GAUSSIANS MODELS IN THE STOCK MARKET

Johnson Lai

DECEMBER 21, 2021

UNIVERSITY OF TORONTO
CSC C11

Final Report

Written December 20, 2021

1 Abstract:

The stock market is a long-term asset utilized by the most successful in society. If someone were able to predict the future, they could gain untold riches trading shares. Many indicators of stock performance exist, from support lines to weekly moving averages. This paper discusses the trends of the fear index as a mixture of Gaussian models. Most significantly, the fear index maintains a steady model, then under a market shock, transfers rapidly to extreme fear or no fear. This rapid change quickly reverses; however, and soon thereafter the fear index approaches the steady model.

2 Introduction:

The stock market is a beast that millions of players try to game every day. Companies make and break in this ruthless area, and hedge funds try to predict the next big winner. The goal of this paper is to identify the underlying machinations of VIX, the fear index. VIX quantifies the general fear of the stock market's performance on any given day. A value of 10-15 is considered low fear, while a value above 20 is concerning. By looking at the trends of fear, it may be possible to predict when there is a high amount of fear and thus a market crash. Being able to look ahead at the stock market can save a stakeholder's life savings - a worthwhile endeavor. Knowing when a crash will occur can also help buyers get in at a time prices are low, multiplying investments in the years to come.

However, one might question, is it possible to predict the fear index? One would think no, it's not possible. Also, Zillow wouldn't be in debt if they could predict the housing market, and surely, they have brilliant minds at work. But there is a silver lining to this- you can predict long term trends and predict probabilities

that fear will increase. Secondly, past data. Is that enough to predict the future? Again, the answer is no. Nothing is guaranteed. The pandemic hit hard, and the fear index shot up within a few months, something that could not be predicted by previous data. But then again, one could argue that the previous data had indicated a likelihood of market crashes every 10 years. There was one in the dot-com boom of 1999 and one in the housing crash of 2008. This once-a-decade phenomena could be a heightened probabilistic trend.

Finally, is it possible to utilize other indicators to forecast the market? Perhaps, yes. Hence we introduce EMV – the concept of "fear" in newspapers, media, and articles. Words that can illicit reactions and form the sentiment of the public. We will delve into both VIX and EMV data to learn what can be forecast, and how accurate these models are.

3 Model Specification:

The data model used to analyze the VIX data was the Mixture of Gaussian (MoG) model. This model utilizes the fact that most repeating events follow a gaussian distribution or process, similar to a random walk. Initially, I had selected a Clustering Model, which takes in a data and attempts to group them due to how close the points are. The problem with this method is that there is only one label for all of the data points. That label is the VIX index. Given a large number of data, with no knowledge of the number of underlying distributions, I used a genetic algorithm to formulate hundreds of random Gaussian distributions.

Each time the model iterated, my Objective Function would determine how accurate the model was, and if it was getting more accurate. With each change in mean, deviation, and probability, the objective function would update to reflect better or worse constraints. After 100 iterations of 1, 2, 3, 4, and 5 Gaussian models, I had to determine which number was the most representative of reality. This was a hard selection as 3, 4, and 5 fit well according to the Chi-squared test. Ultimately, I settled on 3 Gaussians, as simpler can be more accurate (one such example is Runge's phenomenon).

4 Fitting & Diagnostics:

First, I had to determine the number of bins my model would utilize. Since the final distribution was a Chi-squared distribution of M-3k-1 degrees of freedom, and the maximum value of k was 5, I knew I needed at least 17 bins. Looking at the data, the lowest value was around 10 and the highest was below 60. Hence, I chose an M value of 25 - each bin containing 2 units of the fear index, from 10 to 60. From there, I ran the models through 100 iterations and recorded the final values. Below are the results for the models that passed a significant P-value of 0.05 (mean, standard deviation, probability):

[ ]:
```
[ 5.95208039e+01 1.71018123e+01 4.52684667e-02
 1.87215193e+01 4.36981913e+01 3.29446868e-02
 2.93842131e+00 1.99765429e+01 5.47663918e-02
 5.98790364e+01 5.70736000e+01 2.30329889e-01
 -8.03040097e+01 3.89174793e+01 4.17266683e-01]
```

[ ]:
```
[-5.65112356e+00 1.27709321e+01 1.34334970e-01
 5.09710103e+01 3.64707905e+01 6.88385493e-01
 1.81406949e+01 3.66117405e+01 4.12776595e-02
-6.63184198e+01 5.88879531e+01 6.11043860e-02]
```

The model of 3 Gaussians was as followed:

N1 = (6.89, 1.773, 0.111)

N2 = (12.541, 5.826, 0.722)

N3 = (-5.65, 1.227, 0.125)

I identified the K=3 Model as the candidate because it was simpler, and the probabilities summed close to 1 after rounding. The probability sums of K=4 and K=5 did not seem to work out too well, as there was too

much error in the genetic process of creating sums to 1 and a lack of iterations. Additionally, 100 iterations were taking in excess of 20 minutes per run and limited my model creation process.

As for the EMV models, I utilized a mixture of Linear, LASSO, Ridge, and Elastic Net regression. In image 5, you can see just how similar EMV's Overall Tracker is to VIX. Sub-trackers in EMV, such as the Policy-Related tracker, follow a near identical pattern as well. Linear produced a moderate estimate for future forecasts. However, since it could only predict a linear path, it gave an estimation of close to the mean of 30 years of data. That wasn't particularly helpful but is relatively accurate in confirming what I gleaned from VIX data. LASSO, on the other hand, was nearly the same in complexity as OLS. However, there were still some particularities to the model. Namely, it formed a curvy line that was not as pronounced as the EMV data. This curve predicted an upward swing for the future, and soon. While adjusting coefficients for lambda; however, I noticed a noticeable sharp angle in the final function curve. On the other hand, Ridge regression created a similar curve to LASSO, except it continued to create smooth function curves even with increasing lambda values. While LASSO seems easier to interpret and plot, Ridge does not zero-out some variables with its extra complexity. This keeps the curve smooth. Ridge regression also predicted an upward swing for the future, but not as pronounced as LASSO. Finally, Elastic Net regression looked like an in-between variant of LASSO and Ridge regression. Being in-between, it likely contains the presence of multicollinear variables in EMV, and thus stays away from the sharp angles of LASSO. At the same time, it normalizes by adding a Ridge function to equalize variable coefficient weight.

5 Forecasts:

Given my Gaussian models and the EMV regression, I can predict the following: (1.) The current fear index will trend horizontally and down in the long term. (2.) The fear index is due for a major correction near the end of 2019, as it has been over 10 years since the last major correction. (3.) The fear index recently dipped under 10, indicating stakeholders are too comfortable. This never lasts too long, given the MCP table (Image

1). (4.) The Media is stoking fear again, as the EMV data is beginning to spike. This agrees with point number 2. (5.) LASSO, Ridge, and Elastic Net regression are predicting an increase in the EMV data. This agrees with points number 2 and 4. Hence, given the data from 1990 to 2019, I can predict a sharp rise in fear that will raise the VIX floor once again.

6 Discussion:

My Mixture of Gaussian Model offers an insight into broad trends under the VIX fear index. The probabilities of the 3 Gaussian distributions give a chance at each scenario, and show that all things considered, the most common event is no change. A lack of entropy. The EMV regressions show that some data can precede changes in VIX. These things are usually words of fear written in the media- things related to policy, the economy, and bad news. In the long term, I can conclude, it is possible to gauge the direction of the fear index.

However, my data analysis was not perfect. Firstly, my model was not as accurate as it could. I was unable to run 1000 iterations and might have coding issues that propagated throughout the model. One large problem I encountered was that a model of Gaussians doesn't particularly say much of the next event to come, nor the timeframe it will happen in. To identify a timeframe, I had to look at the big picture, by graphing the data over time (Image 2).

In the end, is it possible to understand what causes volatility in the stock market? In a general sense, yes. Look to general sentiment, duration of highs and lows, and where past floors and ceilings have been. Check the media and websites for a spike in fear related imagery. Then what happens in the EMV index can precede the VIX index- and in turn, precede a volatile crash in the stock market.

7 Bibliography:

Godwin, J. Andrew. (2021). “Ridge, LASSO, and Elastic Net Regression”. https://towardsdatascience.com/ridge-lasso-and-elasticnet-regression-b1f9c00ea3a3

Hertzmann, A. Fleet, D. (2020). “CSCC11”. University of Toronto.

Kuepper, J. Scott, G. (2021). “Cboe Volatility Index”. https://www.investopedia.com/terms/v/vix.asp

8 Appendices:

	N1	N2	N3
N1	0.11	0.61	0.28
N2	0.08	0.85	0.07
N3	0.19	0.73	0.08

In this figure, the Row trends to the Column.

Image 1: MCP table

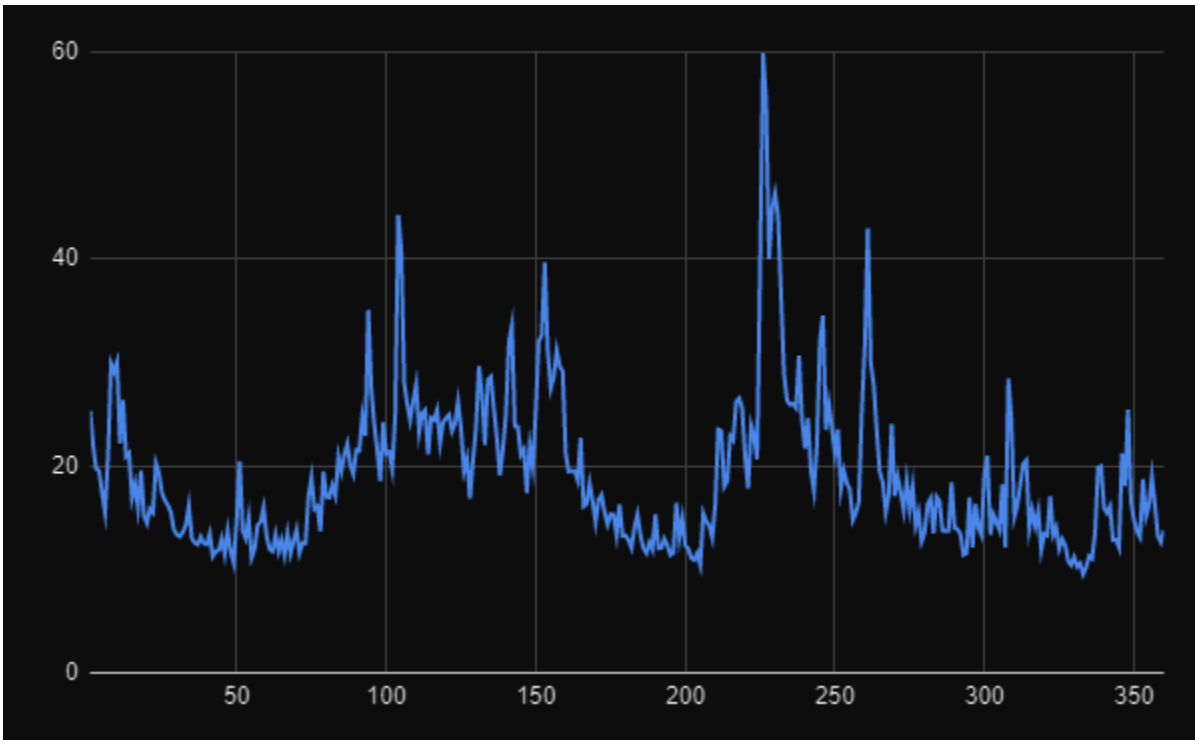

Image 2: VIX over time

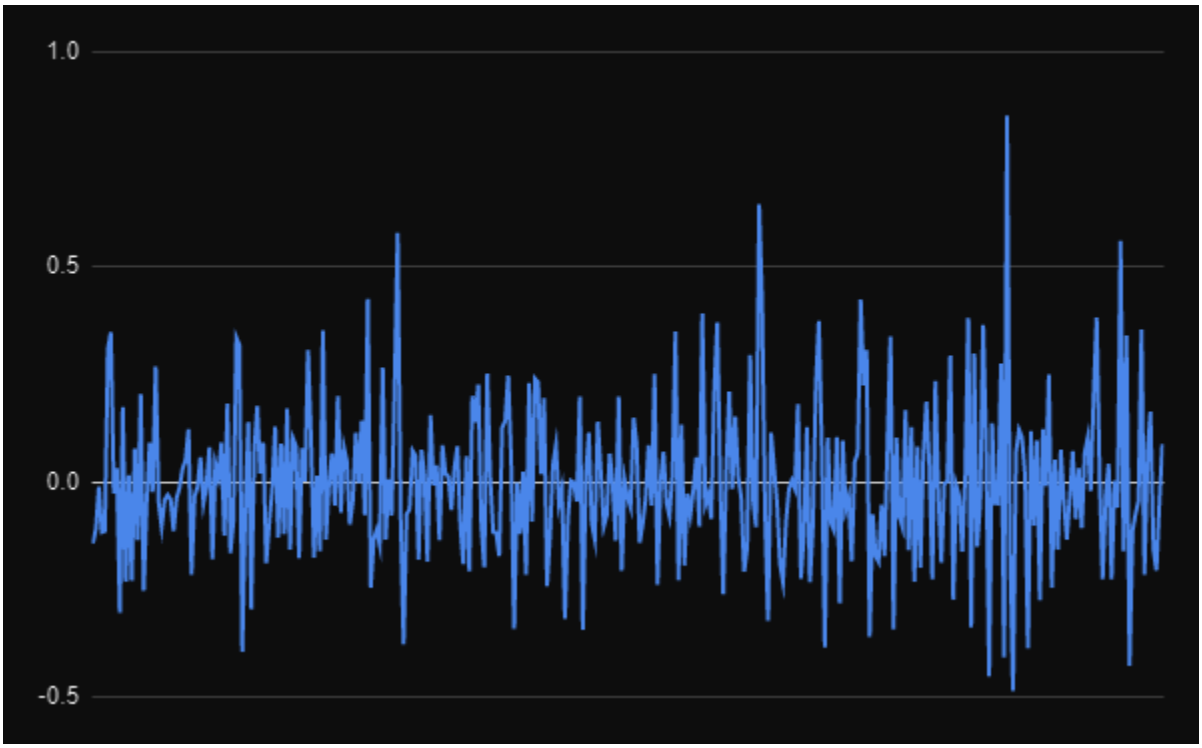

Image 3: Change of VIX over time

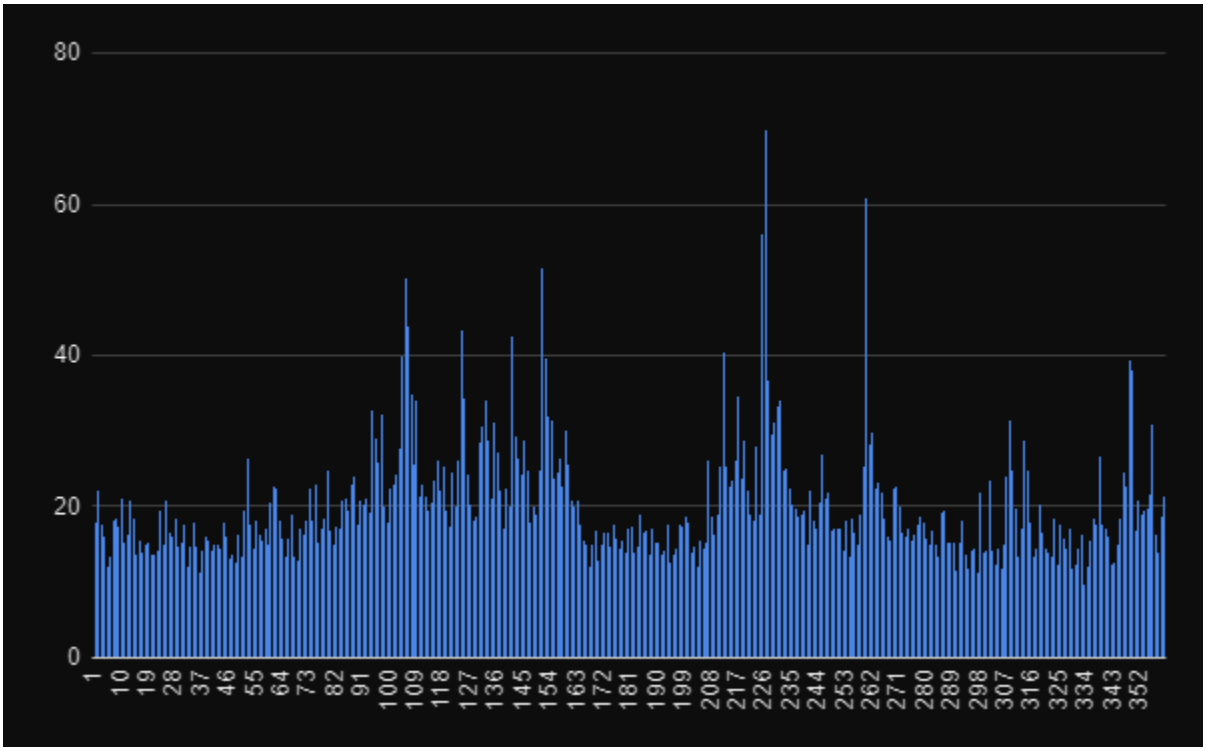

Image 4: EMV Tracker over time

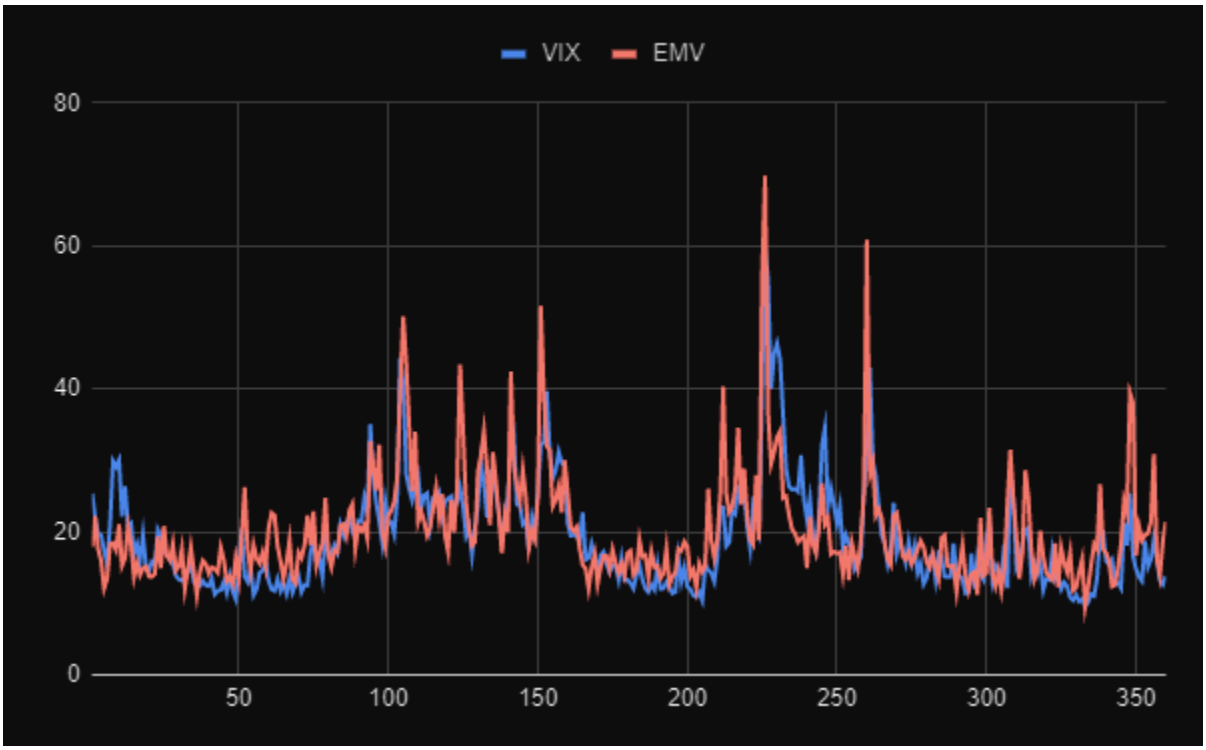

Image 5: VIX and EMV data overlayed

```
[ ]: def integrand(x, m, s):
         return (1/(s*(2*np.pi)**0.5) * np.exp((1/2*1/s*1/s) * (x - m)**2))

     def f(X):

         dataset = "MVIX.pkl"
         dataLoad = load_pickle_dataset(dataset)
         data = dataLoad['Adj Close']

         obj = 0

         #BINs
         k = math.floor(len(X)/3) # num of gaussians
         m = 25
         pHat = [0]*30
         pT = [0]*30

         for i in range(0, 360):
             for j in range(10, 58, 2):
                 if data[i] >= j and data[i] < (j+2):
                     pHat[math.ceil((j-10)/2)] += 1/360

         for i in range(0, m):
             for o in range(10, 59, 2):
                 for nG in range(0,k):
                     pT[i] += X[2+nG*3] * integrate.quad(integrand, o, o+2, args =␣
      ↪(X[nG*3], X[1+nG*3]))[0]

         for j in range(0,m):
             obj += pHat[j] * np.log(pT[j]/pHat[j])

         return obj*-2
```

Image 6: Objective function for GA solver