ISyE 6416 – Computational Statistics – Spring 2016 Bonus Project: "Big" Data Analytics Proposal

Team Member Names: Ting Sun, Shiyu Zheng, Zhe Liu

Project Title: HMM Model in Economic Recession Identification

Problem Statement

Financial markets provide daily analysis and interpretation on a large set of new micro and macroeconomic information leading them to update their expectations on the future growth path. Theoretically, these alterations must be reflected by the dynamics of asset prices or monetary aggregates. We want to infer economic movements of the United States from those data, especially, the turning point of the economy detected by our analysis. Three "macroeconomic" time series have been selected in this case: the industrial production index, the unemployment rate, and GDP growth rate.

In this project, we develop an extension of the Hidden Markov Model (HMM) to find the hidden economic states of the US behind the data. Specifically, the HMM has a Gaussian mixture at each state as the forecast generator. Gaussian mixtures have desirable features in that they can model series that does not fall into the single Gaussian model. Then the parameters of the HMM are updated in each iteration with an add-drop Expectation-Maximization (EM) algorithm. At each timing point, the parameters of the Gaussian mixtures, the weight of each Gaussian component in the output, and the transformation matrix are all updated dynamically to cater to the non-stationary financial time series data.

Using the techniques from the EM Theorem, we can prove that the convergence of the proposed algorithm is guaranteed. Therefore we are guaranteed to find the hidden economic states, recession and boom, from the analysis of "big" data regarding macroeconomic time series.

Data Source

Most economists use the NBER (National Bureau of Economic Research) recession data as a reference. We also compare our outcome with NBER recession data in this project. The three "macroeconomic" time series: the industrial production index, the unemployment rate, and GDP growth rate are downloaded from FRED economic data.

Methodology

The general HMM approach framework is an unsupervised learning technique, which allows new patterns to be found. It can handle variable lengths of input sequences, without having to impose a "template" to learn.

First of all, why do we want to use HMM in this case? Three indexes are chosen as representative of the economic states. Those data can be thought of being generated by underlying recession or boom state and observation generation probability density function, which we denote as Gaussian Mixture in this case. From many other examples, we can see that there is a good match of sequential data analysis and the HMM. Especially in the case of financial time series analysis, where the data can be thought of being influenced by underlying circumstances that is not known to the public. The transition probabilities as well as the observation generation probability density function are both adjustable, which gives us more power to fit the model. Given plenty of data that are generated by some hidden power, we can create an HMM architecture and the EM algorithm allows us to find out the best model parameters that account for the observed data.

We use the Viterbi algorithm to find the most likely hidden state sequence, $S' = (s_1, s_2 \dots s_T)$, given three observed data sets (here we take one as example) $O = (o_1, o_2 \dots o_T)$, from $\text{argmax}_{S'} P(S', O|\mu)$. Denote he following variable $\delta_j(x) = \max_{s_1, s_2 \dots s_{t-1}} P(s_1, s_2 \dots s_{t-1}, o_1, o_2 \dots o_{t-1}, s_t = j|\mu)$. This variable stores the probability of observing $o_1, o_2 \dots o_t$ using the most likely path that ends in state i at time t given the model μ. The corresponding variable $\psi_j(t)$ stores the node of the incoming arc that leads to this most probable path.

The calculation is done by induction, similar to the forward-backward algorithm, except that the forward-backward algorithm uses summing over previous states while the Viterbi algorithm uses maximization. The complete Viterbi algorithm is as follows ($b_i(\mathrm{o}), a_{ij}$ *as in class*):

1. Initialization: $\delta_j(1) = \pi_i b_i(o_1), \psi_j(1) = 0, j = 1 \dots N$.

2. Induction: $\delta_j(t) = b_{ijo_t} \max_{1 \le i \le N} \delta_i(t-1)\, a_{ij}, \psi_j(t) = arg \max_{1 \le i \le N} \delta_i(t-1)\, a_{ij}$.

3. Update time: $t = t + 1$ while $t \le T$ or turn to step 4.

4. Termination: find optimal path $s_T^* = arg \max_{1 \le i \le N} \delta_i(T)$. Read out the path.

When there are multiple paths generated by the Viterbi algorithm, we can pick up one randomly or choose the best n paths.

For three observed data sets, we generate optimal paths and estimate model parameters separately. Now we have three sets of the hidden state of recession or boom, we can compare them with the NBER data and to each other. To make our outcome more precise, analysis can be done here to compare and adjust our conclusions.

Expected Results

As we have applied a multivariate qualitative hidden Markov model to three data sets, we expect that it provides a reliable framework to track the growth cycle and that a financial related model would be helpful to predict economic downturns. For instance, the latest significant growth slowdown could have been detected in early 2007, in the start of the recession. We also expect three macroeconomic variables that we pick generates the same results as identifying recession times, hopefully, similar to NBER recession index (1 as recession and 0 as boom).

However, we can still get unsatisfying results due to these reasons: poorly chosen Gaussian Mixture model as observation generator, or other hidden forces behind the data that disrupt our judgment. Therefore, this work could be extended in two directions. One the one hand, because of some limitations to maximum likelihood procedures, a Bayesian approach could prove promising. On the other hand, this model is a first step towards building a qualitative nonlinear factor analysis framework. Adapting a hierarchical Markovian classification model could then be a future way of research.

Team member Responsibility

Ting will be responsible for finding or collecting data. Zhe will do the coding part, and Shiyu is responsible for problem formulation. The three of us will discuss and analyze the results. Finally, for the final report, everyone will be responsible for his own part.