Text as Instruments

James Nesbit

November 1, 2019

Motivation

Text is an increasing popular input to economics research
Implicit assumption: common latent variables drives text + data

- Policy makers in FOMC meetings
- Economic news in newspapers

Link used **informally** to infer latent variables in quantitative econometric models

- Romer, Romer (2004, 2010): narrative identification of exogenous monetary / fiscal policy
- Baker, Bloom, Davis (2016): Economic Policy Uncertainty (EPU)

This Paper

Model of text generation formalizing link between text and latent variables from quantiative econometric model

- Random utility model where speaker chooses word given observables (**context**)
- Utility of word, context pairs will be function of latent variables

Use this model to

1. Find weak sufficient conditions for utilities to be used as instruments
2. Formally justify the use of conditional probabilities (or transforms thereof) as instruments
3. Compare and contrast previous approaches: EPU

Outline

1. Model
2. Results
3. Past approaches: EPU
4. Empirical Application: Monetary policy shocks

Connection to Literature

Using text to infer latent shocks

- Friedman, Schwartz (1963): monetary policy
- Romer, Romer (2004): monetary; (2010): fiscal
- Baker, Bloom, Davis (2016): EPU

Using random utility to model text

- Groseclose, Milyo (2005): media bias
- Taddy (2013, 2015)
- Gentzkow, Taddy, Shapiro (2019): congressional speech

Latent Variables

Econometric model: two latent variables: $\epsilon_t = [\epsilon_{1,t}, \epsilon_{2,t}]'$

- $\mathbb{E}[\epsilon_t] = 0$
- $var(\epsilon_t) = I_2$
- $\epsilon_{1,t} \perp\!\!\!\perp \epsilon_{2,t}$

$\epsilon_{1,t}$ is our **target** latent variable

$\epsilon_{2,t}$ is a **nuisance** latent variable

Instruments

Instrument Z_t for $\epsilon_{1,t}$ satisfies

- Relevance $\mathbb{E}[Z_t \epsilon_{1,t}] \neq 0$
- Exogeneity $\mathbb{E}[Z_t \epsilon_{2,t}] = 0$

Running example: external instruments identifying monetary policy shock in SVAR

- $y_t = [i_t, \pi_t]'$
- $\epsilon_{1,t}$ is a monetary policy shock
- $\epsilon_{2,t}$ is a inflation shock

Text Model

Collection of T documents

Document t is a collection of N_t words $\{w_1, \cdots, w_{N_t}\}$

Word $\in V$: vocabulary of **terms**

Context c is some observable event

- the previous word w_{n-1}
- term v appears in document t
- element of sub-σ-algebra

A speaker choosing word w, given context c, in document t at (word) position n

$$U_{w,c,t,n} = \alpha_{0,w,c} + \alpha_{1,w,c}\epsilon_{1,t} + \alpha_{2,w,c}\epsilon_{2,t} + e_{w,c,t,n}$$

Flexible Model for Data Generating Text

$$U_{w,c,t,n} = \underbrace{\alpha_{0,w,c} + \alpha_{1,w,c}\epsilon_{1,t} + \alpha_{2,w,c}\epsilon_{2,t}}_{u_{w,c,t}} + e_{w,c,t,n}$$

- Lots of flexibility
- Lots of parameters to estimate (will avoid this)

Assume $e_{w,c,t,n}$ is i.i.d. Gumbel, leads to familiar logit form for conditional probabilities

$$p_t(w|c) = \frac{\exp\{u_{w,c,t}\}}{\sum_{w'=1}^{V} \exp\{u_{w',c,t}\}}$$

An Example

> *The Committee expects inflation to moderate in coming quarters, but it will be necessary to continue to monitor inflation developments carefully.*

FOMC statement - January 30, 2008

An Example

> *the committee expects inflation to moderate in coming quarters, but it will be necessary to continue to monitor inflation developments carefully.*

Lower case

An Example

> *the committee expects inflation to moderate in coming quarters but it will be necessary to continue to monitor inflation developments carefully*

Remove punctuation

An Example

[‘the’, ‘committee’, ‘expects’, ‘inflation’, ‘to’, ‘moderate’, ‘in’, ‘coming’, ‘quarters’, ‘but’, ‘it’, ‘will’, ‘be’, ‘necessary’, ‘to’, ‘continue’, ‘to’, ‘monitor’, ‘inflation’, ‘developments’, ‘carefully’]

Tokenize

An Example

[‘committee’, ‘expects’, ‘inflation’, ‘moderate’, ‘coming’, ‘quarters’, ‘necessary’, ‘continue’, ‘monitor’, ‘inflation’, ‘developments’, ‘carefully’]

Remove stopwords

An Example

[‘committe’, ‘expect’, ‘inflat’, ‘moder’, ‘come’, ‘quarter’, ‘necessari’, ‘continu’, ‘monitor’, ‘inflat’, ‘develop’, ‘care’]

Stem words

An Example

['committe', 'expect', 'inflat', 'moder', 'come', 'quarter', 'necessari', 'continu', 'monitor', 'inflat', 'develop', 'care']

(c = 'monitor', w = 'inflat')

Utility for 'inflat' given 'monitor' is

$$u_{\text{'inflat'},\text{monitor'},t} = \alpha_{0,\text{'inflat'},\text{'monitor'}} + \alpha_{1,\text{'inflat'},\text{'monitor'}}\epsilon_{1,t} + \alpha_{2,\text{'inflat'},\text{'monitor'}}\epsilon_{2,t}$$

We would expect

- Not related to monetary policy shock: $\alpha_{1,\text{'inflat'},\text{'monitor'}} = 0$
- Related to inflation shock: $\alpha_{2,\text{'inflat'},\text{'monitor'}} > 0$

Assumptions

Assumption (Weak)

There exist two words w^* and context c^* such that

1. $\alpha_{1,w^*,c^*} \neq 0$
2. $\alpha_{2,w^*,c^*} = 0$

1. w^* in context c^* is related to the target shock
2. w^* in context c^* is not related to the nuisance shock

Consider

- $w^* =$ 'interest'
- $c^* =$ 'inflat' $\notin \{w_{n-10}, \cdots, w_{n+10}\}$

Utilities are an Instrument

Proposition

Under Assumption (Weak), $u_{w^*,c^*,t}$ is an instrument for $\epsilon_{1,t}$.

Proof:

$$Z_t = u_{w^*,c^*,t} = \alpha_{0,w^*,c^*} + \underbrace{\alpha_{1,w^*,c^*}}_{\neq 0} \epsilon_{1,t} + \underbrace{\alpha_{2,w^*,c^*}}_{=0} \epsilon_{2,t}$$

Relevance:

$$\mathbb{E}[Z_t \epsilon_{1,t}] = \alpha_{1,w^*,c^*} \neq 0$$

Exogenity:

$$\mathbb{E}[Z_t \epsilon_{2,t}] = \alpha_{1,w^*,c^*} \mathbb{E}[\epsilon_{1,t} \epsilon_{2,t}] = 0$$

Slightly Stronger Assumptions

Assumption (Stronger)

There exist two words w^*, w^{**} and shared context c^* such that

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

1. At least one of w^* and w^{**} in context c^*, are functions of target shock
2. Both w^* and w^{**} in context c^*, are not functions of nuisance shock
3. w^* and w^{**} in context c^* cannot have the same response to target shock

Slightly Stronger Assumptions

1. At least one of w^* and w^{**} in context c^*, are functions of target shock
2. Both w^* and w^{**} in context c^*, are not functions of nuisance shock
3. w^* and w^{**} in context c^* cannot have the same response to target shock

Consider

- $w^* =$ 'interest'
- $w^{**} =$ 'reserv'
- $c^* =$ 'inflat' $\notin \{w_{n-10}, \cdots, w_{n+10}\}$

Log Odds are an Instrument

Proposition

Under Assumption (Stronger), $\log \frac{p_t(w^*|c^*)}{p_t(w^{**}|c^*)}$ is an instrument for $\epsilon_{1,t}$.

Proof:

$$
\begin{aligned}
Z_t &= \log \frac{p_t(w^*|c^*)}{p(w^{**}|c^*, \epsilon_t)} \\
&= u_{w^*,c^*,t} - u_{w^{**},c^*,t} \\
&= (\alpha_{0,w^*,c^*} - \alpha_{0,w^{**},c^*}) + \underbrace{(\alpha_{1,w^*,c^*} - \alpha_{1,w^{**},c^*})}_{\neq 0} \epsilon_{1,t} \\
&+ \underbrace{(\alpha_{2,w^*,c^*} - \alpha_{2,w^{**},c^*})}_{=0} \epsilon_{2,t}
\end{aligned}
$$

Probabilities are Not Instruments

Proposition

Under Assumption (Weak) $p_t(w^*|c^*)$ is **not** an instrument for $\epsilon_{1,t}$.

$$
\begin{aligned}
Z_t &= p_t(w^*|c^*) \\
&= \frac{\exp\{\alpha_{0,w^*,c^*} + \alpha_{1,w^*,c^*}\epsilon_{1,t}\}}{\sum_{w'=1}^{V}\exp\{\alpha_{0,w',c^*} + \alpha_{1,w',c^*}\epsilon_{1,t} + \alpha_{2,w',c^*}\epsilon_{2,t}\}}
\end{aligned}
$$

Write $p_t(w^*|c^*) = g(\epsilon_{1,t}, \epsilon_{2,t})$

Exogeneity:

$$\mathbb{E}[Z_t\epsilon_{2,t}] = \mathbb{E}[g(\epsilon_{1,t}, \epsilon_{2,t})\epsilon_{2,t}] \neq 0$$

Intuition

$$p_t(w^*|c^*) = \frac{\exp\{\alpha_{0,w^*,c^*} + \alpha_{1,w^*,c^*}\epsilon_{1,t}\}}{\sum_{w'=1}^{V}\exp\{\alpha_{0,w',c^*} + \alpha_{1,w',c^*}\epsilon_{1,t} + \alpha_{2,w',c^*}\epsilon_{2,t}\}}$$

The reason that we fail to satisfy exogeneity is that $\alpha_{2,w',c^*} \neq 0$, for some w'

- Words in same context that are related to nuisance shock

Given context c^*

- Choose w^* when $\epsilon_{1,t}$ is high $\implies$ choose w' less
- Choose w' when $\epsilon_{2,t}$ is high $\implies$ choose w^* less

$p_t(w^*|c^*)$ is dependent on $\epsilon_{2,t}$

Sufficiently Strong Assumptions

Assumption (Strong)

Suppose exists a word w^* such that

1. $\alpha_{1,w,c^*} \neq 0$ for $w = w^*$
2. $\alpha_{2,w,c^*} = 0$ for all w

1. w^* is relevant for target shock
2. The nuisance shock cannot 'share the same context' as w^*

Would require a context c' such that $\alpha_{2,\text{'inflat'},c'} = 0$

Proposition

Under Assumption (Strong) $p_t(w^*|c^*)$ is an instrument for $\epsilon_{1,t}$.

EPU Index

We search the digital archives of each paper from January 1985 to obtain a 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'). In other words, to meet our criteria, an article must contain terms in all three categories pertaining to uncertainty, the economy, and policy.*

EPU Index

Assume they are estimating $p_t(w^*|c^*)$

- $w^* =$ 'uncertain'
- $c^* =$ 'economy' and at least one of 'congress', 'deficit', 'Federal Reserve', ... appear in document t

Estimate a VAR

- In order: EPU index, S&P 500 index, federal funds rate, employment, industrial production
- Use a Cholesky factorization to identify their shocks
- Implies that EPU index responds only to contemporaneous EPU shocks

EPU Index

Implicit assumption of this ordering has the flavour of Assumption (Strong)

1. $\alpha_{1,w,c^*} \neq 0$ for $w = w^*$
2. $\alpha_{2,w,c^*} = 0$ for all w

One of their nuisance shocks is a monetary policy shock

c^* = ‘economy‘ and “the Fed” appear in document t

- $\alpha_{2,\text{‘rates’},c^*} \neq 0$, which violates 2.

Empirical Application

Simple 3-variable monetary SVAR: Kilian, Lütkepohl (2017)

$$y_t = (\pi_t, y_t, i_t).$$

SVAR is not identified without further restrictions

- Stock, Watson (2008, 2012) show that we can identify structural shock with instrument

Identify monetary policy shock from transcripts of FOMC meetings

Empirical Application

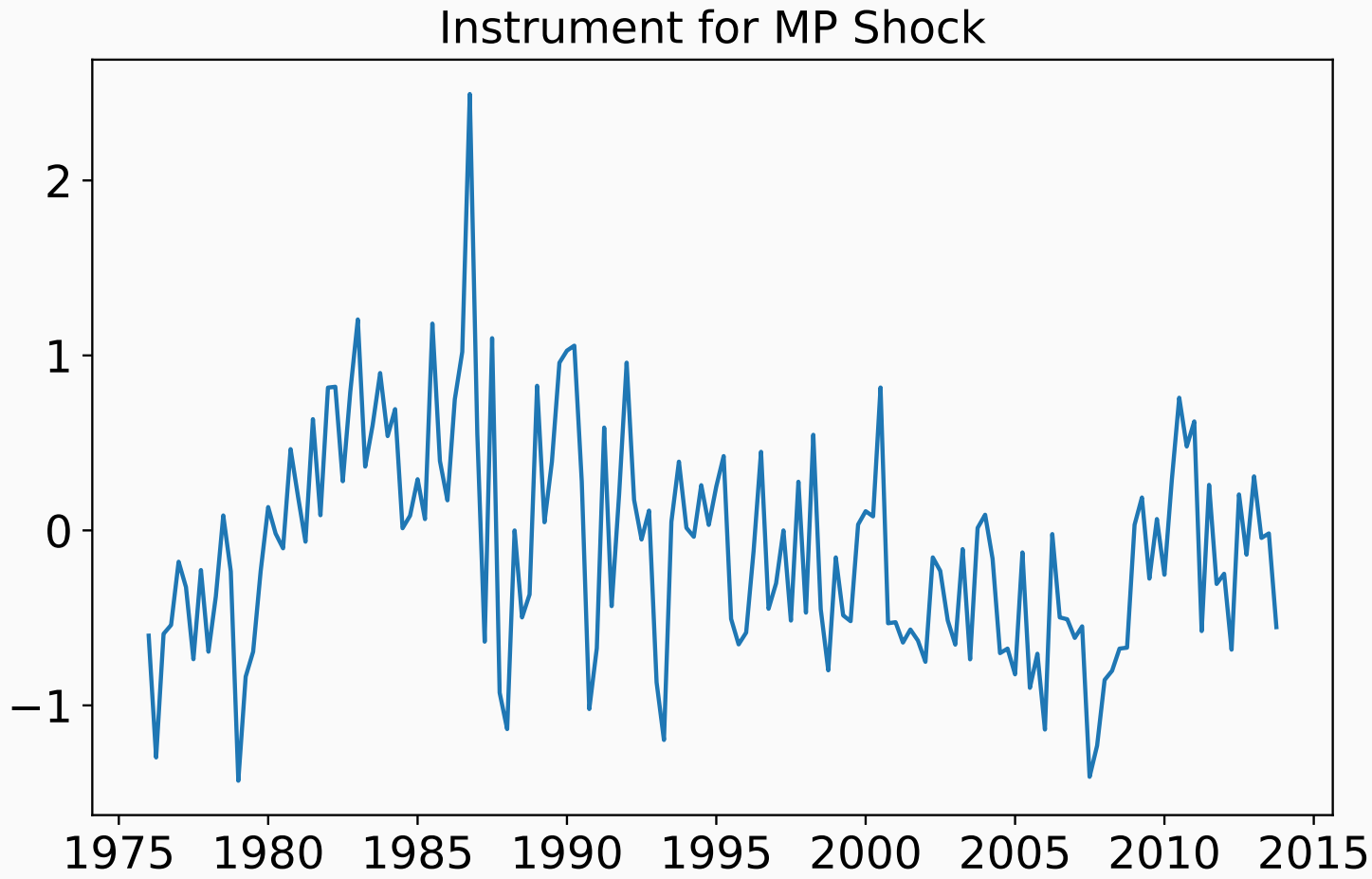
FOMC transcripts from 1976–2013

318 meetings, $5M+$ words, $30K+$ terms

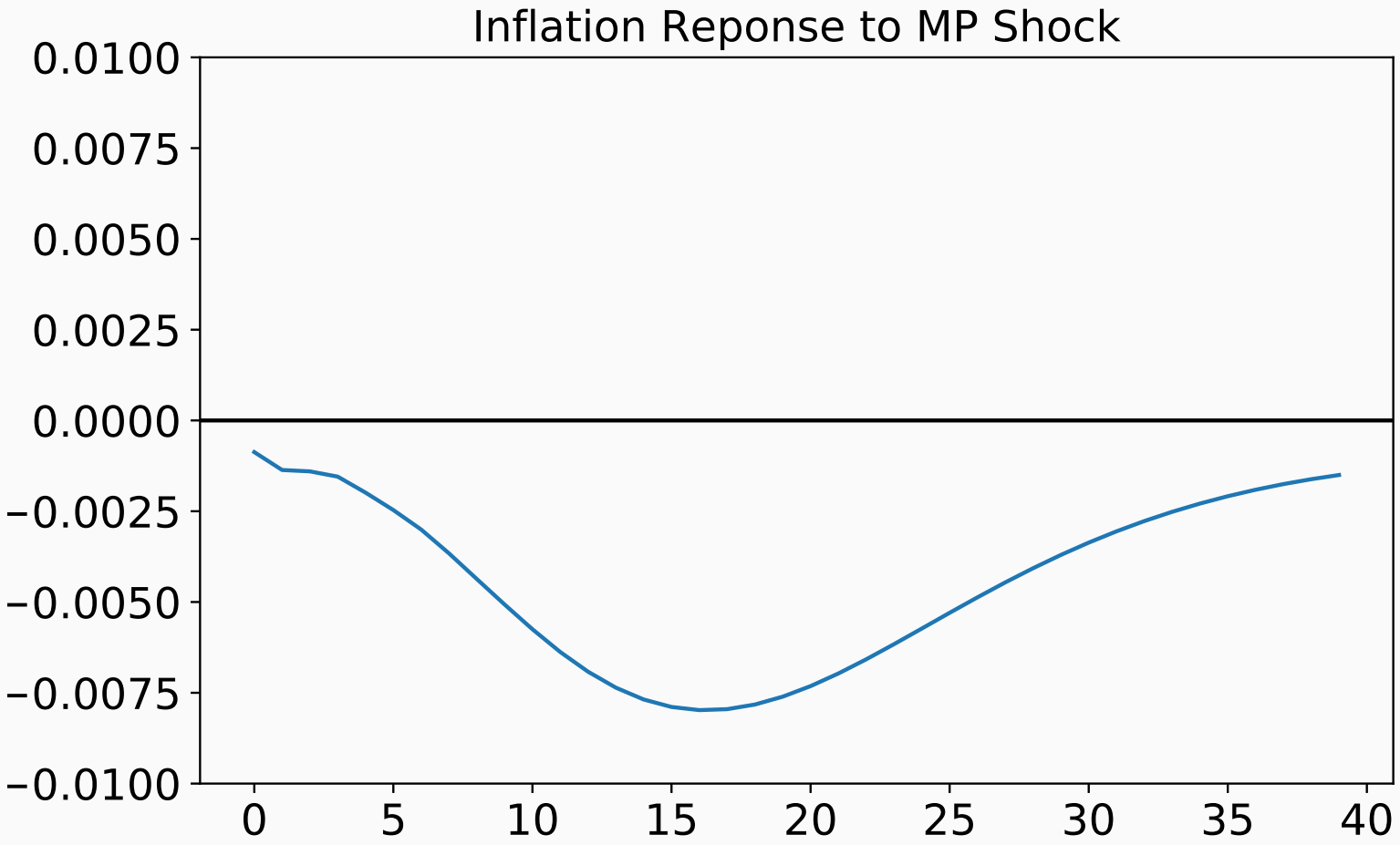
Instrument

- $w^* =$ 'interest'
- $w^{**} =$ 'funds'
- $c^* = \{\text{'output'}, \text{'inflat'}\} \notin \{w_{n-10}, \cdots, w_{n+10}\}$

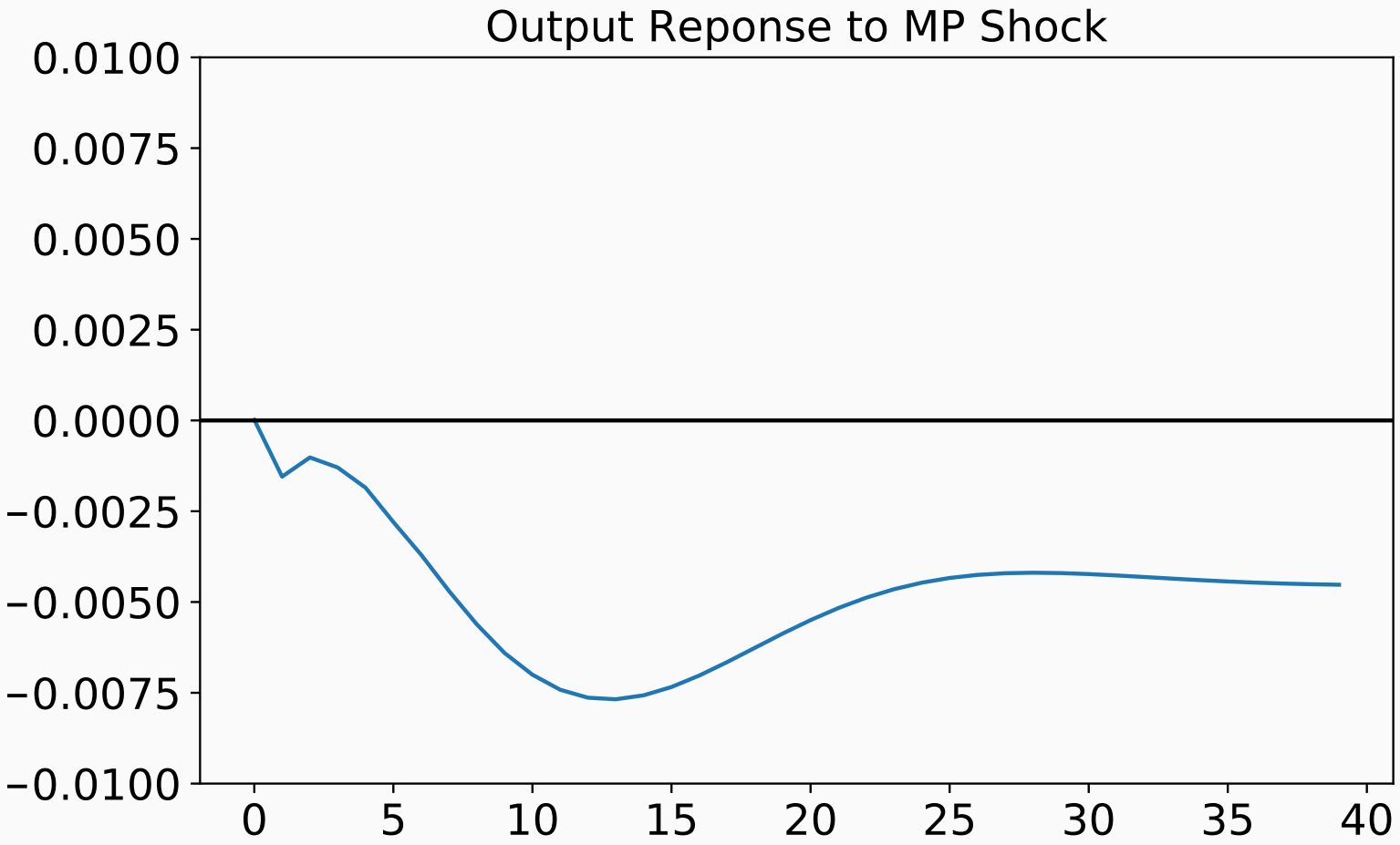
Instrument Time Series



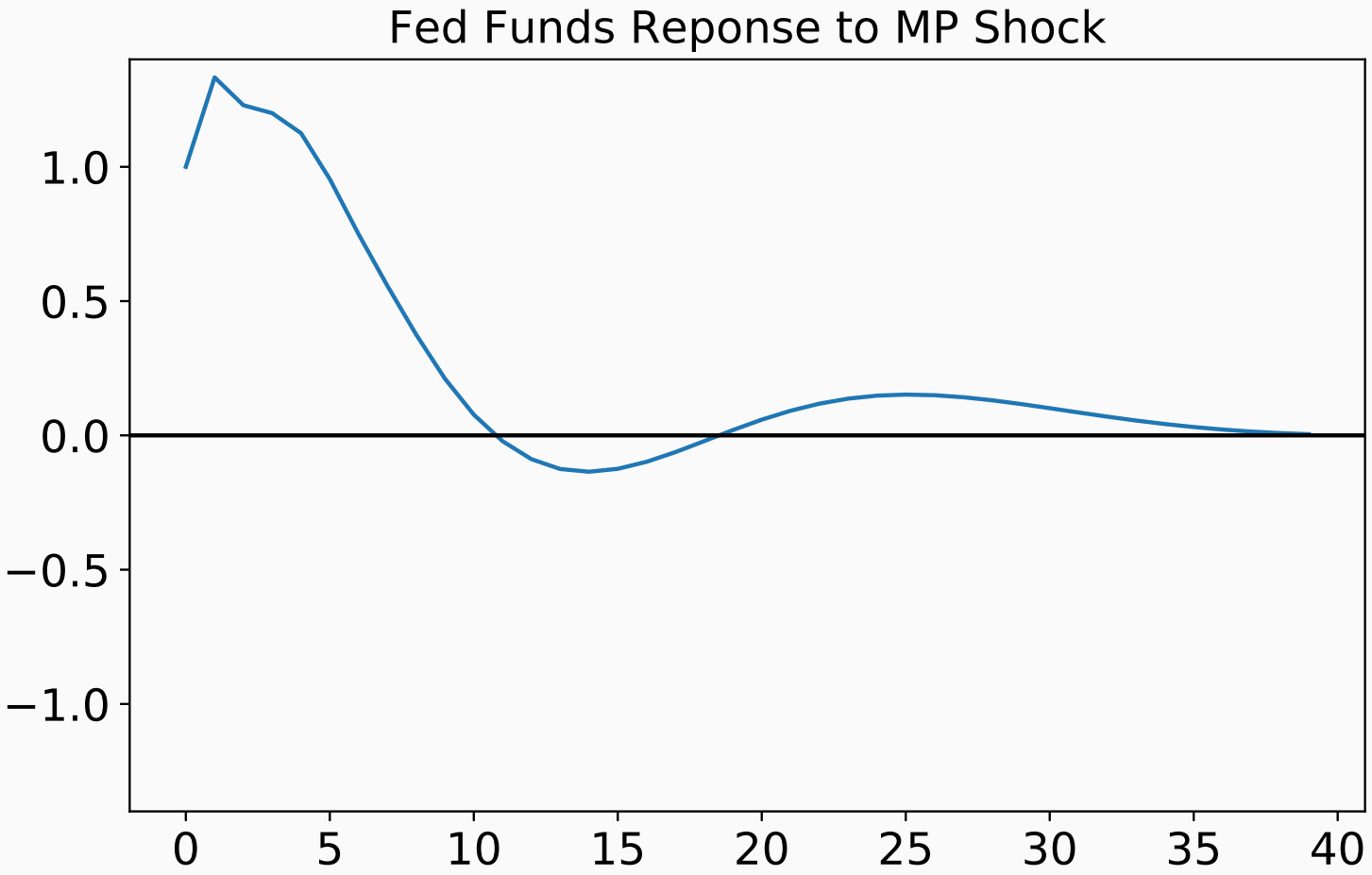
Response of Inflation to MP Shock



Response of Output to MP Shock



Response of Fed Funds to MP Shock



What I've Done

Construct model of text and data

- Discrete choice model where words are chosen given context

Find sufficient conditions for instruments

- Utility of a word, context pair that is related to target shock, orthogonal to nuisance shock

Compare and contrast with existing methods

- EPU index

Identify monetary policy shocks in an SVAR

Some Possible Paths Forward

How to incorporate synonym information

- Nested logit?

Many possible candidates for instruments

- Optimal efficient instruments
- Data-driven approach to find instrument under stronger DGP assumptions?

Relax model's parametric assumptions

- Gumbel errors
- Linear form of utility

Other empirical applications

- Revisit EPU
- Stock news shocks using sentiment