STAT 771 - Project Final Draft

Steven Moen

Sunday, November 29th, 2020

Background and Motivation

My M.S. thesis at the University of Chicago developed a novel method to understand and measure value-at-risk, a commonly accepted method to measure downside risk. The metric is understood as follows: a one-day 1% VaR of -10 million dollars for a portfolio means that the portfolio will lose at least 10 million dollars of its value on the 1% worst trading days. A major advantage of VaR is that it distills a distribution of returns into one number. As such, VaR is often used in stress testing by regulatory agencies (Holton 2014).

There have been many popular approaches in the literature such as modeling the total distribution of returns (Longerstaey and Spencer 1996), and approaches using a semiparametric or a nonparametric historical simulation (Richardson, Boudoukh, and Whitelaw 2005). While modeling the entire distribution is likely too simplistic, Engle and Manganelli in a 2004 paper (Engle and Manganelli 2004) argue that nonparametric methods in the other camp are usually chosen for "empirical justifications rather than on sound statistical theory". To balance these approaches, they propose a framework called CAViaR that directly forecasts the VaR quantile using a conditional autoregressive quantile specification. This approach builds upon the statistical literature that extends linear quantile models to settings amenable to financial modeling, such as with heteroskedastic and nonstationary error distributions (Portnoy 1991).

My thesis extends the model beyond a univariate setting into a multivariate setting using the diffusion index model, originally developed by Stock and Watson for predicting conditional means (Stock and Watson 2002b, 2002a). My model uses exchange-traded funds (ETFs) as explanatory variables that are combined into principal component vectors at the forecast origin. Combining these principal component vectors with transformations of lagged autoregressive response variables produces similar predictive accuracy during periods of relatively low volatility (when compared to the CAViaR model) along with more insight into the drivers of the changes in the response variable.

Intended Work

As encouraging as the results were from my thesis, two important questions remained unanswered. The first is whether some sort of mixture model would be appropriate, that is, aiming to use the basket of ETFs during good times, and use the CAViaR ARMA specification during bad times. The approach of using ETFs allows a prediction based on forward-looking expectations of fundamental factors. Indeed, ETFs are just baskets of individual stocks or bonds, and those securities are (in theory) based on rational expectations about future resources, market conditions, etc - the microfoundations of what drives our economy. The ARMA specification, while practically and statistically sound, is contradicted by economic theory and practice - the weak form of the efficient market hypothesis states that it is impossible to forecast future values of asset prices using past values. But perhaps this view is incomplete.

To combine these ideas, I would fit a Hidden Markov Model to infer the state of the world - the "rational" one, or the "irrational" one. Given the highly non-normal nature of financial data, I suspect there would be many interesting statistical and computational challenges that would arise with this approach. In addition, it is likely worth exploring alternative ensemble methods to further probe into the seemingly enigmatic nature that pervades financial time series.

When I implement this, the “rational” state of the world will refer to the predictions of one of the multivariate CAViaR models where as the “irrational” state of the world will refer to one of the predictions from the univariate CAViaR models. One way to possibly implement this is to say that if the losses from the multivariate models are lower, then the HMM will lean towards the rational state of the world, otherwise, it will lean towards the other state.

Hidden Markov Model Work

Below is arguably the most consequential plot from my M.S. thesis. The reasons for this are because it deals with an important VaR Level - 1%, which in the context of trading days means about the worst day out of 100. The biggest takeaway might be the fact that the four lines dashed lines (corresponding to the multivariate CAViaR model) do not perform as well as the last four lines, which refer to the univariate model. This can also be seen in the table of losses printed below.

Notation

Below is the notation used later in this paper. Items 2 - 5 listed below are new multivariate CAViaR models developed in this thesis; models 6 - 9 are from the established univariate CAViaR model developed by Engle and Manganelli.

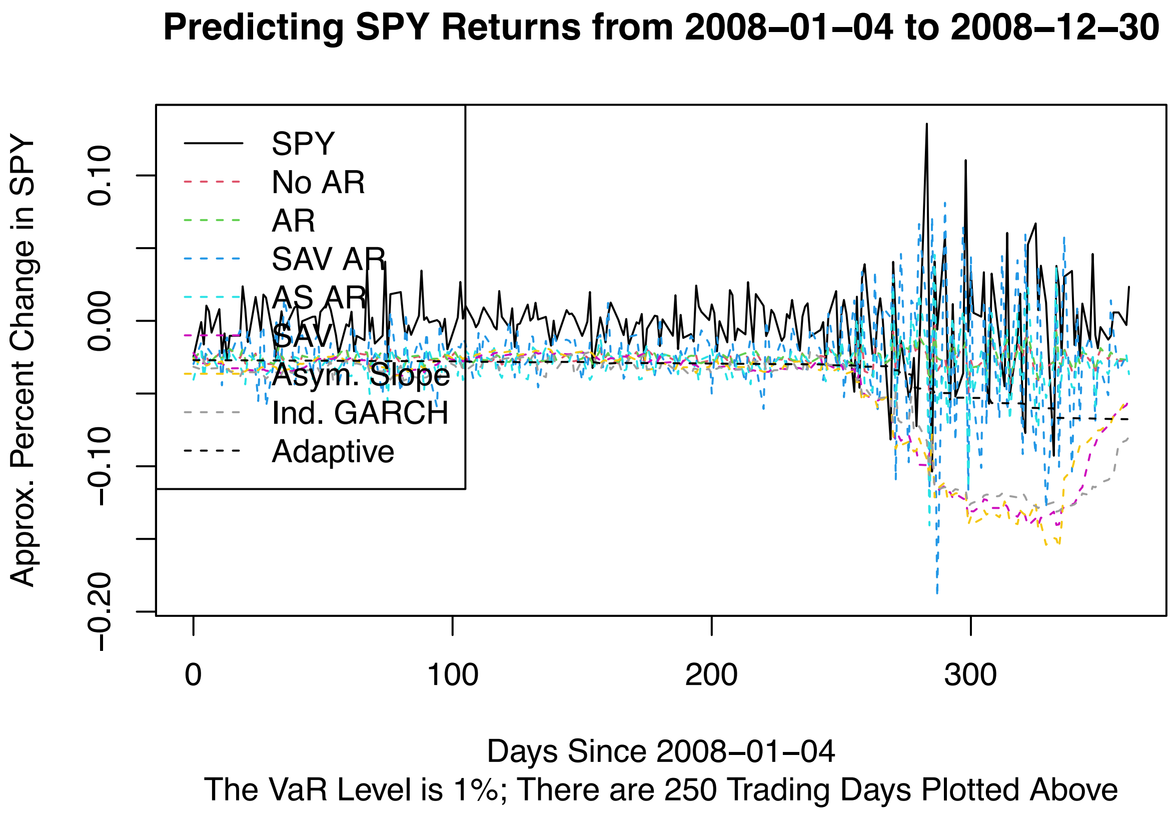
1. SPY: SPY ETF
2. No AR: Multivariate CAViaR Model with no lags
3. AR: Multivariate CAViaR Model with p lags
4. SAV AR: Multivariate CAViaR Model with p absolute value lags
5. AS AR: Multivariate CAViaR Model with $2p$ lags with asymmetric slopes
6. SAV: Univariate CAViaR Model with symmetric absolute framework
7. Asym. Slope: Univariate CAViaR Model with asymmetric slope framework
8. Ind. GARCH: Univariate CAViaR Model with indirect GARCH framework
9. Adaptive: Univariate CAViaR Model with adaptive slope framework

Table 1: Losses Over the Last 250 Trading Days in 2008 for the Multivariate CAViaR Models

	No AR	AR	SAV AR	AS AR
Losses by Model	0.736	0.737	1.733	0.863

Table 2: Losses Over the Last 250 Trading Days in 2008 for the Univariate CAViaR Models

	SAV	Asym. Slope	Ind. GARCH	Adaptive
Losses by Model	0.208	0.213	0.219	0.355



Based on the losses for each model during the last 250 trading days in 2008, it looks like the best options are the multivariate CAViaR model without AR terms for the multivariate model class and the symmetric absolute value model for the univariate model class. The full model specifications can be found in the appendix.

Coincidentally, these are among the simplest models available among the models plotted above. A natural criticism of this approach is that the losses are lower for the CAViaR specifications without lagged predictors. This is a fair point, however, the period of 2008 is a period of extreme crisis, and a simpler, ARMA-style model might seem to work better other things equal. Future work can extend this work to other periods. Since these are the best two options during this period of interest, the next step is to find reasonable parametric distributions to model their values.

Finding Distributions of the Forecasts

To fit the Hidden Markov Model, it is necessary to find distributions that appropriately fit the two models above. While the distribution of predictions from the multivariate model is fairly well-approximated by a

Table 3: Optimal Parameters for the Normal Distribution

	Mean	Standard Deviation
Optimal Parameters	-0.03	0.01

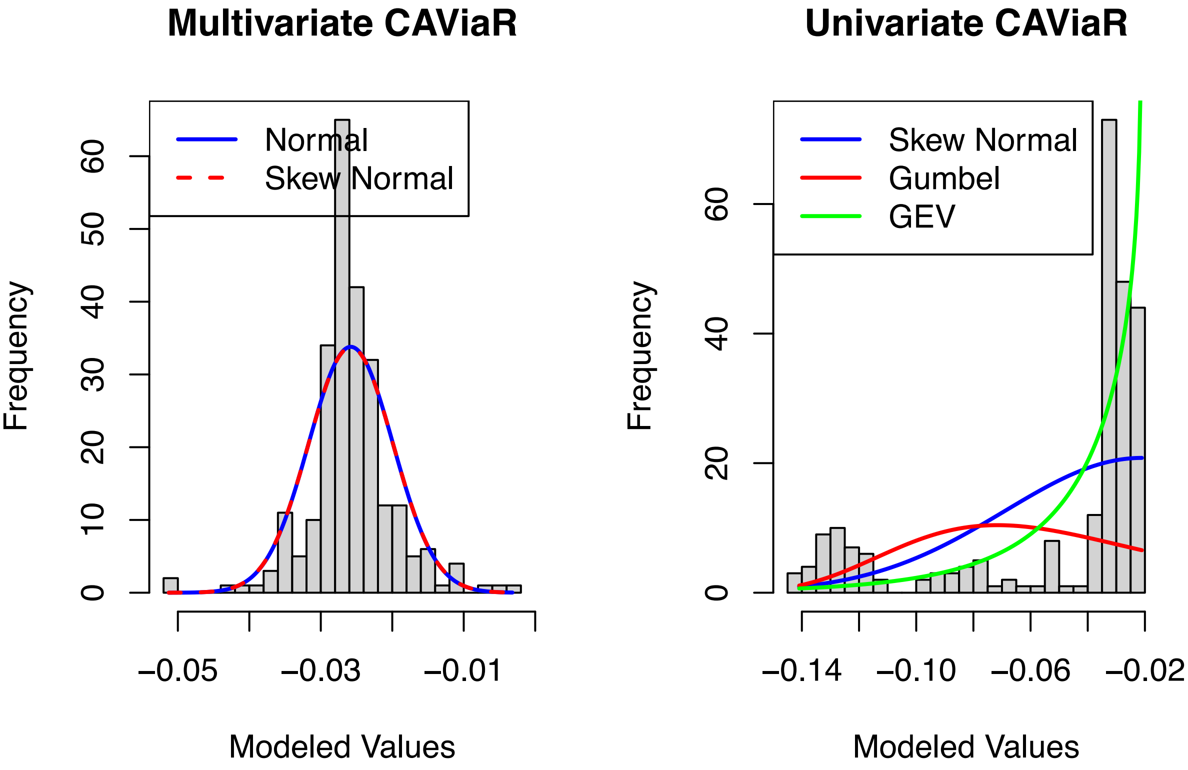
Note:
The Mean Was Estimated Using the Sample Mean, SD was Estimated Using Sample Std. Dev.

normal distribution, the distribution of predictions from the univariate model was highly left-skewed, which makes fitting a distribution difficult.

The first step was to try transformations of the aforementioned predictions from the univariate model, namely $\sqrt{\max(x+1)-x}$, $\log_{10}(\max(x+1)-x)$, and $1/(\max(x+1)-x)$ following the suggestions found here (Kassambara 2018). While these did (in some cases) eliminate the left-skew, it often created a more pronounced right-skew! Thus, I attempted to model the data without transformation.

There are many candidate distributions that could be used to model the empirical distribution of values from the univariate CAViaR model, but the three that stood out are:

1. The Skew Normal Distribution (parameterized by location, scale, and shape parameters)
2. The Gumbel Distribution (parameterized by location and scale parameters)
3. The Generalized Extreme Value Distribution (parameterized by location, scale, and shape parameters)



To fit the above histograms,the optimal parameters were fit using maximum likelihood. Note that while there were convergence issues used in fitting the skew-normal distribution, there were not issues with fitting the Gumbel or the GEV distributions.

To evaluate the model fit more rigorously, I compared the Kullback-Leibler divergence for each theoretical distribution.

Based on the K-L Divergence, it would seem to make sense to use the generalized extreme value distribution, however, the problem with doing this is the fact that this distribution doesn't have support over the entire

Table 4: Optimal Parameters for the Candidate Distributions

	Location	Scale	Shape
Skew Normal	-0.059	0.029	0.002
Gumbel	-0.072	0.044	NA
GEV	-0.045	0.028	-1.174

Note:

Estimated Using Maximum Likelihood

Table 5: Comparing K-L Divergences By Model Fits

	Skew-Normal	Gumbel	GEV
Mean Sum K-L Divergence	22.52	25.4	17.71

real line. Moreover, the Skew-Normal distribution did not work well when fitting the Hidden Markov Model. Therefore, the Gumbel is used in fitting the HMM below.

HMM Background and Results

The motivating idea behind a Hidden Markov Model is that there are 2 unknown latent states k that generate the data that is seen. (The reference for this information is given here (Stephens 2018)). The algorithm implemented here computes forwards probabilities, $\alpha_{tk} = \mathbb{P}(X_1, ..., X_t; Z_t = k)$. To start, one simply multiplies an equally-weighted prior $\pi_k = 0.5$ by the likelihood of the data given each state, given by $\mathbb{P}(X_1|Z_1 = k)$.

The likelihood function for the "rational" state (represented by the multivariate CAViaR model) is represented by a normal distribution whereas the likelihood function for the "irrational" state (represented by the univariate CAViaR model) is represented by the Gumbel distribution. Both use the parameters estimated above.

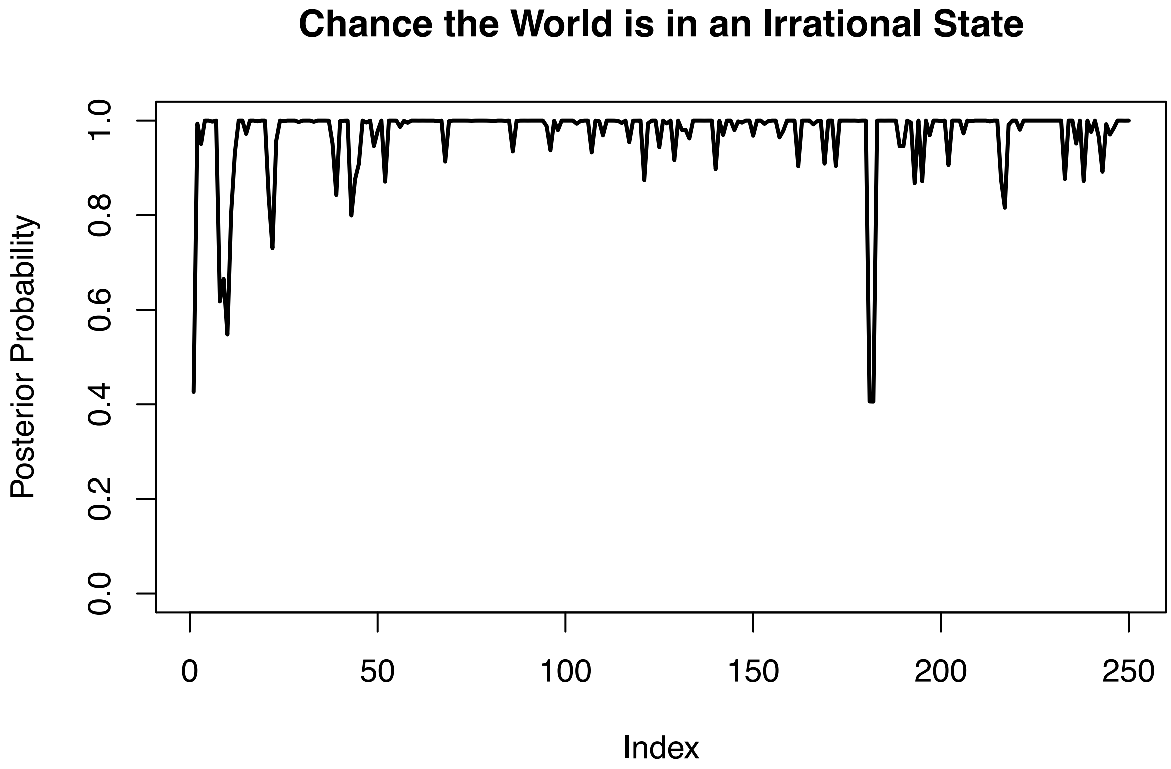
Now, once the α_1 value is calculate, α_2 is calculated as follows, with a similar process for an arbitrary value α_t.

$$\alpha_2 = (\alpha_1 P)_k \mathbb{P}(X_2|Z_2 = k)$$

The P symbol corresponds to a symmetric 2x2 transition matrix where the first row is $(0.9, 0.1)$ and the second row is $(0.1, 0.9)$.

To compute the backwards probabilities, we compute the following $\beta_{tk} = \mathbb{P}(X_{t+1}, ..., X_T; Z_t = k)$, and then the posterior distribution for each state Z_t is given by the following:

$$\mathbb{P}(Z_t = k|X_1, ..., X_T) = \alpha_{tk}\beta_{tk} / \sum_k \alpha_{tk}\beta_{tk}$$



The interpretation of this above graph seems quite clear at first blush - the "hidden" state of the world throughout 2008 is indeed the "irrational" one, marked by the relative success of the univariate CAViaR model. As perhaps with all research, this work generates many more questions than it answers. There are several next steps that are worth exploring.

- Adding more lagged terms into the HMM, or at least tinkering with the transition probability to understand the sensitivity of these parameters to the outcome
- Exploring what happens if the HMM were fit to all 8 candidate models, or using a Neural Network or Random Forest to find the hidden state
- Considering the implicit assumption that these forecasts are based on an asymmetric loss function (see the Appendix), and finding a way to weigh the consideration that an overprediction is a fairly clear indication that the world is not in that state

Changepoint Detection

The second question is to understand shifts in the economy using a changepoint detection algorithm:

1. Using a set of ETFs, perform Principal Component Analysis at T many points for M many factors - $f_{m,t}$
2. At each time point, add the vectors together to get a resultant: $\sum_{m=1}^{M} f_{m,t} = r_t$, giving $r_1, r_2, ..., r_T$.
3. Starting with an arbitrary reference point t_0 with associated r_0 resultant, measure the angle between resultants calculated at different time steps r_t

$$\theta_t = \arccos\left(\frac{r_0 \cdot r_t}{||r_0||||r_t||}\right)$$

The angle θ could be plotted over time, and changepoints could be detected using Monte Carlo simulation, because PCA transformations are non-linear, so calculating an analytical density from the transformed data is intractable. Moreover, the data fed into the PCA transformation is non-normal, which further supports the notion of using Monte Carlo simulation to establish reasonable estimates of uncertainty for detected changepoints. As with the first line of reasoning, there would certainly be interesting challenges, particularly in creating crisp null and alternative hypotheses.

Data Used

The response variable used in this analysis is SPY, which is an exchange-traded fund that aims to track the performance of the S&P 500, which is discussed above. It is broadly used as a bellwether of the U.S. economy, and has the advantage of avoiding survivorship bias - while an individual stock might go bankrupt or merge with another, it is reasonable to assume that these issues do not apply with an ETF.

Following this logic, there are several classes of response variables used in this analysis. The first group is a set of U.S. sector ETFs obtained from Seeking Alpha (NA 2020). As with the response variable, these ETFs were publicly traded throughout the Great Recession of 2008.

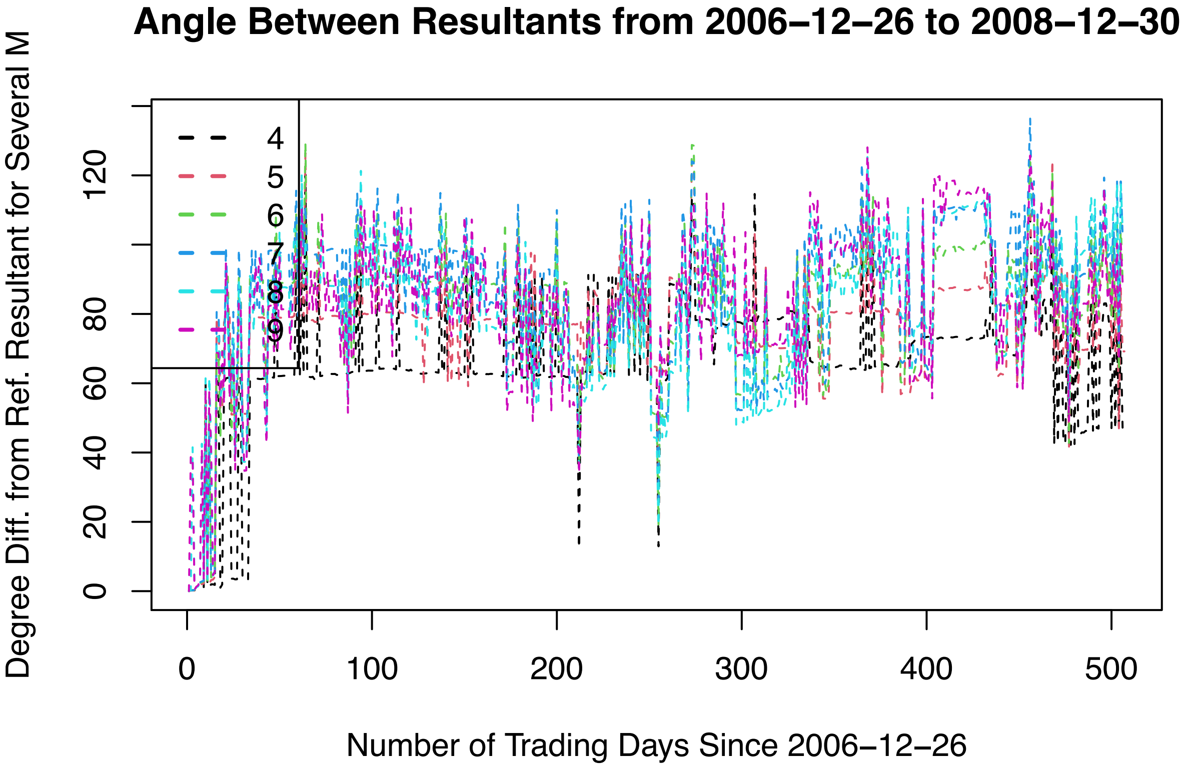
a. Utilities (XLU)
b. Consumer Staples (XLP)
c. Healthcare (XLV)
d. Technology (XLK)
e. Consumer Discretionary (XLY)
f. Industrial (XLI)
g. Financial Services (XLF)
h. Basic Materials (XLB)
i. Energy (XLE)

The second group for this analysis is bond ETFs. Like the previous two groups, these ETFs potentially contain forward-looking information about the stock market. These ETFs were chosen because they were the first fixed-income ETFs available in the United States, and had enough history for this paper (NA 2017).

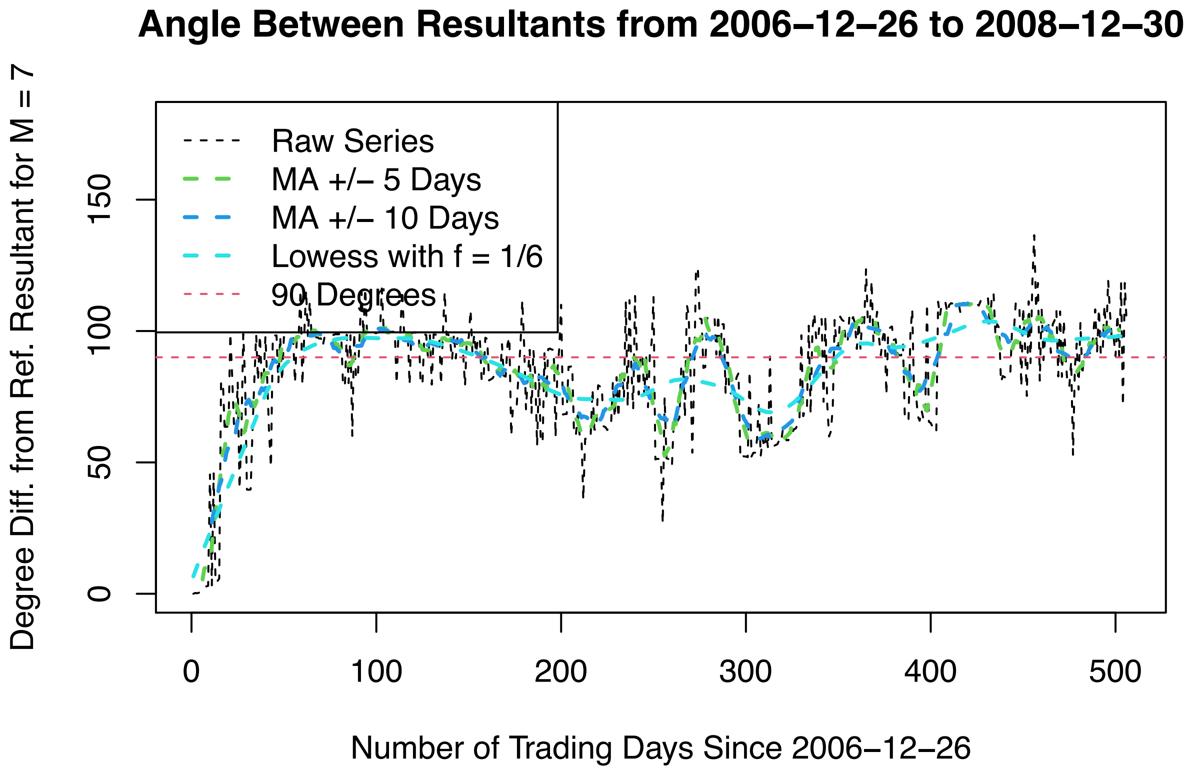
a. iShares 1-3 Year Treasury Bond Fund (SHY)
b. iShares 7-10 Year Treasury Bond Fund (IEF)
c. iShares 20+ Year Treasury Bond Fund (TLT)
d. iShares iBoxx $ Investment Grade Corporate Bond ETF (LQD)

Results

U.S. ETFs Results for 2007 - 2008



A natural criticism of the above fits are that the data is noisy. A complication of this analysis is the fact that there isn't necessarily anything to "pin" the data to, because the problem is unsupervised. As such, I think a decent way to picking the wheat from the chaff (or the signal from the noise) is to apply some smoothing filters to the above data. Two options to do so are the moving average smoother and the Lowess smoothers.



In the above plot, I applied a few filters, namely the moving average filter with equal weights for plus/minus 5 days and 10 days as well as the Lowess with a weight of $f = 1/6$ for higher levels of precision. Below are the specifications for moving average with 5 and 10 days, where x_t is the angle between the resultants.

$$m_{t,5} = \sum_{j=-5}^{5} \frac{1}{11} x_{t-j}, m_{t,10} = \sum_{j=-10}^{10} \frac{1}{21} x_{t-j}$$

The Lowess smoother is a smoother that per Shumway and Stoffer (Shumway and Stoffer 2016), is a technique "based on k-nearest neighbors regression, wherein one uses only the data $[x_{t-k/2}, ..., x_t, ..., x_{t+k/2}]$ to predict the true value of x_t. Based on a visual inspection of the data, it appears that a more precise estimator was in order, so the Lowess function only uses 1/6th of the data.

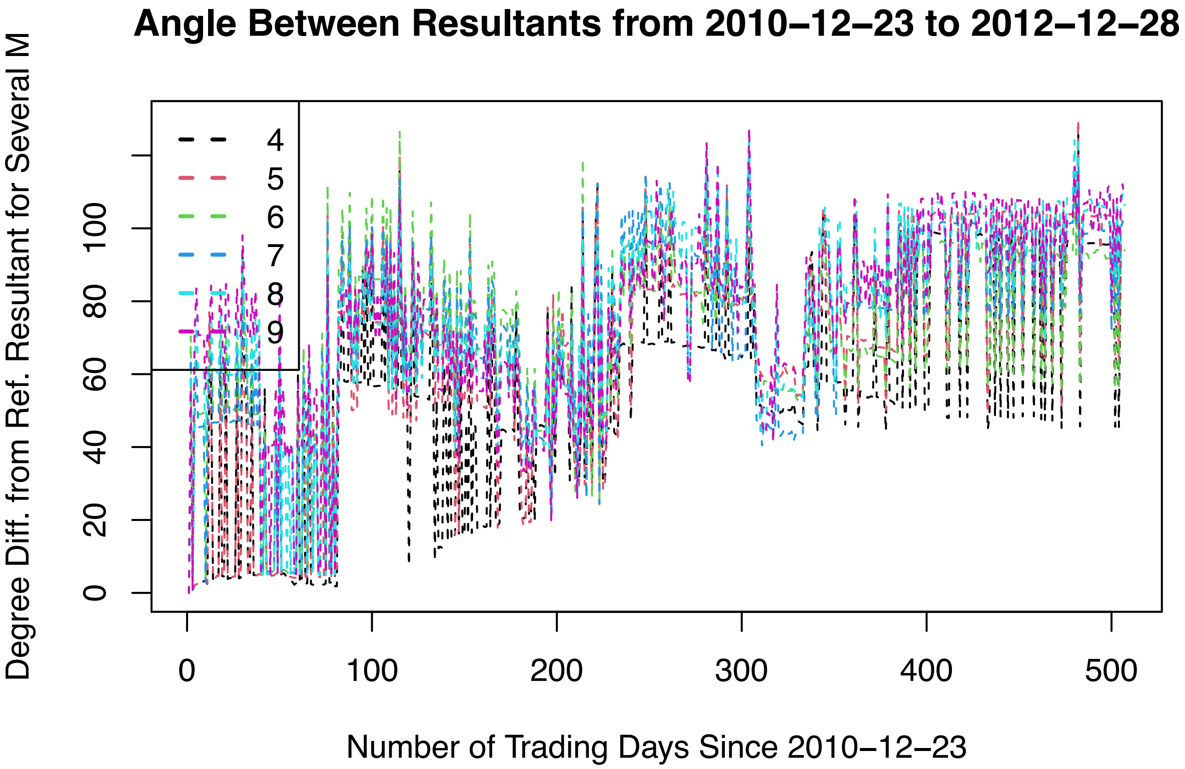
There are some interesting trends that bear discussion. While some of macroeconomics may appear to be little more than a crystal ball, or a science based upon strong priors, I don't believe that to be the case here. In the above graph, the angle between resultant vectors calculated from baskets of equity exchange-traded funds (ETFs) for different sectors of the U.S. economy point, almost as a leading indicator, towards the once-in-a-lifetime tumult that was about to grip the U.S. and global financial markets. Indeed, in the graph above, the angle of the resultant seems to stabilize for the month of August 2008, which corresponds roughly to the 400 – 425 trading days since the reference point at the end of 2006. This is before Lehman Brothers declared bankruptcy, before AIG was bailed out, and before Congress passed TARP. It remains to be seen whether this is coincidental or whether there really is a leading indicator here, but I believe the question is worth asking and worth exploring further.

From a statistical perspective, the most important fact about these trends is that the algorithm that produces them, namely PCA, is *an unsupervised learning algorithm.* At no point in this work is there a model of any kind trying to predict gyrations in the S&P500 or the Dow Jones or the real economy. Indeed, remember that these principal components are based upon sector ETFs, things like Energy, Consumer Staples, and Utilities, which in turn are based upon individual stocks - companies like Exxon, Capital One, General Electric, and Amazon. These stock prices may seemingly be enigmatic and noisy, but ironically they look this way because according to economic theory, they are likely to contain all the relevant information about their company as determined by market participants (Malkiel 2003). What if, in all of their foresight, they saw the most cataclysmic economic event of our time before it happened, without even aiming to do so?

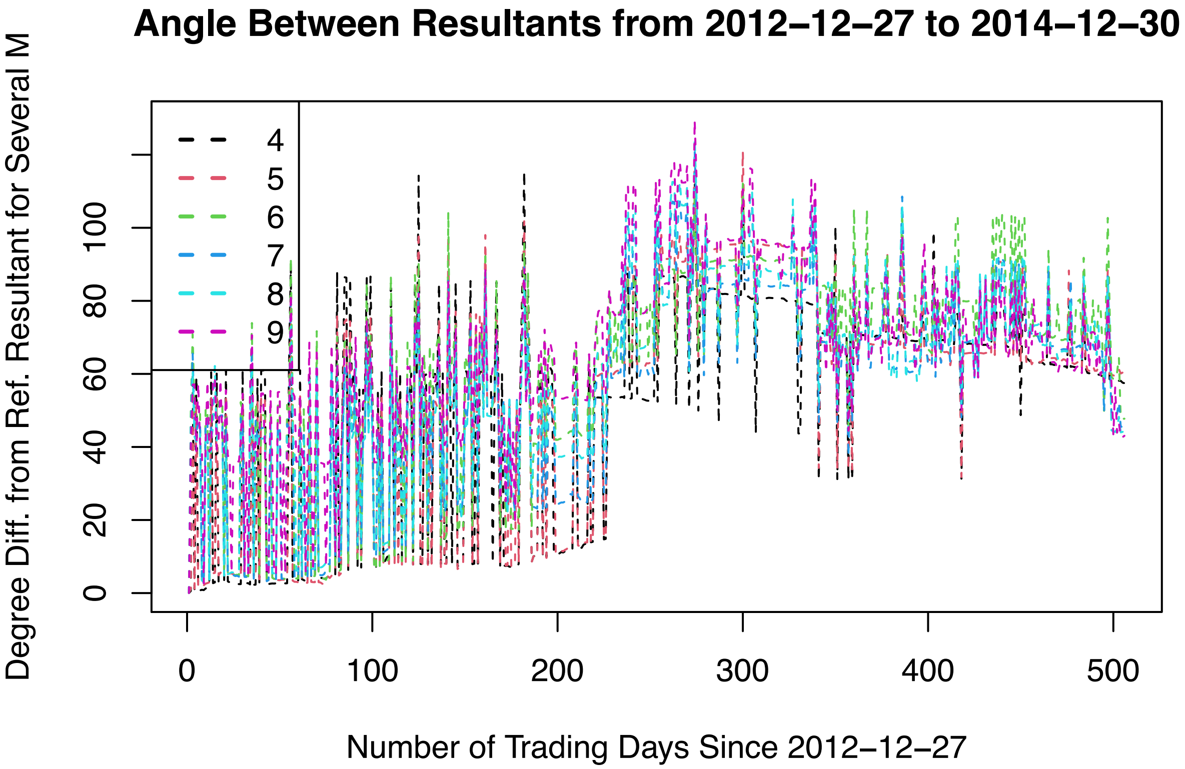
While future work in this area would benefit greatly from a variational autoencoder to be able to figure out what is exactly a meaningful change, it may also benefit from having additional data fed into it.

Appendix

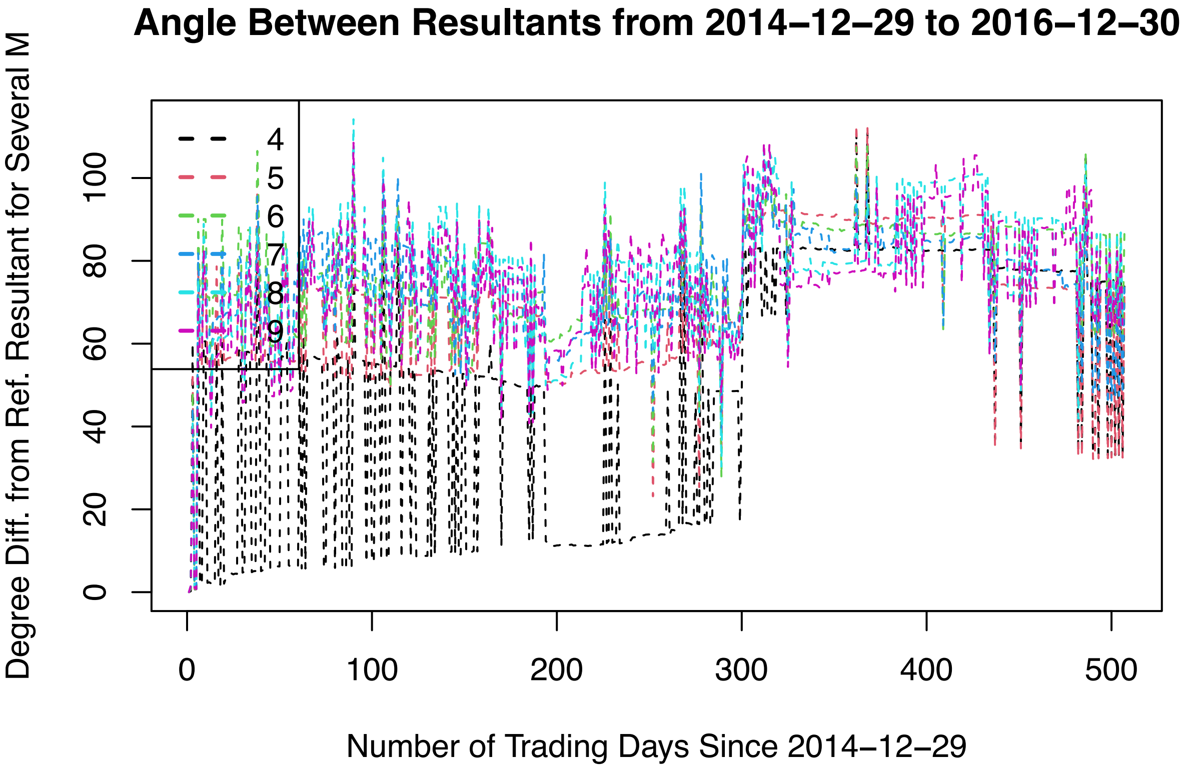
U.S. ETFs Results for the 2011 - 2012



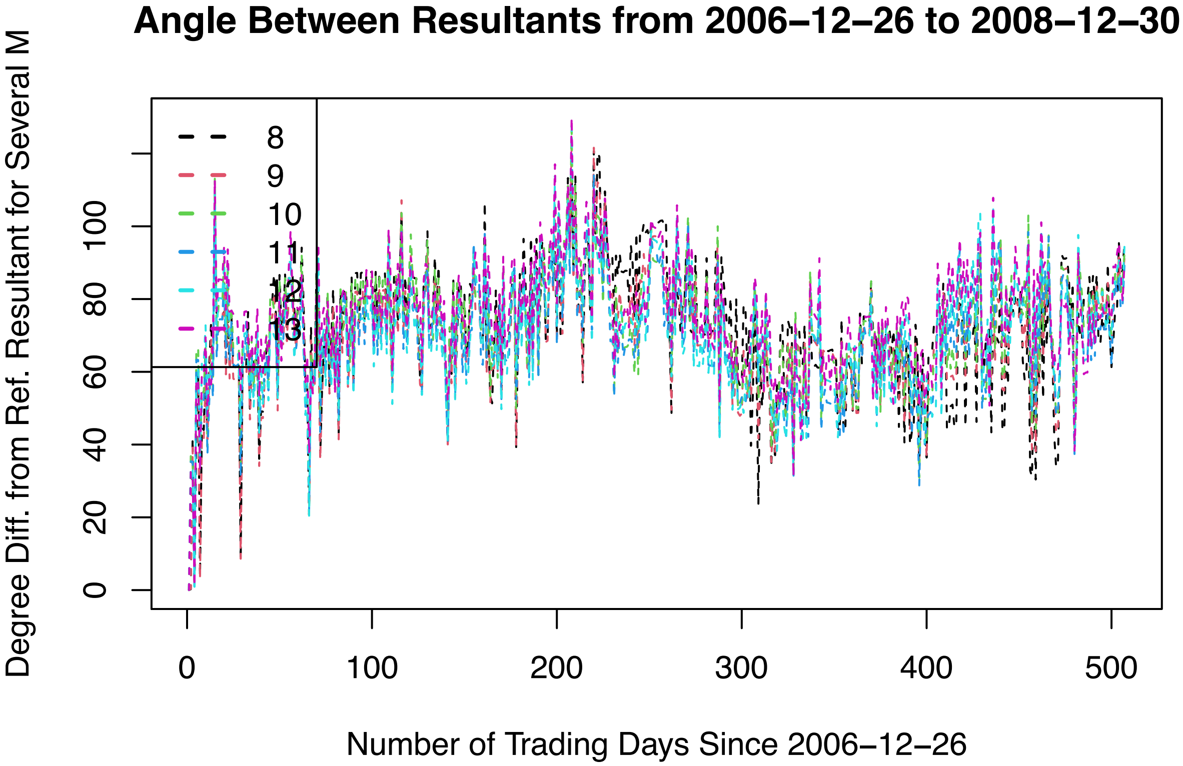
U.S. ETFs Results for the 2013 - 2014



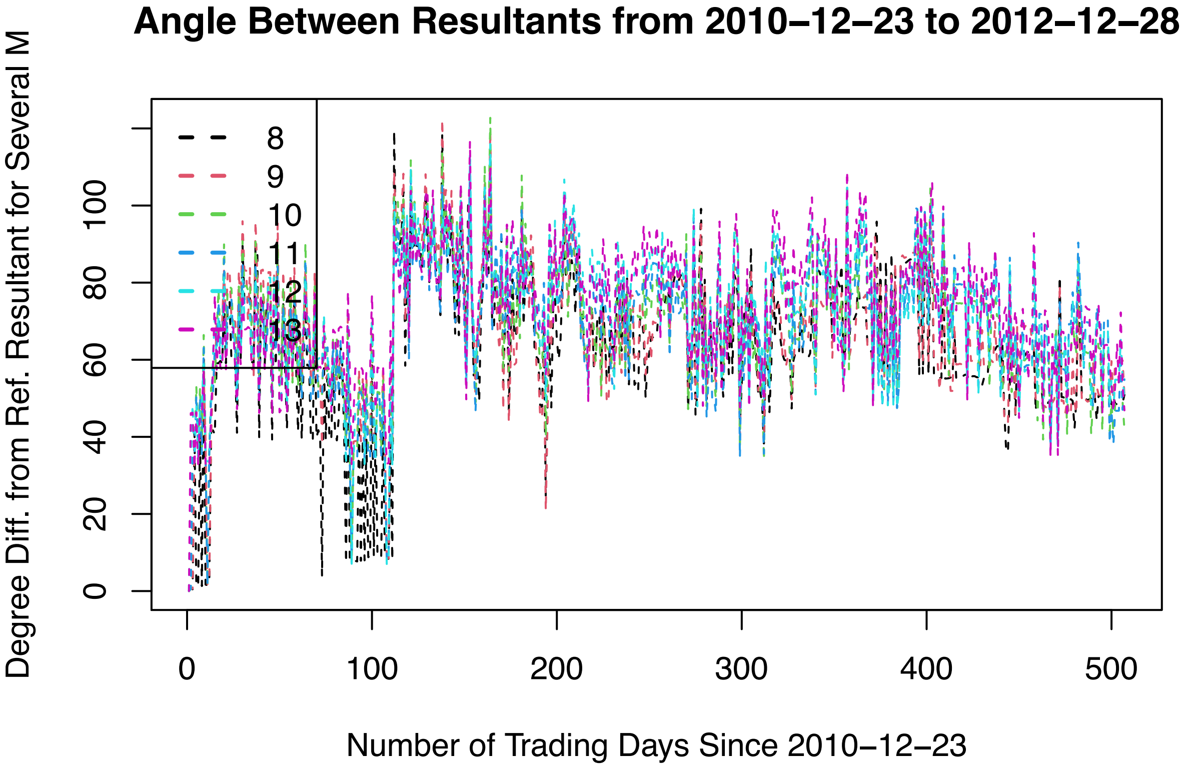
U.S. ETFs Results for the 2015 - 2016



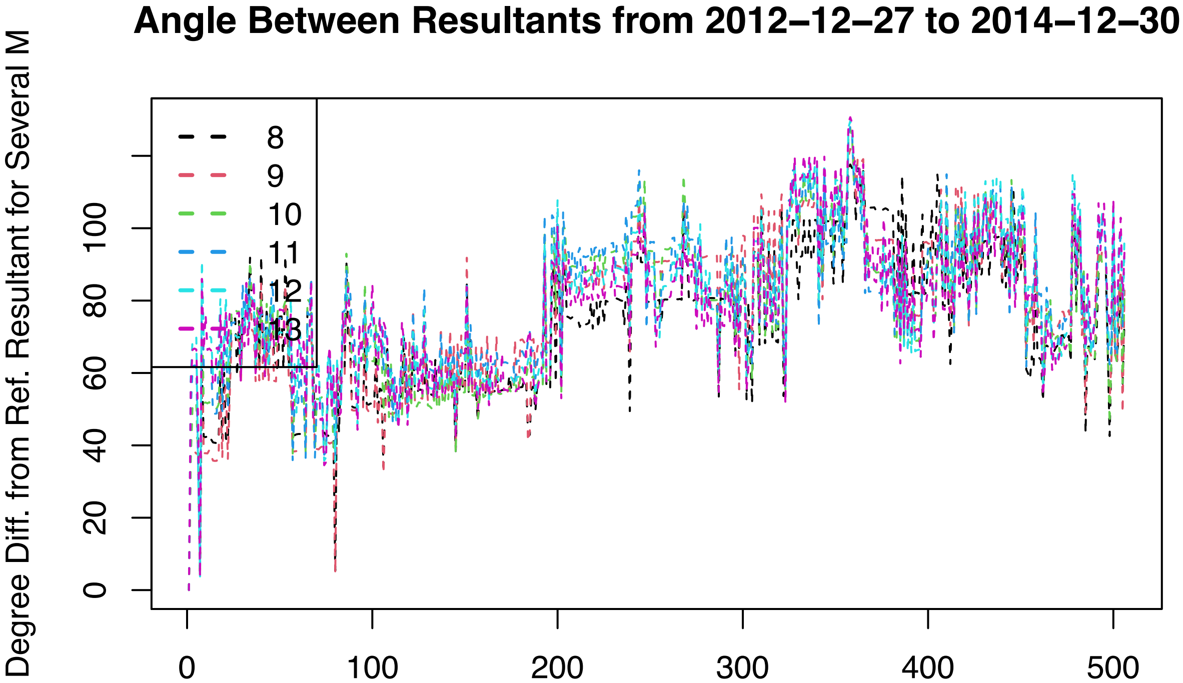
U.S. + Bond ETF Results for 2007 - 2008



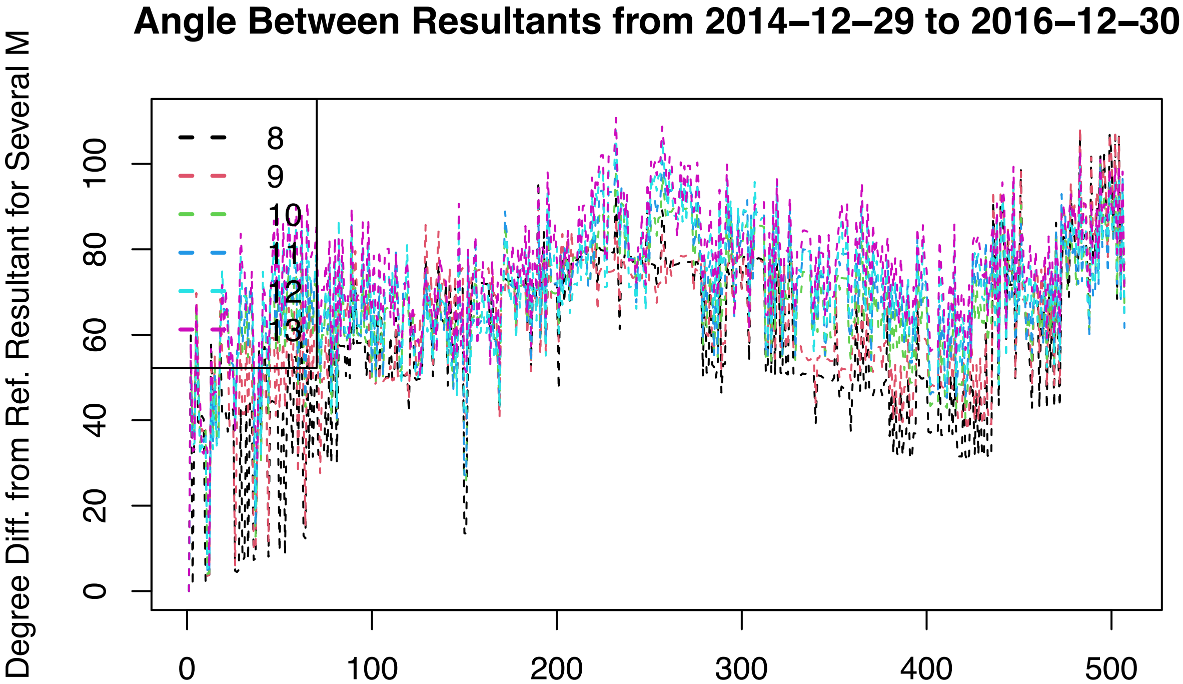
U.S. + Bond ETF Results for the 2011 - 2012



U.S. + Bond ETF Results for the 2013 - 2014



U.S. + Bond ETF Results for the 2015 - 2016



Univariate CAViaR Model Specifications

However, work needed to be done to align the diffusion index model with the CAViaR model, which is defined below. The following variables are required for use in the CAViaR model. For ease of notation, these are sourced directly from the Engle and Manganelli 2004 CAViaR paper (Engle and Manganelli 2004), with some added description:

- $(y_t)_{t=1}^{T}$ is a "vector of portfolio returns"
- θ is the "probability associated with VaR" (a 5% VaR would mean $\theta = 0.05$)
- $\boldsymbol{x_t}$ is a "vector of time t observable variables"
- $f_t(\boldsymbol{\beta}) \equiv f_t(\boldsymbol{x_{t-1}, \beta_\theta})$ is the "time t", "θ quantile of the distribution of portfolio returns formed at time $t-1$"

The authors then describe a "generic CAViaR specification" as follows:

$$f_t(\boldsymbol{\beta}) = \beta_0 + \sum_{i=1}^{q} \beta_i f_{t-1}(\boldsymbol{\beta}) + \sum_{j=1}^{r} \beta_{q+j} l(\boldsymbol{x_{t-j}})$$

What is interesting about the general setup is that there are two main components to the model - lagged observed variables (represented by l) and lagged values of unknown parameters, which in the specification below is used as moving average terms. As such, it is reasonable to generalize the specifications below as nonlinear ARMA models where y_{t-1} terms refer to previous returns, whereas $f_{t-1}(\beta_1)$ terms refer to previous predictions.

Adaptive CAViaR Model

Consider the following model:

$$f_t(\beta_1) = f_{t-1}(\beta_1) + \beta_1 \left[(1 + \exp(G[y_{t-1} - f_{t-1}(\beta_1)]))^{-1} - \theta \right]$$

Following Engle and Manganelli's 2004 paper, we choose $G = 10$, so that is what is used in the results section of this paper. The authors state the reason for the seemingly arbitrary choice is that while "the parameter G itself could be estimated; however, this would go against the spirit of this model, which is simplicity". Previous sensitivity analysis showed that running the adaptive model with $G = 5$ did not materially affect the VaR predictions - the accuracy was not changed. While this model is nonlinear in G and total scale invariance in G would be surprising given the nonlinear relationship, the fact that the other fitted parameters likely adjusted is not surprising.

Symmetric Absolute Value CAViaR Model

Below is the symmetric absolute value CAViaR model:

$$f_t(\boldsymbol{\beta}) = \beta_1 + \beta_2 f_{t-1}(\boldsymbol{\beta}) + \beta_3 |y_{t-1}|.$$

Asymmetric Slope CAViaR Model

Below is the asymmetric slope CAViaR model:

$$f_t(\boldsymbol{\beta}) = \beta_1 + \beta_2 f_{t-1}(\boldsymbol{\beta}) + \beta_3 (y_{t-1})^{+} + \beta_4 (y_{t-1})^{-}.$$

Indirect GARCH (1,1) CAViaR Model

Below is the Indirect GARCH (1,1) model:

$$f_t(\boldsymbol{\beta}) = (\beta_1 + \beta_2 f_{t-1}^2(\boldsymbol{\beta}) + \beta_3 y_{t-1}^2)^{1/2}.$$

Multivariate CAViaR Model Specifications

The multivariate CAViaR model takes inspiration from the models described above in several specifications, as mentioned in the original specifications. The general model form looks like the specification below:

$$f_t(\boldsymbol{\beta}) = \beta_0 + \sum_{i=1}^{p} \beta_i y_{t-i} + \sum_{j=1}^{m} \beta_{j+p} f_{j,t-1} + e_t.$$

As with the univariate CAViaR model, the object of interest is a θ percentile return and the model is fit iteratively to minimize the loss function on the training data. However, there are some notable differences between the univariate model and the multivariate model. First, there are no moving average terms (lagged error terms) - the reasoning for this is because this model aims for a clear economic interpretation, and crisp interpretations of MA models are harder to create. Also, moving average models require recursive estimation since error terms are not observed, and so developing a method to work with these errors in a robust regression framework is challenging.

Second, in some of the specifications below, there are lagged return variables. This is similar to the univariate CAViaR specification, though there is often more than 1 lag as in the univariate model - there are p lags in the dataset. Third, in all of the specifications below, there are m diffusion indices used in each model lagged by one time step to avoid look-ahead bias.

Multivariate CAViaR: No Lags Model

$$f_t(\boldsymbol{\beta}) = \beta_0 + \sum_{j=1}^{m} \beta_j f_{j,t-1} + e_t$$

Multivariate CAViaR with Autoregressive Terms Added

$$f_t(\boldsymbol{\beta}) = \beta_0 + \sum_{i=1}^{p} \beta_i y_{t-i} + \sum_{j=1}^{m} \beta_{j+p} f_{j,t-1} + e_t$$

Multivariate CAViaR with Symmetric Absolute Value Autoregressive Terms Added

$$f_t(\boldsymbol{\beta}) = \beta_0 + \sum_{i=1}^{p} \beta_i |y_{t-i}| + \sum_{j=1}^{m} \beta_{j+p} f_{j,t-1} + e_t$$

Multivariate CAViaR with Asymmetric Slope Autoregressive Terms Added

$$f_t(\boldsymbol{\beta}) = \beta_0 + \sum_{i=1}^{p} \beta_i (y_{t-i})_+ + \sum_{j=p+1}^{2p} \beta_i (y_{t-i})_- + \sum_{k=1}^{m} \beta_{k+2p} f_{k,t-1} + e_t$$

Fitting the Models

To fit the models, an optimal value of m diffusion indices and p autoregressive terms are added (or $2p$ in the case of the asymmetric slope model). The optimal values of these parameters are determined using a validation dataset. In all of the runs below, there are a total of 5 years of trading days, or about 1,260 days assuming 252 trading days a year. The adjusted closing prices are logged and differenced, shortening the dataset by one. After doing this, the last 250 data points are reserved as test data, and the 250 data points before that are used as a validation set. Measured by the loss function written out below, the values of p and m that minimize losses are chosen and the optimal model is refit over both the training and the validation data combined and then evaluated on the test data. Note that there is an optimal model which is chosen for each of the four multivariate CAViaR specifications described above, so there are 4 optimal sets of p and m chosen for each set of models. Thus, there are 8 models compared on the test data - 4 univariate CAViaR models and 4 multivariate CAViaR models.

From the CAViaR paper, the θth regression quantile is defined as any $\hat{\boldsymbol{\beta}}$ that solves the following loss function:

$$\overset{argmin}{\beta} \frac{1}{T}\sum_{t=1}^{T}[\theta - I(y_t < f_t(\boldsymbol{\beta}))][y_t - f_t(\boldsymbol{\beta})]$$

Code

The code can be found at the location listed below in the "STAT_771_Class_Project.Rmd" file.

https://github.com/stevenmoen/stat_771_final_project

Big HMM Function

Literature Cited

Engle, Robert F, and Simone Manganelli. 2004. "CAViaR." *Journal of Business & Economic Statistics* 22 (4). Taylor & Francis: 367–81. https://doi.org/10.1198/073500104000000370.

Holton, Glyn A. 2014. "History of VaR - Value-at-Risk: Theory and Practice." https://www.value-at-risk.net/history-of-value-at-risk/.

Kassambara, Alboukadel. 2018. "Transform Data to Normal Distribution in R: Easy Guide - Datanovia." https://www.datanovia.com/en/lessons/transform-data-to-normal-distribution-in-r/.

Longerstaey, Jacques, and Martin Spencer. 1996. "RiskMetrics - Technical Document." J.P. Morgan/Reuters. http://www.jpmorgan.com/RiskManagement/RiskMetrics/RiskMetrics.html.

Malkiel, Burton Gordon. 2003. *A random walk down Wall Street : the time-tested strategy for successful investing*. New York: W.W. Norton.

NA. 2017. "iShares Institutional Guide to Bond ETFs." https://www.complianceweek.com/download?ac=5780.

———. 2020. "Sector ETFs | Seeking Alpha." https://seekingalpha.com/etfs-and-funds/etf-tables/sectors.

Portnoy, Stephen. 1991. "Asymptotic behavior of regression quantiles in non-stationary, dependent cases." *Journal of Multivariate Analysis* 38 (1): 100–113. https://doi.org/https://doi.org/10.1016/0047-259X(91)90034-Y.

Richardson, Matthew P., Jacob (Kobi) Boudoukh, and Robert F Whitelaw. 2005. "The Best of Both Worlds: A Hybrid Approach to Calculating Value at Risk." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.51420.

Shumway, Robert, and David Stoffer. 2016. *Time Series Analysis and Its Applications With R Examples Fourth Edition*. Fourth. Springer. https://www.stat.pitt.edu/stoffer/tsa4/tsa4.pdf.

Stephens, Matthew. 2018. "HMM example." https://stephens999.github.io/fiveMinuteStats/hmm.html.

Stock, James H, and Mark W Watson. 2002a. "Macroeconomic Forecasting Using Diffusion Indexes." *Journal of Business & Economic Statistics* 20 (2). Taylor & Francis: 147–62. https://doi.org/10.1198/073500102317351921.

———. 2002b. "Forecasting Using Principal Components from a Large Number of Predictors." *Journal of the American Statistical Association* 97 (460). [American Statistical Association, Taylor & Francis, Ltd.]: 1167–79. http://www.jstor.org/stable/3085839.

Group Feedback III

Please note that this feedback was given on Friday, November 20th. I was trying to problem-solve how to fit a Hidden Markov Model – specifically, how do I fit the thing to the data that I had? I was confusing myself by trying to use the distributions of the loss functions (somehow), but the problem is that there's no "data" in that case – it's already factored into the loss function. I sent the group a detailed explanation of what problem I was trying to solve, and I was able to boil it down into a key question – how do I select an optimal model in the presence of a loss function that penalizes overprediction much more severely than underprediction?

My group was able to give thoughtful answers on the background and motivation of my problems – namely whether the I was explaining the economic theory of the efficient market hypothesis correctly – as well as an idea of incorporating additional lagged terms into the HMM.

As it goes, I ended up not incorporating either piece, at least not right now. I think the problem is well-motivated from an economics perspective, but I think it really could use mentorship from someone in the field of financial economics who sees the promise of the models that I am proposing if it were fed data that could help it shine. I also ended up not incorporating the idea of incorporating additional lagged terms simply because the code I was working off of was not set up to do that (at least from my perspective) and I didn't want to plug things into a black-box package without really understanding the guts of the code.

Group Feedback IV

This feedback was given on Friday, November 27th. I was able to show my group the work from the changepoint algorithm with some added smoothing. My group was positive and encouraging about my work, and encouraged me to add more smoothed data and to make sure that the narrative was clear. I did end up working on the narrative and making sure it was smoother, but I chose not to add more plots simply because I wasn't completely sure what they would add besides additional length to the paper. The problem I find with this work is that I fear it's caught in a no-mans-land of not theoretical enough for a statistician to care about while also not generating powerful enough results for an economist to notice. At least not yet, and I have the feeling that this work can really grow into something great.

Group Feedback V

This feedback occurred on Saturday, November 28th and Sunday, November 29th. A group member wrote back with some questions about my thesis – the most pertinent of which was whether the changepoint algorithm was well motivated since adding together PCA vectors doesn't necessarily produce something meaningful since they are mutually orthogonal. Fair point, but I suppose the usefulness or lack thereof of the algorithm must speak for itself based on its predictive accuracy. I’m not sure if it’s there yet, but I believe there is some potential in the algorithm.

Lastly, I ran the narrative by my parents, and they seem to be quite positive about it. I hope you enjoy reading it as much as I enjoyed writing it.

- Steven