Predicting S&P 500 Sector Returns with New York Times Articles

Rujun Han, Yunlin Zhang

DS-1005 Inference and Representation

Abstract and Data

News can provide insight into the market and potentially help to project future movements. We tried to model the news articles from New York Times and combine with classical time series models to see whether it helps in improving prediction accuracy.

Data:

- New York Times news corpus from 1988 to 2008
- Daily returns for each of the S&P500 sectors starting in October, 2001
- Harvard Pyschology Dictionary (HPD)

Data Preprocessing

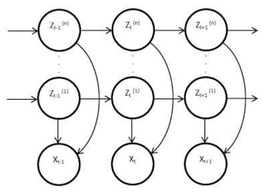
- Transformed raw S&P 500 indexes to log weekly returns to make them stationary.
- Emotion Index
  - Used HPD to build two sets of negative and positive words
  - For every NYT article whose headline has company names in the S&P500 sector list, counted the occurrences of emotion words.
- LDA Topics
  - Articles were tokenized to form bag-of-ngrams
  - Trained LDA model on the entire NYT corpus
  - Aggregated weekly LDA topics $F_{p,t}$ based on following rule:

$$F_{p,t} = \sum_i T_{i,p} * E_{i,t}$$

  - $T_{i,p}$ topic p and $E_{i,t}$ emotion index of article i;

Model I – HMM and FHMM

Factorial Hidden Markov Model proposed by Ghahramani and Jordan (1997). We used the version with Gaussian emissions.



$$P(X_{1:T}, Z_{1:T}) = P(X_1)\prod_{i=1}^{N} P(S_{i,1})\prod_{t=2}^{T}\prod_{i=1}^{N} P(Z_{i,t}|Z_{i,t-1})P(X_t|Z_t)$$

Model II – State Space Model

- The HMM model can be casted into a state space model representation.

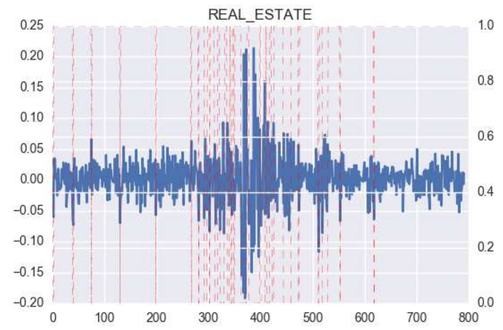
$$X_t = A * S_t + B * F_t + \epsilon_t$$

$$S_t = C * S_{t-1} + \nu_t$$

  where X_t is the observed series, S_t is the hidden states and F_t is the additional predictors (or exogenous variables). ϵ_t and ν_t are assumed to follow MVN(0, Σ) and Σ is a diagonal matrix.
- Python state-space library allows us to incorporate exogenous variables into the standard state space model and uses Kalman filter and smoother to compute MLE estimators

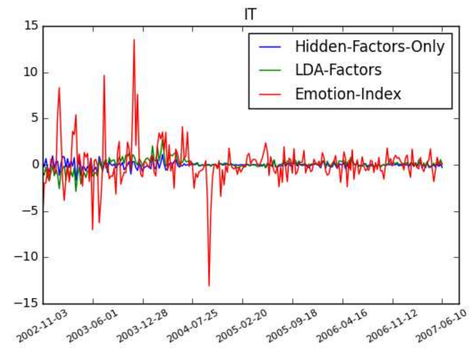
HMM/FHMM Results

HMMs are not as good SSM in predicting the returns. However, HMMs can be used to identify volatility regimes.



State Space Model Results

In general, the simple SSM model is really hard to beat. Both Emotion index and LDA topics specifications are more volatile than SSM.



Model Evaluation

- Used a 54-week rolling window to train the model and forecast one-week-head log returns
- Evaluated against the actual log returns based on two measures
- **Measure I** Mean Square Errors
- **Measure II** Prediction Accuracy: defined as the average 0-1 accuracy based on whether models predicted the same direction as the actual returns

Results

Table (1) Mean Square errors

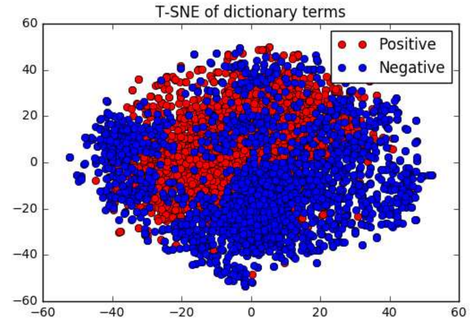
Sectors	State Space	LDA Factors	Emotion Index	HMM	FHMM
Information Technology	7.62	8.07	13.46	5.28	3.80
Industrial	3.71	3.92	5.71	5.74	3.85
Consumer Discretionary	4.23	4.49	7.32	6.85	4.91
Health Care	2.81	2.99	4.63	4.54	3.71
Financial	3.72	3.90	6.19	8.73	6.84
Energy	-	-	-	7.14	5.43
Materials	-	-	-	6.80	4.37
Consumer Staple	-	-	-	3.83	3.06
Telecommunication services	-	-	-	5.49	4.52
Utilities	-	-	-	4.68	3.98
Real Estate	-	-	-	8.05	4.95

Table (2) Prediction Accuracy

Sectors	State Space	LDA Factors	Emotion Index	HMM	FHMM
Information Technology	0.53	0.54	0.52	0.53	0.54
Industrial	0.57	0.56	0.52	0.55	0.54
Consumer Discretionary	0.57	0.57	0.51	0.54	0.53
Health Care	0.56	0.52	0.56	0.53	0.52
Financial	0.56	0.54	0.57	0.54	0.54
Energy	-	-	-	0.55	0.56
Materials	-	-	-	0.53	0.53
Consumer Staple	-	-	-	0.53	0.53
Telecommunication services	-	-	-	0.51	0.51
Utilities	-	-	-	0.54	0.53
Real Estate	-	-	-	0.54	0.52

Other Observations

Created Word2Vec models using the first 13 years of NYT corpus and evaluated how well a sentiment analysis model based on it might compared to keyword lookup



Distributionally dissimilar but not separable

Conclusion and Next Steps

Our method of incorporating the news did not yield improvements to the predictive power of the models.

- Financial time series is inherently difficult to forecast
- New York Times articles are not primarily focused on the market
- Can try to create more sophisticated language models for drawing insights