NLP Final Project Proposal

Eduardo Espino Robles
Meha Sadasivam

What is the project about?

The objective of this project is to create a non-parametric regime switching model that for each period identifies the most similar previous periods based on 3 state variables. Typically, regime switching models use a predefined set of regimes and then map the current period into one of those, however, in this case, the process will be reversed. For every current period the model will output the most probabilistically similar periods in the past and then use that information to forecast US equity and fixed income assets returns. For our state variables, we are going to come up with a summary statistic that describes 3 macroeconomic variables, inflation, growth (GDP) and volatility, based on text analysis.

Initially we were planning to analyze how topics in central bankers speeches changed over time as a way to detect changes in regimes. Since the mandate of the Federal Reserve is both full employment, that directly links to economic growth and keeping inflation on target, going through the speeches of governors of the Federal Reserve System and its board should give us an indicative of what the central bank is thinking with respect to those topics. However, we decided it was better to use the Thomson Reuters data set because it provides more information that encompasses both central bankers and companies information.

The motivation behind this approach is that news articles offer more data to make a more robust analysis based on the fact that news outlets are typically the first place where topics show up. Usually, even before an economic indicator has been published, news articles have already been written about such indicators, which encourage us to use them as a barometer of importance of such topics and the market sentiment about them. We are confident that by analysing how news on the 3 state variables mentioned above changed over time we will be able to characterize its importance and the state of such variables with a number that will be later used in doing the regime switching analysis.

What NLP approach/es will you be utilizing?

1. **Word Similarity**: Cosine/ euclidean distance measures to come up with the most similar words for each of our main themes: ‘Inflation’, ‘Growth’, ‘Volatility’
    a. To do this, we will need to model all the articles in the same space (tf-idf), and then pick out the top 10 words that are most similar to the theme words
    b. For each theme, we will use these (synonymous) words to filter out articles and create a themed subsample of article - only articles that contain at least one of the words will be present in the subsample
2. **Topic Modeling**: For each themed subsample (which consists of articles across the entire time period) we intend to come up with a coherent list of topics that can help us distinguish between talk about high or low environment for our themes (macro variables) inflation/growth/volatility:
    a. To do this, we plan to explore multiple approaches
        i. The basic topic modelling approach, using an existing pre-trained LDA model
        ii. Train a new LDA model based on our sample, where we manually indicate whether a topic relates to high/low environment in the training sample (we will need to further explore how this can be done)
    b. We would also need to explore and come up with the right stop words for the sample, as well as the effective vocabulary (will need to find the optimal word frequency threshold)
    c. Further, we will need to calibrate the hyper parameters so that the topics are most coherent
3. **Sentiment analysis**: One approach to help us identify the market attitude on a given topic that talk about inflation/growth/volatility would be to run sentiment analysis on the top ‘n’ words of a topic, and try and determine what could be the correct “positive”/ “negative” words to use in each subsample to classify them accurately. (Would also explore creating a custom sentiment dictionary)

With the topics generated for each of the themed subsamples, we hope to identify specific topics for each theme, and we will compute the average beta (aggregated monthly) for each of these topics over time. This will be our final measure, and after we obtain the betas for each of the three themes, we plug them into the regime switching model.

What dataset/s will your project involve?

We plan to use the Thomson Reuters News Articles dataset that was made available by Professor Harry Mamaysky on the Research Grid. This dataset runs from 1996-01 to 2020-09, and contains around 300,000 articles each month. Also, for the equity returns we are going to use the S&P 500 index and for the fixed income assets we are using the Barclays Treasury Index, both of them will be obtained from Bloomberg.

What is the MVP (Minimum Viable Product) version of your project?

We think that a potential problem for our project is that we don't get coherent topics that would allow us to easily use the probability of a given topic as our state variable to describe our 3 macroeconomic variables. In that case, using the beta as an input into the regime switching model might not be a good idea, because we won't be able to interpret if the beta is actually capturing what we want. If we get to that situation, we have two alternative approaches, we can do sentiment analysis on the top n-words in the topics we think more directly represent what we want and use either a composite score of a positive/negative/neutral classification to be used as the state variables.

The other approach is to use the top n-words from the most popular topic in each period as a state vector in each period, instead of a state variable. That way, in assigning probabilities to previous periods, we will essentially calculate the distance between the words in the most popular topic in each period against the current period. If the distance is small it means that most likely both periods belong to the same regime, while if the distance is high it can be a sign of a switch to another one. The MVP version of our project is a less refined version of our idea, but one that will ultimately allow us to find the most probabilistically similar periods from a given reference/current month.