Text as Instruments*
(Job Market Paper)

James Nesbit†

June 17, 2020
Click here for latest version

This paper provides a theoretical framework to justify and guide the use of text data in the estimation of quantitative economic models. Previous work utilizing text data has implicitly assumed that the text and traditional data are driven by common latent variables (e.g. news shocks). This link is has been used informally in previous work to infer features of latent variables from text data. In contrast, this paper introduces a model that formalizes an explicit link between text data and the latent variables in the econometric model, and therefore justifies formally and guides the use of text data to augment existing econometric methods. To do so, we develop a random utility framework in which speakers choose a word given observables—notably other words in the document that define that word's "context". This allows us to find instruments for a target latent variable in the presence of other confounding latent variables. We show that under weak sufficient conditions the estimated utilities from this model or the log odds ratio can be used as instruments, whereas raw word counts cannot be used as instruments. We derive rate conditions such that the first stage estimation of the text instruments does not distort second-stage inference, allowing our instruments to be used without needing to adjust standard errors. We use our results to quantify the effects of contemporaneous news shocks to asset returns, controlling for stale news. We also identify monetary policy shocks, in the presence of other macroeconomic shocks, using FOMC transcripts.

Keywords: Machine Learning, Text Analysis, News Shocks, Monetary Policy Shocks.

*I am very grateful to my advisors Timothy Christensen and José Luis Montiel Olea for their support. I would also like to thank David Childers, Jean-Jacques Forneron, Alfred Galichon, Elena Manresa, Konrad Menzel, Anna Mikusheva, Mikkel Plagborg-Møller, Serena Ng, Isabelle Perrigne, Alex Torgovitsky, Quang Vuong, and seminar participants at NYU. The usual disclaimer applies.

†Department of Economics, New York University, NY 10003 (email:jmn425@nyu.edu).

1 Introduction

Text data is used in applied economics research to infer information that is otherwise unobservable using standard data sources. Newspaper articles have been used to infer economic policy uncertainty (Baker et al. (2016)) and climate change risk (Engle et al. (2019)); transcripts of Federal Open Market Committee (FOMC) meetings have been used to quantify the size of monetary policy shocks (Romer and Romer (2004)); and presidential speeches and Congressional reports have been used to separate exogenous from endogenous tax policy (Romer and Romer (2010)).

In order to utilize text data in traditional econometric models, previous work has *implicitly* assumed that quantitative data and text are driven by common latent exogenous variables. In the quantitative data, these latent variables are either completely unobserved, or observed with error, but the text allows features of these variables to be inferred. However, without a formal link between text and data, it is difficult to understand the properties of statistics generated from the text, particularly when there may exist multiple, potentially confounding, latent variables.

As a motivating example, consider the problem of quantifying the effect that contemporaneous news shocks have on asset returns. Assume that the return of an asset is a function of news shocks, the latent variables. The return of an asset may depend on both today's news and stale news—both contemporaneous and past news shocks. As an instrument (proxy) for news shocks, consider the sentiment of newspaper articles reporting on this asset. In order to predict asset returns, Ke et al. (2019) demonstrate that the term '*shortfall*' is a highly charged negative sentiment term, and thus frequent mentions of this term should indicate that there is a negative contemporaneous news shock. However given that current events are often still discussed in subsequent days, we would also imagine that '*shortfall*' could appear in discussion of past news.[1] Given that past news shocks are unobserved but also drive asset returns, this frequency of the term '*shortfall*' would be endogenous, and thus not a valid instrument.

In this paper, we introduce a model that formalizes an explicit link between text data and latent variables from a quantitative econometric model. We embed a random utility model of text choice into a standard simultaneous equations setup. In the text model, speakers choose terms from a finite vocabulary, conditional on covariates which include events within the same document. We call these conditioning events, a word's *context*, as it makes clear that the specific meaning of term is defined by "the company it keeps" (Firth (1957)). This construction gives sufficient flexibility to allow term-context pairs to have very precise

[1] Consider the following example about record high Apple share prices in 2020, (Feuer (2020)): "*If you look at our results, our shortfall is over 100 percent from iPhone and it's primarily in greater China," Apple CEO Tim Cook told CNBC's Josh Lipton in an interview in January 2019.*".

meanings, insofar as the utility of the pair may depend only our unobservable of interest.

The model of text is used to construct instruments; random variables that correlated with our target latent variable of interest, but uncorrelated with the other confounding, nuisance unobservables in the model. We show that it is sufficient to consider 'special' term-context pairs that are a function of the target latent variable, but not of the nuisance latent variables. Consider our running example of news shocks, where unconditionally 'shortfall' would fail to be exogenous as the utility of that term is a function of past news shocks. If we were to condition on the context that terms describing past tense—e.g. '*yesterday*' or '*last week*'—do not occur in that article, then we could argue that the choice of 'shortfall' in this context only depends on contemporaneous news shocks. We show that the log odds ratio of two such terms in the same context are valid instruments. In contrast, we find that the conditional probabilities on these special term-context pairs are not able to be used as instruments under similar conditions.

Estimation is a two stage procedure: instruments are estimated in the first stage used as a plug-in estimator in the second stage. The first stage is consistent and asymptotically normal and we derive a supremum large deviation bound. If the size of each document, N, grows sufficiently quickly relative to the time dimension T, $(T \log(T))/N \to 0$, then the first stage estimation has no effect on the second stage inference. As document size is often large relative to the time, allows applied researchers to use can use estimated instruments (log odds ratios) as data without requiring any adjustments to standard errors.

We present two empirical exercises. The first application is our motivating example, seeking to quantity the effects of contemporaneous news shocks on asset returns, controlling for stale news. We use 'sentiment charged words' from Ke et al. (2019)—positively charged words predict high asset returns and negatively charged words predict low asset returns. In this example, contemporaneous news shocks are the target latent variable, while stale news shocks are the nuisance latent variables. These highly sentiment charged words, like '*shortfall*', are used as instruments by conditioning on the context that discussion of the past, words like '*yesterday*', do not appear in the same article.

The second empirical application is to identify monetary policy in a monetary SVAR using transcripts from the Federal Open Market Committee (FOMC) meetings. The SVAR contains three variables, interest rates, output and inflation, and three structural shocks (our latent variables), a monetary policy shock, an output shock and an inflation shock. Our target shock is the monetary policy shock and the nuisance shocks are the output and inflation shock. The instrument we identify is discussion about monetary policy, that takes place not in the context of discussion about inflation or output.

The remainder of the paper is as follows. Section 2 discusses the related literature. Section 3 introduces the quantitative model and the text model. Section 4 demonstrates how the text

model can be used to find instruments. Section 5 discuss estimation and inference. Section 7 presents the first empirical application, identifying monetary policy shocks in SVAR using FOMC transcripts. Section 6 present the second empirical application, quantifying the effect of news shocks on asset returns. Section 8 concludes. The Appendix contains supplementary results and all proofs.

2 Related Literature

As discussed in the introduction, this paper seeks to formalize the use of text to infer information about latent variables from a quantitative economic model. The use of the 'narrative record' has a long history in macroeconomics, where it has been used to identify structural shocks, for example Friedman et al. (1963); Romer and Romer (1989, 2004, 2010). In these papers, the text is read by humans, who generate measures of target latent variables based on these readings. Humans are very capable of inferring latent information from text, but the resulting classifications are opaque, not replicable, and specific to the particular setting. In contrast, the method presented in this paper is transparent, the analysis is replicable and is able to used in a wide variety of settings.

Other recent work has used statistics derived from text, as opposed to human readings, to infer the state of unobserved variables, e.g. Baker et al. (2016); Engle et al. (2019). These papers include, either implicitly or explicitly, nuisance latent variables in the quantitative model, but do not account for the possibility that these nuisance latent variables may also be driving text choice. As a result, it is difficult to argue that these measures are actually exogenous, and thus may not be suitable as instruments or proxies. Our approach differs in that we construct a formal model where the text and quantitative model are driven by common latent variables. This allows us to make the relationship between conditional (on the context) probabilities and *all* latent variables explicit. With such a model in hand, we can have a clear map to searching for term-context pairs that can be used to identify the model.

The model of text that we develop generates conditional choice probability that take logit form. This form is commonly used in machine learning and statistics and is motivated as a generalized linear model with a logistic link, Taddy (2013); Taddy et al. (2015); Gentzkow et al. (2019b), or a neural language model with a softmax transform Bengio et al. (2003); Mnih and Hinton (2007). These models condition on standard covariates—speaker characteristics and time varying variables—as their goal is to model the probabilities of the entire vocabulary. In contrast, our model is motivated as a choice problem, which helps to make explicit the link between latent variables and word choice. As our goal is to model a small number of words of interest, this gives us the flexibility to condition of a richer range of events, in the particular the 'context', where we condition on events within the same

document.

Conditioning on context is commonly used in the natural language processing and machine learning literatures, Shannon (1948); Mikolov et al. (2013b,a). These models seek to model an entire vocabulary for a prediction task, and use generic conditioning events: n-gram models will condition on the previous $n-1$ words and word vector models will condition of the words that appear in a neighbourhood around a word. Our focus is instead on quantifying the casual link between latent variables and our outcome of interest. As a result, we do not seek to model the entire vocabulary, and can tailor the choice of context to tease out very specific meanings of words. Unlike these models, the choice of context is not at all generic, and will likely be idiosyncratic to the question being studied, and rely on domain expertise.[2]

Out second empirical application—the identification of monetary policy shocks using an external instrument—has been an active area of research for over 40 years. One area that the SVAR literature has struggled with is the 'price puzzle', the counter-intuitive increase in the inflation to an unexpected monetary policy shock, first recognized by Sims (1986). A common approach to solving the prize puzzle is to include extra variables (Sims (1992); Bernanke et al. (2005); Sims and Zha (2006)).[3] Our application is most similar to Gertler and Karadi (2015), who include financial variables, as well as standard macroeconomic variables, and use high frequency surprises around policy announcements as external instruments. This paper only uses standard macroeconomic variables and uses our generated instruments from text as external instruments.

Finally, our inference results connect to the literature on generated regressors and two step estimation, Pagan (1984); Murphy and Topel (1985); Newey and McFadden (1994). Wooldridge (2010) shows that when instruments are generated from a model with a finite dimensional parameter, that under certain moment conditions, estimation of the instrument does not effect second stage inference. Bai and Ng (2008) show first stage estimation of a factor model does not affect second stage inference. These results hold as long as $\sqrt{T}/N \to 0$.[4] Our paper derives similar results but with the rate condition that $T\log(T)/N \to 0$.

[2] We do not see this as a shortcoming of the model. This is common in the identification of much applied economic work. In do find however, that conditioning on the absence of particular terms (rather than the presence) is particularly useful for identification.

[3] There is a very large literature on the price puzzle, see Kilian and Lütkepohl (2017) Chapter 7.

[4] The rate $\sqrt{T}/N \to 0$ is for a linear model. For a non-linear model, the required rate condition is $T^{5/8}/N \to 0$.

3 Setup

This section will lay out the econometric environment. First, we will present our motivating model, a linear regression with an endogenous variable of interest. The departure from the standard IV framework, will be that exogenous variables to be used as instruments are only observed through word choices. This will motivate the model of text, where word choice is driven by these exogenous variables.

3.1 Motivating Model

Denote $y_t \in \mathbb{R}$ as the outcome variable, $x_t \in \mathbb{R}^{d_x}$ a vector of exogenous covariates, $Y_t \in \mathbb{R}$ an endogenous variable and unobservables $e_t \in \mathbb{R}$. Assume that

$$y_t = x_t'\beta_0 + Y_t\theta_0 + e_t,$$

where $\mathbb{E}[x_t'e_t] = 0$, but $\mathbb{E}[Y_te_t] \neq 0$, so that OLS can not deliver a consistent estimate of it. One way to manage to this, is to use an instrument. We assume the existence of such an exogenous variable z_t^*, that can be used as an instrument, but it cannot be directly observed. Instead it must be inferred from auxiliary text data.

3.2 Text Model

At each time period t, there exists a single document.[5] Each document contains N_t words, indexed by an integer $n \in \{1, \ldots, N_t\}$. A word with index n, denoted $w_{t,n}$, can be one of V terms in a vocabulary.[6] To be clear, 'word' refers to the index (or the choice in a discrete choice framework), whereas 'term' refers to which particular member of the vocabulary is chosen (the alternative).

The corpus will be modelled as a discrete choice problem where *speakers*, choose an independent sequence of terms. Speakers will make these choices conditional on covariates, which may include those from the econometric model, x_t, but may also include events within a document—e.g. the presence or absence of certain terms within a window of that particular word. We impose tis by assuming that the utility of the term-context pair depends on both the observables covariates (x_t) and the unobservables (z_t^* and e_t), from our primary specification. Additionally, the utility from choosing word w depends on other events within the same document—referred to as "context"—and that the response to z_t^* and e_t are word,

[5] If there exist a corpus of documents at time t, we will concatenate these into a single document.

[6] A term can be a single word—a unigram—or a combination of n words—an n-gram. It is also common for a number of 'preprocessing' steps that take place to transform a document into usable form. For details see Gentzkow et al. (2019a).

context specific.

The utility that a speaker receives from choosing term $w_{t,n}$, given context is

$$U_{w_{t,n},c} = \underbrace{x_t'\gamma_{w,c} + \alpha_{1,w,c}z_t^* + \alpha_{2,w,c}e_t}_{u_{w_{t,n},c}} + \epsilon_{w_{t,n},c}. \tag{1}$$

The random utility shock $\epsilon_{w_{t,n},c}$ is assumed to be Gumbel (appropriately centred) and is independent across n, w and t. This leads to the familiar result from random utility theory that the probability of choosing term w in document t, conditional on context c is

$$\begin{aligned} p_t(w,c) &:= p_t(w|c, x_t, z_t^*, e_t) \\ &= \frac{\exp\{x_t'\gamma_{w,c} + \alpha_{1,w,c}z_t^* + \alpha_{2,w,c}e_t\}}{\sum_{w'=1}^{V}\exp\{x_t'\gamma_{w,c} + \alpha_{1,w',c}z_t^* + \alpha_{2,w',c}e_t\}}. \end{aligned} \tag{2}$$

This model of text is quite different from others presented in econometrics and machine learning.[7] The object of interest of most NLP models is a function of the entire vocabulary. As a result, they model all the terms within a document. We are instead focusing on modelling terms conditional on a fixed context. Our choice is simplistic, but does capture that in certain contexts, a speaker faces different word choices, and will respond to information that is both observable (x_t) and unobservable (z_t^* and e_t) to the econometrician. Framing this as a choice problem, and having agents making choices in response to movement in these variables is crucial in using text to infer latent information.

4 Instruments

In this section we will show that estimated quantities from our model of text can be used as instruments. In order to find valid instruments we will need to assume the existence of at least 'special' term-context pairs that depend on z_t^*, but not e_t. We show that the log-odds ratio of two 'special' term-context pairs can be used as instruments. In contrast, we will show that under similar conditions, conditional choice probabilities cannot be used as instruments, where instruments are defined in the standard way

Definition 1. Instrument z_t for Y_t satisfies

1. Relevance $\mathbb{E}[z_t Y_t] \neq 0$
2. Exogeneity $\mathbb{E}[z_t e_t] = 0$

We will conclude the section by examining the EPU index of Baker et al. (2016) through the lens of our model.

[7] For a comprehensive list of NLP models used within economics see Gentzkow et al. (2019a).

4.1 Log Odds

Assumption 1. There exist two terms w_1^* and w_2^* in a shared context c^* such that

1. $\alpha_{1,w,c*} \neq 0$ for $w = w_1^*$ **or** $w = w_2^*$.
2. $\alpha_{2,w,c*} = 0$ for $w = w_1^*$ **and** $w = w_2^*$.

Assumption 1.1 states that there exist *two* words w_1^* and w_2^*, in the same same context c^*, and at least one of one them depends on the target latent variable. Assumption 1.1 states that both w_1^* and w_2^* in context c^* do not depend on the nuisance shock.

Example 1. Consider our example from the introduction. We were seeking to find an instrument for contemporaneous news shocks, in the presence of past or stale news. Here our target latent exogenous variable is the contemporaneous news shock and the nuisance exogenous variables are the stale news shocks. In addition to the the term '*shortfall*', that we considered earlier, consider another highly sentiment charged word '*downgrade*'. We argued that these sentiment terms were likely relevance, but without context would likely fail to be exogenous (also being related the stale news shocks). This is because '*shortfall*' or '*downgrade*' could be discussed in the context events that happened the previous day or several days in the past.[8] A solution to this problem is to condition on the context that words relating to the past (like '*yesterday*', '*the day before*', '*last week*' or days of the week or months prior to the day the article was published) do not occur within a small window around '*shortfall*' or '*downgrade*'.

Define the *log odds ratio* of w on w' in context c as

$$z_t(w, w', c) := \log\left(\frac{p_t(w, c)}{p_t(w', c)}\right). \tag{3}$$

The log odds ratio is equal to the difference in utilities between choosing w and w' in context c at time t. Under Assumption 1 and Equation (2), this implies that

$$\begin{aligned} z_t(w, w', c) &= u_{w,c} - u_{w',c} && (4)\\ &= x_t'(\gamma_{w,c} - \gamma_{w',c}) + z_t^*(\alpha_{1,w,c} - \alpha_{1,w',c}). && (5) \end{aligned}$$

The log odds ratio is simply a linear combination of z_t^* and x_t. As both are uncorrelated to e_t by assumption, and z_t^* can be used as an instrument for Y_t, we have the following result.

Proposition 1. Let w_1^*, w_2^*, c^* satisfy Assumption 1, and suppose $\alpha_{1,w_1^*,c^*} \neq \alpha_{1,w^*,c^*}$. Then $z_t(w_1^*, w_2^*, c^*)$ is an instrument for Y_t.

$z_t(w_1^*, w_2^*, c^*)$ is instrument for Y_t because it is difference in utilities between choosing w

[8] Refer to Footnote 1 for an example.

and w' in context c at time t. This implies that we could use the utilities themselves as instruments.

There are in theory multiple pairs of terms (in a fixed context) that satisfy Assumption 1. This is particularly true as there may be different words that convey the same meaning in a given context. Similarly there are multiple possible configurations of log odds ratios such that Proposition 1 and can be used as instruments.[9] We consider the following generalization of the log odds ratio. Define a subset of the vocabulary, that will be our candidate instruments, $\mathcal{J} = \{1, \ldots, V\}$. We will generalize Assumption 1 to account for more than two terms.

Assumption 2. Suppose $|\mathcal{J}| > 2$, and

1. $\alpha_{1,w_j,c^*} \neq 0$ for at least one $j \in \mathcal{J}$.
2. $\alpha_{2,w_j,c^*} = 0$ for all $j \in \mathcal{J}$.

Assumption 1.1 states that least one of the terms in $\mathcal{J}$ must depend on the target latent variable. Assumption 1.2 states that all of the terms in $\mathcal{J}$ must not depend on the nuisance variable.

Additionally define a vector of weights, $\boldsymbol{\omega} = \omega_1, \ldots, \omega_{|\mathcal{J}|}$, $\omega_j \in \mathbb{R}$, and

$$z_t(\mathcal{J}, c, \boldsymbol{\omega}) := \sum_{j \in \mathcal{J}} \omega_j \log(p_t(w_j, c)). \tag{6}$$

One can recover the log odds ratio of w_i on w_j by setting $\omega_i = 1$, $\omega_j = -1$ and $\omega_{j'} = 0$ for all $j' \in \mathcal{J} \backslash \{i, j\}$. Note that if we express $p_t(w_j, c) = \exp(u_{w_j,c}) / \sum_{j'=1}^{V} \exp\{u_{w_{j'},c}\}$, then Equation (6) can be written

$$z_t(\mathcal{J}, \boldsymbol{\omega}, c) = \sum_{j \in \mathcal{J}} \omega_j u_{w_j,c} - \sum_{j \in \mathcal{J}} \omega_j \log \left(\sum_{j'=1}^{V} \exp\{u_{w_{j'},c}\} \right). \tag{7}$$

The *generalized log odds* is defined as $z_t(\mathcal{J}, c, \boldsymbol{\omega})$, such that the weights sum to 0, eliminating the troublesome denominator term $\log \left(\sum_{j'=1}^{V} \exp\{u_{w_{j'},c}\} \right)$.

Proposition 2. Let $\mathcal{J}^*$ in context c^*, satisfy Assumption 2. For any $\boldsymbol{\omega}$ such that $\sum_{j \in \mathcal{J}} \omega_j = 0$, $z_t(\mathcal{J}^*, \boldsymbol{\omega}, c^*)$ is an instrument for Y_t.

The optimal choice of $\boldsymbol{\omega}$ will be examined in Section 5.

4.2 Conditional Probabilities

The results of Proposition 2 demonstrate that under the weak conditions of Assumption 1, we can find valid instruments using the log odds ratio of two special words in a shared

[9] For example $\log(p_t(w_1, c)/p_t(w_2, c))$ and $\log(p_t(w_2, c)/p_t(w_1, c))$ are valid instruments.

context. Can we derive a similar result, under those same assumptions (adapted for a single word), for conditional probabilities?

Proposition 3. Suppose w^* satisfies $\alpha_{1,w^*,c^*} \neq 0$ and $\alpha_{2,w^*,c^*} = 0$. Under these assumptions $p_t(w^*, c^*)$ is an not instrument for Y_t.

Consider the form of the conditional probability of w^* in c^* under the assumptions in Proposition 3

$$\begin{aligned} p_t(w^*, c^*) &:= p_t(w|c, x_t, z_t^*, e_t) \\ &= \frac{\exp\{x_t'\gamma_{w,c} + \alpha_{1,w,c}z_t^*\}}{\exp\{x_t'\gamma_{w,c} + \alpha_{1,w,c}z_t^*\} + \sum_{w' = w^*} \exp\{x_t'\gamma_{w,c} + \alpha_{1,w,c}z_t^* + \alpha_{2,w,c}e_t\}}. \end{aligned} \tag{8}$$

The conditional probabilities are not valid instruments, because they fail to satisfy the exogeneity requirement (although they do satisfy relevance), due to the presence of e_t in the denominator of Equation (8). This is not merely as artifact of the logit form of conditional probability, this represents a tradeoff for speakers.[10]

Consider the choosing between two terms in context c^*, w^* and w'. Suppose that w^* is our special word, relevant to Y_t, $\alpha_{1,w^*,c} > 0$, and exogenous to e_t, $\alpha_{2,w^*,c} = 0$; and w' is the opposite, $\alpha_{1,w',c} = 0$, $\alpha_{2,w',c^*} > 0$. Suppose that at time t, we have $Y_t = 0$ and $e_t < 0$. While the utility from speaking w^* is unaffected by e_t, the utility from w' falls. In relative terms, the speaker gets more utility from uttering w^* and hence $p_t(w^*, c)$ is higher.

This insight gives us a path to solving the endogeneity problem that conditional probabilities face. If place stronger assumptions on words appearing in the same context, we will be able to use conditional probabilities as instruments.

Assumption 3. There exists a term w^* and a context such that

1. $\alpha_{2,w,c^*} = 0$ for all w.
2. $|\alpha_{1,w^*,c^*}| > |\alpha_{1,w',c^*}|$, for all $w' \neq w^*$.
3. $\mathbb{E}[Y_t|e_t] = 0$.

Assumption 3.1 states that no word in the context c^* can be related to the nuisance exogenous variables, which will allow us to remove the effect e_t of the denominator term of $p_w(w^*, c^*)$. Assumption 3.2 assumes that w^* is responded more strongly to Y_t than any other word. This is needed to show $p_t(w^*, c^*)$ is strictly increasing in Y_t, which is needed to show relevance. Assumption 3.3 is a strengthening of the uncorrelated assumption $\mathbb{E}[Y_t e_t] = 0$. This is required to show exogeneity as $p_t(w^*, c^*)$ is a nonlinear function of Y_t.

[10] The endogeneity problem of conditional probabilities depends on the extent to which the denominator in Equation (8) varies with e_t. Under certain, this term, known as the *log partition function* can be well approximated by a constant (known as *self-normalization*, see Andreas and Klein (2015)).

Proposition 4. Let Assumption 3 hold. Then $p_t(w^*, c^*)$ is an instrument for Y_t.

4.3 EPU Index

In this section, we will examine the Economic Policy Uncertainty Index of Baker et al. (2016) through the lens of our model. The EPU index is constructed as the monthly count of articles that contain the following triple: '*uncertainty*' or '*uncertain*'; '*economic*' or '*economy*'; and one of the following policy terms: '*congress*', '*deficit*', '*Federal Reserve*', '*legislation*', '*regulation*' or '*white house*' (including variants like '*uncertainties*', '*regulatory*' or '*the Fed*').

BBD use the resulting time series in a SVAR that contains (in order): EPU index, S&P 500 index, federal funds rate, employment and industrial production. A Cholesky factorization is used to identify the structural impulse response functions. The Cholesky factorization and the ordering of the variable implies that the EPU index responses only to policy uncertainty shocks.

To see this let's consider a simplification of their SVAR model. Assume: i) that we have only two variables, the EPU index and industrial production (in that order); ii) the two structural shocks are a policy uncertainty shock $\epsilon_{1,t}$ and a output shock $\epsilon_{2,t}$; iii) we use a Cholesky factorization to identify the structural shocks; iv) the model is a VAR(0). Therefore we have

$$\begin{aligned} y_t &= B\epsilon_t \\ \begin{bmatrix} EPU_t \\ IP_t \end{bmatrix} &= \begin{bmatrix} B_{1,1}\epsilon_{1,t} & B_{1,2}\epsilon_{2,t} \\ B_{2,1}\epsilon_{1,t} & B_{2,2}\epsilon_{2,t} \end{bmatrix}. \end{aligned}$$

Note that $B_{1,2} = 0$, as a result of imposing that the identifying assumption that B is the Cholesky factorization of the covariance matrix of the reduced form residuals. This implies $EPU_t = B_{1,1}\epsilon_{1,t}$; the constructed EPU measure is a function of only the policy uncertainty shock, $\epsilon_{1,t}$.

One way to implement BBD's identifying assumptions into our model would be to set $w_1^* =$ '*uncertainty*' and c^* = '*economy*' and at least one of '*congress*', '*deficit*','*Federal Reserve*', ... appear in document t. Note that in this case, the EPU measure would be a conditional probability, so we will need to consider other words mentioned in this context. For example, we would likely expect to see the term '*interest rates*' being discussed in the context of '*Federal Reserve*' and '*economy*'. We also know that interest rates respond to fluctuations in output (due to Taylor rule for example), so we expect this discussion to be influenced by the output shock. Hence $\alpha_{2,\text{'interest rates'},c*} \neq 0$. The EPU measure is hence a function of the output shock, implying $B_{1,2} \neq 0$, which violates the Cholesky identification assumption.

This could be very easily remedied by constructing the log odds ratio of w_1^* = '*uncertainty*'

and w_2 = '*uncertain*' in the context c^*. This would remedy the problem of conditional probabilities. This doesn't address the possibility that w_1^* and w_2^* are still related to the output shock in the context c^*.

5 Estimation and Inference

In this section, we will discuss our estimation procedure, as well as how to perform inference. We will use GMM as our framework and perform estimation in 2 steps. In the first step, at each time t, we estimate the generalized log odds using the corpus of documents published at that time. This estimator will be used as a plug-in for instruments in the second stage. Throughout this section we will assume for notational simplicity that $N_t = N$ for all t.

5.1 First stage

In this section we will provide estimators for the generalized odds ratio and show that our estimator is consistent and asymptotically normal at a $\sqrt{N}$ rate. In addition, we will provide a bound on the worst case estimation error over the time dimension. These results will be used in the next section to show that the second stage estimation is unaffected by the plug-in estimators as long as we assume that N grows at a specific rate relative to T.

Given a context c, a set of words $\mathcal{J}$ and weights $\boldsymbol{\omega}$ s.t. $\sum_{j\in\mathcal{J}} \omega_j = 0$, the generalized log odds estimator at time t is

$$\begin{aligned}\hat{z}_t(\mathcal{J}, c, \boldsymbol{\omega}) &= \sum_{j\in\mathcal{J}} \omega_j \log\left(\frac{1}{N}\sum_{w=1}^{N} \mathbf{1}_{w_{t,n}=w_j}\mathbf{1}_{c_{t,n}=c}\right) \\ &= \sum_{j\in\mathcal{J}} \omega_j \log\left(\frac{1}{N}\#(w_{t,n}=w_j, c_{t,n}=c)\right), \end{aligned} \tag{9}$$

where $\#(x)$ is a count of the number of words in a document satisfying condition x.

Proposition 5. Suppose that $\mathcal{J}^*$ in context c^* is a set of valid instruments. For any (ω) such that $\sum_{j\in\mathcal{J}^*} \omega_j = 0$,

1. $\hat{z}_t(\mathcal{J}^*, c^*, \boldsymbol{\omega})$ is consistent,

$$\hat{z}_t(\mathcal{J}^*, c^*, \boldsymbol{\omega}) \xrightarrow{p} z_t(\mathcal{J}^*, c^*, \boldsymbol{\omega}).$$

2. $\hat{z}_t(\mathcal{J}^*, c^*, \boldsymbol{\omega})$ is asymptotically normal

$$\sqrt{N}(\hat{z}_t(\mathcal{J}^*, c^*, \boldsymbol{\omega}) - z_t(\mathcal{J}^*, c^*, \boldsymbol{\omega})) \xrightarrow{d} N(0, G_t),$$

where the asymptotic variance is

$$G_t = \sum_{j \in \mathcal{J}} \frac{\omega_j^2}{p_t(w_j, c)}.$$

In the next section, we will show that we can use estimated instruments as a plug-in estimator, with affecting the asymptotic variance of the second stage, provided that the size of documents N, grows faster than T. The exact rate of at which N grows relate to T is governed by the tail probability of the worst case (over t) estimation error. Define the maximum weight given to a log conditional probability as $\omega_M = \max_{j \in \mathcal{J}}$, define L, as the lower bound of the conditional probabilities on all $j \in \mathcal{J}$.[11] Then we have the following large deviation bound

Proposition 6. For any $\mathcal{J}$, c and $\boldsymbol{\omega}$, such that $\omega_M < \infty$,

1. $P\left(\max_t |\hat{z}_t(\mathcal{J}, c, \boldsymbol{\omega}) - z_t(\mathcal{J}, c, \boldsymbol{\omega})| > \epsilon\right) \leqslant T2|\mathcal{J}| \exp(-N\epsilon^2 L^2/(2|\mathcal{J}|^2 \omega_M^2))$.
2. $\max_t |\hat{z}_t(\mathcal{J}, c, \boldsymbol{\omega}) - z_t(\mathcal{J}, c, \boldsymbol{\omega})| = O_p\left(\sqrt{\frac{\log(T)}{N}}\right)$.

5.2 Second stage

The second stage will be formulated in the GMM framework. Define the data $\mathbf{w} = (y_t, x_t, z_t)$, and the population moments $g : \mathbb{R}^{dim(w)} \times \Theta \to \mathbb{R}^K$, and suppose they are jointly measurable in data $\mathbf{w}_t$. Define $\theta_0 \in \Theta$ as the parameter value that uniquely sets the population moments to zero, $\mathbb{E}[g(\mathbf{w}_t; \theta_0)] = 0$. We will denote z_t as the true data, and $\hat{z}_t$ as the first stage estimate.

The sample objective function is

$$Q_T(\theta, z) = -\frac{1}{2} g_T(\theta, z)' \hat{W} g_T(\theta, z), \text{ where } g_T(\theta, z) = \frac{1}{T} \sum_{t=1}^{T} g(y_t, x_t, z; \theta),$$

where $\hat{W}$ is a $K \times K$ positive definite symmetric weighting matrix that may be a function of the data, and we make the dependence on z explicit.

If we were to observe z_t, the *infeasible* GMM estimator, $\tilde{\theta}$ would be the argmax of the sample objective function, that is

$$\tilde{\theta} = \arg\max_{\theta} Q_T(\theta, z_t).$$

[11] That is $L \geqslant 0$ is such that $L < \min_{j \in \mathcal{J}}\{p(w_j, c)\}$. Given that the conditional probabilities are derived from logit discrete choice model, $L > 0$.

Instead if we do not observe z_t, but only an estimate from our first stage, $\hat{z}_t$, the *feasible* GMM estimator is

$$\hat{\theta} = \arg\max_{\theta} Q_T(\theta, \hat{z}_t).$$

Under standard conditions (Newey and McFadden (1994)), the infeasible estimator is consistent, $\tilde{\theta} \xrightarrow{p} \theta_0$.[12] In order to achieve consistency for the feasible GMM estimator, $\hat{\theta}$, we require an additional assumption.

Assumption 4.

$$\sup_{\theta \in \Theta, z \in N_{\hat{z}_t}} \| \frac{1}{T} \sum_{t=1}^{T} \frac{\partial g(y_t, x_t, z, \theta)}{\partial z} \| = O_p(1),$$

where N_{z_t} is an open neighborhood of z_t.

Assumption 4 requires that the norm of the first derivative of the moment function be bounded uniformly in θ and uniformly in z_t in an open neighborhood of $\hat{z}_t$. This ensures that the feasible sample objective function $Q(\theta, \hat{z}_t)$ is uniformly close to infeasible objective function $Q(\theta, z_t)$.

Proposition 7. Let Assumption 4 and Assumption 7 in Section A.2.1 hold. Additional assume $\sqrt{\frac{T \log(T)}{N}} \to 0$. Then $\hat{\theta}$ is consistent

$$\hat{\theta} \xrightarrow{p} \theta_0.$$

The rate condition $\sqrt{\frac{T \log(T)}{N}} \to 0$, ensures that the maximum deviation of $\hat{z}_t$ from z_t over time, which we established in Proposition 6, vanishes when $T \to \infty$.

The proof involves first showing that $g_T(\theta, \hat{z}_t)$ is uniformly (in θ) close to $g(\theta, z_t)$, then that the feasible sample objective function $Q(\theta, \hat{z}_t)$ is uniformly close to infeasible objective function $Q(\theta, z_t)$. Then proof then follows from standard arguments.

In order to show asymptotic normality we will require an additional assumption.

Assumption 5.

$$\sup_{z \in N_{z_t}} \| \frac{1}{T} \sum_{t=1}^{T} \frac{\partial^2 g_j(y_t, x_t, z, \hat{\theta})}{\partial \theta \partial z'} \| = O_p(1),$$

where N_{z_t} is an open neighbourhood around z_t.

[12] These assumptions are made explicit in Assumption 7 in Section A.2.1.

Assumption 5 requires that second derivative with respect to θ and z_t is bounded uniformly in θ. This assumption (together with our rate assumption and Proposition 6) ensures that the sample average derivatives of the moment conditions, evaluated under $\hat{z}_t$ converge in probability to those evaluated under z_t.

Proposition 8. Under Assumption 5, and $\sqrt{\frac{T\log(T)}{N}} \to 0$,

$$\sqrt{T}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, (G'WG)^{-1}G'WSWG(G'WG)^{-1}).$$

This is the same asymptotic distribution as $\sqrt{T}(\tilde{\theta} - \theta_0)$.

In the appendix, we show that we can consistently estimate the asymptotic variance using our plug-in estimator for our instruments.

Both Assumptions 4 and 5 are easy satisfy in the linear models that will be used in the empirical applications, requiring the just the existence of first moments.

6 News Shocks

In this section, we will present our first empirical application, quantifying the effect of news shocks on equity returns. Denote $y_t \in \mathbb{R}$ as the return of a company at date t, $x_t \in \mathbb{R}^{d_x}$ a vector of covariates, and $\epsilon_t \in \mathbb{R}$ the time t news shock. Suppose that

$$y_t = x_t'\beta + \sum_{j=0}^{p} \theta_j \varepsilon_{t-j} + u_t, \tag{10}$$

with $\mathbb{E}[u_t | x_t, \varepsilon_t, \ldots, \varepsilon_{t-p}] = 0$. Because news may disseminate slowly, the return at time t depends not only on contemporaneous news shocks, but also on past, or stale, news shocks. Suppose that we do not observe ϵ_{t-j} for $j = 0, \ldots, p$, but our object of interest is the effect of a contemporaneous news shock θ_0. We observe a noisy measurement of the contemporaneous news shock, which is corrupted by the past news shocks

$$\tilde{\epsilon}_{1,t} = \sum_{j=1}^{p} \gamma_j \epsilon_{t-j}.$$

This noisy measurement of the contemporaneous news shock is the sentiment of news published about company i at time t. This sentiment measure will contain the contemporaneous news, but may also contain stale news as newspapers report stories from earlier in the week. Mapping this to our motivating model from Section 3.1, $Y_t = \tilde{\epsilon}_{1,t}$ and $e_t = \sum_{j=1}^{p} \theta_j \varepsilon_{t-j} + u_t$.

Our instrument will be the same sentiment measure, but conditioning on the context that

words indicating stale news are not present in the article. We will expand on this in Section 6.2.

6.1 Data

The data set is a corpus of financial and economic news extracted from Reuters, first used in Ding et al. (2014). It spans from 20 October 2006 to 19 November 2013, with the number of documents totalling 105, 375. First, we remove a small number of articles with blank text, and then match the articles with the main firm of interest. This is done by extracting company names from an identifier in the url of the article.[13] We extract any company that has at least 100 occurrences of an identifier (note that some companies have multiple identifiers, "boeing" and "dreamliner" or "apple" and "iphone"), leaving us with a total of 86 companies. We further remove articles that have more than one identifier.

We match these articles with the adjusted close-to-close returns of the identified firms, retrieved from CRSP. We match returns at date t with articles posted on date t. This differs from the strategy of Ke et al. (2019), who match articles posted on day t with a companies three day returns from market close on day $t-2$ to market close on day $t+1$). They follow this strategy because they do not know the timing by which new information is impounded in prices: if prices adjust slowly to news it makes sense to align with future returns, and if articles are a restatement of past news it makes sense to align with past returns. This is exactly the endogeneity issue that we are trying to solve.

Articles posted on weekends and trading holidays are dropped from the sample. Some companies have news articles posted about them but do not have returns associated with the date, as the stock has been removed from trading (for example due to bankruptcy, of which there were several during this time period). These articles are dropped from the sample. The total number of articles that can be matched to company returns of 21, 524. Figure 1a displays the number of articles by hour of the day, summed over the entire sample and Figure 1b displays the number of articles each month over the 7 year period.

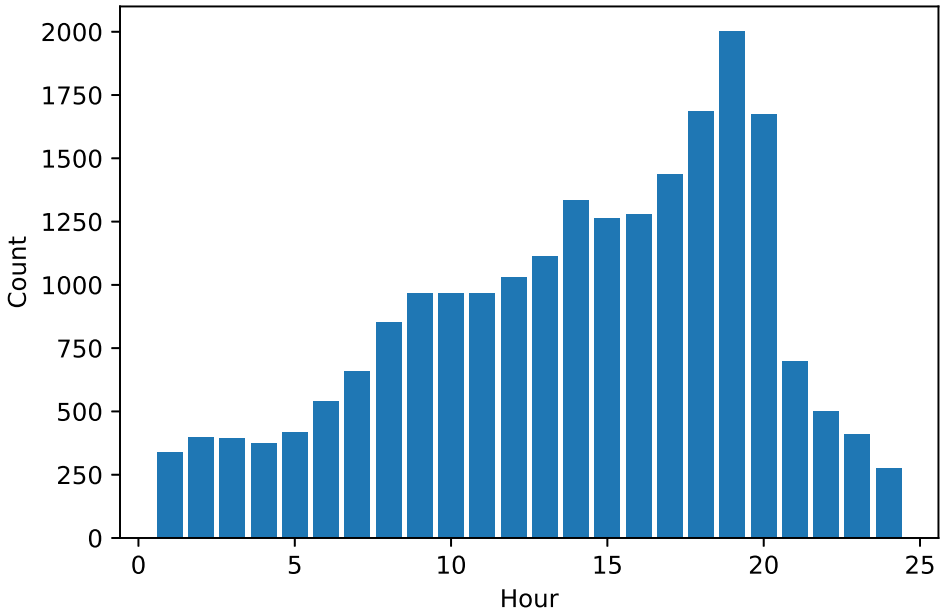
After combining articles about the same company on the same day, we have a total of 13, 740 return-document pairs.

6.2 Identification and results

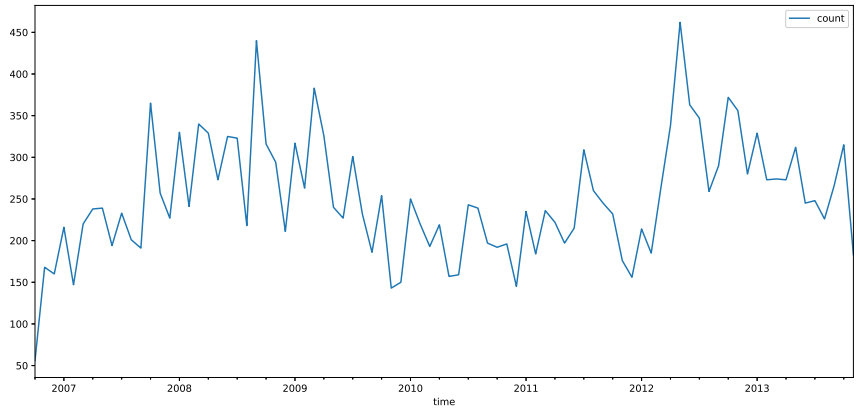
We use the highly charged sentiment terms from Ke et al. (2019)(Table A2), which are replicated in Table 2. We will combine these into 2 categories, positive and negative sentiment

[13] For example the article hosted at https://www.reuters.com/article/us-energy-bp-idUSWLA488420061024, we identify using "us-energy-bp", simplifying to "bp".

Figure 1



(a) The count of the number of articles posted at each hour (in 24 hour EST time).



(b) The number of articles each month from October 2006 to November 2013.

Table 1: Highly charged sentiment words of Ke et al. (2019). We present the stemmed and lemmatized version, but is otherwise identical to Table A2.

Positive Words	Negative Words
undervalu, repurchas, surpass, upgrad, ralli, surg, treasuri, customari, imbal, jump, declar, unsolicit, up, discret, buy, climb, bullish, beat, tender, top, visibl, soar, horizon, tanker, deepwat, reconnaiss, tag, deter, valv, foray, clip, fasten, bracket, potent, unanim, buoy, bake, get, fragment, activist, cardiolog, oversold, bidder, cheer, exceed, terrain, terrif, upbeat, gratifi, armor	shortfal, downgrad, disappoint, tumbl, blame, hurt, plummet, auditor, plung, waiver, miss, slowdown, halt, sluggish, lower, downward, warn, fall, coven, woe, slash, resign, delay, subpoena, lacklust, soften, default, soft, widen, postpon, unfortun, insuffici, unlaw, issuabl, unfavor, regain, deficit, irregular, eros, bondhold, weak, hamper, overrun, ineffici, persist, notifi, allot, wors, setback, grace

words. The context we will be using is that the words 'yesterday', 'last week' ,'day before yesterday', 'in a row', and the three weekdays prior to the article are not mentioned within that article. Of the total $21,524$ documents within the sample, $7,770$ or 36% of articles satisfy this particular context.

We run a regression of returns for company i at time t on the odds ratio of positive words to neutral words (words which are neither positive nor negative), and the odds ratio of negative words to neutral words. Ideally we would construct our sentiment variables as the ratio of the positive words to negative words (or vice versa), but some articles contain either no positive or no negative words (or both), and variable would be undefined for those articles.

Table 2 presents results. The first column presents the reduced form regression, where the odds ratios for our positive and negative sentiment terms are not conditioned on the context. The second column presents the same regression but includes company fixed effects. The third column presents our main specifications, the IV regression, where we instrument for the positive and negative sentiment odds ratios, with positive and negative sentiment odds ratios that are conditioned on the context. The fourth presents the IV regression but including company fixed effects.

We would expect that the coefficients on the IV regression are larger than the reduced form, as our instruments removes articles that are reporting on stale news, and we would expect most of this stale news to already be reflected in returns. We can see that the effect of positive words is similar with or without conditioning on the context designed to remove stale news shocks. However the effect of negative terms is approximately 50% larger when we

Table 2: News shocks

	Reduced Form	Reduced Form	IV	IV
Positive	0.3361	0.3938	0.3181	0.3964
	(0.040)	(0.043)	(0.0785)	(0.0883)
Negative	-0.4175	-0.4257	-0.6061	-0.6335
	(0.050)	(0.053)	(0.0823)	(0.0909)
Company FE	No	Yes	No	Yes

Table reports the effects of shocks of asset returns. The first column is the reduced form regression of the odds ratio of positive terms over neutral and the odds ratio of negative terms over neutral terms, without conditioning on the context. The second column is the same as the first but includes company fixed effects. The three columns is the IV regression, where the sentiment variables are instrumented by sentiment variables that do condition on the context. The fourth column is the same as the third but includes company fixed effects. Standard errors are reported in parenthesis. All coefficient are significant at the 1% level.

focus purely on contemporaneous news shocks. This leads to 2 news insights. First, negative news shocks (as measured by negative sentiment) have significant larger effect than positive shocks, even when controlling for unobserved heterogeneity across firms; Second, the markets doesn't discount positive news shocks are much as negative news shocks. Removing stale news articles that had positive sentiment had little effect on the asset returns. This could also be capturing serial correlation in positive news shocks that we do not model.

7 Monetary Policy Shocks

In this section, we will present our second empirical application, identifying monetary policy shocks in a SVAR using transcripts from FOMC meetings.

Consider a K-dimensional, covariance stationary structural model. The model is given by

$$y_t = \sum_{j=1}^{p} \boldsymbol{A}_j y_{t-j} + \mathbf{B}\epsilon_t$$

where ϵ_t are the structural innovations, distributed i.i.d. according to some unknown distribution F, with $\mathbb{E}_F[\epsilon_t] = 0_{K\times 1}$, $\mathbb{E}_F[\epsilon_t\epsilon_t'] = \mathbb{I}_K$ for all $t = 1\cdots, T$ and $\mathbf{B}$ is an unknown $K \times K$ matrix.

Our objects of interest are structural impulse response functions h^{th} period ahead structural impulse response function of our variables to a target structural shock of interest (which

without loss of generality we assume is the first structural shock, $\epsilon_{1,t}$). The h^{th} period ahead structural impulse response function (to the first shock is) is

$$\theta_h(\mathbf{A}, \mathbf{B}) = C_h(\mathbf{A})\mathbf{B}e_1,$$

where e_1 denotes the 1^{st} column of $\mathbb{I}_K$ and $A \equiv (\mathbf{A}_1, \ldots, \mathbf{A}_p)$.[14] Note that $\theta_0 = \mathbf{B}e_1$, and hence

$$\theta_h(\mathbf{A}, \mathbf{B}) = C_h(\mathbf{A})\theta_0.$$

We can consistently estimate $\mathbf{A}$ and $\Sigma = \mathbf{B}\mathbf{B}'$, but $\mathbf{B}$ and hence θ_0 is not uniquely pinned down. With sufficient restrictions, we can specify a unique θ_0 and hence identify the entire sequence of θ_h. A popular way to identify the impulse response of a target structural shock is to use an 'external instrument' (Stock and Watson (2008, 2012); Mertens and Ravn (2013)).

We will consider a simple 3-variable monetary SVAR, that includes the Federal Funds rate (i_t), GDP Deflator (p_t) and GDP (gdp_t). Hence

$$y_t = (i_t, \ln p_t, \ln gdp_t).$$

Suppose that we are interested in the impact of the monetary policy shock, $\epsilon_{1,t}$, on our three macroeconomic variables. In the langauge of our model, the monetary policy shock is our target latent variable, and the nuisance latent variables are the inflation shock, $\epsilon_{2,t}$, and the output shock, $\epsilon_{3,t}$. In order to construct an external instrument, z_t for our monetary policy shock, we need to satisfy the following assumption:

Assumption 6 (SVAR-IV)**.**

1. $\mathbb{E}[z_t\epsilon_{1,t}] = \alpha \neq 0$.
2. $\mathbb{E}[z_t\epsilon_{j,t}] = 0$ for $j \neq 1$.

Assumption 6 allows us to identify our structural impulse response functions to our target shock. First, we can identify up to scale the covariance between the external instruments and the reduced form residuals, u_t

$$\kappa := \mathbb{E}[z_t u_t] = \mathbb{E}[\mathbf{B}z_t\epsilon_t] = \alpha\theta 0, \tag{11}$$

[14] Y_t has a structural moving average representation $Y_t = \sum_{k=1}^{\infty} C_k(\mathbf{A})\mathbf{B}\epsilon_{t-k}$, where

$$C_h(A) \equiv \sum_{m=1}^{h} C_{h-m}\mathbf{A}_m,\ h \in \mathbb{N},$$

with $C_0 \equiv \mathbb{I}_n$ and $A_m = 0$ if $m > p$; see Lütkepohl (1990).

where the final equality follows from $\mathbf{B}e_1 = \theta_0$.[15] The structural impulse responses can then be identified by imposing a unit normalization of the impulse response of the first variable on impact, i.e. $\theta_{0,1} = \mathbf{B}_{1,1} = 1$. This implies $\kappa_1 = \mathbb{E}[z_t u_{1,t}] = \alpha$, $\theta_0 = \kappa/\kappa_{1,1}$ and $\theta_h = C_h(\mathbf{A})\kappa/\kappa_{1,1}$. We can identify the sequence θ_h from observables, the instrment z_t and the reduced form residuals u_t.

We can use the results from Section 4 to find valid instruments to identify the monetary policy shock. In particular, using the log odds ratio defined in Equation (3) and our result from Proposition 1, we simply need to find two words in a shared context that are relevant to the target latent, our monetary policy shock, but exogenous to the nuisance latent variables, the output and inflation shocks.

7.1 Data

The data are the transcripts of the Federal Open Market Committee (FOMC) meetings from March 29, 1976–December 18, 2013.[16] There are a total of 318 meetings within this time period. In our SVAR we will be using quarterly data, so we concatenate all meeting transcripts within a quarter, resulting in a total of 152 documents.

We perform the following preprocessing steps. First, we separate the spoken text from the name of the speaker, to create a labelled set of 'interjections'. There are an average of 834 interjections per document. We then transform the interjections into lower case and remove stop words.[17] We remove any remaining words of size 1, any non-alphanumeric characters.[18] This gives us an average word count per document of $35,026$, and a total vocabulary of $31,747$ distinct terms. A word cloud of the 200 most commonly spoken terms is presented in Figure 2.

7.2 Identification

In order to use the log odds ratio as a instrument for our monetary policy shock, we need two words, in a shared context, which are relevant to the monetary policy shock, but are exogenous to both the output shock and the interest rate shock.

[15] More specifically,

$$\mathbb{E}[\mathbf{B}z_t\epsilon_t] = [\mathbf{B}e_1, \mathbf{B}e_2, \mathbf{B}e_3]\mathbb{E}[z_t[\epsilon_{1,t}, \epsilon_{2,t}, \epsilon_{3,t}]'] = [\mathbf{B}e_1, \mathbf{B}e_2, \mathbf{B}e_3][\alpha, 0, 0]' = \alpha\mathbf{B}e_1 = \alpha\theta_0$$

Where the second equality follows from Assumption 6.

[16] Full transcripts of the meeting do not exist prior to March 29 1976, only minutes of the meetings. The transcripts can be obtained directly from the Federal Reserve website.

[17] The stop word list is from Hansen et al. (2018).

[18] Removing stop words and short words is not strictly necessary using our method, as we only require a small vocabulary of terms. However if one seeks to use an automated method of generating n-grams, then these proprocessing steps are required.

Figure 2: A word cloud of the 200 most commonly uttered terms after preprocessing. The size of the terms is proportional to their frequencies.

If we were to look at words unconditionally, this would be a difficult task. For example consider a term like '*interest*'. Clearly these will relevant to the monetary policy shock, but it is unlikely that, unconditionally, this term will be exogenous to the nuisance shocks. This is due to the VAR structure. Consider the interest rate equation:

$$i_t = \sum_{j=1}^{p} \mathbf{A}_{j,1,\cdot} y_{t-j} + B_{1,1}\epsilon_{1,t} + B_{1,2}\epsilon_{2,t} + B_{1,3}\epsilon_{3,t},$$

where $A_{j,1,\cdot}$ is the 1^{st} row of A_j and $B_{i,j}$ is the i,j entry of B. Interest rates respond not only to monetary policy shocks, but the output and inflation shocks as well (this is easily rationalized as a Taylor rule). Therefore, if there is an inflation or an output shock, we would except the central bank to adjust interest rates in response. Importantly we would expect that the FOMC members would be discussing '*interest*' in response to these shocks. Hence discussion about '*interest*' is endogenous to inflation and output (our nuisance) shocks.

However by leveraging the idea of a context, we can construct instruments that are plausibly exogenous. In particular, the context we consider is c^* ='*output*' and '*inflation*' are not mentioned within that sentence. We then set w_1^* = '*interest*' and w_2^* = '*funds*'. The log odds ratio of w_1^* and w_2^* in context c^* is plotted in Figure 3.

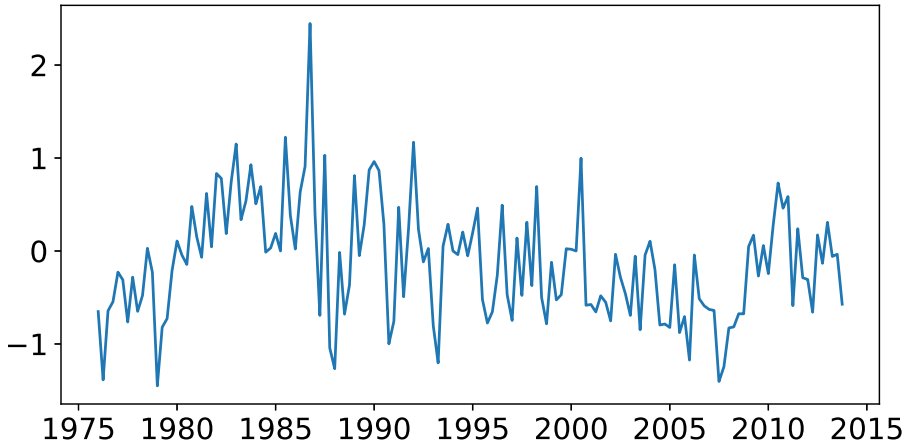


Figure 3: The log odds ratio of '*interest rates*' and '*funds*' in context '*output*' and '*inflation*' are not mentioned within that interjection. Standard errors are extremely small and thus not plotted.

7.3 Results

We estimate the SVAR model with 4 lags and compute impulse responses up to 20 horizons after impact of the monetary policy shock. The impulse responses for output, inflation and interest rates are plotted in Figure 4.

We can see that output and inflation decrease on impact, continue to fall and then gradually returning to baseline. Interest rates on the other hand, increases on impact before returning to baseline after approximately 10 quarters. Our result that inflation falls on impact due to an unexpected monetary policy tightening (a positive monetary policy shock) seems intuitive. However, in SVAR's, an unexpected monetary tightening often leads to a counter intuitive increase in inflation, known in the literature as the prize puzzle (Sims (1986)). We are able to solve the price puzzle using our measure of monetary policy shocks as an external instrument, without needing to expand the model beyond the simple three variable setup.

8 Conclusion

This paper presents a theoretical framework that formalizes an explicit link between text data and latent variables in a quantitative econometric model. Our model justifies formally and guides the use of text data to augment existing econometric methods. In particular, this model is used to find instruments for a target latent variable in the presence of confounding

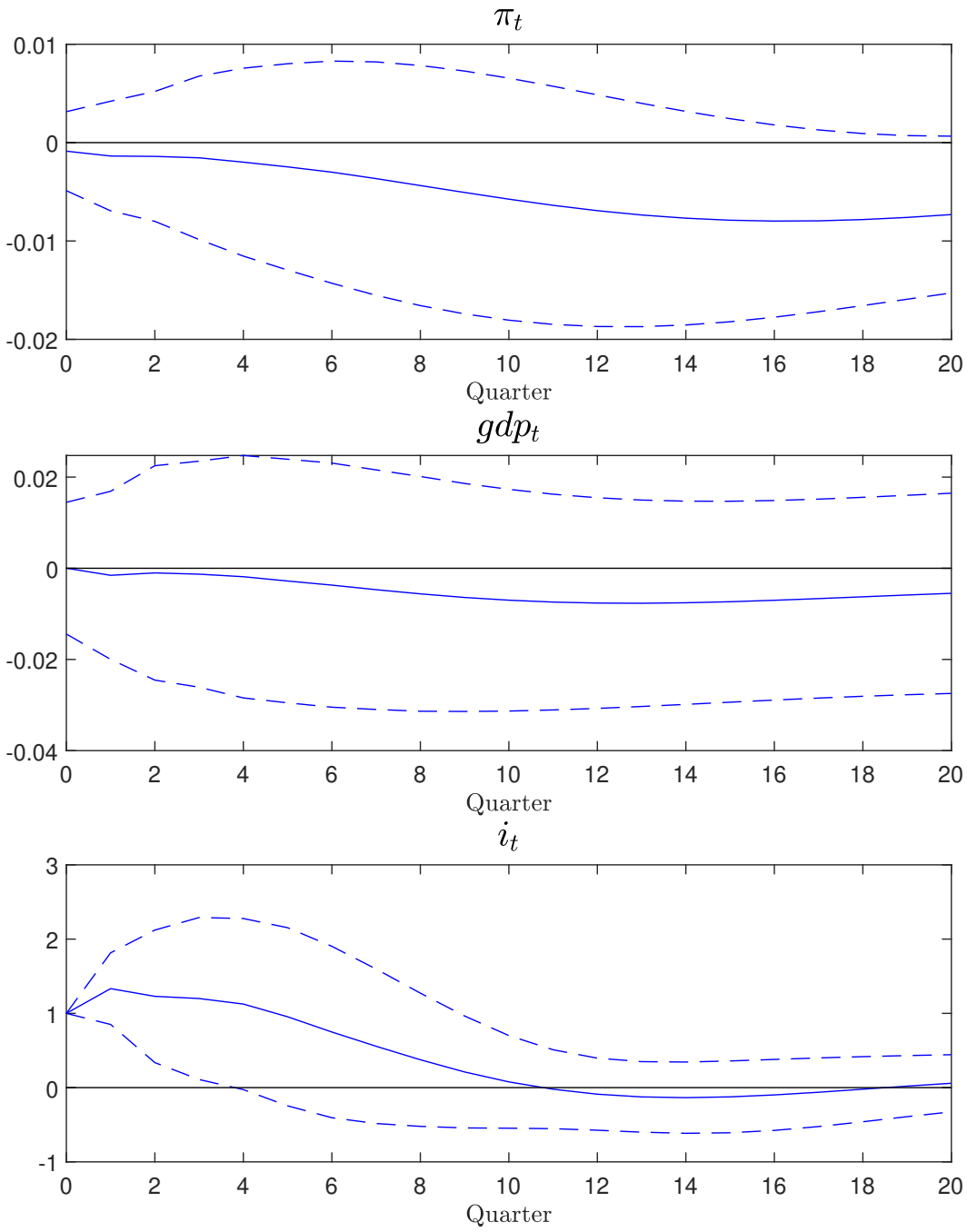


Figure 4: Impulse responses of (in order) inflation, output and interest rates to a contractionary monetary policy shock.

latent variables.

The paper uses the model to provide several new insights about identifying latent variables from text data. First, we show that under weak conditions, the estimated utilities from this model, or the odds ratio, of certain term-context pairs can be used as instruments. Secondly, stronger conditions are needed for raw word counts to be used as instruments. Finally, we demonstrate that the EPU index, which uses raw probabilities, may violate its identifying assumptions.

We develop a two step estimation procedure which first estimates the instruments from the text data, and then plugs these in as instruments in the econometric model. We provides explicit bounds on the worst case estimation error of the logs odds ratio, and use this to show that the estimation of the instruments in the first stage does not affect inference in the second stage as $(T \log(T))/N \to 0$. As documents are often large relative to the time dimension, this implies that researchers can use the estimated instruments as if they were data, not requiring any adjustment to standard errors in their regression models.

We present two empirical applications. The first is to quantify the affect of news shocks on asset returns, whilst controlling for stale news. We find that the effect of negative contemporaneous news shocks is much larger when we use our instrument to remove the effects of stale news. We also find that positive news shocks have a very similar magnitude after instrumenting, suggesting that investors do not discount positive news shocks as quickly as negative news shocks.

The second empirical application is the identification of monetary policy shocks in a monetary SVAR using our text as external instruments. We show that the external instruments assumptions of SVAR-IV are the same as our assumptions for instruments validity, thus allowing our results to be applied immediately. We find that the impulse responses identified using external instruments do not suffer from the price puzzle .

References

Andreas, J. and D. Klein (2015): "When and Why are Log-Linear Models Self-Normalizing?" in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 244–249.

Bai, J. and S. Ng (2008): "Extremum Estimation when the Predictors are Estimated from Large Panels," *Annals of Economics and Finance*, 9, 201–222.

Baker, S. R., N. Bloom, and S. J. Davis (2016): "Measuring Economic Policy Uncertainty," *The Quarterly Journal of Economics*, 131, 1593–1636.

Bengio, Y., R. Ducharme, P. Vincent, and C. Jauvin (2003): "A Neural Probabilistic Language Model," *Journal of Machine Learning Research*, 3, 1137–1155.

Bernanke, B. S., J. Boivin, and P. Eliasz (2005): "Measuring the Effects of Monetary Policy: A Factor-Augmented Vector Autoregressive (FAVAR) Approach," *The Quarterly Journal of Economics*, 120, 387–422.

Ding, X., Y. Zhang, T. Liu, and J. Duan (2014): "Using Structured Events to Predict Stock Price Movement: An Empirical Investigation," *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1415–1425.

Engle, R. F., S. Giglio, B. T. Kelly, H. Lee, and J. Stroebel (2019): "Hedging Climate Change News," *NBER Working Paper w25734*.

Feuer, W. (2020): "Apple stock reaches all-time high after Chinese government data shows iPhone unit sales spike," *CNBC*.

Firth, J. R. (1957): "A Synopsis of Linguistic Theory, 1930-1955," *Studies in Linguistic Analysis*, 1–32.

Friedman, M., A. J. Schwartz, et al. (1963): "A Monetary History of the United States, 1867–1960," *NBER Books*.

Gentzkow, M., B. Kelly, and M. Taddy (2019a): "Text as Data," *Journal of Economic Literature*, 57, 535–74.

Gentzkow, M., J. M. Shapiro, and M. Taddy (2019b): "Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech," *Econometrica*, 87, 1307–1340.

Gertler, M. and P. Karadi (2015): "Monetary Policy Surprises, Credit Costs, and Economic Activity," *American Economic Journal: Macroeconomics*, 7, 44–76.

Hansen, S., M. McMahon, and A. Prat (2018): "Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach," *The Quarterly Journal of Economics*, 133, 801–870.

Ke, Z. T., B. T. Kelly, and D. Xiu (2019): "Predicting returns with text data," *NBER Working Paper w26186*.

Kilian, L. and H. Lütkepohl (2017): *Structural Vector Autoregressive Analysis*, Cambridge University Press.

Lütkepohl, H. (1990): "Asymptotic Distributions of Impulse Response Functions and Forecast Error Variance Decompositions of Vector Autoregressive Models," *The Review of Economics and Statistics*, 72, 116–125.

Mertens, K. and M. O. Ravn (2013): "The Dynamic Effects of Personal and Corporate Income Tax Changes in the United States," *American Economic Review*, 103, 1212–47.

Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013a): "Efficient Estimation of Word Representations in Vector Space," *arXiv preprint arXiv:1301.3781*.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean (2013b): "Distributed Representations of Words and Phrases and their Compositionality," 3111–3119.

Mnih, A. and G. Hinton (2007): "Three New Graphical Models for Statistical Language Modelling," *Proceedings of the 24th International Conference on Machine learning*, 641–648.

Murphy, K. M. and R. H. Topel (1985): "Estimation and Inference in Two-Step Econometric Models," *Journal of Business & Economic Statistics*, 3, 370–379.

Newey, W. K. and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," *Handbook of Econometrics*, 4, 2111–2245.

Newey, W. K., K. D. West, et al. (1987): "A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703–708.

Pagan, A. (1984): "Econometric Issues in the Analysis of Regressions with Generated Regressors," *International Economic Review*, 25, 221–247.

Romer, C. D. and D. H. Romer (1989): "Does Monetary Policy Matter? A New Test in the Spirit of Friedman and Schwartz," *NBER Macroeconomics Annual*, 4, 121–170.

——— (2004): "A New Measure of Monetary Shocks: Derivation and Implications," *American Economic Review*, 94, 1055–1084.

——— (2010): "The Macroeconomic Effects of Tax Changes: Estimates Based on a New Measure of Fiscal Shocks," *American Economic Review*, 100, 763–801.

Shannon, C. E. (1948): "A Mathematical Theory of Communication," *Bell System Technical Journal*, 27, 379–423.

Sims, C. A. (1986): "Are Forecasting Models Usable for Policy Analysis?" *Quarterly Review of the Federal Reserve Bank of Minneapolis*, 10, 2–16.

——— (1992): "Interpreting the Macroeconomic Time Series Facts: The Effects of Monetary Policy," *European Economic Review*, 36, 975–1000.

Sims, C. A. and T. Zha (2006): "Does Monetary Policy Generate Recessions?" *Macroeconomic Dynamics*, 10, 231–272.

Stock, J. H. and M. W. Watson (2008): "What's New in Econometrics-Time Series," *NBER Summer Institute, Lecture*, 7, 2007–09.

——— (2012): "Disentangling the Channels of the 2007-2009 Recession," *NBER Working Paper w18094*.

Taddy, M. (2013): "Multinomial Inverse Regression for Text Analysis," *Journal of the American Statistical Association*, 108, 755–770.

Taddy, M. et al. (2015): "Distributed Multinomial Regression," *The Annals of Applied Statistics*, 9, 1394–1414.

Wooldridge, J. M. (2010): *Econometric Analysis of Cross Section and Panel Data*, MIT press.

A Proofs

A.1 Proofs in Section 4

Proof of Proposition 1

Proof. From Equation (5), under Assumption 1

$$z_t(w, w', c) = x_t'(\gamma_{w,c} - \gamma_{w',c}) + z_t^*(\alpha_{1,w,c} - \alpha_{1,w',c}). \tag{12}$$

We need to show that $z_t(w_1^*, w_2^*, c^*)$ satisfies the conditions in Definition 1.

Relevance:

$$\begin{aligned}\mathbb{E}[z_t Y_t] &= \mathbb{E}[(x_t'(\gamma_{w,c} - \gamma_{w',c}) + z_t^*(\alpha_{1,w,c} - \alpha_{1,w',c}))Y_t] \\ &= \mathbb{E}[Y_t x_t'](\gamma_{w,c} - \gamma_{w',c}) + \mathbb{E}[z_t^* Y_t](\alpha_{1,w,c} - \alpha_{1,w',c}).\end{aligned}$$

By assumption z_t^* can be used as an instrument for Y_t, hence $\mathbb{E}[z_t^* Y_t] \neq 0$. Also $(\alpha_{1,w,c} - \alpha_{1,w',c}) \neq 0$ by assumption. Together these imply $\mathbb{E}[z_t Y_t] \neq 0$.

Exogeneity:

$$\begin{aligned}\mathbb{E}[z_t e_t] &= \mathbb{E}[(x_t'(\gamma_{w,c} - \gamma_{w',c}) + z_t^*(\alpha_{1,w,c} - \alpha_{1,w',c}))e_t] \\ &= \mathbb{E}[e_t x_t'](\gamma_{w,c} - \gamma_{w',c}) + \mathbb{E}[z_t^* e_t](\alpha_{1,w,c} - \alpha_{1,w',c}).\end{aligned}$$

Once again by assumption z_t^* can be used as an instrument for Y_t and hence $\mathbb{E}[z_t^* e_t] = 0$. x_t is exogenous by assumption and therefore $\mathbb{E}[e_t x_t'] = 0$. Together this implies $\mathbb{E}[z_t(\mathcal{J}^*, \boldsymbol{\omega}, c^*)e_t] = 0$. □

Proof of Proposition 2

Proof. Let $\boldsymbol{\omega}$ be such that $\sum_{j\in\mathcal{J}} \omega_j = 0$. We can write the generalized log odds from Equation (7) as

$$\begin{aligned}z_t(\mathcal{J}, \boldsymbol{\omega}, c) &= \sum_{j\in\mathcal{J}} \omega_j u_{w_j,c} \\ &= x_t' \sum_{j\in\mathcal{J}} \gamma_{w_j,c} + z_t^* \sum_{j\in\mathcal{J}} \alpha_{1,w_j,c}\end{aligned}$$

The proof will follow the same strategy of Proposition 1. We need to show that $z_t(\mathcal{J}^*, \boldsymbol{\omega}, c^*)$ satisfies the conditions in Definition 1.

Relevance:

$$\begin{aligned}\mathbb{E}[z_t(\mathcal{J}^*, \boldsymbol{\omega}, c^*)Y_t] &= \mathbb{E}[(x_t' \sum_{j\in\mathcal{J}^*} \gamma_{w_j,c^*} + z_t^* \sum_{j\in\mathcal{J}^*} \alpha_{1,w_j,c^*})Y_t] \\ &= \mathbb{E}[Y_t x_t'] \sum_{j\in\mathcal{J}^*} \gamma_{w_j,c^*} + \mathbb{E}[z_t^* Y_t] \sum_{j\in\mathcal{J}^*} \alpha_{1,w_j,c^*}.\end{aligned}$$

By assumption z_t^* can be used as an instrument for Y_t, hence $\mathbb{E}[z_t^* Y_t] \neq 0$. Also $(\alpha_{1,w,c} - \alpha_{1,w',c}) \neq 0$ by assumption. Together these imply $\mathbb{E}[z_t Y_t] \neq 0$.

Exogeneity:

$$\begin{aligned}\mathbb{E}[z_t(\mathcal{J}^*, \boldsymbol{\omega}, c^*)e_t] &= \mathbb{E}[(x_t' \sum_{j\in\mathcal{J}^*} \gamma_{w_j,c^*} + z_t^* \sum_{j\in\mathcal{J}^*} \alpha_{1,w_j,c^*})e_t] \\ &= \mathbb{E}[e_t x_t'](\gamma_{w,c} - \gamma_{w',c}) + \mathbb{E}[z_t^* e_t](\alpha_{1,w,c} - \alpha_{1,w',c}).\end{aligned}$$

Once again by assumption z_t^* can be used as an instrument for Y_t and hence $\mathbb{E}[z_t^* e_t] = 0$. x_t is exogenous by assumption and therefore $\mathbb{E}[e_t x_t'] = 0$. Together this implies $\mathbb{E}[z_t(\mathcal{J}^*, \boldsymbol{\omega}, c^*)e_t] = 0$. □

Conditional probabilities are nonlinear transformations of the data (x_t, z_t^*, e_t). In order to show the covariance between the conditional probabilities and z_t^* and e_t, under general distributions, we will need the following lemma.

Lemma 1. For a random variable X, not constant almost surely, if $g(x)$ is strictly decreasing, and $\mathrm{Cov}(g(x), x) < \infty$, then $\mathrm{Cov}(g(x), x) < 0$. If $g(x)$ is strictly increasing, and $\mathrm{Cov}(g(x), x) < \infty$, then $\mathrm{Cov}(g(x), x) > 0$.

Proof.

$$\begin{aligned}\mathrm{Cov}(g(x), x) &= \mathbb{E}[Xg(x)] - \mathbb{E}[X]\mathbb{E}[g(x)] \\ &= \mathbb{E}[(X - \mathbb{E}[X])g(X)]\end{aligned}$$

Note that $\mathbb{E}[X - E[X]g(\mathbb{E}[x])] = \mathbb{E}[X - E[X]]g(\mathbb{E}[X]) = 0$, and hence

$$= \mathbb{E}[(X - \mathbb{E}[X])(g(X) - g(\mathbb{E}[X]))].$$

As g is strictly decreasing $(X - \mathbb{E}[X])(g(X) - g(\mathbb{E}[X])) \leqslant 0$, with equality when $X = \mathbb{E}[X]$. As X is not constant almost surely, $\mathrm{Cov}(g(x), x) < 0$.

The proof for when $g(x)$ is strictly increasing is identical. □

Proof of Proposition 3

Proof. It will suffice to show that for certain parameter values that we fail to satisfy one of our two instrument conditions. Suppose

- $\gamma_{w,c^*} = 0$ for all w,
- $\alpha_{1,w,c^*} = 0$ for all w,
- $|\alpha_{2,w^*,c^*}| > |\alpha_{2,w',c^*}|$ for all $w' \neq w^*$,

Under these parameter values

$$p_t(w^*, c^*) = \frac{\exp(\alpha_{2,w^*,c^*} e_t)}{\sum_{w'=1}^{V} \exp(\alpha_{2,w',c^*} e_t)}.$$

If we take the derivative w.r.t. e_t

$$\frac{\partial p_t(w^*, c^*)}{\partial e_t} = \frac{\exp(\alpha_{2,w^*,c^*} e_t) \sum_{w'=1}^{V} (\alpha_{2,w^*,c^*} - \alpha_{2,w',c^*}) \exp(\alpha_{2,w',c^*} e_t)}{(\sum_{w'=1}^{V} \alpha_{2,w',c^*} e_t))^2}.$$

As $|\alpha_{2,w^*,c^*}| > |\alpha_{2,w',c^*}|$, for all $w' \neq w^*$, this derivative is either strictly positive ($\alpha_{2,w^*,c^*} > 0$) or strictly negative (if $\alpha_{2,w^*,c^*} < 0$). Assume w.l.o.g. that $\alpha_{2,w^*,c^*} > 0$ and hence z_t is strictly increasing in e_t. Therefore

$$\mathbb{E}[p_t(w^*, c^*) e_t] > 0,$$

by Lemma 1. z_t is not exogenous. □

Proof of Proposition 4

Proof. Under Assumption 3, we have

$$p_t(w^*, c^*) = \frac{\exp(x_t' \gamma_{w^*,c^*} + \alpha_{1,w^*,c^*} z_t^*)}{\sum_{w'=1}^{V} \exp(x_t' \gamma_{w',c^*} + \alpha_{1,w',c^*} z_t^*)}.$$

Relevance:

If we take the derivative of $p_t(w^*, c^*)$ w.r.t. z_t^*

$$\frac{\partial p_t(w^*, c^*)}{\partial z_t^*} = \frac{\exp(x_t' \gamma_{w^*,c^*} + \alpha_{1,w^*,c^*} z_t^*) \sum_{w'=1}^{V} (\alpha_{1,w^*,c^*} - \alpha_{1,w',c^*}) \exp(x_t' \gamma_{w',c^*} + \alpha_{1,w',c^*} z_t^*)}{(\sum_{w'=1}^{V} \exp(x_t' \gamma_{w',c^*} + \alpha_{1,w',c^*} z_t^*))^2}.$$

By Assumption 3.1, $|\alpha_{1,w^*,c^*}| > |\alpha_{1,w',c^*}|$, for all $w' \neq w^*$, and hence this derivative is either strictly positive ($\alpha_{1,w^*,c^*} > 0$) or strictly negative (if $\alpha_{1,w^*,c^*} < 0$). Assume w.l.o.g.

that $\alpha_{1,w^*,c^*} > 0$ and hence $p_t(w^*, c^*)$ is strictly increasing in z_t^*. Therefore

$$\mathbb{E}[p_t(w^*, c^*)Y_t] = \mathbb{E}[p_t(w^*, c^*)(\delta z_t^* + \nu_t)] => 0,$$

by Lemma 1. Hence z_t is relevant.

Exogeneity:

$$\mathbb{E}[p_t(w^*, c^*)e_t] = \mathbb{E}[\mathbb{E}[p_t(w^*, c^*)e_t|z_t^*] = \mathbb{E}[p_t(w^*, c^*)\mathbb{E}[e_t|z_t^*]] = 0.$$

□

A.2 Proofs in Section 5

Proof of Proposition 5

Proof. **Consistency:**

As the $w_{t,n}$ are i.i.d., we can apply the Law of Large Numbers for each $w_j \in \mathcal{J}$ and in context c^*

$$\begin{aligned}\frac{1}{N}\sum_{n=1}^{N} \mathbf{1}_{w_{t,n}=w_j}\mathbf{1}_{c_{t,n}=c^*} &\xrightarrow{p} \mathbb{E}[\mathbf{1}_{w_{t,n}=w_j}\mathbf{1}_{c_{t,n}=c^*}] \\ &= p_t(w_{t,n} = w_j \cap c_{t,n} = c^*).\end{aligned}$$

The result, then follows from the Continuous Mapping Theorem, noting $\sum_{j\in\mathcal{J}} \omega_j \log(x_j)$ is continuous in each x_j.

Asymptotic Normality: For notational convenience denote $\hat{p}_t(w_j, c) := \frac{1}{N}\#(w_{t,n} = w_j, c_{t,n} = c)$.

Denote $\mathbf{p}_t(c) = [p_t(w_1, c), \ldots, p_t(w_V, c)]'$, the vector of conditional probabilities, and $\hat{\mathbf{p}}_t(c) = [\hat{p}_t(w_1, c), \ldots, \hat{p}_t(w_V, c)]'$, the vector of estimates of the conditional probabilities.

We will show asymptotic normality using the delta method.

Using the Central Limit Theorem we have, $\sqrt{N}(\hat{\mathbf{p}}_t(c) - \mathbf{p}_t(c)) \xrightarrow{d} N(0, M_t)$, where $M_t = P_t - \mathbf{p}_t(c)\mathbf{p}_t(c)'$, and P_t is a diagonal matrix with entries $\mathbf{p}_t(c)$.

The generalized log odds is a function of $\mathbf{p}_t(c)$, with derivative $\nabla(\mathbf{p}_t(c))_j = \omega_j/p_t(w_j, c)$ for $j \in \mathcal{J}$ and 0 otherwise. By the delta method,

$$\sqrt{N}(z_t(\hat{\mathbf{p}}_t(c)) - z_t(\mathbf{p}_t(c)) \xrightarrow{d} N(0, G_t),$$

where

$$G_t = \nabla(\mathbf{p}_t(c))' M_t \nabla(\mathbf{p}_t(c)) = \sum_{j \in \mathcal{J}} \frac{\omega_j^2}{p_t(w_j, c)}.$$

□

Proof of Proposition 6

We will first establish a number of lemmas.

Lemma 2. Let $\hat{p}$ be a multinominal frequency of a word which converges in probability to p. Then $P(|\hat{p} - p| > \alpha) \leqslant 2\exp(-n\alpha^2/2)$.

Proof. $|\hat{p} - p|$ has the bounded differences property. If we change one $w_{t,n}$, $|\hat{p} - p|$ changes by at most $c_n = 2/N$. Therefore using the bounded difference inequality

$$\begin{aligned} P(|\hat{p} - p| > \alpha) &\leqslant \exp\left(\frac{2\alpha^2}{\sum_{i=n}^{N} c_n^2}\right) \\ &= \exp\left(\frac{2\alpha^2}{N4/N^2}\right) \\ &= 2\exp\left(-N\alpha^2/2\right). \end{aligned}$$

□

Lemma 3. Let $\log(x)$ have domain $D = [L, \infty)$, then

$$|\log(x) - \log(y)| \leqslant \frac{1}{L}|x - y|,$$

for all $x, y \in D$.

Proof. Assume without loss of generality that $L \leqslant x \leqslant y$. Then

$$\begin{aligned} |\log(x) - \log(y)| &= \log(y/x) \\ &= \log(1 + (y/x - 1) \\ &\leqslant y/x - 1 \\ &\text{(as } \log(1+u) \leqslant u \text{ for all } u > -1\text{)}, \\ &= 1/x(y - x) \\ &\leqslant \frac{1}{L}(y - x) \\ &\text{(as } L \leqslant x \text{ by assumption)}, \end{aligned}$$

$$= \frac{1}{L}|y - x|$$

□

Lemma 4. Recall $z_t(\mathcal{J}, c, \boldsymbol{\omega}) = \sum_{j \in \mathcal{J}} \omega_j \log(\hat{p}_t(w_j, c))$, where $\hat{p}_t(w_j, c) \xrightarrow{p} p_t(w_j, c)$ for all $j \in \mathcal{J}$. Then

$$P(|\hat{z}_t(\mathcal{J}, c, \boldsymbol{\omega}) - z_t(\mathcal{J}, c, \boldsymbol{\omega})| > \epsilon) \leqslant 2|\mathcal{J}| e^{-\frac{N\epsilon^2 L^2}{2|\mathcal{J}|^2 \omega_M^2}}$$

Proof.

$$\begin{aligned}
P(|\hat{z}_t(\mathcal{J}, c, \boldsymbol{\omega}) - z_t(\mathcal{J}, c, \boldsymbol{\omega})| > \epsilon) &= P\left(|\sum_{j \in \mathcal{J}} \omega_j \log(\hat{p}_t(w_j, c)) - \sum_{j \in \mathcal{J}} \omega_j \log(p_t(w_j, c))| > \epsilon\right) \\
&= P\left(|\sum_{j \in \mathcal{J}} \omega_j (\log(\hat{p}_t(w_j, c)) - \log(p_t(w_j, c)))| > \epsilon\right) \\
&\leqslant P\left(\sum_{j \in \mathcal{J}} |\omega_j| |\log(\hat{p}_t(w_j, c)) - \log(p_t(w_j, c))| > \epsilon\right) \\
&\text{(by the triangle inequality and } |xy| = |x||y|\text{)}, \\
&\leqslant P\left(\bigcup_{j \in \mathcal{J}} |\mathcal{J}||\omega_j| |\log(\hat{p}_t(w_j, c)) - \log(p_t(w_j, c))| > \epsilon\right) \\
&\leqslant \sum_{j \in \mathcal{J}} P\left(|\log(\hat{p}_t(w_j, c)) - \log(p_t(w_j, c))| > \frac{\epsilon}{|\omega_j||\mathcal{J}|}\right) \\
&\text{(by Boole's inequality)}, \\
&\leqslant \sum_{j \in \mathcal{J}} P\left(\frac{1}{L}|\hat{p}_t(w_j, c) - p_t(w_j, c)| > \frac{\epsilon}{|\omega_j||\mathcal{J}|}\right) \\
&\text{(by Lemma 3)}, \\
&\leqslant \sum_{j \in \mathcal{J}} 2\exp\left(-\frac{N\epsilon^2 L^2}{2|\mathcal{J}|^2 \omega_j^2}\right) \\
&\text{(by Lemma 2 with } \alpha = L\epsilon/|\omega_j||\mathcal{J}|\text{)}, \\
&\leqslant \sum_{j \in \mathcal{J}} 2\exp\left(-\frac{N\epsilon^2 L^2}{2|\mathcal{J}|^2 \omega_M^2}\right) \\
&\text{(as } \omega_M \geqslant \omega_j \text{ for all } j \in \mathcal{J}\text{)}, \\
&\leqslant 2|\mathcal{J}| \exp\left(-\frac{N\epsilon^2 L^2}{2|\mathcal{J}|^2 \omega_M^2}\right).
\end{aligned}$$

□

Proof of Proposition 6.

Proof. 1.

$$\begin{aligned} P\left(\max_{t\leqslant T}|\hat{z}_t(\mathcal{J},c,\boldsymbol{\omega})-z_t(\mathcal{J},c,\boldsymbol{\omega})|>\epsilon\right) &= P\left(\bigcup_{t=1}^{T}|\hat{z}_t(\mathcal{J},c,\boldsymbol{\omega})-z_t(\mathcal{J},c,\boldsymbol{\omega})|>\epsilon\right) \\ &\leqslant \sum_{t=1}^{T} P\left(|\hat{z}_t(\mathcal{J},c,\boldsymbol{\omega})-z_t(\mathcal{J},c,\boldsymbol{\omega})|>\epsilon\right) \\ &\text{(by Boole's inequality)} \\ &\leqslant T2|\mathcal{J}|\exp\left(-\frac{N\epsilon^2L^2}{2|\mathcal{J}|^2\omega_M^2}\right). \\ &\text{(by Lemma 4)} \end{aligned}$$

2. Follows if we choose $\epsilon = \frac{|\mathcal{J}|w_M\sqrt{C^2+2\log(2\mathcal{J})+2\log(T)}}{\sqrt{N}L}$ for a sufficiently large constant C.

□

A.2.1 Proof of Proposition 7

We will first make explicit the standard GMM assumptions that we will operate under.

Assumption 7 (Standard Consistency)**.**

1. Let $\hat{\theta}$ satisfy $Q_T(\hat{\theta},z_t) > \sup_{\theta\in\Theta} Q_T(\theta,z_t) - \eta_T$, where η_T is a sequence of positive random variable such that $\eta_T = o_p(1)$.
2. $\mathbf{w}_1,\ldots,\mathbf{w}_t$ be strictly stationary and ergodic.
3. Θ be compact.
4. $g(\mathbf{w}_t;\theta)$ be continuous in θ for all $\mathbf{w}_t$.
5. $\hat{W} \xrightarrow{p} W$, where W is positive definite and symmetric.
6. $\mathbb{E}[g(\mathbf{w}_t;\theta)] = 0$ if and only if $\theta = \theta_0$.
7. $\mathbb{E}[\sup_{\theta\in\Theta}\|g(\mathbf{w}_t;\theta)\|] < \infty$.

Proof. The clean maximum is the same as standard. It will suffice to show uniform convergence.

We want to show that $\sup_{\theta\in\Theta}\|g_T(\theta,\hat{z}_t) - g(\theta,z_t)\| \xrightarrow{p} 0$. By the triangle inequality,

$$\sup_{\theta\in\Theta}\|g_T(\theta,\hat{z}_t)-g(\theta)\| \leqslant \sup_{\theta\in\Theta}\|g_T(\theta,\hat{z}_t)-g_T(\theta,z_t)\| + \sup_{\theta\in\Theta}\|g_T(\theta,z_t)-g(\theta)\|.$$

The second term is $o_p(1)$ by standard arguments, using Assumption 7. For the second term,

a mean value expansion of $g_T(\theta, \hat{z}_t)$ around z_t yields

$$g_T(\theta, \hat{z}_t) = g_T(\theta) + \frac{1}{T}\sum_{t=1}^{T} \frac{\partial g(y_t, x_t, \bar{z}_t, \theta)}{\partial z}(\hat{z}_t - z_t),$$

where $\bar{z}_t$ is in the segment between $\hat{z}_t$ and z_t, and the mean value expansion is formally justified whenever $\hat{z}_t$ is in an open neighbourhood of z_t (which it is w.p.a.1 as $\hat{z}_t \xrightarrow{p} z_t$ and z_t is in the interior of the parameter space). Plugging in this expansion

$$\begin{aligned}
\sup_{\theta\in\Theta}\|g_T(\theta, \hat{z}_t) - g_T(\theta, z_t)\| &= \sup_{\theta\in\Theta}\|\frac{1}{T}\sum_{t=1}^{T} \frac{\partial g(y_t, x_t, \bar{z}_t, \theta)}{\partial z}(\hat{z}_t - z_t)\| \\
&\leqslant \sup_t\|\hat{z}_t - z_t\| \sup_{\theta\in\Theta}\|\sum_{t=1}^{T} \frac{\partial g(y_t, x_t, \bar{z}_t, \theta)}{\partial z}\| \\
&= O_p\left(\sqrt{\frac{\log(T)}{N}}\right) O_p(1) \\
&= o_p(1). \qquad (13)
\end{aligned}$$

The $O_p\left(\sqrt{\frac{\log(T)}{N}}\right)$ term follows from Proposition 6, which is $o_p(1)$ under the rate assumption that $\sqrt{\frac{T\log(T)}{N}} \to 0$, and the second term is $O_p(1)$ by Assumption 4.

Next we want to show that $\sup_{\theta\in\Theta}|Q_T(\theta, \hat{z}_t) - Q(\theta)| \xrightarrow{p} 0$. Using the triangle inequality,

$$\sup_{\theta\in\Theta}|Q_T(\theta, \hat{z}_t) - Q(\theta)| \leqslant \sup_{\theta\in\Theta}|Q_T(\theta, \hat{z}_t) - Q_T(\theta, z_t)| + \sup_{\theta\in\Theta}|Q_T(\theta, z_t) - Q(\theta)|.$$

The second term is $o_p(1)$ by standard arguments. We want to show that the first term is also $o_p(1)$. Note that

$$\begin{aligned}
2Q_T(\theta, \hat{z}_t) - 2Q_T(\theta, z_t) &= g_T(\theta, \hat{z}_t)\hat{W}g_T(\theta, \hat{z}_t) - g_T(\theta, z_t)'\hat{W}g_T(\theta, z_t) \\
&= (g_T(\theta, \hat{z}_t) - g_T(\theta, z_t))'\hat{W}(g_T(\theta, \hat{z}_t) - g_T(\theta, z_t)).
\end{aligned}$$

Taking the sup over $\theta \in \Theta$ yields

$$\sup_{\theta\in\Theta} 2|Q_T(\theta, \hat{z}_t) - Q_T(\theta, z_t)| \leqslant \sup_{\theta\in\Theta}\|g_T(\theta, \hat{z}_t) - g_T(\theta, z_t)\|^2 \times \|\hat{W}\| = O_p(1)$$

$\|g_T(\theta, \hat{z}_t) - g_T(\theta, z_t)\|^2 = o_p(1)$ by Equation (13) and $\|\hat{W}\|$ converges to a constant by Assumption 7.4. □

A.2.2 Proof of Proposition 8

Assumption 8 (Standard Asymptotic Normality)**.** Let

1. $\hat{\theta}$ satisfy $\frac{\partial Q_T(\hat{\theta},\hat{z}_t)}{\partial\theta} = o_p(T^{-1/2})$.
2. $\mathbf{w}_1,\ldots,\mathbf{w}_T$ be strictly stationary and ergodic.
3. $\hat{\theta} \xrightarrow{p} \theta_0$.
4. $\hat{W} \xrightarrow{p} W$ where W is positive definite and symmetric.
5. θ_0 be in the interior of Θ.
6. $g(\mathbf{w}_t;\theta)$ be continuously differentiable in θ for any $\mathbf{w}_t$.
7. $\sqrt{T}g_T(\theta_0,z_t)) \xrightarrow{p} N(0,S)$ with S positive definite.
8. $\mathbb{E}[\sup_{\theta\in N_{\theta_0}} \|\frac{\partial g(\mathbf{w}_t;\theta)}{\partial\theta}\|] < \infty$ for some neighborhood N_{θ_0} of θ_0.
9. G be of full column rank p.

Proof. Define $G_T(\theta,z) = \frac{1}{T}\sum_{t=1}^{T}\frac{\partial g(y_t,x_t,z;\theta)}{\partial\theta}$, $G(\theta) = \mathbb{E}\left[\frac{\partial g(y_t,x_t,z_t;\theta)}{\partial\theta}\right]$ and $G = G(\theta_0)$

By Assumption 8.1,

$$\begin{aligned} o_p(T^{-1/2}) &= \frac{\partial Q_T(\hat{\theta},\hat{z}_t)}{\partial\theta} \\ &= -G_T(\hat{\theta},\hat{z}_t)'\hat{W}G_T(\hat{\theta},\hat{z}_t) \\ &= -G_T(\hat{\theta},\hat{z}_t)'\hat{W}(g_T(\theta_0,\hat{z}_t) + G_T(\bar{\theta},\hat{z}_t)(\hat{\theta}-\theta_0)). \end{aligned}$$

The last line is mean value expansion of $g_T(\theta_0,\hat{z}_t)$ around θ_0. $\bar{\theta}$ is in the segment between $\hat{\theta}$ and θ_0. The mean value expansion is formally justified by $\hat{\theta} \xrightarrow{p}$ and Assumption 8.5.

We can therefore rewrite the above display as

$$G_T(\hat{\theta},\hat{z}_t)'\hat{W}G_T(\hat{\theta},\hat{z}_t)(\sqrt{T}\hat{\theta}-\theta_0) = -G_T(\hat{\theta},\hat{z}_t)'\hat{W}G_T(\bar{\theta},\hat{z}_t)\sqrt{T}g_T(\theta_0,\hat{z}_t) + o_p(1)$$

We show below that $G_T(\hat{\theta},\hat{z}_t) \xrightarrow{p} G$, $G_T(\bar{\theta},\hat{z}_t) \xrightarrow{p} G$ and $\sqrt{T}g_T(\theta_0,\hat{z}_t) \xrightarrow{p} \sqrt{T}g_T(\theta_0,z_t)$. Together with $\hat{W} \xrightarrow{p} W$ (Assumption 8.4), we have

$$G'WG\sqrt{T}(\hat{\theta}-\theta_0) = -G'W\sqrt{T}g_T(\theta_0,z_t) + o_p(1)$$

Finally

$$\sqrt{T}(\hat{\theta}-\theta_0) = -(G'WG)^{-1}G'W\sqrt{T}g_T(\theta_0,z_t) + o_p(1)$$

where $G'WG$ is invertible by positive definiteness of W (Assumption 8.4) and that fact that G has full column rank (Assumption 8.9). By Assumption 8.7 and the continuous mapping

theorem

$$\sqrt{T}(\hat{\theta} - \theta_0) = N(0, (G'WG)^{-1}G'WSWG(G'WG)^{-1}).$$

Now we will show

1. $G_T(\hat{\theta}, \hat{z}_t) \xrightarrow{p} G$.
2. $G_T(\bar{\theta}, \hat{z}_t) \xrightarrow{p} G$.
3. $\sqrt{T}g_T(\theta_0, \hat{z}_t) \xrightarrow{p} \sqrt{T}g_T(\theta_0, z_t)$.

1. $G_T(\hat{\theta}, \hat{z}_t) \xrightarrow{p} G$

By triangle inequality

$$\|G_T(\hat{\theta}, \hat{z}_t) - G\| \leqslant \|G_T(\hat{\theta}, \hat{z}_t) - G_T(\hat{\theta}, z_t)\| + \|G_T(\hat{\theta}, z_t) - G\|$$

The second term is $o_p(1)$ by standard arguments. Consider a mean value expansion of $G_T(\hat{\theta}, \hat{z}_t)$ around z_t,[19] where the j^{th} column is

$$G_{j,T}(\hat{\theta}, \hat{z}_t) = G_{j,T}(\hat{\theta}, z_t) + \frac{1}{T}\sum_{t=1}^{T} \frac{\partial^2 g_j(y_t, x_t, \bar{z}_t, \hat{\theta})}{\partial\theta\partial z'}(\hat{z}_t - z_t)$$

where $\bar{z}_t$ lies in the segment between $\hat{z}_t$ and z_t. Hence

$$\begin{aligned}
\|G_T(\hat{\theta}, \hat{z}_t) - G_T(\hat{\theta}, z_t)\| &= \|\frac{1}{T}\sum_{t=1}^{T} \frac{\partial^2 g_j(y_t, x_t, \bar{z}_t, \hat{\theta})}{\partial\theta\partial z'}(\hat{z}_t - z_t)\| \\
&\leqslant \sup_t |\hat{z}_t - z_t| \, \|\frac{1}{T}\sum_{t=1}^{T} \frac{\partial^2 g_j(y_t, x_t, \bar{z}_t, \hat{\theta})}{\partial\theta\partial z'}\| \\
&= O_p\left(\sqrt{\frac{\log(T)}{N}}\right) O_p(1)
\end{aligned}$$

The $O_p\left(\sqrt{\frac{\log(T)}{N}}\right)$ term follows from Proposition 6, which is $o_p(1)$ under the rate assumption that $\sqrt{\frac{T\log(T)}{N}} \to 0$, and the second term is $O_p(1)$ by Assumption 5.

2. $G_T(\bar{\theta}, \hat{z}_t) \xrightarrow{p} G$

This follows an identical proof to $G_T(\hat{\theta}, \hat{z}_t) \xrightarrow{p} G$.

[19] Formally justified by $\hat{z}_t \xrightarrow{p} z_t$ and z_t is in the interior of the parameter space.

3. $\sqrt{T}g_T(\theta_0, \hat{z}_t) \xrightarrow{p} \sqrt{T}g_T(\theta_0, z_t)$

A mean value expansion of $\sqrt{T}g_T(\theta_0, \hat{z}_t)$ around z_t (formally justified by the fact that $\tilde{z}_t \xrightarrow{p} z_t$ and z_t is in the interior of the parameter space) yields

$$\sqrt{T}g_T(\theta_0, \hat{z}_t) = \sqrt{T}g_T(\theta_0, z_t) + \sqrt{T}\frac{1}{T}\sum_{t=1}^{T}\frac{\partial g(y_t, x_t, \bar{z}_t, \theta_0)}{\partial z}(\hat{z}_t - z_t),$$

where $\bar{z}_t$ is in the segment between $\hat{z}_t$ and z_t. Therefore

$$\begin{aligned}
\|\sqrt{T}g_T(\theta_0, \hat{z}_t) - \sqrt{T}g_T(\theta_0, z_t)\| &= \|\sqrt{T}\frac{1}{T}\sum_{t=1}^{T}\frac{\partial g(y_t, x_t, \bar{z}_t, \theta_0)}{\partial z}(\hat{z}_t - z_t)\| \\
&\leqslant \sqrt{T}\sup_t\|\hat{z}_t - z_t\|\,\|\sum_{t=1}^{T}\frac{\partial g(y_t, x_t, \bar{z}_t, \theta_0)}{\partial z_t}\| \\
&= \sqrt{T}O_p\left(\sqrt{\frac{\log(T)}{N}}\right)O_p(1) \\
&= O_p\left(\sqrt{\frac{T\log(T)}{N}}\right)O_p(1)
\end{aligned}$$

The $O_p\left(\sqrt{\frac{\log(T)}{N}}\right)$ term follows from Proposition 6, which is $o_p(1)$ under the rate assumption that $\sqrt{\frac{T\log(T)}{N}} \to 0$, and the second term is $O_p(1)$ by Assumption 4. □

B Additional Estimation Results

In this appendix, we will discuss additional estimation results. In particular we will show that we can consistently estimate the asymptotic variance using the estimated instruments $\hat{z}_t$, under heteroscedasticity and homoscedasticity.

In Proposition 8 we derived the asymptotic variance for $\hat{\theta}$

$$\operatorname{avar}(\hat{\theta}) = (G'WG)^{-1}G'WSWG(G'WG)^{-1}.$$

This can be estimated using

$$\widehat{\operatorname{avar}}(\hat{\theta}) = (\hat{G}'\hat{W}\hat{G})^{-1}\hat{G}'\hat{W}\hat{S}\hat{W}\hat{G}(\hat{G}'\hat{W}\hat{G})^{-1}$$

Where $\hat{G} = G_T(\hat{\theta}, \hat{z}_t) = \frac{1}{T}\sum_{t=1}^{T}\frac{\partial g(y_t, x_t, \hat{z}_t; \hat{\theta})}{\partial \theta}$. When (i) the data are i.i.d. or when $(g(\mathbf{w}_t; \theta_0), \mathcal{F}_t)_{t\in\mathbb{Z}}$ is a martingale difference sequence (where $\mathcal{F}_t$ denotes the σ-algebra generated by $(\mathbf{w}_t, \mathbf{w}_{t-1}, \dots)$), (ii) the condition $\mathbb{E}[\|g(\mathbf{w}_t; \theta_0)\|^2] < \infty$ then by the central limit theorem (or CLT for stationary and ergodic martingale difference sequences) we obtain

$$\sqrt{T}g_T(\theta_0, z_t) \xrightarrow{d} N(0, S), \text{ where } S = \mathbb{E}[g(\mathbf{w}_t; \theta_0)g(\mathbf{w}_t; \theta_0)'].$$

And hence we can estimate S by $\hat{S} = \frac{1}{T}\sum_t g(y_t, x_t, \hat{z}_t; \hat{\theta})g(y_t, x_t, \hat{z}_t; \hat{\theta})'$. In order to show consistency of variance estimator, we require an additional assumption.

Assumption 9 (Consistent Estimation of Variance)**.**

$$\mathbb{E}\left[\sup_{\theta\in N_{\theta_0}, z\in N_{z_t}} \|g(\mathbf{w}_t; \theta)g(\mathbf{w}_t; \theta)'\|\right] < \infty,$$

where N_{θ_0} is an open neighborhood around θ_0 and N_{z_t} is an open neighbourhood around z_t.

Proposition 9. Let Assumptions 5, 8 and 9 hold. Then

$$\widehat{\operatorname{avar}}(\hat{\theta}) \xrightarrow{p} \operatorname{avar}(\hat{\theta}).$$

If one chooses $\hat{W}$ such that $\hat{W} \xrightarrow{p} S^{-1}$, then we can achieve the optimally weighted GMM variance $(G'S^{-1}G)^{-1}$. One such weighting matrix is

$$\hat{W} = \left[\frac{1}{T}\sum_{t=1}^{T} g(\hat{z}_t, \dot{\theta})g(\hat{z}_t, \dot{\theta})'\right]^{-1}. \tag{14}$$

Where $\dot{\theta}$ is a consistent initial estimate of θ_0.

Proposition 10. Let Assumptions 5, 8 and 9 hold, and $\dot{\theta} \xrightarrow{p} \theta_0$. Then

$$\hat{W} \xrightarrow{p} S^{-1}.$$

$\hat{W}$ could be constructed using the standard 2 step procedure. First estimate $\dot{\theta}$ by maximizing $Q_T(\theta, \hat{z}_t)$ using $\hat{W} = I$. Then estimate $\hat{\theta}$ by maximizing $Q_T(\theta, \hat{z}_t)$ using $\hat{W}$.

If instead the data are not i.i.d. or $(g(\mathbf{w}_t; \theta_0, \mathcal{F}_t)_{t\in\mathbb{Z}}$ is not a martingale difference sequence, then S will be the long run variance of $(g(\mathbf{w}_t; \theta_0, \mathcal{F}_t)_{t\in\mathbb{Z}}$,

$$\sqrt{T} g_T(\theta_0, z_t)) \xrightarrow{p} N(0, S), \text{ where } S = \sum_{j=-\infty}^{\infty} \Gamma_j = \Gamma_0 + \sum_{j=1}^{\infty} (\Gamma_j + \Gamma_j'), \tag{15}$$

and where

$$\Gamma_j = \mathbb{E}[g(\mathbf{w}_t; \theta_0) g(\mathbf{w}_{t-j}; \theta_0)'].$$

One estimator is the Newey et al. (1987) long-run variance estimator

$$\hat{S} = \hat{\Gamma_0} + \sum_{j=1}^{J_n} \left(1 - \frac{j}{J_n + 1}\right) (\hat{\Gamma}_j + \hat{\Gamma}_j') \tag{16}$$

$$\hat{\Gamma}_j = \frac{1}{T} \sum_{t=j+1}^{T} g(y_t, x_t, \hat{z}_t; \hat{\theta}) g(y_t, x_t, \hat{z}_t; \hat{\theta})',$$

where the number of lags $J_n \to \infty$ as $n \to \infty$.

Proposition 11. Let Assumptions 5, 8 and 9 hold, and suppose that S satisfies Equation (15) and $\hat{S}$ satisfies Equation (15). Then

$$(\hat{G}'\hat{W}\hat{G})^{-1}\hat{G}'\hat{W}\hat{S}\hat{W}\hat{G}(\hat{G}'\hat{W}\hat{G})^{-1} \xrightarrow{p} (G'WG)^{-1}G'WSWG(G'WG)^{-1}.$$

Just like when the data are i.i.d. we can obtain the efficient estimator using the standard two step GMM procedure.

Proposition 12. Under the assumptions of Proposition 11, and $\dot{\theta} \xrightarrow{p} \theta_0$

$$\hat{W} \xrightarrow{p} S^{-1},$$

where $\hat{W}$ is defined in Equation (14).

B.1 Proofs

B.1.1 Proof of Proposition 9

Proof. First note that if

1. $\mathbf{w}_1, \ldots, \mathbf{w}_T$ be strictly stationary and ergodic, which is verified by Assumption 8.2.
2. Θ be compact, which is verified by Assumption 7.3.
3. $g(\mathbf{w}_t;\theta)g(\mathbf{w}_t;\theta)'$ is continuous in θ for all $\mathbf{w}_t$, which is verified by Assumption 8.6.
4. $\mathbb{E}\left[\sup_{\theta\in N_{\theta_0}, z\in N_z}\|g(\mathbf{w}_t;\theta)g(\mathbf{w}_t;\theta)'\|\right] < \infty$, which is Assumption 9.

Then

$$\sup_{\theta\in N_\theta, z\in N_z}\|\frac{1}{T}\sum_{t=1}^{T} g(z_t;\theta)g(z_t;\theta)' - \mathbb{E}[g(z_t;\theta)g(z_t;\theta)']\| \xrightarrow{a.s.} 0 \tag{17}$$

Our goal is to show that

$$(\hat{G}'\hat{W}\hat{G})^{-1}\hat{G}'\hat{W}\hat{S}\hat{W}\hat{G}(\hat{G}'\hat{W}\hat{G})^{-1} \xrightarrow{p} (G'WG)^{-1}G'WSWG(G'WG)^{-1}.$$

We have already shown that $\hat{G} \xrightarrow{p} G$ in the proof of Proposition 8, and $\hat{W} \xrightarrow{p}$ by Assumption 8.4. Therefore if we show that $\hat{S} \xrightarrow{p} S$, then the result follow from the continuous mapping theorem.

By the triangle inequality

$$\begin{aligned}\|\hat{S} - S\| &= \|\frac{1}{T}\sum_{t=1}^{T} g(\hat{z}_t;\hat{\theta})g(\hat{z}_t;\hat{\theta})' - \mathbb{E}[g(z_t;\theta_0)g(z_t;\theta_0)']\| \\ &\leqslant \|\frac{1}{T}\sum_{t=1}^{T} g(\hat{z}_t;\hat{\theta})g(\hat{z}_t;\hat{\theta})' - \mathbb{E}[g(z_t;\hat{\theta})g(z_t;\hat{\theta})']\| \\ &+\|\mathbb{E}[g(z_t;\hat{\theta})g(z_t;\hat{\theta})'] - \mathbb{E}[g(z_t;\theta_0)g(z_t;\theta_0)']\|\end{aligned}$$

The second term is $o_p(1)$ by standard arguments. We will focus on bounding the first term. By the triangle inequality

$$\begin{aligned}\|\frac{1}{T}\sum_{t=1}^{T} g(\hat{z}_t;\hat{\theta})g(\hat{z}_t;\hat{\theta})' - \mathbb{E}[g(z_t;\hat{\theta})g(z_t;\hat{\theta})']\| \leqslant \\ \|\frac{1}{T}\sum_{t=1}^{T} g(\hat{z}_t;\hat{\theta})g(\hat{z}_t;\hat{\theta})' - \mathbb{E}[g(\hat{z}_t;\hat{\theta})g(\hat{z}_t;\hat{\theta})']\| \\ +\|\mathbb{E}[g(\hat{z}_t;\hat{\theta})g(\hat{z}_t;\hat{\theta})'] - \mathbb{E}[g(z_t;\hat{\theta})g(z_t;\hat{\theta})']\|\end{aligned}$$

The first term is $o_{a.s.}(1)$ by (17). As for the second term, using Assumption 9, we can apply

the dominated convergence theorem

$$\lim_{n\to\infty} \mathbb{E}[g(\hat{z}_t;\hat{\theta})g(\hat{z}_t;\hat{\theta})'] - \mathbb{E}[g(z_t;\hat{\theta})g(z_t;\hat{\theta})'] =$$
$$\mathbb{E}\left[\lim_{n\to\infty} g(\hat{z}_t;\hat{\theta})g(\hat{z}_t;\hat{\theta})'\right] - \mathbb{E}[g(z_t;\hat{\theta})g(z_t;\hat{\theta})'] = 0$$

Where the last equality follows from $\hat{z}_t \xrightarrow{p} z_t$ and the continuous mapping theorem. Hence

$$\|\frac{1}{T}\sum_{t=1}^{T} g(\hat{z}_t;\hat{\theta})g(\hat{z}_t;\hat{\theta})' - \mathbb{E}[g(z_t;\hat{\theta})g(z_t;\hat{\theta})']\| \leqslant o_{a.s.}(1) + o_p(1).$$

□

B.1.2 Proof of Proposition 10

Proof. This is identical to the proof that $\hat{S} \xrightarrow{p} S$ in Proposition 9, except that we exchange one consistent estimator $\hat{\theta}$ for another $\dot{\theta}$. □

B.1.3 Proof of Proposition 11

Proof. Similar to the proof of Proposition 9 it will suffice to show that $\hat{S} \xrightarrow{p} S$. Define

$$\hat{S}(z) = \hat{\Gamma_0}(z) + \sum_{j=1}^{J_n}\left(1 - \frac{j}{J_n+1}\right)(\hat{\Gamma}_j(z) + \hat{\Gamma}_j(z)')$$
$$\hat{\Gamma}_j(z) = \frac{1}{T}\sum_{t=j+1}^{T} g(y_t, x_t, z;\hat{\theta})g(y_t, x_t, z;\hat{\theta})'.$$

By the triangle inequality

$$|\hat{S}(\hat{z}_t) - S| \leqslant |\hat{S}(\hat{z}_t) - \hat{S}(z_t)| + |\hat{S}(z_t) - S|$$

The second term is $o_p(1)$ by standard arguments. Therefore we need to show that $|\hat{S}(\hat{z}_t) - \hat{S}(z_t)| = o_p(1)$.

$$|\hat{S}(\hat{z}_t) - \hat{S}(z_t)| = |\hat{\Gamma_0}(\hat{z}_t) + \sum_{j=1}^{J_n}\left(1 - \frac{j}{J_n+1}\right)(\hat{\Gamma}_j(\hat{z}_t) + \hat{\Gamma}_j(\hat{z}_t)')$$
$$- \left(\hat{\Gamma_0}(z_t)\sum_{j=1}^{J_n}\left(1 - \frac{j}{J_n+1}\right)(\hat{\Gamma}_j(z_t) + \hat{\Gamma}_j(z_t)')\right)|$$

$$= |\hat{\Gamma_0}(\hat{z}_t) - \hat{\Gamma_0}(z_t) + \sum_{j=1}^{J_n} \left(1 - \frac{j}{J_n + 1}\right) ((\hat{\Gamma}(\hat{z}_t) - \hat{\Gamma}_j(z_t)) + (\hat{\Gamma}(\hat{z}_t) - \hat{\Gamma}_j(z_t)'))|$$

$$\leqslant |\hat{\Gamma_0}(\hat{z}_t) - \hat{\Gamma_0}(z_t)| + \sum_{j=1}^{J_n} \left(1 - \frac{j}{J_n + 1}\right) (|\hat{\Gamma}(\hat{z}_t) - \hat{\Gamma}_j(z_t)| + |\hat{\Gamma}(\hat{z}_t)' - \hat{\Gamma}_j(z_t)'|)$$

Therefore it suffices to show that $\hat{\Gamma}_j(\hat{z}_t) \xrightarrow{p} \hat{\Gamma}_j(z_t)$ for $j = 0, \ldots, J_n$, which follows from the continuous mapping theorem as $\hat{z}_t \xrightarrow{p} z_t$ for each t. □

B.1.4 Proof of Proposition 12

Proof. This is identical to the proof of Proposition 10. □

C Verifying Extra Assumptions

In this appendix, we derive the restrictions placed on the models in our empirical application by Assumptions 4 and 5

C.1 Linear IV

The moment conditions are

$$g(y_t, x_t, z_t, \theta) = z_t(y_t - x_t'\beta_0 - Y_t\theta_0) = 0, \tag{18}$$

and the model parameters are $\theta := [\beta_0, \theta_0]'$.

Recall Assumption 4 was

$$\sup_{\theta\in\Theta, z\in N_{z_t}} \| \frac{1}{T}\sum_{t=1}^{T} \frac{\partial g(y_t, x_t, z, \theta)}{\partial z} \| = O_p(1).$$

Equation (18) has first derivative w.r.t. z, $y_t - x_t'\beta - Y_t\theta$ and hence

$$\|\mathbb{E}\left[y_t - x_t'\beta - Y_t\theta\right]\| \leqslant \mathbb{E}\|y_t\| + \mathbb{E}\|x_t'\|\|\beta\| + \mathbb{E}\|Y_t\|\|\theta\|,$$

which is bounded provided $\mathbb{E}[\|y_t\|] < \infty$, $\mathbb{E}[\|x_t'\|] < \infty$, $\mathbb{E}[\|Y_t\|] < \infty$, $\sup_\beta\|\beta\| < \infty$, $\sup_\theta\|\theta\| < \infty$. Assume that the first moments of the random variables is bounded is not controversial, and is required for consistency and asymptotic normality of IV estimators. The parameters being bounded requires a compactness assumption.

Recall Assumption 4 was

$$\sup_{z\in N_{\hat{z}_t}} \| \frac{1}{T}\sum_{t=1}^{T} \frac{\partial^2 g_j(y_t, x_t, z, \hat{\theta})}{\partial\theta\partial z'} \| = O_p(1),$$

And second cross derivative of Equation (18) is $[-x_t, -Y_t]'$, and therefore we simply require $\mathbb{E}[\|x_t\|] < \infty$ and $\mathbb{E}[\|Y_t\|] < \infty$.

C.2 SVAR

Equation (11) leads to the moment condition

$$\mathbb{E}\left[z_t \left(y_t - \sum_{j=1}^{p} \boldsymbol{A}_j y_{t-j} \right) - \alpha\theta_0 \right] = 0. \tag{19}$$

And the VAR slope parameters, $\text{vec}(\boldsymbol{A}_1, \ldots, \boldsymbol{A}_p)$ satisfy the moment conditions

$$\mathbb{E}\left[(y'_{t-1}, \ldots, y'_{t-p})' \otimes \left(y_t - \sum_{j=1}^{p} \boldsymbol{A}_j y_{t-j}\right)\right] = 0. \tag{20}$$

The model parameters are $\theta := [A_1, \ldots, \boldsymbol{A}_p, \theta_0]'$.

Recall Assumption 4 was

$$\sup_{\theta \in \Theta, z \in N_{\hat{z}_t}} \| \frac{1}{T} \sum_{t=1}^{T} \frac{\partial g(y_t, x_t, z, \theta)}{\partial z} \| = O_p(1).$$

Equation (19) has first derivative w.r.t. z equal to $y_t - \sum_{j=1}^{p} \boldsymbol{A}_j y_{t-j}$. Therefore

$$\| \mathbb{E}\left[y_t - \sum_{j=1}^{p} \boldsymbol{A}_j y_{t-j} \right] \| \leqslant E[\|y_t\|] + \|\boldsymbol{A}_1\| \mathbb{E}[\|y_{t-1}\|] + \|\boldsymbol{A}_p\| \mathbb{E}[\|y_{t-p}\|] < \infty.$$

Where $\mathbb{E}[\|y_{t-j}\|] < \infty$ for $j = 0, \ldots, p$ and $\|\boldsymbol{A}_j\| < \infty$ for $j = 1, \ldots, p$ follow from covariance stationarity. Equation (20) is not a function of z_t, it will trivially satisfy Assumption 4.

Recall Assumption 4 was

$$\sup_{z \in N_{\hat{z}_t}} \| \frac{1}{T} \sum_{t=1}^{T} \frac{\partial^2 g_j(y_t, x_t, z, \hat{\theta})}{\partial \theta \partial z'} \| = O_p(1),$$

The cross derivative for Equation (19) is $[-y_{t-1}, \ldots, y_{t-p}, 0]$, and hence the conditions bounding the first moment of the y_t will suffice. Equation (20) is satisfies this trivially.