LIETUVOS BANKAS
EUROSISTEMA

Text Data Analysis Using Latent Dirichlet Allocation: An Application to FOMC Transcripts

Discussion Paper Series

No. 11 / 2019

ISSN 2345-0835 (online)

Discussion Paper Series

No. 11 / 2019

Hali Edison

(Williams College)

Hector Carcel

(Bank of Lithuania)

April 2019[1]

[1] Special thanks to C. Schwarz for providing the Stata LDA codes and assistance with them.

Gedimino pr. 6, LT-01103 Vilnius

www. lb.lt

Discussion Papers describe research in progress by the author(s) and are published to stimulate discussion and critical comments.

The series is managed by the Applied Macroeconomic Research Division of the Economics Department and the Center for Excellence in Finance and Economic Research.

The views expressed are those of the author(s) and do not necessarily represent those of the Bank of Lithuania.

ABSTRACT

This paper applies Latent Dirichlet Allocation (LDA), a machine learning algorithm, to analyze the transcripts of the U.S. Federal Open Market Committee (FOMC) covering the period 2003 – 2012, including 45,346 passages. The goal is to detect the evolution of the different topics discussed by the members of the FOMC. The results of this exercise show that discussions on economic modelling were dominant during the Global Financial Crisis (GFC), with an increase in discussion of the banking system in the years following the GFC. Discussions on communication gained relevance toward the end of the sample as the Federal Reserve adopted a more transparent approach. The paper suggests that LDA analysis could be further exploited by researchers at central banks and institutions to identify topic priorities in relevant documents such as FOMC transcripts.

Keywords: FOMC, Text data analysis, Transcripts, Latent Dirichlet Allocation.
JEL codes: E52, E58, D78.

1. INTRODUCTION

Text data analysis can be a useful tool for disentangling and analyzing the main topics in different kinds of documents. In a recent study, Hartmann and Smets (2018) analyze the topics addressed by ECB Board members in their public speeches during the period 1999-2017. The authors employ Latent Dirichlet Allocation (LDA), a machine learning algorithm developed by Blei et al. (2003) which allows for the automatic clustering of 1,892 ECB Board speeches. The aim of the current paper is to introduce the LDA methodology as presented in Schwarz (2018) and obtain results using the ldagibbs Stata command. We analyze 45,346 entries or passages of the Federal Open Market Committee (FOMC) during the period 2003-2012, with the goal of identifying the evolution of the different topics discussed by the members of the FOMC. Overall, we find that discussions on economic modelling were dominant during the Great Financial Crisis (GFC), followed by a considerable increase in discussions on the banking system in the following years; discussions on communication gained importance in the most recent years for which FOMC transcripts are available.

Previous papers have applied an array of techniques to analyze the FOMC transcripts addressing a variety of questions, including the use of models (Edison and Marquez, 1998), the use of Phillips curves (Meade and Thornton, 2012) or the Taylor rule (Asso et al., 2010) and the transparency of the transcripts (Hansen et al., 2018). The innovation of this short paper is to clearly explain the LDA algorithm and apply this machine learning technique to the analysis of the FOMC transcripts during the GFC.

2. FOMC MEETINGS

The most detailed record of FOMC meeting proceedings is the transcript. For meetings before 1994, the transcripts were produced from the original, raw transcripts in the FOMC Secretariate files. From 1994 on, the FOMC Secretariat has produced transcripts shortly after each meeting from an audio recording of the proceedings. In these transcripts, the speakers' words are lightly edited, where necessary, to facilitate readability. Meeting participants are then given several weeks to review the transcript for accuracy. The Federal Reserve Act states that the objectives of monetary policy enhanced by the FOMC shall "promote effectively the goals of maximum employment, stable prices and moderate long-term interest rates". There exists considerable debate among economists on how to translate these goals into a coherent description of U.S. monetary policy. A precise account of the discussions that take place during FOMC meetings can shed light on the evolution of this policy.

The FOMC meets eight times a year in order to formulate Federal Reserve policies, including monetary policy. It is composed of nineteen members: seven governors of the Federal Reserve Board, located in Washington DC, one of whom is the chairperson of both the Board of Governors and the FOMC, as well as twelve presidents of the Regional Federal Reserve Banks. The president of the New York Fed traditionally serves as vice-chairperson of the FOMC. The main policy variable of the FOMC is a target for the Federal Funds rate, as well as potential guidance on future monetary policy. All seven governors vote at every meeting; the president of the New York Fed and four of the remaining eleven Fed presidents vote on a rotating basis. With the exception of FOMC meetings that take place before the Monetary Policy Report for the President, which span two days, each FOMC meeting takes place over the course of a single day. Members participate in the discussions independently from their voting right. In this paper we analyze the transcripts of these meetings, concentrating on the conversations held between the FOMC members.

3. METHODOLOGY

As explained in Schwarz (2018), Latent Dirichlet Allocation (LDA) is composed of two parts. The first is a probabilistic model describing the text data as a likelihood function. In the second part, given the unfeasibility of maximizing the likelihood function, LDA utilizes an inference algorithm. The probabilistic model of LDA considers that each document d of the D documents in the whole text can be described as a probabilistic combination of T topics. These probabilities are in a document vector θ_d of length T. The value of T, that is, the number of topics, is decided by the user according to the precision required. The outcome of LDA is a $D \times T$ matrix θ containing $P(t_t|d_d)$, with θ_1,..., θ_D being $1 \times T$ vectors, in such a way that the probability of document d belonging to topic t corresponds to:

$$\theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_D \end{pmatrix} = \begin{pmatrix} P(t_1|d_1) & \cdots & P(t_T|d_1) \\ \vdots & \ddots & \vdots \\ P(t_1|d_D) & \cdots & P(t_T|d_D) \end{pmatrix}.$$

Every topic $t \in T$ is determined by a probabilistic distribution over the vocabulary (the set of words in all documents) of size V. A topic will thus determine how likely it is to detect a word conditional on a topic. In our case, documents regarding forecasting will have a high probability of containing words such as “expectations” or “market-based”, while in documents related to economic modelling there will be a higher probability of finding terms such as “model”, “standard errors” or “shocks”. The word probability vectors of each of the topics can be represented in a matrix φ of dimensions $V \times T$:

$$\varphi \ = \ (\varphi_1, \ldots, \varphi_T) = \begin{pmatrix} P(w_1|t_1) & \cdots & P(w_1|t_T) \\ \vdots & \ddots & \vdots \\ P(w_v|t_1) & \cdots & P(w_v|t_T) \end{pmatrix}.$$

The probabilities $P(w_v|t_T)$ in φ_t describe how probable it is to detect word v from the vocabulary conditional on topic t. Hence, the φ_t vectors permit one to decide the content of each topic and how each topic can eventually be named, since LDA does not provide concrete topic labels. These need to be decided by the users in accordance with their knowledge of the subject under study.

Given the parameters θ and φ, the LDA probabilistic model considers that the whole data text is created by the following procedure. First, a word probability distribution is drawn following $\varphi \sim Dir(\beta)$. For each document d in the text, topic proportions are drawn following $\theta_d \sim Dir(\alpha)$. For each of the N_d words w_d, a topic assignment is drawn such that $z_{d,n} \sim Mult(\theta_d)$ and each word $w_{d,n}$ is drawn from $p(w_{d,n}|z_{d,n}, \varphi)$. In this model α and β, both bigger than 0, are hyperparameters required for the Gibbs sampling process. The likelihood of the whole text with respect to the model parameters is:

$$\prod_{d=1}^{D} P(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{d,n}} P(z_{d,n}|\theta_d) P(w_{d,n}|z_{d,n}, \varphi) \right).$$

$P(\theta_d|\alpha)$ describes how likely it is to observe the topic distribution of θ_d of document d conditional on α. The term $P(z_{d,n}|\theta_d)$ determines how likely the individual topic assignment $z_{d,n}$ of word n in document d is conditional on the topic distribution of the document. Ultimately, $P(w_{d,n}|z_{d,n}, \varphi)$ is the probability to detect a concrete word conditional on the topic assignment of the word and the word probabilities of the given topics which are contained in φ. By calculating the sum over all possible topic assignments, the product over all words in a document and the product over all documents in the text, we obtain the likelihood of observing the texts in the documents.

The LDA procedure is based on finding the optimal topic assignment $z_{d,n}$ for each word in each document and the optimal word probabilities φ for each topic that maximizes this likelihood. This would require adding up all possible topic assignments for all words in all documents, which is computationally impossible. Thus, alternative methods such as the Gibbs sampler have been developed for this purpose. In this work, such method is used following Griffiths and Steyvers (2004), based on the ldagibbs Stata command introduced by Schwarz (2018). Gibbs sampling consists of a Markov Chain Monte Carlo (MCMC) algorithm based on repeatedly drawing new samples conditional on all other data. In the case of LDA, the Gibbs sampler relies on updating the topic assignment of words conditional on the topic assignments of all other words. As Gibbs Sampling is a Bayesian technique, it requires priors for the values of the hyperparameters α and β, which lie within the unit interval. The prior for α is chosen based on the number of topics T while the prior for β depends on the size of the vocabulary.

Firstly, the ldagibbs algorithm divides the document into single words or word tokens. These are randomly assigned to one of the T topics with equal probability, which gives an initial assignment of words and thereby documents to topics for the sampling process. Later, ldaggibs samples new topic assignments for each of the word tokens, with the probability of a word token being assigned to topic t being:

$$P(\mathrm{z}_{d,n} = t | \mathrm{w}_{d,n}, \varphi) \propto P(w_{d,n} | z_{d,n} = t, \varphi) \cdot P(z_{d,n} = t).$$

The Gibbs Sampler makes use of the topic assignment of all other tokens in order to acquire approximate values for $P(z_{d,n} = t | w_{d,n}, \varphi)$ and $P(z_{d,n} = t)$. The probability $P(w_{d,n} | z_{d,n} = t, \varphi)$ is given by the number of words which are identical to $w_{d,n}$ and assigned to topic t divided by the total number of words assigned to that topic. In the first annex, we present a graphical illustration of the LDA with Plate notation based on the work by Srba et al. (2015).

4. ANALYSIS AND RESULTS

We used a total of 80 FOMC meeting transcripts, which covered all the meetings that took place between 2003 and 2012. We chose this time period to examine how the discussions at the FOMC meetings evolved leading up to the GFC, at the height of the GFC, and thereafter. A full set of minutes for each FOMC meeting is published three weeks after each regular meeting, but complete transcripts are published only five years after the meeting. It is these complete transcripts that we use in our analysis. We introduced the text of the FOMC transcripts into the Stata software database, dividing each of the transcripts into data text entries consisting of sentences or paragraphs stated by the governors during the meetings. A total of 45,346 discussion entries were analyzed, which covered all the conversations that took place between the FOMC members. Staff explanations were not included, thus spotlighting the predominant topics discussed by the governors during the meetings. The LDA algorithm was then implemented with the goal of splitting the whole text data into eight distinguishing topics. For this study, 500 burn–in periods and 1000 iterations were required. After a careful analysis of the data texts with highest probability of belonging to each topic, we decided that the topics corresponded to the following themes: Forecasting, Economic Modelling, Statement Language, Risks, Banking, Voting Decisions, Economic Activity and Communication. In the second annex, we provide the two text entries that had the highest probability of belonging to each of the topic groups. The topic labels were assigned after close inspection of these entries.

The evolution of the probability of each of the topics being addressed during this period at each of the meetings is graphically shown in Figure 1. Discussions on Economic Modelling were predominant during the GFC, with an increase in discussion of the banking system in the following years; in the most recent years, discussions on communication gained prominence. Figures 2 and 3 show the equivalent evolution in the number of data entries. A clear upward trend can be detected in the amount of text of the transcripts, showing that FOMC meetings have become more extensive.

5. CONCLUDING COMMENTS

The LDA algorithm can be easily implemented to analyze the different themes and their corresponding evolution in terms of use throughout time. In this paper, we have explained the algorithm, its implementation and estimation, and we have provided an empirical example by analyzing the FOMC transcripts covering the meetings that took place during the most recent ten-year-period for which the material is available. The LDA algorithm, and, in particular, the Stata command Idagibbs introduced by Schwarz (2018), can be readily used to detect which topics receive the most attention in a large number of documents. Here, we have presented the case of FOMC transcripts, applying the algorithm to more than 45,000 text data entries and obtaining the evolution of eight identified different topics. We observed that discussions on economic modelling were central during the GFC, followed by an increase in discussion of the banking system in the following years; discussions on communication gained relevance in the most recent years.

This type of analysis could be further utilized by researchers at central banks and institutions to determine topic priorities in relevant documents such as FOMC transcripts. The authors of this short paper expect to carry out further research to investigate the evolution of concrete economic terminology (e.g., Phillips curve, Taylor rule, etc.) in these transcripts, as well as shifts in their general tone.

REFERENCES

Asso, P.F., G. A. Kahn and R. Leeson (2010) The Taylor rule and the practice of central banking, Research Working Paper 10-05, Federal Reserve Bank of Kansas City.

Blei, D.M., A.Y. Ng and M.I. Jordan (2003) Latent Dirichlet Allocation, *Journal of Machine Learning Research* 3, 993-1022.

Edison, H.J. and J. Marquez (1998) US monetary policy and econometric modeling: tales from the FOMC transcripts 1984-1991, *Economic Modelling* 15, 411-428.

Griffiths, T.L. and M. Steyvers (2004) Finding scientific topics, *Proceedings of the National Academy of Sciences* 101, 5228-5235.

Hansen, S., M. McMahon and A. Prat (2018) Transparency and Deliberation within the FOMC: A Computational Linguistics Approach, *The Quarterly Journal of Economics* 133, 2, 801-870.

Hartmann, P. and F. Smets (2018) The First Twenty Years of the European Central Bank: Monetary Policy, ECB Working Paper No 2219.

Meade, E.E. and D. Thornton (2012) The Phillips curve and US monetary policy: what the FOMC transcripts tell us, *Oxford Economics Papers* 64, 197-216.

Schwarz, C. (2018) ldagibbs: A Command for Topic Modelling in Stata using Latent Dirichlet Allocation, *The Stata Journal* 18, 1, 101-117.

Srba, I., M. Grznar and M. Bielikova (2015) Utilizing Non-QA Data to Improve Questions Routing for Users with Low QA Activity in CQA, IEEE/ACM International Conference.

Figure 1: Average topic share of FOMC transcripts (2003-2012)

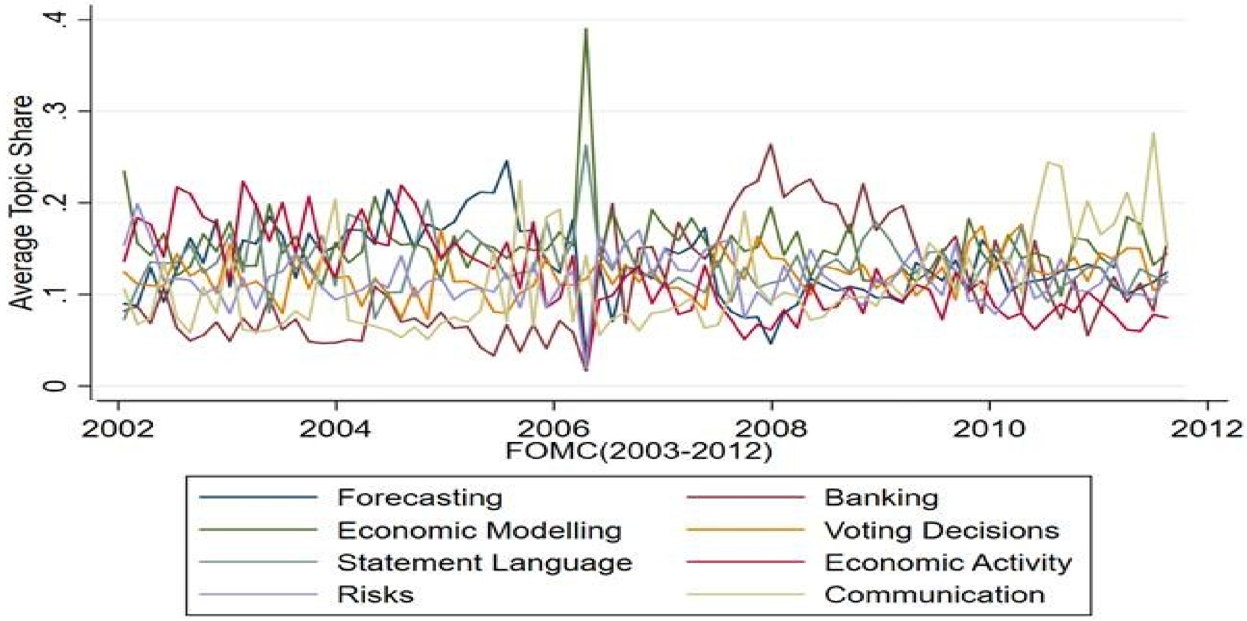


Figure 2: Number of data entries (sentences and paragraphs) in FOMC Transcripts (2003-2012)

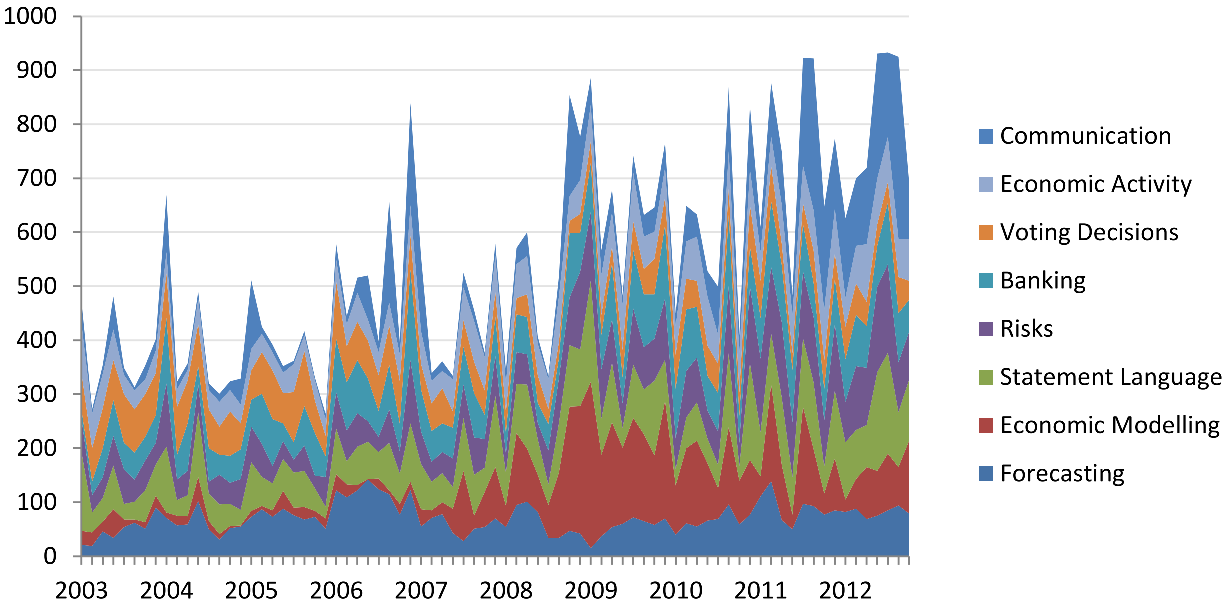
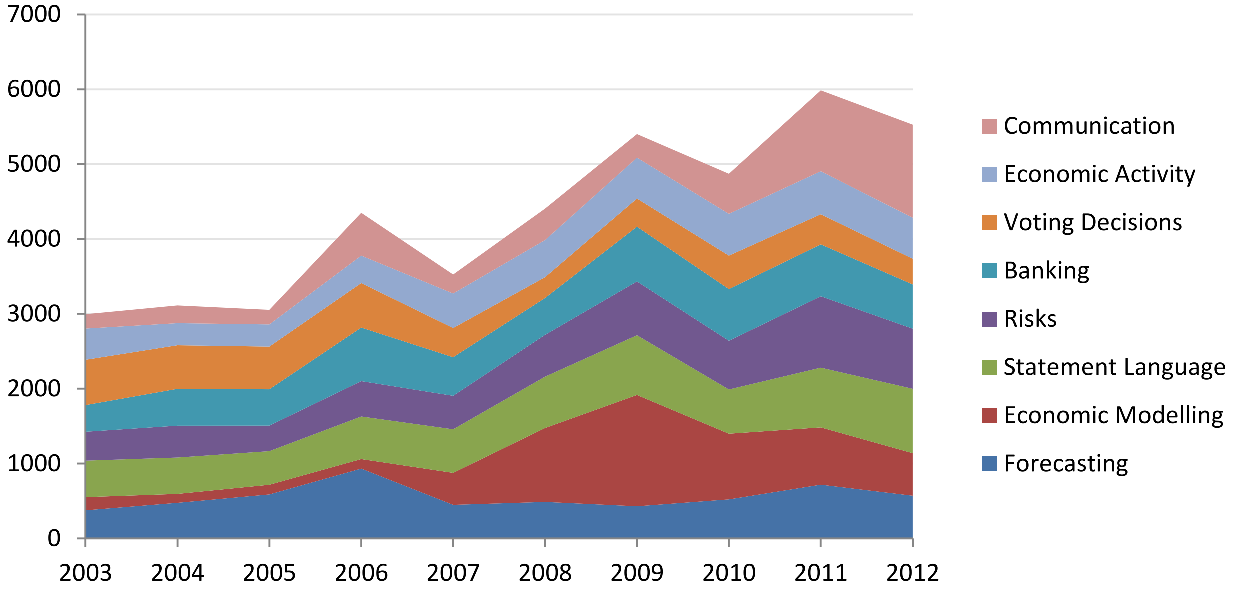
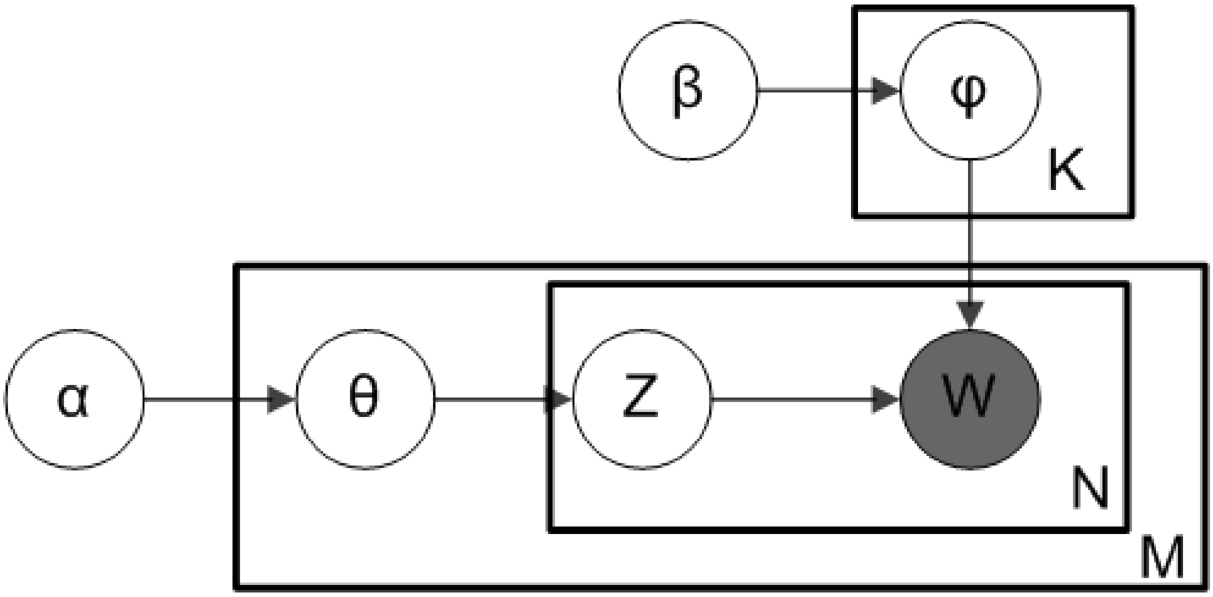


Figure 3: Annual number of data entries (sentences and paragraphs) in FOMC transcripts (2003-2012)



Annex 1: Plate notation for the Latent Dirichlecht allocation. Source: Srba et al. (2015)



Annex 2: Topic selection of text entries

TOPIC 1: Forecasting

Turning to inflation, I have nudged my forecast for both core and headline PCE inflation down a little since April …

When I compare the Board staff's forecast with ours, I find that the Greenbook projection, even the most updated one …

TOPIC 2: Banking System

Wells, Goldman, Bank of New York, Sun Trust, and BB&T, for example—opted out. Whenever a fee is assessed on assets or …

The first thing is that if we had a floor system, there would be more reserves in the banking system, and that might ac …

TOPIC 3: Economic Modelling

Your second question about standard errors is a really good one, and it is hard. There are lots of different models that …

If you ask whether a DSGE model would tell the story differently from, let's say, FRB/US, the answer is "maybe—it depends …

TOPIC 4: Voting Decisions

We had a second vote to approve the two amendments to domestic authorization. Brian gave an explanation. Are there any …

We had a second vote to approve the two amendments to domestic authorization. Brian gave an explanation. Are there any …

TOPIC 5: Statement Language

In terms of the wording of the statement itself, I like alternative B as it is currently worded. I agree with you that …

A small thing regarding wording in the statement. I supported the idea of moving from "will act as needed" to "will emp…

TOPIC 6: Economic Activity

The year-over-year production index edged up to 31 in March and jumped to 46 in April in our region. The year-over-year …

MR. BULLARD. Thank you, Mr. Chairman. The level of economic activity in the Eighth District is slowly improving, although …

TOPIC 7: Risks

I continue to see many downside risks to the outlook. Problems in Europe, fiscal austerity measures, tightening in …

Turning to the national outlook, the data have painted a mixed picture. The quarter opened with promise compared with …

TOPIC 8: Communication

In broad terms, I support both of the subcommittee’s proposals for improving the communications of the Committee. I believe …

Fourth, and finally, with respect to the FOMC responsibilities, is communication to the public. The public doesn’t make …