Modeling Latent Variables in Economics and Finance

PhD Thesis submitted to

Institute of Financial Analysis
University of Neuchâtel
Switzerland

Solvay Business School
Vrije Universiteit Brussel
Belgium

For the joint degree of

Doctor of Finance (UniNE)
&
Doctor in Business Economics (VUB)

by

Keven Bluteau

Accepted by the dissertation committee:

Prof. David Ardia (co-directeur de thèse, Université de Neuchâtel)
Prof. Kris Boudt (co-directeur de thèse, Vrije Universiteit Brussel)
Prof. Tim Kröncke (président du jury, Université de Neuchâtel)
Prof. Steven Vanduffel (Vrije Universiteit Brussel)
Prof. Leopoldo Catania (Aarhus University)
Prof. Marie-Claude Beaulieu (Université Laval)

Defended on May 6, 2019

UNIVERSITÉ DE NEUCHÂTEL

IMPRIMATUR POUR LA THÈSE

(cotutelle avec Vrije Universiteit Brussel)

Modeling Latent Variables in Economics and Finance

Keven BLUTEAU

UNIVERSITÉ DE NEUCHÂTEL

FACULTÉ DES SCIENCES ÉCONOMIQUES

La Faculté des sciences économiques,
sur le rapport des membres du jury

Prof. David Ardia (co-directeur de thèse, Université de Neuchâtel)
Prof. Kris Boudt (co-directeur de thèse, Vrije Universiteit Brussel)
Prof. Tim Kröncke (président du jury, Université de Neuchâtel)
Prof. Steven Vanduffel (Vrije Universiteit Brussel)
Prof. Leopoldo Catania (Aarhus University)
Prof. Marie-Claude Beaulieu (Université de Laval)

Autorise l'impression de la présente thèse.

Neuchâtel, le 27 mai 2019

Le doyen

Mehdi Farsi

■ Téléphone : +41 32 718 15 00 ■ E-mail : secretariat.seco@unine.ch ■ www.unine.ch

There is a philosophy that says that if something is unobservable – unobservable in principle – it is not part of science. If there is no way to falsify or confirm a hypothesis, it belongs to the realm of metaphysical speculation, together with astrology and spiritualism. By that standard, most of the universe has no scientific reality – it's just a figment of our imaginations.

LEONARD SUSSKIND
The Black Hole War: My Battle with Stephen Hawking to Make the World Safe for Quantum Mechanics

Acknowledgment

The completion of my thesis represents a significant milestone in my life. While challenging, the writing of this thesis has been a fruitful, fascinating, and rewarding process. My experience would, of course, not have been the same without the support and friendship of several people. As such, I would like to thank them.

First, I am grateful to my two supervisors, David Ardia and Kris Boudt, for their support throughout the entire process. Notably, I have been fortunate to meet David Ardia, during my master degree at University Laval, as this was the turning point in my life which kick–started my entire career as a researcher. For their trust in me and their commitment to my success, I am forever grateful.

Second, I am very thankful to the dissertation committee, Tim A. Kroencke, Marie–Claude Beaulieu, Leopoldo Catania, and Steven Vanduffel, for accepting to review my thesis and contributing their precious time and effort to the achievement of this work.

I would also like to thank the University of Neuchâtel and Vrije Universiteit Brussel for providing me with the necessary resources and truly outstanding research conditions. Primarily, I am grateful to the academic and administrative staff at the Institute of Financial Analysis, notably, Peter Fiechter, Carolina Salva, and Kira Facchinetti. Moreover, I extend a special thanks to Michel Dubois for which our several conversations and debates have generated several ideas relevant to this thesis and for future research projects.

Additionally, I extend my gratitude to my Ph.D. colleagues at the University of Neuchâtel and Vrije Universiteit Brussel for their support and friendship. They have made enjoyable the low and high moments of this entire journey.

Finally, and most significantly, I am thankful to my family and friends for their tremendous support and friendship throughout my life.

Abstract

The subject of unobservable variables encompasses this thesis. These latent (*i.e.*, unobservable) variables must be inferred using statistical models or observable proxies. The objectives of my doctoral thesis are to develop and test new statistical models to infer these variables and link them to the analysis and improvement of economic and financial decisions.

In my first essay, I tackle the evaluation of volatility models which allow for (latent) structural breaks. It is of utmost importance to capture these breaks in a timely manner, as a precise measure of volatility is crucial for optimal decision–making that requires a trade–off between expected return and risk, as well as for applications in asset pricing and risk management. However, no empirical study has been done to evaluate the overall performance of volatility model considering structural breaks. To that end, I perform a large–scale empirical study to compare the forecasting performance of single–regime and Markov–switching GARCH (MSGARCH) models, from a risk management perspective. I find that, for daily, weekly, and ten–day equity log–returns, MSGARCH models yield more accurate Value–at–Risk, Expected Shortfall, and left–tail distribution forecasts than their single–regime counterpart. Also, my results indicate that accounting for parameter uncertainty improves left–tail predictions, independently of the inclusion of the Markov–switching mechanism.

While my first essay tackles the modeling of latent variables from a statistical point of view, my second and third essay capture a more novel variable, namely the sentiment expressed in written communications.

My second essay addresses the development and testing of new text–based proxies for economic sentiment. More specifically, I introduce a general sentiment engineering framework that optimizes the design for forecasting purposes in a high–dimensional context. I apply the new methodology to the forecasting of the US industrial production, which is usually predicted using available quantitative variables from a large panel of indicators. I find that, compared to the use of high–dimensional forecasting techniques based solely economic and financial indicators, the additional use of optimized news–based sentiment values yield significant forecasting accuracy gains for the nine–month and annual growth rates of the US industrial production.

My third essay focuses on the analysis of the dynamics of abnormal tone or sentiment around the time of events. To do so, I introduce the Cumulative Abnormal Tone (CAT) event study and Generalized Word Power methodologies. I apply these methodologies to media reports

in newswires, newspapers, and web publications about firms' future performance published around the quarterly earnings announcements of non–financial S&P 500 firms over the period 2000–2016. I find that the abnormal tone is more sensitive to negative earnings surprises than positive ones. Additionally, I report that investors overreact to the abnormal tone contribution of web publications at earnings announcement dates, which generates a stock price reversal in the following month. This result is consistent with an overreaction pattern on the abnormal tone and psychological biases such as the representativeness heuristic. Moreover, it highlights that there is heterogeneity in the informational value of different types of media.

Keywords: abnormal return, abnormal tone, earnings announcements, elastic net, expected shortfall, forecasting performance, GARCH, generalized word power, large–scale study, MSGARCH, news media, risk management, sentiment analysis, sentometrics, textual tone, time–series aggregation, topic–sentiment, US industrial production, value–at–risk

Résumé

Le sujet des variables latentes est au coeur de cette thèse. Ces variables latentes (*i.e.*, non observables) doivent être inférées à l'aide de modèles statistiques ou de variables proxy observables. Les objectifs de ma thèse de doctorat sont de développer et de tester de nouveaux modèles statistiques pour déduire ces variables afin de les utiliser pour l'amélioration des décisions économiques et financières.

Dans mon premier chapitre de thèses, je traite de l'évaluation des modèles de volatilité qui intègre de possible changement (latent) structurels dans les paramètres du modèle. Il est d'une importance capitale pour capturer ces changements structurels rapidement, comme une mesure précise de la volatilité est cruciale pour la prise optimale de décision qui nécessite un compromis entre le rendement prévu et le risque, ainsi que pour des applications dans l'évaluation des prix d'actifs et en gestion des risques. Cependant, aucune étude empirique n'a été réalisée pour évaluer la performance globale de modèles de volatilité qui prennent en compte les changements structurels. À cette fin, j'entreprends une étude à grande échelle empirique pour comparer la performance de prévision de modèle GARCH sans changement de régime et de modèle GARCH à changement de régimes Markovien (MSGARCH) du point de vue d'un gestionnaire des risques. Les résultats indiquent que, pour tous les horizons de prédictions considérées, les modèles MS-GARCH génère des prédictions plus précis de la Value-at-risk, d'Expected Shortfall et de la densité que les modèles GARCH sans changement de régime. De plus, mes résultats indiquent que la prise en compte de l'incertitude des paramètres améliore les prévisions de la densité, indépendamment de l'inclusion du mécanisme Markovien.

Tandis que mon premier chapitre de thèses à une emphase sur la modélisation de variables latentes d'un point de vue de la modélisation statistique, le second et troisième chapitres tentent de capturer une variable plus originale: le sentiment exprimé dans les communications écrites.

Mon deuxième chapitre de thèses fait face au développement et l'évaluation de nouveau proxy de sentiment économique basé sur des document textuelles. Spécifiquement, j'introduis une infrastructure générale de développement d'indices de sentiment qui sont optimisés avec l'objectif de faire de la prédiction dans un contexte de régression à grande dimension. J'applique cette nouvelle méthodologie à la prédiction de la production industrielle américaine. Mes résultats indiquent que, comparé à l'utilisation unique de variables économiques et financières, l'ajout de l'utilisation d'indices de sentiments textuelles économiques optimisés ajoute de manière impor-

tante au pouvoir de prédiction de la croissance économique américaine pour des horizons de neuf mois et un ans.

Mon troisième chapitre de thèses à comme emphase l'analyse de la dynamique de sentiment textuelles anormaux près d'évènement financier. J'introduis, en autre, l'analyse d'évènement basé sur le sentiment anormal cumulatif et le Generalized Word Power, une méthode de calcul du sentiment. J'applique ces méthodologies sur des articles médiatiques provenant de journaux, du web et de fil de presse qui discutent des entreprises publiques non-financières près de l'annonce des résultats trimestriels. Les résultats indiquent que le sentiment anormal est plus sensible aux surprises de bénéfices négatives qu'aux surprises de bénéfices positives. De plus, je note que les investisseurs ont une réaction trop forte à la contribution au sentiment anormal des articles provenant du web. Cela, en autre, fait en sorte que l'on observe un inversement du prix de l'action après l'annonce des bénéfices pour le mois qui suit les annonces. Les résultats sont conformes avec des évidences dans le domaine de la psychologie humaine telle que le biais heuristique de représentation. De plus, les résultats mettent en évidence qu'il y a de l'hétérogénéité dans la valeur de l'information de différent types de médias.

Keywords: rendement anormaux, sentiment anormaux, annonce de bénéfices, elastic net, expected shortfall, pouvoir de prédictions, GARCH, generalized word power, étude à grande échelle, MSGARCH, articles médiatiques, gestion des risques, analyse du sentiment, sentometrics, sentiment textuelle, aggregation temporelle, sentiment par sujet, production industrielle américaine, value–at–risk

Contents

List of Tables

List of Figures

General Introduction

The suggestion that unobservable or latent variables drive observable variables is highly important in several areas of social science, including finance, economics, and psychology. A non–formal definition of latent variables is that *"latent variables are hypothetical constructs that cannot be directly measured"* (MacCallum and Austin, 2000).[1] Several variables in economics and finance such as stock return volatility, business cycles or regimes, or economics and market sentiment arguably fit that definition. Therefore, these variables have to be inferred through econometric models or by observable proxies. The objectives of my doctoral thesis are to develop and test statistical models to infer these variables and relate them to analyzing and improving economic and financial decisions. I focus on two types of latent variables: volatility and sentiment. I analyze the identification of these latent variables from an econometric perspective leveraging the recent availability of big data and advances in computer knowledge. I believe that this thesis will have important implications in economics and finance. Below, I introduce the three chapter of my doctoral thesis.

1. Regime–switching in volatility models

Under the regulation of the Basel Accords, risk managers of financial institutions must make use of state–of–the–art methodologies for monitoring financial risks (Board of Governors of the Federal Reserve Systems, 2012). As a consequence, the modeling of volatility is one of the central focus of risk management practitioners. Therefore, researchers have been hard at work to develop volatility models that capture the dynamics and statistical properties of financial market instruments. Indeed, the importance of the volatility modeling field was recognized in 2003 when Robert Engle received the Nobel Prize for his "contribution to methods of analyzing economic time series with time–varying volatility". His contribution encompasses most notably the seminal paper on the AutoRegressive Conditional Heteroskedasticity (ARCH) model (Engle, 1982), which sparked the development of a long and lasting literature on volatility modeling. Four years later, an important generalization to the ARCH model, namely the Generalized ARCH (GARCH) model, was introduced by Bollerslev (1986). The GARCH model takes into account two well–accepted stylized facts of financial returns, that is, volatility vary in time, and high (low) volatility periods tend to be followed by similarly high (low) volatility periods, a

[1] See, Bollen (2002), for various non–formal and formal definition of latent variables.

stylized fact usually referred to as "volatility clustering". From there, multiple extensions of the standard GARCH scedastic function have been proposed in order to capture additional stylized facts of the financial market. These so–called GARCH–type models recognize that there may be skewness (Luca and Loperfido, 2015; Franceschini and Loperfido, 2010; Luca and Loperfido, 2015), excess of kurtosis (Bollerslev, 1987), asymmetries in the volatility dynamic (Nelson, 1991; Zakoian, 1994), and parameters uncertainty (Ardia et al., 2012).

Academics and practitioners using GARCH–type models in a real environment setting realized that the GARCH–type model parameters' estimate were often unrealistic. Indeed, while seemingly good at capturing the observed stylized facts of asset returns, there is often an unrealistic level of volatility persistence in estimated GARCH–type models, even sometimes close to the unit root. Several studies pointed out that the unrealistic level of persistence in GARCH–type model is associated to asset returns series which display regime–changes or structural breaks in their volatility dynamics (see, *e.g.*, Lamoureux and Lastrapes, 1990; Caporale et al., 2003; Bauwens et al., 2014). Thus, volatility predictions by GARCH–type models may fail to capture the true variation in volatility in the case of regime–changes in the true volatility process. A solution to this problem is to allow the parameters of the GARCH–type model to vary over time according to a latent discrete Markov process. This approach is called the Markov–switching GARCH (MSGARCH) model, which leads to volatility forecasts that can quickly adapt to variations in the parameters of a volatility model (see, *e.g.*, Marcucci, 2005; Ardia, 2008).

The objective of the first chapter of my thesis, titled *"Forecasting risk with Markov–switching GARCH models: A large-scale performance study"* (Ardia et al., 2018), is to determine the added value of MSGARCH–type models compared with standard single–regime GARCH–type models in the context of a risk management task. Currently, it is unclear if these models have sufficient added practical forecasting power to outweigh the additional model complexity compared with single–regime GARCH models. I aim to get general recommendations for risk managers and regulators regarding the usage of MSGARCH models. I, therefore, take the perspective of a risk manager evaluating the risk measures and density forecasts of several MSGARCH– and GARCH–type models for a large universe of stocks, equity indices, and foreign exchange rates. Moreover, I investigate the impact of the estimation method for such models. To that end, I conduct the study by comparing the forecasting performance of these models conditional on if the frequentist approach or the Bayesian approach has been used to estimate the models.

The empirical results can be summarized as follows. First, MSGARCH models provide risk

managers with better Value–at–Risk and Expected Shortfall forecasts as well as better left–tail density forecasts than their single–regime counterparts. This result is especially strong for stock return data, less for stock indices, and practically non–existent for currencies. This indicates that MSGARCH models do not have the same utility for risk managers of different trading desks. Second, the added value of taking into account parameter uncertainty is clear and does not depends on the choice of volatility model. The most substantial improvement, however, is seen in single–regime models. Overall, I recommend that risk managers use MSGARCH models in their practice and account for parameter uncertainty by using Bayesian principles.

2. Textual sentiment indices as economic predictors

Forecasts about all aspects of the economy are fundamental for optimal economic policy and business decision–making. Therefore, it seems self–evident that forecasters should make use of all data available for performing their forecasting task. In the last decade, there has been a trend of digitalization and fast distribution of textual information. This has sparked life into the use of alternative data for economic forecasting. Moreover, the rise in the natural language processing field and the development of high–dimensional statistics and machine learning methods has enabled the use of those new potential predictors of the economy.

In practice, however, the dominating approach is to forecast economic variables using large panel of macroeconomic indicators (Stock and Watson, 2002), consumer surveys (Bram and Ludvigson, 1998), and financial variables (Espinoza et al., 2012). Thus, forecasts about the economy currently do not leverage all available information. I posit that this is mostly due to the nature of alternative data. Alternative sources of data, particularly news reports, are highly unstructured and it is not straightforward, in most cases, to extract the relevant information from the texts. Thus, it is of high interest for forecasters to obtain a sound methodological framework aimed at leveraging the information in news reports to forecast economic variables.

The objective of the second chapter of my thesis, titled *"Questioning the news about economic growth: Sparse forecasting using thousands of news–based sentiment values"* (Ardia et al., 2019), is to complement the traditional quantitative variables with predictors obtained from a large set of sentiment values expressed by authors of news discussing a country's economy to obtain timely forecasts of the country's economic growth. These texts need to be selected, transformed into sentiment values, and then aggregated. I thus propose a methodology that structures the sentiment data from news articles in a natural way by considering three dimensions of sentiment and textual data. The dimensions are: (1) the sentiment computation method (*e.g.*, using

various lexicons), (2) the topic of the texts (*e.g.*, "real estate market" or "job creation"), and (3) the time (*e.g.*, short and long–term sentiment indices). The consideration of these dimensions allows creating interpretable indices of sentiment which are then used within high–dimensional forecasting methods.

I illustrate the methodology for the case of forecasting the economic growth for the United States. I find that, for an out–of–sample evaluation window ranging from January 2001 to December 2016, the text–based sentiment indices computed from news in major US newspapers provide additional predictive power for the nine–month and annual growth rates of the US industrial production index, controlling for standard use of macroeconomic, sentiment–survey, and financial variables. Moreover, I test to which extent each dimension of the sentiment index (*i.e.*, sentiment calculation method, topic, and time) matters. I find that the optimization of all dimensions is essential to achieve a high forecasting accuracy, but, in order, the most relevant is the time dimension, followed by the topic, and then the sentiment calculation method. My results are shown to be robust to various choices of implementations.

3. The analysis of abnormal tone in relation to events

The event study methodology serves as a research tool to measure the impact of a specific event on a variable. In financial and accounting contexts, event studies are used to identify how firm–specific and economy–wide events impact firms' valuations (see, *e.g.*, MacKinlay, 1997). For instance, in the case of firms' earnings announcement events, the consensus view is that there is an immediate abnormal return reaction to the earnings result and a significant post–earnings–announcement abnormal return drift (see, *e.g.*, Bernard and Thomas, 1989, 1990).

When the event is important, it is natural to expect that the media discuss it in the days near the event date. It can further be expected that the tone of the media communication about future firm performance has a relationship with the market reaction. Understanding this relationship requires a methodology for the joint analysis of the dynamics of media tone and market returns around events.

The objective of the third chapter of my thesis, titled *'Media and the stock market: A CAT and CAR analysis"*, is to introduce a comprehensive framework for the analysis of the abnormal tone near events. Thus, I introduce the Cumulative Abnormal Tone (CAT) event study methodology and Generalized Word Power tone computation approach. I apply these methodologies to analyze media abnormal tone dynamics about firms' future performance in the daily media textual information written about non–financial S&P 500 firms near quarterly

earnings announcements for the period ranging from 2000 to 2016. I analyze the drivers of the CAT dynamics near earnings announcements as well as the predictive power of CAT over the post–earnings–announcement abnormal returns.

My results suggest that firm– and earnings–specific variables drive the CAT measure at the earnings announcement dates. Moreover, the level of CAT and cumulative abnormal return at the earnings announcement dates induce a post–earnings–announcement abnormal tone momentum. I also find that the CAT measure is more sensitive to negative earnings surprises than positive ones. Additionally, I report that the CAT measure predicts a post–earnings–announcement abnormal return reversal and this effect is stronger for negative CAT. This result suggests an overreaction pattern to news at earnings announcement dates. Using a segregated analysis of the abnormal tone by type of sources, I additionally find that the reversal effect is mainly attributed to the CAT measure contribution of web publications.

References

Ardia, D., 2008. Financial Risk Management with Bayesian Estimation of GARCH Models: Theory and Applications. Springer.

Ardia, D., Baştürk, N., Hoogerheide, L.F., Van Dijk, H.K., 2012. A comparative study of Monte Carlo methods for efficient evaluation of marginal likelihood. Computational Statistics & Data Analysis 56, 3398–3414. doi:10.1016/j.csda.2010.09.001.

Ardia, D., Bluteau, K., Boudt, K., 2019. Questioning the news about economic growth: Sparse forecasting using thousands of news-based sentiment values. International Journal of Forecasting (in press). doi:10.1016/j.ijforecast.2018.10.010.

Ardia, D., Bluteau, K., Boudt, K., Catania, L., 2018. Forecasting risk with Markov–switching GARCH models: A large-scale performance study. International Journal of Forecasting 34, 733 – 747. doi:10.1016/j.ijforecast.2018.05.004.

Bauwens, L., Dufays, A., Rombouts, J.V.K., 2014. Marginal likelihood for Markov–switching and change–point GARCH models. Journal of Econometrics 178, 508–522. doi:10.1016/j.jeconom.2013.08.017.

Bernard, V.L., Thomas, J.K., 1989. Post–earnings–announcement drift: Delayed price response or risk premium? Journal of Accounting Research 27, 1–36. doi:10.2307/2491062.

Bernard, V.L., Thomas, J.K., 1990. Evidence that stock prices do not fully reflect the implications of current earnings for future earnings. Journal of Accounting and Economics 13, 305–340. doi:10.1016/0165-4101(90)90008-R.

Board of Governors of the Federal Reserve Systems, 2012. 99th Annual Report. Technical Report. Board of Governors of the Federal Reserve Systems.

Bollen, K.A., 2002. Latent variables in psychology and the social sciences. Annual Review of Psychology 53, 605–634.

Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics 31, 307–327. doi:10.1016/0304-4076(86)90063-1.

Bollerslev, T., 1987. A conditionally heteroskedastic time series model for speculative prices and rates of return. Review of Economics and Statistics 69, 542–547. doi:10.2307/1925546.

Bram, J., Ludvigson, S., 1998. Does consumer confidence forecast household expenditure? A sentiment index horse race. Economic Policy Review 4, 59–78.

Caporale, G.M., Pittis, N., Spagnolo, N., 2003. IGARCH models and structural breaks. Applied Economics Letters 10, 765–768. doi:10.1080/1350485032000138403.

Engle, R.F., 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica 50, 987–1008. doi:10.2307/1912773.

Espinoza, R., Fornari, F., Lombardi, M.J., 2012. The role of financial variables in predicting economic activity. Journal of Forecasting 31, 15–46. doi:10.1002/for.1212.

Franceschini, C., Loperfido, N., 2010. Mathematical and Statistical Methods for Actuarial Sciences and Finance. Springer Milan, Milano. chapter A skewed GARCH–type model for multivariate financial time series. 143–152. doi:10.1007/978-88-470-1481-7_15.

Lamoureux, C.G., Lastrapes, W.D., 1990. Persistence in variance, structural change, and the GARCH model. Journal of Business & Economic Statistics 8, 225–234. doi:10.2307/1391985.

Luca, G.D., Loperfido, N., 2015. Modelling multivariate skewness in financial returns: A SGARCH approach. European Journal of Finance 21, 1113–1131. doi:10.1080/1351847X.2011.640342.

MacCallum, R.C., Austin, J.T., 2000. Applications of structural equation modeling in psychological research. Annual Review of Psychology 51, 201–226.

MacKinlay, A.C., 1997. Event studies in economics and finance. Journal of Economic Literature 35, 13–39.

Marcucci, J., 2005. Forecasting stock market volatility with regime–switching GARCH models. Studies in Nonlinear Dynamics & Econometrics 9. doi:10.2202/1558-3708.1145.

Nelson, D.B., 1991. Conditional heteroskedasticity in asset returns: A new approach. Econometrica 59, 347–370. doi:10.2307/2938260.

Stock, J.H., Watson, M.W., 2002. Forecasting using principal components from a large number of predictors. Journal of the American Statistical Association 97, 1167–1179. doi:10.1198/016214502388618960.

Zakoian, J.M., 1994. Threshold heteroskedastic models. Journal of Economic Dynamics & Control 18, 931–955. doi:10.1016/0165-1889(94)90039-6.

Forecasting risk with Markov–switching GARCH models: A large–scale performance study

Keven Bluteau – Chapter 1
joint work with David Ardia, Kris Boudt, and Leopoldo Catania
Published in International Journal of Forecasting

Abstract

We perform a large–scale empirical study to compare the forecasting performance of single–regime and Markov–switching GARCH (MSGARCH) models from a risk management perspective. We find that, for daily, weekly, and ten–day equity log–returns, MSGARCH models yield more accurate Value–at–Risk, Expected Shortfall, and left–tail distribution forecasts than their single–regime counterpart. Also, our results indicate that accounting for parameter uncertainty improves left–tail predictions, independently of the inclusion of the Markov–switching mechanism.

Keywords: GARCH, MSGARCH, forecasting performance, large–scale study, Value–at–Risk, Expected Shortfall, risk management

1. Introduction

Under the regulation of the Basel Accords, risk managers of financial institutions need to rely on state–of–the–art methodologies for monitoring financial risks (Board of Governors of the Federal Reserve Systems, 2012). Clearly, the use of a regime–switching time–varying volatility model and Bayesian estimation methods can be considered to be strong candidates for being classified as state–of–the–art methodologies. However, many academics and practitioners also consider the single–regime volatility model and the use of frequentist estimation via Maximum Likelihood (ML) as state–of–the–art. Risk managers disagree whether the computational complexity of a regime–switching model and the Bayesian estimation method pay off in terms of a higher accuracy of their financial risk monitoring system. We study this question for monitoring the individual risks of a large number of financial assets.

Among the various building–blocks of any risk management system, the specification of the conditional volatility process is key, especially for short–term horizons (McNeil et al., 2015). Research on modeling volatility using time series models has proliferated since the creation of the original ARCH model by Engle (1982) and its generalization by Bollerslev (1986). From there, multiple extensions of the GARCH scedastic function have been proposed to capture additional stylized facts observed in financial markets, such as nonlinearities, asymmetries, and

long–memory properties; see Engle (2004) for a review. These so–called GARCH–type models are today essential tools for risk managers.

An appropriate risk model should be able to accommodate the properties of financial returns. Recent academic studies show that many financial assets exhibit structural breaks in their volatility dynamics and that ignoring this feature can have large effects on the precision of the volatility forecast (see, *e.g.*, Lamoureux and Lastrapes, 1990; Bauwens et al., 2014b). As noted by Danielsson (2011), this shortcoming in the individual forecasting systems can have systemic consequences. He refers to these single–regime volatility models as one of the culprits of the great financial crisis: *"(...) the stochastic process governing market prices is very different during times of stress compared to normal times. We need different models during crisis and non–crisis and need to be careful in drawing conclusions from non–crisis data about what happens in crises and vice versa".*

A way to address the *switch* in the return process is provided by Markov–switching GARCH models (MSGARCH) whose parameters can change over time according to a discrete latent (*i.e.*, unobservable) variable. These models can quickly adapt to variations in the unconditional volatility level, which improves risk predictions (see, *e.g.*, Marcucci, 2005; Ardia, 2008).

Initial studies on Markov–switching autoregressive heteroscedastic models applied to financial times series focus on ARCH specifications and thus omit a lagged value of the conditional variance in the variance equation (Cai, 1994; Hamilton and Susmel, 1994). The use of ARCH instead of GARCH dynamics leads to computational tractability in the likelihood calculation. Indeed, Gray (1996) shows that, given a Markov chain with K regimes and T observations, the evaluation of the likelihood of a Markov–switching model with general GARCH dynamics requires the integration over all K^T possible paths, rendering the estimation infeasible. While this difficulty is not present in ARCH specifications, the use of lower order GARCH models tends to offer a more parsimonious representation than higher order ARCH models.

Gray (1996), Dueker (1997) and Klaassen (2002) tackle the *path–dependence problem* of MSGARCH through approximation, by collapsing the past regime–specific conditional variances according to ad–hoc schemes. A further solution is to consider alternatives to traditional Maximum Likelihood estimation. Bauwens et al. (2014b) recommend to use Bayesian estimation methods that are still feasible through the so–called data augmentation techniques and particle MCMC techniques. Augustyniak (2014) relies on a Monte Carlo EM algorithm with importance sampling. In our study, we consider the alternative approach provided by Haas et al. (2004),

who let the GARCH processes of each state evolve independently of the GARCH process in the other states. Besides avoiding the path–dependence problem in traditional Maximum Likelihood estimation, their model allows for a clear–cut interpretation of the variance dynamics in each regime.

The first contribution of our paper is to test if, indeed, MSGARCH models provide risk managers with useful tools that can improve their volatility forecasts.[1] To answer this question, we perform a large–scale empirical analysis in which we compare the risk forecasting performance of single–regime and Markov–switching GARCH models. We take the perspective of a risk manager working for a fund manager and conduct our study on the daily, weekly and ten–day log–returns of a large universe of stocks, equity indices, and foreign exchange rates. Thus, in contrast to Hansen and Lunde (2005), who compare a large number of GARCH–type models on a few series, we focus on a few GARCH and MSGARCH models and a large number of series. For single–regime and Markov–switching specifications, the scedastic specifications we consider account for different reactions of the conditional volatility to past asset returns. More precisely, we consider the symmetric GARCH model (Bollerslev, 1986) as well as the asymmetric GJR model (Glosten et al., 1993). These scedastic specifications are integrated into the MSGARCH framework with the approach of Haas et al. (2004). For the (regime–dependent) conditional distributions, we use the symmetric and the Fernández and Steel (1998) skewed versions of the Normal and Student–t distributions. Overall, this leads to sixteen models.

Our second contribution is to test the impact of the estimation method on the performance of the volatility forecasting model. GARCH and MSGARCH models are traditionally estimated with a frequentist (typically via ML) approach; see Haas et al. (2004), Marcucci (2005) and Augustyniak (2014). However, several recent studies have argued that a Bayesian approach offers some advantages. For instance, Markov chain Monte Carlo (MCMC) procedures can

[1] Our study focuses exclusively on GARCH and MSGARCH models. GARCH is the workhorse model in financial econometrics and has been investigated for decades. It is widely used by practitioners and academics; see for instance Bams et al. (2017) and Herwartz (2017). MSGARCH is the most natural and straightforward extension to GARCH. Alternative conditional volatility models include stochastic volatility models (Taylor, 1994; Jacquier et al., 1994), realized measure–based conditional volatility models such as HEAVY (Shephard and Sheppard, 2010) or Realized GARCH (Hansen et al., 2011), or even combinations of these (Opschoor et al., 2017). Note finally that our study only considers the (1,1)–lag specification for the GARCH and MSGARCH models. While there is a clear computational cost of considering higher orders for (MS)GARCH model specifications, the payoff in terms of improvement in forecasting precision may be low. In fact, several studies have shown that increasing the orders does not lead to a substantial improvement of the forecasting performance in case of predicting the conditional variance of asset returns (see, *e.g.*, Hansen and Lunde, 2005). We tested whether this result also holds for our sample and performed the fit of GARCH(p, q) and GJR$(p, 1, q)$ models over the three universes of stocks, indices and currencies, for rolling windows of 1,500 points, and selected the best in–sample model via BIC. We found that the (1,1) specification is selected in the vast majority of the fits.

explore the joint posterior distribution of the model parameters, and parameter uncertainty is naturally integrated into the risk forecasts via the predictive distribution (Ardia, 2008; Bauwens et al., 2010, 2014a; Geweke and Amisano, 2010; Ardia et al., 2017c).

Combining the sixteen model specifications with the frequentist and Bayesian estimation methods, we obtain 32 possible candidates for the state–of–the–art methodology for monitoring financial risk. We use an out–of–sample evaluation period of 2,000 days, that ranges from (approximately) 2005 to 2016 and consists of daily log–returns. We evaluate the accuracy of the risk prediction models in terms of estimating the Value–at–Risk (VaR), the Expected Shortfall (ES), and the left–tail (*i.e.*, losses) of the conditional distribution of the assets' returns.

Our empirical results suggest a number of practical insights which can be summarized as follows. First, we find that MSGARCH models deliver better VaR, ES, and left–tail distribution forecasts than their single–regime counterpart. This is especially true for stock return data. Moreover, improvements are more pronounced when the Markov–switching mechanism is applied to simple specifications such as the GARCH–Normal model. Second, accounting for parameter uncertainty improves the accuracy of the left–tail predictions, independently of the inclusion of the Markov–switching mechanism. Moreover, larger improvements are observed in the case of single–regime models. Overall, we recommend risk managers to rely on more flexible models and to perform inference accounting for parameter uncertainty.

In addition to showing the good performance of MSGARCH models and Bayesian estimation methods, we refer risk managers to our R package MSGARCH (Ardia et al., 2017a,b), which implements MSGARCH models in the R statistical language with efficient C++ code.[2] We hope that this paper and the accompanying package will encourage practitioners and academics in the financial community to use MSGARCH models and Bayesian estimation methods.

The paper proceeds as follows. Model specification, estimation, and forecasting are presented in Section 2. The datasets, the testing design, and the empirical results are discussed in Section 3. Section 4 concludes.

2. Risk forecasting with Markov–switching GARCH models

A key aspect in quantitative risk management is the modeling of the risk drivers of the securities held by the fund manager. We consider here the univariate parametric framework, that

[2]Our research project was funded by the 2014 SAS/IIF forecasting research grant, to compare MSGARCH vs. GARCH models, and to develop and render publicly available the computer code for the estimation of MSGARCH models.

computes the desired risk measure in four steps. First, a statistical model which describes the daily log–returns (profit and loss, P&L) dynamics is determined. Second, the model parameters are estimated for a given estimation window. Third, the one/multi–day ahead distribution of log–returns is obtained (either analytically or by simulation). Fourth, relevant risk measures such as the Value–at–Risk (VaR) and the Expected Shortfall (ES) are computed from the distribution. The VaR represents a quantile of the distribution of log–returns at the desired horizon, and the ES is the expected loss when the loss exceeds the VaR level (Jorion, 2006). Risk managers can then allocate risk capital given their density or risk measure forecasts. Also, they can assess the quality of the risk model, *ex–post*, via statistical procedures referred to as *backtesting*.

2.1. Model specification

We define $y_t \in \mathbb{R}$ as the (percentage point) log–return of a financial asset at time t. To simplify the exposition, we assume that the log–returns have zero mean and are not autocorrelated.[3] The general Markov–switching GARCH specification can be expressed as:

$$y_t \,|\, (s_t = k, \mathcal{I}_{t-1}) \sim \mathcal{D}(0, h_{k,t}, \boldsymbol{\xi}_k)\,, \tag{1}$$

where $\mathcal{D}(0, h_{k,t}, \boldsymbol{\xi}_k)$ is a continuous distribution with zero mean, time–varying variance $h_{k,t}$, and additional shape parameters (*e.g.*, asymmetry) gathered in the vector $\boldsymbol{\xi}_k$.[4] Furthermore, we assume that the latent variable s_t, defined on the discrete space $\{1, \ldots, K\}$, evolves according to an unobserved first order ergodic homogeneous Markov chain with transition probability matrix $\mathbf{P} \equiv \{p_{i,j}\}_{i,j=1}^{K}$, with $p_{i,j} \equiv \mathbb{P}[s_t = j \,|\, s_{t-1} = i]$. We denote by $\mathcal{I}_{t-1}$ the information set up to time $t-1$, that is, $\mathcal{I}_{t-1} \equiv \{y_{t-i}, i > 0\}$. Given the parametrization of $\mathcal{D}(\cdot)$, we have $\mathbb{E}[y_t^2 \,|\, s_t = k, \mathcal{I}_{t-1}] = h_{k,t}$, that is, $h_{k,t}$ is the variance of y_t conditional on the realization of s_t and the information set $\mathcal{I}_{t-1}$.

As in Haas et al. (2004), the conditional variance of y_t is assumed to follow a GARCH–type model. More precisely, conditionally on regime $s_t = k$, $h_{k,t}$ is specified as a function of past returns and the additional regime–dependent vector of parameters $\boldsymbol{\theta}_k$:

$$h_{k,t} \equiv h(y_{t-1}, h_{k,t-1}, \boldsymbol{\theta}_k)\,,$$

[3] In practice, this means that we apply the (MS)GARCH models to de–meaned log–returns, as explained in Section 3.

[4] For $t = 1$, we initialize the regime probabilities and the conditional variances at their unconditional levels. To simplify exposition, we use henceforth for $t = 1$ the same notation as for general t, since there is no confusion possible.

where $h(\cdot)$ is a $\mathcal{I}_{t-1}$–measurable function, which defines the filter for the conditional variance and also ensures its positiveness. We further assume that $h_{k,1} \equiv \bar{h}_k$ $(k = 1, \ldots, K)$, where $\bar{h}_k$ is a fixed initial variance level for regime k, that we set equal to the unconditional variance in regime k. Depending on the form of $h(\cdot)$, we obtain different scedastic specifications. For instance, if:

$$h_{k,t} \equiv \omega_k + \alpha_k y_{t-1}^2 + \beta_k h_{k,t-1}\,,$$

with $\omega_k > 0$, $\alpha_k > 0$, $\beta_k \geq 0$ and $\alpha_k + \beta_k < 1$ $(k = 1, \ldots, K)$, we obtain the Markov–switching GARCH$(1,1)$ model presented in Haas et al. (2004).[5] In this case $\boldsymbol{\theta}_k \equiv (\omega_k, \alpha_k, \beta_k)'$.

Alternative definitions of the function $h(\cdot)$ can be easily incorporated in the model. For instance, to account for the well–known asymmetric reaction of volatility to the sign of past returns (often referred to as the *leverage effect*; see Black 1976), we specify a Markov–switching GJR$(1,1)$ model exploiting the volatility specification of Glosten et al. (1993):

$$h_{k,t} \equiv \omega_k + (\alpha_k + \gamma_k \mathbb{I}\{y_{t-1} < 0\})\, y_{t-1}^2 + \beta_k h_{k,t-1}\,,$$

where $\mathbb{I}\{\cdot\}$ is the indicator function, that is equal to one if the condition holds, and zero otherwise. In this case, the additional parameter $\gamma_k \geq 0$ controls the asymmetry in the conditional variance process. We have $\boldsymbol{\theta}_k \equiv (\omega_k, \alpha_k, \gamma_k, \beta_k)'$. Covariance–stationarity of the variance process conditionally on the Markovian state is achieved by imposing $\alpha_k + \beta_k + \kappa_k \gamma_k < 1$, where $\kappa_k \equiv \mathbb{P}[y_t < 0 \,|\, s_t = k, \mathcal{I}_{t-1}]$. For symmetric distributions we have $\kappa_k = 1/2$. For skewed distributions, κ_k is obtained following the approach of Trottier and Ardia (2016).

We consider different choices for $\mathcal{D}(\cdot)$. We take the standard Normal $(\mathcal{N})$ and the Student–t $(\mathcal{S})$ distributions. To investigate the benefits of incorporating skewness in our analysis, we also consider the standardized skewed version of $\mathcal{N}$ and $\mathcal{S}$ obtained using the mechanism of Fernández and Steel (1998) and Bauwens and Laurent (2005); see Trottier and Ardia (2016) for more details. We denote the standardized skew–Normal and the skew–Student–t by sk$\mathcal{N}$ and sk$\mathcal{S}$, respectively.

Overall, our model set includes 16 different specifications recovered as combinations of:

- The number of regimes, $K \in \{1, 2\}$. When $K = 1$, we label our specification as single–

[5] We require that the conditional variance in each regime is covariance–stationary. This is a stronger condition than in Haas et al. (2004), but this allows us to ensure stationarity for various forms of conditional variance and/or conditional distributions.

regime (SR), and, when $K = 2$, as Markov–switching (MS);

- The conditional variance specification: GARCH$(1,1)$ and GJR$(1,1)$;
- The choice of the conditional distribution $\mathcal{D}(\cdot)$, that is, $\mathcal{D} \in \{\mathcal{N}, \mathcal{S}, \text{sk}\mathcal{N}, \text{sk}\mathcal{S}\}$.[6]

2.2. Estimation

We estimate the models either through frequentist or Bayesian techniques. Both approaches require the evaluation of the likelihood function.

In order to write the likelihood function corresponding to the MSGARCH model specification (1), we regroup the model parameters into $\boldsymbol{\Psi} \equiv (\boldsymbol{\xi}_1, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\xi}_K, \boldsymbol{\theta}_K, \mathbf{P})$. The conditional density of y_t in state $s_t = k$ given $\boldsymbol{\Psi}$ and $\mathcal{I}_{t-1}$ is denoted by $f_{\mathcal{D}}(y_t \,|\, s_t = k, \boldsymbol{\Psi}, \mathcal{I}_{t-1})$.

By integrating out the state variable s_t, we obtain the density of y_t given $\boldsymbol{\Psi}$ and $\mathcal{I}_{t-1}$ only. The (discrete) integration is obtained as follows:

$$f(y_t \,|\, \boldsymbol{\Psi}, \mathcal{I}_{t-1}) \equiv \sum_{i=1}^{K} \sum_{j=1}^{K} p_{i,j}\, \eta_{i,t-1}\, f_{\mathcal{D}}(y_t \,|\, s_t = j, \boldsymbol{\Psi}, \mathcal{I}_{t-1})\,, \tag{2}$$

where $\eta_{i,t-1} \equiv \mathbb{P}[s_{t-1} = i \,|\, \boldsymbol{\Psi}, \mathcal{I}_{t-1}]$ is the filtered probability of state i at time $t-1$ and where we recall that $p_{i,j}$ denotes the transition probability of moving from state i to state j. The filtered probabilities $\{\eta_{k,t}; k = 1, \ldots, K; t = 1, \ldots, T\}$ are obtained via the Hamilton filter; see Hamilton (1989) and Hamilton (1994, Chapter 22) for details.

Finally, the likelihood function is obtained from (2) as follows:

$$\mathcal{L}(\boldsymbol{\Psi} \,|\, \mathcal{I}_T) \equiv \prod_{t=1}^{T} f(y_t \,|\, \boldsymbol{\Psi}, \mathcal{I}_{t-1})\,. \tag{3}$$

The ML estimator $\widehat{\boldsymbol{\Psi}}$ is obtained by maximizing the logarithm of (3). In the case of the Bayesian estimation, the likelihood function is combined with a prior $f(\boldsymbol{\Psi})$ to build the kernel of the

[6] We also tested the asymmetric EGARCH scedastic specification (Nelson, 1991) as well as alternative fat–tailed distributions, such as the Laplace and GED distributions. The performance results were qualitatively similar.

posterior distribution $f(\boldsymbol{\Psi} \,|\, \mathcal{I}_T)$. We build our prior from diffuse independent priors as follows:

$$
\begin{aligned}
f(\boldsymbol{\Psi}) &\propto f(\boldsymbol{\theta}_1, \boldsymbol{\xi}_1) \cdots f(\boldsymbol{\theta}_K, \boldsymbol{\xi}_K) f(\mathbf{P}) \, \mathbb{I}\{\bar{h}_1 < \cdots < \bar{h}_K\} \\
f(\boldsymbol{\theta}_k, \boldsymbol{\xi}_k) &\propto f(\boldsymbol{\theta}_k) f(\boldsymbol{\xi}_k) \, \mathbb{I}\{(\boldsymbol{\theta}_k, \boldsymbol{\xi}_k) \in \mathcal{CSC}_k\} \quad (k = 1, \ldots, K) \\
f(\boldsymbol{\theta}_k) &\propto f_{\mathcal{N}}(\boldsymbol{\theta}_k; \mathbf{0}, 1{,}000 \times \mathbf{I}) \, \mathbb{I}\{\boldsymbol{\theta}_k > \mathbf{0}\} \quad (k = 1, \ldots, K) \\
f(\boldsymbol{\xi}_k) &\propto f_{\mathcal{N}}(\boldsymbol{\xi}_k; \mathbf{0}, 1{,}000 \times \mathbf{I}) \, \mathbb{I}\{\xi_{k,1} > 0, \xi_{k,2} > 2\} \quad (k = 1, \ldots, K) \\
f(\mathbf{P}) &\propto \prod_{i=1}^{K} \left(\prod_{j=1}^{K} p_{i,j} \right) \mathbb{I}\{0 < p_{i,i} < 1\},
\end{aligned}
$$

where $\mathbf{0}$ and $\mathbf{I}$ denote a vector of zeros and an identity matrix of appropriate sizes, $f_{\mathcal{N}}(\bullet; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the multivariate Normal density with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, $\xi_{k,1}$ is the asymmetry parameter, and $\xi_{k,2}$ the tail parameter of the skewed Student–t distribution in regime k. The prior density for the transition matrix is obtained by assuming that the K rows are independent and follow a Dirichlet prior with all hyperparameters equal to two. Moreover, $\bar{h}_k \equiv \bar{h}_k(\boldsymbol{\theta}_k, \boldsymbol{\xi}_k)$ is the unconditional variance in regime k and $\mathcal{CSC}_k$ denotes the covariance–stationarity condition in regime k; see Trottier and Ardia (2016). As the posterior is of an unknown form (the normalizing constant is numerically intractable), it must be approximated by simulation techniques. In our case, MCMC draws from the posterior are generated with the adaptive random–walk Metropolis sampler of Vihola (2012). We use 50,000 burn–in draws and build the posterior sample of size 1,000 with the next 50,000 draws keeping only every 50th draw to diminish the autocorrelation in the chain.[7] For both the frequentist and the Bayesian estimation, we ensure positivity and stationarity of the conditional variance in each regime during the estimation. Moreover, we impose constraints on the parameters to ensure that volatilities under the MSGARCH specification cannot be generated by a single–regime specification. In the case of the frequentist estimation, these constraints are enforced in the likelihood optimization by using mapping functions. For the Bayesian estimation, this is achieved through the prior.

[7] We performed several sensitivity analyses to assess the impact of the estimation setup. First, we changed the hyper–parameter values. Second, we ran longer MCMC chains. Third, we used 10,000 posterior draws instead of 1,000. Finally, we tested an alternative MCMC sampler based on adaptive mixtures of Student–t distribution (Ardia et al., 2009). In all cases, the conclusions remained qualitatively similar. Note that we choose a long burn–in sample size to rule out the possibility that results are affected by non–convergent MCMC chains. For simpler applications where it is easier to check the convergence of the MCMC algorithm, a lower value for the burn–in phase can be chosen to speed up the computations. In the MSGARCH package (Ardia et al., 2017a,b), the default value is set to 5,000.

2.3. Density and downside risk forecasting

Generating one–step ahead density and downside risk forecasts (VaR and ES) with MSGARCH models is straightforward. First, note that the one–step ahead conditional probability density function (PDF) of y_{T+1} is a mixture of K regime–dependent distributions:

$$f(y_{T+1} \,|\, \boldsymbol{\Psi}, \mathcal{I}_T) \equiv \sum_{k=1}^{K} \pi_{k,T+1} f_{\mathcal{D}}(y_{T+1} \,|\, s_{T+1} = k, \boldsymbol{\Psi}, \mathcal{I}_T)\,, \tag{4}$$

with mixing weights $\pi_{k,T+1} \equiv \sum_{i=1}^{K} p_{i,k}\eta_{i,T}$ where $\eta_{i,T} \equiv \mathbb{P}[s_T = i \,|\, \boldsymbol{\Psi}, \mathcal{I}_T]$ $(i = 1, \ldots, K)$ are the filtered probabilities at time T. The cumulative density function (CDF) is obtained from (4) as follows:

$$F(y_{T+1} \,|\, \boldsymbol{\Psi}, \mathcal{I}_T) \equiv \int_{-\infty}^{y_{T+1}} f(z \,|\, \boldsymbol{\Psi}, \mathcal{I}_T)\mathrm{d}z\,. \tag{5}$$

Within the frequentist framework, the predictive PDF and CDF are simply computed by replacing $\boldsymbol{\Psi}$ by the ML estimator $\widehat{\boldsymbol{\Psi}}$ in (4) and (5). Within the Bayesian framework, we proceed differently, and integrate out the parameter uncertainty. Given a posterior sample $\{\boldsymbol{\Psi}^{[m]}, m = 1, \ldots, M\}$, the predictive PDF is obtained as:

$$f(y_{T+1} \,|\, \mathcal{I}_T) \equiv \int_{\boldsymbol{\Psi}} f(y_{T+1} \,|\, \boldsymbol{\Psi}, \mathcal{I}_T) f(\boldsymbol{\Psi} \,|\, \mathcal{I}_T)\mathrm{d}\boldsymbol{\Psi} \approx \frac{1}{M}\sum_{m=1}^{M} f(y_{T+1} \,|\, \boldsymbol{\Psi}^{[m]}, \mathcal{I}_T)\,. \tag{6}$$

The predictive CDF is given by:

$$F(y_{T+1} \,|\, \mathcal{I}_T) \equiv \int_{-\infty}^{y_{T+1}} f(z \,|\, \mathcal{I}_T)\mathrm{d}z\,. \tag{7}$$

For both estimation approaches, the VaR is estimated as a quantile of the predictive density, by numerically inverting the predictive CDF. For instance, in the Bayesian framework, the VaR at the α risk level equals:

$$\mathrm{VaR}^{\alpha}_{T+1} \equiv \inf\left\{y_{T+1} \in \mathbb{R} \,|\, F(y_{T+1} \,|\, \mathcal{I}_T) = \alpha\right\}\,, \tag{8}$$

while the ES at the α risk level is given by:

$$\mathrm{ES}^{\alpha}_{T+1} \equiv \frac{1}{\alpha}\int_{-\infty}^{\mathrm{VaR}^{\alpha}_{T+1}} z f(z|\mathcal{I}_T)\mathrm{d}z\,. \tag{9}$$

In our empirical application, we consider the VaR and the ES at the 1% and 5% risk levels.

For evaluating the risk at an h–period horizon, we must rely on simulation techniques to obtain the conditional density and downside risk measures, as described, for instance, in Blasques et al. (2016). More specifically, given a model parameter, we generate 25,000 paths of daily log–returns over a horizon of h days.[8] The simulated distribution and the obtained α–quantile then serve as estimates of the density and downside risk forecasts of the h–day cumulative log–return.

3. Large–scale empirical study

We use 1,500 log–returns (in percent) for the estimation and run the backtest over 2,000 out–of–sample log-returns for a period ranging from October 10, 2008, to November 17, 2016 (the full dataset starts on December 26, 2002). Each model is estimated on a rolling window basis, and one–step ahead as well as multi–step cumulative log–returns density forecasts are obtained.[9] From the estimated density, we compute the VaR and the ES at the 1% and 5% risk levels.

3.1. Datasets

We test the performance of the various models on several universes of securities typically traded by fund managers:

- A set of 426 stocks, selected by taking the S&P 500 universe index as of November 2016, and omitting the stocks for which more than 5% of the daily returns are zero, and stocks for which there are less than 3,500 daily return observations.
- A set of eleven stock market indices: (1) S&P 500 (US; SPX), (2) FTSE 100 (UK; FTSE), (3) CAC 40 (France; FCHI), (4) DAX 30 (Germany; GDAXI), (5) Nikkei 225 (Japan; N225), (6) Hang Seng (China, HSI), (7) Dow Jones Industrial Average (US; DJI), (8) Euro Stoxx 50 (Europe; STOXX50), (9) KOSPI (South Korea; KS11), (10) S&P/TSX Composite (Canada; GSPTSE), and (11) Swiss Market Index (Switzerland; SSMI);

[8] With the frequentist estimation, we generate 25,000 paths with parameter $\widehat{\mathbf{\Psi}}$, while in the case of the Bayesian estimation, we generate 25 paths for each of the 1,000 value $\mathbf{\Psi}^{[m]}$ in the posterior sample. We use this number to get enough draws from the predictive distribution as we focus on the left tail. Geweke (1989) shows that the consistent estimation of the predictive distribution does not depend on the number of paths generated from the posterior. So with 25 paths, we indeed converge to the correct predictive distribution. We verified that increasing the number of simulations has no material impact on the results.

[9] Model parameters are updated every ten observations. We selected this frequency to speed up the computations. Similar results for a subset of stocks were obtained when updating the parameters every day. This is also in line with the observation of Ardia and Hoogerheide (2014), who show, in the context of GARCH models, that the performance of VaR forecasts is not significantly affected when moving from a daily updating frequency to a weekly or monthly updating frequency. Note that while parameters are updated every ten observations, the density and downsides risk measures are computed every day.

- A set of eight foreign exchange rates: USD against CAD, DKK, NOK, AUD, CHF, GBP, JPY, and EUR.[10]

Data are retrieved from Datastream. Each price series is expressed in local currency. We compute the daily percentage log–return series defined by $x_t \equiv 100 \times \log(P_t/P_{t-1})$, where P_t is the adjusted closing price (value) on day t. We then de–mean the returns x_t using an AR(1)–filter, and use those filtered returns, y_t, to estimate and evaluate the precision of the financial risk monitoring systems.

In Table 1, we report the summary statistics on the out–of–sample daily, five–day, and ten–day cumulative log–returns for the three asset classes. We report the standard deviation (Std), the skewness (Skew) and kurtosis (Kurt) coefficients evaluated over the full sample as well as the historical 1% and 5% VaR and ES levels. We note the higher volatility in all periods for the universe of stocks, followed by indices and exchange rates. All securities exhibit negative skewness, with larger values for indices and stocks, while exchange rates seem to behave more symmetrically. Interestingly, the negative skewness tends to be more pronounced for indices as the horizon grows. Finally, at the daily horizon, we observe a significant kurtosis for stocks. Fat tails are also present for indices and exchange rates, but less pronounced than for stocks. However, as the horizon grows, the kurtosis of all asset classes tends to diminish.

[Insert Table 1 about here.]

3.2. Forecasting performance tests

We compare the adequacy of the 32 models in terms of providing accurate forecasts of the left tail of the conditional distribution and the VaR and ES levels.

3.2.1. Accuracy of VaR predictions

For testing the accuracy of the VaR predictions, we use the so–called *hit* variable, which is a dummy variable indicating a loss that exceeds the VaR level:

$$I_t^\alpha \equiv \mathbb{I}\{y_t \leq \mathrm{VaR}_t^\alpha\}\,,$$

where VaR_t^α denotes the VaR prediction at risk level α for time t, and $\mathbb{I}\{\cdot\}$ is the indicator function equal to one if the condition holds, and zero otherwise. If the VaR is correctly specified,

[10] In the context of foreign exchange rates, left–tail forecasts aim at assessing the risk for a foreign investor investing in USD and therefore facing devaluation of USD.

then the hit variable has a mean value of α and is independently distributed over time. We test this for the $\alpha = 1\%$ and $\alpha = 5\%$ risk levels using the unconditional coverage (UC) test by Kupiec (1995), and the dynamic quantile (DQ) test by Engle and Manganelli (2004).

The UC test by Kupiec (1995) uses the likelihood ratio to test that the violations have a Binomial distribution with $\mathbb{E}[I_t^\alpha] = \alpha$. Denote by $x \equiv \sum_{t=1}^{T} I_t^\alpha$ the number of observed rejections on a total of T observations, then, under the null of correct coverage, we have that the test statistic:

$$\mathrm{UC}_\alpha \equiv -2\ln\left[(1-\alpha)^{T-x}\alpha^x\right] + 2\ln\left[\left(1-\frac{x}{T}\right)^{T-x}\left(\frac{x}{T}\right)^x\right],$$

is asymptotically chi–square distributed with one degree–of–freedom.

The DQ test by Engle and Manganelli (2004) is a test of the joint hypothesis that $\mathbb{E}[I_t^\alpha] = \alpha$ and that the hit variables are independently distributed. The implementation of the test involves the de–meaned process $\mathrm{Hit}_t^\alpha \equiv I_t^\alpha - \alpha$. Under correct model specification, unconditionally and conditionally, Hit_t^α has zero mean and is serially uncorrelated. The DQ test is then the traditional Wald test of the joint nullity of all coefficients in the following linear regression:

$$\mathrm{Hit}_t^\alpha = \delta_0 + \sum_{l=1}^{L}\delta_l \mathrm{Hit}_{t-l}^\alpha + \delta_{L+1}\mathrm{VaR}_{t-1}^\alpha + \epsilon_t\,.$$

If we denote the OLS parameter estimates as $\widehat{\boldsymbol{\delta}} \equiv (\widehat{\delta}_0, \ldots, \widehat{\delta}_{L+1})'$ and $\mathbf{Z}$ as the corresponding data matrix with, in column, the observations for the $L+2$ explanatory variables, then the DQ test statistic of the null hypothesis of correct unconditional and conditional coverage is:

$$\mathrm{DQ}_\alpha \equiv \frac{\widehat{\boldsymbol{\delta}}'\mathbf{Z}'\mathbf{Z}\widehat{\boldsymbol{\delta}}}{\alpha(1-\alpha)}\,.$$

As in Engle and Manganelli (2004), we choose $L = 4$ lags. Under the null hypothesis of correct unconditional and conditional coverage, we have that DQ_α is asymptotically chi–square distributed with $L+2$ degrees of freedom.[11]

[11] As in Bams et al. (2017), it is possible to add more explanatory variable such as lagged returns and lagged squared returns and jointly test the new coefficients. In our case, results obtained by adding lagged returns or lagged squared returns are qualitatively similar to the simpler specification.

3.2.2. Accuracy of the left–tail distribution

Risk managers care not only about the accuracy of the VaR forecasts but also about the accuracy of the complete left–tail region of the log–return distribution. This broader view of all losses is central in modern risk management, and, consistent with the regulatory shift to using Expected Shortfall as the risk measure for determining capital requirements starting in 2018 (Basel Committee on Banking Supervision, 2013). We evaluate the effectiveness of MSGARCH models to yield accurate predictions of the left–tail distribution in three ways.

A first approach is to compute the weighted average difference of the observed returns with respect to the VaR value, and give higher weight to losses that violate the VaR level. This corresponds to the quantile loss assessment of González-Rivera et al. (2004) and McAleer and Da Veiga (2008). Formally, given a VaR prediction at risk level α for time t, the associated quantile loss (QL) is defined as:

$$\mathrm{QL}_t^{\alpha} \equiv (\alpha - I_t^{\alpha})(y_t - \mathrm{VaR}_t^{\alpha})\,.$$

The choice of this loss function for VaR assessment is appropriate since quantiles are elicited by it; that is, when the conditional distribution is static over the sample, the VaR_t^{α} can be estimated by minimizing the average quantile loss function. Elicitability is useful for model selection, estimation, forecast comparison, and forecast ranking.

Unfortunately, there is no loss function available for which the ES risk measure is elicitable; see, for instance, Bellini and Bignozzi (2015) and Ziegel (2016). However, it has been recently shown by Fissler and Ziegel (2016) (FZ) that, in case of a constant conditional distribution, the couple (VaR, ES) is jointly elicitable, as the values of v_t and e_t that minimize the sample average of the following loss function:

$$FZ(y_t, v_t, e_t, \alpha, G_1, G_2) \equiv (I_t^{\alpha} - \alpha)\left(G_1(v_t) - G_1(y_t) + \frac{1}{\alpha}G_2(e_t)v_t\right) - G_2(e_t)\left(\frac{1}{\alpha}I_t^{\alpha}y_t - e_t\right) - \mathcal{G}_2(e_t)\,,$$

where G_1 is weakly increasing, G_2 is strictly positive and strictly increasing, and $\mathcal{G}_2' = G_2$. In a similar setup as ours, Patton et al. (2017) assume the values of VaR and ES to be strictly negative and recommend setting $G_1(x) = 0$ and $G_2(x) = -1/x$. For a VaR and a ES prediction at risk level α for time t, the associated joint loss function (FZL) is then given by:

$$\mathrm{FZL}_t^{\alpha} \equiv \frac{1}{\alpha\,\mathrm{ES}_t^{\alpha}}\, I_t^{\alpha}\,(y_t - \mathrm{VaR}_t^{\alpha}) + \frac{\mathrm{VaR}_t^{\alpha}}{\mathrm{ES}_t^{\alpha}} + \log(-\mathrm{ES}_t^{\alpha}) - 1\,, \tag{10}$$

for $\mathrm{ES}_t^\alpha \leq \mathrm{VaR}_t^\alpha < 0$. Hence, in order to gauge the precision of both the VaR and ES downside risk estimates, we use the FZL function as our second evaluation criterion.

A third approach that we consider is to compare the empirical distribution with the predicted conditional distribution through the weighed Continuous Ranked Probability Score (wCRPS), introduced by Gneiting and Ranjan (2011) as a generalization of the CRPS scoring rule (Matheson and Winkler, 1976). Following the notation introduced in Section 2, the wCRPS for a forecast at time t is defined as:

$$\mathrm{wCRPS}_t \equiv \int_{\mathbb{R}} \omega(z) \left(F(z \,|\, \mathcal{I}_{t-1}) - \mathbb{I}\{y_t \leq z\}\right)^2 \mathrm{d}z \,,$$

where F is the predictive CDF and $\omega \colon \mathbb{R} \to \mathbb{R}^+$ is a continuous weight function, which emphasizes regions of interest of the predictive distribution, such as the tails or the center. Since our focus is on predicting losses, we follow Gneiting and Ranjan (2011) and use the decreasing weight function $\omega(z) \equiv 1 - \Phi(z)$, where Φ is the CDF of a standard Gaussian distribution. This way, discrepancies in the left tail of the return distribution are weighed more than those in the right tail.[12]

For the QL, FZL and wCRPS approaches, we test the statistical significance of the differences in the forecasting performance of two competing models, say models i and j. We do this by first computing, for each out–of–sample date t, the average performance statistics across all securities in the same asset class. Denote this difference as $\Delta_t^{i-j} \equiv L_t^i - L_t^j$, where L_t^i is the average value of the performance measure (QL, FZL or wCRPS) of all assets within the same asset class. We then test $H_0 : \mathbb{E}[\Delta_t^{i-j}] = 0$ using the standard Diebold and Mariano (1995) (DM) test statistic, implemented with the heteroscedasticity and autocorrelation robust (HAC) standard error estimators of Andrews (1991) and Andrews and Monahan (1992). If the null hypothesis is rejected, the sign of the test statistics indicates which model is, on average, preferred for a particular loss measure.

[12] We follow the implementation of Gneiting and Ranjan (2011) and compute wCRPS with the following approximation:

$$\mathrm{wCRPS}_t \approx \frac{z_u - z_l}{M - 1} \sum_{m=1}^{M} w(z_m) \left(F(z_m \,|\, \mathcal{I}_{t-1}) - \mathbb{I}\{y_t \leq z_m\}\right)^2 \,,$$

where $z_m \equiv z_l + m \times (z_u - z_l)/M$ and z_u and z_l are the upper and lower values, which defines the range of integration. The accuracy of the approximation can be increased to any desired level by M. Setting $z_l = -100$, $z_u = 100$ and $M = 1{,}000$ provides an accurate approximation when working with returns in percentage points. We also tested the triangular integration approach and results were numerically equivalent. Alternative weights specifications, focusing on the right tail, center, of full distribution, lead to similar conclusions at the one–day forecasting horizon. The results are available from the authors upon request.

3.3. Results

We now summarize the results regarding our main research question: *Does the additional complexity of Markov–switching and the use of Bayesian estimation methods lead to more accurate out–of–sample downside risk predictions?* We first present our results regarding the accuracy of the VaR predictions and then use the QL, FZL and wCRPS approaches to evaluate the gains in terms of left–tail predictions.

3.3.1. Effect of model and estimator choice on the accuracy of VaR predictions

We first use the UC test of Kupiec (1995) and the DQ test of Engle and Manganelli (2004) to evaluate the accuracy of each of the 32 methods considered in terms of predicting the VaR at the 5% and 1% level for the daily returns on the 426 stocks, 11 stock indices and 8 exchange rates. For each asset, we obtain the p–value corresponding to the UC and DQ test computed using 2,000 out–of–sample observations. In Table 2, we aggregate the results per asset class by presenting the percentage of assets for which the null hypothesis of correct unconditional and conditional coverage is rejected at the 5% level, by the UC and DQ test, respectively.[13]

[Insert Table 2 about here.]

Consider in Panels A and B of Table 2 the results for the UC test. At both VaR risk levels, we find that the validity of the VaR predictions based on the GARCH and GJR skewed Student–t risk model is never rejected, whatever the use of SR or MS models, or frequentist or Bayesian estimation methods. The result changes drastically when we consider the more powerful DQ test of correct conditional coverage in Panels C and D. Here, we find clear evidence that the use of MS GJR models leads to a lower percentage of rejections of the validity of the VaR prediction for all asset classes. At the 1% risk level, these differences are most often significant.

Overall, the one–day ahead backtest results indicate outperformance of MS over SR models, especially for VaR prediction on equities. Moreover, a GJR specification leads to a substantial reduction in the rejection frequencies. Both for MS and SR specifications, a fat–tailed conditional distribution is of primary importance and delivers excellent results at both risk levels.

Finally, for this analysis, the frequency of rejections are similar between the Bayesian and frequentist estimation methods. More precisely, a t–test for equal average rejections indicates

[13] In the case of stocks, as the universe is large and therefore prone to false positives, the p–values are corrected for Type I error using the false discovery rate (FDR) approach of Benjamini and Hochberg (1995). The FDR correction for a confidence level q proceeds as follows. For a set of m ordered p–values $p_1 \leq p_2 \leq \ldots \leq p_m$ and corresponding null hypotheses $H_1, H_2, \ldots, H_m$, define v as the largest value of i for which $p_i \leq \frac{i}{m}q$, and then reject all hypotheses H_i for $i = 1, \ldots, v$.

that differences are insignificant. We thus conclude that, based on the analysis of VaR forecast accuracy, it is hard to discriminate between the estimation methods.

3.3.2. Effect of model choice on accuracy of left–tail predictions

A further question is how model simplification affects the accuracy of the left–tail return prediction. In Table 3, we report the standardized difference between the average QL, FZL and wCRPS values of the assets belonging to the same asset class, when we switch from a MS specification to a SR specification. The standardization corresponds to the Diebold and Mariano (1995) (DM) test statistic. Negative values indicate out–of–sample evidence of a deterioration in the prediction accuracy when using the SR specification instead of the MS specification. When the standardized value exceeds 2.57 (*i.e.*, the critical value computed using a 1% significance level for a bilateral test based on the asymptotic Normal distribution) in absolute value, the statistical significance is highlighted with a gray shading.[14] We report results obtained with the Bayesian framework only, as the performance obtained with the Bayesian estimation is either similar or better for both MS and SR models compared with the frequentist estimation.[15]

[Insert Table 3 about here.]

One–step ahead results for wCRPS favor MS models with negative values observed for almost all asset classes and model specifications. QL, FZL and wCRPS results are consistent with the backtest results: They confirm the superior performance of the MS specification for the universe of stocks, while outperformance is less clear for indices and exchange rates. Indeed, for indices, MS is required only when a non fat–tailed conditional distribution is assumed, while for exchange rates, MS is generally not required. Note that, for all assets, the improvements tend to be more pronounced when the Markov–switching mechanism is applied to simple specifications such as the GARCH–Normal model.

For stocks, the MS specification significantly outperforms in terms of the FZL and wCRPS measures at the five–day horizon. For the wCRPS measure at the ten–day horizon, and for the QL measure at the five– and ten–day horizons, results are mostly insignificant, except for

[14] We take the standard critical value in Diebold and Mariano (1995) as our Markov–switching specifications do not nest the alternative single–regime model due to parameter constraints imposing that the volatility dynamics are numerically different in each regime, and that each regime has a non–zero probability. The approach by Clark and McCracken (2001) should be used when comparing nested models.

[15] In Section 3.3.3, we find that the gains of Bayesian estimation compared to frequentist estimation are larger in the case of SR models. Our discussion regarding the gains of MS versus SR models based on the Bayesian estimation results is thus conservative in the sense that it gives an advantage to the SR specifications.

the FZL 5% measure, which favors MS models when a non fat–tailed conditional distribution is assumed. MS and SR models perform similarly for the five– and ten–day returns on stock indices. Finally, for exchange rate returns, SR models outperform MS models at the five– and ten–day horizons according to the QL 1% measure, while the differences in QL 5%, FZL, and wCRPS are insignificant.

It is informative to examine if these gains in forecasting precision are stable across the out–of–sample window. To determine this, we display in Figure 1 the cumulative average loss differentials over the whole out–of–sample period for the best performing specification, the GJR skewed Student–t model. Interestingly, we find that MS systematically outperforms SR according to the criteria that are most sensitive to the extreme left tail of the return distribution, namely the FZL (for $\alpha = 1\%$ and $\alpha = 5\%$) and QL (for $\alpha = 1\%$). We also notice that in these cases, the gains of MS over SR increase during the last phase of the turbulent period 2008–2012. With regards to wCRPS and QL at $\alpha = 5\%$, we find that MS starts outperforming SR after the end of the turbulent period 2008–2012. We conjecture that this improvement in performance can be explained by the lack of flexibility of the single–regime GARCH specification. As also evident from the first panel of Figure 1, the market volatility has changed both its unconditional level and its dependence structure between the two periods 2008–2012 and 2012–2015. Since the estimation window is of 1,500 observations (approximately 7 years), observations in the period 2008–2012 affect SR predictions for the whole 2012–2015 forecasting period. Differently, MSGARCH allows the volatility process to adapt more rapidly to changes in regimes, resulting in better risk predictions. This is the case for the first half of the window, ranging from December 2008 to November 2012, and which encompasses the Great Financial Crisis, but also for the second half of the window, ranging from December 2012 to November 2016, which is a more calm market period.

[Insert Figure 1 about here.]

We now consider in Table 4 a complete comparison of the wCRPS performance of all MS models (in row) versus all SR models (in column). The elements in the diagonal correspond to the wCRPS values reported in Table 3. They are informative about the change in wCRPS when switching from a MS model to a SR model, keeping the same specification for the conditional variance and distribution. The analysis of the extra–diagonal elements is informative about the changes in wCRPS when switching from a MS model to a SR model, and changing the

specification of the volatility model or the density function. In this table, an outperforming MS risk model is a model for which all standardized gains when changing the specification are negative. For almost all comparisons, this is the case for the MS GJR model with skewed Student–t innovations. The only exception is for modeling the returns of stock market indices, where it performs similarly as its SR counterpart.

[Insert Table 4 about here.]

Compared to SR models, MS specifications offer the flexibility of a different volatility response to extreme (positive or negative) observations than to moderately large observations. This feature is desirable in case the discretely observed returns are generated by an underlying continuous–time process with jumps. Those jumps usually correspond to one–off events (as in Boudt et al. (2013)) and have a less persistent effect on future volatility (see Andersen et al., 2007). As explained in Laurent et al. (2016), in a SR framework, this effect can be captured through the use of Generalized Autoregressive Score (GAS) models (also referred to as Dynamic Conditional Score (DCS) models) with fat tails, as introduced by Creal et al. (2013) and Harvey (2013). Clearly a SR GAS model is computationally simpler than a MS model. It is therefore relevant to benchmark the MSGARCH model with a GAS model whose specification fits with the assumed density function, which, in case of the skewed Student–t density function, is the Beta–Skew–t–EGARCH(1,1) model introduced by Harvey and Sucarrat (2014). In case of a fat–tailed conditional distribution, this model yields volatility forecasts that are resistant to outliers in the return series (due, for instance, to one–off events causing price jumps) and may therefore deliver better downside risk predictions than GARCH; see, for instance, Bernardi and Catania (2016) and Ardia et al. (2018).

In Table 5, we report the standardized gain in average QL, FZL (at 1% and 5% risk levels) and wCRPS performance when switching from the most flexible MSGARCH model, that is, the MS GJR sk$\mathcal{S}$, to the Beta–Skew–t–EGARCH(1,1) model. For downside risk prediction related to the returns on stocks and stock market indices, results significantly favor the MS specification when focusing on the 1% largest losses, as can be seen in the QL 1% and FZL 1% columns. For exchange rates, results favor the GAS model but are not significant. Overall, we thus find that the MS specification can offer added value in downside risk prediction of equity investments, as compared to the use of GAS models.

[Insert Table 5 about here.]

3.3.3. Effect of estimator choice on accuracy of left–tail predictions

In Table 6, we report the results for the Bayesian versus frequentist estimation methods in the case of one–step ahead QL, FZL and wCRPS measures. Panel A (Panel B) shows the results for MS (SR) models, where a negative (positive) value indicates outperformance (underperformance) of Bayesian against frequentist estimation. In light gray, we emphasize cases of significant outperformance of the Bayesian estimation over the frequentist approach. For stocks, the QL 1% and 5% comparisons indicate that Bayesian is preferred over ML, and it is significant in the majority of the specifications. The same observation can be made when using the FZL and wCRPS evaluation criteria. For stock indices and exchange rates, QL, FZL and wCRPS results are in favor of the Bayesian estimation for both MS and SR models but results are less significant than for stocks. Overall, we recommend to account for parameter uncertainty especially for stock data, and when the interest is on the left tail of the log–returns distribution. The performance gain is especially large for SR models.

[Insert Table 6 about here.]

3.3.4. Constrained Markov–switching specifications

So far, our empirical results have highlighted the need for a MS mechanism in GARCH–type models in the case of stocks. We now refine the analysis by examining whether the same gains are achieved when constraining that the conditional distribution of the MS specifications has the same shape parameter across the regimes. Hence, we apply the MS mechanism only to the conditional variance. The objective is to determine whether, in the context of MS models, the switches in the variance dynamics are the dominant contributor to the gains in risk forecasting accuracy.

In Table 7, we report the performance measures obtained with the constrained MS models for the various horizons, when models are estimated with the Bayesian approach.[16] Results are in line with the non–constrained case of Table 3, but less significant. Hence, accounting for structural breaks in only the variance dynamics improves the risk forecasts at the daily, weekly and ten–day horizons. If we let the shape parameters depend upon the regime, we further improve the performance.

[Insert Table 7 about here.]

[16] Forecasting results obtained via frequentist estimation are qualitatively similar. They are available from the authors upon request.

4. Conclusion

In this paper, we investigate if MSGARCH models provide risk managers with useful tools for improving the risk forecasts of securities typically hold by fund managers. Moreover, we investigate if integrating the model's parameter uncertainty within the forecasts, via the Bayesian approach, improves predictions. Our results and practical advice can be summarized as follows.

First, risk managers should extend their GARCH–type models with a Markov–switching specification in case of investment in equities. Indeed, we find that Markov–switching GARCH models deliver better Value–at–Risk, Expected Shortfall, and left–tail distribution forecasts than their single–regime counterpart. This is especially true for stock return data. Moreover, improvements are more pronounced when the Markov–switching mechanism is applied to simple specifications such as the GARCH–Normal model.

Second, accounting for parameter uncertainty helps for left–tail predictions independently of the inclusion of the Markov–switching mechanism. Moreover, larger improvements are observed when parameter uncertainty is included in single–regime models.

Overall, we recommend risk managers to rely on more flexible models and to perform inference accounting for parameter uncertainty. To help them implementing these in practice, we have released the open–source R package MSGARCH; see Ardia et al. (2017a,b).

Our research could be extended in several ways. First, our study considered single–regime versus two–state Markov–switching specifications. Hence, it would be of interest to see if a third regime leads to superior performance, and if the optimal number of regimes (according to penalized likelihood information criteria) changes over time and is different across data sets. Second, additional universes could be considered, such as emerging markets and commodities. Third, one could extend the set of models and compare the performance of MSGARCH with realized volatility models such as the HEAVY model of Shephard and Sheppard (2010). Fourth, as suggested by a referee, it would be interesting to shed light on the parameter configurations for which the MSGARCH predictions can be expected to yield the higher improvement in risk forecast precision. An exploratory analysis has shown that a high persistence of at least one state seems needed to have a substantial difference in precision between MSGARCH and single–regime GARCH downside risk forecasts. A definite answer to this question is beyond the scope of this paper. Finally, our analysis only considered financial risk monitoring systems for individual financial assets. The new standard for capital requirements for market risk (Basel Committee on Banking Supervision, 2016) calls for backtesting at the individual desk level and the aggregate

level. For this reason, it would be interesting to consider also the impact of choices in modeling dependence. Including these extensions in our current research setup increases further the (already large) number of models included in the comparison. We leave them as a topic for future work.

References

Andersen, T.G., Bollerslev, T., Diebold, F.X., 2007. Roughing it up: Including jump components in the measurement, modeling, and forecasting of return volatility. Review of Economics and Statistics 89, 701–720. doi:10.1162/rest.89.4.701.

Andrews, D., 1991. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. Econometrica 59, 817–858. doi:10.2307/2938229.

Andrews, D., Monahan, J., 1992. An improved heteroskedasticity and autocorrelation consistent covariance matrix estimation. Econometrica 60, 953–966. doi:10.2307/2951574.

Ardia, D., 2008. Financial Risk Management with Bayesian Estimation of GARCH Models: Theory and Applications. Springer, Heidelberg. doi:10.1007/978-3-540-78657-3.

Ardia, D., Bluteau, K., Boudt, K., Catania, L., Peterson, B., Trottier, D.A., 2017a. MSGARCH: Markov-switching GARCH models in R. URL: https://cran.r-project.org/package=MSGARCH. forthcoming in Journal of Statistical Software.

Ardia, D., Bluteau, K., Boudt, K., Catania, L., Trottier, D.A., 2017b. Markov-switching GARCH models in R: The MSGARCH package. URL: https://ssrn.com/abstract=2845809. working paper.

Ardia, D., Boudt, K., Catania, L., 2018. Downside risk evaluation with the R package GAS. URL: https://ssrn.com/abstract=2871444. working paper.

Ardia, D., Hoogerheide, L.F., 2014. GARCH models for daily stock returns: Impact of estimation frequency on Value-at-Risk and Expected Shortfall forecasts. Economics Letters 123, 187–190. doi:10.1016/j.econlet.2014.02.008.

Ardia, D., Hoogerheide, L.F., van Dijk, H.K., 2009. Adaptive mixture of Student-t distributions as a flexible candidate distribution for efficient simulation: The R package AdMit. Journal of Statistical Software 29, 1–32. doi:10.18637/jss.v029.i03.

Ardia, D., Kolly, J., Trottier, D.A., 2017c. The impact of parameter and model uncertainty on market risk predictions from GARCH-type models. Journal of Forecasting 36, 808–823. doi:10.1002/for.2472.

Augustyniak, M., 2014. Maximum likelihood estimation of the Markov-switching GARCH model. Computational Statistics & Data Analysis 76, 61–75. doi:10.1016/j.csda.2013.01.026.

Bams, D., Blanchard, G., Lehnert, T., 2017. Volatility measures and Value-at-Risk. International Journal of Forecasting 33, 848–863. doi:10.1016/j.ijforecast.2017.04.004.

Basel Committee on Banking Supervision, 2013. Fundamental review of the trading book: A revised market risk framework. Technical Report 265. Bank of International Settlements.

Basel Committee on Banking Supervision, 2016. Minimum capital requirements for market risk. techreport 352. Bank of International Settlements.

Bauwens, L., De Backer, B., Dufays, A., 2014a. A Bayesian method of change-point estimation with recurrent regimes: Application to GARCH models. Journal of Empirical Finance 29, 207–229. doi:10.1016/j.jempfin.2014.06.008.

Bauwens, L., Dufays, A., Rombouts, J.V.K., 2014b. Marginal likelihood for Markov-switching and change-point GARCH models. Journal of Econometrics 178, 508–522. doi:10.1016/j.jeconom.2013.08.017.

Bauwens, L., Laurent, S., 2005. A new class of multivariate skew densities, with application to generalized

autoregressive conditional heteroscedasticity models. Journal of Business & Economic Statistics 23, 346–354. doi:10.1198/073500104000000523.

Bauwens, L., Preminger, A., Rombouts, J.V.K., 2010. Theory and inference for a Markov switching GARCH model. Econometrics Journal 13, 218–244. doi:10.1111/j.1368-423X.2009.00307.x.

Bellini, F., Bignozzi, V., 2015. On elicitable risk measures. Quantitative Finance 15, 725–733. doi:10.1080/14697688.2014.946955.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B Vol. 57, 289–300. URL: http://www.jstor.org/stable/2346101.

Bernardi, M., Catania, L., 2016. Comparison of value-at-risk models using the MCS approach. Computational Statistics 31, 579–608. doi:10.1007/s00180-016-0646-6.

Black, F., 1976. Studies of stock price volatility changes, in: Proceedings of the 1976 Meetings of the American Statistical Association, pp. 177–181.

Blasques, F., Koopman, S.J., Łasak, K., Lucas, A., 2016. In-sample confidence bands and out-of-sample forecast bands for time-varying parameters in observation-driven models. International Journal of Forecasting 32, 875–887. doi:10.1016/j.ijforecast.2015.11.018.

Board of Governors of the Federal Reserve Systems, 2012. 99th Annual Report. Technical Report. Board of Governors of the Federal Reserve Systems.

Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics 31, 307–327. doi:10.1016/0304-4076(86)90063-1.

Boudt, K., Daníelsson, J., Laurent, S., 2013. Robust forecasting of dynamic conditional correlation GARCH models. International Journal of Forecasting 29, 244–257. doi:10.1016/j.ijforecast.2012.06.003.

Cai, J., 1994. A Markov model of switching-regime ARCH. Journal of Business & Economic Statistics 12, 309–316. doi:10.2307/1392087.

Clark, T.E., McCracken, M.W., 2001. Tests of equal forecast accuracy and encompassing for nested models. Journal of Econometrics 105, 85–110. doi:10.1016/S0304-4076(01)00071-9.

Creal, D., Koopman, S.J., Lucas, A., 2013. Generalized autoregressive score models with applications. Journal of Applied Econometrics 28, 777–795. doi:10.1002/jae.1279.

Danielsson, J., 2011. Risk and crises. VoxEU.org URL: http://voxeu.org/article/risk-and-crises-how-models-failed-and-are-failing.

Diebold, F.X., Mariano, R.S., 1995. Comparing predictive accuracy. Journal of Business & Economic Statistics 13, 253–263. doi:10.1080/07350015.1995.10524599.

Dueker, M.J., 1997. Markov switching in GARCH processes and mean-reverting stock-market volatility. Journal of Business & Economic Statistics 15, 26–34. doi:10.2307/1392070.

Engle, R.F., 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica 50, 987–1008. doi:10.2307/1912773.

Engle, R.F., 2004. Risk and volatility: Econometric models and financial practice. American Economic Review 94, 405–420. doi:10.1257/0002828041464597.

Engle, R.F., Manganelli, S., 2004. CAViaR: Conditional autoregressive Value at Risk by regression quantiles.

Journal of Business & Economic Statistics 22, 367–381. doi:10.1198/073500104000000370.

Fernández, C., Steel, M.F.J., 1998. On Bayesian modeling of fat tails and skewness. Journal of the American Statistical Association 93, 359–371. doi:10.1080/01621459.1998.10474117.

Fissler, T., Ziegel, J.F., 2016. Higher order elicitability and Osband's principle. The Annals of Statistics 44, 1680–1707. doi:10.1214/16-AOS1439.

Geweke, J., 1989. Exact predictive densities for linear models with ARCH disturbances. Journal of Econometrics 40, 63–86. doi:10.1016/0304-4076(89)90030-4.

Geweke, J., Amisano, G., 2010. Comparing and evaluating Bayesian predictive distributions of asset returns. International Journal of Forecasting 26, 216–230. doi:10.1016/j.ijforecast.2009.10.007.

Glosten, L.R., Jagannathan, R., Runkle, D.E., 1993. On the relation between the expected value and the volatility of the nominal excess return on stocks. Journal of Finance 48, 1779–1801. doi:10.1111/j.1540-6261.1993.tb05128.x.

Gneiting, T., Ranjan, R., 2011. Comparing density forecasts using threshold –and quantile– weighted scoring rules. Journal of Business & Economic Statistics 29, 411–422. doi:10.1198/jbes.2010.08110.

González-Rivera, G., Lee, T.H., Mishra, S., 2004. Forecasting volatility: A reality check based on option pricing, utility function, Value-at-Risk, and predictive likelihood. International Journal of Forecasting 20, 629–645. doi:10.1016/j.ijforecast.2003.10.003.

Gray, S.F., 1996. Modeling the conditional distribution of interest rates as a regime-switching process. Journal of Financial Economics 42, 27–62. doi:10.1016/0304-405x(96)00875-6.

Haas, M., Mittnik, S., Paolella, M.S., 2004. A new approach to Markov-switching GARCH models. Journal of Financial Econometrics 2, 493–530. doi:10.1093/jjfinec/nbh020.

Hamilton, J.D., 1989. A new approach to the economic analysis of nonstationary time series and the business cycle. Econometrica 57, 357–384. doi:10.2307/1912559.

Hamilton, J.D., 1994. Time Series Analysis. First ed., Princeton University Press, Princeton, USA.

Hamilton, J.D., Susmel, R., 1994. Autoregressive conditional heteroskedasticity and changes in regime. Journal of Econometrics 64, 307–333. doi:10.1016/0304-4076(94)90067-1.

Hansen, P.R., Huang, Z., Shek, H.H., 2011. Realized GARCH: A joint model for returns and realized measures of volatility. Journal of Applied Econometrics 27, 877–906. doi:10.1002/jae.1234.

Hansen, P.R., Lunde, A., 2005. A forecast comparison of volatility models: Does anything beat a GARCH(1,1)? Journal of Applied Econometrics 20, 873–889. doi:10.1002/jae.800.

Harvey, A., Sucarrat, G., 2014. Egarch models with fat tails, skewness and leverage. Computational Statistics & Data Analysis 76, 320–338. doi:10.1016/j.csda.2013.09.022.

Harvey, A.C., 2013. Dynamic models for volatility and heavy tails: with applications to financial and economic time series. volume 52. Cambridge University Press.

Herwartz, H., 2017. Stock return prediction under GARCH – An empirical assessment. International Journal of Forecasting 33, 569–580. doi:10.1016/j.ijforecast.2017.01.002.

Jacquier, E., Polson, N.G., Rossi, P.E., 1994. Bayesian analysis of stochastic volatility models. Journal of Business & Economic Statistics 12, 371–389. doi:10.1080/07350015.1994.10524553.

Jorion, P., 2006. Value at Risk – The New Benchmark for Managing Financial Risk. Third ed., McGraw–Hill.

Klaassen, F., 2002. Improving GARCH volatility forecasts with regime-switching GARCH, in: Advances in Markov-Switching Models. Springer-Verlag, pp. 223–254. doi:10.1007/978-3-642-51182-0_10.

Kupiec, P.H., 1995. Techniques for verifying the accuracy of risk measurement models. Journal of Derivatives 3. doi:10.3905/jod.1995.407942.

Lamoureux, C.G., Lastrapes, W.D., 1990. Persistence in variance, structural change, and the GARCH model. Journal of Business & Economic Statistics 8, 225–234. doi:10.2307/1391985.

Laurent, S., Lecourt, C., Palm, F.C., 2016. Testing for jumps in conditionally gaussian ARMA-GARCH models, a robust approach. Computational Statistics & Data Analysis 100, 383–400. doi:10.1016/j.csda.2014.05.015.

Marcucci, J., 2005. Forecasting stock market volatility with regime-switching GARCH models. Studies in Nonlinear Dynamics & Econometrics 9. doi:10.2202/1558-3708.1145.

Matheson, J.E., Winkler, R.L., 1976. Scoring rules for continuous probability distributions. Management Science 22, 1087–1096. doi:10.1287/mnsc.22.10.1087.

McAleer, M., Da Veiga, B., 2008. Single-index and portfolio models for forecasting Value-at-Risk thresholds. Journal of Forecasting 27, 217–235. doi:10.1002/for.1054.

McNeil, A.J., Frey, R., Embrechts, P., 2015. Quantitative Risk Management: Concepts, Techniques and Tools. Second ed., Princeton University Press.

Nelson, D.B., 1991. Conditional heteroskedasticity in asset returns: A new approach. Econometrica 59, 347–370. doi:10.2307/2938260.

Opschoor, A., van Dijk, D., van der Wel, M., 2017. Combining density forecasts using focused scoring rules. Journal of Applied Econometrics 32, 1298–1313. doi:10.1002/jae.2575. in press.

Patton, A.J., Ziegel, J.F., Chen, R., 2017. Dynamic semiparametric models for expected shortfall. URL: https://ssrn.com/abstract=3000465. working paper.

Shephard, N., Sheppard, K., 2010. Realising the future: Forecasting with high-frequency-based volatility (HEAVY) models. Journal of Applied Econometrics 25, 197–231. doi:10.1002/jae.1158.

Taylor, S.J., 1994. Modeling stochastic volatility: A review and comparative study. Mathematical Finance 4, 183–204. doi:10.1111/j.1467-9965.1994.tb00057.x.

Trottier, D.A., Ardia, D., 2016. Moments of standardized Fernández-Steel skewed distributions: Applications to the estimation of GARCH-type models. Finance Research Letters 18, 311–316. doi:10.1016/j.frl.2016.05.006.

Vihola, M., 2012. Robust adaptive Metropolis algorithm with coerced acceptance rate. Statistics and Computing 22, 997–1008. doi:10.1007/s11222-011-9269-5.

Ziegel, J.F., 2016. Coherence and elicitability. Mathematical Finance 26, 901–918. doi:10.1111/mafi.12080.

Table 1: Summary statistics of the return data
The table presents the summary statistics of the (de–meaned) h–day cumulative log–returns for securities in the three asset classes used in our study. We report the standard deviation (Std), the skewness (Skew), the kurtosis (Kurt), and the 1% and 5% historical VaR and ES, on an unconditional basis for the 2,000 out–of–sample observations. For each statistic, we compute the 25th, 50th and 75th percentiles over the whole universe of assets.

h	Percentile	Std	Skew	Kurt	1% VaR	5% VaR	1% ES	5% ES
Panel A: Stocks (426 series)								
	25th	1.48	−0.39	6.89	−6.55	−3.44	−9.30	−5.53
1	50th	1.89	−0.13	9.24	−5.23	−2.85	−7.31	−4.50
	75th	2.33	0.12	14.10	−4.10	−2.25	−5.68	−3.50
	25th	3.29	−0.42	4.93	−14.60	−7.94	−19.14	−12.11
5	50th	4.21	−0.20	5.87	−11.59	−6.55	−14.84	−9.82
	75th	5.19	0.01	7.53	−9.15	−5.17	−12.00	−7.71
	25th	4.54	−0.49	4.47	−19.99	−10.92	−25.42	−16.54
10	50th	5.76	−0.27	5.30	−15.74	−9.02	−20.28	−13.19
	75th	6.98	−0.05	6.92	−12.43	−7.16	−16.08	−10.46
Panel B: Stock market indices (11 series)								
	25th	1.07	−0.40	6.07	−3.70	−2.37	−4.84	−3.30
1	50th	1.15	−0.23	7.29	−3.39	−1.85	−4.31	−2.78
	75th	1.39	−0.17	10.29	−3.05	−1.77	−4.01	−2.58
	25th	2.42	−0.55	5.04	−8.38	−5.09	−10.65	−7.30
5	50th	2.54	−0.47	6.18	−7.60	−4.22	−9.85	−6.17
	75th	3.09	−0.29	8.22	−6.91	−3.86	−9.22	−5.97
	25th	3.29	−0.79	5.47	−12.32	−7.13	−15.96	−10.22
10	50th	3.43	−0.62	6.31	−10.83	−5.70	−13.92	−8.70
	75th	4.19	−0.55	7.04	−9.99	−5.19	−12.90	−8.22
Panel C: Exchange rates (8 series)								
	25th	0.61	−0.53	4.36	−1.73	−1.07	−2.42	−1.60
1	50th	0.62	−0.08	4.51	−1.62	−1.01	−2.10	−1.42
	75th	0.77	0.05	11.60	−1.56	−0.95	−1.92	−1.34
	25th	1.32	−0.36	3.65	−3.72	−2.39	−5.02	−3.36
5	50th	1.39	−0.05	4.05	−3.48	−2.26	−4.33	−3.03
	75th	1.66	0.08	5.91	−3.07	−2.06	−3.82	−2.77
	25th	1.85	−0.31	3.36	−5.00	−3.43	−6.99	−4.55
10	50th	1.93	−0.10	3.52	−4.78	−3.04	−5.72	−4.06
	75th	2.29	0.13	5.12	−4.64	−2.93	−5.41	−3.94

Table 2: Percentage of assets for which the validity of the VaR predictions is rejected
The table presents the percentage of assets for which the unconditional coverage test (UC, Panels A and B) by Kupiec (1995) and the Dynamic Quantile test (DQ, Panels C and D) by Engle and Manganelli (2004) reject the null hypothesis of correct unconditional coverage (UC, DQ) and independence of violations (DQ) for the one–step ahead 1%–VaR (Panels A and C) and 5%–VaR (Panels B and D) at the 5% significance level. The VaR forecasts are obtained for Markov–switching (MS) and single–regime (SR) models for the various universes (426 stocks, 11 indices, and 8 exchange rates) and estimated via Bayesian or frequentist techniques. We highlight in gray the best performing method for the cases in which, for a given asset class and model specification, the percentages of rejections between MS and SR models are significantly different at the 5% level. In the case of stocks, rejections frequencies are corrected for Type I error using the FDR approach of Benjamini and Hochberg (1995).

	Stocks				Stock market indices				Exchange rates			
	Bayesian		Frequentist		Bayesian		Frequentist		Bayesian		Frequentist	
Model	MS	SR	MS	SR	MS	SR	MS	SR	MS	SR	MS	SR
Panel A: UC 1%–VaR												
GARCH $\mathcal{N}$	0.00	26.76	0.23	29.34	72.73	90.91	72.73	90.91	25.00	25.00	25.00	25.00
GARCH sk$\mathcal{N}$	0.00	8.92	0.23	9.62	9.09	63.64	0.00	63.64	0.00	12.50	0.00	12.50
GARCH $\mathcal{S}$	0.00	0.00	0.00	0.00	54.55	45.45	27.27	27.27	25.00	25.00	25.00	12.50
GARCH sk$\mathcal{S}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GJR $\mathcal{N}$	0.00	16.43	0.00	19.48	54.55	90.91	63.64	90.91	25.00	25.00	25.00	37.50
GJR sk$\mathcal{N}$	0.00	3.52	0.00	5.16	0.00	54.55	0.00	45.45	0.00	12.50	0.00	25.00
GJR $\mathcal{S}$	0.00	0.00	0.00	0.00	18.18	36.36	18.18	36.36	12.50	12.50	12.50	12.50
GJR sk$\mathcal{S}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Panel B: UC 5%–VaR												
GARCH $\mathcal{N}$	0.70	39.20	0.70	38.73	36.36	36.36	27.27	36.36	25.00	50.00	25.00	50.00
GARCH sk$\mathcal{N}$	0.00	41.31	0.00	40.38	0.00	0.00	0.00	0.00	12.50	25.00	0.00	25.00
GARCH $\mathcal{S}$	0.94	1.17	0.70	0.70	54.55	54.55	36.36	54.55	25.00	12.50	25.00	12.50
GARCH sk$\mathcal{S}$	0.23	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GJR $\mathcal{N}$	0.47	38.73	0.47	36.15	18.18	18.18	36.36	27.27	25.00	37.50	25.00	37.50
GJR sk$\mathcal{N}$	0.00	40.38	0.00	39.91	0.00	0.00	0.00	0.00	12.50	12.50	0.00	12.50
GJR $\mathcal{S}$	1.64	1.64	0.70	0.47	18.18	27.27	18.18	27.27	37.50	37.50	37.50	37.50
GJR sk$\mathcal{S}$	0.00	0.00	0.00	0.00	0.00	18.18	0.00	18.18	0.00	0.00	0.00	0.00
Panel C: DQ 1%–VaR												
GARCH $\mathcal{N}$	14.08	53.52	14.32	54.69	63.64	90.91	72.73	90.91	25.00	37.50	12.50	37.50
GARCH sk$\mathcal{N}$	14.08	48.36	15.49	50.00	45.45	63.64	45.45	63.64	12.50	37.50	12.50	37.50
GARCH $\mathcal{S}$	19.95	28.64	16.90	29.34	54.55	63.64	63.64	54.55	25.00	25.00	25.00	25.00
GARCH sk$\mathcal{S}$	18.31	23.94	17.37	24.18	45.45	45.45	36.36	36.36	12.50	25.00	12.50	25.00
GJR $\mathcal{N}$	5.87	32.39	6.10	34.74	18.18	90.91	36.36	90.91	12.50	37.50	12.50	37.50
GJR sk$\mathcal{N}$	5.87	27.00	6.10	28.17	9.09	27.27	9.09	45.45	12.50	25.00	0.00	25.00
GJR $\mathcal{S}$	7.04	10.33	4.46	9.86	18.18	27.27	18.18	18.18	12.50	25.00	12.50	25.00
GJR sk$\mathcal{S}$	5.16	10.33	6.57	11.27	0.00	0.00	0.00	0.00	12.50	12.50	12.50	12.50
Panel D: DQ 5%–VaR												
GARCH $\mathcal{N}$	3.52	26.29	3.52	25.82	18.18	9.09	36.36	9.09	0.00	0.00	0.00	0.00
GARCH sk$\mathcal{N}$	3.52	29.81	2.82	30.05	9.09	9.09	9.09	9.09	0.00	0.00	0.00	0.00
GARCH $\mathcal{S}$	1.64	7.75	1.64	8.92	45.45	54.55	36.36	54.55	0.00	0.00	0.00	0.00
GARCH sk$\mathcal{S}$	2.11	6.57	2.82	7.98	9.09	9.09	9.09	9.09	0.00	0.00	0.00	0.00
GJR $\mathcal{N}$	0.00	14.32	0.00	14.55	9.09	9.09	9.09	0.00	0.00	0.00	0.00	0.00
GJR sk$\mathcal{N}$	0.00	15.02	0.00	13.62	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GJR $\mathcal{S}$	0.00	0.00	0.00	1.17	9.09	0.00	9.09	9.09	12.50	12.50	12.50	12.50
GJR sk$\mathcal{S}$	0.00	0.70	0.00	0.70	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 3: Standardized gain in average performance when switching from MS to SR models

This table presents the Diebold and Mariano (1995) test statistic of equal average loss between the MS and SR models for forecasting the distribution of h–day cumulative log–returns ($h \in \{1, 5, 10\}$). As loss functions, we consider the QL and FZL measures (at $\alpha = 1\%$ and $\alpha = 5\%$), and the wCRPS measure. Negative values indicate outperformance of the Markov–switching specification compared with the single–regime models. In light (dark) gray, we report statistics which are significantly negative (positive) at the 1% level (bilateral test). The multi–step cumulative log–returns forecasts are generated using 25,000 simulated paths of daily log–returns. Models are estimated with the Bayesian approach.

		Stocks					Stock market indices					Exchange rates				
Horizon	Model	QL 1%	QL 5%	FZL 1%	FZL 5%	wCRPS	QL 1%	QL 5%	FZL 1%	FZL 5%	wCRPS	QL 1%	QL 5%	FZL 1%	FZL 5%	wCRPS
$h = 1$	GARCH $\mathcal{N}$	-0.60	-4.94	-3.74	-8.54	-9.32	-3.84	0.50	-5.11	-1.59	-4.04	-0.09	0.39	-0.90	0.41	-2.65
	GARCH sk$\mathcal{N}$	-0.25	-4.90	-3.43	-8.39	-9.25	-2.64	0.10	-3.32	-1.18	-3.26	0.95	-0.25	-0.77	0.37	-3.41
	GARCH $\mathcal{S}$	-4.00	-3.55	-3.82	-4.42	-3.41	-1.50	-0.90	-1.70	-1.37	-0.17	1.12	-1.26	1.17	1.08	-2.17
	GARCH sk$\mathcal{S}$	-4.52	-4.20	-3.86	-4.86	-2.79	-2.21	-0.87	-1.69	-0.86	0.22	2.18	-0.56	1.63	1.26	-1.45
	GJR $\mathcal{N}$	-0.63	-6.02	-4.07	-10.74	-9.96	-3.58	0.53	-4.99	-2.3	-4.30	0.64	0.58	-0.98	-0.02	-1.64
	GJR sk$\mathcal{N}$	-0.22	-5.95	-3.76	-10.32	-9.94	-2.04	-0.31	-3.06	-1.40	-3.00	0.79	0.07	-0.87	0.11	-1.88
	GJR $\mathcal{S}$	-3.88	-4.44	-3.23	-4.29	-5.00	-1.80	0.49	-2.26	-1.43	0.11	0.36	-1.03	-0.20	-0.19	-2.35
	GJR sk$\mathcal{S}$	-3.64	-4.17	-3.16	-4.10	-3.44	-0.93	0.61	-1.20	-0.55	0.47	0.92	-0.68	1.32	0.65	-1.66
$h = 5$	GARCH $\mathcal{N}$	-0.66	-1.70	-2.16	-7.24	-2.72	-2.11	-1.47	-4.97	-2.48	-1.19	2.69	1.69	-0.52	0.95	0.73
	GARCH sk$\mathcal{N}$	-0.52	-1.67	-1.93	-7.01	-2.65	-2.14	-1.61	-2.69	-1.40	-0.88	2.59	1.17	-0.70	0.43	0.46
	GARCH $\mathcal{S}$	-1.78	-2.27	-2.86	-3.22	-2.68	-0.77	-1.61	-1.52	-2.18	-0.89	2.00	1.53	1.26	1.55	-1.26
	GARCH sk$\mathcal{S}$	-1.70	-2.37	-2.68	-3.23	-2.39	-1.55	-2.72	-0.44	-1.32	-0.32	3.67	0.85	1.02	1.50	-0.93
	GJR $\mathcal{N}$	-0.53	-2.38	-2.24	-8.29	-2.77	0.10	-0.10	-4.91	-3.29	-0.51	3.27	1.15	-1.02	-0.23	0.30
	GJR sk$\mathcal{N}$	-0.30	-2.37	-1.98	-8.37	-2.74	0.62	-1.10	-2.76	-2.44	-0.30	3.70	1.54	-1.09	-0.40	1.15
	GJR $\mathcal{S}$	-1.32	-2.63	-2.46	-2.74	-4.52	-0.21	-1.76	-2.08	-2.85	-0.49	4.08	0.04	0.41	0.75	-2.07
	GJR sk$\mathcal{S}$	-1.14	-1.61	-2.26	-2.75	-3.37	0.26	-0.65	-1.06	-1.40	-0.04	4.60	1.05	1.46	1.11	-1.09
$h = 10$	GARCH $\mathcal{N}$	-0.18	-0.82	-1.59	-6.36	-1.93	-1.23	-0.66	-3.96	-2.07	-1.91	1.55	1.18	-0.33	0.99	0.89
	GARCH sk$\mathcal{N}$	-0.14	-0.71	-1.42	-6.32	-1.96	-1.76	-0.80	-2.69	-2.12	-1.35	1.23	1.58	-0.72	0.25	0.86
	GARCH $\mathcal{S}$	-0.83	-1.01	-1.77	-1.90	-1.25	-1.05	-0.79	-1.69	-2.01	-1.95	1.12	1.63	1.24	1.55	-0.59
	GARCH sk$\mathcal{S}$	-0.93	-1.22	-1.53	-2.04	-1.02	-1.19	-0.99	-1.52	-1.60	-1.22	2.78	2.12	1.30	1.54	-0.46
	GJR $\mathcal{N}$	-0.15	-1.24	-1.64	-6.73	-1.91	0.48	-1.08	-3.87	-2.81	-0.92	1.16	1.11	-0.99	-0.41	0.62
	GJR sk$\mathcal{N}$	0.02	-1.22	-1.32	-6.76	-1.71	0.21	-1.44	-1.79	-1.79	-0.87	2.47	1.15	-0.97	-0.11	1.03
	GJR $\mathcal{S}$	-0.53	-1.46	-1.47	-2.10	-4.04	1.16	-1.43	-1.47	-3.32	-1.13	2.92	0.34	0.71	0.89	-1.71
	GJR sk$\mathcal{S}$	-0.48	-1.15	-1.56	-2.44	-2.80	0.72	-2.19	-1.44	-2.17	-1.31	4.55	1.12	1.79	1.17	-1.09

Table 4: Standardized gain in average performance when switching from MS to SR and changing the specification

This table presents the Diebold and Mariano (1995) test statistic of equal average wCRPS between a MS implementation (in rows) and a SR implementation (in column), for all considered specifications, when forecasting the distribution of one–day ahead log–returns. We report test statistics computed with robust HAC standard errors. Negative values indicate outperformance of the Markov–switching specification compared with single–regime models. In light (dark) gray, we report statistics which are significantly negative (positive) at the 1% level (bilateral test). Models are estimated with the Bayesian approach.

		SR GARCH				SR GJR			
		$\mathcal{N}$	sk$\mathcal{N}$	$\mathcal{S}$	sk$\mathcal{S}$	$\mathcal{N}$	sk$\mathcal{N}$	$\mathcal{S}$	sk$\mathcal{S}$
Panel A: Stocks									
MS GARCH	$\mathcal{N}$	-9.32	-9.56	3.29	3.30	-6.80	-6.85	3.29	3.38
	sk$\mathcal{N}$	-9.00	-9.25	3.60	3.67	-6.60	-6.65	3.42	3.54
	$\mathcal{S}$	-9.01	-9.20	-3.41	-2.99	-7.29	-7.36	-0.14	-0.13
	sk$\mathcal{S}$	-8.86	-9.07	-2.92	-2.79	-7.15	-7.22	0.01	0.04
MS GJR	$\mathcal{N}$	-10.11	-10.26	0.88	0.93	-9.96	-10.25	3.20	3.18
	sk$\mathcal{N}$	-9.88	-10.06	0.88	0.95	-9.64	-9.94	3.33	3.38
	$\mathcal{S}$	-9.73	-9.88	-2.92	-2.76	-9.48	-9.68	-5.00	-4.79
	sk$\mathcal{S}$	-9.57	-9.74	-2.46	-2.34	-9.24	-9.46	-3.19	-3.44
Panel B: Stock market indices									
MS GARCH	$\mathcal{N}$	-4.04	-0.67	3.09	6.00	4.80	7.15	8.15	9.76
	sk$\mathcal{N}$	-5.25	-3.26	-1.04	3.29	3.06	5.46	6.18	8.55
	$\mathcal{S}$	-5.66	-2.90	-0.17	5.09	3.68	6.13	7.17	9.20
	sk$\mathcal{S}$	-6.08	-4.83	-3.52	0.22	2.00	4.39	4.98	7.71
MS GJR	$\mathcal{N}$	-9.65	-7.81	-6.19	-4.26	-4.30	0.33	2.19	4.76
	sk$\mathcal{N}$	-10.39	-9.41	-7.75	-6.35	-5.21	-3.00	-1.80	1.82
	$\mathcal{S}$	-9.79	-8.28	-6.91	-5.11	-4.66	-1.15	0.11	3.92
	sk$\mathcal{S}$	-10.20	-9.53	-8.29	-7.19	-5.34	-3.80	-2.83	0.47
Panel C: Exchange rates									
MS GARCH	$\mathcal{N}$	-2.65	-3.49	5.38	3.95	-2.06	-2.74	3.52	2.81
	sk$\mathcal{N}$	-2.00	-3.41	4.86	5.74	-1.53	-2.45	3.44	3.78
	$\mathcal{S}$	-6.84	-6.53	-2.17	-2.36	-6.09	-6.03	-2.31	-2.45
	sk$\mathcal{S}$	-5.45	-6.29	-0.99	-1.45	-4.81	-5.61	-1.32	-1.73
MS GJR	$\mathcal{N}$	-1.71	-2.33	4.40	3.59	-1.64	-2.35	5.32	3.89
	sk$\mathcal{N}$	-1.13	-1.95	4.26	4.53	-1.02	-1.88	4.53	5.14
	$\mathcal{S}$	-6.02	-6.03	-1.56	-1.68	-6.38	-6.38	-2.35	-2.46
	sk$\mathcal{S}$	-5.05	-5.49	-0.84	-1.21	-5.21	-5.74	-1.35	-1.66

Table 5: Standardized gain in average performance when switching from the MS GJR sk$\mathcal{S}$ model to the Beta–Skew–t–EGARCH(1,1) model

This table presents the Diebold and Mariano (1995) test statistic of equal average loss between the MS GJR sk$\mathcal{S}$ model and the Beta–Skew–t–EGARCH(1,1) model for forecasting the distribution of one–day ahead log–returns. As loss functions, we consider the QL and FZL measures (at $\alpha = 1\%$ and $\alpha = 5\%$), and the wCRPS measure. Negative values indicate outperformance of the Markov–switching specification compared with the Beta–Skew–t–EGARCH(1,1) model. In light gray, we report statistics which are significantly negative at the 1% level (bilateral test). Critical values of a two–sided (one–sided) test are 2.57 (2.33), 1.96 (1.64), and 1.64 (1.28) at the 1%, 5%, and 10% significance levels, respectively. Models are estimated by Maximum Likelihood.

	QL 1%	QL 5%	FZL 1%	FZL 5%	wCRPS
Stocks	-3.54	-0.29	-3.72	-1.17	-1.37
Stock market indices	-4.63	-1.52	-1.03	-1.03	-1.25
Exchange rates	-0.03	2.40	-0.04	1.04	0.58

Table 6: Standardized gain in average performance when switching from Bayesian to frequentist estimation

This table presents the Diebold and Mariano (1995) test statistic of equal average loss between Bayesian and frequentist estimated models for forecasting the distribution of one–day ahead log–returns. As loss functions, we consider the QL and FZL measures (at $\alpha = 1\%$ and $\alpha = 5\%$), and the wCRPS measure. Panels A and B report the test statistics when comparing Bayesian against frequentist estimation for SR and MS specifications, respectively. Negative values indicate outperformance of the Bayesian estimation method. In light (dark) gray, we report statistics which are significantly negative (positive) at the 1% level (bilateral test).

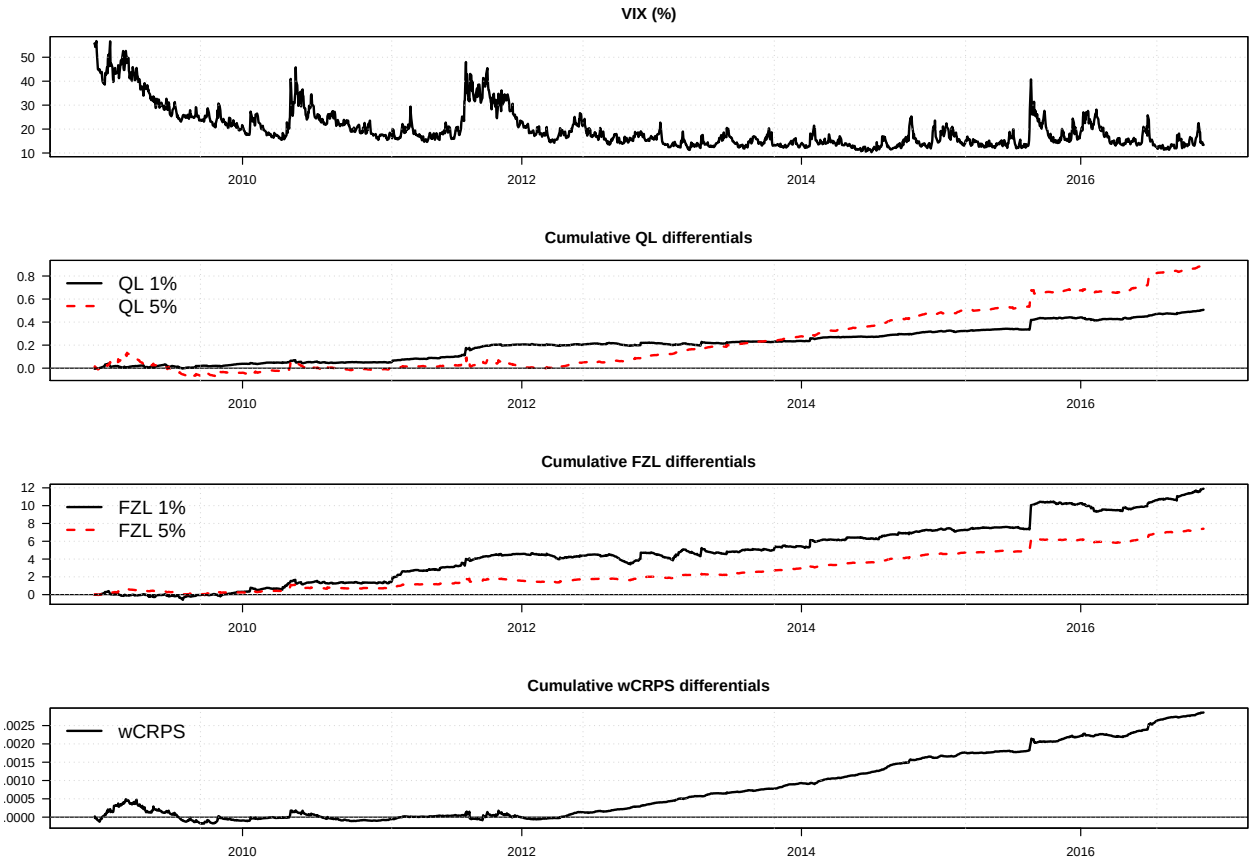
	Stocks					Stock market indices					Exchange rates				
Model	QL 1%	QL 5%	FZL 1%	FZL 5%	wCRPS	QL 1%	QL 5%	FZL 1%	FZL 5%	wCRPS	QL 1%	QL 5%	FZL 1%	FZL 5%	wCRPS
Panel A: Markov–switching GARCH models															
GARCH $\mathcal{N}$	-3.65	-3.26	-2.24	-2.85	-2.24	-0.33	-0.58	0.17	-1.78	-0.25	-1.33	0.99	-1.34	-0.37	-2.02
GARCH sk$\mathcal{N}$	-3.58	-2.93	-2.57	-2.81	-0.60	-1.56	-2.33	-2.05	-2.43	-1.04	-0.82	-1.24	0.97	0.35	-1.04
GARCH $\mathcal{S}$	-2.20	-5.78	-3.12	-4.27	-5.55	0.77	-0.17	-0.89	-0.97	-0.85	-0.78	0.29	-0.69	-1.00	0.35
GARCH sk$\mathcal{S}$	-5.04	-6.88	-6.12	-5.70	-7.04	1.13	-0.52	-0.62	-2.02	-0.58	-1.54	-1.64	-0.35	-1.54	-2.98
GJR $\mathcal{N}$	-1.91	-2.66	-1.59	-2.17	-3.22	-1.21	-2.95	-2.05	-1.81	-2.08	-1.09	-1.38	-0.85	-1.42	-3.61
GJR sk$\mathcal{N}$	-1.83	-3.12	-2.03	-2.44	-2.06	-1.11	-0.84	-2.47	-1.76	-1.40	0.06	-0.32	-0.28	-0.19	-1.17
GJR $\mathcal{S}$	-1.07	-3.11	-2.90	-2.67	-4.48	-1.29	-1.56	-1.66	-2.61	-4.11	-1.75	-2.40	-0.46	-0.51	-4.19
GJR sk$\mathcal{S}$	-3.10	-3.90	-4.67	-2.54	-5.28	-2.95	-2.02	-0.66	-0.46	-3.48	-1.59	-0.38	-0.81	-0.75	-2.50
Panel B: Single–regime GARCH models															
GARCH $\mathcal{N}$	-5.05	-4.23	-6.62	-4.50	-7.84	-2.99	-0.23	-3.40	-3.85	-5.63	-1.59	-0.42	-1.58	0.24	-3.20
GARCH sk$\mathcal{N}$	-4.77	-3.36	-6.59	-4.25	-6.64	-2.55	-1.05	-3.49	-1.98	-4.48	-1.33	-0.86	0.48	-1.00	-4.14
GARCH $\mathcal{S}$	-5.13	-5.08	-5.48	-5.34	-4.93	-1.27	-0.60	-3.01	-1.53	-3.39	-1.41	-1.12	0.15	1.18	-3.76
GARCH sk$\mathcal{S}$	-5.72	-5.40	-5.73	-5.44	-5.18	-2.74	-1.51	0.56	-0.38	-3.87	-2.83	-2.47	-1.74	-1.61	-4.46
GJR $\mathcal{N}$	-5.11	-2.80	-7.64	-3.90	-6.90	-3.65	-2.43	-5.52	-4.87	-5.92	-1.67	-2.70	-1.87	-0.93	-4.14
GJR sk$\mathcal{N}$	-4.55	-2.30	-7.22	-3.02	-5.65	-2.26	-2.03	-2.86	-1.30	-3.94	-0.13	-2.62	-1.53	-0.30	-4.61
GJR $\mathcal{S}$	-3.78	-4.23	-4.90	-4.14	-5.23	-4.13	-2.52	-4.94	-4.00	-4.17	-1.61	-1.96	0.97	-1.29	-4.71
GJR sk$\mathcal{S}$	-3.93	-4.06	-5.32	-4.41	-5.03	-3.82	-1.64	-3.01	-2.01	-3.16	-1.46	-2.24	-0.94	0.74	-4.66

Table 7: Standardized average gain in performance when switching from constrained MS to SR models

This table presents the Diebold and Mariano (1995) test statistic of equal average loss function between the constrained MS and SR models for forecasting the distribution of h–day cumulative log–returns ($h \in \{1, 5, 10\}$). As loss functions, we consider the QL and FZL measures (at $\alpha = 1\%$ and $\alpha = 5\%$), and the wCRPS measure. We report the test statistics computed with robust HAC standard errors, for the time series in the various universes. Negative values indicate outperformance of the shape–parameter constrained Markov–switching specification compared with its single–regime counterpart. In light (dark) gray, we report statistics which are significantly negative (positive) at the 1% level (bilateral test). The multi–step cumulative log–returns forecasts are generated using 25,000 simulated paths of daily log–returns. Models are with the Bayesian approach.

		Stocks					Stock market indices					Exchange rates				
Horizon	Model	QL 1%	QL 5%	FZL 1%	FZL 5%	wCRPS	QL 1%	QL 5%	FZL 1%	FZL 5%	wCRPS	QL 1%	QL 5%	FZL 1%	FZL 5%	wCRPS
$h = 1$	GARCH sk$\mathcal{N}$	-0.44	-5.19	-3.56	-8.80	-9.34	-2.68	0.66	-3.53	-1.00	-3.26	0.27	-0.93	-1.13	-0.31	-3.70
	GARCH $\mathcal{S}$	-2.43	-2.32	-3.49	-3.74	-2.97	-1.53	-1.54	-1.68	-1.51	-1.02	0.73	0.10	-0.08	0.25	-0.53
	GARCH sk$\mathcal{S}$	-2.70	-2.69	-3.83	-4.22	-2.62	-1.70	-0.98	-1.33	-0.65	-1.40	1.62	0.31	-0.01	0.54	-0.23
	GJR sk$\mathcal{N}$	-0.37	-6.39	-3.90	-10.91	-9.92	-1.95	-1.59	-3.15	-2.21	-5.07	0.16	-0.91	-1.23	-0.89	-2.77
	GJR $\mathcal{S}$	-2.64	-2.99	-3.29	-3.80	-4.33	-2.25	-0.48	-2.29	-1.44	-0.64	0.30	-0.60	-0.63	-0.22	-1.04
	GJR sk$\mathcal{S}$	-2.90	-3.20	-3.31	-3.73	-4.10	-1.34	-0.93	-1.38	-0.89	-0.71	0.72	-0.40	0.32	-0.17	-0.97
$h = 5$	GARCH sk$\mathcal{N}$	-0.62	-1.79	-1.98	-7.07	-2.76	-2.05	-0.55	-3.02	-1.67	-0.73	2.75	0.63	-1.08	-0.27	0.18
	GARCH $\mathcal{S}$	-0.48	-0.98	-1.74	-1.93	-1.44	-0.92	-2.23	-1.00	-2.43	-1.72	1.96	0.28	-0.58	0.34	-0.58
	GARCH sk$\mathcal{S}$	-0.81	-1.06	-1.98	-2.07	-1.08	-1.58	-2.34	0.16	-1.11	-1.45	1.25	1.56	-0.54	0.65	0.46
	GJR sk$\mathcal{N}$	-0.44	-2.46	-2.11	-8.61	-2.60	-0.48	-1.49	-3.14	-2.55	-0.49	1.94	0.55	-1.26	-1.00	0.41
	GJR $\mathcal{S}$	-0.41	-1.26	-1.94	-2.29	-3.19	-0.19	-1.03	-2.62	-3.49	-0.87	3.19	1.40	-0.61	0.11	-0.56
	GJR sk$\mathcal{S}$	-0.21	-0.83	-1.71	-1.97	-2.86	0.80	-0.33	-1.59	-1.21	0.02	3.02	1.54	0.63	0.24	-0.92
$h = 10$	GARCH sk$\mathcal{N}$	-0.25	-0.84	-1.48	-6.63	-2.06	-1.92	0.03	-3.43	-2.68	-1.71	1.44	1.55	-0.80	-0.09	0.73
	GARCH $\mathcal{S}$	-0.29	-0.25	-0.86	-1.09	-0.23	-1.74	-1.46	-2.00	-2.85	-2.11	0.17	1.76	-0.48	0.40	0.10
	GARCH sk$\mathcal{S}$	-0.27	0.09	-0.61	-0.93	0.42	-0.40	-1.94	-1.01	-2.13	-2.30	1.20	1.30	1.19	1.70	0.92
	GJR sk$\mathcal{N}$	-0.10	-1.28	-1.43	-6.86	-1.80	-0.74	-1.92	-2.28	-2.49	-1.90	0.66	0.61	-1.06	-0.82	0.58
	GJR $\mathcal{S}$	-0.18	-0.41	-0.97	-1.55	-1.74	-0.53	-1.24	-2.23	-3.21	-1.79	2.36	1.27	-0.04	0.24	-0.25
	GJR sk$\mathcal{S}$	0.01	-0.25	-0.91	-1.29	-1.45	0.12	-1.54	-0.95	-1.60	-1.14	2.86	1.06	-0.36	0.47	-1.21

Figure 1: Cumulative performance
This figure presents the evolution of VIX (the Chicago Board of Exchange's volatility index) in the top panel, together with the cumulative average loss differentials (QL, FZL and wCRPS) for the 2,000 out–of–sample observations (ranging from December 2008 to November 2016). The comparison is done between the Markov–switching and the single–regime GJR skewed Student–t models. A positive value indicates outperformance of the Markov–switching specification. A positive slope indicates outperformance at the corresponding date.



Questioning the news about economic growth: Sparse forecasting using thousands of news–based sentiment values

Keven Bluteau – Chapter 2
joint work with David Ardia and Kris Boudt
Forthcoming in International Journal of Forecasting

Abstract

Modern calculation of textual sentiment involves a myriad of choices for the actual calibration. We introduce a general sentiment engineering framework that optimizes the design for forecasting purposes. It includes the use of the elastic net for sparse data–driven selection and weighting of thousands of sentiment values. These values are obtained by pooling the textual sentiment values across publication venues, article topics, sentiment construction methods, and time. We apply the framework to investigate the added value of textual analysis–based sentiment indices for forecasting economic growth in the US. We find that, compared to the use of high–dimensional forecasting techniques based on only economic and financial indicators, the additional use of optimized news–based sentiment values yields significant accuracy gains in forecasting the nine–month and annual growth rates of the US industrial production.

Keywords: elastic net, sentiment analysis, time–series aggregation, topic–sentiment, US industrial production, sentometrics

1. Introduction

Understanding the current and future state of the economy is crucial for timely and efficient economic policy and business decision–making. Forecasts of economic variables such as the country's gross domestic product, industrial production, consumer spending, and unemployment rate are closely followed by policymakers to assess the state of the economy. It seems self–evident that not only is the readily available quantitative information useful to obtain this assessment, but so is the qualitative information in news reports.

In practice, however, the dominating approach is to exclusively use the available quantitative information for economic growth prediction. In fact, most often, the macroeconomic variables are forecasted using a large panel of macroeconomic indicators which reflects the economic environment; see Stock and Watson (2002). Additionally, surveys such as the University of Michigan Consumer Sentiment Index or the Conference Board's Consumer Confidence Index for the US, and the European Economic Sentiment Index (ESI) for Europe can contain information about the current and future economic growth. The US survey–based sentiment indices are used in Bram and Ludvigson (1998) and Ludvigson (2004) to forecast US household expenditure and

consumer spending, while the ESI is used in Gelper and Croux (2010) to forecast national and aggregated European industrial production growth rates. Finally, financial indicators that reflect economic and financial expectations, as well as credit conditions, are used in Espinoza et al. (2012) to forecast long–term US and Euro area GDP growth.

In this paper, we complement the readily available quantitative information (*i.e.*, macroeconomic, financial, and survey–based indicators) with predictors obtained from a large set of sentiment values expressed by authors of news discussing a country's economy to obtain timely forecasts of the country's economic growth. The approach starts off with a rich (big) data environment of a virtually infinite number of texts. These texts need to be selected, transformed into sentiment values, and then aggregated. The potential high–dimensionality of the data becomes an issue, as we want to only extract the relevant information from the text and create informative indices for predicting economic growth.

To address this challenge, we propose a methodology which first computes thousands of sentiment values capturing the tone expressed by the authors of news discussing topics related to the country's economic growth. It then maps the hordes of sentiment values in a single economic growth prediction using aggregation based on (1) sentiment computation method (*e.g.*, using various lexicons), (2) topic (*e.g.*, "real estate market" or "job creation") and (3) time (*e.g.*, short and long–term sentiment indices). We then use a data–driven calibration approach based on penalized least squares regression to optimally combine the indices to forecast a variable of interest. We refer to the resulting optimized aggregate value of sentiment as a text–based sentiment index. The resulting index is a linear combination of the original sentiment values. This is a choice of design that allows us to perform an attribution analysis of the sentiment prediction to gauge the contribution of the various textual sentiment indices to the prediction.

Besides being flexible, timely, and data–rich, the proposed methodology has the advantage that its design can be backtested. In a real-time setting, its design adapts itself to the changing forecasting environment, that is, the weights attributed to each component of the final sentiment index change according to the economic environment and the targeted variable to forecast. Gelper and Croux (2010) find that letting the aggregation weights of each component of the survey–based ESI be data–driven improve its forecasting performance compared to the ad–hoc weights set by the European Commission. This feature is, by construction, integrated in our textual sentiment index. Furthermore, it also alleviates to a certain degree most of the subjective decisions that a forecaster has to do before the forecasting exercise. Indeed, the optimization

process automatically chooses which sentiment computation methods are used for each topic (topic–specific sentiment calculation), which topic is included in the textual–sentiment index (removal of non–predictive topics), and how past values of each component of the textual–sentiment index are considered (structured lag per component). This adaptive scheme is thus more flexible than text–based (sentiment) indices with a fixed design, like the Economic Policy Uncertainty (EPU) index of Baker et al. (2016). Moreover, the latter is not optimized for forecasting and not aimed at extracting sentiment.

This paper contributes to the increasing literature on the use of text– and news–based measures as sources of information for forecasting and assessing the economy (see, *e.g.*, Thorsrud, 2018, 2016; Baker et al., 2016; Tobback et al., 2018; Shapiro et al., 2018). We exploit the sentiment information in news articles incrementally to the information included in the macroeconomic indicators. Two approaches exist to deal with the high–dimensionality of the latter. First, via dimensionality reduction through (dynamic) factor models (see, *e.g.*, Stock and Watson, 2011, for a review). In this case one assumes that a small number of unobserved factors drives the economy. Many methods have been developed to tackle the problem of estimating the latent factors (see Stock and Watson, 2002; Doz et al., 2011, 2012; Bräuning and Koopman, 2014) and choosing the appropriate number of factors (see Bai and Ng, 2002; Alessi et al., 2010). Second, via penalized regression models used as a replacement or in conjunction to factor models. Bai and Ng (2008) combine penalized regression with factor models to first select a set of predictors and then constructing the factors from them. Different variants of this approach are tested in Kim and Swanson (2014), Kim and Swanson (2018), and Smeekes and Wijler (2018). The proposed optimization of textual sentiment can be applied in conjunction to those traditional methods for a wide set of forecasting problems.

We illustrate the methodology in the case of forecasting the economic growth for the United States. We find that, for an out–of–sample evaluation window ranging from January 2001 to December 2016, the text–based sentiment indices computed from news in major US newspapers provide additional predictive power for the nine–month and annual growth rates of the US industrial production index, controlling for standard use of macroeconomic, sentiment–survey, and financial variables. Moreover, we test to which extent each dimension of the sentiment index (sentiment calculation method, topic, and time) matters. We find that the optimization of all dimensions is important to achieve a high forecasting accuracy but, in order, the most relevant is the time dimension followed by the topic and then the sentiment calculation method. Our

result is shown to be robust to various choices of implementation.

In a wish to disseminate the methodology, and render the results reproducible, we have released the R package **sentometrics** (Ardia et al., 2018, 2017), which implements all the steps described in this paper in the R statistical language with efficient C++ code. We hope that this paper and the accompanying package will encourage practitioners such as government institutions and academics to use and test our framework for optimizing the use of textual sentiment for forecasting their variable(s) of interest.

The rest of the paper proceeds as follows. Section 2 introduces the methodology. Section 3 presents the empirical study. Section 4 concludes.

2. Methodology

The variable to predict is the h–period logarithmic change in the variable Y_t, expressed in percentage points:

$$y_t^h \equiv 100 \times (\ln Y_{t+h} - \ln Y_t)\,, \tag{1}$$

where $t = 1, 2, \ldots, T$ is a time index. We require y_t^h to be covariance stationary. This is typically the case when Y_t represents a country's economic activity (*e.g.*, its gross domestic product or industrial production), its price level (*e.g.*, the consumer price index or the exchange rate), and similarly for corporate variables, like the firm's sales or stock price. In our application, y_t^h is the logarithmic growth in industrial production of the US over horizons ranging from one to twelve months. Note that, due to the publication lag, it may be that Y_t is not known at time t.

Let T be the day for which we need a prediction of y_T^h. Specifically, we want to estimate the expected value of y_T^h given the information available at time T, that is, $\mathbb{E}(y_T^h \,|\, \mathcal{I}_T)$. This is a common problem in time–series forecasting, where, typically, the information set $\mathcal{I}_T$ consists of the usual available quantitative information, such as past values of Y_t as well as macroeconomic and financial metrics (see, *e.g.*, Stock and Watson, 2002; Espinoza et al., 2012). We expand the information set by also including various sentiment values extracted from a corpus of texts published up to date T. We describe below the methodology, as depicted in Figure 1.

[Insert Figure 1 about here.]

2.1. Data preparation

Step 1: Classify texts by topic and use expert opinion to choose a subset of topics to select the potentially relevant texts. We assume that all texts are categorized by a set of topic–markers.

These topic–markers are usually provided by the publishers of the texts or extracted directly from the texts. In our application, we use the corpus of major US newspaper from *LexisNexis* for which topics are readily available using *LexisNexis*' proprietary SmartIndexing™ technology. Alternative techniques for topic identification include the use of likelihood–based techniques using probabilistic models such has the latent Dirichlet allocation (see Liu et al., 2016, for a recent review). Latent Dirichlet allocation has, for example, been used recently by Thorsrud (2016) in conjunction with the dynamic factor model developed in Thorsrud (2018) to nowcast the Norwegian GDP growth. It also includes keywords–based identification such as those keywords used to identify EPU related texts in Baker et al. (2016), or, if topic–labelled news are available for a training set, identification via a support vector machine classifier such as in Tobback et al. (2018).[1] Expert opinion is then used to exclude the topics that, beforehand, can be qualified as being irrelevant for forecasting the variable of interest y_T^h. The resulting topic–markers for our application on forecasting economic growth are reported in Table 1. The corpus consists of the texts that discuss at least one of the selected topics. The corpus is organized in terms of publication date t, with $t = 1, \ldots, T$, where N_t is the number of texts in the corpus of texts published at time t. We use n to index the text available at time t, with $n = 1, \ldots, N_t$.

[Insert Table 1 about here.]

Step 2: Compute for each text n of corpus t the sentiment using L methods. For each text, we compute the underlying sentiment using L different textual sentiment computation methods. For a general review of available methods, we refer the reader to Ravi and Ravi (2015). Methods can differ from each other in terms of the item classified (*e.g.*, word, sentence, paragraph), the method of classification (*e.g.*, supervised or unsupervised), the aggregation method used to obtain a single value per text (*e.g.*, equal–weighting, inverse frequency weighting), among others. In our application, we use the simple bag–of–words approach to compute the net sentiment using L different lexicons to classify the words as positive, negative, or neutral. We thus obtain for each text document $n = 1, \ldots, N_t$, published at time $t = 1, \ldots, T$, L different sentiment values, which we denote by $s_{n,t,l}$, where $l = 1, \ldots, L$.

2.2. Aggregating sentiment into a prediction

At this stage, we have for each day t and for each of the N_t texts, L textual sentiment computation methods and thus L vectors $\mathbf{s}_{t,l} \equiv (s_{1,t,l}, \ldots, s_{N_t,t,l})'$ of size $N_t \times 1$. The next

[1] Other high–accuracy machine–learning classification methods are, of course, viable.

steps aim at reducing the high–dimensionality of the available texts (*i.e.*, the total of texts is $N_1 + \ldots + N_T$). To that end, we first compute the daily sentiment per topic–markers by aggregating across the sentiment of texts published on a given day. We then aggregate over time. We choose a linear mapping as this allows us to perform sentiment attribution. We do not use aggregation to reduce the dimensionality of the number of methods L, as it is small compared to the cross–section and time–series dimensions, and can be handled at the estimation stage through penalized regression.

Step 3: For each corpus n and method l, obtain K topic–based sentiments. We compute sentiment values for each topic–marker by aggregating across the sentiment values of the texts associated with each topic–marker. Formally, we define for each day t the text–to–topic aggregation matrix $\mathbf{W}_t$ of dimension $K \times N_t$ such that the L vectors $\mathbf{W}_t \mathbf{s}_{t,l}$ $(l = 1, \ldots, L)$ of dimension $K \times 1$ capture the daily sentiment for each of the K topics. In the application, each row of $\mathbf{W}_t$ is divided by its total sum, which corresponds to equally weighting the texts for each topic. The equal–weighting approach has the advantage of simplicity. An alternative approach for calibrating the text–to–topic aggregation matrix $\mathbf{W}_t$ could be to use expert opinion or a data–driven procedure to overweight the sources of news (*i.e.*, type of journal or publisher) that are deemed more informative for predicting economic growth.

Step 4: For each topic k and method l, obtain time–series aggregated values. Next, we aggregate through time. We take a maximum time–aggregation lag τ $(0 \leq \tau < T)$, and, for a given l, we stack the vectors column–by–column into $K \times (\tau + 1)$ matrices as follows:

$$\mathbf{V}_{t,l} \equiv \begin{bmatrix} | & & | \\ \mathbf{W}_{t-\tau}\mathbf{s}_{t-\tau,l} & \cdots & \mathbf{W}_t\mathbf{s}_{t,l} \\ | & & | \end{bmatrix}. \tag{2}$$

We do this for $l = 1, \ldots, L$, and then stack the matrices row–by–row into the $LK \times (\tau + 1)$ matrix:

$$\mathbf{V}_t \equiv \begin{bmatrix} \mathbf{V}_{t,1} \\ \vdots \\ \mathbf{V}_{t,L} \end{bmatrix}. \tag{3}$$

Given $\mathbf{V}_t$ and a suitable time aggregation matrix $\mathbf{B}$ of size $(\tau+1)\times B$, we then construct the final vector of size $LKB\times 1$ of textual sentiment predictors $\mathbf{s}_t$ as:

$$\mathbf{s}_t \equiv \text{vec}(\mathbf{V}_t\mathbf{B})\,, \tag{4}$$

where $\text{vec}(\cdot)$ is the vectorization operator.[2]

We use a data–driven calibration of the aggregation matrix $\mathbf{B}$ to strike a balance between a strong decay in weights to obtain timeliness on the one hand, and, on the other hand, an equal–weighting approach to obtain efficiency when all time–lags are equally informative. To do so, we rely on the Beta weighting function, often used in the mixed–data sampling literature (see Ghysels et al., 2007). The approach requires two parameters $a>0$ and $b>0$:

$$c(i;a,b) \equiv \frac{f(\frac{i}{\tau};a,b)}{\sum_{i=1}^{\tau} f(\frac{i}{\tau};a,b)}\,, \tag{5}$$

where $f(x;a,b) \equiv \frac{x^{a-1}(1-x)^{b-1}\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$ is the Beta density function and $\Gamma(\cdot)$ is the Gamma function.

Given a grid $\{a_i,b_i\}_{i=1}^{B}$, the $(\tau+1)\times B$ aggregation matrix is given by:

$$\mathbf{B} \equiv \begin{bmatrix} c(1;a_1,b_1) & & c(1;a_B,b_B) \\ \vdots & \cdots & \vdots \\ c(\frac{i}{\tau};a_1,b_1) & & c(\frac{i}{\tau};a_B,b_B) \\ \vdots & \cdots & \vdots \\ c(0;a_1,b_1) & & c(0;a_B,b_B) \end{bmatrix}. \tag{6}$$

Step 5: Calibration to optimize forecast precision. The next and final aggregation step is to aggregate these textual sentiment indices optimally given a variable of interest. To this end, we define the following model:

$$y_t^h = \alpha + \boldsymbol{\gamma}'\mathbf{x}_t + \boldsymbol{\beta}'\mathbf{s}_t + \varepsilon_t \quad (t=1,\ldots,T)\,, \tag{7}$$

where α is an intercept, $\mathbf{x}_t$ is a $M\times 1$ vector of (non–textual sentiment) variables available at time t, $\boldsymbol{\gamma}$ is the corresponding vector of parameters, $\boldsymbol{\beta}\equiv(\beta_1,\ldots,\beta_P)'$ is a vector of parameters associated with the P textual–sentiment indices ($P=LKB$), and ε_t is an error term at time t. Typically, $\mathbf{x}_t$ includes y_s where y_s is the dependent variable up to time t, that is $s\leq t$. In practice,

[2]The vectorization operator stacks the columns of a matrix one on top of the other into a vector.

we often have that $s < t$ in economics due to the release lag faced by economic indicators. It is also common to include macroeconomic and financial metrics, or the information obtained from surveys.

We use a penalized least squares criterion to estimate regression (7). Penalization is needed to regularize the estimation of the high–dimensional parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$. Given the high correlation between the sentiment variables, we use the elastic net regularization of Zou and Hastie (2005) to deal with both the high degree of collinearity in the regressors and the need for variable selection.[3]

To ease the presentation, let us define $\mathbf{z}_t \equiv (\mathbf{x}_t', \mathbf{s}_t')'$ and $\boldsymbol{\theta} \equiv (\boldsymbol{\gamma}', \boldsymbol{\beta}')'$, both of size $(M+P)\times 1$. In our context, the optimization problem of the elastic net can then be expressed as:

$$\min_{\widetilde{\alpha},\widetilde{\boldsymbol{\theta}}} \left\{ \frac{1}{T}\sum_{t=1}^{T} \left[y_t^h - \left(\widetilde{\alpha} + \widetilde{\boldsymbol{\theta}}'\widetilde{\mathbf{z}}_t \right) \right]^2 + \lambda_1 \left[\lambda_2 \|\widetilde{\boldsymbol{\theta}}\|_1 + (1-\lambda_2)\|\widetilde{\boldsymbol{\theta}}\|_2^2 \right] \right\}, \tag{8}$$

where $\|\cdot\|_p$ is the L^p–norm, $\lambda_1 \geq 0$ is the parameter that sets the level of regularization and $0 \leq \lambda_2 \leq 1$ is the weight between the two types of penalties. The elastic net regularization nests both the Ridge regularization of Hoerl and Kennard (1970) (when $\lambda_2 = 0$) and LASSO regularization (when $\lambda_2 = 1$) introduced by Tibshirani (1996). The variable $\widetilde{\mathbf{z}}_t$ is the standardized version of $\mathbf{z}_t$ with components $\widetilde{z}_{i,t} \equiv (z_{i,t} - \text{av}_i)/\text{std}_i$, where av_i and std_i are the sample mean and standard deviation of $\{z_{i,t}; t = 1, \ldots, T\}$, respectively. The standardization is crucial in penalized regressions as the penalty depends on the scale of the components of $\boldsymbol{\theta}$.

Once the estimation is done, $\widetilde{\boldsymbol{\theta}}$ is rescaled to give the corresponding optimal unstandardized vector $\widehat{\boldsymbol{\theta}}$. The unstandardized regression parameter can be recovered by rescaling each component of $\widetilde{\boldsymbol{\theta}}$; $\widehat{\theta}_i \equiv \frac{\widetilde{\theta}_i}{\text{std}_i}$ $(i = 1, \ldots, M+P)$. An additional value must then be subtracted from the regression intercept to account for the centering of the series:

$$\widehat{\alpha} \equiv \widetilde{\alpha} - \sum_{i=1}^{M+P} \frac{\widetilde{\theta}_i}{\text{std}_i}\, \text{av}_i\,. \tag{9}$$

The implementation of the elastic net in (8) requires the calibration of the penalty parameters λ_1 and λ_2. We follow Zou et al. (2007) and minimize the so–called BIC–like criterion, where BIC stands for Bayesian Information Criterion.[4] Let the vector $\widehat{\mathbf{y}}^h_{\lambda_1,\lambda_2}$ of size $T\times 1$ be the forecast of

[3] All calibrations are performed with the R package **glmnet** (Friedman et al., 2010). Various models with sparsity features exist, such as the adaptive elastic net of Zou and Zhang (2009). However, in our application to forecasting US growth, we find that these methods do not increase significantly the forecasting performance.

[4] In our study, the low sample size and cross–correlation generated by the overlapping data when $h > 1$ make

$\mathbf{y}^h \equiv (y_1^h, \ldots, y_T^h)'$ obtained by fixing λ_1 and λ_2. The BIC–like criterion is defined as:

$$BIC_{\lambda_1,\lambda_2} \equiv \frac{\|\mathbf{y}^h - \widehat{\mathbf{y}}^h_{\lambda_1,\lambda_2}\|_2^2}{T\sigma^2} + \frac{\ln T}{T}\,\widehat{df}(\widehat{\mathbf{y}}^h_{\lambda_1,\lambda_2})\,, \tag{10}$$

where σ^2 is defined as the variance of the forecast error given by the largest $\widehat{df}(\widehat{\mathbf{y}}^h_{\lambda_1,\lambda_2})$. In (10), $\widehat{df}(\widehat{\mathbf{y}}^h_{\lambda_1,\lambda_2})$ is an estimator of the number of degree–of–freedom of the elastic net given $\widehat{\mathbf{y}}^h_{\lambda_1,\lambda_2}$ (see Tibshirani and Taylor, 2012). In the special case where $\lambda_2 = 1$ (*i.e.*, LASSO regularization), $\widehat{df}(\widehat{\mathbf{y}}^h_{\lambda_1,1})$ is equal to the number of non–zero parameters.[5]

Step 6: Forecasting. As the estimator $\widehat{\boldsymbol{\theta}}$ contains the vectors $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\beta}}$, our forecast at time T is then given by:

$$\widehat{y}_T^h \equiv \widehat{\alpha} + \widehat{\boldsymbol{\gamma}}'\mathbf{x}_T + \widehat{\boldsymbol{\beta}}'\mathbf{s}_T\,. \tag{11}$$

2.3. Forecast precision and attribution

Given the predicted values of y_T^h, it is critical to evaluate whether the computational cost of text–based prediction pays off in terms of a higher out–of–sample precision than when the forecast is obtained using a simpler time–series model. Another step in validating the outcome is to attribute the contribution of each topic to the predicted value.

Step 7: Forecast precision evaluation. For the evaluation of the forecasting performance, we can use the Root Mean Squared Forecast Error (RMSFE) and the Mean Absolute Forecast Error (MAFE). Let $e_{i,t}^h \equiv y_t^h - \widehat{y}_{i,t}^h$ be the error term for model i at time t for an horizon h where $\widehat{y}_{i,t}^h$ is the forecast. The RMSFE and MAFE measures of model i at horizon h are defined by:

$$\text{RMSFE}_i^h \equiv \sqrt{\frac{1}{T_F}\sum_{t=T+1}^{T+T_F}(e_{i,t}^h)^2} \qquad \text{MAFE}_i^h \equiv \frac{1}{T_F}\sum_{t=T+1}^{T+T_F}|e_{i,t}^h|\,, \tag{12}$$

where T is the size of the estimation sample and T_F is the number of out–of–sample observations.

Statistical techniques like the Diebold and Mariano (1995) (DM) test or the Model Confidence Set (MCS) procedure of Hansen et al. (2011) can then be used to evaluate the significance of the

the cross–validation calibration methodology unstable. We also test for other BIC–type criterion such as the extended BIC of Chen and Chen (2008) and the high–dimensional BIC of Wang and Zhu (2011). Performance does not increase significantly in our empirical application.

[5] We use grid–search to find the pair $(\hat{\lambda}_1, \hat{\lambda}_2)$ that minimizes $BIC_{\lambda_1,\lambda_2}$. More specifically, we use the elements in the vector $\boldsymbol{\lambda}_2 \equiv (0, 0.1, 0.3, 0.5, 0.7, 0.9, 1)$ as candidate values of λ_2 and, for each value in $\boldsymbol{\lambda}_2$, a vector $\boldsymbol{\lambda}_{1,\lambda_{2,i}}$, where $\lambda_{2,i}$ is the i–th element of $\boldsymbol{\lambda}_2$, of size 100 is generated using the strategy outlined in Friedman et al. (2010). This gives 100 pairs per candidate λ_2 for a total of 700 pairs ($\boldsymbol{\lambda}_2$ is of size 7). The pair (λ_1,λ_2) that gives the largest degree–of–freedom used to compute σ^2 is found by computing the degree–of–freedom given by each pair. Then, the pair $(\hat{\lambda}_1,\hat{\lambda}_2)$ is the pair that minimize $BIC_{\lambda_1,\lambda_2}$.

difference in forecasting precision between models.[6] When comparing nested models, as we do in the application, the p–value of the DM test has a non–standard distribution. We recommend to use the critical values obtained using the bootstrap approach of Clark and McCracken (2001).

Step 8: Attribution. Until now, our exposition has been a bottom–up story of aggregating the sentiment of individual texts through cross–sectional, time–series, and elastic net weighting into a prediction of economic growth. Once this prediction is obtained, it is important to top–down attribute the obtained prediction to the individual texts at various granularity levels. In fact, thanks to the linearity of the methodology, it is straightforward to retrieve the forecast as a function of the individual text sentiment $s_{n,t,l}$:

$$\widehat{y}_T^h = \widehat{\alpha} + \widehat{\boldsymbol{\gamma}}' \mathbf{x}_T + \sum_{t=(T-\tau)}^{T} \sum_{n=1}^{N_t} \sum_{l=1}^{L} \sum_{k=1}^{K} \sum_{b=1}^{B} \widehat{\boldsymbol{\beta}}' \boldsymbol{e}_{l,k,b} \cdot W_{t,k,n} B_{T-t,b} \cdot s_{n,t,l} \,, \tag{13}$$

where $\boldsymbol{e}_{l,k,b}$ is basis vector of size $LKB \times 1$, which extracts the relevant regression parameter in $\widehat{\boldsymbol{\beta}}$ given l, k and b, $W_{t,k,n}$ is the (k,n)–element of $\mathbf{W}_t$, and $B_{T-t,b}$ is the $(T-t,b)$–element of matrix $\mathbf{B}$. It is easy to see from (13) that the weight $\omega_{n,t,l}$ attributed to the sentiment $s_{n,t,l}$ is equal to:

$$\omega_{n,t,l} = \sum_{k=1}^{K} \sum_{b=1}^{B} \widehat{\boldsymbol{\beta}}' \boldsymbol{e}_{l,k,b} \cdot W_{t,k,n} B_{T-t,b} \,, \tag{14}$$

such that:

$$\widehat{y}_T^h = \widehat{\alpha} + \widehat{\boldsymbol{\gamma}}' \mathbf{x}_T + \sum_{t=(T-\tau)}^{T} \sum_{n=1}^{N_t} \sum_{l=1}^{L} \omega_{n,t,l} \cdot s_{n,t,l} \,. \tag{15}$$

Clearly, it is unfeasible to analyze all (n,t,l)–combinations. We thus proceed by grouping them by common attributes, like time or topic. For example, to obtain the attribution of topic g $(1 \leq g \leq K)$, we can fix $k = g$ and compute the attribution by integrating the other dimensions:

$$a_g \equiv \sum_{t=(T-\tau)}^{T} \sum_{n=1}^{N_t} \sum_{l=1}^{L} \sum_{b=1}^{B} \widehat{\boldsymbol{\beta}}' \boldsymbol{e}_{l,g,b} \cdot W_{t,g,n} B_{T-t,b} \cdot s_{n,t,l} \,. \tag{16}$$

3. Application to forecasting US economic growth

We illustrate the complete optimized sentiment calibration framework to forecast economic growth in the United States. Our corpus consists of all articles published in major US news-

[6] In the DM approach, it is standard to implement the test statistic with an heteroscedasticity and autocorrelation robust (HAC) standard error estimator, such as in Andrews (1991) and Andrews and Monahan (1992), while the MCS approach relies on a (block–) bootstrap estimator for the variance.

papers documents available in *LexisNexis*.[7] We quantify the economic value of the sentiment calibration by evaluating the forecasting gains compared to benchmark approaches that use only the readily available quantitative macroeconomic and financial information in the merged datasets of McCracken and Ng (2016) and Goyal and Welch (2008). We first introduce the data and the models that we compare. We then present our main results and interpret the attribution that we obtain.

3.1. Data and descriptive statistics

3.1.1. Quantitative data

We aim at forecasting the log–growth in the US industrial production at the one–month ($h = 1$), three–month ($h = 3$), six–month ($h = 6$), nine–month ($h = 9$), and twelve–month ($h = 12$) horizons. We transform the level of industrial production into the h–month log–growth in percentage points: $y_t^h \equiv 100 \times (\ln IP_{t+h} - \ln IP_t)$, where IP_t is the industrial production realized at time t. Figure 2 presents the industrial production time series from January 1996 to December 2016.

[Insert Figure 2 about here.]

The workhorse approach to forecasting economic growth is the factor model proposed by Stock and Watson (2002). It consists of predicting economic growth using the most important principal components of a large panel of macroeconomic variables. We thus retrieve all economic related time series of in the FRED–MD historical vintage databases for every month from August 1999 to December 2016 (see McCracken and Ng, 2016). For vintages before August 1999, we use the data as of August 1999. FRED–MD is a large publicly available database of economic variables that satisfy the filtering criteria established in Stock and Watson (1996). The number of variables contained in the databases ranges from 105 to 128 for our time period. These variables are divided into various categories; see Table A.7 of the Appendix for an example with the FRED–MD 2016–12 dataset. Using past vintages allows us to get rid of the look–ahead bias.[8]

[7] *LexisNexis* provides an easy way to search and collect relevant news from over 26,000 news sources including online content. Their SmartIndexing™ technology classifies each text for a wide range of meta–information such as subject, company, person, and country, thus simplifying the collection process and reducing the chance of false positive inclusion of news in the dataset or in a particular subject. More information can be found at `https://www.nexis.com`.

[8] Macroeconomic FRED–MD data are available from Michael McCracken's website at `https://research.stlouisfed.org/econ/mccracken/fred-databases`.

In addition to the macroeconomic variables, we also consider financial indicators. We use the dataset of Goyal and Welch (2008) which consists of 16 financial metrics such as dividend ratios, long/short term yields, stock variances, etc. We add to this dataset the Chicago Board of Exchange's forward–looking volatility index (VIX).[9] Finally, we add to the list of variables the media–attention EPU index as well as six survey–based Conference Board indices (CB).[10] We apply standard transformations to render the variables stationary; see Table A.7 of the Appendix for details.

3.1.2. Qualitative data – corpus

To compute textual sentiment indices for the US, we retrieve the set of news consisting of all English articles from "Major US Newspapers" in the *LexisNexis* database with reference to the US. The *LexisNexis* "Major US Newspapers" source category is composed of the Daily News, Journal of Commerce, Los Angeles Times, Orange County Register, Pittsburgh Post–Gazette, St. Louis Post–Dispatch, Star Tribune, Tampa Bay Times, Atlanta Journal–Constitution, Christian Science Monitor, Daily Oklahoman, New York Post, New York Times, Philadelphia Daily News, Philadelphia Inquirer, Tampa Tribune, Washington Post, and USA Today. Dates range from January 1, 1994, to December 31, 2016. We apply the following filters:

- We use the geographic location such that we select only news relevant to the US (relevance score greater or equal to 85 in *LexisNexis*).
- We use the topic filter and filter out non–economic related topics.
- To be assigned to a topic, the news must have a major reference to the topic (relevance score greater or equal to 85 in *LexisNexis*).
- Article must have at least 200 words.

Table 1 presents the topics selected, the number of documents associated with them and a cluster categorization of each topic for a cluster–based attribution analysis. The final corpus amounts to a total of 338,408 articles and 44 topics over six clusters. The six clusters of topics, which have been manually constructed by identifying economic concepts that are closely related, are

[9] Financial data are available from Amit Goyal's website at `http://www.hec.unil.ch/agoyal` and VIX data from the Federal Reserve Bank of St–Louis at `https://fred.stlouisfed.org/series/VIXCLS`.

[10] EPU data are available from `http://www.policyuncertainty.com` and CB data from `https://www.conference-board.org/data/consumerconfidence.cfm`. The CB data include the leading economic index, the coincident economic index, the lagging economic index, the employment trend index, the consumer confidence: present situation index, and the consumer confidence: expectations index.

namely: “GDP Output”, “Job Market”, “Prices & Interest Rate”, “Real Estate”, “Surveys”, and “Others”. The latter is composed of the remaining topics. Note that a news article might refer to more than one topic as the average number of topics per article is 1.50.[11]

3.1.3. Qualitative data – sentiment calculation

We use standard lexicon–based sentiment analysis to measure the textual sentiment. The fundamental of lexicon–based sentiment analysis (also referred to as the bag–of–words approach) is the qualification of linguistic patterns (*e.g.*, words or sentences) as positive, negative, or neutral using predefined lists called lexicons. Most studies use the Harvard General Inquirer lexicon (2,550 positive words and 3,695 negative words).[12] This dictionary is built independently of any particular narrative text and may not be the most suitable choice for text analysis of the economic domain. For the analysis of financial and economic discourses this implies the use of specialized financial dictionaries, such as those developed by Henry (2008) (105 positive words and 85 negative words) and Loughran and McDonald (2011) (354 positive words and 2,355 negative words).[13] We also use four lexicons that are popular in the sentiment analysis literature: (i) the SentiWordNet lexicon of Baccianella et al. (2016) (8,898 positive words and 11,029 negative words), (ii) the SenticNet lexicon of Cambria et al. (2016) (11,775 positive words and 11,852 negative words), (iii) the SO–CAL lexicon of Taboada et al. (2011) (1,643 positive words and 1,647 negative words), and (iv) the NRC lexicon of Mohammad and Turney (2010) (2,227 positive words and 3,241 negative words).[14]

Another aspect of sentiment analysis is valence–shifting words (see Polanyi and Zaenen, 2006). Valence–shifting words are words such as *“very”* or *“barely”* that affect the context of nearby words. We only consider words that deal with negativity by inverting the sentiment of the first word following it from positive to negative and vice versa.[15]

Once the list of positive and negative words is established, we then calculate the (net) sentiment of each text document as the relative spread between the number of positive and

[11] *LexisNexis* does not provide within text topic identification, making it impossible to identify which part of the text discusses which topic. Ideally, one would have a single topic per text allowing for non–contaminated sentiment indices.

[12] The Harvard General Inquire lexicon is available at `http://www.wjh.harvard.edu/~inquirer/homecat.htm`.

[13] The Loughran & McDonald lexicon is available at `https://sraf.nd.edu/textual-analysis/resources`.

[14] The four lexicons are available through the R package **lexicons** (Rinker, 2018). SentiWordNet, SenticNet, and SO–CAL are weighted lexicons, where words are weighted according to their degree of positiveness or negativeness.

[15] The list of negative valence–shifting words considered is: *ain't, aren't, can't, couldn't, didn't, doesn't, don't, hasn't, isn't, mightn't, mustn't, neither, never, no, nobody, nor, not, shan't, shouldn't, wasn't, weren't, won't, wouldn't.*

negative words:

$$s_{n,t,l} \equiv \frac{N^{+}_{n,t,l} - N^{-}_{n,t,l}}{N^{+}_{n,t,l} + N^{-}_{n,t,l} + N^{0}_{n,t,l}}\,, \tag{17}$$

where $N^{+}_{n,t,l}$ is the number of positive words each word in document n at day t for lexicon l, $N^{-}_{n,t,l}$ is the number of negative words, and $N^{0}_{n,t,l}$ is the number of neutral words. $N^{+}_{n,t,l}$ and $N^{-}_{n,t,l}$ can also be defined as the sum of the positive and negative score respectively in case where the lexicon weights the words according to the degree of positiveness and negativeness in contrast to classifying them as positive or negative.[16] This use of the net sentiment measure, computed as the difference in the frequency of the positive words (positive sentiment) and the frequency of the negative words (negative sentiment) normalized by the total number of words, is widespread in the literature (see, *e.g.*, Arslan-Ayaydin et al., 2016, and the references therein). In our application, we use the net sentiment measure from seven lexicons, thus leading to $L = 7$ sentiment calculation methods.

Figure 3 presents the yearly (standardized) averages of the individual news article sentiments computed with the seven lexicons individually.[17] First, we see that the time–variation of the seven lexicon–based sentiment averages coincides with the economic cycle. In particular, we observe a large common drop during the dot–com bubble burst of 2001 and the financial crisis of 2008. These events are preceded by a large, almost linear, increase in the yearly average. In addition to the common behavior of the seven lexicon–based indices, we also observe cross–sectional variability. The cross–sectional variability is to be expected as no single lexicon offers a perfect estimate of the sentiment embedded in the text and the words classified as positive and negative in each lexicon differ. We are thus reducing the risk of selecting the wrong lexicon simply by the reasoning that if no cross–sectional variation was observed, the choice of the lexicon would be irrelevant.

[Insert Figure 3 about here.]

3.1.4. Qualitative data – aggregation of sentiment

We build the aggregation matrices $\mathbf{W}_t$ $(t = 1, \ldots, T)$ such that each of the 44 topics is summarized by a sentiment index. The time–series aggregation matrix $\mathbf{B}$ contains Beta weights generated from the grid $\{1, 3, 4, 7\} \times \{1, 3, 4, 7\}$ for a total of 16 time–aggregation weights; see

[16] This is, for example, the case for the SentiWordNet, SenticNet, and SO-CAL lexicons.

[17] Sentiment values are standardized (*i.e.*, we subtracted the mean and divided the series by their standard deviation) for readability purposes.

Figure 4. We set the value $\tau = 180$ days. This gives a total of $P = LKB = 7 \times 44 \times 16 = 4,928$ sentiment indices.

[Insert Figure 4 about here.]

Figure 5 presents the yearly average of the 44 topic–based sentiment indices calculated with the Loughran & McDonald lexicon.[18] Similarly to the yearly average of the non–aggregated sentiment shown in Figure 3, we see a general decrease in all sentiment indices during the years 2001 and 2008. We also note a significant variability in the cross–section of the yearly averages. This indicates that each topic has possibly different informational contents. Therefore, not considering the topic dimension by simply letting all news be part of an overarching topic could be sub–optimal, as we would lose important cross–sectional information.

[Insert Figure 5 about here.]

3.2. Models

The forecasting models that we consider are nested in the linear framework (7). The benchmark models $\mathcal{M}_{1a}$ and $\mathcal{M}_{2a}$ include the lagged value of the dependent variable and the macroeconomic, survey–based, and financial indicators ($\mathbf{x}_t$), or factors derived from those variables ($\mathbf{f}_t$). The alternative specifications $\mathcal{M}_{1b}$ and $\mathcal{M}_{2b}$ include in addition the 4,928 textual–based sentiment indices ($\mathbf{s}_t$).

More precisely, we study the following specifications:

$$\mathcal{M}_{1a}: \qquad y_t^h = \alpha_0^h + \alpha_1^h y_{t-h}^h + (\boldsymbol{\gamma}^h)' \mathbf{x}_t + \qquad + \varepsilon_t^h \tag{18}$$

$$\mathcal{M}_{1b}: \qquad y_t^h = \alpha_0^h + \alpha_1^h y_{t-h}^h + (\boldsymbol{\gamma}^h)' \mathbf{x}_t + (\boldsymbol{\beta}^h)' \mathbf{s}_t + \varepsilon_t^h \tag{19}$$

and:

$$\mathcal{M}_{2a}: \qquad y_t^h = \alpha_0^h + \alpha_1^h y_{t-h}^h + (\boldsymbol{\gamma}^h)' \mathbf{f}_t + \qquad + \varepsilon_t^h \tag{20}$$

$$\mathcal{M}_{2b}: \qquad y_t^h = \alpha_0^h + \alpha_1^h y_{t-h}^h + (\boldsymbol{\gamma}^h)' \mathbf{f}_t + (\boldsymbol{\beta}^h)' \mathbf{s}_t + \varepsilon_t^h \tag{21}$$

for $t = 1, \ldots, T$ months where $\mathbf{f}_t$ are factors extracted from $\mathbf{x}_t$ using the IC_{p1} criterion of Bai and Ng (2002). This criterion performs well compared to the other candidate information criteria in

[18] We observe the same pattern for other lexicons.

the various Monte Carlo experiments of Bai and Ng (2002). More detail about the construction of the factors is described in Appendix A.1.[19] Note that we are now dealing with a monthly frequency as opposed to the daily frequency used in the construction of the sentiment indices.

All models are estimated using the elastic net procedure in (8). We enforce the inclusion of the lagged dependent variable in the model specification and therefore exclude it from the penalization of the elastic net. Each model is estimated on a rolling window basis of 60 months.

Because of the overlapping nature of y_t^h when $h > 1$, we evaluate each model using the h–month–ahead observations. That is, if the sample window ranges from months $t = 1$ to $t = 60$, we evaluate the out–of–sample performance with the observation for month $t = 60 + h$.

Out–of–sample forecasting performance is evaluated using the RMSFE and MAFE measures. We evaluate $\mathcal{M}_{1b}$ ($\mathcal{M}_{2b}$) against $\mathcal{M}_{1a}$ ($\mathcal{M}_{2a}$) using the Diebold and Mariano (1995) test with the approach of Clark and McCracken (2001) for nested models at the 5% significance level.[20] To account for possible changes in out–of–sample forecasting performances over time, we analyze the full out–of–sample period and three sub–periods: pre–crisis, crisis, and post–crisis. The complete sample ranges from January 2001 (January 2003 for $h = 12$) to December 2016 (192 observations for $h = 1$ and 168 observations for $h = 12$). The pre–crisis period ranges from January 2001 (January 2003 for $h = 12$) to June 2007 (78 observations for $h = 1$ and 54 observations for $h = 12$). The crisis period ranges from July 2007 to December 2009 (30 observations). Finally, the post–crisis period ranges from January 2010 to December 2016 (84 observations).

3.3. Main results

3.3.1. Model's forecasting performance comparison

Table 2 presents the RMSFE and MAFE measures for the four model specifications and the five forecasting horizons over the four time–windows. We focus our analysis on comparing the added value of using sentiment information when forecasting economic growth controlling for readily available predictors, either used as raw inputs (*i.e.*, $\mathcal{M}_{1b}$ vs $\mathcal{M}_{1a}$) or through factors ($\mathcal{M}_{2b}$ vs $\mathcal{M}_{2a}$). A gray cell indicates that the outperformance is statistically significant according to the DM test at the 5% significance level.

[19] We justify the use of principal component in conjunction to the Bai and Ng (2002) information criterion by noting that this method has shown to perform well at forecasting the growth in the US industrial production in Smeekes and Wijler (2018) compared to more complex factor and penalized regression models.

[20] The bootstrapped distribution is computed using 5,000 block bootstrap samples with the optimal block length determined from the fit of an autoregressive model. The variance of the mean loss difference is computed using the HAC standard error estimator of Andrews (1991) and Andrews and Monahan (1992).

[Insert Table 2 about here.]

For the full sample, we see that textual sentiment–related specifications do not add forecasting power over the macroeconomic, surveys–based, and financial indicators at the one– to six–month horizons. However, at the nine– to twelve–month horizons, they exhibit the best performance and results are significant according to the DM test for both the RMSFE and MAFE measures.

This gain in outperformance as the forecasting horizon grows is also observed in Ulbricht et al. (2017) for news–derived economic sentiment indices in the context of forecasting the German industrial production. It is consistent with the "time–lag" effect in economics. While financial markets can react (quasi) instantaneously to the sentiment expressed in the economic news, it takes time for that sentiment to affect economic behaviors (consumption, production, investments) and thus to become visible in the published economic growth figures (see George et al., 1999). This may explain why the sentiment becomes more predictive for economic growth over longer horizons.

Looking at the pre–crisis period, we can observe that the textual–sentiment related specifications outperform their benchmark according to the DM test at the twelve–month horizon. The post–crisis period, however, shows outperformance for the nine– and twelve–month horizons. To the contrary of the other periods, sentiment–related specification only shows outperformance during the crisis period at the six–month and nine–month horizon and that is only according to the RMSFE measure.

Overall, we observe that textual–sentiment related specifications provide additional forecasting power over traditional macroeconomic, financial, and survey indicators at long–term horizons.

3.3.2. Attribution

A common criticism for big data approaches to economic forecasting is that their results seem to come from a "black box". In our setting, this criticism can be easily countered, since the attribution analysis described in Step 8 of Section 2 allows us to pinpoint the contribution of each sentiment value to the growth prediction. Given a large number of sentiment values, we can analyze the attribution at the intermediate level of the grouping per cluster of topics from the categorization shown in Table 1.

Figure 6 presents the normalized attribution of these clusters for the twelve–month forecasts

obtained with model $\mathcal{M}_{1b}$, where we divide each of its elements by the L^2–norm of the attribution vector at that date.[21] This procedure makes it easier to do comparison across different dates. Note first that there is a persistence in the attribution of each cluster over time. This is consistent with the presence of stable information value in the selection and weighting used when engineering the textual sentiment index for predicting economic growth. Over the full sample, we find that "GDP Output", "Price & Interest Rates" and "Survey" contribute the most the predicted growth when compare to the other clusters. They dominate the prediction at different times. In the pre–crisis period, texts published about "GDP Output" are the main predictors. During the crisis, the texts discussing the surveys are selected and weighted to have the biggest impact on the predictions. Finally, post–crisis, the "Price & Interest Rates"–related texts are dominating the predictions.

[Insert Figure 6 about here.]

3.4. Importance of the optimization of each dimension

We now proceed to analyze the impact of some of the modeling choices employed in our study.

We analyze to which extent the optimization of the lexicon–, topic– and time–dimensions are relevant in predicting the industrial production growth. To that aim, we compare the extended specifications $\mathcal{M}_{1b}$ and $\mathcal{M}_{2b}$ with four alternatives in which we (equally–weighted) aggregate: (i) the lexicon–dimension (denoted LEX), (ii) the topic–dimension (denoted TOPIC), (iii) the time–dimension (denoted TIME), and (iv) all dimensions (denoted ALL). The last approach is, therefore, the naive way of calculating a sentiment index and adding it to the set of macroeconomic, surveys, and financial variables. Note that these dimension reductions are only special cases of the methodology. Results are reported in Table 3 for the full out–of–sample period. We can observe that, on a lower RMSFE and MAFE basis, the optimization of all dimension is preferable. This is principally the case at the nine–month and particularly the twelve–month horizons. The time dimension seems to be the most important to optimize, followed by the topic and the lexicon dimensions.

[Insert Table 3 about here.]

[21] Results for model $\mathcal{M}_{2b}$ are similar and available from the authors upon request.

4. Conclusion

Do textual sentiment indices provide any added value to the prediction accuracy of economic growth when compared to the use of information contained in macroeconomic, financial, or survey–based variables? To answer this question, one needs to first capture the relevant sentiment–based growth prediction from a textual analysis of news releases. The latter is a big data problem, given the large number of texts published every day, the number of possible historical dates at which news releases may have predictive value for the future economic activity, and the various methods of calculating sentiment. We show how to overcome this dimensionality issue by introducing a framework that optimizes sentiment aggregation for predicting economic growth using both topics–based aggregation, time–series aggregation, and predictive regressions using the elastic net regularization.

We test the predictive power of text–based sentiment indices by forecasting the growth in US industrial production using major newspapers from the news database *LexisNexis* over the period January 2001 to December 2016. We find that the proposed optimized text–based sentiment analysis can significantly improve the forecasting performance for predicting the nine–month and annual growth rates.

To help practitioners and academics implementing our methodology in practice, we have released the open–source R package **sentometrics** (Ardia et al., 2018, 2017). The package is designed in a way that each step of the methodology, from sentiment calculation to time–series aggregation, can be configured for specific needs. It thus not only allows one to replicate the configuration used in our empirical application but also allows for extensions and modifications.

The scope of applications of the proposed optimized textual sentiment analysis framework goes beyond forecasting economic growth. In future work, we will consider applying the framework to quantify brand reputation when forecasting firm sales, and studying spillover effects between types of news media.

References

Alessi, L., Barigozzi, M., Capasso, M., 2010. Improved penalization for determining the number of factors in approximate factor models. Statistics & Probability Letters 80, 1806–1813. doi:10.1016/j.spl.2010.08.005.

Andrews, D.W.K., 1991. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. Econometrica 59, 817–858. doi:10.2307/2938229.

Andrews, D.W.K., Monahan, J.D., 1992. An improved heteroskedasticity and autocorrelation consistent covariance matrix estimation. Econometrica 60, 953–966. doi:10.2307/2951574.

Ardia, D., Bluteau, K., Borms, S., Boudt, K., 2017. The R package **sentometrics** to compute, aggregate and predict with textual sentiment. doi:10.2139/ssrn.3067734. working paper.

Ardia, D., Bluteau, K., Borms, S., Boudt, K., 2018. **sentometrics**: An integrated framework for textual sentiment time series aggregation and prediction. URL: https://CRAN.R-project.org/package=sentometrics. version 0.4.

Arslan-Ayaydin, Ö., Boudt, K., Thewissen, J., 2016. Managers set the tone: Equity incentives and the tone of earnings press releases. Journal of Banking & Finance 72, S132–S147. doi:10.1016/j.jbankfin.2015.10.007.

Baccianella, S., Esuli, A., Sebastiani, F., 2016. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining, in: LREC, 2200–2204.

Bai, J., Ng, S., 2002. Determining the number of factors in approximate factor models. Econometrica 70, 191–221. doi:10.1111/1468-0262.00273.

Bai, J., Ng, S., 2008. Forecasting economic time series using targeted predictors. Journal of Econometrics 146, 304–317. doi:10.1016/j.jeconom.2008.08.010.

Baker, S.R., Bloom, N., Davis, S.J., 2016. Measuring economic policy uncertainty. Quarterly Journal of Economics 131, 1593–1636. doi:10.1093/qje/qjw024.

Bram, J., Ludvigson, S., 1998. Does consumer confidence forecast household expenditure? A sentiment index horse race. Economic Policy Review 4, 59–78.

Bräuning, F., Koopman, S.J., 2014. Forecasting macroeconomic variables using collapsed dynamic factor analysis. International Journal of Forecasting 30, 572–584. doi:10.1016/j.ijforecast.2013.03.004.

Cambria, E., Poria, S., Bajpai, R., Schuller, B., 2016. SenticNet 4: A semantic resource for sentiment analysis based on conceptual primitives, in: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, 2666–2677.

Chen, J., Chen, Z., 2008. Extended Bayesian information criteria for model selection with large model spaces. Biometrika 95, 759–771. doi:10.1093/biomet/asn034.

Clark, T.E., McCracken, M.W., 2001. Tests of equal forecast accuracy and encompassing for nested models. Journal of Econometrics 105, 85–110. doi:10.1016/S0304-4076(01)00071-9.

Diebold, F.X., Mariano, R.S., 1995. Comparing predictive accuracy. Journal of Business & Economic Statistics 13, 253–263. doi:10.1080/07350015.1995.10524599.

Doz, C., Giannone, D., Reichlin, L., 2011. A two-step estimator for large approximate dynamic factor models based on Kalman filtering. Journal of Econometrics 164, 188–205. doi:10.1016/j.jeconom.2011.02.012.

Doz, C., Giannone, D., Reichlin, L., 2012. A quasi–maximum likelihood approach for large, approximate dynamic factor models. Review of Economics and Statistics 94, 1014–1024. doi:10.1162/REST_a_00225.

Espinoza, R., Fornari, F., Lombardi, M.J., 2012. The role of financial variables in predicting economic activity. Journal of Forecasting 31, 15–46. doi:10.1002/for.1212.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33, 1–22. doi:10.18637/jss.v033.i01.

Gelper, S., Croux, C., 2010. On the construction of the European economic sentiment indicator. Oxford Bulletin of Economics and Statistics 72, 47–62. doi:10.1111/j.1468-0084.2009.00574.x.

George, E., King, M., Clementi, D., Budd, A., Buiter, W., Goodhart, C., Julius, D., Plenderleith, I., Vickers, J., 1999. The transmission mechanism of monetary policy. Bank of England.

Ghysels, E., Sinko, A., Valkanov, R., 2007. MIDAS regressions: Further results and new directions. Econometric Reviews 26, 53–90. doi:10.1080/07474930600972467.

Goyal, A., Welch, I., 2008. A comprehensive look at the empirical performance of equity premium prediction. Review of Financial Studies 21, 1455–1508. doi:10.1093/rfs/hhm014.

Hansen, P.R., Lunde, A., Nason, J.M., 2011. The model confidence set. Econometrica 79, 453–497. doi:10.3982/ECTA5771.

Henry, E., 2008. Are investors influenced by how earnings press releases are written? Journal of Business Communication 45, 363–407. doi:10.1177/0021943608319388.

Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12, 55–67. doi:10.1080/00401706.1970.10488634.

Kim, H.H., Swanson, N.R., 2014. Forecasting financial and macroeconomic variables using data reduction methods: New empirical evidence. Journal of Econometrics 178, 352–367. doi:10.1016/j.jeconom.2013.08.033.

Kim, H.H., Swanson, N.R., 2018. Mining big data using parsimonious factor, machine learning, variable selection and shrinkage methods. International Journal of Forecasting 34, 339–354. doi:10.1016/j.ijforecast.2016.02.012.

Liu, L., Tang, L., Dong, W., Yao, S., Zhou, W., 2016. An overview of topic modeling and its current applications in bioinformatics. SpringerPlus 5, 1–22. doi:10.1186/s40064-016-3252-8.

Loughran, T., McDonald, B., 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. Journal of Finance 66, 35–65. doi:10.1111/j.1540-6261.2010.01625.x.

Ludvigson, S.C., 2004. Consumer confidence and consumer spending. Journal of Economic Perspectives 18, 29–50. doi:10.1257/0895330041371222.

McCracken, M.W., Ng, S., 2016. FRED-MD: A monthly database for macroeconomic research. Journal of Business & Economic Statistics 34, 574–589. doi:10.1080/07350015.2015.1086655.

Mohammad, S.M., Turney, P.D., 2010. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon, in: Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text, Association for Computational Linguistics. 26–34.

Polanyi, L., Zaenen, A., 2006. Computing Attitude and affect in text: Theory and applications. Springer-Verlag. volume 20 of *The Information Retrieval*. chapter Contextual Valence Shifters. 1–10. doi:10.1007/1-4020-4102-0.

Ravi, K., Ravi, V., 2015. A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. Knowledge-Based Systems 89, 14–46. doi:10.1016/j.knosys.2015.06.015.

Rinker, T.W., 2018. **lexicon**: Lexicon data. URL: http://github.com/trinker/lexicon. version 1.0.0.

Shapiro, A.H., Sudhof, M., Wilson, D., 2018. Measuring news sentiment. Working paper.

Smeekes, S., Wijler, E., 2018. Macroeconomic forecasting using penalized regression methods. International Journal of Forecasting 34, 408–430. doi:10.1016/j.ijforecast.2018.01.001.

Stock, J.H., Watson, M., 2011. Dynamic factor models. Oxford Handbook on Economic Forecasting doi:10.1093/oxfordhb/9780195398649.013.0003.

Stock, J.H., Watson, M.W., 1996. Evidence on structural instability in macroeconomic time series relations. Journal of Business & Economic Statistics 14, 11–30. doi:10.1080/07350015.1996.10524626.

Stock, J.H., Watson, M.W., 2002. Forecasting using principal components from a large number of predictors. Journal of the American Statistical Association 97, 1167–1179. doi:10.1198/016214502388618960.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M., 2011. Lexicon-based methods for sentiment analysis. Computational Linguistics 37, 267–307. doi:10.1162/COLI_a_00049.

Thorsrud, L.A., 2016. Nowcasting using news topics. Big data versus big bank. Working paper.

Thorsrud, L.A., 2018. Words are the new numbers: A newsy coincident index of the business cycle. doi:10.1080/07350015.2018.1506344. Forthcoming in Journal of Business & Economic Statistics.

Tibshirani, R., 1996. Regression shrinkage and selection via the LASSO. Journal of the Royal Statistical Society: Series B 58, 267–288. doi:10.1111/j.1467-9868.2011.00771.x.

Tibshirani, R.J., Taylor, J., 2012. Degrees of freedom in LASSO problems. Annals of Statistics 40, 1198–1232. doi:10.1214/12-AOS1003.

Tobback, E., Naudts, H., Daelemans, W., de Fortuny, E.J., Martens, D., 2018. Belgian economic policy uncertainty index: Improvement through text mining. International Journal of Forecasting, 34, 355–365. doi:10.1016/j.ijforecast.2016.08.006.

Ulbricht, D., Kholodilin, K.A., Thomas, T., 2017. Do media data help to predict German industrial production? Journal of Forecasting 36, 483–496. doi:10.1002/for.2449.

Wang, T., Zhu, L., 2011. Consistent tuning parameter selection in high dimensional sparse linear regression. Journal of Multivariate Analysis 102, 1141–1151. doi:10.1016/j.jmva.2011.03.007.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B 67, 301–320. doi:10.1111/j.1467-9868.2005.00503.x.

Zou, H., Hastie, T., Tibshirani, R., 2007. On the "degrees of freedom" of the LASSO. Annals of Statistics 35, 2173–2192. doi:10.1214/009053607000000127.

Zou, H., Zhang, H.H., 2009. On the adaptive elastic-net with a diverging number of parameters. Annals of Statistics 37, 1733–1751. doi:10.1214/08-AOS625.

Table 1: Total number of documents related to a given topic
This table presents the number of articles from Major US newspaper related to a given topic in the corpus. The list of topics is manually selected from the full list of topics identified by the *LexisNexis* SmartIndexing™ classifier, which provides a set of topics to each article in the database. Non–economic related topics have been removed resulting in a corpus that focuses exclusively on the US economy. Documents with less than 200 words are removed. Note that each article may be related to multiple topics. Topics are also organized into clusters of topics. The clusters are constructed manually and identified as: 1: GDP Output, 2: Job Market, 3: Prices & Interest Rate, 4: Real Estate, 5: Surveys, 6: Others.

Topic	#	Cluster	Topic	#	Cluster
ECONOMIC CONDITIONS	25,522	1	IMPORT TRADE	15,709	3
COMPANY EARNINGS	20,116	1	INTEREST RATES	14,018	3
RECESSION	15,907	1	PRICE INCREASES	12,233	3
COMPANY PROFITS	11,075	1	INFLATION	11,841	3
SALES FIGURES	8,051	1	CURRENCIES	10,281	3
ECONOMIC GROWTH	7,904	1	PRICE CHANGES	9,363	3
BUDGET DEFICITS	6,656	1	ECONOMIC POLICY	7,270	3
OUTPUT & DEMAND	6,200	1	BOND MARKETS	4,027	3
MANUFACTURING OUTPUT	4,924	1	COMMODITIES PRICES	1,264	3
ECONOMIC STIMULUS	3,798	1	DEBT CRISIS	841	3
GROSS DOMESTIC PRODUCT	3,541	1	HOUSING MARKET	14,296	4
ECONOMIC DECLINE	2,818	1	REAL ESTATE DEVELOPMENT	11,144	4
CONSUMPTION	530	1	HOME PRICES	10,133	4
WAGES & SALARIES	37,157	2	CONSUMER CONFIDENCE	3,623	5
EMPLOYMENT	23,993	2	ECONOMIC SURVEYS	963	5
EMPLOYMENT GROWTH	11,708	2	BUSINESS CLIMATE & CONDITIONS	790	5
UNEMPLOYMENT RATES	10,070	2	BUSINESS CONFIDENCE	75	5
JOB CREATION	7,846	2	RETAILERS	32,695	6
PRICES	49,207	3	OIL & GAS INDUSTRY	20,384	6
EXPORT TRADE	19,390	3	MANUFACTURING FACILITIES	12,889	6
OIL & GAS PRICES	17,784	3	UTILITY RATES	3,215	6
INTERNATIONAL TRADE	17,029	3	RETAIL SECTOR PERFORMANCE	896	6
Number of topics			44		
Number of articles			338,408		
Average number of topics per article			1.50		

Table 2: Forecasting results
This table presents the Root Mean Squared Forecast Error (RMSFE) and the Mean Absolute Forecast Error (MAFE) for model $\mathcal{M}_{1a}$ (*i.e.*, benchmark model with raw variables), $\mathcal{M}_{1b}$ (*i.e.*, $\mathcal{M}_{1a}$ augmented by textual sentiments), $\mathcal{M}_{2a}$ (*i.e.*, benchmark model with factors), and $\mathcal{M}_{2b}$ (*i.e.*, $\mathcal{M}_{2a}$ augmented by textual sentiments). Lower RMSFE and MAFE values are preferred. We consider the one– ($h = 1$), three– ($h = 3$), six– ($h = 6$), nine– ($h = 9$), and twelve–month ($h = 12$) log–growth in the US industrial production. The full out–of–sample period ranges from January 2001 (January 2003 for $h = 12$) to December 2016 (192 observations for $h = 1$ and 168 observations for $h = 12$). The out–of–sample pre–crisis period ranges from January 2001 to June 2007 (78 observations for $h = 1$ and 54 observations for $h = 12$). The out–of–sample crisis period ranges from July 2007 to December 2009 (30 observations). The out–of–sample post–crisis period ranges from January 2010 to December 2016 (84 observations). A gray cell indicates that the extended model is superior to the benchmark model (*i.e.*, $\mathcal{M}_{1b}$ against $\mathcal{M}_{1a}$ and $\mathcal{M}_{2b}$ against $\mathcal{M}_{2a}$) for a given horizon at the 5% significance level. Testing is based on the Diebold and Mariano (1995) test statistic implemented with the heteroscedasticity and autocorrelation robust (HAC) standard error estimators of Andrews (1991) and Andrews and Monahan (1992) and with p–values computed by bootstrap following Clark and McCracken (2001).

		RMSFE				MAFE			
Period	h	$\mathcal{M}_{1a}$	$\mathcal{M}_{1b}$	$\mathcal{M}_{2a}$	$\mathcal{M}_{2b}$	$\mathcal{M}_{1a}$	$\mathcal{M}_{1b}$	$\mathcal{M}_{2a}$	$\mathcal{M}_{2b}$
	1	0.68	0.70	0.64	0.70	0.49	0.49	0.45	0.49
	3	1.52	1.54	1.59	1.52	0.96	1.01	1.02	1.01
Full sample	6	4.86	3.93	5.01	3.14	2.36	2.35	2.85	2.14
	9	7.01	4.95	8.36	4.58	3.71	3.28	4.89	3.19
	12	6.39	5.19	8.69	5.14	4.25	3.41	6.03	3.32
	1	0.55	0.57	0.56	0.56	0.43	0.42	0.43	0.44
	3	0.99	0.93	1.21	0.93	0.72	0.70	0.87	0.70
Pre–crisis	6	1.67	1.65	2.62	1.62	1.31	1.36	1.80	1.32
	9	2.41	2.42	4.67	2.53	1.96	1.93	3.00	1.98
	12	3.27	2.00	6.07	1.90	2.72	1.67	3.73	1.57
	1	1.19	1.27	1.08	1.27	0.81	0.87	0.69	0.88
	3	3.20	3.19	3.17	3.04	2.46	2.52	2.31	2.29
Crisis	6	11.30	8.54	10.64	6.20	7.63	6.44	7.45	4.99
	9	8.58	7.94	9.92	7.94	6.67	6.20	7.67	6.20
	12	10.43	10.14	9.42	10.12	8.34	7.84	7.49	7.70
	1	0.53	0.50	0.49	0.50	0.42	0.40	0.40	0.41
	3	0.78	0.93	0.89	1.03	0.62	0.74	0.70	0.82
Post–crisis	6	1.72	2.26	2.86	2.32	1.32	1.68	2.05	1.77
	9	8.47	4.93	9.72	4.07	3.93	3.22	5.27	3.00
	12	6.02	3.85	9.81	3.78	3.80	2.98	7.01	2.90

Table 3: Robustness results – Aggregation of dimensions
This table presents the forecasting results when the various dimensions (lexicon, topic, and time) are aggregated. We compare the results of the extended models $\mathcal{M}_{1b}$ and $\mathcal{M}_{2b}$ with four alternative approaches in which we (equally–weighted) aggregate: (i) the lexicon–dimension (denoted LEX), (ii) the topic–dimension (denoted TOPIC), (iii) the time–dimension (denoted TIME), and (iv) all dimensions (denoted ALL). A light (dark) gray cell indicates that the extended model ($\mathcal{M}_{1b}$ or $\mathcal{M}_{2b}$) is superior (inferior) according to the Diebold and Mariano (1995) test statistic at the 5% significance. See Table 2 for details.

		RMSFE					MAFE				
	h	$\mathcal{M}$	LEX	TOPIC	TIME	ALL	$\mathcal{M}$	LEX	TOPIC	TIME	ALL
$\mathcal{M}_{1b}$	1	0.70	0.69	0.68	0.68	0.64	0.49	0.48	0.48	0.49	0.46
	3	1.54	1.50	1.41	1.52	1.58	1.01	0.98	0.93	0.96	0.99
	6	3.93	4.51	4.52	4.86	5.24	2.35	2.42	2.32	2.36	2.55
	9	4.95	5.91	5.57	7.01	8.37	3.28	3.43	3.28	3.71	4.17
	12	5.19	5.85	6.11	6.39	8.24	3.41	4.01	4.09	4.25	5.02
$\mathcal{M}_{2b}$	1	0.70	0.69	0.68	0.65	0.68	0.49	0.49	0.48	0.46	0.48
	3	1.52	1.53	1.50	1.39	1.31	1.01	1.06	1.07	0.95	0.92
	6	3.14	3.72	3.23	3.62	3.34	2.14	2.39	2.17	2.25	2.20
	9	4.58	5.65	5.36	6.99	6.23	3.19	3.74	3.42	4.16	4.06
	12	5.14	6.79	7.14	8.18	7.82	3.32	4.81	5.04	5.30	5.34

Figure 1: Methodology
This figure presents a scheme of the nine steps of the building blocks of the methodology.

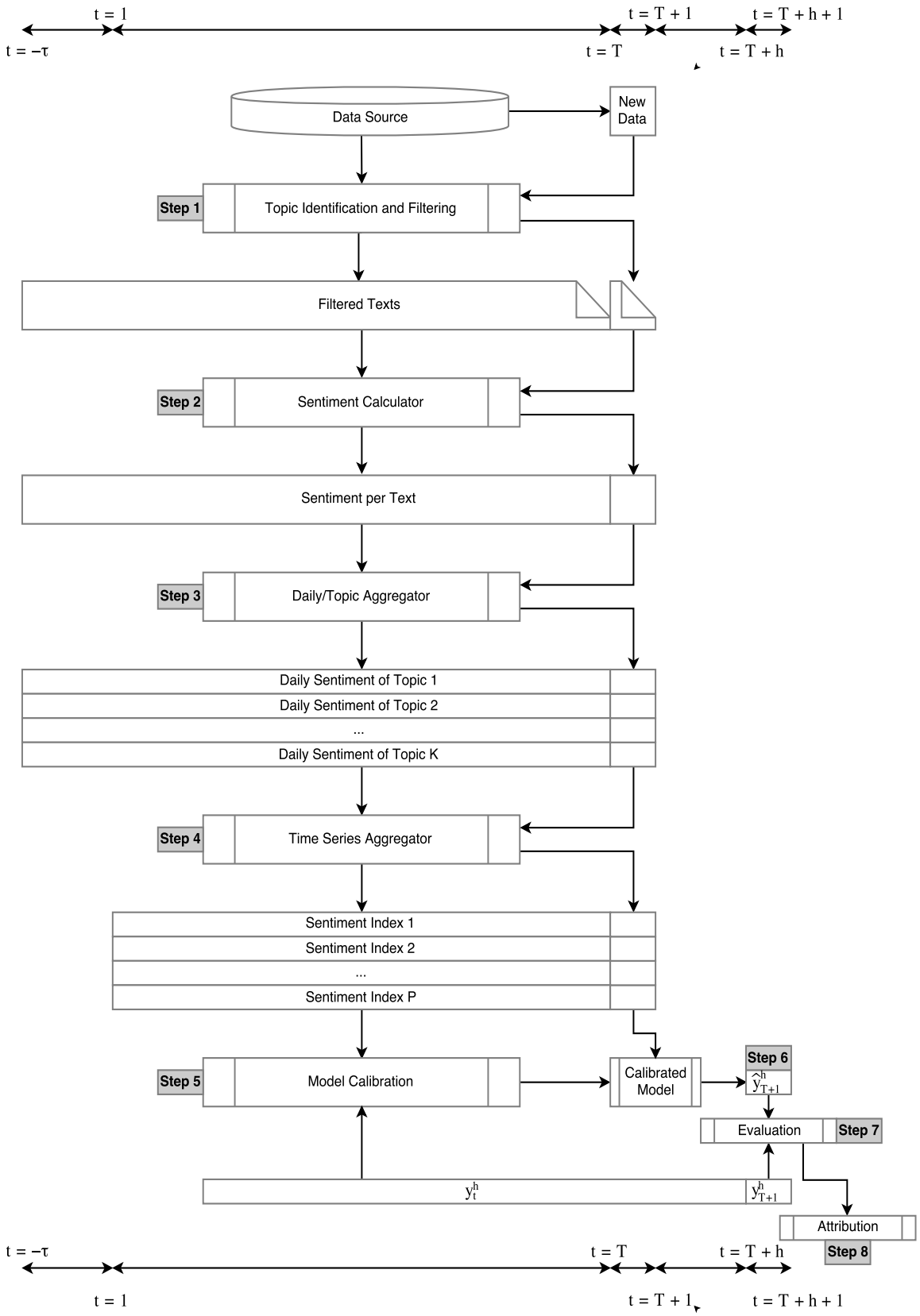


Figure 2: US Industrial production
This figure presents the US industrial production from January 1996 to December 2016 (192 monthly observations). The gray zone indicates the crisis period, which spans from July 2007 to December 2009 (30 months).

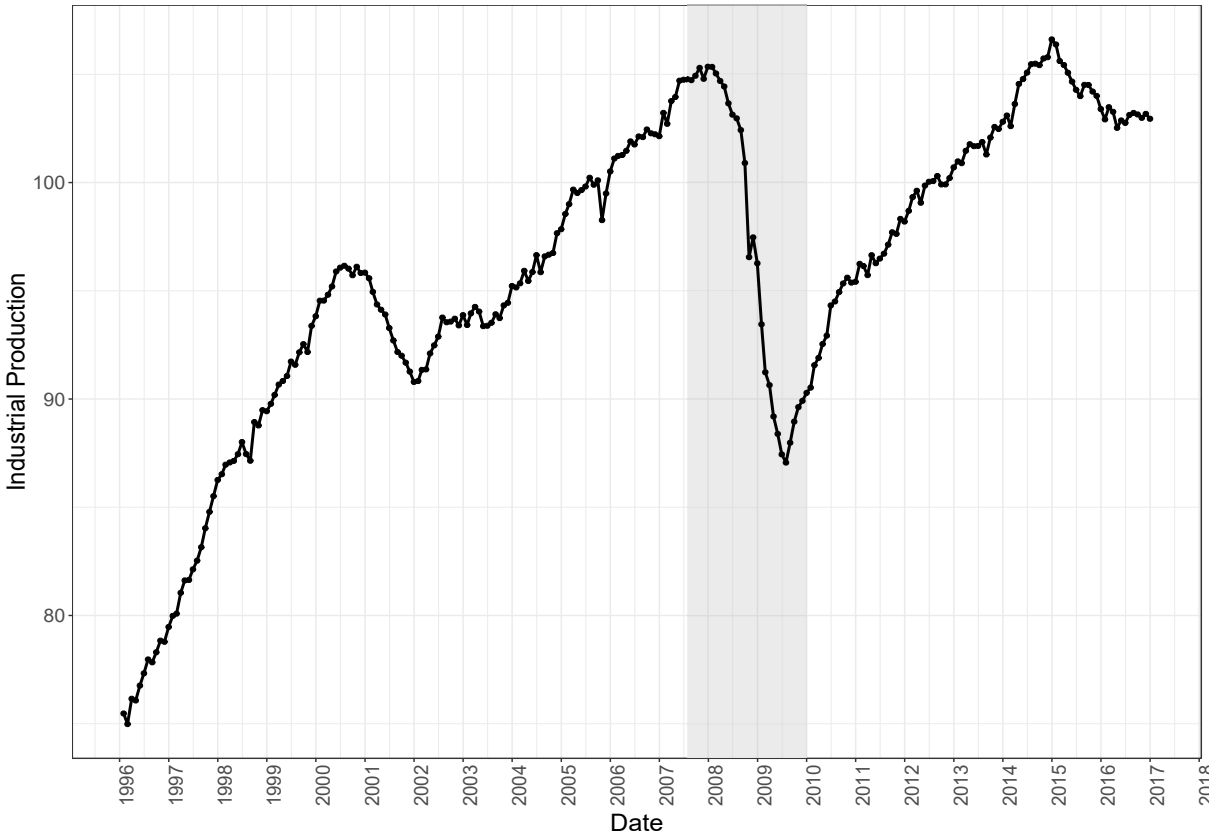


Figure 3: Yearly lexicon–based averages of the individual news articles' sentiments
This figure presents the seven lexicon–based yearly averages of the individual news articles sentiment for the period ranging from 1994 to 2016. Sentiment values are standardized for readability purposes. The gray zone indicates the 2007 to 2009 crisis period.

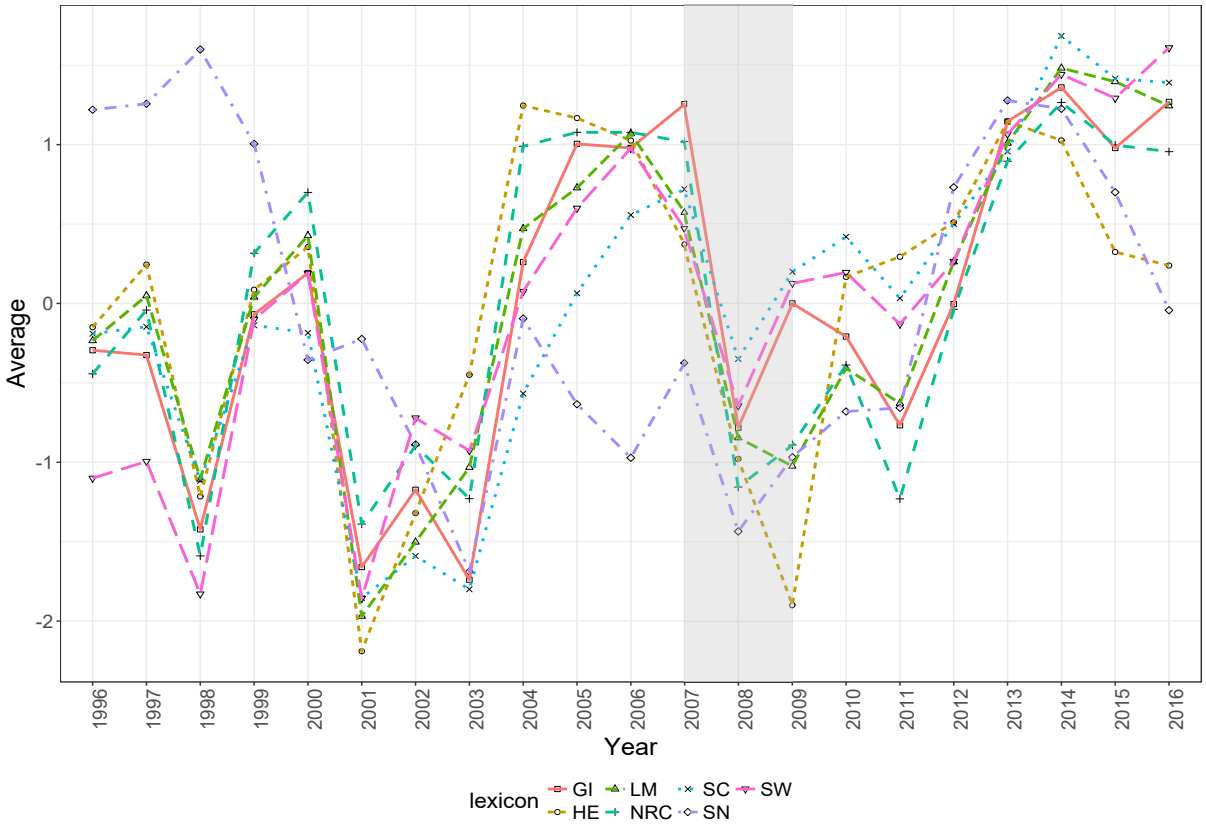


Figure 4: Beta weights
This figure presents the time–aggregation weights of the Beta function for the grid $\{1, 3, 4, 7\} \times \{1, 3, 4, 7\}$ for a total of 16 weighting schemes.

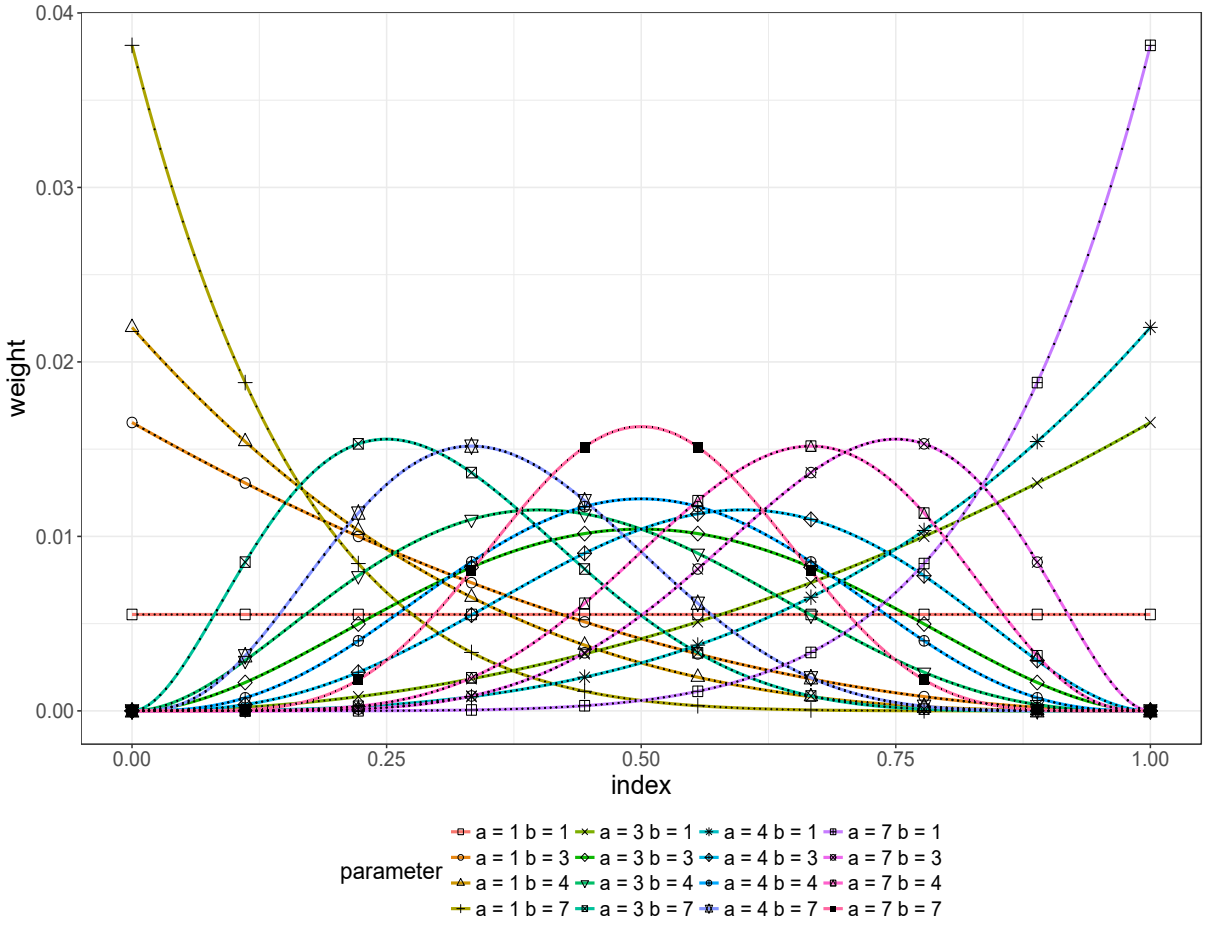
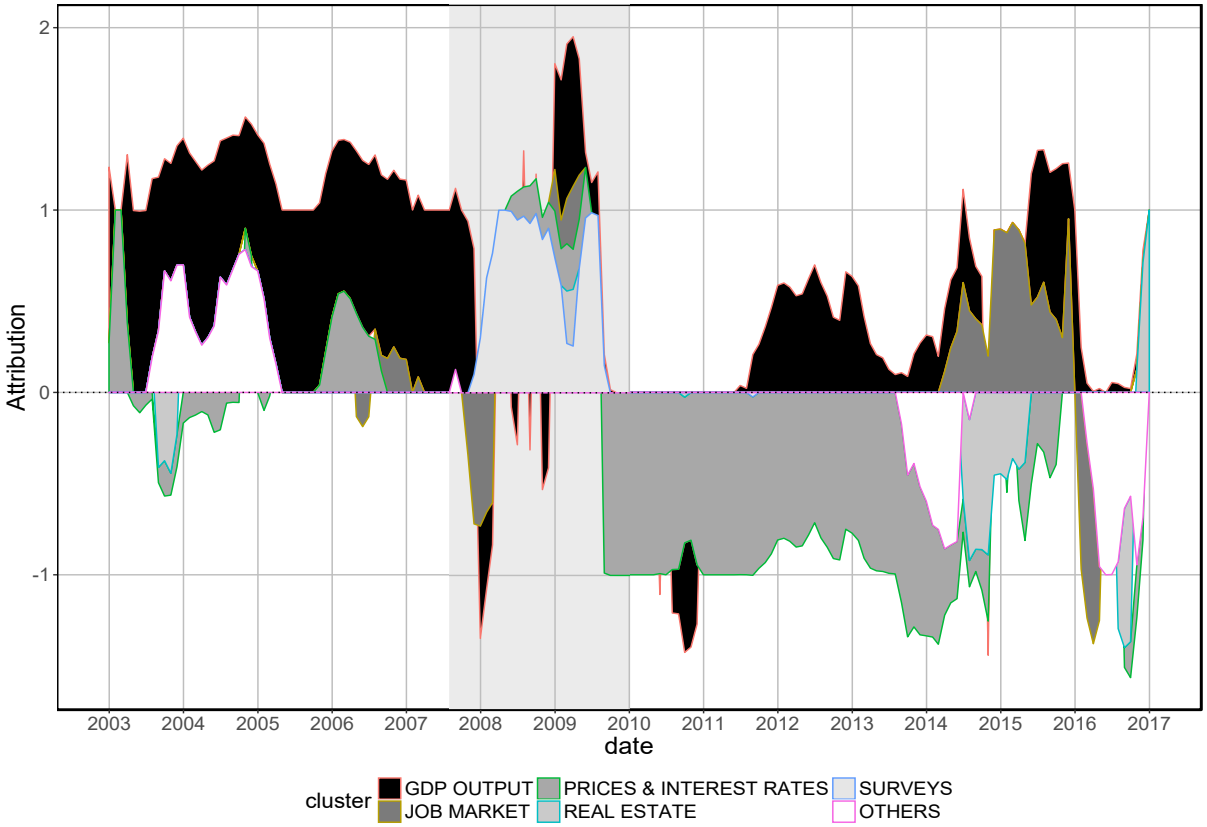


Figure 5: Yearly average of the 44 topic sentiment indices
This figure presents the yearly average of 44 sentiment indices for the period ranging from 1996 to 2016. Sentiment values are computed using the Loughran and McDonald (2011) lexicon. Each time series is standardized for sake of comparability across topics. Topics are organized into clusters on the y–axis and delimited by black and gray text labeling. Black cases indicate that there is no news for that particular topic during that year.

Figure 6: Forecast attribution
This figure presents the cluster attribution of model $\mathcal{M}_{1b}$ for the out–of–sample forecasts of the twelve–month US industrial production log–growth. The period ranges from January 2003 to December 2016 (180 monthly observations). The attribution vector for a given date is scaled by dividing each element of the attribution vector by the L^2–norm of the attribution vector for that date. The gray zone indicates the July 2007 to December 2009 crisis period. A positive (negative) value indicates that the topic contributes positively (negatively) to the forecast and therefore increases (decreases) the forecast of the US industrial production log–growth.



Appendix A. Appendix

Table A.7: List of variables

This table summarizes the macroeconomic, financial, and additional media–attention and survey–based variables used in our study. The column "Code" refers to one of the following data transformations for a time series: 1: no transformation, 2: level–difference, 3: second level–difference, 4: log, 5: log–difference, 6: second log–difference, 7: growth rate. FRED–MD vintage datasets (Groups 1–8) are available from `https://research.stlouisfed.org/econ/mccracken/fred-databases`, financial variables (Group 9) from `http://www.hec.unil.ch/agoyal`, VIX from `https://fred.stlouisfed.org/series/VIXCLS`, EPU index from `http://www.policyuncertainty.com`, and Chicago conference board indices from `https://www.conference-board.org/data/consumerconfidence.cfm` (all in Group 10).

Index	Code	Variable	Description
Group 1: Output and Income			
1	5	RPI	Real Personal Income
2	5	W875RX1	Real personal income ex transfer receipts
3	5	INDPRO	IP Index
4	5	IPFPNSS	IP: Final Products and Nonindustrial Supplies
5	5	IPFINAL	IP: Final Products (Market Group)
6	5	IPCONGD	IP: Consumer Goods
7	5	IPDCONGD	IP: Durable Consumer Goods
8	5	IPNCONGD	IP: Nondurable Consumer Goods
9	5	IPBUSEQ	IP: Business Equipment
10	5	IPMAT	IP: Materials
11	5	IPDMAT	IP: Durable Materials
12	5	IPNMAT	IP: Nondurable Materials
13	5	IPMANSICS	IP: Manufacturing (SIC)
14	5	IPB51222s	IP: Residential Utilities
15	5	IPFUELS	IP: Fuels
16	1	NAPMPI	ISM Manufacturing: Production Index
17	2	CUMFNS	Capacity Utilization: Manufacturing
Group 2: Consumption and Order			
1	2	HWI	Help-Wanted Index for United States Help wanted indx
2	2	HWIURATIO	Ratio of Help Wanted/No. Unemployed
3	5	CLF16OV	Civilian Labor Force
4	5	CE16OV	Civilian Employment
5	2	UNRATE	Civilian Unemployment Rate
6	2	UEMPMEAN	Average Duration of Unemployment (Weeks)
7	5	UEMPLT5	Civilians Unemployed - Less Than 5 Weeks
8	5	UEMP5TO14	Civilians Unemployed for 5-14 Weeks
9	5	UEMP15OV	Civilians Unemployed - 15 Weeks & Over
10	5	UEMP15T26	Civilians Unemployed for 15-26 Weeks
11	5	UEMP27OV	Civilians Unemployed for 27 Weeks and Over
12	5	CLAIMSx	Initial Claims
13	5	PAYEMS	All Employees: Total nonfarm
14	5	USGOOD	All Employees: Goods-Producing Industries
15	5	CES1021000001	All Employees: Mining and Logging: Mining
16	5	USCONS	All Employees: Construction
17	5	MANEMP	All Employees: Manufacturing
18	5	DMANEMP	All Employees: Durable goods
19	5	NDMANEMP	All Employees: Nondurable goods
20	5	SRVPRD	All Employees: Service-Providing Industries
21	5	USTPU	All Employees: Trade, Transportation & Utilities
22	5	USWTRADE	All Employees: Wholesale Trade
23	5	USTRADE	All Employees: Retail Trade
Group 3: Housing			
1	4	HOUST	Housing Starts: Total New Privately Owned
2	4	HOUSTNE	Housing Starts, Northeast
3	4	HOUSTMW	Housing Starts, Midwest
4	4	HOUSTS	Housing Starts, South
5	4	HOUSTW	Housing Starts, West
6	4	PERMIT	New Private Housing Permits (SAAR)
7	4	PERMITNE	New Private Housing Permits, Northeast (SAAR)
8	4	PERMITMW	New Private Housing Permits, Midwest (SAAR)
9	4	PERMITS	New Private Housing Permits, South (SAAR)
10	4	PERMITW	New Private Housing Permits, West (SAAR)
Group 4: Orders and Inventories			
1	5	DPCERA3M086SC	Real personal consumption expenditures
2	5	CMRMTSPLx	Real Manu. and Trade Industries Sales
3	5	RETAILx	Retail and Food Services Sales
4	1	NAPM	ISM : PMI Composite Index
5	1	NAPMNOI	ISM : New Orders Index
6	1	NAPMSDI	ISM : Supplier Deliveries Index
7	1	NAPMII	ISM : Inventories Index
8	5	ACOGNO	New Orders for Consumer Goods
9	5	AMDMNOx	New Orders for Durable Goods
10	5	ANDENOx	New Orders for Nondefense Capital Goods
11	5	AMDMUOx	Unfilled Orders for Durable Goods
12	5	BUSINVx	Total Business Inventories
13	2	ISRATIOx	Total Business: Inventories to Sales Ratio
14	2	UMCSENTx	Consumer Sentiment Index
Group 5: Money and Credit			
1	6	M1SL	M1 Money Stock
2	6	M2SL	M2 Money Stock
3	5	M2REAL	Real M2 Money Stock
4	6	AMBSL	St. Louis Adjusted Monetary Base

Table A.7: List of variables (cont'd)

Index	Code	Variable	Description
Group 5: Money and Credit			
4	6	AMBSL	St. Louis Adjusted Monetary Base
5	6	TOTRESNS	Total Reserves of Depository Institutions
6	7	NONBORRES	Reserves Of Depository Institutions
7	6	BUSLOANS	Commercial and Industrial Loans
8	6	REALLN	Real Estate Loans at All Commercial Banks
9	6	NONREVSL	Total Nonrevolving Credit
10	2	CONSPI	Nonrevolving consumer credit to Personal Income
11	6	MZMSL	MZM Money Stock
12	6	DTCOLNVHFNM	Consumer Motor Vehicle Loans Outstanding
13	6	DTCTHFNM	Total Consumer Loans and Leases Outstanding
14	6	INVEST	Securities in Bank Credit at All Commercial Banks
Group 6: Interest rate and Exchange rates			
1	2	FEDFUNDS	Effective Federal Funds Rate
2	2	CP3Mx	3-Month AA Financial Commercial Paper Rate
3	2	TB3MS	3-Month Treasury Bill:
4	2	TB6MS	6-Month Treasury Bill:
5	2	GS1	1-Year Treasury Rate
6	2	GS5	5-Year Treasury Rate
7	2	GS10	10-Year Treasury Rate
8	2	AAA	Moodyis Seasoned Aaa Corporate Bond Yield Aaa bond
9	2	BAA	Moodyis Seasoned Baa Corporate Bond Yield Baa bond
10	1	COMPAPFFx	3-Month Commercial Paper Minus FEDFUNDS CP-FF spread
11	1	TB3SMFFM	3-Month Treasury C Minus FEDFUNDS 3 mo-FF spread
12	1	TB6SMFFM	6-Month Treasury C Minus FEDFUNDS 6 mo-FF spread
13	1	T1YFFM	1-Year Treasury C Minus FEDFUNDS 1 yr-FF spread
14	1	T5YFFM	5-Year Treasury C Minus FEDFUNDS 5 yr-FF spread
15	1	T10YFFM	10-Year Treasury C Minus FEDFUNDS 10 yr-FF spread
16	1	AAAFFM	Moodyis Aaa Corporate Bond Minus FEDFUNDS Aaa-FF spread
17	1	BAAFFM	Moodyis Baa Corporate Bond Minus FEDFUNDS Baa-FF spread
18	5	TWEXMMTH	Trade Weighted U.S. Dollar Index: Major Currencies Ex rate: avg
19	5	EXSZUSx	Switzerland / U.S. Foreign Exchange Rate
20	5	EXJPUSx	Japan / U.S. Foreign Exchange Rate
21	5	EXUSUKx	U.S. / U.K. Foreign Exchange Rate
22	5	EXCAUSx	Canada / U.S. Foreign Exchange Rate
Group 7: Prices			
1	6	WPSFD49207	PPI: Finished Goods
2	6	WPSFD49502	PPI: Finished Consumer Goods
3	6	WPSID61	PPI: Intermediate Materials
4	6	WPSID62	PPI: Crude Materials
5	6	OILPRICEx	Crude Oil, spliced WTI and Cushing
6	6	PPICMM	PPI: Metals and metal products:
7	1	NAPMPRI	ISM Manufacturing: Prices Index
8	6	CPIAUCSL	CPI : All Items
9	6	CPIAPPSL	CPI : Apparel
10	6	CPITRNSL	CPI : Transportation
11	6	CPIMEDSL	CPI : Medical Care
12	6	CUSR0000SAC	CPI : Commodities
13	6	CUSR0000SAD	CPI : Durables
14	6	CUSR0000SAS	CPI : Services
15	6	CPIULFSL	CPI : All Items Less Food
16	6	CUSR0000SA0L2	CPI : All items less shelter
17	6	CUSR0000SA0L5	CPI : All items less medical care
18	6	PCEPI	Personal Cons. Expend.: Chain Index
19	6	DDURRG3M086S	Personal Cons. Exp: Durable goods
20	6	DNDGRG3M086S	Personal Cons. Exp: Nondurable goods
21	6	DSERRG3M086SI	Personal Cons. Exp: Services
Group 8: Stock market			
1	5	S&P	500 S&P'sCommon Stock Price Index: Composite
2	5	S&P:	indust S&P's Common Stock Price Index: Industrials
3	2	S&P	div yield S&P's Composite Common Stock: Dividend Yield S&P div yield
4	5	S&P	PE ratio S&P' Composite Common Stock: Price-Earnings Ratio S&P PE ratio
5	1	VXOCLSx	VXO
Group 9: Goyal & Welch (2008) financial variables			
1	1	SR	S&P 500 return
2	1	RF	Risk Free rate
3	1	DP	Log dividend on S&P 500 minus log S&P 500
4	1	DY	Log dividend on S&P 500 minus log lagged S&P 500
5	1	EP	Log earnings on S&P 500 minus log S&P 500
6	1	DP	Log dividend on S&P 500 minus log earnings
7	1	SVAR	Sum of square return on the S&P 500
8	1	BM	Book value of Dow Jones over the Dow Jones indus average
9	1	NEE	12-month sum of net issue on NYSE over capitalization of NYSE
10	1	TB	Treasury bills
11	1	LTY	Long term governement yield
12	1	LTR	Long term governement bond rate of return
13	1	TMS	Term spread
14	1	DYS	Difference between Baa and Aaa-rated corporate bond yield
15	1	DRS	Difference between the rate of return of Baa and Aaa-rated corporate bond
16	1	INF	Inflation
Group 10: Others			
1	2	VIX	VIX index
2	2	EPU	Economic policy uncertainty index for the US
3	2	LEI	Conference Board Leading Economic Index
4	2	CEI	Conference Board Coincident Economic Index
5	2	LAG	Conference Board Lagging Economic Index
6	2	CCI	Conference Board Consumer Confidence Index
7	2	PSI	Conference Board Present Situation Index
8	2	EXI	Conference Board Expectations Index

Appendix A.1. Factor model

Due to the high dimensionality of $\mathbf{x}_t$, it is practical to reduce the dimensionality of $\mathbf{x}_t$ by assuming that they are driven by a small number of common factors, see, for instance, Stock and Watson (2011). Let $\mathbf{X}_T \equiv \left[\mathbf{x}_1 \,|\, \cdots \,|\, \mathbf{x}_T\right]'$ be $T \times M$ matrix of covariates and $\mathbf{F}_T \equiv \left[\mathbf{f}_1 \,|\, \cdots \,|\, \mathbf{f}_T\right]'$ the $T \times R$ matrix of latent common factors of $\mathbf{X}_T$. We have the following regression problem:

$$\mathbf{X}_T = \mathbf{F}_T \mathbf{\Lambda} + \varepsilon_t \,, \tag{A.1}$$

where $\mathbf{\Lambda}$ is the $R \times M$ matrix of loadings and ε_t is an error term at time t. To estimate the latent factors, we minimize the following expression:

$$V(\mathbf{F}_T, \mathbf{\Lambda}) \equiv \frac{1}{MT} \sum_{i=1}^{M} \sum_{t=1}^{T} (x_{i,t} - \boldsymbol{\lambda}_i \mathbf{f}_t)^2 \,, \tag{A.2}$$

where $\boldsymbol{\lambda}_i$ is the ith row of $\mathbf{\Lambda}$. Under some assumptions, principal component (PC) analysis provide us with estimates of $\mathbf{\Lambda}$ and $\mathbf{F}_T$ with $R = \min\{M, T\}$. However, with PC, some factors can be considered as pure noise. To estimate the optimal number of factor R, we minimize the information criterion proposed in Bai and Ng (2002):

$$IC_{p1}(k) \equiv \ln\left(V(\widehat{\mathbf{F}}_T^k, \widehat{\mathbf{\Lambda}}^k)\right) + k \left(\frac{M+T}{MT}\right) \ln\left(\frac{MT}{M+T}\right) \,, \tag{A.3}$$

where $\widehat{\mathbf{F}}_T^k$ and $\widehat{\mathbf{\Lambda}}^k$ are the first k columns of the PC estimator of $\mathbf{F}_T$ and the first k rows of the PC estimator of $\mathbf{\Lambda}$. The value $k \in \{1, \ldots, k_{\max}\}$ which leads to the minimal IC_{p1} gives us the number of factors to use in the forecasting models $\mathcal{M}_{2a}$ and $\mathcal{M}_{2b}$ in (20)–(21). We follow Bai and Ng (2002) and set $k_{\max} = 8$. Other values were tested but led to qualitatively similar results.

Media and the stock market: A CAT and CAR analysis

Keven Bluteau – Chapter 3
joint work with David Ardia and Kris Boudt

Abstract

We introduce the Cumulative Abnormal Tone (CAT) event study methodology for analyzing the dynamic relationship between printed media news and cumulative abnormal returns (CAR) around events. We apply the CAT event study methodology to media news about the firm published in a window around the quarterly earnings announcements of non–financial S&P 500 firms over the period 2000–2016. We document that there is an abnormal media tone not only on the three days around the earnings announcement, but that there is a substantial drift in the days following the announcement. We also find that the media tone is more sensitive to negative earnings surprises than positive ones. Finally, we report empirical evidence that the abnormal tone of web publications at the earnings announcement date predicts a stock price reversal in the month following the announcement.

Keywords: abnormal return, abnormal tone, earnings announcements, event study, news media, sentometrics

1. Introduction

Corporate announcements, media news about the firm and stock price reactions are highly interconnected. Understanding this dependence is of utmost importance for firm stakeholders and investors. In economics and finance, most studies ignore the information in the big data set of news publications around the event. In fact, it has become popular to study the relationship between an event and the behavior of the stock market around that event using the event study methodology. For instance, in the case of firms' earnings announcement events, the consensus view is that there is an immediate abnormal return reaction to the earnings result, and a significant post–earnings–announcement abnormal return drift (see, *e.g.*, Bernard and Thomas, 1989, 1990). The event often also triggers the publication of news in the media. In this paper, we argue that by considering the abnormal tone in media news, it becomes possible to jointly analyse the joint effect of events on media and stock market behaviour.

Because of regulatory obligations, we expect a spike in media coverage on the day of the announcement. In fact, in order to be compliant with the Fair Disclosure SEC regulation, firms need to make public disclosure of that information to avoid that the information would be available to only a few enumerated persons. The disclosure of material information by large corporations is typically done through press releases distributed by newswires to other news

media outlets and investors. Those newswires are an efficient way for firms to inform investors as they are widely used by journalists as source material for developing stories for their platform.[1] Thus, when the event is important, it is natural to expect that also printed media and web publication outlets discuss it in the days around the event, and that leads to further spreading, discussing, and analyzing of the material information disclosed by firms. As such, the media may provide additional soft information beyond the press releases on which investor could react. We analyze this by studying the dynamics of the tone in the print media news published around earnings announcements. The tone quantifies the polarity of the journalist's disposition with respect to the firm. It can be expected that the tone of media communications is aligned with the market reaction. Moreover, the effect of the event on the tone of those communications may depend on characteristics of a news publication, and it's publisher, such as the distribution method (*e.g.*, newswires, web publications, and newspapers) and business model (*e.g.*, services–, subscriptions–, advertisement–based revenue). This, in turn, could affect individual investors response to the event depending on where and how they get informed.

The empirical analysis of the relationship between earnings announcements, stock markets, and the tone of print media news requires new methodological tools. Thus, in this study, we introduce the Cumulative Abnormal Tone (CAT) event study methodology. We also note that the effectiveness of the approach depends on the tone estimation method used. We recommend a Generalized Word Power approach, which yields application–specific polarity scores for words through a predictive regression calibration. The approach is a generalization of the Word Power methodology of Jegadeesh and Wu (2013). An important characteristic of these methods is that they allow for the extraction of abnormal tone contribution from features of the news articles, such as the type of source (*e.g.*, newswires, web publications, and newspapers).

In financial market applications, we recommend a joint analysis of cumulative abnormal tone (CAT) and cumulative abnormal return (CAR) dynamics. At a daily frequency, one can expect the CAT and CAR dynamics to be intimately linked, provided that the tone is computed on a suitable corpus of texts discussing the news driving the stock market. From the behavioral finance literature on investors' conservativeness and market momentum (see, *e.g.*, Hong and Stein, 1999; Hong et al., 2000; Chan, 2003; Gutierrez and Prinsky, 2007; Huynh and Smith, 2017), it can further be expected that CAT has predictive power towards the future stock market

[1] See, `http://prnewswire.mediaroom.com/2014-03-26-68-of-Journalists-Check-Newswire-At-Least-Once-Per-Day`

return. We test these hypotheses by applying the CAT methodology to analyze media abnormal tone dynamics about firms' future performance in the daily media textual information written about non–financial S&P 500 firms near quarterly earnings announcements for the period ranging from 2000 to 2016. We proxy the overall media using a large sample of newswire, newspaper, and web publication sources. Specifically, we use the Generalized Word Power method to compute the tone of media documents. Then, we divide the tone into normal and abnormal components to analyze the CAT dynamics around earnings announcements. Consequently, we introduce texts–based factor models of normal tone.

Our results suggest that firm– and earnings–specific variables drive CAT at the earnings announcement date. Moreover, we report that journalists will tend to write abnormally positively (negatively) in the month following the earnings announcement if the earnings event had an abnormally positive (negative) response by journalists (*i.e.*, high values for CAT in the three days around the announcement) or investors (*i.e.*, high value for the corresponding CAR) . We also report that CAT is more sensitive to negative earnings surprises than positive earnings surprises. This result is consistent with Soroka (2006) who finds that the mass media response to negative events is much larger than to positive events.

Additionally, we report that CAT for the previous, at, and following day of the earnings announcement provides investors with incremental predictive power regarding the post–earnings–announcement abnormal returns for two to 20 days after the announcement. Specifically, we find that CAT computed for the three days around the event predicts a post–earnings–announcements abnormal return reversal. This effect is stronger for negative CAT. Therefore, our results are in line with an overreaction pattern to news (see, *e.g.*, Antweiler and Frank, 2004; Tetlock, 2007; Garcia, 2013) at the earnings announcements, and psychological evidence indicating that people tend to overreact to negative events (Taylor, 1991).

Finally, we report that the reversal effect is mainly due to the CAT contribution of web publications. The overreaction effect from web publications could be due to two complementary channels. First, based on the journalistic (see, Karlsson and Strömbäck, 2010; Karlsson, 2011; Kilgo et al., 2018) and financial literature (see, Ahern and Sosyura, 2015; Zhang et al., 2016), this result is consistent with the view that the speculative and sensational aspect of web publications could lead uninformed traders to overreact to earnings news. Second, based on the study of Da et al. (2011), this result is also consistent with the view that web publication coverage of earnings events increases the attention of uninformed traders and, thus, could lead to an increase of the

overreaction effect. Moreover, we document that this relationship is predominantly significant for the last seven years of our sample, namely 2010 to 2016, when web publications outnumber the traditional newspapers in terms of volume. This period is characterized by the fact that web–based news has surpassed newspapers in term of primary sources of news by the US population, being only surpassed by television (Pew Research Center, 2011).

The rest of the paper proceeds as follows. Section 2 introduces the CAT event study methodology. Section 3 presents the Generalized Word Power tone methodology, the tone factors, and how we decompose the abnormal tone into textual–document–specific abnormal tone contributions. Section 4 presents the data for the CAT event study on the quarterly earnings announcements. Section 5 presents the analysis of the drivers of CAT. Section 6 presents the analysis of the predictive power of CAT over CAR. Section 7 concludes. Additional results are reported in supplementary appendices.

2. CAT event study methodology

We present the Cumulative Abnormal Tone (CAT) event study methodology to quantify the abnormal tone related to an aspect of an entity in response to an event, where we extract the tone from textual communications (*e.g.*, news articles, press releases) about the entity.

Specifically, we define the tone as the sentiment of the media towards an aspect of an entity at a certain point in time. Instances of aspects include reputation, market valuation, and risk . Entities include public corporations (*e.g.*, S&P 500 firms, NYSE listed firms, private firms), personalities (*e.g.*, politicians, CEOs, celebrities), or products and concepts (*e.g.*, iPhone, dairy products, renewable energy). Examples of events include corporate communications (*e.g.*, earnings announcement, change of CEO, product recall), political communications (*e.g.*, announcements by central banks, elections, political scandal), natural phenomenon (*e.g.*, earthquakes, flooding, drought), among others.

2.1. Tone decomposition

For a given frequency (*e.g.*, daily, weekly, monthly), we measure the tone about a specific aspect of the entity from articles published about the entity near the event date t_i of the event–entity pair i ($i = 1, \ldots, N$), where N is the total amount of event–entity pairs considered. We then divide the daily tone into two components: the normal and the abnormal tone.

Formally, for an aspect, event–entity pair i, and entity–related text communications published at time τ relative to the event date t_i, we express the textual tone ($tone_{i,\tau}$) as the sum of

a normal tone ($ntone_{i,\tau}$) and an abnormal tone ($atone_{i,\tau}$):

$$tone_{i,\tau} \equiv ntone_{i,\tau} + atone_{i,\tau} \,. \tag{1}$$

In this formulation, the abnormal tone can be interpreted as the degree of surprise from the media point of view. The degree of surprise in this case follows the definition of Teigen and Keren (2003) by which it is mainly determined by the extent to which an event contrasts with the expected alternative (*i.e.*, $ntone_{i,\tau}$) and not uniquely by its low probability of outcome.[2]

It is often of interest to evaluate the effect of an event on the media abnormal tone about an aspect of the entity over a period ranging from τ_1 to τ_2 ($\tau_1 < \tau_2$) relative to the event day. As such, we define the cumulative abnormal tone:

$$CAT_i(\tau_1, \tau_2) \equiv \sum_{\tau=\tau_1}^{\tau_2} atone_{i,\tau} \,. \tag{2}$$

Additionally, as we cannot make inference through a single event–entity pair observation, we aggregate through the cross–section of all event–entity pairs $i = 1, \ldots, N$ and consider the average cumulative abnormal tone $\overline{CAT}(\tau_1, \tau_2)$.

2.2. Estimation of the normal tone model

The separation of the tone into the normal and the abnormal tone, therefore, requires that we have an estimate of the normal tone. To model the normal tone, we consider a linear factor model with factors that are updated at the frequency of the tone observations:

$$tone_{i,\tau} \equiv \alpha_i + \mathbf{f}_\tau' \boldsymbol{\beta}_i + \epsilon_{i,\tau} \,, \tag{3}$$

where $\mathbf{f}_\tau$ are common text–based factors of tone of the targeted aspect across all entities at relative time τ, α_i is an event–entity–specific constant, $\boldsymbol{\beta}_i$ are factor exposures around the event–entity i, and $\epsilon_{i,\tau}$ is an error term. This model is analogous to the model of normal return typically used in abnormal return event studies (MacKinlay, 1997).[3]

We estimate the constant α_i and the factor exposures $\boldsymbol{\beta}_i$ in (3) using data available in an

[2] We note that in Huang et al. (2013) and Arslan-Ayaydin et al. (2016), they use a measure of abnormal tone for firms' earnings press releases where the normal tone is modeled using firms' fundamentals. They define the abnormal tone as a measure intended to capture the discretionary and inflated component of tone.

[3] Note that this model nests the special case of setting the normal tone to an event–specific constant (α_i with $\boldsymbol{\beta}_i = \mathbf{0}$).

estimation window and as a simple ordinary least square regression. We define the estimation window as $\tau \in [t_i - L - K, \ldots, t_i - K - 1]$ where L is the length of the estimation window, and K is the offset of the estimation window relative to the event date t_i. Then, we obtain the abnormal tone over an event window:

$$atone_{i,\tau} \equiv tone_{i,\tau} - \widehat{\alpha}_i - \mathbf{f}'_\tau \widehat{\boldsymbol{\beta}}_i \,, \tag{4}$$

for $\tau > t_i - K - 1$, where $\widehat{\alpha}_i$ is the estimated event–entity–specific constant and $\widehat{\boldsymbol{\beta}}_i$ are the estimated factor exposures.

For illustration, let us consider a daily–frequency CAT event study about a specific earnings announcement event of Abbvie Inc. We focus on the second quarter of 2013 (event date on July 26, 2013). First, we collect all relevant articles published about Abbvie Inc. for each day around July 26, 2013. Then, we compute, for each day, an estimate of the daily tone about the change in market valuation (proxied by stock return) of Abbvie Inc. using a tone model calibrated on data observed up to December 31, 2012. In Panel A of Figure 1, we report the steps for this specific example. Then, we estimate the normal tone model in an estimation window of size $L = 30$ and offset $K = 5$. Finally, we can estimate the abnormal tone in the event window. In Panel B of Figure 1, shows the timeline related to this event study setup.

[Insert Figure 1 about here.]

3. Tone, tone factors, and abnormal tone contribution

The choice of an extraction method for the tone is an important consideration. In this study, we consider the lexicon–based approach. Several lexicon–based methods are available in the literature, but the interpretation of the resulting tone computed with these lexicons is not always obvious. In finance, for example, the reference lexicon is the dictionary developed by Loughran and McDonald (2011). Nowadays, numerous studies rely on this method to capture the tone of firm–related textual documents and study its relationship to stock returns (see, *e.g.*, Kelley and Tetlock, 2013; Chen et al., 2014; Ahmad et al., 2016). However, it is important to note that the lexicon was created to capture the degree of positiveness or negativeness in Form 10–K reports. This raises doubts about the usage of the Loughran and McDonald (2011) lexicon as the appropriate choice to compute the tone of financial news and documents that differ widely

from 10–Ks. We, therefore, propose a general tone computation methodology which relates the tone to an underlying aspect of an entity and thus makes the tone measure easily interpretable.

3.1. Daily tone estimation: The Generalized Word Power methodology

Our goal is to compute the media tone about aspect a of entity k at time t. The aspect a of an entity k can, for instance, be the market change in valuation of a firm (*i.e.*, *aspect* = "market change" and *entity* = "firm"). Consider a list of sentiment words $j \in 1, \ldots, J$, where sentiment words are words deemed essential to compute the tone. We define the tone about the aspect of entity k at time t as:

$$tone_{k,t} \equiv \sum_{j=1}^{J} \zeta_j \frac{1}{D_{k,t}} \sum_{d=1}^{D_{k,t}} \frac{F_{d,j,k,t}}{N_{d,k,t}}\,, \tag{5}$$

where ζ_j is the score regarding the aspect a for the sentiment word j, $D_{k,t}$ is the number of textual documents written about entity k at time t conditional on the presence of at least one sentiment word $j \in 1, \ldots, J$ in each textual document $d \in 1, \ldots, D_{k,t}$, $F_{d,j,k,t}$ is the number of times the jth sentiment word is encountered in the textual document d, and $N_{d,k,t}$ is the number of words in the textual document d.

To compute the sentiment words score, we assume that the aspect of entity k at time t has an observable quantitative proxy defined as $a_{k,t}$. For instance, it would be reasonable to use a firm's stock return to proxy change in market valuation. [4] Then, we estimate the sentiment words score using the following linear regression:

$$a_{k,t} \equiv \gamma + \sum_{j=1}^{J} \lambda_j \frac{1}{D_{k,t}} \sum_{d=1}^{D_{k,t}} \frac{F_{d,j,k,t}}{N_{d,k,t}} + \epsilon_{k,t}\,, \tag{6}$$

where γ is a constant, λ_j is the regression coefficient regarding the aspect a for sentiment word j, and $\epsilon_{k,t}$ is an error term. The parameter γ and λ_j for $j \in 1, \ldots, J$ are estimated by ordinary least square.[5] We note that, if $a_{k,t}$ is equal to the abnormal return of firm k at time t and $D_{k,t} = 1$ for all values of k and t, the tone formulation reduces to the Word Power methodology of Jegadeesh and Wu (2013). Thus, we refer to our tone computation methodology as the Generalized Word Power.

[4] In the case where no observable proxy of the aspect is available, one could manually build a lexicon and corresponding sentiment words score for their specific research question as done in Henry (2008), Loughran and McDonald (2011), and Renault (2017).

[5] The parameters could also be estimated using machine learning methods such as the LASSO (see Pröllochs et al., 2015).

Following Jegadeesh and Wu (2013), $\lambda_j \equiv \zeta_j \lambda$, where λ is the regression coefficient of the overall tone measure, not the individual sentiment words. Thus, we derive the Generalized Word Power score about aspect a for sentiment word j as:

$$\widehat{\zeta}_j \equiv \frac{\widehat{\lambda}_j - \widehat{\mu}_{\boldsymbol{\lambda}}}{\widehat{\sigma}_{\boldsymbol{\lambda}}}\,, \tag{7}$$

where $\widehat{\lambda}_j$ is the estimated coefficient for word j, $\widehat{\mu}_{\boldsymbol{\lambda}}$ and $\widehat{\sigma}_{\boldsymbol{\lambda}}$ are the sample mean and standard deviation of $\widehat{\boldsymbol{\lambda}} \equiv (\widehat{\lambda}_1, \ldots, \widehat{\lambda}_J)'$, respectively. Then, the estimated $tone_{k,t}$, computed using the estimated score, can be interpreted as the media implied expected value of $a_{k,t}$.

Finally, using the above results, we define $tone_{i,\tau}$ as:

$$tone_{i,\tau} \equiv \sum_{j=1}^{J} \widehat{\zeta}_j \frac{1}{D_{i,\tau}} \sum_{d=1}^{D_{i,\tau}} \frac{F_{d,j,i,\tau}}{N_{d,i,\tau}}\,, \tag{8}$$

where $D_{i,\tau}$ is the number of textual documents written about the entity related to the event–entity pair at relative time τ conditional on the presence of at least one sentiment word $j \in 1, \ldots, J$ in each textual document $d \in 1, \ldots, D_{i,\tau}$, $F_{d,j,i,\tau}$ is the number of times the jth sentiment word is encountered in the textual document d, and $N_{d,i,\tau}$ is the number of words in the textual document d.

3.2. Tone factors

Referring to (3), we define $\mathbf{f}_\tau$ as a vector of size Q where each entry is a different text–based factors q about the aspect that is appropriately aligned with relative time τ. We compute a text–based factor q about the aspect at time t using a linear function of the entity k tone measures:

$$f_{q,t} \equiv \sum_{k=1}^{K} \omega_{q,k,t} tone_{k,t}\,, \tag{9}$$

where $\omega_{q,k,t}$ is the entity–k weight for factor q at time t, $\sum_{k=1}^{K} \omega_{q,k,t} = 1$, K is the number of entities, and $tone_{k,t}$ is the entity–k tone about the aspect at time t.

3.3. Abnormal tone contribution of individual documents

Thanks to a large amount of metadata embedded in textual data, it is of interest to compute the contribution of the individual textual document to the tone and abnormal tone for any given day near the event date. By aggregating the contribution of textual documents that have a certain commonality, we can compute the average contribution of the types of media outlet,

the individual media outlets, the subjects, and the authors, for instance. This flexibility provides a range of possible analyses.

Using the Generalized Word Power methodology, it is straightforward to derive the tone about the aspect of a single textual document d discussing event–entity pair i and published at relative time τ:

$$tone_{d,i,\tau} \equiv \sum_{j=1}^{J} \widehat{\zeta}_j \frac{F_{d,j,i,\tau}}{N_{d,i,\tau}} \,. \tag{10}$$

Similarly, we obtain the individual document abnormal tone as:

$$atone_{d,i,\tau} \equiv tone_{d,i,\tau} - \widehat{\alpha}_i - \mathbf{f}'_\tau \widehat{\boldsymbol{\beta}}_i \,. \tag{11}$$

By summing over the number of documents, we obtain:

$$\sum_{d=1}^{D_{i,\tau}} \underbrace{\frac{atone_{d,i,\tau}}{D_{i,\tau}}}_{\equiv atc_{d,i,\tau}} \equiv \sum_{d=1}^{D_{i,\tau}} \underbrace{\frac{tone_{d,i,\tau}}{D_{i,t}}}_{\equiv tc_{d,i,\tau}} - \widehat{\alpha}_i - \mathbf{f}'_\tau \widehat{\boldsymbol{\beta}}_i \,, \tag{12}$$

which is a reformulation of (4). This expression highlights that the contribution to the abnormal tone and tone about the aspect of a single textual document d discussing event–entity pair i and published at time τ are $atc_{d,i,\tau}$ and $tc_{d,i,\tau}$, respectively.

4. Earnings announcement CAT event study: Data

Earnings announcements are major firm information release dates (Basu et al., 2013). As an information intermediary, the media cover substantially earnings announcements, which suggests that they play a critical role in the distribution of the information at earnings announcement dates (see Tetlock et al., 2008). Previous study focuses mostly on the effect of earnings surprises on stock return and documented the subsequent post–earnings–announcement abnormal return drift anomaly; see Ball and Brown (1968), Bernard and Thomas (1989, 1990), Daniel et al. (1998), Sadka (2006), DellaVigna and Pollet (2009), for early and more recent studies. We focus on the analysis of media surprise, defined by the CAT value, and on how it relates to the earnings surprise and CAR. Our empirical analysis focuses on articles related to 597 non–financial firms which were included in the S&P 500 index for the period ranging from 2000–Q1 to 2016–Q4.[6]

[6]We consider only the time at which the firms were included in the S&P 500 index. Our sample thus tracks the S&P 500 index constituents over time. We exclude the financial sector as commonly done in earnings–

4.1. Earnings, accounting, and return data

We collect quarterly earnings dates, values, and analyst forecasts from the I/B/E/S database.[7] We gather the quarterly asset value, net income, book value, market capitalization as well as daily stock prices from the merged CRSP–Compustat database. We match these two data sources using the I/B/E/S CUSIP and the Compustat NCUSIP.

To identify good and bad earnings events, we compute the analysts' standardized unexpected earnings (SUE, also called the "earnings surprise"), as done in Livnat and Mendenhall (2006):

$$SUE_i \equiv \frac{EPS_i - F_i}{P_i}\,, \tag{13}$$

where EPS_i is the reported earning–per–share, F_i is the median of the analysts' forecasts of the earning–per–share for event–entity pair i, and P_i is the price of the firm at the end of the earnings quarter for event i. We consider only the most recent forecast for each analyst that is made at maximum 90 days prior to the earnings announcements.

4.2. Textual data

First, we collect from Compustat the historical company names and tickers corresponding to the 597 non–financial firms. For each historical company name, we retrieve all documents available on LexisNexis.[8] We start in 1999 to build an initial news dataset that we use to compute the first estimate of the sentiment word scores. We use the following LexisNexis search filters:

- We retrieve all relevant textual documents from LexisNexis English sources that are categorized as newswires, newspapers, or web publications. It is worthwhile to obtain news from different type of news media as they differs in terms of offered services, business model, and distribution medium. This could have an impact on how the news are written, which news are covered, and type of readers. Newswires, such as PR Newswire, distribute press releases from organizations to news media companies and generate revenue with member subscriptions and publication fees. Newspapers, distribute news in a printed format and

announcements event studies. Financial firms are identified using the first two digits of the Global Industry Classification Standard code of each firm (*i.e.*, first two digits "40" for banks, diversified financials, and insurance firms, and first two digits "60" for REITs and real estate management & development firms).

[7] We use the Forecast Period Indicator data to obtain the forecasts made for the quarterly figures and reported in the quarter before the publication of the quarterly report (*i.e.*, FPI = 6).

[8] See `https://www.nexis.com`.

generate revenue with prepaid advertisements and reader subscriptions. Web publications distribute news online, on their website, and generate most of their revenue with advertisements or sponsored contents, where the fee to advertisers is either a pay–per–view or a pay–per–click scheme. Some organizations have both a newspaper and a web publication. For instance, the newspaper edition of the The New York Times is identified as The New York Times in LexisNexis while the web publication is identified as The New York Times Blogs.

- We use the historical company name as a search term in the COMPANY search index of LexisNexis, and keep documents with a minimum relevance score of 85.[9]
- We require that each textual document contains at least 100 words.
- We exclude *"Plus Patent News"* and *"Indian Patents News"* as both publish patent announcements. We consider the overall textual information contained in these sources irrelevant for the analysis.

We further filter the documents using methods outside of what LexisNexis offers:

- We remove (near–)duplicated documents using locality sensitive hashing (see, Wang et al., 2014).[10]
- We remove machine–generated textual documents. Examples of machine–generated textual documents are automatic daily stock picks from newswires. These machine–generated textual documents are mainly composed of numbers, highly structured, and not written by human authors thus lacking opinions.
- To isolate the media tone of each firm, we remove documents with more than two firms or tickers with a major reference (*i.e.*, relevance score equal or larger than 85). This procedure allows keeping most documents about the individual firm and documents that potentially discuss agreements between two firms.

[9] LexisNexis indexes each document with metadata information, such as the company or subject referred in the document. These metadata tags are each associated with a relevance score indicating if there is a minor or major reference to the metadata tag in the document. To avoid sampling errors, we manually verify that each historical company name is matched to the LexisNexis equivalent company metadata tag. We remove historical company names without a LexisNexis equivalent company tag.

[10] The methodology is implemented in the R package **textreuse** (Mullen, 2016).

Each textual document is then processed to remove HTML tags, numbers, punctuations, and URLs. We also standardize each textual document by removing extra spaces, capital letters, and non–alphanumeric characters. We then transform all words to their root form using the Porter stemming algorithm.

In Figure 2, we report the total number of documents retrieved per year by publication type. We first observe that newswires dominate the other type of publication outlets over the entire sample. We also observe a substantial growth in the overall number of documents available in our corpus. This growth can be attributed to an increase in the publication frequency of the news covering S&P 500 firms and to the rise in the number of sources available in the sample. For example, the number of documents originating from web publications is insignificant previous to the year 2006 but dominates the newspapers in terms of coverage for the years 2009 and beyond.

[Insert Figure 2 about here.]

In Table 1, we report the summary statistics for the number of textual documents and daily coverage data. We define the daily coverage as the percentage of trading day where there is at least one textual document discussing the firm. Overall, our corpus contains 2,224,505 textual documents with an average of 3,719 textual documents per firm and an average daily coverage of 38% (*i.e.*, about one trading day with a textual document every three trading days). The minimum daily coverage is 1% and the maximum daily coverage is 99%. We further split the sample across four buckets of about 156 firms conditional on the average market capitalization of the firms. We observe a positive relationship between the average number of documents (and the average daily coverage) and the average market capitalization of the firms. There is, therefore, more media coverage for larger firms than smaller firms. Firms with the lowest market capitalization have an average of 1,296 documents related to them and an average daily coverage of 25% (*i.e.*, about one trading day with a textual document every four trading days). Firms with the highest market capitalization have an average of 8,498 documents related to them and an average daily coverage of 54% (*i.e.*, about one trading day with a textual document every two trading days).

[Insert Table 1 about here.]

In Appendix A, we go deeper into the analysis of the corpus. In particular, we analyze the major topics or subject discussed in the textual documents of the corpus. Moreover, we provide

a list of the top 50 sources of textual documents, where we report for each source, the source name, the source type, the number of textual documents published available in the corpus, as well as a brief description of each source.

4.3. Tone computation implementation details

The details of our implementation of the Generalized Word Power methodology for the CAT event study on quarterly earnings announcements are as follows. First, as we focus on earnings announcement events, we want to capture the market reaction to the information published by the media. As such, we set $a_{k,t} = r_{k,t}$, where $r_{k,t}$ is the firm k stock return at time t. Therefore, we interpret the tone as the media implied expected return. Second, we merge the LM (Loughran and McDonald, 2011) and Harvard IV–4 (Stone and Hunt, 1963) dictionaries to construct a set of sentiment words and process them using the Porter stemming algorithm to define root words for our lexicon. Moreover, we remove rare root words which appear less than 0.1% of the time each day in the year 1999 (*i.e.*, the first estimation window). The final lexicon contains 2,035 root words. Finally, we compute the Generalized Word Power scores for each of the sentiment words at the beginning of each year using an expanding window (see Panel A of Figure 1). We use those scores to estimate the tone of the documents of the year immediately following the last year of each expanding window. In Table 2, we report the sentiment words with the most positive and negative scores for the last estimation window (ranging from 1999 to 2015). We observe that the majority of the most positive (negative) sentiment words are root words that reasonably qualify as positive (negative) from a finance point of view. The ten most positive root sentiment words are *"beat"*, *"gain"*, *"outperfom"*, *"opportunist"*, *"dump"*, *"bui"*, *"boost"*, *"upbeat"*, *"strong"*, *"upsid"*. The ten most negative root sentiment words are *"downgrad"*, *"lower"*, *"disappoint"*, *"sank"*, *"weak"*, *"miss"*, *"low"*, *"drop"*, *"weaker"*, *"cut"*.

[Insert Table 2 about here.]

Analyzing the most positive and negative sentiment words of the lexicon can provide an indication about the quality of a lexicon. We provide further analysis for validation of the method in Appendix B.

4.4. Tone factors

We consider a market capitalization–weighted text–based factor, $tone_t^{mkt}$, where the weights are proportional to the market–capitalization value of the outstanding shares of the firms at

time t. Our motivation for this choice is based on two observations. First, as observed in our corpus, high market capitalization firms have, on average, more textual documents written about them by the media than firms with low market capitalization. Thus, firms with high market capitalization should have a more significant impact on the market tone. Furthermore, a market capitalization–weighted text–based normal tone factor model creates a parallel with the market model used for computing the abnormal returns in this event study.

5. Earnings announcement CAT event study: Exploratory data analysis and drivers of CAT

Event studies are used for two purposes: exploratory data analysis and understanding the drivers of the variable of interest. In this section, we are interested in the relationship between the CAT and the level of SUE. We first proceed by using a graphical analysis of the average CAT conditional on SUE quintiles. Then we use a formal panel regression analysis controlling for any confounding factors.

In total, we have 23,502 earnings announcement events across 597 non–financial S&P 500 firms. We consider an estimation window of 30 days with an offset of 5 days (*i.e.*, $L = 30$, $K = 5$, and the estimation window is $[t_i - 35, t_i - 6]$). The event window contains 26 days. It starts five days before the event and ends 20 days after the event (*i.e.*, $[t_i - 5, t_i + 20]$; see Panel B of Figure 1). This way we can analyze the pre– and post–event dynamics of the abnormal tone in addition to the event date (*i.e.*, $\tau = 0$).[11]

As most of our analysis focuses on $CAT_i(-1, 1)$, we require that an event–entity pair has a daily tone observation at, one day before, or one day after the event date. This reduces the number of event–entity pairs to 21,867. Moreover, to reduce potential issues related to missing observations for the normal tone model estimation, we require at least ten daily tone observations in the estimation window, therefore reducing further the sample of event–entity pairs to 17,157. Finally, as we want to analyze the post–event dynamics $CAT_i(2, 20)$, we require at least ten daily tone observations in the estimation window, thus reducing the sample size to 14,896 event–entity pairs.

[11] The sum of the length of the estimation window and event window amounts to 61 days, which is typically the length between two quarterly earnings announcements for the same firm. This allows having non–overlapping events when considering a single firm. However, we do have overlapping events across firms.

5.1. Number of news near the earnings announcement events

We first analyze the efficiency of newswires, web publications, and newspapers at distributing firm–related news near earnings announcements. If the media are efficient at distributing important information, we should observe a spike in the number of news publications at the event date. In Figure 3, we report the average number of documents at each day relative to the event date across the event–entity pairs for newswires, web publications, and newspapers.

[Insert Figure 3 about here.]

Note that, for newswires and web publications, the average number of publications at the event date spikes at 4.1 and 1.88, as compared to the unconditional average of approximately 1.25 and 0.33, respectively, on non–event day. This is in stark contrast to newspapers where the average number of publications spike at 1.27 only the day following the event day. This observation indicates that newswires and web publications are fast at reporting the event while newspapers lag. This is expected as press releases from firms are distributed to newswires for wide distribution purpose to other media services and outlets and thus are the first to spread the news. Web publication outlets can almost immediately write a story about the event and publish it on their online platforms while newspapers, however, need to go through the process of printing and physical distribution of the news.

From the literature on the number of analysts covering S&P 500 firms (see, *e.g.*, Hong et al., 2000), we can expect that the media coverage is positively related to the market capitalization. This would lead to an increase in the level of attention and visibility for large firms' earnings news. In Figure 4, we report the average number of documents at each day relative to the event date across different market capitalization buckets (#1 for the lowest to #4 for the largest).

[Insert Figure 4 about here.]

Results indicate that the largest firms (*i.e.*, bucket #4) are widely more covered than all other firms at earnings announcement dates. Indeed, the firms in the firms in the largest market capitalization quartiles average 8.55 news at earnings announcement date compared to 5.88, 4.70, and 5.01 for buckets #3, #2, and #1, respectively, thus almost doubling the news of news coverage of the lowest two buckets. Moreover, there is substantially more coverage for large market capitalization firms on non–event day averaging about 3.10 news a day compared to 1.53, 1.12, and 1.11 for buckets #3, #2, and #1, respectively. These results indicate that the largest firms generate substantially more media coverage than smaller firms.

Overall, these observations in news release suggest that the print media play an active role in the dissemination of the information at earnings announcement dates. However, there is a substantial cross–sectional differences between types of media or firms market capitalization level.

5.2. Average CAT analysis by level of SUE

Consider a financial market where the media act as an information intermediary and report the earnings event adequately. In such a financial market, there should be a significant relationship between the CAT and the SUE. Thus, we investigate the dynamics of the average CAT conditional on different level of the SUE. Most earnings announcements event studies classify events into buckets based on SUE level (see, *e.g.* Bernard and Thomas, 1990; Mendenhall, 2004; Livnat and Mendenhall, 2006). We split the events into five buckets where the cutoff points for SUE_i is based on the SUE quintiles observed before the quarter associated to the earnings event i.[12] In Figure 5, we report the average CAT dynamics conditional on the SUE quintile–bucket. We test for significance using the traditional t–test from the event study literature (MacKinlay, 1997) (t–test) and the non–parametric generalized rank test (Kolari and Pynnonen, 2011) (t–grank) that takes into account non–normality, serial correlation, cross-correlation, and event–induced volatility.

[Insert Figure 5 about here.]

Before the event date (*i.e.*, $\tau_1 = -5$ and $\tau_2 = -2$), we do not observe a significant drift in the average CAT for any of the SUE bucket. At the event date (*i.e.*, $\tau_1 = -1$ and $\tau_2 = 1$), we observe a clear relationship between the level of SUE and CAT. Specifically, the average CAT is -0.39% (t–stat of -25.58 and t–grank of -14.48), -0.09% (t–stat of -5.30 and t–grank of -2.33), -0.07% (t–stat of 4.87 and t–grank of 3.28), 0.08% (t–stat of 5.65 and t–grank of 5.83), and 0.11% (t–stat of 7.24 and t–grank of 5.07), for buckets #1, #2, #3, #4, and #5, respectively. Moreover, the figure suggests that the CAT reacts more strongly to negative earnings surprises. This result is in line with Soroka (2006) who finds that the mass media response to negative events is much greater than for positive events. He argues, based on the prospect theory of Kahneman and Tversky (1979), that as people are averse to losses, journalists will consider negative information as more important, not only based on their interests but also based on the

[12] We use quintiles instead of deciles as our number of observations is relatively low compared with Livnat and Mendenhall (2006).

interests of their audience. The same reasoning can be applied to the journalists of the media outlets contained in our corpus.

For the post–event time range (*i.e.*, $\tau_1 = 2$ and $\tau_2 = 30$), we observe clear drifts across all SUE buckets. The short–term drift for CAT could be explained by delayed reporting of earnings events by a subset of news sources. Indeed, some outlets such as printed newspapers are restricted to strict release periods, while newswire and web–based news can be distributed in near–real time. Moreover, we observe that for the lowest SUE buckets, which show the sharpest reaction in CAT at the event date, also show the sharpest short–term drift. This finding is in line with the "delayed reporting" explanation. However, delayed reporting is unlikely to explain the long–term drift. Indeed, the graph indicates that highly positive (negative) earnings surprise events generate a long–term positive (negative) momentum in abnormal tone. This highlights that there is abnormally positive (negative) reporting post–event from the media when good (bad) earnings event occur.

5.3. Regression analysis

The previous graphical analysis reports that the CAT reflects, to some degree, the level of earnings surprise. We now test this more formally, via a panel regression framework, to identify possible drivers of CAT at the event date and after the event date. We proceed by regressing $CAT_i(-1,1)$, $CAT_i(2,5)$, and $CAT_i(2,20)$ on two SUE–related variables. Specifically, the two SUE–related variables are the bucket (from 1 to 5) of the SUE level scaled between 0 and 1 and subtracted by 0.5 as done in Livnat and Mendenhall (2006) ($QSUE_i$), and the negative SUE indicator ($SUE_i < 0$). We additionally control for the return–on–asset (ROA_i), the book–to–market ratio (in logarithm, $log(B_i/M_i)$), and the market capitalization (in logarithm, $log(M_i)$). We also include year–quarter and firm fixed–effects, $CAR_i(-1,1)$, and $CAR_i(-5,-2)$ as additional variables. The abnormal returns are computed on same event study parameters as the abnormal tone and we use a market model to compute the normal return (MacKinlay, 1997). The market is defined as the market index from the Kenneth R. French website.[13]

[Insert Table 3 about here.]

In Table 3, we report the regression results. We first note that $QSUE_i$ (and $CAR_i(-1,1)$) and $SUE_i < 0$ are positive and negative significant explanatory variables of $CAT_i(-1,1)$, respectively. This result is in line with our previous analysis, which suggests that CAT reacts to

[13] The data is available at `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`.

the level of earnings surprise and this reaction is stronger for negative earnings surprise events. Moreover, we also find that $CAT_i(-5,-2)$ has a predictive power towards $CAT_i(-1,1)$. This result could either be due to an inertia in the media opinion (*i.e.*, autocorrelation in the abnormal tone), or that the media published earnings relevant information before the events. We find the latter less likely and test this further in the next section.

Regarding $CAT_i(2,20)$, we find that the earnings surprise is not a predictive variable. Indeed, we find instead that the market abnormal reaction, $CAR_i(-1,1)$, as well as the the media surprise, $CAT_i(-1,1)$, drive the post–earnings CAT drift. If we believe that there is an inertia in the media opinion, this result is consistent with the fact that $CAT_i(-5,-2)$ predicts $CAT_i(-1,1)$. Moreover, a large positive (negative) stock price reaction is likely to act as a positive (negative) indicator for journalists and can be a genuine indicator of better (worse) firm financial health. Both these reasons are likely to generate more positive (negative) reporting by the media post–event.

6. Earnings announcement CAT event study: Predictive power of CAT over CAR

We now tackle the question of whether CAT is useful for predicting CAR. We use the same event study setup as for the analysis of the CAT dynamics and drivers. We remove the condition that at least ten abnormal tone observations are needed in the event window as it is an *ex–post* filtering condition. It is essential to have *ex–ante* filtering conditions to make trading strategy based on CAT implementable. Under that set of conditions, we have 17,157 events.

The analysis of interest is regarding the relation between $CAT_i(-1,1)$ and the post–earnings–announcement abnormal returns. Studies of the tone of documents near earnings announcements indicate that the tone predicts a post–earnings–announcement abnormal return drift; see, for instance, Engelberg (2008) for news available in the Dow Jones News Service, Demers et al. (2008) and Arslan-Ayaydin et al. (2016) for earnings press releases, and Price et al. (2012) for conference call transcripts. These results are in line with the slow diffusion of information theory (see, Hong and Stein, 1999; Brav and Heaton, 2002) and conservatism bias (Barberis et al., 1998), both leading to investor underreaction at earnings events. We differentiate from these studies by focusing on the abnormal tone which is computed in a novel way and has a much different form of interpretation, that is the media surprise, and the abnormal component in the media news–based predicted return.

Similarly to Peress (2008), we note that our sample is biased towards events and firms that

will exhibit less post–earnings–announcement abnormal return drift due to low–attention effect (see, *e.g.*, DellaVigna and Pollet, 2009; Hirshleifer et al., 2009). First, our set of firms is constrained to largely followed S&P 500 firms. Second, we require at least one analyst forecast to compute the earnings surprise. Finally, we require at least one news item near the earnings announcement and ten days with news coverage in the estimation window. These are all conditioning factors that most likely increase the level of attention. Therefore, the documented post–earnings–announcement abnormal return drift attributed to the low–attention effect will be weaker.

6.1. Average CAR by level of CAT

We first approach this question from an exploratory data analysis by investigating the dynamics of the average CAR conditional on the level of $CAT_i(-1,1)$. Similar to the previous analysis on the relationship between CAT and SUE, we base the cutoff points for the $CAT_i(-1,1)$ on CAT quintiles observed before the quarter of the earnings event i. In Figure 6, we report the average CAR dynamics conditional the CAT quintile–bucket.

[Insert Figure 6 about here.]

We observe that CAT induces a post–earnings–announcement reversal which is consistent with the overreaction pattern related to news observed in Antweiler and Frank (2004), Tetlock (2007), and Garcia (2013). It is also consistent with the representativeness heuristic in the behavioral model of Barberis et al. (1998). Representativeness, in that case, is thought of as investors overweighting the strength of the evidence (*i.e.*, in our case, the level of $CAT_i(-1,1)$), despite the relatively low reliability of that evidence, thus leading to overreaction. Moreover, the figure suggests that the reversal is stronger for the lowest CAT than the highest CAT buckets, indicating that the reversal is stronger when $CAT_i(-1,1)$ is highly negative. This asymmetric relationship is consistent with evidence from psychological studies. Specifically, negative and more extreme events, where the $CAT_i(-1,1)$ proxies the degree of positiveness and negativeness, attract more attention (Fiske, 1980) and evoke strong and rapid responses when compared to neutral and positive events (Taylor, 1991). This, in turn, exacerbates the overreaction of investors in regard to the level of $CAT_i(-1,1)$.

6.2. Regression analysis

We now render the previous results more robust via a panel regression framework to assess the predictive power of CAT in regards to CAR at the event date and after the event date, while

controlling for confounding factors. Moreover, as the previous results indicate that CAT has an asymmetric effect on CAR, we differentiate between positive and negative CAT. In Table 4, we report the result of various regressions.

[Insert Table 4 about here.]

First, we observe a significant relationship between positive $CAT_i(-1,1)$ and the short–term abnormal return drift, that is, $CAR_i(2,5)$. The regression coefficient is negative and significant (-0.102) indicating that positive CAT at the event date predicts a short–term stock price reversal. Second, we observe a significant relationship between negative $CAT_i(-1,1)$ and the long–term abnormal return drift, that is, $CAR_i(2,20)$. The regression coefficient is negative and significant (-0.322) indicating that the negative CAT at the event date also predicts a long–term stock price reversal. Thus, the CAT at the event date has incremental predictive power regarding the forecast of the post–earnings–announcement abnormal returns. Third, we observe that $CAR_i(-1,1)$ is a significant predictor of $CAR_i(2,5)$ and $CAR_i(2,20)$. The regression coefficients are positive (0.032 and 0.101) indicating that $CAR_i(-1,1)$ predicts a momentum effect. Thus, the overall results suggest that it is possible to have an overreaction or underreaction at earnings announcement events depending on level of $CAT_i(-1,1)$ and $CAR_i(-1,1)$.

6.3. Does the CAR respond differently depending on the source of CAT?

It is possible that the news source of CAT influences the predictive power of CAT over CAR. We test this by decomposing the CAT into newspaper–, newswire–, and web publication–specific CAT contributions. This allows us to analyze whether the informational content of one source–type has more predictive power than another when predicting the CAR. We obtain the source–type contributions in two steps. For a range of relative day $[\tau_1, \tau_2]$, we sum the abnormal tone contributions in (12) attributed to the individual documents belonging to a given source–type. The CAT value then normalizes the CAT contribution per source–type over the same time range. This leads to percentage cumulative abnormal tone contributions for an event–entity pair i over the range τ_1 and τ_2 for newspapers, $c_i(\tau_1,\tau_2)_{newspaper}$, newswire, $c_i(\tau_1,\tau_2)_{newswire}$, and web publication, $c_i(\tau_1,\tau_2)_{web\ publication}$. In Figure 5, we report the regression results.

[Insert Table 5 about here.]

We observe that only the web publication category contains significant predictive power towards the post–earnings–announcement abnormal return. As for the general results, it predicts

a reversal in stock price. This is true for both the short and long–term post–earnings CAR. There are several reasons that might explain why only web publication news predicts a stock price reversal.

First, web publications are characterized by their sense of immediacy compared to traditional news (Karlsson and Strömbäck, 2010).[14] Immediacy, however, comes at the cost of accuracy and quality (Karlsson, 2011). Indeed, some studies suggest that fast reporting of news events impoverishes the quality of journalism (see, *e.g.*, Lewis and Cushion, 2009; Reich, 2016). Also, there are reports that an increasing number of web–based media journalists are paid by clicks, thus increasing sensationalism in web news as journalists compete for attention (Kilgo et al., 2018). Thus, the information contained in web publications tends to be more speculative, more sensational, and therefore less accurate. Moreover, it appeals to a broader readership due to the easy access of web–based news. Under that reasoning, our result is consistent with Ahern and Sosyura (2015), where they report an overreaction pattern on less accurate stories that use ambiguous language and focus on well–known firms with broad readership appeal. This result is also consistent with Zhang et al. (2016), who find that investors overreact to internet news that should have no effect in an efficient market.

Second, our result is related to Da et al. (2011), who find that internet search increases the attention of uninformed traders leading to a significant initial price increase and a subsequent price reversal for IPO events. It would also be reasonable to assume that web–based coverage of an earnings event increases the attention of uninformed traders. Thus, the increased attention of uninformed traders, combined with the sensational nature of web–based publications, introduces two complementary channels that can lead to overreaction at the earnings event.

6.4. Does this effect change in time?

The time–varying composition of textual media news in Figure 2 raises the question of whether the documented effects are stable over time. In fact, web publication coverage was a small subset of the overall sample prior to 2005, and became higher than newspaper from 2009 and onward. This can be attributed to the large declined in newspaper readership at the benefit of web–based outlets. Moreover, according to the State of the News Media report (Pew Research Center, 2011), as of the end of 2010 in the United States, more people got their news

[14]The term "immediacy" refers to the notion that the news cycle as such has become radically shortened and that the time lag between when a news organization publishes new information about new issues has been shortened.

from the web than from printed newspapers. While this study aimed at the general population and not stock market participant, there is no reason to believe that this would not be the case for investors. From this, we conjecture that the effect of web publication will mostly be observed towards the end of our sample. We test this by performing a five–year rolling regression analysis using the same regression specification of Section 6.3. We focus on the effect of $CAT_i(-1,1)$ contribution by source type on $CAR_i(2,20)$. In Figure 7, we report the coefficient for the newswire, newspaper and web publication $CAT_i(-1,1)$ contribution.

[Insert Figure 7 about here.]

Results indicate that the $CAT_i(-1,1)$ contribution of web publication is a significant predictor of $CAR_i(2,20)$ towards the end of the sample. Specifically, the subsamples starting from 2010 to 2012 and ending from 2014 to 2016, respectively. Similarly to the main results of Section 6.3, the CAT contribution of web publication predicts a stock price reversal. Moreover, the CAT contribution of newspapers and newswires are non–significant predictors in all subsamples.

7. Conclusion

The analysis of the media abnormal tone dynamics towards firms around financial events is an important step towards the understanding of the relationship between the information diffused by media sources and the stock market processing of information by investors. We introduce the Cumulative Abnormal Tone (CAT) event study methodology to track the dynamics in the abnormal media tone for a given aspect of an entity. Moreover, we also introduce the Generalized Word Power tone computation methodology to compute the tone of media documents about a particular aspect of an entity.

We apply the CAT event study and Generalized Word Power tone computation methodologies to media reports published around the quarterly earnings announcements of 597 non–financial S&P 500 firms over the period 2000–2016. Our results suggest that CAT is driven by firm and earnings–specific variables, mainly the stock return and the earnings surprise. Moreover, CAT provides investors with incremental predictive information regarding the post–earnings–announcement abnormal returns. Notably, we find that the CAT contribution of web publication sources at the earnings announcement date is the most informative regarding the post–earnings–announcement abnormal returns when compared to newspaper and newswire sources. Our results are consistent with psychological studies, and further suggest that the media reporting is

an information channel influencing investors' decision–making.

The proposed CAT methodology can be used for a wide range of applications outside the financial context. In particular, it could be interesting to analyze how political events shift the narrative towards a specific entity such as a country.

References

Ahern, K.R., Sosyura, D., 2015. Rumor has it: Sensationalism in financial media. Review of Financial Studies 28, 2050–2093. doi:10.1093/rfs/hhv006.

Ahmad, K., Han, J., Hutson, E., Kearney, C., Liu, S., 2016. Media–expressed negative tone and firm–level stock returns. Journal of Corporate Finance 37, 152–172. doi:10.1016/j.jcorpfin.2015.12.014.

Akhtar, S., Faff, R., Oliver, B., Subrahmanyam, A., 2012. Stock salience and the asymmetric market effect of consumer sentiment news. Journal of Banking & Finance 36, 3289–3301. doi:10.1016/j.jbankfin.2012.07.019.

Antweiler, W., Frank, M.Z., 2004. Is all that talk just noise? The information content of internet stock message boards. Journal of Finance 59, 1259–1294. doi:10.1111/j.1540-6261.2004.00662.x.

Arslan-Ayaydin, Ö., Boudt, K., Thewissen, J., 2016. Managers set the tone: Equity incentives and the tone of earnings press releases. Journal of Banking & Finance 72, 132–147. doi:10.1016/j.jbankfin.2015.10.007.

Ball, R., Brown, P., 1968. An empirical evaluation of accounting income numbers. Journal of Accounting Research 6, 159–178.

Barberis, N., Shleifer, A., Vishny, R., 1998. A model of investor sentiment. Journal of Financial Economics 49, 307–343. doi:10.1016/S0304-405X(98)00027-0.

Basu, S., Duong, T.X., Markov, S., Tan, E.J., 2013. How important are earnings announcements as an information source? European Accounting Review 22, 221–256. doi:10.1080/09638180.2013.782820.

Bernard, V.L., Thomas, J.K., 1989. Post–earnings–announcement drift: Delayed price response or risk premium? Journal of Accounting Research 27, 1–36. doi:10.2307/2491062.

Bernard, V.L., Thomas, J.K., 1990. Evidence that stock prices do not fully reflect the implications of current earnings for future earnings. Journal of Accounting and Economics 13, 305–340. doi:10.1016/0165-4101(90)90008-R.

Brav, A., Heaton, J.B., 2002. Competing theories of financial anomalies. Review of Financial Studies 15, 575–606. doi:10.1093/rfs/15.2.575.

Chan, W.S., 2003. Stock price reaction to news and no–news: Drift and reversal after headlines. Journal of Financial Economics 70, 223–260. doi:10.1016/S0304-405X(03)00146-6.

Chen, H., De, P., Hu, Y.J., Hwang, B.H., 2014. Wisdom of crowds: The value of stock opinions transmitted through social media. Review of Financial Studies 27, 1367–1403. doi:10.1093/rfs/hhu001.

Da, Z., Engelberg, J., Gao, P., 2011. In search of attention. Journal of Finance 66, 1461–1499. doi:10.1111/j.1540-6261.2011.01679.x.

Daniel, K., Hirshleifer, D., Subrahmanyam, A., 1998. Investor psychology and security market under– and overreactions. Journal of Finance 53, 1839–1885. doi:10.1111/0022-1082.00077.

DellaVigna, S., Pollet, J.M., 2009. Investor inattention and Friday earnings announcements. Journal of Finance 64, 709–749. doi:10.1111/j.1540-6261.2009.01447.x.

Demers, E., Vega, C., et al., 2008. Soft information in earnings announcements: News or noise? Working Paper.

Engelberg, J., 2008. Costly information processing: Evidence from earnings announcements. Working Paper.

Fiske, S.T., 1980. Attention and weight in person perception: The impact of negative and extreme behavior. Journal of Personality and Social Psychology 38, 889–906. doi:10.1037/0022-3514.38.6.889.

Garcia, D., 2013. Sentiment during recessions. Journal of Finance 68, 1267–1300. doi:10.1111/jofi.12027.

Gutierrez, R.C., Prinsky, C.A., 2007. Momentum, reversal, and the trading behaviors of institutions. Journal of Financial Markets 10, 48–75. doi:10.1016/j.finmar.2006.09.002.

Henry, E., 2008. Are investors influenced by how earnings press releases are written? Journal of Business Communication 45, 363–407. doi:10.1177/0021943608319388.

Hirshleifer, D., Lim, S.S., Teoh, S.H., 2009. Driven to distraction: Extraneous events and underreaction to earnings news. Journal of Finance 64, 2289–2325. doi:10.1111/j.1540-6261.2009.01501.x.

Hong, H., Lim, T., Stein, J.C., 2000. Bad news travels slowly: Size, analyst coverage, and the profitability of momentum strategies. Journal of Finance 55, 265–295.

Hong, H., Stein, J.C., 1999. A unified theory of underreaction, momentum trading, and overreaction in asset markets. Journal of Finance 54, 2143–2184. doi:10.1111/0022-1082.00184.

Huang, X., Teoh, S.H., Zhang, Y., 2013. Tone management. Accounting Review 89, 1083–1113. doi:10.2308/accr-50684.

Huynh, T.D., Smith, D.R., 2017. Stock price reaction to news: The joint effect of tone and attention on momentum. Journal of Behavioral Finance 18, 304–328. doi:10.1080/15427560.2017.1339190.

Jegadeesh, N., Wu, D., 2013. Word power: A new approach for content analysis. Journal of Financial Economics 110, 712–729. doi:10.1016/j.jfineco.2013.08.018.

Kahneman, D., Tversky, A., 1979. Prospect theory: An analysis of decision under risk. Econometrica 47, 263–291.

Karlsson, M., 2011. The immediacy of online news, the visibility of journalistic processes and a restructuring of journalistic authority. Journalism 12, 279–295. doi:10.1177/1464884910388223.

Karlsson, M., Strömbäck, J., 2010. Freezing the flow of online news: Exploring approaches to the study of the liquidity of online news. Journalism Studies 11, 2–19. doi:10.1080/14616700903119784.

Kelley, E.K., Tetlock, P.C., 2013. How wise are crowds? Insights from retail orders and stock returns. Journal of Finance 68, 1229–1265. doi:10.1111/jofi.12028.

Kilgo, D.K., Harlow, S., García-Perdomo, V., Salaverría, R., 2018. A new sensation? An international exploration of sensationalism and social media recommendations in online news publications. Journalism 19, 1497–1516. doi:10.1177/1464884916683549.

Kolari, J.W., Pynnonen, S., 2011. Nonparametric rank tests for event studies. Journal of Empirical Finance 18, 953–971. doi:10.1016/j.jempfin.2011.08.003.

Lewis, J., Cushion, S., 2009. The thirst to be first: An analysis of breaking news stories and their impact on the quality of 24–hour news coverage in the UK. Journalism Practice 3, 304–318. doi:10.1080/17512780902798737.

Livnat, J., Mendenhall, R.R., 2006. Comparing the post–earnings announcement drift for surprises calculated from analyst and time series forecasts. Journal of Accounting Research 44, 177–205. doi:10.1111/j.1475-679X.2006.00196.x.

Loughran, T., McDonald, B., 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. Journal of Finance 66, 35–65. doi:10.1111/j.1540-6261.2010.01625.x.

MacKinlay, A.C., 1997. Event studies in economics and finance. Journal of Economic Literature 35, 13–39.

Mendenhall, R.R., 2004. Arbitrage risk and post–earnings–announcement drift. Journal of Business 77, 875–894.

Mullen, L., 2016. **textreuse**: Detect text reuse and document similarity. URL: https://CRAN.R-project.org/

package=textreuse. R package version 0.1.4.

Peress, J., 2008. Media coverage and investors' attention to earnings announcements Working paper.

Petersen, M.A., 2009. Estimating standard errors in finance panel data sets: Comparing approaches. Review of Financial Studies 22, 435–480. doi:10.1093/rfs/hhn053.

Pew Research Center, 2011. State of the News Media: An Annual Report on American Journalism. URL: https://www.pewresearch.org/wp-content/uploads/sites/8/2017/05/State-of-the-News-Media-Report-2011-FINAL.pdf.

Price, S.M., Doran, J.S., Peterson, D.R., Bliss, B.A., 2012. Earnings conference calls and stock returns: The incremental informativeness of textual tone. Journal of Banking & Finance 36, 992–1011. doi:10.1016/j.jbankfin.2011.10.013.

Pröllochs, N., Feuerriegel, S., Neumann, D., 2015. Generating domain–specific dictionaries using Bayesian learning, in: Twenty-Third European Conference on Information Systems.

Reich, Z., 2016. Comparing news reporting across print, radio, television and online: Still distinct manufacturing houses. Journalism Studies 17, 552–572. doi:10.1080/1461670X.2015.1006898.

Renault, T., 2017. Intraday online investor sentiment and return patterns in the US stock market. Journal of Banking & Finance 84, 25–40. doi:10.1016/j.jbankfin.2017.07.002.

Sadka, R., 2006. Momentum and post–earnings-announcement drift anomalies: The role of liquidity risk. Journal of Financial Economics 80, 309–349. doi:10.1016/j.jfineco.2005.04.005.

Soroka, S.N., 2006. Good news and bad news: Asymmetric responses to economic information. Journal of Politics 68, 372–385. doi:10.1111/j.1468-2508.2006.00413.x.

Stone, P.J., Hunt, E.B., 1963. A computer approach to content analysis: Studies using the general inquirer system, in: Proceedings of the May 21-23, 1963, Spring Joint Computer Conference, ACM. 241–256.

Taylor, S.E., 1991. Asymmetrical effects of positive and negative events: The mobilization–minimization hypothesis. Psychological Bulletin 110, 67. doi:10.1037/0033-2909.110.1.67.

Teigen, K.H., Keren, G., 2003. Surprises: Low probabilities or high contrasts? Cognition 87, 55–71. doi:10.1016/s0010-0277(02)00201-9.

Tetlock, P.C., 2007. Giving content to investor sentiment: The role of media in the stock market. Journal of Finance 62, 1139–1168. doi:10.1111/j.1540-6261.2007.01232.x.

Tetlock, P.C., Saar-Tsechansky, M., Macskassy, S., 2008. More than words: Quantifying language to measure firms' fundamentals. Journal of Finance 63, 1437–1467. doi:10.1111/j.1540-6261.2008.01362.x.

Wang, J., Shen, H.T., Song, J., Ji, J., 2014. Hashing for similarity search: A survey. Working Paper.

Zhang, Y., Song, W., Shen, D., Zhang, W., 2016. Market reaction to internet news: Information diffusion and price pressure. Economic Modelling 56, 43–49. doi:10.1016/j.econmod.2016.03.020.

Table 1: Text and coverage
This table reports the average, the minimum, and the maximum number of documents as well as the daily percentage media coverage for the 597 non–financial historical constituents of the S&P 500. We split the S&P 500 constituents into four buckets, where the 1st (4th) bucket is composed of the firms with the lowest (highest) average market–capitalization value when they were constituents of the S&P 500.

		# of texts			Coverage		
Bucket	# of Firms	Average	Min	Max	Average	Min	Max
#1	149	1,296	6	16,958	25%	1%	90%
#2	149	1,902	2	10,440	34%	1%	87%
#3	149	3,174	51	16,861	40%	2%	82%
#4	150	8,498	28	128,558	54%	6%	99%
All	597	3,719	2	128,558	38%	1%	99%

Table 2: Most positive and negative root sentiment words
This table reports the most positive and negative root words among the 2,035 sentiment words. We use the Generalized Word Power scores estimated for the year 2016, that is, using the data from 1999 to 2015, to sort the sentiment words.

Most Positive				Most Negative			
#	Root Word	#	Root Word	#	Root Word	#	Root Word
1	beat	51	unattract	1	downgrad	51	dissatisfi
2	gain	52	split	2	lower	52	call
3	outperform	53	signifi	3	disappoint	53	admir
4	opportunist	54	regain	4	sank	54	arrog
5	dump	55	cancel	5	weak	55	subscrib
6	bui	56	hell	6	miss	56	junk
7	boost	57	notabl	7	low	57	broken
8	upbeat	58	board	8	drop	58	disclaim
9	strong	59	premier	9	weaker	59	circl
10	upsid	60	trivial	10	cut	60	round
11	ralli	61	joke	11	shortfal	61	unhappi
12	improv	62	eas	12	declin	62	stand
13	stronger	63	encourag	13	warn	63	investig
14	surpass	64	impedi	14	fall	64	grace
15	strength	65	meticul	15	sharpli	65	vivid
16	better	66	catch	16	slower	66	invis
17	cutback	67	hedg	17	hurt	67	pride
18	upgrad	68	cost	18	lost	68	vice
19	fool	69	idl	19	overshadow	69	foresight
20	multitud	70	exception	20	weakest	70	weaken
21	benefit	71	monster	21	best	71	unsaf
22	posit	72	outstand	22	worsen	72	reap
23	special	73	deal	23	motlei	73	exclud
24	share	74	quicken	24	wors	74	strengthen
25	hit	75	guardian	25	loss	75	purport
26	rebound	76	underestim	26	dismal	76	bad
27	hot	77	su	27	downward	77	expens
28	upward	78	definit	28	difficult	78	shark
29	underperform	79	subscript	29	concern	79	decreas
30	highest	80	disproportion	30	competit	80	need
31	interest	81	stabil	31	drunk	81	overestim
32	pleas	82	overdu	32	slowdown	82	unlaw
33	unfound	83	know	33	aggress	83	play
34	advanc	84	audibl	34	maxim	84	nervou
35	thwart	85	unsuspect	35	grim	85	suspicion
36	skill	86	revolution	36	beset	86	conjunct
37	limit	87	foster	37	partner	87	mine
38	lose	88	absurd	38	soft	88	cool
39	ironi	89	pleasantli	39	charg	89	empow
40	avoid	90	eager	40	alert	90	throw
41	pai	91	approv	41	poor	91	contribut
42	upset	92	convict	42	subpoena	92	harsh
43	depreci	93	refug	43	competitor	93	misfortun
44	rumor	94	traumat	44	suspens	94	inaccuraci
45	rampant	95	liabil	45	sluggish	95	shoddi
46	standstil	96	nomin	46	unpleas	96	meet
47	coher	97	downturn	47	suit	97	slowli
48	rival	98	incorrect	48	slow	98	overturn
49	exact	99	plain	49	experi	99	wari
50	close	100	defam	50	agil	100	sour

Table 3: Drivers of CAT
This table reports the panel regression of firm– and event–specific variables on $CAT_{a,i}(\cdot,\cdot)$. Variables are the return–on–asset (ROA_i), the quintile rank of the SUE_i ($QSUE_i$), the negative earnings surprise indicator ($SUE_i < 0$), the logarithm of the book–to–market ratio ($log(B_i/M_i)$), the logarithm of the market–capitalization ($log(M_i)$), $CAR_i(-1,1)$, and $CAR_i(-5,-2)$, as well as year–quarter and firm fixed effects. Our sample consists of $14,896$ earning announcement events for 597 non–financial S&P 500 constituents ranging from 2000 to 2016. We use a market–capitalization text–based normal tone factor model and the market model for the normal return. For each case, we use the estimation window $\tau \in \{t_i-30, t_i-6\}$, where t_i is the event date of event i, to estimate the normal tone and the normal return models. The daily tone is estimated using the Generalized Word Power methodology where the targeted aspect is the future performance of firms, proxied by the firms' stock returns. Significant explanatory variables are highlighted in gray and the level of significance is indicated by: * for 10%, ** for 5%, and *** for 1%. The standard errors are computed using double–clustered standard error (Petersen, 2009) and are reported in parenthesis below the parameter estimates.

Variable / Target	$CAT_i(-1,1)$	$CAT_i(2,5)$	$CAT_i(2,20)$
$QSUE_i$	0.003^{***}	-0.001	-0.001
	(0.001)	(0.0005)	(0.001)
$SUE_i < 0$	-0.002^{***}	-0.0003	0.0004
	(0.0004)	(0.0003)	(0.001)
$CAT_i(-5,-2)$	0.144^{***}		
	(0.013)		
$CAT_i(-1,1)$		0.254^{***}	0.615^{***}
		(0.034)	(0.084)
$CAR_i(-5,-2)$	0.001		
	(0.002)		
$CAR_i(-1,1)$	0.027^{***}	0.009^{***}	0.012^{***}
	(0.002)	(0.002)	(0.004)
ROA_i	0.011^{***}	-0.003	-0.007
	(0.004)	(0.005)	(0.009)
$log(B_i/M_i)$	-0.001^{***}	0.0004	0.001^{**}
	(0.0003)	(0.0003)	(0.001)
$log(M_i)$	0.0003	-0.0001	-0.0004
	(0.0003)	(0.0002)	(0.0005)
Firm Fixed Effect	Yes	Yes	Yes
Year–Quarter Fixed Effect	Yes	Yes	Yes
R^2 ($\times$ 100)	7.4	5.9	6.8

Table 4: CAT and CAR regression results

This table reports the various regression results regarding the predictive power of $CAT_i(\cdot,\cdot)$ on $CAR_i(\cdot,\cdot)$. We also differentiate between negative and positive CAT, that is, $CAT_i^-(-1,1)$ and $CAT_i^+(-1,1)$, respectively. Controls variables are $CAR_i(-1,1)$ for the short and long–term post–earnings–announcement abnormal returns (*i.e.*, $CAR_i(2,5)$ and $CAR_i(2,20)$), $CAR_i(-5,-2)$ for the $CAR_i(-1,1)$ regression, the return–on–asset (ROA_i), the quintile rank of the SUE_i ($QSUE_i$), the negative earnings surprise indicator ($SUE_i < 0$), the logarithm of the book–to–market ratio ($log(B_i/M_i)$), and the logarithm of the market–capitalization ($log(M_i)$), as well as year–quarter and firm fixed effects. Our sample consists of 17,157 earning announcement events for 597 non–financial S&P 500 constituents ranging from 2000 to 2016. We use a market–capitalization text–based normal tone factor model and the market model for the normal return. For each case, we use the estimation window $\tau \in \{t_i - 30, t_i - 6\}$, where t_i is the event date of event i, to estimate the normal tone and the normal return models. The daily tone is estimated using the Generalized Word Power methodology where the targeted aspect is the future performance of firms, proxied by the firms' stock returns. Significant explanatory variables are highlighted in gray and the level of significance is indicated by: * for 10%, ** for 5%, and *** for 1%. The standard errors are computed using double–clustered standard error (Petersen, 2009) and are reported in parenthesis below the parameter estimates.

Variable / Target	$CAR_i(-1,1)$			$CAR_i(2,5)$			$CAR_i(2,20)$		
$CAT_i(-5,-2)$	0.015	−0.013							
	(0.056)	(0.054)							
$CAT_i^-(-5,-2)$			−0.016						
			(0.095)						
$CAT_i^+(-5,-2)$			−0.008						
			(0.093)						
$CAT_i(-1,1)$				−0.046	-0.069^{*}		-0.178^{**}	-0.189^{**}	
				(0.035)	(0.037)		(0.090)	(0.087)	
$CAT_i^-(-1,1)$						−0.041			-0.322^{**}
						(0.068)			(0.145)
$CAT_i^+(-1,1)$						-0.102^{*}			−0.032
						(0.056)			(0.120)
$CAR_i(-5,-2)$		-0.036^{*}	-0.036^{*}						
		(0.021)	(0.021)						
$CAR_i(-1,1)$					0.032^{***}	0.032^{***}		0.101^{***}	0.102^{***}
					(0.010)	(0.010)		(0.021)	(0.021)
$QSUE_i$		0.068^{***}	0.068^{***}		0.0002	0.0003		-0.018^{***}	-0.018^{***}
		(0.003)	(0.003)		(0.002)	(0.002)		(0.005)	(0.005)
$SUE_i < 0$		0.004^{*}	0.004^{*}		0.0005	0.001		−0.002	−0.002
		(0.002)	(0.002)		(0.002)	(0.001)		(0.003)	(0.003)
ROA_i		0.054	0.054		0.057^{**}	0.057^{**}		0.146^{***}	0.146^{***}
		(0.035)	(0.035)		(0.027)	(0.027)		(0.051)	(0.051)
$log(B_i/M_i)$		-0.003^{**}	-0.003^{**}		0.006^{***}	0.006^{***}		0.017^{***}	0.017^{***}
		(0.001)	(0.001)		(0.001)	(0.001)		(0.002)	(0.002)
$log(M_i)$		-0.003^{**}	-0.003^{**}		-0.006^{***}	-0.006^{***}		-0.020^{***}	-0.020^{***}
		(0.001)	(0.001)		(0.001)	(0.001)		(0.003)	(0.003)
Firm Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year–Quarter Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R^2 ($\times$ 100)	0.0	8.2	8.2	0.01	1.5	1.5	0.1	3.6	3.7

Table 5: CAT contribution of source type and CAR regression results
This table reports the various regression results regarding the predictive power of source–type specific $CAT_i(\cdot,\cdot)$ contribution over $CAR_i(\cdot,\cdot)$. We consider newspaper (*i.e.*, $c_i(\cdot,\cdot)_{newspaper}$), newswire (*i.e.*, $c_i(\cdot,\cdot)_{newswire}$), and web publication (*i.e.*, $c_i(\cdot,\cdot)_{web\ publication}$) CAT percentage contribution. Control variables are $CAR_i(-1,1)$ for the short– and long–term post–earnings–announcement abnormal returns (*i.e.*, $CAR_i(2,5)$ and $CAR_i(2,20)$), $CAR_i(-5,-2)$ for the $CAR_i(-1,1)$ regression, the return–on–asset (ROA_i), the quintile rank of the SUE_i ($QSUE_i$), the negative earnings surprise indicator ($SUE_i < 0$), the logarithm of the book–to–market ratio ($log(B_i/M_i)$), and the logarithm of the market–capitalization ($log(M_i)$), as well as year–quarter and firm fixed effects. Our sample consists of 17,157 earning announcement events for 597 non–financial S&P 500 constituents ranging from 2000 to 2016. We use a market–capitalization text–based normal tone factor model and the market model for the normal return. For each case, we use the estimation window $\tau \in \{t_i - 30, t_i - 6\}$, where t_i is the event date of event i, to estimate the normal tone and the normal return models. The daily tone is estimated using the Generalized Word Power methodology where the targeted aspect is the future performance of firms, proxied by the firms' stock returns. Significant explanatory variables are highlighted in gray and the level of significance is indicated by: * for 10%, ** for 5%, and *** for 1%. The standard errors are computed using double–clustered standard error (Petersen, 2009) and are reported in parenthesis below the parameter estimates.

Variable / Target	$CAR_i(-1,1)$	$CAR_i(2,5)$	$CAR_i(2,20)$
$CAT_i(-5,-2) \times c_i(-5,-2)_{newspaper}$	-0.056		
	(0.121)		
$CAT_i(-5,-2) \times c_i(-5,-2)_{newswire}$	-0.074		
	(0.085)		
$CAT_i(-5,-2) \times c_i(-5,-2)_{web\ publication}$	0.098		
	(0.085)		
$CAT_i(-1,1) \times c_i(-1,1)_{newspaper}$		-0.012	-0.081
		(0.080)	(0.193)
$CAT_i(-1,1) \times c_i(-1,1)_{newswire}$		-0.050	-0.155
		(0.050)	(0.129)
$CAT_i(-1,1) \times c_i(-1,1)_{web\ publication}$		-0.110^{**}	-0.264^{***}
		(0.055)	(0.101)
$CAR_i(-5,-2)$	-0.036^{*}		
	(0.021)		
$CAR_i(-1,1)$		0.032^{***}	0.101^{***}
		(0.010)	(0.021)
ROA_i	0.054	0.057^{**}	0.146^{***}
	(0.035)	(0.027)	(0.051)
$QSUE_i$	0.068^{***}	0.0002	-0.018^{***}
	(0.003)	(0.002)	(0.005)
$SUE_i < 0$	0.004^{*}	0.0005	-0.002
	(0.002)	(0.002)	(0.003)
$log(B_i/M_i)$	-0.003^{**}	0.006^{***}	0.017^{***}
	(0.001)	(0.001)	(0.002)
$log(M_i)$	-0.003^{**}	-0.006^{***}	-0.020^{***}
	(0.001)	(0.001)	(0.003)
Firm Fixed Effect	Yes	Yes	Yes
Year–Quarter Fixed Effect	Yes	Yes	Yes
R^2 ($\times$ 100)	8.3	1.5	3.6

Figure 1: Abnormal tone event study timing information
This figure shows the time chart for the Generalized Word Power (*i.e.*, GWP model) tone estimation scheme (Panel A) and the event study methodology (Panel B). Panel A assumes that the tone model is estimated at the end of each year using an expanding window (with the sample data range indicated on each horizontal lines). Then, each tone model is used to estimate the daily tone of firms for the following year (OOS texts boxes). Panel B is based on the Abbvie Inc earnings announcement event for the second quarter of 2013. Similarly to our our empirical application, $L = 30$ days and $K = 5$ days. The abbreviation "TD" in Panel B is for trading days.

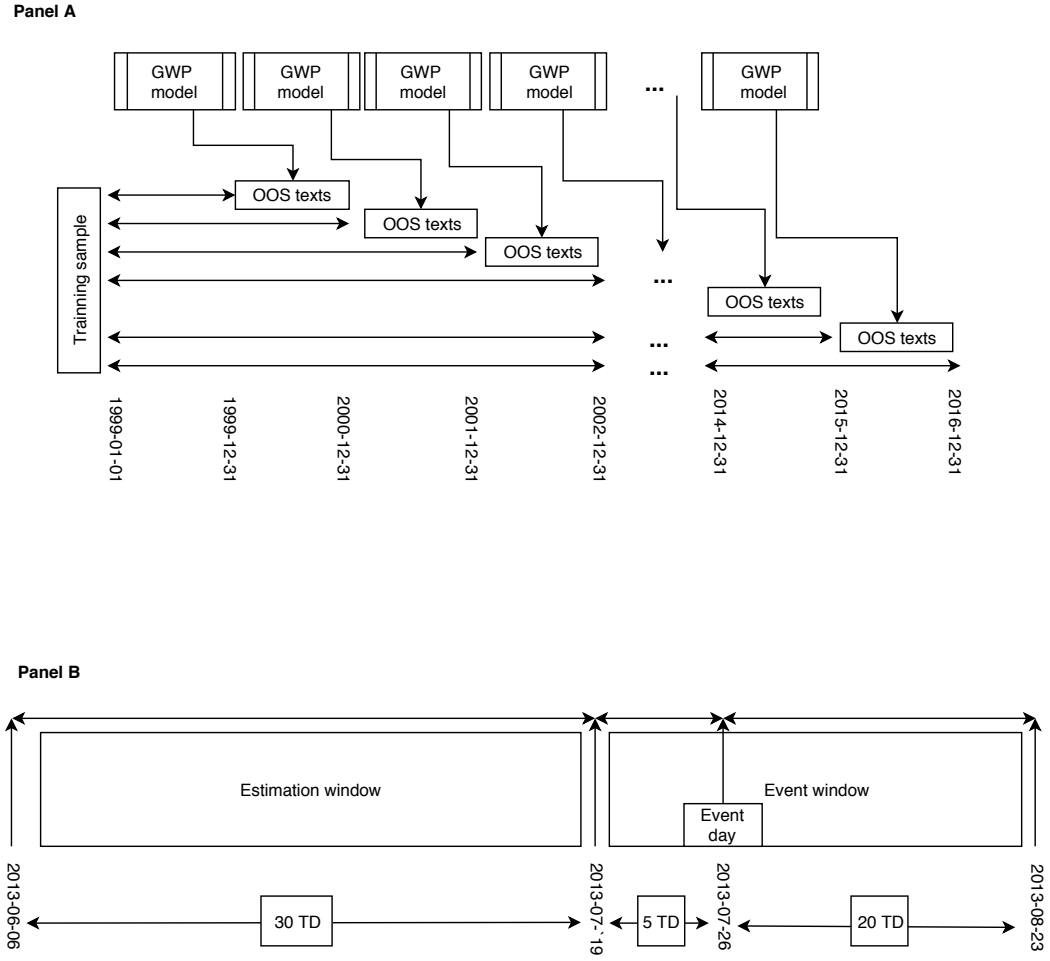


Figure 2: Number of documents per year
This figure shows the number of documents per year by source type.

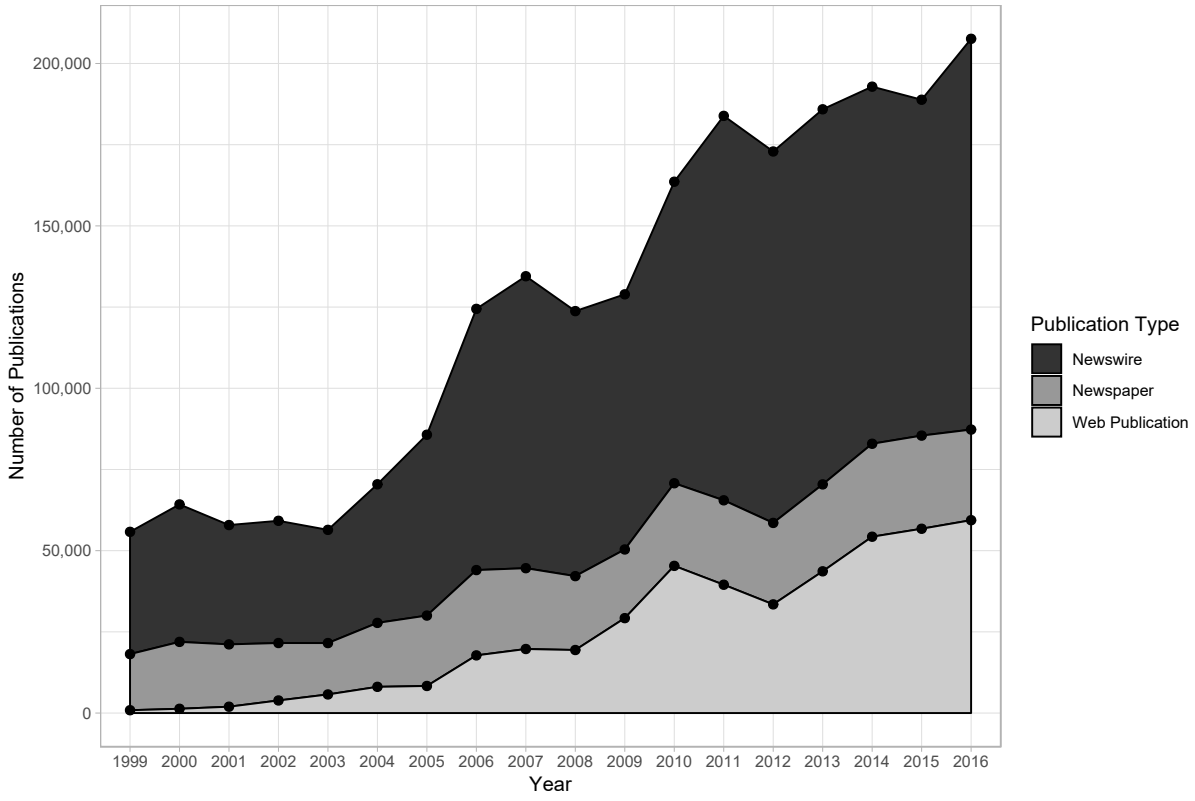


Figure 3: Average number of documents per day relative to the event date by publication type

This figure shows the average number of documents relative to the event date by publication types across all 14,896 quarterly earning announcement events. Day 0 indicates the event day.

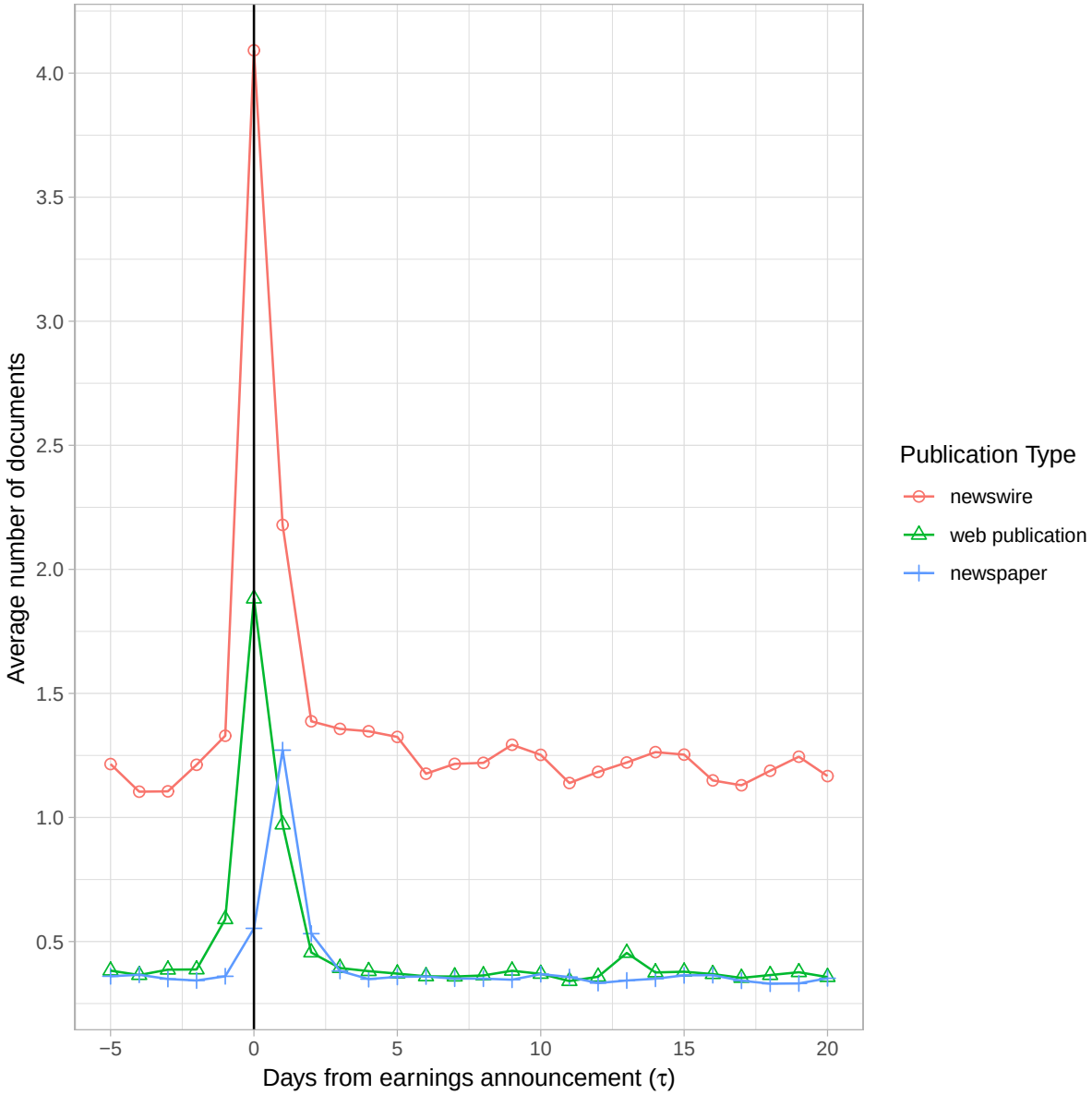


Figure 4: Average number of documents per day relative to the event date by average market capitalization buckets
This figure shows the average number of documents relative to the event date by average market capitalization buckets across all events in their respective buckets. Bucket #1 contains the firms with the lowest market capitalization average while bucket #4 contains the firms with the largest market capitalization average. Day 0 indicates the event day.

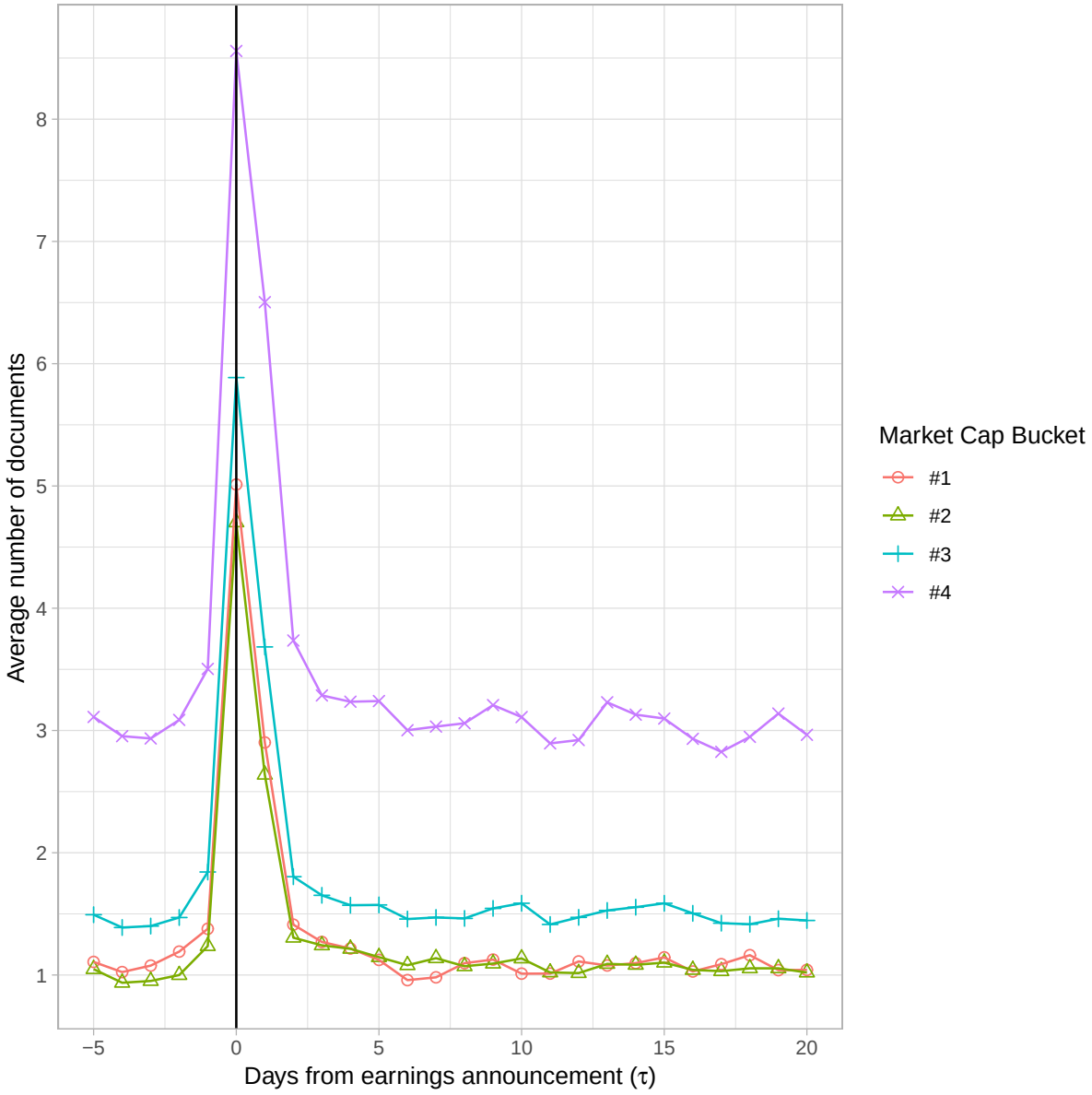


Figure 5: Average CAT per SUE bucket
This figure shows the evolution of the average CAT for five SUE buckets done over $14,896$ quarterly earning announcement events. The buckets are based on the quintiles of the SUE from the lowest (#1) to the highest (#5). The normal tone model is estimated over the estimation window $\tau \in \{t_i - 30, t_i - 6\}$. We use a market–capitalization text–based normal tone factor model. Day 0 indicates the event day.

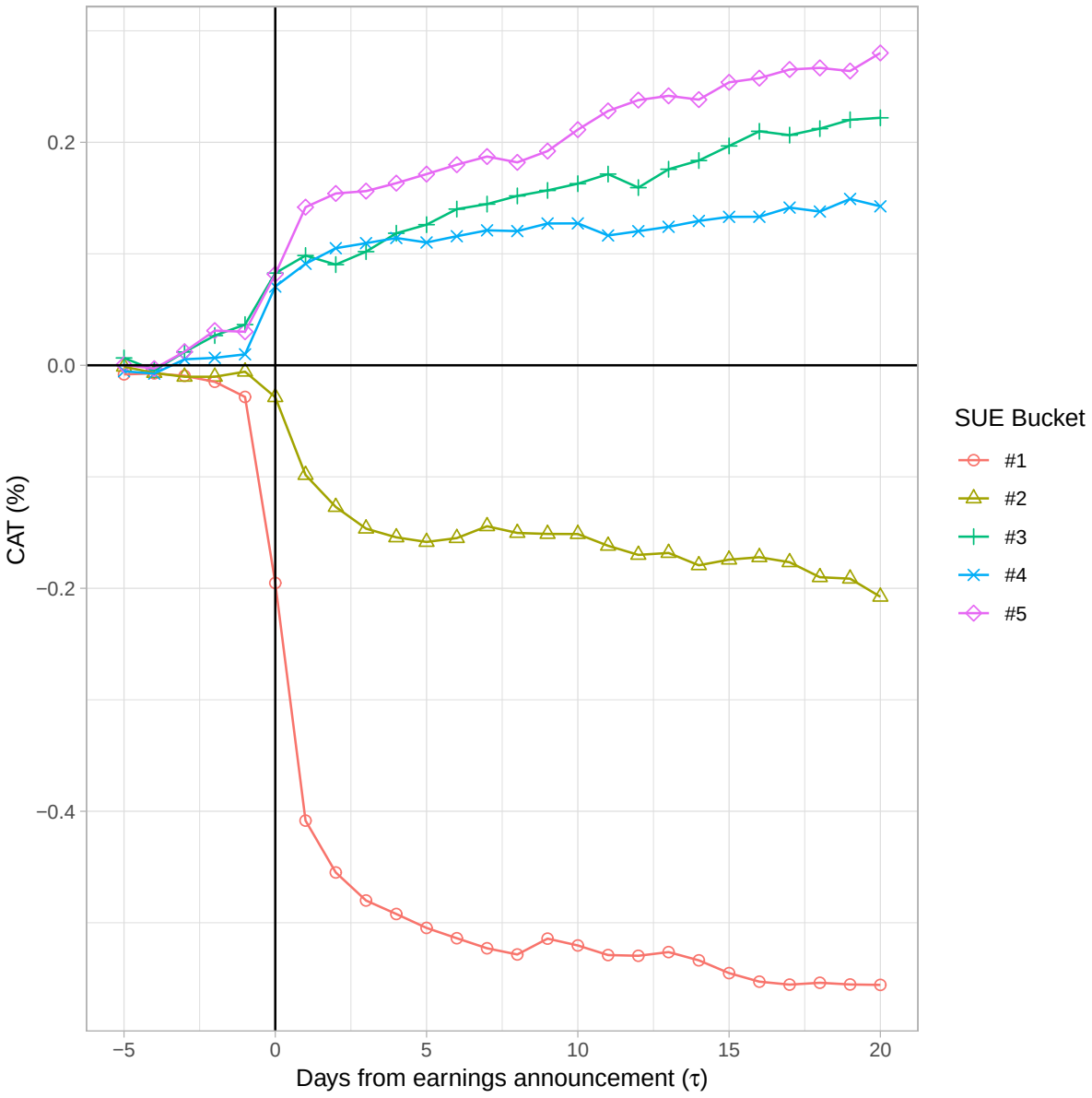


Figure 6: Average CAR per CAT bucket
This figure shows the evolution of the average CAR for five CAT buckets done over $17,157$ quarterly earning announcement events. The buckets are based on the quintiles of the $CAT_i(-1,1)$ from the lowest (#1) to the highest (#5). The normal tone and normal return models are estimated over the estimation window $\tau \in \{t_i - 30, t_i - 6\}$. We use a market–capitalization text–based normal tone factor model. The normal return is estimated using the market model where the market is the market index from the Kenneth R. French website. Day 0 indicates the event day.

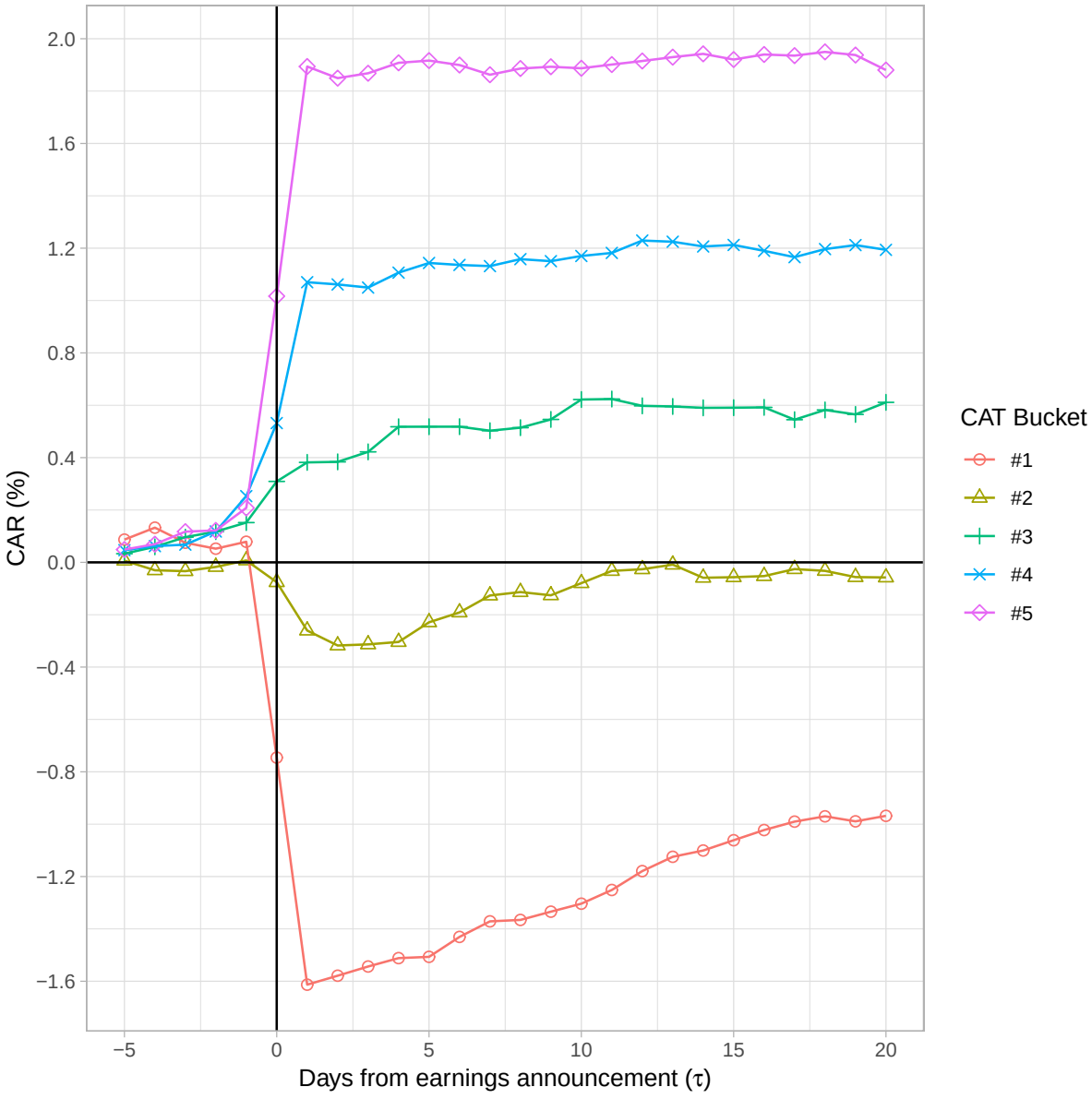
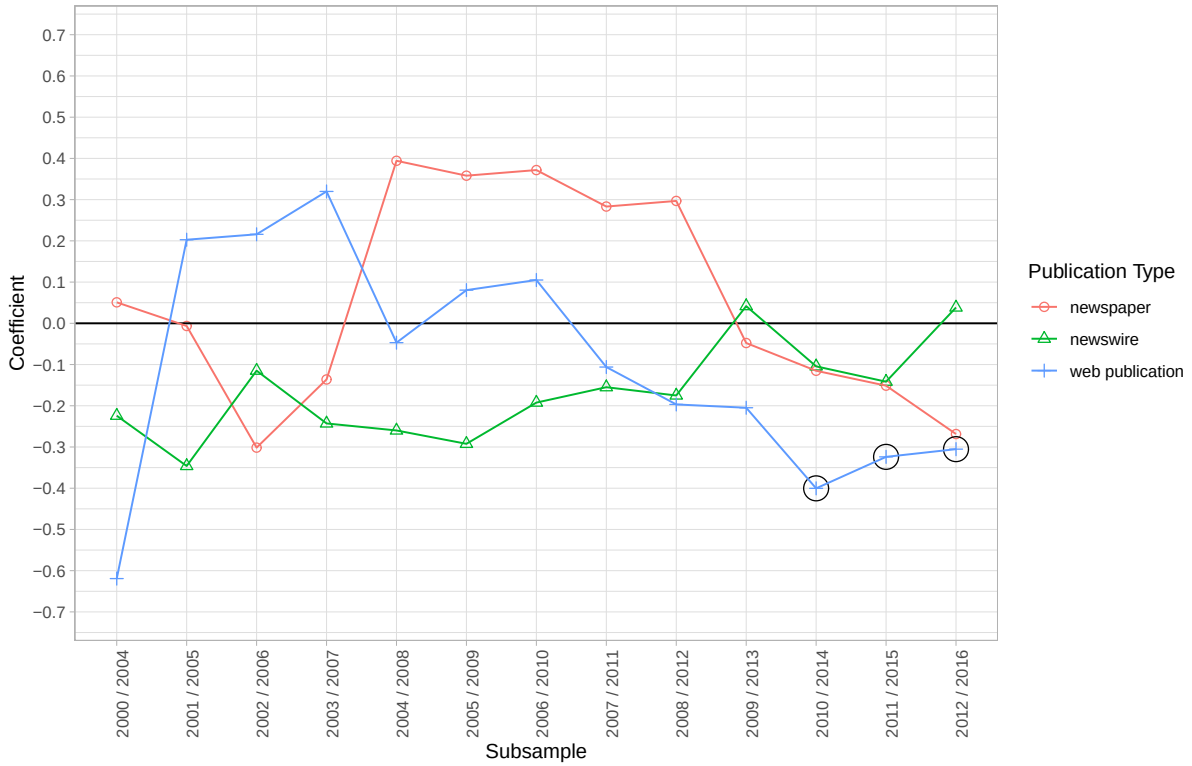


Figure 7: Time–varying CAT contribution coefficients
This figure shows the evolution of the coefficients regarding the regression of the CAT contribution of newswire, newspaper, and web publication on $CAR(2, 20)$(see full detail of the regression specification in Table 5). To extract the time varying–coefficient we perform a yearly rolling regression with a sample of five years. Significant coefficient points at the 5% level are enclosed by a black circle.



Appendix A. Corpus analysis

It is interesting to look at the topics written about in our corpus to validate if the content is relevant to our analysis. Table A.6 reports the most frequently encountered topics present in our corpus. Most of the topics indicate that our corpus is largely composed of documents that are related to earnings announcements. For example, some interesting topics are the *"COMPANY EARNINGS"* (215,728 documents), *"FINANCIAL RESULTS"* (91,787 documents), *"COMPANY PROFITS"* (69,525 documents), and *"INDUSTRY ANALYST"* (42,588 documents).

[Insert Table A.6 about here.]

It is also insightful to look at the publisher of the majority of the texts as it validates if our corpus contains varied and respectable sources. Table A.7 reports the top 50 sources along with their source types, number of publications, and a brief description of the sources. In the newswire category, the top sources are the *"PR newswires"* (184,636 documents) followed by *"Business wire"* (156,246 documents), in the newspaper category the top source is *"The New York Times"* (15,556 documents) followed by the *"Investor's Business Daily"* (15,484 documents), and finally, in the web publication category the top source is *"Comtex News Network, Inc"* (120,361 documents) followed by *"TendersInfo – News"* (44,030 documents).

[Insert Table A.7 about here.]

Table A.6: Top 50 topics in the corpus
This table reports the most frequent article topics along with the number of documents related to them.

Topic	#	Topic	#
PRESS RELEASES	478,104	TELECOMMUNICATIONS SERVICES	61,118
COMPANY EARNINGS	215,728	SHAREHOLDERS	57,266
PATENTS	202,214	HOLDING COMPANIES	56,296
EXECUTIVES	181,946	SALES FIGURES	55,837
COMPUTER SOFTWARE	116,060	BOARDS OF DIRECTORS	54,841
PHARMACEUTICALS INDUSTRY	109,757	UTILITIES INDUSTRY	53,089
AGREEMENTS	108,391	APPOINTMENTS	51,769
STOCK EXCHANGES	108,214	WIRELESS INDUSTRY	51,718
INTERIM FINANCIAL RESULTS	102,982	SUITS & CLAIMS	51,672
FINANCIAL RESULTS	91,787	CONSUMERS	48,924
INTERNET & WWW	89,811	MANAGERS & SUPERVISORS	46,077
FINANCIAL PERFORMANCE & REPORTS	88,470	HEALTH CARE	44,287
MOBILE & CELLULAR TELEPHONES	82,264	NATURAL GAS PRODUCTS	44,154
OIL & GAS INDUSTRY	81,067	WEBCASTS	43,955
RETAILERS	75,874	NATURAL GAS	42,801
ELECTRIC POWER PLANTS	72,979	INDUSTRY ANALYSTS	42,588
NATURAL GAS & ELECTRIC UTILITIES	72,349	APPROVALS	42,359
COMPUTER NETWORKS	71,540	US FEDERAL GOVERNMENT	42,230
PRICES	70,372	INVESTIGATIONS	40,395
COMPANY PROFITS	69,525	ASSOCIATIONS & ORGANIZATIONS	39,591
EARNINGS PER SHARE	69,171	FINANCIAL RATINGS	39,317
STOCK INDEXES	66,264	NEWS BRIEFS	39,311
LITIGATION	62,866	ELECTRIC POWER INDUSTRY	39,115
PHARMACEUTICAL PREPARATION MFG	61,334	ENERGY & UTILITY LAW	39,070
SMARTPHONES	61,142	COMMON STOCK	38,516

Table A.7: Top 50 sources in the corpus

This table report the source, the source type, the number of publications, and the description of the top 50 sources of textual documents in our corpus. Classification of the type of source is given in the LexisNexis database. The sources' descriptions are a summary or the full description reported on the LexisNexis website.

Source	Type	#	Description
PR Newswire	newswire	184,636	PR Newswire plays a key role in the dissemination of time-critical financial information. The newswire delivers full-text, unedited news releases as written by the originators.
Business Wire	newswire	156,246	Business Wire transmits to the media the full text of news releases issued by corporations and other organizations. The news sources can be from the banking industry, entertainment, aviation or many other industries.
US Fed News	newswire	153,592	US Fed News is a daily publication comprising a comprehensive compilation of publicly distributed government information, department press releases, federal/appellate/district court rulings and data related to federal/state business and grant opportunities.
Comtex News Network, Inc.	web publication	120,361	Comtex News Network, Inc. is a news aggregation service that has provides select content from key sources. With a specialization in the financial news and content marketplace, Comtex receives, enhances, combines and filters news and content collected from national and international news bureaus, agencies and publications.
Targeted News Service	newswire	82,785	Targeted News Service is a national news and editorial services company that produces news and information for America's newspapers, databases, and services businesses and consumers directly. Coverage areas include news on federal government and congressional activities, regulation, nation's foundations, corporations, educational institutions, and native Americans.
The Associated Press State & Local Wire	newswire	55,565	The Associated Press State & Local Wire source includes news from all 50 states, drawing news stories from 143 U.S. bureaus and from Associated Press member newspapers and broadcasters. The wire provides coverage on a variety of regional topics such as information on state capitols, legislation and politics, local regional and state sports, cross-state issues, news analysis, and entertainment.
MT Newswires	newswire	55,056	MT Newswires is a source of original, forward-looking, multi-asset class news and analysis of developed capital markets and economies globally.
The Associated Press	newswire	48,559	The Associated Press is the the oldest and largest news service in the world. News collected by the Associated Press republished by more than 1,300 newspapers and broadcasters.
TendersInfo – News	web publication	44,030	TendersInfo – News is a comprehensive intra-daily source of business and industry news from around the globe, keeping its readers updated on joint ventures, memorandum of understanding, project launches, contract awards, budget allocations, financial status, stock updates, etc.
US Official News	newswire	41,398	US Official News is a comprehensive source of major happenings, developments and full–text public announcements made through press releases, statements and other documents issued by various federal and state governments. Coverage includes the latest reports on stock exchange filings, patents, financial reports, economic surveys, inflation index, banking and economic performance reviews, and sector–wise export/import trade as well as parliamentary news including congressional legislation updates, bills, laws, business regulations and local government news.

Continued on next page

Table A.7 – *Continued from previous page*

Source	Type	#	Description
M2 PressWIRE	newswire	41,119	M2 PressWIRE is the world's third-largest electronic press release distribution service and UK/Europe's largest. M2 PressWIRE's breadth and depth of coverage draws acclaim and high degrees of external use throughout the world.
RTTNews	web publication	37,972	RTTNews provides comprehensive corporate news coverage of companies ranging from Blue Chips to Penny Stocks. In addition to reporting financial figures, RTT also provides comments from prominent world leaders, U.S. & International political and general news that can affect the markets, Interest Rate changes, important speeches/comments from US Federal Reserve, ECB, BoE and other central banks from around the world.
ENP Newswire	newswire	33,475	ENP Newswire is a global press release distribution service that allows companies to distribute their news both via the world's largest news agencies and directly to individual journalists.
Marketwire	newswire	27,094	Marketwire is the leading Internet-based distributor of direct company news. Marketwire distributes corporate news, including press releases, financial announcements, and other time-critical business communications.
Benzinga.com	web publication	23,742	Benzinga.com is a news and analysis service that focuses on global markets providing original, accurate and timely global financial content from industry experts and experienced analysts while also covering the news of the day.
News Bites – people in business	newswire	20,376	News Bites – people in business collects announcements for stock exchanges globally concerning company executives (e.g. CEOs, CFOs, Directors & Company Secretaries). These include appointments, resignations, dismissals, addresses to shareholders, chairman's statements, directors' buying selling of shares, financial results, profit warnings, appointment anniversary and others.
Associated Press – Financial News	newswire	19,880	Associated Press – Financial News provides detailed coverage of the top 1,000 US companies plus major international corporations. Coverage includes world market news, quarterly earnings announcements, executive changes, regulatory actions, mergers and acquisitions, and new product developments.
Market News Publishing	newswire	19,248	Market News Publishing supplies news, commentaries, analysis and related information on public companies and exchanges in Canada and the U.S.
M2 EquityBites	newswire	17,440	M2 EquityBites offers articles relating to companies in the FTSE 100, FTSE 250 and the S&P 500 exchanges.
Wireless News	newswire	16,340	Wireless News delivers a daily roundup designed to keep readers on the leading edge of the wireless industry. Special focus includes emerging technologies, convergence, international news, regulations and policy, mergers and acquisitions, legal issues, emerging technologies, trends, and the youth sector.
Global Insight	web publication	16,021	Global Insight is a leading global provider of business-critical information and is relied on by thousands of executives in hundreds of multinational corporations, financial institutions, and governments throughout the world. Their political, economic and security risk analysis of 185 markets.
The New York Times	newspaper	15,557	The New York Times bears the reputation of being the United States' unofficial newspaper of record. Comprehensive coverage of national, foreign, business and local news comes from The Times' extensive foreign news network and bureaus around the United States.

Continued on next page

Table A.7 – *Continued from previous page*

Source	Type	#	Description
Investor's Business Daily	newspaper	15,384	Investor's Business Daily is a national business and financial newspaper that covers all the major business and economic news of the day.
The FinancialWire	newswire	14,207	The FinancialWire news coverage extends to the 100 Most Active NASDAQ and NYSE companies, some 1000 companies in the Investrend platforms, the only daily wrap-up of the largest repository of standards-based professional independent research, forums, webcasts, official filings and events not generally included in other newswires' offerings, as well as socio-political perspectives on today's financial marketplace.
States News Service	newswire	13,336	States News Service reports on events in Washington that affect programs or projects of major interest in individual cities and states. States News Service tracks federal laws that directly impact a state's key industries. States News Service works principally for newspapers, although its Washington data collection and reporting are now available to corporations, government agencies, and databanks.
The Deal Pipeline	web publication	12,842	The Deal Pipeline is a premier news service covering the deal economy. Deal news and analysis combined with coverage of the people and the personalities behind the deals. Extensive coverage of Venture Capital, Mergers and Acquisitions, IPO's, Private Equity and Bankruptcy.
Agence France Presse	newswire	12,629	Agence France Presse is the world's oldest news agency. AFP's Europe coverage is outstanding, it's reporting from Africa is renowned and its Latin American correspondence comprehensive. AFP also covers the Middle East, Asia, and the Pacific Rim.
The Associated Press International	newswire	12,173	The Associated Press International Service covers breaking news from around the world. It provides top international business, general and sports developments continuously, 24 hours each day, seven days a week.
AFX International Focus	newswire	11,675	AFX International Focus covers European economic, financial, corporate, and general news. It includes news reporting from the USA, Japan, and other international centers relevant to European markets.
CashFlowNews	newswire	11,354	CashFlowNews is the primary "cash flow" news source for over 10,000 public companies, monitoring and reporting on EBITDA, Cash Flow from Operations, and Free Cash Flow.
United Press International	newswire	11,206	United Press International provides readers with industry level analytical documents written by in-house experts. United Press International experts explain the meaning of the news as well as reporting it. Covering a cross-section of news, business, health, and politics, UPI covers the day's current issues from multiple angles while looking ahead to major issues of tomorrow.
Progressive Media – Company News	newspaper	10,765	Progressive Media Company News is a collection of all the latest news, comments and industry information. It covers the following industry: Pharmaceutical, Technology, Banking, Insurance, Food, Drinks, Automotive, Logistics, Medical Devices, Clean Technology, Energy, Retail, and Packaging.
The Globe and Mail	newspaper	10,666	The Globe and Mail, and its daily Report on Business section are the definitive sources for Canadian news and business reporting.
National posts; Financial post & FP investing	newspaper	10,213	Contains articles from the Financial Post, Financial Business Post & FP Investing sections of the National Post.

Continued on next page

Table A.7 – *Continued from previous page*

Source	Type	#	Description
Plus Company updates	newswire	9,728	Plus Company updates is a comprehensive source of company information, financial statements, and corporate announcements, etc. It covers Pakistan, India, Bangladesh, Brazil, Singapore, Malaysia, UAE, Philipines, Taiwan, and other countries of the world.
eWEEK.com	web publication	9,587	eWEEK.com provides the first source for breaking news, vendor analysis and critical examination of recent deployments.
Aggregated Regulatory News Service	newswire	9,490	The Aggregated Regulatory News Service (ARNS) provides full regulatory news relating to companies listed with the London Stock Exchange. ARNS provides all the information companies are required to make public according to FSA regulations. This includes results, new issues, shareholdings, board changes and any other information, which may affect the company's share price.
SNL Kagan Media & Communications Report	newswire	9,431	Offers timely and comprehensive news on telecom, satellite, wireless, publishing, cable, entertainment, new media, broadcasting, and advertising.
The Mercury News	newspaper	9,388	The Mercury News is a general circulation daily newspaper providing local, national, and international news coverage.
Entertainment Close-Up	newswire	8,511	Entertainment Close–Up covers the dealmakers, companies, products, start-ups, technologies, and trends that are transforming the world of arts and entertainment. Special focus includes global partnerships, emerging technology, legal issues, the convergence of old and new media, telematics, standards, digital rights issues, and youth culture.
Professional Services Close–Up	newswire	8,108	Professional Services Close–Up covers the firms, products, services, start-ups, technologies, and trends that make up this growing market. Particular focus includes new products, legal services and issues, technology support, brand loyalty, customer satisfaction, and trends as well as emerging technologies and opportunities.
News Bites – US Markets	newswire	7,834	News Bites US Markets reports on the US Stock exchange markets, by index and sector, to identify price and trading volume changes, including bullish and bearish signals and market action tables. It also reports on the US Sectors that have been markedly active in the day, by price, volume and value changes, US Stocks that have been notably active in the day, by price, volume and value changes, substantial shareholder changes and directors' dealings.
Associated Press Online	newswire	7,454	Associated Press Online is a news service tailored specifically for use in databases or similar online environments. The service is comprised of the top national, international, Washington, financial and sports news on a given day. Stories cover various topics including Politics, Business, Wall Street, Sports, Entertainment and Weather, and are transmitted 24 hours a day, seven days a week.
The Washington Post	newspaper	7,249	The Washington Post is one of the few U.S. newspapers with a serious interest in foreign news, deploying correspondents from its 16 foreign bureaus to produce in-depth articles from the world's hot spots.
Class Action Reporter	newswire	7125	Class Action Reporter covers all significant class action litigation throughout the United States.
M&A Navigator	newswire	6,622	M&A Navigator focuses on global merger and acquisition activity, ranging from corporate mergers and acquisitions to private equity sponsored leveraged buyouts, joint ventures, venture capital investments, stake-building, and restructuring.

Continued on next page

Table A.7 – *Continued from previous page*

Source	Type	#	Description
The Daily PAK banker	newspaper	6,394	The Daily PAK banker is published in Pakistan's three major cities, *i.e.*, Karachi, Lahore, and Islamabad. The Daily PAK banker covers banking and financial sectors of Pakistan exclusively and also the major developments of the world banking and financial scenario.
Product News	web publication	6,068	Product News is an information service designed to keep industrial professionals informed on hundreds of new products announced each week.
USA Today	newspaper	5,962	USA TODAY is the one of the US most–read newspaper with more than 6.3 million readers. USA TODAY provides outstanding coverage of issues and events from across the US and the world.
The Australian	newspaper	5,922	Includes The Australian newspaper, The Weekend Australian newspaper, and its inserted Australian Magazine. The Australian is a national morning broadsheet newspaper which is published six days a week.

Appendix B. Generalized Word Power tone analysis

In this section, we analyze the relationship between the tone measure computed using the Generalized Word Power methodology and stock returns. Additionally, we analyze the estimation scheme, in particular, the weight series estimated at the end of each year. Finally, we compare the score of the individual sentiment words to the original lexicon polarity of the sentiment words.

Appendix B.1. Generalized Word Power tone and contemporaneous stock returns

We analyze the relationship between the estimated daily Generalized Word Power tone estimates and the corresponding firm stock returns. To that end, we split our sample of $711,112$ tone–return observations into ten buckets. Each bucket contains 10% of the stock return data. The thresholds for the attribution in each bucket corresponds to Generalized Word Power daily tone deciles computed from the entire sample of Generalized Word Power daily tone observations. Thus, bucket #1 contains the returns with the lowest daily tone observations while bucket #10 contains the returns with the highest daily tone observations. Figure B.8 reports the average return for each bucket.

[Insert Figure B.8 about here.]

We can observe a positive relationship between the average return and the Generalized Word Power tone. The lowest bucket reports an average return of about -0.85% while the highest bucket reports an average return of about 0.40%. Thus, this result provides some evidence that the tone estimate used in our event study can be interpreted as the media implied expected returns. Also, note that the figure indicates that negative tone observations have a bigger effect on stock returns than positive tone observations. A similar asymmetric effect has been observed in Akhtar et al. (2012) using the University of Michigan consumer sentiment index. Particularly, they observe that a negative market effect occurs upon the release of bad sentiment news but no positive market effect is observed for good news. Our results differ as we observe a positive effect on good news as well, but less prominent than for negative news.

Appendix B.2. Generalized Word Power scores compared to the original lexicons

It is of interest to also analyze the deviation between the Generalized Word Power scores and the original polarity (positive or negative) of those words present in the original lexicon. To that

end, we computed the kernel density of the Generalized Word Power scores conditional on the sign of the polarity of the words present in the original lexicon. Figure B.9 reports the results for the Loughran and McDonald (2011) lexicon (top plot) and the General Inquirer lexicon (bottom plot) when using the Word Power weights estimated for the year 2016.

[Insert Figure B.9 about here.]

We observe that the Generalized Word Power methodology provides scores that are more similar to the Loughran and McDonald (2011) lexicon than to the General Inquirer Lexicon. Indeed, the kernel density of the Generalized Word Power weights for positive (negative) polarity words given by the Loughran and McDonald (2011) lexicon is highly positively (negatively) skewed when compared to the density conditional on the polarity of the words present in the General Inquirer lexicon. While the Loughran and McDonald (2011) lexicon is aimed at capturing the polarity of 10–K documents and not aimed at obtaining the media implied expected return, it is still what is considered a financial domain–specific sentiment lexicon. This is not the case for the General Inquirer lexicon. Therefore, the observation that the weights estimated by the Generalized Word Power methodology are more similar to the Loughran and McDonald (2011) lexicon word polarities than to the General Inquirer lexicon word polarities encourages us to believe that the tone measures generated by Generalized Word Power weights capture financial information.

Figure B.8: Generalized Word Power daily tone and stock return relationship
This figure shows the average return for the Generalized Word Power daily tone buckets. Bucket #1 consists of returns with the most negative contemporaneous Generalized Word Power daily tone and bucket #10 consists of returns with the most positive contemporaneous Generalized Word Power daily tone. Buckets thresholds are set according to the Generalized Word Power daily tone deciles computed from our sample of 774, 934 daily tone observations.

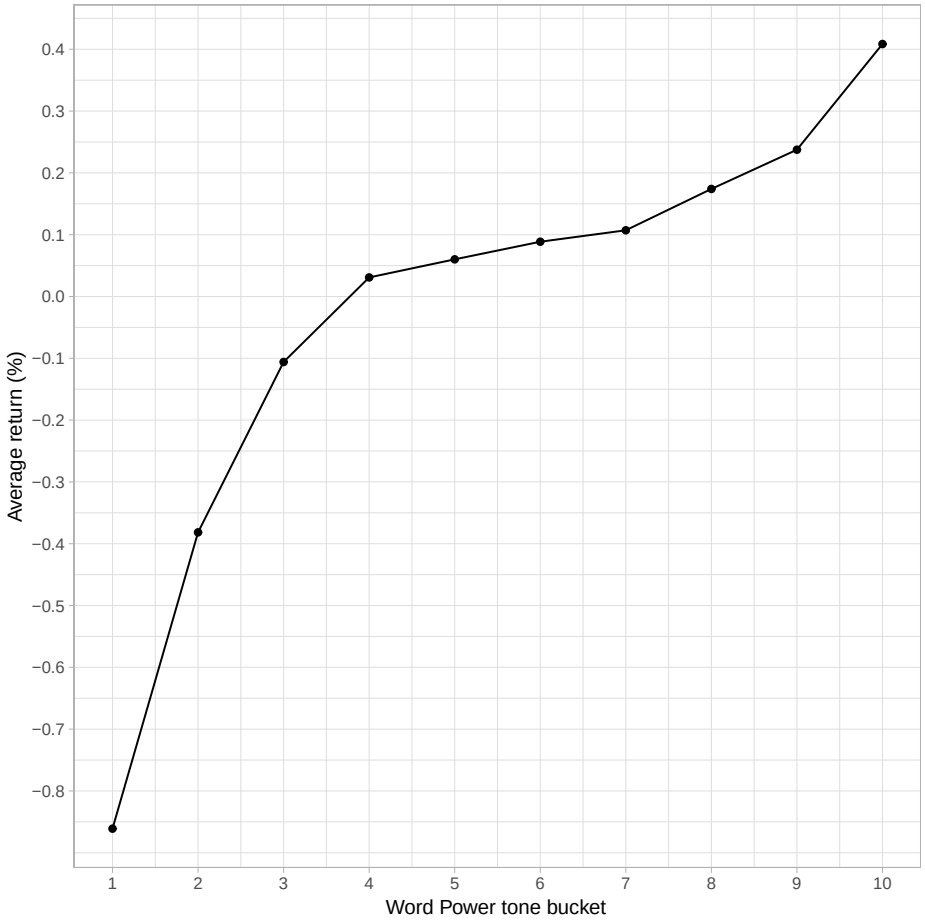


Figure B.9: Generalized Word Power score densities
This figure shows the Generalized Word Power scores densities for the words that are present in the original lexicons. The scores densities are conditional on the original lexicon polarity (*i.e.*, positive or negative). We use the scores estimated for the year 2016, that is, estimated using all data prior to the year 2016. The top plot shows the density for the scores of the sentiment words in the Loughran and McDonald (2011) lexicon while the bottom plot shows the density for the scores of the sentiment words in the General Inquirer lexicon.