in [Equation \(3\)](#), this approximation provides useful intuition for understanding the relation between spreads, default probabilities, and rates of LGD.

Under the standard practice of assuming a constant rate of LGD of 60%, the spread and the CDS implied default probability in this simple model are perfectly correlated by construction, both cross-sectionally and in the time series. However, if the option-implied default probability is a valid estimate of the true default probability, then the ratio of CDS spread and option-implied default probability should provide an (approximate) estimate of LGD that is allowed to vary across firms and through time. Of course, this ratio will also capture effects related to liquidity, mis-specification of the default threshold, and other estimation errors. We begin by calculating this ratio, denoted as \hat{LGD} , and considering its properties below.

3.1.1 Cross-market inferences

In [Figure 2](#), we present the time series of average estimates of \hat{LGD} across all firms in our sample, where option-implied default probabilities are calculated using both a threshold based on credit ratings (labeled Rating-derived) and a constant default threshold of 0.15. To limit the LGDs to 100%, we set values of \hat{LGD} that are above 100% to missing.

The figure shows that the average \hat{LGD} measures obtained using the two default thresholds are highly correlated. In addition, the behavior of both measures of \hat{LGD} through time is consistent with the interpretation that the measure is related to LGD; note that the average \hat{LGD} varies strongly with business conditions, consistent with the relation between recovery rates and market fundamentals documented in [Jankowitsch, Nagler, and Subrahmanyam \(2014\)](#). In particular, the variation in \hat{LGD} implies that average LGDs are high in the early 2000s (at approximately 45%), followed by declines in LGD to values of approximately 10–15% during the economic recovery of the mid-2000s. LGDs again rise sharply to 55% during the financial crisis of 2007–2009, and then gradually decline. The secondary increase in LGD in 2011 is contemporaneous with the downgrade of U.S. debt in 2011. Over the recent period of 2013–2017, the LGDs hover around 15%. Note that the average LGD in economic expansions is substantially lower than the typical assumed

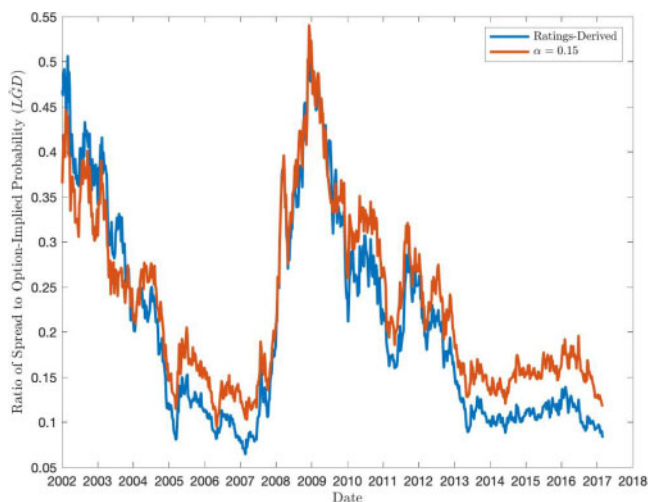


Figure 2. Differences in log CDS spreads and option-implied default probabilities.

Notes: The figure plots the time series of the ratio of one-year CDS spreads and option-implied default probabilities aggregated over firms, denoted as $L\hat{G}D$. Option-implied default probabilities are calculated with a rating-derived default threshold and a constant threshold $\alpha = 0.15$. Options data are from OptionMetrics and CDS data are from Markit. Data cover 276 firms over the period January 2002 through February 2017.

(constant) rate of 60%, although it is consistent with the average historical LGD figures reported for large corporate borrowers [see, e.g., Global Credit Data (2018)].

Regardless of default thresholds, the evidence in Figure 2 that $L\hat{G}D$ varies with economic conditions is consistent with an inference that combining option-implied default probability measures with CDS data provides information about recovery rates, and indicates that the assumption that recovery rates are constant through time is a poor fit to the data. In addition, if both default probabilities and LGDs are countercyclical, note that estimates of default probability taken from CDS under the assumption that LGD are *constant* would result in default probability estimates that are more variable, and more right-skewed, than estimates taken from option prices, consistent with the evidence in Table 2; intuitively, if one assumes that LGD is constant instead of allowing LGD and default probabilities to covary positively, then estimated default probabilities must vary more in order to match the volatility in CDS rates. In the next sections, we examine the variation in $L\hat{G}D$ across sectors, while controlling for other factors such as liquidity.

3.1.2 Loss given default and industry sectors

We analyze variation in $L\hat{G}D$ across sectors. The results so far indicate that the cross-sectional variation in option-implied default probabilities is sensitive to the default threshold chosen, although the time series information in $L\hat{G}D$ across different thresholds is similar. As a consequence, we examine LGD inferences for the two default thresholds presented in Figure 2: a constant $\alpha = 0.15$ and the firm-specific default thresholds based on credit rating (i.e., the Rating-derived α). We utilize the GICS sector definitions, which separate firms

Table 5. Summary statistics for ratio of CDS spreads to option probabilities

Sector	N	Rating-Derived α			$\alpha = 0.15$		
		p5	p50	$L = 100$	p5	p50	$L = 100$
All	276	0.08	0.21	99	0.08	0.32	84
Energy	26	0.12	0.19	100	0.09	0.27	85
Materials	19	0.08	0.20	100	0.10	0.33	95
Industrials	37	0.08	0.16	95	0.07	0.16	84
Discretionary	50	0.11	0.24	100	0.12	0.44	78
Staples	26	0.10	0.20	100	0.09	0.18	81
Healthcare	30	0.10	0.19	100	0.07	0.27	97
Financials	32	0.14	0.30	94	0.10	0.34	91
Technology	24	0.08	0.27	100	0.07	0.41	83
Telecommunications	10	0.08	0.53	90	0.08	0.98	50
Utilities	16	0.11	0.20	100	0.13	0.32	75
Real Estate	6	0.17	0.47	83	0.27	0.68	83

Notes: The table presents summary statistics for the ratio of one-year CDS spreads and default probabilities implied by risk-neutral probabilities of default as measured by options. Option-implied default probabilities are measured using the BKM procedure for computing risk-neutral moments, then computing risk-neutral probabilities of the NIG on the basis of these moments. Risk-neutral moments are computed using the implied volatility surface of options with 365 days to maturity. CDS data are from Markit and options data are from OptionMetrics. The table presents the number of firms, 5th and 50th percentiles of the cross-sectional distribution of average ratio of one-year CDS spreads to option-implied default probabilities, and the percentile at which the ratio exceeds 1.0. Results are presented for all firms in the sample and with GICS sectors and two default thresholds; one based on Rating-derived thresholds and the other a constant threshold of 0.15. Data are sampled at the weekly frequency for 276 firms over the period January 2002 through February 2017.

into eleven sectors; Energy, (EN) Materials (MA), Industrials (IN), Consumer Discretionary (CD), Consumer Staples (CS), Healthcare (HC), Financial (FI), Information Technology (IT), Telecommunications (TC), Utilities (UT), and Real Estate (RE). Sector classifications are obtained from Compustat.

In Table 5, we report the estimates obtained using both default thresholds. For each default threshold measure, we report the number of firms in the sample or sector in the first column, followed by the average $L\hat{G}D$ in the sample or sector; in the remaining columns, we present the 5th percentile and 50th percentile of the cross-sectional distribution of $L\hat{G}D$. This is followed by a column that reflects the percentile in the sample or sector where $L\hat{G}D$ exceeds 100%.

When default thresholds vary by credit rating, the median-implied LGD across the entire sample is 21%, or a recovery rate of 79%; note again that the median $L\hat{G}D$ using this approximation is lower than the typical Markit estimate of 60%. For nine of the eleven sectors, the median $L\hat{G}D$ ranges between 16% (Industrials) to 30% (Financials), with substantially higher $L\hat{G}D$ for Telecommunications and Real Estate. The number of firms with $L\hat{G}D$ greater than 1 is relatively small; there are six firms in total with these extreme $L\hat{G}D$ in four sectors. Overall, 99% of the firms have $L\hat{G}D$ lower than 1.0.

We obtain qualitatively similar results when the default threshold is held at a constant 0.15, with the range of $L\hat{G}D$ increasing in most cases. In the full sample, the median $L\hat{G}D$ is 0.32, and the range of $L\hat{G}D$ in nine out of eleven sectors is within the range of 16–44%.

For both default thresholds, we observe considerable variation in $L\hat{G}D$ across sectors, indicating that recovery rates vary cross-sectionally. Using both measures of default threshold, there are two sectors with higher $L\hat{G}D$: telecommunications and real estate, where the estimated median LGD increases to approximately 0.5 in the ratings-derived default threshold, and to values of median $L\hat{G}D$ at 98% and 68%, respectively, using the constant default threshold. Note that these two sectors also have a relatively small number of firm observations (ten or fewer); as a consequence, these results may be subject to greater estimation error. Overall, however, the correlation across sectors in the median estimate of $L\hat{G}D$ across the two default thresholds is quite high, at 0.97.

3.2 Liquidity, Default Probabilities, and LGD

While variation in LGD is one possible explanation for differences in risk-neutral default probabilities across options and CDS, another possibility is that estimated probabilities in these two markets differ as a result of market frictions. As mentioned above, OTM options used to estimate risk-neutral moments, and then option-implied default probabilities, may be thinly traded; in addition, the liquidity of some CDS contracts is low. As a consequence, the marked increase in LGD around the crisis may reflect instead changes in market liquidity.¹⁰ Note that the approximate relation between CDS rates, option-implied default probabilities, and LGD discussed in Section 4.1 implies that, in the absence of market frictions, the difference between the log CDS spread and the log default probability should be approximately equal to the log LGD. We use this approximation and estimate the extent to which variation in this difference is related to changes in various liquidity measures.

Illiquidity in the CDS and options markets may reflect both security-specific and market-wide variation in liquidity.¹¹ We are limited in measuring security-specific liquidity by the data available for options and CDS. In the case of options, we have information on bid-ask spreads, open interest, and volume. Since the default probabilities recovered from options are likely to depend most on the prices of OTM options close to 365 days to maturity, we construct $SPREAD_t^O$, the average percentage bid-ask spread for the OTM options used in constructing our volatility surface. We also compute VOL_t^O and $OPEN_t^O$, the sum of volume and open interest for these contracts. In the case of CDS, we have a measure of the firm-specific depth for five-year CDS contracts, $DEPTH_t^C$. We assume that depth for the one-year contracts is correlated with the depth of the five-year contracts for each firm and use that as another measure of liquidity.

To capture aggregate liquidity, we use two measures from the fixed income security markets. First, we use the Treasury-Eurodollar spread, TED_t , measured as the difference in 90-day LIBOR and 90-day Treasury Bill yields. An increase in the TED spread can indicate an increase in interbank counterparty credit risk, and a consequent drop in funding liquidity. The second measure is the root mean squared error of the difference in market Treasury

- 10 It is possible that precipitous declines in market liquidity are associated with declines in asset values and thus increases in LGD (see, e.g., Brunnermeier and Pedersen (2009)). If that is the case, our controls for market liquidity will cause the increases in LGD during periods of market illiquidity to be estimated conservatively.
- 11 Note that by using Fama and MacBeth (1973) regressions we are in effect including a time fixed effect in the regression. Thus, the results reported earlier supporting the interpretation of $L\hat{G}D$ as a proxy for LGD are unlikely to be due to aggregate liquidity effects.

security yields from those implied by a Nelson–Siegel–Svensson model. This measure, $NOISE_t$, is investigated in [Hu, Pan, and Wang \(2013\)](#). The authors suggest that $NOISE_t$ is high when there is less arbitrage capital available in the Treasury market, a condition associated with lower liquidity. The TED spread is constructed using data from the Federal Reserve and $NOISE_t$ is obtained from Jun Pan's webpage.¹² Finally, we include a proxy for liquidity in the equity markets; [Nagel \(2012\)](#) suggests that a high level of the VIX index, VIX_t is associated with a high-risk premium, and a consequent large reduction in liquidity provision, in equity markets. Data on the VIX are also obtained from the Federal Reserve.

We examine the extent to which liquidity effects influence the relation between CDS and option-implied default probabilities by estimating the relation between changes in the (log) approximate LGD and changes in the (log) liquidity variables, beginning at the aggregate level. That is, we estimate the parameters of a regression,

$$\begin{aligned} \Delta \hat{lgd}_{a,t} = & a_a + b_{a,1} \Delta ted_t + b_{a,2} \Delta noise_t + b_{a,3} \Delta vix_t + b_{a,4} \Delta spread_{a,t}^O \\ & + b_{a,5} \Delta vol_{a,t}^O + b_{a,6} \Delta open_{a,t}^O + b_{a,7} \Delta depth_{a,t}^C + e_{a,t}, \end{aligned} \quad (6)$$

where a indicates that we are measuring the quantity at the aggregate level, and aggregate variables are calculated as the cross-sectional average of individual time series observations. Lowercase variables are natural logs of their uppercase counterparts. The option-implied probability of default that is used to construct the estimate of LGD uses a rating-derived default threshold. All variables are measured at the weekly horizon.

Results of this regression are reported in [Table 6](#). When all liquidity variables are included, there is some evidence that changes in options market liquidity variables are associated with changes in CDS spreads. Specifically, the coefficient on $\Delta spread^O$ is significant at the 1% level, with an increase in spreads associated with a downward revision in the approximate LGD. In contrast, changes in the VIX are associated with a significant increase in LGD, suggesting that some of the marked increase in LGD observed in the financial crisis may be associated with an increase in market-wide volatility. This result is consistent with the evidence in [Nagel \(2012\)](#), who shows that an increase in the VIX is associated with a reduction of equity arbitrage capital; alternatively, or in addition to a liquidity effect, the increase in market-wide volatility may be associated with a decline in asset values. Together, changes in these liquidity variables explain approximately 17% of the variation in changes in the log LGD measure.

We also report the results of this regression across sectors in [Table 7](#). The results are generally consistent with the results observed in the aggregate. Across all sectors with the exception of real estate, we continue to find evidence that changes in the VIX are positively associated with changes in the average LGD in the sector. Changes in option spreads are negatively and statistically significantly associated with changes in LGD in six out of the eleven sectors at the 5% or 10% critical level (specifically, consumer discretionary, consumer staples, healthcare, financial, information technology, and utilities). Changes in option open interest are significantly negatively related to changes in \hat{lgd} in the real estate sector. Finally, changes in the $NOISE$ measure are statistically significantly and positively related to changes in \hat{lgd} in the energy and consumer staples sectors. In all of these cases,

12 We thank Jun Pan for making these data available at <http://www.mit.edu/junpan/>.

Table 6. Liquidity, CDS spreads, and option-implied default probabilities

	Δted	$\Delta noise$	Δvix	Δvol^O	$\Delta open^O$	$\Delta spread^O$	$\Delta depth^C$	R^2
Aggregate	0.012	0.024	0.152	0.003	0.033	-0.113	-0.027	0.168
SE	(0.019)	(0.016)	(0.015)	(0.006)	(0.030)	(0.029)	(0.033)	
p5	-0.493	-0.215	-0.195	-0.108	-0.724	-0.525	-0.133	0.012
SE	(0.371)	(0.596)	(0.147)	(0.132)	(0.517)	(0.242)	(0.053)	
p25	-0.082	-0.022	0.029	-0.034	-0.133	-0.215	-0.023	0.029
SE	(0.174)	(0.055)	(0.065)	(0.031)	(0.406)	(0.122)	(0.023)	
p50	0.007	0.040	0.165	-0.008	-0.002	-0.079	0.008	0.050
SE	(0.080)	(0.045)	(0.079)	(0.026)	(0.164)	(0.093)	(0.017)	
p75	0.082	0.115	0.259	0.016	0.110	0.031	0.039	0.086
SE	(0.041)	(0.057)	(0.073)	(0.038)	(0.258)	(0.274)	(0.054)	
p95	0.285	0.330	0.457	0.082	0.447	0.446	0.122	0.256
SE	(0.098)	(0.170)	(0.064)	(0.059)	(0.183)	(0.393)	(0.090)	

Notes: The table presents the results of regressions of changes in the log approximate LGD on changes in log liquidity variables. The regressions are specified as

$$\Delta \hat{lgd}_{a,t} = a_a + b_{a,1} \Delta ted_t + b_{a,2} \Delta noise_t + b_{a,3} \Delta vix_t + b_{a,4} \Delta spread_{a,t}^O + b_{a,5} \Delta vol_{a,t}^O + b_{a,6} \Delta open_{a,t}^O + b_{a,7} \Delta depth_{a,t}^C + e_{a,t},$$

where $\hat{lgd}_{a,t}$ is the log of an approximate LGD measure and is constructed by subtracting the log of the option-implied default probability from the log of the one-year CDS spread. ted_{t+1} is the log TED spread, the difference between the yield on 90-day LIBOR and 90-day Treasury Bills, $noise_{t+1}$ is the log of the noise measure from Hu, Pan, and Wang (2013), vix_{t+1} is the log VIX index, $vol_{i,t+1}^O$ is the log of the sum of volume for OTMOTM options on firm i , $open_{i,t+1}^O$ is the log sum of open interest on firm i 's OTM options, $spread_{i,t+1}^O$ is the log of the average percentage bid-ask spread for firm i 's OTM options, and $depth_{i,t+1}^C$ is the depth of five-year CDS contracts for firm i . The table first reports results of regressions of the change in log LGD on the explanatory variables, where the aggregate is constructed as the cross-sectional average of each week's observations. The table also reports results for the 5th, 25th, median, 75th, and 95th percentile coefficient estimates and their associated standard errors for firm-specific regressions. The sample covers 276 firms over the period January 2002 through February 2017.

Sources: Options data is from OptionMetrics, CDS data from Markit, financial market data from the Federal Reserve, and the noise measure from Jun Pan's website.

the R^2 measures indicate that liquidity measures account for less than 10% of the time series variation in changes in the approximate LGD measure.

Using these regression results, we construct an alternative measure of approximate LGD that controls for the effect of these liquidity measures. Specifically, we construct an alternative \hat{LGD} by cumulating the residuals from Equation (6). At time t , this liquidity-adjusted LGD measure is calculated as

$$\hat{LGD}_{a,t}^* = \exp \left(\hat{a}_a + \sum_{j=0}^t \hat{e}_{a,t-j} \right),$$

where we initialize $\hat{LGD}^* = \hat{LGD}$ at time $t = 1$ (in January 2002) and at each period add the residual at time t given by the regression above. We exponentiate this series so that it can be compared to the \hat{LGD} measure calculated previously.

Table 7. Liquidity, CDS spreads, and option-implied default probabilities by sector

Sector	Δted	$\Delta noise$	Δvix	Δvol^O	$\Delta open^O$	$\Delta spread^O$	$\Delta depth^C$	\bar{R}^2
Energy	-0.005 (0.038)	0.065 (0.032)	0.218 (0.031)	-0.010 (0.013)	-0.070 (0.062)	0.001 (0.059)	0.009 (0.037)	0.075
Materials	0.011 (0.045)	0.011 (0.038)	0.192 (0.037)	0.006 (0.015)	-0.015 (0.074)	-0.014 (0.071)	-0.042 (0.036)	0.041
Industrials	0.041 (0.041)	-0.046 (0.035)	0.081 (0.034)	0.018 (0.014)	-0.056 (0.068)	-0.075 (0.065)	-0.017 (0.046)	0.022
Cons. Disc.	0.020 (0.035)	0.012 (0.029)	0.174 (0.028)	0.003 (0.012)	0.058 (0.057)	-0.129 (0.054)	-0.006 (0.042)	0.068
Cons. Staples	-0.006 (0.037)	0.083 (0.031)	0.096 (0.030)	0.027 (0.013)	0.042 (0.061)	-0.144 (0.058)	0.046 (0.041)	0.045
Healthcare	-0.025 (0.034)	0.012 (0.028)	0.132 (0.028)	-0.009 (0.012)	0.039 (0.056)	-0.126 (0.053)	-0.087 (0.039)	0.056
Financials	-0.002 (0.036)	0.053 (0.030)	0.155 (0.029)	-0.003 (0.012)	0.098 (0.059)	-0.158 (0.056)	-0.023 (0.034)	0.062
Info. Tech.	0.047 (0.058)	0.007 (0.051)	0.048 (0.047)	-0.019 (0.021)	0.158 (0.096)	-0.176 (0.092)	-0.077 (0.051)	0.014
Telecom.	0.023 (0.101)	0.115 (0.088)	0.166 (0.083)	0.029 (0.036)	0.111 (0.168)	-0.218 (0.165)	-0.141 (0.060)	0.024
Utilities	0.015 (0.065)	0.077 (0.055)	0.175 (0.054)	-0.020 (0.022)	0.176 (0.110)	-0.182 (0.109)	-0.074 (0.057)	0.033
Real Estate	-0.048 (0.122)	0.001 (0.106)	0.154 (0.103)	-0.002 (0.043)	-0.460 (0.208)	0.147 (0.211)	0.019 (0.075)	0.012

Notes: The table presents the results of regressions of changes in the log approximate LGD on changes in log liquidity variables. The regressions are specified as

$$\Delta \hat{l}gd_{a,t} = a_a + b_{a,1}\Delta ted_t + b_{a,2}\Delta noise_t + b_{a,3}\Delta vix_t + b_{a,4}\Delta spread_{a,t}^O + b_{a,5}\Delta vol_{a,t}^O + b_{a,6}\Delta open_{a,t}^O + b_{a,7}\Delta depth_{a,t}^C + e_{a,t},$$

where $\hat{l}gd_{a,t}$ is the log of an approximate LGD measure and is constructed by subtracting the log of the option-implied default probability from the log of the one-year CDS spread. ted_{t+1} is the log TED spread, the difference between the yield on 90-day LIBOR and 90-day Treasury Bills, $noise_{t+1}$ is the log of the noise measure from Hu, Pan, and Wang (2013), vix_{t+1} is the log VIX index, $vol_{i,t+1}^O$ is the log of the sum of volume for OTM options on firm i , $open_{i,t+1}^O$ is the log sum of open interest on firm i 's OTM options, $spread_{i,t+1}^O$ is the log of the average percentage bid-ask spread for firm i 's OTM options, and $depth_{i,t+1}^C$ is the depth of five-year CDS contracts for firm i . The table reports results for series aggregated across two-digit GICS codes, Energy (EN), Materials (MA), Industrials (IN), Consumer Discretionary (CD), Consumer Staples (CS), Healthcare (HC), Financials (FI), Information Technology (IT), Telecommunications (TC), Utilities (UT) and Real Estate (RE). Data are sampled at the weekly frequency from January 2002 through February 2017.

Sources: Options data is from OptionMetrics, CDS data from Markit, financial market data from the Federal Reserve, and the noise measure from Jun Pan's website.

Figure 3 presents the time series of this variable. The behavior of this liquidity-adjusted LGD measure is broadly similar to the series plotted in Figure 2, with a relatively high implied LGD during the early 2000s recession that declines in the mid-2000s, followed by sharp increases in the financial crisis of 2008–2009. There are two noteworthy differences

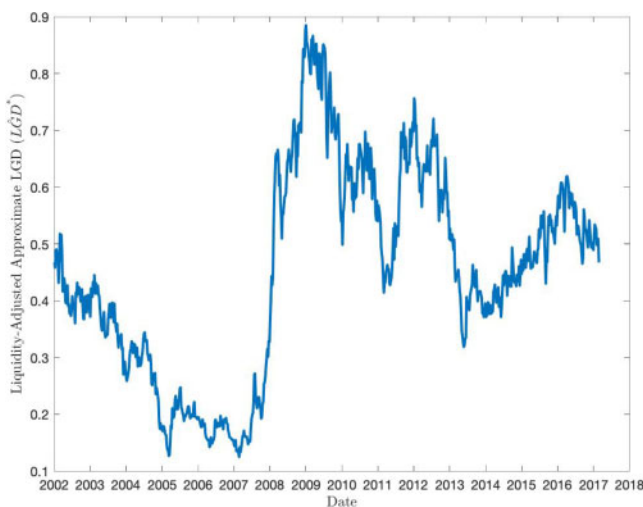


Figure 3. Liquidity-adjusted LGD.

Notes: The figure presents a liquidity-adjusted measure of the LGD, represented by the cumulative residual from a regression of the change in approximate log LGD on a set of variables measuring liquidity. Specifically, we regress changes in the log aggregate ratio of one-year CDS spread to option-implied default probability on changes in the log of the aggregate of seven measures of liquidity; the TED spread, ted , the noise measure from [Hu, Pan, and Wang \(2013\)](#), $noise$, the VIX index, vix , the sum of OTM option volume, vol^O , the sum of OTM option open interest, $open^O$, the average OTM option bid-ask spread, $spread^O$, and the depth of five-year CDS contracts, and $depth^C$:

$$\hat{lgd}_t = a + b_1 \Delta ted_t + b_2 \Delta noise_t + b_3 \Delta vix_t + b_4 \Delta vol_t^O + b_5 \Delta open_t^O + b_6 \Delta spread_t^O + b_7 \Delta depth_t^C + e_{t+1},$$

where aggregates of the variables vol^O , $open^O$, $spread^O$, and $depth^C$ are cross-sectional averages or sums of firm-level variables. The LGD net of liquidity is

$$LGD^* = \exp \left(\hat{lgd}_1 + \hat{a} + \sum_{j=0}^{t-1} \hat{e}_{it-j} \right).$$

Sources: Data for the construction of the TED spread and the VIX are obtained from the Federal Reserve. The noise measure is taken from Jun Pan’s website. Data on option volume, open interest, and bid-ask spreads are from OptionMetrics. CDS contract depth is obtained from Markit. Data are sampled at the weekly frequency over the period January 2002 through February 2017.

in the behavior of these LGD measures. First, LGD^* is substantially higher than LGD beginning approximately in 2009. That is, adjusting for liquidity effects results in an estimate of loss given default LGD^* that is meaningfully higher during the financial crisis. Second, the liquidity-adjusted estimate of LGD remains relatively elevated after the financial crisis, although it does decline, to approximately 0.35 in 2013, and then increases somewhat in the 2014–2017 period. Overall, estimates of LGD that control for liquidity effects increase in economic downturns; indeed, our evidence suggests that liquidity-adjusted LGD measures are more sensitive to economic states than LGD measures that do not control for liquidity.

4 Conclusion

In this article, we propose a new method of estimating default probabilities for firms. Using option prices, we construct an estimate of the risk-neutral density; the default probability for the firm is the mass under the density up to the return that corresponds to a default event. We estimate default probabilities for different levels of default thresholds; we find that, although the level of estimated default probabilities is sensitive to the choice of default threshold, they are very highly correlated with one another and behave very similarly over time.

We examine the relationship of the option-implied default probabilities to default probabilities estimated from CDS prices, as well as their relation to firm characteristics and ratings categories. We find that option-implied default probabilities are strongly, but not perfectly, related to CDS default probabilities that assume constant LGDs, with the latter exhibiting higher variation and higher skewness. With respect to firm characteristics and ratings categories, the option-implied default probabilities behave as one would expect. Specifically, the default probabilities increase monotonically as ratings decline; in addition, default probabilities of nonfinancial firms are significantly and positively related to leverage and volatility; they are negatively and significantly related to profitability, book-to-market equity, and size.

If option-implied default probabilities are valid estimates of the firm's propensity of failure, then an examination of the relation between these probabilities and CDS prices should provide information about LGDs. We construct a simple measure of LGD by calculating the ratio of CDS spreads and option-implied default probabilities. We find significant time series variation in this ratio, which is related to economic conditions. While the ratio is highly correlated across default thresholds, we also find evidence of significant cross-sectional variation in this measure depending on sectors.

When we estimate the relation between CDS spreads and option-implied default probabilities, while controlling for liquidity effects, we find evidence that changes in the VIX are significantly and positively related to changes in log LGD measures; we find weaker evidence that changes in the liquidity of the options and fixed income markets affect changes in LGD. Finally, after controlling for changes in liquidity effects, estimates of LGD inferred from CDS and option prices again show a strong relation to underlying business conditions. Overall, the equity options market may provide useful information with which to infer default probabilities, as well as the LGDs of underlying assets.

Supplementary Data

Supplementary data are available at *Journal of Financial Econometrics* online.

References

- Acharya, V., S. A. Davydenko, and I. A. Strebulaev. 2012. Cash Holdings and Credit Risk. *Review of Financial Studies* 25: 3572–3609.
- Altman, B., Edward, Bradi, A. Resti, and A. Sironi. 2005. The Link Between Default and Recovery Rates: Theory, Empirical Evidence and Implications. *The Journal of Business* 78: 2203–2228.

- Augustin, P. and F. Saleh. 2017. *A Note on CDS Returns*, McGill University and New York University. Unpublished manuscript.
- Augustin, P., M. G. Subrahmanyam, D. Yongjun Tang, and S. Qian Wang. 2016. Credit Default Swaps: Past, Present, and Future. *Annual Review of Financial Economics* (in press).
- Bakshi, G., N. Kapadia, and D. Madan. 2003. Stock Return Characteristics, Skew Laws and the Differential Pricing of Individual Equity Options. *Review of Financial Studies* 16: 101–143.
- Bakshi, G., D. Madan, and F. Zhang. 2006. Understanding the Role of Recovery in Default Risk Models: Empirical Comparisons and Implied Recovery Rates, University of Maryland. Unpublished manuscript.
- Berndt, A., R. Douglas, D. Duffie, and M. Ferguson. 2018. Corporate Credit Risk Premia, National Bureau of Economic Research. Unpublished manuscript.
- Breeden, D. and R. Litzenberger. 1978. Prices of State Contingent Claims Implicit in Options Prices. *The Journal of Business* 51: 621–651.
- Brunnermeier, M. K. and L. H. Pedersen. 2009. Market Liquidity and Funding Liquidity. *Review of Financial Studies* 22: 2201–2238.
- Campbell, J. Y., J. Hilscher, and J. Szilagyi. 2008. In Search of Distress Risk. *The Journal of Finance* 63: 2899–2939.
- Capuano, C. 2008. “The Option-Ipod: The Probability of Default Implied by Option Prices Based on Entropy.” *IMF Working papers*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1266527.
- Carr, P. and L. Wu. 2011. Simple Robust Linkages between American Puts and Credit Protection. *Review of Financial Studies* 24: 473–505.
- Chang, B. Y., P. Christoffersen, and K. Jacobs. 2013. Market Skewness Risk and the Cross-Section of Stock Returns. *Journal of Financial Economics* 107: 46–68.
- Chava, S. and R. Jarrow. 2004. Bankruptcy Prediction with Industry Effects. *European Finance Review* 8: 537–569.
- Chen, L, P. Collin-Dufresne, and R. S. Goldstein. 2009. On the Relation between the Credit Spread Puzzle and the Equity Premium Puzzle. *Review of Financial Studies* 22: 3367–3409. 10.1093/rfs/hhn078
- Das, S. and R. Sundaram. 2007. An Integrated Model for Hybrid Securities. *Management Science* 53: 1439–1451.
- Davydenko, S. A. and I. A. Strebulaev. 2007. Strategic Actions and Credit Spreads: An Empirical Investigation. *The Journal of Finance* 62: 2633–2671.
- Dennis, P. and S. Mayhew. 2002. Risk-Neutral Skewness: Evidence from Stock Options. *The Journal of Financial and Quantitative Analysis* 37: 471–493.
- Doshi, H., R. Elkamhi, and C. Ornthanalai. 2018. The Term Structure of Expected Recovery Rates. *Journal of Financial and Quantitative Analysis* 53: 2619–2661.
- Duffie, D. and K. J. Singleton. 1999. Modeling Term Structures of Defaultable Bonds. *Review of Financial Studies* 12: 687–720.
- Eriksson, A., E. Ghysels, and F. Wang. 2009. The Normal Inverse Gaussian Distribution and the Pricing of Derivatives. *The Journal of Derivatives* 16: 23–37.
- Fama, E. F. and J. D. MacBeth. 1973. Risk, Return, and Equilibrium: Empirical Tests. *Journal of Political Economy* 81: 607–636.
- Global Credit Data. 2018. “LGD Report 2018 - Large Corporate Borrowers.” Discussion paper, Global Credit Data Consortium.
- Hansis, A., C. Schlag, and G. Vilkov. 2010. The Dynamics of Risk-neutral Implied Moments: Evidence from Individual Options, Goethe University Frankfurt. Unpublished manuscript.
- Hovakimian, A., A. Kayhan, and S. Titman. 2012. Are Corporate Default Probabilities Consistent with the Static Trade-off Theory? *Review of Financial Studies* 315–340.

- Hu, X., J. Pan, and J. Wang. 2013. Noise as Information for Illiquidity. *The Journal of Finance* 68: 2341–2382.
- Jankowitsch, R., F. Nagler, and M. G. Subrahmanyam. 2014. The Determinants of Recovery Rates in the u.s. corporate Bond Market. *Journal of Financial Economics* 114: 155–177.
- Le, A. 2015. Separating the Components of Default Risk: A Derivate-Based Approach. *Quarterly Journal of Finance* 5: 1550005.
- Madan, D., and H. Unal. 1998. Pricing the Risks of Default. *Review of Derivatives Research* 2: 121–160.
- Madan, D. and L. Güntay. 2003. Pricing the Risk of Recovery in Default with Absolute Priority Rule Violation. *Journal of Banking & Finance* 27: 1001–1025.
- Merton, R. C. 1974. On the Pricing of Corporate Debt: The Risk Structure of Interest Rates. *The Journal of Finance* 29: 449–470.
- Nagel, S. 2012. Evaporating Liquidity. *Review of Financial Studies* 25: 2005–2039.
- Nelson, C. R. and A. F. Siegel. 1987. Parsimonious Modeling of Yield Curves. *The Journal of Business* 60: 473–489.
- O’Kane, D. and S. Turnbull. 2003. Valuation of Credit Default Swaps. *Lehman Brothers Fixed Income Quantitative Research Quarterly* 1–18.
- Schuermann, T. 2004. “What Do We Know about Loss Given Default?” Wharton Financial Institutions Center Working paper No. 04–01.
- Svensson, L. E. O. 1994. “Estimating and Interpreting Forward Interest Rates: Sweden 1992-1994.” NBER Working Paper 4871.
- Vilsmeier, J. 2014. “Updating the Option Implied Probability of Default Methodology.” Deutsche Bundesbank Discussion Paper No. 43/2014. Unpublished manuscript.

585 Mixed-Frequency Macro–Finance Factor Models:
Theory and Applications

*Elena Andreou, Patrick Gagliardini,
Eric Ghysels and Mirco Rubin*

629 Implied Default Probabilities and Losses Given
Default from Option Prices

*Jennifer Conrad, Robert F. Dittmar and
Allaudeen Hameed*

The goal of *Journal of Financial Econometrics* is to reflect and advance the relationship between econometrics and finance, both at the methodological and at the empirical levels. Estimation, testing, learning, prediction and calibration in the framework of asset pricing or risk management represent the core focus. The scope includes topics relating to volatility processes, return modelling, dynamic conditional moments, machine learning, big data, fintech, extreme values, long memory, dynamic mixture models, endogenous sampling transaction data, and microstructure of financial markets.