Essays in Macroeconomics and Machine Learning

Julian Ashwin

Nuffield College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy in Economics

Michaelmas 2021

Abstract

The unifying theme of this thesis is the use of techniques from machine learning and data science to address questions in macroeconomics. It makes both theoretical contributions by applying neural networks as a learning algorithm in models that are indeterminate under rational expectations, and empirical contributions by developing and applying natural language processing methods to datasets including news media articles and central bank communication.

The first Chapter, **Resolving Indeterminacy with Neural Network Learning: Sinks become Sources**, aims to make a theoretical contribution to the literature on indeterminacy in rational expectations models. Indeterminacy (i.e. non-uniqueness of equilibrium under rational expectations) is a pervasive and often neglected challenge for macroeconomists. Previous literature on learning in macroeconomics has shown that the equilibria in linear indeterminate models are almost always not learnable. This Chapter examines the equilibrium sophisticated learning agents converge to in models where there is indeterminacy, but it is bounded. This neural network learning converges to a stable equilibrium, in which indeterminate regions are sources and determinate regions are sinks. Furthermore, this equilibrium is consistent with Rational Expectations. There are multiple steady states and agents have correct beliefs about transitions between them, which means that transitory shocks can have permanent effects. This is demonstrated with an application to a well-known example of indeterminacy: a New Keynesian model with a Zero Lower Bound in interest rates.

The resolution to the challenge of indeterminacy presented here is plausible, as it's based on learnability, and also appealing because the identified equilibrium is globally stable, can have multiple steady states with well-defined transitions between them, and passes tests for rationality. It also demonstrates the value of using machine learning, as the neural network acts as a very flexible function approximator that is fast to train, allowing the use of learnability as an equilibrium selection device in non-linear models. Alternative learning algorithms like Recursive Least Squares yield qualitatively similar results, but are not sufficiently flexible to pass tests of rationality.

Chapter 2, **Bayesian Topic Regression for Causal Inference** has a more methodological focus, developing Bayesian Topic Regression, a model for causal inference with text data. This methodology is then applied in Chapter 3 to help identify a potentially causal effect of media coverage on stock price volatility. The Bayesian Topic Regression

model jointly estimates topics in text documents and a regression using these topics and associated numerical data to predict a response variable. As well as showing that performing text feature extraction and prediction in separate stages can lead to incorrect inference, we benchmark our model on two real-world customer review datasets and show markedly improved out-of-sample prediction in comparison to competing approaches.

Chapters 3 and 4 use text data to address to empirical questions relating to how agents in the economy acquire information and on what they focus their attention. In Chapter 3 **Financial news media and volatility: is there more to newspapers that news?** identifies a co-movement media coverage in the *Financial Times* newspaper and a firm's intra-day stock price volatility is identified. I argue that part of this co-movement is causal, relying on an identification strategy based on the publication time of the newspaper, controlling for persistence and anticipation effects as well as the content of articles using the Bayesian Topic Regression introduced in Chapter 2. These results are consistent with a salience-based view of the media's role in financial markets: media coverage does not (only) provide information, but influences where investors choose to direct their focus. This identified effect also has interesting spillovers to firms in sectors that are linked by the production network.

In Chapter 4 **The Shifting focus of Central Bankers.** I use an unsupervised topic model to quantify the focus of central bank communication and economic news media. I offer an explanation for the variation of this focus over time, and identify a robust co-movement between central bank and media focus. A model of multidimensional uncertainty and limited attention is proposed to explain the shifting focus of central bank communication. Evidence from the Survey of Professional Forecasters is used to support this explanation, showing that focus shifts to cover variables about which there is greater uncertainty. An event study approach is used to show a potentially causal influence of Federal Reserve communication on the focus of US news media and on the communication of other central banks.

Acknowledgements

This thesis was made possible by the help and support of a great many people.

I would firstly like to thank my supervisor Martin Ellison for his advice and inspiration over the last six years, as well as Paul Beaudry and Stephen Hansen for regular invaluable discussions. Many other members of the Macro group at Oxford also provided hugely useful feedback and encouragement.

I am also indebted to the many people who took time to discuss my various projects with me and provide valuable insight. These include many visiting seminar speakers, discussants and participants at conferences, and colleagues at the Cabinet Office, ECB and Goldman Sachs. A special mention also goes to Max Ahrens as without our countless evenings in Eagle House, powered by leftover sandwiches, Chapter 2 might never have been finished.

I would also like to thank the UK's Economic and Social Research Council and the David Walton Distinguished Doctoral Scholarship for financial support throughout my MPhil and DPhil, without which I would not have had the luxury of devoting my time to research.

I have met many incredible people throughout the MPhil and DPhil, to whom I'm indebted on both a personal and academic level. In particular both Alistair and Merrilyn have been hugely helpful. We started our postgraduate journeys together and I am very grateful for the emotional and academic support they have provided over the years. I would also like to thank the friends from times at Hertford and Exmouth, who's company I continue to enjoy greatly.

My parents Peter and Angela have been an utterly reliable and unconditional source of support and advice throughout my studies. I am eternally grateful for all they have done for me.

Most importantly, meeting Eva has undoubtedly been the best thing that has ever happened to me. I would like to thank her for her love and support in both the good and not so good moments, as well as for regularly correcting my maths.

Contents

List of Tables

List of Figures

Chapter 1

Resolving Indeterminacy with Neural Network Learning: Sinks become Sources

Abstract

This paper uses neural network learning to identify learnable rational expectations equilibria in environments where equilibrium behaviour is indeterminate under rational expectations in some regions of the state space. The identified rational expectations equilibria act as sources in locally indeterminate regions, meaning that endogenous variables are repelled and spend very little time in their neighbourhoods. These results contrast sharply with the perfect-foresight behaviour in these environments, in which locally indeterminate regions act as a sink, attracting endogenous variables to their neighbourhood. Previous work has analysed such systems under perfect foresight or perturbation around steady states, discussing behaviour in the locally indeterminate region as acting as a sink. Such emphasis would appear to be misplaced, since under rational expectations the locally indeterminate region is a source not a sink. It is also shown that more familiar learning algorithms, such as recursive least square will converge to qualitatively similar equilibria, but the flexibility of a neural network is necessary for this equilibrium to be consistent with rational expectations. These results have potentially important implications in a wide range of contexts, as demonstrated by applying neural network learning to a simple New Keynesian model in which monetary policy is constrained by a Zero Lower Bound. If the indeterminacy due to this constraint on policy is bounded, agents can learn a fully-stochastic equilibrium with multiple steady states where transitory shocks can have permanent effects.

1.1 Introduction

The existence of multiple stationary solutions under rational expectations has long been a potential feature of macroeconomic models. A recent increased focus on non-linear models with multiple steady states has brought this sharply into focus. In many cases the behaviour of such a system will be indeterminate under rational expectations in the neighbourhood of some of these steady states. In the absence of a positive predictions about equilibrium behaviour, a model has limited practical value to economists.

This paper uses learning to provide a positive prediction for equilibrium behaviour in indeterminate models: steady states around which rational expectations behaviour is indeterminate should be treated as unstable and explosive. Furthermore, in models with multiple steady states where agents learn using a sufficiently flexible neural network, explosive paths out of these indeterminate steady states can form a "Learnable-Rational Expectations Equilibrium". This Learnable-REE has the properties of a rational expectations solution, as demonstrated by a formal test on agents' forecast errors, but behaves very differently to any of paths identified under perfect foresight or through perturbation analysis. In this equilibrium, agents are aware of multiple steady states and have well-defined and accurate beliefs about the transitions between steady states. This means that transitory shocks can have potentially permanent effects by moving the system from one steady state to another.

Indeterminacy under rational expectations is a feature of macroeconomic models in a wide range of contexts, and the results presented in this paper have important implications in all of these. Indeterminacy will often arise when there is some non-convexity or complementarity across agents. Well known examples include: increasing returns to scale in a production function (Benhabib and Farmer, 1994; Cazzavillan et al., 1998); a passive monetary policy rule (Clarida et al., 2000; Lubik and Schorfheide, 2004); an Effective Lower Bound on monetary policy (Benhabib et al., 2001; Christiano et al., 2018; Eggertsson et al., 2019); externalities in a search and matching framework in the goods market (Huo and Ríos-Rull, 2013; Kaplan and Menzio, 2016), labour market (Eeckhout and Lindenlaub, 2019) or inter-firm market (Fernandez-Villaverde et al., 2020); endogenous markups generating non-convex marginal revenue product of capital (Gali, 1995); complementarities in R&D investment (Greiner and Bondarev, 2017); correlated private information about asset payoffs (Manzano and Vives, 2011); non-convex relationships between economic activity and ecological systems (Mäler et al., 2003); positive externalities in production (Krugman, 1991); feedback between government debt and interest rate

through liquidity constraint (Angeletos et al., 2019).

The non-convexity or complementarity that drives indeterminacy will often be locally powerful, but unlikely to hold once the economy embarks on an explosive path. Many of the papers mentioned in the previous paragraph recognise this and so feature multiple steady states. These models are then typically analysed under perfect foresight, under the assumption that global stochastic REE will be similar. Local REE in the region of one of these steady states can be identified with perturbation analysis and in indeterminate regions sunspots can be invoked to select between the many candidate stationary paths. This paper argues that this approach is misguided as these methods will miss the presence of a learnable REE in which indeterminate regions are sources rather than super stable as they will appear to be under perfect foresight or perturbation analysis. Neural networks offer a flexible non-linear solution method which recognise the presence of the multiple steady states and can characterise transition paths between them.

The steady states around which the complementarity or non-convexity driving indeterminacy is powerful will feature a continuum of converging stable paths. In other words, for any initial state its neighbourhood, multiple paths bring the system back to the steady state. Which path is selected depends on agents' expectations, and all are consistent with rationality and so qualify as rational expectations equilibria (REE). Hence such a model is indeterminate. I will refer to such steady states as "local indeterminacy steady states". When the complementarity or non-convexity ceases to be effective, additional steady states may exist around which only one stable path is consistent with rational expectations. In the neighbourhood of these steady states the model is therefore determinate. I will refer to these as "local determinacy steady states".[1]

Indeterminacy is not always explicitly treated as a local phenomenon, for example the seminal contribution of Benhabib and Farmer (1994) analyses a log-linearised neo-classical growth model. This is understandable as, in pure rational expectations terms, indeterminate systems appears to be super stable so it is natural to assume that a system will remain in the neighbourhood of a local indeterminacy steady state. However, previous literature has shown that *none* of the stationary REE around a local indeterminacy steady

[1]For example, as I will discuss further in Section 1.5 the benchmark New Keynesian model with a Zero Lower Bound on nominal interest rates can be thought of as having one local determinacy steady state and one local indeterminacy steady state. The "target inflation" steady state, where the Taylor Principle holds is local determinate: there is only one stable path consistent with rational expectations in that neighbourhood. The "ZLB" steady state, first identified by Benhabib et al. (2001), is locally indeterminate as there are many stable paths converging to it that are consistent with rational expectations.

state are learnable, except under very restrictive conditions (McCallum, 2007; Ellison and Pearlman, 2011). Linearised model featuring indeterminacy will thus be explosive under learning. If we take learnability seriously as "lending plausibility" to rational expectations (Lucas, 1986), we therefore ought to consider how the forces generating indeterminacy are bounded. In some cases, this could involve the economy hitting an extreme outcome or constraint, but in most cases the mechanism driving indeterminacy will be naturally limited.

This paper takes non-linearity and multiple steady states seriously. By allowing agents to learn using neural networks, a flexible and powerful function approximator, working with non-linear models in a stochastic environment becomes possible. This delivers a positive prediction for equilibrium behaviour around local indeterminacy steady states. While local determinacy steady states are stable under learning (i.e. agents learn the unique stable REE in their neighbourhood), local indeterminacy steady states are unstable and agents learn an explosive path away from these steady states. Furthermore, when agents' learning algorithm is sufficiently flexible there can be a learnable transition path between steady states. A neural network offers them this flexibility, but qualitatively similar results are observed with the more familiar Recursive Least Squares (RLS) learning algorithm when some higher order regressors are included. However, this RLS equilibrium does not have the properties of a REE, as agents' forecast errors are not independent of their information sets.[2]

Modelling agents as learning using a neural network has several advantages. First, as REE in non-linear models with state variables cannot be exactly characterised by a finite number of parameters, the parsimony and flexibility of neural networks ideally suits them to approximating non-linear and potentially discontinuous expectation formation processes. Second, neural networks perform well even in small samples, needing only a few hundred periods to generate outcomes which are qualitatively similar to the long run outcome, as shown in Section 1.3.5. Third, they are efficient to train on even very large datasets and there are well developed techniques to prevent overfitting and deal with collinearity. Finally, by allowing agents to learn using a sophisticated and flexible algorithm, they are given every chance to learn one the stationary REE in the indeterminate region. The fact that they do not strengthens the results from the learning literature which show that indeterminate REE are not generally not learnable with RLS learning. There is some recent work using neural networks as a global solution method, which I

[2]Nevertheless, if agents learn using a simpler algorithm the Learnable-REE identified by the neural network is a good approximation to their behaviour.

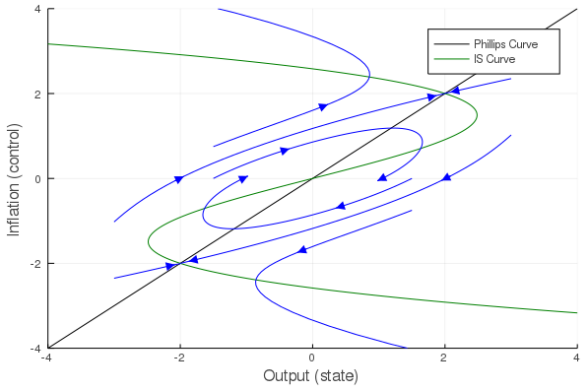
briefly survey below. However, to the best of my knowledge, this paper represents the first attempt to address indeterminacy, and show the existence of fully stochastic learnable equilibria with multiple steady states and accurate beliefs about the transitions between them.

A simple model with one state/pre-determined variable (y_{t-1}), one control/forward-looking variable (π_t) and a complementarity that gives rise to local indeterminacy, is used as an illustrative example throughout the paper. Where the feedback from the control to output is positive, it generates indeterminacy. However, for more extreme values the feedback becomes negative, so there two additional steady states which bound the region of indeterminacy. In Section 1.5, I will show how the lessons learned from this illustrative case can be applied to a particular case of interest, where monetary policy is constrained by a lower bound on nominal interest rates. However, the general lessons are much more widely applicable. For example, if the state variable is unemployment and control variable is the value of employment, the model resembles that of Kaplan and Menzio (2016). Alternatively, if the state variable is capital stock and the control variable is consumption, it resembles the endogenous mark-ups model of Gali (1995).[3]

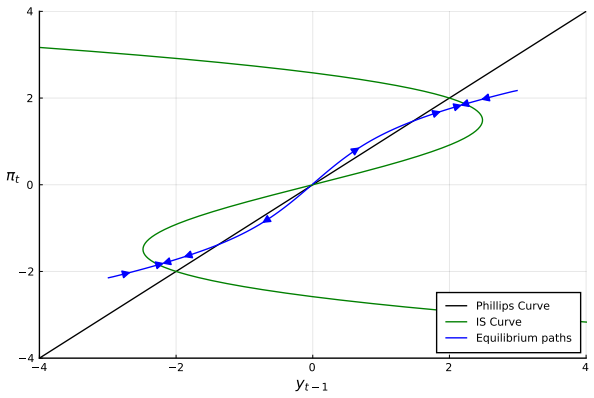
The flavour of the main results can be illustrated by the two phase diagrams in Figure 1.1. In both cases the state variable (y_{t-1}) is on the horizontal axis and the control variable (π_t) is on the vertical axis. The black and green lines represent the two deterministic steady state conditions, i.e. the points at which each equilibrium condition is constant. The intersections of these two lines are deterministic steady states. The blue arrows represent the evolution of the system from a given starting point in π_t, y_{t-1} space.

Figure 1.1: The Paper in Two Diagrams

(a) Perfect foresight



(b) Neural Network Learnable-REE



[3]Higher dimensional models can often also be represented with similar phase diagrams, as in Eggertsson et al. (2019) whose model of secular stagnation also features three steady states, where the outer two or locally determinate and the central one is locally indeterminate.

The left hand panel shows the system's behaviour under perfect foresight. The steady state conditions cross three times, so there are three steady states. The two outer steady states are saddle path stable under perfect foresight, so there is one path which converges into each but all other paths are unstable.[4] The REE around these steady states is determinate in that for any value of the state variable (y_{t-1}) in its neighbourhood, there is only one value of the control variable (π_t) which brings the system onto a path back to the steady state. The central steady state is a sink under perfect foresight, so has a continuum of stable REE paths. In other words, for any value of the state variable in the neighbourhood, there are infinitely many values of the control variable for which the system is on a path back to that steady state. The behaviour of the system under rational expectations around this steady state is therefore indeterminate. The indeterminacy is bounded by the saddle paths into the outer steady states, as paths outside of this region are explosive. There is thus a continuum of stable REE paths, but they are bounded in a finite region.

The right hand panel shows the long run behaviour of the same system under neural network learning. There are still three steady states, but now the central steady state acts as a source to the system, while the two outer steady states are stable. The paths that leave the central steady state are the "Learnable-REE" path. This path does not resemble any one of the perfect foresight paths but has the properties of a REE: agents do not make systematic forecast errors and have accurate beliefs about the laws of motion. Furthermore, transitory shocks can have permanent effects by moving the system from the region of one local determinacy steady state to the other.

The Learnable-REE that learning converges to is intuitively rather simple and suggests that the local indeterminacy steady state should essentially be treated as an unstable equilibrium. Under learning, the system spends little time near the local indeterminacy steady state. Instead, as in a model with hysteresis, the system moves around one of the outer steady states, until a large enough shock moves it into the region of the other. This suggests that much of the interest in these indeterminate equilibria may be largely misplaced as the system will be repulsed by them. This is in line with the previous work on learning in macroeconomics, which cast doubt on the learnability of REE in indeterminate models. However, it goes a step further by generating a positive prediction for equilibrium behaviour under learning in these models, and shows that equilibrium has the desirable properties of a stationary REE.

[4] As there is one control variable and one state variable, we can think of indeterminacy in terms of the eigenvalue conditions of Blanchard and Kahn (1980). The model is locally determinate around a steady state if a first order approximation has one stable and one unstable eigenvalue.

Outline. The rest of the paper is organised as follows: what remains of this Section discusses related literature. Section 1.2 introduces the illustrative model, and discusses the concept of indeterminacy in linear rational expectations models. Section 1.3 introduces neural network learning, presents mean dynamics and simulation results. It is demonstrated that the equilibrium which learning mean dynamics converge to has the properties of a REE and that the system can converge to a qualitatively similar solution very quickly. Section 1.4 explores the robustness of these results by presenting the mean dynamics under RLS learning and under a wide range of alternative parameterisations, as well as discussing the implications for empirical tests for indeterminacy. Section 1.5 then shows how this framework could be applied to the indeterminacy implied by a Zero Lower Bound (ZLB) on nominal interest rates. Section 1.6 concludes.

Related Literature. This paper principally contributes to four areas of research in macroeconomics: indeterminacy, particularly in the context of non-linear models with multiple steady states; learning in macroeconomic models; the use of neural networks as a solution algorithm; and the implications of a lower bound on nominal interest rates for the macroeconomy. The first of these is discussed above and the last at the beginning of Section 1.5, so I briefly review the other two here.

The literature on learning in macroeconomics goes back at least as far as Lucas (1986) who suggested that stability under learning could be used as a criterion to select between REE in an overlapping generations (OLG) model of fiat money. The most influential contribution in this field is undoubtedly Evans and Honkapohja (2001) who provided a rich and powerful analytic framework for analysing learning. The results presented in this paper are entirely consistent with Evans and Honkapohja (2001). Their concept of E-stability can be used to verify that REE around local indeterminacy steady states are often not learnable. For example, Bullard and Mitra (2002) show that E-stability and determinacy are closely related in a benchmark New Keynesian model.

However, although previous work has cast doubt on the learnability of indeterminate equilibria, the general relationship between indeterminacy and learnability is still an open question and appears to be quite complex. For example, Duffy (1994) shows that the OLG model used by Lucas (1986) to argue for learnability as an equilibrium selection device does in fact have an E-stable indeterminate equilibrium when the agents learn to predict an inflation factor rather than the price level. Similarly, Woodford (1990) shows that sunspot equilibria in a similar model can be learnable when agents include the possibility

of a sunspot in their learning rule and sort their observations on the basis of this variable.

McCallum (2007) and Ellison and Pearlman (2011) explicitly examine the correspondence between indeterminacy and E-stability in general linearised dynamic rational expectations models. The former proves that determinacy is a sufficient, but not necessary, condition for E-stability. The latter show that even if agents know which variables are pre-determined, indeterminate equilibria will not be E-stable unless agents know the location of the steady state and all of the system's eigenvalues are real and positive.

Other papers show that sunspot equilibria are learnable in several contexts. Evans and McGough (2005) identify a common factor representation of sunspots under which a sunspot solution can be E-stable in a New Keynesian model if expectations of future variables are included in the policy rule. Arifovic et al. (2013) apply social learning in which multiple agents start with heterogeneous models and learn from one another, showing that determinacy is not required for agents to coordinate on an equilibrium. Benhabib et al. (2015) show that sunspot equilibria are learnable, subject to conditions on the learning rate, in a model where imperfect information leads to sentiment driven aggregate demand fluctuations.

Taking a similar approach to that in this Chapter, recent work by Chen et al. (2021) uses reinforcement learning to solve New Keynesian models which feature indeterminacy and shows that the Pareto optimal steady state can be identified. This contrasts with the results described here, as they are able to show learning converges to steady states that are not E-stable. A key difference in approach that may explain these different results is that in their approach agents learn how to maximise utility rather than to minimise their forecast errors. This difference in the loss/reward function leads to markedly different learning dynamics, a phenomenon that is certainly worth exploring in more depth.

This Chapter should thus not be seen as evidence for the general non-learnability of indeterminate equilibria, as there certainly are some contexts in which they may be learnable. Rather, it proposes a solution in cases where indeterminate equilibria do not appear to be learnable. When indeterminate regions are bounded, we can select a globally stable, stationary and learnable equilibrium by using a flexible non-linear learning algorithm.

There is a comparatively small literature which tackles learning in non-linear models. As REE in a non-linear model with continuous endogenous state variables cannot be expressed in a finite number of parameters (Judd et al., 1998), exact E-stability anal-

ysis in the spirit of Evans and Honkapohja (2001) is not possible. Some papers deal with non-linearity by modelling linear learning in non-linear model (Hommes and Sorger, 1998; Bullard, 1994), often with very interesting results but expectations in these equilibria will never be consistent with rationality. The RLS dynamics explored in Section 1.4 can be thought of as taking a similar approach. Berardi and Duffy (2015) use a PEA (Parameterised Expectations Algorithm) approach to approximate non-linear expected values around a steady state, and show some reasonable convergence results, but do not report any evidence on the accuracy of the learning equilibrium.

Although there is a fairly well-developed empirical literature employing neural networks for forecasting in economics and finance, e.g. Kaastra and Boyd (1996), their use as a function approximator to solve rational expectations models is more recent. Maliar et al. (2019) discuss the general applicability of neural networks to computational economists, and show how many models can be written to allow solution with neural networks. Other work has shown that equilibrium conditions can be directly included in the loss function of a neural network to efficiently find global solutions (Azinovic et al., 2019); the value of neural networks for solving high dimensional models by endogenously dealing with collinearity Fernández-Villaverde et al. (2019); Villa and Valaitis (2019). This paper contributes to this literature by demonstrating that neural networks can provide an intuitive and plausible resolution to indeterminacy in non-linear models with multiple steady states, which builds on the well-known results from the learning literature that indeterminacy often leads to instability under learning.

1.2 An illustrative model

A simple New Keynesian model with one state/pre-determined variable, y_{t-1} and one control/forward-looking variable π_t is used to illustrate the intuition behind the results. For concreteness and clarity of expression a single parametrization are presented here, but Section 1.4 will demonstrate the general applicability of these results. Section 1.5 then applies the same framework to a more extensive model.

Inflation follows a standard New Keynesian Phillips Curve.

$$\pi_t = \beta \mathbb{E}_t \pi_{t+1} + \kappa y_t + \epsilon_{\pi,t} \tag{1.1}$$

Output is determined by a backward-looking IS curve. The IS curve is backward-looking as the illustration is clearer and easier to visualise in two dimensions with one control and

one state, but the key results are the same with a forward-looking consumption Euler Equation.

$$y_t = \eta y_{t-1} - \sigma(r_t - \mathbb{E}_t \pi_{t+1}) + \epsilon_{y,t} \tag{1.2}$$

where $\eta < 1$ and $\epsilon_{\pi/y,t} = \rho_{\pi/y}\epsilon_{\pi/y,t-1} + \nu_{\pi/y,t}$ with $\nu_{\pi/y,t} \sim \mathcal{N}(0, \sigma^2_{\pi/y})$. The interest rate r_t is determined by a Taylor Rule which is indeterminate around the target steady state, but satisfies the Taylor Principle for more extreme values of inflation

$$r_t = \phi_\pi \pi_t + \alpha \pi_t^3 \tag{1.3}$$

where $\phi_\pi < 1$ and $\alpha > 0$. As discussed above, in most plausible cases, the forces that generate indeterminacy are likely to be bounded. The cubic term in the Taylor Rule demonstrates an example of this. For small deviations in inflation the central bank does not react sufficiently aggressively to make the model determinate. However, once inflation gets too extreme, the central bank will react strongly and adjust the interest rate more than proportionally.

As a baseline case, I choose parameter values for which there are three deterministic steady states, at (0,0), (-2,-2) and (2,2).[5]

Table 1.1: Parameter values for baseline illustrative model

	Parameter							
	β	κ	η	σ	ϕ_π	α	$\rho_{\pi/y}$	σ_π/y
Value	0.95	0.05	0.95	0.25	0.5	0.075	0.5	0.3

The two outer steady states are saddle path stable under perfect foresight, so the system is determinate under rational expectations in their neighbourhood. The central steady state is a complex sink under perfect foresight, so the system is indeterminate under rational expectations in its neighbourhood.

1.2.1 Perfect foresight solution

As global REE in non-linear models with shocks and state variables cannot be exactly characterised in finite parameters (Judd et al., 1998), global dynamics in models of this form are therefore often analysed under perfect foresight under the assumption that the REE will resemble these perfect foresight dynamics. This paper argues that this is a mistake: indeterminate regions which appear to be super stable under perfect foresight will in fact be a source under the slightest deviation from full information rational ex-

[5]Stochastic and deterministic steady states will not in general coincide, as shown in Section 1.3.

pectations.

The deterministic steady states of the model are the intersections of the steady state Phillips Curve (Eq. 1.4) and IS Curve (Eq. 1.5), as shown in Figure 1.3. The two conditions cross three times, representing the three deterministic steady states.

$$\pi = \beta\pi + \kappa y \tag{1.4}$$

$$y = \eta y - \sigma(\phi_\pi \pi + \alpha\pi^3 - \pi) \tag{1.5}$$

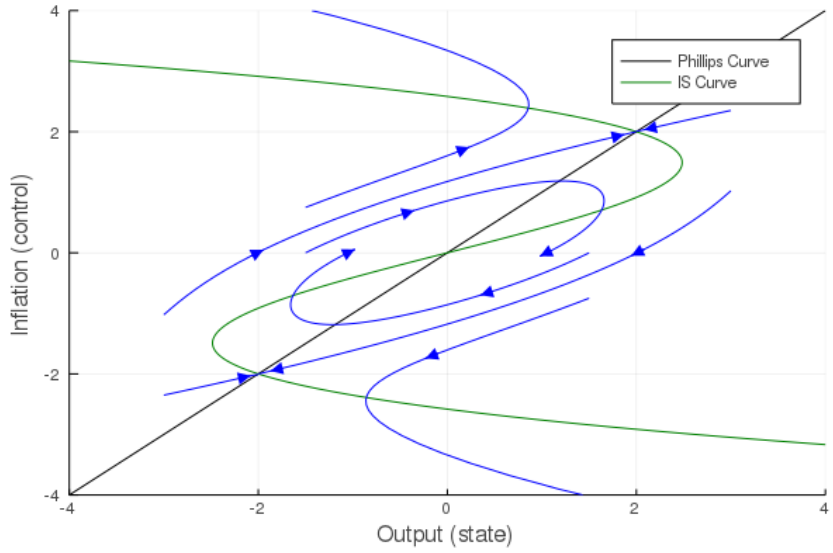
Under perfect foresight, $\mathbb{E}_t\pi_{t+1} = \pi_{t+1}$ so we can express π_t and y_t in terms of π_{t-1} and y_{t-1}. These two difference equations allow us to characterising the dynamics of the model under perfect foresight.

$$\pi_t = \frac{1}{\beta}(\pi_{t-1} - \kappa y_{t-1}) \tag{1.6}$$

$$y_t = \frac{\beta}{\beta + \sigma\kappa}\left(\eta y_{t-1} - \sigma * \left(\frac{\phi_\pi\beta - 1}{\beta}\pi_t + \alpha * \pi_t^3\right)\right) \tag{1.7}$$

Figure 1.3 represents the model under perfect foresight as a phase diagram. The blue arrows represent paths along which the perfect foresight system will evolve from a given starting point in π_t, y_{t-1} space. These paths show that the outer steady states are saddle paths and the central steady state is a complex sink. Therefore, as discussed above, the model is determinate under rational expectations around the outer steady states and indeterminate around the central steady state. This is the classic Taylor Principle result that a New Keynesian model is indeterminate if monetary policy is not sufficiently aggressive.

Figure 1.3: Illustrative model under perfect foresight



1.2.2 Indeterminacy in the neighbourhood of a steady state

A model is indeterminate around a particular steady state if, for given initial conditions, there are multiple candidate paths in to that steady state which satisfy rational expectations and any transversality conditions. The illustrative model is indeterminate under rational expectations because paths near the central steady state are both non-explosive and fully consistent with rational expectations. I linearise the model around each of its steady states in order to demonstrate the intuition behind the concept of indeterminacy.

Denote (π^*, y^*) as one of the model's steady states, linearising around this steady state then gives.

$$\hat{\pi}_t = \beta \mathbb{E}_t \hat{\pi}_{t+1} + \kappa \hat{y}_t + \epsilon_{\pi,t} \tag{1.8}$$

$$\hat{y}_t = \eta \hat{y}_{t-1} - \sigma \left((\phi_\pi + 3\alpha {\pi^*}^2) \hat{\pi}_t - \mathbb{E}_t \hat{\pi}_{t+1} \right) + \hat{\epsilon}_{y,t} \tag{1.9}$$

When $\pi^* = 0$ at the central steady state, this is simply the system without the cubic term, and so monetary policy is passive. However, at the outer steady states where $\pi^* = \pm 2$ monetary policy is active. Defining $\hat{x}_t = (\hat{\pi}_t \ \hat{y}_{t-1})'$ and $\hat{\epsilon}_t = (\hat{\epsilon}_{\pi,t} \ \hat{\epsilon}_{y,t})'$, we can write the linearised model in matrix form.

$$\mathbb{E}_t \hat{x}_{t+1} = \Phi \hat{x}_t + \Psi \hat{\epsilon}_{y,t} \tag{1.10}$$

where

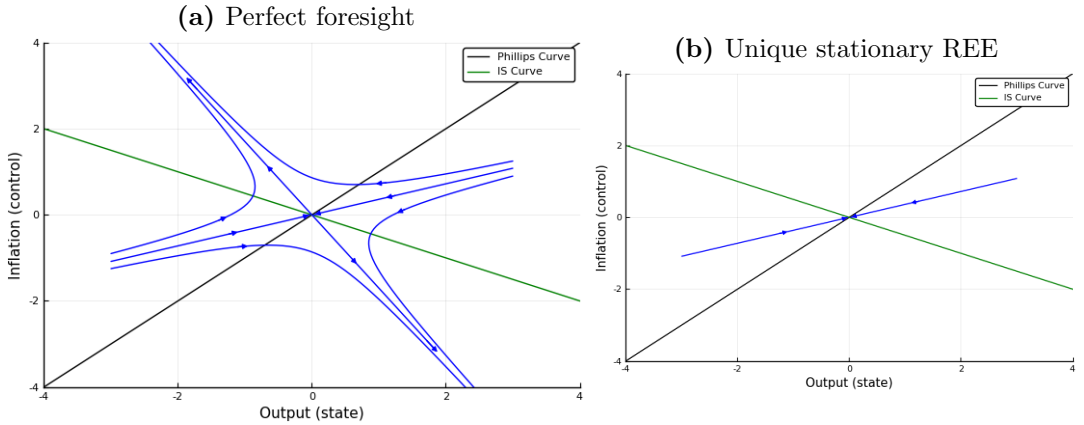
$$\Phi = \begin{pmatrix} \frac{1}{\beta} & \frac{-\kappa}{\beta} \\ \frac{-\sigma(\beta(\phi_\pi + 3\alpha {\pi^*}^2) - 1)}{\beta + \sigma\kappa} & \frac{\beta\eta}{\beta + \sigma\kappa} \end{pmatrix}, \ \Psi = \begin{pmatrix} \frac{1}{\beta} \\ \frac{\beta}{\beta + \sigma\kappa} \end{pmatrix}$$

Following Blanchard and Kahn (1980), the determinacy of the model in the neighbourhood of that steady state depends on the eigenvalues of the Φ matrix. As the system has one pre-determined state variable, y_{t-1}, and one forward-looking control variable, π_t, the Blanchard-Kahn conditions tell us that the system will be determinate if Φ has one stable eigenvalue (inside the unit circle) and one unstable eigenvalue (outside the unit circle). It is straightforward to verify that at the central steady state, where ${\pi^*}^2 = 0$, both eigenvalues of Φ are stable, and at the outer steady states, where ${\pi^*}^2 = 4$, Φ has one stable and one unstable eigenvalue.

For values of the state variable (y_{t-1}) in the neighbourhood of the outer steady states there is only one value for the control variable (π_t) which puts the system on a stationary path leading back to the steady state. Invoking some transversality conditions which rule out the explosive paths selects the unique stationary rational expectations solution, as illustrated by Figure 1.4. Panel (a) shows the linearised system under perfect foresight

(i.e. all candidate RE solutions). Panel (b) shows the unique stationary REE. For any given initial value of the state variable y_{t-1} the economy will jump onto the path back into the steady state.

Figure 1.4: Linearised model around upper steady state



Most solution methods will choose the unique stable REE by assumption (Blanchard and Kahn, 1980; Sims, 2002), but an alternative justification is learnability. As shown by McCallum (2007), in a determinate linear system the only learnable equilibrium will be the unique stable REE.[6]

In the neighbourhood of the central steady state, the complementarity introducedby passive monetary policy means that there are many stationary REE. For a given value of the state variable y_{t-1}, there are infinitely many values of the control variable π_t will place the system on a stable path that converges back to the steady state. In this situation, rational expectations do not give us a unique prediction for equilibrium behaviour so some extra criteria is needed to close the model. One option is to introduce an exogenous sunspot shock that chooses one of these candidate paths, but this paper uses learning to make positive predictions about equilibrium behaviour models with indeterminate environments.

[6] A small field of research has questioned the use of these transversality conditions and argued that these explosive paths may be plausible, at least temporarily (Cochrane, 2009; Ascari et al., 2019). This paper makes similar predictions by showing that local indeterminacy steady states are unstable under learning, so such an explosive path is learnable, even if it is not present under perfect foresight.

1.2.3 Learning and indeterminacy: an intuition

It is well established in the learning literature that models that are indeterminate under rational expectations are unstable under learning. For example Bullard and Mitra (2002) show that learnability and determinacy are tightly linked in a benchmark New Keynesian model similar to that used here. This paper does not contradict this literature but points out that if indeterminacy is bounded, as it plausibly will be in most cases, then this instability under learning is also bounded.

To illustrate the intuitive link between indeterminacy and learnability, consider the Phillips curve in our illustrative model and assume that y_t and $\epsilon_{\pi,t}$ are set exogenously to zero for all periods.

$$\pi_t = \beta \mathbb{E}_t \pi_{t+1}$$

We now have a univariate linear rational expectations model. If $\beta \in (-1, 1)$ it is determinate and there is a single stationary solution: $\mathbb{E}_t \pi_{t+1} = \pi_t = 0$. While if $\beta \notin (-1, 1)$ it is indeterminate and there are many stationary solutions.[7]

Now imagine that the agent does not have rational expectations but instead sets their expectation with an initial guess that is slightly perturbed from zero, and attempts to update this guess until it is consistent with their observations.[8] If $\beta \in (-1, 1)$, then any mistake in their guess will be dampened: if their guess is above zero, the realised π_t will also be closer to zero than their guess and so they will update their guess closer to the stationary solution. However, if $\beta \notin (-1, 1)$ then the realised π_t will be even further from their guess and so the mistake in expectations is amplified and learning cannot converge.

The E-stability concept (Evans and Honkapohja, 2001) formalises this intuition and can be used to analyse the learnability of of REE in the neighbourhood of a steady state. In line with the literature, it is straightforward to verify that the unique stable REE around the *outer* steady states are indeed E-stable. In contrast, none of the REE around the central local indeterminacy steady state are E-stable under the baseline parameterisation. Neither are they under alternative parameterisations which preserve the indeterminacy, except under some very restrictive conditions on the properties of the model and agents' information sets, in line with Ellison and Pearlman (2011).[9] As the linearisation is exact

[7] Under perfect foresight the model becomes $\pi_{t+1} = \frac{1}{\beta}\pi_t$, so if $\beta > 1$ then any finite real value of π_t is consistent with a path that converges to zero.

[8] As there are no state variables trying to learn about current inflation is equivalent to learning about future inflation.

[9] The most important restriction being that agents know the location of the steady state.

in an infinitesimal neighbourhood of the steady state, this E-stability analysis will apply exactly in the neighbourhood of a steady state around which we have linearised. However, it does not make predictions about the global dynamics in non-linear systems, so to analyse the global behaviour of the system under learning, I will analyse mean dynamics under given learning algorithms.[10]

1.3 Neural Network Learning

Neural networks offer a parsimonious functional form which can be used to approximate arbitrarily complex non-linear functions and can be estimated efficiently, so they are a natural choice for learning in non-linear models. In this Section, I introduce neural network learning, show that it converges to an equilibrium in which local indeterminacy steady states become unstable, and that this solution has the properties of a rational expectations solution. This solution is the aforementioned "Learnable-Rational Expectations Equilibrium".

1.3.1 Background: neural networks

This paper allows agents to use a single layer feedforward neural network to form their expectations.[11] This neural network takes a vector of inputs (s) and transforms it into a vector of outputs ($\hat{p}$) through one "hidden layers" of nodes. Between each layer, the inputs are transformed by multiplying them by "weights" (w) and adding them to a "bias" (b). At each node in a hidden layer, the input is also transformed by a non-linear activation function ψ, which is applied to the linear transformation of the inputs.

The neural network can thus be expressed as in Equation 1.11, where w_i^{hid} and b_i^{hid} are the weights and biases between the input and the ith node of the hidden layer, and w_j^{out} and b_j^{out} represent the weights and biases applied between the hidden and output layers (so $w_{j,i}^{out}$ is the weight of the ith node for output j). There are N nodes in the

[10] There have been some attempts to apply the concept of E-stability to non-linear models, but these are either restricted to models without transition dynamics Evans and Honkapohja (2001) or rely on a finite elements approximation resulting in a very high dimensional functions Christiano et al. (2018).

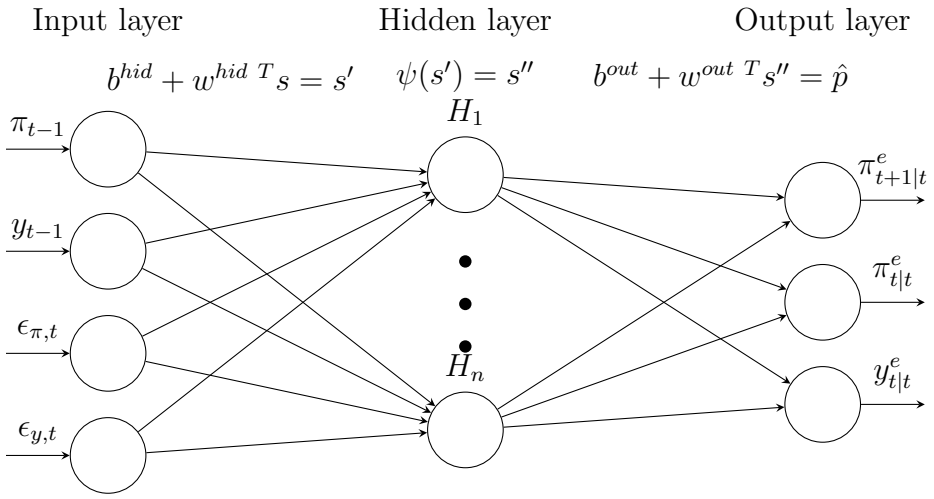
[11] Feedforward neural networks, or multilayer perceptrons, are function approximates which respond to their inputs in a deterministic way. Recurrent neural networks include feedback connections between layers, but are not considered here.

hidden layer and the output is a vector of length M.

$$\hat{p} = F(s) = \begin{pmatrix} b_1^{out} + \sum_{i=1}^{N} w_{1,i}^{out}\ \psi(w_i^{hid\ \prime} s + b_i^{hid}) \\ \vdots \\ b_M^{out} + \sum_{i=1}^{N} w_{M,i}^{out}\ \psi(w_i^{hid\ \prime} s + b_i^{hid}) \end{pmatrix} \tag{1.11}$$

Figure 1.6 shows the network's structure. The observed variables and shocks are the inputs, these are then transformed with the weights and biases into each node in the hidden layer. In the hidden layer, the activation function is applied to each of these linear transformations. Finally, a linear transformation of the activated nodes then creates the outputs, which are predictions for this period's variables and next period's price.

Figure 1.6: Structure of neural network used here



The Universal Approximation Theorem Cybenko (1989) states that any real-valued continuous function can be arbitrarily well approximated by a single layer feedforward neural network with any appropriate activation function and sufficient nodes. A neural network could thus approximate the expectations process of any possible rational expectations equilibrium. However, a given function will not be associated with a unique set of weights and biases which best approximate it.

Any continuous and bounded function can be used as an activation function, but the default for many types of neural network is the Rectified Linear Unit (ReLU) function. The ReLU activation function is a piecewise linear function that will output the input directly if is positive, otherwise, it will output zero, i.e. $\psi(x) = \max(0, x)$. A neural network that uses ReLU activation is both easy to train and often achieves better performance than alternative activation functions such as the sigmoid or tan-sigmoid functions. The Universal Approximation Theorem is also quite intuitive with a ReLU activation func-

tion: each node contains a linear function that can be switched on and off. This allows the network to approximate any function by combining linear functions.

In the baseline case, a neural network with the structure shown in Figure 1.6 is used. There are 64 nodes in the hidden layer and a ReLU activation function. This neural network nests a purely linear model: imagine that there is as single node in the hidden layer, the weights and biases can be set such that the output is a straightforward linear transformation of the inputs (above some arbitrarily low minimum).

1.3.2 Neural network learning and mean dynamics

Agents form expectations using a neural network. They update the parameters (i.e. the weights and biases) of this neural network based on their forecast errors using the Levenberg-Marquardt algorithm. In order to approximate the mean dynamics under neural network learning, a large number of periods are simulated between updates which approximate the distribution of the system given current beliefs. As the model is non-linear, agents form expectations of π_{t+1} directly from observed $t-1$ variables and current shocks, rather than learning a one-period ahead model and then applying it twice.

Training a neural network requires a definition of the loss function. A natural choice here, which demonstrates how neural network learning is a generalisation of least squares learning, is the sum of squared errors.[12] For our neural network parameterised by θ, with output $\hat{p}_{t|\theta}$ and target p, trained on T observations, the loss would thus be

$$L = \sum_{t=1}^{T} (p - \hat{p}_{\theta})^2 = e_t^2 \tag{1.12}$$

The least squared loss function allows to use the Levenberg-Marquardt (LM) algorithm (Hagan and Menhaj, 1994), described in greater detail in Appendix A.1. The LM algorithm uses the Jacobean of the network errors with respect to the parameters $\mathbf{J}$ to update parameters θ

$$\theta^{(i+1)} = \theta^{(i)} + (\mathbf{J}^{(i)\prime} \cdot \mathbf{J}^{(i)} + \mu^{(i)} I)^{-1} \cdot (2\mathbf{J}^{(i)\prime} \cdot e^{(i)}) \tag{1.13}$$

where μ is a "dampening factor" that ensures the approximate Hessian $(\mathbf{J}^{(i)\prime} \cdot \mathbf{J}^{(i)} + \mu^{(i)} I)$ is positive definite. The dampening factor is reduced as the algorithm approaches a minimum, at which point the LM algorithm effectively becomes a Newton method with an approximate Hessian.

[12] Other loss functions are possible and may even be preferable, but given that an accurate solution is learned with squared error approach it is clearly sufficient.

The update equation for the LM algorithm can be expressed as a function of current beliefs, state variables and shocks. The learning process can thus be defined as a stochastic recursive algorithm, as discussed in Evans and Honkapohja (2001).

$$\theta_\tau = \theta_{\tau-1} + \mathcal{H}(\theta_{\tau-1}, X_t) \tag{1.14}$$

where $X_t = (\pi_t \ y_t \ \epsilon_{\pi,t} \ \epsilon_{y,t} \ \pi_{t-1} \ y_{t-1} \ \epsilon_{\pi,t-1} \ \epsilon_{y,t-1})'$ is the state space of the structural equations and $\mathcal{H}$ is some function of the distribution of forecast errors agents make with beliefs θ_{t-1}. In order to characterise the mean dynamics of neural network learning we thus have to find a fixed point of the differential equation $\frac{dh(\theta)}{d\theta}$ which is the expected value of $\mathcal{H}$ for given beliefs θ.

$$h(\theta) = \int \mathcal{H}(\theta, x)\Gamma_\theta(dx) \tag{1.15}$$

$\Gamma_\theta(dy)$ is the invariant distribution of X conditional on θ. The challenge for calculating these mean dynamics is identifying the $h(\theta)$ function and finding $h(\theta^*) = 0$. We can thus think of mean dynamics as identifying the convergence point of learning in an infinitely long simulation.

In the RLS case discussed in Section 1.4, it is possible to characterise this distribution analytically but given the highly non-linear nature of neural networks, a more numerical approach is necessary here. In order to approximate the distribution $\mathcal{H}(\theta_{t-1}, X_t)$ we can simply hold beliefs θ_{t-1} fixed and sample many times from the system. This gives us an empirical distribution for the forecast errors and hence for $\mathcal{H}(\theta_{t-1}, X_t)$. We can then use this empirical distribution to update the parameters by training the model on these samples, using the current beliefs as initial parameter values. In the baseline case, agents recycle some past simulations by periodically updating their network on a rolling window of past observations. This speeds up convergence: updating regularly while maintaining sufficiently large samples to avoid oversampling from part of the distribution, but the eventual result is the same with large non-overlapping samples.

As shown in Figure 1.6, agents predict $\pi^e_{t+1|t}$ directly from last period's variables and the contemporaneous shocks, rather than predicting period t variables and then using these to predict period $t+1$ variables. In a linear model, the two approaches are equivalent, in that both will nest the same REE solutions, as shown in Appendix A.2.1. However, in a non-linear model, Jensen's Inequality means that the two are not equivalent, as shown in Appendix A.2.2. As the $\pi^e_{t+1|t}$ forecast contains non-linear transformations of ϵ_{t+1}, we

cannot simply use $\mathbb{E}_t \epsilon_{t+1}$ in a one-period ahead model to forecast $\pi^e_{t+1|t}$.

As neural networks can predict a vector of outputs, one can allow agents to simultaneously predict period-t values, making it easier to compare agents' beliefs with the implied dynamics. As agents observe the state variables and contemporaneous shocks, they have all the necessary information to predict period t values accurately. This is the case here, as the R^2 of regressions of expected contemporaneous variables $\pi^e_{t|t}$ and $y^e_{t|t}$ on the realised values are 0.99991 and 0.99998 respectively.

After an initial burn-in phase of 50,000 observations during which expectations generated by a draw from a standard normal distribution, the agents update their network every 1,000 periods based on a rolling window of the past 50,000 observations. There are 500,000 simulations in total, taking around 2 hours a single core of a 2.8 GHz processor. They take the previous parameter values of the network as a starting point and update using the LM algorithm. The data is split into training, validation and test sets in order to verify that the model is not overfitting to noise. A stopping criterion that includes performance in a validation set is used to assess convergence, as described in Appendix A.1.[13]

1.3.3 Results

Neural Network learning converges to an equilibrium that is fundamental, stationary and stable under learning, in which the local indeterminacy steady state becomes an unstable source and local determinacy steady states become absorbing. The system will fluctuate around the neighbourhood of one of the local determinacy steady states until there is a sufficiently large fundamental shock to move the economy into the neighbourhood of the other local determinacy steady state. These results are robust to different initial conditions and specifications of the neural network.

Figure 1.7 shows a phase diagram for the actual laws of motion (ALM) for the Neural Network learning equilibrium and compares it to agents' perceived law of motion (PLM). Panel (a) shows the ALM implied by the beliefs contained within the neural network - i.e. how the system will evolve from a given starting point in π_t, y_{t-1} space. In stark contrast to the perfect foresight system shown in Figure 1.3, the central steady state is

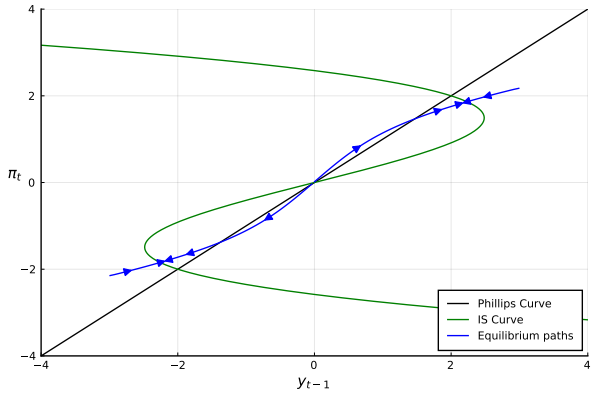
[13] As discussed in Section 1.3.5, an alternative "online" learning framework shows that convergence to qualitatively similar dynamics is possible in a plausible time frame. Thus, the Learnable-REE to which learning can converge to with sufficient data and a sufficiently flexible model may form a good approximation of empirically plausible behaviour in more restricted settings.

now a source rather than a sink. The phase diagram thus shows the key properties of this equilibrium, which is the key result of the paper. There are still three steady states. The outer two steady states, which are saddle paths under perfect foresight are stable: the system will converge to them from a point in their neighbourhood. The analysis of the linearised system around these outer steady states is a good approximation of the behaviour under learning in their neighbourhood.

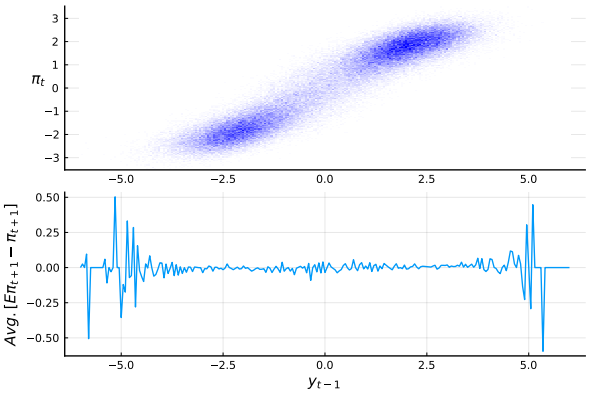
The right hand panel compares this ALM to agents' PLM. The upper part shows the the distribution of the system's endogenous variables over a long simulation (100,000 periods) showing that the system is concentrated in the neighbourhood of the two outer steady states. The lower part shows the average forecast error for the control variable $(\pi^e_{t+1|t} - \pi_{t+1})$ for different values of the state variable y_{t-1}.[14] As we can see, the agents make only very small errors, especially in the areas where the system spends most of it's time. The next Section will show more formally how this learnable equilibrium is consistent with rational expectations.

Figure 1.7: Equilibrium with Neural Network Learning mean dynamics

(a) ALM under Neural Network Learning



(b) Distribution of variables and errors



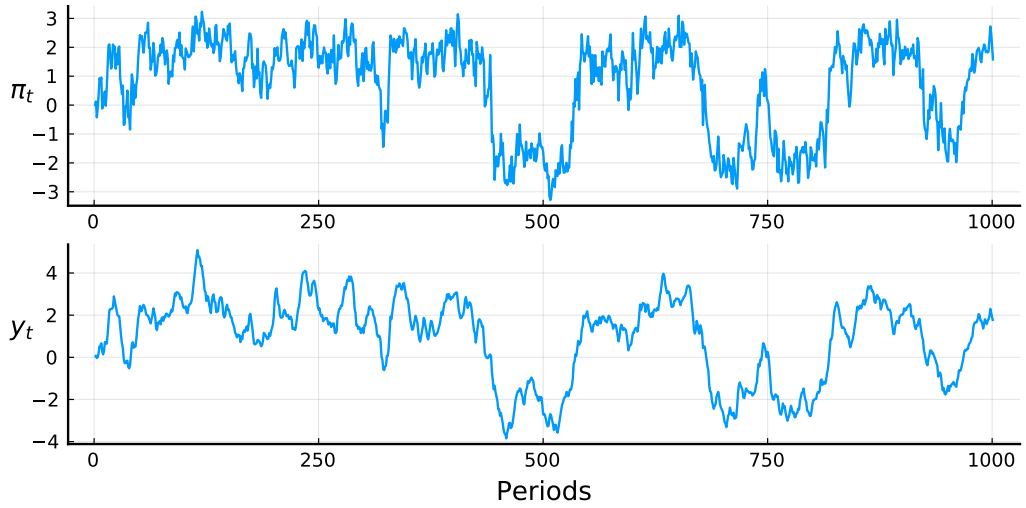
The behaviour around the central steady state is completely different to any of the REE identified by perfect foresight or the linearised system around that steady state. The indeterminacy around this steady state is due to the fact that all perfect foresight paths it its neighbourhood lead back into that steady state - i.e. it is a sink. However, under learning this steady state becomes a source. There is still a steady state at which the system will not move, but any perturbation from this steady state will lead to the system moving further away from it, and towards one of the other two steady states. Neither perfect foresight nor a perturbation around this steady state are thus a good approxima-

[14] As the agent's forecasts are formed using the direct two-step-ahead prediction, mapping from these forecasts to a PLM is not straightforward. Therefore, the one-step ahead forecast, $\pi^e_{t|t}$, which the network also outputs, is used.

tion of the system under learning.

There is thus a single equilibrium with multiple steady states. This equilibrium is fundamental in that no additional exogenous processes are introduced into the model, the dynamics are entirely driven by the two fundamental shocks $\epsilon_{\pi,t}$ and $\epsilon_{y,t}$. It is stationary in that the distributions of the variables are invariant over time, and it is stable under learning in that learning will bring the system back to these beliefs after a small deviation. The outer stochastic steady states are slightly closer to the centre than the deterministic steady states, at ± 1.79 rather than ± 2. This is because higher order moments of the shocks are taken into account. In fact, the noisier the shock processes are, the closer these outer steady states will move to the central steady state, as the cubic term enters negatively into the structural equations.

Figure 1.9: Regime-Switching behaviour
Simulated system under Neural Network learning



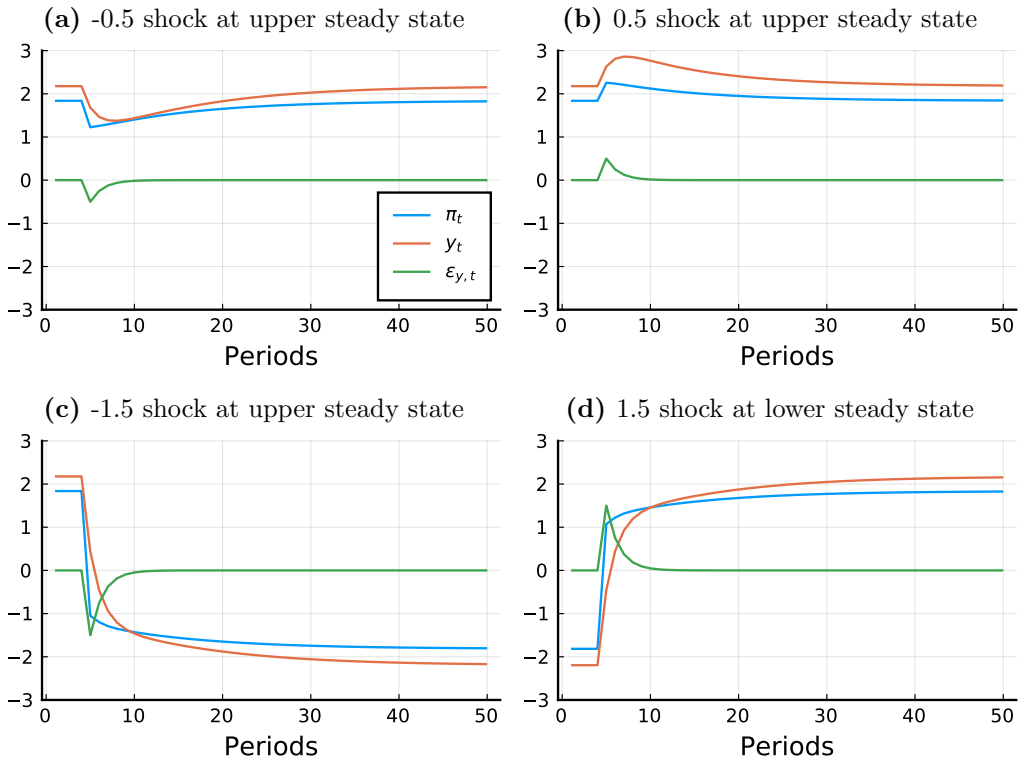
The behaviour of the system will be akin to a regime switching model, but the transition from one "regime" to the other is entirely determined by the fundamental shocks. Figure 1.9 shows 1,000 simulated periods with the beliefs that the mean dynamics of the Neural Network learning converge to. The system will fluctuate around one of the two outer steady states until a sufficiently large shock, or combination of shocks, moves it into the neighbourhood of the other outer steady state. This suggests that that non-linear models with regions of indeterminacy could micro-found this sort of regime switching behaviour without the need to introduce an exogenous transition probability.

The effect of shocks in this equilibrium are state and magnitude contingent: a shock can have very different effects depending on the initial state and its magnitude. Figure

1.10 shows Impulse Response Functions for π_t and y_t to a shock to $\epsilon_{y,t}$ (i.e. a demand shock). If the economy stars at one of the outer steady states then small demand shocks will qualitatively have the same effect as in a standard New Keynesian model: a positive shock temporarily increases inflation and output which then converges back to the (same) steady state. However, if a shock is sufficiently large, it can move the economy from one outer steady state to the other.

Panel (a) shows the response to a 1 unit positive shock to $\epsilon_{y,t}$ when the system starts at (-1.79,-1.79). The shock is large enough to move the system into the neighbourhood of the higher steady state, so has a potentially infinite multiplier. In contrast, panel (b) shows the response to a 0.5 unit positive shock. This shock is too small to move the system to another steady state, so the multiplier on this shock will be much smaller. Finally panel (c) shows the response to the same positive 1 unit shock as in panel (a), but when the system starts at (1.79, 1.79). The effects of an identical shock can thus be very different. This equilibrium displays strong hysteresis effects, as a sufficiently large shock can move the system into the neighbourhood of a different steady state, where will remain unless reversed by another large shock.

Figure 1.10: Impulse Response Functions for π_t and y_t to $\epsilon_{y,t}$: Transitory Shocks can have Permanent Effects



Note: The shock to $\epsilon_{y,t}$ is unanticipated and hits in period 5, before which the system is either in the high or low outer stochastic steady state. The blue line represents the response of π_t, the red line the response of y_t and the green line is $\epsilon_{y,t}$.

The key advantage of Neural Network learning over less flexible algorithms like RLS, which converge to a qualitatively similar equilibrium, is its ability to learn an equilibrium that has the properties of a REE. The next subsection shows this formally, but the phase diagrams in Figure 1.7 provide some evidence of this. The PLM and ALM are very closely matched, which is a necessary but not sufficient condition for a REE.

1.3.4 Equilibrium properties under neural network learning

A fixed point in the evolution of beliefs does not necessarily imply a REE, unless those beliefs are also consistent with rationality. We can test whether the neural network equilibrium has the properties of a rational expectations solution using the accuracy test proposed by Den Haan and Marcet (1994), referred to henceforth as DHM. This test is a demanding standard for any numerical solution, yet the neural network solution passes it comfortably. This shows that there exists a Learnable-REE which looks very different from those identified through perturbation or perfect foresight analysis. Furthermore, a sufficiently flexible learner with sufficient data available to them will reach this Learnable-REE.

In rational expectations equilibrium, agents will still make forecast errors as they do not observe future shocks. However, if expectations are rational, then these forecast errors should have the minimum possible variance and be independent of agents' information sets. Define the forecast errors made by agents beliefs θ as $u_{t+1|\theta}$. If expectations are formed rationally then these forecast errors should therefore satisfy

$$\mathbb{E}[u_{t+1|\theta} \otimes h(s_t)] = 0 \tag{1.16}$$

for any s_t in agents' information sets at time t and any function $h()$. DHM propose a test statistic for whether Equation 1.16 is satisfied for simulated series obtained with given beliefs θ.

The DHM accuracy test consists of obtaining many independent simulations of the process for given beliefs and comparing the empirical distribution of the test statistic with its theoretical distribution under the null hypothesis that forecast errors are independent of the information sets. If roughly 5% of draws from the empirical distribution are in the lower and upper 5% critical regions of the theoretical distribution, then we can accept the null hypothesis of independence and conclude that expectations are rational. Table 1.2 reports the results for 1,000 draws of 500 periods simulated for three specifications of agents' information set $h(s_t)$.[15]

Table 1.2: Den Haan & Marcet test results

$h(s_t)$	Lower-tail	Upper-tail
Constant	0.0461	0.0520
Constant, y_t and π_t	0.0420	0.0635
Extensive	0.0384	0.0781
Note:	1,000 draws of 500 periods	

The equilibrium passes the DHM test comfortably, so has the properties of a REE.

Crucially, this shows that the learnable equilibrium not only has very different properties to the previously analysed REE in indeterminate models, which regard the indeterminate region as super stable, but it is also consistent with rationality. We thus have a response to the challenge of indeterminacy that does not sacrifice the attractive features of rational expectations while being robust to departures from rationality.

[15]Contemporaneous values p_t and y_t are included in the agents' information sets even though they are not observed by agents as, in a rational expectations solution, that period's variables ought to be perfectly predictable.

1.3.5 Convergence time and learning in small samples

While a substantial number of observations is needed for neural network learning to converge to an equilibrium which passes the DHM accuracy test, the qualitative properties of the equilibrium (switching between the two outer steady states) can appear after only a few hundred periods. This suggests that neural networks are not only useful as a non-linear solution method, but might also be able to approximate learning under realistic time-frames. This is a considerable advantage over non-parameteric finite-element approximation approaches such as that of Chen and White (1998), used by Christiano et al. (2018) to model learning in a non-linear setting.

The neural network learning process described above involved a large number of simulations being drawn between updates for the beliefs in order to approximate the distribution of the variables given current beliefs. To investigate convergence time I instead update the network every period. This is computationally efficient if the network is trained "online", so that observations are only considered once, in the order that they appear, analogous to the recursive formulation of OLS regression used in RLS learning.

Figure 1.12: Evolution of system with online neural network learning

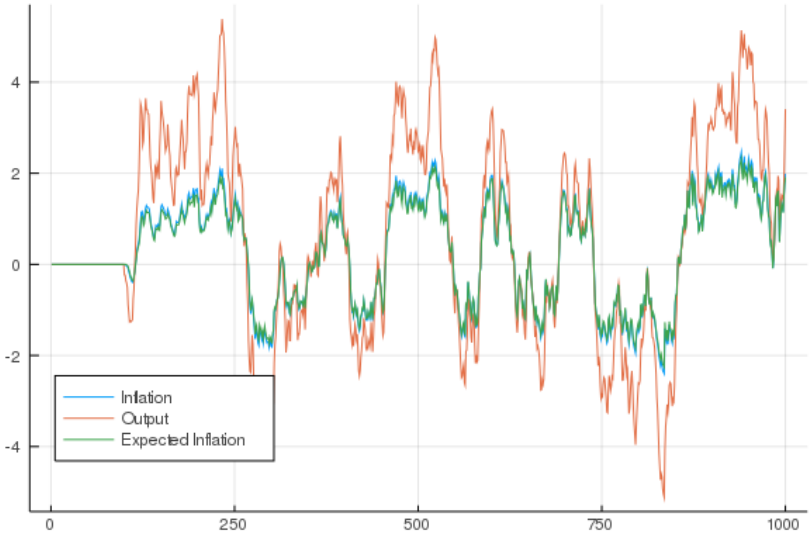


Figure 1.12 shows the evolution of the system with a 64 node neural network that, after the 10th period, is updated every period. The weights and biases of the network are all initialised to zero, so agents' prior belief is effectively that both variables will revert to zero.[16] The two left hand panels show π_t and y_t for the first 500 periods and the two right

[16] Although there are several hundred parameters in the network, the initialisation means that even a single observation can be used to update the network, analogous to how priors allow a Bayesian to form well-identified posteriors for a single observation.

hand panels show the first 5,000 periods. After only 120 or so periods, the system moves to the region of the lower steady state. The neural network is thus not only sufficiently flexible to capture the dynamics of the rational mean dynamics equilibrium, but also robust enough to generate sensible predictions on very small samples. From the right hand panels it can be seen that neural network learning also allows the possibility of switching between the outer steady states in small samples. Although agents may not have well formed beliefs about such a transition so early in the learning process, large shocks will still force the system into the region of the other outer steady state. The mean dynamics Learnable-REE is thus a good approximation of behaviour under learning.

It is worth noting that the convergence under any learning algorithm depends on the variance of inputs. When this is too low the system will converge to one of the two outer steady states, at least initially. If we take learning seriously at a macroeconomic timescale, this connection to volatility may be especially important.

1.4 Robustness

This Section demonstrates the robustness of these results by presenting the mean dynamics under Recursive Least Squares (RLS) and alternative parameterisations of the illustrative model. The equilibrium under RLS is qualitatively similar to that under Neural Network Learning, but it does not pass the DHM accuracy test, so agents' forecast errors are not independent of their information sets and we cannot consider it a rational expectations equilibrium. The results are also robust to alternative parameterisations, including if the central steady state is a source rather than a sink under perfect foresight.

1.4.1 Mean dynamics with RLS learning

As an REE in a non-linear model with state variables cannot be exactly expressed with finite parameters, it is not possible to specify a linear regression that exactly nests an REE. Therefore, I explore the mean dynamics of the system under RLS learning, as described by Evans and Honkapohja (2001), with a simple PLM.[17] When agents learn using RLS with higher order terms as regressors, simulations and mean dynamics converge to an equilibrium that is fundamental, stationary and stable under learning. While this learnable equilibrium is qualitatively similar to the solution under neural network learn-

[17] An alternative approach would be to define the non-linear model as a grid over the state space in which the PLM and ALM are both characterised as transition probability matrices (Chen and White, 1998). This non-parametric learning approach yields less interpretable results though, suffers from the curse of dimensionality and is a less plausible characterisation of learning.

ing, it does *not* have the properties of a REE.

I consider a PLM that is linear in parameters, but includes square and cubic terms for the endogenous state variable y_{t-1}. This will not nest any of the global REE, but the cubic term allows agents to capture the multiple steady states (it also does nest the local REE in a linearised system around any of the steady states). To emphasise the robustness of the qualitative properties of the learnable equilibrium, agents are also not able to directly observe the shocks.

$$\begin{aligned} \pi^e_{t|t} =& \beta_{\pi,1} + \beta_{\pi,2} y_{t-1} + \beta_{\pi,3} \pi_{t-1} + \beta_{\pi,4} y^2_{t-1} + \beta_{\pi,5} y^3_{t-1} \\ y^e_{t|t} =& \beta_{y,1} + \beta_{y,2} y_{t-1} + \beta_{y,3} \pi_{t-1} + \beta_{y,4} y^2_{t-1} + \beta_{y,5} y^3_{t-1} \end{aligned} \tag{1.17}$$

Define agents' beliefs as the parameters of their PLM as θ_t. The evolution of these beliefs under RLS learning is then a stochastic recursive algorithm as in Equation 1.14, and finding the equilibrium to which the mean dynamics converge amounts to identifying the fixed point of the differential equation $\frac{dh(\theta)}{d\theta}$. I characterise $h(\theta)$ (Equation 1.15) by discretising the state space X_t and computing a transition probability matrix for the conditional probability of (π_t, y_t) given (π_{t-1}, y_{t-1}). These transition probabilities are determined by values of the exogenous shocks for which each transition would occur. From this transition probability matrix we can extract the marginal probability distribution of the initial state (π_{t-1}, y_{t-1}) and then applying Bayes' Rule gives the unconditional distribution of X_t. Appendix A.3 describes this process in detail.

The mean dynamics of RLS learning converges to the PLM coefficients in Table 1.3. The corresponding ALM will have a different parameterisation due to the non-linearity, so cannot be compared directly. The existence of this fixed point in $h(\theta)$ tells us that there is a point at which the system is in expectation stable under learning. The stability of this fixed point is equivalent to E-stability in that it tells us that the system will return to this equilibrium after small deviations in beliefs.[18]

Table 1.3: RLS mean dynamics PLM coefficients

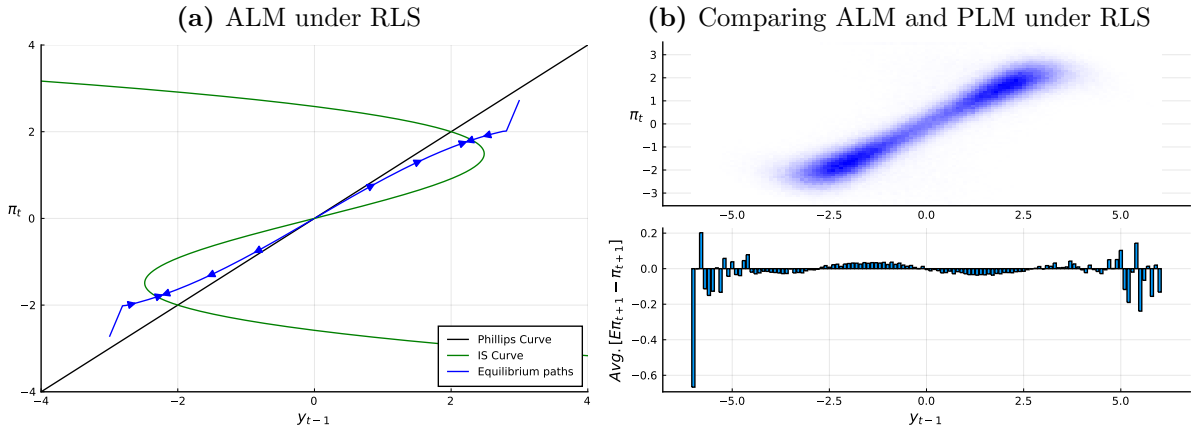
	β_1	β_2	β_3	β_4	β_5
π	0.0000	1.8451	0.0878	0.0000	-0.2120
y	0.0000	1.0677	0.0084	0.0000	-0.0205

The implied phase diagrams, in Figure 1.13, show that this RLS equilibrium is qualita-

[18]The system converges to very similar values in simulations with constant gain learning, verifying that the system converges to something close to the mean dynamics in finite time and that the approximation is sufficiently accurate.

tively similar to that with Neural Network Learning. Panel (a) shows the ALM implied by the beliefs in Table 1.3 and panel (b) compares agents' PLM and the ALM under RLS.. The central steady state has once again become a source, and the outer steady states are stable, so the system will switch between the neighbourhoods of these outer steady states.

Figure 1.13: The Perceived and Actual Laws of Motion with Recursive Least Squares mean dynamics



Unlike the equilibrium under Neural Network Learning, the agents' PLM and the ALM do not match up, so agents make large systematic mistakes in their expectations, as shown in panel(b). The RLS specification examined here is for partial information, so agents do not observe contemporary shocks, but Appendix A.3.1 shows that under full information agents still make systematic errors in predicting current period variables which are much larger than those under a neural network.

Unsurprisingly, this solution therefore fails the DHM accuracy test and cannot be described as a REE. Table 1.4 reports the results for 1,000 draws of 500 periods simulated with the RLS beliefs for three different specifications of the $h(s_t)$ function.[19]

Table 1.4: Den Haan & Marcet test results with mean dynamics beliefs

	Saddle	
$h(s_t)$	Lower-tail	Upper-tail
Constant	0.0060	0.7458
Constant, y_{t-1} and π_{t-1}	0.0000	1.0000
Extensive	0.0000	1.0000
Note:	1,000 draws of 500 periods	

[19] As agents do not observe the shocks in this specification, their information sets only include lagged values π_{t-1} and y_{t-1} as agents will not be able to perfectly predict π_t and y_t.

While the RLS solution is stable and stationary, it does not satisfy the conditions of a REE.

1.4.2 Alternative parameterisations

In the previous pages, the model has been analysed for particular parameter values under which the central steady state is a complex sink under perfect foresight. The results are qualitatively similar whether the central steady state is complex/real or a source/sink. E-stability analysis when the model is linearised around each steady state suggests that similar results will obtain whenever the central steady state is not determinate.

As discussed in Section 1.2.2, the dynamics of the perfect foresight system in the neighbourhood of a steady state is determined by the eigenvalues of the coefficient matrix Φ in Equation 1.10. The model has one forward-looking/control variable (π_t) and one predetermined/state variable (y_{t-1}), so the system is determinate around a particular steady state if Φ has one eigenvalue inside the unit circle, and one outside the unit circle. If both eigenvalues are inside the unit circle, the system is a sink under perfect foresight and thus indeterminate, as there are many candidate locally stationary REE. If both eigenvalues are outside the unit circle, then the system is a source under perfect foresight, and so there are no locally stationary REE at all. However, we can still think of the system as being indeterminate in the sense that there will be many explosive paths, so there is still no unique prediction for behaviour under rational expectations.[20]

There are thus three classes additional classes of parameterisation to consider: when the central steady state is a real sink, a real source and a complex source. Appendix A.4 presents phase diagrams for an example of each of these, showing that there is a qualitatively similar Learnable-REE in these cases where the central steady state is unstable and the two outer steady states are locally stable. Under neural network learning, the equilibria in these cases also pass the DHM accuracy test, so we can conclude that they are at least very close to a REE. Whether a steady state is a source or a sink under perfect foresight, it will be unstable under learning if it is not determinate. The results described for the particular parameterisation of the illustrative model therefore appear to be very general.[21]

[20] Figure A.5 in Appendix A.4 shows the ranges in the parameter space for which the perfect foresight system will be complex/real and a source/sink/saddle around the central steady state. When the positive feedback from π_t to y_t between the two variables is high, then the central steady state will be a sink and so the model is indeterminate. This positive feedback is thus the complementarity behind the indeterminacy in this model.

[21] Whether the system converges to a symmetric equilibrium, in which transitions between the steady states are well defined in beliefs, in a given simulations depends in crucially on the variance and persis-

The results of this paper are entirely consistent with the results for linear models in the learning literature, which suggest that determinacy and learnability are closely linked. They simply build on the negative result that the REE converging into an indeterminate steady state are not learnable, adding a positive prediction for the behaviour of indeterminate RE models under learning. The main contribution of the paper is to show the existence and learnability of transition paths between the steady states of an indeterminate RE model, and that these transition paths are consistent with rational expectations.

The E-stability concept of Evans and Honkapohja (2001) gives us analytic results for learnability in the neighbourhood of a steady state where the approximation is exact, under a broad class of forecast error-based learning algorithms. We therefore know that the REE converging into a local indeterminacy steady state will not be exactly learnable. This paper takes this a step further by showing that, at least in some cases, when indeterminacy is bounded there will be a learnable path between local indeterminacy and local determinacy steady states.

1.4.3 Implications for Empirical Tests for Indeterminacy

Previous literature has sought to provide empirical evidence for indeterminacy. Most notably, Lubik and Schorfheide (2004) propose an empirical test for indeterminacy based on comparing Bayesian estimation of a model with and without a sunspot shock. This test relies on the assumption that agents have fully rational expectations but when there is indeterminacy an exogenous sunspot process selects one of the candidate paths. I show that researchers conducting this sort of test may mistakenly find evidence of sunspots when in fact agents are behaving as in the Learnable-REE, where the local indeterminacy steady state is unstable and there are no self-fulfilling sunspot shocks to expectations.

Linearising the model around the central steady state gives a linear model for which there is thus no unique stationary rational expectations solution. We can therefore define a "sunspot" shock which determines the expected price and will turn out to be self-fulfilling. There is thus an additional stochastic process, on top of the two fundamental shocks, which needs to be calibrated or estimated. Given this sunspot shock we can then solve and estimate our illustrative model, linearised around the central steady state, as a rational expectations model. Appendix A.5 explains this in more detail.

tence of the shock processes. If the volatility of these shocks is too low, then the system will tend to converge to the neighbourhood of one of the outer steady states and stay there.

The Lubik and Schorfheide (2004) test for indeterminacy amounts to estimating a linear model with sunspots and a model without sunspots on the same data and then comparing their fit with a posterior odds ratio. I perform this test on simulated data from the non-linear system under the Neural Network Learnable-REE. I estimate the two models on 2,000 simulated periods of the non-linear model with the neural network Learnable-REE by Maximum Likelihood and by Bayesian inference.

Under both Bayesian and Maximum Likelihood estimation, the test shows a clear preference for the model with sunspots over the standard linear rational expectations model. With Bayesian estimation, the log posterior for the sunspot model is -5,601 and for the standard model is -25,338. So the posterior odds ratio favours the model with sunspots by a factor of $\exp(19,373)$. Similarly, with Maximum Likelihood estimation the log likelihood of the model with sunspots is -5,554 and for the standard model is - 25,314, so the likelihood ratio favours the model with sunspots by a factor of $\exp(19,760)$. In either case, it is clear that the data very strongly prefer the "sunspot" model over a determinate linear rational expectations model. Estimation results for both models are provided in Appendix A.5.

As the true DGP is not a linear model, the estimation does not recover the true parameters. The important point is that a researcher testing for sunspots in this environment might conclude that the data is explained by a linear model with a sunspot shock, when in reality this is not the case. This approach is in line with standard practice in macroeconomic literature based on linearising/perturbing models around a steady state and then solving and estimating these approximated models. The results presented here suggest that this is not an appropriate way to deal with models featuring indeterminacy.

1.5 Indeterminacy and the Zero Lower Bound

Since Benhabib et al. (2001) pointed out the existence of a second deflationary steady state in a New Keynesian model with a lower bound on nominal interest rates, macroeconomists have struggled with how to account for it.[22] This Sections applies the neural network learning framework and the lessons outlined above to this well-known example of indeterminacy. It shows that, provided that the indeterminacy caused by the constraint on monetary policy is bounded, there exists a learnable and stable equilibrium in which the economy is moved between a deflationary and a target inflation steady state by tran-

[22]Policy makers have also taken it seriously, worrying that negative shocks "could push the economy into an unintended, low nominal interest rate steady state" (Bullard, 2010).

sitory shocks.

The implications of a lower bound on nominal interest rates has been an important focus for macroeconomic research in the past two decades. Not only does the discontinuity and second steady state make it difficult to apply standard perturbation solution methods, but the implied indeterminacy is potentially troubling, as there exist multiple stationary solutions under rational expectations. This absence of a positive prediction about equilibrium behaviour, limits the value of models to policy makers. For example, Mertens and Ravn (2014) show that fiscal multipliers can vary substantially across these different equilibria.

In addition to the challenge of multiplicity indeterminacy also introduces instability under learning.Christiano et al. (2018) claim that this unintended steady state should not be a concern as it is not E-stable and so will not be arrived at as a result of learning. But by restrict long run expectations they ignore the instability illustrated by Evans et al. (2008) who show that the intended steady state in a New Keynesian (NK) model is locally but not globally stable under learning, and negative shocks can lead to unbounded deflationary spirals.[23] This lack of stability causes problems for global solution methods. As analysed systematically by Richter and Throckmorton (2015), if error variances are too high and expectations place too high a weight on ZLB episodes, numerical solution methods will not converge. This may be because global solution methods which rely on updating an initial guess based on some measure of inaccuracy will are in practice very similar to learning.[24]

Two common approaches are to invoke an exogenous sunspot which coordinates on one of the rational expectations solutions Aruoba et al. (2018), or to solve forward assume that at some future period the economy will be at the "target rate" steady state and shocks are set to zero Braun and Körber (2011); Christiano et al. (2018).[25] Neither of these approaches are fully satisfactory as they either offer no explanation of what causes ZLB episodes, or they do not allow for agents to factor in future ZLB episodes. Moreover, as shown by Ascari and Mavroeidis (2020), even when "target rate" steady state is assumed to be absorbing there are many candidate REE, which have very different implications, even if we only restrict attention to minimum state variable solutions.

[23]Benhabib et al. (2014) study this instability further and suggest that sufficiently aggressive policy can prevent these deflationary spirals by keeping expectations anchored

[24]Other attempts to use global solution methods are consistent with this. For example, Fernández-Villaverde et al. (2015) report results for a calibration in which the economy is at the ZLB 5.53% of the time.

[25]This latter approach is taken in the popular OccBin toolkit Guerrieri and Iacoviello (2015).

Perhaps the most similar to this paper's approach to indeterminacy in this context are Evans et al. (2016) and Eggertsson et al. (2019). Both these papers place a lower bound on the indeterminate region by introducing a third steady state around which rational expectations are locally determinate. However, neither identify a fully stochastic non-linear equilibrium in which agents are aware of both determinate steady states and have accurate beliefs about the transitions between them. In the remainder of this Section, I apply neural network learning to a version of the former, and show that is able to identify such an equilibrium.

1.5.1 NK model with Lower Bounds (Evans et al., 2016)

Evans et al. (2016) place lower bounds on inflation and output, as well as the nominal interest rate, in a linearised NK model and show that this lower bound equilibrium is also locally stable under learning. As they only allow for agents with linear Perceived Laws of Motion, they are not able to capture an equilibrium in which agents' beliefs allow for more than one steady state at a time. Neural network learning is able to identify such an equilibrium.

As an endogenous state variable is needed for agents' expectations to allow for multiple steady states, I introduce persistence into the benchmark model used by Evans et al. (2016).[26] The model can then be expressed as the following three equilibrium conditions, where 1.18 is the Phillips Curve, 1.19 is the output Euler Equation and 1.20 is a Taylor Rule, all of which are subject to a lower bound.

$$\pi_t = \max\{\beta E_t \pi_{t+1} + \kappa y_t, \pi^{lim}\} + \epsilon_{\pi,t} \tag{1.18}$$

$$y_t = \max\{(1-\eta) E_t y_{t+1} + \eta y_{t-1} - \sigma(r_t - E_t \pi_{t+1}), y^{lim}\} + \epsilon_{y,t} \tag{1.19}$$

$$r_t = \max(\phi_\pi \pi_t + \phi_y y_t, r^{lim}) \tag{1.20}$$

The shocks to the Euler Equation and Phillips curve follow an AR(1) process with a low variance (0.001), but with rare large shocks of ± 0.03 that appear on average every 100 periods.[27] The model's other parameters are calibrated with standard values shown in Table 1.5. The lower bounds are set such that they correspond to an interest rate 2.5% below the level at the "target rate" steady state, output 5% below and inflation 3% below.

[26] These could be micro-founded through external consumption habits as in, for example Smets and Wouters (2007).

[27] Future work will relax this assumption.

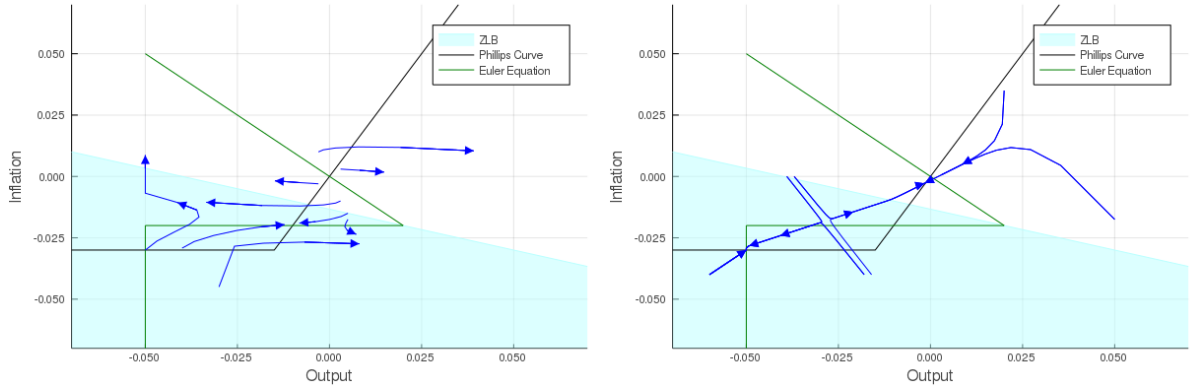
Table 1.5: Parameter values for NK model with lower bounds

	β	θ	κ	η	σ	ϕ_π	ϕ_y	r^{lim}	y^{lim}	π^{lim}
Value	0.99	0.3	0.024	0.5	0.157	1.5	0.5	-0.025	-0.05	-0.02

Figure 1.15 shows the steady state conditions and deterministic steady states of the model. The steady state in the upper right is the "target rate" steady state which is locally determinate and stable under learning. The central steady state is the "ZLB steady state" identified by Benhabib et al. (2001), which is locally indeterminate and unstable under learning. The lower left steady state is the lower bound steady state introduced by Evans et al. (2016), which is determinate and stable under learning.

Figure 1.15: NK model with lower bounds

(a) Perfect foresight **(b)** Neural Network Learnable Equilibrium



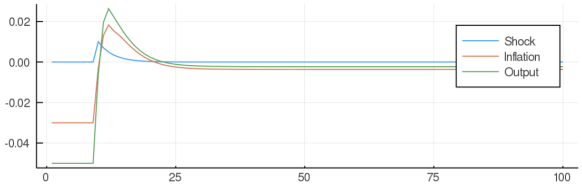
Neural network learning is applied as described in Section 1.3 and, as expected, converges to an equilibrium in which the economy switches between the two local determinacy steady states, based on the large shocks to ϵ_t. Evans et al. (2016) obtain observationally similar results by analysing the model under linear learning where agents have a short memory and so can "forget" the steady state at which the economy was previously and thus converge to a new steady state.

Figure 1.17 shows Impulse Response Functions from the equilibrium which the neural network learning converges to. Panel (a) shows that if the economy starts in the lower steady state, a sufficiently large positive shock will move it permanently to the higher steady state. Panel (b) analogously shows that if the economy starts in the upper steady state, a sufficiently large negative shock will move it permanently to the lower steady state. In Panel (c) we see that if the economy is already in the upper steady state, positive shocks will only have a transitory effect. At the lower steady state, negative shocks will have no effect as the economy is already at it's lower bound.[28]

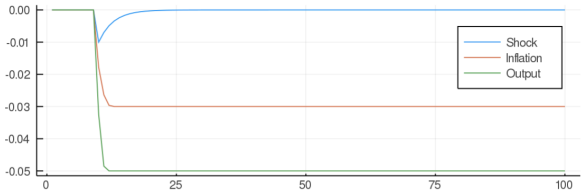
[28]Note that the upper stochastic steady state is slightly lower than the upper deterministic steady

Figure 1.17: NK model with ZLB and lower bounds on y_t and π_t

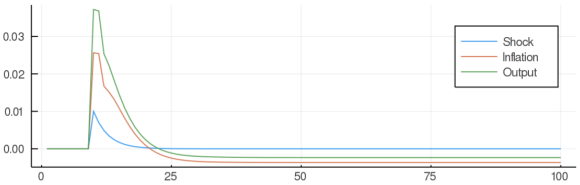
(a)



(b)



(c)



This example highlights a key implication of the the insight that local indeterminacy steady states are unstable under learning. If that indeterminacy is bounded, then transitory shocks can have permanent effects by moving the economy to a new steady state. This may be of some encouragement to policy-makers concerned with secular stagnation, as it implies permanently moving the economy to a more desirable steady state may not require a permanent policy intervention. These lower bounds on inflation and output here are clearly introduced in an ad hoc manner, future work could remedy this by applying the same framework to the OLG model with a ZLB and downward nominal wage rigidity in Eggertsson et al. (2019). The downward nominal wage rigidity introduces a third, determinate steady state very similar to the case discussed above. Gibbs (2018) shows that the upper and lower steady states in this model are E-stable. This would allow the analysis of both fiscal and monetary policy responses to secular stagnation in a rich micro-founded model.

state, as agents' expectations place a small probability on a shock that will move them to the lower steady state.

1.6 Conclusions

This paper presents a resolution to the challenge of indeterminacy. Steady states around which behaviour under rational expectations is indeterminate should be treated as unstable and explosive. Furthermore, this need not come at the cost of rationality. When the forces driving indeterminacy are bounded there exists a Learnable-REE which is fundamental and stationary.

These results have some negative implications. Analysing non-linear models under perfect foresight can be misleading, as they suggest that indeterminate steady states are super stable. Furthermore, learnability casts doubt on the interpretation of indeterminacy as leading to self-fulfilling sunspot shocks to beliefs.

However, these are outweighed by the positive implications. Learning can generate unique and determinate outcomes in models that are indeterminate under rational expectations. If agents learn with a sufficiently flexible algorithm, these learnable equilibria are consistent with rationality. When faced with models that display indeterminacy we ought to therefore consider what forces generate that indeterminacy as well as whether and how they might be bounded. If our model has multiple steady states, we can identify a global Learnable-REE which captures transitions between steady states without relying on some exogenous regime switching process. This can generate potentially important hysteresis effects and state-contingent responses to shocks. The flexibility and efficiency of neural networks for function approximation can help us identify such equilibria.

Chapter 2

Bayesian Topic Regression: Controlling for Text*

Abstract

Quantitative analysis using observational text data is becoming increasingly popular across social science. In many contexts, this require estimating models that include both numerical and text data. This paper presents the Bayesian Topic Regression (BTR) model that uses both text and numerical information to predict an outcome variable. Our model jointly estimates an interpretable representation of the text and the relationship of this text and additional numerical variables to predict an outcome. This joint estimation respects the Frisch-Waugh-Lovell theorem and is therefore less likely to lead to incorrect parameter estimates. We show with both synthetic and semi-synthetic datasets that our joint approach recovers ground truth with lower bias than any benchmark model, when text and numerical features are correlated. Experiments on two real-world datasets demonstrate that jointly estimating a representation of text and a predictive model also yields superior prediction results, compared to strategies that estimate regression weights for text and non-text features separately.

*This chapter is co-authored with Maximillian Ahrens, Jan-Peter Calliess, Vu Nguyen.

2.1 Introduction

In many contexts, such as that which will be presented in Chapter 3, information that is relevant to a causal effect might be captured in text data. A key challenge when working with unstructured data such as text is the representation and selection of text features. In this paper, we present a model that allows this feature extraction to occur jointly with the estimation of the relationship between text documents and associated numerical variables. In particular, we are interested in cases where researchers wish to interpret the relationship between numerical variables while controlling for text data in a transparent and interpretable way.

Our model is based on a combination of a supervised Latent Dirichlet Allocation (sLDA) topic model Blei and McAuliffe (2008) and a Bayesian linear regression.[1] Topic models assign each word in a text document to one of a pre-set number of topics. Each text document can therefore be represented as proportions of each topic, and each topic can be represented as a probability weighting on the unique words that appear across all documents. In our model, observations are thus made up of a text document (represented as mixtures of topics), additional numerical covariates, and a (numerical) outcome variable. The outcome variable is predicted by a linear combination of the topic proportions, numerical covariates and a normally distributed residual. Extensions to multiple documents per observation, observations without documents, and interaction terms between text and numerical features are also provided.

We jointly estimate both this topic representation of the text documents and the parameters of the regression model for two reasons. Firstly, by using information from the regression model in estimating the topics, the model is better able to represent the features of the text that are most relevant to our research question. Secondly, if we wish to interpret the parameters of the regression as causal, then potential correlations between explanatory variables need to be taken into account when

Considering both text and numerical data jointly can be crucial for conducting unbiased statistical inference. When explanatory variables are correlated with each other and with the outcome, the Frisch-Waugh-Lovell (FWL) theorem Frisch and Waugh (1933); Lovell (1963), implies that all regression coefficients should be estimated jointly, or es-

[1] "Supervised" learning refers to algorithms that aim to predict an outcome variable. This is in contrast to "unsupervised" learning which refers to algorithms that aim to describe or cluster data.

timates will be biased.[2] Text features themselves are "estimated data". If they stem from supervised learning which estimated the text features with respect to the outcome variable separately from the numerical features, then the resulting estimated effects will be biased. We show both synthetic and semi-synthetic datasets that our BTR model recovers the ground truth more accurately than a wide range of benchmark models.

To illustrate this, take the example of hotel reviews from *Booking.com* discussed in Section 2.6. Our data here includes (i) the text of the review written by a customer w; (ii) the numerical rating (on a scale of 1-10) that they give the hotel y; and (iii) some additional numerical data on the hotel, in particular the historical average rating x. Our research question is whether customers give higher ratings (y) to hotels that already have high historical average ratings (x). If we want to interpret this relationship causally, we will want to control for confounding factors. An obvious example here would be the quality of the customer's experience, for which we can use the text review (w) as a proxy. Our question thus becomes: do customers give hotels a higher rating if they already have a high rating, even once we control for how they describe their experience in the text review?

However, in order to control for the confounding information in the text, we need to extract the relevant features in a way that can be concisely expressed numerically. As the information in the text is likely correlated both with the historical average rating and this customer's rating, estimating the relationship in multiple steps can lead to incorrectly identifying the relationship between them and the customer's rating. In other words, as better hotels generally get more positive reviews and higher ratings, x and w are positively correlated and this needs to be taken into account to identify the effect of x on y. By extracting the relevant information from w at the same time as estimating the relationship between w, x and y, our model is able to take into account these correlations and so yield more accurate estimates of these relationships.

The remainder of this paper is structured as follows. Section 2.2 discusses related literature both from social science and natural language processing (NLP). Section 2.3 then describes the model in detail and Section 2.4 describes our estimation algorithm. The fact that our model yields accurate parameter estimates compared to other models in the literature is shown on synthetic data in Section 2.5.2 and on semi-synthetic data in 2.6. Section 2.7 then demonstrates on two real-world customer review datasets - *Yelp* and *Booking.com* - that a joint supervised learning strategy, using both text and non-text fea-

[2]Alternatively, all explanatory variables must be orthogonalised from one another, but this is not possible if text features are extracted in a supervised manner.

tures, also improves prediction accuracy of the target variable compared to a 'two-step' estimation approach with the same models.

2.2 Related Work

2.2.1 Text data as Explanatory Variables in Social Science

Topic modelling has been an important tool for researchers using unstructured data, see Vayansky and Kumar (2020) for a technical review of the field and Jelodar et al. (2019) for a survey of applications. Recent applications in social science include analyses of CEO time usage (Bandiera et al., 2020), central bank transcripts (Hansen et al., 2018), political ideology (Preoţiuc-Pietro et al., 2017), socio-economic maps (Hong et al., 2016), and consumer purchases (Athey et al., 2018).

The most common approach in those applications is to perform unsupervised feature extraction on a corpus of documents, such as LDA (Blei et al., 2003) or Probabilistic Latent Semantic Analysis (Hofmann, 1999), and use topic mixtures as features in a subsequent regression task. This approach is consistent with the Frisch-Waugh-Lovell Theorem, if the unsupervised representations of text capture the relevant features. However in many contexts this may not be the case as, particularly when using text as a control variable, it may not be clear a priori which text features contain important information. Our model addresses this by learning topic representations that are relevant to the regression equation of interest, while allowing for interpretable regression parameters. To the best of our knowledge, our paper is first to do this both in the social science and NLP literature.

2.2.2 Related work from Natural Language Processing

Traditionally, topic features are obtained via Markov Chain Monte Carlo (MCMC) or mean-field VI methods, but advances in neural variational inference (Kingma and Welling, 2014; Rezende et al., 2014; Mnih and Gregor, 2014) have led to the development of unsupervised neural topic model counterparts such as Miao et al. (2016), Miao et al. (2017) and Srivastava and Sutton (2017). These often showing the ability to estimate more expressive and interpretable topics. However, as unsupervised topic features are not optimised for the supervised task, they are often outperformed by simple dictionary methods, such as the sentiment indices proposed by Correa et al. (2017) and Dodds et al. (2011).

Given our focus on interpretation, we opt for a Gibbs Sampling implementation. This provides statistical guarantees of providing asymptotically exact samples of the target

density while (neural) variational inference does not (Robert and Casella, 2013). Blei et al. (2017) point out that MCMC methods are preferable over variational inference when the aim of the task is to obtain asymptotically precise estimates. We provide an efficient *Julia* (Bezanson et al., 2017) implementation.

Our model is inspired by the architecture of supervised Latent Dirichlet Allocation (sLDA) model of Blei and McAuliffe (2008). We extend this framework to jointly learn the supervised topic representation alongside the association between the response variable and both topic features and non-text features. Furthermore, we follow a fully Bayesian approach, placing prior distributions on the regression weights and measurement error variance. To the best of our knowledge, none of the existing supervised (neural) topic models offer these characteristics.

There are several approaches to supervised topic modelling for regression. The earliest is the sLDA model of Blei and McAuliffe (2008), which optimizes with respect to the joint likelihood of the document data and the response variable. Building on this, Maximum Entropy Discrimination Latent Dirichlet Allocation (MedLDA) (Zhu et al., 2012) optimizes with respect to the maximum margin principle, Spectral-sLDA (Wang and Zhu, 2014) proposes a spectral decomposition algorithm, and BPsLDA (Chen et al., 2015) uses backward propagation over a deep neural network.[3] Other recent supervised topic models that can handle covariates are for example the Structural Topic Model (STM) (Roberts et al., 2016) and Diagonal Orthant Latent Dirichlet Allocation (DOLDA) (Magnusson et al., 2020). DOLDA was not designed for regression nor for causal inference setups. Topics models in the spirit of STM incorporate document metadata, but in order to better predict the content of documents rather than to predict an outcome. Our approach also differs from latent factor recommender models (Agarwal and Chen, 2010; McAuley and Leskovec, 2013; Elbadrawy and Karypis, 2015), since those models focus on text-based latent meta-data for the prediction of the outcome variable, estimating separate regressions for each individual.

Given that BPsLDA has been found to outperform sLDA, MedLDA and several other models (Chen et al., 2015), we include it in our benchmark list for two-stage models. We also include a Gibbs sampled sLDA, to have a two-stage model in the benchmark list that is conceptually very similar to BTR in the generative topic modelling part. Unsupervised LDA (Blei et al., 2003; Griffiths and Steyvers, 2004) and a neural topic model counterpart

[3]Similarly, supervised topic models such as Lacoste-Julien et al. (2009), Chong et al. (2009) and Ramage et al. (2009) focus on classification instead of regression, but are otherwise similar in setup to regression based topic models.

GSM (Miao et al., 2017) are also added for comparison.

Perhaps the closest approach to BTR in terms of generative model, if not estimation, is the SCHOLAR model proposed by Card et al. (2018). This is a supervised topic model that generalises both sLDA (Blei and McAuliffe, 2008) as it allows for predicting labels, and SAGE Eisenstein et al. (2011) which handles jointly modelling covariates via 'factorising' its topic-word distributions (β) into deviations from the background log-frequency of words and deviations based on covariates. SCHOLAR is solved via neural variational inference Kingma and Welling (2014); Rezende et al. (2014). However, it was not primarily designed for causal inference. We extend SCHOLAR with a linear regression layer (rSCHOLAR) to allow direct comparison with BTR. This regression layer is jointly optimized with the main SCHOLAR model via backpropagation, replacing the original downstream cross-entropy loss with mean squared error loss.

More recently, supervised neural network architectures have been enhanced with topic model components. TopicRNN Dieng et al. (2017) combines an LDA model with a recurrent neural network (RNN) for unsupervised tasks, Cao et al. (2015) propose a supervised neural topic model to learn topic-word and topic-document distributions, and TAM Wang and Yang (2020) enhances the attentional heads in a bidirectional RNN with topic vectors to provide more global context in supervised tasks. We include both a recurrent neural network with attention (Bahdanau et al., 2015) and the topic attention model proposed by Wang and Yang (2020) as benchmarks in our prediction exercise, showing that BTR can compete in prediction with even these state of the art approaches.

2.3 BTR Model

2.3.1 An overview of the model

The general problem is to estimate the relationship between two types of explanatory variable (numerical data $\boldsymbol{X}$ and text data $\boldsymbol{W}$) and an outcome variable ($\boldsymbol{y}$) while expressing the relevant features of $\boldsymbol{W}$ in a concise and interpretable way.

$$\boldsymbol{y} = f(\boldsymbol{X}, g(\boldsymbol{W})) \tag{2.1}$$

The BTR model assumes that the $g()$ function that converts the raw text data into a numerical representation is a topic model based on the well-known LDA model, and that $f()$ is linear in X and $(\bar{Z}) = g(\boldsymbol{W}))$ with a normally distributed idiosyncratic error term.

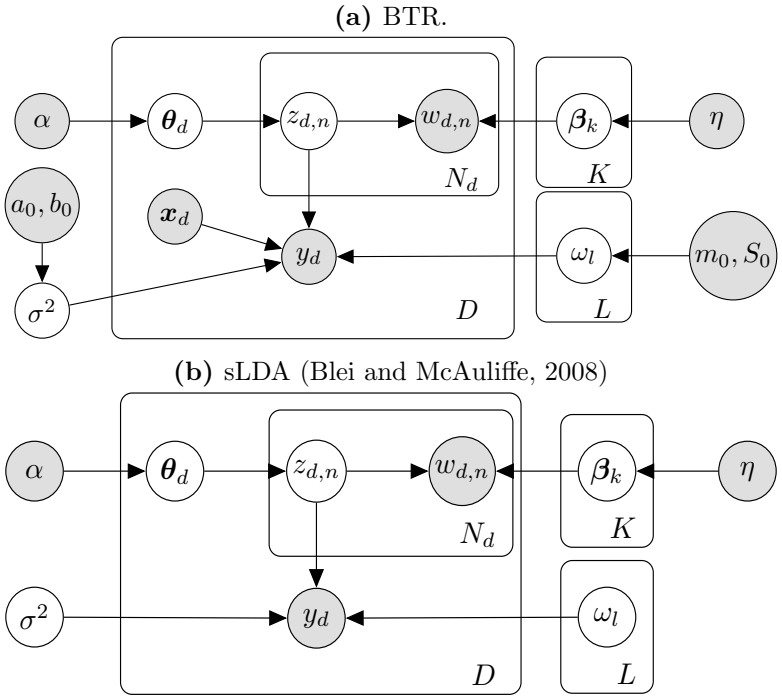
This Section describes the BTR model in detail, describing the regression component in Section 2.3.2 and then the topic model component in Section 2.3.3. Table 2.1 shows all relevant variables and parameters.

Table 2.1: Glossary of notation

Type	Parameter	Description
Dimensions	D	Total number of observations
	K	Number of topics
	L	Number of numerical covariates
	V	Length of vocabulary
	N_d	Number of words in document d
Data	$\boldsymbol{W}$	Text documents (represented as a $D \times V$ matrix)
	$\boldsymbol{X}$	Numerical covariates ($D \times L$ matrix)
	$\boldsymbol{y}$	Outcome variable ($D \times 1$ vector)
Estimated	$\boldsymbol{\theta}$	Distribution of documents over topics ($D \times K$ matrix)
	$\boldsymbol{Z}$	Assignment of words to topics ($N_d \times 1$ vector for each d)
	$\boldsymbol{A}$	Regression design matrix ($D \times (K + L)$ matrix)
	$\boldsymbol{\beta}$	Distribution of topics over vocabulary ($K \times V$ matrix)
	$\boldsymbol{\omega}$	Regression coefficients ($(K + L) \times 1$ vector)
	σ^2	Regression residual variance (scalar)
Prior	α	Dirichlet prior on $\boldsymbol{\theta}$
	η	Dirichlet prior on $\boldsymbol{\beta}$
	m_0, S_0	Mean and variance of Normal prior on $\boldsymbol{\omega}$
	a_0, b_0	Shape and scale of Inverse-Gamma prior on σ^2

Figure 2.1 provides a graphical representation of the generative model for BTR. The grey nodes represent observed data or priors that are chosen by the researcher and the white nodes latent variables/parameters that are estimated. Comparing this the the graphical model for Blei and McAuliffe's sLDA model, we see the addition of additional numerical variables $\boldsymbol{x}_d$ as well as the priors on the regression coefficients $\boldsymbol{\omega}$ and error variance σ^2. The BTR model is thus a fully Bayesian version of the sLDA model, with additional numerical explanatory variables.

Figure 2.1: Graphical models for BTR and sLDA



2.3.2 Regression Model

As described below, our text data is represented as topic proportions $\bar{\boldsymbol{Z}} \in \mathbb{R}^{D \times K}$, where K is the number of topics. Define $\boldsymbol{A} = [\bar{\boldsymbol{Z}}, \boldsymbol{X}]$ as the overall regression design matrix containing all features (optionally including interaction terms between topics and numerical features).[4] Assuming that the relationship between the outcome and explanatory variables is linear, subject to Gaussian iid errors $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$, the model's regression equation is then

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{\omega} + \boldsymbol{\epsilon}, \tag{2.2}$$

such that

$$p(\boldsymbol{y}|\boldsymbol{A}, \boldsymbol{\omega}, \sigma^2) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{A}\boldsymbol{\omega}, \sigma^2 \boldsymbol{I}) = \prod_{d=1}^{D} \mathcal{N}(y_d|\boldsymbol{a}_d\boldsymbol{\omega}, \sigma^2 \boldsymbol{I}), \tag{2.3}$$

where $\boldsymbol{a}_d$ is the dth row of design matrix $\boldsymbol{A}$. We model our prior beliefs about parameter vector $\boldsymbol{\omega}$ by a Gaussian density

$$p(\boldsymbol{\omega}) = \mathcal{N}(\boldsymbol{\omega}|\boldsymbol{m}_0, \boldsymbol{S}_0) \tag{2.4}$$

[4]For example, an interaction between a document's sentiment score and the topic proportions might be of interest in many applications, as topics might have different effects depending on the tone of the document.

where mean $\boldsymbol{m}_0$ and covariance matrix $\boldsymbol{S}_0$ are hyperparameters. Following Bishop (2006), we place an Inverse-Gamma prior on the conditional variance estimate σ^2 with shape and scale hyperparameters a_0 and b_0

$$p(\sigma^2) = \text{Inv-Gamma}(\sigma^2|a_0, b_0). \tag{2.5}$$

Placing priors on all our regression parameters allows us to conduct full Bayesian inference, which not only naturally counteracts parameter over-fitting but also provides us with well-defined posterior distributions over $\boldsymbol{\omega}$ and σ^2 as well as a predictive distribution of our response variable.

Due to the conjugacy of the Normal-Inverse-Gamma prior, the regression parameters' posterior distribution conditional on $\boldsymbol{A}$ has a known Normal-Inverse-Gamma distribution Stuart et al. (1994)

$$p(\boldsymbol{\omega}, \sigma^2|\boldsymbol{y}, \boldsymbol{A}) \propto p(\boldsymbol{\omega}|\sigma^2, \boldsymbol{y}, \boldsymbol{A})p(\sigma^2 \mid \boldsymbol{y}, \boldsymbol{A}) = \mathcal{N}\left(\boldsymbol{\omega}|\boldsymbol{m}_n, \sigma^2\boldsymbol{S}_n^{-1}\right) \text{Inv-Gamma}\left(\sigma^2|a_n, b_n\right). \tag{2.6}$$

where $\boldsymbol{m}_n$, $\boldsymbol{S}_n$, a_n and b_n follow standard updating equations for a Bayesian Linear Regression (Bishop, 2006).

$$\begin{aligned} \boldsymbol{m}_n &= (\boldsymbol{A}^\intercal\boldsymbol{A} + \boldsymbol{S}_0)^{-1}(\boldsymbol{S}_0\boldsymbol{m}_0 + \boldsymbol{A}^\intercal\boldsymbol{y}) \\ \boldsymbol{S}_n &= (\boldsymbol{A}^\intercal\boldsymbol{A} + \boldsymbol{S}_0) \\ a_n &= a_0 + N/2 \\ b_n &= b_0 + \left(\boldsymbol{y}^\intercal\boldsymbol{y} + \boldsymbol{m}_0^\intercal\boldsymbol{S}_0\boldsymbol{m}_0 - \boldsymbol{m}_n^\intercal\boldsymbol{S}_n\boldsymbol{m}_n\right)/2. \end{aligned} \tag{2.7}$$

2.3.3 Topic Model

The estimated topic features $\bar{\boldsymbol{Z}}$, which form part of the design regression matrix $\boldsymbol{A}$, are generated from a supervised model that builds on an LDA-based topic structure Blei et al. (2003).

We have d documents in a corpus of size D, a vocabulary of V unique words and K topics. A document has N_d words, so that $w_{d,n}$ denotes the nth word in document d. The bag-of-words representation of a document is $\boldsymbol{w}_d = [w_{d,1}, \ldots, w_{d,N_d}]$, so that the entire corpus of documents is described by $\boldsymbol{W} = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_D]$. $z_{d,n}$ is the topic assignment of word $w_{d,n}$, where $\boldsymbol{z}_d$ and $\boldsymbol{Z}$ mirror $\boldsymbol{w}_d$ and $\boldsymbol{W}$ in their dimensionality. Similarly, $\bar{\boldsymbol{z}}_d$ denotes the estimated average topic assignments of the K topics across words in document d, such that $\bar{\boldsymbol{Z}} = [\bar{\boldsymbol{z}}_{\boldsymbol{1}}, \ldots, \bar{\boldsymbol{z}}_{\boldsymbol{D}}]^\intercal \in \mathbb{R}^{D\times K}$. $\boldsymbol{\beta} \in \mathbb{R}^{K\times V}$, describes the K topic distributions

over the V dimensional vocabulary. $\boldsymbol{\theta} \in \mathbb{R}^{D \times K}$ describes the K topic mixtures for each of the D documents. $\eta \in \mathbb{R}^V$ and $\alpha \in \mathbb{R}^K$ are the respective hyperparameters of the prior for $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$.

The generative process of our BTR model is thus:

1. $\boldsymbol{\omega} \sim \mathcal{N}(\boldsymbol{\omega}|\boldsymbol{m}_0, \boldsymbol{S}_0)$ and $\sigma^2 \sim$ Inv-Gamma$(\sigma^2|a_0, b_0)$
2. **for** $k = 1, \ldots, K$:
   (a) $\boldsymbol{\beta}_k \sim \text{Dir}(\eta)$
3. **for** $d = 1, \ldots, D$:
   (a) $\boldsymbol{\theta}_d \sim \text{Dir}(\alpha)$
   (b) **for** $n = 1, \ldots, N_d$:
      i. topic assignment $z_{d,n} \sim Mult(\boldsymbol{\theta}_d)$
      ii. term $w_{d,n} \sim Mult(\boldsymbol{\beta}_{z_{d,n}})$
4. $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{\omega}, \sigma^2\boldsymbol{I})$, where $\boldsymbol{A}$ is a design matrix based on $\boldsymbol{Z}$ and $\boldsymbol{X}$.

2.3.4 Observations without documents

A straightforward extension allows for some observations to be associated with an $\boldsymbol{X}$ and $\boldsymbol{y}$, but no document. This is often the case in a social science context, for example time-series may be associated with documents at irregular intervals. If an observation is not associated with any documents, the priors on the document topic distributions suggest that the topic assignment for topic K is set to $\alpha_k / \sum_k \alpha_k$. These observations may still be very useful in estimating the relationship between $\boldsymbol{X}$ and $\boldsymbol{y}$ so may be worth including in the estimation.

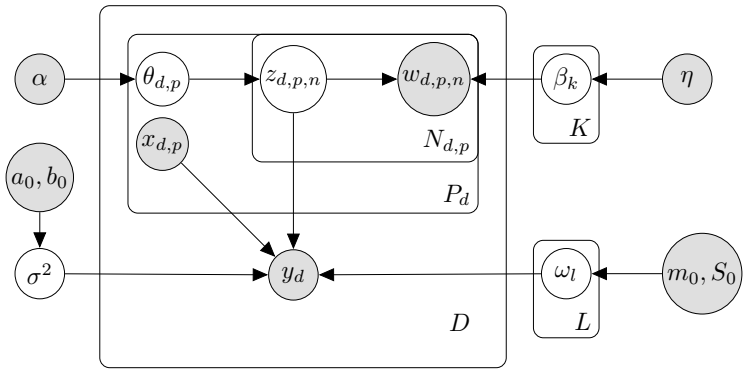
2.3.5 Multiple documents/paragraphs per observation

If, as is often the case in the context of social science applications, we have relatively few observations but the documents associated with those observations are relatively long, we can exploit the structure of the documents by estimating the model at a paragraph level. Splitting up longer documents into paragraphs brings one of the key advantages of topic modelling to the fore: that the same word can have different meanings in different contexts. For example, the word "increase" might have quite a different meaning if it is in a paragraph with the word "risk" than if it is alongside "productivity". Treating the entire document as a single entity makes it hard for the model to make this distinction.

If there are observations with multiple documents, we can treat these as P_d separate

paragraphs of a combined document, indexed by p, each with an independent $\boldsymbol{\theta}_p$ distribution over topics. These paragraphs may also have different associated $\boldsymbol{x}_{d,p}$ that interact with the topics, for example we may wish to interact topics with a paragraph specific sentiment score, but the response variable y_d is common to all paragraphs in the same document and the M-step estimated at the document level. Figure 2.3 shows the extended graphical model. More detail on how this changes the estimation procedure are given in Appendix B.2.

Figure 2.3: Graphical model for BTR with multiple documents per observation



2.4 Estimation

We estimate BTR using a Gibbs EM algorithm (Levine and Casella, 2001), in which the posterior distribution of $\boldsymbol{Z}$ is approximated by Gibbs sampling in the E-step, and the regression parameters are estimated in the M-step.

2.4.1 Posterior Inference

The objective is to identify the latent topic structure and regression parameters that are most probable to have generated the observed data. We obtain the joint distribution for our graphical model through the product of all nodes conditioned only on their parents, which for our model is

$$p(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{y}, \boldsymbol{\omega}, \sigma^2 | \boldsymbol{X}, \alpha, \eta, \boldsymbol{m}_0, \boldsymbol{S}_0, a_0, b_0) = \prod_{d=1}^{D} p(\boldsymbol{\theta}_d|\alpha) \prod_{k=1}^{K} p(\boldsymbol{\beta}_k|\eta) \times$$
$$\prod_{d=1}^{D} \prod_{n=1}^{N_d} p(z_{d,n}|\boldsymbol{\theta}_d) p(w_{d,n}|z_{d,n}, \boldsymbol{\beta}) \prod_{d=1}^{D} p(y_d|\boldsymbol{x}_d, \boldsymbol{z}_d, \boldsymbol{\omega}, \sigma^2) \prod_{l=1}^{L} p(\boldsymbol{\omega}_l|\boldsymbol{m}_0, \boldsymbol{S}_0) p(\sigma^2|a_0, b_0). \tag{2.8}$$

The inference task is thus to compute the posterior distribution of the latent variables ($\boldsymbol{Z}$, $\boldsymbol{\theta}$, $\boldsymbol{\beta}$, $\boldsymbol{\omega}$, and σ^2) given the observed data ($\boldsymbol{y}$, $\boldsymbol{X}$ and $\boldsymbol{W}$) and the priors governed by hyperparameters ($\alpha, \eta, \boldsymbol{m}_0, \boldsymbol{S}_0, a_0, b_0$). We will omit hyperparameters for sake of clarity

unless explicitly needed for computational steps. The posterior distribution is then

$$p(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{Z}, \boldsymbol{\omega}, \sigma^2 | \boldsymbol{W}, \boldsymbol{y}, \boldsymbol{X}) = \frac{p(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\omega}, \sigma^2)}{p(\boldsymbol{W}, \boldsymbol{X}, \boldsymbol{y})}. \tag{2.9}$$

In practice, computing the denominator in equation 2.9, i.e. the evidence, is intractable due to the sheer number of possible latent variable configurations.

We use a Gibbs EM algorithm (Levine and Casella, 2001) set out below, to approximate the posterior. Following Griffiths and Steyvers (2004), we can collapse out the latent variables $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ Griffiths and Steyvers (2004), so only need to identify the sampling distributions for topic assignments $\boldsymbol{Z}$ and regression parameters $\boldsymbol{\omega}$ and σ^2, conditional on their Markov blankets

$$p(\boldsymbol{Z}, \boldsymbol{\omega}, \sigma^2 | \boldsymbol{W}, \boldsymbol{X}, \boldsymbol{y}) = p(\boldsymbol{Z} | \boldsymbol{W}, \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\omega}, \sigma^2) p(\boldsymbol{\omega}, \sigma^2 | \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{y}). \tag{2.10}$$

Once topic assignments $\boldsymbol{Z}$ are estimated, it is straightforward to recover $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. The expected topic assignments are estimated by Gibbs sampling in the E-step, and the regression parameters are estimated in the M-step.

2.4.2 E-Step: Estimate Topic Parameters

In order to sample from the conditional posterior for each $z_{d,n}$ we need to identify the probability of a given word $w_{d,n}$ being assigned to a given topic k, conditional on the assignments of all other words (as well as the model's other latent variables and the observed data)

$$p(z_{d,n} = k | \boldsymbol{Z}_{-(d,n)}, \boldsymbol{W}, \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\omega}, \sigma^2), \tag{2.11}$$

where $\boldsymbol{Z}_{-(d,n)}$ are the topic assignments of all words apart from $w_{d,n}$. This section defines this distribution, with derivations in Appendix B.1. By conditional independence properties of the graphical model, we can split this joint posterior into

$$p(\boldsymbol{Z} | \boldsymbol{W}, \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\omega}, \sigma^2) \propto p(\boldsymbol{Z} | \boldsymbol{W}) p(\boldsymbol{y} | \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{\omega}, \sigma^2). \tag{2.12}$$

Topic assignments within one document are independent from topic assignments in all other documents and the sampling equation for $z_{d,n}$ only depends on it's own response variable y_d, hence

$$\begin{gathered} p(z_{d,n} = k | \boldsymbol{Z}_{-(d,n)}, \boldsymbol{W}, \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\omega}, \sigma^2) \propto \\ p(z_{d,n} = k | \boldsymbol{Z}_{-(d,n)}, \boldsymbol{W}) p(y_d | z_{d,n} = k, \boldsymbol{Z}_{-(d,n)}, \boldsymbol{x}_d, \boldsymbol{\omega}, \sigma^2). \end{gathered} \tag{2.13}$$

The first part of the RHS expression is the sampling distribution of a standard LDA model. We can express it in terms of count variables s (topic assignments across a document) and m (assignments of unique words across topics over all documents). For example, $s_{d,k}$ denotes the total number of words in document d assigned to topic k and $s_{d,k,-n}$ the number of words in document d assigned to topic k, except for word n. Analogously, $m_{k,v}$ measures the total number of times term v is assigned to topic k across all documents and $m_{k,v,-(d,n)}$ measures the same, but excludes word n in document d.

The second part is the predictive distribution for y_d. This is a Gaussian distribution depending on the linear combination $\boldsymbol{\omega}(\boldsymbol{a}_d|z_{d,n}=k)$, where $\boldsymbol{a}_d$ includes the topic proportions $\bar{\boldsymbol{z}}_d$ and $\boldsymbol{x}_d$ variables (and any interaction terms), conditional on $z_{d,n}=k$. We can write this in a convenient form that preserves proportionality with respect to $z_{d,n}$ and depends only on the data and the count variables.

First, we split the $\boldsymbol{X}$ features into those that are interacted, $\boldsymbol{X}_{1,d}$, and those that are not, $\boldsymbol{X}_{2,d}$ such that the generative model for y_d is then

$$y_d \sim \mathcal{N}(\boldsymbol{\omega}_z^\intercal \bar{\boldsymbol{z}}_d + \boldsymbol{\omega}_{zx}^\intercal(\boldsymbol{x}_{1,d} \otimes \bar{\boldsymbol{z}}_d) + \boldsymbol{\omega}_x^\intercal \boldsymbol{x}_{2,d}, \sigma^2), \tag{2.14}$$

where $\otimes$ is the Kronecker product. Define $\tilde{\boldsymbol{\omega}}_{z,d}$ as a length K vector such that

$$\tilde{\omega}_{z,d,k} = \omega_{z,k} + \boldsymbol{\omega}_{zx,k}^\intercal \boldsymbol{x}_{1,d}. \tag{2.15}$$

Noting that $\tilde{\boldsymbol{\omega}}_{z,d}^\intercal \bar{\boldsymbol{z}}_d = \frac{\tilde{\boldsymbol{\omega}}_{z,d}^\intercal}{N_d}(\boldsymbol{s}_{d,-n} + s_{d,n})$, gives us the sampling distribution for $z_{d,n}$ stated in Eq (2.13): a multinomial distribution parameterised by

$$p(z_{d,n}=k|w_{d,n}=v, \boldsymbol{Z}_{-(d,n)}, \boldsymbol{W}, \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\alpha}, \boldsymbol{\eta}, \boldsymbol{\omega}, \sigma^2) \propto (s_{d,k,-n}+\alpha)\frac{m_{k,v,-(d,n)}+\eta}{\sum_v m_{k,v,-(d,n)}+V\eta} \times \exp\left\{\frac{1}{2\sigma^2}\left(\frac{2\tilde{\omega}_{z,d,k}}{N_d}\left(y_d - \boldsymbol{\omega}_x^\intercal \boldsymbol{x}_{2,d} - \frac{\tilde{\boldsymbol{\omega}}_{z,d}^\intercal}{N_d}\boldsymbol{s}_{d,-n}\right) - \left(\frac{\tilde{\omega}_{z,d,k}}{N_d}\right)^2\right)\right\}. \tag{2.16}$$

This defines the probability for each k that $z_{d,n}$ is assigned to that topic k. These K probabilities define the multinomial distribution from which $z_{d,n}$ is drawn.

2.4.3 M-Step: Estimate Regression Parameters

To estimate the regression parameters, we hold the design matrix $\boldsymbol{A} = [\bar{\boldsymbol{Z}}, \boldsymbol{X}]$ fixed. Given the conjugacy of the Normal-Inverse-Gamma prior, this is a standard Bayesian linear regression problem and the posterior distribution of the regression parameters conditional on $\boldsymbol{A}$ has a known Normal-Inverse-Gamma distribution stated in Equation

2.6. To prevent overfitting to the training sample there is the option to randomly split the training set into separate sub-samples for the E- and M-steps, following a Cross-Validation EM approach (Shinozaki and Ostendorf, 2007). We use the prediction mean squared error from the M-step sample to assess convergence across EM iterations.

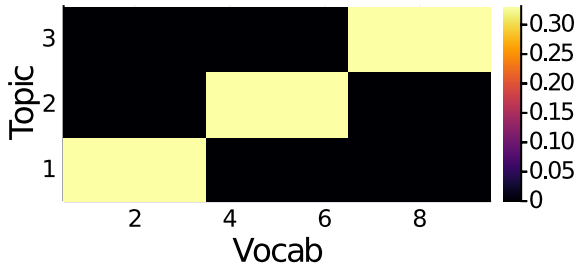
2.5 Experiment: Synthetic Data

To illustrate the benefits of our BTR approach, we generate a synthetic dataset of documents which have explanatory power over a response variable, along with an additional numerical covariate that is correlated with both documents and response. BTR is able to recover the true model parameters, which a range of benchmark models cannot.

2.5.1 Synthetic Data Generation

We generate $10,000$ documents of 50 words each, following an LDA generative process, with each document having a distribution over three topics, defined over a vocabulary of 9 unique terms. Figure 2.4 shows the topic-vocabulary distribution from which the synthetic documents are generated. We thus have underlying "true" topic assignment proportion for each document, represented by $\bar{\boldsymbol{Z}} = \{\bar{\boldsymbol{z}}_1, \bar{\boldsymbol{z}}_2, \bar{\boldsymbol{z}}_3\}$, where $\bar{z}_{1,d}$ is the proportion of words in document d assigned to topic 1.

Figure 2.4: Ground truth topic distribution for synthetic documents.



A numerical feature, $\boldsymbol{x} = [x_1, ..., x_D]^\intercal$, is generated by calculating the document-level frequency of the first word in the vocabulary. As the first topic places a greater weight on the first three terms in the vocabulary, $\boldsymbol{x}$ is positively correlated with $\bar{\boldsymbol{z}}_1$. The response variable $\boldsymbol{y} = [y_1, ..., y_D]$ is generated through a linear combination of the numerical feature $\boldsymbol{x}$ and the average topic assignments, with the proportion of words assigned to the first topic entering with a coefficient of -1 and the numerical covariate entering with a coefficient of 1,

$$\boldsymbol{y} = -\bar{\boldsymbol{z}}_1 + \boldsymbol{x} + \boldsymbol{\epsilon}, \tag{2.17}$$

where $\boldsymbol{\epsilon}$ is an i.i.d. Gaussian white noise term. The regression model to recover the ground truth is then

$$\boldsymbol{y} = \omega_1 \bar{\boldsymbol{z}}_1 + \omega_2 \bar{\boldsymbol{z}}_2 + \omega_3 \bar{\boldsymbol{z}}_3 + \omega_4 \boldsymbol{x}_d + \boldsymbol{\epsilon}. \tag{2.18}$$

The *true* regression weights are thus $\boldsymbol{\omega}^* = [-1, 0, 0, 1]$. In accordance with the FWL theorem, we cannot recover the true coefficients with a two-stage supervised feature extraction process.

2.5.2 Synthetic Data Results

We compare the ground truth of the synthetic data generating process against:

1. **BTR:** our Bayesian model, estimated via Gibbs sampling.
2. **rSCHOLAR**: the regression extension of SCHOLAR, estimated via neural VI.
3. **LR-sLDA:** first linearly regress $\boldsymbol{y}$ on $\boldsymbol{x}$, then use the residual of that regression as the response in an sLDA model, estimated via Gibbs sampling.
4. **sLDA-LR:** First sLDA, then linear regression.
5. **BPsLDA-LR:** replacing sLDA with BPsLDA, which is sLDA estimated via the backpropagation approach of Chen et al. (2015).
6. **LR-BPsLDA:** again replacing sLDA with BPsLDA.

Figure 2.5 shows the true and estimated regression weights for each of the six models on the synthetic dataset . LR-sLDA and sLDA-LR estimate inaccurate regression weights for both the text and numerical features, as do the BPsLDA variants. Similarly, rSCHOLAR fails to recover the ground truth. However, BTR estimates tight posterior distributions around to the true parameter values. The positive correlation between $\boldsymbol{z}_1$ and $\boldsymbol{x}$ makes a joint estimation approach crucial for recovering the true parameters.

Standard supervised topic models estimate the regression parameters for the numerical features separately from the topic proportions and their associated regression parameters, violating the FWL theorem. A key difference between rSCHOLAR and BTR lies in their posterior estimation techniques (neural VI vs Gibbs). rSCHOLAR's approach seems to have a similarly detrimental effect as the two-stage approaches. We suspect further research into (neural) VI assumptions and their effect on interpretability and accurate parameter estimation with text could be fruitful. In addition to estimating inaccurate regression parameters, both sLDA-based models also lead to inferior out-of-sample forecasting. In Section 2.7 we show demonstrate this outcome on two real-world datasets.

Figure 2.5: Comparing recovery of true regression weights across different topic models for synthetic dataset. For each panel, the true regression weights are shown as red points and the estimated 95% posterior credible (or bootstrap, depending on model) interval in blue. Only BTR contains the true weights within the estimated intervals.

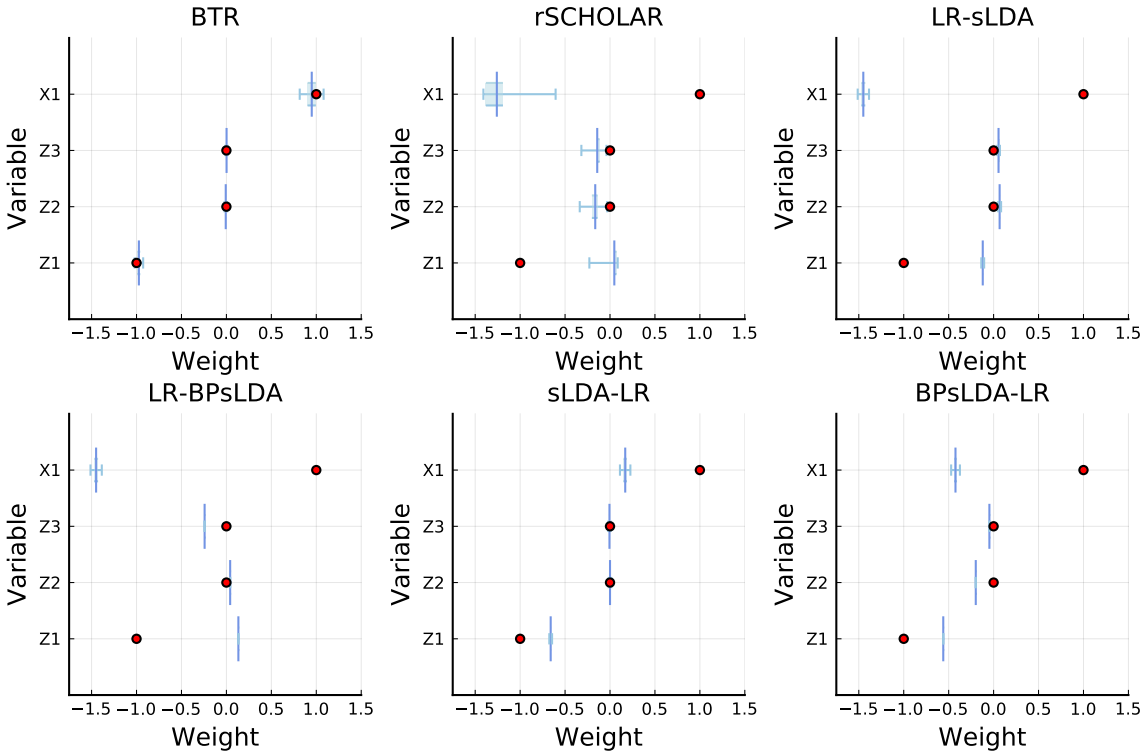


Table 2.2 shows the hyperparameter settings used in the synthetic data section. The results shown above are robust to a wide range of hyperparameters, which is unsurprising given the simplicity and strength of signal in the synthetic dataset.

Table 2.2: Synthetic example hyperparameters

	K	α	η	μ_{ntm}	σ_{ntm}	a_0	b_0	m_0	S_0
LR	-	-	-	-	-	0.2	4	0	2
LDA	3	1.0	1.0	-	-	-	-	-	-
sLDA	3	1.0	1.0	-	-	0.2	4	0	2
BPsLDA	3	1.0	1.0	-	-	-	-	-	-
BTR	3	1.0	1.0	-	-	0.2	4	0	2

2.6 Experiment: Semi-Synthetic Data

We further benchmark our model's abilities to recover the ground truth on two semi-synthetic datasets. In these datasets, we still have access to the ground truth as we either synthetically create or directly observe the correlations between explanatory and

outcome variables. However, the text and some numerical metadata that we use is empirical.

In each case, we estimate a "treatment effect": the effect of one of the $\boldsymbol{X}$ variables on $\boldsymbol{y}$, controlling for the text.

2.6.1 Semi-Synthetic Data Generation

As text data for our semi-synthetic examles, we use customer review datasets from Booking.com[5] and Yelp.[6] For both datasets, we randomly sample $50,000$ observations. Appendix B.4 provides summary statistics for the two datasets, as well as descriptions of the numerical variables. We analyse two different 'mock' research questions, one on each of these datasets.

The Booking.com dataset allows consumers to enter the positive and negative parts of their reviews in separate boxes. We combine these two reviews for all our exercises, but we do use information on the word count in each of these sections as potentially relevant numerical metadata.

Booking: effect of historic rating on review scores

This example can be thought of as asking the question: *Do customers give more critical ratings ($y_{booking,i}$) to hotels that have high historic ratings (hotel_av$_i$), once controlling for review texts?*

Our "treatment" variable here will be the historical average rating of that hotel, hotel_av$_i$, which is a continuous variable with range between 0 and 10, taken from the raw data. We then generate an outcome variable using this treatment variable and a confounder that will not be included in the estimation, but is correlated with the text and the treatment. In the Booking case, we use a "true" variable taken from the raw data, although for the Yelp example below we will generate this confounder ourselves so that we can vary the degree of correlation with the treatment. Our confounder here is the proportion of words in the combined review that belong to the positive part of the review (prop_pos$_i$) and has a correlation coefficient with the treatment variable of 0.22. This confounder can be thought of as representing the fact that generally higher rated hotels will get more positive reviews, and our mock research question wants to control for this.

[5] Available at kaggle.com/jiashenliu/

[6] Available at yelp.com/dataset, to reduce the size and heterogeneity of the Yelp dataset, we select only reviews for establishments based in Toronto, Canada.

The true data generating process (DGP) for the synthetic outcome variable in the Booking case is

$$\text{Booking DGP}: \quad y_{\text{booking},i} = -\text{hotel\_av}_i + 5\text{prop\_pos}_i + \epsilon_i \tag{2.19}$$

The true treatment effect here is thus -1 (i.e. if the historical average rating is increased by 1, the score for review i will be lower by 1, holding the text constant). This treatment effect is confounded by the text, with more positive reviews both increasing the score of this review and being positively correlated with the treatment.

We can illustrate the importance of the text as a confounder in this example by omitting the text from the estimation altogether and just estimating a regression of y_i on the treatment and a constant. This approach estimates a treatment effect of 0.092 (0.022), far removed from the true treatment effect of -1, as shown in the left panel of Figure 2.7. As we show and discuss in Section 2.6.2, our BTR model gets closest to the ground truth compared to a range of benchmark models.

Yelp: effect of reviewer nationality on review score

This example can be thought of as asking the question: *Do people from the US give different Yelp ratings ($y_{yelp,i}$) than customers from Canada, controlling for average business review ($stars\_av\_b_i$) and the review text?*

In this case, our treatment variable is binary and generated synthetically, along with the outcome. This allows us to control the degree of correlation between the confounding text and the treatment. We can therefore show that the greater this correlation between confounder and treatment, the greater the potential bias in the estimated treatment effect if the text is not accounted for properly.

We create the binary treatment variable US_i, we first compute each review's sentiment score ($sent_i$) using the Harvard Inquirer dictionaries.[7] We then generate the treatment variable to be correlated with this sentiment score, as shown in Equation 2.20.

$$\Pr(US_i = 1) = \frac{\exp(\gamma_1 sent_i)}{1 + \exp(\gamma_1 sent_i)}, \tag{2.20}$$

The γ_1 parameter thus controls the degree of correlation between text and treatment: when $\gamma_1 = 1$ the correlation between US_i and $sent_i$ is 0.39, for $\gamma_1 = 0.5$ it is 0.23.

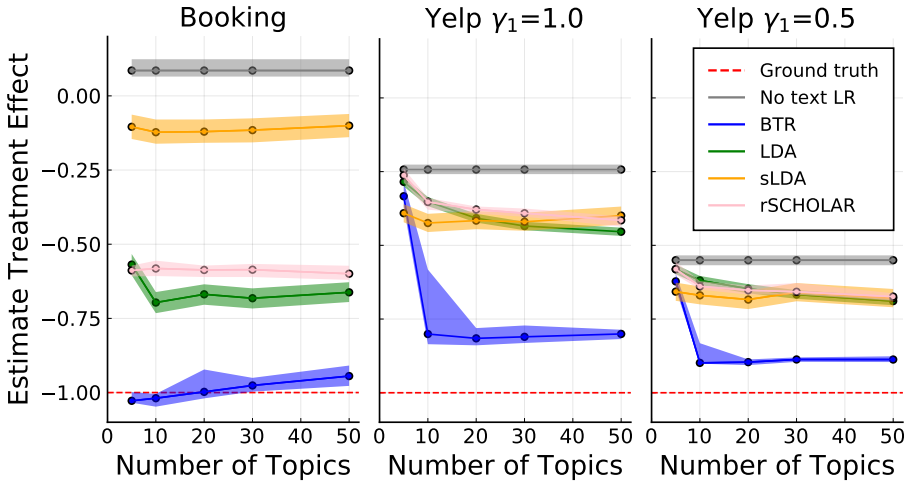
[7]This sentiment score is calculated as the difference between frequency of positive and negative words in the review, divided by the total number of words in the review.

The true DGP for the synthetic outcome variable in the Booking case is

$$\text{Yelp DGP}: \quad y_{\text{yelp},i} = -\text{US}_i + \text{stars\_av\_b}_i + sent_i. \tag{2.21}$$

As in the Booking case, the true treatment effect here is therefore -1. We can think of this synthetic treatment as the reviewer being American, so in this example Americans give lower review scores, but write more positive text in their review. We also include the historical average review for the business as an additional covariate to illustrate a further advantage of our model in being able to include such additional numerical covariates alongside the text.

Figure 2.7: Estimated TE semi-synthetic Booking (left panel), Yelp (middle and right panel). Intervals are either 95% credible interval of posterior distribution, or based on 20 run bootstrap, depending on model.



2.6.2 Semi-Synthetic Data Results

In Figure 2.7 we compare the treatment effects estimated by BTR to a range of benchmarks, across different numbers of topics. The benchmark models are the same as those used in the synthetic example discussed in 2.5 with the addition of an unsupervised LDA model. For the sLDA example we show results for the Gibbs sampled version, where the sLDA is performed first and then these residuals used in a linear regression (the other way around would just estimate the same treatment effect as the linear regression without text). For the LDA model, we first estimate topic proportions for each review and then include these in a linear regression with the numerical variables. This approach does not violate the FWL theorem, but has the disadvantage of the text features being extracted in an unsupervised way and therefore not optimised to the task at hand.

On both semi-synthetic datasets and across all benchmarked models, BTR estimates the regression weights that are the closest to the ground truth. This consistently holds true across all tested numbers of topics K (see Figure 2.7). For Yelp, we also vary the correlation strength between treatment and confounder. The middle panel in Figure 2.7 shows the estimation results with a very high correlation between confounder and treatment ($\gamma_1 = 1$). The RHS panel shows the results when this correlation is lower ($\gamma_1 = 0.5$). As expected, a higher correlation between confounder and treatment increases the bias. If the correlation between confounder and treatment is zero, a two-stage estimation approach no longer violates FWL and all models manage to estimate the ground truth (see Appendix B.3). Since the topic modelling approach is an approximation to capture the true effect of the text and its correlation with the metadata - and since this approximation is not perfect - some bias may remain. Indeed we see that in the Yelp case, none of the models are able to recover the true treatment effect, presumably because some aspects of the relevant text features are not well modelled by topics. Overall, BTR's joint estimation approach allows it to get substantially closer to the ground truth than any other model in all three cases.

2.7 Experiment: Real-World Data

The joint supervised estimation approach using text and non-text features, not only counteracts bias but also improves prediction performance. We use the real-world datasets of Booking and Yelp to illustrate this. For both datasets, we predict customer ratings (response) for a business or hotel given customer reviews (text features) and business and customer metadata (numerical features).

For Yelp, this metadata is *historic avg. rating by user*, *historic avg. rating of business*, and a *Harvard Inquirer* sentiment score. For Booking, the metadata is *historical avg hotel score*, *total neg word counts in customer review*, *total pos word counts in customer review*, *total num of reviews by customer*, *total num of reviews of hotel*. Descriptions for each variable are given in Appendix B.4. For both datasets, we randomly sample $50,000$ observations and select 75% in Yelp, 80% in Booking for training with the remaining observations used as a test set to assess out of sample prediction.

2.7.1 Benchmarks

We compare the out-of-sample performance of our BTR model to a wide range of models, includinadd the following models to the benchmark list from the previous section:

1. **OLS:** linear regression using only the non-text features.
2. **LR+aRNN:** a bidirectional RNN with attention. Bahdanau et al. (2015). Since the model does not allow for non-text features, we use the OLS residuals of the linear regression as the target.
3. **LR+TAM:** a bidirectional RNN using global topic vector to enhance its attention heads (Wang and Yang, 2020) - again using the OLS residuals as target.[8]
4. **LDA+LR**: unsupervised topic proportions estimated by Gibbs sampling (Griffiths and Steyvers, 2004) and then used in a linear regression together with the non-text features.
5. **GSM+LR:** an unsupervised neural topic model Miao et al. (2017), with proportions then used in a linear regression together with the non-text features.
6. **LR-sLDA:** as described in Section 2.5.
7. **LR-BPsLDA:** as described in Section 2.5.
8. **rSCHOLAR**: as described in Section 2.5.
9. **LR+rSCHOLAR**: the two-step equivalent for rSCHOLAR, estimating non-text regression weights in a separate step from the supervised topic model.

We also tested sLDA+LR and a pure sLDA, which performed consistently worse so they are not included for the sake of brevity. For example, for K=50, sLDA+LR achieved pR^2 of 0.420 and 0.564 for Booking and Yelp respectively, compared to 0.432 and 0.571 for LR+sLDA. Standalone sLDA achieves 0.356 and 0.526 respectively.

2.7.2 Prediction and Perplexity Results

We evaluated all topic models on a range from 10 to 100 topics, with results for 50 and 100 in Table 2.3 and results with fewer topics displayed in Appendix B.5 along with extensive robustness checks which show that results are robust across a wide range of hyperparameters. The BTR Dirichlet hyperparameters for displayed results are $\alpha = 0.5$ and $\eta = 0.01$. Hyperparameters of benchmark models that have no direct equivalent in our model were set as suggested in the pertaining papers.

[8]Wang and Yang (2020) use 100-dimensional word embeddings in their default setup for TAM and pre-train those on the dataset. We follow this approach. aRNN and TAM results were very robust to changes in the hidden layer size in these setups, we use a layer size of 64. Full details of all model parametrisations are provided in Appendix B.5.2.

Table 2.3: Mean pR^2 and perplexity, standard deviation in brackets. 20 runs per model. Best model **bold**.

Dataset	Booking		Yelp	
K	50	100	50	100
	pR^2 (higher is better)			
OLS	0.315		0.451	
LR+aRNN	0.479 (0.007)		0.582 (0.008)	
LR+TAM	0.479 (0.014)	0.487 (0.014)	0.585 (0.012)	0.587 (0.008)
LDA+LR	0.426 (0.003)	0.437 (0.002)	0.586 (0.006)	0.606 (0.007)
GSM+LR	0.386 (0.004)	0.395 (0.005)	0.495 (0.004)	0.517 (0.007)
LR+sLDA	0.432 (0.002)	0.438 (0.004)	0.571 (0.002)	0.574 (0.001)
LR+BPsLDA	0.419 (0.009)	0.455 (0.001)	0.603 (0.002)	0.609 (0.001)
LR+rSCHOLAR	0.469 (0.002)	0.465 (0.002)	0.550 (0.034)	0.557 (0.027)
rSCHOLAR	**0.494** (0.004)	**0.489** (0.003)	0.571 (0.01)	0.581 (0.009)
BTR	0.454 (0.003)	0.46 (0.002)	**0.630** (0.001)	**0.633** (0.001)
	Perplexity (lower is better)			
LR+TAM	521 (2)	522 (2)	1661 (7)	1655 (7)
LDA+LR	454 (1)	432 (1)	1306 (4)	1196 (2)
GSM+LR	**369** (8)	**348** (5)	1431 (34)	1387 (14)
LR+sLDA	436 (2)	411 (1)	1294 (5)	1174 (3)
LR+rSCHOLAR	441 (20)	458 (11)	1515 (34)	1516 (30)
rSCHOLAR	466 (19)	464 (9)	1491 (9)	1490 (9)
BTR	437 (1)	412 (1)	**1291 (5)**	**1165** (3)

We assess the models' predictive performance based on predictive R^2 ($pR^2 = 1 - \frac{\text{MSE}}{var(y)}$). The upper part of Table 2.3 shows that BTR achieves the best pR^2 in the Yelp dataset and and very competitive results in the Booking dataset, where our rSCHOLAR extension outperforms all other models. Even the non-linear neural network models aRNN and TAM cannot achieve better results. Importantly, rSCHOLAR and BTR perform substantially better than their counterparts that do not jointly estimate the influence of covariates (LR+rSCHOLAR and LR+sLDA).

To assess document modelling performance, we report the test set perplexity score for all models that allow this (Table 2.3, bottom panel).[9] The joint approach of both rSCHOLAR and BTR does not come at the cost of increased perplexity. If anything, the supervised learning approach using labels and covariates even improves document modelling performance when compared against its unsupervised counterpart (BTR vs

[9]Perplexity is defined as $\exp\left\{-\frac{\sum_{d=1}^{D}\log p(\boldsymbol{w}_d|\boldsymbol{\theta},\boldsymbol{\beta})}{\sum_{d=1}^{D}N_d}\right\}$.

LDA).

Assessing the interpretability of topic models is ultimately a subjective exercise. In Appendix B.5.4 we show topics associated with the most positive and negative regression weights, for each dataset. Overall, the identified topics and the sign of the associated weights seem interpretable and intuitive.

2.8 Conclusions

In this paper, we introduced BTR, a Bayesian topic regression framework that incorporates both numerical and text data for modelling a response variable, jointly estimating all model parameters. Motivated by the FWL theorem, this approach is designed to avoid potential bias in the regression weights, and can provide a sound regression framework for statistical and causal inference when one needs to control for both numerical and text based explanatory variables in observational data. We demonstrate that our model recovers the ground truth with lower bias than any other benchmark model on synthetic and semi-synthetic datasets. Experiments on real-world data show that a joint and supervised learning strategy also suggest that BTR yields superior prediction performance compared to 'two-stage' strategies, even competing with deep neural networks.

Chapter 3

Financial News Media and Volatility: is there more to Newspapers than News?

Abstract

It is an open question what role the media plays in financial markets, whether it is a causal one, and if so whether this effect is of aggregate importance. Using a text mining approach, this paper identifies a robust link between media coverage in the *Financial Times* newspaper from 1998 to 2017 and a firm's intra-day stock price volatility. By exploiting the timings associated with this effect, I argue that media coverage causes stock price volatility. I show that the effect is not driven by persistence in volatility or by the media anticipating future newsworthy events. Using a topic modelling framework, I also show that it is not driven by the content of the coverage. Finally, the identified effects are used to investigate whether this volatility propagates across the stock market, showing that while volatility may spill over into firms related by the structure of the production network, the volatility due to media coverage does not. The paper concludes that while this effect of media coverage is potentially important at a firm-level, it has limited aggregate implications.

3.1 Introduction

In a textbook model of financial markets, public information is quickly and efficiently incorporated into asset prices, leaving little role for the media. However, in a world where investors have limited attention and information diffuses slowly, the media may play a crucial role. Although many studies establish Granger causality between media content and market activity, few studies distinguish market reactions to media reporting per se from reactions to the underlying new information (Tetlock, 2015). There is some evidence of a causal effect at a granular level, but confined to specific contexts such as fund holdings Solomon et al. (2014) or earnings announcements Engelberg and Parsons (2011). Furthermore, whether this granular media coverage effect is of aggregate importance is unclear. This paper shows that print media coverage of a firm increases the volatility of its stock price beyond the effect of any new information. This affects other firms in the same sector and may propagate through the production network. I do not argue for a precise causal mechanism, but I find that the evidence is consistent with an explanation based on a salience view of the media's role in financial markets. The salience view argues that media coverage affects financial markets not by providing information, but by bringing companies more to investors' attention (Fan et al., 2020).

For the sake of clarity, it will be important to define the difference between media coverage and new information. When relevant new information becomes available, i.e. when there is what Tetlock (2015) calls an "information event", it certainly affects financial markets. However, holding the new information constant, media coverage may or may not have a causal effect. For example, a large firm's earnings announcement is new information, which we would expect markets to price in fairly quickly. Whether the media covers could influence how this new information is priced in. I will avoid using the term "news" as this may cause unnecessary confusion, referring as it does to both new information and the media.

I construct a novel dataset by combining articles from the *Financial Times* newspaper (FT) from May 1998 to December 2017 with data on stock prices traded on the London Stock Exchange over the same period. I use print media as my sample of media coverage because the fixed publication time makes it easier to separate out a media coverage effect from the effect of new information. Furthermore, it is unlikely that investors use the FT as their principal information source during trading hours rather relying on more high frequency newswires, but it likely that many of them are exposed to its coverage throughout the day. This makes it an ideal sample for which to test the potential causal

effects of media coverage as distinct from the effects of new information.

In this paper, articles are matched to the names of firms using a joint approach: string-matching the names to the headlines of articles and applying a Named Entity Recognition algorithm to the text of the article. Matching to the headlines ensures that coverage which is principally about that firm is identified. The Named Entity Recognition algorithm tags words which are used to refer to a named entity. This ensures that the mention of a firm's name indeed refers to the firm, and that the firm is the principal focus of the article.

The primary result of this paper is that an article in the *Financial Times* (FT) increases the intra-day volatility, as measured by the percentage difference between the highest and lowest price traded at that day, by around 10 basis points. The sample covers all articles in the FT from May 1998 to December 2017, and all firms who were a constituent of the FTSE 100 Index at some stage over this period. This is identified as a causal effect by exploiting the timing of publication relative to trading hours and then controlling for potential confounding factors. As the newspaper is published before the market opens, the predictive effect of the media coverage on volatility is not due to reverse causality, i.e. the media reporting on volatility.

I examine two alternative explanations for this effect which do not assign a causal role to media coverage: persistence and new information. Firstly, I rule out that the effect is driven by reporting on persistent past volatility, showing that media coverage Granger causes volatility and that forward looking coverage has a relatively greater effect. Secondly, I show that the increase in volatility is not driven by a reaction to new information, either available prior to publication in the media coverage or relating to events that were anticipated and so reported on pre-emptively. I control for this using the absolute intra-day return as the broadest possible measure of realised new information; using an implied volatility measure derived from options prices as a measure of expected new information; and showing that the sentiment of articles does not predict subsequent returns.

Having shown that there is a robust, substantial and plausibly causal effect of media coverage on a firm's stock price volatility, I also show that this effect is not explained by the *content* of that coverage. In order to account for the content in a systematic and theoretically consistent way, I use the Bayesian Topic Regression framework presented in 2.

Finally, I investigate whether the effect of media coverage is of aggregate important by

showing that the volatility predicted by media coverage spreads to other firms linked by the production network. I use input-output tables for the UK from the Office of National Statistics to compute a priori links between sectors. I then show that a firm's stock price volatility is affected by media coverage of firms in sectors that are downstream in the production network (i.e. firms that are their potential customers). This volatility spillover survives controlling for absolute return implied volatility and realised volatility with the firm's own sector, so can perhaps be interpreted as causal. An aggregate measure of coverage for all firms in the FTSE index does not appear to have an impact on index-level volatility, however. This suggests that the media coverage effect may also affect firms that are fundamentally related in investor's minds, but not the market as a whole. This interpretation would give further support to the salience view over the information view.

The remainder of the paper is organised as follows. Section 3.2 reviews the literature concerning the causal role of media in financial markets, and provides some wider context for this paper. Section 3.3 describes the dataset and explains how media coverage is measured. Section 3.4 then presents the primary result, that media coverage increase a firm's stock price volatility, and explains the identification strategy. Section 3.5 explains how I control for the content of articles, and shows that the media coverage effect is not explained by this content. Section 3.6 investigates whether this effect has aggregate implications, and in particular shows potential propagation through the production network. Finally, Section 3.7 concludes.

3.2 Literature Review

This Section provides a brief review of the literature surrounding the role of the media and information in financial markets. This paper contributes to the limited literature assessing the causal impact of media coverage on financial markets, using a novel dataset and identification strategy to show that media coverage of a firm increases the volatility of its stock price. Furthermore, it explore whether coverage of individual firms can have aggregate implications.

This paper uses the Bayesian Topic Regression (BTR) described in Chapter 2, so this method is put in context of the wider literature there. Many studies argue that the sentiment of news is the most important element of its information content (Tetlock, 2007), but there is more limited range of work looking at other dimensions of news. Prior work has also largely focused on professional investing platforms where news is explicitly linked to stock tickers. By focusing on a print edition of a newspaper, this paper argues

that media coverage which is not necessarily directly linked to new information can have potentially important economic effects.

3.2.1 Causal effect of media coverage

Although there is widespread speculation that the news media may have the power to influence financial markets beyond from simply reporting events, showing such a causal relation is non-trivial. Media coverage is the product of profit maximization by media companies and so will cover stories their readers are interested in, as shown by Gentzkow and Shapiro (2010). As a number of the unobservable factors which influence these coverage decisions also affect investor behaviour, an identification problem arises. For a given "news" event, it is difficult to separate whether the media's coverage changed the market's response, or whether some unobserved aspect of the event simultaneously drove both media coverage and the market reaction.

Previous work has generally followed one of two approaches to identify this causal effect: either focusing on case studies of specific events, or relying on some exogenous variation in the level of media coverage. Perhaps the best known example of the former approach, documenting specific instances in which media coverage appears to have had an effect on markets, is Huberman and Regev (2001). The authors detail how how a feature story in the New York Times (NYT) caused the stock of Entremed Inc. to increase fourfold overnight. Yet, as the authors carefully show, virtually all of the facts reported in the NYT story had been previously published in scientific journals. Similarly, in 2008 several websites mistakenly posted a 6-year-old story about United Airlines' bankruptcy, which lead to United's stock price falling by 76% within minutes (Carvalho et al., 2011; Marshall et al., 2014). While these individual cases may be convincing, this approach is not amenable to a more systematic analysis.

The main alternative to this case-study approach is to identify exogenous variation in media exposure. For example, Dougal et al. (2012) show that the exogenous scheduling of *Wall Street Journal* columnists can predict excess returns of the Dow Jones Industrial Average. Peress (2014) uses newspaper strikes to identify exogenous variation in media coverage and shows that lower media coverage leads to lower trading volume and intraday volatility. Larsen and Thorsrud (2017) also use newspaper strikes as an exogeneous source of variation, and argue that between 20 to 40 percent of media coverage's predictive power on returns is due to a causal media effect. Similarly, Engelberg and Parsons (2011) use regional differences in media coverage to show that local media coverage pre-

dicts local trading volume in response to earnings announcements of S&P 500 firms.[1]

This paper is different in that it uses a temporal separation between media coverage and market reactions, and examines the effect on intraday volatility rather than returns. Therefore, in controlling for realised returns, we can control for new information in the broadest and most stringent way possible. Furthermore, as the majority of literature focuses on US financial markets, this paper provides useful additional evidence of a link between news media and stock markets from the UK. UK participation among households in stock markets is substantially lower than in the US Guiso et al. (2003), so it is not immediately obvious that results across the two countries should be similar.

There is no consensus as to the effect that media coverage has on the efficiency of financial markets, or whether it provides new information. Solomon et al. (2014) argue that media coverage can exacerbate investor biases, finding that mutual fund holdings with high past returns attract extra flows, but only if these stocks were recently featured in the media. Conversely, Bushee et al. (2010) find that press coverage reduces asymmetry around earnings announcements (i.e., lower spreads and greater depth) suggesting that the press helps reduce information problems. Drake et al. (2014) find further evidence that greater press coverage mitigates mispricing after annual earnings announcements, arguing that this impact is driven primarily by the press disseminating the information more broadly, rather than by the creation of new content that helps investors understand the implications of accounting information. Kerl and Walter (2007) show that German Personal Finance magazine recommendations can help investors earn excess positive returns.

3.2.2 New information and financial markets

There is an extensive empirical literature analysing how new information affects financial markets. This literature will often use media coverage data, but in an attempt to identify new information, rather than to identify a causal effect of the media coverage itself. Historically, attempts to link news about future cash-flows and aggregate stock market movements have not been able to explain more than around one third of index-level volatility (Roll, 1988; Cutler et al., 1989).[2] A significant effect of new information

[1]There is also a related literature which uses a similar approach to quantify the media's impact on political attitudes. For instance, DellaVigna and Kaplan (2007) show that voting patterns can be predicted by whether an area's local cable company carries Fox News and Gerber et al. (2009) conduct a field experiment which shows that exposure to different media may have a causal effect on political attitudes.

[2]That public information can only explain a portion of financial market volatility is also the case in much older data. Koudijs (2016) uses the arrival sailboats in 18th century Amsterdam to identify

on volatility at an aggregate level is often found, but its explanatory power is limited. For example, Jiang et al. (2012) examine the spillover of volatility across international stock markets and find that unscheduled news releases lead to an increase in implied volatility, but that this does not explain the spillovers.

While new information does not appear to be able to explain more than a fraction of aggregate volatility. The link between firm-specific news and equity prices is however fairly robust and well-documented Hendershott et al. (2015); Alanyali et al. (2013); Jiao et al. (2020). Sprenger et al. (2014) use Twitter to identify company specific news events and show that different news events lead to different stock market reactions. Chen et al. (2014) show that activity on investor social network platforms can predict future stock returns and earnings surprises. Similarly, Groß-Klußmann and Hautsch (2011) use the Reuters NewsScope Sentiment Engine, commercially available data which provides sentiment, novelty and relevance indicators for Reuters newswire stories. They find clear responses in returns volatility, trading volume and bid-ask spreads after news announcements. They find that the classification of news according to indicated relevance is required to identify a significant effect, which is not the case for the *Financial Times* data which is used here. As the FT produces only a small number of stories each day, it appears that a higher proportion of their coverage is relevant to financial markets.

This paper finds evidence of a media coverage effect increasing volatility at the firm level, and potentially having some aggregate implications by spilling over to related firms. On this latter point, the paper contributes to the empirical literature on the importance of the input-output network to the macroeconomy (Acemoglu et al., 2016; Ozdagli and Weber, 2017; Barrot and Sauvagnat, 2016) and financial markets in particular (Cohen and Frazzini, 2008). This suggests that a media coverage effect, if focussed on a systemically important firm, may play a role in explaining the excess aggregate level volatility which the literature on new information and financial markets has been unable to explain.

Solomon et al. (2014) state that there are broadly two theories of how media coverage influences investors: the information view (lowers cost of acquisition and reduces asymmetry), and the salience view (brings a company more investors attention). By focusing on the volatility effect of media coverage and abstracting from the effect of new information, this paper provides evidence for the salience view. The role of the media in financial markets has been linked to attention since Merton (1987) presented a model

the arrival of public information. Only around 40-50% of volatility can be attributed to this public information.

of capital markets with incomplete information, in which media reporting could direct investor attention to certain assets and thus raise their price (and lower their returns).

There is a small theoretical literature linking explicitly linking investor attention to increased volatility. This effect is often explained through reference to investor attention. For instance, Andrei and Hasler (2015) show that in a pure exchange economy with time-varying attention and a representative investor with recursive preferences, stock-return volatility and risk premia will increase with attention. Peng and Xiong (2006) consider a model with multiple risky assets whose payoffs are determined by market, sector, and firm-specific factors. They show that investors with limited attention prioritize market news over firm-specific news.

There is also a growing empirical literature attempting to use measures that proxy for investor attention to show a link to volatility. Schmidt (2013) uses Google searches for international sports events to shows that an increase in investor attention to sports reduces market volatility and trading volume. Dimpfl and Jank (2016) find evidence that realised volatility of the Dow Jones can be predicted by the previous day's Google searches, which they show is consistent with an attention hypothesis.[3] Goddard et al. (2015) find that investor attention, also by Google search queries, co-moves with and predicts future FX market volatility even after controlling for macroeconomic fundamentals. Peress (2008) compares earnings announcements which are or are not covered in the Wall Street Journal, finding that announcements with media coverage generate stronger price and trading volume reactions.

Similarly, social media can be used as a proxy measure of investor attention. At an aggregate level, Zhang et al. (2011) show that there is a significant relationship between tweet sentiment and US stock market indices. At an individual firm level, Fan et al. (2020) show that mentions of a firm on Twitter appear to affect the volatility and trading volume of individual stocks, and show that messages generate by "bots" have different effects to those generated by humans.[4]

[3]High volatility encourages high retail investor attention, which is in turn followed by high (excess) volatility.

[4]Gorodnichenko et al. (2018) show evidence that bots increased the polarization of public opinion during the 2016 US Presidential Election and 2016 Brexit Referendum.

3.3 Data

I construct a novel dataset by combining articles from the *Financial Times* print newspaper with financial data on firms which are listed on the London Stock Exchange. As most firms do not have an article devoted to them in most editions of the newspaper, these articles provide a natural and intuitive measure of media coverage. I also introduce the UK input-output tables which I use to measure the UK production network in Section 3.6.

The stock price data, which is downloaded from Refinitiv's Datastream, is daily data covering every firm which appeared in the FTSE 100 Index from May 1998 to December 2017. This restricts focus to large firms who receive a considerable amount of media coverage, but avoids any survivorship bias by including firms both before and after they are a constituent of the Index. The data includes opening and closing prices for each trading day, as well as the highest and lowest price traded at one the exchange, trading volume, total turnover and sector codes for the firm. For around half the sample, options data for the individual stocks is also available, which allows the construction of an implied volatility variable.

I focus primarily on three financial variables which allow the separation of a media coverage effect on volatility from the effect of new information: intra-day volatility, absolute return and implied volatility. Intra-day volatility, $vol_{i,t}$, is measured as the percentage difference between the highest and lowest price traded at that day.

$$vol_{i,t} = 100 \times \frac{p_{i,t}^{high} - p_{i,t}^{low}}{p_{i,t}^{low}} \tag{3.1}$$

Absolute return, $|\Delta p|_{i,t}^{o,c}$, is measured as the absolute percentage difference between the opening and closing price that day, which provides a very broad measure of the markets' assessment of new information that day.

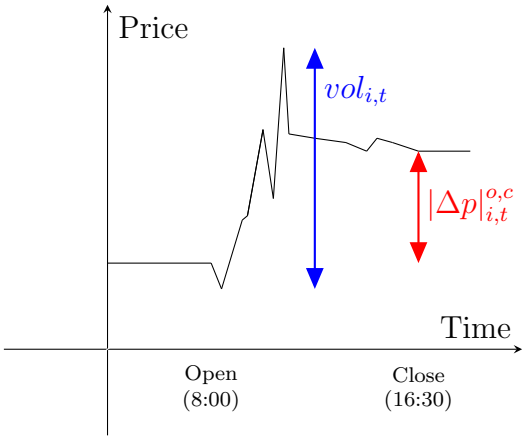
$$|\Delta p|_{i,t}^{o,c} = 100 \times \frac{|p_{i,t}^{close} - p_{i,t}^{open}|}{p_{i,t}^{open}} \tag{3.2}$$

A measure of anticipated news is given by the implied volatility calculated from the prices of options on the underlying stock, $VI_{i,t}$. This implied volatility measure is provided directly by Datastream where a sufficiently rich options market is available and can be interpreted as a measure of market expectations of volatility of the underlying asset price.

Figure 3.1 illustrates the difference between this intra-day volatility measure and the

absolute return. This shows how although they are related the volatility measure gives a measure of how extreme price movements have been on a given day, distinct from any new information which is priced in over the course of trading hours.

Figure 3.1: Intra-day volatility is measured as the difference between the highest and lowest price that day



All articles from the *Financial Times* (FT) print newspaper between May 1998 and December 2017 are used as the news media sample, articles from the website which is updated throughout the day are excluded.[5] Various meta-data for the articles is also extracted, importantly the date on which the article is published as well as the headline. This comprises of around 1.6 million articles, many of which are unlikely to be relevant in this context, so the sample is restricted to articles which explicitly mention the firm in the following way.

The name for each firm is taken from the Thomson Reuters Eikon database, which also includes previous names. In addition, alternative names are created by dropping common abbreviations such as "PLC" or "Limited" as well as replacing non-standard characters (e.g. replacing "&" with "and"). A full list of the alternative names for each firm is given in Appendix C.1. This list should provide coverage of most references to the firms in the sample, but it will likely miss a substantial proportion because they are referred to by abbreviations or nicknames. However, in keeping the list of possible names fairly short, I ensure that the rate of false positive matches will be low, compared to the rate of false negatives, thereby biasing my estimates of the media coverage effect downwards rather than upwards.

The FT articles are then matched to the firm stock price data using these names through

[5]These are downloaded from the Nexis service, as shown in Appendix C.1.

a combination of two methods. String-matching the firm names to the article headlines, and matching the firm names to a list of named entities identified in the main text of the article using the Apache OpenNLP *Named Entity Recognition* algorithm (NER).[6] The NER algorithm locates and classifies mentions of "named entities" within a body of text and classifies them as referring to a person, location or organisation. Although there are some entities that are not picked up by the algorithm, so there are likely to be false negatives, it very rarely leads to false positives (i.e. match a firm to an irrelevant article). The use of the NER algorithm is crucial as purely string-matching would lead to articles being falsely matched to firms because their name has a different meaning in different context. The retailer *Next* is a good example here, as the word "Next" features in many headlines that have no relation whatsoever to the retailer.

Figure 3.2 shows an example of this matching, demonstrating why a joint approach is preferable, the full articles are shown in Appendix C.1. Panel (a) shows an article which is matched to the retailer *Next* in both the headline and through the NER algorithm. As is evident, the article's principal focus is indeed the retailer. Panel (b) shows an example where *Next* is identified as a named entry, but is not matched to the headline. Although this article does mention the firm, it is not the article's principal focus. Matching purely based on the NER algorithm would lead to the tagging of firms to articles in which they are only mentioned in passing. This occurs especially often for firms in Financial Services, representatives of which are often asked to comment on developing stories. Matching purely based on the headline would encounter a different problem because, as mentioned above, many articles would be falsely matched to firms. As the structure of headlines is not that of standard prose, running the NER algorithm on the headlines does not prove effective. An alternative approach would be to follow Larsen and Thorsrud (2017) in matching firms to articles based on the topics of those articles and firm descriptions. This would match firms with articles which are likely to be relevant, but does not highlight *firm-specific* coverage, and so is not as well suited as the headline-NER approach used here.

[6] http://incubator.apache.org/opennlp/

Figure 3.2: An example of the firm-to-article matching

(a) Article matched by both headline and NER

2003-01-09: Next confident of hitting its full-year target - Retailers company calms fears of poor pre-Christmas trading by announcing a rise in underlying sales

Next yesterday dispersed the worst fears about its performance with news of a small increase in underlying sales and promised it would achieve profit forecasts. The group, which had been the focus of continued rumours of poor trading in the run up to Christmas, admitted that growth was slowing and that it was still suffering because of some mistakes in its fashion ranges.
...
The group said that taken together, sales for the Next brand as a whole were ahead by 14.3 per cent. Sales at full price in the period up to Christmas were up by 9.5 per cent - but down by 0.8 per cent like-for-like. Full-price sales for stores and the directory were ahead by 11.8 per cent.

(b) Article matched with only NER on main text

2001-01-24: Buoyant retail sales data defy bleak forecasts - Economy analysts urge caution over 'inconclusive' December figures

Retail sales growth surged in December to its fastest rate in eight months, defying economists' forecasts and apparently dispelling talk of a slowdown in consumer demand. While admitting that recent pessimism appeared to have been overdone, economists said the figures - which included a week of post-Christmas sales - were not conclusive.
...
Surveys conducted in the first half of the month suggested sales were weak, verging on stagnant, and big retailers such as Dixons and Next said Christmas trade was disappointing.
...
The stronger than expected retail figures made it likely that today's growth data would also outperform economists' expectations of a 0.4 per cent expansion in the fourth quarter.

By matching firms to articles in this way, I obtain a measure of whether a given firm is the principal focus of an article in the FT print newspaper. For a total of 860,911 firm-day observations, 73,671 are matched to an article through the NER method, while 38,252 are matched through the headline. Of these 17,027 are matched using both methods. Using the overlap between the two measures, ensures that the possibility of false positives is very low, although as mentioned above there are likely many articles which are not matched. This will only serve to attenuate any effect.

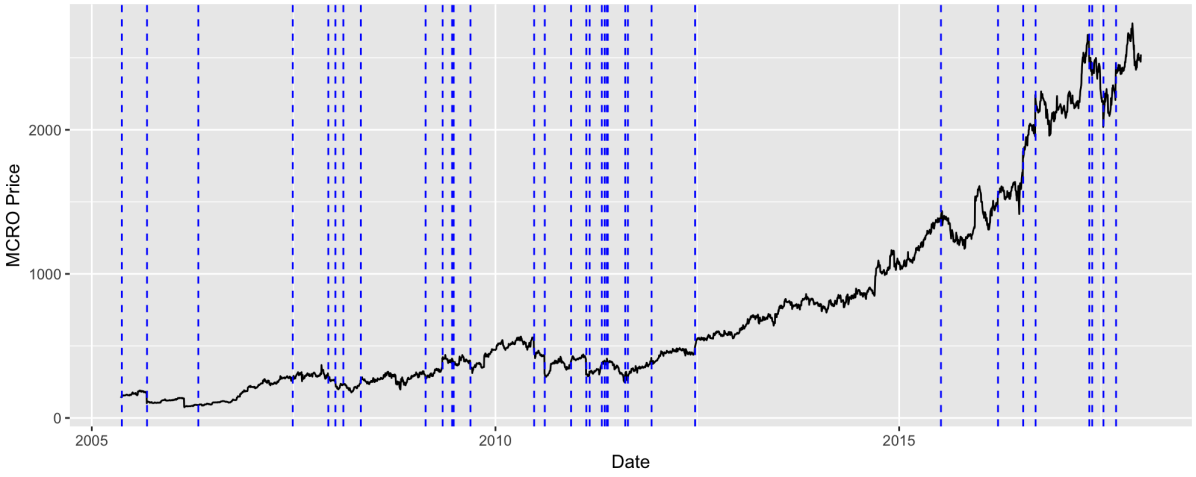
Table 3.1: Number of matches with alternative matching methods

	NER	
headline	0	1
0	765,401	56,644
1	21,225	17,027

Section 3.5 will analyse the content of these articles in more detail, but a natural measure of media coverage which will clearly demonstrate the lower bound of any effect is given by a simple dummy variable capturing whether there is a matched article for a given firm on a given day. Figure 3.4 shows this for the software provider *Micro Focus*. As the next

Section will set out, this "mention" variable can be used to identify a subsequent increase in the volatility of that firm's stock price.

Figure 3.4: Micro focus stock price, with vertical lines indicating a "mention" in the FT



Section 3.6 of the paper examines potential spillovers of the media coverage effect to other firms linked by the production network, showing that the aggregate implications of the media coverage effect are limited. In order to do this, I construct a measure weighting the importance of each sector as either an iate consumer or an intermediate supplier to every other sector. The Office of National Statistics provide yearly input-output (I-O) tables for sectors defined by NACE codes, which can be matched to the sector codes provided with the firm-level financial data from Datastream. I average the I-O matrices over the last 20 years, giving an estimate of the links between sectors provided by the UK production network.

3.4 A media coverage effect

An article in the FT increases the intra-day volatility of a firm's stock price by around 10 basis points. I exploit the timing structure of trading hours and newspaper publication and control for potential confounding factors to show that this is a causal effect. I show that the effect is not due to reverse causality, persistence or new information independent of media coverage.

I corroborate that this is a causal effect in two steps. First, I explain how my identification strategy, through the timing of newspaper publication relative to intra-day volatility, shows that it is not driven by reverse causality, i.e. by the media reporting

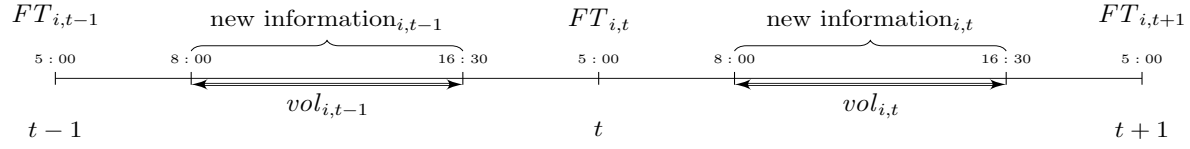
on volatility. Second, I account for two alternative explanations for the effect: persistence and new information. I rule out that the effect is driven by reporting on the past by showing that media coverage Granger causes volatility (in a very robust sense) and that forward looking coverage has a relatively greater effect. I show that the increase in volatility is not driven by a reaction to new information, by controlling for anticipated and unanticipated news shocks using absolute intra-day returns and an implied volatility measure derived from option prices.

3.4.1 Identification strategy: timings are key

Media coverage's positive predictive effect on volatility, as documented here, is not due to reverse causality. The FT print edition is published at 5:00 on a given day (Alanyali et al., 2013), and the London Stock Exchange is open from 8:00 until 16:30.[7] As the newspaper is published hours before the London Stock Exchange opens, it cannot be the case that media coverage in the morning newspaper is directly caused by price movements that day. These timings ensure that any media coverage effect on volatility is not due to the media simply reporting on that day's volatility.

The timings of FT's publication and trading hours are shown in Figure 3.5, showing the clear separation between publication and opening hours. The timeline also implies that volatility will be to some extent driven by new information which arrives that day, which is discussed at some length in subsection 3.4.2 below.

Figure 3.5: Publication of FT is temporally separated from trading hours



Being the focus of an article in the morning edition of the newspaper increases a firm's intra-day volatility by around 10 basis points. I show this in a dynamic panel regression of the following form

$$vol_{i,t} = \alpha_i + \mu_t + \beta mention_{i,t} + \gamma_1 vol_{i,t-1} + \epsilon_{i,t} \tag{3.3}$$

The $mention_{i,t}$ variable is simply the binary indicator which indicates whether firm i is featured in the FT newspaper at time t, as demonstrated for the firm Micro Focus in Fig-

[7] https://en.wikipedia.org/wiki/London_Stock_Exchange#Opening_times

ure 3.4. The firm fixed effects α_i removes any effect due to intrinsic differences between the volatilities of different firms' stock prices which may be correlated to the amount of media coverage that firm receives. For example, it could be that in general larger firms have more volatile prices and receive more media coverage for unrelated reasons. The fact that the distribution of mentions across firms is quite heterogeneous suggests that this is a factor that ought to be accounted for. Similarly the μ_t term is a time fixed effect which removes any market-wide effect. This rules out that the effect is just driven by there being more articles about listed firms in times when the overall stock market is particularly volatile. As the auto-regressive conditional heteroskedasticity of financial variables is well documented, I also control for a lag of the dependent variable. This does substantially reduce the magnitude of the effect, but it remains quantitatively and statistically significant. In subsection 3.4.2 I will look in more detail at the importance of persistence for this effect.

Table 3.2 shows the results of this simple dynamic panel regression, showing the statistically and quantitatively significant effect of mentions in the FT print newspaper on a firm's intra day stock price volatility. Column (1) shows the results of a simple pooled regression over the entire sample on a constant and the mention dummy. Column (2) then includes firm fixed effects to account for firm-level heterogeneity. Column (3) adds in time fixed effects, which account for any market level factors which might drive volatility. Finally, Column (4) includes the lagged dependent variable, taking into account persistence in the volatility measure. Appendix C.2.1 presents the results under of a range of robustness tests, including alternative matching methodologies and subsamples of the data.

Table 3.2: An article in the *Financial Times* leads to an increase in a firm's stock price volatility

	Dependent variable:			
	$vol_{i,t}$ (in percentage points)			
	(1)	(2)	(3)	(4)
$mention_{i,t}$	0.501***	0.597***	0.517***	0.121***
	(0.020)	(0.020)	(0.017)	(0.016)
$vol_{i,t-1}$				0.390***
				(0.001)
Constant	3.086***			
	(0.003)			
Firm fixed effects		✓	✓	✓
Time fixed effects			✓	✓
Observations	862,974	862,974	862,974	849,061
R^2	0.001	0.123	0.330	0.433
Adjusted R^2	0.001	0.123	0.326	0.430
Residual Std. Error	2.616	2.451	2.149	1.973
Note:				*p<0.1; **p<0.05; ***p<0.01

As shown by the constant term in column (1) of Table 3.2, the mean value of the volatility measure on days without a mention is around 3%, so the increase of around 10 basis points shown in column (4) is quantitatively as well as statistically significant. This estimate is likely to be a lower bound due to any attenuation bias caused by imperfect matching and the crudeness of the media coverage measure. Section 3.5 will take the content of articles into account to create a richer measure of media coverage.

3.4.2 Reporting on the past and anticipating the future

I consider two alternative explanations for the empirical connection between media coverage and volatility that do not imply a causal effect, and reject both of these. These alternative explanations are based on (i) the persistence of volatility and (ii) new information. Neither appear to be driving the observed link between media coverage and

stock price volatility.

Persistence

Volatility in financial markets is persistent. If this persistence is not adequately accounted for, then the observed correlation between the morning newspaper and intra-day volatility may simply be due to the newspaper reporting on past volatility. In order to show that the persistence of volatility is not what drives the media coverage effect, I demonstrate Granger causality, show that forward-looking articles have a greater effect and condition on future media coverage.

I take three approaches to show that the media coverage effect is not driven by reporting on past price movements. Firstly, I control for various specifications of past price movements and show that these do not remove the effect. Secondly, I isolate articles which are more forward looking and show that these have a greater effect. Thirdly, I show that the effect of today's media coverage survives controlling for tomorrow's media coverage, thus accounting for persistence in media coverage as well as in volatility.

Controlling directly for past price movements is perhaps the most straightforward way to show that the media coverage effect is not driven by the media reporting on the past. We have already seen, in Table 3.2 above, that adding a single lag of the dependent variable reduces the magnitude of the estimated media coverage effect considerably. This suggests that the media does indeed report on the past and that part of the predictive effect of media coverage is explained by this. However, as shown in Table 3.3 below, controlling for a wide range of additional information about past price movements and trading activity does not lead to further substantial reductions in the size of this estimated coefficient. Controlling for a small number of lags of the dependent variable thus appears to be sufficient to take the media's reporting on the past into account.

Table 3.3 shows that accounting for past price movements and market activity in a very broad way does not change the media coverage effect much more than controlling for a single lag of volatility. I include 10 and 20 lags of the absolute return, trading volume and turnover, as well as lagged polynomial of order 1/2, 2, 3/2, 3 and 4. Even with these additional 479 covariates capturing past events, the magnitude of the estimated media coverage effect remains around the same.[8]

[8]There are four lagged variables (volatility, absolute return, volume and turnover. Each of these appears 5 times as a different polynomial (1/2, 1, 2, 3/2, 3, 4). Finally there are 20 lags of each, giving 480 variables in total (minus the original lagged value of volatility).

Table 3.3: Controlling for a wide set of past price movements and trading activity does not reduce the effect

	Dependent variable:					
	$vol_{i,t}$					
	(1)	(2)	(3)	(4)	(5)	(6)
$mention_{i,t}$	0.160***	0.156***	0.137***	0.136***	0.123***	0.118***
	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)	(0.017)
$vol_{i,t}$ lags	✓	✓	✓	✓	✓	✓
$\lvert\Delta p\rvert_{i,t}^{o,c}$ lags		✓	✓	✓	✓	✓
Volume lags			✓	✓	✓	✓
Turnover lags			✓	✓	✓	✓
$vol_{i,t}$ polynomial lags				✓	✓	✓
$\lvert\Delta p\rvert_{i,t}^{o,c}$ polynomial lags					✓	✓
Volume polynomial lags					✓	✓
Turnover polynomial lags					✓	✓
Firm fixed effects	✓	✓	✓	✓	✓	✓
Time fixed effects	✓	✓	✓	✓	✓	✓
No. of lags	10	10	10	10	10	20
Observations	735,517	735,426	670,247	670,247	670,247	576,128
R^2	0.498	0.499	0.517	0.519	0.520	0.531
Adjusted R^2	0.495	0.495	0.513	0.516	0.517	0.527
Residual Std. Error	1.870	1.867	1.814	1.810	1.809	1.778
Note:						*p<0.1; **p<0.05; ***p<0.01

An alternative test of whether the media coverage effect is driven by reporting on past news is to examine whether forward and backward looking articles have a similar effect. If backward looking coverage is driving the effect, then this would support the idea that this media coverage effect is due to the media reporting on the past. This could still constitute a causal effect, as investor's attention could still be influenced by backward looking articles. However, if forward looking articles have a greater effect then this would cast doubt on the view that the effect is simply due to reporting on the past and the persistence of volatility.

I use two strategies to identify articles as forward looking: using a measure based on the article text and by focusing on articles that were published on a Monday. Both approaches show that forward-looking articles have a greater effect, corroborating that media coverage's predictive content for volatility is not a result of reporting on the past.

I use the Linguistic Inquiry Word Count (LIWC) dictionaries to classify the text of the articles as forward looking (Pennebaker et al., 2001). LIWC is a widely used proprietary software based on dictionaries developed by psycholinguists to score documents based on the proportion of words that "reflect different emotions, thinking styles, social concerns, and even parts of speech."[9] This includes scores for "future focus" and "past focus" which I use here. As the word lists on which LIWC is based are not publicly observable, I also select short dictionaries of my own which yields similar results, as shown in Appendix C.2.2 alongside some further explanation and examples of LIWC classifications.

I include the $future_{i,t}$ and $past_{i,t}$ variable from the LIWC, i.e. the proportion of words which are from the "future" dictionary, in the dynamic panel regression described previously. As Table 3.4 shows, articles with a higher proportion of forward-looking future tense words have a greater effect on volatility, and backward-looking articles have a smaller effect. Column (1) presents the baseline case, with firm and day fixed effects and 10 lags of the dependent variable. Column (2) adds the $future_{i,t}$ and $past_{i,t}$ variables.[10] We observe that forward looking coverage has a greater effect, casting doubt on the hypothesis that the media coverage effect is driven by reporting on a persistent past.

In addition to the text-based measure described above, we can also use the day of the week and article is published to identify articles that are more likely to be forward looking (and less likely to be reporting on the past). Articles are more likely to be forward looking on a Monday as any events that occurred on Friday will have been reported on in the weekend edition of the newspaper and considerably less new information about firms becomes available over the weekend. Therefore, by isolating the effect of media coverage on Mondays separately from other days we have can identify coverage that is more likely to be forward looking without looking at the contents of the text, providing an alternative test to that based on the LIWC dictionaries.

Column (3) of Table 3.4 shows that mentions on a Monday have a larger effect that mentions on other days. The $mention(Monday)_{i,t}$ is simply the interaction term of the

[9] http://liwc.wpengine.com/how-it-works/

[10] As these variables are only non-zero for days in which there articles, we can think of this as being interacted with the mention dummy.

mention dummy variable and a Monday dummy variable. The point estimates thus imply that a mention on a Monday has a total effect of 0.348. Once again, this provides evidence that the media coverage effect is not driven by reporting on the past.

Finally, we can also test the persistence hypothesis by explicitly controlling for future media coverage. It is certainly the case that the FT does report on past events, so we would expect to see a strong empirical link between future coverage and volatility. Column (4) in Table 3.4 shows results when a mention dummy for coverage in the next (trading) day's newspaper is also included. The effect of being mentioned in the next day's newspaper is, unsurprisingly, considerably larger than the effect of being mentioned in today's newspaper. Importantly, however, the effect of today's coverage is not simply a product of persistence and reporting on the past as even when future media coverage is included, the baseline effect of around 10 basis points remains highly significant.

Table 3.4: Forward looking articles have a greater effect and the effect is not explained by future coverage

	Dependent variable:			
	$vol_{i,t}$			
	(1)	(2)	(3)	(4)
$mention_{i,t}$	0.160^{***}	0.173^{***}	0.135^{***}	0.095^{***}
	(0.016)	(0.030)	(0.017)	(0.016)
$future_{i,t}$		2.132^{**}		
		(1.025)		
$past_{i,t}$		-1.010^{*}		
		(0.536)		
$mention(Monday)_{i,t}$			0.213^{***}	
			(0.049)	
$mention_{i,t+1}$				0.855^{***}
				(0.016)
$vol_{i,t}$ lags	✓	✓	✓	✓
Firm fixed effects	✓	✓	✓	✓
Time fixed effects	✓	✓	✓	✓
Observations	735,517	735,517	735,517	723,882
R^2	0.498	0.498	0.498	0.501
Adjusted R^2	0.495	0.495	0.495	0.497
Residual Std. Error	1.870	1.870	1.870	1.867
Note:				*p<0.1; **p<0.05; ***p<0.01

New information

If the co-movement between media coverage and volatility is driven by new information, then there may not be a causal relationship. This new information could either have been available at the time of publication, but only subsequently priced in, or it arrives during the day but its arrival was anticipated by the media. I control for the arrival and anticipation of new information in three different ways and show that none of these can explain the media coverage effect.

First, I use absolute intra-day and overnight returns, as the broadest possible measure of realised new information that day. Second, I use the implied volatility derived from options on the underlying stocks, which provides a measure of market expectations of future volatility, to demonstrate that the effect is not driven by the arrival of new information that is anticipated by the media. Third, I test whether the effect is driven by information which is available at the time of publication by showing that the sentiment of articles does not predict subsequent returns. In Section 3.5 I then control for the content of the articles in a more systematic way using a Bayesian Topic Regression.

An approach I do not follow here is to explicitly control for information events, and for example examine the effect of media coverage on market activity surrounding earnings announcements as Peress (2008) does. This approach necessarily limits analysis to media coverage of a specific type of information event. Furthermore, it does not account for the fact that the media's decision to report on a event will depend on its nature - i.e., coverage is probably more likely for events that are thought likely to be surprising. My approach, of controlling for the realised absolute return and the markets expectation of volatility, provides the broadest possible measure of the new information which is realised and expected.

That the media anticipates newsworthy events to some extent is beyond doubt. Figure 3.6 shows an example of this, where an article on the 19th of January 2009 described how RBS was expected to reveal a loss of £20bn that day. An article from the following day, the 20th January, shows that the losses were in fact £28bn. The stock price of RBS fell by 66% on the 19th January 2009 and the intra-day volatility, as measured by the percentage difference between the highest and lowest price, was 77%.

Figure 3.6: An example of the *Financial Times* anticipating a newsworthy event

2009-01-19: RBS set to reveal it suffered £20bn loss last year

Royal Bank of Scotland is today expected to announce it suffered a £20bn loss last year as the state-controlled lender unveiled further damage as a result of the credit crunch. RBS executives were last night finalising a trading statement that shows it made the biggest loss in British corporate history, exceeding the £15bn deficit reported by Vodafone, the mobile phone group, in 2006...

2009-01-20: RBS counts £28bn cost of past ambition

Stephen Hester, chief executive of Royal Bank of Scotland, yesterday resisted all attempts to persuade him to blame his predecessor for the bank's travails. But Monday's trading statement told its own story about Sir Fred Goodwin's stewardship. The problems, which will be detailed when RBS reports expected losses of £28bn ($40.6bn) for 2008 next month, have been caused by two main factors...

In the RBS example, the high volatility on the day of the first article was at least in large part due to the new information that was anticipated by the media coverage, rather than the media coverage itself. Controlling for every new piece of information explicitly would be impossible as it would not only require data on the date of every "newsworthy event", but also a measure of how significant that new information was and whether it was positive or negative. However, a natural proxy for new information that day is the absolute intra-day return. It measures how investors beliefs about the profitability of a firm have changed over the course of the day. The absolute returns explain a large portion of the variation in intra-day volatility - the R^2 in a regression of just the volatility measure on absolute intra-day and overnight returns is 0.581.

The absolute intra-day return is thus an extremely broad and holistic measure of the new information that has been priced in that day. By controlling for this we can effectively place a lower bound on the effect of media coverage on volatility. It may of course be the case that media coverage changes the price reaction to a given piece of new information, i.e. it could be that the price of RBS stock would have fallen by less than 66% had there not been media coverage of the announcement. However, identifying an effect on volatility greater than that expected by the absolute return shows that media coverage increases the volatility of a firm's stock price even taking into account the market's assessment of new information which has arrived that day. This could be explained if, for example, events which have received pre-emptive media coverage are more likely to be initially overreacted to.

Table 3.5 shows results for the baseline panel model when the absolute intra-day return and its lags are controlled for. Comparing columns (1) and (2) shows that although the absolute intra-day (and its lags) return explains a large portion of the variation in the volatility measure, the media coverage effect remains highly significant and although it is reduced is still just under 10 basis points. In column (3) I also control for the absolute overnight return which, as indicated by the sentiment regressions in Table 3.6, may also be affected by new information available at the time of publication. Controlling for the absolute intra-day and overnight returns therefore suggests that while some of the media coverage effect on volatility is driven by media anticipation of new information, the majority of it is not. It is worth bearing in mind once again that as returns may also be made more extreme by media coverage, the effect that survives controlling for absolute returns should be regarded as a lower bound.

As well as controlling for the realised new information in the form of absolute returns, we

can also control for the market's expectation of volatility. If the media is able to anticipate days which are likely to contain newsworthy events, this is also possible for investors. Investors therefore have an incentive to hedge against potential large movements in the stock price, and can do so through the options market. Any expected increase in volatility, for example on a day in which a firm releases an earnings report, will be reflected in the price of options as investors seek to hedge against large movements. The price of options on a particular stock can therefore give a measure of the market's expectation of the future volatility of that stock.

For 152 of the 275 firms, making up 362,833 of 863,641 observations of the total data, an implied volatility measure linked to the individual equity price series are available through Thomson Reuters Datastream. This implied volatility measure is calculated from the values of the nearest expiry month options, priced on the NYSE Liffe exchange, so a continuous series is available for as long as the options class continues to exist. Implied volatility is calculated as the volatility parameter of an at-the-money option, which is to say an option whose strike price is equal to the current price of the underlying security (i.e. the stock price). These options have no intrinsic value, but can have time value as it allows investors to hedge against future movements in the price.[11] I include the implied volatility for both put and call options, which are both highly significant predictors of realised volatility. In fact, the R^2 in a regression of just the volatility measure on the two implied volatility measures is 0.525.

Table 3.5 shows that controlling for this implied volatility measure of anticipated price changes does not affect the media coverage effect much at all. Column (4) includes 10 lags of each implied volatility (based on put and call options). Column (5) takes into account potential non-linear relationships between implied volatility and the empirical measure of volatility used by including polynomials of the implied volatility measures (1/2, 1, 2, 3, 4).

[11] The documentation of the variable in Datastream states that the implied volatility is calculated using the Black-Scholes formula. See https://datateamoftheeur.wordpress.com/category/datastream/options/ for more details on the variables available through Datastream.

Table 3.5: Neither realised nor anticipated new information explain the media coverage effect

	Dependent variable:				
	$vol_{i,t}$				
	(1)	(2)	(3)	(4)	(5)
$mention_{i,t}$	0.160***	0.088***	0.090***	0.084***	0.079***
	(0.016)	(0.012)	(0.012)	(0.013)	(0.013)
$\lvert\Delta p\rvert_{i,t}^{o,c}$		0.716***	0.708***	0.702***	0.699***
		(0.001)	(0.001)	(0.002)	(0.002)
$\lvert\Delta p\rvert_{i,t}^{c,o}$			0.089***	0.124***	0.128***
			(0.001)	(0.002)	(0.002)
$VI_{i,t-1}^{put}$				0.768***	0.906
				(0.054)	(1.115)
$VI_{i,t-1}^{call}$				0.211***	5.973***
				(0.038)	(0.902)
$vol_{i,t}$ lags	✓	✓	✓	✓	✓
Lags of all covariates	✓	✓	✓	✓	✓
Polynomial lags of VI					✓
Observations	735,517	730,836	719,292	306,741	306,741
R^2	0.498	0.707	0.712	0.768	0.770
Adjusted R^2	0.495	0.705	0.710	0.765	0.767
Residual Std. Error	1.870	1.421	1.408	1.184	1.180
Note:					*p<0.1; **p<0.05; ***p<0.01

Perhaps the most intuitive, but by no means the most plausible, explanation of the media coverage effect is that it the FT newspaper provides information which is new to investors and then priced in the following day. A priori, we would be surprised if this was the case precisely because of the timings set out in the previous subsection and because of the time lags inherent with any print newspaper (i.e. the time taken to print and distribute). Therefore, once information appears in the print newspaper it is essentially "old news" from the perspective of financial markets. However, it could still be that as the newspaper is a respected institution, its view is taken seriously by investors who update their beliefs in response to it.

Measuring the sentiment of media coverage provides a useful initial test for whether the FT provides new information to investors. If the sentiment of FT articles can predict subsequent returns, this would imply that the FT does provide, or at least makes accessible, information that investors take seriously. Sentiment is measured as the proportion of positive words from the economics and finance dictionaries provided by Loughran and McDonald (2011), as described in Appendix C.2.3. This measure is designed to capture positive and negative news in the context of financial news.[12]

Column (1) in Table 3.6 shows that the media coverage effect on volatility is not explained by the sentiment of that coverage. Column (2) shows that sentiment of coverage the next day also does not explain this effect. In both cases, we see that negative coverage has a greater effect on volatility than positive.[13] This could be because investors are more likely to pay attention to negative media coverage, suggesting a salience based explanation of the media coverage effect.

Table 3.6 shows that the sentiment of articles does not significantly predict intra-day returns, although tomorrow's media coverage does. The purpose of showing the strong predictive effect of the next day's sentiment on today's returns is to demonstrate that the Loughran and McDonald (2011) measure captures economically relevant information in the expected way. Column (3) shows that neither sentiment nor the indicator of media coverage have a significant effect on intra-day returns, while column (4) shows that sentiment of the next day's coverage is a strong positive predictor of today's intra-day returns. This is in line with the view that any information contained in the FT is priced in by the time the market opens several hours after publication.

In Appendix C.2.3 I also show that sentiment does significantly predict effect on overnight returns, and that media coverage in general predicts positive overnight returns. This validates the sentiment dictionary approach and indicates that the morning print newspaper does reflect some events that occurred outside of market hours. Furthermore, it shows the importance of using an intra-day measure of volatility (i.e. bounded by trading hours) rather than the close-to-close return as this avoids endogeneity. It could be that the positive sentiment effect on overnight returns is due to new information which is provided by the newspaper, but it could also be that the newspaper is simply reporting on trading in

[12]So while sentiment is often thought of as capturing opinions expressed in text rather than content, in this case we can think of it as a measure of the content of the articles.

[13]The point estimate for the effect of $sentiment_{i,t}$ on $vol_{i,t}$ of -0.277 which therefore implies that very positive news may lead to a decrease in volatility the following day.

overseas markets. Therefore, while the effect of media coverage on the overnight returns is interesting, the associated timings make arguing for causation difficult.

Table 3.6: Sentiment does not predict future returns, but negative articles have a greater effect on volatility

	Dependent variable:			
	$vol_{i,t}$		$\Delta p_{i,t}^{o,c}$	
	(1)	(2)	(3)	(4)
$mention_{i,t}$	0.161^{***}	0.161^{***}	0.015	0.023
	(0.016)	(0.016)	(0.019)	(0.019)
$sentiment_{i,t}$	-0.277^{***}		0.115	
	(0.064)		(0.073)	
$sentiment_{i,t+1}$		-0.137^{**}		1.130^{***}
		(0.065)		(0.074)
$\Delta p_{i,t}^{c,o}$			-0.531^{***}	-0.528^{***}
			(0.002)	(0.002)
$\Delta p_{i,t}^{o,c}$ lags			✓	✓
$vol_{i,t}$ lags	✓	✓		
Firm fixed effects	✓	✓	✓	✓
Time fixed effects	✓	✓	✓	✓
Observations	735,517	723,882	719,337	707,836
R^2	0.498	0.499	0.387	0.385
Adjusted R^2	0.495	0.496	0.382	0.380
Residual Std. Error	1.870	1.871	2.114	2.118
Note:				*p<0.1; **p<0.05; ***p<0.01

This Section has used a measure of media coverage - a dummy variable capturing whether a firm is the focus of an article in the FT print newspaper - to show a robust and plausibly causal effect of media coverage increasing volatility. It has shown that neither reverse causality, reporting on persistent past volatility or new information explain this effect. Using the simple dummy variable has the advantage of making the interpretation of the coefficients, and thereby the magnitude of the effect, simple and intuitive. However, it

does not make use of the potentially rich source of information contained in the contents of the articles themselves. Section 3.5 sets out a novel topic modelling framework with which the content of the articles can be quantified in a systematic way.

3.5 Controlling for content of articles

This Section uses the Bayesian Topic Regression (BTR) framework presented in **Chapter 2** to broadly control for the content articles. More specific approaches to exploring the content of articles are possible, and in the last Section we saw that forward looking and negative articles have a slightly greater effect on volatility. In what follows, I take a more systematic approach to analysing the content of articles using a model which decomposes articles into topics that best explain the volatility, while taking into account media coverage and other control variables.

3.5.1 Model set up

The Bayesian Topic Regression (BTR) framework allows researchers to extract topics while simultaneously estimating a regression in which those topics are explanatory variables. This allows us to learn topics which directly explain the effect of interest, as well as conduct inference jointly over the topics and the regression parameters. Furthermore, in estimating coefficients on the topics and covariates jointly, it respects the Frisch-Waugh-Lovell Theorem. The model is explained in **Chapter 2** together with example applications that illustrate its value, so I will not discuss it in detail here.

Before estimating the model, the articles are cleaned following standard procedures. All stopwords are removed along with special characters and words shorter than 3 characters in length. Words are then stemmed using a Porter stemmer, so that they are reduced to their root. Words that appear very frequently or very rarely, as measured by their term frequency-inverse document frequency (tf-idf) are removed as well as any words that appear in fewer than 2 articles are also removed. When multiple articles for the same firm appear on a day, these articles are combined into one. This leaves a total of 17,027 documents and a vocabulary of 13,209 unique terms.

BTR assigns each word in each article to one of a pre-set number of topics, K. This therefore generates a document-level proportion for each of these K topics (as well as a distribution over the terms in the vocabulary for each topic which can be used for interpretation). These document-level topic proportions are then included in a regression

alongside additional numerical covariates, to predict an outcome variable, which in this case is intra-day volatility. This regression is then estimated jointly alongside the topic assignments of each word.

In line with the discussion of identification of a casual effect presented in Section 3.4, I include a number of covariates to control for possible persistence and new information effects. I include 5 lags of the dependent variable, as this reflects a full week and adding further lags to the panel model did not materially change the results. I also include the contemporaneous values of the absolute intra day return and the implied volatility measures. These two variables control for realised and anticipated new information, and adding further controls (e.g. overnight returns or polynomial lags) did not change the results of panel models. In the place of firm and day fixed effects I also include the The vector of numerical covariates for firm i at time t is shown in Equation 3.4.

$$x_{i,t} = \begin{pmatrix} mention_{i,t} \\ sentiment_{i,t} \\ |\Delta p|^{o,c}_{i,t} \\ \overline{vol}_t \\ \overline{vol}_i \\ vol_{i,t-1} \\ vol_{i,t-2} \\ vol_{i,t-3} \\ vol_{i,t-4} \\ vol_{i,t-5} \\ VI^{put}_{i,t} \\ VI^{call}_{i,t} \end{pmatrix}' \tag{3.4}$$

The regression part of the model can thus be expressed as in Equation

$$vol_{i,t} = \omega'_x x_{i,t} + \omega'_z \bar{z}_{i,t} + \epsilon_{i,t}, \tag{3.5}$$

where $\bar{z}_{i,t}$ is the $1 \times K$ vector of topic proportions firm i on day t and $\epsilon_{i,t}$ is normally distributed residual. Note that as the topic proportions will sum to one by construction, we will not need to include a constant in the regression. I include the average volatility for each firm over time, $\overline{vol}_i$ and for each day across firms $\overline{vol}_t$ in place of firm and day fixed effects.

The estimated media coverage effect is then the coefficient on $mention_{i,t}$, so the first

element of ω_x. The posterior distribution of this parameter will therefore be the primary focus of the next Section.

As the majority of observations are not associated with any media coverage, but these observations are still useful for learning the relationship between $x_{i,t}$ and $vol_{i,t}$. Furthermore, they are crucial for identifying the media coverage effect by comparing days with media coverage to days without. The topic proportions for these observations are set according to the Dirichlet prior on topic distributions. In practice, given that we use a symmetric prior, these means that the observations without documents will have an equal proportion of each topic.

3.5.2 Results

Alongside the standard BTR described above I also estimate three alternative topic models, the first two act as a robustness check and the third demonstrates the importance of BTR's joint estimation approach.

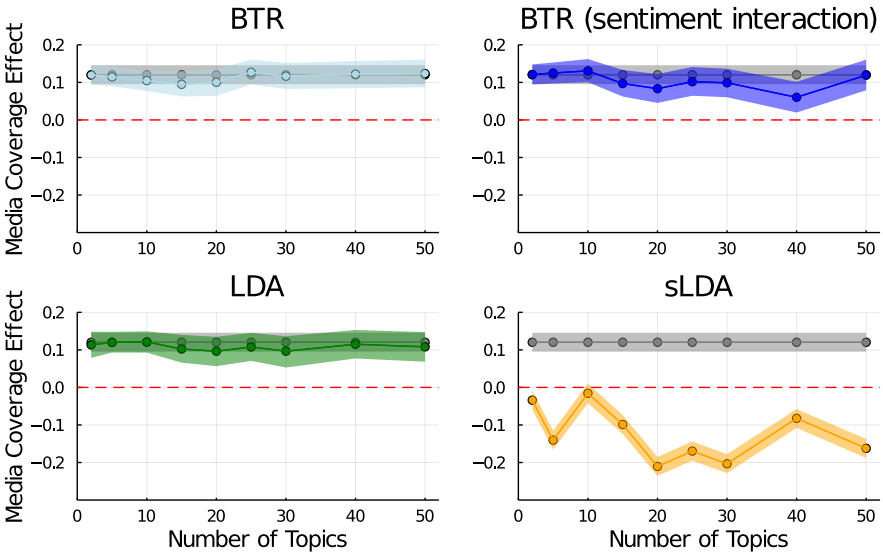
1. **BTR (sentiment interaction)**: adds interaction terms between the score sentiment of each article and the topic proportions. This accounts for the fact that the effect of a certain topic on volatility may differ if the article has a positive or negative tone. This acts as a robustness check on the results given by the standard BTR without interactions.
2. **LDA**: estimates topic proportions with an unsupervised Latent Dirichlet Allocation (LDA) model (Blei et al., 2003). These topic proportions are then included as covariates in a linear regression alongside the same numerical covariates as in the BTR case, shown in Equation 3.4. This also acts as a robustness check to BTR as the topic proportions are estimated in an unsupervised manner it is theoretically consistent to include them in a second-stage regression.
3. **sLDA**: estimates a standard supervised LDA (sLDA) model with intraday volatility as the target variable (Blei and McAuliffe, 2008). The *residuals* from this sLDA model are then used as the independent variable in a seond stage regression on the numerical variables in Equation 3.4. As this approach estimates regression coefficients in two stages and the two sets of regressors are not orthogonal, this approach violates the Frisch Waugh Lovell theorem. It is therefore included not as a robustness check, but to demonstrate the value of BTR's joint estimation approach.

Each model is estimated for a range of number of topics K from 2 to 50, with the model estimated 10 times for each value of K with different initial conditions and random seeds. The posterior distributions for the media coverage effect are then generated by combin-

ing samples from these 10 independent runs. Further details on the hyperparameters and estimation options chosen in each case are given in C.2.4.

Figure 3.7 shows the estimated media coverage effects for each of these four models, alongside the estimated effect for a linear regression without any topics for reference. These results are shown numerically in Appendix C.2.4.

Figure 3.7: Media coverage effect, controlling for article content



Note: Grey points and lines in each panel represent a linear regression on just the numerical covariates, as a reference point for the text models. The shaded bands represent 95% posterior credible intervals for the estimated media coverage effect.

In the first three cases (BTR, BTR with sentiment interactions, and LDA) the posterior distribution for the estimated media coverage effect is concentrated above zero and not dissimilar from the results shown in Section 3.4. Indeed the estimated media coverage effects with the topic models largely overlap those for a linear regression without topics. We can therefore conclude that the content of articles, at least within the structure of these topic models, does not explain the media coverage effect. In the sLDA case, we get a very different estimate of the media coverage effect. This can be explained by the fact that the topic model is estimated in isolation from the media coverage indicator or indeed any of the numerical control variables.

By controlling for the content of articles in a broad and systematic way, we can see that the identified media coverage effect on volatility appears to be due to the mere presence of a media coverage of a firm, rather than what that coverage contains. Once again, this provides support for a salience-based view of the media's role in financial markets.

3.6 Spillovers and aggregate implications

The media coverage effect on volatility appears to spread to other firms linked by the production network. However, an aggregation of the firm-level media coverage measures described in the previous Section does not predict index-level volatility. This provides further evidence for a salience-based interpretation of the causal mechanism, where investors attention is towards some firms and those connected to them, and consequently away from others, by media coverage.

3.6.1 Sector and network-weighted measures

As explicit data on firm-to-firm links is not available for the UK, I rely on the links between sectors implied by the production network. I use the input-output tables for the UK from the Office of National Statistics to compute a priori links between sectors. I then show that media coverage of one firm in a sector has a volatility effect on other firms in the same sector, and that this survives controlling for absolute return and implied volatility, so could be interpreted as a media coverage effect. Conversely, while a volatility effect spreads to other sectors connected by the production network, this does not survive controlling for absolute return. This suggests that while news effects move through the production network, media coverage linked volatility does not, giving further support a causal mechanism based on salience.

Given the NACE sector classification of each firm, I construct a sector level media coverage measure by averaging the mention dummy across all firms in that sector in the sample.[14] To obtain a measure, for each firm, of sector-level media coverage that this is not contaminated by coverage of the firm itself, I remove the contribution of firm i when evaluating the average news in i's own sector. The sector-level media coverage measure for firm i in sector s, with n_s constituent firms, is given by

$$sector\_mentions_{i,t} = \frac{1}{n_s}\sum_{j\in s} mention_{j,t} - \frac{1}{n_s} mention_{i,t} \tag{3.6}$$

As I will distinguish any network effect from an own-sector effect, I also compute the average *realised* volatility of other firms in the same sector $\overline{voll}^s_{j\neq i,t}$. This is calculated analogously to the $\overline{sector\_article}_{i,t}$ effect, by finding the average across the sector less the

[14] Results are qualitatively similar if sector coverage is the maximum of mention dummies across the sector, so sector coverage is 1 if any firm from that sector has an article on that day (see Appendix C.3 for results with this alternative specification).

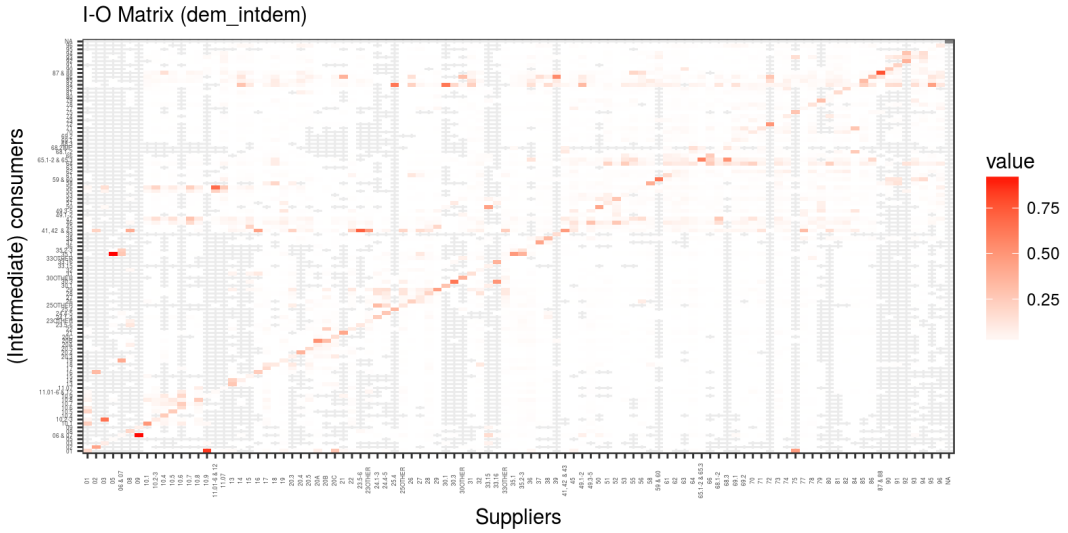
contribution of that firm

$$sector\_vol_{i,t} = \frac{1}{n_s}\sum_{j \in s} vol_{j,t} - \frac{1}{n_s} vol_{i,t} \tag{3.7}$$

This within sector volatility will provide a stringent test for the relevance of media coverage as separate from fundamental correlations between sectors.

For each sector, I also construct a measure media coverage in other sectors weighted by the importance of those sector as either intermediate suppliers or intermediate consumers. Figure 3.8 shows the matrix with each cell adjusted so that it is a proportion of each industry's intermediate demand. The columns of this matrix sum to one, so for example the row for sector 41, 42 and 43 (Construction) show that this sector is generally a high proportion of the customers of many sectors. Construction is thus an important consumer.

Figure 3.8: Input-Output table, as proportion of intermediate demand



The analogous matrix for intermediate consumption is shown in Appendix C.3. This matrix is adjusted so that it gives a proportion of each industry's intermediate consumption. That is to say, how important each sector is as a supplier to other sectors. The rows of this matrix sum to one, so for example the column for Financial services has high values, showing that it is generally a high proportion of the inputs used by many sectors, so an important supplier.

These matrices are used to construct the weighted measure of media coverage from sectors which are important intermediate suppliers, and sectors which are important intermediate consumers. Let s and ς be two of the S sectors. The amount that sector s supplies

to sector ς is $m^{sell}_{s,\varsigma}$ (i.e. the intermediate demand matrix) and the amount that sector s consumes from sector ς is $m^{buy}_{s,\varsigma}$ (i.e. the intermediate consumption matrix).[15] We thus have two measures of media coverage in relevant sectors according to these intermediate demand and intermediate consumption matrices

$$downstream\_mentions_{i,t} = \sum_{\varsigma=1}^{S} m^{sell}_{s,\varsigma} \frac{1}{n_\varsigma} \sum_{j \in \varsigma} mention_{j,t} \tag{3.8}$$

$$upstream\_mentions_{i,t} = \sum_{\varsigma=1}^{S} m^{buy}_{s,\varsigma} \frac{1}{n_\varsigma} \sum_{j \in \varsigma} mention_{j,t} \tag{3.9}$$

The $downstream\_mentions_{i,t}$ variable is thus a measure of media coverage in sectors that are downstream of firm i's sector, while $upstream\_mentions_{i,t}$ is a measure of media coverage in sectors that are upstream of firm i's sector. In other words, "downstream" and "upstream" refer to sectors' positions with respect to sector s.

The I-O tables allow distinguishing between potential effects from downstream and upstream sectors by weighting sectors according to their importance as intermediate consumers or intermediate suppliers. The results presented below suggest that firms' stock price volatility is affected by media coverage of a firm that are in sectors downstream from them (i.e. their customers). This provides some initial evidence that media coverage may have aggregate effects, if it is focused on highly connected firms.

3.6.2 Cross and within sector spillovers

A growing macroeconomic literature studies the importance of the production network in amplifying idiosyncratic shocks Gabaix (2011); Acemoglu et al. (2016). Although the links implied by the production network are relevant for the market reaction to new information Cohen and Frazzini (2008), it is not clear whether this should be the case for a causal media coverage effect. In order to test whether there are downstream and/or upstream effects we can include these production network weighted measures of media coverage in panel regressions of the form discussed in Section 3.4. I show that a firm's stock price volatility is increased by media coverage in sectors that are downstream of them. This effect survives controlling for any coverage and even the realised volatility in that firm's own sector, as well as the previously discussed persistence and anticipation controls.

[15] Intuitively, $m^{sell}_{s,\varsigma}$ is thus the proportion of it's (intermediate) output that sector s *sells* to sector ς, and $m^{buy}_{s,\varsigma}$ is thus the proportion of its (intermediate) inputs that sector s *buys* from sector ς.

Table 3.7 summarises the results. Column (2) adds only the production network weighted measures to the baseline regression with persistence and anticipation controls. We see that the coverage of firms in downstream sectors (i.e. firm i's potential consumers) significantly predicts firm i's intra-day volatility.

As can be seen in the diagonal elements of Figure 3.8, trade within a sector is an important feature of the production network. I therefore also control for news within the firm's own sector, as shown in column (3). This own sector effect is significant but does not remove the downstream sector effect, suggesting that this latter effect is not just driven by co-movement with firms in the same sector. The effect from downstream sectors also survives controlling for the *realised* volatility within firm i's sector, as shown in Column (4), further implying that this spillover effect is not simply generated by co-movement of similar firms.[16]

To further verify the relevance of the network implied by the input-output matrix, I generate a placebo matrix by shuffling the rows of the intermediate demand matrix. Using weights from this placebo matrix does not generate any significant spillovers. Results for this exercise are shown in Appendix C.3.

[16]Note that we also control for firm i's own media coverage throughout, so the effect is unlikely to be driven by correlations in the media coverage of similar firms.

Table 3.7: Spillover effect of media coverage with article variable

	Dependent variable:			
	$vol_{i,t}$			
	(1)	(2)	(3)	(4)
$mention_{i,t}$	0.090***	0.083***	0.093***	0.085***
	(0.013)	(0.014)	(0.014)	(0.014)
$\lvert\Delta p\rvert_{i,t}^{o,c}$	0.723***	0.723***	0.723***	0.717***
	(0.001)	(0.001)	(0.001)	(0.001)
$VI_{i,t}^{put}$	1.059***	1.060***	1.059***	0.983***
	(0.024)	(0.024)	(0.024)	(0.024)
$VI_{i,t}^{call}$	0.027***	0.027***	0.028***	0.025***
	(0.008)	(0.008)	(0.008)	(0.008)
$sector\_mentions_{i,t}$			0.125***	0.059
			(0.046)	(0.046)
$sector\_vol_{i,t}$				0.094***
				(0.002)
$downstream\_mentions_{i,t}$		0.161***	0.140**	0.168***
		(0.059)	(0.059)	(0.059)
$upstream\_mentions_{i,t}$		−0.002	−0.121	−0.134
		(0.113)	(0.121)	(0.121)
$vol_{i,t}$ lags	✓	✓	✓	✓
Firm fixed effects	✓	✓	✓	✓
Time fixed effects	✓	✓	✓	✓
Observations	312,499	312,499	312,499	311,925
R^2	0.766	0.766	0.766	0.767
Adjusted R^2	0.763	0.763	0.763	0.764
Residual Std. Error	1.194	1.194	1.194	1.191
Note:				*p<0.1; **p<0.05; ***p<0.01

These results show that firms stock price volatility may be affected by media coverage

of firms that are downstream of them in the production network. This effect survives controlling for both the average realised volatility of other firms in that sector and the absolute return that day. This suggests that, insofar as there is a causal volatility effect, it may affect firms that are fundamentally related in investor's minds. This is consistent with the effect being driven by increased investor attention due to media coverage.

It is worth noting that these results are based on a very simple measure of media coverage from a single news source, and for a single measure of the linkages between firms. Further work on the potential spillovers of any media coverage effect are therefore certainly necessary before coming to any conclusions.

3.6.3 Media coverage and index volatility

As a test of the aggregate importance of this media coverage effect, I compute an aggregate measure of media coverage from the firm-level measure. This measure does not predict index-level volatility once persistence and anticipation are accounted for.

The $mention_{i,t}$ variable is aggregated up to daily frequency by calculating the mean across all firms in the sample for that day.

$$\overline{mention_t} = \frac{1}{N}\sum_{i=1}^{N} mention_{i,t} \tag{3.10}$$

I then regress this on the realised volatility of the entire FTSE 100 index, as measured by the percentage change between the highest and lowest price that day. Following the same logic as that used to identify the firm level effect, I also control for the absolute index return as a broad measure of new information which may have been anticipated by the media coverage

Table 3.8 shows that although media coverage of listed firms is strongly correlated with index-level volatility, column (1), this effect is not significant once past volatility or the absolute return are controlled for. The aggregate level of media coverage across the index cannot be interpreted as having a causal impact on index-level volatility. The lack of an index-level effect is again consistent with a salience-based theory of the media's role in financial markets: media coverage affects which firms or sectors market participants chose to devote their (limited) attention too.

Table 3.8: The aggregate effects of firm-level media coverage

	Dependent variable:		
	vol_t^{FTSE}		
	(1)	(2)	(3)
$\overline{mention}_t$	3.893***	0.100	0.475
	(1.290)	(0.934)	(0.617)
$\lvert\Delta p\rvert_t^{FTSE}$			0.643***
			(0.010)
Constant	1.329***	0.127***	0.113***
	(0.029)	(0.027)	(0.018)
vol_t^{FTSE} lags		✓	✓
Observations	4,291	3,690	3,807
R^2	0.002	0.574	0.805
Adjusted R^2	0.002	0.573	0.804
Residual Std. Error	0.975	0.651	0.439
F Statistic	9.115***	450.366***	1,565.109***
Note:		*p<0.1; **p<0.05; ***p<0.01	

3.7 Conclusion

This paper provides evidence for a robust and substantial causal effect of financial news media coverage on stock price volatility. It shows while this effect is potentially important at a firm-level, it has limited aggregate effects. This is consistent with a theory of media coverage in which the media's editorial decisions direct the attention of investors towards firms with more coverage. This leads to an increase in the volatility of that firm's stock price over and above the pricing in of any new information that is either anticipated by the media are provided in the content of articles. However, as the attention that is drawn towards these firms is directed away from others, any spillover effects of media coverage on volatility are likely limited to related firms rather than the market more broadly. We can therefore think of this media coverage effect as reallocating volatility across the market rather than creating it from nothing.

Chapter 4

The Shifting Focus of Central Bankers

Abstract

This paper quantifies the focus of central bank communication and news media, offers an explanation for its variation over time, and shows a robust co-movement in this focus. A model of multidimensional uncertainty and limited communication is proposed to explain the shifting focus of central bank communication. Evidence from the Survey of Professional Forecasters is used to support this explanation, suggesting that the focus of the Federal Reserve's communication shifts to cover variables about which there is greater uncertainty. An event study approach is used to show a potentially causal influence of Federal Reserve communication on the focus of US news media, implying that central banks have some power to inform the public even if their own communication does not reach agents directly. Finally, we show that the communication of three different central banks (Federal Reserve, Bank of England and European Central Bank from 1997 to 2014) co-move, and that the focus of the Federal Reserve's communication appears to lead that of other central banks.

4.1 Introduction

The content and focus of a central banks' communication changes over time in response to the state of the economy. This paper aims to quantify and explain this shifting focus, and document co-movement in this focus in the communication of different central banks and the media. It suggests that a significant driver of this variation may be a result of central banks devoting this limited volume of their communication to aspects of the economy where it can most benefit the private sector. An implication of this explanation is that the focus of different central banks will co-move, because of shared global aspects of their respective economies, but also that the communication of one central bank may have a direct influence on that of one another.

Central bank communication seen as an increasingly important tool for policy makers. There are two principal channels through which central bank communication is thought to influence the economy. A "monetary" information channel working through private sector expectations about future policy, and a "non-monetary" information channel working through the private sector's assessment of the state of the economy. This paper joins the growing literature focussed on this second channel, e.g. Cieslak and Schrimpf (2019); Jarociński and Karadi (2020), showing how it can lead to measurable changes in the focus of a central bank's communication over time.

The aim of this paper is to quantify and explain which aspects of the economy a central bank chooses to focus on, rather than measure the information they convey (i.e. what they talk about, rather than what they say about it). Therefore, we work directly with the text of central bank communication and news media articles, rather than market reactions to this communication. We use the Latent Dirichlet Allocation algorithm to decompose the text of central bank communication into topic proportions (Blei et al., 2003). These topic proportions are then used as direct measures of the focus of the central bank's communication on different aspects of the economy.

For a central bank to have a non-monetary information effect it must either have some private information, or be able to present public information in a way that is useful to agents in the economy. There is extensive empirical evidence that this is the case, which includes both studies showing that central bank forecasts are more accurate than those of the private sector such as Romer and Romer (2000), and evidence from market reactions to central bank communication (Andersson et al., 2006; Born et al., 2014; Cieslak and Schrimpf, 2019). In limiting the volume of its communication and only presenting the

most pertinent and useful information, a central bank could improve the private sector understanding of the economy and thus improve welfare.

In order to clarify this intuition, that central bank communication ought to focus on aspects of the economy around which it can have a positive impact on private sector expectations, we set out a model of this communication decision. Rather than focus on how a central bank should structure its communication, as previous literature such as Haldane et al. (2020) has done, our focus here is on what the central bank ought to focus on given the communication channels it has available.

The shifting focus of central bank communication reflects multi-dimensional uncertainty about the current state of the economy. The central bank receives signals about various economic state variables, and wishes to improve the decision making of the private sector by sharing this information. However, as the volume of the central bank's communication is limited, it faces a trade-off and so has to chooses the focus of its communication so that it most improves the private sector's assessment of the economy. The model shows that central bank's communication will therefore focus more on state variables about for which its signal is more informative, and for which it receives more extreme signals. This explanation for the shifting focus of central bank communication based on multidimensional uncertainty also has implications for the focus of the media. The media's editorial decisions are made, in part, on a profit maximising basis and so will aim to provide their readers with information which they find most useful. The media's incentives are thus similar to those of the central bank, so we would expect to see similar co-movement patterns between media focus and uncertainty, and central bank focus and uncertainty.

Empirical support for this explanation is provided by linking the topic proportion series to the dispersion of forecasts given in the Survey of Professional Forecasters (SPF). The topics learned by the LDA model have fairly clear interpretations and in many cases can be linked to an economic time series. We show that focus on topics in both the media and central bank communication is correlated with the dispersion of private sector expectations of related economic variables. A panel data set based on the forecast dispersion, central bank communication focus and media focus on different economic variables is constructed to show that central bank communication and media focus co-move with this forecast dispersion.

While it may be plausible that participants in the SPF in the private sector (for whom forming expectations may be part of their job description) are directly exposed to central

bank communication, this is not the case for households. Survey evidence suggests that households have very limited if any direct exposure to central banks (Kumar et al., 2015; Binder, 2017b), so will rely instead on the media for information about the state of the economy. Previous literature has shown that media can affect households' expectation formation Lamla and Lein (2014). Using an event study approach, we show that the focus of central bank communication plausibly has a causal impact on the focus of economic news articles. This suggests that in setting the focus of its communication, the central bank not only influences the information sets of those who are directly exposed to the communication, but also those who rely on the media for information.

The final contribution of this paper is to take the implications of the single economy case and apply them to multiple economies. Extending the model of multi-dimensional uncertainty to include multiple economies and multiple central banks, illustrates that if central banks can observe one another's communication and the shocks hitting their economies are correlated, we would not only expect their focus to co-move, but also that they might have a direct influence on each other.

We find significant and robust co-movement of topic proportions across the communication of the Federal Reserve, Bank of England and European Central Bank over a sample period of 1997 to 2014. The Federal Reserve's communication appears to be the most influential, as it leads the communication of other central banks at a quarterly frequency, and a shortening in its publication lag in 2005 led to an increase in the observed co-movement. This is consistent with the prediction that the focus of one central bank's communication will follow that of other central banks who's communication is informative for its own economy. This is in line with evidence from Armelius et al. (2020) who find that the sentiment (rather than focus) of Federal Reserve communication has the greatest spillover to other central banks' communication.

The remainder of the paper is structured as follows. Section 4.2 reviews relevant literature and points out how this paper contributes. Section 4.3 then describes the data used and explains the LDA topic model that is applied to the text documents. Section 4.4 sets out a model in which the focus of a central bank's communication responds to multi-dimensional uncertainty, and an extension of this model to multiple central banks. Section 4.5 show's that the Federal Reserve's communication focuses more on variables for which the private sector's forecasts are more dispersed, and that the focus of central bank communication may influence the focus of the news media. Section 4.6 documents a robust co-movement across the communication of the Federal Reserve, Bank of England

and European Central Bank. Finally, Section 4.7 concludes.

4.2 Contributions and Related Literature

This section highlights this paper's contribution by reviewing two sets of relevant literature: that on central bank communication and the use of text as data in monetary economics. The principal contributions of this paper are to document a co-movement in the the *focus* of communication across central banks and identifying a potential influence across central banks, and provide an explanation for this co-movement. In this sense, it is close to the work of Armelius et al. (2020) and Gonzalez et al. (2021), who document a similar co-movement in the *tone* of communication across central banks. Furthermore, it verifies that the focus of central bank communication and the news media is measurably correlated and provides evidence that the central bank may be able to influence the focus of the media.

Central Bank Communication. The increasing focus on transparency and communication among central banker in recent decades has been matched by wide-ranging literature on how central banks should and do communicate with the wider public. Blinder et al. (2008) and de Haan and Sturm (2019) provide wide-ranging reviews of this literature. There are two parts of this literature which this paper is closest to: central banks' private, non-monetary information, and central bank communication beyond financial markets.

While much of the literature focuses on communication about current and future policy, a separate issue is the central bank's potentially important additional dimension is communication about the current and future state of the economy. We follow Cieslak and Schrimpf (2019) in referring to this as non-monetary information, to distinguish this from the central bank's communication about its reaction function and intentions for the future policy.

For CBC to have non-monetary information effects, it must be the case that central banks have some relevant private information about the state of the economy. As discussed by Miranda-Agrippino and Ricco (2021), there is plenty of empirical evidence of informational asymmetry between central banks and financial markets. For example Romer and Romer (2000) show that Federal Reserve staff forecasts are more accurate than private sector forecasts, Andersson et al. (2006) find that markets react to information on economic outlook. More recently, high-frequency event studies have shown

financial market reactions reactions to non-monetary information Nakamura and Steinsson (2018); Cieslak and Schrimpf (2019); Jarociński and Karadi (2020). Hansen et al. (2019) use the separate publication of the Bank of England's Inflation Report from any policy decision to identify signals which drive long-run but not short-run rates, providing evidence that central bank communication can shape markets' long run expectations. Similarly, Leombroni et al. (2018) use the temporal separation of ECB policy announcements and press conferences to show yield curve reactions consistent with a risk premium channel of central bank communication, where an expansionary shock is interpreted as bad news about potential break-up of the eurozone.

Relatedly, several papers examine whether the optimal level of transparency may involve central banks hiding certain information. Reis (2013) argues that central banks face a "cacophony" trade-off between revealing more information and ensuring a common understanding. Haldane et al. (2020) take a rational inattention approach to this problem, showing that central banks ought to structure their communication in a way that can reach a wider audience but that overly simple communication can have negative effects by eroding trust in the central bank. This paper takes for granted that central banks have a limited capacity for communication, and does not take a stance on *how* information should be conveyed. Instead, it explores how central banks should choose which of the many aspects of the economy they should (and do) focus on in their communication.

Although most research on central bank communication focuses on the reaction of financial markets, but as argued by Haldane et al. (2020), communication with the wider public is both desirable and feasible. For example, the Federal Reserve communication strategy separates political authorities, financial markets and the general public (Bernanke, 2003). Binder (2017a) surveys this growing body of work which recognises that central bank's aim to communicate with different target audiences, emphasising the potentially crucial but academically neglected role of the media. For example, the rational inattention literature in a monetary policy communication context predicts that agents pay attention to information when the benefits of doing so (improved decision making) outweigh the costs, of using their scarce information processing capacity (Sims, 2010). The costliness of paying attention depends on several factors, some of which the central bank has control over: agents' economic and financial literacy; the format and complexity of communications; and crucially media coverage.

Although many central banks make efforts to increase the financial literacy of their popu-

lations[1] and to make their communication more readable (Bholat et al., 2018), it remains the case that details of their communication are only accessible to expert audiences. Bulíř et al. (2013) compare the readability of inflation reports and press statements from different central banks and show that even the clearest (UK and Sweden) require around 12 years of schooling to read. Similarly, under Janet Yellen, the FOMC post-meeting statements required a reading level several years beyond college education (Hernández-Murillo et al., 2014). Furthermore, Kumar et al. (2015) find that very few households get information directly from central banks or are even aware of how the central bank operates and studies which aim to test households' understanding of monetary policy generally find that they are inconsistent with the Taylor Rule (Van Der Cruijsen et al., 2010; Carvalho and Nechio, 2014; Dräger et al., 2016).

Given the often technical nature of central bank communication, households can be expected to rely on media coverage of central bank communication to learn about the state of the economy. Carroll (2003) shows that expectational dynamics are well captured by a model in which households' views derive from news reports. Similarly, Lamla and Lein (2014) provide empirical evidence that media coverage can affect the expectation formation of consumers. Furthermore, the well-documented discrepancies between the expectations of households and professional forecasters, e.g. Coibion and Gorodnichenko (2015a) and Coibion and Gorodnichenko (2015b), suggest that the role of the media is an important avenue for research.

While there is some work addressing the communication of political leaders with the public through the media (Fairclough et al., 2011; Lee, 2014), how and whether central banks and the media interact has received very little attention. Binder (2017b) uses the Pew Economic Journalism News Coverage Index, which code stories from a variety of news sources by broad topic and with up to two "newsmakers", to show that press conferences do result in significantly higher coverage of the Federal Reserve, but this coverage is still orders of magnitude less than coverage of the US President. This paper thus represents one of the first attempt to measure the interaction between news media and central bank communication in a systematic way.

The extent to which central banks are able to communicate with both households and the private sector, as well as the role that the media plays in this communication, is still very much an open question. This paper makes a contribution to addressing this by

[1] An example of this is the Bank of England's econoME resource which aims to give young people greater economic and financial awareness.

showing that central bank communication can direct the focus of news media, and provides evidence that central banks tailor their communication to take the private sector's limited attention into account. Furthermore, we also present some suggestive evidence of an effect of communication across different central banks, the implications of which may be worth exploring in more detail.

Text as Data in Monetary Economics. As a large portion of economic communication is done through text and speech, recent years have seen an increase in the use of techniques from natural language processing as well as traditional qualitative work to analyse the text of key economic documents. Romer and Romer (1989) is perhaps the first attempt to do so systematically in the context of monetary policy, using the minutes of FOMC meetings to identify monetary policy shocks. More recently, researchers have moved away from the "manual" approach employed by Romer and Romer (1989) towards automated text analysis, towards more automated methods. An early example of this is Lucca and Trebbi (2009).

In particular, this paper makes use of the popular Latent Dirichlet Allocation (LDA) model introduced by Blei et al. (2003) to decompose the text of central bank communication and media articles into topics.[2] This model has been used fairly widely in the monetary policy literature, due to its intuitive simplicitly and interpretable results. For instance, Hansen and McMahon (2016) and Hansen et al. (2018) use LDA, alongside dictionary and simple word-counting approaches in order to analyse the effect of communication on the wider economy, and to chow the effect of transparency on the behaviour of FOMC members. Baerg and Lowe (2020) use topics in FOMC minutes to estimate a "textual" Taylor Rule and Azqueta-Gavaldon et al. (2020) use topic analysis of news media to develop a economic policy uncertainty index for the euro area.

This paper uses the *New York Times* (NYT) as its sample of media articles for the US. In doing so it follows the political communication literature where the NYT is often used as a representative news outlet (Lee, 2014; Blood and Phillips, 1995; Brown et al., 1987; De Boef and Kellstedt, 2004; Sigal, 1973). Numerous studies have also found that the NYT tends to lead other media's news selection and stories (Gans, 2004; Bartels, 1996; Hess, 1981). This makes it an ideal sample for investigating potential transmission of focus from central bank communication to the media.

[2]The model is described in more detail in Section 4.3 and Appendix D.2.

4.3 Data and Topic Modelling

This Section describes the text data which is used to quantify central bank and media focus, as well as macroeconomic time series. It also briefly describes the Latent Dirichlet Allocation topic model which is used to quantify the focus of the text documents.

4.3.1 Text data

This paper uses four distinct corpora of text documents: three bodies of central bank communication text and news articles from the *New York Times* (NYT) newspaper.

1. Published minutes of the Federal Reserve's Open Market Committee (FOMC) from Jan 1993 to May 2014.
2. Published minutes of the Bank of England's Monetary Policy Committee (MPC) from Jan 1997 to May 2014.
3. The introductory press conference statement given after each meeting of the European Central Bank's Governing Council (GC) from June 1998 to May 2014.
4. Articles are taken from the print edition of the *New York Times* (NYT) between January 1993 to May 2014. These are downloaded from the Nexis service, and are restricted to articles tagged as "economic news".[3]

These central bank communication texts perform a similar function in that they both communicate the central bank's monetary policy decision as well as explaining their reasoning. Armelius et al. (2020) who show co-movement in the tone of communication across different central banks focus on speeches rather than minutes or press conferences. Extending the analysis of this paper to include speeches is certainly an interesting avenue for future work.

Each central bank document is associated with a "meeting date", the date at which the meeting took place, and a "publication date", the date at which the document was published. In the case of the ECB statements, these coincide as they are transcripts of a press conference given directly after the meeting. However in the case of the Federal Reserve and Bank of England minutes, the publication date can be up to 2 months after the meeting date. The Bank of England and ECB communication are only available over a shorter period than for the Federal Reserve. Thus when analysing the co-movement between the media and the Federal Reserve a longer sample is used than when comparing the co-movement across central banks.

[3]Restricting to "economic news" narrows the focus to articles which are relevant to the economy, which should reduce the measurement error in measuring the economically-relevant tone and topics of media coverage.

Some standard preparation and cleaning is performed on the text documents to facilitate the extraction of meaningful machine-readable information, details of which are given in Appendix D.1. Most notably, all names of months and seasons are removed as the obvious seasonality of these terms might generate spurious co-movement. Table 4.1 shows some summary statistics for the four corpora after this process. The central bank corpora are of a comparable size as well as having a similar purpose, so comparing their contents is reasonable. The size of the documents is also roughly constant over time.[4] The NYT corpus, on the other hand, is unsurprisingly much larger than that of the central banks. When identifying co-movement with FOMC communication, we therefore restrict the sample to windows around meetings and publication dates.

Table 4.1: Summary statistics for each corpus

Corpus	Total documents	Total paragraphs	Total words	Total vocabulary
Federal Reserve	171	5,944	459,706	6,439
Bank of England	209	6,925	420,344	3,888
European Central Bank	194	2,888	165,074	2,754
New York Times	72,763	NA	30,655,062	143,259

4.3.2 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a probabilistic topic model which assigns each word in a set of documents to one of a set number of topics. Although there are several forms of probabilistic topic model, they are characterised by the use of "statistical methods that analyze the words of the original texts to discover the themes that run through them" (Blei, 2012). They this allow the annotation of large corpora while avoiding the biases and time constraints of human evaluation. The seminal and best-known topic model is the LDA model introduced by Blei et al. (2003). The aim of the LDA model is to summarise and condense a large corpus of qualitative text data into a quantitative measure of content. In this sense it is analogous to principal components analysis, as it summarises a large data set in terms of a manageable number of topic proportions.

Topic models have key advantages over simpler content analysis methods such as raw word counts or dictionary methods. Dictionary methods define a list of key words that captures content of interest (for example, words relating uncertainty), allowing the rep-

[4]As shown in Appendix D.1, there does appear to be a countercyclical element to the number of "economic news articles" appearing in the NYT. This echoes the measure of uncertainty produced by Baker et al. (2016), as economic uncertainty is generally higher in recessions which will be reflected in increased coverage of economic issues and events.

resentation of documents as the (normalized) frequency of words in the dictionary. Dictionary methods are useful when we have a strong prior on which words are linked to the concept that we want to measure but, in the case of this paper which examines the co-movement in focus across corpora, a topic model is more appropriate.

The basis of the LDA model is a generative model of text, summarised below. It assigns each document a distribution over K distinct topics, where each of these topics is a distribution over all unique words that appear across all documents. The model therefore estimates a distribution over topics for each document (i.e. how prevalent each topic is in a given document) and a distribution of words for each topic (i.e. how prevalent each word is in a given topic).

Consider a corpus of D documents with a vocabulary of V unique terms. In the LDA framework, each *word* in a given document is assigned to one of K topics, with each of these topics being a probability vector $\beta_k \in \Delta^{V-1}$. The topics can thus be thought of as weighting vectors giving the importance of each term in the vocabulary in that topic. As topic assignment is done at the word level, a document can feature multiple topics and thus has a distribution over topics given by a probability vector $\theta_d \in \Delta^{D-1}$. The vectors θ_d and β_k are drawn from Dirichlet priors, and the topic assignment and realisation of each word are drawn from a multinomial parameterised by the relevant vector of θ and β respectively.[5] The number of topics K is chosen by the researcher. We present results for 30 topic models in this paper, but 20 and 40 topic models generated similar results. The full generative process for LDA is set out in Figure 4.1. The LDA model are typically by estimated either through MCMC sampling or variational inference. We use the collapsed Gibbs sampling algorithm presented by Griffiths and Steyvers (2004), described and derived in detail in Appendix D.2.

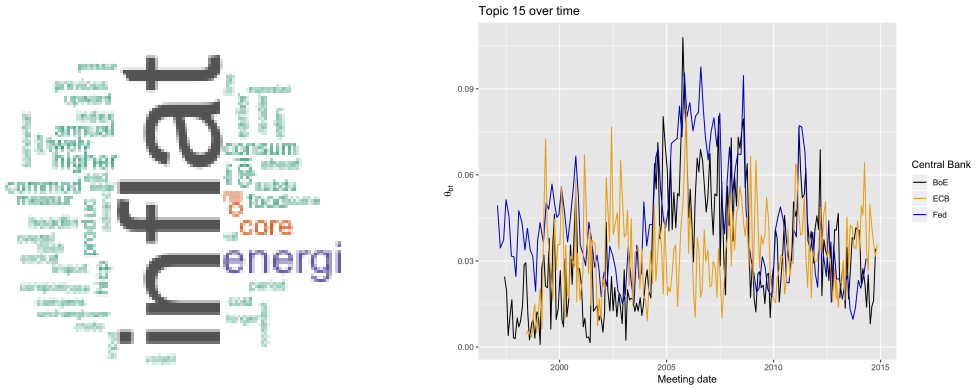
Figure 4.1: Generative model for LDA

1. For each of K topics, draw $\beta_k \sim \text{Dir}(\eta)$
2. For each of D documents, draw $\theta_d \sim \text{Dir}(\alpha)$
3. For each word n in document d:
   Draw topic assignment $z_{d,n}$ from $\text{Mult}(\theta_d)$
   Draw $w_{d,n}$ from $\text{Mult}()\beta_{z_{d,n}})$

[5]The prior parameters α and η determine how concentrated the topic definition and distribution vectors are. Throughout this paper we chose standard values proposed by Griffiths and Steyvers (2004) of $\alpha = 50/K$ and $\eta = 200/V$. Experimentation suggests that general results are robust to some variation in these priors.

Figure 4.2 gives an example of a topic which appears to deal with inflation, showing that the focus of central banks' communication on this topic varies over time. The β_k vector is represented as a word cloud in the left hand panel, and the it's prevalence in the corpora of the three central banks, as defined by θ, is shown in the right hand panel. The β parameters thus allow us to give an economic interpretation to each of the topics which are learned from the corpus, and the θ parameters give us an indication of how the prevalence of these topics varies over time and across corpora. These θ topic proportions will form the basis of the analysis of central banks' and the media's shifting focus and its co-movement throughout this paper.

Figure 4.2: LDA Topic 15 word cloud and topic proportion for central bank corpora



4.3.3 Macroeconomic and SPF data

As a proxy for private sector uncertainty about different aspects of the economy, we use dispersion on expectations in the Survey of Professional Forecasters (SPF) produced by the Federal Reserve Bank of Philadelphia (Croushore, 1993). The SPF provides measures of cross-sectional forecast dispersion for several economic series, measured as the difference between the 75th and 25th percentile of the individual forecasts. In order to keep the units of this dispersion measure fairly consistent across series, we use dispersion in predicted growth rates rather than dispersion in levels. The 25th and 75th percentiles for the growth rate of a range of economic variables are constructed by computing the growth rate for each panellist from her forecast for the level of the variable. The percentiles are then calculated on theses growth rates.[6] This of course does not apply for variables which are reported in growth rates, namely the CPI and PCE inflation rates, where the forecasts themselves are of the annualised growth rates.

[6]This differs from the method used to construct the survey's official measures for mean and median growth were the mean and median levels forecast is first calculated, and the growth rate is backed out from this.

Table 4.3 displays the economic time series for which the SPF dispersion measures are available and matches these to topics in the FOMC/NYT corpora. These cover the series which receive the majority of the attention of both academic economists, professional economists and the news media.

When measuring the co-movement in the content of central bank communication across different central banks in Section 4.6, we use CPI inflation, GDP growth and a policy rate series as a crude control for macroeconomic conditions in the three economies of interest. These macroeconomic time are taken from the St Louis FRED website, with the exception of the policy rates which are downloaded from the BIS website. The policy rate variable is included both as a level and change variable..

4.4 A Model of Central Bank Communication and Multi-Dimensional Uncertainty

This Section sets out a model to illustrate central banks how might shift the focus of their communication in response to multi-dimensional uncertainty about the current/short term state of the economy. The intuition behind this model is straightforward. The central bank has private information about various economic state variables, and wishes to improve the decision making of the private sector by sharing this information. However, the central bank is aware that it only has a limited capacity for communication so chooses the focus of its communication to provide information where it is most effective. The central bank's communication will therefore focus more on state variables about for which its signal is more informative relative to the public's, and for which it receives more extreme signals.

We will first set out the model in a case with a single central bank and show what affects the focus of central bank communication in this case. We then present an extension to a case with two central banks and show how the focus of their communication will co-move and may even be direct influence from one to the other. In this extension, the second central bank can also be thought of as the news media.

4.4.1 State Variables and Information Structure

The central bank communicates to the private sector about the fundamental states of the economy. This state is encapsulated in a fixed number state variables, which are

only observed with a lag. In Section 4.5 we will interpret them as different economic variables such as output, inflation and unemployment, but we can also think of them as the structural shocks driving business cycle fluctuations, as in Haldane et al. (2020). we assume N such state variables that evolve according to exogenous AR(1) processes.

$$X_{i,t} = \mu_1 + \rho_1 X_{i,t-1} + \epsilon_{i,t} \quad \text{where} \quad \epsilon_{i,t} \sim \mathcal{N}(0, \sigma_{\epsilon,i}) \quad \text{and} \quad i \in \{1, ..., N\} \tag{4.1}$$

The central bank observes last period's state variables and a *private* signal, $s_{i,t}$, for each of the shocks.

$$s_{i,t} = \epsilon_{i,t} + \nu_{i,t} \quad \text{where} \quad \nu_{i,t} \sim \mathcal{N}(0, \sigma_{\nu,i}) \tag{4.2}$$

The conditional distribution of the structural shocks, given the central bank's private signal is therefore

$$\begin{gathered} \epsilon_{i,t} | s_{i,t} \sim \mathcal{N}\left(\tilde{\mu}_{i,t}, \tilde{\Sigma}_{i,t}\right) \\ \text{where} \\ \tilde{\mu}_{i,t} = \frac{\sigma^2_{\epsilon,i}}{\sigma^2_{\epsilon,i} + \sigma^2_{\nu,i}} s_{i,t} \quad \text{and} \quad \tilde{\Sigma}_{i,t} = \sigma^2_{\epsilon,i}\left(1 - \frac{\sigma^2_{\epsilon,i}}{\sigma^2_{\epsilon,i} + \sigma^2_{\nu,i}}\right) \end{gathered} \tag{4.3}$$

The central bank has a finite volume communication each period, which we normalise to 1.[7] They can therefore control the proportion of this communication that is devoted to each of the state variables $1 \leq a_{i,t} \geq 0$. Through their communication, the central bank creates a public version of its conditional expectation of the signal that can be observed by the private sector

$$\hat{s}_{i,t} = s_{i,t} + (1 - a_{i,t})\eta_{i,t} \quad \text{where} \quad \eta_{i,t} \sim \mathcal{N}(0, \sigma_{\eta,i}) \tag{4.4}$$

Crucially, the extra noise with which the private sector observes the central bank's signal, depends on the communication's focus. In a two state variable case, the central bank can devote its focus to two topics. Intuitively, the more of its communication the central bank devotes to discussing and explaining its signal of state variable i, the greater the precision with which the private sector will observe the central bank's signal of $\epsilon_{i,t}$.

In addition to the central bank's communication (i.e. the public signal) the private sector observes the past values of the state variables, as well as the constant parameters μ_i and ρ_i. Given the system's linearity and the independent Gaussian shocks, their expectation

[7] This constraint could be in response to the limited amount of attention the private sector devotes to digesting central bank communication.

of $X_{i,t}$ is given by

$$\mathbb{E}_t[X_{i,t}|\hat{s}_{i,t}] = \mu_i + \rho_i X_{i,t-1} + \lambda_{i,t}\hat{s}_{i,t} \tag{4.5}$$

where

$$\lambda_{i,t} = \frac{\sigma^2_{\epsilon,i} - \tilde{\Sigma}_{i,t}}{\sigma^2_{\epsilon,i} - \tilde{\Sigma}_{i,t} + (1-a_{i,t})^2\sigma^2_{\eta,i}} \tag{4.6}$$

The private sector thus weights the public signal more heavily if the central bank increases its precision by focusing on it in communication. Therefore, $\lambda_{i,t}$ will vary over time. It will also weight signals more heavily if the underlying shock has a higher variance (higher $\sigma^2_{\epsilon,i}$), if the central bank's own signal is more accurate (lower $\tilde{\Sigma}_{i,t}$), and if the public signal is less noisy (lower $\sigma^2_{\eta,i}$).

4.4.2 Central Bank Problem

The central bank's objective is minimise the mean squared error of the private sector's expectations, given their limited communication capacity. Note that while the private conditions on the public signal $\hat{s}_{i,t}$, while the central bank conditions on their, more accurate, private signal $s_{i,t}$.

$$L(\mathbf{a}_t, \mathbf{s}_t) = \mathbb{E}\left[\sum_i \left(\mathbb{E}[X_{i,t}|\hat{s}_{i,t}] - X_{i,t}\right)^2 |s_{i,t}\right] = \mathbb{E}\left[\sum_i \left(\lambda_{i,t}\hat{s}_{i,t} - \epsilon_{i,t}\right)^2 |s_{i,t}\right] \tag{4.7}$$

In allocating the focus of their communication, we can think of the central bank as making its own signals observable to the private sector with differing levels of precision for the different state variables. Given that $\mathbf{s}_t$ is observed and the shocks are all independent and zero mean, we can use Equations (4.1-4.4) write the central bank's loss function in terms of known parameters and variables (see Appendix D.3.1).

$$\begin{gathered}
\min_{\mathbf{a}_t} L(\mathbf{a}_t, \mathbf{s}_t) = \sum_i \left(\left(\lambda_{i,t} - \frac{\sigma^2_{\epsilon,i}}{\sigma^2_{\epsilon,i} + \sigma^2_{\nu,i}} \right)^2 s^2_{i,t} + \lambda^2_{i,t}(1-a_{i,t})^2\sigma^2_{\nu,i} + \sigma^2_{\epsilon,i}\left(1 - \frac{\sigma^2_{\epsilon,i}}{\sigma^2_{\epsilon,i} + \sigma^2_{\nu,i}}\right)\right) \\
\text{s.t.} \\
\sum_i a_{i,t} = 1, \\
a_{i,t} \geq 0, \quad \text{for} \quad i \in \{1, ..., N\} \\
\lambda_{i,t} = \frac{\sigma^2_{\epsilon,i}}{\sigma^2_{\epsilon,i} + \sigma^2_{\nu,i} + (1-a_{i,t})^2\sigma^2_{\eta,i}}, \quad \text{for} \quad i \in \{1, ..., N\}.
\end{gathered} \tag{4.8}$$

The relationship between $\lambda_{i,t}$ and $a_{i,t}$ captures that central bank communication focus $a_{i,t}$ changes the private sector's expectation in two distinct but complementary ways. Firstly, it makes the private sector's signal $\hat{s}_{i,t}$ more accurate by reducing impact of the

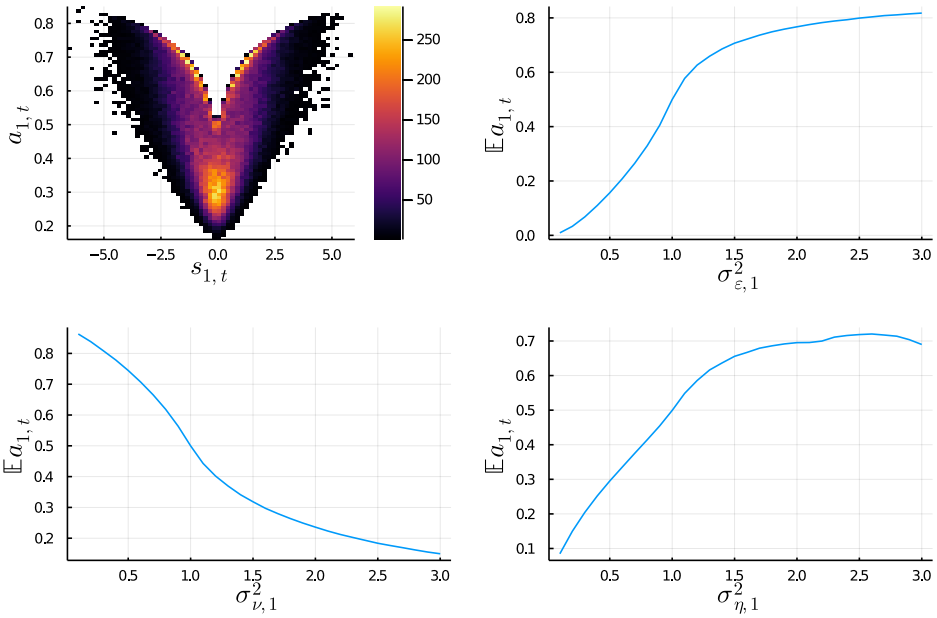
noise term $\eta_{i,t}$. Secondly, it increases the weight that the private sector places on this signal, as its precision has increased.

4.4.3 Determinants of Central Bank Focus

Even in the two state case, the solution to this constrained optimization problem is a ninth-order polynomial and so cannot be expressed in closed form. However, it can be solved numerically for given values of $\mathbf{s}_t$ and variance terms. For a given value of the variance parameters, we can therefore approximate the moments of $\mathbf{s}_t$ and $\mathbf{a}_t$ through Monte-Carlo simulation. To assess the effect of the variance parameters themselves, we set these to default values and then vary one, again approximating the distributions by Monte Carlo simulation. More details on the solution and simulation algorithm are given in Appendix D.3.3.

We can thus investigate how the central bank's focus in a two-state case depends on the signals they receive and the variance parameters. As a default, all variance terms are set to 1 and all covariances to 0. Figure 4.4 shows how focus on the first state variable, $a_{1,t}$ varies along each of four dimensions. In general, these results support the intuition that the central bank should focus their communication where it will be most useful to the private sector.

Figure 4.4: Central bank focus in one bank case



In the upper left panel we see the distribution of $s_{1,t}$ and $a_{1,t}$ across 200,000 simulated

periods. The more extreme this signal is (i.e. the further from zero) the greater focus will be placed on this variable in communication. When the two signals are equal (at zero) attention is split evenly between the two (given that all variances are symmetric). As the signal moves further from this mean value, and because it is unbiased, the central bank will focus more on the corresponding state variable. Intuitively, the more unexpected the information a central bank has about a given variable is, the more effort we would expect them to devote to communicating this information.

The upper right panel shows the effect of $\sigma^2_{\epsilon,1}$, the variance of the first state variable, on the expectation of $a_{1,t}$ across 200,000 simulated periods. The more volatile a state variable is, the greater focus will, in expectation, be devoted to it in communication. Again, when the variance terms of the two state variables are equal (at one), attention will on average be split evenly between the two.

Similarly, the bottom two panels show the effect of $\sigma^2_{\nu,1}$ (the variance of noise in the central bank's private signal) and $\sigma^2_{\eta,1}$ (the variance of noise the their public signal) on the expectation of $a_{1,t}$. The less noisy the private signal for the first state variable is, the greater focus should be devoted to it. In contrast, the noisier the public signal is, the greater focus will be devoted to it. This is because the communication can reduce the importance of this noise. Intuitively, if a signal is harder to explain, a greater proportion of communication should be devoted to explaining it.

The central intuition that this model illustrates is thus that a central bank's communication will focus more on the variables for which it has recently received a more extreme signal, and on variables for which its signals are more informative.

In Section 4.5, we use the dispersion of private sector forecasts provided by the Survey of Professional forecasters as a proxy for how informative central bank communication can be for that variable. We then show that the focus on a variable in the FOMC's minutes co-moves with its forecast dispersion. Further work could explore this relationship in more detail, perhaps using the distance between the Federal Reserve Greenbook projections and the SPF values as a potential test of the different channels.

4.4.4 Extension to two central banks

Extending the model to include two central banks, we can illustrate that if central banks can observe one another's communication and the shocks hitting their economies are correlated, we would not only expect their focus to co-move, but also that they might have

a direct influence on each other. Furthermore, we can straightforwardly interpret one of the central banks in this model as the news media and therefore think of the results in terms of co-movement in the focus of central bank communication and news media.

Assume that there are now two central banks, b and c. Both banks can communicate with their domestic private sectors about N state variables. These state variables are distinct to each economy, but the shocks hitting them may be correlated.

$$\begin{pmatrix} \epsilon_{b,i,t} \\ \epsilon_{c,i,t} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{b,\epsilon,i} & \sigma_{bc,i} \\ \sigma_{bc,i} & \sigma^2_{c,\epsilon,i} \end{pmatrix} \right) \quad \text{and} \quad i \in \{1, ..., N\} \tag{4.9}$$

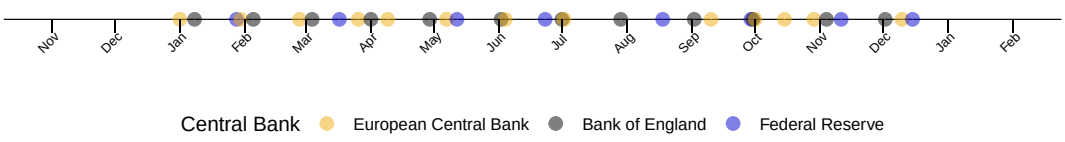
In addition to observing its own private signal (as in Eq. 4.2), central bank b also observes the public signal of central bank c.

$$\hat{s}_{c,i,t} = \epsilon_{c,i,t} + \nu_{c,i,t} + (1 - a_{c,i,t})\eta_{c,i,t} \tag{4.10}$$

where $a^c_{i,t}$ is the proportion of its communication which central bank c devotes to $X^c_{i,t}$.

If central banks produce multiple pieces of communication in a given period, or at least that the publication schedule is not lined up, central bank b may be able to observe $\hat{s}_{c,i,t}$, the public signal communication of central bank c earlier in period t, when setting $a_{b,i,t'}$ (where t' is a time within the same period as t but slightly later). Figure 4.5 shows the distribution of meetings, all of which are accompanied by communication, throughout a representative year. Each central bank has multiple meetings in each quarter, and the order in which they occur is not necessarily consistent. As communication within a period is sequential and minutes are published with a lag, c's communication at t will not be influenced by b's later communication at t', and we will assume that central banks do not account for the fact that they influence one another.[8]

Figure 4.5: Central Bank Policy Committee meetings throughout the year 2000



[8]If we were to allow for this, it might introduce a strategic incentive whereby one central bank may deliberately increase its focus on a particular state variable in order to induce a different central bank to also increase its focus on that variable.

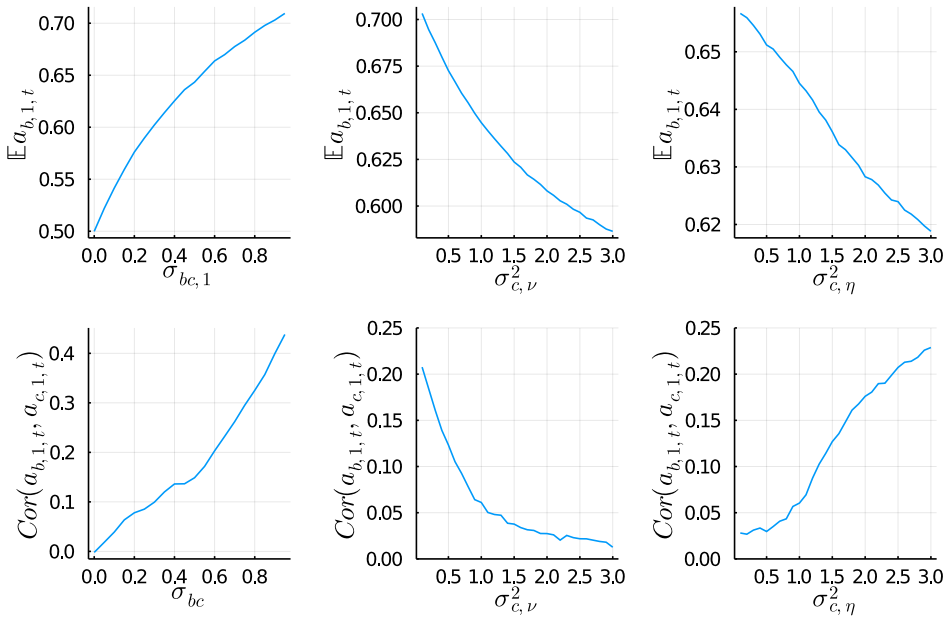
If $\sigma_{bc,i} \neq 0$, then $\hat{s}_{c,i,t}$ includes useful information for central bank b. We can therefore define the conditional distribution of $\epsilon_{b,i,t}$ given $s_{b,i,t}$ and $\hat{s}_{c,i,t}$. Given this we can define the central bank's problem as in the one-bank case. As central bank b's communication is published at t', after that of central bank c, they can include information from c's communication in their own. The public signal of central bank b at t' is therefore

$$\hat{s}_{b,i,t'} = \Sigma_{12}(b,i,t')\Sigma_{22}(b,i,t')^{-1} \begin{pmatrix} s_{b,i,t} + (1 - a_{b,i,t'})\eta_{b,i,t} \\ \hat{s}_{c,i,t} + (1 - a_{b,i,t'})\eta_{b,i,t} \end{pmatrix} \tag{4.11}$$

where $\Sigma_{12}(b,i,t')$ and $\Sigma_{22}(b,i,t')$ are derived and set out in Appendix D.3.2, along with a full statement of the central banks' problems in the two-bank case.

Figure 4.6 shows the effect of variance parameters for central bank c on the communication of central bank b. In each case, we examine the effect of varying the parameter on the expected value of $a_{b,1,t}$ and the correlation between $a_{c,1,t}$ based on a simulation of 200,000 periods. All variance parameters that are not being varied are set to a default of one, and $\sigma_{bc,1}$ (the covariance between shocks to the first state variable across the two economies) is set to a default value of 0.5. This ensures that there is one state variable for which central bank c's public signal includes information that is useful to central bank b. In each case, we assume that central bank c communicate first and central bank b decide on their communication having observed c's.

Figure 4.6: Central bank focus in two bank case



The two left-most panels show the effect of $\sigma_{bc,1}$, the covariance between the shock to the first state variable in each economy. We see that increasing this covariance both increases the proportion of communication that b will devote to the first variable, and increases the correlation between $a_{b,1,t}$ and $a_{c,1,t}$. The increase in proportion of communication devoted to the first variable is due to b now effectively having a more accurate signal for this shock (so intuitively similar to varying $\sigma^2_{b,\nu,1}$). The increase in co-movement of communication is due to both the correlation of the two shocks, but also because b is directly learning from c.

The two central panels show the effect of $\sigma^2_{c,\nu}$, the variance of noise in central bank c's *private* signal.[9] As c's private signal becomes less reliable, and therefore their public signal also gets less reliable, b will place less weight on it and so both devotes a smaller proportion of it's communication to the first state variable and their focus will co-move less with that of c. It is important to note that in this example $\sigma_{bc,1}$ is held constant at 0.5

The two right-most panels show the effect of $\sigma^2_{c,\eta}$, the variance of the additional noise in central bank c's *public* signal. As with the $\sigma^2_{c,\nu,2}$ case, increasing the noise of c's signal will decrease the proportion of b's communication devoted to the first state variable, as the first is the only one correlated with b's economy. The relationship with co-movement is somewhat less intuitively obvious. As the noise in c's public signal scales with the focus of their communication, the greater $\sigma^2_{c,\eta}$, the greater an impact an increase in $a_{c,1,t}$ has on the informativeness of $\hat{s}_{c,i,t}$ for b. In other words, if c's public signal is very hard to interpret unless it has been explained in detail, b will only take this signal into account if c did indeed explain it in great detail. This increases the degree of co-movement as b reacts not only to the signal it observes from c, but also to how much c chose to focus on each variable.

In the Section 4.6, we will show that the focus of the Federal Reserve's communication leads that of the Bank of England and the European Central Bank, and we use a change in publication policy to show that there may be a direct influence. The discussion above suggests that three potential explanations for this, each of which are intuitively appealing.

1. The shocks concerning the Federal Reserve are in general of greater global relevance.
2. The Federal Reserve's signals are more precise than those of other central banks, i.e. it has a smaller σ^2_{ν}.
3. The Federal Reserve's signal are hard to interpret unless they are explained in greater detail, i.e. it has a larger σ^2_{η}.

[9] In this and the $\sigma^2_{c,\eta}$ example, we vary $\sigma^2_{c,\nu,1}$ and $\sigma^2_{c,\nu,2}$ together so that they are always equal.

Further work could aim to identify some robust testable implications of each channel which could be brought to the data. However, it is nevertheless useful to conceptualise why central bank communication might co-move, before we go on to show that it does.

4.5 Central Bank Communication, Private Sector Forecasts and the Media

This Section investigates empirically how the focus of central bank communication news media and private sector forecast dispersion co-move over time. The remainder of the Section is structured as follows. Section 4.5.1. Section 4.5.2 then provides some suggestive empirical evidence for this model of multi-dimensional uncertainty by showing that the FOMC minutes (and to a lesser extent the NYT articles) focus more on variables around which there is greater forecast uncertainty. Finally, Section 4.5.3 formally shows that the focus of FOMC minutes and NYT articles co-moves and presents an event study which shows that the publication of FOMC minutes predicts articles in the following week, even when articles in the week prior to publication are controlled for. This suggests that central bank communication can influence the focus of the media, giving support to the model in Section 4.4 as central banks do not have to rely on agents directly reading their communication.

4.5.1 Measuring Central Bank and Media Focus

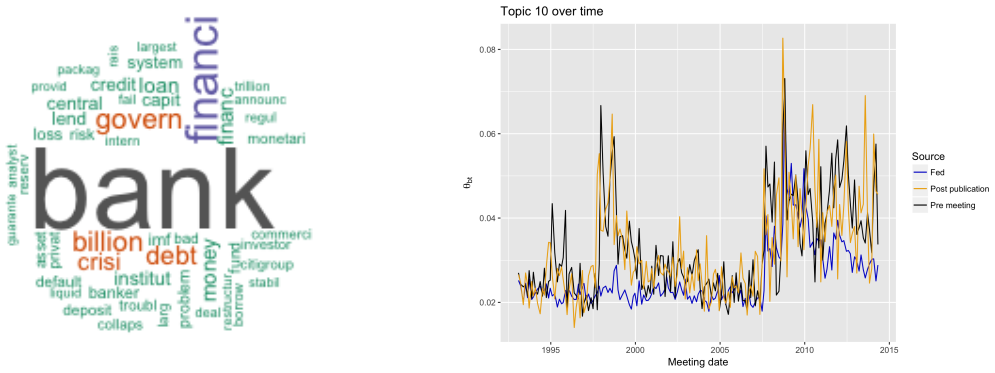
We estimate a 30 topic LDA model over the combined FOMC minutes and NYT articles corpora. To reduce the size and dimensionality of the NYT corpus relative to the FOMC corpus, we use only the subset of articles that are published in the weeks either side of an FOMC meeting and the weeks either sides of the publication of those minutes, which selects 36,983 of the 72,763 articles. This subsample will allow the exploration of co-movement in the focus of the two corpora over time. Further details on the estimation with this corpus are given in Appendix D.1.1.

This thus constructs five sets of 30 time series covering the sample period of January 1993 through to May 2014, one for the FOMC minutes and one for the NYT articles. For now, we will work with each of these aggregated to a quarterly frequency, with FOMC minutes assigned to the quarter of their *meeting* date, rather than their publication.[10]

[10]In Section 4.5.3 we will decompose the media articles into separate series for the weeks either side of the meeting date the weeks either side of the publication date. This will allow the event study approach we will take.

We verify that these series are stationary with Augmented Dickey-Fuller tests reported in Appendix D.5. Figure 4.7 shows the β vector for topic 10 as a word cloud in the left panel, and the proportion of each corpus devoted to this topic over time in the right panel. The word cloud shows us that this is a topic dealing with financial crises, and we can see that in both the NYT and FOMC corpora, there is more focus on this topic in the wake of the 2008 financial crisis.

Figure 4.7: LDA Topic 10 word cloud and topic proportion for FOMC-NYT corpus



Note: The black and yellow lines represent the proportion in the NYT articles around the meeting date and the publication date of an FOMC meeting respectively, while the blue line represents the proportion the FOMC minutes.

These topic proportions capture the shifting focus of the FOMC and the media, and will be the subject of the remainder of this Section. Table 4.2 summarises the topics learned, in most cases an interpretation for the topic is clear from just the 5 most highly weighted words in the corresponding β vector. It also reports the average proportion of each meeting period for both the NYT articles and the FOMC minutes, across the sample period. We see that in most cases the average proportion is similar across the two corpora, suggesting that they are sufficiently similar to make comparisons meaningful. The two exceptions are Topic 2 and Topic 30, which can be interpreted as discussion of economic data and the policy committee. Unsurprisingly, these two topics appear much more heavily in the FOMC minutes, as they correspond to relatively formulaic passages that set out the latest key economic indicators and describe the policy decision. We account for this in what follows by both de-meaning these series and including fixed effects that account for the fact that some topics will be more prominent in some corpora compared to others.

Table 4.2: Topics estimated on the NYT & Fed documents and their average in each

Topic	Description	Top 5 words	$\bar{\theta_k}^{Fed}$	$\bar{\theta_k}^{NYT}$
Topic 1	Foreign policy	nation, unit, stat, american, russia	0.0191	0.0318
Topic 2	Economic data I	quarter, growth, spend, continu, busi	0.1384	0.0227
Topic 3	Fiscal policy	tax, budget, cut, spend, billion	0.0228	0.0348
Topic 4	Explanations	seem, problem, good, differ, fact	0.0230	0.0397
Topic 5	US domestic politics	presid, administr, hous, senat, congress	0.0205	0.0362
Topic 6	International trade I	dollar, trade, japan, currenc, foreign	0.0327	0.0277
Topic 7	Bond markets	bond, yield, rate, treasuri, market	0.0295	0.0329
Topic 8	Economic data II	percent, report, month, rose, show	0.0286	0.0445
Topic 9	Inflation	price, inflat, oil, inceas, energi	0.0493	0.0278
Topic 10	Financial crisis	bank, financi, govern, debt, billion	0.0261	0.0326
Topic 11	Regulation	law, agenc, case, rule, regul	0.0245	0.0335
Topic 12	Public goods	school, health, people, care, educ	0.0198	0.0342
Topic 13	Europe	european, europ, countri, euro, germani	0.0214	0.0308
Topic 14	Elections	govern, polit, elect, power, leader	0.0199	0.0391
Topic 15	Investment	fund, invest, market, stock, investor	0.0269	0.0314
Topic 16	Labour market	job, worker, unemploy, labor, employ	0.0273	0.0293
Topic 17	States	state, said, million, york citi	0.0193	0.0311
Topic 18	Stock market	stock, percent, point, market, index	0.0255	0.0448
Topic 19	Retail sales	percent, sale, quarter, said, retail	0.0257	0.0370
Topic 20	Economic growth	economi, growth, recess, recoveri, grow	0.0313	0.0336
Topic 21	International trade II	china, countri, state, world, unit	0.0213	0.0330
Topic 22	Housing market	house, home, said, market, real	0.0284	0.0297
Topic 23	Business	compani, busi, execut, corpor, said	0.0202	0.0362
Topic 24	Interest rates	rate, fed, federal, interest, reserve	0.0349	0.0341
Topic 25	Infrastructure/construction	citi, develop, build, project, plan	0.0207	0.0373
Topic 26	Mortgage market	rate, mortgag, loan, credit, interest	0.0278	0.0285
Topic 27	Quotations	said, offici, meet, week, announc	0.0224	0.0362
Topic 28	Lifestyle	people, said, day, work, live	0.0189	0.0420
Topic 29	Industrial production	industri, product, manufactur, car, factori	0.0260	0.0283
Topic 30	Monetary policy committee	committee, polici, member, inflat, expect	0.1481	0.0193

An active point of the monetary policy literature is whether central banks have an informational advantage over the private sector (Romer and Romer, 2000; Reis, 2013; Nakamura and Steinsson, 2018; Miranda-Agrippino and Ricco, 2021). Although this paper takes this informational asymmetry for granted, we can compare the informational content of the FOMC minutes and the NYT articles to provide some support for this assumption. Appendix D.4 does this, showing that the FOMC minutes contain information that could improve private sector forecats of GDP growth and that this information is not simply contained in media coverage.

4.5.2 Central Bank Focus and Private Sector Forecast Dispersion

In order to test whether there is an empirical link between the focus of central bank communication and this multi-dimensional uncertainty over different aspects of the economy, we show that topic proportions in both the media and central bank communication co-move with the dispersion of private sector expectations of corresponding economic variables. We then create a panel data set to explore this relationship systematically.

The SPF provides measures of cross-sectional dispersion for several economic series, measured as the difference between the 75th and 25th percentile of the individual forecasts. Most of these economic series can quite naturally be matched to topics in the media/minutes corpus. The attention devoted to these topics moves together with the dispersion of the corresponding SPF forecasts in many cases. Crucially, the topic proportions are more correlated with the forecast dispersions of the variable they are related to, than to the dispersion of forecasts for other variables. As the SPF dispersion measures are expressed in the units of the series, we normalise them to have zero mean and unit variance.

The SPF dispersion series are matched to topics which most closely relate to the economic variable of interest, as set out in Table 4.3. This involves a certain amount of subjective judgement, so the 5 most highly weighted words from each topic are shown here and the word clouds for each topic are shown in Appendix D.6. The Table also reports the (Pearson's product-moment) correlation of this dispersion series and the proportions of the corresponding topic in the FOMC minutes and the NYT articles. In many cases the dispersion of the SPF forecasts are positively and significantly correlated with the corresponding topic in both the NYT articles and the FOMC minutes.

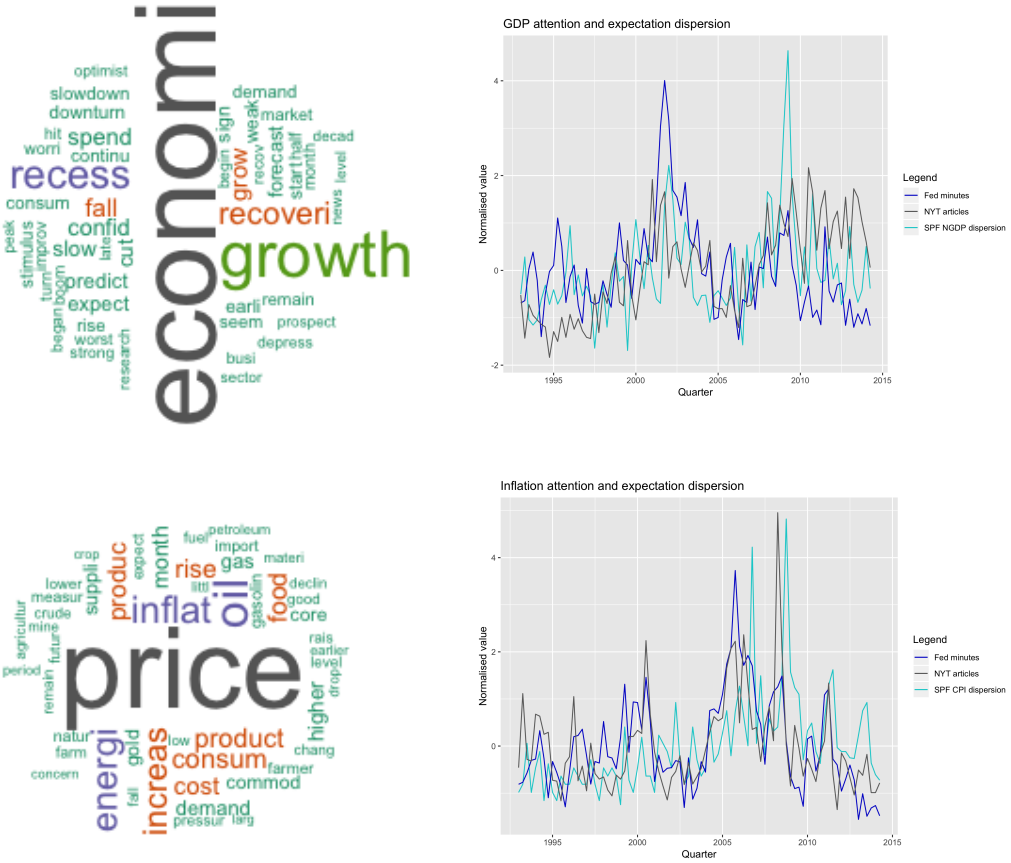
Table 4.3: Series covered by SPF and corresponding topic

SPF series	Description	Topic	Top 5 words	Fed/SPF	News/SPF
NGDP	Nominal GDP growth	20	economi, growth, recess, recoveri, grow	0.332***	0.293***
RGDP	Real GDP growth	20	economi, growth, recess, recoveri, grow	0.360***	0.278***
CPI	CPI inflation	9	price, inflat, oil, increas, energi	0.305***	0.211*
TBILL	3-month Treasury Bill rate	24	rate, fed, federal, interest, reserve	0.171	0.293***
EMP	Nonfarm Employment	16	job, worker, unemploy, labor, employ	0.363**	0.410***
UNEMP	Unemployment Rate	16	job, worker, unemploy, labor, employ	0.105	0.331***
CPROF	Corporate Profits (after tax)	23	compani, busi, execut, corpor, said	0.273**	0.012
INDPROD	Industrial Production Index	29	industri, product, manufactur, car, factori	0.089	0.273**
HOUSING	Housing starts	26	mortgag, loan, rate, credit, interest	0.625***	0.273**
RRESINV	Residential Investment	22	house, home, said, market, real	0.340***	-0.012
RCONSUM	Personal Consumption Exp	19	percent, sale, quarter, said, retail	-0.119	0.275**
RNRESIN	Nonresidential Investment	15	fund, invest, market, stock, investor	0.337***	0.017
RFEDGOV	Federal Government Exp	3	tax, budget, cut, spend, billion	0.281***	0.070
RSLGOV	State Government Exp	3	tax, budget, cut, spend, billion	0.124	0.246**

Note: *p<0.1; **p<0.05; ***p<0.01

To illustrate this co-movement we use the series for nominal GDP growth and CPI inflation. Figure 4.9 plots the quarterly θ topic proportions for topics 20 and 9, which are respectively interpreted as economic growth and inflation topics, alongside the dispersion of the SPF forecasts for that quarter's nominal GDP growth and CPI inflation. Analogous Figures for the other SPF series and corresponding topics are shown in Appendix D.6.

Figure 4.9: Inflation and nominal GDP topics



As reported in Table 4.4, the SPF dispersion for both inflation and GDP growth is significantly and positively correlated with the corresponding topic proportion in the FOMC minutes (0.305 for inflation and 0.332 for GDP growth). Perhaps even more importantly, the SPF dispersion for inflation is not significantly correlated with the FOMC growth topic (-0.007, with p-value 0.951) and neither is the SPF dispersion for GDP growth significantly correlated with the FOMC inflation topic (-0.136, with p-value 0.212). This provides supporting evidence for the theory that central bank's tailor their communication to provide more information about state variables around which there is more uncertainty. The central bank allocates the limited space in its published minutes to reflect where private sector uncertainty is greatest, as this is where it can have the greatest benefit.

The correlation of the SPF dispersion with their related topics is systematically higher than the correlation with other topics, reflecting the trade-off that comes with the central bank's limited communication capacity. A full tables of cross-correlations of the SPF dispersion series and the 11 topics which are linked to them, in both the FOMC and

NYT documents, are presented in Appendix D.6. The average raw correlation of an SPF dispersion series with its corresponding topic in the FOMC corpus is 0.256, while the average raw correlation of the SPF series with the other topics is 0.049. As also reported in Table 4.3, the topics in the NYT articles also appear to be correlated with the SPF series in a similar manner to the FOMC minutes. However, the link appears to be weaker for the NYT articles than for the FOMC minutes, with an average correlation of 0.212. This is somewhat surprising as we might expect news to be less formulaic than central bank communication. However, even when restricted to economics, news coverage likely contains more noise than central bank communication.

In order to test the relationship between these quarterly series for media/central bank focus and private sector forecasts in a more systematic fashion, we construct a panel data set from the 14 cases set out in Table 4.3. For each of these 14 individuals we thus have three variables: a quarterly topic proportion in the FOMC minutes ($\theta^{\text{Fed}}_{k,t}$), a quarterly topic proportion in the the NYT articles ($\theta^{\text{NYT}}_{k,t}$) and the SPF forecast dispersion for the associated economic variable ($\text{disp}^{\text{SPF}}_{k,t}$). We standardise all variables so that they have a zero mean and unit variance. This allows direct comparison of the coefficients, in particular across the FOMC minutes and NYT articles. Results for the unstandardised variables are reported in Appendix D.7, the signs and significance of the coefficients are similar, but the interpretation of their magnitudes is unclear.

We then show the co-movement in these quarterly series by regressing each of the three on contemporaneous and past values of the other two, as well as its own lags and fixed effects. For example, for the SPF forecast dispersion the regression would be of the following form.

$$\text{disp}^{\text{SPF}}_{k,t} = \alpha_k + \mu_t + \sum_{q=0}^{Q} \left[\beta_{\text{Fed},q}\theta^{\text{Fed}}_{k,t-q} + \beta_{\text{NYT},q}\theta^{\text{NYT}}_{k,t-q}\right] + \sum_{p=1}^{P} \rho_p \, \text{disp}^{\text{SPF}}_{k,t-p} + \varepsilon_{k,t} \qquad (4.12)$$

Analogous regressions are also estimated with $\theta^{\text{Fed}}_{k,t-q}$ and $\theta^{\text{NYT}}_{k,t-q}$ as the dependent variables. Table 4.4 shows results with two specifications for each of the three variables. Columns 1, 3 and 5 show a simpler specification without a time fixed effect, no lags ($P = 0$ and $Q = 0$). Columns 2, 4 and 6 show a fuller specification with two-way fixed effects and lags ($P = 3$ and $Q = 1$).

Table 4.4: Federal Reserve minutes, NYT articles and SPF forecast dispersion

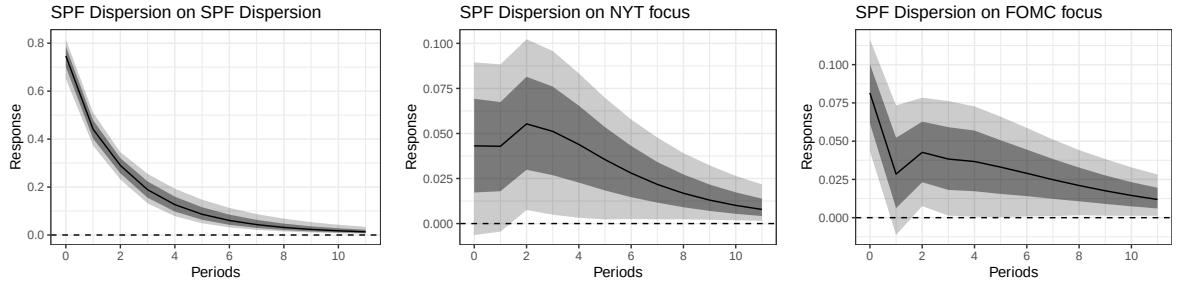
	Dependent variable:					
	$\text{disp}_{k,t}^{\text{SPF}}$		$\theta_{k,t}^{\text{NYT}}$		$\theta_{k,t}^{\text{Fed}}$	
	(1)	(2)	(3)	(4)	(5)	(6)
$\theta_{k,t}^{\text{Fed}}$	0.198***	0.118***	0.218***	0.134***		
	(0.030)	(0.032)	(0.030)	(0.035)		
$\theta_{k,t-1}^{\text{Fed}}$		0.021		−0.031		
		(0.032)		(0.035)		
$\theta_{k,t}^{\text{NYT}}$	0.169***	0.028			0.218***	0.127***
	(0.030)	(0.029)			(0.030)	(0.028)
$\theta_{k,t-1}^{\text{NYT}}$		0.115***				0.070**
		(0.029)				(0.029)
$\text{disp}_{k,t}^{\text{SPF}}$			0.171***	0.036	0.200***	0.112***
			(0.030)	(0.035)	(0.030)	(0.031)
$\text{disp}_{k,t-1}^{\text{SPF}}$				−0.014		−0.058*
				(0.035)		(0.031)
Dep variable lags		✓		✓		✓
Topic fixed effects	✓	✓	✓	✓	✓	✓
Time fixed effects		✓		✓		✓
Observations	1,105	1,063	1,105	1,065	1,105	1,065
R^2	0.089	0.429	0.099	0.320	0.112	0.467
Adjusted R^2	0.077	0.371	0.087	0.251	0.100	0.413
Residual Std. Error	0.953	0.793	0.958	0.874	0.957	0.773
Note:					*p<0.1; **p<0.05; ***p<0.01	

An alternative, but complimentary, approach to the separate regressions shown in Table 4.4 is a panel VAR which treats the three variables as a system of simultaneous equations rather than analysing each in isolation.[11] We estimate a panel VAR with two lags of each variable and identify impulse response functions with a Cholesky decomposition.

[11] A panel VAR extends the basic VAR framework to include a cross-sectional dimension, see Canova and Ciccarelli (2013) for a survey of theory and application of these models.

The ordering of variables is SPF dispersion, NYT articles and then FOMC minutes. We therefore assume that the SPF dispersion contemporaneously affects the two other series, but is itself only affected with a lag. Figure 4.11 shows the impulse response functions to a shock to SPF dispersion, which significantly increases both the focus of both the NYT and FOMC on the corresponding topic.[12]

Figure 4.11: Impulse Response Functions to SPF dispersion shock



Note: Impulse Response Functions from a panel VAR with 2 lags and Cholesky decomposition. Confidence Intervals are bootstrapped by 5,000 iterations of blockwise sampling of individuals. The darker band represents the 70% confidence interval and the lighter the 95% confidence interval.

In Table 4.4 and Figure 4.11, we see strong and robust evidence of co-movement across the three variables. While we cannot necessarily make any claims about causation, we can conclude that these co-movement patterns support the idea that central bank focus and multi-dimensional uncertainty are linked. This relationship also appears to be stronger for the FOMC minutes than for the NYT articles.

As the SPF is a slow-moving variable only available at a quarterly frequency, and it is difficult to know precisely when participants make their forecasts. An event study style identification of the effect of this dispersion on communication or vice versa is therefore difficult. In any case, we think of the SPF more as an indicator of where uncertainty is greatest, rather than a variable that directly impacts communication and media coverage. However, media coverage appears continuously over time so we can take advantage of this to identify a causal effect of FOMC communication on media coverage. This is the focus of the remainder of this Section.

[12]The complete IRFs and results with alternative specifications and identification strategies are shown in Appendix D.8

4.5.3 Central Bank Influence on News Media

The explanation for the shifting focus of central bank communication set out above rests on the central intuition that central banks can and do use their communication to combat uncertainty in multiple dimensions of the economy. We now show that the focus of central bank communication not only co-moves with the focus of the media, but also may have a direct influence on it. This is relevant for two reasons. Firstly, it shows that the central bank may be able to provide information to agents in the economy even if they are not directly exposed to its communication, by influencing the focus of the media. This is particularly important given the literature discussed in Section 4.2 which suggests that, beyond financial market participants, agents have little to no direct exposure to the central bank. Secondly, it provides further evidence to support the intuition behind the model of Section 4.4. The media's editorial decisions are made, in part to provide their readers with information which they find more useful. The media's incentives are thus similar to those of the central bank, so we would expect to see similar same co-movement patterns.

In order to identify a potentially causal effect of FOMC communication on the focus of media coverage, we use one week windows around the meeting and publication dates for each set of minutes. We thus separate out the NYT topic proportions into four separate time series. For a given meeting, each topic will therefore have five series: a proportion of the minutes, proportions of the media articles for the weeks either side of the meeting date and proportions of media articles for the weeks either side of the publication date. We standardise all series so that they have zero mean and unit variance.[13]

A given meeting indexed by meeting date m and the publication date of its minuts m_p. In a slight abuse of notation, we will refer to the weeks before and after the meeting date as $m-w$ and $m+w$ respectively. Similarly, the weeks either side of the publication date will be referred to as $m_p - w$ and $m_p + w$. For each meeting there are two variables of interest:

1. $\theta^{\text{NYT}}_{m-w,k}$, the proportions of topic k in NYT articles in the week prior to the meeting.
2. $\Delta\theta^{\text{NYT}}_{m_p,k} = \theta^{\text{NYT}}_{m_p+w,k} - \theta^{\text{NYT}}_{m_p-w,k}$, the change proportions of topic k in NYT articles in the from the week prior to the week after the publication of the minutes.

Note that the m does not line up neatly to months or quarters, given the irregular meeting schedule. However, the time periods between the θ^{Fed} and θ^{NYT} variables for a given m are consistent. Furthermore, the m subscripts are sequential, so we can include lagged values of, for example, the topic proportions in the minutes to account for persistence in

[13] Once again, results with the unstandardised series and shown in Appendix D.7.

the communication over time.

This implies two questions: whether media articles the week prior to a meeting predicts the content of minutes, and whether the content of minutes predicts the content of media articles in the week following the publication of the minutes.

To address the first of these questions, whether media articles predict minutes, we estimate panel regressions as shown in Equation 4.13.

$$\theta^{\text{Fed}}_{m,k} = \alpha_k + \beta\theta^{\text{NYT}}_{m-w,k} + \sum_{q=1}^{Q} \rho_q \theta^{Fed}_{m-q,k} + \text{further controls} + \varepsilon_{m,k} \tag{4.13}$$

Columns 1 and 2 in Table 4.5 show results for two specifications of this regression. We see that the focus of media coverage in the week prior to an FOMC meeting predicts the focus of the minutes. This effect survives controlling for persistence in the contents of minutes. Whether this is a causal effect is unclear, as it is possible both the minutes and media coverage are driven by the same underlying factors. However, in our answer to the second question we can be more confident that an identified effect could be causal.

To address the second of these questions, whether central bank communication influences media coverage, we estimate panel regressions as shown in Equation 4.14.

$$\Delta\theta^{\text{NYT}}_{m_p,k} = \alpha_k + \beta\theta^{\text{Fed}}_{m,k} + \sum_{q=1}^{Q} \rho_q \Delta\theta^{\text{NYT}}_{m_p,k} + \text{further controls} + \varepsilon_{m,k} \tag{4.14}$$

Columns 3 and 4 of Table 4.5 show the results: the focus on FOMC minutes significantly predicts the change in focus from the week prior to publication to the week after. This effect survives controlling for persistence in both media coverage and the minutes. Columns 5 and 6 show that this effect is also present on the level of focus post-publication.

This effect, of FOMC focus on the change in media coverage around publication, could be interpreted as causal. As the minutes are typically published around a month after the corresponding meeting, see Figure 4.19, an external factor driving both the minutes and media coverage would have to be both persistent and not captured by previous media coverage. By controlling for the contents of articles in the week prior to publication, and in the weeks around the meeting itself, we can therefore identify what is plausibly a causal effect of the publication of FOMC minutes on the topics covered by the NYT.

Table 4.5: NYT and FOMC results

	Dependent variable:					
	$\theta^{\text{Fed}}_{m,k}$		$\Delta\theta^{\text{NYT}}_{m_p,k}$		$\theta^{\text{NYT}}_{m_p+w,k}$	
	(1)	(2)	(3)	(4)	(5)	(6)
$\theta^{\text{NYT}}_{m-w,k}$	0.159***	0.067***		0.161***		0.131***
	(0.014)	(0.013)		0.015)		(0.017)
$\theta^{\text{Fed}}_{m,k}$			0.146***	0.064***	0.087***	0.060***
			(0.014)	(0.015)	(0.013)	(0.015)
$\theta^{\text{NYT}}_{m_p-w,k}$				0.266***	0.369***	0.241***
				(0.014)	(0.013)	(0.014)
$\theta^{\text{NYT}}_{m+w,k}$				0.158***		0.100***
				(0.015)		(0.017)
$\theta^{\text{NYT}}_{m-w,k}$ lags		✓				
$\theta^{\text{Fed}}_{m,k}$ lags				✓		✓
Dep variable lags		✓		✓		✓
Topic fixed effect	✓	✓	✓	✓	✓	✓
Time fixed effect		✓		✓		✓
Observations	4,920	4,830	4,920	4,830	4,920	4,830
R^2	0.025	0.283	0.021	0.212	0.154	0.228
Adjusted R^2	0.019	0.252	0.015	0.179	0.149	0.196
Residual Std. Error	0.987	0.863	0.989	0.906	0.920	0.897
Note:						*p<0.1; **p<0.05; ***p<0.01

The focus of the FOMC's minutes and the NYT articles thus robustly co-move over time, which provides supporting evidence for the idea that this focus is a response to the distribution of uncertainty across the different dimensions of the economy. The fact that the minutes predict the focus of articles immediately after their publication, which is several weeks after the meeting itself, suggests that there may be direct influence from the focus of the minutes to the media's focus.

4.6 Co-movement across Central Banks

This Section extends the analysis of co-moving focus to additional central banks. We find significant and robust co-movement of topic proportions across the communication of the Federal Reserve's Federal Open Market Committee (FOMC), Bank of England's Monetary Policy Committee (MPC) and European Central Bank's Governing Council (GC). We show that the communication of one central bank can predict the communication of other central banks, in two different ways. Firstly, we aggregate the topic proportions into quarterly variables and show that FOMC communication Granger causes that of the MPC and GC. Secondly, we show that the proportions of the most recently published communication has similar cross-central bank effects. Finally, we show that a change in the publication policy of the FOMC's minutes can be used to show that they may have a causal influence on the MPC minutes.

4.6.1 Measuring Central Bank Focus

We estimate a 30 topic LDA model on the combined paragraphs of the FOMC minutes, MPC minutes and GC statements.[14]. This process constructs three sets of 30 time series covering the sample period of January 1997 through to May 2014. These topic proportions capture the shifting attention of the different central banks, and will be the focus of the remainder of this Section. Table 4.6 summarises the topics learned, in most cases an interpretation for the topic is clear from just the 5 most highly weighted words in the corresponding β vector. It also reports the average proportion of each meeting devoted to each topic for each of the three central banks, across the sample period, showing that the average proportion is sufficiently similar to make comparisons meaningful.

As would be expected, some topics are much more prevalent in one central bank's communications than the others. For example, Topic 13 appears to discuss the FOMC directly and so appears much more in the FOMC corpus than the other two corpora. Similarly, Topic 14 which covers fiscal policy and structural reforms features much more prevalently in the ECB corpus, as due to the international nature of its monetary union this issue is of more concern to the ECB that either the Federal Reserve or Bank of England.

[14]Details of the estimation are given in Appendix D.1.2. We also verify that these series are stationary in Appendix D.5.

Table 4.6: Topics estimated on the three central bank corpora and their average in each

Topic	Description	Top 5 words	$\bar{\theta_k}^{BoE}$	$\bar{\theta_k}^{ECB}$	$\bar{\theta_k}^{Fed}$
Topic 1	Economic data	fallen, sinc, risen, fall, average	0.0575	0.0126	0.0091
Topic 2	Growth expectations	seem, might, prospect, slowdown, recoveri	0.0641	0.0143	0.0131
Topic 3	Staff projections	project, forecast, report, staff, central	0.0348	0.0233	0.0240
Topic 4	International trade	trade, import, export, foreign, net	0.0310	0.0144	0.0428
Topic 5	Cost push factors	pressure, cost, product, wage, capac	0.0497	0.0212	0.0225
Topic 6	Inflation expectations	inflat, risk, target, committee, view	0.0641	0.0114	0.0180
Topic 7	Hypotheticals	might, possibl, earn, pay, one	0.0646	0.0106	0.0106
Topic 8	GDP data	quarter, first, second, gdp, estim	0.0394	0.0253	0.0335
Topic 9	Household consumption	consum, spend, household, consumpt, incom	0.0330	0.0112	0.0417
Topic 10	Credit conditions	credit, bank, loan, financi, lend	0.0360	0.0510	0.0220
Topic 11	Business investment	busi, invest, inventori, spend, capit	0.0212	0.0099	0.0680
Topic 12	Market expectations	particip, econom, note, improve, longer	0.0120	0.0124	0.0674
Topic 13	FOMC	comitte, feder, percent, consist, reserve	0.0073	0.0090	0.0737
Topic 14	Fiscal reforms	fiscal, countri, govern, reform, structur	0.0168	0.1310	0.0126
Topic 15	Core inflation	inflat, energi, oil, core, cpi	0.0326	0.0330	0.0431
Topic 16	Committee expectations	member, expans, prospect, factor, persist	0.0199	0.0179	0.0741
Topic 17	Output data	survey, data, output, manufactur, servic	0.0674	0.0130	0.0112
Topic 18	Interest rate	interest, point, short, basi, reduct	0.0459	0.0221	0.0178
Topic 19	Labour market	labour, employ, unemploy, measur, privat	0.0302	0.0173	0.0381
Topic 20	Policy committee	polici, member, committe, monetari, econom	0.0235	0.0150	0.0685
Topic 21	Bond market	period, yield, bond, spread, fund	0.0252	0.0115	0.0518
Topic 22	Policy decision	polici, financi, committe, decis, discuss	0.0322	0.0159	0.0248
Topic 23	Exchange rates	unit, state, sterl, dollar, exchang	0.0484	0.0102	0.0212
Topic 24	Industrial production	product, industri, moder, rose, manufactur	0.0097	0.0079	0.0805
Topic 25	Quantitative easing	bank, purchas, asset, committe, vote	0.0403	0.0111	0.0158
Topic 26	Housing market	hous, mortgag, home, sale, new	0.0237	0.0090	0.0463
Topic 27	ECB GC	govern, council, will, meet, ecb	0.0091	0.1040	0.0090
Topic 28	Eurozone	euro, area, econom, recoveri, global	0.0234	0.1121	0.0075
Topic 29	Monetary stability	monetari, medium, stabil, develop, econom	0.0121	0.1731	0.0113
Topic 30	Risk	risk, develop, uncertainti, downsid, global	0.0249	0.0694	0.0199

4.6.2 Co-movement of Focus across Central Banks

The content of central bank communication across the three central banks in the sample displays a co-movement which is robust to different specifications and controlling for local macroeconomic conditions. The Federal Reserve's communication appears to be the most influential, as it leads the communication of other central banks at a quarterly frequency, and a shortening in its publication lag lead to an increase in the observed co-movement. This is consistent with the prediction that the focus of one central bank's communication will follow that of other central banks who's communication is informative for its own economy.

Some aggregation of the communication documents is necessary to statistically measure the co-movement of the central bank communication series. This highlights an

initial difficulty when assessing the co-movement across the different central banks, as the meetings do not line up perfectly with one another, as was shown in Figure 4.5. Although each central bank has multiple meetings in each quarter, the number of meetings per year/quarter is not consistent across central banks or across time. Furthermore, the time between meetings is also not consistent.

We therefore present results here under two different specifications: a straightforward quarterly panel VAR, and panel regressions of a central bank's focus on recently published communication of its peers (either the single most recent document, or an average over a three month window prior to the meeting date).

Quarterly Panel VAR

Aggregating each series to quarterly frequency makes testing for Granger causality straightforward with a reduced form panel VAR. The panel is made up of the 90 topic series (30 topics $\times$ 3 central banks). The topic proportions of the three central banks for each topic and quarter are stacked such that $\theta_{k,t} = (\theta_{k,t}^{\text{Fed}}, \theta_{k,t}^{\text{BoE}}, \theta_{k,t}^{\text{ECB}})'$. Indexing over topics with k, and quarters with t the panel VAR estimated is

$$\theta_{k,t} = \alpha_k + \sum_{l=1}^{p} A_l \theta_{k,t-l} + \varepsilon_{k,t}$$

The α_k vector of intercept terms accounts for central bank-specific differences in topic proportions, and the A_l matrices capture the effects of lags across the three central banks. At the quarterly level, there is no natural ordering of the variables to allow structural identification (i.e. through a Cholesky decomposition) so we report the results of a reduced form VAR and Generalised Impulse Response Functions as proposed by Pesaran and Shin (1998). However, the specifications later in this Section will make use of the chronology of meetings in more detail, which implies a natural ordering.

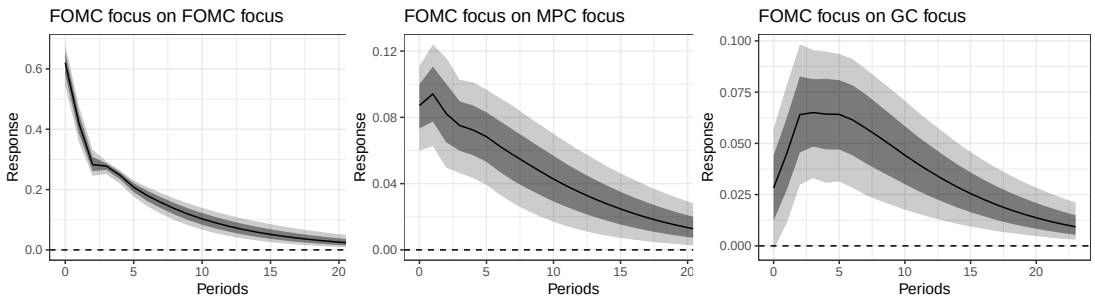
The panel VAR on the quarterly data reveals that the topic proportions in the Federal Reserve's communication Granger cause those of the Bank of England and the European Central Bank. Table 4.7 shows the results for this panel VAR estimated with 1 lag in Columns (1), (3), (5) and with 3 lags in Columns (2), (4), (6). As before, variables are standardised to have zero mean and unit variance, the results for unstandardised data reported in Appendix D.7 are qualitatively the same.

Table 4.7: Panel VAR results on quarterly (standardised) topic proportions

	Dependent variable:					
	$\theta^{\text{Fed}}_{k,t}$		$\theta^{\text{BoE}}_{k,t}$		$\theta^{\text{ECB}}_{k,t}$	
	(1)	(2)	(3)	(4)	(5)	(6)
$\theta^{\text{Fed}}_{k,t-1}$	0.627***	0.463***	0.099***	0.108***	0.110***	0.080***
	(0.018)	(0.023)	(0.020)	(0.025)	(0.020)	(0.025)
$\theta^{\text{BoE}}_{k,t-1}$	0.022	0.007	0.526***	0.380***	0.032	0.006
	(0.018)	(0.022)	(0.020)	(0.023)	(0.020)	(0.024)
$\theta^{\text{ECB}}_{k,t-1}$	0.028	0.007	0.023	0.027	0.443***	0.319***
	(0.018)	(0.021)	(0.019)	(0.023)	(0.020)	(0.023)
Number of lags	1	3	1	3	1	3
Topic fixed effects	✓	✓	✓	✓	✓	✓
Observations	1,950	5,100	1,950	5,060	1,950	5,060
Number of groups	30	30	30	30	30	30
Obs per group	65	65	65	65	65	65
Note:			*p<0.1; **p<0.05; ***p<0.01			

The Generalised Impulse Response Functions tell a similar story, shown in Figures 4.13 - 4.17, with the focus of FOMC communication having persistent and highly significant effects on the focus of both the MPC's and GC's communication. These IRFs are based on a panel VAR using all 30 topic series and 3 lags. We also see a smaller and less persistent effect of MPC and GC communication on that of the other two central banks.

Figure 4.13: Generalised IRFs for shock to FOMC focus.



Note: Confidence Intervals are bootstrapped by 5,000 iterations of blockwise sampling of individuals. The darker band represents the 70% confidence interval and the lighter the 95% confidence interval.

Figure 4.15: Generalised IRFs for shock to MPC focus.

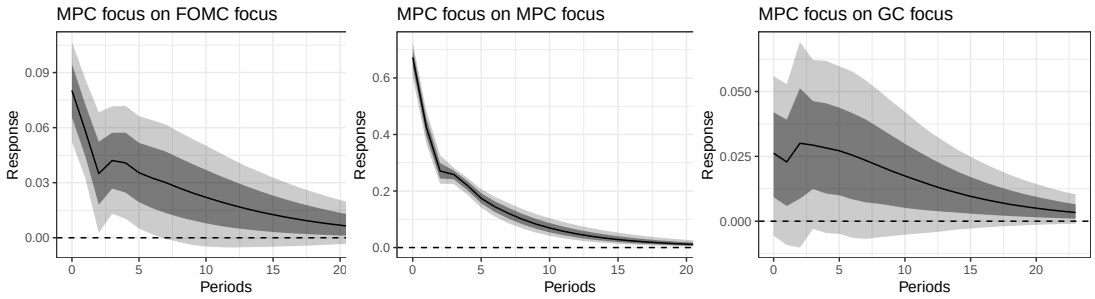
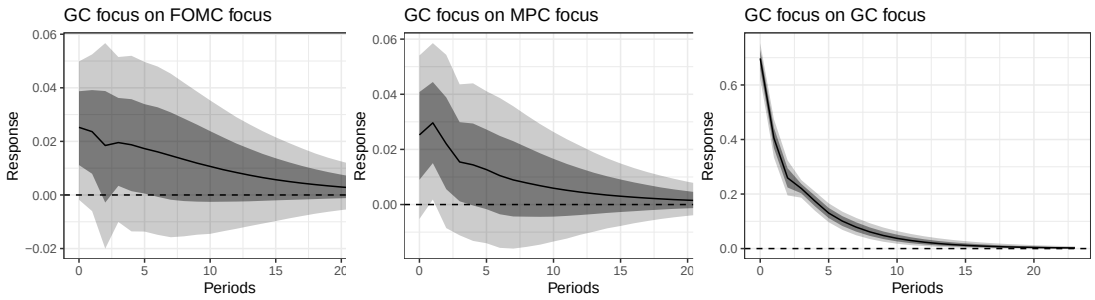


Figure 4.17: Generalised IRFs for shock to GC focus.

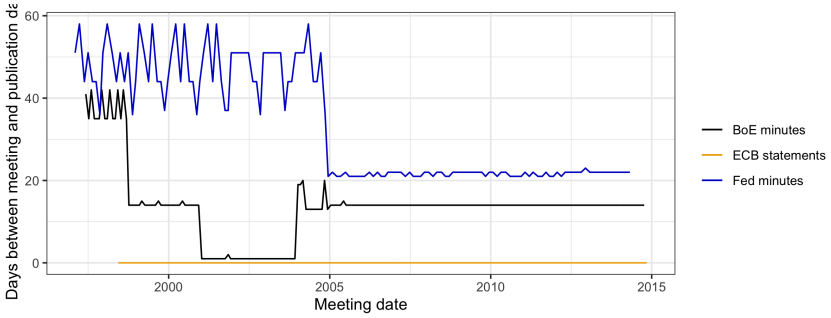


The focus of the Federal Reserve's communication thus leads that of the other central banks. In the context of the discussion in Section 4.4.4, this could be because the state of the US economy is more relevant to the UK and Eurozone than vice versa, and/or because the Federal Reserve has more accurate signals about its economy so its communication is a more useful signal for other central banks. An alternative explanation is also possible, where the Federal Reserve is a "thought leader" which influences the way other central banks process information and structure their communication, but beyond this is the scope of this paper.

Recently Published Communication

The quarterly panel VAR presented above does not make use of the more fine-grained information on the meeting and publication date associated with each document, which would allow a richer analysis of co-movement. Unfortunately, official central bank communication does not occur regularly or consistently enough to allow an event-study type approach such as that presented in Section 4.5.3. However, we can identify communication that was published shortly before a meeting and test whether this has a predictive effect. As Figure 4.19 shows, this window between meeting and publication is not consistent across central banks or constant over time. It is therefore necessary to take a stance on the time-scale at which cross-central bank influence is possible.

Figure 4.19: Gap between meeting and publication of minutes



We use two alternative approaches to define "recently published" communication:

1. Most recent: other central banks' most recently *published* piece of communication, prior to the meeting date.[15]
2. 3 month window: an average of a central bank's communication published in a rolling window of three months prior to a meeting.

As the meetings and subsequent communications of the three central banks do not line up neatly, a simple timing notation will be inappropriate. We therefore use the (b, m) subscript to indicate meeting m of central bank b.[16] The recently published communication of central bank c prior to central bank b's meeting m is denoted by the superscript c. For example $\theta^{\text{ECB}}_{\text{ECB},1,k}$ denotes the proportion of the first ECB statement in the sample that is devoted to topic k. $\theta^{\text{Fed}}_{\text{ECB},1,k}$ denotes the proportion of the FOMC meeting while precedes the first ECB statement in the sample, that is devoted to topic k. This essentially gives a three dimensional panel data set which is indexed by central bank, topic and meeting.

The two different approaches described above will give different values for the potentially influential documents of other central banks, $\theta^{c}_{b,m,k}$, but the regression specification will be the same for both.

$$\theta^{b}_{b,m,k} = \alpha_{b,k} + \sum_{c \neq b} \gamma_{b,c} \theta^{c}_{b,m,k} + \sum_{p=1}^{P} \rho_{b,p} \theta^{b}_{b,m-p,k} + \text{controls} + \varepsilon_{b,m,k}$$

[15]Federal Reserve minutes from the meeting of 12th February 2019 would thus potentially be influenced by Bank of England minutes for a meeting on the 1st January which was published on 1st February and an ECB statement on 29th January (published on 29th January). These Fed minutes are published on 2nd March and so potentially influence Bank of England minutes of a meeting on 6th March and an ECB statement on 29th March.

[16]Note that the m will not line up across the different central banks, so $(\text{Fed}, 1)$ does not necessarily correspond to $(\text{BoE}, 1)$. The cross-central bank effects are instead explicitly defined by the meeting and publication dates as described. However, within one central bank's topic proportions the m subscripts are sequential, so we can include lagged values of the topics to account for persistence in the communication over time.

The primary coefficients of interest are $\gamma_{b,c}$ which capture the predictive content of central bank c's recently published communication on that of central bank b. There are 6 of these coefficients, one for each central bank's effect on each of the others, so $\gamma_{\text{ECB, Fed}}$ captures the effect of recent FOMC communication on the statements of the ECB Governing Council. The $\sum_{p=1}^{P} \rho_{b,p}\theta^{b}_{b,m-p,k}$ terms control for persistence in central bank b's communication. The $\rho_{b,p}$ are allowed to vary across the central banks to capture the fact that the persistence of communication may differ (not least because the meeting schedules differ). The δ_k coefficients capture the topic-specific effects of the control variables (GDP growth, inflation and the change in the policy rate). Note that the these control variable effects are topic specific rather than central bank specific, so allow for the fact that different topics will have different relationships with the same control variable. The central bank specific fixed effects ($\alpha_{b,k}$) account for inherent differences in language and communication structure across the central banks and the lags of previous meetings ($\rho_{b,p}\theta^{b}_{b,m-p,k}$) account for potential persistence in communication content. The regression is estimated by OLS, with the topic and central bank specific coefficients calculated by creating a sparse matrix where variables are set to zero for topics/central banks they do not apply to.

Table 4.8 shows the results, which indicate that the Federal Reserve communication is the most "influential" in that it's communication has the greatest predictive effect on that of the other central banks, but that recently published Bank of England communication also predicts Federal Reserve and European Central Bank communication. The standardised series are used, and results for the unstandardised shown in Appendix D.7. Columns (3) and (6) present the estimates of the $\gamma_{b,c}$ coefficients with central-bank specific lags and topic-specific macroeconomic controls. The largest and most significant coefficients are those for the Federal Reserve's influence on the other two central banks.

Table 4.8: CBC can be predicted by recently published communication of other central banks.

	Empirical strategy					
	3 month window			Most recent		
	(1)	(2)	(3)	(4)	(5)	(6)
$\gamma_{\text{Fed,BoE}}$	0.101***	0.042***	0.035*	0.068***	0.041***	0.035**
	(0.020)	(0.019)	(0.019)	(0.015)	(0.014)	(0.014)
$\gamma_{\text{Fed,ECB}}$	0.075***	0.018	0.018	0.048***	0.015	0.013
	(0.019)	(0.018)	(0.018)	(0.014)	(0.013)	(0.013)
$\gamma_{\text{BoE,Fed}}$	0.129***	0.053***	0.054***	0.108***	0.054***	0.052***
	(0.014)	(0.013)	(0.013)	(0.012)	(0.011)	(0.012)
$\gamma_{\text{BoE,ECB}}$	0.047***	0.007	0.009	0.030***	0.006	0.005
	(0.015)	(0.014)	(0.015)	(0.012)	(0.012)	(0.012)
$\gamma_{\text{ECB,Fed}}$	0.107***	0.068***	0.054***	0.822***	0.051***	0.039***
	(0.014)	(0.014)	(0.014)	(0.012)	(0.012)	(0.012)
$\gamma_{\text{ECB,BoE}}$	0.068***	0.030*	0.029**	0.030*	0.043***	0.037***
	(0.016)	(0.016)	(0.016)	(0.016)	(0.018)	(0.018)
Number of CB-specific lags	1	10	10	1	10	10
Topic-specific macro controls			✓			✓
CB-topic fixed effect	✓	✓	✓	✓	✓	✓
Observations	15,330	15,060	15,060	15,330	15,060	15,060
R^2	0.204	0.298	0.309	0.202	0.299	0.309
Adjusted R^2	0.199	0.293	0.299	0.196	0.293	0.299
Residual Std. Error	0.885	0.823	0.819	0.886	0.822	0.819

Note: *p<0.1; **p<0.05; ***p<0.01

Note: The $\gamma_{i,j}$ coefficients measure the effect of central bank j's recently published communicationon the communication of central bank i. For example, $\gamma_{\text{Fed,BoE}}$ indicates the effect of the BoE's recently published communication on that of the Fed.

An indicative test of whether this co-movement might be driven by some influence of the Federal Reserve's communication on that of the other two central banks, rather than just an independent reaction to similar economic conditions, is given by the dramatic shortening of the FOMC's publication policy at the start of 2005. As can be seen in Figure 4.19, the FOMC switched from around a 50 day gap between a meeting and the publication of the corresponding minutes to around 20 days.

If the published communication of the Federal Reserve is indeed influencing the future communication of other central banks, then we would expect the effect to be greater from 2005 onwards, as published minutes will be from more recent meetings and so more likely to contain relevant and new information. To test whether this is the case, the $\theta^{Fed}_{BoE,t,k}$ and $\theta^{Fed}_{ECB,t,k}$ terms are interacted with a dummy variable with value one from 2005 onwards and zero prior to 2005. We report the results for the "most recent" specification (with 10 lags and topic-specific macro controls) as this is where the change in publication policy has the greatest effect. Table 4.9 shows that for the effect of the Federal Reserve communication on the Bank of England, this increases after the change in Federal Reserve publication policy, although there is no change for the effect on the ECB. Full results for the test are shown in Appendix D.7.

Table 4.9: The FOMC influence on MPC minutes is slightly greater after 2005

Regressor			
$\theta^{Fed}_{BoE,t,k}$	$\theta^{Fed}_{ECB,t,k}$	$\theta^{Fed}_{BoE,t,k}\mathbb{I}_{\{t\geq 2005\}}$	$\theta^{Fed}_{ECB,t,k}\mathbb{I}_{\{t\geq 2005\}}$
0.020	0.050**	0.065**	0.001
(0.019)	(0.020)	(0.025)	(0.026)
Note:			*p<0.1; **p<0.05; ***p<0.01

We thus have evidence of a robust co-movement in the focus of the communication of the Federal Reserve, Bank of England and European Central Bank, over the period from 1998-2014. The Federal Reserve's communication appears to lead that of the other two central banks. Moving from a documentation of this co-movement, and suggestive evidence of influence, to concrete evidence for a specific causal mechanism is a challenge left for further work.

4.7 Conclusion

This paper proposes an explanation for the shifting focus of central bank communication which is observed in the data. Central bank's devote more of their communication to dimensions of the economy about which there is greater uncertainty. This explanation implies that the focus of central bank communication may co-move across different central banks not only because of similar conditions in their domestic economies, but also because they learn from one another's signals.

We provide a variety of novel empirical evidence in support of this theory. Firstly, the FOMC's communication is shown focuses more on topics related to variables about which the private sector is more uncertain. Secondly, the focus of central bank communication and that of the media co-moves, and central bank communication plausible has a direct influence on the focus of the media. Finally, the focus of the communication of the Federal Reserve, the Bank of England and the European Central Bank co-move, and the Federal Reserve appears to lead the other two.

Bibliography

Daron Acemoglu, Ufuk Akcigit, and William Kerr. Networks and the macroeconomy: An empirical exploration. *NBER Macroeconomics Annual*, 30(1):273–335, 2016.

Deepak Agarwal and Bee-Chung Chen. flda: matrix factorization through latent dirichlet allocation. In *Proceedings of the third ACM International Conference on Web Search and Data Mining*, pages 91–100. ACM, 2010.

Merve Alanyali, Helen Susannah Moat, and Tobias Preis. Quantifying the relationship between financial news and the stock market. *Scientific reports*, 3:3578, 2013.

Malin Andersson, Hans Dillén, and Peter Sellin. Monetary policy signaling and movements in the term structure of interest rates. *Journal of Monetary Economics*, 53(8): 1815–1855, 2006.

Daniel Andrei and Michael Hasler. Investor attention and stock market volatility. *Review of Financial Studies*, 28(1), 2015.

George-Marios Angeletos, Fabrice Collard, and Harris Dellas. Public debt as private liquidity: Optimal policy. Technical report, National Bureau of Economic Research, 2019.

Jasmina Arifovic, James Bullard, and Olena Kostyshyna. Social learning and monetary policy rules. *The Economic Journal*, 123(567):38–76, 2013.

Hanna Armelius, Christoph Bertsch, Isaiah Hull, and Xin Zhang. Spread the word: International spillovers from central bank communication. *Journal of International Money and Finance*, 103:102116, 2020.

S Borağan Aruoba, Pablo Cuba-Borda, and Frank Schorfheide. Macroeconomic dynamics near the zlb: A tale of two countries. *Review of Economic Studies*, 85(1), 2018.

Guido Ascari and Sophocles Mavroeidis. The unbearable lightness of equilibria in a low interest rate environment. *arXiv preprint arXiv:2006.12966*, 2020.

Guido Ascari, Paolo Bonomolo, and Hedibert F Lopes. Walk on the wild side: Temporarily unstable paths and multiplicative sunspots. *American Economic Review*, 109 (5):1805–42, 2019.

Susan Athey, David Blei, Robert Donnelly, Francisco Ruiz, and Tobias Schmidt. Estimating heterogeneous consumer preferences for restaurants and travel time using mobile location data. In *AEA Papers and Proceedings*, volume 108, pages 64–67, 2018.

Marlon Azinovic, Luca Gaegauf, and Simon Scheidegger. Deep equilibrium nets. *Available at SSRN 3393482*, 2019.

Andres Azqueta-Gavaldon, Dominik Hirschbühl, Luca Onorante, and Lorena Saiz. Economic policy uncertainty in the euro area: An unsupervised machine learning approach. 2020.

Nicole Baerg and Will Lowe. A textual taylor rule: estimating central bank preferences combining topic and scaling methods. *Political Science Research and Methods*, 8(1): 106–122, 2020.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Scott R Baker, Nicholas Bloom, and Steven J Davis. Measuring economic policy uncertainty. *The Quarterly Journal of Economics*, 131(4):1593–1636, 2016.

Oriana Bandiera, Andrea Prat, Stephen Hansen, and Raffaella Sadun. Ceo behavior and firm performance. *Journal of Political Economy*, 128(4):1325–1369, 2020.

Jean-Noël Barrot and Julien Sauvagnat. Input specificity and the propagation of idiosyncratic shocks in production networks. *The Quarterly Journal of Economics*, 131(3): 1543–1592, 2016.

Larry M Bartels. Politicians and the press: Who leads, who follows. In *annual meeting of the American Political Science Association, San Francisco, CA*. Citeseer, 1996.

Jess Benhabib and Roger EA Farmer. Indeterminacy and increasing returns. *Journal of Economic Theory*, 63(1):19–41, 1994.

Jess Benhabib, Stephanie Schmitt-Grohé, and Martin Uribe. The perils of taylor rules. *Journal of Economic Theory*, 96(1-2):40–69, 2001.

Jess Benhabib, George W Evans, and Seppo Honkapohja. Liquidity traps and expectation dynamics: Fiscal stimulus or fiscal austerity? *Journal of Economic Dynamics and Control*, 45:220–238, 2014.

Jess Benhabib, Pengfei Wang, and Yi Wen. Sentiments and aggregate demand fluctuations. *Econometrica*, 83(2):549–585, 2015.

Michele Berardi and John Duffy. Real-time, adaptive learning via parameterized expectations. *Macroeconomic Dynamics*, 19(2):245–269, 2015.

Ben Bernanke. A perspective on inflation targeting: why it seems to work. *Business Economics*, 38(3):7–16, 2003.

Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B Shah. Julia: A fresh approach to numerical computing. *SIAM review*, 59(1):65–98, 2017.

David Bholat, Nida Broughton, Alice Parker, Janna Ter Meer, and Eryk Walczak. Enhancing central bank communications with behavioural insights. 2018.

Lorenz T Biegler and Victor M Zavala. Large-scale nonlinear programming using ipopt: An integrating framework for enterprise-wide dynamic optimization. *Computers & Chemical Engineering*, 33(3):575–582, 2009.

Carola Binder. Fed speak on main street: Central bank communication and household expectations. *Journal of Macroeconomics*, 52:238–251, 2017a.

Carola Binder. Federal reserve communication and the media. *Journal of Media Economics*, 30(4):191–214, 2017b.

Christopher M Bishop. *Pattern recognition and machine learning*. Information Science and Statistics. Springer, New York, 2006. ISBN 9780387310732.

Olivier Jean Blanchard and Charles M Kahn. The solution of linear difference models under rational expectations. *Econometrica: Journal of the Econometric Society*, pages 1305–1311, 1980.

David M Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

David M Blei and Jon D McAuliffe. Supervised topic models. In *Advances in Neural Information Processing Systems*, pages 121–128, 2008.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

Alan S Blinder, Michael Ehrmann, Marcel Fratzscher, Jakob De Haan, and David-Jan Jansen. Central bank communication and monetary policy: A survey of theory and evidence. *Journal of Economic Literature*, 46(4):910–945, 2008.

Deborah J Blood and Peter CB Phillips. Recession headline news, consumer sentiment, the state of the economy and presidential popularity: A time series analysis 1989–1993. *International Journal of Public Opinion Research*, 7(1):2–22, 1995.

Benjamin Born, Michael Ehrmann, and Marcel Fratzscher. Central bank communication on financial stability. *The Economic Journal*, 124(577):701–734, 2014.

R Anton Braun and Lena Mareen Körber. New keynesian dynamics in a low interest rate environment. *Journal of Economic Dynamics and Control*, 35(12):2213–2227, 2011.

Jane Delano Brown, Carl R Bybee, Stanley T Wearden, and Dulcie Murdock Straughan. Invisible power: Newspaper news sources and the limits of diversity. *Journalism Quarterly*, 64(1):45–54, 1987.

Aleš Bulíř, Martin Čihák, and David-Jan Jansen. What drives clarity of central bank communication about inflation? *Open Economies Review*, 24(1):125–145, 2013.

James Bullard. Learning equilibria. *Journal of Economic Theory*, 64(2):468–485, 1994.

James Bullard. Seven faces of" the peril". *Federal Reserve Bank of St. Louis Review*, 92 (September/October 2010), 2010.

James Bullard and Kaushik Mitra. Learning about monetary policy rules. *Journal of monetary economics*, 49(6):1105–1129, 2002.

Brian J Bushee, John E Core, Wayne Guay, and Sophia JW Hamm. The role of the business press as an information intermediary. *Journal of Accounting Research*, 48(1): 1–19, 2010.

Fabio Canova and Matteo Ciccarelli. Panel vector autoregressive models: A survey. In *VAR Models in Macroeconomics–New Developments and Applications: Essays in Honor of Christopher A. Sims*, pages 205–246. Emerald Group Publishing Limited, 2013.

Ziqiang Cao, Sujian Li, Yang Liu, Wenjie Li, and Heng Ji. A novel neural topic model and its supervised extension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Dallas Card, Chenhao Tan, and Noah A. Smith. Neural models for documents with metadata. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2031–2040. Association for Computational Linguistics, jul 2018.

Christopher D Carroll. Macroeconomic expectations of households and professional forecasters. *The Quarterly Journal of Economics*, 118(1):269–298, 2003.

Carlos Carvalho and Fernanda Nechio. Do people understand monetary policy? *Journal of Monetary Economics*, 66:108–123, 2014.

Carlos Carvalho, Nicholas Klagge, and Emanuel Moench. The persistent effects of a false news shock. *Journal of Empirical Finance*, 18(4):597–615, 2011.

Guido Cazzavillan, Teresa Lloyd-Braga, and Patrick A Pintus. Multiple steady states and endogenous fluctuations with increasing returns to scale in production. *Journal of Economic Theory*, 80(1):60–107, 1998.

Hailiang Chen, Prabuddha De, Yu Jeffrey Hu, and Byoung-Hyoun Hwang. Wisdom of crowds: The value of stock opinions transmitted through social media. *The Review of Financial Studies*, 27(5):1367–1403, 2014.

Jianshu Chen, Ji He, Yelong Shen, Lin Xiao, Xiaodong He, Jianfeng Gao, Xinying Song, and Li Deng. End-to-end learning of lda by mirror-descent back propagation over a deep architecture. *arXiv preprint arXiv:1508.03398*, 2015.

Mingli Chen, Andreas Joseph, Michael Kumhof, Xinlei Pan, Rui Shi, and Xuan Zhou. Deep reinforcement learning in a monetary model. *arXiv preprint arXiv:2104.09368*, 2021.

Xiaohong Chen and Halbert White. Nonparametric adaptive learning with feedback. *Journal of Economic Theory*, 82(1):190–222, 1998.

Wang Chong, David Blei, and Fei-Fei Li. Simultaneous image classification and annotation. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1903–1910. IEEE, 2009.

Lawrence Christiano, Martin S Eichenbaum, and Benjamin K Johannsen. Does the new keynesian model have a uniqueness problem? Technical report, National Bureau of Economic Research, 2018.

Anna Cieslak and Andreas Schrimpf. Non-monetary news in central bank communication. *Journal of International Economics*, 118:293–315, 2019.

Richard Clarida, Jordi Gali, and Mark Gertler. Monetary policy rules and macroeconomic stability: evidence and some theory. *The Quarterly journal of economics*, 115(1):147–180, 2000.

John H Cochrane. Can learnability save new-keynesian models? *Journal of Monetary Economics*, 56(8):1109–1113, 2009.

Lauren Cohen and Andrea Frazzini. Economic links and predictable returns. *The Journal of Finance*, 63(4):1977–2011, 2008.

Olivier Coibion and Yuriy Gorodnichenko. Information rigidity and the expectations formation process: A simple framework and new facts. *American Economic Review*, 105(8):2644–78, 2015a.

Olivier Coibion and Yuriy Gorodnichenko. Is the Phillips curve alive and well after all? Inflation expectations and the missing disinflation. *American Economic Journal: Macroeconomics*, 7(1):197–232, 2015b.

Ricardo Correa, Keshav Garud, Juan M Londono, Nathan Mislang, et al. Constructing a dictionary for financial stability. *Board of Governors of the Federal Reserve System (US)*, 6(7):9, 2017.

Dean D Croushore. Introducing: the survey of professional forecasters. *Business Review-Federal Reserve Bank of Philadelphia*, 6:3, 1993.

DM Cutler, JM Poterba, and LH Summers. What moves stock prices? *Journal of Portfolio Management*, 15(2), 1989.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

Suzanna De Boef and Paul M Kellstedt. The political (and economic) origins of consumer confidence. *American Journal of Political Science*, 48(4):633–649, 2004.

Jakob de Haan and Jan-Egbert Sturm. Central bank communication. In *The Oxford handbook of the economics of central banking*, page 231. Oxford University Press, 2019.

Stefano DellaVigna and Ethan Kaplan. The fox news effect: Media bias and voting. *The Quarterly Journal of Economics*, 122(3):1187–1234, 2007.

Wouter J Den Haan and Albert Marcet. Accuracy in simulations. *The Review of Economic Studies*, 61(1):3–17, 1994.

David A Dickey and Wayne A Fuller. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74(366a): 427–431, 1979.

Adji B. Dieng, Chong Wang, Jianfeng Gao, and John W. Paisley. Topicrnn: A recurrent neural network with long-range semantic dependency. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

Thomas Dimpfl and Stephan Jank. Can internet search queries help to predict stock market volatility? *European Financial Management*, 22(2):171–192, 2016.

Peter Sheridan Dodds, Kameron Decker Harris, Isabel M Kloumann, Catherine A Bliss, and Christopher M Danforth. Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PloS one*, 6(12):e26752, 2011.

Casey Dougal, Joseph Engelberg, Diego Garcia, and Christopher A Parsons. Journalists and the stock market. *The Review of Financial Studies*, 25(3):639–679, 2012.

Lena Dräger, Michael J Lamla, and Damjan Pfajfar. Are survey expectations theory-consistent? the role of central bank communication and news. *European Economic Review*, 85:84–111, 2016.

Michael S Drake, Nicholas M Guest, and Brady J Twedt. The media and mispricing: The role of the business press in the pricing of accounting information. *The Accounting Review*, 89(5):1673–1701, 2014.

John Duffy. On learning and the nonuniqueness of equilibrium in an overlapping generations model with fiat money. *Journal of Economic Theory*, 64(2):541–553, 1994.

Jan Eeckhout and Ilse Lindenlaub. Unemployment cycles. *American Economic Journal: Macroeconomics*, 11(4):175–234, 2019.

Gauti B Eggertsson, Neil R Mehrotra, and Jacob A Robbins. A model of secular stagnation: Theory and quantitative evaluation. *American Economic Journal: Macroeconomics*, 11(1):1–48, 2019.

Jacob Eisenstein, Amr Ahmed, and Eric P Xing. Sparse additive generative models of text. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 1041–1048. Citeseer, 2011.

Asmaa Elbadrawy and George Karypis. User-specific feature-based similarity models for top-n recommendation of new items. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 6(3):1–20, 2015.

Martin Ellison and Joseph Pearlman. Saddlepath learning. *Journal of Economic Theory*, 146(4):1500–1519, 2011.

Joseph E Engelberg and Christopher A Parsons. The causal impact of media in financial markets. *The Journal of Finance*, 66(1):67–97, 2011.

George W Evans and Seppo Honkapohja. *Learning and Expectations in Macroeconomics*. Princeton University Press, 2001.

George W Evans and Bruce McGough. Monetary policy, indeterminacy and learning. *Journal of Economic Dynamics and Control*, 29(11):1809–1840, 2005.

George W Evans, Eran Guse, and Seppo Honkapohja. Liquidity traps, learning and stagnation. *European Economic Review*, 52(8):1438–1463, 2008.

George W Evans, Seppo Honkapohja, and Kaushik Mitra. Expectations, stagnation and fiscal policy. *CEPR Discussion Paper No. DP11428*, 2016.

Norman Fairclough, Jane Mulderrig, and Ruth Wodak. Critical discourse analysis. in van dijk, ta. *Discourse as social interaction*, 2011.

Rui Fan, Oleksandr Talavera, and Vu Tran. Social media bots and stock markets. *European Financial Management*, 26(3):753–777, 2020.

Roger EA Farmer, Vadim Khramov, and Giovanni Nicolò. Solving and estimating indeterminate dsge models. *Journal of Economic Dynamics and Control*, 54:17–36, 2015.

Jesús Fernández-Villaverde, Grey Gordon, Pablo Guerrón-Quintana, and Juan F Rubio-Ramirez. Nonlinear adventures at the zero lower bound. *Journal of Economic Dynamics and Control*, 57:182–204, 2015.

Jesús Fernández-Villaverde, Samuel Hurtado, and Galo Nuno. Financial frictions and the wealth distribution. Technical report, National Bureau of Economic Research, 2019.

Jesus Fernandez-Villaverde, Federico Mandelman, Yang Yu, and Francesco Zanetti. Search complementarities, aggregate fluctuations, and fiscal policy. Technical report, National Bureau of Economic Research, 2020.

Ragnar Frisch and Frederick V Waugh. Partial time regressions as compared with individual trends. *Econometrica: Journal of the Econometric Society*, pages 387–401, 1933.

Xavier Gabaix. The granular origins of aggregate fluctuations. *Econometrica*, 79(3): 733–772, 2011.

Jordi Gali. Product diversity, endogenous markups, and development traps. *Journal of Monetary Economics*, 36(1):39–63, 1995.

Herbert J Gans. *Deciding what's news: A study of CBS evening news, NBC nightly news, Newsweek, and Time*. Northwestern University Press, 2004.

Matthew Gentzkow and Jesse M Shapiro. What drives media slant? evidence from us daily newspapers. *Econometrica*, 78(1):35–71, 2010.

Alan S Gerber, Dean Karlan, and Daniel Bergan. Does the media matter? a field experiment measuring the effect of newspapers on voting behavior and political opinions. *American Economic Journal: Applied Economics*, 1(2):35–52, 2009.

Christopher G Gibbs. Learning to believe in secular stagnation. *Economics Letters*, 163: 50–54, 2018.

John Goddard, Arben Kita, and Qingwei Wang. Investor attention and fx market volatility. *Journal of International Financial Markets, Institutions and Money*, 38:79–96, 2015.

Mario Gonzalez, Raul Cruz Tadle, et al. Monetary policy press releases: An international comparison. Technical report, Central Bank of Chile, 2021.

Yuriy Gorodnichenko, Tho Pham, and Oleksandr Talavera. Social media, sentiment and public opinions: Evidence from# brexit and# uselection. Technical report, National Bureau of Economic Research, 2018.

Alfred Greiner and Anton Bondarev. Optimal r&d investment with learning-by-doing: Multiple steady states and thresholds. *Optimal Control Applications and Methods*, 38 (6):956–962, 2017.

Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235, 2004.

Axel Groß-Klußmann and Nikolaus Hautsch. When machines read the news: Using automated text analytics to quantify high frequency news-implied market reactions. *Journal of Empirical Finance*, 18(2):321–340, 2011.

Luca Guerrieri and Matteo Iacoviello. Occbin: A toolkit for solving dynamic models with occasionally binding constraints easily. *Journal of Monetary Economics*, 70:22–38, 2015.

Pablo A Guerrón-Quintana and James M Nason. Bayesian estimation of dsge models. *Handbook of research methods and applications in empirical macroeconomics*, page 486, 2013.

Luigi Guiso, Michael Haliassos, and Tullio Jappelli. Household stockholding in europe: where do we stand and where do we go? *Economic Policy*, 18(36):123–170, 2003.

Martin T Hagan and Mohammad B Menhaj. Training feedforward networks with the marquardt algorithm. *IEEE transactions on Neural Networks*, 5(6):989–993, 1994.

Martin T Hagan, Howard B Demuth, and Mark Beale. *Neural network design*. PWS Publishing Co., 1997.

Andrew Haldane, Alistair Macaulay, and Michael McMahon. The 3 e's of central bank communication with the public. *CEPR Discussion Paper No. DP14265*, 2020.

Stephen Hansen and Michael McMahon. Shocking language: Understanding the macroeconomic effects of central bank communication. *Journal of International Economics*, 99:S114–S133, 2016.

Stephen Hansen, Michael McMahon, and Andrea Prat. Transparency and deliberation within the fomc: a computational linguistics approach. *The Quarterly Journal of Economics*, 133(2):801–870, 2018.

Stephen Hansen, Michael McMahon, and Matthew Tong. The long-run information effect of central bank communication. 2019.

Terrence Hendershott, Dmitry Livdan, and Norman Schürhoff. Are institutions informed about news? *Journal of Financial Economics*, 117(2):249–287, 2015.

Rubén Hernández-Murillo, Hannah Shell, et al. The rising complexity of the fomc statement. *Economic Synopses*, (23), 2014.

Stephen Hess. Washington reporters. *Society*, 18(4):55–66, 1981.

Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57, 1999.

Cars Hommes and Gerhard Sorger. Consistent expectations equilibria. *Macroeconomic Dynamics*, 2(3):287–321, 1998.

Lingzi Hong, Enrique Frias-Martinez, and Vanessa Frias-Martinez. Topic models to infer socio-economic maps. In *AAAI*, pages 3835–3841, 2016.

Gur Huberman and Tomer Regev. Contagious speculation and a cure for cancer: A nonevent that made stock prices soar. *The Journal of Finance*, 56(1):387–396, 2001.

Zhen Huo and José-Víctor Ríos-Rull. Paradox of thrift recessions. Technical report, National Bureau of Economic Research, 2013.

Marek Jarociński and Peter Karadi. Deconstructing monetary policy surprises—the role of information shocks. *American Economic Journal: Macroeconomics*, 12(2):1–43, 2020.

Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 78(11):15169–15211, 2019.

George J Jiang, Eirini Konstantinidi, and George Skiadopoulos. Volatility spillovers and the effect of news announcements. *Journal of Banking & Finance*, 36(8):2260–2273, 2012.

Peiran Jiao, Andre Veiga, and Ansgar Walther. Social media, news media and the stock market. *Journal of Economic Behavior & Organization*, 176:63–90, 2020.

Kenneth L Judd et al. Numerical methods in economics. *MIT Press Books*, 1, 1998.

Iebeling Kaastra and Milton Boyd. Designing a neural network for forecasting financial and economic time series. *Neurocomputing*, 10(3):215–236, 1996.

Greg Kaplan and Guido Menzio. Shopping externalities and self-fulfilling unemployment fluctuations. *Journal of Political Economy*, 124(3):771–825, 2016.

Alexander G Kerl and Andreas Walter. Market responses to buy recommendations issued by personal finance magazines: effects of information, price-pressure, and company characteristics. *Review of Finance*, 11(1):117–141, 2007.

Robert G King and Mark W Watson. The solution of singular linear difference systems under rational expectations. *International Economic Review*, pages 1015–1026, 1998.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

Paul Klein. Using the generalized schur form to solve a multivariate linear rational expectations model. *Journal of economic dynamics and control*, 24(10):1405–1423, 2000.

Peter Koudijs. The boats that did not sail: Asset price volatility in a natural experiment. *The Journal of Finance*, 71(3):1185–1226, 2016.

Paul Krugman. History versus expectations. *The Quarterly Journal of Economics*, 106 (2):651–667, 1991.

Saten Kumar, Hassan Afrouzi, Olivier Coibion, and Yuriy Gorodnichenko. Inflation targeting does not anchor inflation expectations: Evidence from firms in new zealand. Technical report, National Bureau of Economic Research, 2015.

Simon Lacoste-Julien, Fei Sha, and Michael I Jordan. Disclda: Discriminative learning for dimensionality reduction and classification. In *Advances in neural information processing systems*, pages 897–904, 2009.

Michael J Lamla and Sarah M Lein. The role of media for consumers' inflation expectation formation. *Journal of Economic Behavior & Organization*, 106:62–77, 2014.

Vegard Larsen and Leif Thorsrud. Asset returns, news topics, and media effects. Technical report, Centre for Applied Macro-and Petroleum Economics (CAMP), BI Norwegian ..., 2017.

Han Soo Lee. Analyzing the multidirectional relationships between the president, news media, and the public: Who affects whom? *Political Communication*, 31(2):259–281, 2014.

Matteo Leombroni, Andrea Vedolin, Gyuri Venter, and Paul Whelan. Central bank communication and the yield curve. *Available at SSRN 2873091*, 2018.

Richard A Levine and George Casella. Implementations of the monte carlo em algorithm. *Journal of Computational and Graphical Statistics*, 10(3):422–439, 2001.

David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5 (Apr):361–397, 2004.

Tim Loughran and Bill McDonald. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65, 2011.

Michael C Lovell. Seasonal adjustment of economic time series and multiple regression analysis. *Journal of the American Statistical Association*, 58(304):993–1010, 1963.

Thomas A Lubik and Frank Schorfheide. Testing for indeterminacy: an application to us monetary policy. *American Economic Review*, 94(1):190–217, 2004.

Robert Lucas. Adaptive behavior and economic theory. *The Journal of Business*, 59(4): S401–26, 1986.

David Lucca and Francesco Trebbi. Measuring central bank communication: An automated approach with application to fomc statements. Technical report, National Bureau of Economic Research, Inc, 2009.

Måns Magnusson, Leif Jonsson, and Mattias Villani. Dolda: a regularized supervised topic model for high-dimensional multi-class regression. *Computational Statistics*, 35 (1):175–201, 2020.

Karl-Göran Mäler, Anastasios Xepapadeas, and Aart De Zeeuw. The economics of shallow lakes. *Environmental and resource Economics*, 26(4):603–624, 2003.

Lilia Maliar, Serguei Maliar, and Pablo Winant. Will artificial intelligence replace computational economists any time soon? *CEPR Discussion Paper No. DP14024*, 2019.

Carolina Manzano and Xavier Vives. Public and private learning from prices, strategic substitutability and complementarity, and equilibrium multiplicity. *Journal of Mathematical Economics*, 47(3):346–369, 2011.

Ben R Marshall, Nuttawat Visaltanachoti, and Genevieve Cooper. Sell the rumour, buy the fact? *Accounting & Finance*, 54(1):237–249, 2014.

Julian McAuley and Jure Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM Conference on Recommender Systems*, pages 165–172. ACM, 2013.

Bennett T McCallum. E-stability vis-a-vis determinacy results for a broad class of linear rational expectations models. *Journal of Economic dynamics and control*, 31(4):1376–1391, 2007.

Karel RSM Mertens and Morten O Ravn. Fiscal policy in an expectations-driven liquidity trap. *The Review of Economic Studies*, 81(4):1637–1667, 2014.

Robert C Merton. A simple model of capital market equilibrium with incomplete information. *The journal of finance*, 42(3):483–510, 1987.

Yishu Miao, Lei Yu, and Phil Blunsom. Neural variational inference for text processing. In *International conference on machine learning*, pages 1727–1736. PMLR, 2016.

Yishu Miao, Edward Grefenstette, and Phil Blunsom. Discovering discrete latent topics with neural variational inference. In *International Conference on Machine Learning*, pages 2410–2419. PMLR, 2017.

Silvia Miranda-Agrippino and Giovanni Ricco. The transmission of monetary policy shocks. *American Economic Journal: Macroeconomics*, 13(3):74–107, 2021.

Andriy Mnih and Karol Gregor. Neural variational inference and learning in belief networks. In *International Conference on Machine Learning*, pages 1791–1799. PMLR, 2014.

Emi Nakamura and Jón Steinsson. High-frequency identification of monetary non-neutrality: the information effect. *The Quarterly Journal of Economics*, 133(3):1283–1330, 2018.

Ali Ozdagli and Michael Weber. Monetary policy through production networks: Evidence from the stock market. Technical report, National Bureau of Economic Research, 2017.

Lin Peng and Wei Xiong. Investor attention, overconfidence and category learning. *Journal of Financial Economics*, 80(3):563–602, 2006.

James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001, 2001.

Joel Peress. Media coverage and investors' attention to earnings announcements. *Available at SSRN 2723916*, 2008.

Joel Peress. The media and the diffusion of information in financial markets: Evidence from newspaper strikes. *The Journal of Finance*, 69(5):2007–2043, 2014.

H Hashem Pesaran and Yongcheol Shin. Generalized impulse response analysis in linear multivariate models. *Economics letters*, 58(1):17–29, 1998.

Daniel Preoţiuc-Pietro, Ye Liu, Daniel Hopkins, and Lyle Ungar. Beyond binary labels: political ideology prediction of twitter users. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 729–740, 2017.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D Manning. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 248–256. Association for Computational Linguistics, 2009.

Ricardo Reis. Central bank design. *Journal of Economic Perspectives*, 27(4):17–44, 2013.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1278–1286, Bejing, China, 22–24 Jun 2014. PMLR.

Alexander W Richter and Nathaniel A Throckmorton. The zero lower bound: frequency, duration, and numerical convergence. *The BE Journal of Macroeconomics*, 15(1):157–182, 2015.

Christian Robert and George Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

Margaret E Roberts, Brandon M Stewart, and Edoardo M Airoldi. A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, 111(515):988–1003, 2016.

Richard Roll. R2. *Journal of Finance*, 43(3):541–566, 1988.

Christina D Romer and David H Romer. Does monetary policy matter? A new test in the spirit of Friedman and Schwartz. *NBER macroeconomics annual*, 4:121–170, 1989.

Christina D Romer and David H Romer. Federal reserve information and the behavior of interest rates (digest summary). *American Economic Review*, 90(3):429–457, 2000.

Daniel Schmidt. Investors' attention and stock covariation. Technical report, Working paper, HEC Paris, 2013.

Takahiro Shinozaki and Mari Ostendorf. Cross-validation em training for robust parameter estimation. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4, pages IV–437. IEEE, 2007.

Leon V Sigal. *Reporters and officials: The organization and politics of newsmaking*. DC Heath, 1973.

Christopher A Sims. Solving linear rational expectations models. *Computational Economics*, 20(1):1–20, 2002.

Christopher A Sims. Rational inattention and monetary economics. In *Handbook of Monetary Economics*, volume 3, pages 155–181. Elsevier, 2010.

Frank Smets and Rafael Wouters. Shocks and frictions in us business cycles: A bayesian dsge approach. *American economic review*, 97(3):586–606, 2007.

David H Solomon, Eugene Soltes, and Denis Sosyura. Winners in the spotlight: Media coverage of fund holdings as a driver of flows. *Journal of Financial Economics*, 113(1): 53–72, 2014.

Timm O Sprenger, Philipp G Sandner, Andranik Tumasjan, and Isabell M Welpe. News or noise? using twitter to identify and understand company-specific news flow. *Journal of Business Finance & Accounting*, 41(7-8):791–830, 2014.

Akash Srivastava and Charles Sutton. Autoencoding variational inference for topic models. *arXiv preprint arXiv:1703.01488*, 2017.

Alan Stuart, Steven Arnold, J Keith Ord, Anthony O'Hagan, and Jonathan Forster. *Kendall's advanced theory of statistics*. Wiley, 1994.

Paul C Tetlock. Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3):1139–1168, 2007.

Paul C Tetlock. The role of media in finance. In *Handbook of media Economics*, volume 1, pages 701–721. Elsevier, 2015.

Carin Van Der Cruijsen, David-Jan Jansen, and Jakob De Haan. How much does the public know about the ECB's monetary policy? Evidence from a survey of Dutch households. 2010.

Ike Vayansky and Sathish AP Kumar. A review of topic modeling methods. *Information Systems*, 94:101582, 2020.

Alessandro T Villa and Vytautas Valaitis. Machine learning projection methods for macro-finance models. *Economic Research Initiatives at Duke (ERID) Working Paper Forthcoming*, 2019.

Xinyi Wang and Yi Yang. Neural topic model with attention for supervised learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1147–1156. PMLR, 2020.

Yining Wang and Jun Zhu. Spectral methods for supervised topic models. In *Advances in Neural Information Processing Systems*, pages 1511–1519, 2014.

Michael Woodford. Learning to believe in sunspots. *Econometrica: Journal of the Econometric Society*, pages 277–307, 1990.

Xue Zhang, Hauke Fuehres, and Peter A Gloor. Predicting stock market indicators through twitter "i hope it is not as bad as i fear". *Procedia-Social and Behavioral Sciences*, 26:55–62, 2011.

Jun Zhu, Amr Ahmed, and Eric P Xing. Medlda: maximum margin supervised topic models. *Journal of Machine Learning Research*, 13(Aug):2237–2278, 2012.

Appendix A

Appendix to Chapter 1

A.1 Levenberg Marquardt Algorithm

Consider a neural network parameterised by θ with output $\hat{x}_t$, input z_{t-1} and trained on T observations. The sum of squared error loss for this network would thus be

$$L = \sum_{t=1}^{T} (x_t - \hat{x}_{t\theta})^2 = e_t^2$$

As the loss function is a non-linear function of the parameters, a closed form training algorithms is not possible. Instead, training consists of searching through the parameter space from some initialised point, adjusting parameters to decrease loss at each step.

The Levenberg-Marquardt algorithm, also known as dampened least-squares, uses an approximate Hessian matrix to achieve close-to second order training speed without the computational cost of computing the Hessian matrix. Define $\mathbf{J}$, where $\mathbf{J}_{t,j} = \frac{\partial e_t}{\partial \theta_j}$, as the Jacobian matrix of errors w.r.t. the parameters. If there are T observations in the data and J parameters in the neural network, this Jacobean is an $T \times J$ matrix. The Jacobean can be computed through a backpropagation technique described in Hagan and Menhaj (1994) that is much less computationally expensive than computing the Hessian matrix. The gradient of the loss function can then be computed as

$$\nabla f = 2\mathbf{J}' \cdot e$$

where e is the vector of all error terms. The Hessian matrix can be approximated with the following expression

$$\mathbf{H}f \approx 2\mathbf{J}' \cdot \mathbf{J} + \mu I$$

where μ is a damping factor that ensures the positive definiteness of the Hessian and I is the identity matrix.

The Levenberg-Marquardt update formula is conceptually similar to a Newton method in that it uses an approximation of the second derivative of the loss function to find better training directions. The parameter update process is

$$\theta^{(i+1)} = \theta^{(i)} + (\mathbf{J}^{(i)\ \prime} \cdot \mathbf{J}^{(i)} + \mu^{(i)} I)^{-1} \cdot (2\mathbf{J}^{(i)\ \prime} \cdot e^{(i)}) \tag{1.13}$$

When the damping parameter μ is zero, this is just Newton's method with an approximate Hessian matrix. When μ is large, this becomes gradient descent with a small step size. As Newton's method is faster and more accurate in the region of a minimum, the aim of the damping is to shift towards Newton's method as quickly as possible. Note that this is remarkably similar to the update process for Recursive Least Squares.

$$\beta_t = \beta_{t-1} + \gamma^{-1}\Sigma_t^{-1} z_{t-1}(x_t - \beta'_{t-1} z_{t-1}) \tag{A.1}$$

We can therefore interpret the neural network as a generalisation of ordinary least squares regression.

The dampening parameter μ is initialised to be large so that the first updates are small steps in the gradient descent direction and if any iteration results in failure then μ is increased by some factor, I choose a default value of 10. After each successful update, μ is decreased by some factor, I choose a default of 0.1. This means that the Levenberg-Marquardt algorithm approaches the Newton method as it gets closer to a minimum, which accelerates convergence.

The training process proceeds as follows.

1. Initialise the neural network's adaptive parameters $\theta^{(0)}$ and the dampening parameter μ.
2. Repeat the following until the stopping criteria are met
   (a) Adjust the dampening parameter
   (b) Calculate the gradient and Hessian approximation
   (c) Improve the neural network parameters
   (d) Evaluate the loss index

Levenberg-Marquardt is ideal for medium-scale neural networks that are evaluated by sum of squared error and typically outperforms the other algorithms on speed, but not

necessarily on memory requirements (Hagan et al., 1997).[1]

In order to reduce computational time, while maximising performance, I adopt a multi-conditional stopping criteria. This allows for the fact that early in the learning process there may not be a model that matches the data well as it generated by an evolving expectations formation process. The simulated data for a given updated is divided into a training sample (60%) a validation sample (20%) and a test sample (20%). Training then stops if any of the following five conditions are met: the maximum number of iterations (1,000) is reached; the loss function is minimized to zero; the gradient $2\mathbf{J}^T \cdot e$ falls below $1 \times e^{-7}$; μ exceeds a maximum level of $1 \times e^{10}$; or the loss function evaluated on the validation set has increased more than 6 times since the last time it decreased. The last of these will perhaps be least familiar to economists, but is common in the machine learning literature. In using a validation to stop training, the algorithm prevents overfitting and wasting computational time with small updates in the region of the minimum.

A.2 General results for one-step vs two-step ahead solution

In linear models, forecasting with a one-step ahead PLM such as Equation A.2 which is rolled forward nests the same REE as forecasting with a two-step ahead PLM such as Equation A.3. In a non-linear model however, this equivalence breaks down. The one-step ahead approach is not able to capture non-linear transformations of future (and therefore unobserved) errors. The two-step ahead approach is therefore used in neural network learning.

$$x_t = \begin{cases} a_1 + b_1 x_{t-1} + c_1 \epsilon_t & \text{if linear} \\ f_{p,1}(x_{t-1}, \epsilon_t) & \text{otherwise} \end{cases} \tag{A.2}$$

$$x_{t+1} = \begin{cases} a_2 + b_2 x_{t-1} + c_2 \epsilon_t + d_2 \epsilon_{t+1} & \text{if linear} \\ f_{p,2}(x_{t-1}, \epsilon_t, \epsilon_{t+1}) & \text{otherwise} \end{cases} \tag{A.3}$$

[1] Levenberg-Marquardt is tailored to be computationally efficient for functions with sum-of-squared type errors, but the algorithm does have some drawbacks. It cannot be applied to alternative loss functions, it is not compatible with regularization terms and for high dimensional problems the Jacobean becomes prohibitively large.

A.2.1 Equivalence in linear case

Consider a linear model of the form given in Equation A.4.[2]

$$x_t = A + B\mathbb{E}x_{t+1} + Cx_{t-1} + D\epsilon_t \tag{A.4}$$

As the shocks are mean-zero, the one-step ahead PLM will be of the form $x^e_{t+1|t} = a_1 + b_1 x_t$. By substituting this in for the expectation to find the ALM and then solving using the method of undetermined coefficients gives

$$a_1 = (I - Bb_1)^{-1}(A + Ba_1), \quad b_1 = (I - Bb_1)^{-1}C, \quad c_1 = (I - Bb_1)^{-1}D \tag{A.5}$$

which defines the MSV solution, as well as the T-map used to assess E-stability. In the two-step ahead case, substituting A.3 into A.4 and solving by undetermined coefficients gives

$$\begin{aligned} a_2 = (I + Bb_2 + C)(A + Ba_2), \quad b_2 = (Bb_2 + C)^2, \\ c_2 = (Bb_2 + C)(Bc_2 + D), \quad d_2 = (Bc_2 + D) \end{aligned} \tag{A.6}$$

To compare the ALMs implied by the two different methods, iterate forward A.2 to give the mapping from x_{t-1} to x_{t+1} implied by the 1 step ahead method.

$$x_{t+1} = a_1 + b_1 a_1 + b_1^2 x_{t-1} + b_1 c_1 \epsilon_t + c_1 \epsilon_{t+1}$$

Therefore, for the two to be equivalent

$$a_2 = a_1 + b_1 a_1, \quad b_2 = b_1^2, \quad c_2 = b_1 c_1, \quad d_2 = c_1 \tag{A.7}$$

If these implicit solutions in A.5 and A.6 are consistent with the equivalency conditions A.7 then it is possible to show that the two-step solutions nest the one-step ahead solutions. Premultiplying the definition of b_1 in A.2 and rearranging gives

$$(I - Bb_1)b_1 = C \quad \rightarrow \quad b_1 = Bb_1^2 + C$$

Squaring this both sides of this equation yields the equivalency condition for b. The other conditions then follow after some simple algebra. The two-step ahead approach will likely yield some extra solutions which are not possible with the one-step ahead approach, but it nests all possible one-step ahead solutions, including the MSV solution.

[2]This matches that used by Evans and Honkapohja (2001) and shown by McCallum (2007) to be equivalent to the form used by King and Watson (1998) and Klein (2000) which can include any number of lags

A.2.2 Non-equivalence in the non-linear case

In general, the 2-step ahead and iterative approaches to learning will not be equivalent in a non-linear model. Suppose an economic model takes the form

$$x_t = g(\mathbb{E}_t(x_{t+1}), x_{t-1}, \epsilon_t)$$

and the PLM takes the form

$$x^e_{t|t} = f_1(x_{t-1}, \epsilon_t)$$

The one-step ahead approach would then assume that

$$x^e_{t+1|t} = f_1(x^e_{t|t}, \mathbb{E}_t\epsilon_{t+1}) = f_1(f_1(x_{t-1}, \epsilon_t), \mathbb{E}_t\epsilon_{t+1})$$

Under REE, it should be the case that $\mathbb{E}_t x_{t+1} = x^e_{t+1|t}$. Assuming that the PLM is correct in the first step,

$$x_t = \mathbb{E}_t x_t = x^e_{t|t} = f_1(x_{t-1}, \epsilon_t)$$

Then by Jensen's Inequality

$$\mathbb{E}_t x_{t+1} = \mathbb{E}_t f_1(x_t, \epsilon_{t+1}) \neq f_1(x_t, \mathbb{E}_t\epsilon_{t+1})$$

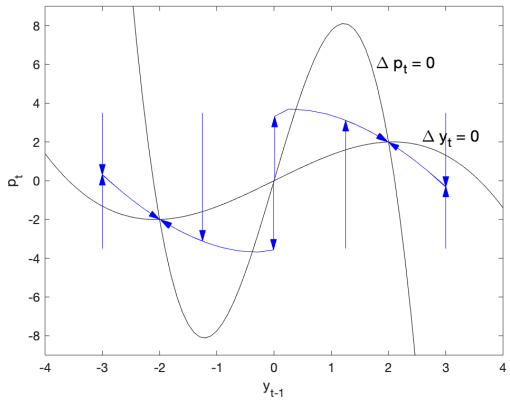
However, if the agents' PLM explicitly predicts two steps ahead, then this is not a problem as

$$f_2(x_{t-1}, \epsilon_t) = \mathbb{E}[x_{t+1}|x_{t-1}, \epsilon_t] \neq f_1(x_t, \mathbb{E}[\epsilon_{t+1}])$$

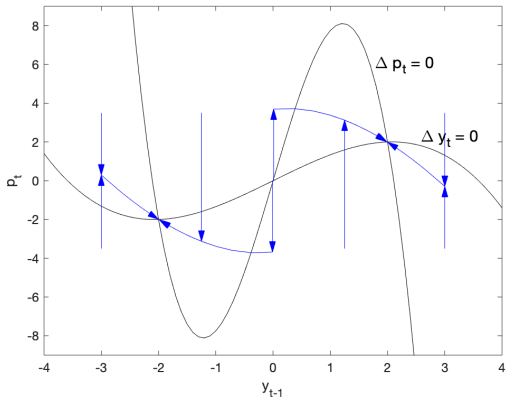
Figure A.1 demonstrates how the two different approaches will lead to different equilibria, showing phase diagrams for the PLM and ALM which neural network learning converges to in the one-step ahead forecast case.

Figure A.1: Neural network simulation results for complex sink case

(a) PLM with one-step ahead forecasts



(b) ALM with two-step ahead forecasts



The solution in this one-step ahead case does not pass the DHM test for accuracy of solution, as the agents are not taking into account the higher order moments of future shocks in their forecasting.

A.3 Mean dynamics under RLS in illustrative model

As the $\pi^e_{t+1|t}$ depends only on π_{t-1} and y_{t-1}, it can be treated as fixed at the beginning of a period and the probability of (π_t, y_t) being in a given range is simply a function of the instantiation of the two shocks, as expressed by Equations 1.1, 1.2 and 1.3.

$$\pi_t = \beta\pi^e_{t+1|t} + \kappa y_t + \epsilon_{\pi,t} \tag{A.8}$$

$$y_t = \eta y_{t-1} - \sigma(\phi_\pi \pi_t + \pi_t^3 - \pi^e_{t+1|t}) + \epsilon_{y,t} \tag{A.9}$$

By discretising the state space we can use these to obtain an arbitrarily accurate approximation of the distribution of $X_t = (\pi_t \ y_t \ \pi_{t-1} \ y_{t-1})$ given θ_{t-1}. This discretisation obviates the need to solve these as a non-linear system of equations, as it is possible to simply plug in values for $\bar{\pi}_t$, $\bar{y}_t$, $\bar{y}_{t-1}$ and $\bar{\pi}^e_{t+1|t}$ and then calculate the $\epsilon^*_{\pi,t}$ and $\epsilon^*_{y,t}$ necessary to achieve them.

$$\epsilon^*_{\pi,t} = \bar{\pi}_t - \beta\bar{\pi}^e_{t+1|t} - \kappa\bar{y}_t \tag{A.10}$$

$$\epsilon^*_{y,t} = \bar{y}_t - \eta\bar{y}_{t-1} + \sigma(\phi_\pi\bar{\pi}_t + \bar{\pi}_t^3 - \bar{\pi}^e_{t+1|t}) \tag{A.11}$$

As $\epsilon_{\pi,t}$ and $\epsilon_{y,t}$ are independently distributed Gaussian distributions, and assuming for illustrative purposes that $\rho_\pi = \rho_y = 0$, the conditional probability of given values of ϵ_y

and ϵ_p is

$$f_{\epsilon_\pi,\epsilon_y}(\epsilon_{\pi,t},\epsilon_{y,t}) = \frac{1}{2\pi\sigma_\pi\sigma_y} e^{\left(\frac{\epsilon_{\pi,t}^2}{2\sigma_\pi^2}+\frac{\epsilon_{y,t}^2}{2\sigma_y^2}\right)} \tag{A.12}$$

To obtain a probability for this transition we can compute the the probability of $\epsilon_{\pi,t}$ and $\epsilon_{y,t}$ being in some range centred at ϵ^*, i.e. $\epsilon^*_{\pi/y,t} - \omega \leq \epsilon_{\pi/y,t} \geq \epsilon^*_{\pi/y,t} + \omega$. The ω parameter determining the size of this range is scaled with the fineness of the grid into which the state space is discretised. This approximation will become arbitrarily good as the grid becomes finer. The probability of $(\epsilon_{\pi,t},\ \epsilon_{y,t})$ being in a certain range is thus

$$F_{ds}(\epsilon_{\pi,t} \leq \epsilon^*_{\pi,t} + \omega, \epsilon_{y,t} \leq \epsilon^*_{y,t} + \omega) = F_d(\epsilon_{\pi,t} \leq \epsilon^*_{\pi,t} + \omega) F_s(\epsilon_{y,t} \leq \epsilon^*_{y,t} + \omega)$$

This can then be used to obtain a conditional probability of each π_t, y_t pair for a given π_{t-1} and y_{t-1}.

These conditional probabilities can be represented as a transition probability matrix. If we were to discretise the state space into only two values -1 and +1, we would then have four (2^2) possible states, i.e. $s_1 \to (\pi_{t-1}, y_{t-1}) = (-1,-1)$ and $s_2 \to (\pi_{t-1}, y_{t-1}) = (-1,+1)$. Using the joint normal distribution above we can calculate transition probabilities between states. For example ,z_{21} is the probability of transitioning from s_2 to s_1.

Given the $n^m \times n^m$ transition probability matrix P, where m is the number of observed state variables and n the number of bins in the grid, we can compute the marginal probability of being in a given initial state (π_{t-1}, y_{t-1}) for a given θ_{t-1}. The stationary distribution of a Markov chain with transition probability matrix P is given by the eigenvector associated with a unit eigenvalue of P. Define M as this eigenvector, which gives us the marginal probability of each (initial) state. We thus have conditional distribution of $\pi_t, y_t | \pi_{t-1}, y_{t-1}$ from the transition probability matrix P and the unconditional distribution of π_{t-1}, y_{t-1} from the eigenvector M.

Applying Bayes Rule will thus give the invariant distribution of $\pi_t, y_t, \pi_{t-1}, y_{t-1}$ which is the $\Gamma_\theta(dx)$ object we need to compute $h(\theta)$.

$$\Gamma_\theta(dx) = P \otimes M$$

This gives us the unconditional probability of observing each possible value of X_t, i.e. the probability of observing $(\pi_t, y_t, \pi_{t-1}, y_{t-1})$ within a particular range, given the beliefs θ_{t-1}.

Recall that

$$h(\theta) = \int \mathcal{H}(\theta, x)\Gamma_\theta(dx) \tag{1.15}$$

where $\mathcal{H}(\theta, x)$ described the updating of θ for a given initial θ and observations x_t. This is derived from the definition of RLS learning

$$\beta_t = \beta_{t-1} + \gamma^{-1}\Sigma_t^{-1} z_{t-1}(x_t - \beta'_{t-1} z_{t-1})$$

$$\Sigma_t = \Sigma_{t-1} + \gamma^{-1}(z_t z'_t - \Sigma_{t-1})$$

For each element of the transition probability matrix, we can calculate the corresponding forecast error given the beliefs θ_{t-1}. This allows us to define $\mathcal{H}(\theta, x)$, i.e. the updating process for a given x_t and z_{t-1}. Multiplying this the unconditional distribution of x given θ gives us the *expected* update of θ.

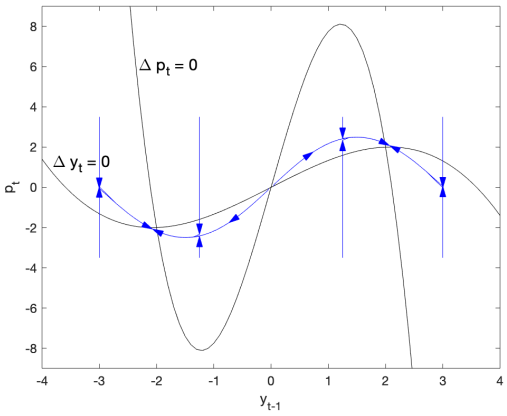
We thus have (an approximation of) the differential equation $\frac{h(\theta)}{\theta}$ defining the mean dynamics of the system under RLS learning. Finding the fixed point of the differential equation using a standard differential equation solver tells us where the system will converge to.

A.3.1 RLS simulation under full information

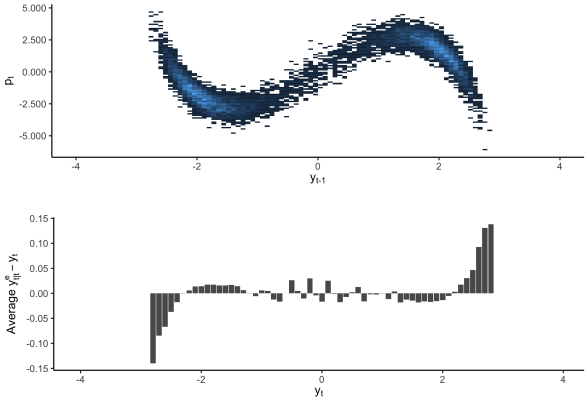
A closed form characterisation of $h(\theta)$ is harder to obtain in the full information case where agents observe the contemporary shocks, but simulations suggest the results are very similar. FigureA.3 shows that agents still make substantially larger forecast errors than in the neural network case and the equilibrium fails DHM test so cannot be characterised as a REE.

Figure A.3: The Perceived and Actual Laws of Motion with Recursive Least Squares mean dynamics

(a) ALM under RLS



(b) Comparing ALM and PLM under RLS



A.4 Alternative parameterisations

To show that results are qualitatively similar whether the central steady state is complex/real or a source/sink under perfect foresight, this Section presents results for a range of cases in a more abstract model. As before, we have one control variable and one state variable, a central steady state around which the model is indeterminate and two outer steady states that bound this indeterminate region.

$$p_t = \phi_{p,p}\mathbb{E}_t p_{t+1} + \phi_{p,y} y_t - \alpha y_t^3 + \epsilon_{p,t} \tag{A.13}$$

$$y_t = \phi_{y,y} y_{t-1} + \phi_{y,p} p_t + \epsilon_{y,t} \tag{A.14}$$

We can identify the deterministic steady states by setting $\Delta p_t = 0$ and $\Delta y_t = 0$ in the perfect foresight model. These two steady state conditions are shown as the black lines in Figures A.6 to A.12.

$$\Delta p_t = \frac{1-\phi_{p,p}}{\phi_{p,p}} p_{t-1} + \frac{\alpha}{\phi_{p,p}} y_{t-1}^3 - \frac{\phi_{p,y}}{\phi_{p,p}} y_{t-1} = 0 \tag{A.15}$$

$$\Delta y_t = \frac{\phi_{y,p}}{\phi_{p,p}} p_{t-1} + \frac{\phi_{y,y}\phi_{p,p} - \phi_{p,p} - \phi_{y,p}\phi_{p,y}}{\phi_{p,p}} y_{t-1} + \frac{\phi_{y,p}}{\phi_{p,p}} y_{t-1}^3 = 0 \tag{A.16}$$

Denoting (p^*, y^*) as one of the model's steady states, linearising around this steady state gives

$$\hat{p}_t = \phi_{p,p}\mathbb{E}_t \hat{p}_{t+1} + (\phi_{p,y} - 3y^{*2}\alpha)\hat{y}_t + \epsilon_{p,t}. \tag{A.17}$$

$$\hat{y}_t = \phi_{y,y}\hat{y}_{t-1} + \phi_{y,p}\hat{p}_t + \epsilon_{y,t} \tag{A.18}$$

As in the New Keynesian example in Section 1.2, at the outer steady states the positive feedback from output to price is reversed, so linearising around these outer steady states creates a determinate model. Defining $\hat{x}_t = (\hat{p}_t \ \hat{y}_{t-1})'$, and $\epsilon_t = (\epsilon_{p,t} \ \epsilon_{y,t})$ we then have a model in the general matrix form.

$$\mathbb{E}_t\hat{x}_{t+1} = \Phi\hat{x}_t + \Psi\epsilon_t \tag{A.19}$$

where

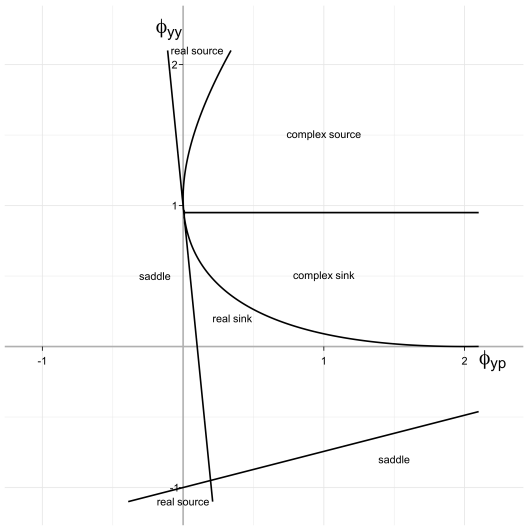
$$\Phi = \begin{pmatrix} \frac{1-(\phi_{p,y}-3y^{*2}\alpha)\phi_{y,p}}{\phi_{p,p}} & \frac{-(\phi_{p,y}-3y^{*2}\alpha)\phi_{y,y}}{\phi_{p,p}} \\ \phi_{y,p} & \phi_{y,y} \end{pmatrix}, \ \Psi = \begin{pmatrix} -\frac{1}{\phi_{p,p}} & -\frac{\phi_{p,y}-3y^{*2}\alpha}{\phi_{p,p}} \\ 0 & 1 \end{pmatrix}$$

Dynamics in the neighbourhood of a steady state are determined by the eigenvalues of the coefficient matrix Φ in Equation A.19. As this is a 2-by-2 matrix, we can find expressions for its eigenvalues in terms of the four parameters $\phi_{p,p}$, $\phi_{p,y}$, $\phi_{y,p}$ and $\phi_{y,y}$.

$$\lambda_{1,2} = \frac{1-(\phi_{p,y}-3y^{*2}\alpha)\phi_{y,p}+\phi_{p,p}\phi_{y,y}}{\phi_{p,p}} \pm \frac{1}{2}\sqrt{\left(\frac{1-(\phi_{p,y}-3y^{*2}\alpha)\phi_{y,p}+\phi_{p,p}\phi_{y,y}}{\phi_{p,p}}\right)^2 - 4\frac{\phi_{y,y}}{\phi_{\pi,\pi}}} \tag{A.20}$$

Figure A.5 shows the ranges in the parameter space for which the perfect foresight system will be complex/real and a sink/source/saddle around the central steady state for different values of $\phi_{y,y}$ and $\phi_{y,p}$, with fixed $\phi_{p,p} = 0.95$ and $\phi_{p,y} = 0.5$.

Figure A.5: Properties of the central steady state in $(\phi_{y,p}, \phi_{y,y})$ space ($\phi_{p,p} = 0.95$, $\phi_{p,y} = 0.5$)



Figures A.6-A.12 shows the perfect foresight system (left hand column) and the Learnable-REE (right hand column) for different parameterisations. More specifically, $\phi_{y,p}$ and $\phi_{y,y}$ are varied so that the central steady state is a complex sink, real sink, complex source and

real source. The phase diagrams show that in all cases the central steady state becomes a source and, depending on the starting point, the system converges to one of the outer two steady states.

Figure A.6: Complex sink ($\phi_{y,p} = 0.1$, $\phi_{y,y} = 0.9$)

(a) Perfect foresight **(b)** Learnable-REE

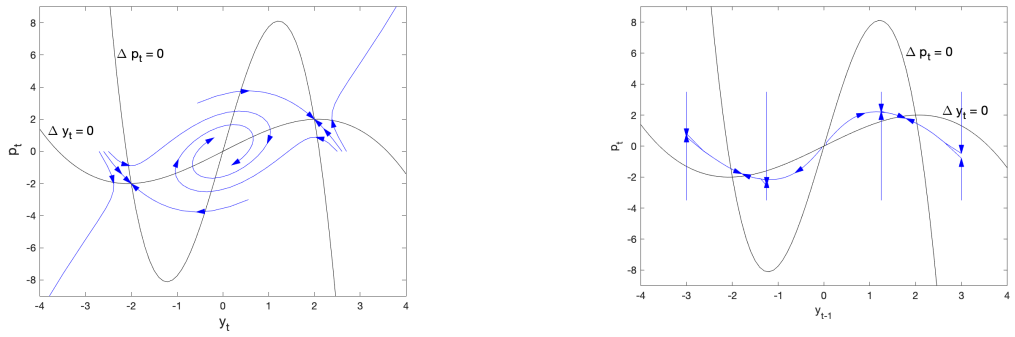


Figure A.8: Real sink ($\phi_{y,p} = 0.9$, $\phi_{y,y} = 0.1$)

(a) Perfect foresight **(b)** Learnable-REE

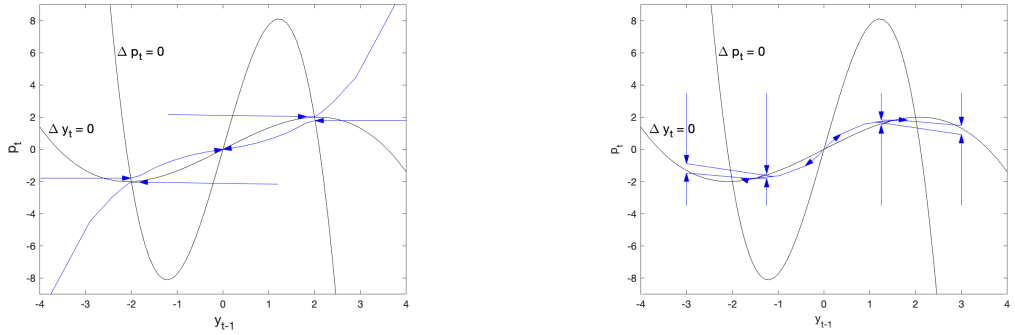


Figure A.10: Complex source ($\phi_{y,p} = 0.1$, $\phi_{y,y} = 1.5$)

(a) Perfect foresight **(b)** Learnable-REE

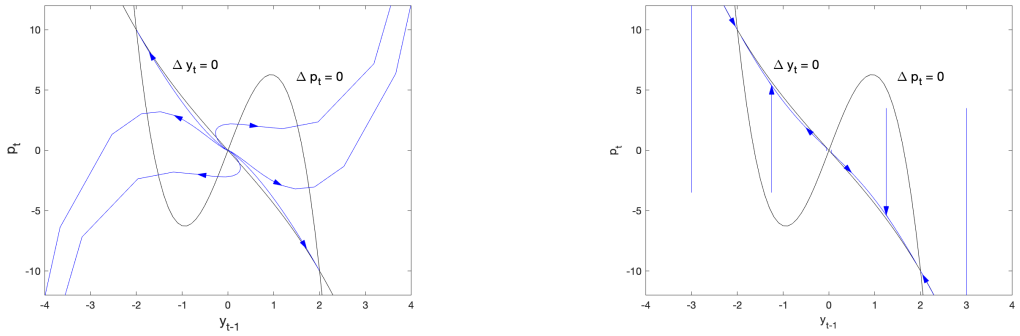
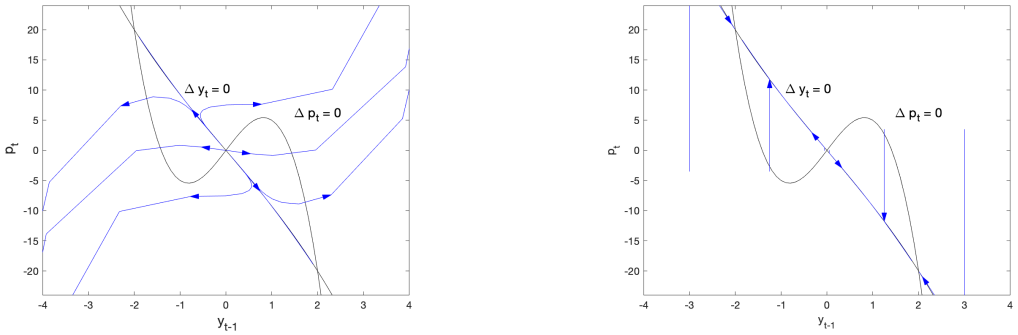


Figure A.12: Real source ($\phi_{y,p} = 0.1$, $\phi_{y,y} = 2$)

(a) Perfect foresight **(b)** Learnable-REE



A.5 Testing for indeterminacy

The linear model with and without sunspots is estimated on data obtained from the simulated neural network learning process.

A.5.1 Specifying a model with sunspots

The linearised illustrative model can be defined in matrix form as in Equation 1.10.

$$\mathbb{E}_t \hat{x}_{t+1} = \Phi \hat{x}_t + \Psi \epsilon_t \tag{1.10}$$

where one variable in $\hat{x}_t$ is forward-looking ($\hat{\pi}_t$) and one is pre-determined ($\hat{y}_{t-1}$). Writing this in terms of forecast errors and using the Jordan decomposition of the matrix $\Phi = V \Lambda V^{-1}$, we can write this as

$$V^{-1}\hat{x}_{t+1} = \Lambda V^{-1}\mathbb{E}_{t-1}\hat{x}_t + \Lambda V^{-1}(\hat{x}_t - \mathbb{E}_{t-1}\hat{x}_t) + V^{-1}\Psi\epsilon_t$$

Which gives two equations in an appropriately defined $\tilde{x}_{1,t}$, $\tilde{x}_{2,t}$, $\tilde{\epsilon}_{1,t}$ and $\tilde{\epsilon}_{2,t}$.[3]

$$\mathbb{E}_t \tilde{x}_{1,t+1} = \lambda_1 \mathbb{E}_{t-1}\tilde{x}_{1,t} + \lambda_1(\tilde{x}_{1,t} - \mathbb{E}_{t-1}\tilde{x}_{1,t}) + \tilde{\epsilon}_{1,t} \tag{A.21}$$

$$\mathbb{E}_t \tilde{x}_{2,t+1} = \lambda_2 \mathbb{E}_{t-1}\tilde{x}_{2,t} + \lambda_2(\tilde{x}_{2,t} - \mathbb{E}_{t-1}\tilde{x}_{2,t}) + \tilde{\epsilon}_{2,t} \tag{A.22}$$

When the system is indeterminate, both λ_1 and λ_2 are inside the unit circle, so neither of these equations are unstable so we cannot pin down a unique solution using the approach of Blanchard and Kahn (1980) or Sims (2002). Farmer et al. (2015) show that we can

[3]Where $\Lambda = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$, $V^{-1} = \begin{pmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{pmatrix}$, $\tilde{x}_{1,t} = v_{11}\hat{p}_t - v_{12}\hat{y}_{t-1}$, $\tilde{x}_{2,t} = v_{21}\hat{p}_t - v_{22}\hat{y}_{t-1}$, $\tilde{\epsilon}_{1,t} = (v_{11}\psi_{11} + v_{12}\psi_{21})\epsilon_{p,t} + (v_{11}\psi_{12} + v_{12}\psi_{22})\epsilon_{y,t}$ and $\tilde{\epsilon}_{2,t} = (v_{21}\psi_{11} + v_{22}\psi_{21})\epsilon_{p,t} + (v_{21}\psi_{12} + v_{22}\psi_{22})\epsilon_{y,t}$.

close the model by defining one of the endogenous forecast errors as exogenous. So we can define

$$\tilde{x}_{1,t} - \mathbb{E}_{t-1}\tilde{x}_{1,t} = \zeta_t$$

This ζ_t is thus a "sunspot" shock: a process that determines expectations and is at least partially independent of fundamentals, but still consistent with rational expectations.

$$\zeta_t = \pi_t - \mathbb{E}_{t-1}\pi_t$$

If we assume that the sunspot shock is also normally distributed, we can solve or estimate the model using standard techniques, just with some additional terms in the covariance matrix. The "exogenous" errors are now $(\tilde{\epsilon}_{1,t} \;\; \tilde{\epsilon}_{1,t} \;\; \zeta_t)'$, so the covariance matrix (assuming $\rho_{\pi/y} = 0$) is

$$\Omega = \mathbb{E}_{t-1}\left[\begin{pmatrix}\tilde{\epsilon}_{1,t}\\ \tilde{\epsilon}_{1,t}\\ \zeta_t\end{pmatrix}\begin{pmatrix}\tilde{\epsilon}_{1,t}\\ \tilde{\epsilon}_{1,t}\\ \zeta_t\end{pmatrix}'\right] = \begin{pmatrix}\sigma_\pi & 0 & \omega_{\pi,\zeta}\\ 0 & \sigma_y & \omega_{y,\zeta}\\ \omega_{\pi,\zeta} & \omega_{y,\zeta} & \sigma_\zeta\end{pmatrix}$$

This nests the cases where the ζ_t "sunspot" error contains either present or past (or future) values of the ϵ_t shock if the covariance matrix is left unrestricted.[4] For given values of $\omega_{\pi,\zeta}$, $\omega_{y,\zeta}$ and σ_ζ, this model has a unique stationary rational expectations solution so can be straightforwardly solved, simulated and estimated.

A.5.2 Bayesian estimation

The models estimated are estimated with Metropolis Hastings MCMC using the Dynare software package, as described in Guerrón-Quintana and Nason (2013). The coefficient parameters are given normal priors, the standard deviations inverse gamma priors, and the correlations between shocks beta priors. Table A.1 displays the estimation results for the two models, and Figure A.14 shows the posterior distributions for each of the estimated parameters.

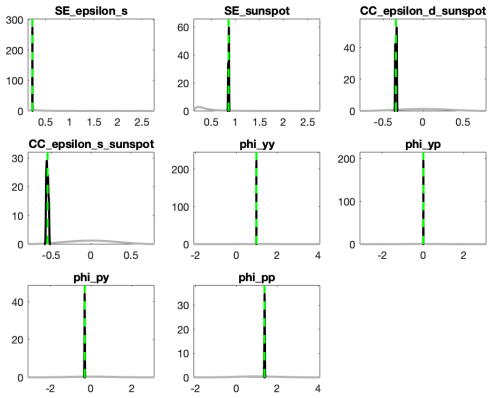
[4]This set-up therefore nests the specification in Lubik and Schorfheide (2004), who explicitly model an AR(1) dependence of forecast errors on the fundamental shocks.

Table A.1: Bayesian estimation of sunspot model on learning data

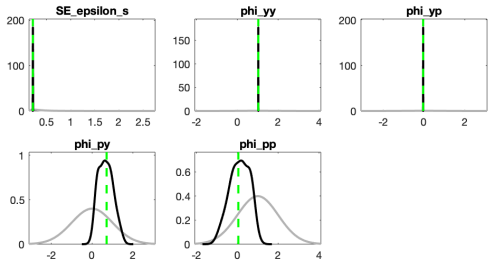
	Sunspot model Log posterior = -5,601.2413			Standard model Log posterior = -25,338.0941		
Par	Post. mean	90 % interval	prior	Post. mean	90 % interval	prior
$\phi_{\pi,p}$	1.383	[1.364, 1.402]	$\mathcal{N}(1,1)$	0.107	[-0.627, 0.789]	$\mathcal{N}(1,1)$
$\phi_{\pi,y}$	-0.315	[-0.332, -0.300]	$\mathcal{N}(0,1)$	0.666	[0.161, 1.214]	$\mathcal{N}(0,1)$
$\phi_{y,\pi}$	0.017	[0.014, 0.021]	$\mathcal{N}(0,1)$	-0.041	[-0.044, -0.038]	$\mathcal{N}(0,1)$
$\phi_{y,y}$	0.980	[0.977, 0.982]	$\mathcal{N}(1,1)$	1.024	[1.021, 1.028]	$\mathcal{N}(1,1)$
σ_y	0.225	[0.223, 0.227]	$\mathcal{IG}(0.5,2)$	0.215	[0.211, 0.218]	$\mathcal{IG}(0.5,2)$
σ_ζ	0.859	[0.846, 0.869]	$\mathcal{IG}(0.5,2)$	-	-	-
$\omega_{\pi,\zeta}$	-0.339	[-0.351, -0.328]	$\mathcal{B}(0.5,2)$	-	-	-

Figure A.14: Posterior distributions for the two models

(a) Model with sunspots



(b) Standard model



A.5.3 Maximum Likelihood estimation

Both models are estimated using the continuous simulated annealing global optimization algorithm built in to the Dynare software package.

Table A.2: Maximum Likelihood estimation of sunspot model on learning data

	Sunspot model Log likelihood = -5554.4820			Standard model Log likelihood = -25,314.3184		
Parameter	Estimate	s.d.	t-stat	Estimate	s.d.	t-stat
$\phi_{p,p}$	1.4075	0.0226	62.3944	0.0892	0.0000	0.0000
$\phi_{p,y}$	-0.3387	0.0208	16.2974	0.6790	0.0000	0.0000
$\phi_{y,p}$	0.0189	0.0028	6.7574	-0.0431	0.0020	20.7762
$\phi_{y,y}$	0.9786	0.0027	356.9166	1.0247	0.0022	462.6641
σ_y	0.2142	0.0022	85.5491	0.2142	0.0022	98.9930
σ_ζ	0.8733	0.0123	71.1330	-	-	-
$\omega_{p,\zeta}$	-0.4106	0.0342	11.9958	-	-	-
$\omega_{y,\zeta}$	-0.5579	0.0148	37.6007	-	-	-

Appendix B

Appendix to Chapter 2

B.1 Gibbs-EM algorithm

The probability of a given word $w_{d,n}$ being assigned to a given topic k (such that $z_{d,n} = k$), conditional on the assignments of all other words (as well as the model's other latent variables and the data) is

$$p(z_{d,n} = k|\boldsymbol{Z}_{-(d,n)}, \boldsymbol{W}, \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\omega}, \sigma^2), \tag{B.1}$$

where $\boldsymbol{Z}_{-(d,n)}$ are the topic assignments for all words apart from $w_{d,n}$. By the conditional independence properties implied by the graphical model, we can split this joint posterior into

$$p(\boldsymbol{Z}|\boldsymbol{W}, \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\omega}, \sigma^2) \propto p(\boldsymbol{Z}|\boldsymbol{W})p(\boldsymbol{y}|\boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{\omega}, \sigma^2). \tag{B.2}$$

As topic assignments within one document are independent from topic assignments in all other documents, the sampling equation for the nth word in document d should only depend it's own response variable, y_d, such that (as the parameters ω and σ^2 are conditioned on)

$$\begin{aligned} &p(z_{d,n} = k|\boldsymbol{Z}_{-(d,n)}, \boldsymbol{W}, \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\omega}, \sigma^2) \propto \\ &p(z_{d,n} = k|\boldsymbol{Z}_{-(d,n)}, \boldsymbol{W})p(y_d|z_{d,n} = k, \boldsymbol{Z}_{-(d,n)}, \boldsymbol{x}_d, \boldsymbol{\omega}, \sigma^2). \end{aligned} \tag{2.13}$$

B.1.1 $p(z_{d,n} = k|\mathbf{Z}_{-(d,n)}, \mathbf{W})$

The first part of the RHS in Equation 2.13 is just the sampling distribution of a standard LDA model. This sampling distribution is derived in detail in Appendix D.2.

It can be expressed in terms of the count variables $\boldsymbol{s}$ (the topic assignments across a

document) and $\boldsymbol{m}$ (the assignments of unique words across topics over all documents). $s_{d,k}$ measures the total number of words in document d assigned to topic k and $s_{d,k,-n}$ the number of words in document d assigned to topic k, except for word n. Analogously, $m_{k,v}$ measures the total number of times term v is assigned to topic k across all documents and $m_{k,v,-(d,n)}$ measures the same, but excludes word n in document d.

$$p(z_{d,n} = k|\boldsymbol{Z}_{-(d,n)}, \boldsymbol{W}) \propto (s_{d,k,-n} + \alpha)\frac{m_{k,v,-(d,n)} + \eta}{\sum_v m_{k,v,-(d,n)} + V\eta}. \tag{B.3}$$

B.1.2 $p(y_d|z_{d,n} = k, \mathbf{Z}_{-(d,n)}, \mathbf{x}_d, \omega, \sigma^2)$

Given that the residuals are Gaussian, the probability of the response variable for a given document d is

$$p(y_d|\boldsymbol{z}_d, \boldsymbol{x}_d, \boldsymbol{\omega}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_d - \boldsymbol{\omega}^\intercal \boldsymbol{a}_d)^2}{2\sigma^2}\right\}. \tag{B.4}$$

We can write this in a convenient form that preserves proportionality with respect to $z_{d,n}$ such that it depends only on the data and count variables used in the other two terms. First, we split the $\boldsymbol{x}_d$ features into those that are interacted, $\boldsymbol{x}_{1,d}$, and those that are not, $\boldsymbol{x}_{2,d}$. The generative model for y_d is then

$$y_d \sim \mathcal{N}(\boldsymbol{\omega}_z^\intercal \bar{\boldsymbol{z}}_d + \boldsymbol{\omega}_{zx}^\intercal(\boldsymbol{x}_{1,d} \otimes \bar{\boldsymbol{z}}_d) + \boldsymbol{\omega}_x^\intercal \boldsymbol{x}_{2,d}, \sigma^2). \tag{B.5}$$

where $\otimes$ is the Kronecker product. Noting that $\boldsymbol{X}$ is observed, so we can think of this as a linear model with document-specific regression parameters. Define $\tilde{\boldsymbol{\omega}}_{z,d}$ as a length K vector such that

$$\tilde{\omega}_{z,d,k} = \omega_{z,k} + \boldsymbol{\omega}_{zx,k}^\intercal \boldsymbol{x}_{1,d}. \tag{B.6}$$

Noting that $\tilde{\boldsymbol{\omega}}_{z,d}^\intercal \bar{\boldsymbol{z}}_d = \frac{\tilde{\boldsymbol{\omega}}_{\boldsymbol{z,d}}^\intercal}{N_d}(\boldsymbol{s}_{d,-n} + \boldsymbol{s}_{d,n})$, the probability density of y conditional on $z_{d,n} = k$ is therefore proportional to

$$p(y_d|z_{d,n} = k, \boldsymbol{z}_{-(d,n)}, \boldsymbol{x}_d, \boldsymbol{\omega}, \sigma^2) \propto$$
$$\exp\left\{\frac{1}{2\sigma^2}\left(\frac{2\tilde{\omega}_{z,d,k}}{N_d}\left(y_d - \boldsymbol{\omega}_x^\intercal \boldsymbol{x}_d - \frac{\tilde{\boldsymbol{\omega}}_{z,d}^\intercal}{N_d}\boldsymbol{s}_{d,-n}\right) - \left(\frac{\tilde{\omega}_{z,d,k}}{N_d}\right)^2\right)\right\}. \tag{B.7}$$

This gives us the sampling distribution for $z_{d,n}$ stated in Equation (2.13): a multinomial distribution parameterised by

$$p(z_{d,n} = k|\boldsymbol{Z}_{-(d,n)}, \boldsymbol{W}, \boldsymbol{X}, \boldsymbol{y}, \alpha, \eta, \boldsymbol{\omega}, \sigma^2) \propto (s_{d,k,-n} + \alpha)\frac{m_{k,v,-(d,n)} + \eta}{\sum_v m_{k,v,-(d,n)} + V\eta} \times \exp\left\{\frac{1}{2\sigma^2}\left(\frac{2\tilde{\omega}_{z,d,k}}{N_d}\left(y_d - \boldsymbol{\omega}_x^\intercal \boldsymbol{x}_{2,d} - \frac{\tilde{\boldsymbol{\omega}}_{z,d}^\intercal}{N_d}\boldsymbol{s}_{d,-n}\right) - \left(\frac{\tilde{\omega}_{z,d,k}}{N_d}\right)^2\right)\right\}. \tag{B.8}$$

This defines for each $k \in \{1, ..., K\}$ the probability that $z_{d,n}$ is assigned to that topic. These K probabilities define the multinomial distribution from which $z_{d,n}$ is drawn.

θ and β

Given topic assignments z, we can recover the latent variables θ and β from their predictive distributions via

$$\hat{\theta}_{d,k} = \frac{s_{d,k} + \alpha}{\sum_k (s_{d,k} + \alpha)} \tag{B.9}$$

and

$$\hat{\beta}_{k,v} = \frac{m_{k,v} + \eta}{\sum_v (m_{k,v} + \eta)}. \tag{B.10}$$

B.2 Multiple documents per observation

If $\boldsymbol{x_{d,p}}$ only enters linearly into the regression then some document-level average will have to be used and this transformation can be performed prior to estimation, converting it into an $\boldsymbol{x}_{1,d}$, and so the algorithm will remain unchanged. However, if any of the $\boldsymbol{x}_{d,p}$ variables are interacted with $\bar{\boldsymbol{z}}_{d,p}$ then we may wish for this interaction to be at the paragraph level. For example, if we think that a topic might have a different effect depending on the sentiment of the surrounding paragraph. In this case, we still need to aggregate the interaction to the document level, but aggregate after interacting rather than interacting after aggregating. We therefore define

$$\overline{\boldsymbol{x}_{d,p} \otimes \boldsymbol{z}_{d,p}} = \frac{1}{N_d} \sum_{p \in [P_d]} \sum_{n \in [N_{d,p}]} [\boldsymbol{x}_{d,p} \otimes \boldsymbol{s}_{d,p,n}] \tag{B.11}$$

where $[N]$ denotes the set of integers $\{1, ..., N\}$ and $\otimes$ represents the Kronecker product. The design matrix $\boldsymbol{A}$ is then

$$\boldsymbol{A} = \begin{bmatrix} \bar{\boldsymbol{z}_1} & \overline{\boldsymbol{x}_{1,1,p} \otimes \boldsymbol{z}_{1,p}} & \boldsymbol{x}_{2,1} \\ \vdots & \vdots & \vdots \\ \bar{\boldsymbol{z}}_1 & \overline{\boldsymbol{x}_{1,d,p} \otimes \boldsymbol{z}_{d,p}} & \boldsymbol{x}_{2,d} \\ \vdots & \vdots & \vdots \\ \bar{\boldsymbol{z}}_1 & \overline{\boldsymbol{x}_{1,D,p} \otimes \boldsymbol{z}_{D,p}} & \boldsymbol{x}_{2,D} \end{bmatrix} \tag{B.12}$$

and the predictive model for y_d will be

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{\omega}, \sigma^2) \text{ where } \boldsymbol{\omega} = (\boldsymbol{\omega}_z, \boldsymbol{\omega}_{zx}, \boldsymbol{\omega}_x) \tag{B.13}$$

The simplest way to aggregate from paragraphs to documents is simply to give each word in the document equal weight as above. This will mean that longer paragraphs have greater weight than shorter ones.

As before, we can collapse out the latent variables $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ so that we only need to sample for the topic assignments $\boldsymbol{z}$ in an E-step and then for $\boldsymbol{\omega}$ and σ^2 in an M-step.

In the E-step, we need to sample from the conditional posterior for the topic assignment of each word

$$\Pr[z_{d,p,n} = k | \boldsymbol{Z}_{d,-(p,n)}, \boldsymbol{W}, \alpha, \eta, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\omega}, \sigma^2]. \tag{B.14}$$

By the conditional independence properties of the graphical model, we can split this into $p(\boldsymbol{Z}|\boldsymbol{W}, \alpha, \eta)$ and $p(\boldsymbol{y}|\boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{\omega}, \sigma^2)$. The sampling equation for the nth token in the pth paragraph of the dth document d will have the form

$$\begin{aligned} &\Pr[z_{d,p,n} = k | \boldsymbol{Z}_{d,-(p,n)}, \boldsymbol{W}, \alpha, \eta, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\omega}, \sigma^2] \propto \\ &\Pr[z_{d,p,n} = k | \boldsymbol{Z}_{d,p,-(n)}, \boldsymbol{W}, \alpha, \eta] \quad \times \quad \Pr[y_d | z_{d,p,n} = k, \boldsymbol{Z}_{d,-(p,n)}, \boldsymbol{x}_d, \omega, \sigma^2]. \end{aligned} \tag{B.15}$$

The topic assignment each document is independent, but there are dependencies across paragraphs. Crucially, these paragraphs have are independent with respect to $\boldsymbol{\theta}$, so $p(\boldsymbol{Z}|\boldsymbol{W}, \alpha, \eta)$ is paragraph specific.

$$\Pr[z_{d,p,n} = k | \boldsymbol{Z}_{d,p-(n)}, \boldsymbol{W}, \alpha, \eta] \propto (s_{d,p,k,-n} + \alpha) \frac{m_{k,v,-(d,p,n)} + \eta}{\sum_v m_{k,v,-(d,p,n)} + V\eta} \tag{B.16}$$

However, the regression part is at the document level to $p(\boldsymbol{y}|\boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{\omega}, \sigma^2)$ will condition on all the paragraphs in a given document. Given that the residuals are Gaussian, the probability of the outcome variable for a given document d is

$$p(\boldsymbol{y}_d|\boldsymbol{z}_d, \boldsymbol{x}_d, \boldsymbol{\omega}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y_d - \boldsymbol{\omega}_z^\intercal \bar{\boldsymbol{z}}_d - \boldsymbol{\omega}_{zx}^\intercal(\overline{\boldsymbol{x}_{1,d,p} \otimes \boldsymbol{z}_{d,p}}) - \boldsymbol{\omega}_x^\intercal \boldsymbol{x}_{2,d})^2}{2\sigma^2}\right] \tag{B.17}$$

We can write this in a convenient form that preserves proportionality with respect to $z_{d,p,n}$ such that it depends only on the data and count variables used in the other two terms and the document-wide counts. First we can break the prediction for y_d into the section that depends on paragraph p and the section that depends on other paragraphs and document wide $\boldsymbol{x}_{1,d}$.

$$\begin{aligned} y_d - \boldsymbol{\omega}_z^\intercal \bar{\boldsymbol{z}}_d - \boldsymbol{\omega}_{zx}^\intercal(\overline{\boldsymbol{x}_{1,d,p} \otimes \boldsymbol{z}_{d,p}}) = \left(y_d - \boldsymbol{\omega}_x^\intercal \boldsymbol{x}_{2,d} - \frac{\boldsymbol{\omega}_z^\intercal}{N_d}\boldsymbol{s}_{d,-p} - \frac{\boldsymbol{\omega}_{zx}^\intercal}{N_d}\sum_{q\in\{[P_d]\setminus p\}}[\boldsymbol{x}_{1,d,q} \otimes \boldsymbol{s}_{d,q}]\right) - \\ \left(\frac{\boldsymbol{\omega}_z^\intercal}{N_d}(\boldsymbol{s}_{d,p,-n} + \boldsymbol{s}_{d,p,n}) - \frac{\boldsymbol{\omega}_{zx}^\intercal}{N_d}\boldsymbol{x}_{1,d,p} \otimes (\boldsymbol{s}_{d,p,-n} + \boldsymbol{s}_{d,p,n})\right) \end{aligned} \tag{B.18}$$

where N_d is the total number of words in the *document*.

Define $\hat{y}_{d,-p}$ as the predicted y_d without paragraph p.

$$\hat{y}_{d,-p} = \boldsymbol{\omega}_x^\intercal \boldsymbol{x}_{2,d} + \frac{\boldsymbol{\omega}_z^\intercal}{N_d}\boldsymbol{s}_{d,-p} + \frac{\boldsymbol{\omega}_{zx}^\intercal}{N_d}\sum_{q\in\{[P_d]\setminus p\}}[\boldsymbol{x}_{1,d,q} \otimes \boldsymbol{s}_{d,q}] \tag{B.19}$$

We then have a predictive distribution that depends only on paragraph p.

$$y_d \sim \mathcal{N}\left(\hat{y}_{d,-p} - \frac{\boldsymbol{\omega}_z^\intercal}{N_d}(\boldsymbol{s}_{d,p,-n} + \boldsymbol{s}_{d,p,n}) - \frac{\boldsymbol{\omega}'_{zx}}{N_d}\boldsymbol{x}_{1,d,p} \otimes (\boldsymbol{s}_{d,p,-n} + \boldsymbol{s}_{d,p,n}), \sigma^2\right) \tag{B.20}$$

We can then follow the same steps as for the single paragraph document case to derive the third term in the sampling distribution, defining $\tilde{\boldsymbol{\omega}}_{z,d,p,k} = \boldsymbol{\omega}_{z,k} + \boldsymbol{\omega}'_{zx,k}\boldsymbol{x}_{1,d,p}$ analogously to $\tilde{\boldsymbol{\omega}}$ defined for the single paragraph case.

This gives us the sampling distribution for z, which is a Multinomial parameterised by

$$\begin{aligned} \Pr[z_{d,n} = k|\boldsymbol{Z}_{-(d,n)}, \boldsymbol{W}, \boldsymbol{y}, \alpha, \eta, \boldsymbol{\omega}, \sigma^2] \propto (s_{d,p,k,-n} + \alpha)\frac{m_{k,v,-(d,p,n)} + \eta}{\sum_v m_{k,v,-(d,p,n)} + V\eta} \times \\ \exp\left[\frac{1}{2\sigma^2}\left(\frac{2\tilde{\omega}_{z,d,p,k}}{N_d}\left(y_d - \hat{y}_{d,-p} - \frac{\tilde{\boldsymbol{\omega}}'_{z,d,p}}{N_d}\boldsymbol{s}_{d,-n}\right) - \left(\frac{\tilde{\omega}_{z,d,p,k}}{N_d}\right)\right)^2\right] \end{aligned} \tag{B.21}$$

In the M-step we can then still use the average $\bar{z}_{d,p}$ estimated in the E-step, but we need to weight each paragraph by the number of words in that paragraph to be consistent with

the E-step.

$$\bar{z}_d = \frac{1}{N_d} \sum_{p \in [P_d]} [N_{d,p} \bar{z}_{d,p}] \tag{B.22}$$

$$\overline{(x_{1,d,p} \otimes z_{d,p})} = \frac{1}{N_d} \sum_{p \in [P_d]} [N_{d,p} x_{1,d,p} \otimes z_{d,p}] \tag{B.23}$$

B.3 Semi-Synthetic Data Experiments

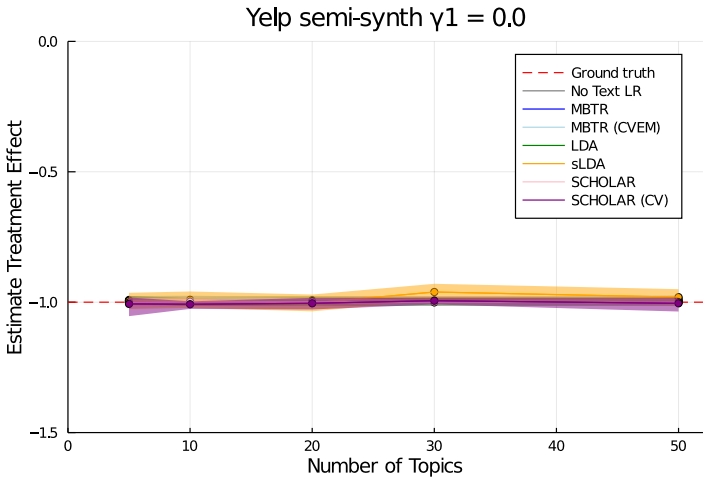


Figure B.1: Without correlation between confounders and treatments, the regression can be dissected into two separate parts (supervised topic estimation and regression weight estimation of the non-text features) without inducing bias in the estimators, as described in the section on the Frisch-Waugh-Lovell theorem. In such a case, all models manage to recover the ground truth.

B.4 Real-World Datasets and Data Pre-Processing

The Yelp dataset contains over 8 million customer reviews of businesses, which we restrict to reviews for businesses in Toronto in order to have relatively homogenous use of language throughout, and randomly sample $50,000$ reviews from this subset. The **Booking dataset** contains around $500,000$ hotel reviews, from which we randomly sample $50,000$ observations.

For our prediction experiments, we randomly select 75% in Yelp, 80% in Booking of our sample for training, holding out the remainder for testing. We then further split the training set equally for training in the E-step and validation in the M-step. The features are normalized on the training data statistics and the response variable is de-meaned. We do this because the K topic features sum to one and therefore implicitly already add

a constant to the regression (Blei and McAuliffe, 2008). We preprocess the text corpora by removing stopwords and then tokenizing and stemming the data.

Table B.1: Summary statistics of the review datasets

Statistics	#train	#val	#test	#vocab	#max words	#avg words
Yelp	18,750	18,750	12,500	24,680	572	61.2
Booking	20,000	20,000	10,000	6,968	305	18.7

Table B.2 shows the numerical metadata we use in order to construct our semi-synthetic examples

Table B.2: Numerical variables used for semi-synthetic experiments

Dataset	Variable	Description
Yelp	stars_av_u	historic avg. rating by user
	stars_av_b	historic avg. rating of business
	sentiment	*Harvard Inquirer* sentiment score
Booking	Average_Score	historical averge hotel score
	Review_Total_Negative_Word_Counts	total number of words in the negative part of review
	Review_Total_Positive_Word_Counts	total number of words in the positive part of review
	Total_Number_of_Reviews_Reviewer_Has_Given	total num of reviews by customer
	Total_Number_of_Reviews	total num of reviews of hotel

The Booking.com dataset allows consumers to enter the positive and negative parts of their reviews in separate boxes. We combine these two reviews for all our exercises, but we do use information on the word count in each of these sections (see below).

For the prediction exercises in Section 2.7, we use the number of stars associated with each review as the target variable. We also use the numerical metadata described in Table B.3as covariates.

Table B.3: Numerical covariates for prediction experiments

Dataset	Variable	Description
Yelp	stars_av_u	historic avg. rating by user
	stars_av_b	historic avg. rating of business
	sentiment	*Harvard Inquirer* sentiment score
Booking	Average_Score	historical averge hotel score
	Total_Negative_Word_Counts	total number of words in the negative part of review
	Total_Positive_Word_Counts	total number of words in the positive part of review
	Total_Number_of_Reviews_Reviewer_Has_Given	total num of reviews by customer
	Total_Number_of_Reviews	total num of reviews of hotel

For the semi-synthetic exercise on the Booking data, we construct

$$pos\_prop_i = \frac{Total\_Positive\_Word\_Counts_i}{Total\_Positive\_Word\_Counts_i + Total\_Negative\_Word\_Counts_i} \quad \text{(B.24)}$$

This variable is correlated with the treatment ($Average\_Score_i$) and with the outcome, and so the text can act as a confounder.

B.5 Real-World Data Experiments

B.5.1 Empirical data evaluation across different K

Table B.4: Mean pR^2 and perplexity over 20 runs per model, standard deviation in brackets

Dataset	Booking				Yelp			
K	10	20	30	50	10	20	30	50
				pR^2 (higher is better)				
LDA+LR	0.400 (0.003)	0.410 (0.004)	0.417 (0.005)	0.426 (0.003)	0.498 (0.005)	0.530 (0.009)	0.561 (0.010)	0.586 (0.006)
GSM+LR	0.387 (0.003)	0.390 (0.004)	0.389 (0.006)	0.386 (0.004)	0.502 (0.013)	0.505 (0.011)	0.503 (0.008)	0.495 (0.004)
LR+sLDA	0.416 (0.007)	0.426 (0.003)	0.430 (0.004)	0.432 (0.002)	0.533 (0.007)	0.564 (0.003)	0.567 (0.006)	0.571 (0.002)
LR+BPsLDA	0.394 (0.004)	0.396 (0.005)	0.400 (0.005)	0.419 (0.009)	**0.593 (0.003)**	*0.597* (0.002)	*0.597* (0.002)	*0.603* (0.002)
rSCHOLAR	**0.494** (0.005)	**0.495** (0.003)	**0.495** (0.003)	**0.494** (0.004)	0.520 (0.02)	0.548 (0.02)	0.563 (0.01)	0.571 (0.01)
BTR	*0.439* (0.008)	*0.447* (0.005)	*0.453* (0.003)	*0.454* (0.003)	*0.586* (0.007)	**0.615** (0.006)	**0.627** (0.004)	**0.630** (0.001)
				Perplexity (lower is better)				
LDA+LR	538 (3)	498 (2)	476 (2)	454 (1)	1544 (5)	1447 (4)	1388 (4)	1306 (4)
GSM+LR	**371 (6)**	**359 (11)**	**356 (14)**	**369 (8)**	**1500 (52)**	*1444* (29)	1463 (21)	1431 (34)
LR+sLDA	*535* (2)	491 (1)	*463* (1)	*436* (2)	1544 (6)	*1444* (6)	1382 (5)	*1294* (5)
rSCHOLAR	941 (134)	1429 (163)	2110 (396)	5014 (1314)	1744 (158)	1918 (138)	2216 (164)	2814 (383)
BTR	*535* (2)	*490* (1)	*463* (2)	437 (1)	*1540* (5)	**1443 (4)**	**1379 (4)**	**1291 (5)**

Table B.5: Best model in **bold**. Second best model in *italics*.

B.5.2 Model parametrisations

This section provides an overview over all used and tested hyperparameter settings across all models in our benchmark list. Table B.6 lists all hyperparameter settings pertaining to topic model components. Table B.7 provides an overview over all used neural network hyperparameters. B.8 summarises the iteration and stopping criteria for all models.

Table B.6: Topic model hyperparameters

	K	α	η	μ_{ntm}	σ_{ntm}	a_0	b_0	m_0	S_0
LDA	[10,20,30,50]	[0.1,0.5,1]	[0.001,0.01,0.1]	-	-	-	-	-	-
sLDA	[10,20,30,50]	[0.1,0.5,1]	[0.001,0.01,0.1]	-	-	-	-	-	-
BPsLDA	[10,20,30,50]	[0.1,0.5,1]	[0.001,0.01,0.1]	-	-	-	-	-	-
BTR	[10,20,30,50,100]	[0.1,**0.5**,1]	[0.001,**0.01**,0.1]	-	-	[0,1.5,**3**,**4**]	[0,**2**,4]	0	2
GSM	[10,20,30,50]	-	-	0	1	-	-	-	-
TAM	100	-	-	0	1	-	-	-	-

Bold parameter specifications were used for reported results in paper, unless stated otherwise. For Booking default $a_0 = 3$, for Yelp $a_0 = 4$.

Table B.7: Neural network hyperparameters

	HidLaySize θ	BatchSize	LearnRate	DropOut KeepRate	EmbedSize	HidLaySize RNN	TAM-thresh	nHidLayers BPsLDA
GSM	64	64	1.00E-03	[0.5,0.8,1]*	-	-	-	-
TAM	64	64	1.00E-03	0.8	100	64	1/K	-
aRNN	-	64	1.00E-03	0.8	100	64	-	-
BPsLDA	-	1050	1.00E-02	-	-	-	-	10

** best results (which occurred under no dropout) were reported in benchmarks*

Table B.8: Iteration parameters

	E-step iters	M-step iters	max. EM-iters	burn-in	max. epochs	Gibbs iters (thinning)
LDA	-	-	50***	100	-	1000 (5)
sLDA	[100,250,500]**	2500	50***	20	-	-
BTR	[100,250,500]**	2500	50***	20	-	-
GSM	-	-	-	-	100***	-
TAM	-	-	-	-	100***	-
BPsLDA	-	-	-	-	50***	-

*** no noticeable performance difference observed, therefore all results reported based on 100 E-step.*
**** best model achieved substantially before max. iterations reached.*

Further notes on benchmark model specifications:

For TAM and aRNN, the sequence length in the RNN component (ie. the maximum number of words per document) is 305 for Booking and 572 for Yelp which corresponds to the longest review in each respective data set. We therefore work with the full text of each review.

BPsLDA changes its behaviour quite drastically when α is set in an area $1 \leq \alpha \leq 2$, where it strongly increases its predictive performance (pR^2) at the cost of its document modelling performance (perplexity). This can be seen in the original paper (Chen et al. 2015). We included $\alpha = 1$ in the robustness test range and BTR is still generally on par with BPsLDA in this specific case for low K and does better for $K > 30$. Even when including $\alpha = 1$ in the robustness test range, BTR still outperforms BPsLDA and all other models across all hyperparameter settings, except $K = 10$ in the Yelp dataset, where BTR is a close second.

B.5.3 Robustness Tests

Robustness test across all topic models with LDA-like structure and Dirichlet hyperparameters for document-topic and word-topic distributions.

We assess the robustness of our findings to changes in the Dirichlet hyperparameters α and η. These hyperparameters act as priors on the topic-document distributions ($\boldsymbol{\beta}$) and word-topic distributions ($\boldsymbol{\theta}$), respectively. Table B.9 shows the results.

In terms of pR^2, BTR continues to perform best for all settings. We generally find that the BTR prediction performance is robust to hyperparameter changes. Evaluating the perplexity scores, we see more fluctuation across all models, which is unsurprising since those hyperparameter directly affect the generative topic modelling processes. BTR remains on par with its sLDA counterpart.

Table B.9: Sensitivity to hyperparameters α and β (K = 20)

		α			η		
Metric	*Model*	0.1	0.5	1	0.001	0.01	0.1
Yelp pR^2	LR-LDA	0.473	0.530	0.550	0.316	0.530	0.521
	LR-sLDA	0.558	0.564	0.559	0.562	0.564	0.568
	LR-BPsLDA	0.602	0.597	0.608	0.607	0.597	0.608
	BTR	**0.611**	**0.615**	**0.613**	0.**611**	**0.615**	**0.624**
Yelp perplexity	LR-LDA	1511	1448	1445	1472	1447	**1470**
	LR-sLDA	1497	1444	**1431**	**1441**	1444	1491
	BTR	**1490**	**1443**	1441	1456	**1443**	1478
Booking pR^2	LR-LDA	0.397	0.410	0.409	0.405	0.410	0.406
	LR-sLDA	0.430	0.426	0.432	0.422	0.426	0.433
	LR-BPsLDA	0.409	0.396	**0.453**	0.395	0.396	0.393
	BTR	**0.451**	**0.447**	0.452	**0.443**	**0.447**	**0.455**
Booking perplexity	LR-LDA	515	498	514	505	498	**512**
	LR-sLDA	**502**	**491**	504	**484**	**491**	516
	BTR	503	**491**	**503**	489	**491**	515

Further Robustness Tests - Booking

Table B.10 provides an extended robustness test on the predictive performance of the benchmark topic models across hyperparameters. BTR continues to be the best performing model throughout. Table B.11 summarises robustness tests in terms of perplexity scores. BTR achieves almost identical perplexity scores as sLDA whilst achieving higher pR^2 throughout.

Table B.10: Booking - pR^2 for different hyperparameter settings across topic benchmark models, best model in bold.

(K=10)	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 1$	$\eta = 0.001$	$\eta = 0.01$	$\eta = 0.1$	a,b = (3,2)	a,b = (0,0)	a,b = (3,4)	a,b = (1.5,4)
LR-LDA	0.378	0.4	0.408	0.397	0.4	0.398				
LR-sLDA	0.42	0.416	0.403	0.401	0.416	0.422				
LR-BPsLDA	0.396	0.394	0.439	0.393	0.394	0.396				
BTR	0.446	0.439	0.435	0.418	0.439	**0.452**	0.439	0.435	0.437	0.446

(K=20)	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 1$	$\eta = 0.001$	$\eta = 0.01$	$\eta = 0.1$	a,b = (3,2)	a,b = (0,0)	a,b = (3,4)	a,b = (1.5,4)
LR-LDA	0.397	0.41	0.409	0.405	0.41	0.406				
LR-sLDA	0.43	0.426	0.432	0.422	0.426	0.433				
LR-BPsLDA	0.409	0.396	0.453	0.395	0.396	0.393				
BTR	0.451	0.447	0.452	0.443	0.447	**0.455**	0.447	0.45	0.45	0.443

(K=30)	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 1$	$\eta = 0.001$	$\eta = 0.01$	$\eta = 0.1$	a,b = (3,2)	a,b = (0,0)	a,b = (3,4)	a,b = (1.5,4)
LR-LDA	0.399	0.417	0.423	0.417	0.417	0.413				
LR-sLDA	0.434	0.43	0.428	0.417	0.43	0.427				
LR-BPsLDA	0.424	0.4	0.451	0.401	0.4	0.402				
BTR	0.455	0.453	0.455	0.444	0.453	**0.459**	0.453	0.453	0.447	0.449

(K=50)	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 1$	$\eta = 0.001$	$\eta = 0.01$	$\eta = 0.1$	a,b = (3,2)	a,b = (0,0)	a,b = (3,4)	a,b = (1.5,4)
LR-LDA	0.415	0.426	0.428	0.418	0.426	0.420				
LR-sLDA	0.434	0.432	0.43	0.429	0.432	0.436				
LR-BPsLDA	**0.461**	0.419	0.449	0.411	0.419	0.418				
BTR	**0.461**	0.454	0.459	0.446	0.454	0.459	0.454	0.455	0.452	0.451

Default model was $\alpha = 0.5$, $\eta = 0.01$, $a_0 = 3$, $b_0 = 2$.
Robustness tests kept all hyperparameters at default, then changing one hyperparameter at a time.

Table B.11: Booking - perplexity scores for different hyperparameter settings across topic benchmark models, best model in bold.

K=10	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 1$	$\eta = 0.001$	$\eta = 0.01$	$\eta = 0.1$	a,b = (3,2)	a,b = (0,0)	a,b = (3,4)	a,b = (1.5,4)
LDA	562	538	539	539	538	545				
LR-sLDA	557	535	539	534	535	554				
BTR	556	535	538	**528**	535	548	535	535	537	536

K=20	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 1$	$\eta = 0.001$	$\eta = 0.01$	$\eta = 0.1$	a,b = (3,2)	a,b = (0,0)	a,b = (3,4)	a,b = (1.5,4)
LDA	515	498	514	505	498	512				
LR-sLDA	502	491	504	**484**	491	516				
BTR	503	490	503	489	490	515	490	491	490	491

K=30	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 1$	$\eta = 0.001$	$\eta = 0.01$	$\eta = 0.1$	a,b = (3,2)	a,b = (0,0)	a,b = (3,4)	a,b = (1.5,4)
LDA	479	476	502	484	476	492				
LR-sLDA	471	463	486	**454**	463	499				
BTR	470	463	483	457	463	500	463	463	463	463

K=50	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 1$	$\eta = 0.001$	$\eta = 0.01$	$\eta = 0.1$	a,b = (3,2)	a,b = (0,0)	a,b = (3,4)	a,b = (1.5,4)
LDA	442	454	494	460	454	476				
LR-sLDA	431	436	466	**421**	436	492				
BTR	430	437	467	423	437	492	437	436	439	437

Default model was $\alpha = 0.5$, $\eta = 0.01$, $a_0 = 3$, $b_0 = 2$.
Robustness tests kept all hyperparameters at default, then changing one hyperparameter at a time.

Further Robustness Tests - Yelp

Table B.12 provides an extended robustness test on the predictive performance of the benchmark topic models across hyperparameters. BTR continues to be the best performing model throughout, apart from the K=10 case, where it is a close second. Table B.13 summarises robustness tests in terms of perplexity scores. BTR achieves almost identical perplexity scores as sLDA whilst achieving higher pR^2 throughout.

Table B.12: Yelp - pR^2 for different hyperparameter settings across topic benchmark models, best model in bold.

K=10	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 1$	$\eta = 0.001$	$\eta = 0.01$	$\eta = 0.1$	a,b = (4,2)	a,b = (0,0)	a,b = (3,4)	a,b = (1.5,4)
LR-LDA	0.476	0.498	0.515	0.503	0.498	0.49				
LR-sLDA	0.523	0.533	0.539	0.52	0.533	0.527				
LR-BPsLDA	0.596	0.593	**0.606**	0.595	0.593	0.592				
BTR	0.592	0.586	0.593	0.575	0.586	0.596	0.586	0.588	0.578	0.59

K=20	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 1$	$\eta = 0.001$	$\eta = 0.01$	$\eta = 0.1$	a,b = (4,2)	a,b = (0,0)	a,b = (3,4)	a,b = (1.5,4)
LR-LDA	0.473	0.53	0.55	0.483	0.53	0.521				
LR-sLDA	0.558	0.564	0.559	0.562	0.564	0.568				
LR-BPsLDA	0.602	0.597	0.608	0.607	0.597	0.608				
BTR	0.611	0.615	0.613	0.611	0.615	**0.624**	0.615	0.62	0.593	0.621

K=30	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 1$	$\eta = 0.001$	$\eta = 0.01$	$\eta = 0.1$	a,b = (4,2)	a,b = (0,0)	a,b = (3,4)	a,b = (1.5,4)
LR-LDA	0.499	0.561	0.563	0.547	0.561	0.565				
LR-sLDA	0.565	0.567	0.563	0.567	0.567	0.56				
LR-BPsLDA	0.609	0.597	0.607	0.599	0.597	0.599				
BTR	0.624	**0.627**	0.612	0.608	**0.627**	0.622	**0.627**	0.623	**0.627**	0.626

K=50	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 1$	$\eta = 0.001$	$\eta = 0.01$	$\eta = 0.1$	a,b = (4,2)	a,b = (0,0)	a,b = (3,4)	a,b = (1.5,4)
LR-LDA	0.523	0.586	0.591	0.571	0.586	0.582				
LR-sLDA	0.573	0.571	0.564	0.556	0.571	0.573				
LR-BPsLDA	0.612	0.603	0.606	0.604	0.603	0.604				
BTR	**0.632**	0.630	0.623	0.621	0.630	**0.632**	0.630	0.629	0.629	0.628

Table B.13: Yelp - perplexity scores for different hyperparameter settings across topic benchmark models, best model in bold.

K=10	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 1$	$\eta = 0.001$	$\eta = 0.01$	$\eta = 0.1$	a,b = (4,2)	a,b = (0,0)	a,b = (3,4)	a,b = (1.5,4)
LDA	1586	1544	1532	1557	1544	1552	1544			
LR-sLDA	1583	1544	**1530**	1561	1544	1554	1544			
BTR	1588	1540	1534	1565	1540	1546	1540	1539	1548	1547

K=20	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 1$	$\eta = 0.001$	$\eta = 0.01$	$\eta = 0.1$	a,b = (4,2)	a,b = (0,0)	a,b = (3,4)	a,b = (1.5,4)
LDA	1511	1447	1445	1472	1447	1469	1447			
LR-sLDA	1497	1444	**1431**	1441	1444	1491	1444			
BTR	1490	1443	1441	1456	1443	1478	1443	1443	1445	1441

K=30	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 1$	$\eta = 0.001$	$\eta = 0.01$	$\eta = 0.1$	a,b = (4,2)	a,b = (0,0)	a,b = (3,4)	a,b = (1.5,4)
LDA	1434	1388	1390	1412	1388	1415	1388			
LR-sLDA	1436	1382	1383	1395	1382	1442	1382			
BTR	1434	**1379**	1385	1390	**1379**	1448	**1379**	1378	1389	1379

K=50	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 1$	$\eta = 0.001$	$\eta = 0.01$	eta = 0.1	a,b = (4,2)	a,b = (0,0)	a,b = (3,4)	a,b = (1.5,4)
LDA	1352	1306	1325	1334	1306	1356	1306			
LR-sLDA	1349	1294	1310	1309	1294	1404	1294			
BTR	1338	**1291**	1303	1288	**1291**	1405	**1291**	1293	1294	1292

B.5.4 Estimated Topics

The below tables are an extended version of the corresponding table in the paper. The show the top 3 negative and positive topics for K=[10,30,100]. Inspecting the top words

in each of these topics compared with its regression coefficient, BTR models highly interpretable topics - at least as interpretable as LDA or sLDA. At the same time BTR achieves substantially better prediction performances throughout all model specifications (see previous section).

Table B.14: Top 3 positive and negative topics for *Yelp* (K = 10)

	Pos1	Pos2	Pos3	Neg3	Neg2	Neg1
BTR topics	food	restaur	time	store	food	order
	great	dish	hair	locat	chicken	us
	place	lobster	back	like	good	ask
	servic	menu	work	can	order	servic
	friend	food	will	price	rice	wait
	love	order	day	go	dish	food
BTR regr. weights	4.3	1.7	1.5	-0.1	-0.5	-8.8
sLDA topics	food	time	locat	coffe	us	like
	place	hair	store	tea	order	place
	great	back	can	tri	ask	go
	servic	work	find	place	servic	much
	good	will	place	ice	tabl	im
	time	day	staff	cream	time	realli
sLDA regr. weights	2.7	1.7	1.2	0.1	-3.7	-4.5
LDA topics	food	place	place	store	fri	order
	great	coffe	great	like	burger	us
	servic	good	good	locat	order	food
	restaur	tri	friend	can	like	servic
	dish	tea	bar	find	good	time
	menu	great	drink	go	chees	ask
LDA regr. weights	1.2	0.7	0.5	-0.1	-0.4	-2.6

Table B.15: Top 3 positive and negative topics for *Booking* (K = 10)

	Pos1	Pos2	Pos3	Neg3	Neg2	Neg1
BTR topics	hotel	room	room	room	check	hotel
	stay	locat	great	bed	book	room
	staff	staff	love	shower	room	locat
	would	good	hotel	bathroom	us	small
	help	clean	view	small	hotel	good
	everyth	comfort	bar	clean	arriv	price
BTR regr. weights	2.8	1.7	1.5	-1.0	-1.1	-5.7
sLDA topics	hotel	room	room	room	room	room
	stay	good	great	bed	night	hotel
	staff	locat	love	bathroom	window	locat
	would	staff	hotel	shower	work	small
	help	clean	view	small	floor	staff
	like	breakfast	nice	comfort	air	posit
sLDA regr. weights	2.4	1.3	1.3	-0.4	-0.6	-5.6
LDA topics	hotel	hotel	neg	check	room	room
	stay	great	staff	room	shower	hotel
	staff	love	locat	book	bathroom	good
	help	room	friendli	hotel	work	locat
	would	view	great	us	bed	breakfast
	noth	locat	help	time	air	price
LDA regr. weights	1.3	1.2	1.0	-1.3	-1.4	-2.0

Table B.16: Top 3 positive and negative topics for *Yelp* (K = 30)

	Pos1	Pos2	Pos3	Neg3	Neg2	Neg1
BTR topics	best	great	restaur	us	ask	like
	plac	friend	menu	order	said	disappoint
	alway	servic	dish	tabl	custom	better
	love	staff	wine	food	told	tast
	ever	recommend	steak	server	never	noth
	toronto	amaz	perfect	came	say	bad
BTR regr. weights	6.9	6.0	2.1	-3.9	-8.4	-13.3
sLDA topics	great	time	im	seem	ask	like
	love	alway	review	like	never	food
	amaz	go	place	much	custom	good
	recommend	year	star	make	said	place
	servic	never	go	think	servic	tast
	friend	everi	give	thing	told	better
sLDA regr. weights	3.7	3.1	3.1	-2.3	-6.4	-7.1
LDA topics	great	toronto	restaur	us	ask	like
	friend	visit	menu	tabl	custom	tast
	love	make	dish	order	said	disappoint
	amaz	love	wine	food	servic	better
	place	made	dessert	came	told	bad
	servic	best	dinner	server	manag	noth
LDA regr. weights	3.0	1.6	1.3	-1.2	-4.9	-8.3

Table B.17: Top 3 positive and negative topics for *Booking* (K = 30)

	Pos1	Pos2	Pos3	Neg3	Neg2	Neg1
BTR topics	us	stay	room	room	ask	room
	staff	would	locat	small	us	hotel
	made	hotel	great	bed	day	old
	upgrad	staff	staff	size	recept	poor
	stay	love	bit	locat	call	star
	welcom	recommend	littl	bathroom	back	bad
BTR regr. weights	2.7	2.5	2.3	-2.5	-3.0	-9.0
sLDA topics	staff	hotel	us	book	room	hotel
	friendli	love	upgrad	charg	need	room
	great	beauti	staff	hotel	locat	bad
	help	decor	room	pay	old	star
	locat	modern	stay	check	look	poor
	neg	great	love	day	smell	posit
sLDA regr. weights	2.1	1.8	1.7	-1.4	-2.1	-9.7
LDA topics	stay	stay	hotel	us	room	hotel
	hotel	hotel	love	ask	locat	like
	made	would	beauti	one	good	star
	like	recommend	great	recept	need	realli
	feel	definit	decor	day	old	much
	realli	love	staff	call	valu	best
LDA regr. weights	2.4	2.1	1.9	-2.5	-2.7	-3.4

Table B.18: Top 3 positive and negative topics for *Yelp* (K = 100)

	Pos1	Pos2	Pos3	Neg3	Neg2	Neg1
BTR topics	love	definit	best	custom	never	disappoint
	delici	ever	amaz	ask	worst	tast
	definit	toronto	everi	said	ever	bland
	perfect	citi	friend	manag	money	dri
	tri	far	free	rude	bad	better
	super	amaz	alway	servic	terribl	lack
BTR regr. weights	5.9	5.2	5.0	-8.4	-12.5	-14.5
sLDA topics	alway	will	amaz	ask	tast	disappoint
	time	definit	definit	said	like	bad
	usual	servic	love	told	felt	cold
	come	friend	great	back	disappoint	worst
	never	return	place	went	better	dri
	everi	back	everyth	want	wasnt	lack
sLDA regr. weights	4.1	4.0	3.9	-7.0	-8.0	-11.3
LDA topics	love	best	experi	money	never	tast
	amaz	toronto	made	go	bad	like
	delici	ever	make	will	ever	disappoint
	place	citi	feel	never	worst	meat
	absolut	far	first	pay	terribl	bland
	super	visit	felt	spend	experi	dri
LDA regr. weights	5.4	4.6	3.8	-5.9	-10.8	-10.9

Table B.19: Top 3 positive and negative topics for *Booking* (K = 100)

	Pos1	Pos2	Pos3	Neg3	Neg2	Neg1
BTR topics	staff	hotel	love	old	room	poor
	help	wonder	great	look	small	posit
	friendli	beauti	staff	carpet	tini	servic
	excel	love	littl	tire	bathroom	bad
	especi	experi	fab	furnitur	noisi	never
	wonder	fabul	especi	need	far	rude
BTR regr. weights	4.3	4.1	3.3	-6.1	-7.3	-14.2
sLDA topics	love	room	great	old	hotel	bad
	beauti	small	locat	dirti	star	poor
	amaz	posit	neg	bathroom	expect	recept
	fantast	size	perfect	carpet	rate	posit
	fabul	bit	awesom	wall	thi	even
	wonder	expect	super	look	basic	never
sLDA regr. weights	3.7	3.6	2.8	-5.8	-6.0	-14.3
LDA topics	love	great	bit	hotel	old	recept
	amaz	locat	littl	star	dirti	manag
	everyth	neg	nice	rate	carpet	rude
	noth	staff	locat	expect	look	receptionist
	perfect	awesom	breakfast	disappoint	wall	check
	absolut	perfect	good	thi	furnitur	guest
LDA regr. weights	3.6	3.1	2.8	-5.3	-6.6	-7.6

B.5.5 Computation Times

Table B.20 shows the time taken for 100 E-step iterations on a single 2.8GHz processor on the Booking data and 300-400 seconds on the Yelp data. We found that 100 E-step iterations is typically sufficient for the best performance and the model typically converges after between 10-25 EM iterations. A typical 30 topic model on Yelp data thus took around 1 hour to converge, and around 20 minutes for Booking. Computation time scales roughly linearly in the number of topics and total number of words across all documents. This is because the evaluation of the K-dimensional multinomial distribution for each $z_{d,n}$ (Equation 2.13) is the principle computational challenge.

Table B.20: Computational time

Dataset	K	100 E-step iters
Yelp	10	50s
	20	110s
	30	200s
	50	320s
	100	740s
Booking	10	18s
	20	33s
	30	50s
	50	79s
	100	200s

Note: Yelp data has roughly 3 times as many words as Booking.com data

Appendix C

Appendix to Chapter 3

C.1 Data

Figure C.1: Article matched with both headline and NER

Next confident of hitting its full-year target (FT, 2003-01-09)

Next yesterday dispersed the worst fears about its performance with news of a small increase in underlying sales and promised it would achieve profit forecasts. The group, which had been the focus of continued rumours of poor trading in the run up to Christmas, admitted that growth was slowing and that it was still suffering because of some mistakes in its fashion ranges. Underlying sales for the 23 weeks to January 4 were up just 1.7 per cent on last year - with full-price sales for the period August 12 to December 24 down by 0.8 per cent. However, Next said it was forecasting annual pre-tax profits of Pounds 293m to Pounds 303m, against Pounds 265.8m last year. Consensus City forecasts were for Pounds 293m. Next shares, which have declined by more than 17 per cent since the start of October, gained 24p to close at 770p. Simon Wolfson, chief executive, said he was confident that Next had largely corrected its ranging mistakes for the new spring/summer clothing. "Obviously sales are disappointing but they are not as bad as they were in October," said Mr Wolfson, referring to the group's surprise autumn profits warning, blamed on unseasonably warm weather. "We are cautious but not pessimistic about the outlook for consumer spending." He said investors should look at total sales growth - claiming there was too much of a focus on like-for-like measures that strip out the effects of new store openings. The group is extending many stores and moving others into larger premises - a trend Mr Wolfson said depressed the like-for-like measure by 1 percentage point. "Our new stores are all exceeding expectations and that is really what has driven our profits and means we will meet expectations," he said. Next sales in the 23 weeks to January 4 rose by 12.4 per cent. Like-for-like sales rose by just 1.7 per cent. However, sales through the Next Directory remained strong - up 22.3 per cent over last year. The group said that taken together, sales for the Next brand as a whole were ahead by 14.3 per cent. Sales at full price in the period up to Christmas were up by 9.5 per cent - but down by 0.8 per cent like-for-like. Full-price sales for stores and the directory were ahead by 11.8 per cent.

Named Entities

Entities identified: Simon Wolfson; Next
Entities missed: Mr Wolfson

H

Figure C.2: Article matched with only NER on main text

Buoyant retail sales data defy bleak forecasts (FT, 2001-01-24)

Retail sales growth surged in December to its fastest rate in eight months, defying economists' forecasts and apparently dispelling talk of a slowdown in consumer demand. While admitting that recent pessimism appeared to have been overdone, economists said the figures - which included a week of post-Christmas sales - were not conclusive. The true state of consumer demand would not be revealed for another month, they said. The Office for National Statistics said retail sales volumes grew a seasonally adjusted 1.1 per cent from November, to stand 6.4 per cent higher than in December 2001. Economists had expected a 0.2 per cent monthly contraction and year-on-year growth of 4.3 per cent. However, the ONS warned that there were problems in adjusting for spending around Christmas, exacerbated by the timing of the December trading period. The figures covered December 1 to January 4 - six days longer than was measured the previous year and including the frantic new year sales. John Butler, UK economist at HSBC, was suspicious of the figures, saying they "totally contradicted" recent company reports, survey evidence and even money data. "The obvious question is do we believe this degree of strength? Quite simply -no," he said."Predicting the ONS number in December is like playing roulette, and we expect January to show a fall." However, he expected underlying demand would hold up, and quarterly figures - which provide a more accurate picture of the trend - suggested consumer demand remained robust. Sales in the December quarter were 1.8 per cent higher than in the previous three months and 5.5 per cent higher than in the last quarter of 2001. There had been tentative signs that shoppers tightened their belts in the crucial December shopping period. Surveys conducted in the first half of the month suggested sales were weak, verging on stagnant, and big retailers such as Dixons and Next said Christmas trade was disappointing. The Bank of England's monetary policy committee noted at its January meeting that sales appeared to have been sluggish in the first half of December, although it said they were thought to have strengthened later in the month. While volumes were up, Simon Rubinsohn, chief economist at Gerrard, the fund manager, said underlying values fell, with prices down 1.5 per cent on the previous December. The stronger than expected retail figures made it likely that today's growth data would also outperform economists' expectations of a 0.4 per cent expansion in the fourth quarter.

Named Entities

Entities identified: Economists; John Butler; Next; Simon Rubinsohn; Office for National Statistics; ONS; Bank of England
Entities missed: HSBC, Dixons, Gerrard

Table C.1: Full list of firm names (A-C)

Code	Company	Previous names
AAL.L	Anglo American	Anglo American; Anglo American PLC; Anglo American plc;
AB.L˜G07	Alliance Boots Holdings Ltd	Alliance Boots Holdings Ltd; Alliance Boots; Boots; Alliance Boots Holdings; Boots;
ABF.L	ABF	ABF; Associated British Foods PLC; Associated British Foods plc; George Weston Holdings plc; Associated British Foods; George Weston Holdings; George Weston Holdings plc; George Weston;
ACAA.L	Acacia Mining	Acacia Mining; Acacia Mining PLC; Acacia Mining plc; African Barrick Gold plc; African Barrick Gold limited; African Barrick Gold; African Barrick Gold plc; African Barrick Gold limited;
ADML.L	Admiral Group	Admiral Group; Admiral Group PLC; Admiral Group plc; Admiral Group limited; Admiral Group limited;
ADN.L˜H17	Aberdeen Inv	Aberdeen Inv; Aberdeen Asset Management PLC; Abadin Asetto Manejimento; ; Aberdeen Asset Management;
ADZ.L˜J00	Allied Zurich	Allied Zurich; Allied Zurich Ltd; Allied Zurich limited;
AG.L˜L02	Arcadia Group	Arcadia Group; Arcadia Group Ltd; Arcadia;
AGGK.L	Aggreko	Aggreko; Aggreko PLC; Aggreko plc;
AHT.L	Ashtead Group	Ashtead Group; Ashtead Group PLC; Needroy limited; Ashtead Group plc; Needroy; Ashtead Group plc;
ALLD.L˜G05	Allied Domecq	Allied Domecq; Allied Domecq Ltd; NA;
ALLL.L˜J08	Alliance UK	Alliance UK; Alliance& Leicester Ltd; Alliance & Leicester limited; Alliance & Leicester plc; Alliance & Leicester; Alliance & Leicester; Alliance & Leicester plc;
AMFW.L˜J17	Amec Foster	Amec Foster; Amec Foster Wheeler Ltd; Amekku Fosuta; Amec Foster Wheeler;
AML.L˜B16	MS Amlin	MS Amlin; MS Amlin PLC; MS Amlin plc; Amlin; Amlin;
ANTO.L	Antofagasta	Antofagasta; Antofagasta PLC; Antofagasta plc; Antofagasta Holdings plc; Antofagasta Holdings; Antofagasta Holdings plc;
ARM.L˜T16	Svf Holdco Uk	Svf Holdco Uk; SVF Holdco Uk Ltd; ARM Holdings; SVF Holdco; ARM; SVF Holdco Uk; ARM; SVF Holdco;
ASHM.L	Ashmore Group	Ashmore Group; Ashmore Group PLC; Ashmore Group plc; Ashmore Group limited; Ashmore Group limited;
ASSD.L˜J99	ASDA Group	ASDA Group; ASDA Group Ltd; Asda;
ATST.L	Alliance Trust plc	Alliance Trust plc; Alliance Trust PLC; The Alliance Trust plc; Alliance Trust; The Alliance Trust;
AUN.L˜G06	WBA Holdings 1	WBA Holdings 1; WBA Holdings 1 Ltd; Alliance Boots Holdings; Alliance UniChem plc; Walgreens Boots Alliance; Boots; Boots Chemist; WBA 1; Alliance Boots; Alliance UniChem; Alliance UniChem plc; Walgreens Boots Alliance; Boots; Boots Chemist;
AUTN.L˜K11	Autonomy	Autonomy; Autonomy Corporation Ltd; Autonomy Corp; Autonomy Corporation; ; Autonomy Corporation; ;
AV.L	Aviva GB	Aviva GB; Aviva PLC; Aviva plc; CGNU plc; CGU plc; Aviva; CGNU; CGNU plc; CGU; CGU plc;
AWA.L˜H00	Windward	Windward; Windward Prospects Ltd; Windward Prospects; Windward Prospects limited; Arjo Wiggins Appleton; Windward Prospects limited; Arjo Wiggins Appleton;
AZN.L	AstraZeneca	AstraZeneca; AstraZeneca PLC; Astrazeneca plc; Zeneca Group plc; Hackplimco plc; Astrazeneca; Zeneca Group; Zeneca Group plc; Hackplimco; Hackplimco plc;
BAA.L˜H06	LHR Airports	LHR Airports; LHR Airports Ltd; London Heathrow Airports; BAA Airports; Heathrow; BAA Airports; Heathrow;
BAB.L	Babcock Intl	Babcock Intl; Babcock International Group PLC; Babcock International Group plc; Flowplus plc; Babcock International Group; Flowplus; Flowplus plc;
BAES.L	BAE Systems	BAE Systems; BAE Systems PLC; BAE Systems plc; British Aerospace plc; British Aerospace limited; British Aerospace; British Aerospace plc; British Aerospace limited;
BALF.L	Balfour Beatty	Balfour Beatty; Balfour Beatty PLC; Balfour Beatty plc; BICC plc; BICC; BICC plc;
BARC.L	Barclays	Barclays; Barclays PLC; Barclays plc; Barclays Bank plc; Barclays Bank limited; Barclays Bank; Barclays Bank plc; Barclays Bank limited;
BATS.L	Brit Am Tobacco	Brit Am Tobacco; British American Tobacco PLC; British American Tobacco plc; Measureprofit plc; British American Tobacco; Measureprofit; Measureprofit plc;
BCI.L˜H01	Tarmac Cement An	Tarmac Cement An; Tarmac Cement and Lime Ltd; Tarmac Cement and Lime limited; Tarmac Cement & Lime limited; Lafarge Cement UK limited; Lafarge Cement UK; Lafarge Tarmac; Blue Circle Inds; Blue Circle; Blue Circle Industrials Tarmac Cement and Lime; Tarmac Cement & Lime limited; Tarmac Cement & Lime; Lafarge Cement UK limited; Lafarge Cement UK; Lafarge Tarmac; Blue Circle Inds; Blue Circle; Blue Circle Industrials;
BDEV.L	Barratt Developments	Barratt Developments; Barratt Developments PLC; Barratt Developments limited;
BG.L˜B16	BG Grp	BG Grp; BG Group Ltd; Biji Gutupu; British Gas; Mahanagar Gas; BG Group; British Gas; Mahanagar Gas;
BGY.L˜B09	EDF Nuclear	EDF Nuclear; EDF Energy Nuclear Generation Group Ltd; EDF Energy Nuclear Generation Group; EDF Energy; EDF Energy;
BIL˜L06	BIP Industries	BIP Industries; BIP Industries Ltd; Brambles Industries plc; BIP Industries limited; Brambles Industries; BIP Industries limited;
BKGH.L	Berkeley Group	Berkeley Group; Berkeley Group Holdings PLC; The Berkeley Group Holdings plc; Sandnumber limited; Berkeley Group Holdings; The Berkeley Group Holdings; The Berkeley Group; Sandnumber limited; Sandnumber;
BLM.L˜B05	Baltimore Capital	Baltimore Capital; Baltimore Capital PLC; Baltimore Capital plc;
BLND.L	British Land	British Land; British Land Company PLC; British Land Company plc; The British Land Company plc; British Lead co plc; British Land Company; The British Land Company; The British Land Company plc; British Lead co; British Lead co plc;
BLT.L	BHP	BHP; BHP Billiton PLC; BHP Billiton plc; Billiton plc; Hackplimco plc; BHP Billiton; Billiton; Billiton plc; Hackplimco; Hackplimco plc;
BMAH.L˜I00	Burmah Castrol	Burmah Castrol; Burmah Castrol PLC; Burmah Castrol plc;
BNZL.L	Bunzl	Bunzl; Bunzl plc; Bunzl Pulp & Paper limited; Bunzl Pulp & Paper limited; Bunzl Pulp & Paper;
BOC.L˜I06	BOC Grop	BOC Grop; BOC Group Ltd; BOC; BOC Group; BOC Group;
BP.L	BP	BP; BP PLC; BP plc; BP Amoco plc; The British Petroleum Company plc; BP Amoco; BP Amoco plc; The British Petroleum Company; The British Petroleum Company plc;
BPB.L˜A06	BPB	BPB; BPB Ltd; BPB limited; British Plaster Board; British Plaster Board;
BPTY.L˜B16	bwin.party	bwin.party; bwin.party digital entertainment plc; bwin.party Digital Entertainment plc; bwin; bwin.party digital entertainment; bwin.party Digital Entertainment; bwin;
BRBY.L	Burberry Group	Burberry Group; Burberry Group PLC; Burberry Group plc; Burberry Group limited; Burberrys Group limited; Burberry Group limited; Burberrys Group;
BSCT.L˜I01	Bank of Scotland	Bank of Scotland; Bank of Scotland PLC; NA;
BT.L	BT Group	BT Group; BT Group PLC; BT Group plc; BT; Newgate Telecommunications limited; British Telecommunications plc; ; BT; Newgate Telecommunications limited; Newgate Telecommunications; British Telecommunications; British Telecommunications plc;
BTR.L˜B99	Invensys Intl UK	Invensys Intl UK; Invensys International Holdings Ltd; Invensys International Holdings; Invensys International; Invensys; Invensys International; Invensys;
CAPC.L˜J15	Clean Air	Clean Air; Clean Air Power Ltd; Clean Air Power limited; Clean Air Power;
CBRY.L˜C10	Cadbury	Cadbury; Cadbury Ltd; Cadbury plc; Cadbury Schweppes plc; Cadbury Schweppes; Cadbury Schweppes plc;
CCH.L	Coca-Cola HBC AG	Coca-Cola HBC AG; Coca Cola HBC AG; Cocal Cola HBC; Coca Cola Hellenic; Coca-Cola HBC; Coca Cola HBC; Cocal Cola HBC; Coca Cola Hellenic;
CCH.L˜H04	Celltech Group	Celltech Group; Celltech Group Ltd; Celltech;
CCL.L	Carnival	Carnival; Carnival PLC; Carnival plc; P&O Princess Cruises plc; Carnival Corporation; P&O Princess Cruises; P&O Princess Cruises plc; Carnival Corporation;
CCM.L˜A04	Carlton Comm	Carlton Comm; Carlton Communications Ltd; Carlton Communications limited; Carlton Communications; Carlton Communications;
CHB.L˜H03	Chubb UK	Chubb UK; Chubb Ltd; Chubb limited; Chubb plc; Exchangestart plc; Chubb; Chubb plc; Exchangestart; Exchangestart plc;
CMG.L˜L02	CMG PLC	CMG PLC; Cmg PLC; CMG plc; CMG; Cmg;
CNA.L	Centrica	Centrica; Centrica PLC; Centrica plc; Yieldtop plc; Yieldtop; Yieldtop plc;
CNE.L	Cairn Energy	Cairn Energy; Cairn Energy PLC; Cairn Energy plc; New Cairn Energy plc; Randotte limited; Cairn; New Cairn Energy; New Cairn Energy plc; Randotte limited; Randotte; Cairn;
CO.L˜F03	Coats Holdings	Coats Holdings; Coats Holdings Ltd; Coats plc; Coats Group plc; Coats; Coats Group; Coats Group plc;
COB.L	Cobham	Cobham; Cobham PLC; Cobham plc; FR Group plc; Flight Refuelling plc; FR Group; FR Group plc; Flight Refuelling; Flight Refuelling plc;
COLT.F˜K15	Colt Group	Colt Group; Colt Group SA; COLT Group SA; COLT Telecom Group SA; COLT GRP S.A.; COLT Telecom Group S.A.; Colt Group; Colt S.A.; Colt Luxembourg; Colt Technology Services; COLT Telecom Group SA; COLT GRP S.A.; COLT GRP COLT Telecom Group S.A.; COLT Telecom Group; Colt S.A.; Colt; Colt Luxembourg; Colt Technology Services;
CPG.L	Compass Group	Compass Group; Compass Group PLC; Compass Group plc; Compass Demerger limited; Factory Canteens limited; Compass Demerger limited; Compass Demerger; Factory Canteens limited; Factory Canteens;
CPI.L	Capta	Capta; Capita PLC; Capita plc; The Capita Group plc; Capita Holdings limited; Capita; The Capita Group; The Capita Group plc; Capita Holdings limited; Capita Holdings;
CRDA.L	Croda	Croda; Croda International PLC; Croda International plc; Croda International;
CRH.L	CRH	CRH; CRH PLC; CRH plc;
CS.L˜D07	Corus Group	Corus Group; Corus Group Ltd; Corus Group limited;
CTEC.L	ConvaTec Group	ConvaTec Group; ConvaTec Group PLC; Convatec Group plc; ConvaTec Group limited; Convatec Group; ConvaTec Group limited;
CTLD.L˜I98	Akzo Nobel UK	Akzo Nobel UK; Cloudburst II Ltd; Cloudburst; Cloudburst II; Cloudburst II;
CWC.L˜E16	C&W	C&W; Cable and Wireless Communications Ltd; Cable and Wireless Communications limited; Cable & Wireless Communications; CWC; Cable and Wireless Communications ; Cable and Wireless Communications; Cable & Wireless Communications; CWC;
CWP.L˜G12	Cable & Wireles	Cable & Wireles; Cable & Wireless Worldwide Ltd; Cable and Wireless Communications limited; Cable & Wireless Communications; CWC; Cable & Wireless Worldwide; Cable and Wireless Communications; Cable & Wireless Communications; CWC;
CWZ.L˜E00	NTL CWC	NTL CWC; NTL (CWC) Ltd; NTL (CWC) limited; Cable & Wireless Communications limited; NTL; ntl:; NTL (CWC); Cable & Wireless Communications limited; Cable & Wireless Communications; NTL; ntl:;

Table C.2: Full list of firm names (D-J)

Code	Company	Previous names
DC.L	Dixons Carphone	Dixons Carphone; Dixons Carphone PLC; Dixons Carphone plc; Carphone Warehouse Group plc; New Carphone Warehouse plc; Carphone Warehouse Group; Carphone Warehouse Group plc; New Carphone Warehouse; New Carphone Warehouse plc;
DCC.L	DCC	DCC; DCC PLC; DCC plc;
DDT.L~L10	Dimension Data	Dimension Data; Dimension Data Holdings PLC; NA; Dimension Data Holdings;
DGE.L	Diageo	Diageo; Diageo PLC; Diageo plc; Guiness plc; Guiness; Guiness plc;
DLAR.L	De La Rue	De La Rue; De La Rue PLC; De La Rue plc; New De La Rue plc; Precis limited; Precis 1809 limited; ; New De La Rue; New De La Rue plc; Precis limited; Precis; Precis 1809 limited; Precis 1809;
DLGD.L	DIRECT LINE INS	DIRECT LINE INS; Direct Line Insurance Group PLC; Direct Line Insurance Group plc; RBS Insurance Group limited; Direct Line; Direct Line Insurance Group; RBS Insurance Group limited; RBS Insurance Group; Direct Line;
DMGOa.L	DMGT	DMGT; Daily Mail and General Trust P L C; Daily Mail and General Trust plc; Daily Mail & General Trust plc; Daily Mail; Daily Mail and General Trust; Daily Mail & General Trust plc; Daily Mail;
DRX.L	Drax Group	Drax Group; Drax Group PLC; Drax Group plc;
DVR.L~I06	Ereved Group	Ereved Group; Ereved Group Holdings Ltd; De Vere Group plc; De Vere Group Holdings; De Vere; Ereved Group Holdings; De Vere Group; De Vere Group Holdings; De Vere;
DXNS.L~H14	Dixons Retail	Dixons Retail; Dixons Retail Group Ltd; Dixons Retail plc; Dixons; Dixons Retail Group; Dixons;
ECM.L	Electrocomponent	Electrocomponent; Electrocomponents PLC; Electrocomponents plc; Electrocomponents; Electrocomponents;
EGS.L~I02	Energis	Energis; Energis PLC; Energis plc;
EIGE.L	EI Grp	EI Grp; EI Group PLC; EI Group plc; Enterprise Inns plc; Midcounties Inns limited; ; EI Group; Enterprise Inns; Enterprise Inns plc; Midcounties Inns limited; Midcounties Inns; ;
EMA.L~C08	Ascential Group	Ascential Group; Ascential Group Ltd; EMAP; Top Right Group limited; Top Right Group limited; Top Right Group;
EMG.L	Man Group	Man Group; Man Group PLC; Man Group plc; Man Strategic Holdings plc; Man Strategic Holdings; Man Strategic Holdings plc; Man Strategic;
EMI.L~I07	EMI Group	EMI Group; EMI Group Ltd; EMI Group limited;
ENQ.L	EnQuest	EnQuest; EnQuest PLC; Enquest plc; EnQuest plc; Enquest; EnQuest plc;
ENRC.L~K13	ENRC	ENRC; Eurasian Natural Resources Corporation Ltd; Eurasian Natural Resources Corporation limited; Eurasian Resources; Eurasian Natural Resources Corporation; Eurasian Resources;
ESSR.L~F14	Essar Engr	Essar Engr; Essar Energy Ltd; Essar Energy limited; Essar Energy plc; Essar Energy; Essar Energy; Essar Energy plc;
ETP.L~F02	Enterprise Oil	Enterprise Oil; Enterprise Oil Ltd; Enterprise Oil limited; Enterprise Oil plc; British Gas North Sea Oil Holdings limited; Enterprise Oil plc; British Gas North Sea Oil Holdings limited; British Gas North Sea Oil Holdings; British Gas North Sea Oil;
EVRE.L	EVRAZ	EVRAZ; EVRAZ plc; Profect Savannah plc; Evraz plc; Profect Savannah; Profect Savannah plc; Evraz; Evraz plc;
EXL.L~F00	Exel Plc	Exel Plc; Exel Plc Pre Merger; Exel plc; Exel plc Pre Merger; Exel; Exel Pre Merger; Exel plc Pre Merger;
EXL.L~L05	Exel	Exel; Exel Ltd; Exel limited; Exel plc; Ocean Group plc; Exel plc; Ocean Group; Ocean Group plc;
EXPN.L	Experian	Experian; Experian PLC; Experian plc; Experian Group limited; New Gemini limited; Jerseyco; Experian Group limited; Experian Group; New Gemini limited; New Gemini; Jerseyco;
EZJ.L	easyJet	easyJet; easyJet plc; Easyjet plc; EasyJet plc; easyjet; Easyjet; EasyJet; EasyJet plc; easyjet;
FERG.L	Ferguson	Ferguson; Ferguson Plc; Ferguson plc; Wolseley plc; Wolseley; Wolseley plc;
FGP.L	FirstGroup	FirstGroup; FirstGroup PLC; FirstBus plc; FirstGroup plc; FirstBus; FirstGroup plc;
FLG.L~D15	Friends Life	Friends Life; Friends Life Group Ltd; Friends Life Group;
FP.L~K09	Friends Life FPG	Friends Life FPG; Friends Life FPG Ltd; Friends Life; Friends Life;
FRE.L~C01	Orange Home UK	Orange Home UK; Orange Home UK Ltd; Freeserve; Freeserve plc; Wanadoo UK; Orange Home UK limited; ; Freeserve plc; Wanadoo UK; Orange Home UK limited; ;
FRES.L	Fresnillo	Fresnillo; Fresnillo PLC; Fresnillo plc; Fresnillo limited; Hackremco limited; Hackremco No. 2516; Fresnillo limited; Hackremco limited; Hackremco; Hackremco No. 2516;
FXPO.L	Ferrexpo	Ferrexpo; Ferrexpo PLC; Ferrexpo plc;
GACC.L~H98	General Accident	General Accident; General Accident PLC; General Accident plc; Aviva plc; Aviva; Aviva; Aviva plc;
GARD.L~G99	Guardian Ryl Ex	Guardian Ryl Ex; Guardian Royal Exchange PLC; Guardian Royal Exchange plc; Guardian Royal Exchange; Guardian Royal Exchange;
GB0274753.L~D04	GE Hlthcare	GE Hlthcare; GE Healthcare Ltd; GE Healthcare limited; GE Healthcare; General Electric Healthcare; GE Healthcare; General Electric Healthcare;
GFS.L	G4S	G4S; G4S PLC; G4F plc; Group 4 Securicor plc; Group 4 Securicor limited; G4F; Group 4 Securicor; Group 4 Securicor plc; Group 4 Securicor limited;
GKN.L	GKN	GKN; GKN PLC; GKN plc; New GKN plc; Mistycove plc; ; New GKN; New GKN plc; Mistycove; Mistycove plc; ;
GLEN.L	Glencore	Glencore; Glencore PLC; Glencore plc; Glecore Xstrata; Glencore International; Glecore Xstrata; Glencore International;
GLH.L~D07	Gallaher Group	Gallaher Group; Gallaher Group Ltd; Gallaher Group limited;
GME.L~B01	Granada Media	Granada Media; Granada Media Ltd; Granada Media limited; Granada Media plc; G.M.I limited; GMI; Granada Media plc; G.M.I limited; G.M.I; GMI;
GRM_pa.N~K04	Grand Metro	Grand Metro; Grand Metropolitan Ltd; Grand Metropolitan; Grand Metropolitan limited; Grand Metropolitan limited;
GSK.L	GSK	GSK; GlaxoSmithKline PLC; GlaxoSmithKline; Glaxo SmithKline; Trushelfco limited; Glaxosmithkline; Glaxo Smithkline; Glaxo SmithKline; Trushelfco; Glaxosmithkline; Glaxo Smithkline;
GUS.L~J06	Experian	Experian; Experian PLC; Experian plc; Experian Group limited; New Gemini limited; Jerseyco; Experian Group limited; Experian Group; New Gemini limited; New Gemini; Jerseyco;
HAYS.L	Hays	Hays; Hays PLC; Hays plc; Chainpoint limited; Climbasset limited; Chainpoint limited; Chainpoint; Climbasset limited; Climbasset;
HBOS.L~A09	HBOS	HBOS; HBOS PLC; HBOS plc;
HIBUgb.ISD~B16	Hibu	Hibu; Hibu PLC; Hibu plc; Yell Group plc; Yell Group limited; hibu; hibu plc; Yell Group; Yell Group plc; Yell Group limited; hibu; hibu plc;
HIK.L	Hikma Pharma	Hikma Pharma; Hikma Pharmaceuticals PLC; Hikma Pharmaceuticals plc; Hikma Pharmaceuticals; Hikma Pharma plc; Hikma; Hikma Pharmaceuticals; Hikma Pharma plc; Hikma;
HLMA.L	Halma	Halma; Halma PLC; Halma plc; Halma Investments limited; Halma Investments limited; Halma Investments;
HMSO.L	Hammerson	Hammerson; Hammerson PLC; Hammerson plc; Hammerson Property Investment and Development Corporation; Hammerson Property and Investment Trust; Hammerson Property Investment and Development Corporation; Hammerson Property and Investment Trust;
HNS.L~H07	Hanson	Hanson; Hanson Ltd; Hanson limited; Hanson Aggregates; Hanson Aggregates;
HOME.L~I16	Home Retail Grp	Home Retail Grp; Home Retail Group Ltd; Home Retail; Home Retail Group limited; Argos; Home Retail Group; Home Retail Group limited; Argos;
HRGV.L	Hargreaves	Hargreaves; Hargreaves Lansdown PLC; Hargreaves Lansdown plc; Quayshelfco 170 limited; Hargreaves Lansdown; Quayshelfco 170 limited; Quayshelfco 170;
HSBA.L	HSBC	HSBC; HSBC Holdings PLC; HSBC Holdings plc; Silom limited; Vernat Eastern Agencies limited; HSBC Holdings; Silom limited; Silom; Vernat Eastern Agencies limited; Vernat Eastern Agencies;
IAP.L~L16	ICAP	ICAP; ICAP PLC; ICAP plc;
ICAG.L	IAG	IAG; International Consolidated Airlines Group SA; International Consolidated Airlines Group; Global Baibco; International Consolidated Airlines; ICAG; British Airways; Iberia; Vueling; Aer Lingus; Avios; Global Baibco; International Consolidated Airlines; ICAG; British Airways; Iberia; Vueling; Aer Lingus; Avios;
ICI.L~A08	ICI	ICI; Imperial Chemical Industries Ltd; Imperial Chemicals Industries; Imperial Chemicals; Imperial Chemical; Imperial Chemical Industries plc; Imperial Chemical Industries; Imperial Chemicals; Imperial Chemical; Imperial Chemical Industries plc;
ICIn.L~E02	ICI	ICI; Imperial Chemical Industries Ltd; Imperial Chemicals Industries; Imperial Chemicals; Imperial Chemical; Imperial Chemical Industries plc; Imperial Chemical Industries; Imperial Chemicals; Imperial Chemical; Imperial Chemical Industries plc;
IHG.L	InterContinental	InterContinental; InterContinental Hotels Group PLC; InterContinental Hotels Group plc; Intercontinental Hotels; New Intercontinental Hotels Group plc; ; InterContinental Hotels Group; Intercontinental Hotels; New Intercontinental Hotels Group New Intercontinental Hotels Group plc; ;
III.L	3i Group	3i Group; 3i Group PLC; Finance for Industry plc; Investors in Industry Group plc; Finance for Industry; Investors in Industry Group; Investors in Industry Group plc;
IMB.L	Imperial Brands	Imperial Brands; Imperial Brands PLC; Imperial Brands plc; Imperial Brands; Imperial Tobacco; Imperial Tobacco Group plc; Imperial Tobacco Group; Imperial Tobacco; Imperial Tobacco Group; Imperial Tobacco Group plc;
IMI.L	IMI	IMI; IMI PLC; IMI plc; IMI Critical Engineering; Imperial Metals Industries limited; IMI Critical Engineering; Imperial Metals Industries limited; Imperial Metals Industries;
INCH.L	Inchcape	Inchcape; Inchcape PLC; Inchcape plc;
INDV.L	Indivior	Indivior; Indivior PLC; Indivior plc;
INF.L	Informa	Informa; Informa PLC; Informa plc; Informa limited; Informa limited;
INTUP.L	Intu Prop	Intu Prop; Intu Properties PLC; Intu Properties plc; Capital Shopping Centres Group plc; Liberty International plc; intu properties; intu Properties; intu; Intu Properties; Capital Shopping Centres Group; Capital Shopping Centres Group plc Liberty International; Liberty International plc; intu properties; intu Properties; intu;
INVP.L	Investec	Investec; Investec PLC; Investec plc; Regatta Services limited; Regatta Services limited; Regatta Services;
IOG.L~G02	Rwe Generation U	Rwe Generation U; Rwe Generation UK Holdings PLC; RWE Generation UK Holdings plc; Rwe Generation; RWE Generation; Innogy Holdings plc; Rwe Generation UK Holdings; Rwe Generation UK; RWE Generation UK Holdings; RWE Generation UK Rwe Generation; RWE Generation; Innogy Holdings; Innogy Holdings plc; Innogy;
IPRgb.PZ~A12	Internationl Pwr	Internationl Pwr; International Power Ltd; International Power limited; International Power plc; National Power plc; International Power; International Power plc; National Power; National Power plc;
ISA.L	Inmarsat	Inmarsat; Inmarsat PLC; Inmarsat plc; Inmarsat Group Holdings limited; Inmarsat Group; Duchessgrove limited; Inmarsat Group Holdings limited; Inmarsat Group Holdings; Inmarsat Group; Duchessgrove limited; Duchessgrove;
ISYS.L~A14	Invensys	Invensys; Invensys Ltd; Invensys International Holdings; Invensys International; Invensys International;
ITRK.L	Intertek Group	Intertek Group; Intertek Group PLC; Intertek Group plc; Intertek Testing Services plc; Intertek Testing Services limited; Intertek Testing Services; Intertek Testing Services plc; Intertek Testing Services limited;
ITV.L	ITV	ITV; ITV PLC; ITV plc; itv; 2232nd Single Member Shelf Investment Company limited; ; itv; 2232nd Single Member Shelf Investment Company limited; 2232nd Single Member Shelf Investment Company;
IVZ.L~L07	Invesco Holding	Invesco Holding; Invesco Holding Company Ltd; Invesco Holding Company limited; Invesco; Invesco Holding Company; Invesco;
JE.L	Just Eat	Just Eat; Just Eat PLC; Just Eat plc; Just-Eat Group Holdings; Just Eat Limited; Just-Eat Group Holdings; Just-Eat Group; Just Eat Limited;
JMAT.L	Johnson Matthey	Johnson Matthey; Johnson Matthey PLC; Johnson Matthey plc;

Table C.3: Full list of firm names (K-R)

Code	Company	Previous names
KAZ.L	Kaz Minerals	Kaz Minerals; Kaz Minerals PLC; Kaz Minerals plc; Kazakhmys plc; KCC International plc; Kazakhmys; Kazakhmys plc; KCC International; KCC International plc;
KCOM.L	KCOM Group	KCOM Group; KCOM Group PLC; KCOM Group plc; Kingston Communications plc; Kingston Communications (Hull); kcom group; Kingston Communications; Kingston Communications plc; Kingston Communications (Hull); kcom group;
KEL.L~B08	Kelda Group	Kelda Group; Kelda Group Ltd; Kelda Group limited; Kelda; Kelda;
KGF.L	Kingfisher	Kingfisher; Kingfisher PLC; Kingfisher plc;
LAND.L	Land Secs Group	Land Secs Group; Land Securities Group PLC; Land Securities Group plc; Hackplimco plc; Hackplimco (No. 104); Land Securities Group; Hackplimco; Hackplimco plc; Hackplimco (No. 104);
LANDn.L~C09		NA; Land Securities Group PLC; Land Securities Group;
LAT.L~J02	Lattice Grp	Lattice Grp; Lattice Group Ltd; Lattice Group limited; Lattice Group; National Grid; Lattice Group; National Grid;
LCL.L~C18	Ladbrokes Coral	Ladbrokes Coral; Ladbrokes Coral Group Ltd; Ladbrokes Coral Group; Ladbrokes; Coral; Radoburokusu Koraru Gurupu; Ladbrokes; Coral; Radoburokusu Koraru Gurupu;
LGEN.L	Legal General	Legal General; Legal & General Group PLC; Legal & General Group plc; Legal & General Group; Legal and General; Legal & General Group; Legal & General Group; Legal and General;
LLOY.L	Lloyds Bank UK	Lloyds Bank UK; Lloyds Banking Group PLC; Lloyds Banking Group; Lloyds TSB Group plc; Lloyds Bank; Lloyds TSB; Lloyds TSB Bank; Lloyds TSB Group; Lloyds TSB Group plc; Lloyds Bank; Lloyds TSB; Lloyds TSB Bank;
LMI.L	Lonmin	Lonmin; Lonmin PLC; Lonmin plc; Lonrho plc; ; Lonrho; Lonrho plc; ;
LOG.L~H12	Logica	Logica; Logica Ltd; Logica plc; CGI Group; CGI Group;
LSE.L	LSE	LSE; London Stock Exchange Group PLC; London Stock Exchange Group plc; London Stock Exchange; Milescreen limited; LSE Group; London Stock Exchange Group; London Stock Exchange; Milescreen limited; Milescreen; LSE Group;
LSMRx.L~E11	ENI Lasmo	ENI Lasmo; ENI Lasmo PLC; ENI Lasmoa plc; Eni Lasmoa; Eni Lasmo; Lasmo; Lasmoa; ENI Lasmoa; Eni Lasmoa; Eni Lasmo; Lasmo; Lasmoa;
LVA.L~E99	LucasVar	LucasVar; LucasVarity; Lucavarity; Lucas Varity plc; Lucavarity; Lucas Varity; Lucas Varity plc;
MAB.L	Mitchells Butler	Mitchells Butler; Mitchells & Butlers PLC; Mitchells & Butlers plc; Mitchells & Butlers; Mitchells and Butlers; Hackplimco (No. 111) plc; Mitchells & Butlers; Mitchells & Butlers; Mitchells and Butlers; Hackplimco (No. 111); Hackplimco (No. 111) plc;
MCRO.L	Micro Focus Inte	Micro Focus Inte; Micro Focus International PLC; Micro Focus; Micro Focus International plc; Micro Focus International; Hackremco limited; Hackremco (No. 2158) limited; Micro Focus International; Micro Focus International plc;
MDCM.L	Mediclinic Intl	Mediclinic Intl; Mediclinic International PLC; Mediclinic International plc; Mediclinic; Al Noor Hospitals Group plc; Al Noor Hospitals plc; Mediclinic International; Mediclinic; Al Noor Hospitals Group; Al Noor Hospitals Group plc; Al Noor Hospitals;
MEPC.L~J00	MEPC	MEPC; MEPC Ltd; MEPC limited; MEPC Business Space Management limited; MEPC Business Space Management limited; MEPC Business Space Management;
MERL.L	Merlin Ent	Merlin Ent; Merlin Entertainments PLC; Merlin Entertainments plc; Merlin Entertainments; Arthur Entertainments plc; Merlin Entertainments; Arthur Entertainments; Arthur Entertainments plc;
MGGT.L	Meggitt	Meggitt; Meggitt PLC; Meggitt plc; Meggitt Holdings plc; Meggitt Holdings; Meggitt Holdings plc;
MKS.L	M&S	M&S; Marks and Spencer Group PLC; Marks and Spencer Group plc; M & S; Marks and Spencer; Conker Retail Limited; Trushelfco limited; Trushelfco (No. 2827) limited; Marks and Spencer Group; M & S; Marks and Spencer; Conker Retail Limited; Conker Retail
MNDI.L	Mondi	Mondi; Mondi PLC; Mondi plc; Hackplimco No. 119 plc; Hackplimco No. 119; Hackplimco No. 119 plc;
MONI.L~E03		NA; M 2003 PLC; M 2003;
MRON.L	Melrose Inds	Melrose Inds; Melrose Industries PLC; Melrose; Melrose Industries; Melrose Industries plc; Melrose Industries; Melrose Industries plc;
MRW.L	Morrison Supermk	Morrison Supermk; WM Morrison Supermarkets PLC; Morrisons; WM Morrison; Morrison Supermarkets; WM Morrison Supermarkets; WM Morrison Supermarkets; WM Morrison; Morrison Supermarkets;
MSY.L~F12	Finastra Group	Finastra Group; Finastra Group Holdings Ltd; Finastra; Misys plc; Misys Newco; Finastra Group Holdings; Misys; Misys plc; Misys Newco;
NAMt.L~F98	GE Healthcare limited	GE Healthcare limited; GE Healthcare Ltd; Amersham plc; Nycomed Amersham plc; GE Healthcare; Amersham; Nycomed Amersham; Nycomed Amersham plc;
NG.L	National Grid	National Grid; National Grid PLC; National Grid plc; National Grid Transco plc; National Grid Group; National Grid Transco; National Grid Transco plc; National Grid Group;
NGn.L~F10		NA; National Grid PLC; National Grid;
NMC.L	NMC Health	NMC Health; NMC Health PLC; NMC Health plc;
NRK.L~C08	Landmark Mart	Landmark Mart; Landmark Mortgages Ltd; Landmark Mortgages; NRAM; NRAM plc; NRAM; NRAM plc;
NU.L~E00	Norwich Union	Norwich Union; Norwich Union Life Insurance Co Ltd; Norwich Union Life Insurance co; Norwich Union; The Norwich Union Life Insurance; Norwich Union Life Insurance Co; The Norwich Union Life Insurance;
NVR.L~D05	Novar	Novar; Novar Ltd; Novar limited; Novar plc; Novar plc;
NWB.L~D00	Natwest	Natwest; NATIONAL WESTMINSTER BANK PLC; National Westminster Bank plc; Natwest; Natwest Bank; NATIONAL WESTMINSTER BANK; National Westminster Bank; Natwest Bank;
NXT.L	Next UK	Next UK; Next PLC; Next plc; Next Group plc; Next; Bronzejasper plc; NEXT; Next; Next Group; Next Group plc; Bronzejasper; Bronzejasper plc; NEXT;
OML.L	Old Mutual	Old Mutual; Old Mutual PLC; Old Mutual plc; Old Mutual Group plc; Laudum (No. 1) limited; Old Mutual Group; Old Mutual Group plc; Laudum (No. 1) limited; Laudum (No. 1);
OOM.L~C06	Telefonica Erop	Telefonica Erop; Telefonica Europe PLC; Telefonica; O2 plc; mmO2 plc; Telefonica Europe; Telefonica Europe; O2; O2 plc; mmO2; mmO2 plc;
ORA.L~B00	Orange UK	Orange UK; Orange Ltd; Orange; Orange limited; Orange plc; Orange limited; Orange plc;
PFC.L	Petrofac	Petrofac; Petrofac Ltd; Petrofac limited;
PFG.L	Provident Fin	Provident Fin; Provident Financial PLC; Provident Financial; Provident Financial plc; Provident Financial Group plc; Provident Financial plc; Provident Financial Group; Provident Financial Group plc;
PILK.L~F06	Pilkington Group	Pilkington Group; Pilkington Group Ltd; Pilkington Group limited; Pilkington; NSG Group; Nippon Sheet Glass Co.; Pilkington; NSG Group; Nippon Sheet Glass Co.;
PNN.L	Pennon Group	Pennon Group; Pennon Group PLC; Pennon Group plc;
PO_p.L~C06	Penins Oriental	Penins Oriental; Peninsular and Oriental Steam Navigation Co; Peninsular and Oriental Steam Navigation co; The Peninsular and Oriental Steam Navigation; Peninsular and Oriental; Peninsular & Oriental; The Peninsular and Oriental Steam Navigation;
POLYP.L	Polymetal Intl	Polymetal Intl; Polymetal International PLC; Polymetal International plc; Polymetal; Polymetal International; Polymetal International; Polymetal;
PPB.L	Paddy Power	Paddy Power; Paddy Power Betfair PLC; Paddy Power Betfair plc; Paddy Power plc; Paddy Power Betfair; Paddy Power Betfair; Paddy Power plc;
PRU.L	Prudential	Prudential; Prudential PLC; Prudential plc; Prudential Corporation plc; Prudential Corporation; Prudential Corporation plc;
PSN.L	Persimmon	Persimmon; Persimmon PLC; Persimmon plc;
PSON.L	Pearson	Pearson; Pearson PLC; Pearson plc;
PUB.L~H17	Punch Taverns	Punch Taverns; Punch Taverns Ltd; Punch Tabenzu; Punch Taverns limited; Punch Taverns limited;
PWG.L~G02	EO UK	EO UK; E.ON UK PLC; E.ON UK plc; E.ON; Powergen UK; Powergen plc; E.ON Energy; E.ON UK; E.ON; Powergen UK; Powergen; Powergen plc; E.ON Energy;
RB.L	Reckitt Benc Grp	Reckitt Benc Grp; Reckitt Benckiser Group PLC; Reckitt Benckiser Group plc; Reckitt Benckiser; Reckitt Benckiser Group limited; Reckitt Benckiser Group; Reckitt Benckiser; Reckitt Benckiser Group limited;
RBS.L	Royal Bank Scot	Royal Bank Scot; Royal Bank of Scotland Group PLC; RBS; Royal Bank of Scotland; The Royal Bank of Scotland Group plc; The Royal Bank of Scotland; Royal Bank of Scotland Group; The Royal Bank of Scotland Group;
RDSa.L	Shell	Shell; Royal Dutch Shell PLC; Royal Dutch Shell; Royal Dutch Shell plc; Royal Dutch Shell plc;
RDSb.L	Shell	Shell; Royal Dutch Shell PLC; Royal Dutch Shell; Royal Dutch Shell plc; Royal Dutch Shell plc;
REL.L	Relx	Relx; Relx PLC; RELX plc; Reed Elsevier plc; Reed International; Relx plc; RELX; Reed Elsevier; Reed Elsevier plc; Reed International; Relx plc;
REX.L~G16	Rexam	Rexam; Rexam Ltd; Rexam limited; Rexam plc; Rexam plc;
RIIG.L~A14	Resources	Resources; Resources in Insurance Group PLC; Resources in Insurance Group; Resources in Insurance Group plc; Resources in Insurance Group plc;
RIO.L	Rio Tinto PLC	Rio Tinto PLC; Rio Tinto;
RMC.L~C05	Cemex Invsts	Cemex Invsts; Cemex Investments Ltd; Cemex; Cemex Investments limited; Cemex Investments; Cemex Investments; Cemex Investments limited;
RMG.L	Royal Mail	Royal Mail; Royal Mail PLC; Royal Mail plc; Royal Mail limited; Royal Mail Group; Royal Mail limited; Royal Mail Group;
RNK.L	Rank Group GB	Rank Group GB; Rank Group PLC; The Rank Group plc; Megastorm plc; Rank Group; The Rank Group; Megastorm; Megastorm plc;
RR.L	Rolls-Royce Hldg	Rolls-Royce Hldg; Rolls-Royce Holdings PLC; Rolls-Royce; Rolls-Royce Holdings plc; Rolls-Royce Holdings limited; Rolls Royce; Rolls-Royce Holdings; Rolls-Royce Holdings plc; Rolls-Royce Holdings limited; Rolls Royce;
RRS.L	Randgold Rsrcs	Randgold Rsrcs; Randgold Resources Ltd; Randgold; Randgold Resources limited; Randgold Resources; Randgold Resources; Randgold Resources limited;
RSA.L	RSA Ins Grp	RSA Ins Grp; RSA Insurance Group PLC; RSA Insurance Group; Royal & Sun Alliance Insurance Group plc; Sun Alliance Group plc; RSA Group; Royal & Sun Alliance Insurance Group; Sun Alliance Group;
RSL.L~E08	Perl Grp No 1	Perl Grp No 1; Pearl Group Holdings (No. 1) Ltd; Pearl Group Holdings limited; Pearl Group; Resolution plc; Pearl Group Holdings No. 1; Pearl Group Holdings (No. 1); Pearl Group (No. 1); Pearl Group Holdings; Pearl Group; Resolution; Resolution plc;
RTK.L~L02	RT Group	RT Group; Rt Group PLC; RT Group plc; Network Raill Railtrack Group plc; Rt Group; Network Raill Railtrack Group; Network Raill Railtrack Group plc;
RTO.L	Rentokil Initial	Rentokil Initial; Rentokil Initial PLC; Rentokil Initial plc; Rentokil Initial 2005 plc; Rentokil; Rentokil Initial 2005; Rentokil Initial 2005 plc; Rentokil;

Table C.4: Full list of firm names (S-Z)

Code	Company	Previous names
SAB.L^J16	Abi Sab Group	Abi Sab Group; Abi Sab Group Holding Ltd; Sabmiller Limited; SABMiller plc; Abi Sab Group Holding; Sabmiller; SABMiller; SABMiller plc;
SB.L^L00	SmithKline Beech	SmithKline Beech; Beecham Group P L C; Beecham Group plc; Smithkline Beecham; SB; Beecham Group; Smithkline Beecham; SB;
SBRY.L	Sainsbury	Sainsbury; J Sainsbury PLC; J Sainsbury plc; J Sainsbury; J Sainsbury;
SCTN.L^D08	Scottish Ltd	Scottish Ltd; Scottish & Newcastle Ltd; Scottish & Newcastle limited; Scottish & Newcastle plc; Scottish and Newcastle; Scottish; Scottish & Newcastle; Scottish & Newcastle; Scottish & Newcastle plc; Scottish and Newcastle;
SDR.L	Schroders	Schroders; Schroders PLC; Schroders plc; Schroder Holdings plc; New Schroders plc; Schroder Holdings; Schroder Holdings plc; Schroder; New Schroders; New Schroders plc;
SEAR.L^C99	Sear	Sear; Sears Ltd; Sears limited; Sears; Sears;
SEL.L^B99	Sse Services	Sse Services; Sse Services PLC; SSE Services plc; Southern Electric plc; Sse Services plc; SSE Services; Southern Electric; Southern Electric plc; Sse Services plc;
SEM.L^E01	SEMA	SEMA; Sema PLC; Sema plc; Sema; SchlumbergerSema; Sema; SchlumbergerSema;
SFW.L^C04	Safeway	Safeway; Safeway Ltd; Safeway Limited; Safeway plc; Argyll Group plc; Safeway plc; Argyll Group; Argyll Group plc;
SGC.L	Stagecoach Grp	Stagecoach Grp; Stagecoach Group PLC; Stagecoach Group plc; Stagecoach Holdings plc; Stagecoach; Stagecoach Group; Stagecoach Holdings; Stagecoach Holdings plc; Stagecoach;
SGE.L	Sage Group	Sage Group; Sage Group PLC; The Sage Group; Ployfinal plc; The Sage Group plc; Ployfinal; Ployfinal plc; The Sage Group plc;
SGRO.L	SEGRO	SEGRO; SEGRO PLC; Segro plc; Slough Estates plc; Slough Estates limites; Segro; SEGRO plc; Segro; Slough Estates; Slough Estates plc; Slough Estates limites; SEGRO plc;
SHP.L	Shire	Shire; Shire PLC; Shire plc; Shire limited; MDFJ 2034773; Shire limited; MDFJ 2034773;
SJP.L	St James's Place	St James's Place; St. James's Place PLC; St James's Place plc; St James's Place Capital plc; St James Place; St. James's Place; St James's Place Capital; St James's Place Capital plc; St James Place;
SKG.L	Smurfit Kapa Grp	Smurfit Kapa Grp; Smurfit Kappa Group PLC; Smurfit Kappa Group; Smurfit Kappa Group plc; Smurfit Kappa; Smurfit Kappa Group plc; Smurfit Kappa;
SKYB.L	Sky	Sky; Sky PLC; Sky plc; British Sky Broadcasting Group plc; British Sky Broadcasting; BSkyB; British Sky Broadcasting Group; British Sky Broadcasting Group plc; British Sky Broadcasting; BSkyB;
SLA.L	Standard Life	Standard Life; Standard Life Aberdeen PLC; Standard Life Aberdeen; Standard Life plc; SLGC limites; Standard Life plc; SLGC limites;
SLP.L^G00	AXA UK	AXA UK; AXA UK PLC; Sun Life and Provincial Holdings plc; UAP Holdings limited; Provincial Group Holdings limited; Rockleigh limited; Sun Life and Provincial Holdings; Sun Life and Provincial; UAP Holdings limited; UAP Holdings; UAP; Provincial Group Holdings limited; Provincial Group Holdings; Provincial Group; Rockleigh limited; Rockleigh;
SMDS.L	DS Smith	DS Smith; DS Smith PLC; DS Smith plc; David S. Smith (Holdings); David S. Smith (Packaging); David S. Smith (Holdings); David S. Smith (); David S. Smith (Packaging);
SMIN.L	Smiths Group	Smiths Group; Smiths Group PLC; Smiths Industries plc; Smiths Industries;
SN.L	Smith & Nephew	Smith & Nephew; Smith & Nephew PLC; Smith and Nephew; Smith & Nephew;
SPD.L	Sports Direct	Sports Direct; Sports Direct International PLC; Sports Direct International plc; Sports Direct International; Sports Direct International;
SPT.L	Spirent	Spirent; Spirent Communications plc; Spirent Communications; Spirent plc; Bowthorpe plc; Spirent plc; Bowthorpe; Bowthorpe plc;
SPW.L^F07	Scottish Power	Scottish Power; Scottish Power Ltd; Scottish Power limited; SCOTTISHPOWER Scottishpower; SCOTTISHPOWER Scottishpower;
SRP.L	Serco Group	Serco Group; Serco Group PLC; Serco; Serco Group plc; Serco; Serco Group plc;
SSE.L	SSE	SSE; SSE PLC; SSE plc; Scottish and Southern Energy plc; Scottish Hydro-electric plc; Scottish Hydro-Electric; Scottish and Southern Energy; Scottish and Southern Energy plc; Scottish Hydro-electric; Scottish Hydro-electric plc; Scottish Hydro-Electric;
STAN.L	StanChart	StanChart; Standard Chartered PLC; Standard Chartered plc; Standard Chartered;
SVT.L	Severn Trent	Severn Trent; Severn Trent PLC; Severn Trent plc;
SXC.L^D03	Six Continents	Six Continents; Six Continents Ltd; Six Continents limited; Six Continents plc; Six Continents plc;
TATE.L	Tate & Lyle	Tate & Lyle; Tate & Lyle PLC; Tate & Lyle plc; Tate and Lyle; Tate & Lyle; Tate and Lyle;
TCG.L	Thomas Cook Grp	Thomas Cook Grp; Thomas Cook Group plc; Thomas Cook; Thomas Cook Group; Thomas Cook Group;
TEG.L^F98	ENERG 3	ENERG 3; Energy Holdings (No.3) Ltd; Energy Holdings (No. 3) limited; Energy Group; Peabody; Eastern Group; Energy Holdings (No.3); Energy (No.3); Energy Holdings (No. 3); Energy (No. 3); Energy Group; Peabody; Eastern Group;
THN.L^K98	Thorn	Thorn; Thorn PLC; Thorn limited; Thorn plc; ; Thorn plc; ;
THUS.L^K08	Thus Group	Thus Group; Thus Group Ltd; Thus Group limited; Vodafone Group; Vodafone Group;
TI.L^L00	T I Group	T I Group; T I Group Ltd; T I Group limited; Tube Investments; T I Group plc; Tube Investments; T I Group plc;
TLW.L	Tullow	Tullow; Tullow Oil PLC; Tullow Oil plc; Tullow Oil; Tullow Oil;
TOMK.L^I10	Gates Worldwide	Gates Worldwide; Gates Worldwide Ltd; Gates Worldwide limited; ; ;
TPK.L	Travis Perkins	Travis Perkins; Travis Perkins PLC; Travis Perkins plc;
TRIL.L^I09	Thomson Reut UK	Thomson Reut UK; Thomson Reuters UK Ltd; Thomson Reuters; Thomson Reuters UK; Thomson Reuters UK;
TSCO.L	Tesco	Tesco; Tesco PLC; Tesco plc;
TT.L^L14	TUI	TUI; TUI Travel Ltd; TUI Travel limited; TUI AG; Tui AG; TUI Travel; Tui AG; Tui;
TUIJe.L^B15		NA; Tui AG; Tui;
TUIT.L	TUI	TUI; Tui AG; TUI Travel limited; TUI AG; Tui AG; Tui; TUI Travel;
TW.L	Taylor PLC	Taylor PLC; Taylor Wimpey PLC; Taylor Wimpey plc; Taylor Wimpey; Taylor Woodrow plc; Taylor; Taylor Wimpey; Taylor Woodrow; Taylor Woodrow plc;
TW.L^B00	Thames Wtr	Thames Wtr; Thames Water Ltd; Thames Water; Thames Water limited; Thames Water limited;
TWT.L^G04	Telewest Com Net	Telewest Com Net; Telewest Communications Networks Ltd; Telewest; Telewest Communications Networks limited; Telewest Communications; Titan Cable plc; Telewest Communications Networks; Telewest Communications Networks limited; Titan Cable;
UBIS.L^F00	Utd Biscuits	Utd Biscuits; United Biscuits (Holdings) Ltd; United Biscuits; United Biscuits (Holdings) limited; United Biscuits (Holdings); United Biscuits (); United Biscuits (Holdings) limited;
UBM.L	UBM	UBM; UBM PLC; UBM plc; United Business Media limited; Project Peach limited; United Business Media limited; United Business Media; Project Peach limited; Project Peach;
ULVR.L	Unilever	Unilever; Unilever PLC; Unilever plc;
UU.L	United UT Grp	United UT Grp; United Utilities Group PLC; United Utilities; United Utilities Group; United Utilities Group;
UUa.L^G05		NA; United Utilities Group PLC; United Utilities Group;
UUn.L^I03		NA; United Utilities Group PLC; United Utilities Group;
VED.L	Vedanta Res PLC	Vedanta Res PLC; Vedanta Resources PLC; Vedanta Resources plc; Vedanta Resources; Vedanta; Angelchange limited; Vedanta Res; Vedanta Resources; Vedanta; Angelchange limited; Angelchange;
VOD.L	Vodafone Group	Vodafone Group; Vodafone Group PLC; Vodafone; Racal Telecom; Racal Telecom;
VSVS.L	Vesuvius	Vesuvius; Vesuvius PLC; Vesuvius plc; Vesuvius Technologies; Vesuvius Technologies;
WEIR.L	Weir Group	Weir Group; Weir Group PLC; Weir Group plc; The Weir Group ; The Weir Group;
WG.L	John Wood	John Wood; John Wood Group PLC; John Wood Group; John Wood Group plc; John Wood Group plc;
WPG.L^A18	Worldpay Group	Worldpay Group; Worldpay Group Ltd; Worldpay; Worldpay Group limited; Worldpay Group; Worldpay Group limited;
WPP.L	WPP	WPP; WPP PLC; WPP Group plc; WPP plc; WPP 2012; WPP Group; WPP Group plc; WPP plc; WPP 2012;
WTB.L	Whitbread	Whitbread; Whitbread PLC; Whitbread plc; Whitbread Holdings; Whitbread Holdings;
WWH.L^J00	Woolwich	Woolwich; Woolwich Ltd; Woolwich limited; ; ;
XTA.L^E13	Xstrata Ltd	Xstrata Ltd; Xstrata;

Table C.5: Number of firm-day observations matched to at least one article

	Ticker	Company	x
38	BP.L	BP	1169
31	BARC.L	Barclays	907
42	BT.L	BT Group	887
221	RDSa.L	Shell	790
222	RDSb.L	Shell	790
160	LSE.L	London Stock Exchange	668
205	RBS.L	Royal Bank of Scotland	610
125	HSBA.L	HSBC	583
213	SDR.L	Schroders	419
35	BLT.L	BHP	384
225	SKYB.L	Sky	359
143	ITV.L	ITV	335
114	GSK.L	GSK	334
268	WPP.L	WPP	295
236	STAN.L	StanChart	285
242	TW.L	Taylor PLC	263
194	PRU.L	Prudential	237
261	VOD.L	Vodafone Group	226
185	PSON.L	Pearson	215
157	LLOY.L	Lloyds Bank UK	196
245	TSCO.L	Tesco	167
203	RIO.L	Rio Tinto PLC	158
128	ICI.L^A08	ICI	149
121	HBOS.L^A09	HBOS	142
24	AV.L	Aviva GB	126
212	SBRY.L	Sainsbury	126
210	SFW.L^C04	Safeway	125
110	GKN.L	GKN	124
162	MKS.L	M&S	123
156	BAA.L^H06	LHR Airports	121
118	HNS.L^H07	Hanson	120
237	SLA.L	Standard Life	118
28	BALF.L	Balfour Beatty	111
75	DXNS.L^H14	Dixons Retail	110
174	NG.L	National Grid	108
9	AB.L^G07	Alliance Boots Holdings Ltd	106
47	CWC.L^E16	C&W	104
126	ICAG.L	IAG	104
12	ALLD.L^G05	Allied Domecq	102
247	TCG.L	Thomas Cook Grp	99
231	SPD.L	Sports Direct	97
262	AUN.L^G06	WBA Holdings 1	96
27	BAES.L	BAE Systems	95
63	COLT.F^K15	Colt Group	95
34	BG.L^B16	BG Grp	90
50	CNE.L	Cairn Energy	89
37	BOC.L^I06	BOC Grop	83
163	EMG.L	Man Group	82
256	ULVR.L	Unilever	82
49	CBRY.L^C10	Cadbury	79
199	REL.L	Relx	78
87	ENRC.L^K13	ENRC	76
76	DMGOa.L	DMGT	74
3	ABF.L	ABF	72
67	CRH.L	CRH	70
255	UBM.L	UBM	70
206	RMG.L	Royal Mail	69
111	GLEN.L	Glencore	68
124	HOME.L^I16	Home Retail Grp	68
22	AZN.L	AstraZeneca	65
145	JMAT.L	Johnson Matthey	65
53	CCL.L	Carnival	63
204	RR.L	Rolls-Royce Hldg	63
201	RIIG.L^A14	Resources	59
131	IMB.L	Imperial Brands	57
254	TLW.L	Tullow	57
181	OML.L	Old Mutual	56
269	XTA.L^E13	Xstrata Ltd	56
240	BCI.L^H01	Tarmac Cement An	55
41	BLND.L	British Land	52
251	TPK.L	Travis Perkins	52
259	VED.L	Vedanta Res PLC	52
127	IAP.L^L16	ICAP	51
56	CNA.L	Centrica	47
70	DLAR.L	De La Rue	47
79	EZJ.L	easyJet	47
257	UU.L	United UT Grp	46
71	DGE.L	Diageo	45
97	FERG.L	Ferguson	45
130	IMI.L	IMI	45
223	SHP.L	Shire	45
244	TWT.L^G04	Telewest Com Net	44
264	WTB.L	Whitbread	42
10	ATST.L	Alliance Trust plc	41
159	LMI.L	Lonmin	41
136	IHG.L	InterContinental	40
183	ORA.L^B00	Orange UK	40
39	BPB.L^A06	BPB	39
200	RTO.L	Rentokil Initial	35
169	MCRO.L	Micro Focus Inte	34
4	SAB.L^J16	Abi Sab Group	33
151	LCL.L^C18	Ladbrokes Coral	33
235	SGC.L	Stagecoach Grp	33
150	KGF.L	Kingfisher	32
219	SRP.L	Serco Group	31
232	SSE.L	SSE	31
176	NXT.L	Next UK	30
220	SVT.L	Severn Trent	30
139	ISYS.L^A14	Invensys	29
104	GFS.L	G4S	28
119	HRGV.L	Hargreaves	28
227	SB.L^L00	SmithKline Beech	28
228	SMIN.L	Smiths Group	28
252	TT.L^L14	TUI	28
99	MSY.L^F12	Finastra Group	27
180	CWZ.L^E00	NTL CWC	27
61	COB.L	Cobham	26
166	MRON.L	Melrose Inds	26
184	PPB.L	Paddy Power	26
144	WG.L	John Wood	25
178	NU.L^E00	Norwich Union	23
191	PILK.L^F06	Pilkington Group	22
17	AG.L^L02	Arcadia Group	21
59	CMG.L^L02	CMG PLC	21
90	DVR.L^I06	Ereved Group	21
120	HAYS.L	Hays	21
196	RRS.L	Randgold Rsrcs	21
202	REX.L^G16	Rexam	21
230	SPT.L	Spirent	21
85	EGS.L^I02	Energis	20
188	RSL.L^E08	Perl Grp No 1	20
214	SCTN.L^D08	Scottish Ltd	20
217	SGRO.L	SEGRO	20
158	LOG.L^H12	Logica	19
69	DCC.L	DCC	18
215	SPW.L^F07	Scottish Power	18
60	CO.L^F03	Coats Holdings	16
149	KEL.L^B08	Kelda Group	16
189	PSN.L	Persimmon	16
263	WEIR.L	Weir Group	16
54	CCH.L^H04	Celltech Group	15
93	EXL.L^L05	Exel	15
100	FGP.L	FirstGroup	15
216	SEAR.L^C99	Sear	15
51	CPI.L	Capta	14
117	HMSO.L	Hammerson	14
147	KAZ.L	Kaz Minerals	13
182	FRE.L^C01	Orange Home UK	13
190	PFC.L	Petrofac	13
243	OOM.L^C06	Telefonica Erop	13
15	AAL.L	Anglo American	12
74	DC.L	Dixons Carphone	12
95	EXPN.L	Experian	12
123	HIK.L	Hikma Pharma	12
16	ANTO.L	Antofagasta	11
23	AUTN.L^K11	Autonomy	11
43	BNZL.L	Bunzl	11
81	EIGE.L	EI Grp	11
134	INF.L	Informa	11
248	TRIL.L^I09	Thomson Reut UK	11
57	CHB.L^H03	Chubb UK	10
73	DLGD.L	DIRECT LINE INS	10
88	ETP.L^F02	Enterprise Oil	10
253	TUIT.L	TUI	10
78	SMDS.L	DS Smith	9
135	ISA.L	Inmarsat	9
258	UBIS.L^F00	Utd Biscuits	9
7	AGGK.L	Aggreko	8
19	ASSD.L^J99	ASDA Group	8
55	RMC.L^C05	Cemex Invsts	8
142	INVP.L	Investec	8
154	LAT.L^J02	Lattice Grp	8
179	NVR.L^D05	Novar	8
2	ADN.L^H17	Aberdeen Inv	7
98	FXPO.L	Ferrexpo	7
165	MGGT.L	Meggitt	7
226	SN.L	Smith & Nephew	7
14	AMFW.L^J17	Amec Foster	6
91	ESSR.L^F14	Essar Engr	6
116	HLMA.L	Halma	6
148	KCOM.L	KCOM Group	6
161	LVA.L^E99	LucasVar	6
173	AML^B16	MS Amlin	6
175	NWB.L^D00	Natwest	6
238	ARM.L^I16	Svf Holdco Uk	6
1	III.L	3i Group	5
64	CPG.L	Compass Group	5
138	INTUP.L	Intu Prop	5
167	MEPC.L^J00	MEPC	5
197	RNK.L	Rank Group GB	5
267	WPG.L^A18	Worldpay Group	5
5	ACAA.L	Acacia Mining	4
33	BKGH.L	Berkeley Group	4
52	CCM.L^A04	Carlton Comm	4
82	ECM.L	Electrocomponent	4
132	INCH.L	Inchcape	4
141	IVZ.L^L07	Invesco Holding	4
171	MNDI.L	Mondi	4
172	MRW.L	Morrison Supermk	4
192	PLYP.L	Polymetal Intl	4
207	RSA.L	RSA Ins Grp	4
211	SGE.L	Sage Group	4
218	SEM.L^E01	SEMA	4
224	SXC.L^D03	Six Continents	4
30	BSCT.L^I01	Bank of Scotland	3
48	CWP.L^G12	Cable & Wireles	3
58	CAPC.L^J15	Clean Air	3
72	DDT.L^L10	Dimension Data	3
102	FLG.L^D15	Friends Life	3
164	MDCM.L	Mediclinic Intl	3
198	RB.L	Reckitt Benc Grp	3
233	SEL.L^B99	Sse Services	3
246	TW.L^B00	Thames Wtr	3
21	AHT.L	Ashtead Group	2
45	BMAH.L^I00	Burmah Castrol	2
68	CRDA.L	Croda	2
86	ENQ.L	EnQuest	2
89	PWG.L^G02	EO UK	2
92	EVRE.L	EVRAZ	2
101	FRES.L	Fresnillo	2
112	GME.L^B01	Granada Media	2
155	LGEN.L	Legal General	2
170	MAB.L	Mitchells Butler	2
187	PNN.L	Pennon Group	2
193	PFG.L	Provident Fin	2
229	SKG.L	Smurfit Kapa Grp	2
249	THN.L^K98	Thorn	2
260	VSVS.L	Vesuvius	2
6	ADML.L	Admiral Group	1
13	ADZ.L^J00	Allied Zurich	1
25	SLP.L^G00	AXA UK	1
26	BAB.L	Babcock Intl	1
32	BDEV.L	Barratt Developments	1
77	DRX.L	Drax Group	1
80	BGY.L^B09	EDF Nuclear	1
83	EMI.L^I07	EMI Group	1
84	TEG.L^F98	ENERG 3	1
94	EXL.L^F00	Exel Plc	1
96	GUS.L^J06	Experian	1
105	GLH.L^D07	Gallaher Group	1
115	GARD.L^G99	Guardian Ryl Ex	1
133	INDV.L	Indivior	1
146	JE.L	Just Eat	1
168	MERL.L	Merlin Ent	1
209	IOG.L^G02	Rwe Generation U	1
250	THUS.L^K08	Thus Group	1
265	AWA.L^H00	Windward	1
8	CTLD.L^I98	Akzo Nobel UK	0
11	ALLL.L^J08	Alliance UK	0
18	EMA.L^C08	Ascential Group	0
20	ASHM.L	Ashmore Group	0
29	BLM.L^B05	Baltimore Capital	0
36	BI.L^L06	BIP Industries	0
40	BATS.L	Brit Am Tobacco	0
44	BRBY.L	Burberry Group	0
46	BPTY.L^B16	bwin.party	0
62	CCH.L	Coca-Cola HBC AG	0
65	CTEC.L	ConvaTec Group	0
66	CS.L^D07	Corus Group	0
103	FP.L^K09	Friends Life FPG	0
106	TOMK.L^I10	Gates Worldwide	0
107	NAMt.L^F98	GE Healthcare limited	0
108	GB0274753.L^D04	GE Hlthcare	0
109	GACC.L^H98	General Accident	0
113	GRM_pa.N^K04	Grand Metro	0
122	HIBUgb.ISD^B16	Hibu	0
129	ICIn.L^E02	ICI	0
137	ITRK.L	Intertek Group	0
140	BTR.L^B99	Invensys Intl UK	0
152	LAND.L	Land Secs Group	0
153	NRK.L^C08	Landmark Mart	0
177	NMC.L	NMC Health	0
186	PO_p.L^C06	Penins Oriental	0
195	PUB.L^H17	Punch Taverns	0
208	RTK.L^L02	RT Group	0
234	SJP.L	St James's Place	0
239	TI.L^L00	T I Group	0
241	TATE.L	Tate & Lyle	0
266	WWH.L^J00	Woolwich	0

C.2 Robustness of media coverage effect

C.2.1 Alternative matching strategies

Relying on just one of the two matching methods does not materially change results, neither does excluding outliers ($vol_{i,t} > 25\%$) or any observation with $vol_{i,t} > 10\%$. Splitting the sample into 1998-2006, 2007-2009 and 2010-2007 in order to ensure that the Great Recession is not driving results, also does not qualitatively change the results. A

regression on trading volume rather than volatility also yields similar results.

Table C.6: Robustness of the media coverage effect

	Dependent variable:								
	$vol_{i,t}$								Volume
	(baseline)	(NER)	(headline)	($vol < 25\%$)	($vol < 10\%$)	(98-06)	(07-09)	(10-17)	(Volume)
mention	0.160***			0.146***	0.104***	0.139***	0.255***	0.129***	1.206m***
	(0.016)			(0.015)	(0.012)	(0.028)	(0.043)	(0.020)	(0.085m)
ner_mention		0.101***							
		(0.009)							
head_mention			0.157***						
			(0.011)						
$vol_{i,t}$ lags	✓	✓	✓	✓	✓	✓	✓	✓	
Volume lags									✓
Firm fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓
Time fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓
Observations	735,517	735,517	735,517	730,790	648,267	319,482	113,973	299,832	727,050
R^2	0.498	0.498	0.498	0.496	0.415	0.457	0.571	0.418	0.745
Adjusted R^2	0.495	0.495	0.495	0.492	0.411	0.453	0.567	0.414	0.743
Residual Std. Error	1.870	1.870	1.870	1.704	1.292	2.202	1.995	1.357	9,762,422.000

Note: *p<0.1; **p<0.05; ***p<0.01

10 lags of the dependent variable are included in each specification

C.2.2 Forward looking articles

I use the LIWC software to classify articles based on their past, present and future focus. Some examples are given in Table C.7, showing that articles which feature words such as "will" and "expected" are more likely to be classified as forward looking.

Table C.7: Examples of LWIC assignment

	LIWC classification		
Text	$past_{i,t}$	$present_{i,t}$	$future_{i,t}$
Logica said revenue in its telecoms business had jumped 69 per cent.	0.167	0.000	0.000
GlaxoSmithKline will consider selling part of its research operations if their discovery of new drugs does not accelerate.	0.000	0.160	0.040
Strong earnings growth is expected when oil group BP Amoco posts its first-quarter results tomorrow.	0.000	0.039	0.115

I also run the same regression as in Table 3.4, but with alternative word lists for future and past tense. As shown in Table C.8, the results are qualitatively very similar as those using LIWC. The alternative word lists I use are:

- future(alt): "will", "expect", "set to", "future", "might", "could", "should", "uncertain", "tomorrow", "today", "later", "soon"
- past(alt): "fell", "rose", "increased", "decreased".

Table C.8: Alternative measure for forward looking articles also has a greater effect that backward looking articles.

	Dependent variable:
	$vol_{i,t}$
	(1)
$mention_{i,t}$	0.125^{***}
	(0.029)
$future(alt)_{i,t}$	6.029^{***}
	(1.079)
$past(alt)_{i,t}$	-0.510
	(2.943)
$vol_{i,t}$ lags	✓
Firm fixed effects	✓
Time fixed effects	✓
Observations	735,517
R^2	0.498
Adjusted R^2	0.495
Residual Std. Error	1.870
Note:	*p<0.1; **p<0.05; ***p<0.01

C.2.3 Sentiment measure

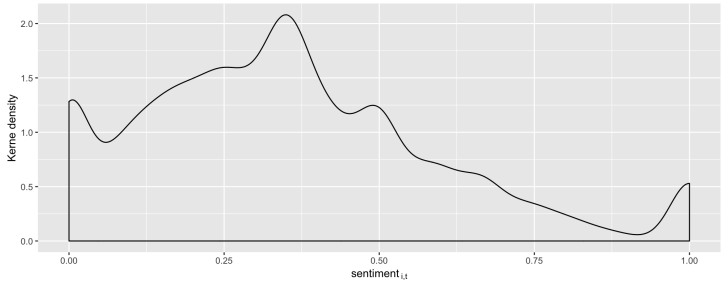
The sentiment measure is then constructed by computing the proportion of total "sentiment" words that are classified as positive by Loughran and McDonald (2011).

$$sentiment_{i,t} = \frac{positive_{i,t}}{positive_{i,t} + negative_{i,t}}$$

where $positive_{i,t}$ is the number of words in an article for firm i at time t which feature in the "positive" list, and $negative_{i,t}$ the number of words that feature in the "negative" list. The measure is thus bounded between zero and one with an average of 0.35, and a

higher value implies a more "positive" sentiment.

Figure C.3: The sentiment measure is bounded by 0 and 1, with a mean of 0.35



As shown in Section 3.4.2, sentiment of media coverage does not predict intra-day returns (i.e. open to close) on that day. However, it does predict the overnight returns (i.e. close previous day to open today), presumably as the morning newspaper may report on overnight events. Overnight returns are defined as the percentage change from the previous day's closing price to the opening price that day:

$$\Delta p_{i,t}^{c,o} = 100 \times \frac{p_{i,t}^{open} - p_{i,t-1}^{close}}{p_{i,t-1}^{close}}$$

Distinguishing between the intra-day and overnight return is important as the FT is published during the overnight window, so it helps demonstrate the importance of the timings described in the identification strategy.

Table C.9: Sentiment predicts overnight returns

	Dependent variable:	
	$\Delta p_{i,t}^{c,o}$	
	(1)	(2)
$mention_{i,t}$	0.035^{***}	0.041^{***}
	(0.013)	(0.013)
$sentiment_{i,t}$	0.101^{**}	
	(0.050)	
$sentiment_{i,t+1}$		0.866^{***}
		(0.050)
$\Delta p_{i,t}^{o,c}$	-0.244^{***}	-0.242^{***}
	(0.001)	(0.001)
$\Delta p_{i,t}^{c,o}$ lags	✓	✓
Firm fixed effects	✓	✓
Time fixed effects	✓	✓
Observations	707,836	
R^2 0.578	0.576	
Adjusted R^2 0.575	0.573	
Residual Std. Error 1.432	1.433	
Note:	*p<0.1; **p<0.05; ***p<0.01	

C.2.4 BTR specification

Table C.10 shows various hyperparameter and estimation choices made for the topic models in Section 3.5. The BTR and sLDA models are estimated using the EM-Gibbs algorithm described in **Chapter 2**. the unsupervised LDA is estimated with the collapsed Gibbs sampler from Griffiths and Steyvers (2004).

Table C.10: Topic model hyperparameters

	K	α	η	a_0	b_0	m_0	S_0	E-step iter	Gibbs iter
LDA	$\{2, 5, 10, 15, 20, 25, 30, 40, 50\}$	0.5	0.1	-	-	-	-	-	10,000
sLDA	$\{2, 5, 10, 15, 20, 25, 30, 40, 50\}$	0.5	0.1	2	2	0	2	100	-
BTR	$\{2, 5, 10, 15, 20, 25, 30, 40, 50\}$	0.5	0.1	2	2	0	2	100	-

Table C.11: Topic model results

K	BTR	BTR (sentiment interaction)	LDA	sLDA
2	0.0937	0.1213	0.1199	0.1081
	(0.0483)	(0.0136)	(0.01368)	(0.0131)
5	0.0934	0.1249	0.1210	-0.0270
	(0.0483)	(0.0139)	(0.0141)	(0.0130)
10	0.0912	0.13058	0.113	-0.0881
	(0.0479)	(0.0157)	(0.0179)	(0.0181)
15	0.0824	0.0967	0.1022	-0.1777
	(0.0444)	(0.0185)	(0.0184)	(0.0129)
20	0.0891	0.0833	0.0963	-0.2158
	(0.0481)	(0.0196)	(0.0197)	(0.0132)
25	0.0942	0.1015	0.1076	-0.2181
	(0.0510)	(0.0193)	(0.0190)	(0.0129)
30	0.0927	0.0989	0.0968	-0.1747
	(0.0497)	(0.0196)	(0.0212)	(0.0130)
40	0.0993	0.0601	0.1151	-0.1611
	(0.0531)	(0.0202)	(0.0193)	(0.0131)
50	0.0996	0.1188	0.1083	-0.0779
	(0.0538)	(0.0204)	(0.0205)	(0.0134)

C.3 Spillover results

Figure C.4 shows the I-O matrix for the UK (averaged over the last 20 years) with each cell adjusted so that it is a proportion of each industry's intermediate consumption. That is to say, how important each sector is as a supplier to other sectors. The rows of this matrix sum to one, so for example the column for sector 64 (Financial services) shows that it is generally a high proportion of the inputs used by many sectors, so an important supplier.

Figure C.4: Input-Output table, as proportion of intermediate consumption

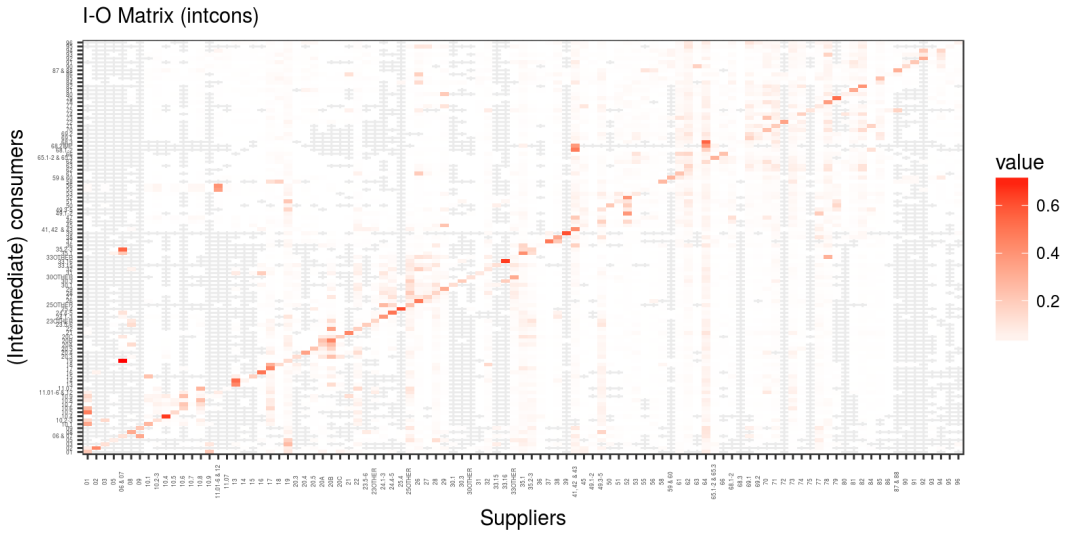


Table C.12 shows results for spillover effects calculated with the a simple indicator for whether any firm in a sector is mentioned, rather than the average mentions across firms in each sector. We can see that the results are qualitatively similar to the average-mention case.

Table C.12: Spillover effects of any media coverage in connected sectors (max, not average)

	Dependent variable:				
	$vol_{i,t}$				
	(1)	(2)	(3)	(4)	(5)
$mention_{i,t}$	0.090***	0.082***	0.091***	0.082***	0.154***
	(0.013)	(0.013)	(0.014)	(0.014)	(0.017)
$downstream\_mentions_{i,t}$		0.056**	0.043*	0.055**	0.057**
		(0.022)	(0.022)	(0.022)	(0.023)
$upstream\_mentions_{i,t}$		0.001	−0.021	−0.034	−0.059*
		(0.026)	(0.027)	(0.027)	(0.032)
$sector\_mentions_{i,t}$			0.021**	0.005	−0.002
			(0.009)	(0.009)	(0.010)
$sector\_vol_{i,t}$				0.094***	0.175***
				(0.002)	(0.002)
$\lvert\Delta p\rvert_{i,t}^{o,c}$	0.723***	0.723***	0.723***	0.717***	
	(0.001)	(0.001)	(0.001)	(0.001)	
$VI_{i,t}^{put}$	1.059***	1.060***	1.060***	0.983***	
	(0.024)	(0.024)	(0.024)	(0.024)	
$VI_{i,t}^{call}$	0.027***	0.027***	0.027***	0.025***	
	(0.008)	(0.008)	(0.008)	(0.008)	
$vol_{i,t}$ lags	✓	✓	✓	✓	✓
Firm fixed effects	✓	✓	✓	✓	✓
Time fixed effects	✓	✓	✓	✓	✓
Observations	312,499	312,499	312,499	311,925	729,795
R^2	0.766	0.766	0.766	0.767	0.502
Adjusted R^2	0.763	0.763	0.763	0.764	0.498
Residual Std. Error	1.194	1.194	1.194	1.191	1.852
Note:					*p<0.1; **p<0.05; ***p<0.01

As a sense check, I randomly generate a placebo matrix by shuffling the rows of the intermediate demand matrix.. We see that media coverage weighted by this matrix has

no significant effect on volatility.

Table C.13: Spillover effects, with placebo matrices

	Dependent variable:	
	$vol_{i,t}$	
	(1)	(2)
$mention_{i,t}$	0.161^{***}	0.087^{***}
	(0.016)	(0.013)
$placebo\_mentions_{i,t}$	-0.093	-0.147
	(0.092)	(0.095)
$sector\_mentions_{i,t}$		0.067
		(0.042)
$sector\_vol_{i,t}$		0.094^{***}
		(0.002)
$\lvert\Delta p\rvert_{i,t}^{o,c}$		0.717^{***}
		(0.001)
$VI_{i,t}^{put}$		0.982^{***}
		(0.024)
$VI_{i,t}^{call}$		0.025^{***}
		(0.008)
$vol_{i,t}$ lags	✓	✓
Firm fixed effects	✓	✓
Time fixed effects	✓	✓
Observations	732,433	311,925
R^2	0.496	0.767
Adjusted R^2	0.493	0.764
Residual Std. Error	1.863	1.191
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table C.14: Spillover effects, with placebo matrices

	Dependent variable:	
	$vol_{i,t}$	
	(1)	(2)
$mention_{i,t}$	0.973^{***}	1.965^{***}
	(0.012)	(0.016)
$\lvert\Delta p\rvert_{i,t}^{o,c}$	0.744^{***}	
	(0.001)	
$VI_{i,t}^{put}$	0.907^{***}	
	(0.021)	
$\sum_{\vartheta=1}^{S} \omega_{\varsigma,\vartheta}^{placebo} mention_{\vartheta,t}$	0.139	-0.085
	(0.095)	(0.130)
$vol_{i,t}$ lags	✓	✓
Firm fixed effects	✓	✓
Time fixed effects	✓	✓
Observations	299,146	299,146
R^2	0.785	0.595
Adjusted R^2	0.782	0.590
Residual Std. Error	1.093	1.499
Note:	*p<0.1; **p<0.05; ***p<0.01	

Appendix D

Appendix to Chapter 4

D.1 Data Preparation

Each CBC document is split into paragraphs, as the entire documents are rather long and will deal with a variety of different topics. Some formulaic preamble and administrative details are also removed. The Bank of England minutes include an Annex summarising the data presented to the Committee by Bank staff prior to the meeting, which are labelled as separate. This Annex was no longer provided after 2004, so in order to maintain comparability across the sample period we exclude it from the analysis. The NYT articles are left intact as they are comparatively short and each article will focus on a small number of topics.

The text data is cleaned to remove any stylistic features and capture only the content. The documents are stripped of any additional white-spaces so that each term is separated by a single space. All numerical characters are removed, as is any punctuation. All characters are transformed to lower case and common stop-words are removed using the list provided by Lewis et al. (2004). The remaining terms are then stemmed using the Porter stemming algorithm, reducing each word to its root, and all terms fewer than three characters in length are removed. We also remove all names of months and seasons as the obvious seasonality of these terms might generate spurious co-movement.[1]

When combining the four copora in different combinations we also remove terms which appear in only one of the corpora, and terms which appear to frequently to be useful in classifying the documents. Figure D.1 shows an example of a paragraph from The Federal Reserve minutes of the meeting on 16th December 1997 (published on 5th February 1998)

[1] As this is done after stemming, the following terms are removed: months "januari", "februari", "march", "april", "may", "june", "july", "juli", "august", "septemb", "octob", "novemb", "decemb", "summer", "autumn", "spring" and "winter".

before and after this cleaning.

Figure D.1: An example paragraph from Federal Reserve minutes for meeting on 16th December 1997

Raw text

Short term interest rates have registered small mixed changes since the day before the Committee meeting on November 12 1997 while bond yields have fallen somewhat. Share prices in U.S. equity markets recorded mixed changes over the period; equity markets in other countries notably in Asia have remained volatile. In foreign exchange markets the value of the dollar has risen over the intermeeting period in terms of both the trade weighted index of the other G 10 countries and the currencies of a number of Asian countries.

Clean text

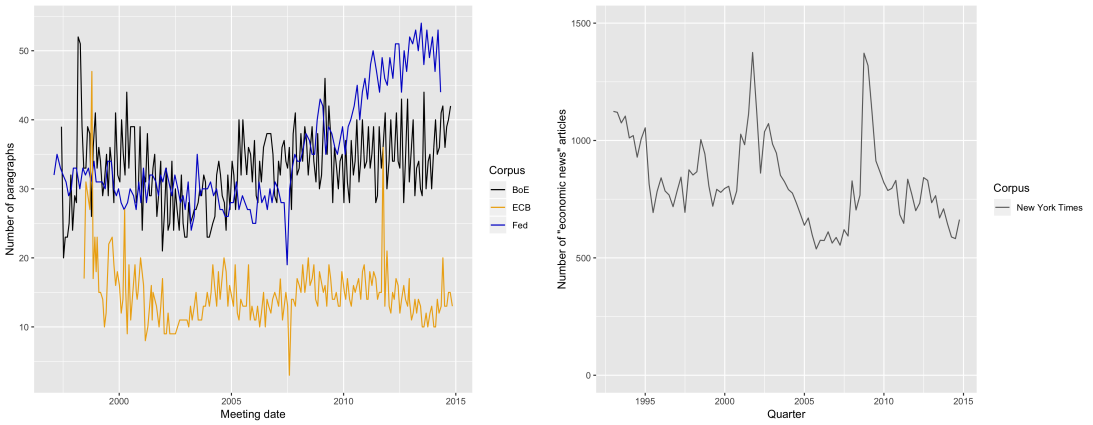
short term interest rate regist small mix chang sinc day committe meet novemb bond yield fallen somewhat share price equiti market record mix chang period equiti market countri notabl asia remain volatil foreign exchang market valu dollar risen intermeet period term trade weight index countri currenc number asian countri

Clean text with corpus-specific and seasonal terms removed

short term interest rate regist small mix chang sinc day committe meet bond yield fallen somewhat share price equiti market record mix chang period equiti market countri notabl asia remain volatil foreign exchang market valu dollar risen period term trade weight index countri currenc number asian countri

The size of the documents is roughly constant over time, and the three central bank corpora are of comparable size. The NYT corpus is considerably larger but the number of articles is also roughly constant over time.

Figure D.2: Number of paragraphs in each central bank corpus over time



D.1.1 FOMC-NYT corpus

In order to allow for meaningful comparison across the two corpora, we remove any words which appear in only one of the two and words which might generate spurious co-movement. Given that the FOMC minutes are already split into paragraphs, combining the two corpora gives 42,927 documents and a total vocabulary of 89,632 unique terms. To facilitate comparison across the corpora, we remove any term which appears only in one of the corpora. Removing this corpus-specific terminology both reduces the dimensionality of the LDA model, making it easier to estimate, and reduces the noisy variation in the documents. The final panel of Figure D.1 shows an example of this data, as we can see the term "novemb" is removed as it is inherently seasonal, as is the term "intermeet" as this appears only in the FOMC text.

The term frequency-inverse document frequency (tf-df) for each term across the combined corpus is calculated as follows

$$\bar{f}_v \times \log_2 \left(\frac{|D|}{|\{d \in D : v \in d\}|} \right) \tag{D.1}$$

where $\bar{f}_v$ is the average number of times term v appears across the D documents and $|\{d \in D : v \in d\}|$ is the number of documents in which term v appears at least once. Thus is a document appears very infrequently, the tf-idf value will be low as the term-frequency will be small. Moreover, if the term appears in many documents, then the inverse document frequency will be small, and so the tf-idf will also be low. Terms with a low tf-idf score therefore will not help to distinguish documents. We remove terms in the bottom 1% of the tf-idf distribution and then remove any documents which are now empty.[2] After this we have 41,873 documents with 8,501,689 words from a vocabulary of 3,032 unique terms.

Estimation over the combined corpus is carried out using the Griffiths and Steyvers (2004) collapsed Gibbs sampler set out in more detail in Appendix D.2.[3] After 10,000 iterations comprising a "burn-in" phase, we run 200,000 iterations with a thinning interval

[2]Most of the (stemmed) terms which are removed because they have a low tf-idf score commonly used words which do not give any indication of the economic topic: "also", "anoth", "big", "come", "econom", "end", "enough", "even", "face", "far", "first", "hard", "help", "hope", "includ", "just", "keep", "last", "less" , "like", "long", "make", "mani", "much", "near", "need", "new", "next", "now", "one", "part", "say", "set", "sinc", "thing", "two", "want", "way", "well", "will", "without", "year", "among", "becom", "call", "high", "made", "put", "recent", "take", "time", "tri", "back", "can", "get", "least", "still", "three", "use", "alreadi", "see", "think", "clear", "move", "yet", "look", "mean", "better", "find", "whether", "sext".

[3]The results are qualitatively similar if the topics are estimated only on the FOMC minutes and then queried over the combined corpus.

of 100. The collapsed Gibbs sampler estimates the latent topic assignment of each word and the θ and β parameters are then backed out from those topic assignments.

The model is originally estimated at an article/paragraph level, as these are more likely to be devoted to a smaller set of topics, giving the model power to estimate β. However, as we are interested in the shifting attention at a time series level, we combine the FOMC paragraphs back into meeting-level documents and the NYT articles into pre-meeting and post-publication documents. We then hold the β vectors estimated at the article/paragraph level constant and estimate the topic proportions at the meeting-level. This process is known as "querying".

D.1.2 FOMC-MPC-GC corpus

Combining the three corpora gives 14,674 paragraphs and a total vocabulary of 5,143 unique terms. In order to reduce the dimensionality of the data and facilitate meaningful comparison across the three corpora, we remove any term which appears only in one of the corpora. This removes terminology specific to a particular central bank which could obscure any co-movement, in particular the names of committee members which comprise a substantial proportion of this corpus-specific vocabulary. As before, we also remove all names of months and quarters as the obvious seasonality of these terms might generate spurious co-movement. After removing these corpus-specific and inherently seasonal terms we are left with 961,232 terms over the 14,674 paragraphs and a vocabulary of 3,032 unique terms. The final panel of Figure D.1 shows an example of this data, as we can see the term "novemb" is removed, as is the term "intermeet" as this appears only in the Federal Reserve text. As described for the FOMC-NYT corpus, term frequency-inverse document frequency (tf-df) for each term across the combined corpus is calculated and terms in the bottom 1% of the tf-idf distribution are removed.[4] This leaves 14,638 paragraphs over 2,979 unique terms.

As before, we estimate a 30 topic model over the combined central bank corpora using the collapsed Gibbs sampler of Griffiths and Steyvers (2004), with a "burn-in" phase of 10,000 iterations followed by 50,000 iterations with a thinning interval of 100. The model is estimated at the paragraph level. Once the β parameters are learned, the paragraphs

[4]The (stemmed) terms which are removed because they have a low tf-idf score are commonly used words which are so widely used that they do not give a firm indication of the economic topic: "growth", "outlook", "time", "activ", "also", "although", "continu", "demand", "economi", "increas", "market", "pace", "price", "reflect"‘ "suggest", "level", "rate", "remain", "year", "declin", "part", "term", "indic", "recent", "condit", "howev", "month", "relat", "rise", "expect", "like", "high", "current", "effect", "support", "signific", "months", "well" "instruments" "objective" "stable".

are aggregated back into meeting-level documents and, holding the β vectors constant, the θ topic proportions are estimated at the meeting-level.

D.2 The Griffiths and Steyvers (2004) collapsed Gibbs sampling algorithm for LDA

The underlying generative model of test for a corpus of D documents over V unique terms and K topics is

1. For each of K topics, draw $\beta_k \sim \text{Dir}(\eta)$
2. For each of D documents, draw $\theta_d \sim \text{Dir}(\alpha)$
3. For each word n in document d:
   Draw topic assignment $z_{d,n}$ from Mult(θ_d)
   Draw $w_{d,n}$ from Mult$\beta_{z_{d,n}}$

Given the generative model of the text assumed by the Latent Dirichlet Allocation model, the likelihood of the observed data (w) and the unobserved data (z) can be expressed in terms of the prior hyperparameters and counts of the observed data:

$$\Pr[W, Z|\alpha, \eta] = \Pr[W|Z, \eta] \Pr[Z|\alpha]$$

In order to estimate z, we thus need to find the probability of the words given the topic assignments and η then the probability of the topic assignments given α.

Both of these are the distribution of some data drawn from a multinomial distribution whose parameters are drawn from a Dirichlet. The distribution of the data given the Dirichlet hyperparameters is derived below.

D.2.1 $\Pr[Z|\alpha]$

The topic assignments z for document d are drawn from a multinomial distribution with parameters θ_d, which is drawn from Dir(α), noting that α is a scalar and Dirichlet is symmetric. Therefore,

$$\Pr[z_d|\alpha] = \frac{\Gamma(K\alpha)}{\Gamma(N_d + K\alpha)} \prod_k \frac{\Gamma(s_{d,k} + \alpha)}{\Gamma(\alpha)}$$

Where $s_{d,k}$ is the number of words in document d assigned to topic k and N_d is the number of words in document, as before. Note that this can also be written as

$$\Pr[z_d|\alpha] = \frac{B(s_{1,d}+\alpha, ..., s_{K,d}+\alpha)}{B(\alpha)}$$

So for the entire corpus, the probability is the product of the probabilities for each individual document.

$$\Pr[Z|\alpha] = \prod_d \frac{\Gamma(K\alpha)}{\Gamma(N_d+K\alpha)} \prod_k \frac{\Gamma(s_{d,k}+\alpha)}{\Gamma(\alpha)}$$

This simplifies to

$$\Pr[Z|\alpha] = \left[\frac{\Gamma(K\alpha)}{\Gamma^K(\alpha)}\right]^D \prod_d \frac{\prod_k \Gamma(s_{d,k}+\alpha)}{\Gamma(N_d+K\alpha)}$$

or alternatively

$$\Pr[Z|\alpha] = \prod_d \frac{B(s_{1,d}+\alpha, ..., s_{K,d}+\alpha)}{B(\alpha)}$$

D.2.2 $\Pr[W|Z,\eta]$

Conditional on the topic assignments Z, we know the words which are assigned to each topic, so we can calculate the likelihood of the words given this and the priors. Let W_k be the words assigned to topic k, and $m_{k,v}$ is the number of times (over the entire corpus) that the token v is associated with topic k. For a given topic k, the problem is directly analogous to that of $Z|\alpha$: we are choosing a multinomial parameter β_k from a $\text{Dir}(\eta)$. The words W_k are then drawn from a multinomial distribution with parameter β_k. Let us therefore

$$\Pr[W_k|Z,\eta] = \frac{\Gamma(V\eta)}{\Gamma(\sum_v m_{k,v}+V\eta)} \prod_v \frac{\Gamma(m_{k,v}+\eta)}{\Gamma(\eta)}$$

The joint probability of all K topics is therefore

$$\Pr[W|Z,\eta] = \prod_k \frac{\Gamma(V\eta)}{\Gamma(\sum_v m_{k,v}+V\eta)} \prod_v \frac{\Gamma(m_{k,v}+\eta)}{\Gamma(\eta)}$$

This can also be expressed as

$$\Pr[W|Z,\eta] = \left[\frac{\Gamma(V\eta)}{\Gamma^V(\eta)}\right]^K \prod_k \frac{\prod_v \Gamma(m_{k,v}+\eta)}{\Gamma(\sum_v m_{k,v}+V\eta)}$$

Where $m_{k,v}$ is the number of time word v is assigned to topic k, and V is the total number of words in the vocabulary.

Or alternatively,

$$\Pr[W|\eta] = \prod_k \frac{B(m_{1,k}+\eta, ..., m_{V,k}+\eta)}{B(\eta)}$$

D.2.3 $\Pr[W, Z|\alpha, \eta]$

Recall that $\Pr[W, Z|\alpha, \eta] = \Pr[W|Z, \eta] \Pr[Z|\alpha]$, so the above sections give us the probability of the data W and the latent topic assignments Z just in terms of the prior hyperparameters, eliminating θ and β. So

$$\Pr[W, Z|\alpha, \eta] = \left[\frac{\Gamma(K\alpha)}{\Gamma^K(\alpha)}\right]^D \prod_d \left(\frac{\prod_k \Gamma(s_{d,k}+\alpha)}{\Gamma(N_d+K\alpha)}\right) \times \left[\frac{\Gamma(V\eta)}{\Gamma^K(\eta)}\right]^K \prod_k \left(\frac{\prod_v \Gamma(m_{k,v}+\eta)}{\Gamma(\sum_v m_{k,v}+V\eta)}\right)$$

Or alternatively

$$\Pr[W, Z|\alpha, \eta] = \prod_d \frac{B(s_{1,d}+\alpha, ..., s_{K,d}+\alpha)}{B(\alpha)} \prod_k \frac{B(m_{1,k}+\eta, ..., m_{V,k}+\eta)}{B(\eta)} \tag{D.2}$$

D.2.4 Factorising $\Pr[z_{d,n} = k|Z_{-(d,n)}, W, \alpha, \eta]$

The Gibbs sampler iterates through each token in turn, drawing a new topic assignment from a multinomial distribution. Therefore, we need to find the probabilities of each topic assignment for a given token, holding all the other assignments fixed (and conditional on the data and hyperparameters). In other words, we need to find $\Pr[z_{d,n} = k|Z_{-(d,n)}, W, \alpha, \eta]$.

We can break this into the joint probability of all the data, given the priors, divided by the probability of the $-(d, n)$ data (application of Bayes' Rule)

$$\Pr[z_{d,n} = k|Z_{-(d,n)}, W, \alpha, \eta] = \frac{\Pr[z_{d,n} = k, Z_{-(d,n)}, W|\alpha, \eta]}{\Pr[Z_{-(d,n)}, W|\alpha, \eta]}$$

We can use the conditional independence property of the LDA model to make variables only depend on their Markov blankets and thereby split the numerator into a W and a Z term:

$$\Pr[z_{d,n} = k, Z_{-(d,n)}, W|\alpha, \eta] = \Pr[W|z_{d,n} = k, Z_{-(d,n)}, \eta] \Pr[z_{d,n} = k, Z_{-(d,n)}|\alpha]$$

The same logic applied to the denominator yields

$$\Pr[Z_{-(d,n)}, W|\alpha, \eta] = \Pr[W|Z_{-(d,n)}, \eta] \Pr[Z_{-(d,n)}|\alpha]$$

Note that this denominator term includes W, not just $W_{-(d,n)}$. However, given that $w_{d,n}$ is generated by $z_{d,n}$, which is independent of $Z_{-(d,n)}$, it follows that

$$\Pr[W|Z_{-(d,n)},\eta] = \Pr[w_{d,n}]\Pr[W_{-(d,n)}|Z_{-(d,n)},\eta] \propto \Pr[W_{-(d,n)}|Z_{-(d,n)},\eta]$$

Therefore, we can abstract from the word choice itself in the denominator and use the result that

$$\Pr[z_{d,n} = k|Z_{-(d,n)},W,\alpha,\eta] \propto \frac{\Pr[W|z_{d,n}=k,Z_{-(d,n)},\eta]\Pr[z_{d,n}=k,Z_{-(d,n)}|\alpha]}{\Pr[W_{-(d,n)}|Z_{-(d,n)},\eta]\Pr[Z_{-(d,n)}|\alpha]}$$

The expression above is now in terms of small transformations of the distributions we found above, $\Pr[W|Z,\eta]$ and $\Pr[Z|\alpha]$.

Making use of the conditional independence property will make it convenient to think of this distribution as having two parts, one relating to the topic assignments and the other to the word choices.

$$\Pr[z_{d,n} = k|Z_{-(d,n)},W,\alpha,\eta] \propto \frac{\Pr[W|z_{d,n}=k,Z_{-(d,n)},\eta]}{\Pr[W_{-(d,n)}|Z_{-(d,n)},\eta]} \times \frac{\Pr[z_{d,n}=k,Z_{-(d,n)}|\alpha]}{\Pr[Z_{-(d,n)}|\alpha]}$$

We will find that quite a few terms cancelling out in each of these two fractions, so construct them separately now

Identifying $\Pr[\mathbf{z_{d,n}} = \mathbf{k}, \mathbf{Z_{-(d,n)}}|\alpha]$

As shown above, because each topic assignment is independent of the others, $\Pr[Z|\alpha]$ is simply the product of each $\Pr[Z_d|\alpha]$

$$\Pr[Z|\alpha] = \Pr[Z_{-d}|\alpha]\Pr[Z_d|\alpha]$$

Therefore, we can begin to rearrange the expression to isolate the terms for document d

$$\Pr[Z_{-d}|\alpha] = \prod_{\delta\neq d}\frac{\Gamma(K\alpha)}{\Gamma(N_\delta + K\alpha)}\prod_k \frac{\Gamma(s_{\delta,k}+\alpha)}{\Gamma(\alpha)}$$

Isolating the nth token in document d requires a slightly subtler step. As we are aiming to find the probability that $z_{dn} = k$ for some given k, we know the topic assignment of that token. First, notice that

$$\Pr[Z_d|\alpha] = \frac{\Gamma(K\alpha)}{\Gamma(N_d+K\alpha)}\prod_k\frac{\Gamma(s_{d,k}+\alpha)}{\Gamma(\alpha)} = \frac{\Gamma(K\alpha)}{\Gamma(N_d+K\alpha)}\left[\prod_{\kappa\neq k}\frac{\Gamma(s_{d,\kappa}+\alpha)}{\Gamma(\alpha)}\right]\frac{\Gamma(s_{d,k}+\alpha)}{\Gamma(\alpha)}$$

Define $s_{d,k,-n}$ as the number of token in document d assigned to topic k, excluding the nth token. As we are calculating the probability of token n being assigned to a particular k we know that

$$s_{d,k} = s_{d,k,-n} + 1$$

Putting these parts together gives us an expression for $\Pr[z_{d,n} = k, Z_{-(d,n)}|\alpha]$ as follows

$$\Pr[z_{d,n} = k, Z_{-(d,n)}|\alpha] =$$
$$\left[\prod_{\delta \neq d} \frac{\Gamma(K\alpha)}{\Gamma(N_\delta + K\alpha)} \prod_k \frac{\Gamma(s_{\delta,k} + \alpha)}{\Gamma(\alpha)}\right] \frac{\Gamma(K\alpha)}{\Gamma(N_d + K\alpha)} \left[\prod_{\kappa \neq k} \frac{\Gamma(s_{d,\kappa} + \alpha)}{\Gamma(\alpha)}\right] \frac{\Gamma(s_{d,k,-n} + 1 + \alpha)}{\Gamma(\alpha)}$$

Identifying $\Pr[\mathbf{Z}_{-(\mathbf{d},\mathbf{n})}|\alpha]$

Note that the only differences between $\Pr[z_{d,n} = k, Z_{-(d,n)}|\alpha]$ and $\Pr[Z_{-(d,n)}|\alpha]$ will be that there is now one fewer word in document d, so N_d becomes $N_d - 1$, and there will be one fewer word assigned to topic k in document d, so n_d, k becomes $n_{d,k,-n}$. Therefore

$$\Pr[Z_{-(d,n)}|\alpha] =$$
$$\left[\prod_{\delta \neq d} \frac{\Gamma(K\alpha)}{\Gamma(N_\delta + K\alpha)} \prod_k \frac{\Gamma(s_{\delta,k} + \alpha)}{\Gamma(\alpha)}\right] \frac{\Gamma(K\alpha)}{\Gamma(N_d - 1 + K\alpha)} \left[\prod_{\kappa \neq k} \frac{\Gamma(s_{d,\kappa} + \alpha)}{\Gamma(\alpha)}\right] \frac{\Gamma(s_{d,k,-n} + \alpha)}{\Gamma(\alpha)}$$

Identifying $\frac{\Pr[\mathbf{z}_{\mathbf{d},\mathbf{n}}=\mathbf{k},\mathbf{Z}_{-(\mathbf{d},\mathbf{n})}|\alpha]}{\Pr[\mathbf{Z}_{-(\mathbf{d},\mathbf{n})}|\alpha]}$

Combining the $\Pr[z_{d,n} = k, Z_{-(d,n)}|\alpha]$ and $\Pr[Z_{-(d,n)}|\alpha]$ expressions, note that the following common factor will cancel

$$\left[\prod_{\delta \neq d} \frac{\Gamma(K\alpha)}{\Gamma(N_\delta + K\alpha)} \prod_k \frac{\Gamma(s_{\delta,k} + \alpha)}{\Gamma(\alpha)}\right] \left[\prod_{\kappa \neq k} \frac{\Gamma(s_{d,\kappa} + \alpha)}{\Gamma(\alpha)}\right] \frac{\Gamma(K\alpha)}{\Gamma(\alpha)}$$

So

$$\frac{\Pr[z_{d,n} = k, Z_{-(d,n)}|\alpha]}{\Pr[Z_{-(d,n)}|\alpha]} = \frac{\Gamma(s_{d,k,-n} + 1 + \alpha)}{\Gamma(N_d + K\alpha)} \frac{\Gamma(N_d - 1 + K\alpha)}{\Gamma(s_{d,k,-n} + \alpha)}$$
$$= \frac{\Gamma(N_d - 1 + K\alpha)}{\Gamma(N_d + K\alpha)} \frac{\Gamma(s_{d,k,-n} + 1 + \alpha)}{\Gamma(s_{d,k,-n} + \alpha)}$$

We can now crucially use the identity that

$$\frac{\Gamma(x+1)}{\Gamma(x)} \equiv x$$

To note that

$$\frac{\Gamma(N_d - 1 + K\alpha)}{\Gamma(N_d + K\alpha)} = \frac{1}{N_d - 1 + K\alpha}$$

and

$$\frac{\Gamma(s_{d,k,-n}+1+\alpha)}{\Gamma(s_{d,k,-n}+\alpha)} = s_{d,k,-n}+\alpha$$

So therefore

$$\frac{\Pr[z_{d,n}=k, Z_{-(d,n)}|\alpha]}{\Pr[Z_{-(d,n)}|\alpha]} = \frac{s_{d,k,-n}+\alpha}{N_d - 1 + K\alpha}$$

Identifying $\Pr[\mathbf{W}|\mathbf{z_{d,n}} = \mathbf{k}, \mathbf{Z_{-(d,n)}}, \eta]$

For a given (observed) token assignment v for token n in document d, we can work out this likelihood and express it in a way that separates out the contribution of $z_{d,n}$. Recall that

$$\Pr[W|Z,\eta] = \prod_k \frac{\Gamma(V\eta)}{\Gamma(\sum_v m_{k,v} + V\eta)} \prod_v \frac{\Gamma(m_{k,v}+\eta)}{\Gamma(\eta)}$$

As discussed above, this is the product of the probability of each of the K topics, so can be written as

$$\Pr[W|Z,\eta] = \left[\prod_{\kappa\neq k} \frac{\Gamma(V\eta)}{\Gamma(\sum_v m_{\kappa,v} + V\eta)} \prod_v \frac{\Gamma(m_{\kappa,v}+\eta)}{\Gamma(\eta)}\right] \frac{\Gamma(V\eta)}{\Gamma(\sum_v m_{k,v} + V\eta)} \prod_v \frac{\Gamma(m_{k,v}+\eta)}{\Gamma(\eta)}$$

The $\sum_v m_{k,v}$ term is just the total number of words in a given topic, so doesn't vary across v (it's the analogue of N_d), therefore we can treat this as a constant (given Z). We also know, as it is observed, that term (d,n) is v, and so $m_{k,v} = m_{k,v,-(d,n)} + 1$, so we can express the above as

$$\Pr[W|Z,\eta] = \left[\prod_{\kappa\neq k} \frac{\Gamma(V\eta)}{\Gamma(\sum_v m_{\kappa,v} + V\eta)} \prod_v \frac{\Gamma(m_{\kappa,v}+\eta)}{\Gamma(\eta)}\right] \times$$
$$\left[\prod_{\nu\neq v} \frac{\Gamma(m_{k,\nu}+\eta)}{\Gamma(\eta)}\right] \frac{\Gamma(V\eta)}{\Gamma(\sum_v m_{k,v} + V\eta)} \frac{\Gamma(m_{k,v,-(d,n)}+1+\eta)}{\Gamma(\eta)}$$

Identifying $\Pr[\mathbf{W_{-(d,n)}}|\mathbf{Z_{-(d,n)}}, \eta]$

As in the $\Pr[Z_{-(d,n)}|\alpha]$ case, we remove the (d,n) term, which again relies on the fact that each $z_{d,n}$ is drawn independently, conditional on α.

$$\Pr[W_{-(d,n)}|Z_{-(d,n)},\eta] = \left[\prod_{\kappa\neq k} \frac{\Gamma(V\eta)}{\Gamma(\sum_v m_{\kappa,v} + V\eta)} \prod_v \frac{\Gamma(m_{\kappa,v}+\eta)}{\Gamma(\eta)}\right] \times$$
$$\left[\prod_{\nu\neq v} \frac{\Gamma(m_{k,\nu}+\eta)}{\Gamma(\eta)}\right] \frac{\Gamma(V\eta)}{\Gamma(\sum_v m_{k,v} - 1 + V\eta)} \frac{\Gamma(m_{k,v,-(d,n)}+\eta)}{\Gamma(\eta)}$$

Identifying $\frac{\Pr[\mathbf{W}|\mathbf{z_{d,n}}=\mathbf{k},\mathbf{Z}_{-(\mathbf{d},\mathbf{n})},\eta]}{\Pr[\mathbf{W}_{-(\mathbf{d},\mathbf{n})}|\mathbf{Z}_{-(\mathbf{d},\mathbf{n})},\eta]}$

Again, we have a common factor that happily cancels out:

$$\left[\prod_{\kappa\neq k}\frac{\Gamma(V\eta)}{\Gamma(\sum_v m_{\kappa,v}+V\eta)}\prod_v\frac{\Gamma(m_{\kappa,v}+\eta)}{\Gamma(\eta)}\right]\left[\prod_{\nu\neq v}\frac{\Gamma(m_{k,\nu}+\eta)}{\Gamma(\eta)}\right]\frac{\Gamma(V\eta}{\Gamma(\eta)}$$

This give the expression

$$\frac{\Pr[W|z_{d,n}=k,Z_{-(d,n)},\eta]}{\Pr[W_{-(d,n)}|Z_{-(d,n)},\eta]}=\frac{\Gamma(m_{k,v,-(d,n)}+1+\eta)}{\Gamma(m_{k,v,-(d,n)}+\eta)}\frac{\Gamma(\sum_v m_{k,v}-1+V\eta)}{\Gamma(\sum_v m_{k,v}+V\eta)}$$

The identity mentioned above, $\frac{\Gamma(x+1)}{\Gamma(x)}\equiv x$, can thus again be used here

$$\frac{\Pr[W|z_{d,n}=k,Z_{-(d,n)},\eta]}{\Pr[W_{-(d,n)}|Z_{-(d,n)},\eta]}=\frac{m_{k,v,-(d,n)}+\eta}{\sum_v m_{k,v}+V\eta}$$

D.2.5 Gibbs sampling distribution

As discussed above, the Gibbs sampling algorithm requires multinomial sampling from a distribution defined by $\Pr[z_{d,n}=k|Z_{-(d,n)},W,\alpha,\eta]$ which can be broken down into a probability of topic assignment and token assignment parts

$$\Pr[z_{d,n}=k|Z_{-(d,n)},W,\alpha,\eta]\propto\frac{\Pr[W|z_{d,n}=k,Z_{-(d,n)},\eta]}{\Pr[W_{-(d,n)}|Z_{-(d,n)},\eta]}\times\frac{\Pr[z_{d,n}=k,Z_{-(d,n)}|\alpha]}{\Pr[Z_{-(d,n)}|\alpha]}$$

The results derived above give

$$\Pr[z_{d,n}=k|Z_{-(d,n)},W,\alpha,\eta]\propto\frac{s_{d,k,-n}+\alpha}{N_d-1+K\alpha}\times\frac{m_{k,v,-(d,n)}+\eta}{\sum_v m_{k,v}+V\eta}$$

The $N_d-1+K\alpha$ term does not vary with d, n or k so this isn't important for the sampling distribution.

$$\Pr[z_{d,n}=k|Z_{-(d,n)},W,\alpha,\eta]\propto(s_{d,k,-n}+\alpha)\frac{m_{k,v,-(d,n)}+\eta}{\sum_v m_{k,v,-(d,n)}+V\eta}$$

The hyperparameter η is chosen beforehand, V is observed in the data and the m and n terms are either being chosen or based on a previous iteration.

The Gibbs sampling algorithm itself starts with a randomly allocated topic assignment, then for each token sequentially draws a topic assignment from the multinomial distribution defined by the above expression. Usually, a burn in phase is used to allow the process to converge and reduce dependence on the starting values, and then a thinning

interval to make the draws approximately *iid*.

As the θ and β parameters are not explicitly sampled from in the collapsed Gibbs algorithm, these are backed out from the predictive distributions. The θ parameter is simply derived from the proportion of words which are assigned to each topic in the given documents.

$$\hat{\theta}_{d,k} = \frac{s_{d,k} + \alpha}{\sum_k (s_{d,k} + \alpha)} \tag{D.3}$$

The β parameter is derived from the proportion of each token in the vocabulary assigned to each topic.

$$\hat{\beta}_{k,v} = \frac{m_{k,v} + \eta}{\sum_v (m_{k,v} + \eta)} \tag{D.4}$$

D.3 Model

D.3.1 Deriving CB loss for one bank case

As shown in Eq. 4.8, the central bank's problem is the constrained minimisation of their loss function.

$$L(\mathbf{a}_t, \mathbf{s}_t) = \mathbb{E}\left[\sum_i \left(\lambda_{i,t}\hat{s}_{i,t} - \epsilon_{i,t}\right)^2 | s_{i,t}\right] \tag{D.5}$$

Substituting in for $\hat{s}_{i,t} = s_{i,t} + (1 - a_{i,t})\eta_{i,t}$ and expanding the quadratic inside the expectation gives us

$$\begin{aligned} L(\mathbf{a}_t, \mathbf{s}_t) &= \mathbb{E}\left[\sum_i \left((\lambda_{i,t}s_{i,t} + \lambda_{i,t}(1-a_{i,t})\eta_{i,t})^2 - 2\epsilon_{i,t}(\lambda_{i,t}s_{i,t} + \lambda_{i,t}(1-a_{i,t})\eta_{i,t}) + \epsilon_{i,t}^2\right) | s_{i,t}\right] \\ &= \mathbb{E}\Bigg[\sum_i \Big(\lambda_{i,t}^2 s_{i,t}^2 + 2\lambda_{i,t}^2 s_{i,t}(1-a_{i,t})\eta_{i,t} + \lambda_{i,t}^2(1-a_{i,t})^2\eta_{i,t}^2 - \\ &\qquad 2\epsilon_{i,t}\lambda_{i,t}s_{i,t} - 2\epsilon_{i,t}\lambda_{i,t}(1-a_{i,t})\eta_{i,t} + \epsilon_{i,t}^2\Big) | s_{i,t}\Bigg] \end{aligned} \tag{D.6}$$

The linearity of the expectations operator then allows us to separate out these terms.

$$\begin{aligned} L(\mathbf{a}_t, \mathbf{s}_t) = \sum_i \Big(&\mathbb{E}\big[\lambda_{i,t}^2 s_{i,t}^2 | s_{i,t}\big] + 2\mathbb{E}\big[\lambda_{i,t}^2 s_{i,t}(1-a_{i,t})\eta_{i,t} | s_{i,t}\big] + \\ &\mathbb{E}\big[\lambda_{i,t}^2(1-a_{i,t})^2\eta_{i,t}^2 | s_{i,t}\big] - 2\mathbb{E}\big[\epsilon_{i,t}\lambda_{i,t}s_{i,t} | s_{i,t}\big] - 2\mathbb{E}\big[\epsilon_{i,t}\lambda_{i,t}(1-a_{i,t})\eta_{i,t} | s_{i,t}\big] + \mathbb{E}\big[\epsilon_{i,t}^2 | s_{i,t}\big]\Big) \end{aligned} \tag{D.7}$$

Using the distributions of the shocks, we can then write this in terms of known parameters and variables.

$$\begin{aligned} L(\mathbf{a}_t, \mathbf{s}_t) = \sum_i \Bigg(&\lambda_{i,t}^2 s_{i,t}^2 + \lambda_{i,t}^2(1-a_{i,t})^2\sigma_{\eta,i}^2 - 2\lambda_{i,t}\frac{\sigma_{\epsilon,i}^2}{\sigma_{\epsilon,i}^2 + \sigma_{\nu,i}^2}s_{i,t}^2 + \\ &\left(\frac{\sigma_{\epsilon,i}^2}{\sigma_{\epsilon,i}^2 + \sigma_{\nu,i}^2}\right)^2 s_{i,t}^2 + \sigma_{\epsilon,i} - \frac{(\sigma_{\epsilon,i}^2)^2}{\sigma_{\epsilon,i}^2 + \sigma_{\nu,i}^2}\Bigg) \end{aligned} \tag{D.8}$$

and this can be conveniently be expressed as a three term loss function.

$$L(\mathbf{a}_t, \mathbf{s}_t) = \sum_i \left(\left(\lambda_{i,t} - \frac{\sigma^2_{\epsilon,i}}{\sigma^2_{\epsilon,i} + \sigma^2_{\nu,i}} \right)^2 s^2_{i,t} + \lambda^2_{i,t}(1 - a_{i,t})^2 \sigma^2_{\nu,i} + \sigma^2_{\epsilon,i} \left(1 - \frac{\sigma^2_{\epsilon,i}}{\sigma^2_{\epsilon,i} + \sigma^2_{\nu,i}} \right) \right) \quad \text{(D.9)}$$

D.3.2 Deriving CB loss for two bank case

In the two central bank case, central bank b observes the public signal of central bank c, in addition to its own private signal. It will therefore condition its expectation on both $s_{b,i,t}$ and $\hat{s}_{c,i,t}$. To define the conditional distribution of $\epsilon_{b,i,t}$ given $s_{b,i,t}$ and $\hat{s}_{c,i,t}$, define the following multivariate normal distribution.[5]

$$\begin{pmatrix} \epsilon_{b,i,t} \\ s_{b,i,t} \\ \hat{s}_{c,i,t} \end{pmatrix} \sim \mathcal{N}\left(\mu_{b,i,t'}, \mathbf{\Sigma}_{b,i,t'} \right)$$

where (D.10)

$$\mu_{b,i,t'} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \quad \text{and} \quad \mathbf{\Sigma}_{b,i,t'} = \begin{pmatrix} \sigma^2_{b,\epsilon,i} & \sigma^2_{b,\epsilon,i} & \sigma_{bc,i} \\ \sigma^2_{b,\epsilon,i} & \sigma^2_{b,\epsilon,i} + \sigma^2_{b,\nu,i} & \sigma_{bc,i} \\ \sigma_{bc,i} & \sigma_{bc,i} & \sigma^2_{c,\epsilon,i} + \sigma^2_{c,\nu,i} + (1 - a_{c,i,t})^2 \sigma^2_{c,\eta,i} \end{pmatrix}$$

Partitioning this distribution so that $\mathbf{y}_{b,i,t'} = \epsilon_{b,i,t'}$ and $\mathbf{x}_{b,i,t'} = (s_{b,i,t}, \quad \hat{s}_{c,i,t})$ we have

$$\begin{pmatrix} \mathbf{y}_{b,i,t'} \\ \mathbf{x}_{b,i,t'} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu}_1(b,i,t') \\ \boldsymbol{\mu}_2(b,i,t') \end{pmatrix}, \begin{pmatrix} \mathbf{\Sigma}_{11}(b,i,t') & \mathbf{\Sigma}_{12}(b,i,t') \\ \mathbf{\Sigma}_{21}(b,i,t') & \mathbf{\Sigma}_{22}(b,i,t') \end{pmatrix} \right)$$

where

$$\boldsymbol{\mu}_1(b,i,t') = 0, \quad \boldsymbol{\mu}_2(b,i,t') = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \text{(D.11)}$$

$$\mathbf{\Sigma}_{11}(b,i,t') = \sigma^2_{b,\epsilon,i}, \quad \mathbf{\Sigma}_{12}(b,i,t') = \begin{pmatrix} \sigma^2_{b,\epsilon,i} & \sigma_{bc,i} \end{pmatrix}, \quad \mathbf{\Sigma}_{21}(b,i,t') = \begin{pmatrix} \sigma^2_{b,\epsilon,i} \\ \sigma_{bc,i} \end{pmatrix},$$

$$\mathbf{\Sigma}_{22}(b,i,t') = \begin{pmatrix} \sigma^2_{b,\epsilon,i} + \sigma^2_{b,\nu,i} & \sigma_{bc,i} \\ \sigma_{bc,i} & \sigma^2_{c,\epsilon,i} + \sigma^2_{c,\nu,i} + (1 - a_{c,i,t})^2 \sigma^2_{c,\eta,i} \end{pmatrix}$$

We can then use this to calculate the conditional distribution of $\epsilon_{b,i,t}$

$$\epsilon_{b,i,t} | s_{b,i,t}, \hat{s}_{c,i,t} \sim \mathcal{N}\left(\tilde{\mu}_{b,i,t'}, \tilde{\Sigma}_{b,i,t'} \right)$$

where

$$\tilde{\mu}_{b,i,t'} = \mathbf{\Sigma}_{12}(b,i,t') \mathbf{\Sigma}_{22}(b,i,t')^{-1} \begin{pmatrix} s_{b,i,t} \\ \hat{s}_{c,i,t} \end{pmatrix} \quad \text{(D.12)}$$

$$\tilde{\Sigma}_{b,i,t'} = \mathbf{\Sigma}_{11}(b,i,t') - \mathbf{\Sigma}_{12}(b,i,t') \mathbf{\Sigma}^{-1}_{22}(b,i,t') \mathbf{\Sigma}_{21}(b,i,t')$$

[5] As there is no information on $\hat{s}_{b,i,t}$ in $\hat{s}_{c,i,t}$ that is not contained in $s_{b,i,t}$, we do not need to compute the conditional distribution of $(\hat{s}_{b,i,t} | s_{b,i,t}, \hat{s}_{c,i,t})$, and can simply substitute in for $\hat{s}_{c,i,t} = s_{b,i,t} + (1 - a_{b,i,t}\eta_{b,i,t})$.

As the central bank now observes two distinct signals for each structural shock, their public signal will also be a combination of the two, with some added noise that can be mitigated by increasing focus on a particular state variable.

$$\hat{s}_{b,i,t'} = \Sigma_{12}(b,i,t')\Sigma_{22}(b,i,t')^{-1}\begin{pmatrix} s_{b,i,t} + (1-a_{b,i,t'})\eta_{b,i,t} \\ \hat{s}_{c,i,t} + (1-a_{b,i,t'})\eta_{b,i,t} \end{pmatrix} \tag{D.13}$$

We further assume that the private sector in economy b does not observe the public signal of central bank c, so they condition on $\hat{s}_{b,i,t'}$ only. The central bank's loss function is therefore

$$\begin{aligned} L(\mathbf{a}_{b,t'}, \mathbf{s}_{b,t}) &= \mathbb{E}\left[\sum_i \left(\mathbb{E}[\epsilon_{b,i,t}|\hat{s}_{b,i,t'}] - \epsilon_{b,i,t}\right)^2 |s_{b,i,t}, \hat{s}_{c,i,t}\right] = \mathbb{E}\left[\sum_i \left(\lambda_{b,i,t'}\hat{s}_{b,i,t'} - \epsilon_{b,i,t}\right)^2 |s_{b,i,t}, \hat{s}_{c,i,t}\right] \\ &\text{where} \\ \lambda_{b,i,t'} &= \mathbb{E}[\hat{s}_{b,i,t'}\epsilon_{b,i,t}]\mathbb{E}[\hat{s}^2_{b,i,t'}]^{-1} = \Sigma_{12}(b,i,t')\Sigma_{22}(b,i,t')^{-1}\begin{pmatrix}\sigma^2_{b,\epsilon,i} \\ \sigma_{bc,i}\end{pmatrix}\Bigg(\Sigma_{12}(b,i,t')\Sigma_{22}(b,i,t')^{-1} \\ &\begin{pmatrix}\sigma^2_{b,\epsilon,i} + \sigma^2_{b,\nu,i} + (1-a_{b,i,t'})^2\sigma^2_{b,\eta,i} & \sigma_{bc,i} + (1-a_{b,i,t'})^2\sigma^2_{b,\eta,i} \\ \sigma_{bc,i} + (1-a_{b,i,t'})^2\sigma^2_{b,\eta,i} & \sigma^2_{c,\epsilon,i} + \sigma^2_{c,\nu,i} + (1-a_{c,i,t})^2\sigma^2_{c,\eta,i} + (1-a_{b,i,t'})^2\sigma^2_{b,\eta,i}\end{pmatrix} \\ &\left(\Sigma_{12}(b,i,t')\Sigma_{22}(b,i,t')^{-1}\right)^T\Bigg)^{-1} \end{aligned} \tag{D.14}$$

As in the one bank case, we can then write the loss function of central bank b in terms of known parameters and observed variables. This gives us the necessary conditional moments for the loss function.

$$\begin{aligned} L(\mathbf{a}_{b,t'}, \mathbf{s}_{b,t}) &= \sum_i \left(\lambda^2_{b,i,t'}\mathbb{E}\left[\hat{s}^2_{b,i,t'}|s_{b,i,t}, \hat{s}_{c,i,t}\right] - 2\lambda_{b,i,t'}\mathbb{E}\left[\epsilon_{b,i,t}\hat{s}_{b,i,t'}|s_{b,i,t}, \hat{s}_{c,i,t}\right] + \mathbb{E}\left[\epsilon^2_{b,i,t}|s_{b,i,t}, \hat{s}_{c,i,t}\right]\right) \\ &= \lambda^2_{b,i,t'}\Bigg(\Sigma_{12}(b,i,t')\Sigma_{22}(b,i,t')^{-1}\begin{pmatrix} s^2_{b,i,t} + (1-a_{b,i,t'})^2\sigma^2_{b,\eta,i} & s_{b,i,t}\hat{s}_{c,i,t} + (1-a_{b,i,t'})^2\sigma^2_{b,\eta,i} \\ s_{b,i,t}\hat{s}_{c,i,t} + (1-a_{b,i,t'})^2\sigma^2_{b,\eta,i} & \hat{s}^2_{c,i,t} + (1-a_{b,i,t'})^2\sigma^2_{b,\eta,i}\end{pmatrix} \\ &\left(\Sigma_{12}(b,i,t')\Sigma_{22}(b,i,t')^{-1}\right)^T\Bigg) - 2\lambda_{b,i,t'}\tilde{\mu}_{b,i,t'}\left(\Sigma_{12}(b,i,t')\Sigma_{22}(b,i,t')^{-1}\begin{pmatrix}s_{b,i,t} \\ \hat{s}_{c,i,t}\end{pmatrix}\right) + \tilde{\mu}^2_{b,i,t'} + \tilde{\Sigma}_{b,i,t'} \end{aligned} \tag{D.15}$$

Central bank b's problem at time t' is therefore to chose the focus of their communication $(\mathbf{a}_{b,t'})$ in order to minimise this loss.

D.3.3 Solution algorithm

The central bank's problem is highly non-linear, with the focus of their communication affecting not only the accuracy of their public signal, but also the weight that the private sector puts on that signal. It is therefore not possibly to solve for their optimal communication mix in closed form. However, for given parameter values and instantiations of the shocks, the central bank's problem is straightforward to solve numerically.

For given values of the variance parameters, we can therefore draw a large number of

values for each of the shocks and solve the central banks' problems numerically to find $\mathbf{a}_{b,t'}$ and $\mathbf{a}_{c,t}$. We do this using the Ipopt optimiser Biegler and Zavala (2009) in the programming language *Julia* Bezanson et al. (2017). We assume that central bank c moves first and so solves the one-bank problem. Central bank b then observes $\mathbf{a}_{c,t}$ and solves the two-bank problem to determine $\mathbf{a}_{b,t'}$. For each value of the variance parameters, we draw 50,000 values of the shocks to approximate the appropriate distributions. These empirical distributions can then be used to identify the moments of central bank focus, and how these vary with the value of their signals.

D.4 Central Bank and Private Sector Information Asymmetry

In addition to the forecast dispersion measures we also use "nowcasts" of GDP growth provided by the SPF, as a measure of the private sector's real time assessment of the state of the economy. The survey's timing is lined up with the release of the Bureau of Economic Analysis' advance report of the national income and product accounts, which contains the first estimate of GDP (and components) for the previous quarter. The survey's questionnaires report recent historical values of the data from the BEA's advance report, as well any other recent reports from government statistical agencies.

Table D.1: Survey of Professional Forecasters Timings

Quarter	Survey sent out	Last Q in Information Sets	Submission Deadline
Q1	End of January	Q4	Mid February
Q2	End of April	Q1	Mid May
Q3	End of July	Q2	Mid August
Q4	End of October	Q3	Mid November

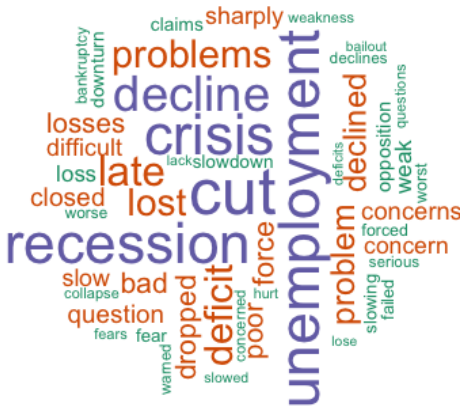
Median growth rates are those for quarter-over-quarter growth, expressed in annualized percentage points, beginning with the forecast for the current quarter. These median forecast growth rates are constructed by first computing the median level of the variable and then compute the rate of growth from these median levels. The "nowcast" thus reflects the private sector's median expectation of this period's GDP growth, based on all the information available halfway through the quarter.

In order for central bank communication to play a non-monetary information role, i.e.

provide information that influences the economy other than through current and future short term interest rate expectations, it must either have some private knowledge, or be able to present public knowledge in a way that is useful to agents in the economy. To show that this is plausibly the case, we here present evidence that the minutes of the FOMC are more informative about contemporaneous GDP growth than the NYT articles. We show that both the sentiment and topic proportions of the FOMC minutes have greater predictive content for contemporaneous US GDP growth that the NYT articles, controlling for contemporaneous forecasts from the SPF.

A sentiment score for each article/paragraph is constructed using the economic-specific sentiment dictionaries provided by Loughran and McDonald (2011). These widely used dictionaries provide lists of words associated with different sentiments. Figure D.4 represents the "positive" and "negative" word lists with word clouds, where the prominence of each word is determined by how many times it appears in the NYT articles and FOMC minutes

Figure D.4: Commonly used sentiment scored words in the NYT and FOMC corpora **(a)** "Negative" words, weighted by occurrence **(b)** "Positive" words, weighted by occurrence



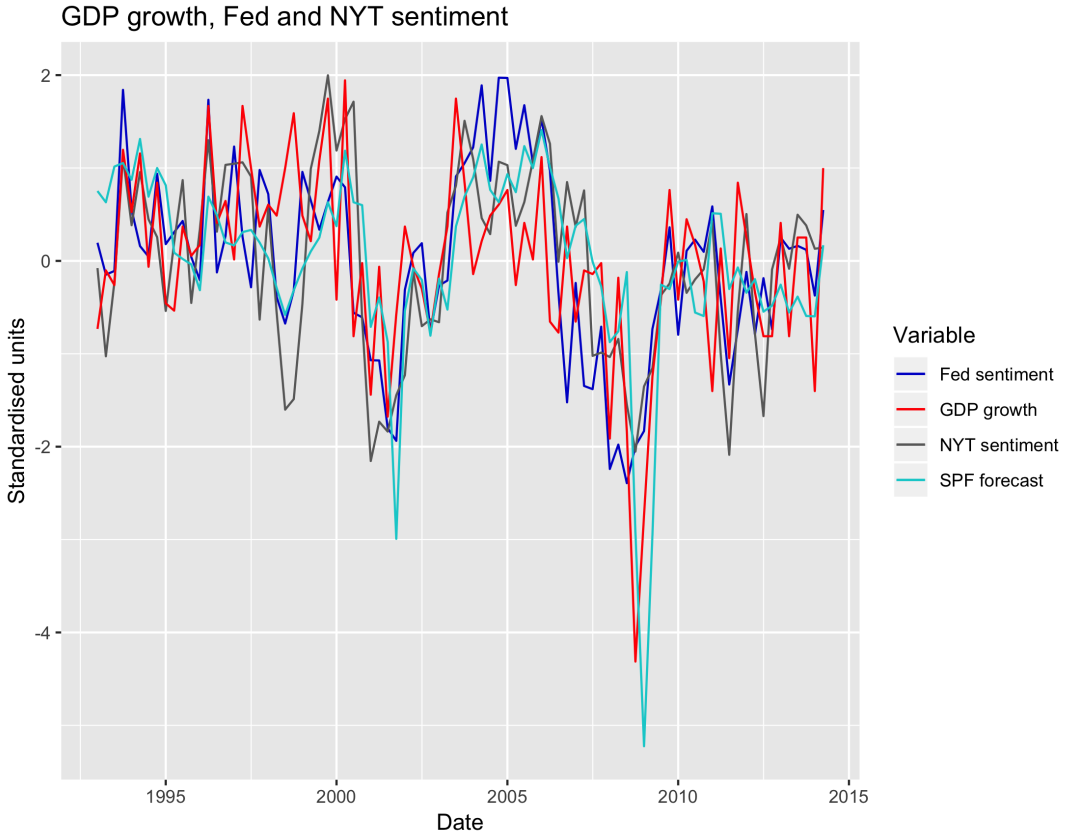
These word lists are used to count the number of "positive" and "negative" terms in each article/paragraph, and the sentiment score of each article/paragraph is calculated as the percentage of scored words which are positive

$$sentiment_d = \frac{positive_d}{positive_d + negative_d} \tag{D.16}$$

where $positive_d$ is the number of words in document d that appear in the positive word list, and $negative_d$ is the number of words that appear in the negative word list. These sentiment scores are then standardised to have mean zero and unit variance, and aggregated to the quarterly level to provide a quarterly sentiment measure for both the media and FOMC corpora. Similarly, the topic proportions for each corpus, calculated as described in the previous subsection, are aggregated to quarterly level in order to match the frequency of the GDP growth data.

The quarterly sentiment score of both the FOMC minutes and the New York Times articles tracks the US business cycle remarkably closely, as shown in Figure D.6. Bearing in mind that these articles and minutes are published well before the GDP data is released, this provides some initial evidence that economic agents can gain some useful information from the two sources of text data. We also include the median value of the SPF forecasts of GDP growth in the same quarter in which the survey is conducted, which we will refer to as a "nowcast". The survey responses have to be submitted in the middle of the quarter, as described in Section 4.3, so these responses give a measure of the private sector's real-time assessment of the state of the economy.

Figure D.6: Sentiment, private sector forecasts and US GDP growth



The SPF median "nowcast" of GDP growth thus forms a natural benchmark against which to compare the predictive content of the text. Table D.2 shows that the NYT articles and the FOMC minutes contain information that could be of use to private sector agents, and that the FOMC minutes sentiment is a stronger predictor. Not only is the magnitude of the coefficient on the FOMC minutes sentiment compared to the NYT articles sentiment (both are standardised, so the coefficients can be compared) and the contribution to R^2 greater, but when both are included as predictors only the FOMC minutes sentiment remains significant.

Table D.2: Fed minutes sentiment have greater predictive content for GDP growth NYT article sentiment

	Dependent variable:			
	GDP growth			
	(1)	(2)	(3)	(4)
$sentiment^{Fed}$		0.924***		0.690**
		(0.277)		(0.315)
$sentiment^{NYT}$			0.789***	0.462
			(0.272)	(0.305)
SPF GDP growth nowcast	0.894***	0.571***	0.654***	0.512***
	(0.175)	(0.192)	(0.187)	(0.194)
Lagged GDP growth	0.003	−0.028	−0.041	−0.046
	(0.115)	(0.109)	(0.111)	(0.108)
Constant	−1.394**	0.120	−0.217	0.425
	(0.680)	(0.785)	(0.767)	(0.805)
Observations	85	85	85	85
R^2	0.342	0.421	0.404	0.437
Adjusted R^2	0.326	0.400	0.382	0.409
Residual Std. Error	2.092	1.974	2.004	1.958
F Statistic	21.272***	19.653***	18.272***	15.548***
Note:				*p<0.1; **p<0.05; ***p<0.01

While these sentiment measures are only one possible feature of the articles and minutes, this provides further evidence that there is information in central bank communication which would be of use the private sector, and that this information may not be completely captured by the media.

D.5 Stationarity test results

We verify that any identified co-movement is not spuriously driven by non-stationarity, we run an Augmented Dickey-Fuller test (Dickey and Fuller, 1979) on each of the series, which tests the null hypothesis of a unit root process against the alternative hypothesis of a stationary autoregressive process. At a 5% significance level, only four of the 60 series fail to reject non-stationarity, so we are justified in treating the series as stationary. Further details of ADF tests on each of the series are given in Appendix D.5.[6]

The equation estimated in the ADF tests on a topic proportion series $\{y_t\}_{t=1}^{T}$ is:

$$\Delta y_t = \alpha_0 + \alpha_1 t + (\gamma - 1)y_{t-1} + \delta_1 \Delta y_{t-1} + ... + \delta_p \Delta y_{t-p} + \varepsilon_t$$

There are therefore 3 relevant test statistics as three hypotheses are tested:

- τ: $(\gamma - 1) = 0$. Failing to reject this implies that there may be a unit root.
- ϕ_1: $(\gamma - 1) = \alpha_1 = 0$. Failing to reject this implies that there is a unit root and no time trend. If rejected, then there is at least either no unit root or no time trend.
- ϕ_2: $(\gamma - 1) = \alpha_0 = \alpha_1 = 0$. Failing to reject this implies that there is a unit root, no drift and no time trend.

Failing to reject at least one of these hypotheses suggests that the series may be well modelled by a non-stationarity unit root process. The null hypothesis is non-stationarity, so if the test statistic is greater than the critical value and we reject the null hypothesis then we have evidence that the series is stationary (or at least not a unit root). Table D.3 shows the relevant test statistics and 5% critical values for an ADF test with a trend, drift and one lag for each of the topic attention series.

Table D.3: Augmented Dickey-Fuller unit root test critical values

Parameter	1%	5%	10%
τ	-3.99	-3.43	-3.13
ϕ_1	6.22	4.75	4.07
ϕ_2	8.43	6.49	5.47

Table D.4 reports results for the jointly estimated topic proportions for the NYT articles and Federal Reserve minutes. Note that the asterisks to indicate significance have the reverse of their usual interpretation, due to the nature of the ADF test. We see here that only four of the 60 series fail to reject all three null hypotheses at the 5%

[6]The only series which comfortable fails to reject the null hypothesis of non-stationarity is Topic 30, which makes up around 15% of the FOMC minutes but under 2% of the NYT articles, and appears to deal explicitly with the committee policy decision.

critical value. Table D.5 reports results for the jointly estimated Federal Reserve, Bank of England and European Central Bank communication documents topic proportions, showing that only four of the 60 series fail to reject all three null hypotheses at the 5% critical value. It therefore appears to be that non-stationarity is not really a concern here.

For the FOMC-MPC-GC case, we also run an Augmented Dickey-Fuller test on each of the series, which tests the null hypothesis of a unit root process against the alternative hypothesis of a stationary autoregressive process. At a 5% significance level, only four of the 90 series fail to reject non-stationarity, so we are justified in treating the series as stationary. Further details of ADF tests on each of the series are shown in Table D.5. The series for which stationarity is (marginally) rejected are from the Federal Reserve corpus which has the fewest observations, as the FOMC meet less frequently than the MPC or GC.

Table D.4: Augmented Dickey-Fuller unit root test results for FOMC-NYT topic series

	Federal Reserve			*New York Times*		
	τ (-3.43)	ϕ_1 (4.75)	ϕ_2 (6.49)	τ (-3.43)	ϕ_1 (4.75)	ϕ_2 (6.49)
Topic						
T1	-5.75	11.05	16.55	-7.24	17.64	26.4
T2	-4.06	5.69*	8.49	-7.28	17.67	26.5
T3	-4.55	7.01	10.43	-7.02	16.44	24.64
T4	-6.19	12.78	19.17	-7.75	20.03	30.03
T5	-7.35	18.02	27.03	-9.47	29.92	44.87
T6	-5.45	9.95	14.83	-7.72	19.9	29.84
T7	-3.82*	4.98*	7.36*	-6.8	15.59	23.23
T8	-7.62	19.38	29.06	-7.82	20.41	30.6
T9	-3.49*	4.16**	6.24**	-5.54	10.25	15.38
T10	-3.64*	4.43**	6.64*	-4.66	7.27	10.87
T11	-8.11	21.93	32.89	-8.96	26.83	40.2
T12	-7.31	17.81	26.72	-7.43	18.52	27.7
T13	-4.57	6.98	10.47	-3.62*	4.38**	6.56*
T14	-6.01	12.05	18.07	-7.77	20.13	30.17
T15	-6.93	15.99	23.98	-6.94	16.06	24.08
T16	-4.63	7.16	10.74	-7.66	19.65	29.45
T17	-7.22	17.38	26.03	-7.65	19.52	29.26
T18	-7.72	19.88	29.81	-6.14	12.56	18.85
T19	-8.19	22.35	33.53	-7.21	17.39	26.08
T20	-4.12	5.72*	8.57	-6.68	14.89	22.33
T21	-5.64	10.64	15.96	-6.48	14.03	21.03
T22	-4.71	7.41	11.11	-7.65	19.53	29.28
T23	-6.09	12.37	18.56	-5.44	9.9	14.83
T24	-7.35	18.03	27.04	-6.13	12.54	18.81
T25	-6.77	15.3	22.94	-7.36	18.07	27.11
T26	-4.58	7.08	10.61	-4.86	7.89	11.83
T27	-4.98	8.28	12.41	-7.73	19.97	29.95
T28	-5.78	11.2	16.8	-6.24	12.97	19.45
T29	-6.22	12.89	19.32	-7.1	16.81	25.2
T30	-1.99***	1.89***	2.8***	-6.8	15.4	23.11

Note: *p>0.01; **p>0.05; ***p>0.1

Table D.5: Augmented Dickey-Fuller unit root test results for FOMC-MPC-GC topic series

Topic	*Bank of England*			*Federal Reserve*			*European Central Bank*		
Topic	τ (-3.43)	ϕ_1 (4.75)	ϕ_2 (6.49)	τ (-3.43)	ϕ_1 (4.75)	ϕ_2 (6.49)	τ (-3.43)	ϕ_1 (4.75)	ϕ_2 (6.49)
T1	-5.96	11.86	17.77	-4.39	6.45	9.65	-6	12.45	18.61
T2	-8.59	24.63	36.93	-4.66	7.26	10.88	-6.97	16.26	24.39
T3	-7.31	17.82	26.72	-4.29	6.32	9.21	-6.16	12.77	19.15
T4	-5.35	9.55	14.32	-3.9*	5.23*	7.67*	-6.41	14.01	21.02
T5	-5.21	9.04	13.56	-4.53	6.88	10.31	-7.61	19.33	28.98
T6	-7.17	17.15	25.72	-5.27	9.38	13.92	-6.65	14.75	22.13
T7	-6.3	13.31	19.88	-5.29	9.32	13.98	-4.73	7.47	11.19
T8	-7.63	19.44	29.15	-6.26	13.07	19.57	-8.14	22.09	33.14
T9	-9.06	27.36	41.03	-6.52	14.2	21.29	-4.43	7.84	11.68
T10	-7.87	20.67	30.96	-5.51	10.2	15.24	-7.33	17.97	26.95
T11	-5.34	9.54	14.31	-6.37	13.54	20.3	-8	21.36	32.03
T12	-5.47	9.98	14.96	-3.47*	4.03***	6.02**	-6.49	14.13	21.14
T13	-5.24	9.19	13.77	-4.94	8.15	12.23	-8.25	22.77	34.13
T14	-10.56	37.18	55.77	-7.63	19.44	29.13	-10.25	35	52.49
T15	-8.25	22.71	34.07	-5.33	9.49	14.23	-8.95	26.69	40.04
T16	-9.32	28.95	43.42	-5.88	11.57	17.3	-8.61	24.73	37.08
T17	-6.73	15.11	22.66	-4.03	5.42*	8.13	-7.08	16.85	25.28
T18	-8.3	23.02	34.52	-5.8	11.22	16.83	-4.82	8.15	12.15
T19	-8.41	23.59	35.38	2.91***	3.2***	4.8***	-7.97	21.2	31.73
T20	-6.19	12.83	19.17	-6.78	15.36	22.99	-5.85	11.79	17.66
T21	-7.28	17.73	26.58	-7.69	19.78	29.66	-6.76	15.25	22.88
T22	-7.22	17.4	26.05	-6.38	13.61	20.36	-7.91	20.91	31.36
T23	-7.52	18.87	28.28	-3.61*	4.41**	6.56*	-9.21	28.26	42.39
T24	-5.94	11.83	17.66	-4.29	6.31	9.44	-6.95	16.1	24.14
T25	-7.1	16.86	25.29	-5.47	10	14.98	-5.07	8.78	13.14
T26	-7.66	19.58	29.34	-4.8	7.7	11.54	-7.95	21.12	31.67
T27	-6.27	13.17	19.71	-5.01	8.43	12.64	-6.76	15.27	22.88
T28	-6.73	15.12	22.68	-3.26**	3.57***	5.31***	-6.9	15.98	23.96
T29	-8.74	25.49	38.22	-4.48	6.99	10.18	-8.69	25.19	37.78
T30	-6.84	15.62	23.42	-4.52	6.91	10.34	-5.77	11.09	16.63

Note: *p>0.01; **p>0.05; ***p>0.1

D.6 Topic proportions and forecast dispersion series

Figure D.7: Growth topic attention and RGDP forecast dispersion

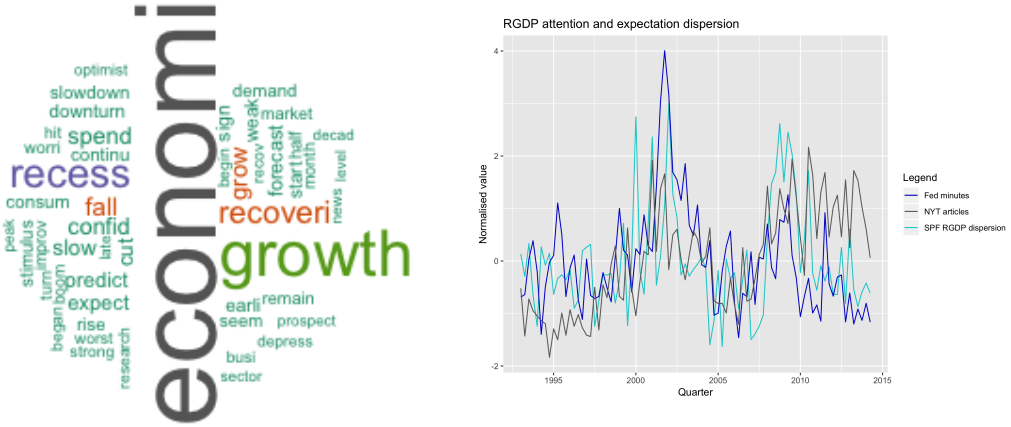


Figure D.9: Growth topic attention and NGDP forecast dispersion

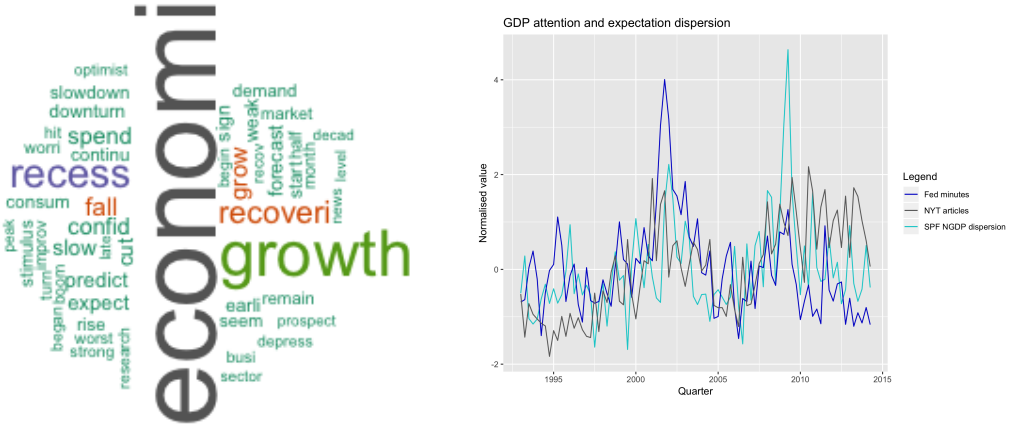


Figure D.11: Inflation topic attention and CPI forecast dispersion

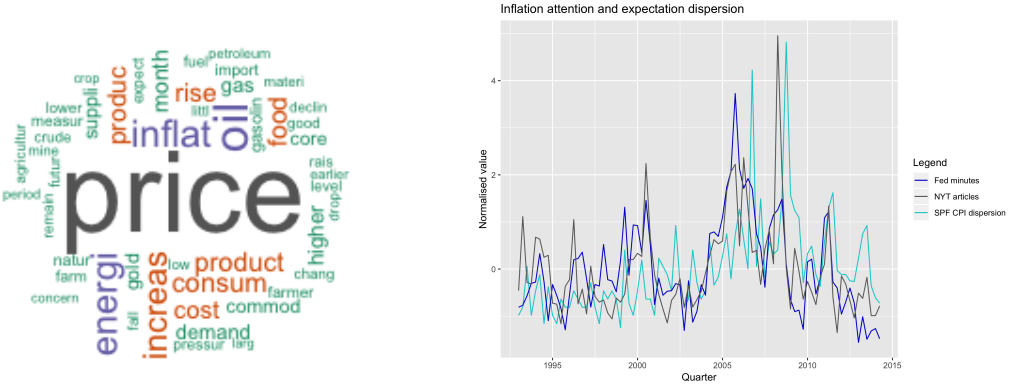


Figure D.13: Interest rate topic attention and TBILL forecast dispersion

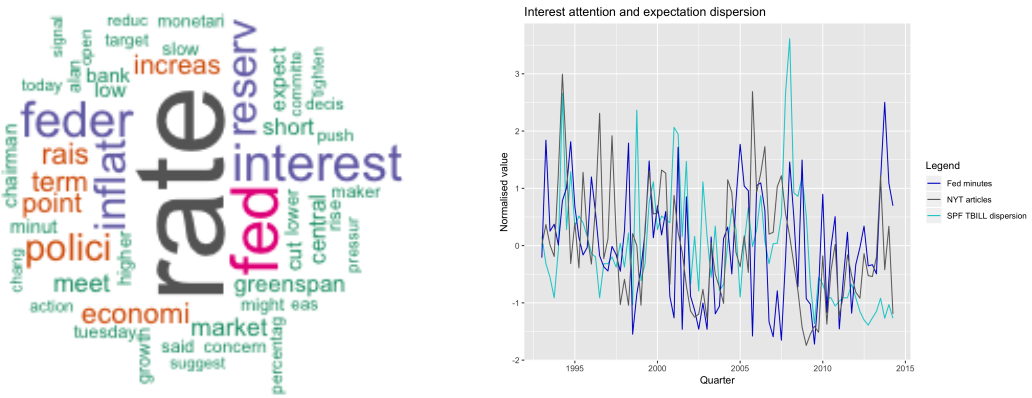


Figure D.15: Employment topic attention and EMP forecast dispersion

Figure D.17: Employment topic attention and UNEMP forecast dispersion

Figure D.19: Profit topic attention and CPROF forecast dispersion

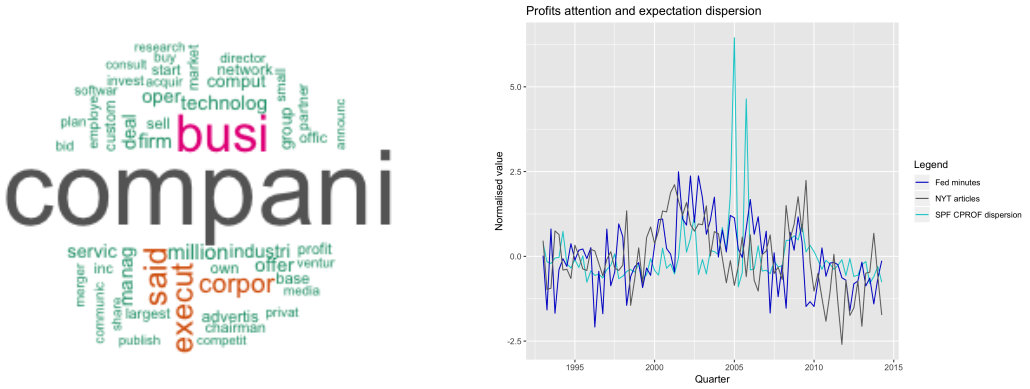


Figure D.21: Production topic attention and INDPROD forecast dispersion

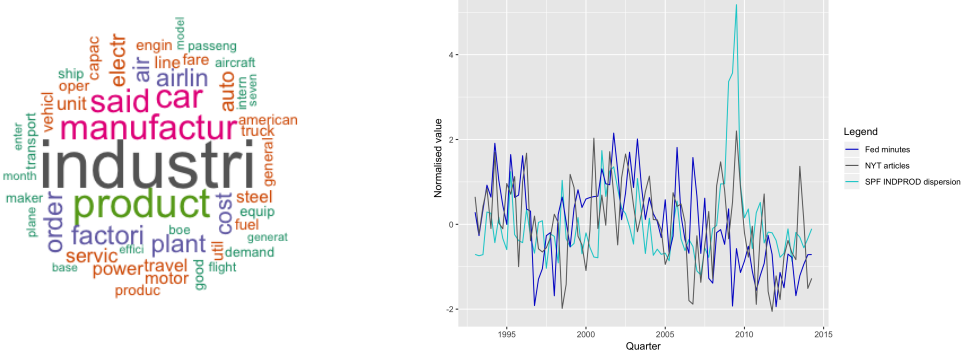


Figure D.23: Mortgage topic attention and HOUSING forecast dispersion

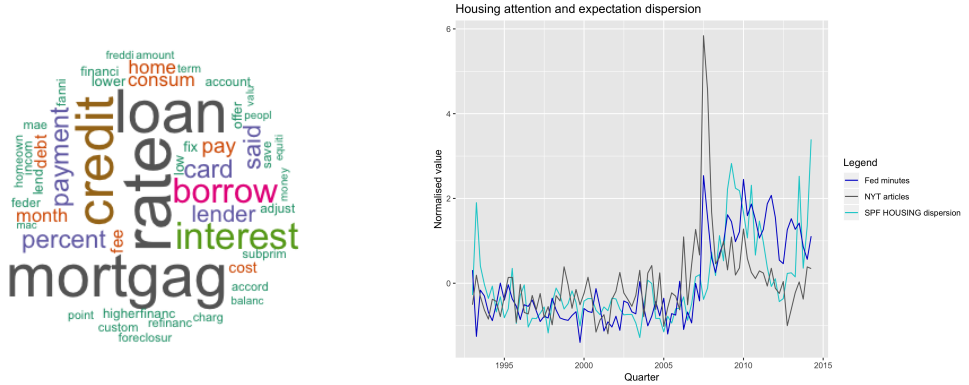


Figure D.25: Housing topic attention and RRESINV forecast dispersion

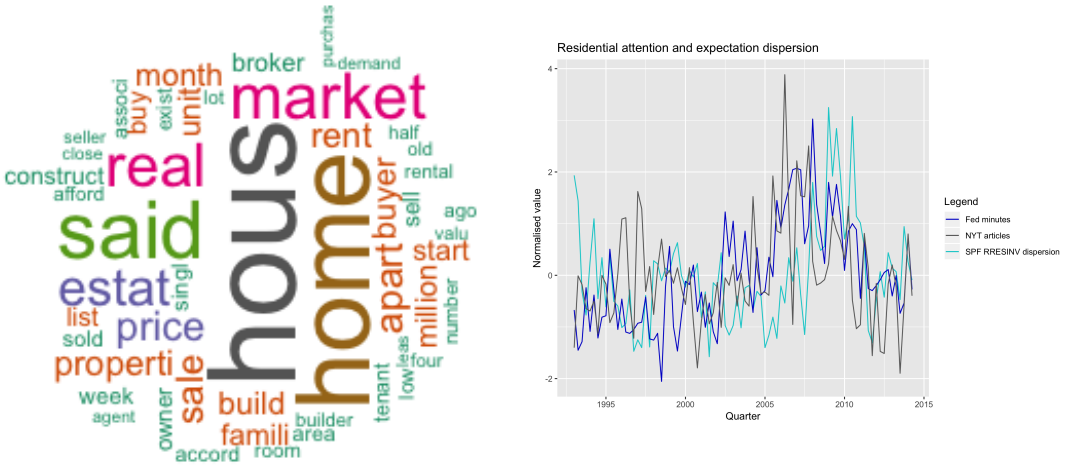


Figure D.27: Retail sales topic attention and RCONSUM forecast dispersion

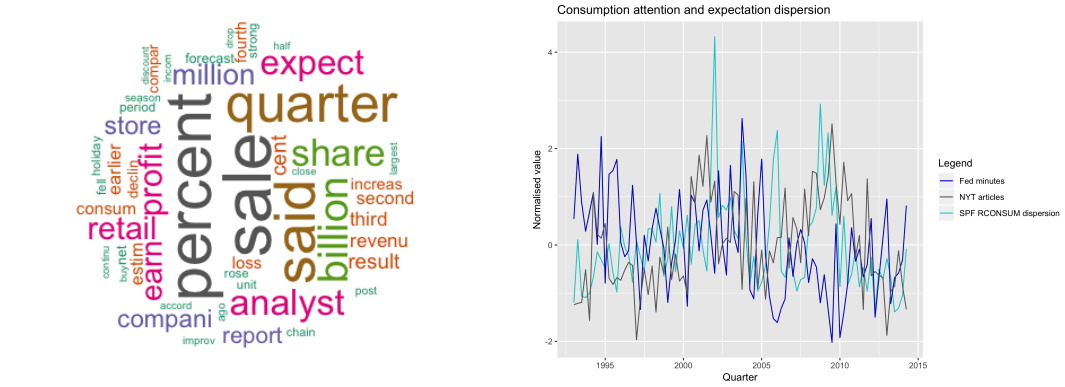


Figure D.29: Investor topic attention and RNRESIN forecast dispersion

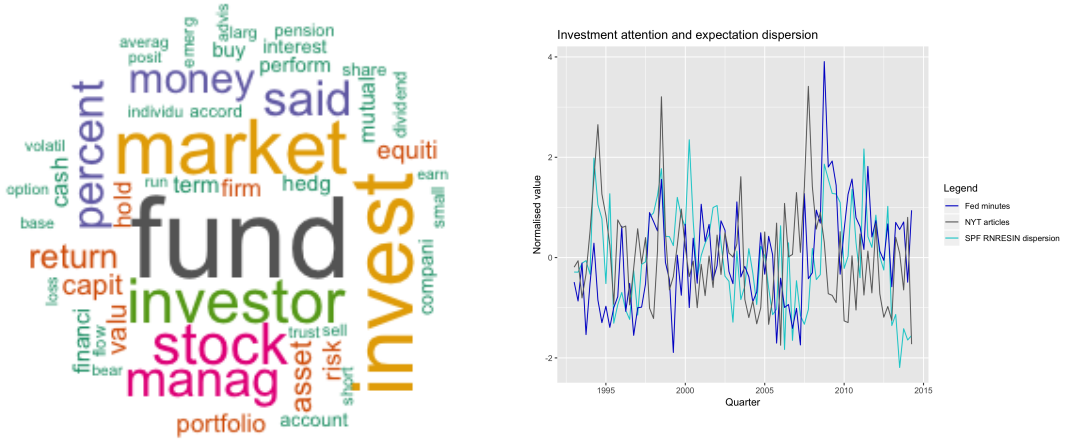


Figure D.31: Fiscal policy topic attention and RFEDGOV forecast dispersion

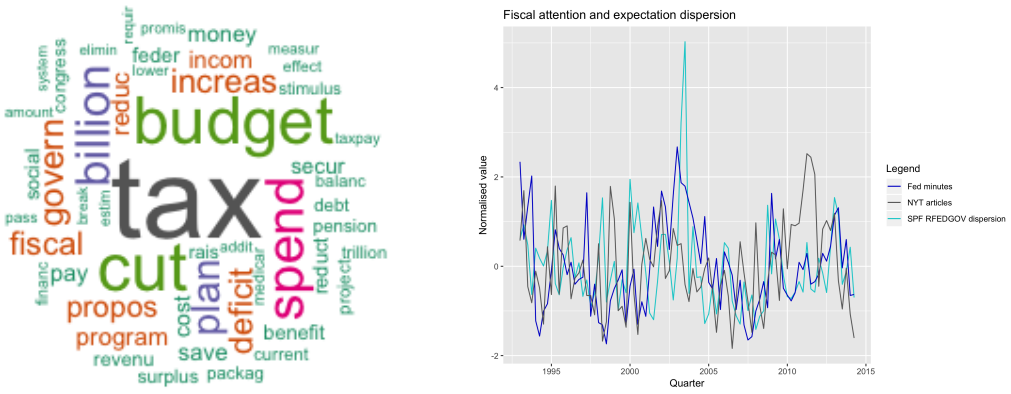


Figure D.33: Fiscal policy topic attention and RSLGOV forecast dispersion

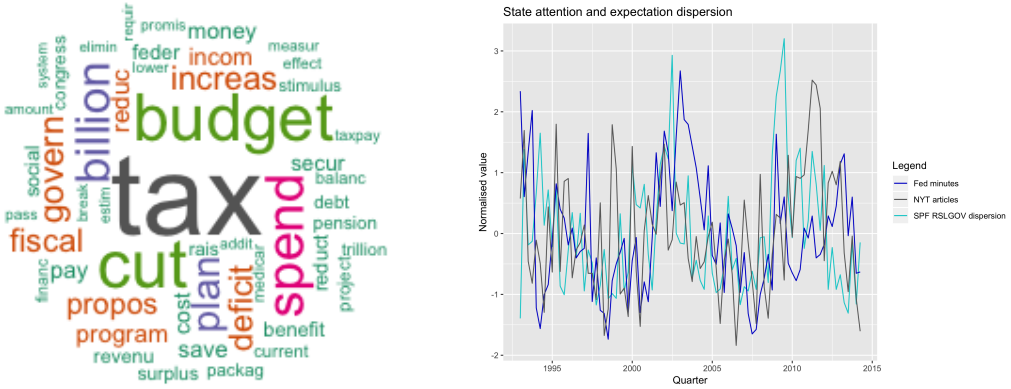


Table D.6: SPF dispersion and FOMC topic correlation matrix (matched series in bold)

	Federal Reserve Topic										
SPF	9 (Inflation)	20 (Growth)	16 (Labour)	23 (Business)	29(Production)	26 (Mortgage)	24 (Interest)	19 (Retail)	15 (Investment)	22 (Housing)	3 (Fiscal)
CPI	**0.305*****	−0.007	−0.029	0.056	−0.119	0.285***	−0.071	−0.341***	0.346***	0.435***	−0.085
	(0.004)	(0.951)	(0.792)	(0.606)	(0.273)	(0.008)	(0.518)	(0.001)	(0.001)	(0.000)	(0.435)
NGDP	−0.136	**0.332*****	−0.079	0.098	−0.072	0.284***	−0.236**	−0.158	0.375***	0.332***	0.115
	(0.212)	**(0.002)**	(0.470)	(0.372)	(0.508)	(0.008)	(0.029)	(0.146)	(0.000)	(0.002)	(0.293)
RGDP	−0.089	**0.360*****	-0.154	0.145	0.043	0.117	−0.157	−0.101	0.372***	0.108	0.116
	(0.417)	**(0.001)**	(0.158)	(0.184)	(0.692)	(0.283)	(0.149)	(0.355)	(0.000)	(0.321)	(0.288)
EMP	−0.100	0.281*	**0.363****	0.177	−0.111	0.277*	−0.266*	−0.178	0.424***	0.080	0.193
	(0.523)	(0.068)	**(0.017)**	(0.256)	(0.478)	(0.072)	(0.085)	(0.253)	(0.005)	(0.611)	(0.216)
UNEMP	−0.230**	0.183*	**0.105**	−0.034	−0.094	0.348***	−0.138	0.077	0.287***	0.121	0.143
	(0.033)	(0.092)	**(0.335)**	(0.759)	(0.387)	(0.001)	(0.205)	(0.479)	(0.007)	(0.267)	(0.191)
CPROF	0.315***	0.109	0.139	**0.273****	0.213**	−0.073	−0.015	0.047	−0.025	0.141	0.013
	(0.003)	(0.319)	(0.201)	**(0.011)**	(0.049)	(0.506)	(0.894)	(0.665)	(0.822)	(0.195)	(0.905)
INDPROD	−0.203*	0.363***	−0.026	0.057	**0.089**	0.179*	−0.294***	−0.246**	0.421***	0.164	0.185*
	(0.061)	(0.001)	(0.812)	(0.6)	**(0.415)**	(0.099)	(0.006)	(0.022)	(0.000)	(0.130)	(0.088)
HOUSING	−0.271**	−0.177	0.237**	−0.298***	−0.423***	**0.625*****	0.018	−0.202*	0.383***	0.270**	−0.050
	(0.012)	(0.103)	(0.028)	(0.005)	(0.000)	**(0.000)**	(0.869)	(0.062)	(0.000)	(0.012)	(0.645)
TBILL	0.262**	0.251**	−0.419***	0.191*	0.332***	−0.295***	**0.171**	0.084	−0.102	0.115	−0.260**
	(0.015)	(0.020)	(0.000)	(0.078)	(0.002)	(0.006)	**(0.116)**	(0.440)	(0.352)	(0.292)	(0.016)
RCONSUM	0.131	0.524***	−0.218**	0.351***	0.232**	−0.129	−0.221**	**−0.119**	0.223**	0.114	0.105
	(0.230)	(0.000)	(0.044)	(0.001)	(0.032)	(0.238)	(0.041)	**(0.274)**	(0.039)	(0.297)	(0.334)
RNRESIN	−0.014	0.175	−0.038	0.092	0.081	0.004	0.031	−0.018	**0.337*****	−0.107	−0.100
	(0.895)	(0.107)	(0.729)	(0.401)	(0.459)	(0.972)	(0.775)	(0.871)	**(0.002)**	(0.327)	(0.358)
RRESINV	−0.108	−0.109	0.101	−0.206*	−0.253**	0.394***	0.034	−0.149	0.301***	**0.340*****	−0.037
	(0.322)	(0.317)	(0.356)	(0.057)	(0.019)	(0.000)	(0.753)	(0.170)	(0.005)	**(0.001)**	(0.738)
RFEDGOV	−0.258**	0.063	−0.157	0.060	0.113	−0.046	0.044	0.034	0.096	−0.152	**0.281*****
	(0.017)	(0.567)	(0.148)	(0.586)	(0.298)	(0.672)	(0.687)	(0.756)	(0.379)	(0.162)	**(0.009)**
RSLGOV	−0.117	0.312***	−0.010	0.104	0.053	0.193*	−0.125	−0.069	0.323***	0.149	**0.124**
	(0.283)	(0.003)	(0.928)	(0.34)	(0.625)	(0.075)	(0.251)	(0.525)	(0.002)	(0.171)	**(0.257)**

Table D.7: SPF dispersion and NYT topic correlation matrix (matched series in bold)

	New York Times Topic										
SPF	9 (Inflation)	20 (Growth)	16 (Labour)	23 (Business)	29(Production)	26 (Mortgage)	24 (Interest)	19 (Retail)	15 (Investment)	22 (Housing)	3 (Fiscal)
CPI	**0.211***	0.317***	0.163	0.073	−0.056	0.269**	−0.122	0.183*	−0.047	0.145	−0.018
	(0.051)	(0.003)	(0.133)	(0.505)	(0.607)	(0.012)	(0.264)	(0.091)	(0.666)	(0.181)	(0.869)
NGDP	−0.03	**0.293*****	0.226**	0.227**	−0.049	0.251**	−0.4***	0.311***	−0.069	0.028	0.31***
	(0.787)	**(0.006)**	(0.037)	(0.035)	(0.654)	(0.020)	(00.000)	(0.004)	(0.527)	(0.795)	(0.004)
RGDP	−0.002	**0.278*****	0.17	0.342***	0.221**	−0.043	−0.378***	0.384***	−0.047	−0.125	0.19*
	(0.983)	**(0.010)**	(0.117)	(0.001)	(0.041)	(0.694)	(0.000)	(0.000)	(0.668)	(0.252)	(0.080)
EMP	−0.173	0.325**	**0.410*****	0.117	0.153	−0.071	−0.324**	0.354**	−0.357**	−0.025	0.339**
	(0.266)	(0.033)	**(0.006)**	(0.454)	(0.327)	(0.652)	(0.034)	(0.020)	(0.019)	(0.873)	(0.026)
UNEMP	−0.255**	0.279***	**0.331*****	0.153	0.023	0.059	−0.446***	0.186*	−0.262**	−0.177	0.211*
	(0.018)	(0.009)	**(0.002)**	(0.160)	(0.833)	(0.590)	(0.000)	(0.086)	(0.015)	(0.103)	(0.052)
CPROF	0.239**	0.018	−0.055	**0.012**	0.096	−0.044	0.053	0.123	0.07	0.016	−0.025
	(0.026)	(0.867)	(0.615)	**(0.910)**	(0.381)	(0.689)	(0.627)	(0.260)	(0.524)	(0.883)	(0.821)
INDPROD	−0.092	0.357***	0.269**	0.422***	**0.273****	0.008	−0.319***	0.493***	−0.042	0.028	0.061
	(0.399)	(0.001)	(0.012)	(0.000)	**(0.011)**	(0.945)	(0.003)	(0.000)	(0.704)	(0.797)	(0.574)
HOUSING	−0.046	0.432***	0.460***	−0.151	0.042	**0.273****	−0.317***	0.142	−0.161	0.051	0.112
	(0.671)	(0.000)	(0.000)	(0.166)	(0.703)	**(0.011)**	(0.003)	(0.192)	(0.139)	(0.639)	(0.304)
TBILL	0.271**	−0.112	−0.317***	0.308***	0.173	0.24**	**0.293*****	0.241**	0.375***	0.143	−0.112
	(0.012)	(0.303)	(0.003)	(0.004)	(0.111)	(0.026)	**(0.006)**	(0.025)	(0.000)	(0.188)	(0.304)
RCONSUM	0.038	0.069	0.145	0.413***	0.223**	−0.044	−0.307***	**0.275****	−0.016	0.113	0.026
	(0.731)	(0.528)	(0.182)	(0.000)	(0.039)	(0.690)	(0.004)	**(0.011)**	(0.885)	(0.302)	(0.816)
RNRESIN	−0.105	0.176	−0.052	0.196*	0.085	−0.144	−0.196*	0.27**	**0.017**	−0.144	0.206*
	(0.334)	(0.105)	(0.634)	(0.070)	(0.439)	(0.187)	(0.070)	(0.012)	**(0.879)**	(0.185)	(0.058)
RRESINV	0.009	0.369***	0.362***	0.072	0.053	0.156	−0.268**	0.287***	−0.066	**−0.012**	0.161
	(0.936)	(0.000)	(0.001)	(0.508)	(0.626)	(0.152)	(0.013)	(0.007)	(0.547)	**(0.909)**	(0.139)
RFEDGOV	−0.1	0.003	0.010	0.191*	0.162	−0.125	−0.094	0.03	0.039	−0.071	**0.070**
	(0.358)	(0.979)	(0.931)	(0.078)	(0.135)	(0.252)	(0.390)	(0.781)	(0.721)	(0.518)	**(0.519)**
RSLGOV	0.009	0.203*	0.276***	0.273**	0.304***	0.013	−0.317***	0.36***	−0.054	−0.046	**0.245****
	(0.934)	(0.061)	(0.010)	(0.011)	(0.004)	(0.907)	(0.003)	(0.001)	(0.622)	(0.672)	**(0.023)**

D.7 Additional Regression Results

Table D.8: FOMC minutes, NYT articles and SPF forecast dispersion (unstandardised data)

	Dependent variable:					
	$disp_{k,t}^{SPF}$		$\theta_{k,t}^{NYT}$		$\theta_{k,t}^{Fed}$	
	(1)	(2)	(3)	(4)	(5)	(6)
$\theta_{k,t}^{Fed}$	148.147***	65.008*	0.0.642***	0.419***		
	(31.526)	(35.556)	(0.054)	(0.068)		
$\theta_{k,t-1}^{Fed}$		−8.332		−0.148**		
		(35.090)		(0.069)		
$\theta_{k,t}^{NYT}$	5.626	−20.037			0.178***	0.091***
	(16.766)	(16.085)			(0.015)	(0.014)
$\theta_{k,t-1}^{NYT}$		42.478***				0.026*
		(16.351)				(0.014)
$disp_{k,t}^{SPF}$			0.00001	−0.0001*	0.0001***	0.0001**
			(0.0001)	(0.0001)	(0.00003)	(0.00002)
$disp_{k,t-1}^{SPF}$				0.00001		−0.00001
				(0.0001)		(0.00002)
Dep variable lags		✓		✓		✓
Topic fixed effects	✓	✓	✓	✓	✓	✓
Time fixed effects		✓		✓		✓
Observations	1,105	1,063	1,092	1,065	1,105	1,065
R^2	0.639	0.777	0.173	0.419	0.796	0.894
Adjusted R^2	0.634	0.754	0.161	0.360	0.793	0.884
Residual Std. Error	4.155	3.390	0.008	0.007	0.004	0.003
Note:						*p<0.1; **p<0.05; ***p<0.01

Table D.9: NYT and FOMC results (unstandardised data)

	Dependent variable:					
	$\theta^{\text{Fed}}_{m,k}$		$\Delta\theta^{\text{NYT}}_{m_p,k}$		$\theta^{\text{NYT}}_{m_p+w,k}$	
$\theta^{\text{NYT}}_{m-w,k}$	0.095***	0.039***		0.154***		0.114***
	(0.009)	(0.007)		(0.015)		(0.014)
$\theta^{Fed}_{k,t}$			0.178***	0.080***	0.093***	0.088***
			(0.021)	(0.026)	(0.018)	(0.026)
$\theta^{\text{NYT}}_{m_p-w,k}$				0.308***	0.444***	0.278***
				(0.014)	(0.012)	(0.014)
$\theta^{\text{NYT}}_{m+w,k}$				0.182***		0.113***
				(0.015)		(0.017)
$\theta^{\text{NYT}}_{m-w,k}$ lags		✓				
$\theta^{\text{Fed}}_{m,k}$ lags				✓		✓
Dep variable lags		✓		✓		✓
Topic fixed effect	✓	✓	✓	✓	✓	✓
Time fixed effect		✓		✓		✓
Observations	4,920	4,830	4,920	4,830	4,920	4,830
R^2	0.959	0.982	0.200	0.421	0.368	0.436
Adjusted R^2	0.959	0.981	0.196	0.396	0.364	0.412
Residual Std. Error	0.006	0.004	0.009	0.008	0.008	0.008
Note:						*p<0.1; **p<0.05; ***p<0.01

Table D.10: Panel VAR results on quarterly (unstandardised) topic proportions

	Dependent variable:					
	$\theta_{k,t}^{Fed}$		$\theta_{k,t}^{BoE}$		$\theta_{k,t}^{ECB}$	
	(1)	(2)	(3)	(4)	(5)	(6)
$\theta_{k,t-1}^{Fed}$	0.894***	0.613***	0.033***	0.099***	0.031**	0.052*
	(0.011)	(0.024)	(0.012)	(0.026)	(0.014)	(0.030)
$\theta_{k,t-1}^{BoE}$	−0.015	−0.001	0.728***	0.547***	0.034*	0.064**
	(0.015)	(0.021)	(0.016)	(0.023)	(0.019)	(0.027)
$\theta_{k,t-1}^{ECB}$	−0.000	−0.003	0.003	0.015	0.734***	0.504***
	(0.009)	(0.019)	(0.010)	(0.020)	(0.011)	(0.023)
Number of lags	1	3	1	3	1	3
Topic fixed effects	✓	✓	✓	✓	✓	✓
Observations	1,950	5,100	1,950	5,060	1,950	5,060
Number of groups	30	30	30	30	30	30
Obs per group	65	65	65	65	65	65
Note:	*p<0.1; **p<0.05; ***p<0.01					

Table D.11: The contents of published FOMC minutes robustly predicts other central banks' communication (unstandardised)

	Empirical strategy					
	3 month window			Most recent		
	(1)	(2)	(3)	(4)	(5)	(6)
$\gamma_{Fed,BoE}$	0.030*	0.019		0.020	0.016	0.009
	(0.016)	(0.014)		(0.014)	(0.012)	(0.013)
$\gamma_{Fed,ECB}$	0.010	−0.001		0.008	−0.001	−0.004
	(0.011)	(0.009)		(0.010)	(0.009)	(0.009)
$\gamma_{vBoE,Fed}$	0.066***	0.022**		0.070***	0.028***	0.0025***
	(0.010v)	(0.009)		(0.010)	(0.009	(0.009)
$\gamma_{BoE,ECB}$	0.013	−0.003		0.010	−0.004	−0.007
	(0.009)	(0.008)		(0.008)	(0.007)	(0.008)
$\gamma_{ECB,Fed}$	0.032***	0.025***		0.025**	0.018**	0.014
	(0.010)	(0.009)		(0.010)	(0.009)	(0.009)
$\gamma_{ECB,BoE}$	0.039***	0.039***		0.060***	0.053***	0.044***
	(0.013)	(0.012)		(0.011)	(0.010)	
Number of CB-specific lags	1	10	10	1	10	10
Topic-level macro controls			✓			✓
CB-topic fixed effect	✓	✓	✓	✓	✓	✓
Observations	15,000	14,760	15,060	15,330	15,060	15,060
R^2	0.819	0.842	0.856	0.817	0.842	0.859
Adjusted R^2	0.817	0.840	0.853	0.815	0.838	0.857
Residual Std. Error	0.015	0.014	0.014	0.016	0.014	0.0014
Note:	*p<0.1; **p<0.05; ***p<0.01					

Table D.12: Full results for publication policy change test

	Dependent variable:	
	var_value	
	(1)	(2)
$\gamma_{Fed,BoE}$	0.047***	0.035**
	(0.014)	(0.014)
$\gamma_{Fed,ECB}$	0.024*	0.013
	(0.014)	(0.013)
$\gamma_{BoE,Fed}$	0.022	0.018
	(0.019)	(0.019)
$\gamma_{BoE,ECB}$	0.013	0.005
	(0.011)	(0.011)
$\gamma_{ECB,Fed}$	0.075***	0.053**
	(0.021)	(0.021)
$\gamma_{ECB,BoE}$	0.045***	0.038***
	(0.012)	(0.012)
$\theta^{Fed}_{BoE,t,k}\mathbb{I}_{\{t\geq 2005\}}$	0.068***	0.056**
	(0.025)	(0.025)
$\theta^{Fed}_{ECB,t,k}\mathbb{I}_{\{t\geq 2005\}}$	−0.018	−0.022
	(0.027)	(0.027)
Number of CB-specific lags	4	10
Topic-specific macro controls		✓
CB-topic fixed effect	✓	✓
Observations	15,120	15,060
R^2	0.274	0.309
Adjusted R^2	0.268	0.299
Residual Std. Error	0.839 (df = 15010)	0.819 (df = 14842)
Note:		*p<0.1; **p<0.05; ***p<0.01

D.8 Additional Impulse Response Functions

Figure D.35: Impulse Response Functions with Cholesky identification and 2 lags

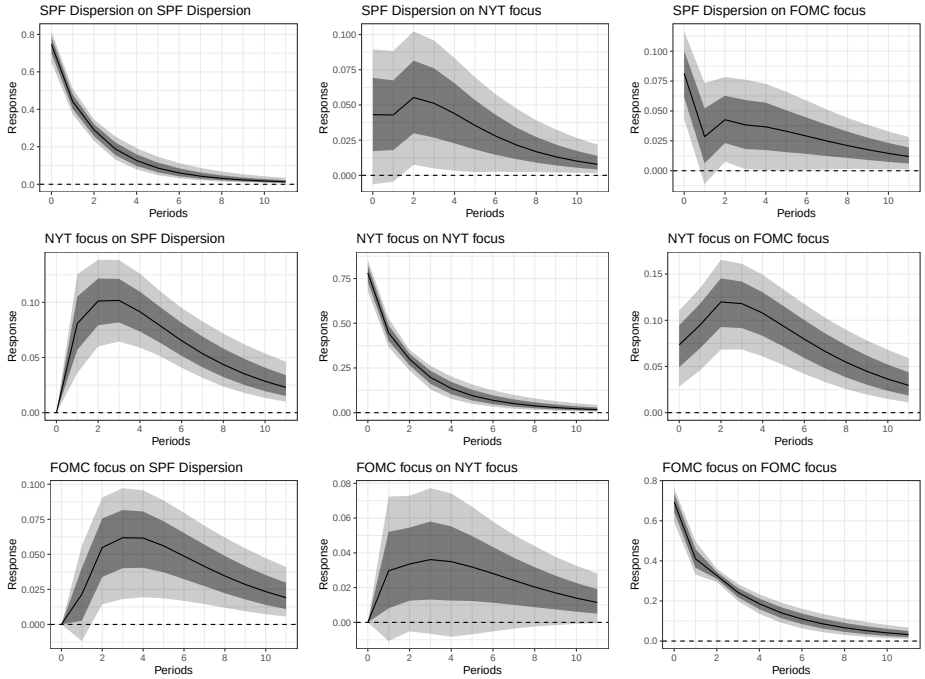


Figure D.37: Impulse Response Functions with Cholesky identification and 4 lags

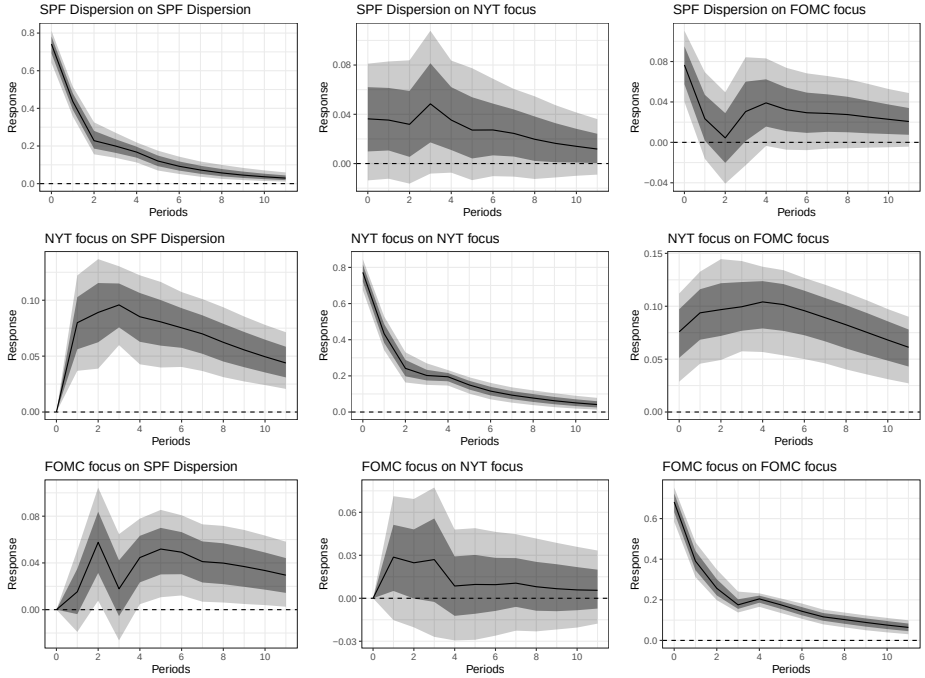


Figure D.39: Generalised Impulse Response Functions (Pesaran and Shin, 1998) with 4 lags