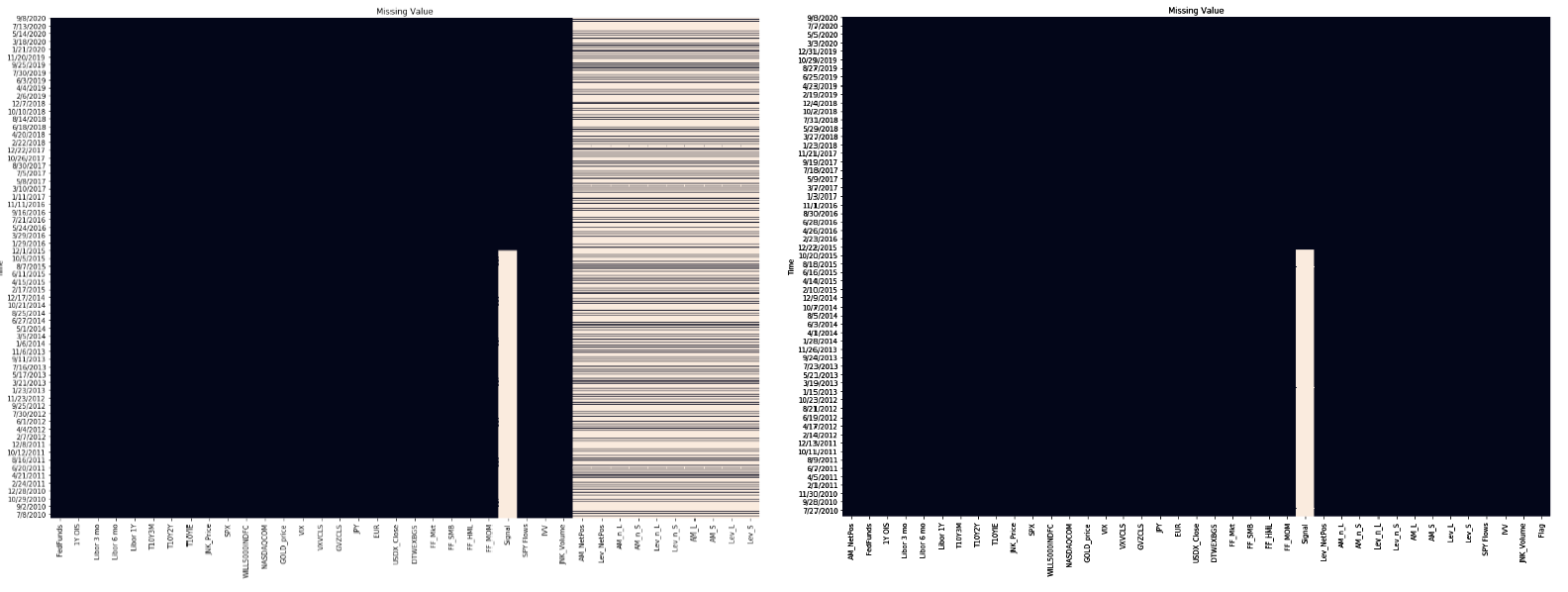
Part I: Merging Data

We extended the date range of the provided data set from January 2012 to June 2010, which is the first date we were able to locate reliable CoT data. We also included additional predictors for interest rates, foreign exchange, credit, equity, and commodity markets.[[1]](#footnote-1) We merged the dataset which aligns the different date in different datasets. The initial dataset and the extended dataset are visualized as followed.

Chart

Description automatically generated

After merging the dataset we cleaned the dataset by filling the missing values by filling with the average of the most recent 5 data points. To this end, we have a complete dataset except for the ‘signal’ column which doesn’t have data tracking back prior to last time stamp of the original provided dataset.



Part II: HMM

The motivation behind using hidden markov model (HMM) is due to the serial correlation between positioning, which we are trying to predict, and other variables. Take the two-month rolling correlation with eight major variables for example, we can see the clear pattern that the correlation bounces between 1 and -1 and same correlation signs with different variables cluster around the same time.

Graphical user interface, application

Description automatically generated

We used a simpler model to see the potential of HMM. The HMM we employed use the positioning data to identify two states of positioning: high and low. We used the R package dthmm to label the states first and the result is shown below.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Given the labelled positioning states, we used a logistic regression with all variables (lagged variables for those contemporary with positioning) to predict the states in train and test set. And we used two different regression for two states respectively. The prediction versus actual in test set is shown below.

Chart, line chart, histogram

Description automatically generated

Model performs well in the 'high' state with a 5.3% test AAPE as labelled in shaded green area but the performance drops significantly in the 'low' state with 38% test AAPE. Overall the model produces a test 32% AAPE. This indicated that identifying the states which the positioning is in can be potentially helpful but the states needs more specific recognition. The residue plots are listed as followed.

Chart

Description automatically generated

The prediction error has serial correlation up to lag 3. Residues are mostly positive indicating the model tends to over-estimate.

Part II: Conclusion

The positioning data obviously has state pattern. Identifying the states which the positioning is in can be potentially helpful but the states needs more specific recognition. Due to the time constraint and technical difficulty, better HMM model with better state identification especially regarding the low state as described in the previous model is yet to be achieved.

1. Midterm report [↑](#footnote-ref-1)