A qualitative text analysis of ECB communication strategy

Davide Viviano

1 Introduction

This article analyzes the evolution of communication of the European Central Bank from the beginning of the year 2000 until the end of 2016 and its impact on market expectation. The core motivation behind this study is to identify with quantitative and qualitative methods the change in communication of the ECB before and after the beginning of the European Debt Crisis, with particular attention to the role played by the current president Mario Draghi. For our analysis we use published speeches and interviews of the ECB Presidents from 1999 to 2016. We conduct an original analysis focusing our attention on the semantic structure of text data; in addition, we study its effect on the European stock market. Together with well-known techniques coming from the Machine Learning and Statistical literature, we introduce a novel statistical approach for interpreting text data based on Sparse Principal Component Analysis.

2 Literature Review

The role of central banks communication is becoming increasingly important over the years. As explained by J.L. Yellet, vice-chair of FED, in a famous speech given in 2012, "... the revolution in central bank communication is not driven by technological advances. Rather, it is the product of advances in our understanding of how to make monetary policy most effective". In fact, economic theories stress the high correlation of credibility and ex-post effects of monetary policies. In 2007, Bini Smaghi, ex-member of ECB executive board, underlined the importance of considering communication as a crucial instrument for the Central Bank in order to achieve both short and long term goals. Little efforts have been devoted to properly study the change in the attitude of central banks towards communication during this last two decades. Among others, [4] studied the long lasting effects of communication on stock market returns and volatility. They used a black-box software for scoring documents on an optimism scale and they used this score as the main variable of interest for their study. In a different paper[6] the author adopted a classical econometric approach to show that the language have become more "aggressive" during the period of the crisis. On the other hand, new methodologies for text analysis have not been fully exploited in this context. The article is organized as follow: in the first part we

describe the data; in the second part we provide the theoretical description of some of the methodology used; in the last part we provide information about the data analysis.

3 Data Description

The text data analyzed contains:

- Speeches of the European central bankers from 1999 to 2017;
- Press Conferences from 1999 to 2017;
- Interviews with the central bankers from 1999 to 2017;
- Press releases excluding the monetary policy decisions release from 1999 to 2017.

The total number of documents analyzed after the data cleaning process exceeds one thousand documents. We have excluded from our analysis all interviews and speeches that have not been given by central bankers. The reasons are two: as expressed by Sbighi, in order to avoid confusion the main source of information for the market always comes from the central banker; we were interested in studying the change of the communication policy of the ECB with particular attention to the role played by the central banker. Monetary policy decision statements have not been included because they were uninformative in terms of qualitative semantic features. To represent the data we have adopted a bag of words representation of the documents. Therefore, we have transformed the corpus into a document term matrix containing the counts per each word in each document. As a main drawback of this approach, we loose the context in which each single word is used, while we gain in terms of computations. Before transforming the text data into the document term matrix we do the following steps:

- All capital letters are transformed in lower case;
- Removal of stop words: we use an English dictionary for stop words together with a self-made dictionary. Self made dictionary contains common and uninformative terms in ECB speeches such as "mario", "draghi", "ecb", "european", "europe", etc;
- Stemming and Lemmatization: we transform words by removing plurals, third persons, etc. This process was very helpful for reducing the dimensionality of the data;
- Tf-Idf treshold: this process is generally used to remove very common and very rare words whose presence in the data might augment the amount of noise. We have considered this step necessary due to the high dimensionality of the data. With this technique we were able to drop many uninformative words that appear in almost all the documents and that we did not include in the stop-words, such as for example the term "financ". We set the cut-off to be 0.1. After this process the overall number of different words is reduced to around 1200.

- Different documents on the same day have been collapsed in a single documents by summing the words count.

As additional meta-data, we consider the day in which the document have been published. Finally, we have also used the daily change rate market stock returns from 2000 to 2016. As a proxy of the European market returns, we have collected data from the Spanish, Italian, French and German stock market index. As a measure of the whole European stock market returns we have computed the first principal component over this four index. The principal component as an approximation of market return is a common technique adopted in some of the Finance literature.

4 Main Contribution

This paper wants to analyze the change in the way European central bankers have communicated. Key questions that we want to tackle are:

- Which words have the highest effect on market return volatility?
- Have central bankers used different communication strategy during different periods?

Due to the difficulty of capturing the content of the dialogues, we focus our attention on qualitative semantic differences between documents, and in particular different words used in different time periods. First, we focus our attention on the communication strategies used by the ECB during different periods. We use Dynamic Topic Models and Sparse PCA to identify the key source of variation over different periods. More details are given in the next section. In addition, we study the reaction of central bankers during high and low volatility periods. High and low volatility period have been estimated using Hidden Markov Model.

5 Statistical Techniques

In this section we review all the methodologies that we have used to perform our analysis. For sake of brevity we have skipped technical details and we have focused our attention on key concepts useful for the understanding of the analysis that we performed.

Before entering into the discussion, let $X \in \mathcal{R}^{n \times p}$ be the document term matrix and y be any output variable of interest.

5.1 Dynamic Topic Model

In a first phase of the analysis we used Dynamic Topic Models(DTM)[2] in order to identify three different topics and to study the change of words used per each topic over three

different periods. In the same spirit of Latent Dirichlet Allocation, DTM is a generative model that assumes conditional independence of words across documents once you condition for the latent variables. Key differences with LDA, DTM impose a further layer in the hierarchical model by assuming that each per topic document distribution at time t depends on the previous document distribution at time $t-1$ and the same per each topic word distribution. A simple version of the generative model is the following.

- Draw word distribution of topic k at time t: $\beta_{k,t}|\beta_{k,t-1} \sim N(\beta_{k,t-1}, \sigma^2 I)\ \forall k$;
- Draw per document topic distribution: $\alpha_t|\alpha_{t-1} \sim N(\alpha_{t-1}, \delta^2 I)$
- Per each Document:
  - Draw $\eta_t \sim N(\alpha_t, a^2 I)$
  - Per each word:
    - Draw topic: $z_{t,d,n} \sim Mult(\pi(\eta_{t,d}))$
    - Draw word: $w_{t,d,n} \sim Mult(\pi(\beta_{t,z_{t,d,n}}))$

with $\pi(x_i) = \frac{\exp(x_i)}{\sum_i x_i}$

5.2 Sparse PCA

Principal Component Analysis is a well known technique used in many context, including dimensionality reduction. By performing the Spectral Decomposition of the input matrix, PCA identifies the components that explain most of the variance within the input space, corresponding to the eigenvectors with largest eigenvalues. One key property of the components is their orthogonality. The main drawback of PCA is the loss of interpretability of each component. In fact, components correspond to a weighted sum of all the input variables - named loadings - and their interpretation might be hard or even impossible. Sparse PCA finds those components that maximize the variance while forcing $p-s$ of the loadings in each component to be equal to zero, with $0 < s < p$ being a given constant. With this approach we gain in terms of interpretability, whereas components are not perfectly orthogonal anymore. From a mathematical perspective, Sparse PCA tackles the following problem:

$$\begin{aligned} &max_u u'\Sigma u \\ &\text{Subject to } ||u||_2 \leq 1 \\ &\text{Subject to } ||u||_0 \leq s \end{aligned}$$

Where $||u||_0$ is the number of non zero loadings for the component u. Importantly, the problem is NP-Hard and the solution can be computed by doing a semi-definite programming relaxation[7].

5.3 Hidden Markov Model with ECB Text Data

Let y_t be the market return at time t and let $z_t \in \{1, 2, .., K\}$ be a latent variable. We specify the following model:

$$y_t|z_t \sim N(\mu_{z_t}, \phi_{z_t})$$
$$z_t \sim Mult(\pi)$$

Posterior probabilities can be computed using EM algorithm. In our case, using AIC as model selection criterion, we identified only two states with equal a priori probability, with one state corresponding to high volatility periods and the other to low volatility periods, as we will show in the next section. Once we have computed the posterior distribution $z_t|y_t$ we treated the posterior states as an additional meta data of the document term matrix. In particular, we matched each document X_{t+1} to the day-before market state $z_t|y_t$ and we use this as additional input feature.

6 Data Analysis

6.1 Dynamic Topic Models

As a first preliminary analysis we identify 3 main topics and we use Dynamic Topic Models in order to study the change of words over time per each different topic. We identify 3 main periods, corresponding to the presence of a different Central Banker - Duisemberg, Trichet and Draghi. Results are shown in Figure 1.

Results show that very similar words tend to be constantly the most influential for topic identification. Interestingly, in the last column, we observe the word *crisi* and the word *challeng* to become relevant during the last period, corresponding to 2011-2016.

6.2 Interpreting Key Words Through Sparse PCA

We repeat a similar analysis using sparse PCA. As a preliminary analysis, we impose that only 5 loadings in each component are different from zero and we compute the first three principal components on the whole corpus. Results are shown in Table 1. As we might expect area, *economi*, *growth*, *inflat* and *polici* are key words. In addition, all words associated with monetary policy decisions explain most of the variability within the whole corpus.

We repeat the same exercise over time. We divide the observations over three time periods and we perform sparse PCA on each period. We compute the first principal component imposing that only 15 loadings are non-zero. Results are shown in Figure 2. The first table from left to right correspond to the period 1999-2004, the second 2004-2011, the third to 2011-2016. The method captures interesting key differences over the periods. The last column, corresponding to the period 2011-2016, is the only column having words such as

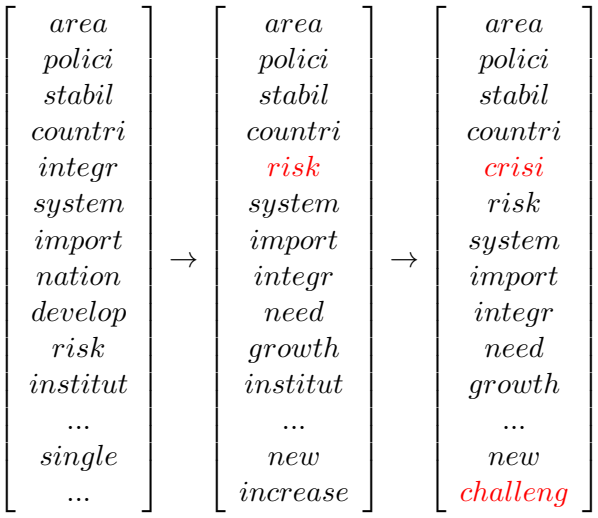

Figure 1: Different topics at time 2011-2016 using DTM.

	PC1	PC2	PC3
allot	0	0.206	0
area	-0.013	0	0
asset	0	0	-0.079
collater	0	0	-0.257
council	0	0.151	0
decid	0	0.268	0
economi	-0.042	0	0
elig	0	0	-0.934
growth	-0.527	0	0
inflat	-0.818	0	0
oper	0	0.917	0
polici	-0.225	0	0
programm	0	0	-0.233
purchas	0	0	-0.017
refinanc	0	0.146	0

Table 1: Key words as main source of variation on the whole corpus computed by using Sparse PCA and imposing only five active loadings per each component.

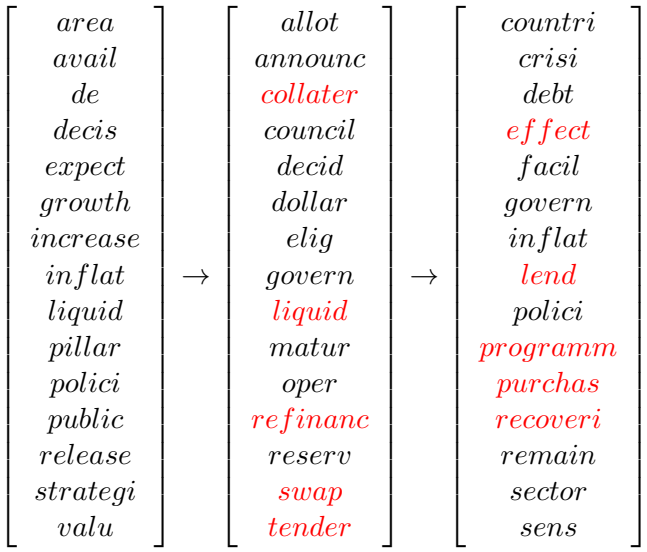

Figure 2: Evolution of most used words from 2000-2004, 2004-2011, 2011-2016. Each column represent the 15 active loadings of the first Sparse Principal Component for each period.

countri, *crisi* and *debt*, in line with what we would expect. In addition, whereas the second column seems to emphasize mostly technical concepts such as *swap*, *liquid*, *refinanc*, etc., the third column has less technical words. A possible and non-exhaustive interpretation is that the same concepts are communicated with a more direct language during the last period.

6.3 Considering additional metadata: EU stock Market Volatility

Market Parameters Estimation

We estimate the whole EU market stock by considering the stock index of Spain, Italy, France and Germany and computing the first Principal Component over these four index. The estimated market returns are shown in Figure 3. In order to estimate different states of the market (e.g. high and low volatility states) over time we specify the following model:

$$y_t|z_t \sim N(\mu_{z_t}, \sigma^2_{z_t})$$

with y_t being returns at time t and $z_t \in \{1, 2\}$ being a latent variable with binomial distribution which specifies the state of the economy. Table 2 show the posterior mean and standard deviation of the returns per each state. State 1 can be identified as the

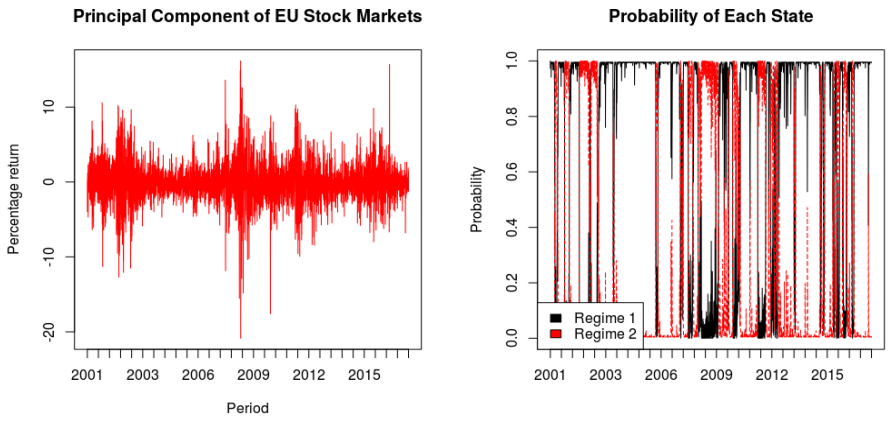

Figure 3: European stock Market returns and estimated probability of each of the two states of the economy (high and low variance state).

low volatility state and State 2 as the high volatility state. In Figure 3, we plot how the probability of being in one of the two states has changed over time. To interpret the figure, when at the top of the y axis there is a predominant color, this correspond to the most probable state of the economy during that period.

	μ_k(Intercept)	σ_k
State 1	-0.092	1.761
State 2	0.242	4.265

Table 2: Results after running Hidden Markov Model with two states on EU stock Market Returns

Effect of Market Volatility on Central Banker Speeches

As explained in the previous section, we used the estimated state of the economy at time t as an additional metadata for the document at time $t+1$ in order to study the effect of the day before market behaviour on central bankers communication. We do following steps:

- Divide the documents between documents during high and low volatility periods;
- Split each sub-group of documents between before and after 2011. The year was chosen because of the new presidency of Mario Draghi at ECB;
- Perform Sparse PCA on each sub-set of documents to study the words that explain most of the variance within each group of documents.

The main reason behind this approach is to check whether the attitude of the ECB has changed before and after 2011 after controlling for the state of the market.

Whereas we do not find interesting differences over time under the regime of low volatility, we do find some differences in the words used by central bankers when we compare the period before and after 2011 under the regime of high volatility. Results are shown in Table 3. Whereas the analysis is purely qualitative, during the 2011-2016 words such as *programm*, *measur*, *instrument*, *media* play a crucial role. They might partially explain a new interest of the central banker in capturing the attention of the whole public opinion through a more direct vocabulary. On the other hand, we recognize that this statement has not strong theoretical foundation and we leave to the reader to derive its own conclusions.

High-volatility State	2011-2016 PC1	2011-2016 PC2	High-volatility State	2000-2011 PC1	2000-2011 PC2
announc	0.288	0	billion	0	-0.035
bond	0.217	0	decis	0.113	0
claim	0	-0.424	dollar	0	-0.211
collater	0	-0.407	economi	0.494	0
debt	0	-0.042	expect	0.099	0
decid	0.893	0	fund	0	-0.048
document	0	-0.183	govern	0.048	-0.169
elig	0	-0.782	growth	0.458	0
govern	0.173	0	inflat	0.403	0
instrument	0	-0.064	oper	0	−0.862
measur	0.098	0	polici	0.561	0
media	0	−0.060	reserv	0	-0.187
programm	0.184	0	risk	0.212	0
swap	0.001	0	swap	0	-0.381

Table 3: Loadings on first two sparse PCA under Regime of high volatility for period 2011-2016 and 2000-2011.

7 Conclusion

Along this article we have shown how new methodologies coming from the machine learning literature might become helpful for analyzing ECB textual data. In particular, we have shown that sparse PCA together with other techniques offer a interesting qualitative representation of the data. We have observed that there is a change in the words used for communicating from the central banker before and after 2011. Most effective words on

stock volatility seem to belong to a non-technical vocabulary. A more direct vocabulary seems to be preferred after 2011, although no statistical test has been performed on this claim. Finally, whereas we have focused our attention on short term market effect of central banker communication, long term effect remains an open research question.

References

[1] Biau, Grard, and Erwan Scornet. "A random forest guided tour." Test 25.2 (2016): 197-227.

[2] Blei, David M., and John D. Lafferty. "Dynamic topic models." Proceedings of the 23rd international conference on Machine learning. ACM, 2006.

[3] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning research 3.Jan (2003): 993-1022.

[4] Born, Benjamin, Michael Ehrmann, and Marcel Fratzscher. "Central bank communication on financial stability." The Economic Journal 124.577 (2014): 701-734.

[5] Hamilton, James D. "Regime switching models." Macroeconometrics and Time Series Analysis. Palgrave Macmillan UK, 2010. 202-209.

[6] Siklos, Pierre L. "The global financial crisis and the language of central banking: Central bank guidance in good times and in bad." (2013).

[7] Zou, Hui, Trevor Hastie, and Robert Tibshirani. "Sparse principal component analysis." Journal of computational and graphical statistics 15.2 (2006): 265-286.