Economic Inferences Using Language Models

Daniel Mittendorf

MPhil

Submitted in fulfilment of the requirements of the degree of

Doctor of Philosophy

Adam Smith Business School

College of Social Sciences

University of Glasgow

March 2025

Abstract

I consider how causal and predictive economic inference problems can be solved using textual data. In doing so, I particularly investigate opportunities created by recent advances in natural language modelling, which appear to remain underexplored as an econometric methodology—a gap this thesis aims to help fill. Through this lens, I revisit three economic inference problems: estimation of the dynamic causal effects of monetary policy, forecasting of macroeconomic aggregates, and predicting financial risk premia. Re-examining the dynamic causal effects of UK monetary policy, I find that orthogonalising conventional instrumental variables with respect to large language model-extracted information appears to help resolve longstanding puzzles within empirical monetary economics. Examining the incremental value of the text of the Federal Reserve System's 'Beige Book' for forecasting US macroeconomic aggregates, I find apparent improvements in forecast accuracy, but also show how the inclusion of language model-generated predictors can make bias-free forecast evaluation challenging. Considering an investor's inference problem, I make a formal connection between microeconomic theory regarding rankings of information structures and the practical choice of which words or tokens should be prioritised in financial language modelling. Taken together, these findings underscore that language models are a powerful addition to the econometric instrumentarium available to researchers and practitioners. Whilst there are pitfalls to avoid, there can be little doubt that language models will become essential tools for drawing economic inferences.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank my supervisors, Dimitris Korobilis and Cathy Yi-Hsuan Chen, for their generous guidance throughout this project, as well as numerous discussants and seminar participants for helpful comments and suggestions. I am forever indebted to my family and partner for their support, love, and encouragement. This work was supported by the Economic and Social Research Council (ESRC) [project reference 2448498].

Author's declaration

I declare that, except where explicit reference is made to the contribution of others, that this dissertation is the result of my own work and has not been submitted for any other degree at the University of Glasgow or any other institution.

Chapter 1

Introduction

1.1 Motivation

Between 1762–63, Adam Smith delivered a lecture on natural language at the University of Glasgow (Land, 1977). Eventually published as ‘Considerations Concerning the First Formation of Languages’ as an appendix to his ‘Theory of Moral Sentiments’ in 1767, Smith investigates questions of linguistic structure. In other words, Smith was concerned with language modelling before he published ‘The Wealth of Nations’ in 1776. It appears reasonable, then, to hypothesise that Smith might have been interested in how natural language data can support economic decision-making.

Kleinberg et al. (2015) note that in situations where a decision maker has to choose X_0 so as to maximise a known function $\pi(X_0, Y)$ that depends on X_0 as well as stochastic Y, the choice X_0 depends on

$$\frac{d\pi(X_0,Y)}{dX_0} = \frac{\partial\pi}{\partial X_0} \underbrace{(Y)}_{\text{Solution to prediction problem}} + \frac{\partial\pi}{\partial Y} \underbrace{\frac{\partial Y}{\partial X_0}}_{\text{Solution to causal inference problem}} . \tag{1.1}$$

That is, choosing X_0 requires solving a prediction problem (Kleinberg et al., 2015), a causal inference problem (Imbens, 2024), or both. In this thesis, I refer to both types of statistical problems collectively as economic inference problems and explore the extent to which textual data can help solve them. With around 80 to 90% of all data estimated to be unstructured information (Harbert, 2021), one might expect it to be possible to improve economic inferences by using text. However, the inherently high dimensionality of natural language makes harnessing textual data non-trivial. Luckily, language models have improved since Smith's days, with recent advances creating countless new opportunities to solve economic inference problems using textual data. These opportunities appear to remain largely unexplored, a gap this thesis aims to help fill.

1.2 Existing literature

A number of surveys explore the potential use of textual data for economic inference, discussing methods that vary greatly in complexity. Methodological approaches range from word counting of researcher-selected phrases to large language models with billions of parameters estimated on large swathes of the internet. Gentzkow et al. (2019) survey the use of textual data in economic research, covering different options for representing documents numerically and methods for modelling the relationship between numerical representations of text and concepts of interest. Algaba et al. (2020) provide an overview of econometric methodologies to capture sentiment from unstructured data in general and text in particular. Ash and Hansen (2023) survey specific algorithms for using textual data within economics. They also define four distinct measurement problems faced by applied researchers that text algorithms can help solve.

These surveys cover numerous examples of textual data being leveraged for economic inferences, although most of these applications involve simple language representations, such as word counts, sentiment scores (Tetlock, 2007), and topic models (Blei et al., 2003). They also provide examples of economic inference problems that textual data can help solve, such as estimating the impact of central bank communications, predicting economic conditions, and estimating financial market risk premia.

Recently, attention has turned specifically towards the striking effectiveness with which autoregressive deep neural network architectures that include so-called attention mechanisms can model natural language text (Vaswani et al., 2017). Beyond the emerging impacts of this technology across the economy (Eisfeldt and Schubert, 2024), a new line of research is focused on how economic researchers and practitioners can use large language models to solve economic inference problems more effectively using textual data (Korinek, 2023; Dell, 2025; Korinek, 2024; Hoberg and Manela, 2025; Ludwig et al., 2025). Language models have been successfully leveraged for measurement problems, for example Hansen et al. (2023), who use a transformer-architecture language model to label automatically 250 million job adverts with minimal error.

However, due to the recency of the latest significant advances in natural language processing, the use of advanced language models for solving economic inference problems remains underexploited and understudied. Firstly, in the context of observational causal inference, large language models can be leveraged to obtain measurements of otherwise unobserved confounding variables from text. These measurements can then be used to arrive at deconfounded causal effect estimates. While the generation of such conditioning information from raw text was possible prior to the advent of transformers, the cost of labelling textual documents is now substantially lower—effectively expanding the set of economic inference problems that can be tackled. An important observational causal inference problem where confounding remains a concern is the estimation of the dynamic effects of monetary policy (Romer and Romer, 2000; Gürkaynak et al., 2005; Reeves and Sawicki, 2007; Gertler and Karadi, 2015; Miranda-Agrippino, 2016; Ramey, 2016; Nakamura and Steinsson, 2018; Stock and Watson, 2018; Cieslak and Schrimpf, 2019; Cesa-

Bianchi et al., 2020; Jarociński and Karadi, 2020; Andrade and Ferroni, 2021; Miranda-Agrippino and Ricco, 2021; Bauer et al., 2022; Bauer and Swanson, 2023a,b; Aruoba and Drechsel, 2024; Braun et al., 2025; Schmitt-Grohé and Uribe, 2024).

Secondly, in the context of predictive inference, much of the existing literature on prediction using text relies on generic high-dimensional regression methods (Giannone et al., 2021) and tends to assume simple text representations such as word or phrase counts (Taddy, 2013, 2015; Kelly and Pruitt, 2015; Kelly et al., 2021), or low-dimensional sentiment scores. There are few studies so far on the extent to which predictors derived from text using language model embeddings can improve performance on economic prediction problems. This generally includes the literature on economic forecasting using text (Armesto et al., 2009; Sadique et al., 2013; Larsen and Thorsrud, 2019; Ardia et al., 2019; Aromi, 2020; Bybee et al., 2020; Thorsrud, 2020; Kelly et al., 2021; Ellingsen et al., 2022; Borup and Schütte, 2022; Kalamara et al., 2022; Filippou et al., 2024). Carriero et al. (2025) focus specifically on forecasting economic time series using transformers, although without augmenting the predictor set with text-derived variables. As Carriero et al. (2025) note, valid evaluation of forecast performance is challenging when using pre-trained language models, due to the difficulty of simulating different information set vintages (Sarkar and Vafa, 2024; Ludwig et al., 2025). Estimates of the extent to which language model lookahead effects bias out-of-sample performance evaluations are still rare.

Thirdly, the extent to which the architecture of language models can be optimised for specific economic inference problems is yet to be explored comprehensively. In the context of predicting financial risk premia, existing studies tend to either rely on simple text representations (Tetlock, 2007; Loughran and McDonald, 2015; Manela and Moreira, 2017; Feuerriegel and Gordon, 2018; Ke et al., 2019) or fine-tune off-the-shelf language models such as BERT (Devlin et al., 2019) or BLOOM (Scao et al., 2022) on financial corpora whilst maintaining the use of generic model architectures (Huang et al., 2023; Wu et al., 2023).

1.3 Aims & thesis structure

Having identified gaps in the literature as discussed in the previous section, the overall objective of this thesis is to contribute to filling them.

My first aim is to revisit an important causal inference problem by using a large language model to obtain measurements of otherwise unobserved confounding variables from relevant documents. In particular, I reassess the effects of the Bank of England's monetary policy on the UK economy. This investigation is necessitated by mounting evidence that a standard approach to identify the dynamic causal effects of monetary policy shocks—so-called high-frequency identification—is vulnerable to producing biased effect estimates. An important assumption underpinning this standard identification strategy is that intraday bond market movements around monetary events are valid instruments, in the sense of being exogenous with respect to economic conditions. However, previous studies have found correlations between these so-called 'monetary policy surprises' and information about the macroeconomic situation available before each monetary event. Correlations of this kind cast doubt on existing results about the direction and magnitude of monetary non-neutrality. Indeed, impulse responses can show puzzling dynamic effects such as an increase in real activity following a contractionary monetary policy shock. I aim to deconfound these causal effect estimates by orthogonalising the instruments used with respect to economic conditions prior to each event, with measures thereof being extracted from pre-event news articles using a large language model. This aim is pursued in Chapter 2.

My second aim is to investigate the potential and pitfalls when using language models to incorporate textual data into economic forecasts. In particular, I explore the extent to which additional predictive macroeconomic information can be extracted from 8,580 Federal Reserve System 'Beige Book' reports. That is, the objective is to extract information from text that is not already contained in non-textual variables that are standard in the forecasting literature. I also aim to compare the relative performance of different text representation methods in extracting such 'marginal' predictive signals, as well as quantifying the impact of language model lookahead bias affecting out-of-sample forecast evaluations using a knowledge cutoff experiment. This aim is pursued in Chapter 3.

My third aim is to explore how language models can be optimised for solving economic inference problems. In particular, I focus on the prediction of financial market risk premia, a central challenge for financial market participants. I seek to determine whether the resource cost of predicting risk premia using large language models could be reduced by studying the general problem of how a data-driven investor can extract parsimonious sets of predictive features from high-dimensional data. This aim is pursued in Chapter 4.

1.4 Contributions

In pursuing the aims described in the previous section, this thesis makes a number of methodological, empirical, and theoretical contributions.

The first methodological contribution, a novel approach to orthogonalise monetary policy surprises against pre-event textual data, is introduced in Chapter 2. The approach utilises a large language model (LLM) to generate structured measures of economic conditions from unstructured text. For the particular implementation in this thesis, this involves assembling a corpus of pre-event newswires relating to UK monetary events. That said, the approach is general and the textual data it requires are available for central banks across jurisdictions. In Chapter 3, I propose a novel methodology for high-dimensional time series regression tailored to the characteristics of text-derived predictors. The proposed method is designed for the large sets of predictor variables that tend to arise when representing text numerically using language model embeddings, enables efficient forecast evaluation, and can be specified to allow for time-varying parameters. Specifically, the approach involves modelling the distribution of prediction targets using a switching Gaussian state space model with compressed predictors. I also describe how a knowledge cutoff experiment can be conducted to assess the impact of language model lookahead bias on forecast evaluation results.

Empirical contributions in Chapter 2 include evidence of ex post correlations between state-of-the-art measures of UK monetary policy surprises and structured pre-event information extracted from text by the large language model, as well as new evidence regarding the effects and transmission of UK monetary policy. Chapter 3 contributes a quantification of the incremental value of Beige Book text for forecasting US macroeconomic aggregates, as well as measurements of language model temporal lookahead bias.

Finally, a theoretical connection is made in Chapter 4 between the choice of which words or tokens to include in the vocabulary of a task-specific financial language model and the microeconomic literature on rankings of information structures. Simulations and an illustrative empirical exercise are performed to gather evidence regarding a microfounded approach to feature selection.

1.5 Findings & implications

In Chapter 2, I find that the proposed text-orthogonalisation approach using a large language model makes a material difference to the estimated dynamic causal effects of UK monetary policy on macroeconomic aggregates. In both small and large Bayesian vector autoregression (BVAR) specifications, the sign of estimated effects can depend on whether or not conventional monetary policy surprises are orthogonalised with respect to measurements of public, pre-event information regarding UK economic conditions. Text-orthogonalisation tends to yield effect estimates that are aligned with theoretical consensus and international evidence. For example, it resolves both a real activity puzzle and an employment puzzle that arise when using conventional monetary policy surprises to identify monetary policy shocks in small BVAR specifications. The findings lend empirical support to the 'central bank response'-mechanism proposed in Bauer and Swanson (2023a), since it appears not to be necessary to orthogonalise monetary policy surprises with respect to private, central bank-internal information to resolve real activity and employment puzzles. As such, the findings are inconsistent with the hypothesis that central banks must have significant 'inside information' about the state of the economy. Instead, my findings are consistent with the idea that it is market participants' imperfect knowledge of the monetary authority's reaction function to publicly-accessible news that explains the observed endogeneity of conventional, unorthogonalised monetary policy surprises.

Quantifying the incremental value of Beige Book text for forecasting US macroeconomic aggregates in Chapter 3, I find moderate improvements in forecast accuracy. That said, there is significant heterogeneity in the extent to which the addition of text-based predictors improves forecasts for different targets and at different horizons. At shorter horizons, for instance, text-augmented forecasts appear to near-uniformly perform as good or better than the non-text benchmark. Having quantified the strength of the predictive information that different methodologies are able to recover from raw text, I also explore a pitfall of using language models for economic forecasting: temporal lookahead bias during forecast evaluation. Using a knowledge cutoff experiment to simulate the training of a language model at different points in time, I find evidence of language model temporal lookahead bias. In this specific empirical context, the findings suggest that up to half of the estimated forecast accuracy gain due to the addition of predictors generated using language models is illusory. Avoiding this temporal lookahead bias whilst using state-of-the-art models can be costly, as doing so requires retraining large language models. For the largest language models, training costs are reported to exceed $40m Cottier et al. (2025), making complete avoidance of temporal lookahead bias through expanding window estimation of language model parameters practically infeasible. Given this constraint, further work regarding the size of temporal lookahead biases across different economic forecasting settings, as well as approaches to avoid them altogether, could be fruitful. Finally, while there appears to be some information about the economic outlook within Beige Book text that is not fully captured by traditional, non-textual predictor variables, future research could investigate further optimisation of how textual predictors are incorporated into forecasts. This could potentially result in forecasts being improved more uniformly across targets and horizons.

In Chapter 4, I find that in investor-relevant settings, ranking features using mutual information has a strong theoretical justification. Through simulations, I observe that while different mutual information estimators enable different choices along the bias-variance trade-off, they generally yield similar feature rankings. In the context of financial large language modelling, ranking tokens by their mutual information with investment returns to select language model vocabularies could enhance both computational efficiency and accuracy. An illustrative experiment supports this idea. These findings indicate that economic theory can guide the design and application of large language models in empirical asset pricing. Using theory-based rankings to pre-select text tokens for inclusion in a model could offer significant benefits. Such microfounded token filtering could enable the training of asset-specific language models tailored to predict the risk premia of particular financial assets. Thus, selecting vocabularies for financial large language models in a theory-based manner could reduce the resource costs of financial languauge models, improve market efficiency, or both.

Chapter 2

A reassessment of the dynamic causal effects of UK monetary policy: new evidence from LLM-orthogonalised high-frequency instruments

This paper reassesses the effects of the Bank of England's monetary policy shocks on the UK economy, necessitated by mounting evidence that a standard approach to identify the dynamic causal effects of monetary policy shocks is vulnerable to producing biased effect estimates. I develop a novel approach to orthogonalise conventional high-frequency monetary policy surprises with respect to economic conditions using a large language model. I find that conventional UK 'surprises' are predictable ex post and that text-orthogonalisation tends to yield effect estimates that are aligned with theoretical consensus and international evidence. These results highlight how text and language models can be used to address instrument invalidity problems of an important identification strategy within empirical macroeconomics.

2.1 Introduction

This paper investigates the dynamic causal effects of the Bank of England's monetary policy on macroeconomic variables. Within empirical monetary economics, the standard identification strategy employed to draw inferences of this kind uses high-frequency yield curve movements around monetary events, such as policy announcements and publications. These movements are then used as external instrumental variables (so-called 'monetary policy surprises') in the structural vector autoregression (SVAR) or linear projection (LP) frameworks. A key assumption underpinning this standard identification strategy is that the monetary policy surprise measures used are valid instruments, in the sense of being exogenous with respect to economic conditions. However, previous studies have found correlations between monetary policy surprises and pre-event information about the macroeconomic situation. Such correlations call into question the validity of high-frequency identification strategies, casting doubt on the resulting estimates of monetary non-neutrality. Indeed, impulse responses produced using high-frequency identification can show puzzling effects—such as an increase in real activity following a contractionary monetary policy shock.

Different explanations for the observed correlations and puzzling dynamic causal effect estimates have been explored in the literature. For example, the observed endogeneity of monetary policy surprises could be rationalised if the central bank acquires private information about the state of the economy ('central bank information'). During a monetary event, implicit or explicit revelation of this information would be incorporated into asset prices by market participants, contaminating monetary policy surprises with endogenous variation and rendering them invalid as external instrumental variables for monetary policy shocks. Another explanation rests on the observation that market participants are unlikely to have perfect information about the central bank's reaction function ('central bank response'). In particular, Bauer and Swanson (2023a) show how market participants' imperfect knowledge of the central bank's monetary policy reaction function can induce *ex post* correlations—i.e. correlations that emerge in retrospect—between the instruments and other variables. Their prescription is to regress conventional monetary policy surprises on information that was publicly available prior to the monetary event, and use the residuals (so-called 'orthogonalised monetary policy surprises') as external instruments in the SVAR or LP frameworks. In related work, Bauer and Swanson (2023b) call for further investigation of relevant information sources, as well as analyses of the extent to which their findings generalise to monetary policy conducted in jurisdictions outside of the United States. This paper heeds their call.

Contributions of this study include the development and implementation of a novel methodology to measure economic conditions from unstructured textual data. The method developed utilises a large language model (LLM) to generate structured measures. The approach is general and the textual data it requires are available for central banks in many jurisdictions. My methodology involves assembling a corpus of pre-event newswires relating to UK monetary events since the Bank of England was granted operational independence from the British government in 1997. I demonstrate that there are ex post correlations between state-of-the-art measures of UK monetary policy surprises and structured pre-event information extracted from text by the LLM. I then contrast the dynamic causal effects of monetary policy estimated using conventional and text-orthogonalised monetary policy surprises, which yields new evidence regarding the effects and transmission of UK monetary policy.

As such, this paper contributes to answering three of the priority topics in the Bank of England's 2024 Agenda for Research, namely: *'How can machine learning and artificial intelligence be deployed by supervisors and central banks?'*, *'How do central bank policy rates affect inflation and has this relationship changed over the recent past?'*, and *'What are the transmission mechanisms of conventional monetary policy and central bank balance sheet* *adjustments?'*.[1] This study also creates new research data assets for the empirical study of UK monetary policy. Firstly, it assembles a corpus of raw newswire text relating specifically to the set of UK monetary events in the Braun et al. (2025) database. Secondly, it produces structured data about UK economic conditions prior to each monetary event. Finally, it generates a set of text-orthogonalised monetary policy surprise series that, in standard SVAR-IV and LP-IV specifications, yield puzzle-free dynamic causal effect estimates.

Orthogonalising state-of-the-art conventional UK monetary surprises with respect to LLM-extracted pre-event information makes a material difference to the estimated dynamic causal effects of the Bank of England's monetary policies on key UK macroeconomic aggregates. My findings lend empirical support to the 'central bank response'-mechanism proposed in Bauer and Swanson (2023a). In particular, I find that it is not necessary to orthogonalise monetary policy surprises with respect to private, central bank-internal information[2] to resolve real activity and employment puzzles and arrive at dynamic effect estimates that align with theoretical consensus. As such, my findings are inconsistent with the hypothesis that the Bank of England must have significant 'inside information' about the state of the economy. Instead, my findings are consistent with the idea that it is market participants' imperfect knowledge of the Bank of England's reaction function to publicly-accessible news that explains the observed endogeneity of conventional, unorthogonalised monetary policy surprises.

My methodology involves compiling a corpus of newswire articles that would have been available to market participants just prior to each of the monetary events in the database of UK monetary policy surprises compiled by Braun et al. (2025). The key advantage of using same-day newswire articles is that in theory these should have a close relationship with the informa-

[1] https://www.bankofengland.co.uk/research/bank-of-england-agenda-for-research

[2] For example greenbook forecasts that are only published with a lag as in Miranda-Agrippino and Ricco (2021).

tion sets of market participants, and include the latest data points available at the time of publication.[3] Having collected these raw data, I take measurements of perceived UK economic conditions as described in each article just before each monetary event. I do this by prompting a large language model to answer a set of multiple choice questions about each article. I then create text-orthogonalised monetary policy surprises by regressing conventional monetary policy surprises on the multiple choice answers and keeping the residuals, consistent with the recommendation in Bauer and Swanson (2023b).

Related literature

My primary contribution relates to the literature on high-frequency reactions to central bank policy communications and their macroeconomic effects (Reeves and Sawicki, 2007; Miranda-Agrippino, 2016; Ramey, 2016; Nakamura and Steinsson, 2018; Cieslak and Schrimpf, 2019; Jarociński and Karadi, 2020; Cesa-Bianchi et al., 2020; Miranda-Agrippino and Ricco, 2021; Kaminska and Mumtaz, 2022; Braun et al., 2025). I rely on the state-of-the-art UK monetary policy surprise dataset constructed in Braun et al. (2025), which is a key input to producing my text-orthogonalised series. In so doing, I engage directly with the challenges of high-frequency identification raised in Nakamura and Steinsson (2018), Bauer and Swanson (2023b) and Bauer and Swanson (2023a). My approach is similar in principle to Aruoba and Drechsel (2024), who also use textual data to orthogonalise monetary shocks, although I focus on the UK rather than the US and rely on a different natural language processing methodology to transform unstructured text into structured data. This study more generally contributes to the emerging literature on the uses of natural language processing for economic research (Ash and Hansen, 2023), particularly ‘concept detection’ using large language models (Hansen

[3] Bauer and Swanson (2023a) (p.675) find that in the US such intra-month data releases often contain significant information regarding economic conditions.

et al., 2023; Korinek, 2023). Econometrically, I benefit from recent advances in understanding dynamic causal effect identification and estimation (Stock and Watson, 2018; Plagborg-Møller and Wolf, 2021; Miranda-Agrippino and Ricco, 2021; Li et al., 2024).

Paper structure

Section 2.2 discusses the potential endogeneity of conventional monetary policy surprises. Section 2.3 details the methodology used to measure economic conditions prior to UK monetary events from textual data using large language model inference. Section 2.4 discusses how text-orthogonalised monetary policy surprises are obtained from conventional monetary policy surprises and the measures obtained in Section 2.3. Section 2.5 investigates how the choice of identification affects estimated dynamic causal effects of UK monetary policy on key macroeconomic aggregates. Section 2.6 uses text-orthogonalised monetary policy surprises to study the transmission mechanism of UK monetary policy in a large VAR. Section 2.7 concludes.

Appendix A.1 describes the assembly of the corpus of pre-event newswires relating to UK monetary events. Appendix A.2 discusses how the LLM is prompted to turn unstructured text into structured information regarding UK economic conditions. Appendix A.3 discusses and validates the measurements obtained using large language model inference. Appendix A.4 explores the ex post empirical relationships between measured UK economic conditions and the financial market reactions to UK monetary events compiled by Braun et al. (2025). Appendix A.5 provides details on the orthogonalisation methods used. Appendix A.6 presents the results of sensitivity analyses. Appendix A.7 compares the results of this study with those in the literature.

2.2 Endogeneity of external instruments constructed from high-frequency monetary policy surprises

This section discusses the potential endogeneity issues that may arise when attempting to estimate the effects of monetary policy using conventional monetary policy surprises, beginning with an overview of the high-frequency identification approach. I then consider empirical findings that cast doubt on the key assumption that high-frequency financial market reactions to monetary events are inherently valid instrumental variables. This is followed by a discussion of different theoretical explanations for the observed patterns, and their implications for dynamic causal effect estimation.

2.2.1 High-frequency identification strategies

Stock and Watson (2018) discuss the exogeneity and invertibility conditions under which external instruments can be used to identify and estimate the dynamic causal effects of structural economic shocks—including monetary policy shocks. In the context of monetary policy, high-frequency interest rate market movements around central bank communications ('monetary policy surprises') are increasingly used as external instruments to identify monetary policy shocks and estimate impulse response functions of monetary non-neutrality (e.g. Gertler and Karadi (2015); Nakamura and Steinsson (2018)). Considering changes over a tight time window around specific monetary events, it is probable that a substantial amount of the variation in yield curve rates during this short period is driven by information released during the monetary event. Conversely, in the absence of arbitrage opportunities, the information released must be surprising for it to move bond prices. That is, any predictability of the content of monetary communications should be 'arbitraged away' due to the informational efficiency of financial markets.

Monetary policy surprises, the argument goes, should therefore be exogenous with respect to the state of the economy. As such, they should satisfy the instrument validity conditions in Stock and Watson (2018), enabling proper identification and estimation of the dynamic causal effects of monetary policy shocks.

2.2.2 Instrument validity concerns

There are concerns around the key assumption underpinning high-frequency identification strategies that monetary policy surprise measures used are exogenous with respect to economic conditions. In particular, there is growing evidence that casts doubt on the hypothesis that conventional high-frequency monetary policy surprises satisfy the exogeneity conditions required for them to be valid external instruments. This includes evidence that monetary policy surprises are correlated with pre-event information (e.g. Nakamura and Steinsson (2018); Miranda-Agrippino and Ricco (2021); Bauer and Swanson (2023a,b); Aruoba and Drechsel (2024)). Moreover, dynamic causal effects estimated using conventional monetary policy surprises as external instruments can appear inconsistent with the theoretical consensus. Examples of puzzling results include real activity puzzles (i.e. contractionary monetary policy increasing real activity or reducing unemployment) and price puzzles (i.e. contractionary monetary policy increasing the price level).[4] As such, the standard approach to identify the dynamic causal effects of monetary policy shocks appears vulnerable to producing biased effect estimates.

[4] For example, using an identification strategy that relies on conventional monetary policy surprises for Bank of England monetary events, Braun et al. (2025) identify both real activity and price puzzles.

2.2.3 Theoretical explanations for observed endogeneity

Previous studies have explored different explanations for the observed endogeneity of monetary policy surprises and puzzling dynamic effect estimates. For example, one theoretical setting in which the puzzling results could be rationalised is if the central bank had private information about the state of the economy. During a monetary event, the central bank or monetary authority may deliberately or inadvertently release some of its private information about economic conditions. The monetary policy surprise measure constructed from bond market reactions to this information release would then represent those aspects of the central bank's assessment of the economic outlook that had not previously been incorporated into asset prices by market participants. To the extent that the central bank's assessment is more accurate than the market's pre-event assessment, the release of central bank information can, in theory, induce correlations between monetary surprises and economic conditions. These resulting 'contaminated' monetary policy surprises would then be invalid external instruments, leading to biased dynamic causal effect estimates. Variations of this scenario are referred to in the literature as 'central bank information effect', 'Fed information effect', or 'Delphic shocks' (see for instance Nakamura and Steinsson (2018); Cieslak and Schrimpf (2019); Hansen et al. (2019); Jarociński and Karadi (2020); Miranda-Agrippino and Ricco (2021); Andrade and Ferroni (2021); Schmitt-Grohé and Uribe (2024)).

Another rationalisation regarding the puzzling dynamic effect estimates specifically could be that the estimates are not necessarily biased, but instead reflect a real but counterintuitive 'neo-Fisher' effect that is yet to be recognised in the theoretical consensus (Schmitt-Grohé and Uribe, 2024). That is, it may be the case that price puzzle-type impulse response estimates are a true reflection of real macroeconomic dynamics. This rationalisation is partial, in that it does not explain the observed correlations between monetary policy surprises and pre-event measures of economic conditions.

In contrast, Bauer and Swanson (2023a,b) rationalise the puzzling empirical results by observing that there is no reason to think that market participants have perfect knowledge of the central bank's reaction function to news that is observed equally by both central bank and the private sector. They show how the private sector's need to learn the reaction function from central bank actions over time can induce *ex post* correlations between monetary policy surprises and publicly-available pre-event measures of economic conditions. This rationalisation has been referred to as the 'Fed/central bank response' to news mechanism. It is important to note that no claim of *ex-ante* predictability is made: the argument is consistent with the absence of arbitrage opportunities in bond markets. Moreover, the argument does not require assuming that central banks have better information about economic conditions than the private sector.

Bauer and Swanson (2023a,b) illustrate their argument using the following stylised model

$$x_t = \rho x_{t-1} - \theta i_{t-1} + \eta_t \qquad \text{(Output gap)}$$

$$i_t = \alpha_t x_t + \epsilon_t \qquad \text{(Monetary policy rule)}$$

$$\alpha_t = \alpha_{t-1} + u_t \qquad \text{(Time-varying responsiveness)}$$

where $|\rho| < 1$, $\theta \geq 0$, are fixed parameters, and η_t, ϵ_t, and u_t are Gaussian with zero mean and fixed variances. The central bank is assumed to have perfect information about parameters and observes realisations without error. The private sector, whilst observing x_t without error (i.e. there is no private central bank information), has imperfect information about α_t. Each period plays out as follows. First, x_t realises and is observed by the private sector and the central bank. Second, the private sector forms a rational expectation of the interest rate $E[i_t|x_t, H_{t-1}] = E[\alpha_t|H_{t-1}]x_t$, where H_{t-1} is the information set at time $t-1$. Third, the central bank sets and announces the interest rate i_t. The monetary policy surprise is then

$$\begin{aligned} mps_t &\equiv i_t - E[i_t|x_t, H_{t-1}] \qquad (2.1) \\ &= \underbrace{(\alpha_t - E[\alpha_t|H_{t-1}])}_{\neq 0 \text{ if imperfect information}} \; x_t + \underbrace{\epsilon_t}_{\text{exogenous policy shock}} . \qquad (2.2) \end{aligned}$$

Bauer and Swanson note two important implications of this stylised model. Firstly, Equation (2.1) implies that monetary policy surprises in this model are unpredictable ex ante (i.e. "no money can be made")

$$E[mps_t|x_t, H_{t-1}] = 0. \qquad (2.3)$$

Secondly, Equation (2.1) implies that

$$Cov(mps_t, x_t) = \alpha_t - E[\alpha_t|H_{t-1}] \neq 0 \qquad (2.4)$$

unless $\alpha_t = E[\alpha_t|H_{t-1}]$. That is, for there to be no correlation between publicly observed economic conditions and monetary policy surprises, the private sector would have to anticipate the central bank's time-varying responsiveness α_t perfectly.

In reality the set of publicly-observed measures of economic conditions is high-dimensional rather than scalar, with x_t being a large and time-varying vector of variables the central bank may consider in its decision-making. For example, Bernanke (2024) summarises:

> *'[T]raditional forecasting methods are increasingly being supplemented by methods based on new technologies or data sources. Many central banks already make use of large data sets ('big data'), such as (anonymised) credit card or mortgage records, to get more timely and granular information about the state of the economy. During the pandemic, many central bank staffers (including at the Bank of England) consulted closely with epidemi-*

> *ologists and other public health professionals to better understand Covid-19's economic consequences. Artificial intelligence tools, which can extract information from immense bodies of qualitative and quantitative data, seem certain to be increasingly important for monitoring the economy and forecasting in the future. Central banks are already preparing for that eventuality.'*

As such, it seems unlikely that $Cov(mps_t, x_t) = 0$ for all x_t that the central bank may consider in its decision-making in the real world. In other words, real-life monetary policy surprises may well be correlated with public pre-event information, given that the private sector is unlikely to have full information about the central bank's time-varying reaction function. Conventional monetary policy surprises would, in this scenario, not be valid external instruments for SVAR-IV or LP-IV estimation of the dynamic causal effects of monetary policy shocks. This is due to them not satisfying Stock and Watson (2018)'s contemporaneous exogeneity requirement—a necessary condition for identification.

2.2.4 Implications of theoretical explanations for dynamic causal effect estimation

Given their diagnosis, Bauer and Swanson propose projecting conventional monetary policy surprises on information that was publicly available prior to the monetary event. That is, they recommend regressing mps_t onto measures of pre-event information X_{t-} and using the residuals (so-called 'orthogonalised monetary policy surprises') as external instruments in the SVAR-IV and LP-IV frameworks

$$mps_t^{\perp} = mps_t - \hat{\alpha} - \hat{\beta}' X_{t-}. \tag{2.5}$$

In Equation (2.5) X_{t-} can be any measures regarding economic conditions that are observed publicly prior to each monetary event and $\hat{\beta}$ are estimated parameters capturing the systematic component of monetary policy. Bauer and Swanson argue that, following this orthogonalisation step, the residuals $mps_t^{\perp}$ ('orthogonalised monetary policy surprises') should be valid external instruments in the SVAR and LP frameworks.

This paper implements the Bauer and Swanson (2023b) recommendation in the UK context. Doing so requires access to public information only. That is, orthogonalisation with respect to internal Bank of England assessments is not required. Implementing this suggestion constitutes an empirical test of the hypotheses discussed in Section 2.2.3. If orthogonalising with respect to public pre-event information is sufficient to resolve 'puzzles' and obtain impulse response estimates that are consistent with theoretical consensus, the central bank information effect may be empirically weak or nonexistent in the UK context. That is, the evidence would suggest that it is market participants' imperfect knowledge of the Bank of England's monetary policy rule rather than private central bank information that explains the empirical puzzles. If instead, orthogonalisation with respect to public information is insufficient, that would lend support to the central bank information hypothesis, suggesting that during monetary events the Bank of England may be revealing its 'inside information' about the state of the economy.[5]

In implementing the Bauer and Swanson (2023b) recommendation, I aim to capture publicly-accessible pre-event information about UK economic conditions as comprehensively as possible. To do so, I assemble a corpus of textual documents that relate specifically to Bank of England monetary events. This set of newswire documents is then analysed systematically by a large language model to extract measures of economic conditions as they were just prior to each monetary event. These measures (and lags thereof) are then used as feature set X_{t-} to project mps_t onto, in order to construct orthogonalised monetary policy surprises for the UK. In the next section, both data

[5] This test assumes that the true impulse responses contain no 'puzzles'. For example, it assumes that the neo-Fisher effects discussed in Schmitt-Grohé and Uribe (2024) are empirically weak or nonexistent.

collection and measurement of potential arguments to the Bank of England's reaction function, X_{t-}, are described.

2.3 Measuring economic conditions prior to UK monetary events from textual data using large language model inference

The Bank of England controls interest rates directly through a number of formal tools, principally the Bank Rate, quantitative easing (QE), and quantitative tightening (QT).[6] In addition, it can influence interest rate markets by releasing information that may cause market participants to update their beliefs regarding the future path of these tools ('forward guidance'). Braun et al. (2025) compile a list of Monetary Policy Committee (MPC) decision announcements as well as significant non-MPC events since the Bank of England became operationally independent of the British government in 1997. I refer to this set of events as 'monetary events' throughout this paper. Furthermore, the Braun et al. (2025) database includes measurements of conventional—in the sense of Bauer and Swanson (2023b)—monetary policy surprises mps_t, obtained by observing high-frequency reactions of various financial markets, such as UK gilt yields, during a tight time window around each monetary event.

In this section, I aim to measure economic conditions relevant to the Bank of England's implicit monetary policy rule as they were just before each UK monetary event between January 1997 and June 2024. The purpose of doing so is to obtain a set of predictors X_{t-} that capture information that is potentially relevant to mps_t. Having obtained this set of variables, I am then able to generate orthogonalised monetary policy surprises $mps_t^{\perp}$. That is, the aim of measuring economic conditions is to enable the creation

[6] https://www.bankofengland.co.uk/monetary-policy/quantitative-easing

of external instruments that are as exogenous as possible with respect to pre-event information, in order to address the concerns around instrument validity discussed in Section 2.2.

To measure economic conditions just before each UK monetary event, I harness a large language model to extract and structure information contained in unstructured textual data. The textual data gathered and analysed in this study are intraday newswire articles. These articles are succinct summaries of economic conditions and events composed for market participants and disseminated at a high frequency via subscription services such as Bloomberg Terminal or Refinitiv Eikon. Appendix A.1 describes how a corpus of these newswires was assembled for the 398 UK monetary events contained in the Braun et al. (2025) database. Having collected these textual data, I use a large language model to take measurements of the perceived UK economic conditions and outlook, as described in each article just prior to each monetary event. The appeal of using an LLM for this 'concept measurement' task[7] is that doing so can be seen as less labour-intensive, more consistent, and more replicable than manual human labelling (Korinek, 2023).

I implement concept measurement of pre-event perceptions about the UK economy by prompting the language model[8] to answer a set of multiple choice questions about each newswire article. Prompts are targeted at extracting information about economic conditions that may be arguments of the Bank of England's implicit monetary policy reaction function. In particular, I prompt the language model to capture three distinct facets of perceived UK economic conditions. The first set of prompts is designed to capture whether the readings or outturns of UK economic indicators were surprisingly high, surprisingly low, or as expected. For example, if a recent unemployment figure was perceived to be surprisingly high, the aim is to capture this perceived surprise from raw newswire text and structure the information using

[7] As defined in (Ash and Hansen, 2023)

[8] Specifically, a 46.7 billion parameter model called 'Mixtral-8x7B-Instruct-v0.1'—see `https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1` for model details.

a multiple-choice answer format. The second set of prompts aims to capture perceived changes in risks to UK economic activity. If the newswire text included a discussion regarding the impact of geopolitical developments on supply chains, for instance, there is a prompt that aims to capture this perception encoded in raw text. The third set of prompts aims to capture the overall economic context for the upcoming Bank of England communication, such as whether the event was perceived to be more highly-anticipated than usual or taking place during an immediate crisis. The model's responses covering perceived surprises in pre-event economic indicator readings, perceived changes in economic risks, and the perceived overall economic context are then treated as discrete measurements of pre-monetary event economic conditions. The multiple choice format chosen means that these discrete measurements are of low cardinality, with there being either two or three possible answers per prompt.

2.3.1 Measuring perceived surprises in pre-event economic indicator readings

The UK economic calendar includes release dates for a wide range of economic indicators, such as retail sales, the consumer price index, and the unemployment rate. While time series on these data are readily available in a structured format, indicator values in themselves do not indicate how surprising readings were. This section discusses how a large language model can be prompted to measure the perceived newsworthiness of UK economic indicator readings. Prompts are engineered with the objective of turning unstructured textual data into structured categorical information. In particular, the prompts used are constructed using the following pattern

> "Does the newswire state that the most recent {type of indicator} reading was HIGHER or LOWER than expected? RESPOND WITH ONE OF THE FOLLOWING PHRASES (in-

> cluding square brackets): [YES, HIGHER] OR [YES, LOWER] OR [NO]".

This prompt pattern begins by making explicit reference to the specific newswire the question relates to. This is to reduce the risk of the model hallucinating a response that is not based on the information provided. The pattern also asks specifically about what the newswire states, rather than requesting reasoning as to what the newswire may imply or what may be deduced from the newswire. The reason for phrasing the question as a basic labelling task rather than an abstract reasoning task is that I found basic labelling to be more reliable during my initial experiments. The pattern is also focused on how the actual reading compared to prior expectations. This is to avoid capturing instances where an economic indicator moved up or down exactly as had been expected. That is, the purpose is to capture how indicator readings were perceived at the time, not the indicator readings themselves. The latter are already available in a structured format, so extracting them from text would not be necessary. In order to maximise the utility of the measurements for subsequent statistical analyses, answers are constrained to be of low-cardinality. Specifically, three possible answers are specified: "[YES, HIGHER]", "[YES, LOWER]", and "[NO]". This set of permissible answers was defined to be mutually exclusive and collectively exhaustive, in order to avoid the language model generating output that cites multiple options or none of the provided options. The specific request to include square brackets is added to reduce variability of responses. Without this specific requirement, the model would sometimes return answers with and sometimes without square brackets, creating the need for manual post-processing. Being as explicit as possible reduces or even eliminates the need for manual postprocessing. I use the prompt pattern to generate measurements regarding the expectedness of pre-monetary event readings of the UK's inflation rate, GDP growth, purchasing manager index, unemployment rate, wage growth, consumer confidence index, business confidence index, retail sales, trade balance, mortgage approvals, house price index, public sector borrowing, exchange rate, as well as the volatility index. The exact phrases used to prompt the language model to extract perceived surprises in the readings of these economic indicators are displayed in Table 3 of Appendix A.2, corresponding to prompts with numbers 1 to 14.

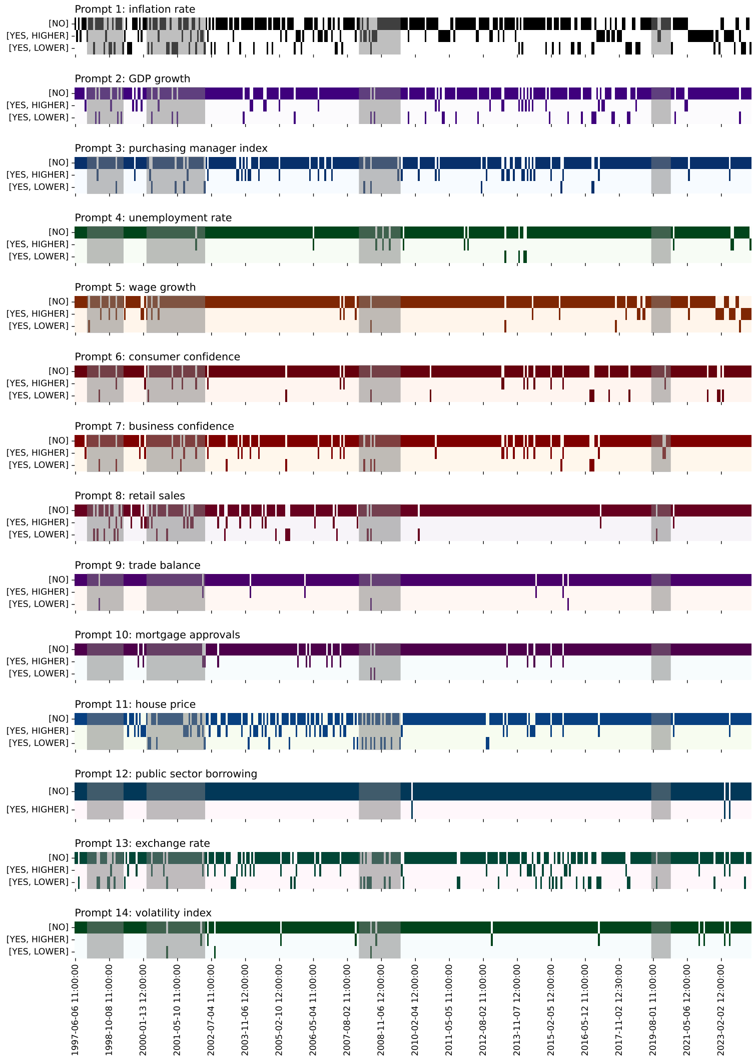
The resulting measurements are presented in Figure 1. Figure 2 summarises the observed statistical associations among the different discrete variables, measured using Cramér's V. It shows that, intuitively, there are strong pairwise associations between business confidence and consumer confidence, business confidence and the purchasing manager index, mortgage approvals and house prices, mortgage approvals and business confidence, and public sector borrowing and the volatility index. Since accuracy and the potential for hallucination are real concerns when using an LLM, a detailed discussion and validation of the measurements obtained is provided in Appendix A.3.1.

2.3.2 Measuring perceived pre-event changes in economic risks

In addition to perceived surprises in recent indicator readings, measurements regarding perceived changes in economic risks the UK economy was exposed to at each point in time are extracted from the newswire corpus in a similar fashion. As before, prompts are engineered with the objective of turning unstructured textual data into structured categorical information. For perceived changes in economic risks, the prompts used are constructed in the following pattern

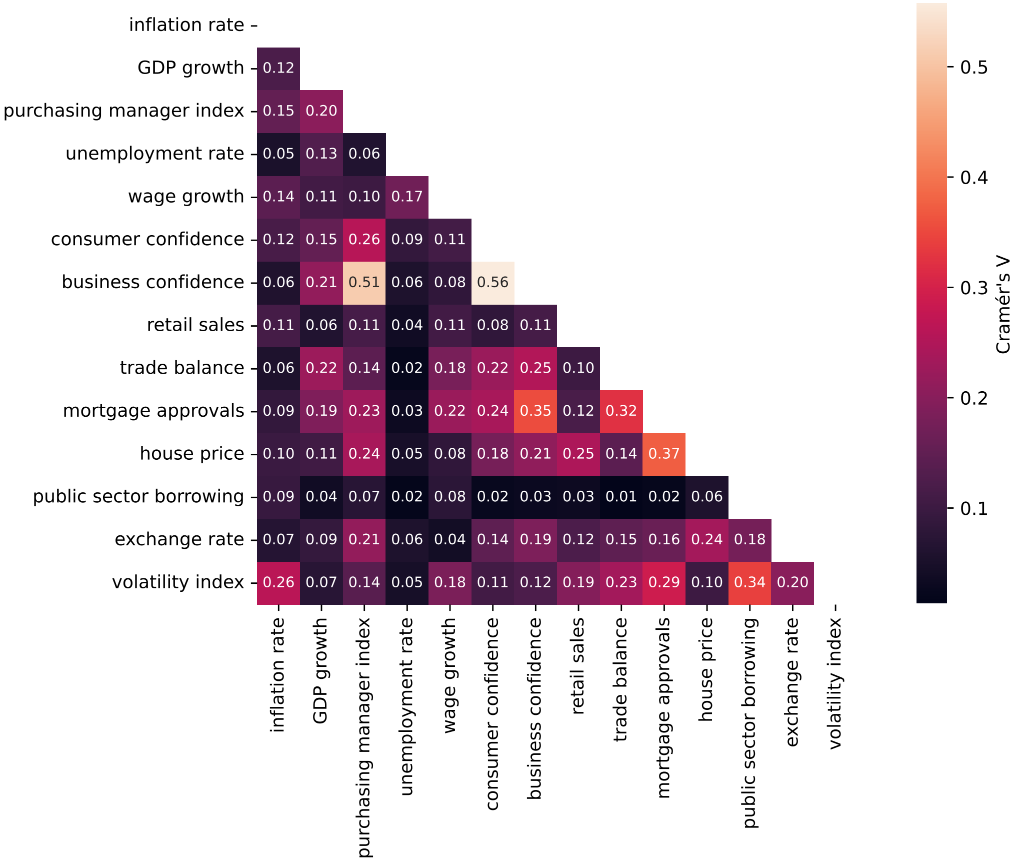
> "Does the newswire state that {type of risk} risk INCREASED or DECREASED recently? RESPOND WITH ONE OF THE FOLLOWING PHRASES (including square brackets): [YES, INCREASED] OR [YES, DECREASED] OR [NO]".

Figure 1: Measured perceived surprises in pre-event economic indicator readings



Notes: All events contained in the database compiled by Braun et al. (2025) are arranged in chronological order, covering observations from January 1997 to June 2024 on the horizontal axis. For each event, answers shown were generated by prompting a large language model with the queries listed in Table 3 to measure information contained in a relevant pre-event public newswire. Grey shaded time periods are UK "peak-to-trough" recessions based on the Federal Reserve Bank of St. Louis's Recession Indicators Series (`https://fred.stlouisfed.org/series/GBRRECDM`).

Figure 2: Associations among measured perceived surprises in pre-event economic indicator readings

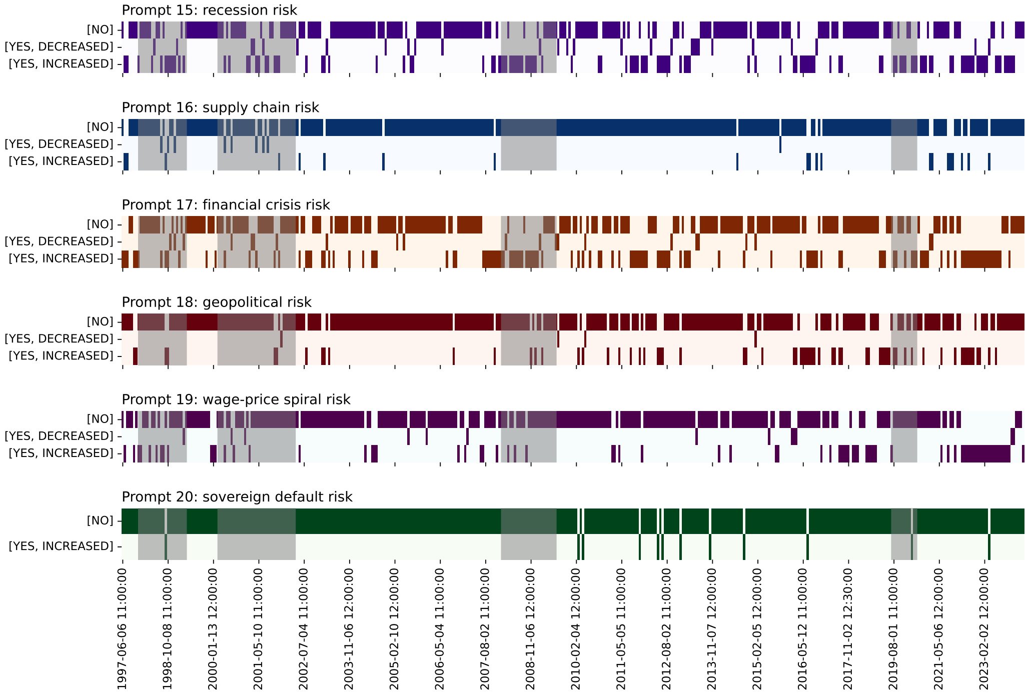


Notes: For each pair of prompts regarding pre-event economic indicator readings, Cramér's V is shown based on discrete prompt responses regarding events between January 1997 and June 2024. Cramér's V can take values between 0 and 1 with higher values indicating a stronger statistical association. Prompts used are detailed Table 3.

Similarly to indicator-related measurements, the prompt pattern regarding risks makes explicit reference to the specific newswire article in order to reduce the risk of the LLM hallucinating a response that is not based on the specific text provided. The pattern also follows the approach of asking specifically about what the newswire states, as opposed to what it might imply, in order to maximise reliability. The prompt is also focused on recent changes in perceived risk of each type rather than levels of risk. This is done to introduce a focus on perceived risk trends just prior to each monetary event. The pattern also constrains answers to be of low-cardinality to maximise the utility of the measurements for subsequent statistical analyses. The three possible answers specified are "[YES, INCREASED]", "[YES, DECREASED]", and "[NO]". As before, this set of permissible answers was also defined to be mutually exclusive and collectively exhaustive and the output is requested to include square brackets to reduce variability of response formatting. The particular risk types considered for this measurement exercise were chosen to be most relevant to business cycle fluctuations and monetary policy decisions. In particular, measurements regarding recession risk, supply chain risk, financial crisis risk, geopolitical risk, wage-price spiral risk, and sovereign default risk are taken from newswire text using the large language model prompts. The exact phrases used to prompt the language model are displayed in Table 4 of Appendix A.2, corresponding to prompts numbered 15 to 20.

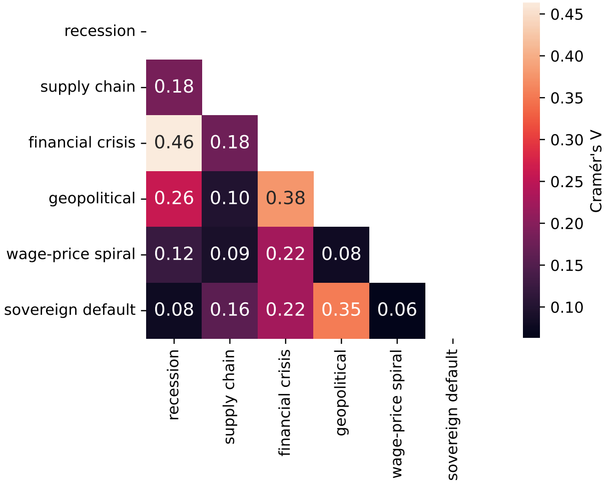
The resulting measurements are presented in Figure 3. Figure 4 displays the observed statistical associations (again using Cramér's V) among the different discrete variables measured. Figure 4 shows that Cramér's V is highest for the pairwise associations between financial crisis risk and recession risk, financial crisis risk and geopolitical risk, and geopolitical risk and sovereign default risk. A detailed discussion and validation of the series of measurements in Figure 3 is provided in Appendix A.3.2.

Figure 3: Measured perceived changes in economic risks



Notes: All events contained in the database compiled by Braun et al. (2025) are arranged in chronological order, covering observations from January 1997 to June 2024 on the horizontal axis. For each event, answers shown were generated by prompting a large language model with a subset of the queries listed in Table 4 to measure information contained in a relevant pre-event public newswire. Grey shaded time periods are UK "peak-to-trough" recessions based on the Federal Reserve Bank of St. Louis's Recession Indicators Series (https://fred.stlouisfed.org/series/GBRRECDM).

Figure 4: Associations among measured perceived changes in economic risks



Notes: For each pair of prompts regarding pre-event changes in economic risks, Cramér's V is shown based on discrete prompt responses regarding events between January 1997 and June 2024. Cramér's V can take values between 0 and 1 with higher values indicating a stronger statistical association. Prompts used are detailed Table 4.

2.3.3 Measuring the perceived overall pre-event economic context

The newswire corpus and language model prompts are also used to measure the perceived overall economic context prior to each monetary event in the Braun et al. (2025) database. The set of prompts used is generated using the following pattern

> "Does the newswire state that the upcoming Bank of England communication was {aspect of perceived context}? RESPOND WITH ONE OF THE FOLLOWING PHRASES (including square brackets): [YES] OR [NO]".

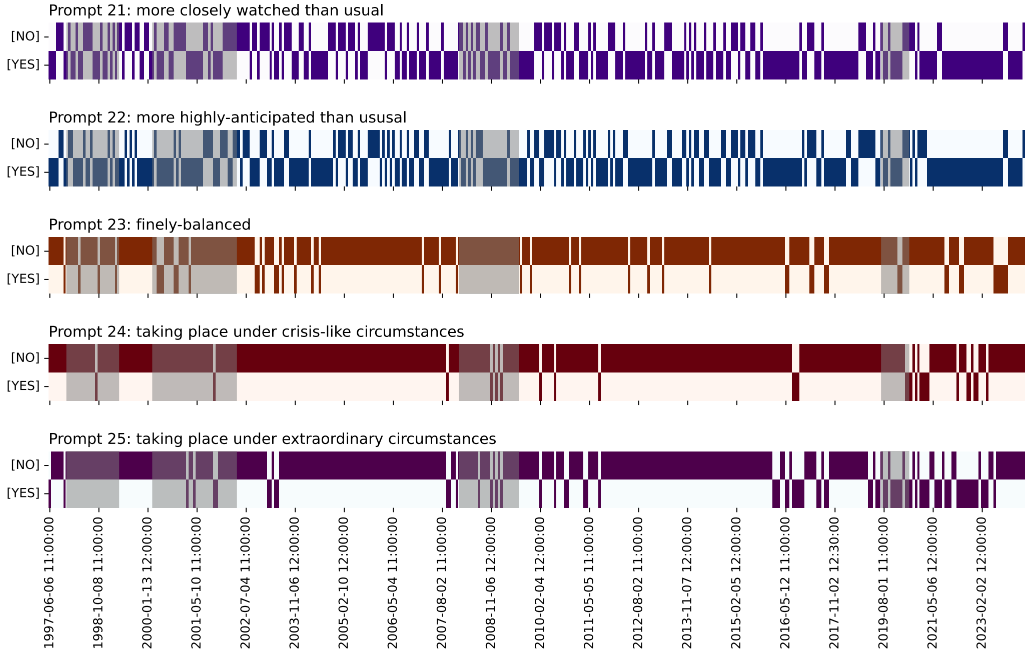
As was done for the other two prompt patterns, the question makes explicit reference to the specific newswire, asks specifically about what the newswire states, and constrains answers to be of low-cardinality to maximise the utility of the measurements for subsequent statistical analyses. A difference compared to the previous two patterns is that in this case there are only two possible answers, specifically "[YES]" and "[NO]". The pattern makes explicit reference to "the upcoming Bank of England communication", which is deliberately worded in a general way to be appropriate for all types of monetary events covered in the Braun et al. (2025) database such as MPC decisions, press conferences, and Inflation Report publications. The first aspect of the economic context that is measured is whether the upcoming monetary policy was perceived to be more closely watched than usual. Similarly, the second set of measurements relates to whether the event was perceived to be more highly-anticipated than usual. The third set of measurements aims to capture whether the upcoming monetary decision/communication was finely-balanced. Finally, further measurements aim to capture whether the monetary event was taking place during a crisis or other extraordinary circumstances. The exact phrases used to prompt the language model to extract aspects of the perceived overall economic context around each monetary event are displayed in Table 4 of Appendix A.2, corresponding to prompts with numbers 21 to 25.

Figure 5 shows a time series of the resulting binary measurements, while Figure 6 shows the pairwise Cramér's V-associations between all pairs of variables. Among the measures for perceived overall economic context, Cramér's V is highest for the pairwise associations between the two indicators for an event being more 'closely-watched' and 'highly-anticipated' than usual, the two indicators for an event being perceived to take place under 'crisis-like' and 'extraordinary' circumstances, and the two indicators for an event being perceived to be more closely watched and to be taking place under extraordinary circumstances. As is the case for the other two sets of prompts, a detailed discussion and validation of the series of measurements in Figure 5 is provided in Appendix A.3.3.

2.4 Constructing orthogonalised UK monetary policy surprises

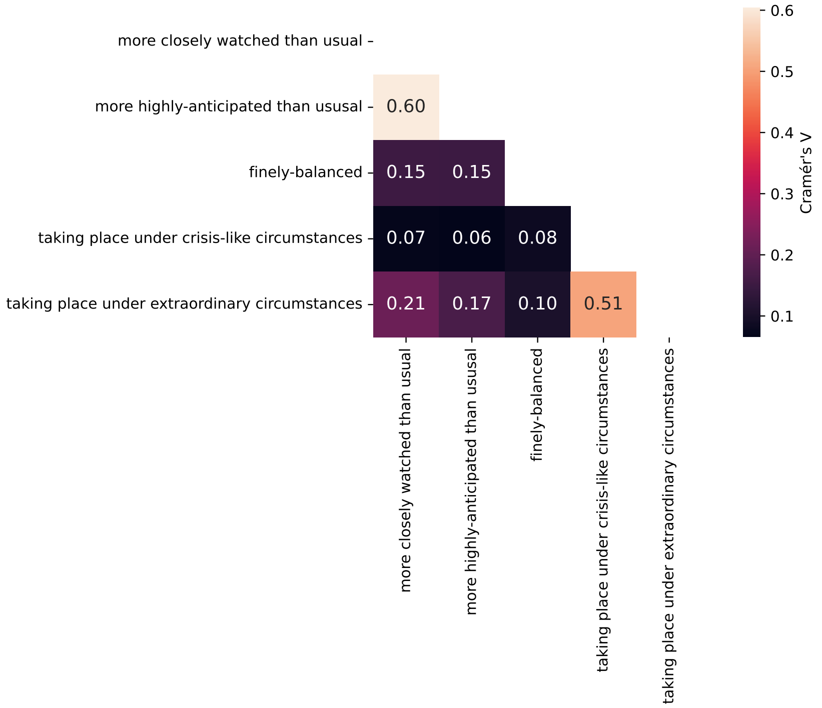
Having created text-derived measures of pre-event UK economic conditions in Section 2.3, Section A.4 of the appendix establishes that the information extracted from newswire text is correlated—ex post—with financial markets' high-frequency reactions to Bank of England communications collected by Braun et al. (2025). This includes yield curve movements, which are conventionally used as monetary policy surprises for use in high-frequency identification strategies to estimate the dynamic causal effects of monetary policy. Given this finding, the Braun et al. (2025) high-frequency yield curve reactions may not satisfy Stock and Watson (2018)'s contemporaneous exogeneity requirement — a necessary condition for identification. Using them as external instruments for SVAR-IV or LP-IV estimation of the dynamic causal effects of monetary policy shocks could therefore result in biased inferences.

Figure 5: Measured perceived overall economic context prior to monetary policy communications



Notes: All events contained in the database compiled by Braun et al. (2025) are arranged in chronological order, covering observations from January 1997 to June 2024 on the horizontal axis. For each event, answers shown were generated by prompting a large language model with a subset of the queries listed in Table 4 to measure information contained in a relevant pre-event public newswire. Grey shaded time periods are UK "peak-to-trough" recessions based on the Federal Reserve Bank of St. Louis's Recession Indicators Series (`https://fred.stlouisfed.org/series/GBRRECDM`).

Figure 6: Associations among measured perceived overall economic context prior to monetary policy communications



Notes: For each pair of prompts regarding pre-event economic context, Cramér's V is shown based on discrete prompt responses regarding events between January 1997 and June 2024. Cramér's V can take values between 0 and 1 with higher values indicating a stronger statistical association. Prompts used are detailed Table 4.

This section implements the recommendation of Bauer and Swanson (2023a) and Bauer and Swanson (2023b), by orthogonalising conventional monetary policy surprises curated by Braun et al. (2025) with respect to the text-derived measures of pre-event UK economic conditions created in Section 2.3. This 'projecting out', as detailed in Equation (2.5) is performed with the aim of constructing valid external instruments to be used within the SVAR-IV and LP-IV dynamic causal effect estimation frameworks. In doing so, there are some choices as to how Equation (2.5) is operationalised. Firstly, of the high-frequency interest rate reactions collated by Braun et al. (2025), gilt yields are available for the longest timeframe and are therefore used henceforth. Secondly, high-frequency movements during event windows are available for several maturities along the gilt yield curve. As a result there is a choice to be made as to how external instruments are constructed from movements of interest rates across the yield curve. Braun et al. (2025) employ a factor approach, where assumptions are made to identify 'Target', 'Path', and 'QE' factors of UK monetary policy. This study departs from the factor approach and instead treats distinct yield movements of each gilt maturity (i.e. 1, 2, 5, and 10-years) as separate external instruments. Where necessary, the conventional monetary policy surprise series are distinguished notationally as $mps1_t$ for high-frequency movements in the 1-year gilt yield, $mps2_t$ for high-frequency movements in the 2-year gilt yield, $mps5_t$ for high-frequency movements in the 5-year gilt yield, and $mps10_t$ for high-frequency movements in the 10-year gilt yield. The maturities of orthogonalised monetary policy surprise series are identified analogously. Thirdly, there are different approaches to estimate $\hat{\beta}$ within Equation (2.5), which are discussed in Section 2.4.1.

In addition to implementing orthogonalisation as recommended in Bauer and Swanson (2023a) and Bauer and Swanson (2023b), this section also implements the so-called 'poor man's sign restrictions'-approach proposed in Jarociński and Karadi (2020) to isolate the exogenous component of monetary policy surprises. This section concludes by comparing the time series of the resulting set of candidate external instruments. In the later sections, the resulting orthogonalised monetary policy surprises are then used as external instruments for SVAR-IV or LP-IV estimation of the dynamic causal effects of monetary policy shocks on the UK economy.

2.4.1 Orthogonalising conventional UK monetary policy surprises with respect to LLM-measured pre-event economic conditions

Bauer and Swanson (2023b)'s recommendation to orthogonalise conventional monetary policy surprises with respect to pre-event information can be stated generally as

$$mps_t^{\perp} = mps_t - \hat{f}(X_{t-}) \tag{2.6}$$

where X_{t-} is the pre-announcement information set.

A key decision when implementing (2.6) concerns the choice of functional form and estimation method for $\hat{f}(X_{t-})$. In what follows, a linear functional form is assumed. That is, the target of estimation is $f(.) = X_{t-}\beta$. It should be noted that, in general, one can consider non-linear functional forms for $f(.)$ as done in Aruoba and Drechsel (2024), although doing so is beyond the scope of this study. That said, the approach taken in this study is not strictly linear, in that the LLM-prompting step involves applying highly non-linear transformations to raw textual data. The dummy variables resulting from this non-linear transformation are then modelled as having a linear relationship with high-frequency gilt yield movements.

A second choice relates to the operational definition of the information set X_{t-}. For each monetary event, there are measures of UK economic conditions as of the day of the event. However, the lagged measures from previous events may also carry information about mps_t. In addition to functional form, there is therefore a choice about the number of lags that X_{t-} is defined to include.

There are also different ways of estimating β. The most common approach is to perform standard linear regression with parameter vector β estimated through ordinary least squares (OLS)/maximum likelihood

$$mps^{\perp}_{t,OLS} = mps_t - X_{t-}\beta_{OLS}. \tag{2.7}$$

Alternatively, one can estimate the parameters with regularised regression. Different options for regularisation include regularised least squares with ridge/L2 penalty

$$mps^{\perp}_{t,ridge} = mps_t - X_{t-}\beta_{ridge} \tag{2.8}$$

and regularised least squares with lasso/L1 penalty

$$mps^{\perp}_{t,lasso} = mps_t - X_{t-}\beta_{lasso}. \tag{2.9}$$

The different specifications used for orthogonalisation are now discussed in turn.

OLS regression

The simplest approach is to use a linear functional form and estimate it with OLS as follows

$$\hat{\beta}_{OLS} = \underset{\beta}{argmin}\{\sum_{t}(mps_t - \alpha - X_{t-}\beta)\}. \tag{2.10}$$

When doing so, there is a limit as to how many lags can be included in the explanatory variable set X_{t-} before the number of variables exceeds the number of observations. For this reason, in what follows, OLS specifications are estimated with lagged predictors up to four lags included in the set of explanatory variables. The estimation results using OLS are presented in Appendix A.5.1. The results suggest that across gilt maturities and lag specifications, text-derived measures of higher-than-expected unemployment rate readings and lower-than-expected consumer confidence readings are among the most significant ex post predictors of "dovish"—i.e. negative—yield curve reactions to UK monetary events. Higher-than-expected public sector borrowing and business confidence readings are significantly predictive of "hawkish"—i.e. positive—yield curve reactions.

Regularised regression

One way to make the inclusion of additional lags feasible is to use regularised least squares with a ridge (or "L2") penalty to estimate the linear model parameters

$$\hat{\beta}_{ridge} = \underset{\beta}{argmin}\{\sum_{t}(mps_t - \alpha - X_{t-}\beta)\} \tag{2.11}$$

subject to

$$||\beta||_2 \leq \alpha_r \tag{2.12}$$

where $||\beta||_2 = [\sum(|\beta_i|^2)]^{\frac{1}{2}}$ is the L2-norm and α_r is an arbitrary parameter that determines the amount of L2-regularisation. As per standard practice, the predictor matrix X_{t-} is standardised prior to estimation.

Another way to make the inclusion of additional lags feasible is to use the lasso (or "L1" penalty to estimate linear parameters as follows

$$\hat{\beta}_{lasso} = \underset{\beta}{argmin}\{\sum_{t}(mps_t - \alpha - X_{t-}\beta)\} \tag{2.13}$$

subject to

$$||\beta||_1 \leq \alpha_l \tag{2.14}$$

where $||\beta||_1 = [\sum_i(|\beta_i|^1)]^{\frac{1}{1}}$ is the L1-norm and α_l is an arbitrary parameter that determines the amount of L1-regularisation. As was done for ridge regression, the predictor matrix is standardised prior to estimation.

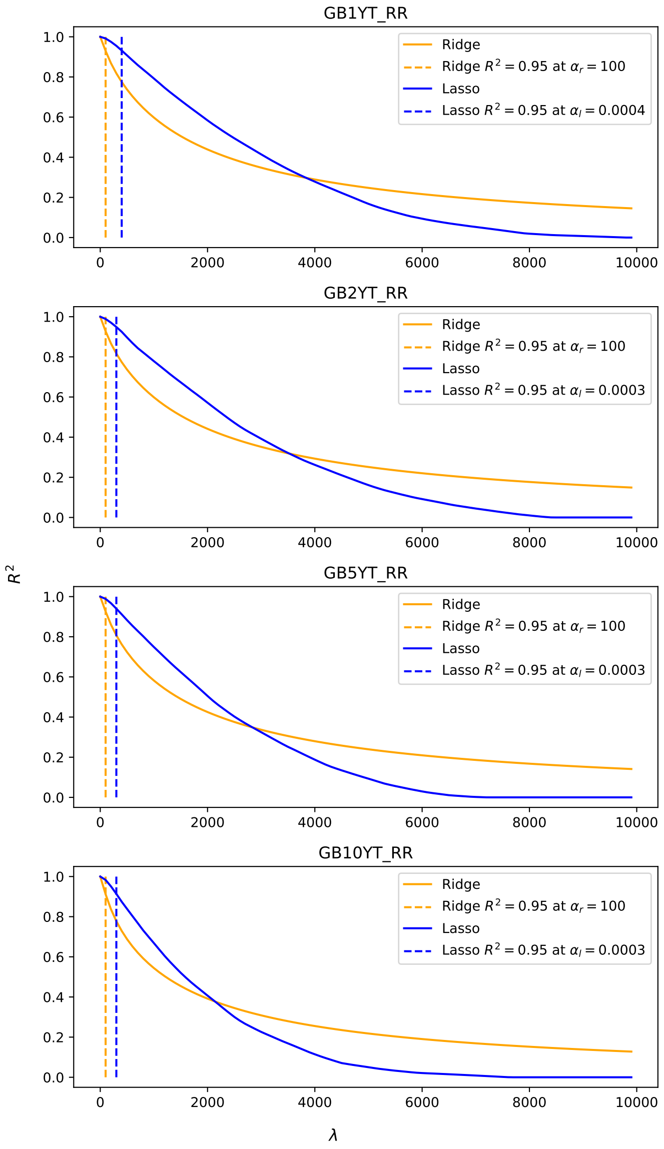
For both regularised regression specifications, X_{t-} is defined to include

15 lags, corresponding to approximately 12 months' worth of preceding monetary events. This is aligned with the standard lag length used in vector autoregressions. As a result of this lag-length choice, there are 688 independent variables and associated parameters to estimate. That is, there are many more predictors than observations—the models are overparameterised. A consequence of overparameterisation is that the choice of regularisation parameter (α_r and α_l) implies a choice about the R-squared of the estimated model. Figure 7 illustrates this implication: depending on the strength of the regularisation penalty (λ), both specifications can explain an arbitrarily high or low share of the variance in gilt yield reactions to monetary events. In what follows, I choose α_r and α_l such that R-squared = 0.95. That is, the regularisation penalty is chosen to be relatively small resulting in model parameters being 'allowed' to explain almost all the variance in gilt yield reactions. The rationale behind applying only light regularisation (or, equivalently, selecting a high R-squared) is an argument in Ramey (2016)

> "Monetary policy is being conducted more systematically, so true monetary policy shocks are now rare".

In particular, Ramey argues that since monetary policy is typically based on monetary policy rules, a large share of variation in the use of policy tools should be expected to be due to the *systematic* component of monetary policy—i.e. the component that ought to be removed from external instrumental variables to be left with exogenous variation in policy. It is hoped that regularising the ridge and lasso regressions lightly and attaining a high R-squared, minimises the likelihood that the remaining variation in text-orthogonalised monetary policy surprises is contaminated by variations due to systematic (i.e. endogenous) monetary policy variations. An alternative approach would be to follow Aruoba and Drechsel (2024), who select α_r and α_l through cross-validation, such that the out-of-sample prediction error is minimised. This involves selecting the amount of regularisation that balances the bias-variance trade-off. For the purpose of prediction error-minimisation,

Figure 7: Selection of regularisation parameters for regularised linear regression



Notes: Each panel shows the relationship between model R-squared (y-axis) and the amount of regularisation (λ). One panel per gilt maturity (1, 2, 5, and 10-years). $\alpha_r = \lambda$ and $\alpha_l = \lambda/1,000,000$.

some amount of bias in the estimation of $\hat{f}(X_{t-})$ can be acceptable in order to reduce the amount of estimation error in the parameters of $\hat{f}(X_{t-})$. It is worth noting, however, that the purpose of estimating $\hat{f}(X_{t-})$ in Equation (2.6) is not to minimise the out-of-sample prediction error. Instead, the rationale for orthogonalisation is to eliminate the in-sample covariance between mps_t and X_{t-}. As such, there is a risk that determining the amount of regularisation through cross-validation would result in under-orthogonalisation due to over-regularisation. An analysis of how orthogonalisation using regularised regression differs from that using OLS regression is provided in Appendix A.5.2.

Key predictors of conventional 'surprises'

In line with the findings of the prompt-by-prompt analysis in the appendix' Section A.4, Appendix A.5 demonstrates that a large share of monetary policy "surprises" is ex post predictable using pre-event information. For instance, OLS analysis shows that a significant share of variation in "dovish surprises" is explained by deteriorating economic conditions such as higher-than-expected unemployment rate readings or lower-than-expected consumer confidence readings. Moreover, higher-than-expected public sector borrowing before a monetary event is associated with a "hawkish" surprise afterwards, suggesting that market participants may have underestimated the extent to which the Bank of England responds to expansionary fiscal policies. Higher-than-expected business confidence is also significantly predictive of "hawkish" yield curve reactions. A comparison of Shapley values across OLS, ridge, and lasso specifications reveals that that measures regarding recession risk are the highest-ranked predictor group in nearly every specification, with financial crisis risk also being consistently ranked highly. Measures of perceived surprises in inflation and GDP readings are also shown to be consistently relevant predictor variables.

2.4.2 "Poor man's sign-restrictions"

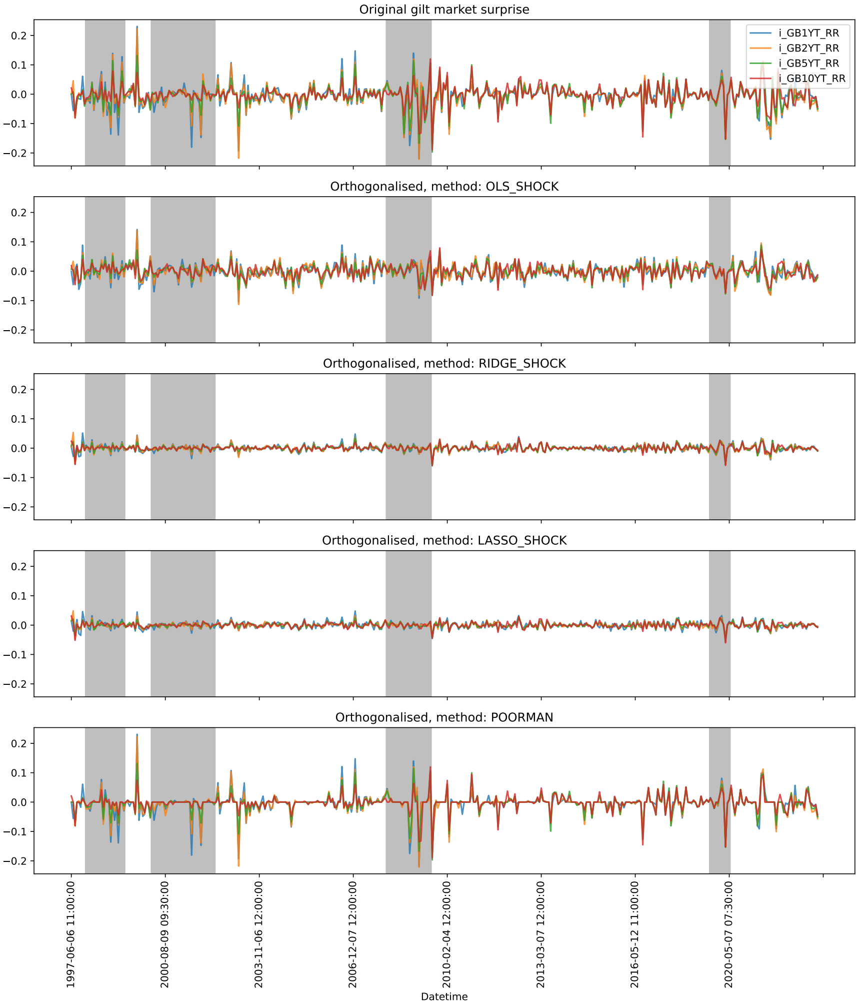
As a benchmark method to address potential endogeneity in conventional monetary policy surprises, I also implement the so-called 'poor man's sign restrictions' proposed in Jarociński and Karadi (2020), which identify exogenous monetary policy shocks by requiring a sign-restriction that mps_t and a broad equity market index move in opposite directions during the measurement window of each event. That is, if during a monetary event the gilt yield goes up and the FTSE250 goes down that would be considered a contractionary monetary policy shock. For all events where the sign on both reactions is equal, the value of the 'orthogonalised' instrument $mps^{\perp}_{t,poorman}$ is set to 0. The hypothesis behind this restriction is that these must be events where information released about economic conditions dominated any monetary policy shock.

2.4.3 Orthogonalised UK monetary policy surprises

For each of the four gilt maturities, I construct five different series to use as external instruments for dynamic causal effect estimation. A time series plot of each of these 20 series is visible in Figure 8.

The first panel shows the mps_t series, the original high-frequency reactions measured by Braun et al. (2025). Second, there are the residuals from OLS regressions of original high-frequency reactions on text-derived variables and four lags thereof, denoted by $mps^{\perp}_{t,OLS}$. The series in the third panel are constructed by taking the residuals from ridge regressions of original high-frequency reactions on text-derived variables and 15 lags thereof, denoted by $mps^{\perp}_{t,ridge}$. As discussed, when generating these series, the penalisation parameters are set in such a way that the fitted values account for 95% of variation in the original high-frequency reactions. The fourth panel shows the residuals from lasso regressions $mps^{\perp}_{t,lasso}$ with the same predictors as those in the ridge regression. The penalisation parameters are set analogously to

Figure 8: Time series of candidate external instruments

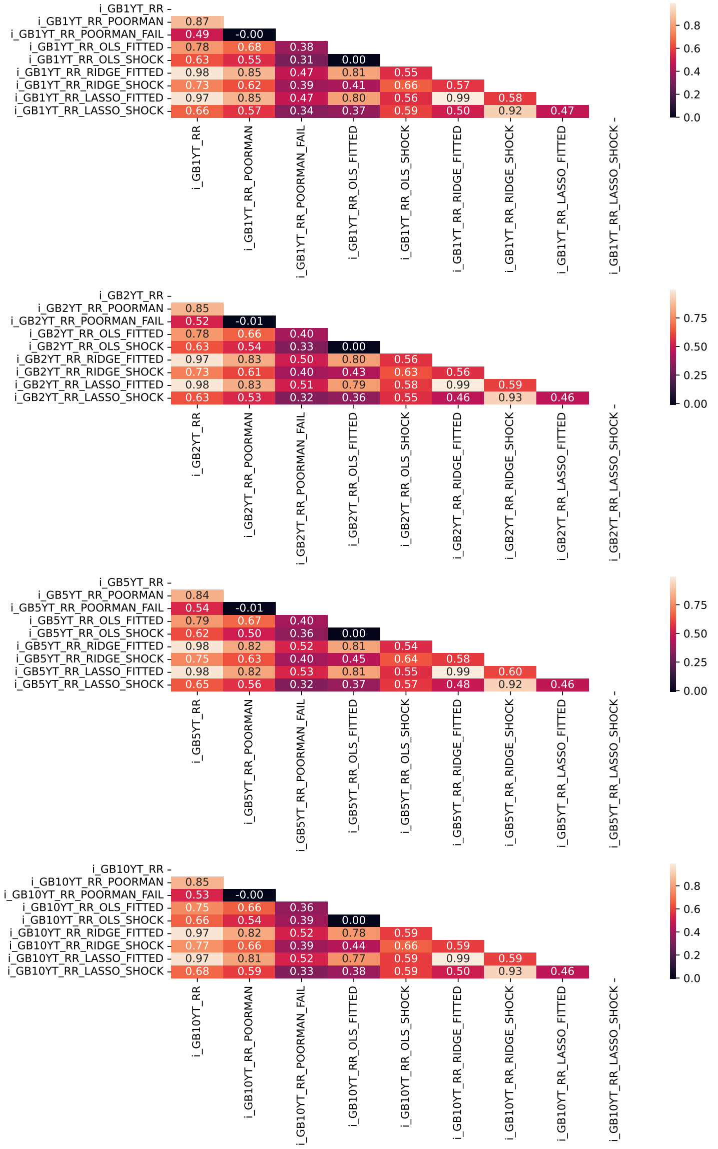


Notes: Each panel displays a group of external instruments based on high-frequency gilt yield movements around UK monetary events, with each group consisting of four instruments with one instrument for each gilt yield maturity (1-year, 2-year, 5-year and 10-year). The top panel shows original Braun et al. (2025) high-frequency reactions, the three middle panels show the residuals from the orthogonalisation procedures described in Section 2.4.1, and the bottom panel shows the reactions after imposing the Jarociński and Karadi (2020) sign restrictions. Grey shaded time periods are UK "peak-to-trough" recessions based on the Federal Reserve Bank of St. Louis's Recession Indicators Series (`https://fred.stlouisfed.org/series/GBRRECDM`).

those of the ridge regressions. The fifth panel shows, $mps^{\perp}_{t,poorman}$, the original high-frequency reactions measured by Braun et al. (2025) subject to the Jarociński and Karadi (2020) 'poor man's sign restrictions'. The equity market index used to implement the sign restriction is the FTSE250, an index containing medium-sized listed companies whose value is more specifically linked to the UK economy compared to the value of larger FTSE100 firms that tend to generate a significant share of revenues in economies other than the UK's.

The unorthogonalised series in the top panel of Figure 8 have the most variance by construction. While parts of the OLS series $mps^{\perp}_{t,OLS}$ in the second panel retain a resemblance to the unorthogonalised series in the first panel, a substantial amount of variation has been removed through orthogonalisation, particularly during recessionary periods. The ridge and lasso series ($mps^{\perp}_{t,ridge}$ and $mps^{\perp}_{t,lasso}$) have lower variance than the OLS series, due to a larger share of variance being explained by these specifications with a larger set of predictors. The series that obtain from imposing the Jarociński and Karadi (2020) sign restrictions in the bottom panel retain most of the largest spikes of the original, unorthogonalised series. To further examine the relationships between these candidate instruments, Figure 9 presents Pearson correlations among these 20 time series for each gilt yield maturity. The figure also includes the fitted values of the OLS, ridge, and lasso regressions, as well as a time series of gilt yield reactions for which the Jarociński and Karadi (2020) condition fails. Interestingly, despite the ridge and lasso residuals only accounting for 5% of the variance of the original series, their correlations with the original reactions mps_t are still high at around 0.6-0.7. The correlations between lasso and ridge series are around 0.9 across all maturities. The lasso series are universally less correlated with the original series.

Figure 9: Correlations among candidate external instruments



Notes: For each gilt yield maturity (1-year, 2-year, 5-year and 10-year) a heatmap of Pearson correlation coefficients among the external instrument series in Figure 8 is diplayed. For the instruments resulting from the orthogonalisation approaches in Section 2.4.1, fitted values of the OLS, ridge, and lasso regressions are included. Similarly, the time series of surprises that do not satisfy the Jarociński and Karadi (2020) sign restrictions (i.e. presumed central bank information shocks) are also included.

2.5 Dynamic causal effects of UK monetary policy on key macroeconomic indicators

This section combines the outputs of the previous sections in order to generate new evidence regarding the dynamic causal effects of Bank of England policies on the UK economy in aggregate. In particular, Section 2.2 discussed how the potential endogeneity of conventional monetary policy surprises can result in invalid inferences about the dynamic causal effects of monetary policy. The section also covers Bauer and Swanson (2023a) and Bauer and Swanson (2023b)'s recommendation as to how endogeneity can be removed through orthogonalisation with respect to pre-event economic indicators. In Section 2.3, I gather and structure—using a large language model— textual data that were created just before each UK monetary event. Section 2.4 uses the structured information generated in Section 2.3 to create orthogonalised monetary policy surprises for the UK. In this section, I estimate the impact of UK monetary policy on key macroeconomic variables and, in doing so, investigate the extent to which the orthogonalisation in Section 2.4 affects dynamic effect estimates. The section also explores the extent to which the choice of statistical orthogonalisation methodology matters.

With regards to the choice of econometric technique, the results of Plagborg-Møller and Wolf (2021) and Li et al. (2024) show that the choice between vector autoregression (VAR) and linear projection (LP) inference about dynamic causal effects comes down to choosing a point along a bias-variance trade-off, with linear projections typically having lower bias but higher variance than VAR estimators. Watson (2023) summarise that a mean square error criterion would typically favour the VAR approach, and that Bayes estimators can result in significantly lower mean squared error than unrestricted estimators. Given the small sample size for UK macro time series since Bank of England independence was granted in 1997, I therefore choose a Bayesian VAR (BVAR) as default technique, as implemented with standard normal-inverse-Wishart priors in Miranda-Agrippino and Ricco (2021).[9] Sensitivity to this choice is explored Section A.6, where BVAR results are compared to Bayesian LP (BLP) estimates. Consistent with the findings of Li et al. (2024), I find BLP estimates to be noisier than BVAR estimates.

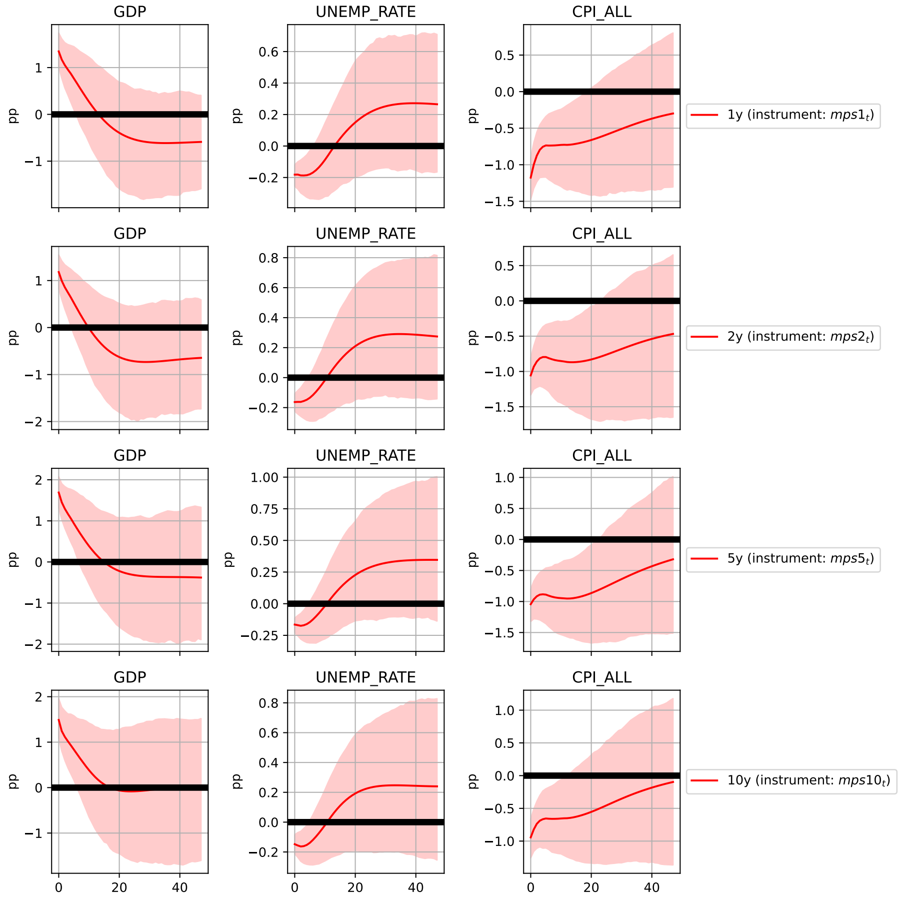
2.5.1 Baseline: identification with conventional monetary policy surprises

As a baseline I estimate four monthly 7-variable BVARs, one for each high-frequency gilt yield reaction series measured by Braun et al. (2025): $mps1_t$, $mps2_t$, $mps5_t$, and $mps10_t$. The set of endogenous variables is the same for all four BVARs, with the exception of the interest rate used for normalisation of the shock. For example, the BVAR where $mps1_t$ is used as the external instrument to identify the monetary policy shock includes the 1-year gilt yield as an endogenous variable. The shock is normalised such that the 1-year yield increases by 100 basis points. Analogously, the 2-year gilt yield is included as endogenous variable and used for normalisation in the BVAR where $mps2_t$ is used as external instrument to identify the monetary policy shock. The same logic applies to the two BVARs for $mps5_t$, and $mps10_t$. As listed in Table 1, the remaining six endogenous variables are GDP, unemployment rate, consumer price index, Favara et al. (2016) excess bond premium, FTSE250 index, and the BIS Sterling broad effective exchange rate index. The BVAR is estimated in levels using 12 lags, after taking the logarithms of variables not already expressed in percentage points (i.e. GDP, consumer price index, FTSE250 index, and the exchange rate). This specification is estimated using all months between 1997M1 and 2019M12.

Figure 10 shows the resulting impulse response functions. Irrespective of the gilt maturity included in the BVAR, there is a significant 'real activity puzzle' in that GDP is estimated to respond positively to a contractionary

[9] I thank Miranda-Agrippino and Ricco for making their estimation procedures available to the research community.

Figure 10: Impulse responses to monetary policy shocks identified using Braun et al. (2025)'s conventional monetary policy surprises



Notes: Monetary policy shocks are identified using external instrumental variables $mps1_t$, $mps2_t$, $mps5_t$, and $mps10_t$ (i.e. conventional gilt yield monetary policy surprises measured by Braun et al. (2025)) and estimated using a monthly 12-lag Bayesian structural vector autoregression (BSVAR-IV) with normal-inverse-Wishart priors as in Miranda-Agrippino and Ricco (2021). There is one row per VAR, with different rows showing results corresponding to different gilt yields used as external instrument and normalisation variable included in the VAR. Monetary policy shocks are normalised such that the corresponding gilt yield increases by 100 basis points. The shaded areas indicate the 90% posterior coverage bands. Variables included in VAR (other those shown and the relevant gilt yield): Favara et al. (2016) excess bond premium, FTSE250 index, and BIS Pound Sterling broad effective exchange rate index. Estimation sample: 1997M1-2019M12.

Table 1: Variables included in BVAR specifications

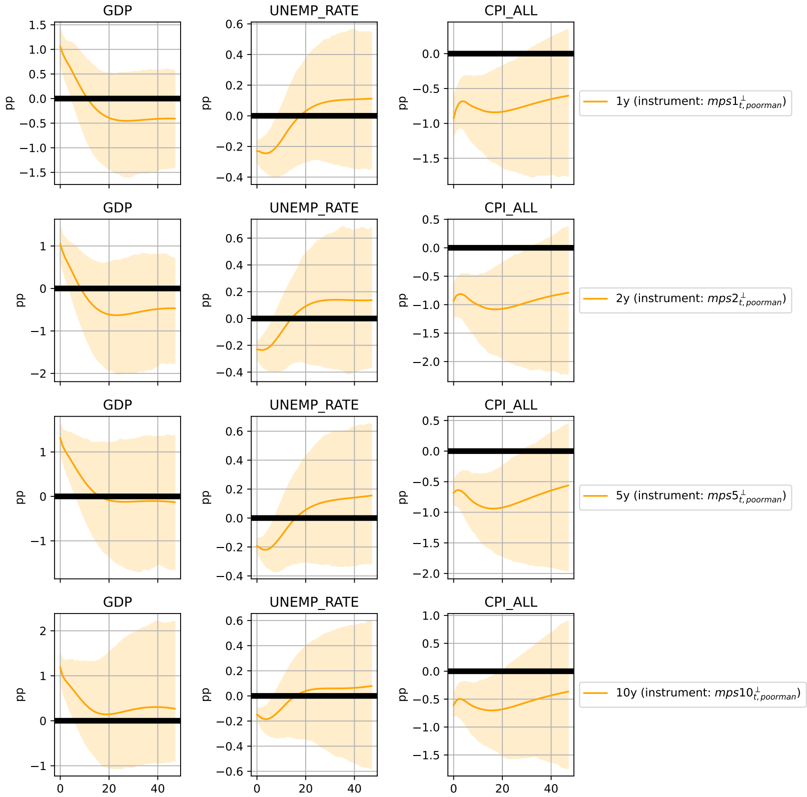
Variable	Description	Source
GB1YT=RR*	1 Year reference gilt yield	Refinitiv
GB2YT=RR*	2 Year reference gilt yield	Refinitiv
GB5YT=RR*	5 Year reference gilt yield	Refinitiv
GB10YT=RR*	10 Year reference gilt yield	Refinitiv
GDP	Monthly GDP and main sectors to 4 decimal places: Monthly GDP (A-T)	ONS
UNEMP_RATE	Unemployment rate (aged 16 and over, seasonally adjusted)	ONS via UK-MD (Coulombe et al., 2021)
CPI_ALL	CPI INDEX 00: ALL ITEMS 2015=100	ONS via UK-MD (Coulombe et al., 2021)
EBP	Gilchrist and Zakrajšek (2012) excess bond premium	Favara et al. (2016)
FTSE250	FTSE250 (ˆFTMC)	YAHOO! via UK-MD (Coulombe et al., 2021)
GBP_BROAD	Monthly average Broad Effective exchange rate index, Sterling (Jan 2005 = 100) (XUMABK82)	BOE via UK-MD (Coulombe et al., 2021)

*depending on specification.

monetary policy shock. Similarly, there is an ‘employment puzzle’ in that the unemployment rate is estimated to respond negatively to a contractionary monetary policy shock. Both of these impact effects are surprising and at odds with theoretical consensus and empirical synthesis regarding monetary non-neutrality. The price level response is negative upon impact but diminishes over time. Given these findings, the endogeneity concerns regarding conventional monetary policy surprise instruments discussed in Section 2.2 may apply in the UK context.

As discussed in Section 2.4, I also include the sign restrictions proposed in Jarociński and Karadi (2020) as a benchmark for existing mitigations of potential endogeneity problems with high frequency external instruments. The estimated impulse response functions when using the sign-restricted monetary policy surprises as instruments are displayed in Figure 11. The results are nearly identical to those resulting from the use of conventional monetary policy surprises. In particular, across all four gilt maturities there is a significant ‘real activity puzzle’ with GDP responding positively to a contractionary shock and an employment puzzle with the unemployment rate responding negatively to a contractionary shock. The magnitude of impact effects for these two variables is marginally smaller compared to the estimates in Figure 10. As such, it appears that applying the Jarociński and Karadi (2020) sign restrictions does not result in estimated impulse responses that are aligned with prior expectations.

Figure 11: Impulse responses to monetary policy shocks identified using Jarociński and Karadi (2020) sign restrictions applied to Braun et al. (2025)'s conventional monetary policy surprises



Notes: Monetary policy shocks are identified using external instrumental variables $mps1^{\perp}_{t,poorman}$, $mps2^{\perp}_{t,poorman}$, $mps5^{\perp}_{t,poorman}$, *and* $mps10^{\perp}_{t,poorman}$ *(i.e. conventional monetary policy surprises measured by Braun et al. (2025) subject to the Jarociński and Karadi (2020) sign restrictions) and estimated using a monthly 12-lag Bayesian structural vector autoregression (BSVAR-IV) with normal-inverse-Wishart priors as in Miranda-Agrippino and Ricco (2021). There is one row per VAR, with different rows showing results corresponding to different gilt yields used as external instrument and normalisation variable included in the VAR. Monetary policy shocks are normalised such that the corresponding gilt yield increases by 100 basis points. The shaded areas indicate the 90% posterior coverage bands. Variables included in VAR (other those shown and the relevant gilt yield): Favara et al. (2016) excess bond premium, FTSE250 index, and BIS Pound Sterling broad effective exchange rate index. Estimation sample: 1997M1-2019M12.*

2.5.2 Identification with text-orthogonalised monetary policy surprises

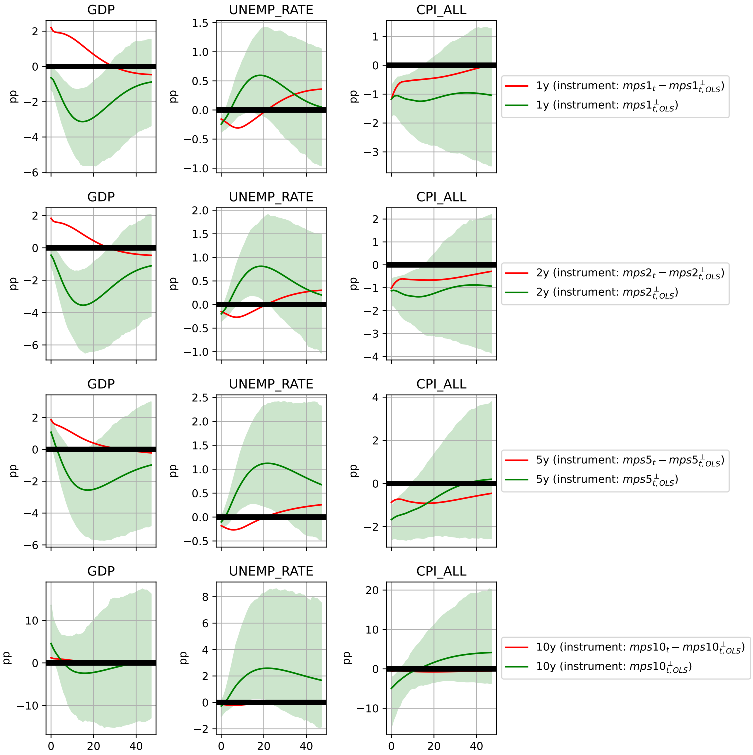
Having established baseline and benchmark results, I now use my text-orthogonalised monetary policy surprises to instrument for monetary policy shocks across the four BVARs (one for each gilt maturity).

OLS

The first set of orthogonalised external instruments are the $mps1^{\perp}_{t,OLS}$, $mps2^{\perp}_{t,OLS}$, $mps5^{\perp}_{t,OLS}$, and $mps10^{\perp}_{t,OLS}$ series, which, as discussed, are generated by taking the residuals from OLS regressions of $mps1_t$, $mps2_t$, $mps5_t$, and $mps10_t$ on text-derived variables generated in Section 2.3 (and four lags thereof).

Figure 12 shows in green the estimated impulse response functions to contractionary monetary policy shocks that are identified using $mps1^{\perp}_{t,OLS}$, $mps2^{\perp}_{t,OLS}$, $mps5^{\perp}_{t,OLS}$, and $mps10^{\perp}_{t,OLS}$. Using orthogonalised monetary policy surprises for identification leads to markedly different results. For the 1 and 2-year gilt models (i.e. the top two rows) the real activity puzzle disappears. Instead, there is a significantly negative GDP response, peaking at around -3 percentage points around 15 months after the shock. For the 5-year model there is still a marginally positive impact effect, though it is followed by a sustained contraction. The employment puzzle is similarly reversed for the 1, 2, and 5-year models, with the positive impact peaking slightly later than the GDP response with a peak around +0.6 percentage points. The price response is significantly negative across the 1, 2, and 5-year models, with an impact effect of around -1 to -2 percentage points that diminishes over time. In the 10-year model, this identification leads to imprecise estimates across all three key indicators. The GDP response continues to be subject to a real activity puzzle, although the unemployment rate responds positively to a contractionary shock identified using the 10-year instrument and the initial price response is negative.

Figure 12: Impulse responses to monetary policy shocks identified using OLS-orthogonalised monetary policy surprises and systematic monetary policy identified using OLS fitted values



Notes: Monetary policy shocks are identified using external instrumental variables $mps1^{\perp}_{t,OLS}$, $mps2^{\perp}_{t,OLS}$, $mps5^{\perp}_{t,OLS}$, and $mps10^{\perp}_{t,OLS}$ (i.e. the residuals from OLS regressions of original high-frequency reactions on text-derived variables) and systematic monetary policy shocks are identified using fitted values of the same OLS regressions. Estimated using a monthly 12-lag Bayesian structural vector autoregression (BSVAR-IV) with normal-inverse-Wishart priors as in Miranda-Agrippino and Ricco (2021). There is one row per VAR, with different rows showing results corresponding to different gilt yields used as external instrument and normalisation variable included in the VAR. Shocks are normalised such that the corresponding gilt yield increases by 100 basis points. The shaded areas indicate the 90% posterior coverage bands. Variables included in VAR (other those shown and the relevant gilt yield): Favara et al. (2016) excess bond premium, FTSE250 index, and BIS Pound Sterling broad effective exchange rate index. Estimation sample: 1997M1-2019M12.

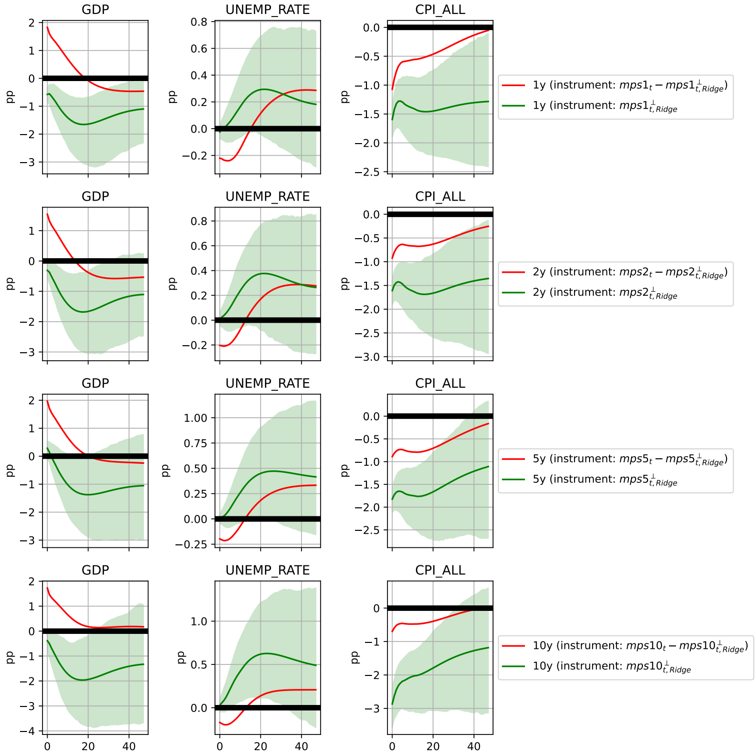
The same figure also shows the impulse responses that result from instrumenting for changes in the gilt yields using the *fitted values* of the OLS regression. That is, the components of high-frequency gilt yield changes that are explained by pre-event public information extracted from text. This series could be interpreted as systematic monetary policy changes, i.e. interest rate changes that are a consequence of the central bank following its (implicit or explicit) policy rule. The dynamic effects are very similar to those resulting from the use of conventional monetary surprises as external instruments.

Ridge

Figure 13 displays the impulse response functions to monetary policy shocks identified using the ridge-orthogonaliation approach, i.e. the $mps1^{\perp}_{t,ridge}$, $mps2^{\perp}_{t,ridge}$, $mps5^{\perp}_{t,ridge}$, and $mps10^{\perp}_{t,ridge}$ orthogonalised monetary policy surprise series. In this specification, the real activity and employment puzzles disappear for all four gilt maturities. The GDP response to a contractionary shock identified using the ridge instrument is significantly negative, with the impact peaking at around -1 to -2 percentage points around 15 months after the shock. The unemployment response is positive in all four models, although only marginally significant. As in the OLS specification, the employment impact peaks slightly later than the GDP impact, around 20 months after the shock, with a peak impact of between +0.3 percentage points to +0.6 percentage points. The price level response is significantly negative in all four models. The response is stronger and more prolonged compared to estimates resulting from conventional monetary policy surprise instruments, with an impact effect of between -1.5 percentage points and -3 percentage points.

As is the case for the OLS instruments, the figure also shows the impulse responses resulting from using the *fitted values* of the ridge regressions as instruments. The effects of these systematic monetary policy changes—as captured by ridge regression fitted values—are very similar to those resulting

Figure 13: Impulse responses to monetary policy shocks identified using ridge-orthogonalised monetary policy surprises and systematic monetary policy identified using ridge fitted values



Notes: Monetary policy shocks are identified using external instrumental variables $mps1^{\perp}_{t,ridge}$, $mps2^{\perp}_{t,ridge}$, $mps5^{\perp}_{t,ridge}$, and $mps10^{\perp}_{t,ridge}$ (i.e. the residuals from ridge regressions of original high-frequency reactions on text-derived variables) and systematic monetary policy shocks are identified using fitted values of the same ridge regressions. Estimated using a monthly 12-lag Bayesian structural vector autoregression (BSVAR-IV) with normal-inverse-Wishart priors as in Miranda-Agrippino and Ricco (2021). There is one row per VAR, with different rows showing results corresponding to different gilt yields used as external instrument and normalisation variable included in the VAR. Shocks are normalised such that the corresponding gilt yield increases by 100 basis points. The shaded areas indicate the 90% posterior coverage bands. Variables included in VAR (other those shown and the relevant gilt yield): Favara et al. (2016) excess bond premium, FTSE250 index, and BIS Pound Sterling broad effective exchange rate index. Estimation sample: 1997M1-2019M12.

from the use of conventional monetary surprises as external instruments.

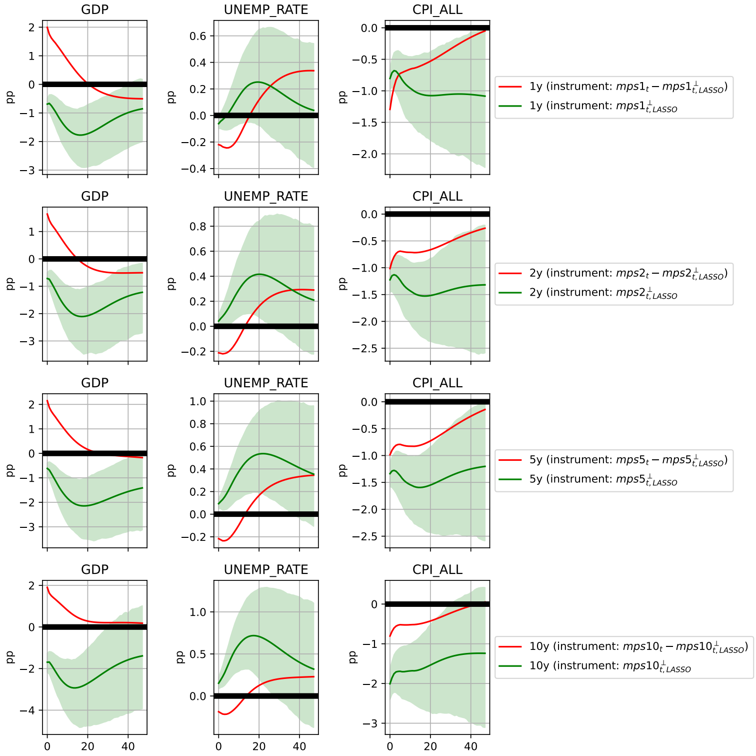
Lasso

Figure 14 shows impulse responses to monetary policy shocks identified using the lasso-orthogonalised monetary policy surprises $mps1^{\perp}_{t,lasso}$, $mps2^{\perp}_{t,lasso}$, $mps5^{\perp}_{t,lasso}$, and $mps10^{\perp}_{t,lasso}$.

Across all four models, the GDP response to a contractionary monetary policy shock is significantly negative, peaking around 15 months after the shock at between -2 percentage points to -3 percentage points. This magnitude is slightly stronger than the response in the ridge identification. Similarly, the employment response is significantly positive in most models, peaking around 20 months after the shock at between +0.2 percentage points to +0.6 percentage points—very close to the estimates resulting from the ridge identification. The price level response is significantly negative and similar to those resulting from the other identifications. Compared to the OLS and ridge-orthogonalisation, the lasso-orthogonalisation appears to be the most precisely-estimated, with eight out of nine impulse response functions in the plot having significant effects at the 90% level. The figure also shows the impulse responses resulting from using the lasso fitted values as instruments, with estimates being similar to those of the OLS and ridge fitted values.

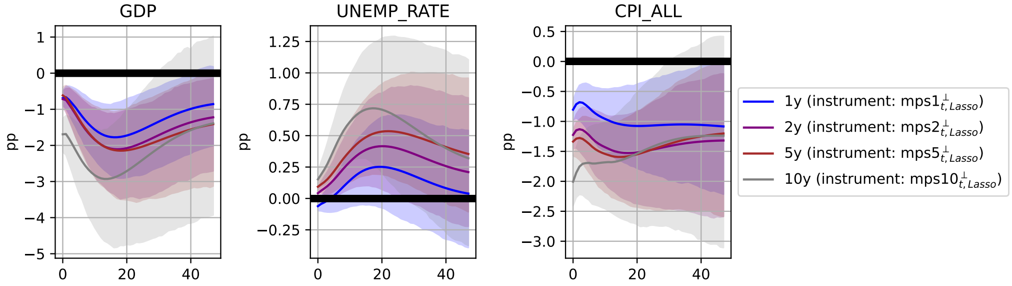
Overall, the impulse responses identified using lasso-orthogonalised monetary policy surprises appear to be the most precisely-estimated. As such, I proceed with using the $mps1^{\perp}_{t,lasso}$, $mps2^{\perp}_{t,lasso}$, $mps5^{\perp}_{t,lasso}$, and $mps10^{\perp}_{t,lasso}$ series as default instruments in subsequent analyses. The summary view in Figure 15 enables a comparison of estimated impulse responses by maturity of the gilt used as instrument in the BVAR by collecting the policy shock responses of Figure 14. Across GDP, unemployment, and price level, the magnitudes of the impulse responses increase with the maturity of the gilt yield used as instrument. In other words, shocks to longer-term rates are

Figure 14: Impulse responses to monetary policy shocks identified using lasso-orthogonalised monetary policy surprises and systematic monetary policy identified using lasso fitted values



Notes: Monetary policy shocks are identified using external instrumental variables $mps1^{\perp}_{t,lasso}$, $mps2^{\perp}_{t,lasso}$, $mps5^{\perp}_{t,lasso}$, and $mps10^{\perp}_{t,lasso}$ (i.e. the residuals from lasso regressions of original high-frequency reactions on text-derived variables) and systematic monetary policy shocks are identified using fitted values of the same lasso regressions. Estimated using a monthly 12-lag Bayesian structural vector autoregression (BSVAR-IV) with normal-inverse-Wishart priors as in Miranda-Agrippino and Ricco (2021). There is one row per VAR, with different rows showing results corresponding to different gilt yields used as external instrument and normalisation variable included in the VAR. Shocks are normalised such that the corresponding gilt yield increases by 100 basis points. The shaded areas indicate the 90% posterior coverage bands. Variables included in VAR (other those shown and the relevant gilt yield): Favara et al. (2016) excess bond premium, FTSE250 index, and BIS Pound Sterling broad effective exchange rate index. Estimation sample: 1997M1-2019M12.

Figure 15: Summary of impulse responses to monetary policy shocks identified using lasso-orthogonalised monetary policy surprises



Notes: Monetary policy shocks are identified using external instrumental variables $mps1^{\perp}_{t,lasso}$, $mps2^{\perp}_{t,lasso}$, $mps5^{\perp}_{t,lasso}$, and $mps10^{\perp}_{t,lasso}$ (i.e. the residuals from lasso regressions of original high-frequency reactions on text-derived variables). This figure shows the rows in Figure 14 aggregated into a single row for ease of comparison.

estimated to be more potent.

2.6 The transmission mechanism of UK monetary policy

While Section 2.5 considered the effect of UK monetary policy on key macroeconomic indicators, this section investigates the *mechanism* through which the Bank of England's communications and policy announcements influence these aggregate quantities. That is, I aim to estimate the dynamic causal effects of UK monetary policy shocks on a wide set of quantities that capture the main channels of the transmission mechanism. As outlined in Chart 1 of Mann (2023), channels through which monetary shocks are thought to percolate through the real economy include (expectations about) interest rates, asset prices, exchange rates, credit, labour and goods markets, and trade.

2.6.1 Large BVAR specification

To explore these channels empirically, I estimate large monthly BVARs with 21 endogenous variables. As in Section 2.5, the specifications include a range of variables from the August 2024 real-time vintage of the monthly 'UK-MD' database curated by Coulombe et al. (2021). I also add monthly GDP, gilt yields, and the latest update by Favara et al. (2016) of the Gilchrist and Zakrajšek (2012) excess bond premium series. The resulting variable set is detailed in Table 2. As before, these specifications are estimated on data from 1997M1 to 2019M12, using the Miranda-Agrippino and Ricco (2021) implementation of Bayesian structural vector autoregression with external instrument (BSVAR-IV) and normal-inverse-Wishart priors. I estimate four versions of this large BVAR—one for each gilt maturity. Analogous to Section 2.5, the interest rate included in each BVAR specification is matched to the maturity of the monetary policy surprise instrument used. Based on the findings of Sections 2.4 and 2.5, I choose the lasso-orthogonalised monetary policy surprises as external instruments for the following analyses. Moreover, it is shown in Section A.6 that there is little difference in the impulse response functions estimated using a 9-lag VAR compared to the default 12-lag specification. As such, the large BVAR specifications in this section are estimated using nine lags. This choice is made due to the computational intensity of estimating a BVAR of this size and to reduce the number of parameters to be estimated from a relatively short sample period.

2.6.2 Implications of text-orthogonalisation

I now compare the dynamic causal effects of UK monetary policy surprises in a large BVAR under different identifications. In particular, monetary policy shocks are identified using either mps_t (i.e. conventional gilt yield monetary policy surprises measured by Braun et al. (2025)) or $mps^{\perp}_{t,lasso}$ (i.e. the residuals from lasso regressions of original high-frequency reactions on

Table 2: Variables included in large BVAR specifications

Variable	Description	Source
GB1YT=RR*	1 Year reference gilt yield	Refinitiv
GB2YT=RR*	2 Year reference gilt yield	Refinitiv
GB5YT=RR*	5 Year reference gilt yield	Refinitiv
GB10YT=RR*	10 Year reference gilt yield	Refinitiv
GDP	Monthly GDP and main sectors to 4 decimal places: Monthly GDP (A-T)	ONS
UNEMP_RATE	Unemployment rate (aged 16 and over, seasonally adjusted)	ONS via UK-MD (Coulombe et al., 2021)
CPI_ALL	CPI INDEX 00: ALL ITEMS 2015=100	ONS via UK-MD (Coulombe et al., 2021)
EBP	Gilchrist and Zakrajšek (2012) excess bond premium	Favara et al. (2016)
FTSE250	FTSE250 (ˆFTMC)	YAHOO! via UK-MD (Coulombe et al., 2021)
GBP_BROAD	Monthly average Broad Effective exchange rate index, Sterling (Jan 2005 = 100) (XUMABK82)	BOE via UK-MD (Coulombe et al., 2021)
AVG_WEEK_HRS	LFS: Avg actual weekly hours of work: UK: All workers in main & 2nd job: SA	ONS via UK-MD (Coulombe et al., 2021)
AWE_ALL	(Average Weekly Earning) AWE: Whole Economy Level (£): Seasonally Adjusted Total Pay Excluding Arrears	ONS via UK-MD (Coulombe et al., 2021)
IOP_PROD	(Index of Production) IOP: B-E: PRODUCTION: CVMSA	ONS via UK-MD (Coulombe et al., 2021)
IOS	(Index of Services) IoS: Services: Index-1dp	ONS via UK-MD (Coulombe et al., 2021)
RSI	(Retail sales index) RSI:Volume Seasonally Adjusted:All Retailers inc fuel:All Business Index	ONS via UK-MD (Coulombe et al., 2021)
RETAIL_TRADE_INDEX	Total Retail Trade in the United Kingdom, Index 2015=100, Monthly, Seasonally Adjusted	FRED via UK-MD (Coulombe et al., 2021)
EXP_TOT	Total Trade (TT): WW: Exports: BOP: CVM: SA	ONS via UK-MD (Coulombe et al., 2021)
IMP_ALL	Total Trade (TT): WW: Imports: BOP: CVM: SA	ONS via UK-MD (Coulombe et al., 2021)
CONS_CREDIT_ex_student_loan	Monthly amounts outstanding of total (excluding the SLC) sterling consumer credit lending to individuals SA (LPMBI2O)	BOE via UK-MD (Coulombe et al., 2021)
TOT_HOUSE_APP	Monthly number of total sterling approvals for house purchase to individuals seasonally adjusted (LPMVTVX)	BOE via UK-MD (Coulombe et al., 2021)
MORT_FIXED_RATE_5YRS	Monthly interest rate of UK MFIs sterling 5 year (75% LTV) fixed rate mortgage to households NSA (IUMBV42)	BOE via UK-MD (Coulombe et al., 2021)
MORT_FIXED_RATE_2YRS	Monthly interest rate of UK MFIs sterling 2 year (75% LTV) fixed rate mortgage to households NSA (IUMBV34)	BOE via UK-MD (Coulombe et al., 2021)
BCI	Business confidence index (BCI)Amplitude adjusted, Long-term average = 100	OECD via UK-MD (Coulombe et al., 2021)
CCI	Consumer confidence index (CCI)Amplitude adjusted, Long-term average = 100	OECD via UK-MD (Coulombe et al., 2021)

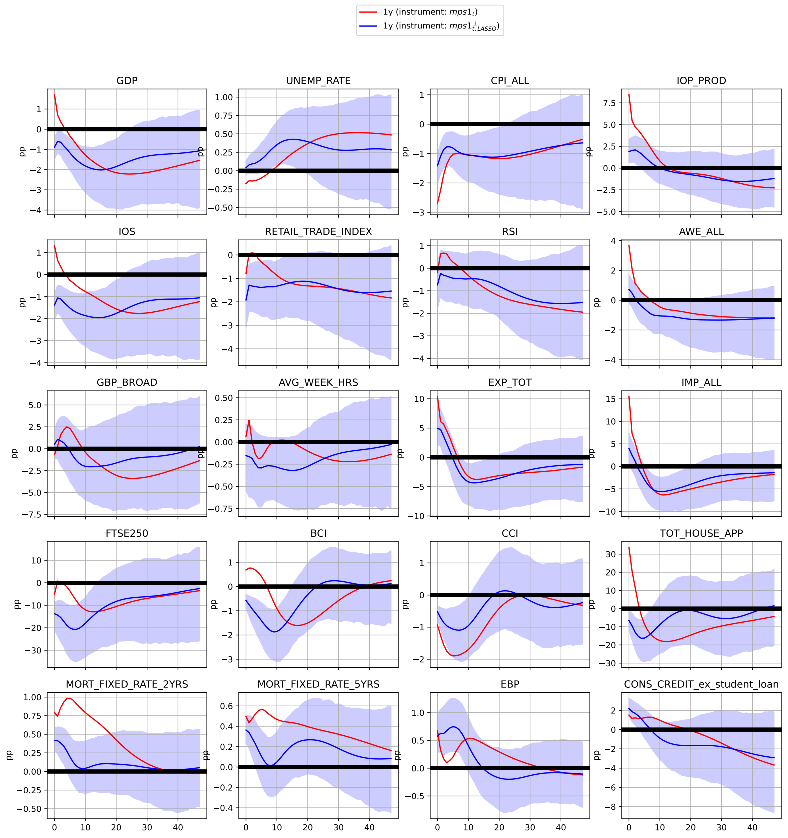
*depending on specification.

text-derived variables) as external instrumental variables in the BSVAR-IV procedure. The results for each maturity specification are now discussed in turn.

1-year gilt specification

I begin by considering the BVAR specification that includes the 1-year gilt. The estimated impulse response functions are shown in Figure 16. Considering the effect on GDP, there is a significant difference between the two identifications. In particular, the identification based on conventional monetary policy surprises results in a significant real activity puzzle of +1.5 percentage points on impact. In contrast, the identification using the text-orthogonalised instrument results in a significant negative impact effect on real activity of -1 percentage point. This latter result is in line with the results of Section 2.5. There is a similar difference with regards to the unemployment rate, where the $mps1_t$ instrument leads to an employment puzzle with a negative effect on unemployment upon impact of around -0.2 percentage points. This is compared to a near-zero impact effect when using the $mps1^{\perp}_{t,lasso}$ instrument. Again, this difference is consistent with the difference observed across Figures 10 and 15. As in Section 2.5, the price level impacts are qualitatively similar but quantitatively different in the large BVAR specification. In par-

Figure 16: Impulse responses of 20 variables to monetary policy shocks identified using conventional and lasso-orthogonalised 1-year gilt yield monetary policy surprise series



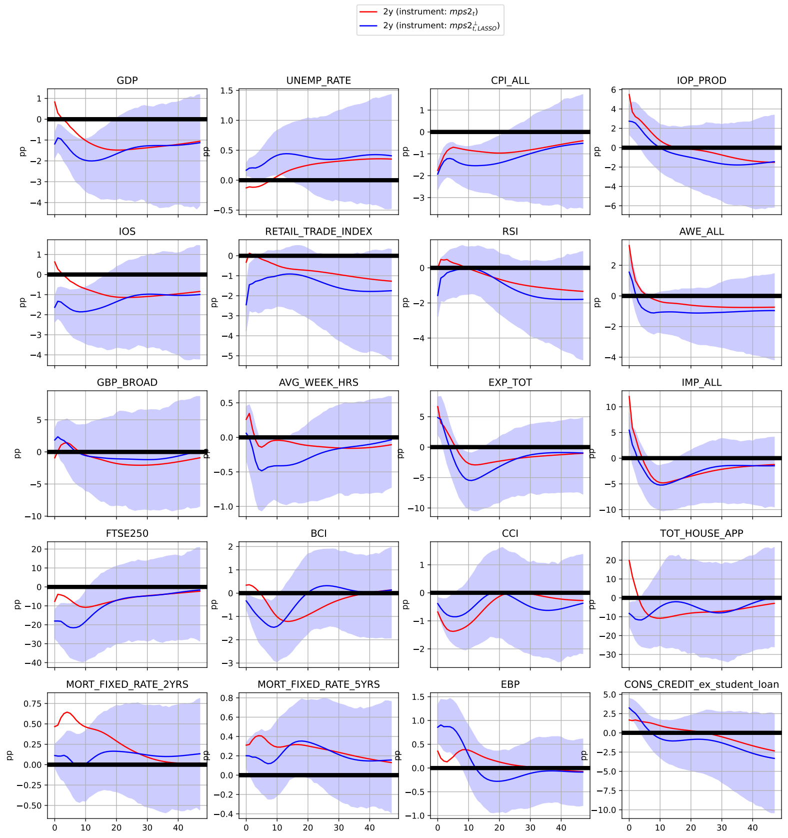
Notes: Monetary policy shocks are identified either using $mps1_t$ *(i.e. conventional gilt yield monetary policy surprises measured by Braun et al. (2025)) or* $mps1^{\perp}_{t,lasso}$ *(i.e. the residuals from lasso regressions of original high-frequency reactions on text-derived variables). Estimated using a monthly 21-variable, 9-lag Bayesian structural vector autoregression (BSVAR-IV) with normal-inverse-Wishart priors as in Miranda-Agrippino and Ricco (2021). Monetary policy shocks are normalised such that the 1-year gilt yield increases by 100 basis points. The shaded areas indicate the 90% posterior coverage bands. Estimation sample: 1997M1-2019M12.*

ticular, a monetary policy shock identified using the conventional monetary policy surprise instrument is estimated to result in a -2.5 percentage points effect on the price level upon impact. This is in contrast to a -1.5 percentage points effect when using the text-orthogonalised instrument. Considering the index of production indicator, which was not included in the specifications in Section 2.5, a significant real activity puzzle is evident for the conventional identification. In particular, in this identification a contractionary monetary policy shock is estimated to have an implausible impact effect of +7.5 percentage points on industrial production. This puzzle is weakened substantially by text-orthogonalisation, with the identification using the $mps1^{\perp}_{t,lasso}$ instrument instead estimating a nearly insignificant +2 percentage points impact effect. The estimated impulse responses for the index of services are qualitatively similar. In particular, there is an activity puzzle when using the conventional monetary policy surprises as instrument with an impact effect of more than +1 percentage point. When instead using the text-orthogonalised instrument, the impact effect switches sign to an effect slightly stronger than -1 percentage point. Considering the impact on the retail trade index under the different identifications, using the text-orthogonalised instrument leads to a stronger negative impact during the first 10 months after the shock. In particular, the impact effect is estimated to be just under -2 percentage points, which is similar to the estimated effect on the index of services and GDP. For the retail sales index, there is an activity puzzle under the conventional identification, with a peak impact of nearly +1 percentage point around three months after the shock. This is compared to an impact effect of -0.8 percentage points when using the text-orthogonalised instrument. While the latter result is broadly aligned with the impacts on GDP, index of services, and the retail trade index, the effect is not statistically significant. Similar to the estimated impulse responses of GDP and the index of production, the effect on average weekly earnings that obtains when using conventional monetary policy surprises contains a significant puzzle. Upon impact, a contractionary monetary policy shock is estimated to result in a surprising +4 percentage points increase in weekly earnings. This puzzle disappears completely when instead using the text-orthogonalised instrument $mps1^{\perp}_{t,lasso}$ for identification. Under the conventional identification, the immediate impact response on the sterling exchange rate index is marginally negative. In contrast, using text-orthogonalised instrument $mps1^{\perp}_{t,lasso}$ results in a positive (albeit statistically insignificant) impact effect on sterling. Under both identifications, the immediate effect is followed by a depreciation that is aligned with the estimated medium-term reduction in GDP shown in the first panel. Considering the estimated response of average weekly hours under both identifications, text-orthogonalisation results in a sustained negative impact of about -0.2 percentage points (albeit this effect is statistically insignificant). This is in contrast to the conventional identification which results in an inconclusive impulse response estimate. Similar to industrial production, the conventional identification results in an activity puzzle response being estimated for total exports. In particular, the immediate impact effect of a contractionary monetary policy shock identified using the conventional instrument $mps1_t$ amounts to over +10 percentage points. The puzzle remains statistically significant under the identification with the text-orthogonalised instrument, although the puzzling impact effect is more moderate at +5 percentage points. Nevertheless, this result provides further evidence that, even after text-orthogonalisation, monetary policy surprises might be subject to some residual endogeneity with respect to industrial production and trade activities. The results for total imports are similar to those for exports and industrial production, in that there is a puzzling impact effect under the conventional identification. In this case, a contractionary monetary policy shock normalised to 100 basis points is estimated to increase imports by more than +15 percentage points on impact. Text orthogonalisation results in a weaker impact effect of +4 percentage points. Under both identifications, the impact effect is followed by a significant reduction in imports that is aligned with the medium-term reduction in GDP shown in the first panel. Under both identifications there is a negative impact on the FTSE250 equity index, although the effect is estimated to be stronger when using $mps1^{\perp}_{t,lasso}$. In particular, the impact effect of a contractionary monetary policy shock normalised to increase the 1-year rate by 100 basis points is estimated to be around -8 percentage points under the conventional identification and around -18 percentage points using the text-orthogonalised external instrument. Regarding the business confidence index, the impact effects have opposite signs across the two specifications. Under the conventional identification, a contractionary monetary policy shock is estimated to have a positive impact on business confidence of around +0.8 percentage points. This is in contrast the identification relying on $mps1^{\perp}_{t,lasso}$, which results in a significantly negative impact effect of around -0.5 percentage points, peaking at -2 percentage points 10 months after the shock. Both identifications result in qualitatively similar effects on consumer confidence, although the negative effect is around half as strong when using the text-orthogonalised instrument. There is a material difference between the two identifications with respect to the estimated impact on the UK housing market. While the conventional identification results in a very strong +30 percentage points impact effect on total house purchase approvals, the text-orthogonalised identification results in an impact effect of around -8 percentage points followed by a peak effect of around -15 percentage points. The estimated impacts on the 2-year and 5-year fixed mortgage rates are qualitatively similar across the two identification, with significant positive responses estimated in response to a contractionary monetary policy shock. The impact effect of the text-orthogonalised identification is somewhat weaker and dissipates more quickly as real activity contracts, while the conventional identification results in a stronger, longer-lasting response being estimated. The Gilchrist and Zakrajšek (2012) excess bond premium responds positively under both identifications, although the estimate that obtains when using the text-orthogonalised instrument stays significantly positive for longer. It is worth noting that this quantity is not specific to the UK economy. As such, the significant positive impact observed here could be the result of both the Bank of England's impact on global quantities or evidence of international correlations of central bank actions. Finally, there is a significantly positive impact effect of around +3 percentage points on consumer credit under the text-orthogonalised identification, compared to no significant impact effect when conventional monetary policy surprises are used.

2-year gilt specification

The results of using the 2-year gilt specification are shown in Figure 17. As in the BVAR specification that includes the 1-year gilt yield and uses 1-year gilt yield monetary surprises, the sign of the impact on GDP differs across the two specifications. This is also the case for the unemployment rate, while the estimated responses of the price level remain qualitatively similar under both identifications and are closer than in the 1-year BVAR. The response of the index of production to a monetary policy shock is subject to the same activity puzzles as in the 1-year specification, with a significant real activity puzzle under the conventional identification that is weakened by instead using the text-orthogonalised instrument $mps2^{\perp}_{t,lasso}$. The responses of the index of services have different signs under different identifications as in the previous section. As in the 1-year specification, the estimated responses of the retail trade and retail sales indices differ somewhat across the two specifications. Average weekly earnings are subject to the same positive impact response as in the 1-year specification, with the puzzling impact being half as strong when identifying the monetary policy shock using the text-orthogonalised instrument $mps2^{\perp}_{t,lasso}$. As before the exchange rate impact effects have opposite signs across the two identifications, with the positive effect when using the text-orthogonalised instrument being slightly stronger (although still only marginally significant) in the 2-year specification. The difference

Figure 17: Impulse responses of 20 variables to monetary policy shocks identified using conventional and text-orthogonalised 2-year gilt yield monetary policy surprise series



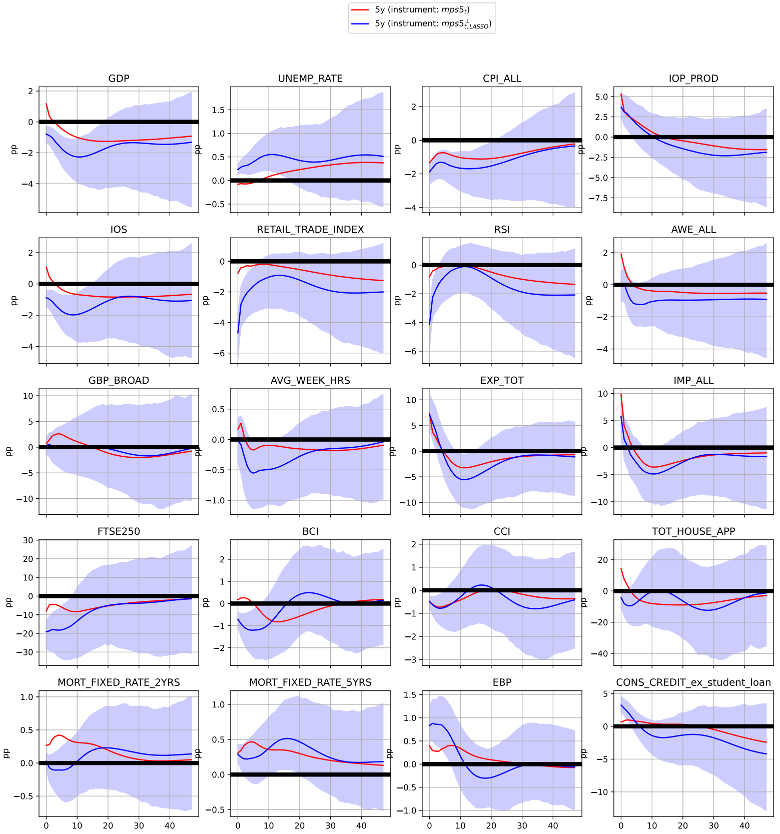
Notes: Monetary policy shocks are identified either using $mps2_t$ *(i.e. conventional gilt yield monetary policy surprises measured by Braun et al. (2025)) or* $mps2^{\perp}_{t,lasso}$ *(i.e. the residuals from lasso regressions of original high-frequency reactions on text-derived variables). Estimated using a monthly 21-variable, 9-lag Bayesian structural vector autoregression (BSVAR-IV) with normal-inverse-Wishart priors as in Miranda-Agrippino and Ricco (2021). Monetary policy shocks are normalised such that the 2-year gilt yield increases by 100 basis points. The shaded areas indicate the 90% posterior coverage bands. Estimation sample: 1997M1-2019M12.*

across identifications in the response of average weekly hours in the 1-year specification persists in the 2-year specification, although impulse responses are imprecisely estimated across both identifications. Both exports and imports respond similarly in this specification, across both the conventional and the text-orthogonalised specification. That said, the strong positive impact on exports under the conventional identification is weaker in the 2-year specification—resulting in little difference regarding the export response between the two identifications. The impact response of imports continues to be twice as strong under the conventional identification. The equity market responses are similar to the 1-year specification with the same differences between identifications persisting in the 2-year specification. The same is true for estimated responses of business confidence, consumer confidence, and housing approvals. Regarding the 2-year and 5-year fixed mortgage rates, the responses in Figure 17 show less pass-through of the monetary policy shock than in Figure 16 under both the conventional and the text-orthogonalised identification. Using the $mps2^{\perp}_{t,lasso}$ instrument results in a muted response of the 2-year mortgage rate in particular. The excess bond premium and consumer credit responses are qualitatively similar across Figures 17 and 16.

5-year gilt specification

Figure 18 shows the results of the 5-year specification. There are some differences compared to Figures 16 and 17. For instance, in response to a monetary policy shock scaled to induce a 100 basis point increase in the 5-year gilt yield, there is a significant difference in the impact response of the retail trade index and retail services index. In particular, the impact response is estimated to be around -4 percentage points under text-orthogonalised identification, compared to less than -1 percentage point when using conventional monetary policy surprises. The response of average weekly earnings is also persistently negative in this specification when using $mps5^{\perp}_{t,lasso}$ as external instrument.

Figure 18: Impulse responses of 20 variables to monetary policy shocks identified using conventional and text-orthogonalised 5-year gilt yield monetary policy surprise series



Notes: Monetary policy shocks are identified either using $mps5_t$ *(i.e. conventional gilt yield monetary policy surprises measured by Braun et al. (2025)) or* $mps5^{\perp}_{t,lasso}$ *(i.e. the residuals from lasso regressions of original high-frequency reactions on text-derived variables). Estimated using a monthly 21-variable, 9-lag Bayesian structural vector autoregression (BSVAR-IV) with normal-inverse-Wishart priors as in Miranda-Agrippino and Ricco (2021). Monetary policy shocks are normalised such that the 5-year gilt yield increases by 100 basis points. The shaded areas indicate the 90% posterior coverage bands. Estimation sample: 1997M1-2019M12.*

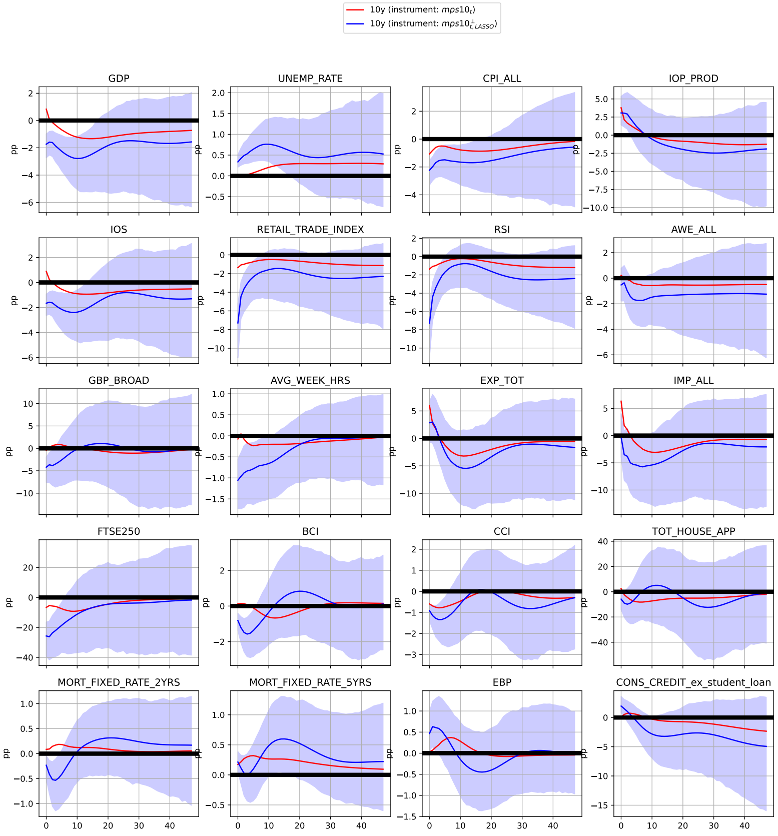
10-year gilt specification

These differences across the two identification approaches are echoed in the results of the 10-year specification shown in Figure 19. Additionally, the 10-year specification also shows a significantly negative impact on average weekly hours of -1 percentage point under the text-orthogonalised identification, compared to no impact when using conventional monetary policy surprise. In this specification there is also a significantly negative response of 2-year fixed mortgage rate. This finding could be a sign that there is some residual endogeneity. For example, to the extent that the monetary policy surprises affecting long-term interest rates occurred primarily in recessionary circumstances, rising mortgage default rates may have led lenders to incorporate a larger risk premium.

2.6.3 Estimated transmission channels of UK monetary policy

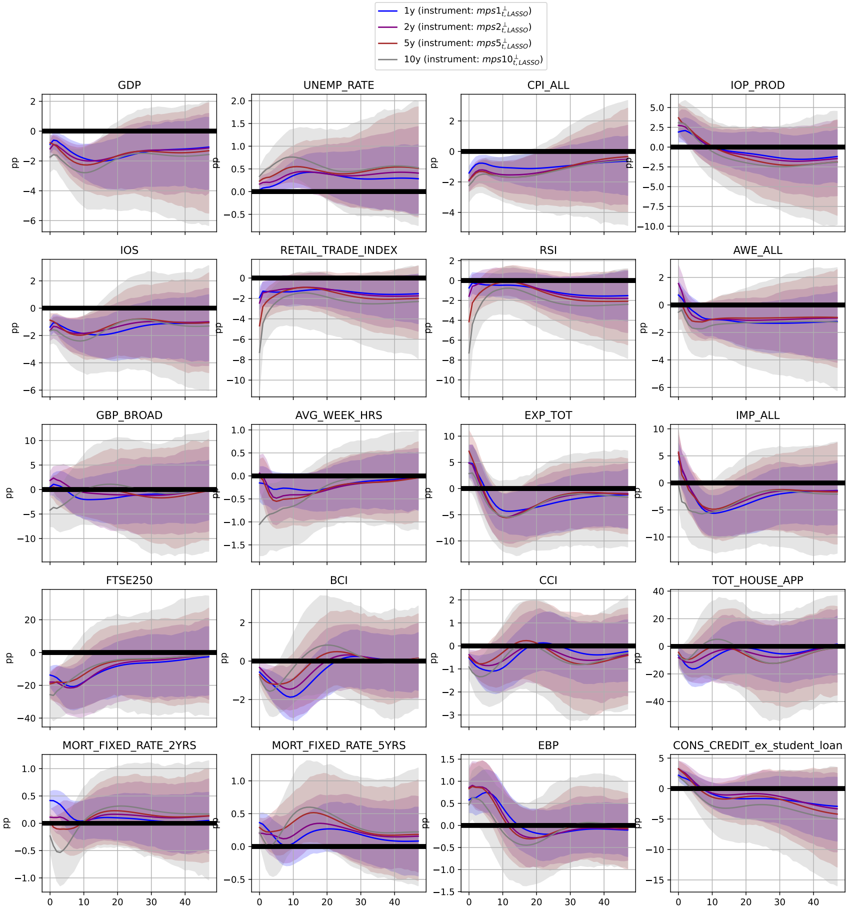
Given the analysis in Section 2.6.2, it appears that overall text-orthogonalised monetary policy surprises result in estimated impulse responses that are more aligned with theoretical consensus. In some cases, identification with conventional monetary policy surprises leads to arguably implausibly strong impact effects with atypical signs. Given this finding and the recommendation in Bauer and Swanson (2023b), I focus my discussion of the UK monetary policy transmission mechanism on the impulse responses estimated using the text-orthogonalised instrument rather than those estimated using conventional monetary policy surprises. To aid comparison across different BVAR specifications, the impulse responses arising from text-orthogonalised identifications reported in Figures 16 to 19 are collected in Figure 20. In what follows, these are discussed in the context of findings in the existing literature.

Figure 19: Impulse responses of 20 variables to monetary policy shocks identified using conventional and text-orthogonalised 10-year gilt yield monetary policy surprise series



Notes: Monetary policy shocks are identified either using $mps10_t$ *(i.e. conventional gilt yield monetary policy surprises measured by Braun et al. (2025)) or* $mps10^{\perp}_{t,lasso}$ *(i.e. the residuals from lasso regressions of original high-frequency reactions on text-derived variables). Estimated using a monthly 21-variable, 9-lag Bayesian structural vector autoregression (BSVAR-IV) with normal-inverse-Wishart priors as in Miranda-Agrippino and Ricco (2021). Monetary policy shocks are normalised such that the 10-year gilt yield increases by 100 basis points. The shaded areas indicate the 90% posterior coverage bands. Estimation sample: 1997M1-2019M12.*

Figure 20: Summary of impulse responses of 20 variables to monetary policy shocks identified using lasso-orthogonalised monetary policy surprise series



Notes: Monetary policy shocks are identified using external instrumental variables $mps1^{\perp}_{t,lasso}$, $mps2^{\perp}_{t,lasso}$, $mps5^{\perp}_{t,lasso}$, $mps10^{\perp}_{t,lasso}$ (i.e. the residuals from lasso regressions of original high-frequency reactions on text-derived variables). This figure aggregates results from Figures 16, 17, 18, and 19 for ease of comparison.

GDP

Aligned with Section 2.5, the impact of a contractionary monetary policy shock on GDP is significantly negative across specifications. The impact effect is estimated to lie between -1 percentage point and -2 percentage points depending on the interest rate maturity used as instrument and in the VAR. The negative impact peaks between 10 to 15 months after the shock at between -2 percentage points for the 1-year specification to -3 percentage points in the 10-year specification. The estimated response in this study is slightly stronger compared to the peak effect of around -1.25 percentage points that Aruoba and Drechsel (2024) find after around 25 months in response to a shock normalised to increase the 1-year rate by 100 basis points. Similarly, Cesa-Bianchi et al. (2020) estimate that GDP responds with a peak effect of around -1.4 percentage points 24 months after the shock. This is also similar to Braun et al. (2025)'s estimated impulse response to a 'Target' shock of around -2 percentage points, while their estimated responses to 'Path' and 'QE' shocks differs substantially.

Unemployment rate

The effect on the unemployment rate is significantly positive across specifications, with the impact effect ranging from near-zero for the 1-year specification to around +0.7 percentage points for the 10-year specification. For both GDP and unemployment, the peak impact comes soonest in the 10-year specification. Peak effect magnitudes range from +0.3 percentage points to +0.8 percentage points. This estimate is similar to the +0.75 percentage points unemployment response to a monetary policy shock scaled to induce a 100 basis point increase in the 1-year US bond yield estimated by Aruoba and Drechsel (2024). It is also in line with Miranda-Agrippino and Ricco (2021)'s estimate of a near-zero impact effect followed by a peak effect of around +0.3 percentage points around 18 months after a shock normalised to a 100 basis point increase in the 1-year rate. The estimated responses in

Figure 15 are also similar to that in Cesa-Bianchi et al. (2020), which finds a near-zero impact effect and a peak effect of around +0.4 percentage points around 20 months after a shock that is scaled appropriately.

Consumer price index

The price level impact is significantly negative across specifications, with an impact response of around -1.5 percentage points to -2 percentage points that dissipates gradually thereafter. The response is strongest in the 10-year specification. In contrast to this paper's estimate, Braun et al. (2025) find a significant price puzzle. Consumer prices are estimated to peak at +1 percentage point six months after a 'Target' shock. In response to a 'Path' shock, Braun et al. find an impact effect of around -0.7 percentage points growing to about -2.8 percentage points after three years. The response in Figure 15 is strong relative to the estimate in Cesa-Bianchi et al. (2020), who find a -0.28 percentage points impact effect with a peak response 10 months after the shock of around -0.4 percentage points. This is also a strong negative response compared to the -0.25 percentage points peak effect estimated in Aruoba and Drechsel (2024) and the Miranda-Agrippino and Ricco (2021) estimate an impact effect of -0.3 percentage points and a peak effect of -0.7 percentage points in response to a similarly-scaled shock.

Index of production

There is a slight 'production puzzle' across specifications, with the impact effect on index of production ranging between around +2 percentage points and +3.5 percentage points. In contrast, Miranda-Agrippino and Ricco (2021) estimate an impact effect of -1 percentage point with a peak effect of -1.6 percentage points. Similarly, Bauer and Swanson (2023b) estimate an impact effect of around -0.8 and a peak effect of -1.6 percentage points. It should be noted that the finding in Figure 15 is a milder puzzle compared to the identification with conventional monetary policy surprises and that the impact effect is marginally significant across specifications. Nevertheless, this finding could indicate that the text-orthogonalised instrument is subject to some residual endogeneity with respect to industrial production.

Index of services

The estimated effects on the index of services is aligned with the impact on GDP, with an impact effect of around -1 percentage point in the 1-year specification to -2 percentage points in the 10-year specification. The peak effect is around -2 percentage points across specifications, although it is strongest and soonest in the 10-year specification.

Retail trade & retail sales indices

Both the retail trade index and the retail sales index are estimated to respond strongly to a contractionary monetary policy shock, with the magnitude increasing with the maturity of the interest rate used as instrument and for normalisation. For both variables, the impact effect ranges from -2 percentage points to -7 percentage points across specifications.

Average weekly earnings

Considering the response of average weekly earnings, results regarding the impact effect differ across specifications, although neither impact nor peak effects are statistically significant in any specification. This result is qualitatively similar the insignificant impulse response estimated for the US in Miranda-Agrippino and Ricco (2021).

Sterling exchange rate

Similar to the findings regarding average weekly earnings, the results regarding the impact effect on the broad exchange rate index have different signs depending on the specification. In particular, impact effects are positive for short-term specifications, but are significantly negative for the 10-year specification. This latter finding may be evidence of some residual endogeneity in the text-orthogonalised instrument. The result for the 2-year specification is similar in magnitude to the +2 percentage points peak effect estimated for the US economy in Miranda-Agrippino and Ricco (2021). Braun et al. (2025) estimate stronger peak effects of +4 percentage points in response to a 'Target' and +10 percentage points in response to a 'Path' shock. Similarly, Cesa-Bianchi et al. (2020) estimate a peak effect of around +6 percentage points.

Average weekly hours

Regarding weekly hours, the impact effect ranges from near-zero to -1 percentage point in the 10-year specification. For the 10-year specification the impact dissipates gradually following the impact, while the peak effect of -0.5 percentage points in other specifications is reached around five months after the shock. Most effects are marginally significant, with the exception of the strongly significant impact effect in the 10-year specification. This is a slightly stronger response compared to a peak effect of -0.2 percentage points estimated by Miranda-Agrippino and Ricco (2021).

Exports

Estimated impulse responses of exports are broadly similar across specifications. In particular, there is a significant impact effect of around +5 percentage points, followed by a peak negative effect of -5 percentage points around 10 months after the shock. Similarly to industrial production, this could be evidence of some residual endogeneity within the text-orthogonalised instruments. The immediate impact effects are weaker, however, than under the conventional identifications (as discussed above). The somewhat puzzling positive impact effect aside, the negative peak effects are similar to the estimates in Miranda-Agrippino and Ricco (2021), who in the US find a peak impact of -4 percentage points after around six months.

Imports

The estimated impulse responses of imports are similar to those of exports, with comparable peak and impact effects. Miranda-Agrippino and Ricco (2021) find a similar peak effect for imports as well.

Equity market index

The estimated effect on the equity market is economically and statistically significantly negative across all specifications. The impact effects range from around -17 percentage points to around -25 percentage points, with the response strongest in the 10-year specification. This is similar to the estimated impulse response in Aruoba and Drechsel (2024), who find that stock prices fall by around 15 percentage points in response to a shock normalised to increase the 1-year rate by 100 basis points. In response to similarly-scaled 'Target' and 'Path' shocks, Braun et al. (2025) estimate similar peak effects of around -10 percentage points and -7 percentage points respectively. For a 'QE shock', their estimated response of around -14 percentage points is also similar in magnitude. Studying the US economy, Miranda-Agrippino and Ricco (2021) find a somewhat weaker negative equity market response of around -6 percentage points to a similarly-scaled shock. Similarly, Cesa-Bianchi et al. (2020) estimate a weaker peak response of around -4 percentage points in the UK context.

Business confidence index

The business confidence index is estimated to respond negatively to a contractionary monetary policy shock. The peak effect reaches around -1.8 percentage points across specifications after around 10 months, although the peak is reached sooner in the 10-year specification.

Consumer confidence index

The results for consumer confidence are similar to those regarding the business confidence index, although the peak effects are somewhat weaker at about -1 percentage point across specifications. While not statistically significant, Miranda-Agrippino and Ricco (2021) estimate a similarly-sized peak effect of around -1 percentage point in the US.

Total housing approvals

The total housing approvals response is estimated to respond strongly negatively across specifications, with an impact effect of around -5 percentage points and peak effects between around -5 percentage points to -15 percentage points. Statistical significance of this response is strongest in the 1-year specification. This is similar to Miranda-Agrippino and Ricco (2021)'s result on US housing starts and building permits, which have peak responses of around -10 percentage points to a similarly-scaled monetary policy shock.

Mortgage rates

Estimated two-year and five-year mortgage rate responses vary across specifications and many are not generally statistically significant. In specifications where there is significance, impact effects are estimated to be around +0.25 percentage points.

Excess bond premium

Across specifications, the impact on the Gilchrist and Zakrajšek (2012) excess bond premium is significantly positive, with an impact effect of around +0.8 percentage points across specifications that dissipates 10 months after the shock. This is somewhat larger compared to the +0.6 percentage points effect in response to a 100 basis point shock (normalised using the 1-year rate) shortly after impact estimated in Aruoba and Drechsel (2024) and the +0.5

percentage points effect Miranda-Agrippino and Ricco (2021). Bauer and Swanson (2023b) estimate a peak effect of around +0.2 percentage points.

Consumer credit

Finally, the initial effect on consumer credit is estimated to be significantly positive across specifications, although this dissipates around five months after the shock and turns negative. The overall shape of this response is similar to that in Figure 3 of Cesa-Bianchi et al. (2020). Miranda-Agrippino and Ricco (2021) also estimate a positive, albeit insignificant, response of US consumer loans to a similarly-scaled shock.

2.7 Conclusion

This paper provides a reassessment of the effects of Bank of England policy on the UK economy. This is in light of mounting evidence that conventional high-frequency identification schemes may lead to biased effect estimates due to invalidity of conventional monetary policy surprises as external instrumental variables in SVAR or LP frameworks. I compile a corpus of newswire articles that would have been available to market participants just prior to monetary events and use an LLM to take measurements of perceived UK economic conditions described in each article. I then orthogonalise conventional monetary policy surprises with respect to these LLM-generated measures in an effort to mitigate instrument invalidity issues.

I find that my text-orthogonalisation approach makes a material difference to the estimated dynamic causal effects of UK monetary policy on macroeconomic aggregates. In both small and large BVAR specifications, the sign of estimated effects can depend on whether or not conventional monetary policy surprises are orthogonalised with respect to measurements of public, pre-event information regarding UK economic conditions. Text-orthogonalisation tends to yield effect estimates that are aligned with theoretical consensus and international evidence. For example, it resolves both a real activity puzzle and an employment puzzle that arise when using conventional (i.e. non-orthogonalised) monetary policy surprises to identify monetary policy shocks in small BVAR specifications.

Overall, these findings can be taken as support of the 'central bank response'-hypothesis proposed in Bauer and Swanson (2023a), in that it is not necessary to orthogonalise monetary policy surprises with respect to private, central bank-internal information to arrive at dynamic effect estimates that are aligned with theoretical consensus. My findings are not consistent with the theory that a significant component of conventional monetary policy surprises are driven by the Bank of England revealing its 'inside information' about the state of the economy. Instead, my results are consistent with market participants having imperfect knowledge about the Bank of England's reaction function to publicly-accessible news. This study's empirical support for the 'central bank response'-hypothesis of Bauer and Swanson (2023a) over the 'central bank information'-hypothesis, suggest that central bankers are not at risk of inadvertently releasing private central bank information that unleashes unintended information effects. In the words of Bauer and Swanson (2023b)

> *'[P]olicymakers have little need to fear that information effects might attenuate the effects of their announcements.'*

Another finding is that monetary policy appears to be more potent when it is targeted at the far end of the yield curve. This has implications for the use of formal tools that target longer-term interest rates (e.g. QE-type interventions) as well as longer-term forward guidance. Regarding forward guidance specifically, the empirical results of this study underscore the importance of resolving commitment problems regarding credibility. If the Bank could make credible commitments regarding the policy rate for the next 10 years, such commitments could be a potent policy tool that avoids large scale asset transaction programmes like QE or QT.

This study makes a number of contributions. It develops and implements a novel methodology to measure economic conditions from unstructured textual data, using a large language model to generate structured information. It also demonstrates that there are ex post correlations between state-of-the-art UK monetary policy surprises and LLM-extracted pre-event information. It studies the sensitivity of the dynamic causal effects of monetary policy estimated with respect to different identifications and provides new estimates of the effects and transmission of Bank of England policies. Overall, this study contributes to answering three of the priority topics in the Bank of England's 2024 Agenda for Research: *'How can machine learning and artificial intelligence be deployed by supervisors and central banks?'*, *'How do central bank policy rates affect inflation and has this relationship changed over the recent past?'*, and *'What are the transmission mechanisms of conventional monetary policy and central bank balance sheet adjustments?'*.[10] This study also creates three research data assets for future empirical investigations into UK monetary policy. The first is a corpus of raw newswire text relating specifically to UK monetary events in the Braun et al. (2025) database. The second is a structured dataset about UK economic conditions prior to each monetary event. The third is a set of text-orthogonalised monetary policy surprise series that yield largely puzzle-free dynamic causal effect estimates.

A limitation of this study is that it does not distinguish between 'formal' and 'informal' monetary policy surprises. That is, the effects of a monetary policy shock due to a formal policy tool (e.g. the bank rate) being set in a way that was not fully expected may be different from the effects of a shock due to there being, for example, a surprising phrase in the latest inflation report that led market participants to revise their forecasts regarding the path of future policy. Another limitation of this study is that dynamic causal effect estimates are generally imprecise due to the relatively short sample size accrued since the Bank of England's operational independence in 1997. Re-

[10] https://www.bankofengland.co.uk/research/bank-of-england-agenda-for-research

garding methodological limitations of this study, there is no reason to think that the information extracted from newswires via LLM-prompts capture the pre-event information set *optimally.* To the extent that the approach in this paper fails to capture all of the relevant information, the text-orthogonalised monetary policy surprises may lead to estimated impulse responses with some residual bias due to residual instrument endogeneity. Indeed, there is some evidence from the large BVAR specifications that suggest there may be some residual endogeneity present—although these could also be driven by the small sample size relative to the parameters to be estimated. An alternative approach could be to tackle the regression task directly by developing and fine-tuning a task-specific LLM to predict monetary policy surprises directly based on raw text. Another methodological shortcoming is that the LLM 'knows' what is going to happen due to it being likely that coverage of events following a specific monetary event was part of the corpus of text that the LLM was trained on. To mitigate this risk of hindsight bias, the prompt includes the phrase 'Does the newswire state ....' in order to retrieve information contained within the newswire itself. Mitigating this risk more fully would require iterated re-training of the LLM with expanding information sets, which is beyond the scope of this study. Finally, the methodology relies on the assessments made by the financial journalists that write newswire articles.

This paper adds to a growing literature on the uses of textual data for economic research, by showing how large language models can be prompted to take meaningful measurements of economic concepts from unstructured text. That is, the language model essentially performed the work of a team of fast yet diligent research assistants. As such, the emergence of capable large language models has expanded the economic research production possibility frontier generally, making previously infeasible projects feasible. Potential areas for future research include addressing one of the above limitations by refining the prediction technology used to orthogonalise with respect to text. In particular, this could involve tackling the regression task directly by modifying and fine-tuning an LLM to predict monetary policy surprises based on raw text. Another extension would involve applying the text-orthogonalisation method developed in this study to the study of central banks in other countries. Newswires cover central bank communications across many jurisdictions, so this approach is readily generalisable. Future work could also aim to differentiate the impact of 'formal' and 'informal' monetary policy surprises.

A.1 Textual data collection

Starting with the monetary events contained in the Braun et al. (2025) database, a corpus of 398 intraday newswire articles dating back to 1997 was assembled as follows. For each event, same-day articles that were published just before the Bank of England's information release were collected, to approximate the pre-event information set as closely as possible. All textual data were retrieved from Refinitiv's Eikon database, via the 'News Monitor' app. Relevant articles were selected manually from the full history of "Important + Most Recent" newswire outputs. To simplify identification of coverage relating specifically to the Bank of England, the following search query was used within the search bar of Eikon's News Monitor: ("bank of england" OR "boe"). There were typically multiple articles covering each monetary event, from a number of different newswire providers. For consistency of editorial style across events, coverage by the 'Reuters News [RTRS]' newswire was selected whenever it was available. Generally, there were articles from this source covering UK monetary events throughout 1997–2024. In a handful of instances other newswire sources were used to fill gaps. For non-MPC monetary events it was sometimes not possible to identify same-day pre-communication coverage. In these instances, relevant coverage that was published during the preceding days was selected to ensure that the corpus includes exclusively pre-event information. In some other cases it was not clear from the article's timestamp alone whether publication occurred prior to or after the monetary event. However, in all instances it was straightforward to determine the article's timing relative to the monetary event from the article's content. The selected newswire articles tend to cover recent economic news, financial market movements, recent central bank communications, and quotes from market participants and other sources. As an example, the pre-event newswire published at 09:37am on 14 December 2023, two hours before the Bank of England's Monetary Policy Committee announcement that day, stated:

> *'Markets are all but certain neither will move their main policy rates but the focus is on how firmly rate-setters push back against market pricing of substantial interest rate cuts next year. . . . While the developments are unlikely to stop the BoE from pushing back again today against earlier rate cut expectations, if the softening growth inflation data continues into early next year it will encourage the BoE more quickly into a dovish policy pivot ...'*

A.2 Prompt engineering

To measure these economic concepts of interest in a reliable way, I followed emerging advice relating to prompt engineering.[11] For instance, each query includes a description of the concept of interest as well as explicit instructions regarding the desired output format. The latter is to avoid generation of unstructured free-text answers, which would not be usable in empirical analysis without further manual processing. Moreover, the set of multiple choice answers the language model is asked to select from is worded with the aim

[11] https://platform.openai.com/docs/guides/prompt-engineering

of providing a mutually exclusive and collectively exhaustive set of options. The model used to execute prompts is ‘Mixtral-8x7B-Instruct-v0.1’. As per the model card[12], the instruction part of all prompts are contained within ‘[INST]’ and ‘[/INST]’ tags for optimal results. One potential pitfall is that the language model, having been trained on a vast corpus of in recent years, ‘remembers’ what happened after each monetary event. To reduce the risk and/or extent of such ‘lookahead’ biasing results, all prompts are prefaced with ‘does the newswire state’. The resulting list of prompts is displayed in Tables 3 and 4.

A.3 Discussion of measurements of economic conditions obtained using large language model inference

The purpose of this section is to discuss and evaluate the measurements of UK economic conditions prior to monetary events obtained by prompting a large language model to answer multiple choice questions about newswire articles.

A.3.1 Perceived surprises in pre-event economic indicator readings

Inflation rate

Based on the measurements taken by prompting the language model using prompt number 1, the newswires for 151 out of the total 398 monetary events (38%) state that the most recent inflation rate reading was either higher or lower than expected. The inflation rate is therefore the indicator for which most surprises have been measured. Inspecting the first row of Figure

[12] https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1

Table 3: Prompts to measure perceived surprises in pre-event economic indicator readings

Prompt id	Concept measured	Prompt
1	Surprise indicator reading: inflation rate	<s>[INST] Beginning of newswire text: {text} End of newswire text Question: Does the newswire state that the most recent inflation rate reading was HIGHER or LOWER than expected? RESPOND WITH ONE OF THE FOLLOWING PHRASES (including square brackets): [YES, HIGHER] OR [YES, LOWER] OR [NO] [/INST]
2	Surprise indicator reading: GDP growth	<s>[INST] Beginning of newswire text: {text} End of newswire text Question: Does the newswire state that the most recent GDP growth reading was HIGHER or LOWER than expected? RESPOND WITH ONE OF THE FOLLOWING PHRASES (including square brackets): [YES, HIGHER] OR [YES, LOWER] OR [NO] [/INST]
3	Surprise indicator reading: purchasing manager index	<s>[INST] Beginning of newswire text: {text} End of newswire text Question: Does the newswire state that the most recent purchasing manager index reading was HIGHER or LOWER than expected? RESPOND WITH ONE OF THE FOLLOWING PHRASES (including square brackets): [YES, HIGHER] OR [YES, LOWER] OR [NO] [/INST]
4	Surprise indicator reading: unemployment rate	<s>[INST] Beginning of newswire text: {text} End of newswire text Question: Does the newswire state that the most recent unemployment rate reading was HIGHER or LOWER than expected? RESPOND WITH ONE OF THE FOLLOWING PHRASES (including square brackets): [YES, HIGHER] OR [YES, LOWER] OR [NO] [/INST]
5	Surprise indicator reading: wage growth	<s>[INST] Beginning of newswire text: {text} End of newswire text Question: Does the newswire state that the most recent wage growth reading was HIGHER or LOWER than expected? RESPOND WITH ONE OF THE FOLLOWING PHRASES (including square brackets): [YES, HIGHER] OR [YES, LOWER] OR [NO] [/INST]
6	Surprise indicator reading: consumer confidence	<s>[INST] Beginning of newswire text: {text} End of newswire text Question: Does the newswire state that the most recent consumer confidence reading was HIGHER or LOWER than expected? RESPOND WITH ONE OF THE FOLLOWING PHRASES (including square brackets): [YES, HIGHER] OR [YES, LOWER] OR [NO] [/INST]
7	Surprise indicator reading: business confidence	<s>[INST] Beginning of newswire text: {text} End of newswire text Question: Does the newswire state that the most recent business confidence reading was HIGHER or LOWER than expected? RESPOND WITH ONE OF THE FOLLOWING PHRASES (including square brackets): [YES, HIGHER] OR [YES, LOWER] OR [NO] [/INST]
8	Surprise indicator reading: retail sales	<s>[INST] Beginning of newswire text: {text} End of newswire text Question: Does the newswire state that the most recent retail sales reading was HIGHER or LOWER than expected? RESPOND WITH ONE OF THE FOLLOWING PHRASES (including square brackets): [YES, HIGHER] OR [YES, LOWER] OR [NO] [/INST]
9	Surprise indicator reading: trade balance	<s>[INST] Beginning of newswire text: {text} End of newswire text Question: Does the newswire state that the most recent trade balance reading was HIGHER or LOWER than expected? RESPOND WITH ONE OF THE FOLLOWING PHRASES (including square brackets): [YES, HIGHER] OR [YES, LOWER] OR [NO] [/INST]
10	Surprise indicator reading: mortgage approvals	<s>[INST] Beginning of newswire text: {text} End of newswire text Question: Does the newswire state that the most recent mortgage approvals reading was HIGHER or LOWER than expected? RESPOND WITH ONE OF THE FOLLOWING PHRASES (including square brackets): [YES, HIGHER] OR [YES, LOWER] OR [NO] [/INST]
11	Surprise indicator reading: house price	<s>[INST] Beginning of newswire text: {text} End of newswire text Question: Does the newswire state that the most recent house price reading was HIGHER or LOWER than expected? RESPOND WITH ONE OF THE FOLLOWING PHRASES (including square brackets): [YES, HIGHER] OR [YES, LOWER] OR [NO] [/INST]
12	Surprise indicator reading: public sector borrowing	<s>[INST] Beginning of newswire text: {text} End of newswire text Question: Does the newswire state that the most recent public sector borrowing reading was HIGHER or LOWER than expected? RESPOND WITH ONE OF THE FOLLOWING PHRASES (including square brackets): [YES, HIGHER] OR [YES, LOWER] OR [NO] [/INST]
13	Surprise indicator reading: exchange rate	<s>[INST] Beginning of newswire text: {text} End of newswire text Question: Does the newswire state that the most recent exchange rate reading was HIGHER or LOWER than expected? RESPOND WITH ONE OF THE FOLLOWING PHRASES (including square brackets): [YES, HIGHER] OR [YES, LOWER] OR [NO] [/INST]
14	Surprise indicator reading: volatility index	<s>[INST] Beginning of newswire text: {text} End of newswire text Question: Does the newswire state that the most recent volatility index reading was HIGHER or LOWER than expected? RESPOND WITH ONE OF THE FOLLOWING PHRASES (including square brackets): [YES, HIGHER] OR [YES, LOWER] OR [NO] [/INST]

Table 4: Prompts to measure perceived changes in economic risks and economic context

Prompt id	Concept measured	Prompt
15	Change in risk: recession	<s>[INST] Beginning of newswire text: {text} End of newswire text Question: Does the newswire state that recession risk INCREASED or DECREASED recently? RESPOND WITH ONE OF THE FOLLOWING PHRASES (including square brackets): [YES, INCREASED] OR [YES, DECREASED] OR [NO] [/INST]
16	Change in risk: supply chain	<s>[INST] Beginning of newswire text: {text} End of newswire text Question: Does the newswire state that supply chain risk INCREASED or DECREASED recently? RESPOND WITH ONE OF THE FOLLOWING PHRASES (including square brackets): [YES, INCREASED] OR [YES, DECREASED] OR [NO] [/INST]
17	Change in risk: financial crisis	<s>[INST] Beginning of newswire text: {text} End of newswire text Question: Does the newswire state that financial crisis risk INCREASED or DECREASED recently? RESPOND WITH ONE OF THE FOLLOWING PHRASES (including square brackets): [YES, INCREASED] OR [YES, DECREASED] OR [NO] [/INST]
18	Change in risk: geopolitical	<s>[INST] Beginning of newswire text: {text} End of newswire text Question: Does the newswire state that geopolitical risk INCREASED or DECREASED recently? RESPOND WITH ONE OF THE FOLLOWING PHRASES (including square brackets): [YES, INCREASED] OR [YES, DECREASED] OR [NO] [/INST]
19	Change in risk: wage-price spiral	<s>[INST] Beginning of newswire text: {text} End of newswire text Question: Does the newswire state that wage-price spiral risk INCREASED or DECREASED recently? RESPOND WITH ONE OF THE FOLLOWING PHRASES (including square brackets): [YES, INCREASED] OR [YES, DECREASED] OR [NO] [/INST]
20	Change in risk: sovereign default	<s>[INST] Beginning of newswire text: {text} End of newswire text Question: Does the newswire state that sovereign default risk INCREASED or DECREASED recently? RESPOND WITH ONE OF THE FOLLOWING PHRASES (including square brackets): [YES, INCREASED] OR [YES, DECREASED] OR [NO] [/INST]
21	Context for monetary communication: more closely watched than usual	<s>[INST] Beginning of newswire text: {text} End of newswire text Question: Does the newswire state that the upcoming Bank of England communication was more closely watched than usual? RESPOND WITH ONE OF THE FOLLOWING PHRASES (including square brackets): [YES] OR [NO] [/INST]
22	Context for monetary communication: more highly-anticipated than usual	<s>[INST] Beginning of newswire text: {text} End of newswire text Question: Does the newswire state that the upcoming Bank of England communication was more highly-anticipated than usual? RESPOND WITH ONE OF THE FOLLOWING PHRASES (including square brackets): [YES] OR [NO] [/INST]
23	Context for monetary communication: finely-balanced	<s>[INST] Beginning of newswire text: {text} End of newswire text Question: Does the newswire state that the upcoming Bank of England communication was finely-balanced? RESPOND WITH ONE OF THE FOLLOWING PHRASES (including square brackets): [YES] OR [NO] [/INST]
24	Context for monetary communication: taking place under crisis-like circumstances	<s>[INST] Beginning of newswire text: {text} End of newswire text Question: Does the newswire state that the upcoming Bank of England communication was taking place under crisis-like circumstances? RESPOND WITH ONE OF THE FOLLOWING PHRASES (including square brackets): [YES] OR [NO] [/INST]
25	Context for monetary communication: taking place under extraordinary circumstances	<s>[INST] Beginning of newswire text: {text} End of newswire text Question: Does the newswire state that the upcoming Bank of England communication was taking place under extraordinary circumstances? RESPOND WITH ONE OF THE FOLLOWING PHRASES (including square brackets): [YES] OR [NO] [/INST]

1, these inflation surprises appear approximately evenly spread throughout the sample period. Between 06/1997 and 06/2008 the output suggests that inflation readings were regularly higher and lower than expected. Between 07/2008 and 10/2013 all detected surprises relate to inflation readings that were higher than expected. From 11/2013 to 08/2016 readings were regularly lower than expected. This was followed by regular upward surprises between 11/2016 and 12/2017. The most notable pattern of regular upward inflation surprises was measured between 06/2021 and 09/2023, capturing the recent inflation surge.

GDP growth

For 58 out of the total 398 monetary events (15%) the language model measurements indicate GDP growth readings that differed from expectations. Row two of Figure 1 shows that these surprises were roughly evenly distributed across the sample period. An even split of upward and downward GDP growth surprises were measured. Compared to inflation rate readings there are fewer notable patterns, although regular upward surprises were measured between 11/2012 and 11/2014.

Purchasing manager index

The language model output indicates 49 readings of the purchasing manager index that differed from expectations (corresponding to 14% of monetary events). More than three in four of these surprises were labelled as upward surprises, as shown in row three of Figure 1. Virtually all surprises detected between 05/2002 and 12/2010 were measured to be purchasing manager index readings that were higher than expected. The only exceptions to this pattern were two downward surprises in 03/2008 and 06/2008.

Unemployment rate

Unexpected unemployment rate readings were rarely detected by the language model (4% monetary events in the sample), with most corresponding to higher-than-expected indicator readings. That is, newswires appear more likely to cover unemployment when it turned out higher than had been expected. As shown in row four of Figure 1, between 06/1997 and 01/2012 all detected surprises were upward surprises. This was followed by downward surprises between 05/2013 and 02/2014 and further upward surprises since 08/2020.

Wage growth

There were 40 detected surprises regarding wage growth readings, with almost all relating to higher growth readings than had been expected. A notable cluster of upward surprises is detected between 11/2022 and 06/2024, following the regular upward inflation surprises that were measured between 06/2021 and 09/2023. Row five of Figure 1 shows that further periods of stronger-than-expected wage growth were measured between 05/1998 and 08/2000, as well as between 03/2018 and 02/2019. Notably, wage growth is perceived as higher than expected in 2022 and 2023, following the 2021-2022 inflation surge.

Consumer confidence index

Based on the measurements taken, consumer confidence index readings were perceived to be surprising in 30 cases (7% of monetary events), with an even split between higher-than-expected and lower-than-expected readings. One pattern in row six of Figure 1 are repeated readings that were lower than expected in 2016, following immediately after the EU referendum.

Business confidence index

Row seven of Figure 1 shows that business confidence index readings were perceived as surprising in 40 cases, prior to one in 10 monetary events. There are more than twice as many upward surprises than downward surprises, with many lower than expected readings happening during recession periods. Similar to consumer confidence, business confidence shows a pattern of lower than expected readings following the EU referendum in June 2016.

Retail sales

There are 41 instances of surprising retail sales readings, as shown in row eight of Figure 1. There are slightly more higher-than-expected surprises compared to lower-than-expected ones. Most downside surprises tend to occur during the middle of recession periods, while most upside surprises are measured during non-recession periods.

Trade balance

Row nine of Figure 1 shows that there are few examples of surprising readings of the trade balance, with only eight instances being recorded throughout the sample period. The five higher-than-expected readings are measured outside of recessions, with the exception of one observation at the very end of the 2001 recession. Two out of the three lower-than-expected readings are measured during recessions.

Mortgage approvals

Mortgage approvals readings were measured to be surprising in 18 instances, with row 10 of Figure 1 showing that all but two were upside surprises. Both downward surprises were measured during the 2007–2008 recession, while nearly all upside surprises occurred outside of recession periods.

House price index

Surprising house price index values are detected 70 times in newswires published prior to monetary events, making it one of the most-discussed indicators. Over 50 of these perceived surprises were higher-than-expected house price index readings. Most instances of the house price index being lower than expected occurred during recessions, particularly the 2007–2008 recession as row 11 of Figure 1 shows.

Public sector borrowing

There are only three measured surprises regarding public sector borrowing outurns, with all three being surprises as to borrowing being higher than expected. As shown in row 12 of Figure 1, one higher-than-expected public sector borrowing surprise follows the 2007–2008 recession and two positive surprises follow the pandemic.

Exchange rate

Sterling exchange rate developments are discussed often, with a total of 78 instances of a value being discussed as surprising. There is a roughly even split between upside and downside surprises, as shown in row 13 of Figure 1. Weaker-than-expected readings are measured predominantly in recessions and around the 2016 referendum period. Upside surprises as to the strength of the pound tend to occur outside of recessionary periods.

Volatility index

Row 14 of Figure 1 shows measured surprises regarding the value of the volatility index. There are a total of 14 perceived surprises, with all but three relating to the volatility index being higher than had been expected. Interestingly, the lower-than-expected surprises regarding the index appear to occur around recessions.

A.3.2 Perceived surprises in pre-event changes in economic risks

Recession risk

Changes in the perceived risk of recession are discussed frequently, with a total of 150 instances being measured from newswire text. The majority of changes mentioned relate to increases in risk. As can be seen in row one of Figure 3, recession risk increases are measured towards the beginning of recessionary periods. For the pandemic period since 2020, multiple increases in recession risk were measured.

Supply chain risk

In contrast, perceived changes in supply-chain risks are less commonly detected—only 30 times during the sample period as shown in row two of Figure 3. Most changes mentioned relate to increases in risk. Notable periods of increased supply chain risk include the pandemic period, and the period following the EU referendum.

Financial crisis risk

Similarly to recession risk, changes in financial crisis risk are discussed frequently in the newswires. As row three of Figure 3 shows, most of these relate to financial crisis risk having increased. Notable periods of successive increases in financial crisis risks include the 2007–2008 and 2020 recessions, the eurozone crisis, as well as the Asian financial crisis.

Geopolitical risk

Perceived changes in geopolitical risks are detected in 71 newswires. Row four of Figure 3 shows that most of these instances relate to perceived increases in geopolitical risk. The frequency of the detected risk changes increases over time, with most falling into the period from 2016 onwards.

Wage-price spiral risk

Row five of Figure 3 shows 91 instances of a perceived change in wage-price spiral risk being mentioned in a newswire. The majority of these risk changes relate to perceived risk increases. A notable period of successive increases of this risk type being perceived is the inflationary period during the recent pandemic.

Sovereign default risk

Changes in sovereign default risk are detected infrequently, with only 12 examples being found during the sample period. All of these instances are perceived increases in sovereign default risk, as shown in row six of Figure 3. Several of these increases are detected for monetary events around the time of the eurozone crisis.

A.3.3 Perceived overall pre-event economic context

More closely-watched than usual

Row one of Figure 5 shows that more than half of the newswires are classified as mentioning that the monetary event was more closely-watched than usual. During the pandemic period, almost all events fall into this category.

More highly-anticipated than usual

Similarly, as per row two of Figure 3, there are many instances of the monetary event being perceived to be more highly-anticipated than usual. Again, most events during the pandemic are seen to be highly anticipated.

Finely-balanced

In 48 cases, the monetary communication was perceived to be finely-balanced beforehand. Row three of Figure 5 shows the incidence of these events over time, several of which fall into recessionary periods.

Taking place under crisis-like circumstances

Row four of the same figure shows the distribution of the 26 instances where the context prior to the monetary event was perceived to be ‘crisis-like’. The indicators align with recessions, periods of political uncertainty, and, most notably, the pandemic period.

Taking place under extraordinary circumstances

Similarly, row five shows the 76 instances where the monetary event was perceived to be taking place under extraordinary circumstances. As was the case for the previous prompt, these observations align with recessions, periods of political uncertainty, and the pandemic period.

A.4 Ex post predictability of high-frequency market reactions using measured economic conditions

In Section 2.3, the measurement of economic conditions—specifically perceived surprises in pre-event economic indicator readings, perceived changes in economic risks, and perceived overall economic context—prior to each UK monetary event between 1997–2024 is described. In this section, the statistical relationships between these new text-derived discrete/dummy variables and the high-frequency reactions collected by Braun et al. (2025) of financial markets—gilt yields, interest rate futures, overnight indexed swaps, equity indices, and exchange rates—are analysed. In Section A.4.1, the analysis focuses on the extent to which the text-derived measurements of economic conditions are, ex post, predictive of the *magnitude* of market reactions. Section A.4.2 studies the extent to which there is ex post predictability of the *direction* of market reactions. Considering the text-derived dummy variables individually, many are found to have significant ex post predictive power for both the magnitude and direction of interest rate markets' reactions to UK monetary events. This is the case regardless of whether these interest rates are measured using gilt yields, interest rate futures, or overnight indexed swaps. Perceived pre-event surprises in public sector borrowing, for example, appear to carry statistically and economically significant information about interest rate markets' subsequent reactions.

A.4.1 Ex post correlations between measured economic conditions and monetary policy surprises

As discussed in Section 2.3, each prompt gives rise to a discrete variable with low cardinality—there are either two or three possible answers per prompt. These discrete measurements can be converted via 'one-hot encoding' to multiple binary dummy variables, one for each possible answer for a given prompt. Having generated the set of dummy variables, they are used as predictors in the following linear regressions

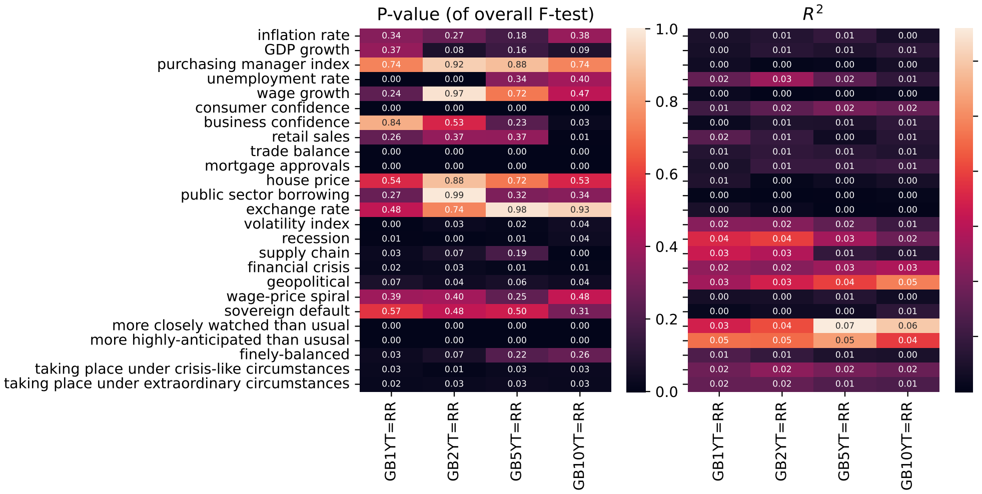
$$|mps_t| = \sum \beta_a 1_{answer=a} + u_t \tag{15}$$

where $|mps_t|$ is the absolute value of the financial market reaction, β_a is the coefficient on the dummy variable for answer a and $1_{answer=a}$ is the dummy variable for answer a. An intercept is not needed by construction. Regressions of this form are estimated for each prompt-financial market pair. The p-value of the F-test of overall significance and the R-squared for each regression are presented in a heatmap to expose patterns of ex post predictability.

Gilt yields

In Figure 21, results for regressions with the *magnitude* of reactions of 1, 2, 5, and 10-year gilt yields as dependent variables are shown. Braun et al. (2025) label these as GB1YT=RR, GB2YT=RR, GB5YT=RR, and GB10YT=RR, respectively—corresponding to the Refinitiv Instrument Code (RIC). The left panel shows p-values of the overall F-test, while the right panel shows the R-squared for each regression.

Figure 21: Ex post relationship with gilt market reaction magnitudes



P-value (of overall F-test)

	GB1YT=RR	GB2YT=RR	GB5YT=RR	GB10YT=RR
inflation rate	0.34	0.27	0.18	0.38
GDP growth	0.37	0.08	0.16	0.09
purchasing manager index	0.74	0.92	0.88	0.74
unemployment rate	0.00	0.00	0.34	0.40
wage growth	0.24	0.97	0.72	0.47
consumer confidence	0.00	0.00	0.00	0.00
business confidence	0.84	0.53	0.23	0.03
retail sales	0.26	0.37	0.37	0.01
trade balance	0.00	0.00	0.00	0.00
mortgage approvals	0.00	0.00	0.00	0.00
house price	0.54	0.88	0.72	0.53
public sector borrowing	0.27	0.99	0.32	0.34
exchange rate	0.48	0.74	0.98	0.93
volatility index	0.00	0.03	0.02	0.04
recession	0.01	0.00	0.01	0.04
supply chain	0.03	0.07	0.19	0.00
financial crisis	0.02	0.03	0.01	0.01
geopolitical	0.07	0.04	0.06	0.04
wage-price spiral	0.39	0.40	0.25	0.48
sovereign default	0.57	0.48	0.50	0.31
more closely watched than usual	0.00	0.00	0.00	0.00
more highly-anticipated than ususal	0.00	0.00	0.00	0.00
finely-balanced	0.03	0.07	0.22	0.26
taking place under crisis-like circumstances	0.03	0.01	0.03	0.03
taking place under extraordinary circumstances	0.02	0.03	0.03	0.03

R^2

	GB1YT=RR	GB2YT=RR	GB5YT=RR	GB10YT=RR
inflation rate	0.00	0.01	0.01	0.00
GDP growth	0.00	0.01	0.01	0.01
purchasing manager index	0.00	0.00	0.00	0.00
unemployment rate	0.02	0.03	0.02	0.01
wage growth	0.00	0.00	0.00	0.00
consumer confidence	0.01	0.02	0.02	0.02
business confidence	0.00	0.01	0.01	0.01
retail sales	0.02	0.01	0.00	0.01
trade balance	0.01	0.01	0.01	0.01
mortgage approvals	0.00	0.01	0.01	0.01
house price	0.01	0.00	0.00	0.00
public sector borrowing	0.00	0.00	0.00	0.00
exchange rate	0.00	0.00	0.00	0.00
volatility index	0.02	0.02	0.02	0.01
recession	0.04	0.04	0.03	0.02
supply chain	0.03	0.03	0.01	0.01
financial crisis	0.02	0.02	0.03	0.03
geopolitical	0.03	0.03	0.04	0.05
wage-price spiral	0.00	0.00	0.01	0.00
sovereign default	0.00	0.00	0.00	0.01
more closely watched than usual	0.03	0.04	0.07	0.06
more highly-anticipated than ususal	0.05	0.05	0.05	0.04
finely-balanced	0.01	0.01	0.01	0.00
taking place under crisis-like circumstances	0.02	0.02	0.02	0.02
taking place under extraordinary circumstances	0.02	0.02	0.01	0.01

Notes: One linear regression relating the magnitude of the gilt market reaction to prompt answer dummy variables as in Equation 15 is estimated for each gilt maturity and prompt. The heatmap on the left shows the resulting p-values of the overall F-test, while the heatmap on the right shows the R-squared for each regression.

Considering the left panel, many of the measures of economic conditions extracted from text have statistically significant ex post predictive power regarding the magnitude of gilt yield market reactions to monetary events. Perceived surprises in pre-event readings of consumer confidence, trade balance, mortgage approvals, and the volatility index, for instance, appear relevant irrespective of gilt maturity. Similarly, perceived changes in recession, financial crisis, and geopolitical risks have significant ex post associations with all gilt maturities. The perceived overall economic context, such as whether the monetary event was more closely watched than usual, more highly-anticipated than usual, or taking place under crisis-like circumstances is also significant at the 5% level for all gilt yield maturities considered. Considering the right panel, the share of variance in gilt yield reactions explained by different prompt responses ranges from 0 to 7%. Of the measures relating to overall perceived context, how closely-watched or highly-anticipated the monetary event was appears most relevant. Perceived changes in geopolitical, recession, and financial crisis risks also appear relevant. Perceived surprises in indicator readings appear to explain a smaller share of gilt yield reactions, with the exception of unemployment rate and consumer confidence surprises.

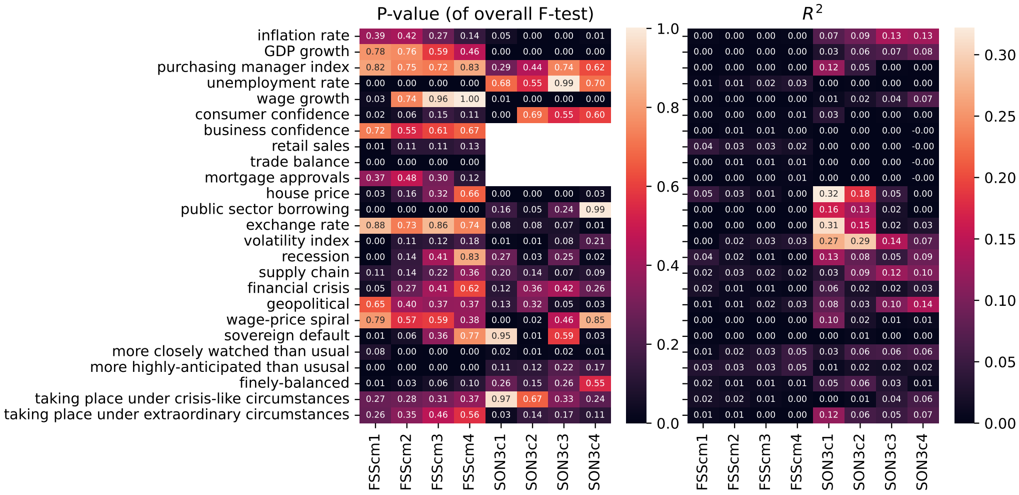
Interest rate futures

In Figure 22, results for regressions with the *magnitude* of reactions of the first, second, third and fourth 3-month quarterly LIBOR/SONIA[13] interest rate futures as dependent variables are shown. Aligned with Braun et al. (2025), these are labelled with their RICs FSScm1–4 and SON3c1–4.

Considering the left panel, during both the LIBOR and the SONIA periods, several of the measures of economic conditions extracted from text have statistically significant ex post predictive power regarding the magnitude of interest rate futures' reactions to monetary events. However, which of the readings are significant varies by subsample. For instance, perceived surprises in pre-event readings of the unemployment rate, trade balance, and public sector borrowing are significant at the 1% level for all four futures contracts during the LIBOR period. During the SONIA period, in contrast, it is perceived surprises in GDP growth, wage growth, and house price readings that

[13] Due to the transition away from LIBOR during the sample period, the LIBOR reaction data are available for 1997–2021, while the SONIA contracts are available from 2021 onwards in the Braun et al. (2025) database.

Figure 22: Ex post relationship with interest rate futures market reaction magnitudes



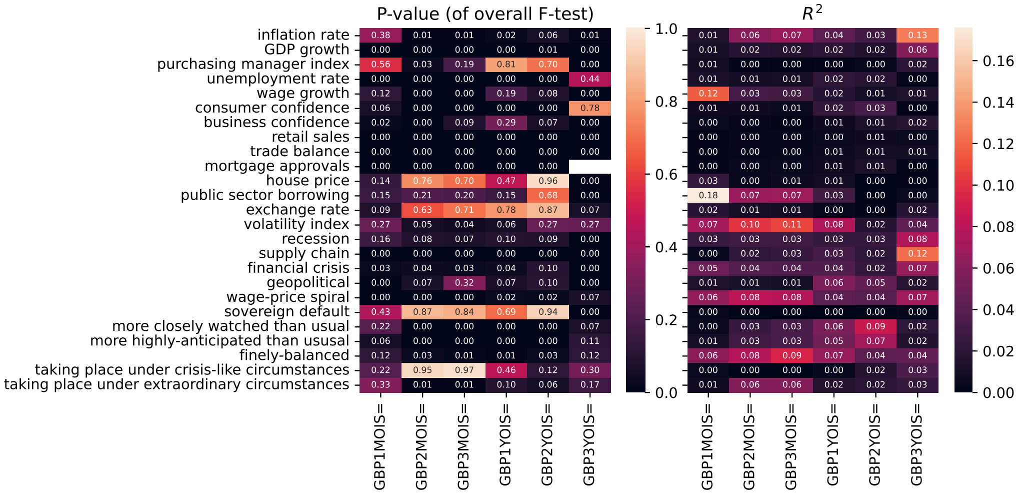
Notes: One linear regression relating the magnitude of the interest rate futures market reaction to prompt answer dummy variables as in Equation 15 is estimated for each futures contract and prompt. The heatmap on the left shows the resulting p-values of the overall F-test, while the heatmap on the right shows the R-squared for each regression. Empty cells are the result of certain dummy variables having zero variance for the period where SONIA futures data are available.

are significant predictors of the magnitude of interest rate futures reactions. It should be noted that due to the relatively small sample size for the SONIA period, for some prompts (business confidence, retail sales, trade balance, and mortgage approvals) there is zero variation in the dummy variables. Changes in perceived recession, sovereign default, and financial crisis risk appear to be significant predictors (at the 5% level) of the magnitude of front-month futures during the LIBOR period but not the SONIA period. The absence of statistical significance could be a result of low statistical power due to a relatively small sample size for the SONIA period. Across both the LIBOR and SONIA period, the magnitude of all futures contracts' reactions to monetary events is significantly associated (at 5% level) with the dummy variable measuring whether the event was perceived to be more closely-watched than usual. Other measures of perceived overall economic context are less consistently associated with interest rate futures reaction magnitudes. Considering the right panel, the share of variance in interest rate futures reactions explained by different prompt responses ranges from 0 to 32%. However, for LIBOR data (which cover the longer time period), R-squared ranges from 0-5%. During the LIBOR sample, most variation is explained by the 'more highly-anticipated than usual'-dummy variable. Perceived surprises in retail sales and house price readings, as well as perceived changes in recession risk appear to have some ex post association during the same period. During the shorter SONIA period, some text-derived measures appear to explain a substantial share of variation ex post. For instance, dummy variable measures of perceived surprises in house price, exchange rate, volatility index, and public sector borrowing readings explain between 16-32% of variation in front-month SONIA contract reaction magnitudes.

Overnight indexed swaps

In Figure 23, results for regressions with the *magnitude* of reactions of the 1, 2, 3, 12, 24, and 36-month overnight indexed swap (OIS) rates[14] as dependent variables are shown. Aligned with Braun et al. (2025), these are labelled with their RICs GBP1–3MOIS=RR (for 1, 2, and 3-month contracts) and GBP1-3YOIS=RR (for the 1, 2, and 3-year contracts).

Figure 23: Ex post relationship with overnight indexed swap market reaction magnitudes



Notes: One linear regression relating the magnitude of the overnight indexed swap market reaction to prompt answer dummy variables as in Equation 15 is estimated for each swap maturity and prompt. The heatmap on the left shows the resulting p-values of the overall F-test, while the heatmap on the right shows the R-squared for each regression. Empty cells are the result of certain dummy variables having zero variance for the period where swap data are available.

Considering the left panel, most of the measures of economic conditions extracted have statistically significant ex post predictive power regarding the

[14] In the Braun et al. (2025) database these reaction data are available from 2009 onwards, with the exception of the 36-month swap which is covered from 2015 onwards.

magnitude of OIS market reactions to monetary events. This includes, for instance, perceived surprises in indicator readings regarding GDP growth, retail sales, trade balance, unemployment rate, and mortgage approvals. Perceived changes in supply chain, wage-price spiral, financial crisis, and geopolitical risks also have sigificant ex post associations. Measures of the perceived overall economic context, such as how closely-watched or highly-anticipated the monetary event was, are also significant for most swap maturities. Considering the right panel, the share of variance in OIS reactions explained by different prompt responses ranges from 0 to 18%. For short maturities, such as the 1-month swap, perceived surprises in the readings of public sector borrowing, wage growth, and the volatility index have the greatest ex post explanatory power for the magnitude of reactions. For longer maturities, such as the 3-year swap, perceived surprises in the inflation rate readings, and perceived changes in supply chain and recession risks explain the most variance.

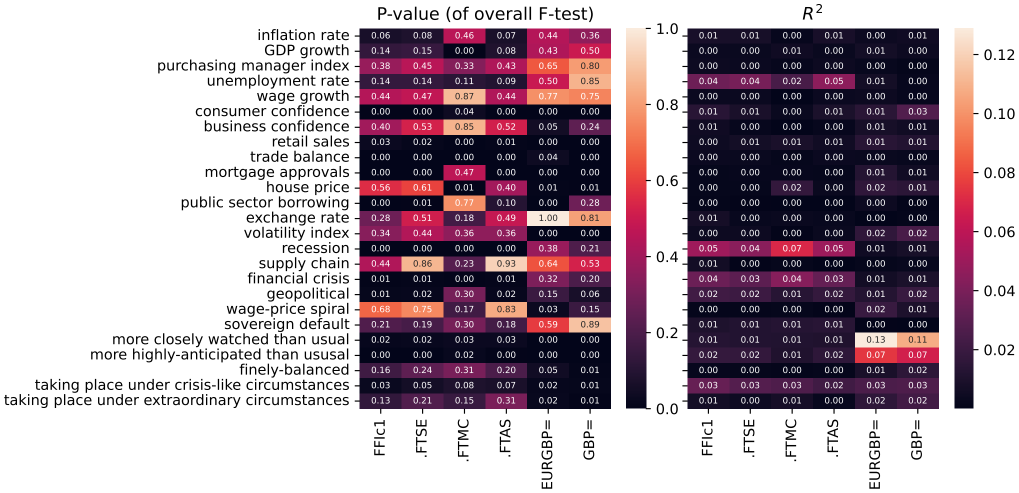
Equity and foreign exchange markets

In Figure 24, results for regressions with the *magnitude* of the Braun et al. (2025) reactions of a range of stock market indices/futures and exchange rates as dependent variables are shown. Specifically covered are the FTSE100 future first month contract (RIC: FFIc1), the FTSE100 index (RIC: .FTSE), the FTSE250 index (RIC: .FTMC), the FTSE All Share index (RIC: .FTAS), the EUR/GBP exchange rate (RIC: EURGBP=), and the GBP/USD exchange rate (RIC: GBP=).[15]

Considering the left panel, several of the measures of economic conditions extracted have statistically significant ex post predictive power regarding the magnitude of both equity and exchange rate market reactions to monetary events. This includes, for example, perceived surprises in consumer con-

[15] These reaction data are available from 1997 onwards, with the exception of the EUR/GBP exchange rate which is covered from 1998 onwards.

Figure 24: Ex post relationship with equity and foreign exchange market reaction magnitudes



Notes: One linear regression relating the magnitude of the equity and foreign exchange markets' reaction to prompt answer dummy variables as in Equation 15 is estimated for each market and prompt. The heatmap on the left shows the resulting p-values of the overall F-test, while the heatmap on the right shows the R-squared for each regression.

fidence, retail sales, and trade balance readings. Overall economic context measures, such as whether the upcoming monetary event was perceived to be more closely-watched or highly anticipated than usual, are also significantly associated. Equity market reaction magnitudes specifically have significant ex post associations with perceived changes in recession, financial crisis and geopolitical risk. Exchange rate reaction magnitudes have significant ex post associations with half of the text-extracted measures of UK economic conditions. Considering the right panel, perceived changes in recession and financial crisis risk, as well as perceived surprises in pre-event unemployment rate readings, explain the greatest share in equity market reaction magnitudes. Measures of the overall context, specifically whether the event was more closely-watched or highly-anticipated than usual, explain around 10% of variance in exchange rate reaction magnitudes.

A.4.2 Ex post predictability of market reactions

The analysis in this section proceeds analogously to that in Section A.4.1, apart from the difference that the dependent variable is now the raw market reaction as opposed to its absolute value. As such, the regressions estimated for each prompt-financial market pair are now

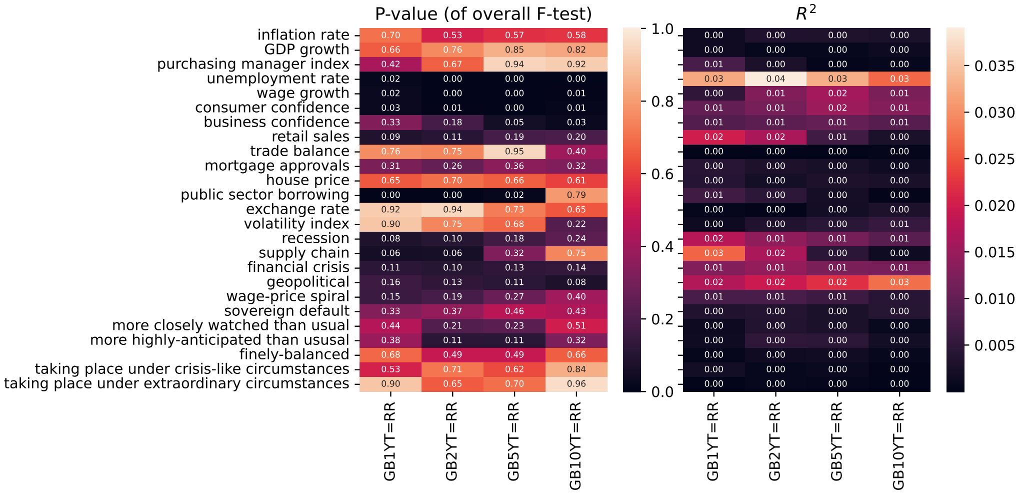
$$mps_t = \sum \beta_a 1_{answer=a} + u_t \tag{16}$$

where mps_t is the raw financial market reaction, β_a is the coefficient on the dummy variable for answer a and $1_{answer=a}$ is the dummy variable for answer a. An intercept is not needed by construction. Again, the p-value of the F-test of overall significance and the R-squared for each regression are presented in a heatmap to expose patterns of ex post predictability.

Gilt yields

In Figure 25, p-values of the overall F-test and the R-squared for regressions with the *raw directional* reactions of 1, 2, 5, and 10-year gilt yields (RICs: GB1–10YT=RR) as dependent variables are shown.

Figure 25: Ex post relationship with gilt market reactions



Notes: One linear regression relating the gilt market reaction to prompt answer dummy variables as in Equation 16 is estimated for each gilt maturity and prompt. The heatmap on the left shows the resulting p-values of the overall F-test, while the heatmap on the right shows the R-squared for each regression.

Considering the left panel, perceived surprises in unemployment rate, wage growth, consumer confidence, business confidence, retail sales, and public sector borrowing readings all have statistically significant ex post associations with the reaction of gilt yields to the monetary events studied. Perceived changes in recession, geopolitical, financial crisis, and supply chain risks also appear relevant but are not significant at conventional levels. In contrast to findings regarding the predictability of *magnitudes* of gilt yield reactions, measures of the perceived overall economic context are not statistically significant ex post predictors. Considering the right panel, the share of variance in gilt yield reactions explained by different prompt responses ranges from 0 to 4%. Among measures of perceived surprises of indicator readings, unemployment rate, wage growth, consumer confidence, and retail sales explain the largest share of variance. Perceived changes in geopolitical, supply chain, recession, and financial crisis risks also appear to have some explanatory power ex post. Measures of the perceived overall economic context do not appear to be relevant.

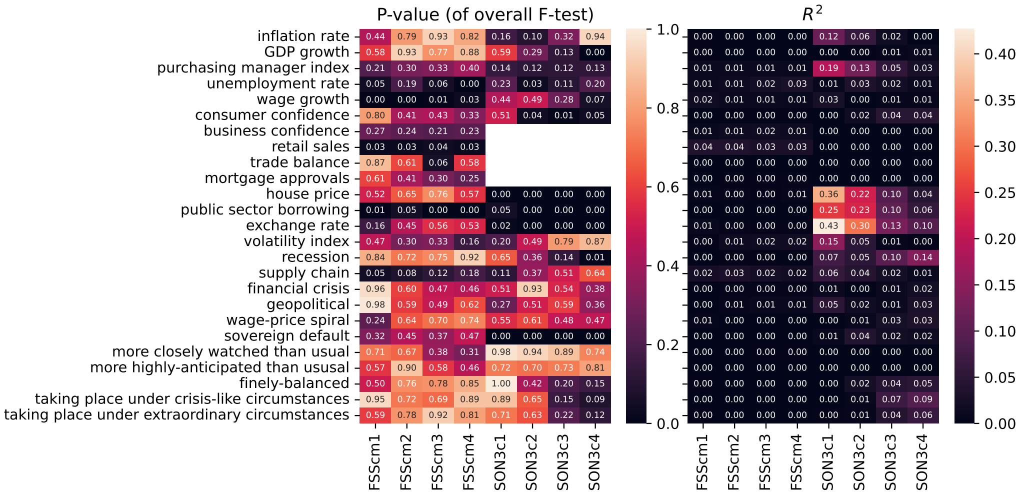
Interest rate futures

In Figure 26, p-values of the overall F-test and the R-squared for regressions with the *raw directional* reactions of first, second, third and fourth 3-month quarterly LIBOR/SONIA interest rate futures (RICs: FSScm1–4 and SON3c1–4) as dependent variables are shown.

Considering the left panel, perceived surprises in public sector borrowing are a significant predictor during both the LIBOR and SONIA periods. For the period of time where LIBOR data are available, perceived surprises in wage growth, retail sales, and unemployment rate readings also have a statistically significant relationship with futures market reactions to monetary events. During the SONIA period, perceived changes in sovereign default risk are significant predictors, as are perceived surprise readings of house prices and exchange rates.

Considering the right panel, the share of variance in interest rate futures reactions explained by different prompt responses ranges from 0 to 43%. During the LIBOR period, however, the maximum R-squared is only 4%, corresponding to the variance explained by perceived retail sales surprises. As in the previous section, some of the predictors explain a substantial share of variation in SONIA responses ex post. For example, measures of perceived surprises in exchange rate, house price, public sector borrowing, and purchasing manager index explain between 19-43% of variation in front-month

Figure 26: Ex post relationship with interest rate futures market reactions



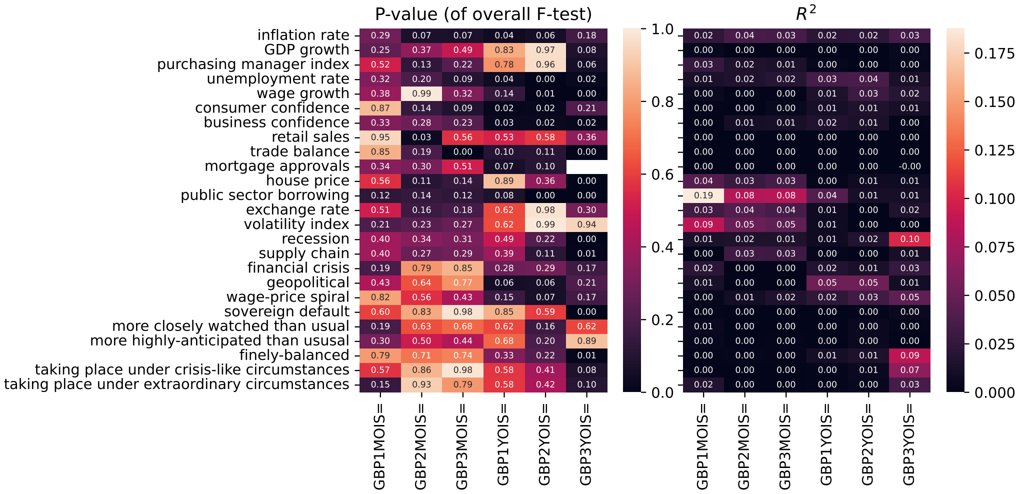
Notes: One linear regression relating the interest rate futures market reaction to prompt answer dummy variables as in Equation 16 is estimated for each futures contract and prompt. The heatmap on the left shows the resulting p-values of the overall F-test, while the heatmap on the right shows the R-squared for each regression. Empty cells are the result of certain dummy variables having zero variance for the period where SONIA futures data are available.

SONIA contract reactions.

Overnight indexed swaps

In Figure 27, p-values of the overall F-test and the R-squared for regressions with the *raw directional* reactions of the 1, 2, 3, 12, 24, and 36-month OIS rates (RICs: GBP1–3MOIS=RR and GBP1-3YOIS=RR) as dependent variables are shown.

Figure 27: Ex post relationship with overnight indexed swap market reactions



Notes: One linear regression relating the overnight indexed swap market reaction to prompt answer dummy variables as in Equation 16 is estimated for each swap maturity and prompt. The heatmap on the left shows the resulting p-values of the overall F-test, while the heatmap on the right shows the R-squared for each regression. Empty cells are the result of certain dummy variables having zero variance for the period where swap data are available.

Considering the left panel, none of the measures of perceived economic conditions extracted from text have strongly significant (i.e. at 5% level) ex post associations with *shorter-term* overnight indexed swap reactions to monetary events. For 1, 2, and 3-year swaps, however, several measures are significant. This includes perceived surprises in inflation rate, unemployment rate, business confidence, and public sector borrowing readings. For 3-year swaps measures perceived changes in recession, sovereign default, and supply chain risks also have significant ex post associations with reactions, as does the ‘finely-balanced decision’-measure.

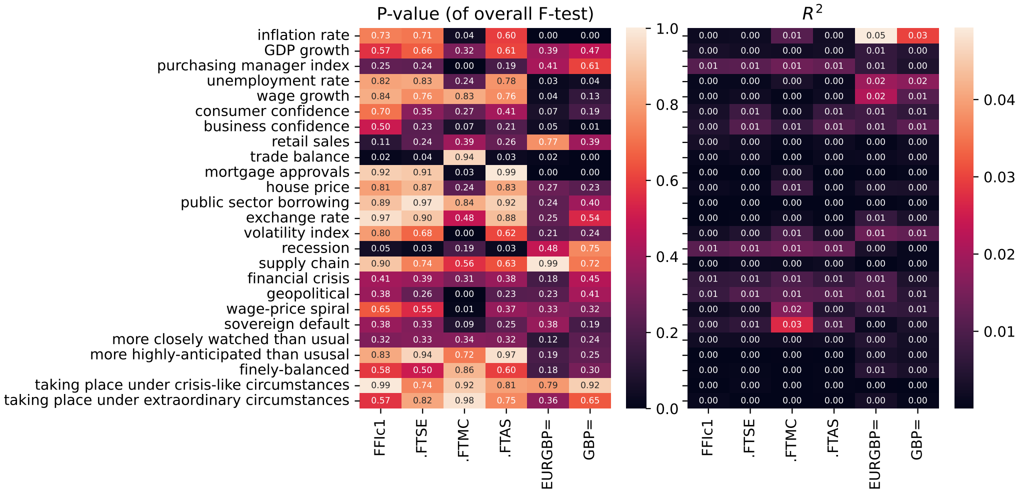
Considering the right panel, the share of variance of swap markets’ reactions explained by different prompt responses ranges from 0 to 19%. For shorter-term OIS (1-3 months), perceived surprises in public sector borrowing, volatility, and house price readings appear to explain the greatest amount of variation—although, as mentioned above, none are statistically significant. For longer-term OIS—especially the 3-year swap—perceived changes in recession, wage-price spiral, and geopolitical risks appear most relevant. Measures of the perceived overall context, such as whether the monetary decision was anticipated to be finely balanced or taking place under unusual circumstances also appear to have some ex post explanatory power for 3-year swap responses.

Equity and foreign exchange markets

In Figure 28, p-values of the overall F-test and the R-squared for regressions with the *raw directional* reactions of the FTSE100 future first month contract (RIC: FFIc1), the FTSE100 index (RIC: .FTSE), the FTSE250 index (RIC: .FTMC), the FTSE All Share index (RIC: .FTAS), the EUR/GBP exchange rate (RIC: EURGBP=), and the GBP/USD exchange rate (RIC: GBP=) as dependent variables are shown.

Considering the left hand side, both equity and exchange rate markets’ reactions have ex post statistical associations with some of the text-derived measures. For instance, perceived surprises in the pre-event trade balance reading are significantly associated with all but one of the markets considered (with the exception being the FTSE250 index). For equity market reactions,

Figure 28: Ex post relationship with equity and foreign exchange market reactions



Notes: One linear regression relating the equity and foreign exchange markets' reaction to prompt answer dummy variables is as in Equation 16 estimated for each market and prompt. The heatmap on the left shows the resulting p-values of the overall F-test, while the heatmap on the right shows the R-squared for each regression.

dummy variables indicating perceived changes in recession risks are also significant indicators. For exchange rate market reactions, perceived surprises in inflation rate, unemployment rate, business confidence, and mortgage approvals are statistically signficiant.

Considering the right hand side, perceived surprises in inflation rate, unemployment rate, wage growth, as well as perceived changes in recession risk account for the greatest amount of variation in exchange rate markets' reactions to monetary events. For equity market reactions, dummy variables measuring surprising purchasing manager index readings, perceived changes in recession, financial crisis, geopolitical, and sovereign default risk are among the most relevant predictors.

A.5 Details of orthogonalisation

A.5.1 Detailed OLS regression results

Table 5 shows the results of estimating five different model specifications to predict the 1-year gilt yield reactions, using between 0-4 lags of text-derived dummy variables capturing pre-event information. The number of variables in the different specifications ranges from 43 (no lags) to 215 (four lags). For brevity, coefficients on lagged variables are not shown. The R-squared ranges from 15% for the specification with no lags to over 60% for the specification with four lags. More than half of the variation in 1-year gilt yield reactions to UK monetary events is explained by pre-event information and three lags thereof. Despite the large number of variables relative to the sample size, some dummy variables are sigificant across several specifications. Firstly, the coefficient on the dummy variable indicating that public sector borrowing had been higher than expected before the monetary event is significantly positive across the 0, 1, 2, and 3-lag specifications. At between 5-7 basis points, the coefficient is also economically significant. In the 4-lag specification the coefficient remains similar but is not statistically signifi-

Table 5: Estimated $\hat{\beta}_{OLS}$ for 1-year yield reactions

	Dependent variable: 1-year gilt yield reactions compiled by Braun et al. (2025) [mps1$_t$]				
	(1)	(2)	(3)	(4)	(5)
GDP growth_[YES, HIGHER]	-0.010	-0.007	-0.008	-0.013	-0.019
GDP growth_[YES, LOWER]	-0.000	-0.004	-0.001	0.002	-0.006
business confidence_[YES, HIGHER]	0.013	0.014	0.025	0.030*	0.033
business confidence_[YES, LOWER]	0.001	0.001	0.009	0.024	0.017
consumer confidence_[YES, HIGHER]	0.004	0.003	-0.017	-0.031	-0.023
consumer confidence_[YES, LOWER]	-0.020	-0.023	-0.023	-0.024	-0.025
exchange rate_[YES, HIGHER]	-0.009	-0.014	-0.017	-0.021*	-0.016
exchange rate_[YES, LOWER]	0.012	0.013	0.005	-0.003	-0.003
financial crisis_[YES, DECREASED]	0.013	0.017	0.021	0.014	0.014
financial crisis_[YES, INCREASED]	0.001	-0.004	-0.001	-0.002	-0.007
finely-balanced_[YES]	-0.001	0.001	-0.000	0.000	0.000
geopolitical_[YES, DECREASED]	-0.042	-0.045	-0.061*	-0.074**	-0.074**
geopolitical_[YES, INCREASED]	-0.013*	-0.013	-0.010	-0.012	-0.009
house price_[YES, HIGHER]	-0.006	-0.004	0.003	-0.001	0.001
house price_[YES, LOWER]	0.012	0.013	0.002	0.009	0.010
inflation rate_[YES, HIGHER]	-0.002	-0.000	-0.002	-0.008	-0.007
inflation rate_[YES, LOWER]	-0.002	0.000	0.006	0.005	0.012
more closely watched than usual_[YES]	-0.004	-0.001	0.001	0.002	0.004
more highly-anticipated than ususal_[YES]	-0.001	-0.002	-0.004	-0.006	-0.003
mortgage approvals_[YES, HIGHER]	0.007	0.005	0.010	0.018	0.012
mortgage approvals_[YES, LOWER]	-0.008	-0.019	-0.062	-0.060	-0.082
public sector borrowing_[YES, HIGHER]	0.069***	0.059**	0.068**	0.052**	0.046
purchasing manager index_[YES, HIGHER]	0.004	-0.000	0.003	0.004	-0.004
purchasing manager index_[YES, LOWER]	-0.005	0.011	0.012	-0.002	0.010
recession_[YES, DECREASED]	-0.004	-0.001	-0.009	-0.006	-0.010
recession_[YES, INCREASED]	-0.005	-0.001	-0.002	-0.003	-0.004
retail sales_[YES, HIGHER]	-0.019	-0.015	-0.024	-0.026*	-0.021
retail sales_[YES, LOWER]	-0.022	-0.017	-0.014	-0.003	0.001
sovereign default_[YES, INCREASED]	-0.009	-0.009	-0.009	-0.002	-0.002
supply chain_[YES, DECREASED]	-0.030	-0.035*	-0.035**	-0.028	-0.049**
supply chain_[YES, INCREASED]	-0.025*	-0.025	-0.023	-0.015	-0.018
taking place under crisis-like circumstances_[YES]	-0.006	-0.004	-0.011	-0.004	0.017
taking place under extraordinary circumstances_[YES]	0.008	0.005	0.006	0.008	0.007
trade balance_[YES, HIGHER]	-0.002	0.008	0.008	0.021	0.023
trade balance_[YES, LOWER]	0.009	0.027	0.040	0.003	0.019
unemployment rate_[YES, HIGHER]	-0.049**	-0.044**	-0.051**	-0.049**	-0.029
unemployment rate_[YES, LOWER]	-0.010	-0.004	0.011	0.014	0.019
volatility index_[YES, HIGHER]	-0.009	-0.002	-0.007	-0.003	-0.013
volatility index_[YES, LOWER]	-0.003	-0.007	0.001	0.022	0.030
wage growth_[YES, HIGHER]	-0.005	-0.001	-0.004	-0.003	-0.004
wage growth_[YES, LOWER]	0.011	0.019	0.019	0.012	0.017
wage-price spiral_[YES, DECREASED]	0.021	0.016	0.021	0.006	-0.003
wage-price spiral_[YES, INCREASED]	0.003	0.005	0.004	0.006	0.007
1st lags	No	Yes	Yes	Yes	Yes
2nd lags	No	No	Yes	Yes	Yes
3rd lags	No	No	No	Yes	Yes
4th lags	No	No	No	No	Yes
Observations	398	398	398	398	398
R^2	0.152	0.261	0.368	0.516	0.606
F Statistic	1.415** (df=43; 354)	1.648*** (df=86; 311)	1.854*** (df=129; 268)	11.378*** (df=172; 225)	4.549*** (df=215; 182)

*p<0.1; **p<0.05; ***p<0.01

Notes: Results of ordinary least squares regression of 1-year gilt reaction to UK monetary events within Braun et al. (2025) database on text-derived dummy variables capturing pre-event perceived UK economic conditions. Coefficients shown are for the level of the extracted measures. Coefficients for lags (if applicable) are omitted for brevity. Significance levels are based on heteroscedasticity-consistent standard errors.

cant. This finding suggests that the Bank tended to act more hawkishly to expansionary fiscal policies than markets anticipated. Secondly, the coefficient on the dummy variable indicating that the unemployment rate had been higher than expected prior to the event is consistently negative across the first four specifications. Again, the 4-lag specification results in a qualitatively similar estimate that is not signficiant at conventional levels. This suggests that markets tended to underestimate the Bank's responsiveness to the labour market loosening. The coefficients on this variable are around five basis points. Thirdly, the coefficient on the measure of reduced geopolitical risks is consistently negative, although it is statistically significant only for the 2, 3, and 4-lag specifications.

Table 6 shows the results of using the same specifications to predict the 2-year gilt yield reactions to UK monetary events. The number of variables in the different specifications is unchanged compared to Table 5. Similar to Table 5, in Table 6 the R-squared ranges from 14% for the specification with no lags to over 60% for the specification with four lags. Again, more than half of the variation in 2-year gilt yield reactions to UK monetary events is explained by pre-event information and three lags thereof. There are similar patterns of significance for higher-than-expected unemployment rate readings and higher-than-expected public sector borrowing as in Table 5, with coefficients around -5 basis points and +5 basis points, respectively. In addition, the dummy variable indicating that supply chain risks decreased has a significantly negative coefficient in all but one specification (around -3 basis points), suggesting that market participants were surprised by the Bank's responsiveness to the easing of such risks.

Table 7 provides the estimates resulting from regressing the 5-year gilt yield responses on the dummy variable measures. R-squared ranges from 13% to 62% across the specifications, naturally increasing with the number of lags included in the explanatory variable set. Similar to Tables 5 and 6, three lags are sufficient to explain more than 50% of variation in the high-

Table 6: Estimated $\hat{\beta}_{OLS}$ for 2-year yield reactions

	Dependent variable: 2-year gilt yield reactions compiled by Braun et al. (2025) [$mps2_t$]				
	(1)	(2)	(3)	(4)	(5)
GDP growth_[YES, HIGHER]	-0.008	-0.006	-0.009	-0.016	-0.024*
GDP growth_[YES, LOWER]	0.000	-0.005	-0.003	-0.001	-0.009
business confidence_[YES, HIGHER]	0.014	0.014	0.019	0.025*	0.027
business confidence_[YES, LOWER]	-0.012	-0.011	-0.002	0.004	0.002
consumer confidence_[YES, HIGHER]	0.006	0.007	-0.007	-0.022	-0.015
consumer confidence_[YES, LOWER]	-0.019	-0.022	-0.024	-0.021	-0.025
exchange rate_[YES, HIGHER]	-0.006	-0.013	-0.014	-0.020*	-0.018
exchange rate_[YES, LOWER]	0.009	0.010	0.004	-0.002	-0.002
financial crisis_[YES, DECREASED]	0.007	0.012	0.015	0.009	0.009
financial crisis_[YES, INCREASED]	-0.001	-0.005	-0.005	-0.006	-0.011
finely-balanced_[YES]	-0.002	-0.001	-0.002	-0.001	0.001
geopolitical_[YES, DECREASED]	-0.019	-0.022	-0.042	-0.058*	-0.061*
geopolitical_[YES, INCREASED]	-0.014*	-0.013	-0.010	-0.011	-0.007
house price_[YES, HIGHER]	-0.003	-0.001	0.004	-0.000	0.002
house price_[YES, LOWER]	0.009	0.007	-0.006	-0.002	-0.004
inflation rate_[YES, HIGHER]	-0.001	0.000	-0.001	-0.006	-0.004
inflation rate_[YES, LOWER]	-0.002	0.001	0.007	0.007	0.014
more closely watched than usual_[YES]	-0.003	-0.000	0.002	0.004	0.004
more highly-anticipated than ususal_[YES]	-0.003	-0.003	-0.005	-0.008	-0.005
mortgage approvals_[YES, HIGHER]	0.002	-0.000	0.004	0.010	0.006
mortgage approvals_[YES, LOWER]	-0.009	-0.019	-0.038	-0.043	-0.063
public sector borrowing_[YES, HIGHER]	0.061***	0.057**	0.061**	0.045*	0.040
purchasing manager index_[YES, HIGHER]	-0.001	-0.005	0.000	0.000	-0.005
purchasing manager index_[YES, LOWER]	0.011	0.026	0.026	0.014	0.022
recession_[YES, DECREASED]	0.001	0.003	-0.003	-0.000	-0.004
recession_[YES, INCREASED]	-0.001	0.003	0.003	0.002	0.002
retail sales_[YES, HIGHER]	-0.018	-0.014	-0.024	-0.024*	-0.019
retail sales_[YES, LOWER]	-0.015	-0.011	-0.011	0.002	0.005
sovereign default_[YES, INCREASED]	-0.011	-0.012	-0.011	-0.009	-0.014
supply chain_[YES, DECREASED]	-0.028*	-0.034**	-0.036**	-0.028	-0.047**
supply chain_[YES, INCREASED]	-0.016	-0.016	-0.014	-0.003	-0.007
taking place under crisis-like circumstances_[YES]	0.001	0.005	-0.005	0.002	0.024
taking place under extraordinary circumstances_[YES]	0.003	0.000	0.001	0.003	0.002
trade balance_[YES, HIGHER]	-0.004	0.003	0.004	0.019	0.019
trade balance_[YES, LOWER]	0.016	0.036*	0.044*	0.004	0.029
unemployment rate_[YES, HIGHER]	-0.050**	-0.045**	-0.055***	-0.053***	-0.037**
unemployment rate_[YES, LOWER]	-0.003	-0.002	0.010	0.010	0.011
volatility index_[YES, HIGHER]	-0.014	-0.011	-0.016	-0.010	-0.019
volatility index_[YES, LOWER]	-0.010	-0.013	-0.006	0.018	0.023
wage growth_[YES, HIGHER]	-0.010	-0.005	-0.008	-0.005	-0.008
wage growth_[YES, LOWER]	0.014	0.022	0.018	0.014	0.021
wage-price spiral_[YES, DECREASED]	0.015	0.012	0.017	0.001	-0.007
wage-price spiral_[YES, INCREASED]	0.002	0.005	0.005	0.007	0.009
1st lags	No	Yes	Yes	Yes	Yes
2nd lags	No	No	Yes	Yes	Yes
3rd lags	No	No	No	Yes	Yes
4th lags	No	No	No	No	Yes
Observations	398	398	398	398	398
R^2	0.142	0.254	0.367	0.509	0.602
F Statistic	1.789*** (df=43; 354)	1.924*** (df=86; 311)	1.930*** (df=129; 268)	2.269*** (df=172; 225)	26.909*** (df=215; 182)

*p<0.1; **p<0.05; ***p<0.01

Notes: Results of ordinary least squares regression of 2-year gilt reaction to UK monetary events within Braun et al. (2025) database on text-derived dummy variables capturing pre-event perceived UK economic conditions. Coefficients shown are for the level of the extracted measures. Coefficients for lags (if applicable) are omitted for brevity. Significance levels are based on heteroscedasticity-consistent standard errors.

Table 7: Estimated $\hat{\beta}_{OLS}$ for 5-year yield reactions

	Dependent variable: 5-year gilt yield reactions compiled by Braun et al. (2025) [mps5$_t$]				
	(1)	(2)	(3)	(4)	(5)
GDP growth_[YES, HIGHER]	-0.006	-0.004	-0.005	-0.013	-0.022**
GDP growth_[YES, LOWER]	-0.001	-0.007	-0.005	-0.004	-0.014
business confidence_[YES, HIGHER]	0.020*	0.018	0.022*	0.027**	0.032*
business confidence_[YES, LOWER]	-0.004	-0.001	0.011	0.019	0.018
consumer confidence_[YES, HIGHER]	-0.001	0.001	-0.012	-0.024	-0.017
consumer confidence_[YES, LOWER]	-0.023*	-0.026*	-0.030**	-0.027*	-0.029*
exchange rate_[YES, HIGHER]	-0.003	-0.010	-0.010	-0.018*	-0.017
exchange rate_[YES, LOWER]	0.013*	0.015*	0.008	0.005	0.005
financial crisis_[YES, DECREASED]	0.009	0.016	0.017	0.014	0.016
financial crisis_[YES, INCREASED]	-0.001	-0.003	-0.003	-0.005	-0.008
finely-balanced_[YES]	-0.002	-0.001	-0.002	0.002	0.004
geopolitical_[YES, DECREASED]	-0.040	-0.042	-0.057	-0.075*	-0.080*
geopolitical_[YES, INCREASED]	-0.011*	-0.010	-0.008	-0.006	-0.001
house price_[YES, HIGHER]	-0.003	-0.002	0.001	0.000	-0.001
house price_[YES, LOWER]	0.010	0.009	0.001	0.005	0.003
inflation rate_[YES, HIGHER]	-0.002	-0.001	-0.001	-0.005	-0.004
inflation rate_[YES, LOWER]	0.000	0.003	0.009	0.011	0.017*
more closely watched than usual_[YES]	-0.002	0.001	0.003	0.004	0.004
more highly-anticipated than ususal_[YES]	-0.004	-0.004	-0.007	-0.009	-0.008
mortgage approvals_[YES, HIGHER]	0.004	0.003	0.006	0.009	0.009
mortgage approvals_[YES, LOWER]	-0.005	-0.015	-0.009	0.000	-0.007
public sector borrowing_[YES, HIGHER]	0.048**	0.043**	0.046**	0.025	0.014
purchasing manager index_[YES, HIGHER]	-0.008	-0.010	-0.003	-0.004	-0.009
purchasing manager index_[YES, LOWER]	0.003	0.015	0.015	-0.002	-0.001
recession_[YES, DECREASED]	-0.001	0.000	-0.007	-0.002	-0.008
recession_[YES, INCREASED]	-0.002	0.001	0.001	0.000	-0.001
retail sales_[YES, HIGHER]	-0.009	-0.005	-0.012	-0.015	-0.007
retail sales_[YES, LOWER]	-0.014	-0.012	-0.012	-0.003	-0.003
sovereign default_[YES, INCREASED]	-0.009	-0.013	-0.013	-0.013	-0.020
supply chain_[YES, DECREASED]	-0.013	-0.020	-0.022*	-0.007	-0.017
supply chain_[YES, INCREASED]	-0.008	-0.010	-0.009	0.000	-0.005
taking place under crisis-like circumstances_[YES]	-0.002	0.003	-0.007	0.001	0.023
taking place under extraordinary circumstances_[YES]	0.005	0.002	0.002	0.003	0.001
trade balance_[YES, HIGHER]	-0.009	0.001	-0.000	0.009	0.012
trade balance_[YES, LOWER]	0.005	0.020	0.020	-0.024	0.001
unemployment rate_[YES, HIGHER]	-0.037**	-0.035**	-0.045***	-0.044***	-0.030**
unemployment rate_[YES, LOWER]	0.015	0.008	0.018	0.019	0.022
volatility index_[YES, HIGHER]	-0.014	-0.011	-0.012	-0.006	-0.015
volatility index_[YES, LOWER]	-0.008	-0.009	-0.011	0.013	0.013
wage growth_[YES, HIGHER]	-0.010*	-0.004	-0.008	-0.007	-0.009
wage growth_[YES, LOWER]	0.018	0.022	0.019	0.012	0.011
wage-price spiral_[YES, DECREASED]	0.006	0.003	0.011	-0.007	-0.011
wage-price spiral_[YES, INCREASED]	0.002	0.003	0.002	0.005	0.005
1st lags	No	Yes	Yes	Yes	Yes
2nd lags	No	No	Yes	Yes	Yes
3rd lags	No	No	No	Yes	Yes
4th lags	No	No	No	No	Yes
Observations	398	398	398	398	398
R^2	0.129	0.237	0.351	0.511	0.618
F Statistic	1.529** (df=43; 354)	1.716*** (df=86; 311)	2.075*** (df=129; 268)	3.400*** (df=172; 225)	14.894*** (df=215; 182)

*p<0.1; **p<0.05; ***p<0.01

Notes: Results of ordinary least squares regression of 5-year gilt reaction to monetary events within Braun et al. (2025) database on text-derived dummy variables capturing pre-event perceived UK economic conditions. Coefficients shown are for the level of the extracted measures. Coefficients for lags (if applicable) are omitted for brevity. Significance levels are based on heteroscedasticity-consistent standard errors.

frequency responses of 5-year gilts. The coefficent on the dummy variable indicating higher-than-expected unemployment rate readings is significantly negative (around four basis points) across all five specifications, including the 4-lag specification with 215 variables. Higher than expected public sector borrowing is significantly positive, as in Tables 5 and 6, at around 4.5 basis points. The dummy variable indicating lower-than-expected consumer confidence readings is also (marginally) significant across specifications, with a negative coefficient of around three basis points. Analogously to the unemployment coefficients, this suggests that the Bank was more dovish than had been expected in circumstances when the economy was cooling. Conversely, higher than expected business confidence is associated with significantly positive 5-year gilt yield reaction in four out of five specifications (around 2-3 basis points).

Finally, Table 8 shows the coefficients of regressions with 10-year gilt reactions to monetary events as dependent variable. R-squared ranges from 12% to 56% across the specifications, naturally increasing with the number of lags included in the explanatory variable set. Four lags are required to explain more than 50% of variation in the high-frequency responses of 10-year gilts. As in Table 7, the coefficent on the dummy variable indicating higher-than-expected unemployment rate readings is significantly negative (around 3-4 basis points) across all five specifications, including the 4-lag specification with 215 variables. As was the case for the 5-year yield, significance of the positive coefficent relating to the higher-than-expected public sector borrowing dummy variable diminishes in the higher-lag specifications. While not consistently significant, the dummy variable indicating lower-than-expected consumer confidence readings has a consistently negative estimated coefficient around 2 basis points. As was the case in Table 7, Table 8 shows that higher-than-expected business confidence is associated with a significantly positive 10-year gilt yield reaction in four out of five specifications (around two basis points).

Table 8: Estimated $\hat{\beta}_{OLS}$ for 10-year yield reactions

	Dependent variable: 10-year gilt yield reactions compiled by Braun et al. (2025) [$mps10_t$]				
	(1)	(2)	(3)	(4)	(5)
GDP growth_[YES, HIGHER]	-0.004	-0.003	-0.002	-0.010	-0.016*
GDP growth_[YES, LOWER]	-0.005	-0.010	-0.008	-0.009	-0.017
business confidence_[YES, HIGHER]	0.017**	0.015*	0.016*	0.017*	0.024
business confidence_[YES, LOWER]	-0.003	-0.001	0.005	0.013	0.017
consumer confidence_[YES, HIGHER]	-0.001	0.001	-0.008	-0.011	-0.006
consumer confidence_[YES, LOWER]	-0.018	-0.021*	-0.023*	-0.023*	-0.023
exchange rate_[YES, HIGHER]	-0.004	-0.009	-0.009	-0.016*	-0.016
exchange rate_[YES, LOWER]	0.009	0.011	0.007	0.006	0.006
financial crisis_[YES, DECREASED]	0.009	0.015	0.013	0.014	0.016
financial crisis_[YES, INCREASED]	-0.001	-0.002	-0.002	-0.004	-0.006
finely-balanced_[YES]	0.000	0.002	0.002	0.006	0.009
geopolitical_[YES, DECREASED]	-0.039	-0.040	-0.049	-0.067*	-0.071*
geopolitical_[YES, INCREASED]	-0.010*	-0.009	-0.009	-0.005	-0.001
house price_[YES, HIGHER]	-0.006	-0.005	-0.004	-0.005	-0.008
house price_[YES, LOWER]	0.002	0.002	-0.001	0.003	-0.003
inflation rate_[YES, HIGHER]	-0.001	-0.002	-0.001	-0.004	-0.002
inflation rate_[YES, LOWER]	0.001	0.004	0.009	0.012*	0.018**
more closely watched than usual_[YES]	-0.001	0.002	0.004	0.005	0.004
more highly-anticipated than ususal_[YES]	-0.002	-0.002	-0.003	-0.004	-0.005
mortgage approvals_[YES, HIGHER]	0.006	0.005	0.007	0.008	0.011
mortgage approvals_[YES, LOWER]	0.004	-0.002	0.007	0.023	0.045
public sector borrowing_[YES, HIGHER]	0.028*	0.029*	0.030	0.021	0.016
purchasing manager index_[YES, HIGHER]	-0.009	-0.010	-0.005	-0.005	-0.009
purchasing manager index_[YES, LOWER]	-0.000	0.011	0.010	-0.005	-0.009
recession_[YES, DECREASED]	-0.002	0.000	-0.007	-0.005	-0.010
recession_[YES, INCREASED]	-0.003	-0.001	0.000	-0.002	-0.003
retail sales_[YES, HIGHER]	-0.002	0.001	-0.003	-0.006	-0.001
retail sales_[YES, LOWER]	-0.010	-0.009	-0.010	-0.002	-0.002
sovereign default_[YES, INCREASED]	-0.009	-0.013	-0.012	-0.014	-0.019
supply chain_[YES, DECREASED]	-0.001	-0.009	-0.008	0.006	-0.001
supply chain_[YES, INCREASED]	-0.004	-0.005	-0.005	0.001	-0.004
taking place under crisis-like circumstances_[YES]	0.001	0.007	-0.002	0.003	0.018
taking place under extraordinary circumstances_[YES]	0.005	0.001	0.002	0.003	0.002
trade balance_[YES, HIGHER]	-0.010**	-0.002	-0.002	0.000	0.003
trade balance_[YES, LOWER]	0.002	0.013	0.013	-0.026	-0.012
unemployment rate_[YES, HIGHER]	-0.028**	-0.031**	-0.040***	-0.040***	-0.030**
unemployment rate_[YES, LOWER]	0.013	0.004	0.011	0.010	0.013
volatility index_[YES, HIGHER]	-0.012	-0.012	-0.010	-0.009	-0.013
volatility index_[YES, LOWER]	-0.013	-0.013	-0.018	-0.000	-0.010
wage growth_[YES, HIGHER]	-0.005	-0.001	-0.004	-0.002	-0.002
wage growth_[YES, LOWER]	0.023*	0.024	0.021	0.019	0.014
wage-price spiral_[YES, DECREASED]	-0.001	-0.004	0.001	-0.012	-0.010
wage-price spiral_[YES, INCREASED]	-0.001	0.001	-0.000	0.002	0.003
1st lags	No	Yes	Yes	Yes	Yes
2nd lags	No	No	Yes	Yes	Yes
3rd lags	No	No	No	Yes	Yes
4th lags	No	No	No	No	Yes
Observations	398	398	398	398	398
R^2	0.116	0.209	0.305	0.450	0.561
F Statistic	1.726*** (df=43; 354)	1.552*** (df=86; 311)	1.737*** (df=129; 268)	13.163*** (df=172; 225)	2.347*** (df=215; 182)

*p<0.1; **p<0.05; ***p<0.01

Notes: Results of ordinary least squares regression of 10-year gilt reaction to UK monetary events within Braun et al. (2025) database on text-derived dummy variables capturing pre-event perceived UK economic conditions. Coefficients shown are for the level of the extracted measures. Coefficients for lags (if applicable) are omitted for brevity. Significance levels are based on heteroscedasticity-consistent standard errors.

A.5.2 Comparison of orthogonalisation methods

Unlike ordinary least squares, where inference about coefficients is standard, inference about ridge and lasso coefficients is not entirely straightforward. To facilitate comparison of the different estimation methods, the Shapley values of each model[16] are computed for each gilt yield maturity. Shapley values for all three regression approaches are shown in Figure 29.

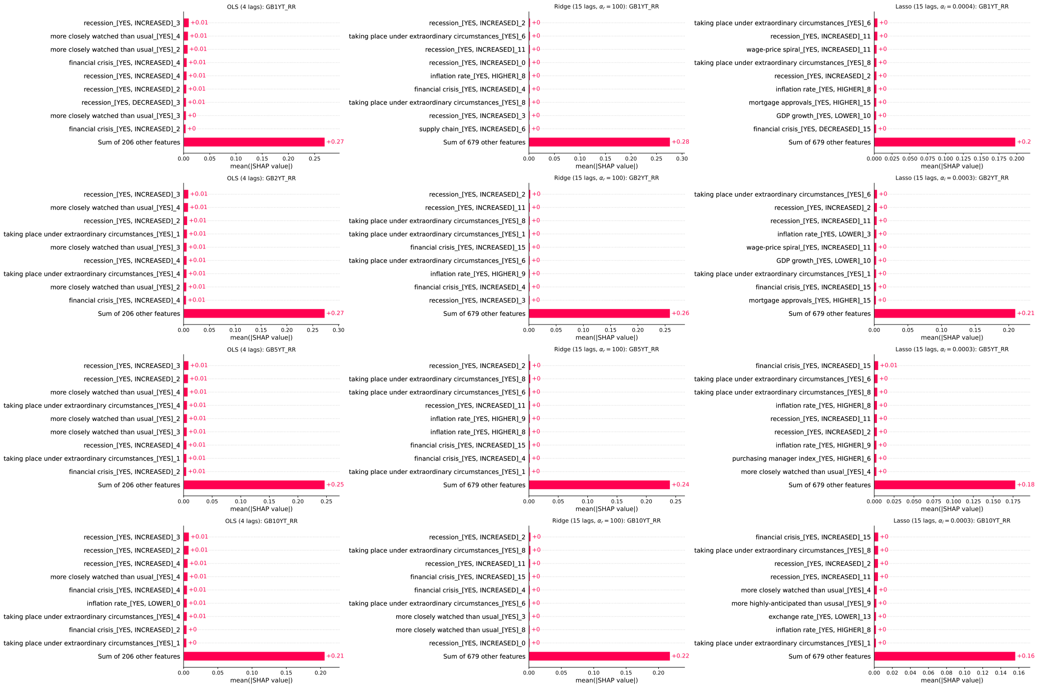
The left column shows the results for linear models with four lags, estimated with ordinary least squares. The middle column shows results for ridge regressions with 15 lags. The right column shows Shapley values for the lasso regression, also with 15 lags. There is one row for each dependent variable, i.e. the 1-year, 2-year, 5-year, and 10-year gilt reaction.

Across all specifications and gilt maturities, the top nine predictors account for a relatively small share of the sum of absolute Shapley values. That is, many of the variables appear to have a role in predicting gilt yield reactions[17] For the 4-lag OLS prediction of 1-year yield reactions, the predictor with the highest Shapley value is the third lag of the dummy variable indicating that recession risk was perceived to have increased prior to the monetary event. This is followed by further lags of the increased-recession-risk indicator, of the more-closely-watched-than-usual dummy, and of the increased-financial-crisis-risk indicator. Interestingly, it appears that the lags of text-derived variables are more relevant than the latest information prior to the monetary event. For the longer-dated yield reactions, the Shapley values of the OLS regression are broadly similar, with the third lag of the higher-than-expected-recession-risk indicator having the largest absolute value. The middle column shows the rankings of the Shapley values of the independent variables in the ridge specifications with 15 lags.

[16] These are estimated using the *shap* Python package. See Buckmann et al. (2022) for an introduction to Shapley values as a method for generic feature importance.

[17] This finding is partly driven by the disaggregated predictor set consisting of many dummy variables and its lags. In Figure 30 the Shapley values of individual dummy variable are aggregated into groups to facilitate interpretation of relative importance.

Figure 29: Shapley values for individual dummy variables



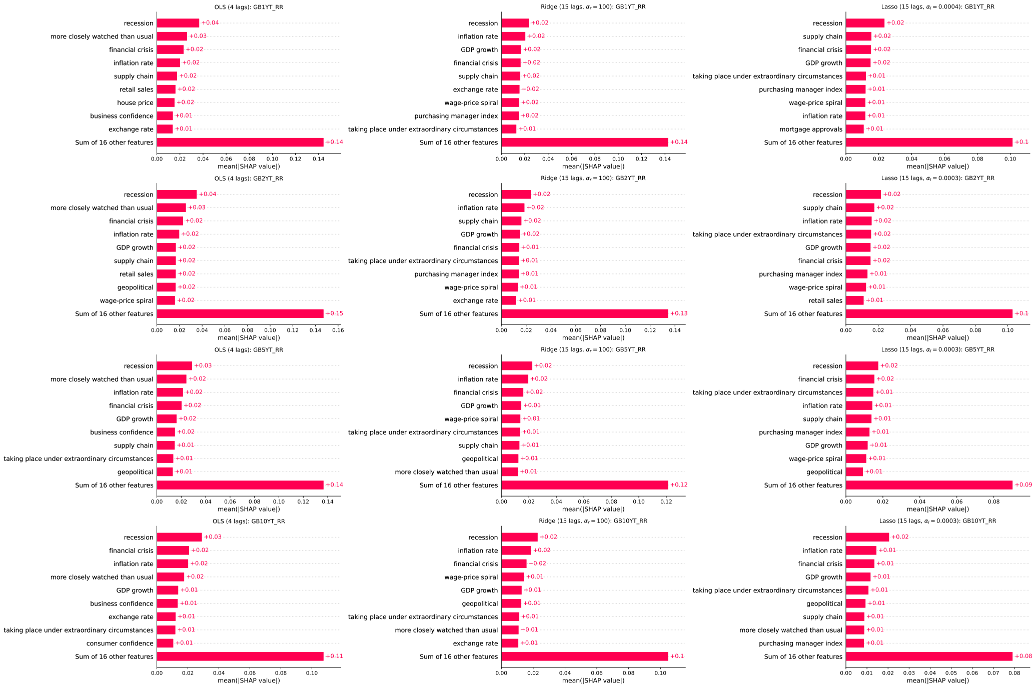
Notes: Shapley values shown were estimated using the shap *Python package. The left column shows the results for the OLS specification with four lags. The middle column shows results for ridge regressions with 15 lags. The right column shows Shapley values for the lasso regression with 15 lags. Rows correspond to different dependent variables, i.e. 1-year, 2-year, 5-year, and 10-year gilt reactions measured by Braun et al. (2025).*

These rankings are similar to the 4-lag OLS specification. The higher-than-expected-recession-risk dummy, for instance, is a highly-ranked predictor in both specifications that is within the top nine predictors with different lags. One difference is that the 'taking place under extraordinary circumstances' indicator appears within the top nine predictors across different gilt yield specifications. Interestingly, the highest-ranked predictors in the ridge specification include several lags greater than four. The ranking of predictors in the lasso specification has some overlap with both the OLS and ridge specifications. However, the top nine rankings for the different gilt maturities include variables that do not appear in the other two specifications, such as the indicator for higher-than-expected purchasing manager index readings, lower-than-expected GDP growth readings, and lower-than-expected exchange rate readings. As in the ridge specfication, the highest ranked variables in the lasso specification include several lags greater than four. One shortcoming of Table 29 is that different lags appear as distinct variables, making interpretation about the ranking of economic concepts that have the most ex post predictive power challenging.

To mitigate this limitation, Figure 30 aggregates the Shapley values by concept, across dummy variables for different answers to the same prompt and across different lags for the same dummy variable. This grouping aims facilitate interpretation of relative importance. A key result is that measures regarding recession risk are the highest-ranked predictor group in every specification and for every gilt yield maturity. The summed absolute Shapley values of the recession group of predictor variables accounts for more than 10% of total sum of Shapley values across specifications. Financial crisis risk is consistently ranked highly as well. Another key group of variables relates to the measures of perceived surprises in inflation and GDP readings. Apart from these similarities there is some variation, both across specifications and gilt maturities.

Another lens on the ex post relationships captured by the different re-

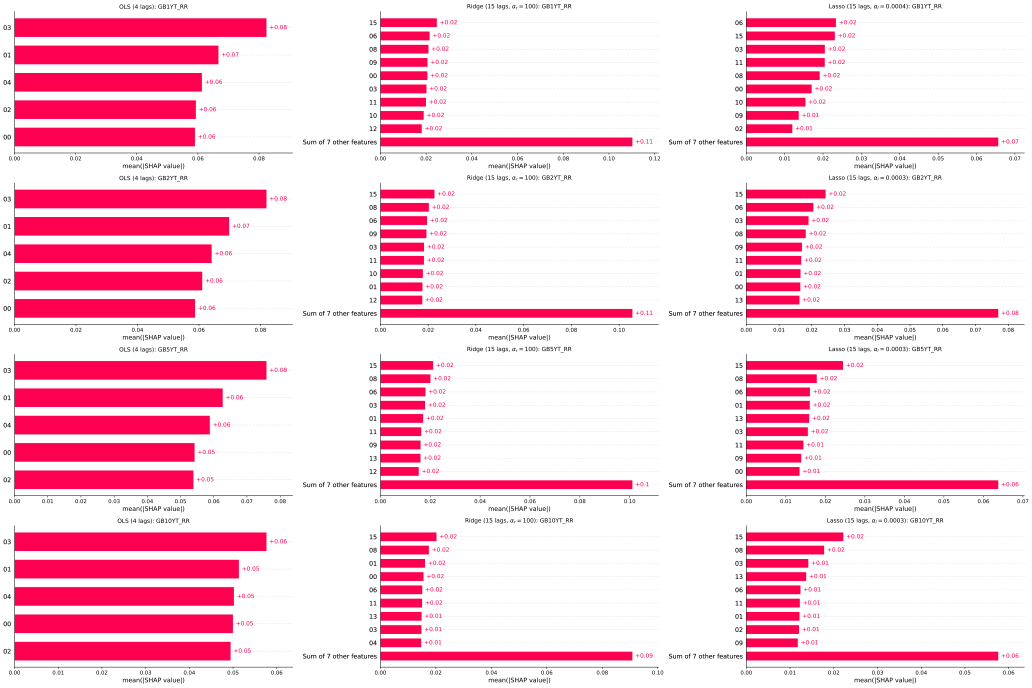
Figure 30: Shapley values (aggregated to prompt-level)



Notes: Values shown are the Shapley values in Figure 29, summed to prompt-level concepts over different lags and different response dummys for the same prompt. As in Figure 29, columns correspond to different specifications while rows correspond to different dependent variables.

gression specifications is to aggregate Shapley values across prompts to the lag length, in order to provide an indication of which lags are most predictive (ex post) of market reactions to monetary events. Figure 31 presents the results of this exercise. For the OLS specification in the left column there are only five Shapley values per regression, due to this specification having only four lags. One notable finding is that across specifications it is not the contemporaneous ('00') features that have the highest aggregate Shapley value. Instead it is the third lag in the OLS specifications and the 15th lag for all of the regularised least squares specifications.

Figure 31: Shapley values (aggregated across variable types to lags)



Notes: Values shown are the Shapley values in Figure 29, summed accross concepts to lag-length. As in Figures 29 and 30, columns correspond to different specifications while rows correspond to different dependent variables.

A.6 Sensitivity analyses

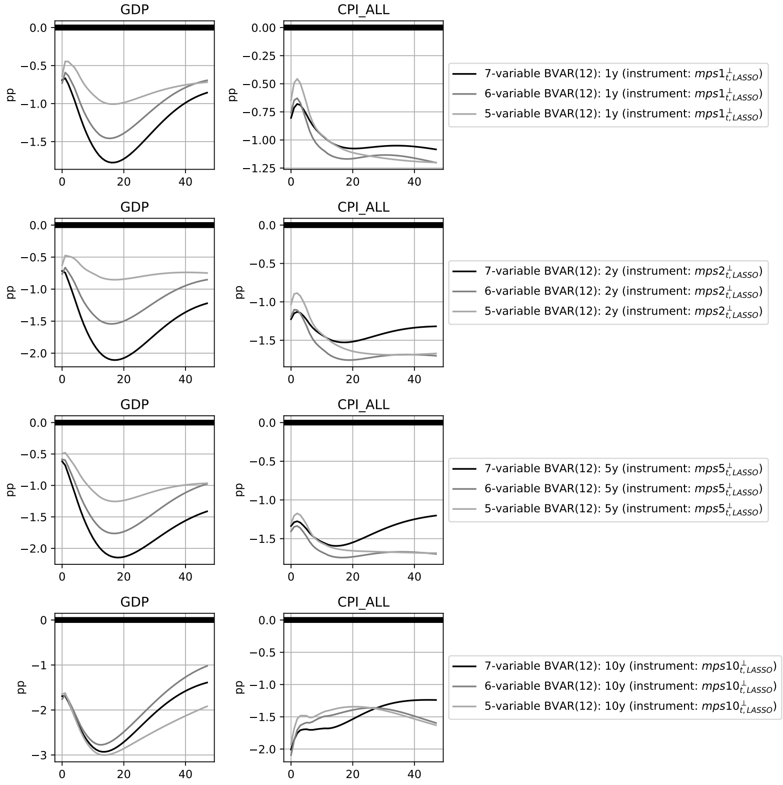
In this subsection I explore how sensitive the results in Figures 14 and 15 are to specification choices.

A.6.1 Variables included in VAR

I first examine the sensitvity of the results in Figure 15 to changes in the variables included in the BVARs. In particular, Figure 32 shows the impulse responses for three BVARs with different sets of endogenous variables included.

The first BVAR includes the same seven variables as in Figures 14 and 15. The second specfication removes the unemployment rate, resulting in a six-variable BVAR. Finally, the third BVAR additionally excludes the Favara et al. (2016) excess bond premium. For both GDP and consumer price index, the impact effects are virtually identical across the different BVAR specifications. The impact on price level is also robust to dropping both variables across gilt yield maturities. The impact on real activity (i.e. GDP) is qualitatively unchanged as well, although removal of both the unemployment rate and the excess bond premium reduce the peak magnitude of the effect. In particular, in the VAR specifications that include the 1-year, 2-year, and 5-year gilt yields as endogenous variable used for normalisation, the peak effect is reduced by removal of the unemployment rate and further dampened by dropping the excess bond premium—with the latter making the bigger difference. As a result, the magnitude of the peak effect on real activity is reduced by around 50% for the specifications involving the three shortest gilt maturities. In the VAR specification for the 10-year gilt, the peak GDP response is robust irrespective of whether five, six, or seven variables are included in the BVARs.

Figure 32: Sensitivity to different BVAR specifications



Notes: Monetary policy shocks are identified using external instrumental variables $mps1^{\perp}_{t,lasso}$, $mps2^{\perp}_{t,lasso}$, $mps5^{\perp}_{t,lasso}$, and $mps10^{\perp}_{t,lasso}$ (i.e. the residuals from lasso regressions of original high-frequency reactions on text-derived variables). Estimated using a monthly 12-lag Bayesian structural vector autoregression (BSVAR-IV) with normal-inverse-Wishart priors as in Miranda-Agrippino and Ricco (2021). There is one row per VAR, with different rows showing results corresponding to different gilt yields used as external instrument and normalisation variable included in the VAR. Monetary policy shocks are normalised such that the corresponding gilt yield increases by 100 basis points. The 7-variable specification is as in Figure 15. In the 6-variable specification the unemployment rate is excluded. In the 5-variable specification the Favara et al. (2016) excess bond premium is also excluded. Estimation sample: 1997M1-2019M12.

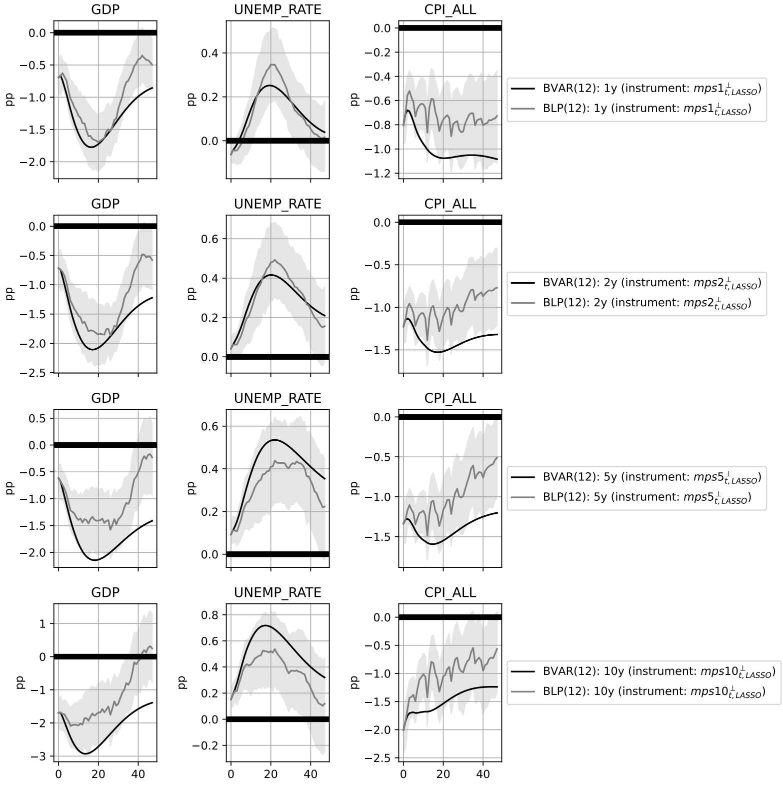
A.6.2 Econometric model specifications

A second sensitvity check is to examine dependence on the econometric model specfication. In theory, Plagborg-Møller and Wolf (2021) show that both VARs and linear projections estimate the same impulse responses. Based on monte carlo evidence, Li et al. (2024) find that the choice between VAR and LP inference involves choosing a point along a bias-variance trade-off, with linear projections typically having lower bias but higher variance than VAR estimators. As a result they advocate the use of shrinkage using either Bayesian VARs or penalised linear projections.

I explore the extent to which changing from the default 12-lag BVAR to a 12-lag Bayesian Linear Projection (BLP) specification results in different impulse response estimates. Both are implemented using the same estimation procedures as in Miranda-Agrippino and Ricco (2021). The resulting impulse responses are shown in Figure 33.

Impact effects appear identical across the two specifications. The shape of impulse responses is unaffected by the modelling framework, with results being qualitatively identical regardless of whether BVAR or BLP is used. At longer horizons, there are some quantitative differences in the estimated impulse responses. For example, the peak unemployment rate response in the bottom row is around +0.7 percentage points if estimated using the BVAR approach but closer to +0.5 percentage points when using the BLP procedure. This is consistent with the second 'implication for empirical practice' in Plagborg-Møller and Wolf (2021), which states that the impulse response estimates obtained using VAR and LP estimators should be approximately the same for the first p horizons of the impulse response function, where p is the number of lags included in the VAR and LP. Indeed, where estimated responses diverge between BVAR and BLP estimates in Figure 33, this tends to happen around response horizon 12—the lag length used to generate the figure. Another observation is that BLP tends to estimate impulse responses that have somewhat smaller effect magnitudes, although this is not a univer-

Figure 33: Sensitivity to different model specifications



Notes: Monetary policy shocks are identified using external instrumental variables $mps1^{\perp}_{t,lasso}$, $mps2^{\perp}_{t,lasso}$, $mps5^{\perp}_{t,lasso}$, *and* $mps10^{\perp}_{t,lasso}$ *(i.e. the residuals from lasso regressions of original high-frequency reactions on text-derived variables). Estimated using a monthly 12-lag Bayesian structural vector autoregression (BSVAR-IV) with normal-inverse-Wishart priors and Bayesian linear projections (BLP-IV) as in Miranda-Agrippino and Ricco (2021). There is one row per VAR, with different rows showing results corresponding to different gilt yields used as external instrument and normalisation variable included in the VAR. Monetary policy shocks are normalised such that the corresponding gilt yield increases by 100 basis points. The shaded areas indicate the 90% posterior coverage bands. Variables included in VAR/LP (other those shown and the relevant gilt yield): Favara et al. (2016) excess bond premium, FTSE250 index, and BIS Pound Sterling broad effective exchange rate index. Estimation sample: 1997M1-2019M12.*

sal pattern. Universally, BVAR estimates are smoother, with BLP estimates having a more ragged shape. This is consistent with Li et al. (2024), who find LP estimators to have higher variance than VAR estimators.

A.6.3 Lag length

A third sensitvity check I perform is to vary the lag length for the BVAR specfication. The results of this exercise are shown in Figure 34.

As can be seen, reducing the lag length from 12 to nine makes little difference to the estimated impulse responses. I leverage this finding in Section 2.6, where I estimate a large BVAR with nine rather than 12 lags to reduce the number of parameters that have to be estimated. The finding in Figure 34 suggests that doing so should not lead to significantly different effect estimates.

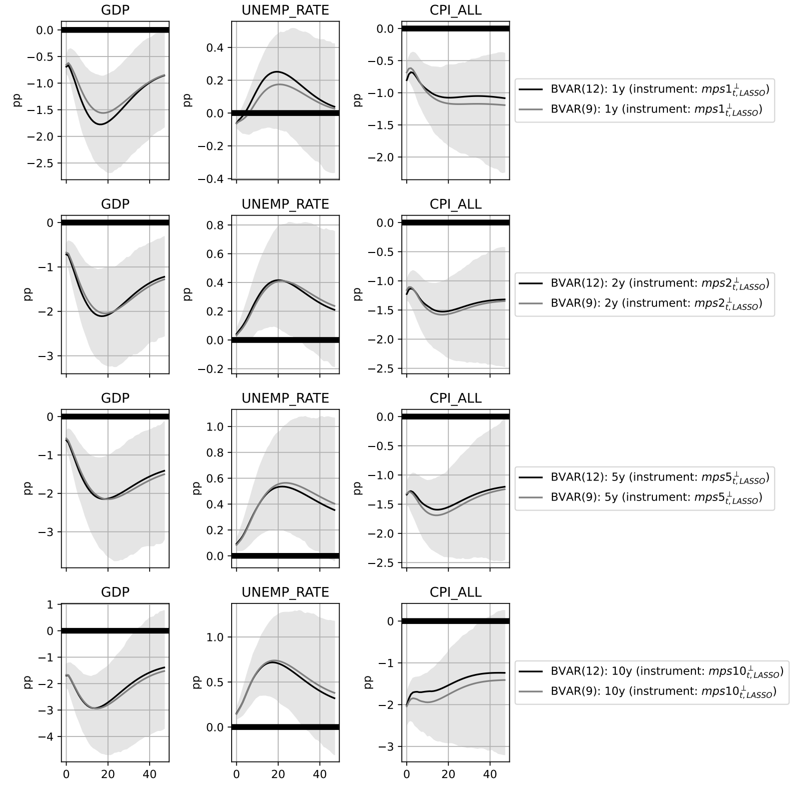
A.7 Comparison to other studies

I now compare the findings based on the lasso identification with results obtained by other empirical studies of the macroeconomic effects of monetary policy. Results covering both UK and US monetary policy are considered in what follows. Throughout, the maturity of the interest rate used for normalising the monetary shock size is matched to enable an appropriate comparison.

A.7.1 Braun et al. (2025)

In addition to curating a database of UK monetary events and conventional monetary policy surprises, Braun et al. (2025) investigate the dynamic effects of Bank of England policies. To do so, they transform conventional high-frequency monetary policy surprises into distinct 'Target', 'Path', and 'QE' factors that are constructed with the aim of distinguishing the effects of the

Figure 34: Sensitivity to different lag specifications



Notes: Monetary policy shocks are identified using external instrumental variables $mps1^{\perp}_{t,lasso}$, $mps2^{\perp}_{t,lasso}$, $mps5^{\perp}_{t,lasso}$, and $mps10^{\perp}_{t,lasso}$ (i.e. the residuals from lasso regressions of original high-frequency reactions on text-derived variables). Estimated using a monthly 9-lag and 12-lag Bayesian structural vector autoregressions (BSVAR-IV) with normal-inverse-Wishart priors as in Miranda-Agrippino and Ricco (2021). There is one row per VAR, with different rows showing results corresponding to different gilt yields used as external instrument and normalisation variable included in the VAR. Monetary policy shocks are normalised such that the corresponding gilt yield increases by 100 basis points. The shaded areas indicate the 90% posterior coverage bands. Variables included in VAR (other those shown and the relevant gilt yield): Favara et al. (2016) excess bond premium, FTSE250 index, and BIS Pound Sterling broad effective exchange rate index. Estimation sample: 1997M1-2019M12.

Bank of England's different monetary policy instruments. This study instead orthogonalises the same conventional monetary policy surprises collected by Braun et al. without the factor transformations with respect to the information extracted from newswires. The sample period (1997M1-2019M12) is the same across both studies. As a benchmark, Figure 10 shows significant real activity and employment puzzles that obtain when the conventional 1-year gilt monetary policy surprises by Braun et al. (2025) are used. Figure 11 shows that implementation of the sign restrictions of Jarociński and Karadi (2020) does not resolve these puzzles. Considering the GDP response to a 'Target' shock scaled to induce a 100 basis point increase in the 1-year rate in Figure 5 of Braun et al. (2025), there is a mild immediate impact. The estimated impulse response peaks around 10 months after the shock at around -2 percentage points—similar to the peak effect of -1.8 percentage points arising from text-orthogonalisation in Figure 15. The GDP reponse to a 'Path' shock estimated by Braun et al. (2025) is short-lived and only marginally significant. This is in contrast to the more long-lived effect on real activity shown in Figure 15. With regards to the price level response to a 'Target' shock scaled to induce a 100 basis point increase in the 1-year rate, Figure 5 of Braun et al. (2025) shows a signficiant price puzzle. In particular, the peak effect on consumer prices is estimated to peak at +1 percentage point six months after the shock before reverting to the pre-shock level. This puzzle remains under alternative factor specifications, as shown in Figure D4 in the online appendix of Braun et al. (2025). This is in contrast to the strong negative response to a contractionary shock shown in Figure 15. Interestingly, it appears that this puzzle could be a result of the factor approach taken by Braun et al. (2025). In particular, Figure 10 shows that in my VAR specification no such price puzzle obtains when using the raw (i.e. non-factor) monetary policy surprises as external instruments. In response to a 'Path' shock scaled to induce a 100 basis point increase in the 1-year rate, Braun et al. (2025) find a significant negative effect on the consumer price index, with an impact effect of around -0.7 percentage points to a 100 basis point shock and a large effect of -2.8 percentage points after three years. That is, the impact effect is similar to that in Figure 15, but the peak impact of a 'Path' shock on the price level is larger than the -1 percentage point after 20 months estimated by this study. Figure D5 of the online appendix of Braun et al. (2025) reports the responses to a contractionary QE shock. There is a signficiant real activity puzzle, with an impact effect of around +2.8 percentage points on GDP for every 100 basis points movement in the 10-year rate. The puzzle remains significant up to 10 months after the shock. There is also a price puzzle, although this is not significant. Both puzzles remain in an alternative factor specfication. In contrast, Figure 15 in this study shows a significantly negative GDP and price response to a contractionary monetary policy shock identfied using text-orthogonalised 10-year gilt yield monetary policy surprises. In summary, some of the estimates of the effects of Bank of England policies on key macroeconomic aggregates in Braun et al. (2025) are materially different to those generated in this study, with the latter tending to be aligned with the theoretical consensus on monetary non-neutrality.

A.7.2 Kaminska and Mumtaz (2022)

Studying UK quantitative easing specifically, Kaminska and Mumtaz (2022) investigate the effect of monetary policy surprises relating to long-term interest rates. In particular, they create separate instruments to identify monetary policy shocks due to (i) 'signalling' and (ii) 'QE-specific term premia'. Figure 4 in Kaminska and Mumtaz (2022) shows the impulse responses to a monetary policy shock identified using the 'signalling'-channel instrument. The shock is normalised to induce a 100 basis point *decrease* in the 10-year gilt yield. That is, the figure displays the impulse responses following an *expansionary* monetary policy shock. There is no measure of real activity included in the VAR specification (e.g. neither GDP nor industrial production are included). Regarding effects on the unemployment rate, there is no significant unemployment reponse until 30 months after the signalling shock. The peak unemployment impact is a reduction in the unemployment rate of about -0.5 percentage points. These estimates are broadly similar to those in Figure 15, where there is little immediate impact followed by a somewhat earlier peak impact of 0.6 percentage points (-0.6 percentage points for an expansionary shock) after around 20 months. With regards to the price level response to a signalling shock, Kaminska and Mumtaz (2022) find only a mild immediate inflation impact, followed by a peak inflation impact of +2 percentage points around 15 months after the shock. While not straightforward to compare impulse responses of inflation with impulse responses of the price level, it appears that the price level impact in Figure 15 materialises more immediately. The impacts of a 'QE-specific term premia'-shock in Figure 6 of Kaminska and Mumtaz (2022) include an unemployment impact of -0.2 percentage points, materialising after around 20 months. This is a weaker peak impact compared to that in Figure 15. Regarding the impact of a 'QE-specific term premia' shock on the price level, there is a moderate price puzzle. In particular, there is a negative impact on inflation following an expansionary monetary policy shock—although the effect is not statistically significant.

A.7.3 Cesa-Bianchi et al. (2020)

Cesa-Bianchi et al. (2020) use a high-frequency identification approach with a 7-variable VAR to estimate impulse responses of a monetary policy shock scaled to induce a 25 basis point increase in the 1-year UK gilt rate. Both the sample period (1992M1 to 2015M1) and the set of variables included in their VAR (e.g. mortgage spread and two different corporate spreads) are different to the specfication used to estimate the impulse respones in Figure 15. Due to diffences in significance levels (68% vs 90%) a direct comparison of significance is not possible. Regarding the impact on employment, the immediate impact shown in Figure 2 of Cesa-Bianchi et al. (2020) is near-zero as in this study. The peak effect magnitude is equivalent to a +0.4 percentage points increase in unemployment in response to a 100 basis point shock, compared to the +0.25 percentage points increase in Figure 15. The impact effect on the price level in response to 25 basis point shock estimated by Cesa-Bianchi et al. (2020) is -0.07 percentage points (or a -0.28 percentage points impact in response to a 100 basis point shock), with a peak response 10 months after the shock of around -0.1 percentage points (-0.4 percentage points for a 100 basis point shock). This is somewhat smaller than the -0.7 percentage points impact effect and peak of -1 percentage point after around 20 months in Figure 15. Both estimated impulse responses are long-lived. Figure 3 in Cesa-Bianchi et al. (2020) reports the results for an extended specification, where GDP has a peak response after around 24 months of -1.4 percentage points to a monetary policy shock that induces a 100 basis point increase in the 1-year gilt yield. Both timing and magnitude of this impact are similar to the estimated responses in Figure 15.

A.7.4 Bauer and Swanson (2023b)

Watson (2023) refers to Figure 8 in Bauer and Swanson (2023b) as *'the new benchmark impulse response functions for the effect of Federal Reserve monetary policy shocks on the US macroeconomy'*. In particular, Bauer and Swanson (2023b) analyse the dynamic causal effects of orthogonalised monetary policy surprises in a six-variable structural vector autoregression with external instruments. There are some differences in specification. Firstly, the estimation window differs, ranging from 1973M1 to 2020M2 in Bauer and Swanson (2023b) compared to 1997M1-2019M12 in this study. Secondly, monthly real activity is measured using industrial production rather than GDP. Thirdly, my specification does not include commodity prices but does include the exchange rate and equity market index. Subject to these caveats, I compare Figure 8 in Bauer and Swanson (2023b) to Figure 15 by quadrupling their estimated effects. This is done to account for their normalisation of the monetary policy shock to be a 25 basis point increase in the 2-year bond rate compared to the 100 basis point normalisation in Figure 15. Regarding the real activity response, Figure 15 shows that GDP contracts just under -1 percentage point on impact of a 100 basis point 2-year rate shock. This is similar to Bauer and Swanson (2023b)'s estimate of the impact effect of a 25 basis point shock on industrial production of just under 20 basis points, corresponding to around -0.8 percentage points in response to a 100 basis point shock. The peak effect in Figure 15 reaches -2 percentage points within around 18 months of a 100 basis point shock. This is compared to a peak effect to US industrial production of -0.4 percentage points (-1.6 percentage points) in response to a 25 (100) basis point shock estimated by Bauer and Swanson (2023b). Whilst this is an imperfect comparison due to monthly real activity being measured using GDP for the UK and industrial production in the US, it illustrates that both impact and peak effects to real activity appear to be of similar magnitude. That said, the peak response of real activity is estimated by Bauer and Swanson (2023b) to materialise within 10 months in the US, compared to a peak effect of closer to 20 months in Figure 15. In both cases the effect is estimated to be long-lasting, remaining significantly negative until more than 40 months after the shock. Considering the effect on the unemployment rate, Figure 15 shows a near-zero impact effect in response to a monetary policy shock identified using the 2-year gilt yield surprise, which is similar to the result in Figure 8 of Bauer and Swanson (2023b). Regarding the peak effect, I find that a monetary policy shock normalised to a 100 basis point increase in the 2-year gilt yield peaks at +0.4 percentage points after around 20 months. In comparison, Bauer and Swanson (2023b) find a peak effect of +0.05 percentage points (+0.2 percentage points) in response to a monetary policy shock scaled to increase the 2-year government bond rate 25 (100) basis points. This study estimates that the effect on the unemployment rate peaks after 20 months, compared to after around 10 months in Bauer and Swanson (2023b). Both studies find the effect on unemployment to be significant only around the peak effect point. Turning to the effect on the price level, Figure 15 shows an impact effect of -1.3 percentage points in response to a 100 basis point shock. This is a larger impact estimate compared to the around -0.4 percentage points (after scaling to 100 basis point shock) in Bauer and Swanson (2023b). Within 40 months their effect increases to about -0.8 percentage points (after scaling), while the effect in Figure 15 is around the impact level of -1.3 percentage points 40 months after the shock. Both estimated impulse responses are significant throughout, beyond 40 months after the shock. In summary, while there are some differences in specification, I find that my estimated impact and peak effects of a monetary policy shock on real activity are similar in magnitude to Bauer and Swanson (2023b), although the peak effect is estimated to materialise nearly twice as fast in Bauer and Swanson (2023b) compared to this study's estimates. The employment response is similar on impact, but Bauer and Swanson (2023b) estimate a more immediate yet somewhat weaker peak effect. Similarly, the effect on the price level estimated by this study is somewhat stronger than that of Bauer and Swanson (2023b).

A.7.5 Aruoba and Drechsel (2024)

Leveraging natural language processing, Aruoba and Drechsel (2024) estimate the effects of monetary policy shocks on US macroeconomic aggregates using a six-variable Bayesian VAR. The set of variables included in their VAR is broadly equivalent to those used to estimate the impulse response functions in Figure 15, although I additionally include an exchange rate variable. Differences in specification include the sample period used to estimate the VAR (1984M2 to 2016M12), which differs from the period used for Figure 15. Key results of Aruoba and Drechsel (2024) are shown in the left panel of their Figure 6. Comparing the panel to the 1-year rate results in Figure 15 requires adjusting the magnitude of the shock, which is scaled to a 100 basis points increase in the 1-year rate in this study and around eight basis points in Aruoba and Drechsel (2024). Considering the effect of a monetary policy shock on real activity, the impact effect on real GDP in Aruoba and Drechsel (2024) is near-zero, compared to a significantly negative impact effect found in this study. The peak effect is reached around 25 months after the shock, which, after scaling the monetary policy shock to induce a 100 basis points increase in the 1-year rate, peaks at around -1.25 percentage points—similar to the peak effect in Bauer and Swanson (2023b) but somewhat below the 1-year estimate in this study of -1.8 percentage points. Compared to strong significance in this study, the effect on real activity in Aruoba and Drechsel (2024) is marginally significant throughout. Regarding the impact on unemployment, the initial impact effect is near-zero as in Figure 15. The peak effect is reached around 25 months after the shock, which, as with the effect on GDP, is around half a year later than in the impulse response shown in Figure 15. The magnitude of the peak effect is equivalent to a +0.75 percentage points unemployment response to a monetary policy shock scaled to induce a 100 basis point increase in the 1-year US bond yield—significantly stronger than the +0.25 percentage points estimated for the UK in this study and the +0.2 percentage points US in Bauer and Swanson (2023b). The effect estimated by Aruoba and Drechsel (2024) is also more long-lasting and significant than that in Bauer and Swanson (2023b) and this study. In Aruoba and Drechsel (2024), the effect on the price level is muted with an impact effect near zero and an insignificant peak effect of -0.25 percentage points in response to a 100 basis point shock reached after serveral years. This is in contrast to the -1 percentage point effect estimated in this study and -0.8 percentage points estimated in Bauer and Swanson (2023b). In summary, despite differences in specification, the effects of US monetary policy on real activity estimated by Aruoba and Drechsel (2024) is of similar magnitude as this study's estimates. The impact effect on employment is similar, but the peak employment response is stronger in Aruoba and Drechsel (2024) compared to Bauer and Swanson (2023b) and this study. In contrast, the impact on the price level is insignificant and weaker than in Bauer and Swanson (2023b) and this study.

A.7.6 Miranda-Agrippino and Ricco (2021)

Figure 3 in Miranda-Agrippino and Ricco (2021) displays the effect of a 100 basis point monetary policy shock on the 1-year US government bond rate in a six-variable VAR. The estimation sample used in their study ranges from 1979M1 to 2014M12. The variables included in their VAR are broadly equivalent to the ones used to estimate the impulse responses in Figure 15, although my specification does not include commodity prices but does include the exchange rate and equity market index. Moreover, rather than monthly GDP, industrial production is used as monthly measure of real activity as in Bauer and Swanson (2023b). The real activity impact effect estimated by Miranda-Agrippino and Ricco (2021) is around -1 percentage point, which is slightly stronger than the -0.8 percentage points in response to a 1-year shock in Figure 15. The peak effect materialised around 12 months after the shock, at around -1.6 percentage points. That is, the effect peak is slightly milder and sooner than the -1.8 percentage points estimated in Figure 15 reached after around 18 months. As in this study, the real activity effect is significant and long-lived. Miranda-Agrippino and Ricco (2021) estimate a near-zero impact effect on the unemployment rate, with their estimated impulse response peaking at around +0.3 percentage points around 1.5 years after the shock. This is very similar to the unemployment impulse response in Figure 15. That said, the Miranda-Agrippino and Ricco response is significant throughout, possibly due to the longer sample size. The price level impulse response in Figure 3 of Miranda-Agrippino and Ricco has an immediate impact effect of -0.3 percentage points, with the effect growing to -0.7 percentage points 24 months after the shock. This is marginally weaker compared to an impact effect of -0.7 percentage points and a peak of -1 percentage point after around 20 months in Figure 15. Both estimated impulse responses are long-lived an significant throughout. In summary, Miranda-Agrippino and Ricco (2021) estimate real activity, employment, and price level effects in the US that are similar in magnitude to the UK estimates in this study.

Chapter 3

Can language models extract predictive information from the Federal Reserve's Beige Book reports? A quantification of marginal benefits and pitfalls

This study explores the macroeconomic information content of 8,580 of the Federal Reserve System's 'Beige Book' reports, which contain official national and regional commentary about the state of the US economy. To do so, I draw on recent advances in natural language processing, using language models to convert raw text into semantically rich quantitative representations. I then model the distribution of prediction targets using switching Gaussian state space models in a way that permits time-variation in the data generating process whilst keeping inference simple and fast, even when there are thousands of predictors. I estimate the marginal benefit of the predictive signal different methodologies are able to recover from Beige Book text to be moderate, with an average root mean squared forecast error (RMSFE) reduction of 1-3% relative to a non-text benchmark model. This masks significant heterogeneity across forecast targets and horizons, as well as text representation methods. At shorter horizons, text-augmented forecasts appear to near-uniformly perform as good or better than the non-text benchmark. These results are subject to an important caveat that applies to all forecast evaluations using pre-trained language models: the possibility of temporal lookahead bias. To quantify its impact on my results, I perform a knowledge cutoff experiment and find that up to half of the apparent RMSFE reduction resulting from the addition of language model-generated predictors could be illusory.

3.1 Introduction

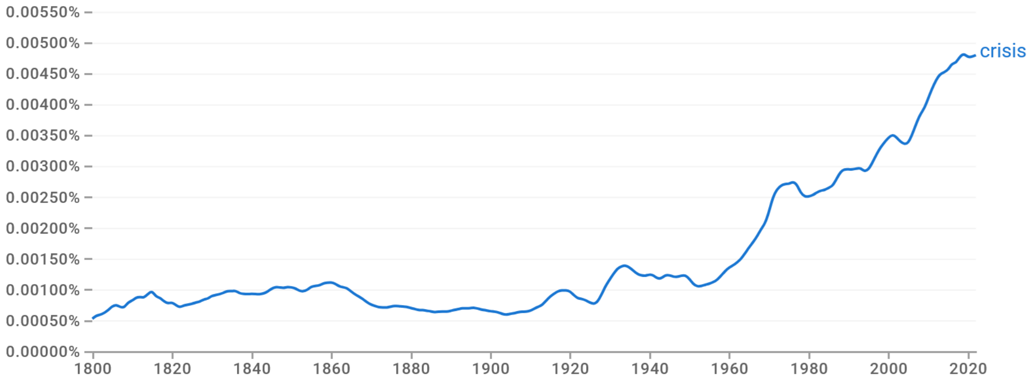
For policymakers and businesses alike, obtaining an accurate picture of the economic outlook is as crucial as it is challenging. Quantitative methods relying on official statistics are vulnerable to miss critical qualitative developments. In an attempt to close gaps left by quantitative economic data, the Federal Reserve has been preparing its 'Beige Book'—formally called 'Summary of Commentary on Current Economic Conditions by Federal Reserve District'—since 1970. Published eight times per year, it continues to attract attention. Popular media coverage includes National Public Radio's 'Beigie Awards' which regularly highlight newsworthy Beige Book anecdotes to a general audience. Each of the 12 regional Federal Reserve Banks[1] employs a team of economists that gather information from a range of sources including local business contacts in a wide range of industries. In addition to district banks' reports, a national summary that provides a national picture of economic conditions based on the regional reports from each district of the Federal Reserve System is prepared. The resulting Beige Book is then shared with the Federal Open Market Committee (FOMC) ahead of each set

[1] Specifically: Atlanta, Boston, Chicago, Cleveland, Dallas, Kansas City, Minneapolis, New York, Philadelphia, Richmond, San Francisco, and St. Louis.

of monetary policy meetings. Despite its long history of consistent preparation, range of sources included, official status, and the public attention paid to it, its value for predicting macroeconomic aggregates is yet to be fully explored.

This paper aims to fill that gap. To do so, I tackle the challenges of predicting macroeconomic aggregates using qualitative reports, which include time-varying text-generating processes, heteroskedastic prediction targets, and the high dimensionality of textual data. To illustrate why time-variation may be a particular issue when using text-derived predictors, Figure 35 shows how the relative frequency of the word 'crisis' in Google's n-gram corpus quadrupled over the last century. This secular change dominates smaller fluctuations that may be attributed to crisis periods of interest, such as the Great Depression in the 1930s or the 1970s energy crisis. Economic forecast targets are typically also highly heteroskedastic, complicating the modelling of their relationship to textual features.

Figure 35: 'Crisis' share of all words in large corpus of books



Notes: Sourced using Google Books Ngram Viewer (`https://books.google.com/ngrams/`).

Finally, as noted by Gentzkow et al. (2019), textual data are intrinsically high dimensional: a lossless numerical representation of word sequences of length W, where words are selected from a dictionary of length L, would require a matrix with $D = L^W$ columns. Even for moderate L and W, this quantity exceeds the number of elementary particles estimated to comprise the observable universe. One way of substantially reducing the dimensionality of a numerical representation of text is to limit a word's context in a document to immediately neighbouring words. Under this simplification, the columns of the matrix would correspond to the counts of words ("unigrams"), word pairs ("bigrams"), or higher "n-gram" tokens in each period. Even after limiting the representation of a document to the counts of its n-gram tokens—resulting in a less rich text representation—the resulting matrix has a sizeable column count.

To overcome these issues, I follow a two-step approach. First, I use language models to obtain dense, fixed-length embeddings of raw Beige Book text in continuous vector space. Such dense embeddings are an alternative to the highly-sparse n-gram representation of text, which map raw tokens into a latent space of "meanings". A reason to think that such richer representations of text may be useful in forecasting is that they may better capture the semantic content of raw text. This is because pre-trained word embeddings are available, whose latent "meaning space" is estimated using massive textual datasets such as the entire corpus of Wikipedia articles. These pre-trained embeddings harness information beyond the dataset at hand to determine how text should be represented. This kind of "transfer learning" may be of particular use in macroeconometrics where the number of observations is typically small.

Secondly, I model the prediction target conditional on low-dimensional compressions of the dense but still high-dimensional language model embeddings obtained in the first step. In particular, I specify a switching linear Gaussian state space model, which enables swift recursive estimation and prediction even for very large numbers of predictors. The estimation procedure involves sampling of compressions, followed by recursive parameter updating and Bayesian model averaging to obtain filtered approximations of functionals of the predictive distribution. In prediction exercises for various simulated data-generating processes, the proposed method outperforms a range of alternative text regression procedures in a squared forecast error sense. Due to recursive estimation, the computational effort to carry out forecast evaluation does not increase materially with the number of evaluation periods—in contrast to existing approaches that require repeated re-estimation to evaluate out-of-sample predictions.

Finally, I consider an important observation by Sarkar and Vafa (2024), Ludwig et al. (2025), and Hoberg and Manela (2025): pre-trained language models can make valid evaluation of temporal prediction tasks challenging. For example, if a language model is trained on a corpus of documents available as of the year 2025, it may—thanks to the benefit of hindsight—learn to associate the concept of mortgage-backed securities with the concept of a severe recession. However, were the language model instead trained on a corpus of documents including information up until the year 2005 this association may not be made. As a result, there may be an illusory boost in predictive performance of forecasts when using the year 2025 embeddings as opposed to the year 2005 embeddings when performing an out-of-sample forecasting experiment. To quantify the impact of this pitfall on my results, I perform a knowledge cutoff experiment.

Following these steps, I find that the marginal benefit of including Beige Book text-derived predictors is moderate, with an average root mean squared forecast error (RMSFE) reduction of 1-3% relative to a non-text benchmark model. There is significant heterogeneity in the extent to which the addition of text-based predictors improves forecasts for different targets and at different horizons. At shorter horizons, for instance, text-augmented forecasts appear to near-uniformly perform as good or better than the non-text benchmark. Considering the extent to which these findings are subject to temporal lookahead bias, I find that up to half of the apparent RMSFE reduction could be illusory.

Related Literature

Empirically, this study contributes directly to the literature examining Beige Book text and its relationship with economic variables (Balke and Petersen, 2002; Zavodny and Ginther, 2005; Armesto et al., 2009; Sadique et al., 2013; Filippou et al., 2024). My primary contribution of quantifying the marginal benefits of text for predicting US macroeconomic aggregates also relates to the literature on macroeconomic forecasting using text (Ardia et al., 2019; Larsen and Thorsrud, 2019; Thorsrud, 2020; Bybee et al., 2020; Kelly et al., 2021; Babii et al., 2022; Larsen et al., 2021; Kalamara et al., 2022; Ellingsen et al., 2022). Methodologically, this study contributes the literature on the general problem of predicting economic quantities when the set of potential predictor variables is large relative to the sample size (Giannone et al., 2021), as well as to wider literature on using text and natural language processing for economic research, as surveyed in (Gentzkow et al., 2019; Algaba et al., 2020; Ash and Hansen, 2023; Korinek, 2023; Hoberg and Manela, 2025). Finally, I contribute to an emerging literature on temporal lookahead bias that can arise when using pre-trained large language models for time series prediction (Sarkar and Vafa, 2024; Ludwig et al., 2025).

Paper structure

Section 3.2 introduces the Beige Book, validates its economic content, and discusses different methods for numerical representation. Section 3.3 proposes a modelling approach to relate the variables generated by numerical text representations to forecast targets, while allowing for time-varying parameters. Section 3.4 quantifies the marginal benefits of incorporating text-derived predictors into forecasts. Section 3.5 describes the setup and results of a knowledge cutoff experiment to quantify the extent to which temporal lookahead issues can bias forecast evaluations. Section 3.6 concludes. Appendix B.1 details the results of monte carlo simulations relating to the method proposed in Section 3.3. Appendix B.2 presents the results of sensitivity analyses of the results in Section 3.4. Appendix B.3 presents the results of subsample analyses.

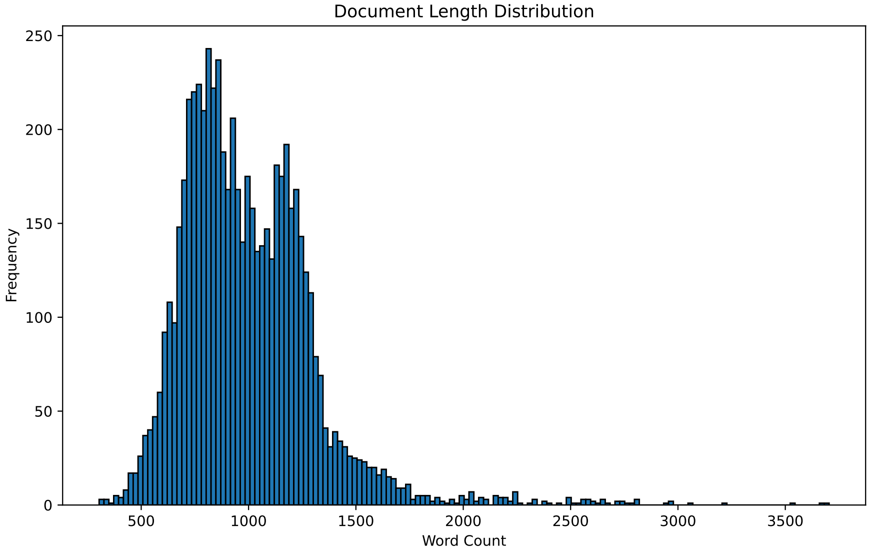
3.2 Representing Beige Book text numerically using document embeddings

In this section, I first analyse the contents of the complete set of Beige Book reports since 1970, exploring variation in content both across time and across districts. Second, I create document embeddings to convert raw Beige Book text into quantitative representations that are of fixed length. These fixed-length representations enable the use of text-derived variables as predictors of future economic conditions.

3.2.1 The Beige Book and its contents

Raw Beige Book text was obtained from the 'Beige Book Archive' on the Federal Reserve Bank of Minneapolis' website. All available reports published between 1970 and 2024 were extracted, yielding a corpus with a total of 8,580 raw textual documents. Since data were obtained directly from the website, some processing was needed to remove characters that do not correspond to Beige Book text such as website navigation components, HTML tags, and whitespace characters. After these steps, the average length of a Beige Book report across the whole sample is 996 words or 6,870 characters. Figure 36 shows that there is some variation in document length, with some Beige Book reports being shorter than 500 words while others exceed 3,500 words. That said, the majority of documents is between 500 and 1,500 words in length.

Figure 36: Distribution of Beige Book report length (in words)



Notes: Histogram based on Beige Book reports for all regional districts and national summary.

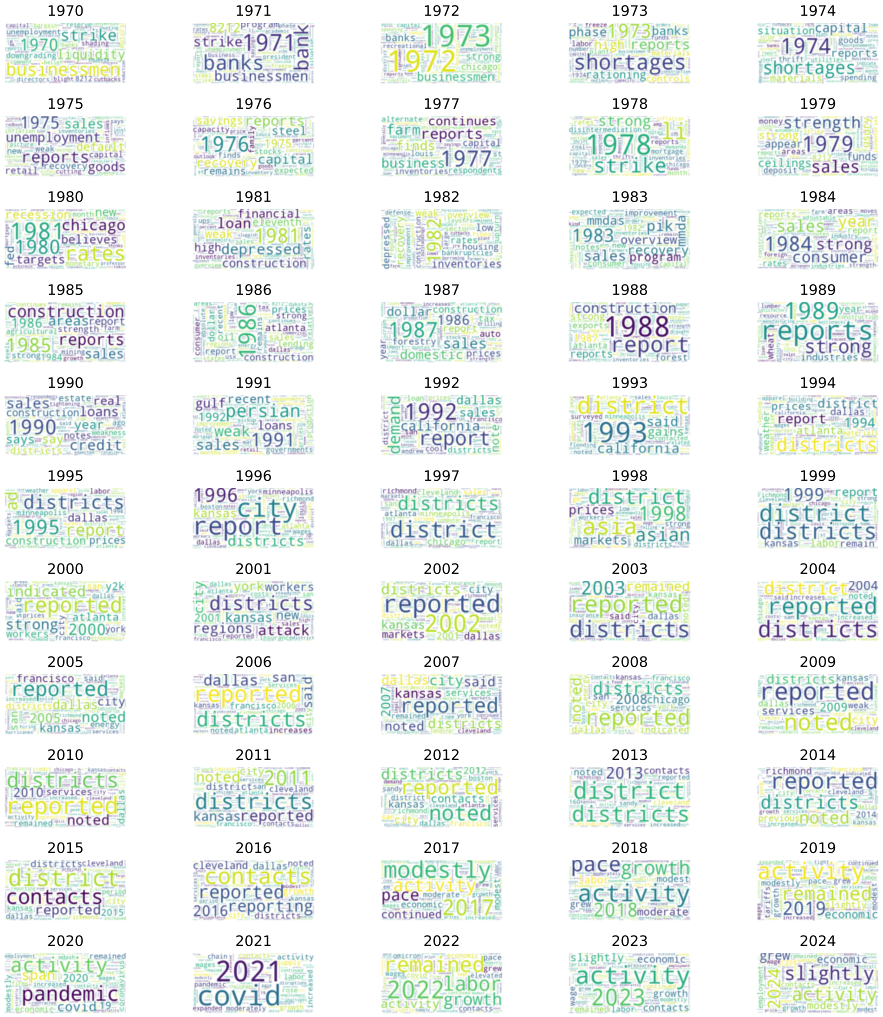
In order to understand the variation and economic content of this corpus, one can use standard natural language processing techniques to identify 'characteristic words'. In particular, focusing on the subset of the corpus corresponding to "national summary" reports between 1970 and 2024, I define the 'characteristicness' of a word to a specific year within the overall corpus as follows

$$tf\tilde{i}df_{w,y} = tfidf_{w,y} - \frac{1}{N_y}\sum_{y} tdidf_{w,y} \tag{3.1}$$

where $tfidf_{w,y}$ is the term frequency-inverse document frequency of word w in year y and N_y is the number of years. That is, I define characteristic words using the difference between a word's year-specific term frequency-inverse document frequency compared to the average across all of the years. The characteristic words resulting from this definition are presented in Figure 37.

One pattern that can be seen is that the words corresponding to the year number or adjacent year numbers appear among the characteristic words. This finding provides assurance that the approach for identifying characteristic words is valid, given that it is likely that these year numbers were mentioned more often during the year in question compared to all other years. More importantly, the characteristic words in Figure 37 appear to include concepts with significant economic content. For example, the word 'strike' is prominent in 1970—the year of an eight-day postal strike involving hundreds of thousands of federal workers (Shannon, 1978). Another characteristic word in 1970 is 'liquidity', coinciding with a commercial paper market liquidity crisis (Nygaard, 2020). During the expansion between 1972 and 1974, characteristic words include 'strong' and 'shortages'. In 1975, a year that saw an unemployment peak of 9%, characteristic words include 'unemployment', 'default', and 'weak'. Frequent mentions of 'recovery', 'strong', 'strike', and 'strength', over the period from 1975 to 1979 align with a recovery in em-

Figure 37: Characteristic words by year (defined as tfidf deviations)



1970 1971 1972 1973 1974 1975 1976 1977 1978 1979 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023 2024

Notes: Each panel displays words characteristic to Beige Book "national summary" reports published within the year indicated above the panel. Figure created using Python's 'wordcloud' package, with $\tilde{tfidf}_{w,y}$ values used as weights.

ployment seen during the same years. In 1980, 1981, and 1982—years during which unemployment rose sharply— ‘recession’, ‘depressed’, ‘weak’, and ‘bankruptcies’ were among the characteristic words. This is followed by ‘recovery’ and ‘improvement’ in 1983. The expansion in the following years until 1990 is reflected in characteristic words such as ‘strong’, ‘construction’, and ‘strength’. The Gulf war in 1991 is reflected in the words ‘persian’ and ‘gulf’. It appears that during the expansion in the following years Beige Book reports focused on regional issues with words such as ‘district’, and district names being among the characteristic words. The Asian financial crisis is also reflected in the characteristic words of Beige Book reports in 1998, for example ‘asia’, ‘asian’, and ‘markets’. The peak of the economic cycle in 2000 is reflected in the word ‘strong’ being prominent. In 2001, unusually frequent use of the word ‘attack’ in Beige Book reports coincided with the events of 9/11. The great recession is reflected by the word ‘weak’ in 2009. That said, in contrast to other years, Figure 37 contains few other characteristic words during 2007 to 2009 that would indicate a significant contraction. As was the case for the 1990s expansion, in the expansion years until 2020 Beige Book reports appear to have focused on regional issues with words such as ‘district’, and district names being among the characteristic words, as well as ‘activity’ and ‘growth’. The first two years of the COVID-19 period are reflected with words such as ‘pandemic’ and ‘covid’. This is followed by characteristic terms ‘growth’, ‘activity’, and ‘grew’ between 2022 and 2024.

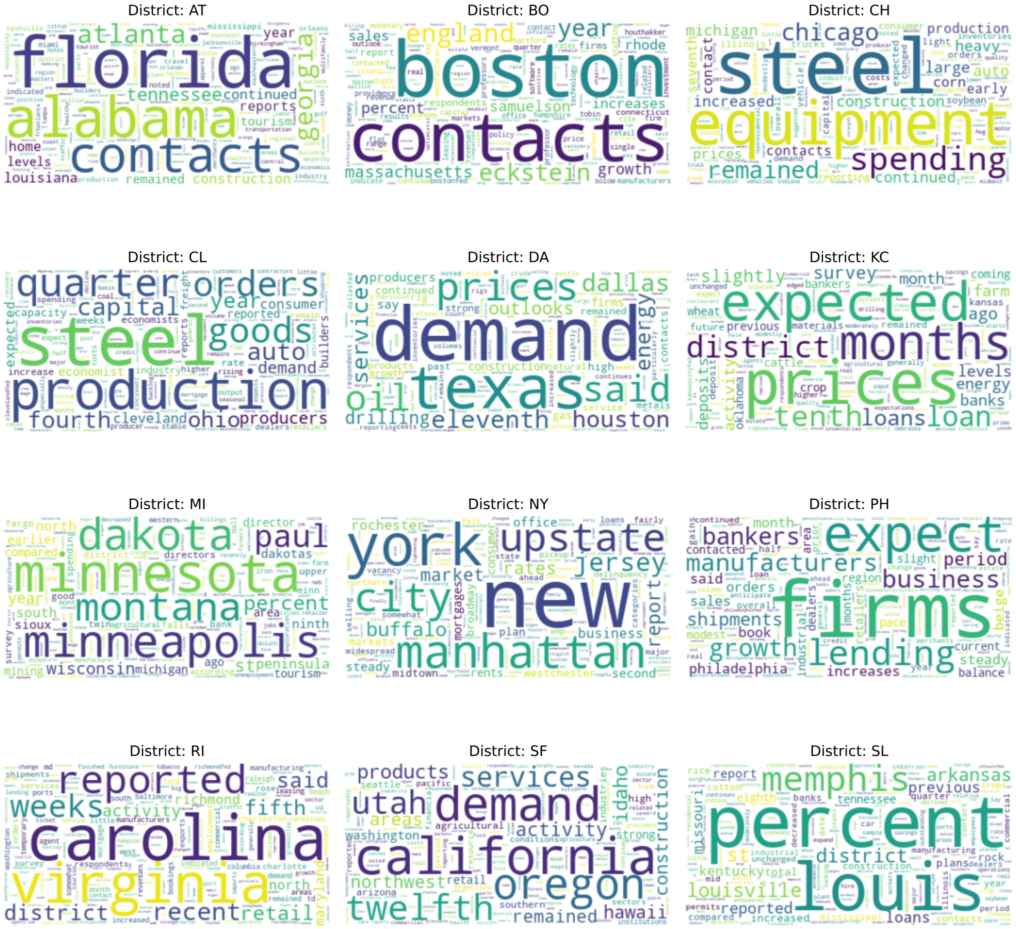
While Figure 37 visualises the heterogeneity of the contents of “national summary” documents across time, Figure 38 displays geographic heterogeneity across reports from the district banks using an analogous definition

$$t\widetilde{fidf}_{w,d} = tfidf_{w,d} - \frac{1}{N_d}\sum_{y} tdidf_{w,d} \tag{3.2}$$

where $tfidf_{w,d}$ is the term frequency-inverse document frequency of word w in district d and $N_d = 12$ is the number of districts. Characteristic words in the reports from each district clearly reflect a local focus, such as state and city names, as well as words relating to sectors that are relatively more important to each district. For example, reports by the Atlanta Fed tend to mention 'florida' 'alabama', 'atlanta', 'tennessee', and 'tourism'. Similarly, characteristic words in the Boston Fed's reports include 'boston', 'massachusetts', and 'rhode [island]'. 'Steel', 'equipment', 'corn', 'auto', 'production' are prominent words in the Chicago and Cleveland banks' reports. The Dallas Fed's reports appear to include frequent mentions of 'oil', 'drilling' and 'energy', in addition to place names such as 'texas', 'houston', and 'dallas'. The Kansas City Fed's reports frequently mention 'crop', 'wheat', 'cattle', 'farm', and 'oklahoma'. Characteristic words in Minneapolis Fed reports include 'mining', 'tourism', 'agricultural', and 'minnesota'. New York Fed reports tend to mention words such as 'rents', 'mortgages', 'delinquency', 'rates', and 'manhattan'. The Philadelphia Fed appears to mention terms such as 'manufacturers', 'industrial', and 'shipments' more commonly than other districts. 'Virginia', 'retail', and 'shipments' are characteristic words for reports from the Federal Reserve Bank of Richmond. San Francisco Fed reports appear to mention terms such as 'agricultural', 'agriculture', 'industries', and 'construction' more than other banks. Finally, the St. Louis Fed tends to mention 'industrial', 'manufacturing', and 'rice'—a significant share of US production of which is concentrated within the bank's district. In sum, the characteristic words in Figure 38 appear to capture plausible differences in the focus of district-specific Beige Book reports.

Taken together, Figures 37 and 38 suggest that the assembled corpus of Beige Book documents contains information about a range of concepts of economic importance at both the national and regional level. This appears to include information about the stage of the business cycle, although it is not immediately clear whether this information is leading, contemporaneous with, or lagging the business cycle.

Figure 38: Characteristic words by district (defined as tfidf deviations)



Notes: Each panel displays words characteristic to Beige Book reports by one of the 12 regional Federal Reserve Banks, specifically Atlanta (AT), Boston (BO), Chicago (CH), Cleveland (CL), Dallas (DA), Kansas City (KC), Minneapolis (MI), New York (NY), Philadelphia (PH), Richmond (RI), San Francisco (SF), and St. Louis (SL). Figure created using Python's 'wordcloud' package, with $\widetilde{tfidf}_{w,d}$ *values used as weights.*

3.2.2 Numerical representation

To investigate the predictive power of Beige Book information, I now discuss how the information contained within the unstructured textual data can be represented in a data format that lends itself to quantitative modelling of the economic outlook.

Gentzkow et al. (2019) discuss how a corpus of raw text $\mathcal{D}$ can be represented in a numerical matrix C. Raw text $\mathcal{D}$ is typically subdivided into a set of documents $\{D_i\}$, depending on the relevant unit of observation. For the purposes of this paper, each individual report (including both regional reports and the national summary) is considered a document. Since the Beige Book is published eight times per year rather than monthly, the first step is to create a balanced monthly panel of report text from each of the 12 district banks as well as the national summary. This is done by forward filling text, such that missing values are replaced by the most recent previous value. This method of filling gaps is chosen to avoid temporal lookahead bias.

Having created a balanced panel of 13 reports per month, the next task is to represent each document in a fixed-length format—that is, the number of variables created from a document should be the same for all documents. One desirable property of such representations is that semantically similar documents have small distances and dissimilar documents have large distances within the corresponding continuous vector space.

A number of modelling approaches can be taken to construct numerical vector c_i for each document D_i, ranging from term frequencies and simple transformations thereof to contextual document embeddings using large, transformer-architecture language models with millions or even billions of parameters. Ash and Hansen (2023) survey a number of numerical text representation approaches, distinguishing between dimensionality reduction through topic modelling, bag-of-word model representations, word embeddings with local context, and sequence embeddings using attention functions. A key difference between these different options for representing text documents numerically include the extent to which word or subword order matters. In a bag-of-word model, reversing or randomising the words within a document would not change the document's numerical representation. This is in contrast to the language models used to obtain word and sequence embeddings, where the context within which a word appears matters for its representation. Moreover, while language models used to generate word embeddings incorporate a small context window around a word when modelling its occurrence, sequence embeddings based on transformer-architecture models mean that the numerical representation of a word or subword can be sensitive to any part of the document. It is this last approach to language modelling that has recently resulted in notable improvements in performance on a range of natural language processing benchmark tasks.

In this study, I consider a number of text modelling approaches, focusing on more recent techniques such as sequence embeddings using attention functions and word embeddings with local context. As a benchmark, I also include two rule-based sentiment analysis techniques that can be used to convert each document into numerical scores.

The resulting document-level representations are listed in Table 9. As Table 9 indicates, the methods differ practically in how embeddings for a document are obtained. Some methods naturally yield embeddings at the document level—such as doc2vec—while others yield embeddings at the word level. In the latter case, embeddings are aggregated to document-level through simple averaging. Another difference relates to the dimensionality of the document representations, ranging from a scalar sentiment score to a vector of length 768. I now discuss each of the methods in Table 9 in detail.

3.2.3 Longformer

The first method I use to obtain numerical representations of Beige Book documents is 'longformer', a transformer-architecture language model with approximately 149 million parameters proposed in Beltagy et al. (2020). As

Table 9: Text representation methods

Method	Text modelling approach	Source	Document representation	Size
longformer	Deep neural network (transformer)	Beltagy et al. (2020)	Direct (using [CLS] token)	768
word2vec	Shallow neural network	Mikolov et al. (2013)	Mean of word embeddings	300
fasttext	Shallow neural network	Mikolov et al. (2018)	Mean of word embeddings	300
doc2vec	Shallow neural network	Le and Mikolov (2014)	Direct	300
vader	Rule-based sentiment	Hutto and Gilbert (2014)	Direct	4
textblob	Rule-based sentiment	Loria et al. (2013)	Direct	1

Ash and Hansen (2023) detail, transformer-based model architectures like the method proposed by Beltagy et al. (2020) have been shown to be effective at a range of natural language processing tasks.[2] With large transformer models, obtaining embeddings for long documents can be challenging, due to computation and memory requirements that increase quadratically with document length. An advantage of the approach proposed in Beltagy et al. (2020) over the widely-known 'BERT' model proposed in Devlin et al. (2019) or the 'FinBERT' sentiment scorer used in Filippou et al. (2024) is that it can handle longer documents—BERT-type models typically impose a document-length limit of 512 tokens (equivalent to around 380 words). This limit is problematic for the forecasting task considered in this paper, where the number of words in Beige Book reports routinely exceeds 1,000 as shown in Figure 36.[3] The longformer method proposed by Beltagy et al. (2020) instead allows for document length of 4,096 tokens (around 3,000 words) and is therefore able to accommodate virtually all Beige Book documents without truncation. I use the 'longformer-base-4096'-model, to obtain embeddings for all Beige Book documents.[4] This model has been trained on a large corpus of text, includ-

[2] Transformer architectures with even larger parameters counts—typically in the billions— account for the remarkable recent advances in 'generative AI' generally (Korinek, 2023).

[3] Filippou et al. (2024) overcome this limitation of FinBERT by applying it to Beige Book reports one sentence at a time. The overall sentiment of a report is then computed by considering the number of positive and negative sentences. While this approach works in the context of aggregating sentiment scores, in general text embeddings do not have a binary interpretation. As such, this aggregation method would not be applicable to 768-dimensional BERT embeddings, for example.

[4] `https://huggingface.co/allenai/longformer-base-4096`

ing the entirety of English Wikipedia articles. An important implication of this is that the embeddings may contain information up until the 'knowledge cutoff' date - i.e. the most recent information included when the transformer model was trained. In the context of a simulated out-of-sample forecasting experiment (as performed in Section 3.4), predictions based on embeddings that rely on information from 'the future' may create subtle temporal lookahead bias (Sarkar and Vafa, 2024) that could invalidate the evaluation of forecast performance. For each document, I use the longformer tokeniser to convert the raw text into a sequence of tokens. A special token ('[CLS]') is then appended at the beginning of the sequence. As the model processes the resulting sequence, information about the whole document is accumulated within the embedding of the [CLS] token. Finally, I obtain this 'document summary' embedding from the output of the last layer of the longformer model. The resulting set of embeddings is of dimension 768. That is, each Beige Book (with its 13 constituent reports) is represented as a vector of length 9,984.

3.2.4 Word2vec

Word2vec, developed in Mikolov et al. (2013), is a method for obtaining embeddings for individual words such that related or similar words have small distances in the embedding space. Compared to transformer language models, which model the text-generating process using deep neural networks, word2vec models text using a shallow neural network architecture. Ash and Hansen (2023) provide a detailed explanation of how word2vec models are pre-trained using self-supervised learning. While word context is considered when embedding vectors are pre-trained, embeddings obtained using pre-trained embedding vectors are static in the sense that the vectors representing words do not depend on the context the word appears in. The specific set of word2vec embeddings used in this study is 'word2vec-google-news-300'[5].

[5] https://huggingface.co/fse/word2vec-google-news-300

These embedding vectors were estimated on a large corpus of Google News articles of around 100 billion words, resulting in a mapping of three million common words/phrases into 300-dimensional continuous vector space. I use this mapping to obtain document embeddings for the Beige Book reports as follows. First, I map every word in the Beige Book document into 300-dimensional vector space. Second, the arithmetic mean of these embeddings is computed to obtain a single embedding for the entire document. This process is repeated for every Beige Book report in the sample, resulting in each Beige Book (i.e. all 13 reports) being represented as a vector of length 3,900.

3.2.5 Fasttext

Bojanowski et al. (2016) propose an extension of word2vec called 'fasttext', which treats words as collections of subwords (i.e. n-grams for character strings). Instead of representing a word as a single vector, fasttext represents it as a sum of vectors for the word's character n-grams, which helps capture subword-level information. Other than this key difference, fasttext works similarly to word2vec, in that embeddings do not depend on the context a word or subword appears in. Embeddings are also estimated using a shallow neural network model. The specific pre-trained fasttext embedding vectors I use are 'fasttext-wiki-news-subwords-300'[6], which map one million words into vectors of size 300. Similar to the word2vec embeddings, these mappings were estimated using a corpus of Wikipedia articles as described in Mikolov et al. (2018). To obtain document embeddings for the Beige Book reports I proceed analogously as is done for word2vec, first mapping every word in the Beige Book document into 300-dimensional vector space and then computing the arithmetic mean of these embeddings for each Beige Book document in the sample. Analogously to word2vec, fasttext document embeddings result in each Beige Book (i.e. all 13 reports) being represented as a vector of length

[6] https://huggingface.co/fse/fasttext-wiki-news-subwords-300

3,900.

3.2.6 Doc2vec

Le and Mikolov (2014) extend the word2vec architecture to include a document (or 'paragraph') vector that represents an overall document rather than single words. This addition eliminates the need to average the word embeddings of all words in a document when using word2vec or fasttext to obtain document-level representations. Similar to word2vec and fasttext, the doc2vec approach models language using a shallow neural network. However, instead of obtaining pre-trained embedding vectors as is done for longformer, word2vec and fasttext, I train the doc2vec model for the specific Beige Book corpus of documents from scratch using the corpus of specific documents at hand.[7] The embedding size for each document is chosen to be 300, which aligns with that of word2vec and fasttext methods, and therefore also results in a vector of length 3,900 for each set of Beige Book reports. Following model training, one can obtain embeddings from the estimated model parameters. As the resulting embeddings are natively at the document level, no additional aggregation step is required.

A key advantage of estimating embeddings using exclusively Beige Book documents rather than using pre-trained embeddings is that I can control the 'knowledge cutoff'. That is, I can obtain embeddings as they would have been obtained at a specific point in time. I exploit this advantage in Section 3.5 to explore the extent to which the temporal lookahead bias highlighted in Sarkar and Vafa (2024) affects forecast results when using different vintages of doc2vec embeddings. In particular, I generate one set of embeddings using the entire corpus up until the end of 2024 and one set of embeddings that only uses Beige Books prepared between 1970 until the end of 2005.

[7] I do so using the 'gensim' Python library with settings $window = 2$ and $min\_count = 1$.

3.2.7 Vader

Hutto and Gilbert (2014) develop ‘valence-aware dictionary for sentiment reasoning’ (VADER), a simple and parsimonious rule-based lexicon technique that maps documents of any length into 4-dimensional vector space. The four scores aim to capture the extent to which the sentiment of the document is positive, neutral, or negative, as well as net sentiment. Despite its simplicity, Kalamara et al. (2022) find that the resulting embedding, while low-dimensional, can capture significant information about macroeconomic aggregates. To explore whether this finding generalises, I apply this technique to all Beige Book reports resulting in a vector of length 52 for each complete Beige Book.

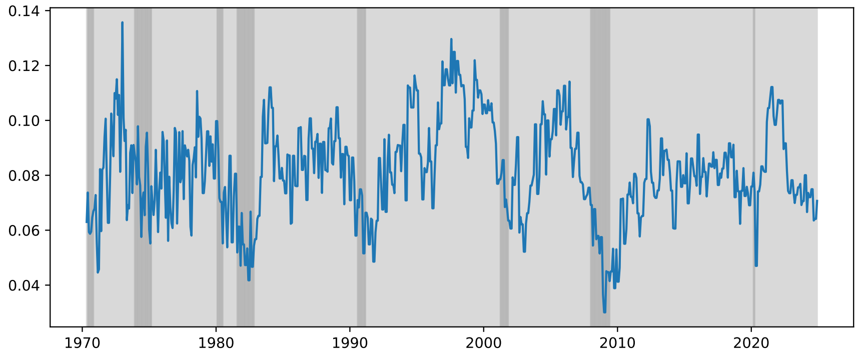
3.2.8 Textblob

Loria et al. (2013) provide another implementation of a simple rule-based technique that aims to capture the sentiment of documents using a sentiment lexicon. In particular, the model outputs a ‘polarity score’[8] which can be between -1 and +1. This overall score is computed by simply averaging the polarity scores of all the words in the documents. As a result, each Beige Book is represented as a vector of length 13. I include the resulting scores in the following analyses as a benchmark for VADER sentiments and the other, more sophisticated, document embedding techniques. Figure 39 shows the time series resulting from this one-dimensional text representation.

The sentiment time series shown in Figure 39 is similar to Figure 1 of Filippou et al. (2024), which suggests that the FinBERT method used in Filippou et al. (2024) and the method by Loria et al. (2013) used in this study yield similar results. One difference between the two figures is that the sentiment time series in Figure 39 is less smooth, due to it not being a rolling average.

[8] Textblob also outputs a ‘subjectivity’ score, but this is not used in this study due to the official nature of Beige Book reports.

Figure 39: “National summary” sentiment over time



Notes: Based on “national summary” reports for each year. Dark shaded periods are NBER peak-to-trough recession indicators for the United States (USREC).

3.3 Time series text regression: Modelling the time-varying relationship between text representations and forecast targets

This section introduces a method for economic forecasting using the quantitative text representations of qualitative Beige Book data obtained through the methods discussed in Section 3.2 as predictors. In particular, the goal is to predict realisations of a univariate time series $\{y_t\}_{t=1}^T$, using realisations of a D-dimensional time series $\{x_t\}_{t=1}^T$ of text-derived predictors. Of interest are functionals of the predictive density $p(y_t|\mathcal{F}_{t-h})$, where h is the prediction horizon and $\mathcal{F}_{t-h}$ is the time $t-h$ filtration/information set generated by the process $\{(y_t, x_t)\}$. I refer to the task of estimating the predictive density $p(y_t|\mathcal{F}_{t-h})$ as *time series text regression.* This terminology is inspired by Gentzkow et al. (2019) who suggest the term *text regression* for the cross-sectional case. A key challenge in performing the time series text regression task is that the dimensionality of text-derived predictors obtained using the representations discussed in Section 3.2 is high relative to the low number of monthly time series observations T that are typically available for key macroeconomic aggregates. Moreover, the relationship between text representations and forecast targets is likely to be time-varying.

In what follows, I propose a relatively simple model and estimation procedure to obtain estimates of a functional of $p(y_t|\mathcal{F}_{t-h})$, namely the expectation $E(y_t|\mathcal{F}_{t-h})$. I tackle the dimensionality issue by assuming that the forecast target depends on low-dimensional *compressed predictors* $\{x_t\Phi\}_{t=1}^{T}$ as in Guhaniyogi and Dunson (2015). The original predictors are compressed in the sense that they are multiplied by $\Phi \in \mathbb{R}^{D\times P}$, a matrix that projects each x_t from high-dimensional $\mathbb{R}^D$ into much lower-dimensional $\mathbb{R}^P$. The assumption of compressed predictors is motivated by the Johnson-Lindenstrauss lemma (Johnson and Lindenstrauss, 1984), which shows that the T textual observations in $\mathbb{R}^D$ can be mapped into $\mathbb{R}^P$ without distorting pairwise squared Euclidean distances by more than a factor of $1 \pm \epsilon$. This factor depends on the subspace dimension P and various bounds exist. Dasgupta and Gupta (2003), for instance, prove that this result holds for $P \geq 4(\epsilon^2/2 - \epsilon^3/3)^{-1} lnT$. In line with the literature on parameter instability in economic time series (e.g. Yousuf and Ng (2021)), my specification allows for sudden regime switches, gradual parameter drift within regimes, and regime-specific stochastic volatility processes. In the parlance of the natural language processing literature, the compressed predictors can be considered an additional embedding layer. Since I allow for regime-switching compressions, this embedding layer is allowed to vary over time.

My estimation approach begins similarly to the Bayesian compressed regression methodology of Guhaniyogi and Dunson (2015), who propose a sampling-based estimator for the cross-sectional case that is shown to converge to the non-compressed predictive density $p(y_i|x_i)$. For the time series case considered here, I sample Φ from a proposal distribution similar to Guhaniyogi and Dunson (2015) and use these samples to obtain compressed predictors. Conditional on these low-dimensional compressed predictors, I use the forgetting factor approach of Raftery et al. (2010) and Koop and Korobilis (2012) to obtain filtered approximations of the predictive density using Kalman-style recursions and Bayesian model averaging.

Rather than specifying a generative process as in Kelly et al. (2021) to model time series of phrase counts, I instead model the target time series $\{y_t\}_{t=1}^T$ directly, *conditional on* the textual time series $\{x_t\}_{t=1}^T$ constructed in Section 3.2. The key advantage of doing this through compressed predictors $\{x_t\Phi\}_{t=1}^T$ is that the number of parameters to be estimated is reduced significantly. This makes the estimation procedure computationally feasible even for extremely large numbers of predictors, where off-the-shelf methods such as penalised linear regression or neural networks can suffer from convergence issues or be computationally impractical. Moreover, evaluating forecasts generated by the proposed methodology is computationally straightforward due to its recursive estimation procedure.

I next introduce my modelling approach, followed by sections on the estimation procedure and computational considerations.

3.3.1 Model

The data-generating process of prediction target $y_t \in \mathbb{R}$ given a high-dimensional set of text-derived predictors $x_t \in \mathbb{R}^{1\times D}$ is modelled as a switching linear Gaussian state space model of the form

$$y_t = \sum_{k=1}^{K} 1_{L_t=k}(x_t\Phi^{(k)}\theta_t^{(k)} + e_t^{(k)}),\ \textit{where}\ e_t^{(k)} \sim N(0, H_t^{(k)}) \tag{3.3}$$

$$\theta_t^{(k)} = \theta_{t-1}^{(k)} + \delta_t^{(k)},\ \textit{where}\ \delta_t^{(k)} \sim N(0, W_t^{(k)}). \tag{3.4}$$

In this specification, $L_t \in \{1, ..., K\}$ is a regime switching process that indicates which of the K possible regime models determines y_t in period t. The key component of each regime model is the projection matrix $\Phi^{(k)} \in$

$\mathbb{R}^{D \times P^{(k)}}$. Conditional on $L_t = k$, I therefore assume that y_t depends on the *compressed* regressors $x_t \Phi^{(k)}$. I also assume throughout that $P^{(k)}$ is much smaller than D, so that the size of the time-varying coefficient vector $\theta_t^{(k)}$ remains manageable even when D is extremely large.

In the special case where $L_t = L \ \forall \ t$ (no regime switching), $W_t^{(k)} = 0$ (constant coefficients) and $H_t^{(k)} = H$ (constant innovation variance), this model reduces to the compressed regression model considered in Guhaniyogi and Dunson (2015). I here consider the compression approach in the general case, where L_t, $W_t^{(k)}$, and $W_t^{(k)}$ are time-varying rather than constant or zero. Instead of specifying these processes explicitly, however, I proceed by using the forgetting-factor approximations of Raftery et al. (2010) and Koop and Korobilis (2012). With forgetting factors, the temporal evolution of L_t, $W_t^{(k)}$, and $H_t^{(k)}$ is characterised implicitly by a combination of the forgetting parameters and past realisations.

In contrast to the approaches of Taddy (2015) and Kelly et al. (2021), the text-derived predictors x_t are not required to be counts. As such, one can accommodate the various numerical representations of text discussed in Section 3.2, as well as all arbitrary transformations thereof—such as moving averages or period-on-period changes.

Finally, one may have a strong prior belief that a small number of non-textual features z_t are also relevant, for instance an intercept or lagged values of 'hard' economic time series. Predictors of this kind can be incorporated by augmenting the model with a non-compressed term $z_t \beta_t^{(k)}$ for each regime k, where $\beta_t^{(k)}$ are time-varying coefficients that evolve analogously to $\theta_t^{(k)}$. As each regime term would include these predictors, they would not be subject to shrinkage.

3.3.2 Recursive estimation and prediction

The goal here is to extend the estimation approach of Guhaniyogi and Dunson (2015) to the general case introduced above. Guhaniyogi and Dunson begin by drawing a single candidate $\Phi^{(k)}$ for each model k from a proposal distribution. Conditional on the model being $L = k$ in all periods, an estimate of the joint posterior distribution of the time-invariant $\theta^{(k)}$ and $H^{(k)}$, and of the posterior predictive distribution can then be obtained analytically under a standard conjugate normal-inverse gamma prior. The model-conditional predictive densities are then aggregated using Bayesian model averaging with a uniform prior over models.

My estimation strategy follows this general approach, but updates posterior distributions recursively as the specification allows $\theta_t^{(k)}$, $H_t^{(k)}$, and L_t to vary over time. Similar to Guhaniyogi and Dunson, I do not attempt to to estimate each regime's projection matrix. Instead, I obtain a single draw of candidate regime projections $\{\hat{\Phi}^{(1)}, ..., \hat{\Phi}^{(K)}\}$. Conditional on these projections, one can then obtain filtered estimates of the predictive distribution of the switching linear Gaussian state space model (Equations 3.3-3.4) using the approximate inference procedures proposed in Raftery et al. (2010) and Koop and Korobilis (2012).

Drawing Φ

For each regime model, I first draw the number of dimensions P of the subspace that the textual features are to be projected to as follows[9]

$$P^{(k)} \sim Uniform(2log(D), UB).$$

While Guhaniyogi and Dunson (2015) suggest an upper bound for the subspace dimension count of $UB = min(T, D)$, I note that high draws for UB can lead to computational issues such as running out of working memory. To avoid such issues in practice, I suggest choosing UB depending on machine specifications. With the number of columns $P^{(k)}$ of $\Phi^{(k)}$ determined, the raw elements $\phi_{i,j}^{(k)*}$ are then drawn independently exactly as in Guhaniyogi and

[9] Rounded to the nearest natural number.

Dunson (2015)

$$\phi_{i,j,k}^{(k)*} \sim \begin{cases} -\sqrt{\frac{1}{s}} & \text{with probability } s^2 \\ 0 & \text{with probability } 2(1-s)s \\ -\sqrt{\frac{1}{s}} & \text{with probability } (1-s)^2 \end{cases}$$

where

$$s^{(k)} \sim Uniform(0.1, 0.9).$$

Finally, a finished draw of $\hat{\Phi}^{(k)}$ is obtained by applying Gram-Schmidt orthonomalisation to the matrix of raw elements $\{\phi_{i,j}^{(k)*}\}$.

Parameter inference

Given K projection matrix draws $\{\hat{\Phi}^{(k)}\}$, one can use Kalman-type recursions for inference about each regime's parameter vector $\theta_t^{(k)}$. Raftery et al. (2010) show that under a forgetting factor specification, regime-conditional filtering densities have the following form

$$\theta_t^{(k)} | L_t = k, y_{1:t-1} \sim N(\hat{\theta}_{t|t-1}^{(k)}, \Sigma_{t|t-1}^{(k)}) \text{ (parameter prediction equation)}$$

$$\theta_t^{(k)} | L_t = k, y_{1:t} \sim N(\hat{\theta}_{t|t}^{(k)}, \Sigma_{t|t}^{(k)}) \text{ (parameter updating equation).}$$

In the prediction step, one can predict next period's value of the state density mean and variance

$$\hat{\theta}_{t|t-1}^{(k)} = \hat{\theta}_{t-1|t-1}^{(k)}$$

$$\Sigma_{t|t-1}^{(k)} = \frac{1}{\lambda}\Sigma_{t-1|t-1}^{(k)}$$

where λ is a forgetting factor a little smaller than one that implicitly defines the evolution of the state innovation variance W_t. Once one observes y_t, one can update the time t filtered estimates as follows

$$\hat{\theta}^{(k)}_{t|t} = \hat{\theta}^{(k)}_{t|t-1} + \Sigma^{(k)}_{t|t-1}(x_t\hat{\Phi}^{(k)})^T(H^{(k)}_t + x_t\hat{\Phi}^{(k)}\Sigma_{t|t-1}(x_t\hat{\Phi}^{(k)})^T)^{-1}(y_t - x_t\hat{\Phi}^{(k)}\hat{\theta}^{(k)}_{t|t-1})$$

$$\Sigma^{(k)}_{t|t} = \Sigma^{(k)}_{t|t-1} - \Sigma^{(k)}_{t|t-1}(x_t\hat{\Phi}^{(k)})^T(H^{(k)}_t + x_t\hat{\Phi}^{(k)}\Sigma^{(k)}_{t|t-1}(x_t\hat{\Phi}^{(k)})^T)^{-1}x_t\hat{\Phi}^{(k)}\Sigma^{(k)}_{t|t-1}.$$

This updating step requires knowledge of $H^{(k)}_t$, the observation innovation variance. I here follow Koop and Korobilis (2012)'s use of Exponentially Weighted Moving Average (EWMA) estimation. That is, one can plug into the above update steps the following EWMA-implied forecast for $H^{(k)}_t$

$$\hat{H}^{(k)}_{t|t-1} = \kappa\hat{H}^{(k)}_{t-1|t-2} + (1-\kappa)(y_{t-1} - x_{t-1}\hat{\Phi}^{(k)}\hat{\theta}^{(k)}_{t-1|t-1})$$

where κ is the EWMA decay/forgetting factor. With this approximation, the regime-conditional predictive distributions become

$$y_t|L_t = k, y_{1:t-1} \sim N(x_t\hat{\Phi}^{(k)}\hat{\theta}^{(k)}_{t|t-1}, \hat{H}^{(k)}_{t|t-1} + x_t\hat{\Phi}^{(k)}\Sigma^{(k)}_{t|t-1}(x_t\hat{\Phi}^{(k)})^T). \qquad (3.5)$$

To operationalise the above recursions, one needs to decide initial values $\hat{H}^{(k)}_{0|-1}$, $\hat{\theta}^{(k)}_{0|0}$, and $\Sigma^{(k)}_{0|0}$ for each regime. There should generally be sufficient information to judge the order of magnitude of the dependent variable. As such, my implementation simply sets $\hat{H}^{(k)}_{0|-1} = \hat{Var}(y_t)$. The latter two initialisations concern the slope coefficients on compressed regressors, which are harder to reason about. I here follow Raftery et al. (2010) who suggest $\hat{\theta}^{(k)}_{0|0} = 0$ and $\Sigma^{(k)}_{0|0} = diag(\hat{Var}(y_t)/\hat{Var}(x_t\hat{\Phi}^{(k)}))$. If there is an intercept term, one can set the corresponding diagonal entry of $\Sigma^{(k)}_{0|0} = \hat{Var}(y_t)$. To avoid lookahead biases in a forecast evaluation exercise, these quantities should be estimated on the in-sample observations only.

Regime averaging

Raftery et al. (2010) show how another forgetting factor $\alpha \in [0, 1]$ can be used to avoid having to specify a $K \times K$ transition matrix for regime indicator L_t. Under this approach, one can recursively predict and update regime probabilities as follows

$$P(L_t = k|y_{1:t-1}) = \frac{P(L_{t-1} = k|y_{1:t-1})^{\alpha} + c}{\sum_{i=1}^{K} P(L_{t-1} = i|y_{1:t-1})^{\alpha} + c} \text{ (regime prediction equation)}$$

$$P(L_t = k|y_{1:t}) = \frac{max(f_t^{(k)}, c)P(L_t = k|y_{1:t-1})}{\sum_{i=1}^{K} max(f_t^{(i)}, c)P(L_t = i|y_{1:t-1})} \text{ (regime updating equation)}$$

where $f_t^{(k)}$ is the regime-conditional predictive distribution in Equation 3.5, evaluated at the true y_t, and $c > 0$ is a small constant that serves to ensure numerical stability.[10] Under a uniform prior over regimes, the above recursions can be initialised with $P(L_0 = k) = 1/K$.

Following Raftery et al. (2010) and Koop and Korobilis (2012), one can obtain the unconditional predictive distribution as a mixture of the regime-conditional predictive distributions weighted by the predicted regime probabilities, which implies the following point predictions[11]

$$\hat{y}_t = \sum_{k=1}^{K} x_t \hat{\Phi}^{(k)} \hat{\theta}^{(k)}_{t|t-1} P(L_t = k|y_{1:t-1}).$$

[10] I here follow Raftery et al. who suggest $c = 0.001/K$.

[11] For forecast horizons h different from one, I use information up to time $t - h$, i.e. $\hat{\theta}^{(k)}_{t|t-h}$ and $P(L_t = k|y_{1:t-h})$.

3.3.3 Computational implementation

The computational steps of the proposed procedure are displayed as pseudocode in Algorithm 1. Computationally expensive steps are the drawing of candidate projection matrices and compressing the predictor matrix X. Both of these steps are part of the first K-loop and can be parallelised straightforwardly. Once a text predictor matrix is compressed, the remaining steps deal with low-dimensional compressed predictors and are therefore computationally inexpensive. Conditional on the density evaluations obtained from each regime model, the model averaging step is independent of the parameter estimation step, making it nearly instantaneous.

A key advantage of the proposed method for time series prediction/forecasting using textual data is that it is estimated recursively. This enables faster evaluation of predictive performance and hyperparameter optimisation relative to methods that require batch estimation. For a faithful evaluation of forecasting performance, batch methods such as the topic model regressions of Ellingsen et al. (2022), the hurdle inverse regression of Kelly et al. (2021), or the non-linear models of Kalamara et al. (2022) need to be re-estimated repeatedly at considerable computational cost. Over a 15-year evaluation horizon with monthly data, for instance, this implies re-estimating the model 180 times. Such repeated estimation can be cumbersome when the number of textual predictors is large.

Appendix B.1 contains a simulation exercise in which the method developed in this section is compared to alternative methods for time series text regression. Compared to alternatives, the method appears particularly useful when the number of text-derived predictors is large, text density is low, and time-variation strong. The estimation procedure is computationally convenient even for the very large predictor counts seen with numerical representations of textual data. This is in contrast to methods relying on batch estimation, which are more computationally complex for time series prediction problems involving many text-derived predictors.

Algorithm 1 Time series text regression using Bayesian model averaging

```
Inputs
    Prediction target data: $y \in \mathbb{R}^{T \times 1}$
    Text-derived predictor data: $X \in \mathbb{R}^{T \times D}$
    Compression count: $K$
    Upper bound for $P$ draw: $UB$
    Coefficient forgetting factor: $\lambda$
    Regime forgetting factor $\alpha$
    EWMA forgetting factor: $\kappa$
for $k \in \{1, ..., K\}$ do
    Draw $\hat{\Phi}^{(k)}$
    Compress X (using draw $\hat{\Phi}^{(k)}$)
    Initialise $\hat{\theta}^{(k)}_{0|0}$
    Initialise $\Sigma^{(k)}_{0|0}$
    Initialise $\hat{H}^{(k)}_{0|-1}$
    Initialise $P(L_0 = k)$
end for
for $t \in \{1, ..., T\}$ do
    for $k \in \{1, ..., K\}$ do
        Predict $\hat{\theta}^{(k)}_{t|t-1}$
        Predict $\Sigma^{(k)}_{t|t-1}$
        Predict $\hat{H}^{(k)}_{t|t-1}$
        Predict $P(L_t = k|y_{1:t-1})$
    end for
    Normalise $P(L_t = k|y_{1:t-1})$ $\forall k$
    Predict $\hat{y}_t$
    for $k \in \{1, ..., K\}$ do
        Update $\hat{\theta}^{(k)}_{t|t}$
        Update $\Sigma^{(k)}_{t|t}$
        Update $P(L_t = k|y_{1:t})$
    end for
    Normalise $P(L_t = k|y_{1:t})$ $\forall k$
end for
Return $\{\hat{y}_t\}_t$
```

3.4 Estimating the incremental signal strength of Beige Book text

Having introduced a method for dealing with the high dimensionality and time-variation of textual data in Section 3.3, this section aims to estimate the value of Beige Book text for forecasting US macroeconomic aggregates. In other words, I aim to estimate the strength of predictive signals within the Beige Book corpus. Of particular interest is determining the extent to which the different text representations of Section 3.2, in conjunction with the modelling approach developed in Section 3.3, can enhance forecasting performance relative to standard methods that do not incorporate text-derived information. The approach developed in Section 3.3 is used to test the marginal value of text-derived predictors directly, by augmenting a linear model with uncompressed standard predictors with compressed textual variables.

3.4.1 Forecast targets

I focus on forecasting key monthly US macroeconomic time series in the FRED-MD monthly database[12] introduced in McCracken and Ng (2016). This is an increasingly standard database for studying US macroeconomic dynamics, especially forecasting. In what follows, I use the '2024-12' vintage of the database. That is, I use the latest available revisions of the economic time series available at the end of December 2024. An alternative approach to estimating the marginal predictive power of textual data would involve iterating through the set of historical vintages of the database, using only those data points that were available at particular points in time. The reason for working with a recent vintage, as opposed to simulating past information sets, is that there were changes in the set of variables included in the database. Implementing an out-of-sample exercise that takes the chang-

[12] https://fred.stlouisfed.org/

ing shape of the information set into account would require re-estimating all models repeatedly in response to variable changes. In this particular context with high-dimensional predictors, many forecast targets and horizons, as well different hyperparameters, doing so would be computationally costly. An exercise that accounts for revisions between the original vintage and the final revised value is therefore left for future work.

I consider the same prediction targets as Kelly et al. (2021), with the exception of a manufacturing index that is no longer published. The remaining forecast target series include the industrial production index, total nonfarm employment, the S&P 500 equity market index, the number of housing starts, the consumer price index, average hourly earnings in goods-producing sectors, the unemployment rate, the effective federal funds rate, as well as the consumer sentiment index, and are summarised in Table 10.

I consider forecast horizons h ranging from one month to 36 months ahead, specifically 1, 3, 6, 12, 24, and 36 months ahead. Predictions are made using direct forecasting, using a separate model for each forecast horizon h. This is approach is taken to avoid accumulation of forecast errors, particularly at longer forecast horizons. To stationarise forecast targets, I follow McCracken and Ng (2016) and transform series to average annualised monthly growth rates as detailed in Table 10. In particular, for the industrial production ('INDPRO'), nonfarm employment ('PAYEMS'), equity market index ('S&P 500'), and housing starts ('HOUST'), series I define the h-month ahead forecast target as follows

$$y_{t+h} = (1200/h)log(var_{t+h}/var_t).$$

For the nominal variables, such as the consumer price index ('CPIAUCSL') and average hourly earnings ('CES0600000008') I also follow McCracken and Ng (2016) and define target series as

$$y_{t+h} = (1200/h)log(var_{t+h}/var_t) - 1200log(var_t/var_{t-1}).$$

Finally, for the unemployment rate ('UNRATE'), effective federal funds rate ('FEDFUNDS'), and consumer sentiment ('UMCSENTx') I define the forecast target as annualised monthly level changes as follows

$$y_{t+h} = (1200/h)(var_{t+h} - var_t).$$

Table 10: Forecast targets

FRED mnemonic	Description	Transformation
INDPRO	Industrial production index	$(1200/h)log(var_{t+h}/var_t)$
PAYEMS	All employees: total nonfarm	$(1200/h)log(var_{t+h}/var_t)$
S&P 500	S&P's common stock price index: composite	$(1200/h)log(var_{t+h}/var_t)$
HOUST	Housing starts: total new privately owned	$(1200/h)log(var_{t+h}/var_t)$
CPIAUCSL	Consumer price index: all items	$(1200/h)log(var_{t+h}/var_t) - 1200log(var_t/var_{t-1})$
CES0600000008	Average hourly earnings: Goods-Producing	$(1200/h)log(var_{t+h}/var_t) - 1200log(var_t/var_{t-1})$
UNRATE	Civilian unemployment rate	$(1200/h)(var_{t+h} - var_t)$
FEDFUNDS	Effective federal funds rate	$(1200/h)(var_{t+h} - var_t)$
UMCSENTx	Consumer sentiment index	$(1200/h)(var_{t+h} - var_t)$

A key advantage of the method introduced in Section 3.3 is that it allows for recursive estimation, simplifying the task of forecast evaluation considerably. Using the '2024-12' vintage of McCracken and Ng (2016)'s FRED-MD database, I estimate models on data between 1970 and 2019, and evaluate forecasts for the observations between December 1999 and November 2019. Missing data are generally replaced using the 'last observation carried forward' technique. In order to plug gaps at the beginning of the sample 'first observation carried backward' is used. These early missing values do not fall into the evaluation window.

3.4.2 Benchmark model specification

The specification I use to obtain forecasts based on the standard, non-text variables follows the benchmark approach in Kelly et al. (2021). In particular, the first five principal components (denoted by $pc1, ..., pc5$) are computed once from all series in the McCracken and Ng (2016) database. Prior to forming the principal components, all series in the database are stationarised by following the authors' recommendations. For each prediction target y and forecast horizon h, benchmark predictions are then generated as follows

$$\hat{y}_{t+h} = [1, y_t, pc1_t, pc2_t, pc3_t, pc4_t, pc5_t]'\hat{\beta}_h \tag{3.6}$$

where $\hat{\beta}_h$ is the estimate obtained using ordinary least squares estimation. To account for often strongly autoregressive patterns, the benchmark specification also includes the last available value of the forecast target, y_t, in addition to an intercept and the principal components.

3.4.3 Time series text regression model specifications

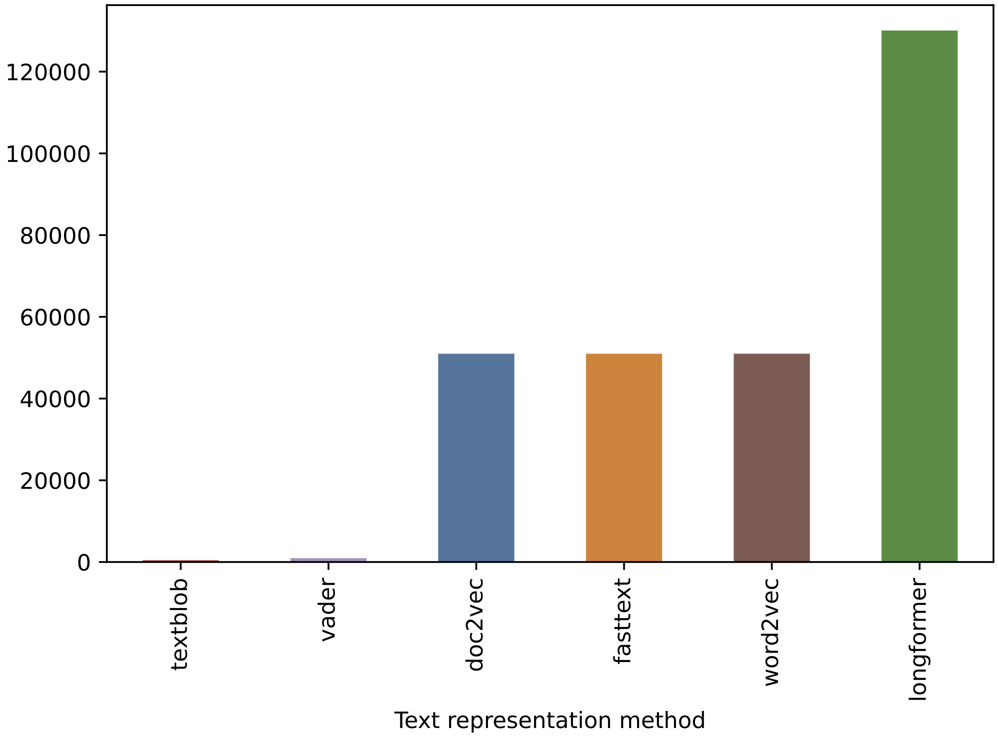
Text-augmented predictions for each forecast target y at every horizon h are generated using the empirical technique introduced in Section 3.3. In particular, as model coefficients are updated recursively, forecasts are obtained as follows

$$\hat{y}_{t+h} = \sum_{k=1}^{K} \hat{P}(L_{t+h} = k|y_{1:t})(x_t\hat{\Phi}^{(k)}\hat{\theta}^{(k)}_{t+h|t}+[1, y_t, pc1_t, pc2_t, pc3_t, pc4_t, pc5_t]'\hat{\beta}^{(k)}_{t+h|t}) \tag{3.7}$$

where x_t contains sets of text-derived predictors as described in Section 3.2. That is, the predictors of the text-augmented model are a superset of the predictors of the benchmark model, with the set difference being made up of exclusively text-derived variables. Only the text-derived variables are compressed. The non-textual features $z_t = [1, y_t, pc1_t, pc2_t, pc3_t, pc4_t, pc5_t]$ enter the model without compression as discussed in Section 3.3.

To be able to compare different text representations, I implement one specification for each method discussed in Section 3.2. In addition to the levels of the numerical text representations, both their lags and differences are computed using 1, 3, 6, 12, 24, and 36-month offsets and included within x_t. Figure 40 shows the resulting number of features for each text representation method.

Figure 40: Number of text-derived features by text representation method



Notes: The total number of text-derived features is comprised of the dimensions of the numerical text representations themselves, as well as their 1, 3, 6, 12, 24, and 36-month lags and differences.

For each dimension shown in Table 9, there are 169 features, as a result of there being 13 Beige Book reports each month (which are represented separately) and there being six lags, six differences, as well as the untransformed text representation variables. As a result, the sentiment methods have by far the smallest number of predictors, with 169 features being constructed from the scalar textblob sentiment and 676 features for the 4-dimensional VADER sentiment score. The language model-based embeddings of size 300 give rise to 50,700 text-derived features, while the longformer embeddings of size 768 give rise to a text-derived feature set of size 129,792.

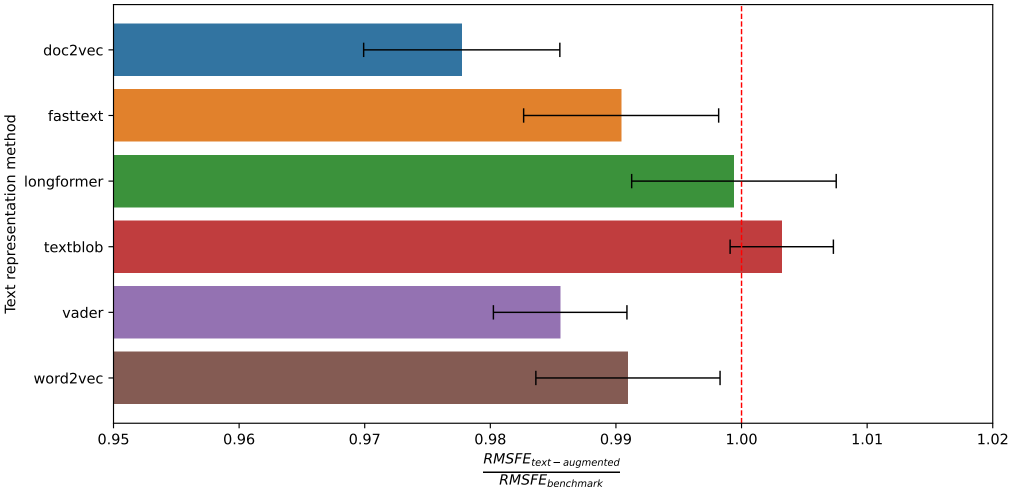
A number of parameter choices need to be made to implement the estimation procedure proposed in Section 3.3 and studied in Appendix B.1. The first parameter to be chosen is the compression count (i.e. the number of regimes of the switching state-space model) K. Based on simulation results, I estimate each specification with both $K = 16$ and $K = 32$. The second parameter is the upper bound for the number of subspace dimensions UB. While Guhaniyogi and Dunson (2015) recommend a value of $UB = min(T, D)$, this is not practicable for the large predictor sets resulting from the text representations discussed in Section 3.2. As such, performance for smaller maximum subspace dimensions of 2, 4, 8, 16, and 32 are explored. Finally, the forgetting factors λ and α are set to either $\lambda = \alpha = 1$ (i.e. no forgetting) or $\lambda = \alpha = 0.999$. With a sample size of around 550 observations (depending on the forecast horizon), the latter forgetting factor value can be interpreted as observations at the beginnning of the sample receiving half the weight of the most recent observation.

3.4.4 Results

Figure 41 shows the overall relative performance of the text-augmented specifications.

The relative performance displayed is the average across all forecast targets and forecast horizons. Error bars indicate the 95% confidence interval—representing variation across forecast targets and forecast horizons – for the mean relative performance. Results were obtained using time series text regression parameters $K = 16$ and $UB = 8$, although Figure 53 in the Appendix illustrates that results are not particularly sensitive to these parameter choices. There is variation in the extent to which different text representations result in lower RMSFE than the benchmark specification. The lowest RMSFE is attained by the specification using document embeddings

Figure 41: Relative performance of text-augmented forecasts by representation method: overall



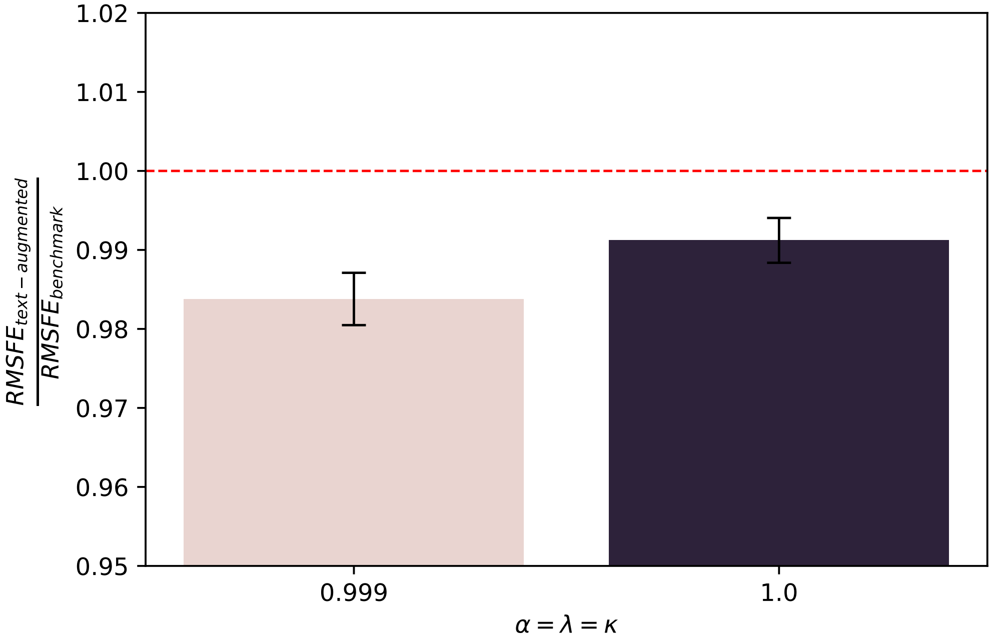
Notes: Relative RMSFE of text-augmented forecast relative to benchmark forecast. For each text representation method, mean ratios are computed across the different forecast targets in Table 10 and across horizons. Error bars represent 95% confidence intervals for the mean ratio based on the standard error of the mean computed across all ratios. Parameter settings: $K = 16$, $UB = 8$, $\alpha = \lambda = \kappa = 1$.

obtained using the doc2vec method. Considering all forecast targets and horizons, this specification achieves an average RMSFE 2.5 percent below that of the benchmark. Interestingly, the longformer specification appears to generate forecasts about as accurate as the benchmark model. This is also the case for the textblob sentiment score, while the VADER sentiment score appears to result in an average RMSFE reduction of around 1.5 percent. Forecasts using word embeddings to obtain document representations perform similarly with an average RMSFE reduction of around one percent.

While the results in Figure 41 were obtained under the constant parameter restriction ($\alpha = \lambda = \kappa = 1$), Figure 42 illustrates how results differ when allowing for time-varying parameters. As discussed above, parameter setting $\alpha = \lambda = \kappa = 0.999$ can be interpreted as observations at the beginning of the sample receiving around half the weight as the most recent observation. Figure 42 shows that this setting appears to improve average performance across forecast targets and horizons. That is, the flexibility of the time series text regression approach introduced in Section 3.3 appears to translate into a reduction in out-of-sample RMSFE of around one percent on average. Figure 43 shows disaggregated results by individual forecast target and horizon.

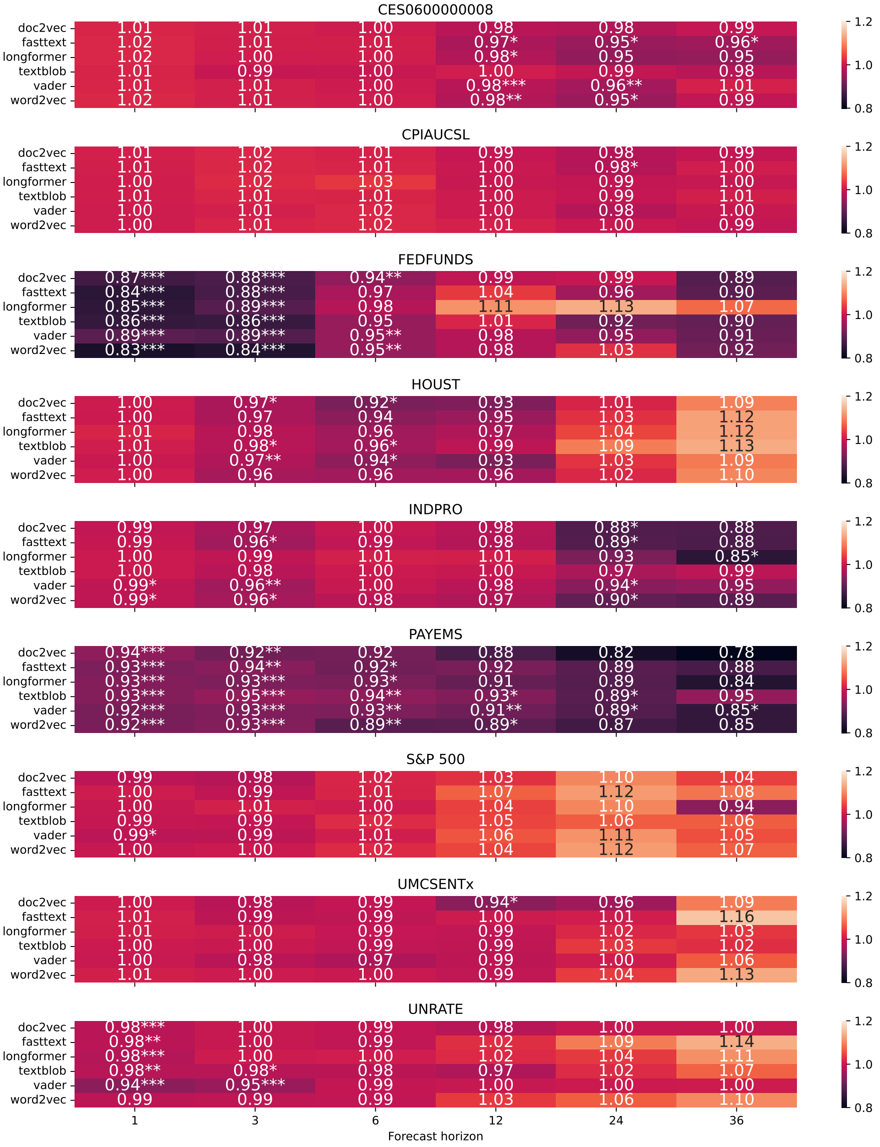
Forecasts are evaluated using the standard RMSFE metric, combined with Diebold and Mariano (1995) p-values indicating whether differences in RMSFE-accuracy between the benchmark specification and each of the text-augmented specifications are statistically significant in a frequentist sense. Given the observation in Figure 42, Figure 43 shows results for the specification with time-varying parameters ($\alpha = \lambda = \kappa = 0.999$). Detailed results for the constant-parameter text-augmented specification are shown in Figure 54 in Appendix B.2. There is considerable heterogeneity in gains of forecast accuracy across forecast targets and horizons. In particular, the use of time series text regression results in significantly lower RMSFE when forecasting the effective Federal Funds Rate, nonfarm employment, and the unemployment rate at shorter horizons. For the first two, the RMSFE re-

Figure 42: Relative performance of text-augmented forecasts by parameter rigidity: overall



Notes: Relative RMSFE of text-augmented forecast relative to benchmark forecast. For each set of forgetting factors, mean ratios are computed across the different forecast targets in Table 10 and across horizons. Error bars represent 95% confidence intervals for the mean ratio based on the standard error of the mean computed across all ratios. Parameter settings: $K = 16$, $UB = 8$.

Figure 43: Relative performance of text-augmented forecasts: detail



Notes: Relative RMSFE of text-augmented forecast relative to benchmark forecast. Asterisks indicate Diebold and Mariano (1995) p-values, with a single asterisk indicating $p <= 0.1$, two asterisks indicating $p <= 0.05$ and three asterisks indicating $p < 0.01$. Parameter settings: $K = 16$, $UB = 8$, $\alpha = \lambda = \kappa = 0.999$.

duction amounts to around 5 to 15 percent. At longer horizons, forecasts for hourly earnings, industrial production, and nonfarm employment appear to be improved by incorporating text-derived information, although the significance of the Diebold and Mariano (1995) p-values is generally weaker. Relative to the benchmark forecasts, RMSFE reductions can be up to 20 percent for these longer-horizon targets. In contrast, forecasts relating to the consumer confidence index, consumer price index, and equity market index, do not appear to be improved by the inclusion of textual predictors. While a lack of predictability of equity markets returns is generally unsurprising, one may have expected consumer sentiment or inflation to be predictable by Beige Book information. There is also some variation across different text representations. For example, while predictors derived from the VADER sentiment score predict the unemployment rate significantly better than the benchmark model at the 1 and 3-month horizons, the word2vec representation performs similar to the benchmark. At longer horizons, word2vec performs worse than the benchmark specification.

Another observation is that, at short horizons, text-augmented forecasts are almost universally at least as accurate as the benchmark predictions. The highest relative RMSFE is 2% higher, for example, in the case of the fasttext prediction for the 2-month ahead consumer price index. In contrast, there are several short-term targets that benefit significantly from the inclusion of text-derived predictors. As such, the risks and rewards of using text-augmented forecasts for shorter horizons appear asymmetric and in support of the use of text. The same cannot be said of longer-term forecasts, where performance of text-augmented predictions is more variable. This finding that the marginal benefit of augmenting forecasts with Beige Book information is more reliable at shorter horizons holds across subsamples, as is evident from Figures 55, 56, and 57 in Appendix B.3.

3.5 Quantifying the impact of temporal lookahead bias via text embeddings

A concern when using text representations, such as the ones introduced in Section 3.2, as predictor variables in a forecasting model is that unbiased evaluation of the forecast model's predictive performance can be challenging (Sarkar and Vafa, 2024; Ludwig et al., 2025; Hoberg and Manela, 2025). A language model trained on a corpus of documents available as of the year 2025 may learn to associate the concept of mortgage-backed securities with severe recessions while the same association may not be made if only documents up until the year 2005 were available during training. As a consequence, a simulated out-of-sample forecasting experiment using year-2025 embeddings as predictor variables may result in overestimating the predictive power of these text-derived variables. In other words, forecast evaluation results are likely to exhibit a downward bias of its loss function—making text-derived predictors appear more valuable than they would have been in genuine real-time forecasting. Sarkar and Vafa (2024) refer to this bias as the 'temporal lookahead bias' of pre-trained language models.

In this section, I investigate the extent to which temporal lookahead bias is present in the forecast exercise presented in Section 3.4. To do so, I perform a knowledge cutoff experiment by modifying the corpus of documents available during language model training.

3.5.1 Knowledge cutoff experiment

For pre-trained embeddings or sentiment lexica, the knowledge cutoff—typically defined as the point in time the last document used to generate embeddings came into being— is often not transparent. Even when there is a stated cutoff date, Ludwig et al. (2025) note that subsequent fine-tuning can result in the effective cutoff date being later than the stated one.

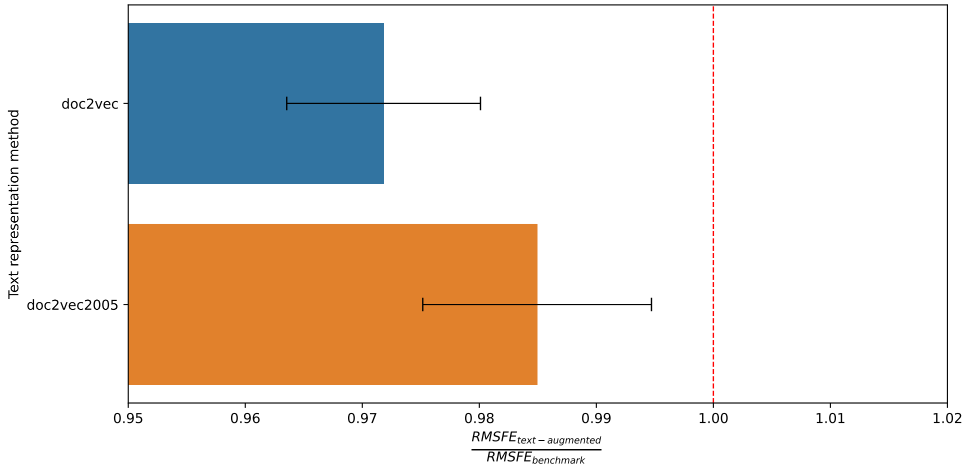
All but one of the methods listed in Table 9, are 'pre-trained' in the sense that the embeddings are a function of information external to the Beige Book corpus being represented. For example, longformer has been trained on the entirety of English Wikipedia. The textblob and VADER sentiment scores are also pre-trained in the sense that their sentiment lexica were constructed using information other than Beige Book reports. The one exception is the doc2vec method, which I implement by estimating the parameters of its language model using only the Beige Book corpus.

The implementation described in Section 3.2 and the forecast evaluation in Section 3.4 use embeddings obtained using the entire corpus of Beige Book documents between 1970 and 2024 to estimate the doc2vec model. In this section, I re-estimate the doc2vec language model on a reduced corpus that excludes all Beige Book reports published after the year 2005. That is, I generate a new vintage of embeddings that is not a function of any information available after 2005. I then repeat the out-of-sample forecasting experiment of Section 3.4 with this '2005 vintage' of doc2vec embeddings. It should be noted that the resulting comparison is subject to a joint hypothesis problem. In particular, the 2024 vintage and the 2005 vintage of doc2vec embeddings differ in two ways. Firstly, they have a different knowledge cutoffs. Secondly, the sample size available to estimate the 2005 embeddings is smaller than that available to estimate 2024 embeddings. The cutoff of 2005 is chosen as being before the great recession, whilst keeping the sample size available for language model training large. Nevertheless, any differences in performance could be due to the difference in sample size rather than the due to the different knowledge cutoffs. Under the assumption that a larger sample size used to estimate the doc2vec model would result in embeddings with better forecast performance, the difference between forecast performance of the 2005 embeddings and the 2025 embeddings can be considered an *upper bound* estimate of the temporal lookahead bias.

3.5.2 Estimated upper bound of temporal lookahead bias

Figure 44 shows the relative performance of the 2005 and 2024 vintages of doc2vec embeddings in a forecasting exercise analogous to that detailed in Section 3.4.

Figure 44: Relative performance of text-augmented forecasts using different vintages of doc2vec embeddings: overall



Notes: Relative RMSFE of text-augmented forecast relative to benchmark forecast. For each vintage of doc2vec embeddings, mean ratios are computed across the different forecast targets in Table 10 and across horizons. Error bars represent 95% confidence intervals for the mean ratio based on the standard error of the mean computed across all ratios. Parameter settings: $K = 16$, $UB = 8$,$\alpha = \lambda = \kappa = 0.999$.

As before, the relative performance displayed for each set of embeddings is an average across all forecast targets and forecast horizons. The error bars indicate the 95% confidence interval for the mean relative performance across these forecast targets and forecast horizons. As in Section 3.4, results were obtained using text regression parameters $K = 16$ and $UB = 8$, and forgetting factors $\alpha = \lambda = \kappa = 0.999$.

Both sets of embeddings appear to outperform the benchmark model, but there is a notable decrease in forecast accuracy when using embeddings that were estimated with a knowledge cutoff in 2005 compared to those with a knowledge cutoff in 2024. In particular, the embeddings estimated on the full set of Beige Book reports between 1970 and 2024 result in a RMSFE that is on average three percent lower than the benchmark model. In contrast, the embeddings obtained using the subset of reports published before 2006 result in forecasts that have an average RMSFE only 1.5 percent lower than that of the non-text benchmark. That is, the marginal benefit of including text-derived predictors is estimated to be 100% greater when including textual data published between 2005 and 2024 in the doc2vec model training dataset. However, this difference in forecast performance cannot neccessarily be interpreted as an unbiased estimate of the temporal lookahead bias—due to the joint hypothesis problem discussed above. While the two sets of embeddings have different knowledge cutoffs, the sample size available to estimate the 2005 embeddings is smaller than that available to estimate 2024 embeddings. This sample size difference may affect forecast performance independently of any knowledge cutoff effect. To the extent that there is signal in the textual data, one might expect this signal to be better recovered by embeddings that were estimated using a larger sample. If one makes this assumption, both the knowledge cutoff and the sample size effect would give the 2024 vintage of doc2vec embeddings with an advantage over the 2005 vintage. The 1.5 percent difference in relative RMSFE between the specification using 2005 embeddings and that using the 2024 embeddings could then be considered an *upper bound* estimate of the temporal lookahead bias.

It should be noted that this upper bound estimate is specific to the set of forecast targets and horizons, the included non-textual predictors, as well as the use of doc2vec as a method for obtaining embeddings. Similar knowledge cutoff experiments for more sophisticated pre-trained large language models such as longformer or BERT-family models may lead to different estimates, but would require significant resources to perform. Due to the computational intensity of repeatedly pretraining these large models on vast corpora of text, such an exercise is beyond the scope of this study and left for future research.

3.6 Conclusion

In this study, I explored the use of language models to extract predictive information about US macroeconomic aggregates from the corpus of the Federal Reserve System's Beige Book reports. To do so, I compared a number of methods for representing raw textual documents in a fixed-length numerical format that enables the use of text-derived variables as predictors in quantitative statistical models. While simple sentiment scores typically result in a relatively small number of predictors that could be added to standard models, embeddings obtained using small or large language models can result in thousands of text-derived variables, even before constructing engineered predictors such as lags. After feature engineering, the number of text-derived predictors obtained using a large language model-based embedding method exceeds 100,000 predictors, for instance. This is a large number of variables relative to the less than 1,000 time series observations available for US macroeconomic aggregates. I therefore propose a method that can handle predictor sets of this size, enables efficient forecast evaluation, and can be specified to handle time-varying parameters. Specifically, I model the distribution of prediction targets using a novel switching Gaussian state space model with compressed predictors. Using this method, I estimate the incremental value of Beige Book text for macroeconomic forecasting—that is, the improvement in forecast accuracy that obtains when including text-derived variables in forecast models—to be moderate overall. I find this gain in accuracy to be distributed heterogeneously across forecast targets and horizons. Interestingly, there is no clear relationship between the complexity of the method used to represent text and the resulting forecast performance. Having quantified the strength of the predictive information different methodologies are able to recover from raw text, I also explore a pitfall of using language models for economic forecasting: temporal lookahead bias during forecast evaluation. In this specific empirical context, my findings suggest that up to half of the estimated forecast accuracy gain due to the addition of predictors generated using language models is illusory. Avoiding this temporal lookahead bias whilst using state-of-the-art models can be costly, as doing so requires retraining large language models. For the largest language models, training costs are reported to exceed \$40m (Cottier et al., 2025), making complete avoidance of temporal lookahead bias practically infeasible. Given this constraint, further work to at least estimate the size of temporal lookahead biases across different economic forecasting settings could be worthwhile. Finally, this paper has shown that there appears to be some information about the economic outlook within Beige Book text that is not fully captured by traditional, non-textual predictor variables. Further research could investigate further optimisation of how textual predictors are incorporated into forecasts, potentially resulting in forecasts being improved more uniformly across targets and horizons.

B.1 Simulations

To study the performance of the proposed method, I generate multiple synthetic time series of prediction targets and predictors. After discarding 100 burn-in periods, I generate $T = 500$ time series observations to mimic a typical sample size encountered in forecasting applications. The simulated forecasting exercise uses expanding window estimation with an initial estimation window of 250.

B.1.1 Data generating processes

I first draw D predictors from a Poisson distribution, with the aim of obtaining time series of counts that resemble n-gram counts of real textual data

$$x_{d,t} \sim^{iid} Poisson(\mu_{text}).$$

The Poisson's mean parameter μ_{text} can be interpreted as modifying the text density of the predictors. When documents in a corpus correspond to short headlines, for instance, very few of the n-gram counts will be non-zero. In contrast, when documents correspond to full articles or even books, words will tend to appear more than once.

Having generated the predictors, I now generate time-varying coefficients for each pseudo n-gram from the following autoregressive process with parameter ρ_β

$$\beta_{d,t} = \mu_{\beta,d} + \rho_\beta(\beta_{d,t-1} - \mu_{\beta,d}) + \eta_{d,t}$$

where the long-run mean for each coefficient is drawn from a continuous uniform distribution

$$\mu_{\beta,d} \sim^{iid} Uniform(-1/D, 1/D)$$

and the innovation is drawn from a normal distribution with variance σ_η^2 (scaled by the number of predictors)

$$\eta_{d,t} \sim^{iid} Normal(0, (\frac{\sigma_\eta}{D})^2).$$

To simulate heteroskedasticity often observed in real-world economic time series, I then generate a time series of stochastic volatility by taking draws from a log-normal process with long-run mean μ_{σ^2} and autoregressive parameter ρ_{σ^2} as follows

$$log(\sigma_t^2) = \mu_{\sigma^2} + \rho_{\sigma^2}(\sigma_t^2 - \mu_{\sigma^2}) + \xi_t$$

where

$$\xi_t \sim^{iid} N(0, \frac{1}{T}).$$

I then draw a time series of raw error terms as follows

$$\tilde{\epsilon_t} \sim N(0, \sigma_t^2).$$

and scale each element in this raw time series of error terms $\tilde{\epsilon_t}$ in such a way as to achieve a desired signal-to-noise ratio[13]

$$\epsilon_t = \frac{1}{SNR_{target}} \frac{Var(\{x_t\beta_t\})}{Var(\{\tilde{\epsilon}_t\})}\tilde{\epsilon_t}.$$

This standardisation step allows one to study how predictive performance depends on signal strength. Having standardised the time series of error terms in this way, I finally obtain the prediction target time series as follows

$$y_t = x_t\beta_t + \epsilon_t.$$

B.1.2 Process parameterisations and hyperparameters

Using the class of data introduced above, I now examine how the choice of hyperparameters λ, α, and K, depends on process parameters. Overall, I find that the hyperparameter values $\lambda = \alpha = 0.99$ and $\kappa = 0.97$ suggested by Raftery et al. (2010) and Koop and Korobilis (2012), as well as $K < 100$ suggested by Guhaniyogi and Dunson (2015) work well for different parameterisations of the class of data-generating processes considered. Unless stated otherwise, these values are used as defaults. To avoid running out of working

[13] The signal-to-noise ratio is here taken to be the ratio of the variance of the regression component to the variance of the noise component of the data generating process $SNR = \frac{Var(\{x_t\beta_t\})}{Var(\{\epsilon_t\})}$.

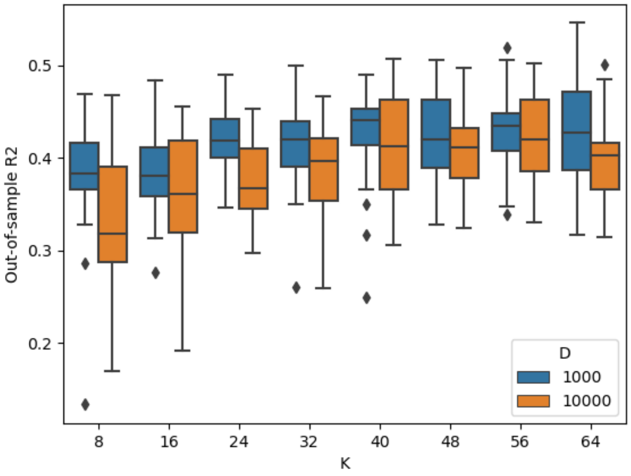
memory, I set $UB = 100$.

Figure 45 reveals the variation in predictive performance across 25 sampled processes. This variation is due to two sources of randomness. The first source is the sampling variation of the data generating process. The second source is the variation due to the estimation procedure involving the sampling of random projection matrices Φ. The figure shows that the variation of predictive performance of my method stabilises as I increase the number of projection matrix draws, suggesting that the remaining variance is due to differences in the sampled processes. Overall, the predictive problem is harder as the number of predictors increases, as shown by the orange boxplots tending to be shifted downwards compared to the blue ones. Nevertheless, even with 10,000 predictors the proposed approach is able to recover most of the signal in the predictors.

Figure 46 shows the predictive performance of the proposed approach for different values of the variance of the time-varying coefficient innovations σ_η and forgetting factor hyperparameters $\lambda = \alpha$. It can be seen how more variation in coefficients makes the predictive problem more challenging. Across different values of σ_η, the higher choices for the forgetting factors λ, α tend to perform better. That is, even when the data-generating process has highly-variable coefficients, gentle forgetting appears preferable to strong forgetting. For mild time-variation in coefficients moderate forgetting ($\lambda = \alpha = 0.96$) appears to improve predictive performance somewhat relative to higher forgetting factors.

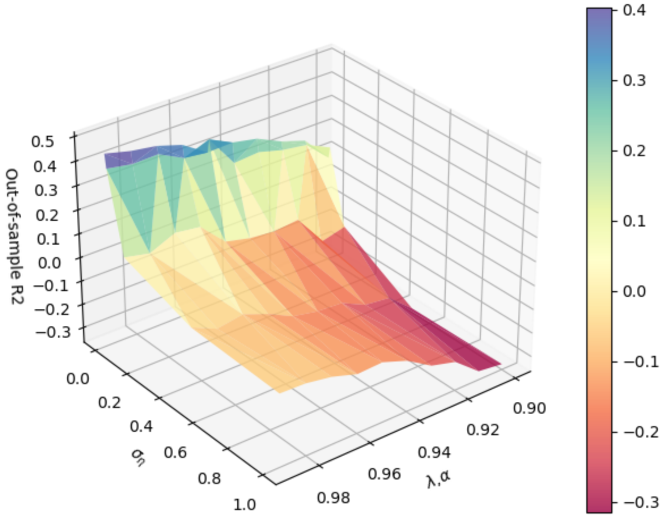
Finally, Figure 47 shows that for a wide range of the text densities, a higher compression counts achieve better predictive performance. This is particularly so when the text density is high, and the number of compressions should be chosen to exceed 20 in such cases.

Figure 45: Predictor count and compression count



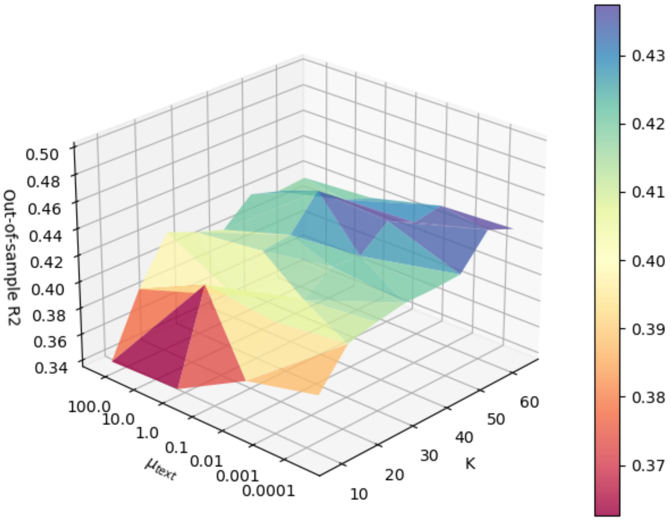
Notes: For each value of predictor count D and hyperparameter choice K (compression count), this chart shows the out-of-sample R-squared attained for 25 sampled processes. The other process parameters are fixed at $\mu_{text} = 1$ $SNR_{target} = 1$, $\sigma_\eta = 0$, $\rho_\beta = 0$, $\rho_{\sigma^2} = 0$, $exp(\mu_{\sigma^2}) = 1$. *The other hyperparameters are fixed at* $\lambda = 0.99$, $\alpha = 0.99$, $\kappa = 0.97$, $UB = 100$.

Figure 46: Time-variation and forgetting factors



Notes: For each value of coefficient innovation variance σ_η and hyperparameter choices λ, α (forgetting factors), this chart shows the out-of-sample R-squared attained (median of 25 sampled processes). The other process parameters are fixed at $D = 1,000$, $SNR_{target} = 1$, $\mu_{text} = 1$, $\rho_\beta = 0.99$, $\rho_{\sigma^2} = 0.99$, $exp(\mu_{\sigma^2}) = 1$. The other hyperparameters are fixed at $K = 64$, $\kappa = 0.97$, $UB = 100$.

Figure 47: Text density and compression count



Notes: For each value of text density μ_{text} and hyperparameter choice K (compression count), this chart shows the out-of-sample R-squared attained (median of 25 sampled processes). The other process parameters are fixed at $D = 1,000$, $SNR_{target} = 1$, $\sigma_\eta = 0$, $\rho_\beta = 0$, $\rho_{\sigma^2} = 0$, $exp(\mu_{\sigma^2}) = 1$. The other hyperparameters are fixed at $\lambda = 0.99$, $\alpha = 0.99$, $\kappa = 0.97$, $UB = 100$.

Table 11: Selected text regression approaches

Model	Description	Literature
TR	Bayesian compressed time series text regression	This paper
EN	Elastic net regression	Pedregosa et al. (2011); Algaba et al. (2020)
RIDGE	Ridge regression	Pedregosa et al. (2011); Algaba et al. (2020)
SVR	Support vector regression	Pedregosa et al. (2011); Manela and Moreira (2017)
DMR-LM	Distributed multinomial regression + OLS	Taddy (2013, 2015); Kelly et al. (2021)
DMR-EN	Distributed multinomial regression + elastic net	Taddy (2013, 2015); Kelly et al. (2021)
HDMR-LM	Hurdle regression + OLS	Kelly et al. (2021)
HDMR-EN	Hurdle regression + elastic net	Kelly et al. (2021)

B.1.3 Comparison with alternative methods for text regression

The proposed approach is now compared to alternative text regression methods, as displayed in Table 11. I consider the penalised least squares methods ridge regression and elastic net, as suggested for text regression problems by Gentzkow et al. (2019). I do not consider a lasso penalty since the data generating process described above is not sparse. I also run a support vector regression as used by Manela and Moreira (2017). For the inverse regression approaches of Taddy (2015) and Kelly et al. (2021) I implement two versions each: one with ordinary least squares and one with elastic net least squares as forward regression step.

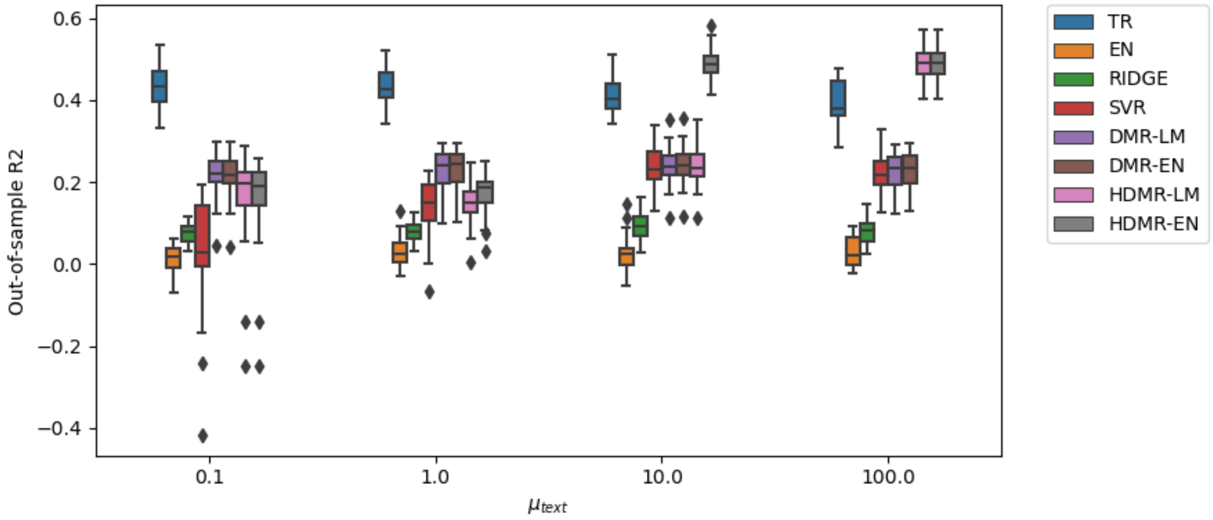
The penalised least squares methods and support vector regression are implemented in Python using scikit-learn (Pedregosa et al., 2011).[14] For the inverse regression models I use the Julia package HurdleDMR (Kelly et al., 2021).

Text density

I first simulate data for different values of the text density μ_{text} of the predictors. The results of applying each text regression method to 25 simulated

[14] Grid search cross-validation is used once on the entire sample to determine the hyperparameter values for these methods. This gives a slight advantage to these methods, as data from the 'future' is used, but re-optimising the hyperparameter at each expanding window is computationally prohibitive.

Figure 48: Text density



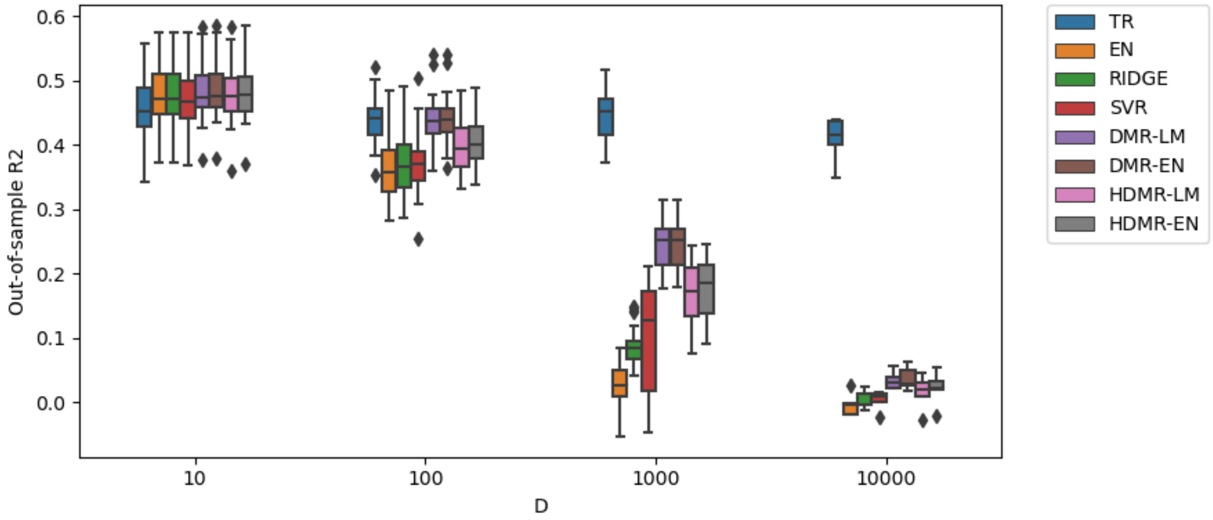
Notes: For each value of text density μ_{text}, this chart shows the out-of-sample R-squared attained for 25 sampled processes. The other process parameters are fixed at $D = 1,000$, $SNR_{target} = 1$, $\sigma_\eta = 0$, $\rho_\beta = 0$, $\rho_{\sigma^2} = 0$, $exp(\mu_{\sigma^2}) = 1$.

data-generating processes are shown in Figure 48. The proposed approach is competitive across text density values, attaining an out-of-sample R-squared close to the ground truth of 0.5 implied by the signal-to-noise ratio of one. For low values of μ_{text}, it outperforms all of the alternatives considered, while the Hurdle method of Kelly et al. (2021) attains higher out-of-sample R-squared values for higher densities.

Predictor count

Figure 49 shows the results of a similar exercise for the number of predictors D, keeping the text density fixed at $\mu_{text} = 1$. Again, the proposed approach is among the most effective methods across different parameterisations of the data-generating process. There is a particularly pronounced advantage for high predictor counts exceeding the sample size ($T = 500$), where alternative

Figure 49: Predictor count



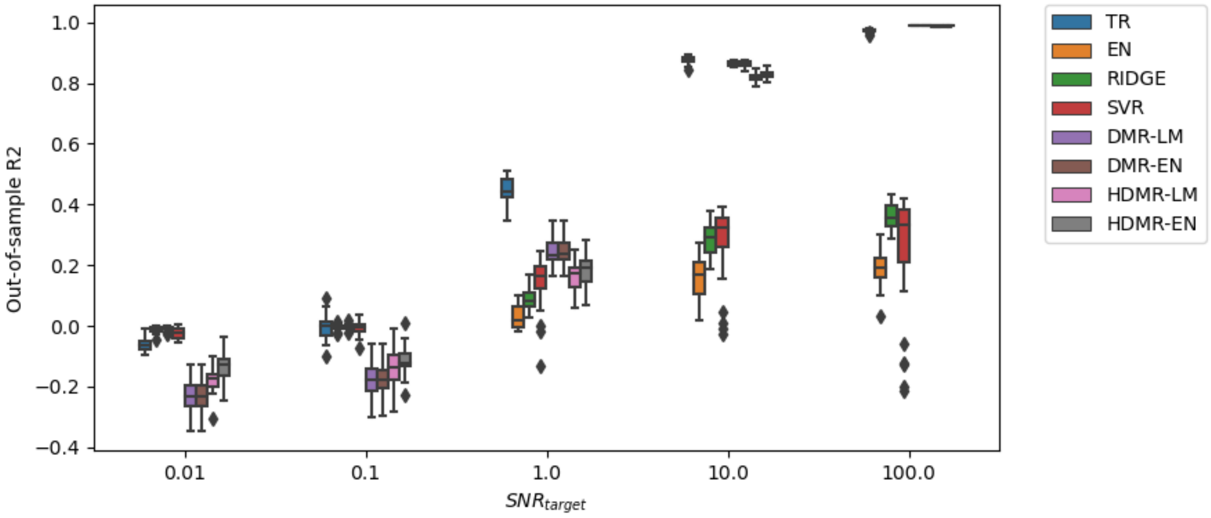
Notes: For each value of predictor count D, this chart shows the out-of-sample R-squared attained for 25 sampled processes. The other process parameters are fixed at $\mu_{text} = 1$, $SNR_{target} = 1$, $\sigma_\eta = 0$, $\rho_\beta = 0$, $\rho_{\sigma^2} = 0$, $exp(\mu_{\sigma^2}) = 1$.

methods appear ill-suited.

Signal-to-noise ratio

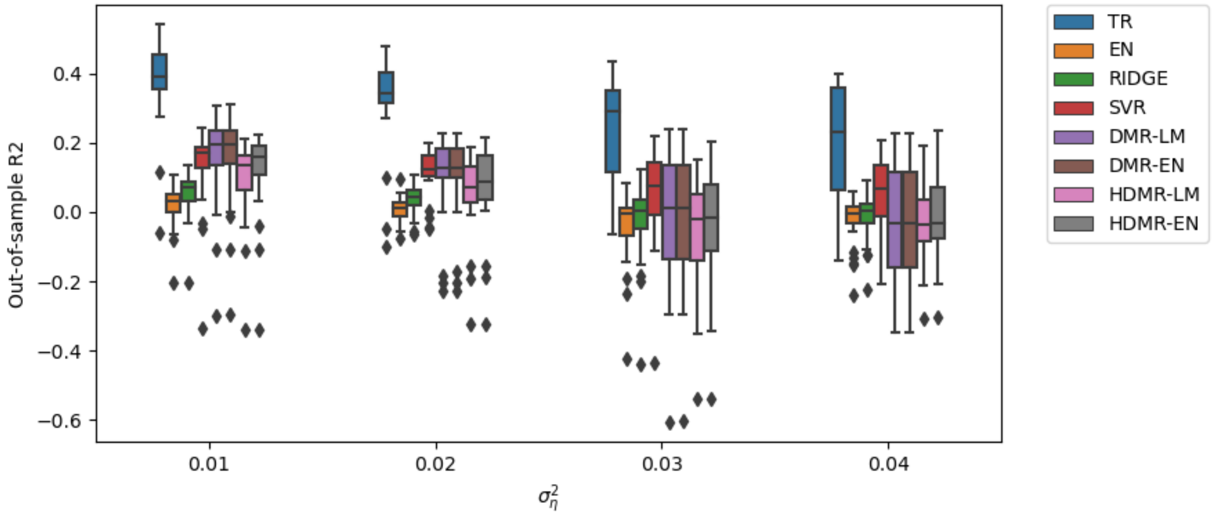
Keeping the text density and predictor count fixed at $\mu_{text} = 1$ and $D = 1,000$, respectively, I now examine results for different values of the signal-to-noise ratio SNR_{target} of the data-generating process. The increasing gradient form left to right in Figure 50 reflects the mechanical relationship between signal-to-noise ratio and the maximum possible 'ground truth' R-squared. The proposed method works particularly well when signal and noise strength are balanced, and is competitive with inverse regression approaches when signal strength is very high.

Figure 50: Signal-to-noise ratio



Notes: For each value of signal-to-noise ratio SNR_{target}, this chart shows the out-of-sample R-squared attained for 25 sampled processes. The other process parameters are fixed at $D = 1,000$, $\mu_{text} = 1$, $\sigma_\eta = 0$, $\rho_\beta = 0$, $\rho_{\sigma^2} = 0$, $exp(\mu_{\sigma^2}) = 1$.

Figure 51: Coefficient innovation variance

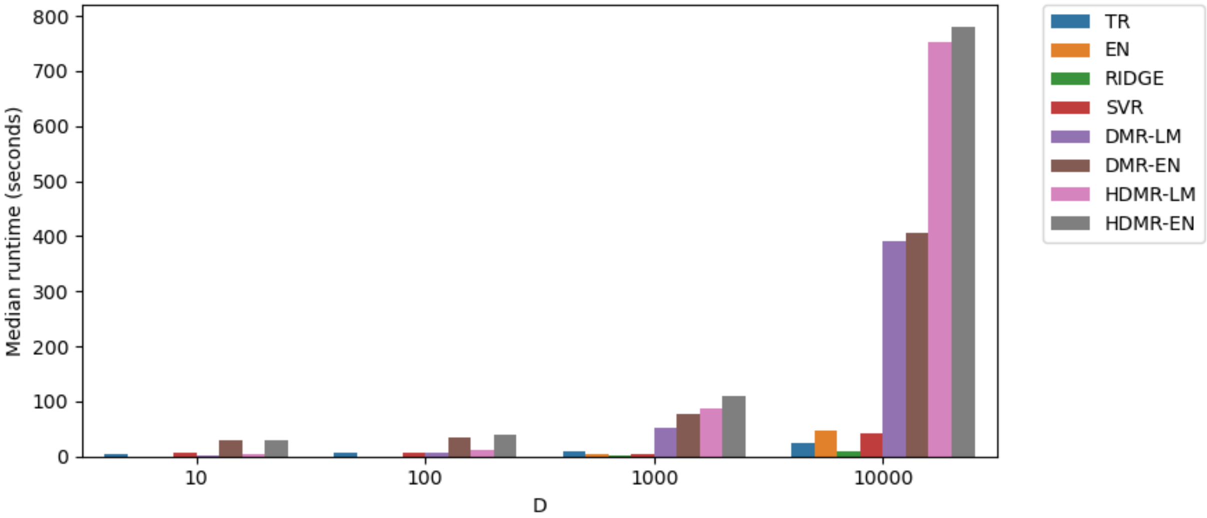


Notes: For each value of coefficient innovation variance σ_η, this chart shows the out-of-sample R-squared attained for 25 sampled processes. The other process parameters are fixed at $D = 1,000$, $\mu_{text} = 1$, $SNR_{target} = 1$, $\rho_\beta = 0.99$, $\rho_{\sigma^2} = 0$, $exp(\mu_{\sigma^2}) = 1$.

Coefficient innovation variance

Echoing the results shown in Figure 46, Figure 51 demonstrates that time-varying parameters make estimation materially more challenging. Nevertheless, the proposed approach outperforms the other methods across the range of coefficient innovation variance parameters considered. For $\sigma_{\eta^2} = 0.04$, for instance, most other methods attain a median out-of-sample R-squared around 0 compared to a median of 0.2 for the compression approach. This advantage is unsurprising, given that my modelling differs from the alternatives in that it allows for parameter evolution.

Figure 52: Runtime



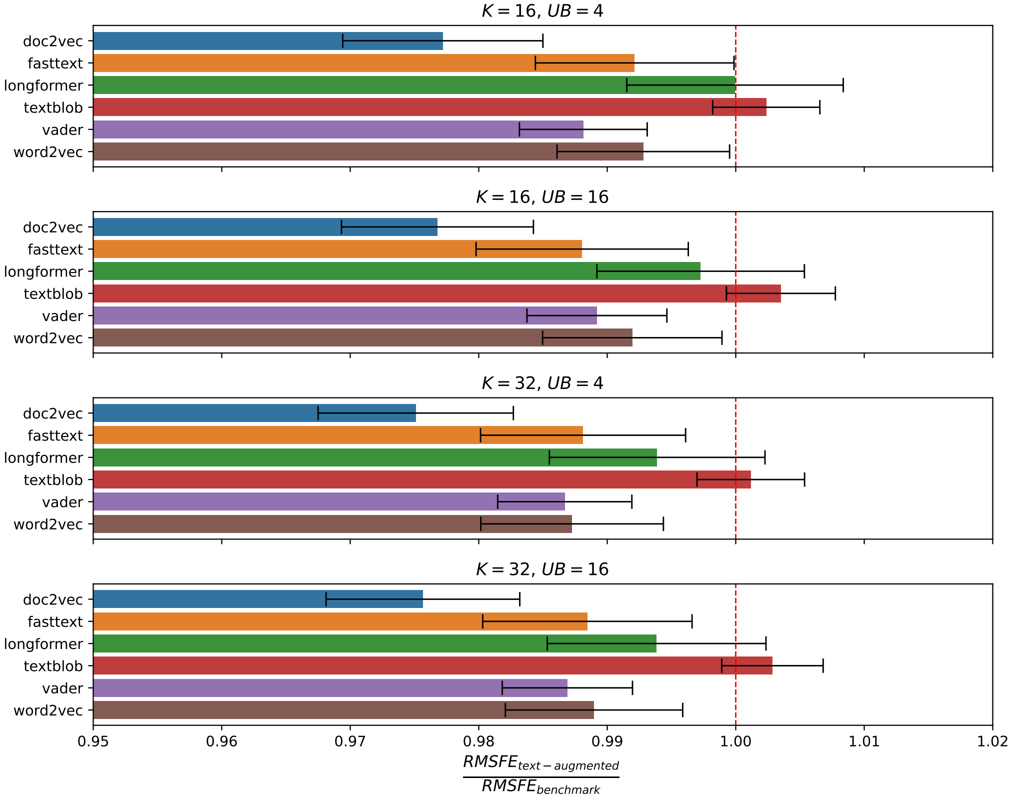
Notes: For each value of predictor count D, this chart shows the number of seconds required for an expanding-window out-of-sample forecasting exercise with $T = 500$ and initial estimation window of size 250.

Runtime

The amount of time required to estimate and evaluate each method for 500 time periods and 250 evaluation periods is shown in Figure 52, which displays the median number of seconds required for 25 sampled data-generating processes. As the proposed method can be estimated in a recursive fashion, evaluation of time series predictive performance comes at virtually no computational cost. All other methods need to be re-estimated 250 times on expanding windows to enable time series evaluation. The results in Figure 52 reflect this difference.

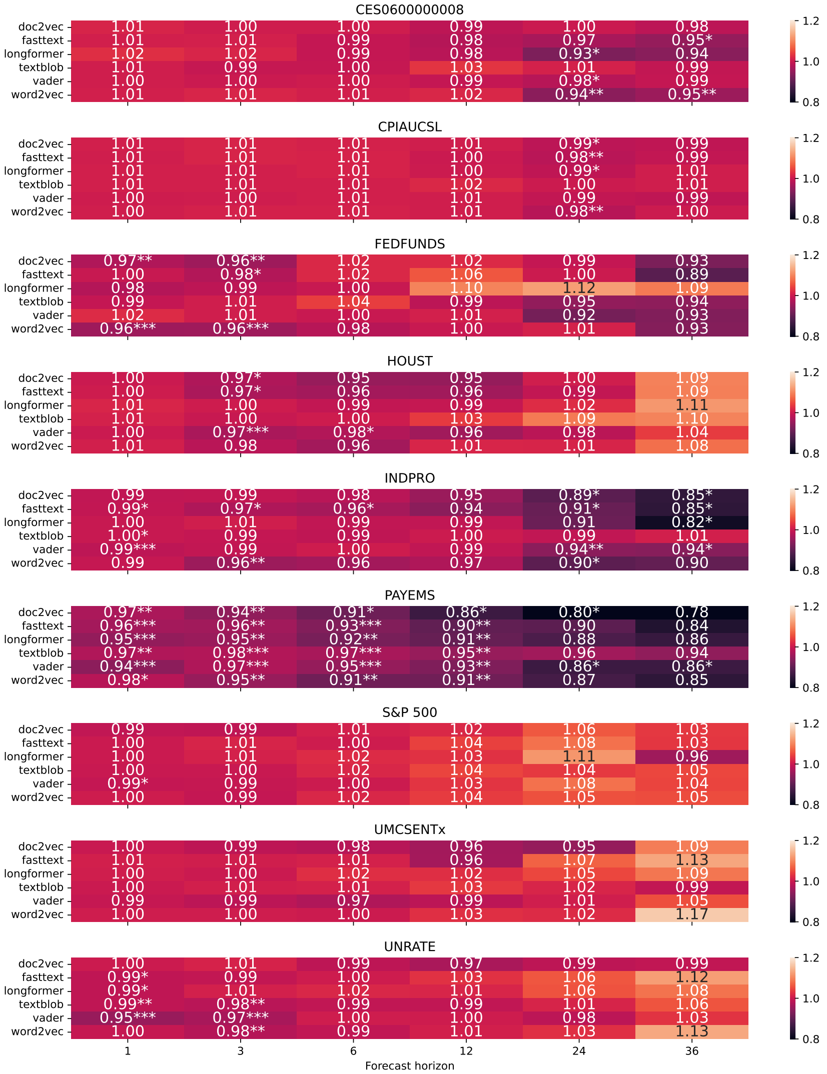
B.2 Sensitivity analyses

Figure 53: Sensitivity of overall results to choice of K and UB



Notes: Relative RMSFE of text-augmented forecast relative to benchmark forecast. For each text representation method and parameter setting, mean ratios are computed across the different forecast targets in Table 10 and across horizons. Error bars represent 95% confidence intervals for the mean ratio based on the standard error of the mean computed across all ratios. Parameter settings: $\alpha = \lambda = \kappa = 1$.

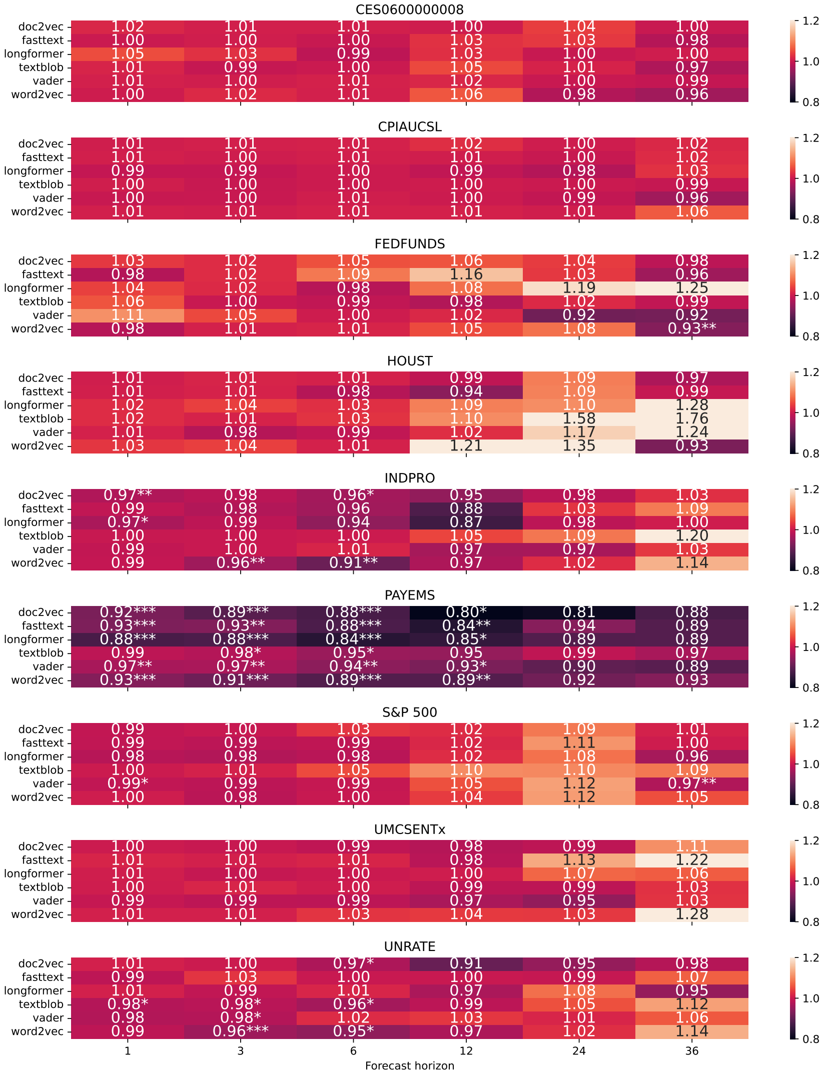
Figure 54: Sensitivity of detailed results to choice of α, λ, and κ



Notes: Relative RMSFE of text-augmented forecast relative to benchmark forecast. Asterisks indicate Diebold and Mariano (1995) p-values, with a single asterisk indicating $p <= 0.1$, two asterisks indicating $p <= 0.05$ and three asterisks indicating $p < 0.01$. Parameter settings: $K = 16$, $UB = 8$, $\alpha = \lambda = \kappa = 1$.

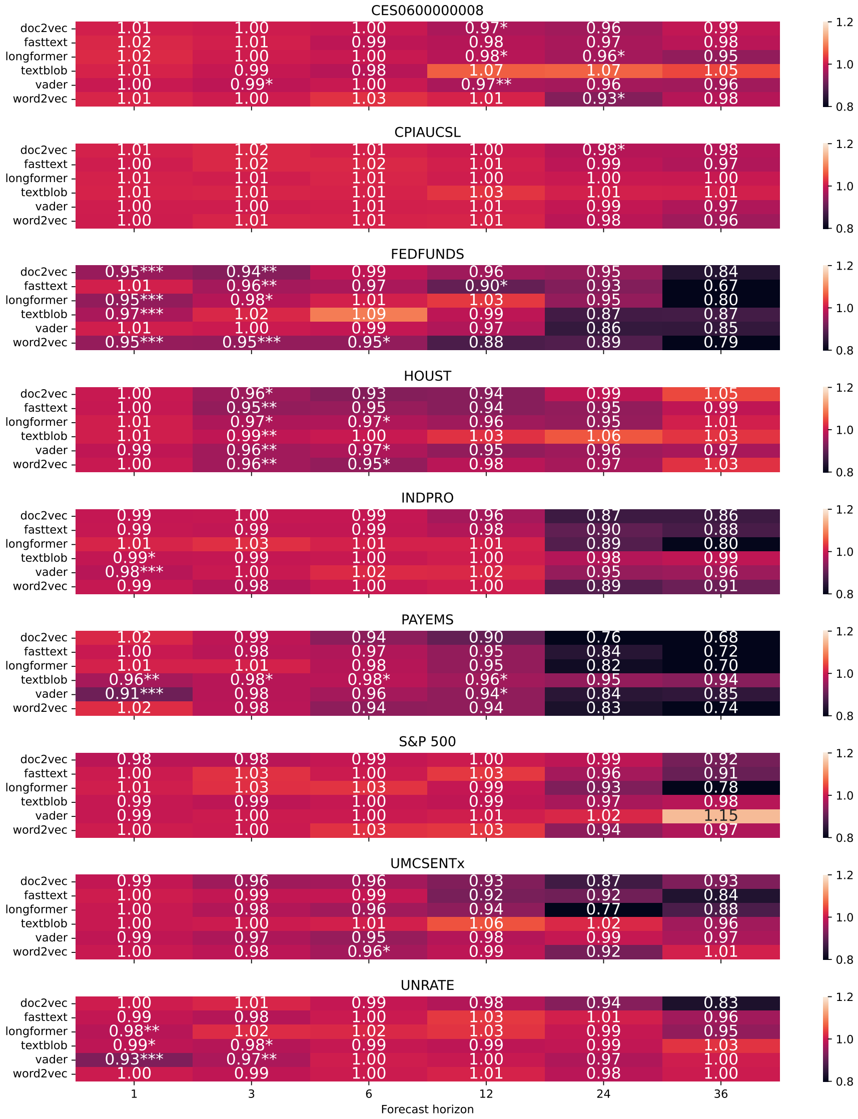
B.3 Subsample analyses

Figure 55: Subsample analysis of detailed results: 1999-12 to 2006-12



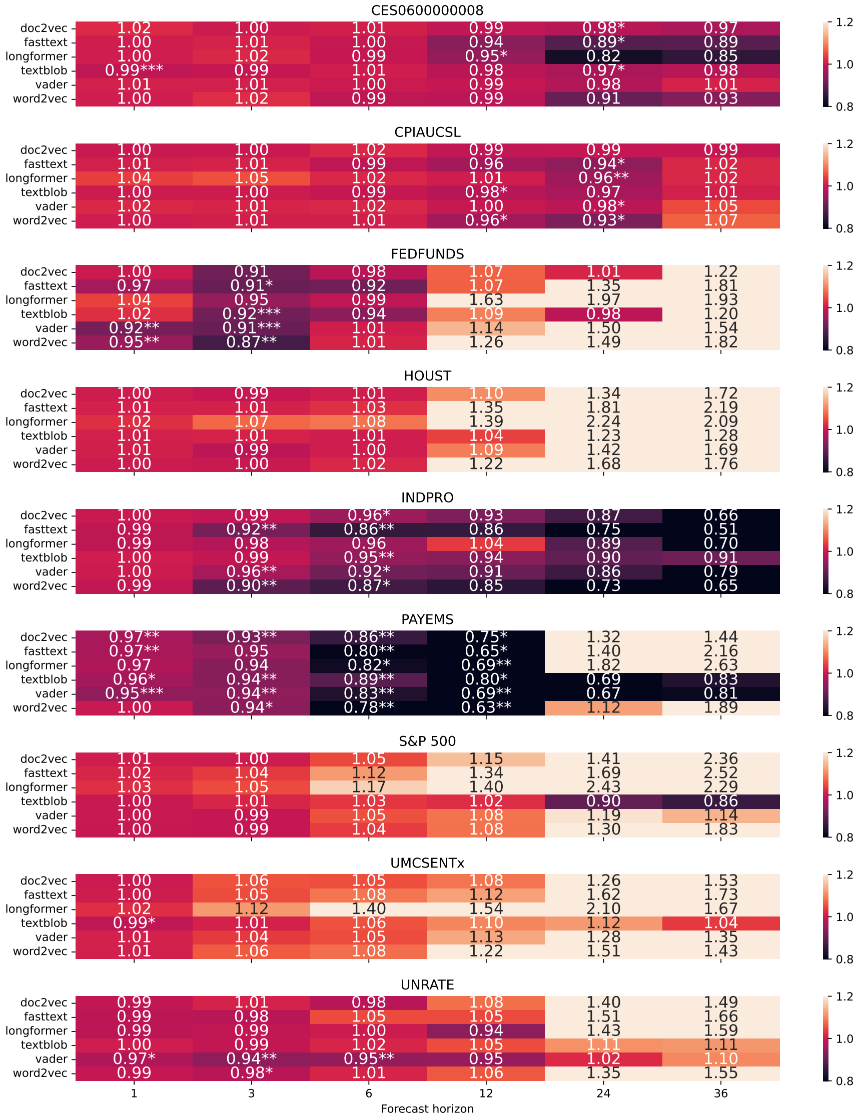
Notes: Relative RMSFE of text-augmented forecast relative to benchmark forecast. Asterisks indicate Diebold and Mariano (1995) p-values, with a single asterisk indicating $p <= 0.1$, two asterisks indicating $p <= 0.05$ and three asterisks indicating $p < 0.01$. Parameter settings: $K = 16$, $UB = 8$, $\alpha = \lambda = \kappa = 0.999$.

Figure 56: Subsample analysis of detailed results: 2007-01 to 2014-01



Notes: Relative RMSFE of text-augmented forecast relative to benchmark forecast. Asterisks indicate Diebold and Mariano (1995) p-values, with a single asterisk indicating $p <= 0.1$, two asterisks indicating $p <= 0.05$ and three asterisks indicating $p < 0.01$. Parameter settings: $K = 16$, $UB = 8$, $\alpha = \lambda = \kappa = 0.999$.

Figure 57: Subsample analysis of detailed results: 2014-02 to 2019-11



CES0600000008

	1	3	6	12	24	36
doc2vec	1.02	1.00	1.01	0.99	0.98*	0.97
fasttext	1.00	1.01	1.00	0.94	0.89*	0.89
longformer	1.00	1.02	0.99	0.95*	0.82	0.85
textblob	0.99***	0.99	1.01	0.98	0.97*	0.98
vader	1.01	1.01	1.00	0.99	0.98	1.01
word2vec	1.00	1.02	0.99	0.99	0.91	0.93

CPIAUCSL

	1	3	6	12	24	36
doc2vec	1.00	1.00	1.02	0.99	0.99	0.99
fasttext	1.01	1.01	0.99	0.96	0.94*	1.02
longformer	1.04	1.05	1.02	1.01	0.96**	1.02
textblob	1.00	1.00	0.99	0.98*	0.97	1.01
vader	1.02	1.01	1.02	1.00	0.98*	1.05
word2vec	1.00	1.01	1.01	0.96*	0.93*	1.07

FEDFUNDS

	1	3	6	12	24	36
doc2vec	1.00	0.91	0.98	1.07	1.01	1.22
fasttext	0.97	0.91*	0.92	1.07	1.35	1.81
longformer	1.04	0.95	0.99	1.63	1.97	1.93
textblob	1.02	0.92***	0.94	1.09	0.98	1.20
vader	0.92**	0.91***	1.01	1.14	1.50	1.54
word2vec	0.95**	0.87**	1.01	1.26	1.49	1.82

HOUST

	1	3	6	12	24	36
doc2vec	1.00	0.99	1.01	1.10	1.34	1.72
fasttext	1.01	1.01	1.03	1.35	1.81	2.19
longformer	1.02	1.07	1.08	1.39	2.24	2.09
textblob	1.01	1.01	1.01	1.04	1.23	1.28
vader	1.01	0.99	1.00	1.09	1.42	1.69
word2vec	1.00	1.00	1.02	1.22	1.68	1.76

INDPRO

	1	3	6	12	24	36
doc2vec	1.00	0.99	0.96*	0.93	0.87	0.66
fasttext	0.99	0.92**	0.86**	0.86	0.75	0.51
longformer	0.99	0.98	0.96	1.04	0.89	0.70
textblob	1.00	0.99	0.95**	0.94	0.90	0.91
vader	1.00	0.96**	0.92*	0.91	0.86	0.79
word2vec	0.99	0.90**	0.87*	0.85	0.73	0.65

PAYEMS

	1	3	6	12	24	36
doc2vec	0.97**	0.93**	0.86**	0.75*	1.32	1.44
fasttext	0.97**	0.95	0.80**	0.65*	1.40	2.16
longformer	0.97	0.94	0.82*	0.69**	1.82	2.63
textblob	0.96*	0.94**	0.89**	0.80*	0.69	0.83
vader	0.95***	0.94**	0.83**	0.69**	0.67	0.81
word2vec	1.00	0.94*	0.78**	0.63**	1.12	1.89

S&P 500

	1	3	6	12	24	36
doc2vec	1.01	1.00	1.05	1.15	1.41	2.36
fasttext	1.02	1.04	1.12	1.34	1.69	2.52
longformer	1.03	1.05	1.17	1.40	2.43	2.29
textblob	1.00	1.01	1.03	1.02	0.90	0.86
vader	1.00	0.99	1.05	1.08	1.19	1.14
word2vec	1.00	0.99	1.04	1.08	1.30	1.83

UMCSENTx

	1	3	6	12	24	36
doc2vec	1.00	1.06	1.05	1.08	1.26	1.53
fasttext	1.00	1.05	1.08	1.12	1.62	1.73
longformer	1.02	1.12	1.40	1.54	2.10	1.67
textblob	0.99*	1.01	1.06	1.10	1.12	1.04
vader	1.01	1.04	1.05	1.13	1.28	1.35
word2vec	1.01	1.06	1.08	1.22	1.51	1.43

UNRATE

	1	3	6	12	24	36
doc2vec	0.99	1.01	0.98	1.08	1.40	1.49
fasttext	0.99	0.98	1.05	1.05	1.51	1.66
longformer	0.99	0.99	1.00	0.94	1.43	1.59
textblob	1.00	0.99	1.02	1.05	1.11	1.11
vader	0.97*	0.94**	0.95**	0.95	1.02	1.10
word2vec	0.99	0.98*	1.01	1.06	1.35	1.55

Notes: Relative RMSFE of text-augmented forecast relative to benchmark forecast. Asterisks indicate Diebold and Mariano (1995) p-values, with a single asterisk indicating $p <= 0.1$, two asterisks indicating $p <= 0.05$ and three asterisks indicating $p < 0.01$. Parameter settings: $K = 16$, $UB = 8$, $\alpha = \lambda = \kappa = 0.999$.

Chapter 4

Microfounded investor feature selection for efficient financial language model inference

This paper investigates how a data-driven investor can extract parsimonious sets of predictive features from high-dimensional data, a challenge that gained in importance recently with the advent of resource-intensive financial language modelling. I propose a general feature selection framework for investors that is rooted in Cabrales et al. (2013)'s result that Blackwell's partial ordering of information structures (Blackwell, 1953) can be completed by considering a finance-specific set of decision problems and utility functions. Through simulations and an empirical illustration, I find that there are both opportunities and limitations when operationalising this theory-based approach in practice. One particular area of promise I identify is that feature selection methods, such as the one proposed in this study, could help to reduce the resource cost of predicting risk premia using large language models.

4.1 Introduction

In this paper, I investigate a practical problem faced by any data-driven investor: the selection of features that carry information about future asset risk premia. An investor trying to make capital allocation decisions using data faces a challenging task. Whether the goal is to allocate across or within asset classes, a vast number of predictive features could be relevant to generating trading signals. The St. Louis Fed's freely-accessible FRED database[1] alone contains 816,000 economic time series—including around 1,000 daily, 210,000 monthly, and 490,000 annual variables. Representations of textual data—which have been shown to have predictive power for future asset returns (Antweiler and Frank, 2004; Ke et al., 2019)—can give rise to even higher counts of potentially relevant variables.

Parsimonious modelling of the statistical relationships between all these variables and the risk premia of a large number of financial assets requires some way of identifying the most important predictors for each asset. As Gu et al. (2020) show, it is often non-linear interactions of individual features that account for a significant share of predictability of asset risk premiums. However, an exhaustive search of the non-linear model space for all possible combinations of predictive features is computationally impractical. It can therefore be useful to filter the feature set prior to the training and evaluation of language models used for empirical asset pricing. Many model-agnostic feature selection methodologies have been proposed in the literature. While investors could empirically evaluate which methodology performs best at predicting a particular risk premium, practitioners may find it more cost-effective to choose a feature selection technique based on heuristics.

This paper explores the idea of using insights from microeconomic theory to avoid having to rely on heuristics to select one of the many feature selection procedures available. Establishing which feature selection methodology has

[1] https://fred.stlouisfed.org/

a theoretical claim to optimality could be useful to practitioners, because it reduces the need for repeated, exhaustive evaluations of the performance of different feature selection methodologies.

While the ‘investor feature selection’ challenge is general and longstanding, it has recently gained new relevance through the emergence of large financial language models (Wu et al., 2023; Huang et al., 2023). Survey evidence from the UK suggests that large language models already account for around a fifth of algorithmic decision-making within financial services generally, with adoption expected to increase over the next three years (Gharbawi et al., 2024). Similarly to the ‘high-frequency trading arms race’ and associated investments in microwave communication infrastructure (Aquilina et al., 2021), it therefore seems predictable that significant resources will be deployed by market participants to win the ‘natural language inference race’. Auguring the scale of resources likely to be involved, a single trading firm is reported to be investing in excess of €1bn to build a deep learning data centre in Finland that will consume 22.5 megawatts—equivalent to the energy consumption of thousands of households.[2] Even marginal improvements in efficiency during the training, fine-tuning, and inference procedures of large financial language models could therefore be economically and environmentally significant. It seems reasonable to hypothesise that large models that are architected and trained with the broad objective of predicting natural language text could, through a reduction in dataset size and model parameter count, be streamlined for the specific purpose of predicting risk premia.

A key contribution of this paper is to show how the general investor feature selection problem can be guided by microeconomic theory. I also explore different ways this microfounded approach can be operationalised in practice, comparing the performance of both frequentist and Bayesian estimation approaches using simulations. A third contribution is to articulate the different

[2] https://www.marketsmedia.com/xtx-markets-to-invest-over-e1bn-in-data-centre-in-finland/

ways in which feature selection can be used to reduce the resource cost of large financial language models, as well as illustrate the proposed approach in an empirical exercise with real data.

I find that in settings relevant to investors, ranking features using the information-theoretic quantity mutual information has a strong theoretical justification. The resulting order of features can be used by the investor to select the highest-ranking predictors to use for predictive modelling of risk premia. Examining different estimation approaches, I find that while the choice of mutual information estimator is subject to the bias-variance trade-off, different estimators return similar feature rankings in simulations. In the context of financial large language modelling, ranking tokens by their mutual information with the investment returns of interest to decide which tokens should be included in the model's vocabulary could result in both computational and accuracy benefits. This is because a smaller token vocabulary could lead to computational efficiency gains throughout the model lifecycle whilst removing noise. An illustrative exercise in which equity returns are predicted using news headlines provides suggestive support for the hypothesis that model size reductions need not result in the loss of predictive performance.

Related literature

In addition to Blackwell's foundational work on rankings of information structures (Blackwell, 1953), this paper relates to the literatures on the value of information for investors (e.g. Cabrales et al. (2013); Kadan and Manela (2018); Frankel and Kamenica (2019)), asset pricing using machine learning (e.g. Ke et al. (2019); Chinco et al. (2019); Gu et al. (2020); Giglio et al. (2022); Gu et al. (2021)), feature selection (e.g. Zaffalon and Hutter (2002); Peng et al. (2005); Dhal and Azad (2022)), estimation of mutual information (e.g. Hutter (2001); Goebel et al. (2005)), and big data in finance more generally (e.g. Farboodi and Veldkamp (2023)). This paper also relates to the recent literature on the computational efficiency of large language models (e.g. Kaplan et al. (2020); Ding et al. (2023); Yang et al. (2024); Nozaki et al. (2025)), as well as the use of such models in finance (e.g. Wu et al. (2023)).

Paper structure

Section 4.2 investigates how the general investor feature selection problem can be guided by microeconomic theory. Section 4.3 explores different ways the proposed theory-based approach can be operationalised in practice. Section 4.4 explores how the proposed feature selection method could be applied to reduce the resource cost of financial language modelling, including an illustrative experiment. Section 4.5 concludes.

4.2 Theoretical ranking of features

The core argument of this section is that an investor facing the general practical problem of selecting features that are predictive of risk premia—as well as the specific problem of selecting which tokens to include in financial language modelling—can find guidance in formal results within the microeconomic literature on rankings of information structures. This argument is rooted in Cabrales et al. (2013)'s refinement of Blackwell's partial ordering of information structures (Blackwell, 1953).

By considering a specific subset of decision problems and utility functions that are relevant to financial investors, Cabrales et al. (2013) are able to complete Blackwell's ordering of information structures. In particular, Cabrales et al. consider a general class of ruin-averse agents. An agent has to choose between at least one risky, no-arbitrage investment and cash. Prior to investing, the agent can pay to observe the realisation of a discrete random variable ('the feature state') that may carry information about discrete risky investment outcomes.[3] Cabrales et al. demonstrate that an agent in this class would rank their willingness to pay for observing different feature realisations according to their *entropy informativeness*. Entropy informativeness is defined as the difference between the entropy of the agent's prior over outcome states and the (expected) entropy of the agent's posterior after observing a feature state. While the result relates to the discrete case, the continuous case can be approximated arbitrarily closely through fine 'change-of-variable' discretisation.

Using Cabrales et al.'s result, I demonstrate how an agent with knowledge of the joint distribution of asset risk premia and features (or tokens) would rank their willingness to pay to observe the realisation of a feature or token. While this distribution will in practice be unknown to the investor, the statistical problem of estimating from data the specific functional of this distribution that matters for the feature ranking is well-studied.

4.2.1 States of nature

Much of the traditional empirical asset pricing literature (e.g. Gu et al. (2020)), adopt a regression-based framework to predict risk premia. In the regression approach, investment outcomes are considered to be continuous and statistical performance is typically measured as the out-of-sample R-squared. The approach in this paper departs from this, instead taking a (multi-class) classification approach. That is, the aim is to predict which of the historically-observed quantiles the future excess return is likely to fall into, rather than aiming to predict its exact value. Statistical performance can be evaluated using a well-established set of evaluation metrics for classifiers. This classification approach can be justified by observing that minimum tick sizes mean that returns are not strictly continuous in practice.

I begin by specifying the possible states of nature for realisations of both the investment outcome state and the feature (or token) state. Let K be a

[3] Cabrales et al. (2013) refer to these random variables as 'signals' instead of 'features'.

random variable whose sample space $\Omega_K = \{1, ..., N_K\}$ represents all possible investment outcome states. Let S be a random variable whose sample space $\Omega_S = \{1, ..., N_S\}$ represents all possible feature states of a feature or token. The joint probability mass function $P_{KS}(k, s)$ of these two discrete random variables is a categorical (or multinoulli) distribution defined on sample space $\Omega = \Omega_K \times \Omega_S$.

It is important to note that the random variable S can represent either a single feature or multiple discrete features. For example, if there are two discrete features, each with two possible states, then these two features can be represented by a single discrete random variable with $N_S = 4$. As such, the ordering of information structures enables ranking *sets of features* rather than just ranking possibly highly-correlated features *individually.* It therefore, in principle, enables the selection of parsimonious, low-redundancy feature sets.

4.2.2 Information structures

In the context of this probabilistic model of the possible states of nature, what Cabrales et al. refer to as an information structure corresponds to a finite set of investment outcome states Ω_K, a finite set of feature states Ω_S, and, for each investment outcome state $k \in \Omega_K$, a probability distribution that specifies the probability of observing each signal in that outcome state. That is, the information structure relating to a feature (or set of features) is defined as the tuple $(\Omega_K, \Omega_S, \{P(S = s|K = k)\}_{k \in \Omega_K, s \in \Omega_S})$.

To illustrate this definition of an information structure, Table 12 displays the outcome-contingent feature state probabilities for the case of there being $N_K = 2$ possible investment outcome states and $N_S = 3$ possible realisations of feature state S. Each row in the table is a conditional distribution over feature states, given a realisation $k \in \{1, 2\}$ of the investment outcome state K. The first row of Table 12 is a conditional distribution over feature states given investment outcome state $K = 1$ so that $\sum\limits_{s} \frac{\pi_{1s}}{\pi_{1\bullet}} = 1$, where $\pi_{1s} = P(S = s, K = 1)$ and $\pi_{1\bullet} = P(K = 1)$. Similarly, the second row is a conditional

Table 12: Example of an information structure with $N_K = 2$ and $N_S = 3$

	$S = 1$	$S = 2$	$S = 3$
$K = 1$	$P(S = 1 \vert K = 1) = \frac{\pi_{11}}{\pi_{1\bullet}}$	$P(S = 2 \vert K = 1) = \frac{\pi_{12}}{\pi_{1\bullet}}$	$P(S = 3 \vert K = 1) = \frac{\pi_{13}}{\pi_{1\bullet}}$
$K = 2$	$P(S = 1 \vert K = 2) = \frac{\pi_{21}}{\pi_{2\bullet}}$	$P(S = 2 \vert K = 2) = \frac{\pi_{22}}{\pi_{2\bullet}}$	$P(S = 3 \vert K = 2) = \frac{\pi_{23}}{\pi_{2\bullet}}$

Notes: Each row is a conditional distribution over feature states, given an investment outcome state $k \in \{1, 2\}$ so that $\sum\limits_s \frac{\pi_{1s}}{\pi_{1\bullet}} = 1$ and $\sum\limits_s \frac{\pi_{2s}}{\pi_{2\bullet}} = 1$.

distribution over feature states, given investment outcome state $K = 2$ so that $\sum\limits_s \frac{\pi_{2s}}{\pi_{2\bullet}} = 1$, where $\pi_{2s} = P(S = s, K = 2)$ and $\pi_{2\bullet} = P(K = 2)$.

4.2.3 Ranking by entropy informativeness

Cabrales et al. (2013) establish an ordering of information structures according to *investment dominance*. Informally, an information structure α is understood to investment dominate another structure γ, if an agent's non-purchase of α implies non-purchase of γ.[4]

The main result of Cabrales et al. (2013) is that for a general class of ruin-averse agents choosing one from a set of no-arbitrage investments (including a risk-free option), with prior beliefs about the probability of each investment outcome state β_k, a ranking by investment dominance is equivalent to a ranking by each information structure's *entropy informativeness*.[5] Definition 1 provides a formal description of this quantity in the context of the investor's decision problem introduced above.

Definition 1 (Entropy informativeness). *The entropy informativeness EI of an information structure α is defined as*

$$EI(\alpha) = -\sum_k \beta_k log_2(\beta_k) - \sum_s \pi_{\bullet s} H(q^s(k)) \tag{4.1}$$

[4] See Definition 1 of Cabrales et al. (2013).

[5] See Theorem 1 of Cabrales et al. (2013).

where

$$q^s(k) = \beta_k * P(S = s|K = k)/P(S = s) = \beta_k * \frac{\pi_{ks}}{\pi_{k\bullet}} * \frac{1}{\pi_{\bullet s}}$$

is the agent's posterior belief in the investment outcome state K being k, after observing feature state $S = s$, and

$$H(q(k)) = -\sum_k q(k) log_2(q(k))$$

is the entropy of a probability distribution over states (with $0log_2 0 := 0$).[6]

Definition 1 illustrates that the entropy informativeness ranking by Cabrales et al. is dependent on the investor's beliefs $\{\beta_k\}$ about the probabilities of each of the investment outcome states. An information structure's entropy informativeness value is also dependent on the investor's beliefs about the conditional probability mass functions that make up each information structure. As such, in the context of ranking features or tokens, a feature or token's ranking score is a functional of all elements of the investor's beliefs regarding $P_{KS}(k, s; \boldsymbol{\pi})$.

4.2.4 The relationship between entropy informativeness and mutual information

In the absence of Knightian uncertainty—that is, $P_{KS}(k, s; \boldsymbol{\pi})$ is known to the investor or can be estimated by the investor with arbitrary accuracy—it can be established that there is an equivalence between ranking information structures by entropy informativeness and by a well-understood information-theoretic quantity. In particular, suppose that the investor's prior for the probability of each investment outcome β_k coincides with the true state probability $P(K = k)$. Proposition 1 demonstrates that, in this scenario,

[6] The base of the logarithm is chosen as 2 here so that the unit is shannons (informally: bits), rather than nats.

the entropy informativeness and the mutual information of the information structure would be identical quantities.

Proposition 1. *Let categorical random variables $K \in \Omega_K$ and $S \in \Omega_S$ be jointly distributed according to $P_{KS}(k, s; \boldsymbol{\pi})$, and the investor's beliefs be equal to $\boldsymbol{\pi}$. The entropy informativeness of the associated information structure is equal to the mutual information of K and S.*

Outline of proof. The mutual information of K and S is defined as

$$MI(K,S) = \sum_k \sum_s \pi_{ks} log_2 \frac{\pi_{ks}}{\pi_{\bullet s}\pi_{k\bullet}}. \tag{4.2}$$

Substituting $\beta_k = \pi_{k\bullet}$ into the definition of entropy informativeness (1) yields

$$EI(\alpha) = -\sum_k \pi_{k\bullet} log_2(\pi_{k\bullet}) + \sum_s \pi_{\bullet s} \sum_k \pi_{k\bullet} \frac{\pi_{ks}}{\pi_{\bullet s}\pi_{k\bullet}} log_2(\pi_{k\bullet} \frac{\pi_{ks}}{\pi_{\bullet s}\pi_{k\bullet}}) \tag{4.3}$$

which simplifies to

$$= -\sum_k \pi_{k\bullet} log_2(\pi_{k\bullet}) + \sum_s \sum_k \pi_{ks} log_2(\frac{\pi_{ks}}{\pi_{\bullet s}}). \tag{4.4}$$

Expanding first term using definition of $\pi_{k\bullet}$ and swapping summations of the second term yields

$$= -\sum_k \sum_s \pi_{ks} log_2(\pi_{k\bullet}) + \sum_k \sum_s \pi_{ks} log_2(\frac{\pi_{ks}}{\pi_{\bullet s}}) \tag{4.5}$$

which can be simplified to the RHS of (4.2).

□

This observation, while straightforward, is useful because it allows drawing on the literature on the estimation of mutual information to operationalise the theoretical ranking of Cabrales et al. (2013) for the purpose of ranking (sets of) features (or tokens).

4.3 Ranking score estimation

Proposition 1 shows that the investor would rationally rank features (or sets of features) based on each feature's mutual information with the investment outcome state. This depends on the investor having full knowledge of the probability mass function parameter vector $\boldsymbol{\pi}$. In practice, $\boldsymbol{\pi}$ and, by implication, $MI(\boldsymbol{\pi})$ will not be known by the investor. As such, this section examines how the investor can estimate $MI(\boldsymbol{\pi})$ for each feature (or set of features) from realised values of each period's investment outcome state k_t and each period's feature state s_t. I consider two cases. In the first case, the investor has no prior beliefs and wishes to estimate a ranking from the data alone. In contrast, in the second case, the investor has a Dirichlet prior over states and wishes to update beliefs in light of the data. For both cases, the estimators for these estimands are well-studied in the literature. It may be the case that $\boldsymbol{\pi}$ is unlikely to be constant over time. I therefore also consider how $MI(\boldsymbol{\pi})$ can be estimated when the data-generating process is time-varying. Finally, while sets of features can be ranked in principle, there is a hard practical constraint that prevents doing so in practice. I discuss the nature of this constraint and a potential workaround.

4.3.1 Frequentist ranking score estimation

If the investor is willing to assume that the random variables (k_t, s_t) are iid, the parameters of each information structure's categorical distribution with $N_K * N_S$ possible states of the world can straightforwardly be estimated from a sample using maximum likelihood estimation. The maximum likelihood estimators are the following intuitive count ratios

$$\hat{\pi}_{ks}^{ML} = \frac{\sum_t 1_{\{k_t=k, s_t=s\}}}{T} \; \forall \; k, \; s. \tag{4.6}$$

The estimators for the marginal distributions of K and S are defined analogously, since the marginal distributions are also categorical. Substituting these estimators into (4.2), yields a widely-used plug-in maximum likelihood estimator

$$\hat{I}_{plug-in} = \sum_k \sum_s \hat{\pi}_{ks}^{ML} log_2 \left(\frac{\hat{\pi}_{ks}^{ML}}{(\sum_k \hat{\pi}_{ks}^{ML}) * (\sum_s \hat{\pi}_{ks}^{ML})} \right). \tag{4.7}$$

As highlighted by Goebel et al. (2005), since mutual information is a non-linear and non-injective function of random variables K and S, deriving an exact sampling distribution for the plug-in estimator is nontrivial. In the special case of two discrete random variables whose mutual information is relatively small, the finite sample bias of the plug-in estimator can be approximated using a Taylor expansion approach. When K and S are independent (so that $MI(K, S) = 0$), Theorem 2 in Goebel et al. (2005) shows that the plug-in mutual information estimator is approximately distributed according to a gamma distribution with expected value $\frac{(N_K-1)(N_S-1)}{2*T*ln2}$. If instead K and S are dependent, Theorem 4 of Goebel et al. (2005) shows that the plug-in estimator is approximately distributed according to a noncentral gamma distribution with expected value $MI(K, S) + \frac{(N_K-1)(N_S-1)}{2*T*ln2}$.

In either case, the approximate bias is $\frac{(N_K-1)(N_S-1)}{2*T*ln2}$, although the approximation only works well for $MI(K, S) < 0.2$ shannons. While the bias term diminishes with T, the presence of N_S in the bias term can distort comparisons of features with different cardinality levels. Moreover, sample size differences for different features may also distort the resulting feature rankings.

4.3.2 Bayesian ranking score estimation

Instead of assuming that the investor relies on purely data-driven estimation, one could also assume that the investor forms beliefs through recursive Bayesian updating. In this approach, it is assumed that the investor begins with a symmetric Dirichlet prior $p(\boldsymbol{\pi_1}|\boldsymbol{\alpha_1}) = Dir(\boldsymbol{\alpha_1})$ for each information structure's categorical distribution parameter. It is well-known that the concentration parameter vector $\boldsymbol{\alpha}$ of the Dirichlet prior can be interpreted as 'pseudocounts'. That is, specifying a concentration of $\alpha^{(n)}$ for each of the N possible states of the world is akin to having observed $\alpha^{(n)}$ realisations of state n. As the Dirichlet distribution is the conjugate prior for categorical likelihoods, the investor's posterior distribution over $\boldsymbol{\pi_1}$ after observing the first realisation is also Dirichlet with new concentration parameter vector $\boldsymbol{\alpha_1'}$. The element of this vector corresponding to the state n_1 that has occurred is increased by one while all others are unchanged. This posterior distribution then becomes the investor's prior distribution for the next period according to the following iteration

$$p(\boldsymbol{\pi_t}|\boldsymbol{\alpha_t}) = Dir(\boldsymbol{\alpha_t}) \text{ (period t prior)}$$

$$p(\boldsymbol{\pi_t}|n_t, \boldsymbol{\alpha_t}) = Dir(\boldsymbol{\alpha_t'}) \text{ (period t posterior)}$$

$$\boldsymbol{\alpha_{t+1}} = \boldsymbol{\alpha_t'} \text{ (iterate forward)}$$

$$p(\boldsymbol{\pi_{t+1}}|\boldsymbol{\alpha_{t+1}}) = Dir(\boldsymbol{\alpha_{t+1}}) \text{ (period t+1 prior).}$$

When an investor forms beliefs over information structures in this way, they can obtain an exact analytical expression for the posterior mean and an approximate analytical expression for the posterior variance of the entropy informativeness for each information structure at each point in time. To see this, suppose the investor has posterior belief $Dir(\boldsymbol{\alpha'_t})$ about the categorical parameter $\boldsymbol{\pi_t}$ of an information structure, with pseudocounts $\alpha_t'^{(k,s)}$, marginal pseudocounts $\alpha_t'^{(k,\bullet)}$, $\alpha_t'^{(\bullet,s)}$, and total pseudocounts $\alpha_t'^{(\bullet,\bullet)}$. As shown in Hutter (2001), the investor's posterior distribution of the entropy informativeness corresponding to this information structure then has mean

$$E[I] = \frac{1}{\alpha_t'^{(\bullet,\bullet)}} \sum_{k,s} \alpha_t'^{(\bullet,\bullet)} [\psi(\alpha_t'^{(k,s)}+1) - \psi(\alpha_t'^{(k,\bullet)}+1) - \psi(\alpha_t'^{(\bullet,s)}+1) + \psi(\alpha_t'^{(\bullet,\bullet)}+1)] \tag{4.8}$$

where $\psi(\bullet)$ is the digamma function. In what follows, the posterior mean under the Dirichlet prior with all pseudocounts $\alpha_t'^{(k,s)} = 1$ is referred to as the 'Bayes-1' estimator.

The posterior variance can be approximated as follows

$$Var[I] = \frac{K - J^2}{\alpha_t'^{(\bullet,\bullet)}} + \frac{M + (N_K - 1)(N_S - 1)(\frac{1}{2} - J) - Q}{(\alpha_t'^{(\bullet,\bullet)} + 1)(\alpha_t'^{(\bullet,\bullet)} + 2)} + O({\alpha_t'^{(\bullet,\bullet)}}^{-3}) \tag{4.9}$$

where K, J, M, and Q are simple functions of the pseudocounts.[7] Zaffalon and Hutter (2002) describe how the expressions above can be used for mutual information-based feature selection. In particular, the mean and variance can be used to approximate the full posterior distribution of the estimand. In addition to simply ranking features (or sets of features) by posterior mutual information means, one can therefore also exclude features that, according to their credible intervals, are unlikely to be above a minimum relevance threshold.

4.3.3 Forgetting

The recursive belief formation iteration introduced above assumed that all observations of the state of the world are weighted equally. However, one

[7] See Hutter (2001).

may wish to allow for the data generating process to evolve over time. This can be accommodated straightforwardly by placing less weight on older observations. This can be achieved by placing an exponential forgetting factor $\lambda \leq 1$ on the pseudocounts when iterating forward as follows

$$\boldsymbol{\alpha_{t+1}} = \boldsymbol{\alpha_t'}^{\lambda} \text{ (iterate forward with forgetting).}$$

The effect of this forgetting factor is that at each iteration pseudocounts are shrunk towards unity, such that older observations receive less weight. Note that this is possible because the pseudocounts that comprise the Dirichlet concentration parameter are not restricted to the set of natural numbers. Such forgetting approaches are well-known in the literature in general, although they do not appear to have been proposed in the context of mutual information estimation.

4.3.4 Ranking sets of features

Section 4.2.1 highlights that both individual features and sets of features can, in principle, be ranked using the Cabrales et al. (2013) ordering. However, a practical challenge is that the cardinality of the sample space of random variable S increases exponentially with the size of the set of features to be ranked. For example, if S represented 100 binary variables, the cardinality of Ω_S would equal 2^{100}—around one quintillion. As such, an exact solution to the problem of ranking sets of predictors according to their *joint* entropy informativeness is practically infeasible. As a feasible approximation, Peng et al. (2005) propose a minimal-redundancy-maximal-relevance criterion (mRMR) which selects features sequentially in order to avoid the combinatorial explosion of the cardinality of S.

4.3.5 Simulation evidence

I now detail simulations conducted to gather evidence regarding the relative performance of the plug-in maximum likelihood and Dirichlet-prior Bayes estimators of mutual information. The key finding is that while the mean squared estimation error differs depending on the choice of estimator and prior (in the Bayesian case), the quality of the resultant ranking of features is largely unaffected by these choices. Since the ultimate objective of the investor is feature selection rather than parameter estimation, the evidence suggests that the standard plug-in maximum likelihood estimator is likely to be adequate for feature selection in settings that are similar to the ones simulated.

Data generating process

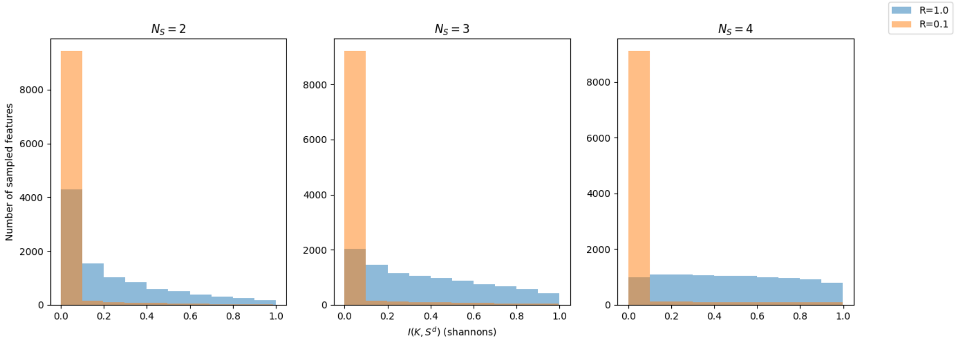
The data generating process should satisfy the following desiderata. Firstly, the aim is to draw a single investment outcome time series $\{k_t\}$ of length T and a set of N_D feature time series $\{s_t^d\}$ that have varying degrees of relevance to the investment outcome. Secondly, the 'ground truth' value of mutual information between each generated feature S and investment outcome state K needs to be known. Thirdly, the share R of the N_D features that has any relevance at all to K needs to be controllable. That is, the sparsity of the feature set needs to be modifiable. The exercise is limited to binary investment outcome states (e.g. 'up' and 'down'), but could easily be generalised.

Algorithm 2 outlines the resulting sampling procedure. In particular, it begins by defining the marginal distribution over investment outcome states, $P(K)$. Following this, it draws the conditional probabilities for each of the N_D features with cardinality N_S, each from a symmetric Dirichlet distribution with concentration parameter $\alpha_d = \frac{1}{N_S} \forall d$. Specifically, for each feature two distributions are drawn: the first conditional on $K = 1$ and the second conditional on $K = 0$. To achieve the desired level of sparsity, these two conditional distributions are set equal for some share R of the features—resulting in these features carrying no information about K. The true mutual information values between the feature and investment outcome random variables can now be obtained. Finally, T realisations of K and S^d are drawn from its marginal and conditional distributions.

Algorithm 2 Data generating process

```
Require: $T \in \mathbb{N}, N_D \in \mathbb{N}, N_S \in \mathbb{N}, R \in (0,1]$

  Define investment outcome distribution:
  $P(K=1) \Leftarrow 0.5$
  $P(K=0) \Leftarrow 0.5$

  For each feature, sample conditional feature distributions:
  for $d \in 1...N_D$ do
      $P(S^d = s|K=1) \;\sim\; \mathrm{Dir}(\alpha = \frac{1}{N_S})$
      $P(S^d = s|K=0) \;\sim\; \mathrm{Dir}(\alpha = \frac{1}{N_S})$
      if $d < (1-R) * N_D$ then
          $P(S^d = s|K=1) \Leftarrow P(S^d = s|K=0)$ ▷ Make feature irrelevant to $K$
      end if
  end for

  For each feature, obtain true mutual information with investment outcome state $I(K, S^d)$:
  for $d \in 1...N_D$ do
      $I(K,S^d) \Leftarrow \sum_k \sum_s P(K=k, S^d=s) \log\left(\frac{P(K=k,S^d=s)}{P(K=k)P(S^d=s)}\right)$
  end for

  Draw investment outcome and feature time series:
  for $t \in 1...T$ do
      $k_t \sim P(K)$
      for $d \in 1...N_D$ do
          $s_t^d \sim P(S^d|K=k_t)$
      end for
  end for
```

Figure 58: Sampling distribution of $I(K, S^d)$ for different values of N_S and R



Notes: $I(K, S^d)$ is the true mutual information measured in shannons. The number of features is $N_D = 10,000$. N_S is the cardinality of features and R is the share of relevant features.

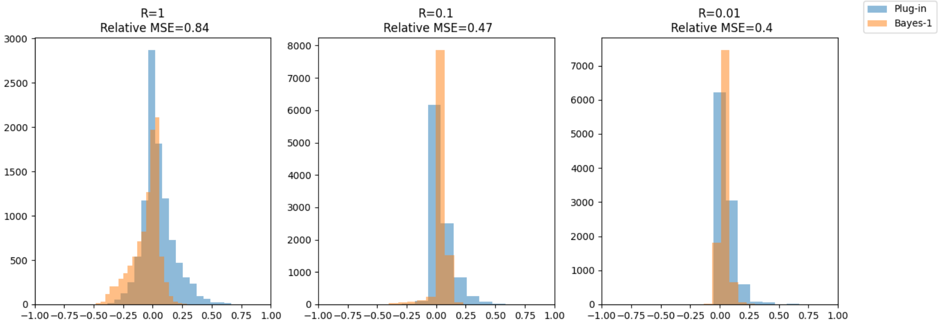
This data generating process is flexible enough to cover a range of different feature cardinality and sparsity scenarios. This is illustrated in Figure 58, which shows sampling distributions of mutual information $I(K, S^d)$ measured in shannons. Different values of feature cardinality N_S and relevance (i.e. inverse sparsity) R are shown for $N_D = 10,000$ sampled feature distributions.

Estimation error

Next, T observations are sampled from the joint feature-outcome distribution $P(K, S^d)$. Obtaining synthetic time series in this way enables a comparison of how well different estimators of mutual information recover the true value.

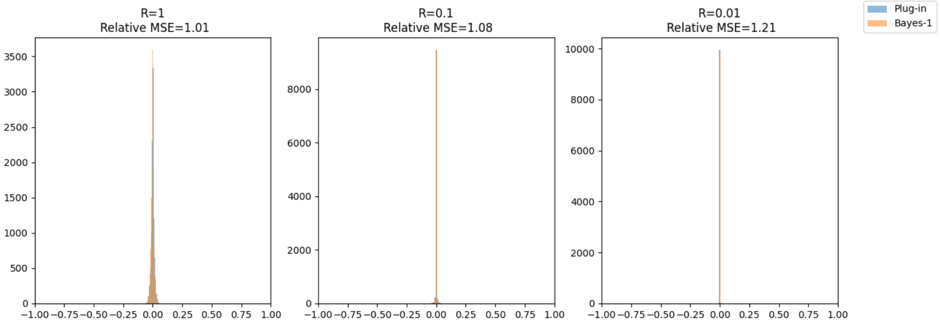
As an illustration, Figure 59 displays the estimation error of each feature's mutual information for both the plug-in and the Bayes-1 estimator. For very small sample sizes, the Bayes-1 estimator achieves significantly lower mean squared errors. As the number of observations T increases, the mean squared error of the Bayes-1 estimator begins to exceed that of the plug-in estimator—as can be seen in Figure 60.

Figure 59: $\hat{I}(K, S^d) - I(K, S^d)$ for different estimators ($T = 10, N_D = 10,000, N_S = 2$)



Notes: $\hat{I}(K, S^d)$ is the estimated mutual information and $I(K, S^d)$ is the true mutual information (both measured in shannons). Results for the Plug-in and Bayes-1 estimators are shown. 'Relative MSE' refers to the mean squared error corresponding to the Bayes-1 estimator relative to that of the Plug-in estimator.

Figure 60: $\hat{I}(K, S^d) - I(K, S^d)$ for different estimators ($T = 1,000, N_D = 10,000, N_S = 2$)



Notes: $\hat{I}(K, S^d)$ is the estimated mutual information and $I(K, S^d)$ is the true mutual information (both measured in shannons). Results for the Plug-in and Bayes-1 estimators are shown. 'Relative MSE' refers to the mean squared error corresponding to the Bayes-1 estimator relative to that of the Plug-in estimator.

Ranking quality

It can be argued that what matters for feature selection is not the estimation error, but the quality of the resulting ranking. To evaluate the performance of different mutual information estimators from this perspective, I use *recall@k*[8], a standard evaluation metric in the literature on recommender systems (Jégou et al., 2011), which is defined as follows

$$recall@k = \frac{\sum_{d \in 1 \ldots D} 1_{d\ selected=True} * 1_{d\ true\ relevance\ in\ top\ k}}{k}. \tag{4.10}$$

Informally, if one selects the k features with the highest estimated mutual information values, *recall@k* corresponds to the share of these selected features whose true mutual information values rank in the 'ground truth' of the k features with the highest true mutual information values. This type of evaluation is only possible in simulations, where the true mutual information values are known.

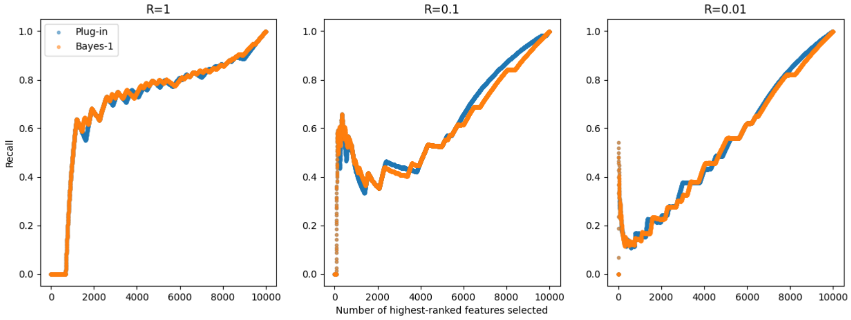
Similar to the mean squared estimation error evaluation above, I present results for the $T = 10$ and $T = 1,000$ case, shown in Figures 61 and 62. Across sample sizes T, sparsity levels R, and the number k of highest-ranking features selected, there is little difference between the two estimators. That is, the difference in mean squared estimation error does not translate into a noticeable difference in ranking quality.

4.4 Reducing the resource cost of financial language models using token rankings

In this section, I analyse how the general microfounded ranking of features derived in Section 4.2 and the estimation approaches analysed in Section 4.3

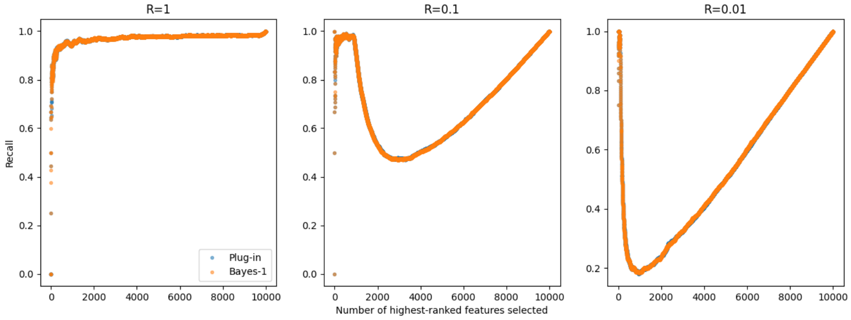
[8] Note that, in this context, k refers to the number of highest-ranked features, rather than the realisation of the random variable K.

Figure 61: Recall@k for different estimators ($T = 10, D = 10,000, N_S = 2$)



Notes: Recall is shown for different values of k along the horizontal axis. R is the share of relevant features.

Figure 62: Recall@k for different estimators ($T = 1,000, D = 10,000, N_S = 2$)



Notes: Recall is shown for different values of k along the horizontal axis. Results for the Plug-in and Bayes-1 estimators are shown.

can be leveraged to reduce the resource cost of financial language models. The specific objective of this section is to explore how such models can be made more efficient at the specific task of predicting financial risk premia using text. In the parlance of the language modelling literature, this is a 'downstream' task.

The general literature on large language model efficiency tends to focus on the 'language modelling' task, where the objective is to model language itself. For example, in order to operationalise the concept of efficiency, Ding et al. (2023) survey a range of approaches that have been proposed in the literature to increase the efficiency of large language models. In this literature efficiency is defined using metrics such as number of parameters of the language model, floating point operations needed, time required for inference, as well as carbon emissions caused. Approaches to increase the efficiency of large models include training data filtering and changes to the model architecture. It is important to note that performance on the general 'language modelling task' is not the same as performance on 'downstream' tasks. In particular, there is little existing literature on the specific 'downstream' task of interest to an investor: the prediction of financial risk premia using textual data.

One strategy to optimise a large financial language model specifically for the 'downstream' task of predicting the risk premium of a financial asset or set of financial assets of interest would be to experiment with different training datasets and model architectures, training and testing large models with different configurations. However, Ding et al. note that a 'trial-and-error' approach to optimising large language models is generally costly due to the resource intensity of the training process. The large number of potential assets whose risk premia an investor may wish to predict using text make the trial and error approach unviable.

The appeal of a principled way of reducing the resource cost of financial language models is that such costly experimentation would be avoided. This section explores how, instead of using costly trial and error, the microfounded investor feature ranking discussed in Section 4.2 can be used to identify those tokens that are most relevant for predicting the risk premium of a particular financial asset or set of financial assets. I then show how doing so can reduce the investor's resource cost of training, fine-tuning, and using transformer-architecture financial language models.

4.4.1 The resource cost of financial language models

Kaplan et al. (2020) estimate the number of floating point operations C needed for *training* a transformer model as

$$C \approx 6ND_{train} \tag{4.11}$$

where N is the number of model parameters and D_{train} is the size of the training dataset in tokens. A token can correspond to words, subwords, characters, or punctuation. For example, Wu et al. (2023) report that the training dataset used for the BloombergGPT model contained 363bn tokens and that 1.3m GPU-hours[9] were used to train its 50bn parameters. With the carbon intensity of US electricity generation at 369gCO2/kwH (Ritchie et al., 2023), and the A100 GPUs used by Wu et al. drawing 400 watts of power, the training of this model can be estimated to have resulted in 192 tons of CO2 being emitted. For *fine-tuning*, the computational effort would generally be smaller than C, although it would vary depending on the share of parameters being selected for fine-tuning and the size of the fine-tuning dataset. Based on Ding et al. (2023), the number of floating point operations needed for transformer-model *inference* (i.e. using the model) is $O(D_{infer})$ where D_{infer} is the length of the input sequence (in tokens) being passed into the language model for each prediction.

[9] A Graphics Processing Unit (GPU) is a specialised processor optimised for parallel processing tasks, such as large language model training.

4.4.2 Vocabulary size reduction

The vocabulary size of large language models can be sizeable. For example, there are around 200,000 different tokens that are modelled within OpenAI's GPT-4o (Yang et al., 2024). In the context of the specific downstream task of predicting financial risk premia, it is not obvious that all of these tokens are necessary. An investor training a large financial language model from scratch, could use the mutual information ranking described in Sections 4.2 and 4.3 to reduce the size of the vocabulary of tokens that are modelled. That is, only those tokens which are among the most predictive for the risk premium of interest would be modelled.

The intuition of this approach is a hypothesis that there may be a 'Pareto principle'-type power law with respect to the tokens' usefulness to the investor. For example, it may be the case that around 20% of the tokens account for around 80% of the downstream task performance relative to a noise baseline. The purpose of a principled, microfounded ranking is to identify those 20% of tokens reliably. Crucially, because the ranking proposed in this paper is theory-based, one can identify these tokens without incurring the cost of repeatedly training and testing a model to find out which sets of tokens are most useful for the downstream task of predicting risk premia.

There are at least three ways in which vocabulary size reduction would affect the resource cost of training a financial language model in Equation 4.11. Firstly, the training corpus size would be smaller. For instance, if 80% of tokens were to be excluded from the vocabulary on the basis of not being among the highest-ranked 20% of tokens the size of the training data—measured in number of tokens—could be reduced substantially. The exact reduction would depend on the relative incidence of the selected compared to the excluded tokens. Secondly, the vocabulary reduction would directly result in a model architecture with fewer parameters in the embedding layer (Nozaki et al., 2025). Thirdly, and more indirectly, the number of parameters in the model more generally may be able to be reduced in response to the smaller training dataset size without significant loss of predictive performance.

Analogously to the training cost reduction, the cost of fine-tuning a model with reduced vocabulary size would be lower due to the dataset used for fine-tuning having a lower token count. Moreover, the smaller number of parameters in the embedding layer may also result in lower fine-tuning costs.

Similar to training and fine-tuning costs, the cost of passing textual data through the language model to generate predictions of risk premia would be reduced due to smaller vocabulary size and the textual data being passed having lower token counts.

4.4.3 Partial parameter tuning

Irrespective of whether the investor has trained the financial language model from scratch or is fine-tuning an existing model, a token ranking such as the one discussed in Sections 4.2 and 4.3 could be used to reduce fine-tuning costs. In particular, only those tokens which have the highest mutual information with the investment outcome of interest could be selected for fine-tuning.

For example, suppose the investor has already selected a pretrained language model that they wish to fine-tune to predict the risk premia of different assets. The pretrained model could be the BloombergGPT model of Wu et al. (2023), which is a general financial language model that could be fine-tuned to predict returns of specific financial assets 'downstream'. Partial parameter tuning involves only adjusting a subset of parameters within the transformer architecture to optimise performance on such specific downstream tasks. Ding et al. (2023) note that partial parameter tuning can be effective, but such approaches often 'lack [a] detailed principle to guide how to select a subset of parameters for further tuning'. The ranking in Section 4.2 can provide such guidance as to which parameters are worth fine-tuning and which are not.

Since the financial language model would already be trained at the time of fine-tuning, there would be no direct reduction in training costs as a result of using a token ranking to identify parameters to tune selectively. There would, however, be a potential reduction in fine-tuning costs. Rather than fine-tuning all parameters in the model, partial parameter tuning would only vary the subset of language model parameters that is most relevant to the specific risk premium prediction problem the investor faces. As such, the resource cost of fine-tuning could potentially be reduced substantially.

With regards to inference costs, there would be no immediate impact on the resource cost of making risk premium predictions as a result of using a partial parameter tuning approach. That is because both data and model size would be unaffected.

4.4.4 Input sequence pruning

Finally, regardless of whether the model has been trained on the full token set and whether selective fine-tuning was used, the investor could choose to prune the input token sequence. For example, say the investor is trying to predict whether the equity risk premium will be positive using newspaper headlines. The input sequence pruning approach would involve removing all but the highest-ranked tokens from the headline text before passing the headline to the model to generate a prediction.

Neither training effort nor fine-tuning costs would be reduced by this approach. However, inference costs—that is, the costs associated with using a model—could be reduced through pruning input tokens based on each token's mutual information with the risk premium of interest.

4.4.5 Empirical example: market timing based on newspaper headlines

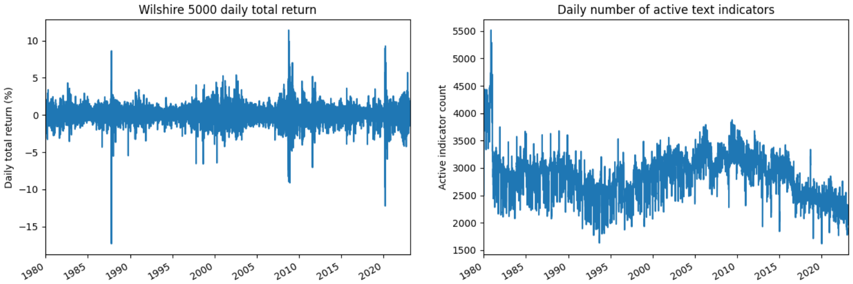
The above discussion identifies a number of ways in which a microfounded, mutual information-based ranking of features could be used to reduce the resource cost of financial language models. Of the approaches discussed, utilising the ranking to reduce the size of a language model's vocabulary appears to be particularly promising. The reason for this is that a smaller token vocabulary would lead to computational efficiency gains throughout the model lifecycle—from training to inference. An additional reason for using feature selection to reduce a financial language model's vocabulary size is that it may enhance the performance of model-driven trading strategies.

A detailed empirical investigation of the tradeoff between possible resource cost savings through vocabulary size reduction and associated impacts on predictive performance when using large financial language models is beyond the scope of this paper. This section does, however, provide an illustrative example. In particular, I consider the problem faced by an investor making daily investment decisions about whether to be exposed to the aggregate equity market based on newspaper headline text. In machine learning parlance, this problem can be considered a time series classification task.

Various statistical techniques have been proposed for this class of problem, most of which require selecting a manageable set of features prior to fitting a supervised machine learning model to predict the daily risk premium. The problem of having too many features is aggravated when the investor wishes to use textual data to classify the sign of the daily equity risk premium. The machine learning literature has proposed a wide range of feature selection methodologies to enable prediction in such high-dimensional contexts. This raises the question of which of these many methods is most likely to identify a strong set of predictors.

The appeal of using the mutual information-based ranking approach proposed in Section 4.2 is that it has a theoretical claim to microeconomic optimality. In what follows, I detail an experiment where the theoretical ranking of information structures proposed is put into practice to train a simple financial language model for classifying the sign of the daily US equity risk premium. The model trained is one of the simplest-possible language models.

Figure 63: Market returns and textual features



Notes: The left panel shows daily percentage changes of the Wilshire 5000 Total Market Full Cap Index (FRED mnemonic: WILL5000INDFC). The right panel shows daily counts of non-zero token indicator variables extracted from New York Times headlines.

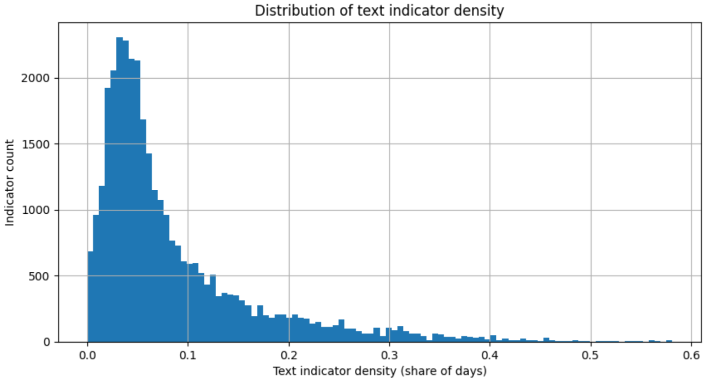
In particular, I adopt an n-gram representation of text and train a single-layer classifier on the resulting text-derived features.

Data

The US equity market risk premium is measured using daily percentage changes of the Wilshire 5000 Total Market Full Cap Index (FRED mnemonic: WILL5000INDFC), obtained from the Federal Reserve Bank of St. Louis' FRED database. The reason for choosing this particular broad market index is that it is constructed to cover the total return to equity, including dividends. The resulting time series is shown in the left panel of Figure 63. In order to convert the raw returns into investment outcome states, I discretise the daily returns such that there are two possible states: positive market return or negative market return.

As mentioned, the investor is trying to predict these daily binary investment outcome states using newspaper headline text. For the purpose of this illustrative exercise, New York Times headlines obtained via the NYT

Figure 64: Sparsity distribution of textual features



Notes: Histogram of the share of days each of the 10,000 text-derived token indicators is non-zero.

Archive API are used as a textual corpus.[10] Adopting an n-gram representation of text, token counts of all 1-grams (words), 2-grams (word pairs), and 3-grams (word triplets) are then computed. These counts are then summed across all headlines on each publication day and converted to a Boolean indicator equal to one if a token newly appeared on a given day, and 0 otherwise. While this binary discretisation constitutes a loss of information, repetition of tokens across headlines on a single day are relatively rare—apart from very common words such as ‘the’ and ‘and’ which are unlikely to have economic content. The text representation is limited to the 10,000 most common indicators. Each of these phrase indicators represents a binary feature, so that the number of feature states for each feature is $N_S = 2$.

Time series of daily counts of non-zero token indicator variables are shown in the right panel of Figure 63. Figure 64 displays a histogram of the share of days each of the 10,000 text-derived token indicators is non-zero. For each indicator, I create five lags, representing one week’s worth of trading days. To ensure that the experiment simluates a feasible trading strategy,

[10] https://developer.nytimes.com/docs/archive-product/1/overview

only lagged text-derived n-gram features are used to predict the investment outcome indicator.

Using investor feature selection to identify a subset of informative headline text tokens

Following the generation of the set of 10,000 candidate predictors, I use the Bayes-1 estimator introduced in Section 4.3 to compute the ranking score for each text-derived feature. These scores are computed using only the training period of 1979–1999 to prevent temporal lookahead bias (see Chapter 3). Based on these ranking scores, three feature sets—covering the 100, 1,000, and 10,000 highest-ranked features—are constructed.

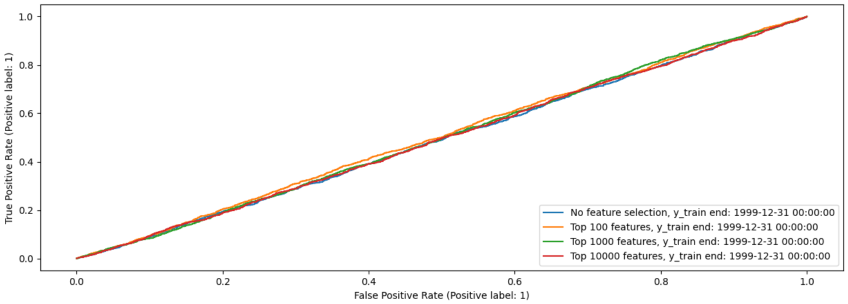
Classifier training

For each feature set, I train a standard naïve Bayes classifier to predict the investment outcome state using the training sample period of period of 1979–1999. Specifically, I use the BernoulliNB estimator of the *scikit-learn* Python library with default parameters to predict whether the one-day ahead market return is positive or negative. One classifier is trained for each set of selected features.

Having trained each model, the naïve Bayes classifiers are used to predict the probability that the US equity risk premium is positive on unseen data covering the time period from 2000 to 2023. The overall timeframe is driven by data availability and the train-test split is chosen to be approximately 50:50.

Figure 65 displays the possible true positive and false positive rates that the trained classifiers could attain, depending on the choice of the classifier threshold. It illustrates two points. Firstly, predicting aggregate market movements is challenging even with textual data, as reflected in the near-diagonal lines for all classifiers. That is, none of the classifiers are much better than a coin flip—in line with the efficient market hypothesis. How-

Figure 65: Classifier evaluation



Notes: True positive and false positive rates that could be attained on the test set for different choices of the classification threshold.

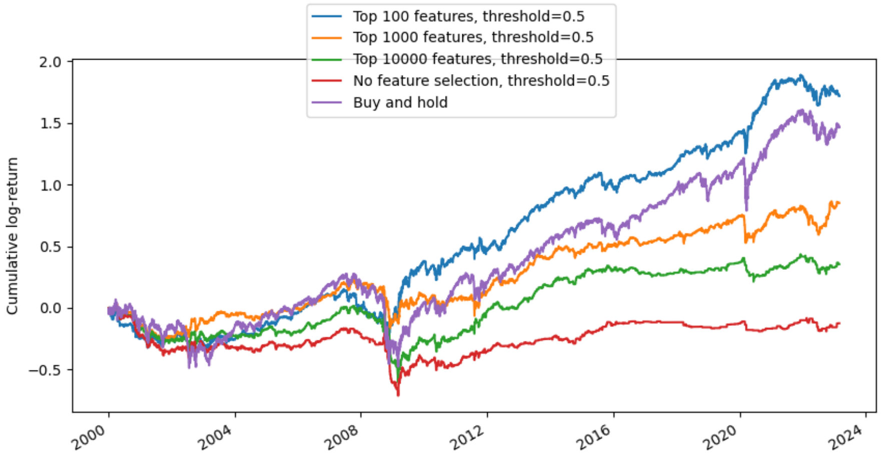
ever, feature selection appears to improve performance above that of the no-selection benchmark across the true positive-false positive tradeoff curve.

Trading strategies

For each model, the trading strategy involves investing in the equity market when the predicted probability of a positive risk premium is above 50% and holding cash otherwise. That is, the models are used to make market-timing decisions, using the naïve Bayes classifier's default probability threshold of 0.5. Each of the portfolios constructed in this way is compared to a market portfolio that is invested in the market at all times—corresponding to a buy-and-hold trading strategy. Figure 66 shows the results of this exercise for each of the models with different feature counts.

The results shown in Figure 66 suggest that reducing the vocabulary size of a 'language model' can not only reduce computational effort but also improve the economic performance of financial language models. It is important to note, however, that these findings are illustrative. For example, the performance of the model corresponding to the 100 highest-ranked tokens is sensitive to the choice of classifier threshold, as Figure 67 shows. The

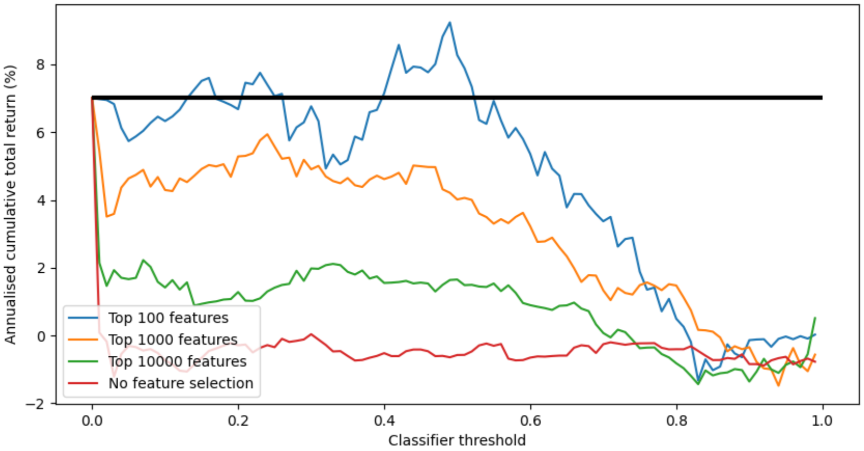
Figure 66: Cumulative total return of classifier portfolios



Notes: Cumulative log returns attained by trading strategies based on different classifiers with default classification threshold of 0.5.

classifier based on the top 100 features happens to outperform at the default classification threshold of 0.5, but not at other thresholds. That is, only for certain threshold choices is the blue line in Figure 67 above the solid black line indicating the buy and hold return of around 7% p.a. That said, in practice data-driven investors are likely to regularly refresh their prediction pipelines covering the preprocessing, feature selection, model training, and threshold optimisation. In addition, sophisticated investors are also likely to have access to proprietary datasets that may contain stronger predictive signals than low-frequency public newspaper headlines. As such, this illustrative exercise may understate the extent to which the principled, theory-based investor feature selection approach developed in Sections 4.2 and 4.3 can reduce the resource cost of data-driven capital allocation decisions and enhance their effectiveness.

Figure 67: Threshold choice and total return



Notes: Annualised cumulative total returns by trading strategies based on different classifiers across different choices of the classification threshold.

4.5 Conclusion

In this paper, I explore how a data-driven investor's approach to feature selection can be informed by economic theory and information theory. I also investigate how to estimate a theory-implied ranking in practice, considering both frequentist and Bayesian estimation. I also consider how the proposed feature selection methodology could be applied to increase the computational efficiency of large financial language models. The performance of different estimators is explored in simulations, and an empirical exercise illustrates how the proposed feature selection methodology could be incorporated into investors' use of large language models.

I conclude that for an investor mutual information-based feature selection has a strong theoretical justification. Examining different estimation approaches using simulations, I find that while the choice of mutual information estimator is subject to the bias-variance trade-off, different estimators return similar feature rankings. In the context of financial large language modelling, ranking tokens by their mutual information with the investment returns of interest to decide which tokens should be included in the model's vocabulary could result in both computational and accuracy benefits. In an illustrative exercise, I find suggestive support for the hypothesis that such model size reductions need not result in reduced predictive performance.

A limitation of this paper is that both features and the investment outcome state are assumed to be discrete throughout. In theory, this assumption can be relaxed by arbitrarily fine discretisation of continuous returns and features. In practice, finer discretisation makes estimation more challenging statistically. Similarly, while it is theoretically possible to rank *sets of features* rather than individual features, this is computationally prohibitive in practice. Approximate feature set selection methods such as mRMR (Peng et al., 2005) could enable the selection of parsimonious, low-redundancy predictor sets, but examining this hypothesis is beyond the scope of the present study. Finally, significant resources would be needed to quantify the extent to which 'pruning' large financial language models in the way proposed in this paper affects computational costs and predictive performance.

With considerable computational resources bound to be dedicated to the training and use of financial large language models, further research into these open questions appears likely to provide valuable insights for practitioners and could result in reduced aggregate resource usage, higher levels of market efficiency, or both.

Chapter 5

Conclusion

In this thesis, I consider the extent to which textual data can contribute to solving causal and predictive economic inference problems. In doing so, I specifically investigate opportunities created by recent advances in natural language modelling, which appear to remain underexplored as an econometric methodology—a gap this thesis aims to help fill. I first revisit an important causal inference problem by using a large language model to obtain measurements of otherwise unobserved confounding variables from relevant documents. Following this, I investigate the potential and pitfalls when using language models to incorporate textual data into economic forecasts. Finally, I study how language models can be optimised for solving specific economic inference problems.

5.1 Summary of findings

Reassessing the effects of the Bank of England's monetary policy on the UK economy in Chapter 2, I find that the proposed text-orthogonalisation approach using a large language model makes a material difference to the estimated dynamic causal effects of UK monetary policy on macroeconomic aggregates. The sign of estimated effects can depend on whether or not conventional monetary policy surprises are orthogonalised with respect to measurements of public, pre-event information regarding UK economic conditions extracted by prompting a large language model. Orthogonalisation with respect to these text-derived measures resolves both a real activity puzzle and an employment puzzle that arise when using conventional (i.e. non-orthogonalised) monetary policy surprises to identify monetary policy shocks. The resulting causal effect estimates are aligned with theoretical consensus and international evidence on monetary non-neutrality.

In Chapter 3, I investigate the potential and pitfalls when using language models to incorporate textual data into economic forecasts. I find that the 8,580 Federal Reserve 'Beige Book' reports appear to contain predictive information about US macroeconomic aggregates, although the gain in forecast accuracy relative to a standard forecasting model without text-derived predictors appears to be generally modest with some variation across forecast targets and horizons. I find that temporal lookahead bias—a key pitfall when using language models for economic forecasting—appears to have a measurable impact on out-of-sample evaluation metrics. In particular, using a knowledge cutoff experiment to simulate the training of a language model at different points in time, I estimate that up to half of the estimated forecast accuracy gain due to the addition of text-derived predictors generated using language models is illusory.

In Chapter 4, I find that in settings relevant to investors, ranking features using the information-theoretic quantity mutual information has a strong theoretical justification. Examining different estimation approaches using simulations, I find that while the choice of mutual information estimator is subject to the bias-variance trade-off, different estimators return similar feature rankings. In the specific context of financial large language modelling, textual tokens could be ranked by their mutual information with the investment returns of interest to decide which tokens should be included in the model's vocabulary. This could result in both computational and accuracy benefits as a result of a smaller, more focused token vocabulary. The results of an illustrative experiment are consistent with this hypothesis.

5.2 Implications

The findings in Chapter 2 demonstrate that economically-meaningful information can be extracted from unstructured text by prompting a large language model. In the context of the literature on the dynamic causal effects of monetary policies, the findings lend empirical support to the 'central bank response'-mechanism proposed in Bauer and Swanson (2023a). That is, my findings are consistent with the idea that it is market participants' imperfect knowledge of the monetary authority's reaction function to publicly-accessible news that explains the observed endogeneity of conventional, unorthogonalised monetary policy surprises.

Chapter 3 finds that textual data that were generated with the explicit aim of providing a forward-looking view of economic conditions appear to contain only modest incremental information over and above that contained within conventional predictors used for macroeconomic forecasting. This finding is robust to different representations of textual data. The result may indicate that the sample size of macroeconomic time series is too small for their relationships to high-dimensional textual data to be estimated with sufficient precision. An alternative, and possibly complementary, interpretation is that the relationship between text and forecast targets evolves too rapidly to be estimated. The propensity of using the word 'crisis', for instance, may evolve from one recession to the next. The apparent presence of temporal language model lookahead bias suggests that practitioners would be well-advised to only use models where the knowledge cutoff is known and to have regard to this date in their analyses of time series. However, as Ludwig et al. (2025) note, knowledge cutoffs are not foolproof due to the possibility of language models being fine-tuned on datasets that include information generated after the knowledge cutoff of the training dataset. As such, determining the date that a particular language model version was created on may be the only feasible way to ensure that temporal lookahead biases can be avoided.

Chapter 4 suggests that there is a role for economic theory in guiding the design and use of large language models for empirical asset pricing. The benefits of using theory-based rankings to determine—before training the model—which text tokens are worth including in a large language model for the purpose of empirical asset pricing are potentially significant. Such microfounded token filtering could even enable the training of asset-specific language models that are specialised to predicting the risk premia of particular financial assets. As such, selecting vocabularies of financial large language models in a theory-based way could result in reduced aggregate resource usage, higher levels of market efficiency, or both.

5.3 Contributions to the literature

Due to the latest significant advances in natural language processing being recent, the use of advanced language models as econometric tools remains yet to be comprehensively studied. This thesis makes a number of methodological, empirical, and theoretical contributions to help close this gap. Methodological contributions include a novel approach to generate structured measures of economic conditions from unstructured text. This approach is general and the textual data that are required to implement it are available across jurisdictions. I also propose a novel methodology for high-dimensional time series regression, which is designed for the large sets of predictor variables that tend to arise when representing text numerically using language model embeddings. The method enables efficient forecast evaluation and can be specified to allow for time-varying parameters. I also show how a knowledge cutoff experiment can be conducted to assess the impact of language model temporal lookahead bias on forecast evaluation results.

Empirical contributions include evidence of ex post correlations between state-of-the-art measures of UK monetary policy surprises and structured pre-event information extracted from text by the large language model, as well as new evidence regarding the effects and transmission of UK monetary policy, in Chapter 2. Chapter 3 contributes a quantification of the incremental value of Beige Book text for forecasting US macroeconomic aggregates, as well as measurements of language model temporal lookahead bias.

Finally, Chapter 4 makes a theoretical connection between the choice of which words or tokens to include in the vocabulary of a task-specific financial language model and the microeconomic literature on rankings of information structures. It also explores and demonstrates how this theoretical finding can be applied to reduce the resource costs of financial language modelling in practice.

5.4 Limitations

The prompt engineering approach taken in Chapter 2 may not be optimal in the sense of extracting the maximum amount of predictive information about monetary policy surprises from text. Moreover, while prompts were worded carefully, it is possible that at least part of the apparent predictability identified in Chapter 2 is driven by a lookahead bias of the kind measured in Chapter 3. To the extent that lookahead issues affect Chapter 2, this would have implications for the theoretical implications of the findings without necessarily invalidating causal effect estimates.

A key constraint throughout this project has been the computational intensity of handling and modelling high-dimensional textual data. For instance, state-of-the-art methods for language modelling require specialist hardware. As a result, the exhaustive exploration of different modelling choices, parameter choices, and task-specific fine-tuning of models in Chapter 3 is computationally prohibitive. For the same reason, performing realistic out-of-sample forecast evaluations when predictors are obtained from language models is left for future research. To sidestep the language model retraining requirement, the knowledge cutoff experiment used to estimate temporal lookahead bias in Chapter 3 involves text embeddings obtained from a moderately-sized language model rather than a large language model. The resources required to perform a similar experiment for a cutting-edge large language model would be substantial. The estimation of temporal lookahead bias is also based on the assumption that a larger sample size used to train the language model would not decrease out-of-sample forecast performance.

The empirical exercise in Chapter 4 is limited to a minimalistic language model due to the significant computational resources needed to perform a similar experiment regarding the training process of a transformer-architecture model. An extension to larger models is left for future work.

5.5 Suggestions for future research

The methodology in Chapter 2 could be applied in other jurisdictions. It could also be refined to fine-tune a language model directly, rather than using prompts to measure economic conditions and then using these measures to predict monetary policy surprises. A direct comparison of both methods could provide valuable insight into which approach extracts stronger predictive signals, although care must be taken that, in the process of phrasing prompts, the researcher does not introduce lookahead biases.

Findings regarding temporal lookahead biases in Chapter 3 suggest that further work to estimate the size such biases across different economic forecasting settings could be worthwhile. Other avenues for future research include alternative modelling approaches to relate language model embeddings to forecast targets, as well as more cutting-edge text embedding models. Further research into forecasting using textual data in general and the Beige Book in particular could explore the use of prompt engineering, as done in

Chapter 2, instead of text embeddings to represent text numerically. Similar studies in other jurisdictions would also aid in understanding the incremental value of text.

In addition to extending the illustrative empirical experiment in Chapter 4 to larger models, further work could explore different approximations to enable the use of mutual information to rank and select parsimonious sets of features rather than ranking features individually. Doing so would provide practitioners with a clearer picture of the cost-performance tradeoff when using large financial language models for empirical asset pricing. Any such studies need to be conducted carefully to avoid evaluations being biased by temporal lookahead issues of the kind identified in Chapter 3.

5.6 Concluding remarks

This thesis shows that economic inferences—be they causal or predictive—can be enhanced by incorporating the information contained within unstructured textual data. Advances in language modelling have created new opportunities to leverage vast quantities of currently-unused textual data, although I find that it is not always the most complex language model that performs best. More research is needed to better understand pitfalls and to realise fully the transformational potential of large language models for economic decision-making.

Bibliography

Algaba, A., Ardia, D., Bluteau, K., Borms, S., and Boudt, K. (2020). Econometrics meets sentiment: An overview of methodology and applications. *Journal of Economic Surveys*, 34(3):512–547.

Andrade, P. and Ferroni, F. (2021). Delphic and odyssean monetary policy shocks: Evidence from the euro area. *Journal of Monetary Economics*, 117:816–832.

Antweiler, W. and Frank, M. Z. (2004). Is all that talk just noise? the information content of internet stock message boards. *The Journal of Finance*, 59(3):1259–1294.

Aquilina, M., Budish, E., and O'Neill, P. (2021). Quantifying the high-frequency trading "arms race". *The Quarterly Journal of Economics*, 137(1):493–564.

Ardia, D., Bluteau, K., and Boudt, K. (2019). Questioning the news about economic growth: Sparse forecasting using thousands of news-based sentiment values. *International Journal of Forecasting*, 35(4):1370–1386.

Armesto, M. T., Hernández-Murillo, R., Owyang, M. T., and Piger, J. (2009). Measuring the information content of the beige book: A mixed data sampling approach. *Journal of Money, Credit and Banking*, 41(1):35–55.

Aromi, J. D. (2020). Linking words in economic discourse: Implications for macroeconomic forecasts. *International Journal of Forecasting*, 36(4):1517–1530.

Aruoba, S. B. and Drechsel, T. (2024). Identifying monetary policy shocks: A natural language approach. Working Paper 32417, National Bureau of Economic Research.

Ash, E. and Hansen, S. (2023). Text algorithms in economics. *Annual Review of Economics*, 15:659–688.

Babii, A., Ghysels, E., and and, J. S. (2022). Machine learning time series regressions with an application to nowcasting. *Journal of Business & Economic Statistics*, 40(3):1094–1106.

Balke, N. S. and Petersen, D. (2002). How well does the beige book reflect economic activity? evaluating qualitative information quantitatively. *Journal of Money, Credit and Banking*, 34(1):114–136.

Bauer, M. D., Pflueger, C., and Sunderam, A. (2022). Perceptions about monetary policy. Working Paper 30480, National Bureau of Economic Research.

Bauer, M. D. and Swanson, E. T. (2023a). An Alternative Explanation for the "Fed Information Effect". *American Economic Review*, 113(3):664–700.

Bauer, M. D. and Swanson, E. T. (2023b). A reassessment of monetary policy surprises and high-frequency identification. *NBER Macroeconomics Annual*, 37(1):87–155.

Beltagy, I., Peters, M. E., and Cohan, A. (2020). Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Bernanke, B. (2024). Forecasting for monetary policy making and communication at the Bank of England: a review. *Bank of England Independent Evaluation Office*.

Blackwell, D. (1953). Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24(2):265–272.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Borup, D. and Schütte, E. C. M. (2022). In Search of a Job: Forecasting Employment Growth Using Google Trends. *Journal of Business & Economic Statistics*, 40(1):186–200.

Braun, R., Miranda-Agrippino, S., and Saha, T. (2025). Measuring monetary policy in the UK: The UK monetary policy event-study database. *Journal of Monetary Economics*, 149.

Buckmann, M., Joseph, A., and Robertson, H. (2022). An interpretable machine learning workflow with an application to economic forecasting. Technical report, Bank of England.

Bybee, L., Kelly, B. T., Manela, A., and Xiu, D. (2020). The structure of economic news. Working Paper 26648, National Bureau of Economic Research.

Cabrales, A., Gossner, O., and Serrano, R. (2013). Entropy and the value of information for investors. *American Economic Review*, 103(1):360–77.

Carriero, A., Pettenuzzo, D., and Shekhar, S. (2025). Macroeconomic forecasting with large language models. *arXiv preprint arXiv:2407.00890*.

Cesa-Bianchi, A., Thwaites, G., and Vicondoa, A. (2020). Monetary policy transmission in the United Kingdom: A high frequency identification approach. *European Economic Review*, 123.

Chinco, A., Clark-Joseph, A. D., and Ye, M. (2019). Sparse signals in the cross-section of returns. *The Journal of Finance*, 74(1):449–492.

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

Cottier, B., Rahman, R., Fattorini, L., Maslej, N., Besiroglu, T., and Owen, D. (2025). The rising costs of training frontier AI models. *arXiv preprint arXiv:2405.21015*.

Coulombe, P. G., Marcellino, M., and Stevanović, D. (2021). Can machine learning catch the Covid-19 recession? *National Institute Economic Review*, 256:71–109.

Dasgupta, S. and Gupta, A. (2003). An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65.

Dell, M. (2025). Deep learning for economists. *Journal of Economic Literature*, 63(1):5–58.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dhal, P. and Azad, C. (2022). A comprehensive survey on feature selection in the various fields of machine learning. *Applied Intelligence*, 52(4):4543–4581.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263.

Ding, T., Chen, T., Zhu, H., Jiang, J., Zhong, Y., Zhou, J., Wang, G., Zhu, Z., Zharkov, I., and Liang, L. (2023). The efficiency spectrum of large language models: An algorithmic survey. *arXiv preprint arXiv:2312.00678*.

Eisfeldt, A. L. and Schubert, G. (2024). AI and Finance. Working Paper 33076, National Bureau of Economic Research.

Ellingsen, J., Larsen, V. H., and Thorsrud, L. A. (2022). News media versus FRED-MD for macroeconomic forecasting. *Journal of Applied Econometrics*, 37(1):63–81.

Farboodi, M. and Veldkamp, L. (2023). Data and Markets. *Annual Review of Economics*, 15:23–40.

Favara, G., Gilchrist, S., Lewis, K. F., and Zakrajšek, E. (2016). Updating the recession risk and the excess bond premium. Technical report, Board of Governors of the Federal Reserve System.

Feuerriegel, S. and Gordon, J. (2018). Long-term stock index forecasting based on text mining of regulatory disclosures. *Decision Support Systems*, 112:88–97.

Filippou, I., Garciga, C., Mitchell, J., and Nguyen, M. T. (2024). Regional Economic Sentiment: Constructing Quantitative Estimates from the Beige Book and Testing Their Ability to Forecast Recessions. Technical report, Federal Reserve Bank of Cleveland.

Frankel, A. and Kamenica, E. (2019). Quantifying information and uncertainty. *American Economic Review*, 109(10):3650–80.

Gentzkow, M., Kelly, B., and Taddy, M. (2019). Text as data. *Journal of Economic Literature*, 57(3):535–74.

Gertler, M. and Karadi, P. (2015). Monetary policy surprises, credit costs, and economic activity. *American Economic Journal: Macroeconomics*, 7(1):44–76.

Gharbawi, M., Ward, E., Bratt, E., Diver, L., Mueller, H., Quartu, R., and Robinson, H. (2024). Artificial intelligence in UK financial services. Technical report, Bank of England.

Giannone, D., Lenza, M., and Primiceri, G. E. (2021). Economic predictions with big data: The illusion of sparsity. *Econometrica*, 89(5):2409–2437.

Giglio, S., Kelly, B., and Xiu, D. (2022). Factor models, machine learning, and asset pricing. *Annual Review of Financial Economics*, 14:337–368.

Gilchrist, S. and Zakrajšek, E. (2012). Credit spreads and business cycle fluctuations. *American Economic Review*, 102(4):1692–1720.

Goebel, B., Dawy, Z., Hagenauer, J., and Mueller, J. (2005). An approximation to the distribution of finite sample size mutual information estimates. In *IEEE International Conference on Communications, 2005. ICC 2005. 2005*, volume 2, pages 1102–1106.

Gu, S., Kelly, B., and Xiu, D. (2020). Empirical Asset Pricing via Machine Learning. *The Review of Financial Studies*, 33(5):2223–2273.

Gu, S., Kelly, B., and Xiu, D. (2021). Autoencoder asset pricing models. *Journal of Econometrics*, 222(1, Part B):429–450.

Guhaniyogi, R. and Dunson, D. B. (2015). Bayesian compressed regression. *Journal of the American Statistical Association*, 110(512):1500–1514.

Gürkaynak, R. S., Sack, B. P., and Swanson, E. T. (2005). Do actions speak louder than words? the response of asset prices to monetary policy actions and statements. *International Journal of Central Banking (May 2005).*

Hansen, S., Lambert, P. J., Bloom, N., Davis, S. J., Sadun, R., and Taska, B. (2023). Remote work across jobs, companies, and space. Working Paper 31007, National Bureau of Economic Research.

Hansen, S., McMahon, M., and Tong, M. (2019). The long-run information effect of central bank communication. *Journal of Monetary Economics*, 108:185–202.

Harbert, T. (2021). Tapping the power of unstructured data. Technical report, MIT Sloan.

Hoberg, G. and Manela, A. (2025). The natural language of finance. Technical report, USC Marshall School of Business.

Huang, A. H., Wang, H., and Yang, Y. (2023). FinBERT: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 40(2):806–841.

Hutter, M. (2001). Distribution of mutual information. *Advances in neural information processing systems*, 14.

Hutto, C. and Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Imbens, G. W. (2024). Causal inference in the social sciences. *Annual Review of Statistics and Its Application*, 11.

Jarociński, M. and Karadi, P. (2020). Deconstructing monetary policy surprises—the role of information shocks. *American Economic Journal: Macroeconomics*, 12(2):1–43.

Johnson, W. B. and Lindenstrauss, J. (1984). Extensions of Lipschitz mappings into a Hilbert space. *Contemporary mathematics*, 26:28.

Jégou, H., Douze, M., and Schmid, C. (2011). Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):117–128.

Kadan, O. and Manela, A. (2018). Estimating the Value of Information. *The Review of Financial Studies*, 32(3):951–991.

Kalamara, E., Turrell, A., Redl, C., Kapetanios, G., and Kapadia, S. (2022). Making text count: Economic forecasting using newspaper text. *Journal of Applied Econometrics*, 37(5):896–919.

Kaminska, I. and Mumtaz, H. (2022). Monetary policy transmission during qe times: role of expectations and term premia channels. Technical report, Bank of England.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Ke, Z. T., Kelly, B. T., and Xiu, D. (2019). Predicting returns with text data. Working Paper 26186, National Bureau of Economic Research.

Kelly, B., Manela, A., and Moreira, A. (2021). Text selection. *Journal of Business & Economic Statistics*, 39(4):859–879.

Kelly, B. and Pruitt, S. (2015). The three-pass regression filter: A new approach to forecasting using many predictors. *Journal of Econometrics*, 186(2):294–316.

Kleinberg, J., Ludwig, J., Mullainathan, S., and Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review*, 105(5):491–495.

Koop, G. and Korobilis, D. (2012). Forecasting inflation using dynamic model averaging. *International Economic Review*, 53(3):867–886.

Korinek, A. (2023). Generative AI for economic research: Use cases and implications for economists. *Journal of Economic Literature*, 61(4):1281–1317.

Korinek, A. (2024). Generative AI for Economic Research: LLMs Learn to Collaborate and Reason. Working Paper 33198, National Bureau of Economic Research.

Land, S. K. (1977). Adam Smith's "Considerations Concerning the First Formation of Languages". *Journal of the History of Ideas*, 38(4):677–690.

Larsen, V. H. and Thorsrud, L. A. (2019). The value of news for economic developments. *Journal of Econometrics*, 210(1):203–218.

Larsen, V. H., Thorsrud, L. A., and Zhulanova, J. (2021). News-driven inflation expectations and information rigidities. *Journal of Monetary Economics*, 117:507–520.

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196.

Li, D., Plagborg-Møller, M., and Wolf, C. K. (2024). Local projections vs. vars: Lessons from thousands of dgps. *Journal of Econometrics*, 244(2).

Loria, S., Keen, P., Honnibal, M., and Yankovsky, R. (2013). Textblob: Simplified text processing.

Loughran, T. and McDonald, B. (2015). The use of word lists in textual analysis. *Journal of Behavioral Finance*, 16(1):1–11.

Ludwig, J., Mullainathan, S., and Rambachan, A. (2025). Large language models: An applied econometric framework. Working Paper 33344, National Bureau of Economic Research.

Manela, A. and Moreira, A. (2017). News implied volatility and disaster concerns. *Journal of Financial Economics*, 123(1):137–162.

Mann, C. L. (2023). Expectations, lags, and the transmission of monetary policy. *Speech given at the Resolution Foundation*.

McCracken, M. W. and Ng, S. (2016). FRED-MD: A Monthly Database for Macroeconomic Research. *Journal of Business & Economic Statistics*, 34(4):574–589.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In Bengio, Y. and LeCun, Y., editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., and Joulin, A. (2018). Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Miranda-Agrippino, S. (2016). Unsurprising shocks: information, premia, and the monetary transmission. Technical report, Bank of England.

Miranda-Agrippino, S. and Ricco, G. (2021). The transmission of monetary policy shocks. *American Economic Journal: Macroeconomics*, 13(3):74–107.

Nakamura, E. and Steinsson, J. (2018). High-frequency identification of monetary non-neutrality: the information effect. *The Quarterly Journal of Economics*, 133(3):1283–1330.

Nozaki, Y., Nakashima, D., Sato, R., and Asaba, N. (2025). Efficient vocabulary reduction for small language models. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 771–783.

Nygaard, K. B. (2020). 1970 Commercial Paper Market Liquidity Crisis. *Journal of Financial Crises*, 2(3):101–115.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 27(8):1226–1238.

Plagborg-Møller, M. and Wolf, C. K. (2021). Local projections and VARs estimate the same impulse responses. *Econometrica*, 89(2):955–980.

Raftery, A. E., Kárný, M., and Ettler, P. (2010). Online prediction under model uncertainty via dynamic model averaging: Application to a cold rolling mill. *Technometrics*, 52(1):52–66.

Ramey, V. A. (2016). Macroeconomic shocks and their propagation. *Handbook of macroeconomics*, 2:71–162.

Reeves, R. and Sawicki, M. (2007). Do financial markets react to bank of england communication? *European Journal of Political Economy*, 23(1):207–227.

Ritchie, H., Rosado, P., and Roser, M. (2023). Energy. `https://ourworldindata.org/grapher/carbon-intensity-electricity`.

Romer, C. D. and Romer, D. H. (2000). Federal reserve information and the behavior of interest rates. *American Economic Review*, 90(3):429–457.

Sadique, S., In, F., Veeraraghavan, M., and Wachtel, P. (2013). Soft information and economic activity: Evidence from the Beige Book. *Journal of Macroeconomics*, 37:81–92.

Sarkar, S. K. and Vafa, K. (2024). Lookahead bias in pretrained language models. *Available at SSRN*.

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., et al. (2022). Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Schmitt-Grohé, S. and Uribe, M. (2024). Central bank information or neo-fisher effect? Working Paper 33136, National Bureau of Economic Research.

Shannon, S. C. (1978). Work stoppage in government: the postal strike of 1970. *Monthly Labor Review*, 101(7):14–22.

Stock, J. H. and Watson, M. W. (2018). Identification and Estimation of Dynamic Causal Effects in Macroeconomics Using External Instruments. *The Economic Journal*, 128(610):917–948.

Taddy, M. (2013). Multinomial Inverse Regression for Text Analysis. *Journal of the American Statistical Association*, 108(503):755–770.

Taddy, M. (2015). Distributed multinomial regression. *The Annals of Applied Statistics*, 9(3):1394–1414.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3):1139–1168.

Thorsrud, L. A. (2020). Words are the new numbers: A newsy coincident index of the business cycle. *Journal of Business & Economic Statistics*, 38(2):393–409.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Watson, M. W. (2023). Comment. *NBER Macroeconomics Annual*, 37:161–166.

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., and Mann, G. (2023). BloombergGPT: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Yang, J., Wang, Z., Lin, Y., and Zhao, Z. (2024). Problematic tokens: Tokenizer bias in large language models. In *2024 IEEE International Conference on Big Data (BigData)*, pages 6387–6393. IEEE.

Yousuf, K. and Ng, S. (2021). Boosting high dimensional predictive regressions with time varying parameters. *Journal of Econometrics*, 224(1):60–87.

Zaffalon, M. and Hutter, M. (2002). Robust feature selection by mutual information distributions. *arXiv preprint cs/0206006*.

Zavodny, M. and Ginther, D. K. (2005). Does the Beige Book move financial markets? *Southern Economic Journal*, 72(1):138–151.