36-723 Hidden Markov Models : Homework 2

Rodu, Section A

Michael Rosenberg, mmrosenb@andrew.cmu.edu

09 May, 2016

1. Introduction

During the 2008 financial crisis, the real estate bubble caused such a devaluation in houses that the construction market was heavily disrupted. While we saw that certain construction company stocks were generally safe investments as real estate valuation increased and new homes became more demanded, but as the housing market crashed and failed to recover during and after 2008, the construction industry has had to deal with a weak construction market within developed countries and a volatile construction market within the emerging markets.

We are interested in seeing if there was any meaningful difference in the volatility of important construction company stock prices before the financial crisis and after the financial crisis. Our dataset is the daily stock price of Caterpillar (CAT) from 2000 to 2015. Caterpillar is a multinational equipment company that generates much of the machinery used for constructing both residential and industrial establishments. We will fit two different Hidden Markov Models (HMMs): one for the data between 2000 and 2006, and one for the data between 2010 and 2015. We will compare these two selected models to see if they predict relatively different data-generating processes. The measurement we will fit upon is relative price change, which we measure at time t as

$$RelPriceChange_t = \frac{Price_t - Price_{t-1}}{Price_{t-1}},$$

where $Price_t$ is the price of Caterpillar's stock at time t. We choose to represent relative price change as a measure of volatility.

2. Exploratory Data Analysis

We see that betwen January 1st, 2000 and December 31st, 2015, we have 4025 observations. That being said, given that we are only making comparisons between observations in the 2000-2006 timeframe and observations in the 2010-2015 timeframe, we will not be considering many of the observations found in the 2006-2009 time frame.

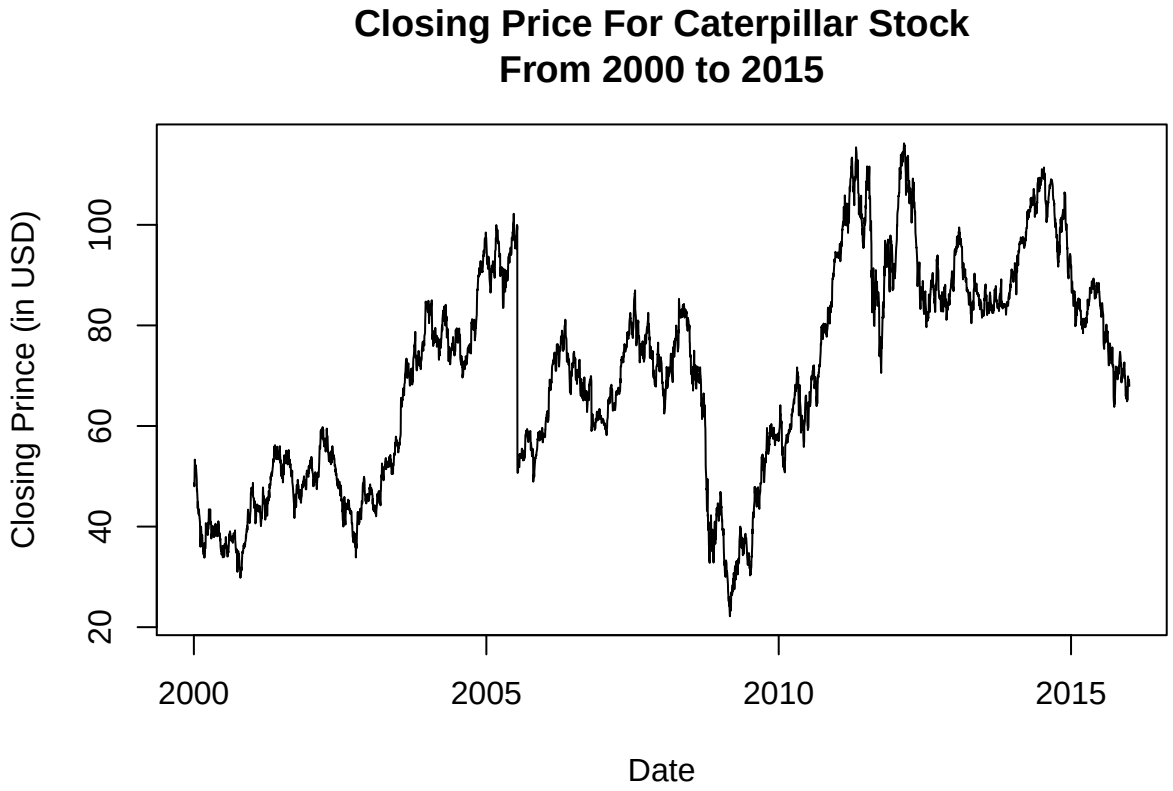

Figure 1: Figure Price for the Caterpillar (CAT) stock over time.

We see when we plot out the time series a very obvious dip in between 2008 and 2010. This is the last recession shocking construction prices in our data, and thus I found this time period to be an anomaly that should be ignored when studying regime-switching within the dataset. We see get general positve trends between 2000 and 2006 and between 2010 and 2014, although we see an apparent dive at the end of 2015.

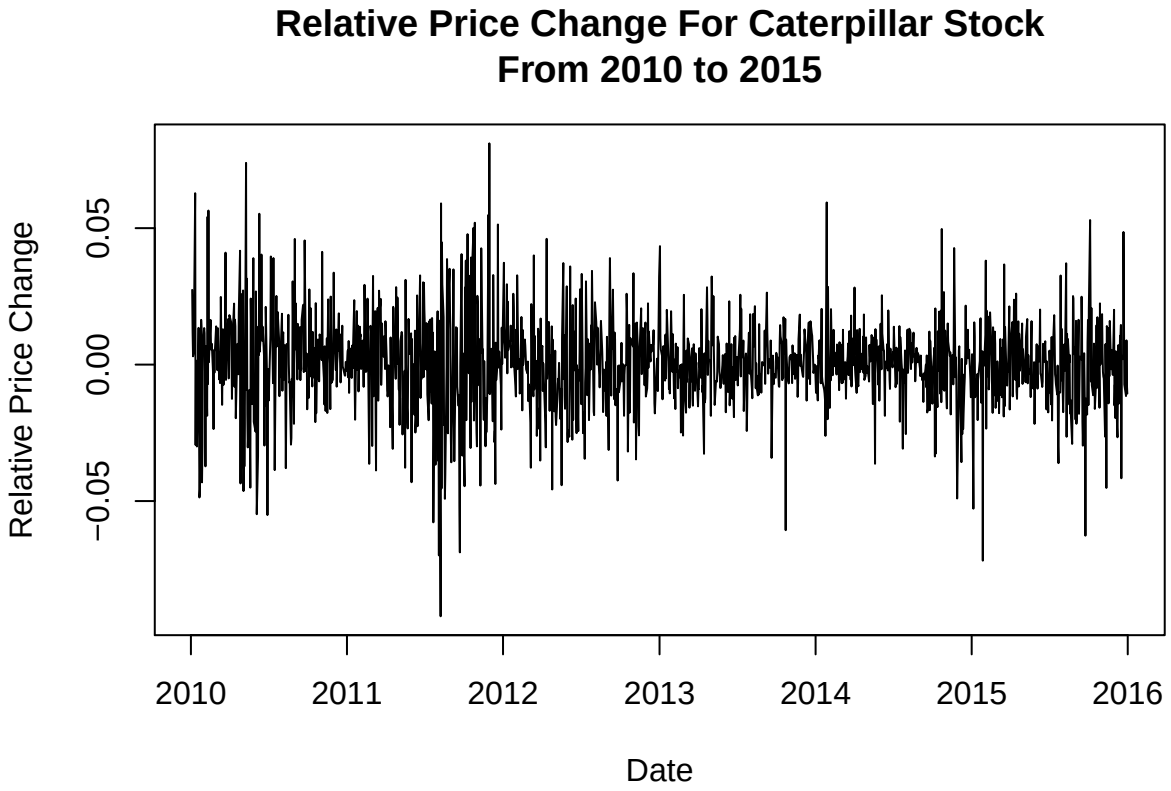

Figure 2: Daily Relative Price Change for the Caterpillar Stock from 2010 to 2015.

We see that in general, prices tend to not have severe daily relative change, with most relative changes being rather close to 0. That being said, there are instances of high volatility during certain time periods, in particular between 2011 and 2012 and near the end of 2015.

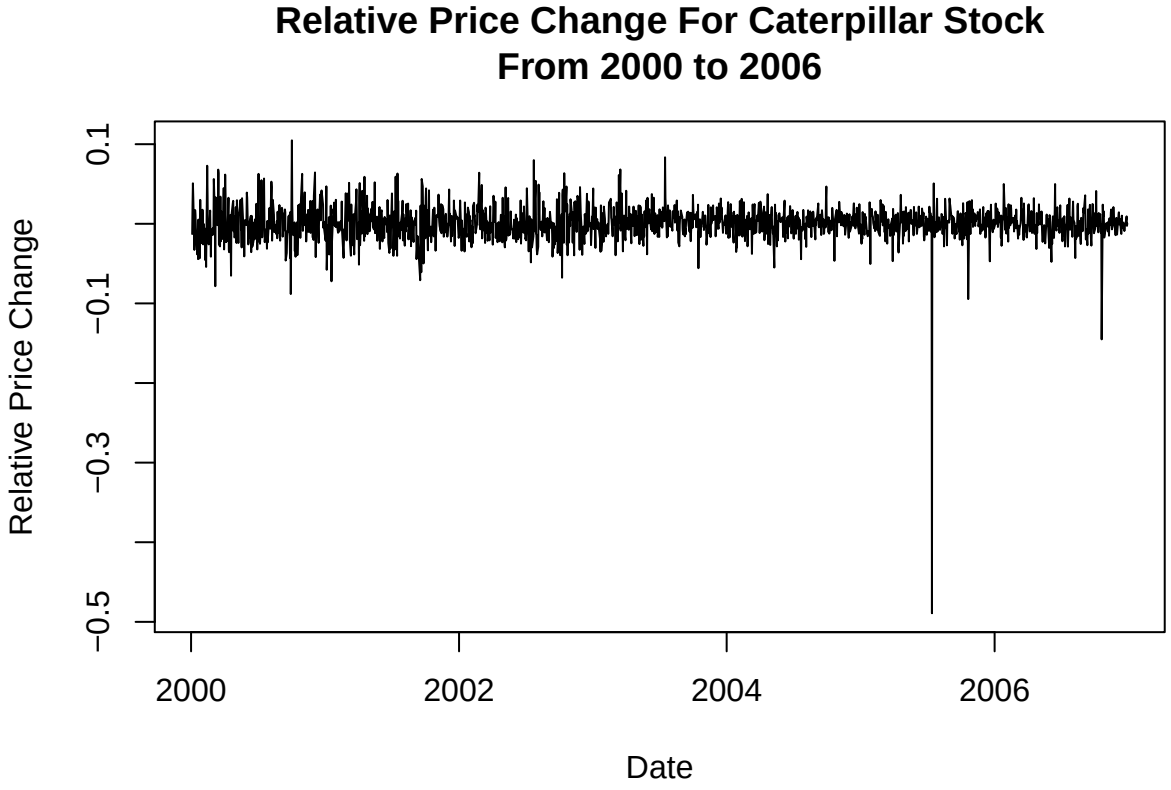

Figure 3: Daily Relative Price Change for the Caterpillar Stock from 2000 to 2006.

We can see that there is an apparent outlier in the data. The date at which this occurs is 2005-07-14, Which was a day in which Caterpillar took on a major stock split[1]. Due to the fact that this event is not a very strong representation of the changing value of Caterpillar and more of a representation of a mechanical choice with the stock, it seems worthwhile to remove this outlier from our analysis in order to discount this severe stock-splitting anomaly.

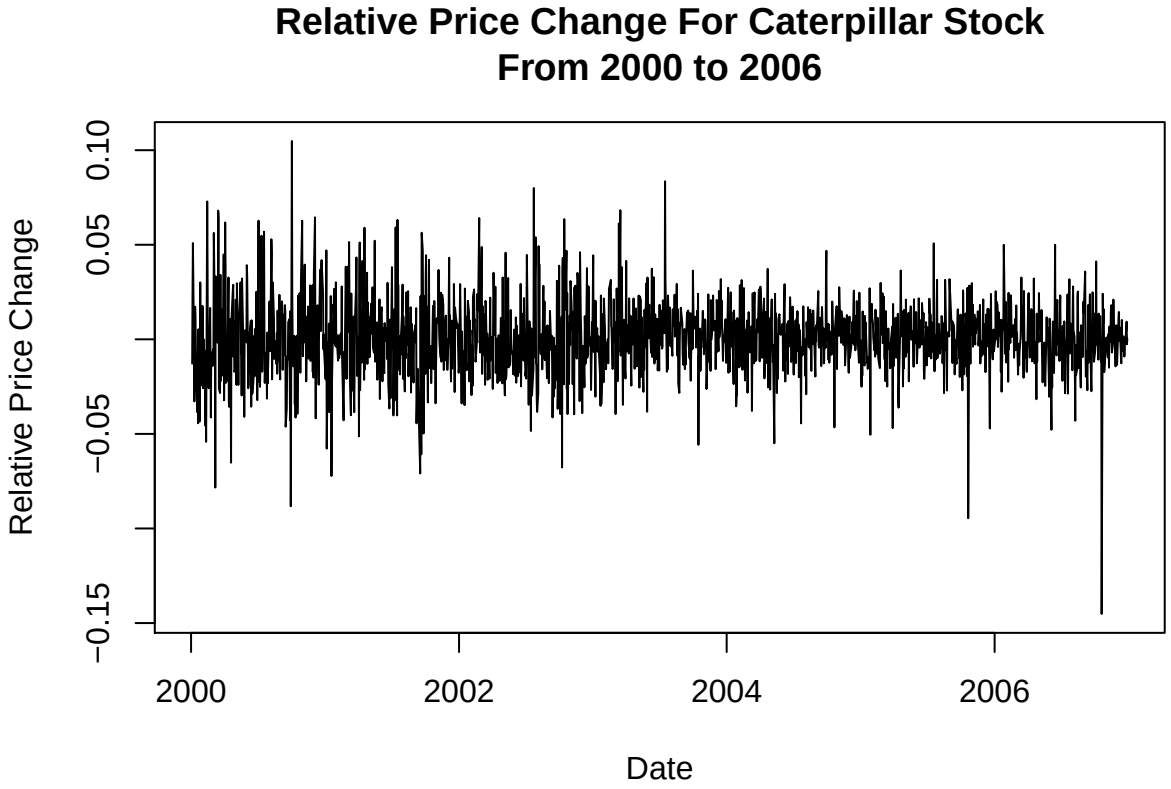

Figure 4: Daily Relative Price Change for Caterpillar Stock from 2000 to 2006, with our sizable outlier removed.

We see a sizable amount of volatility in the early 2000s, but general stability in the middle 2000s with a few outliers creating sizable changes in relative price. These regime shifts between high volatility and low volatility regimes are similar to what we saw with stocks in the current time frame (Figure 2). This suggests that the current data-generating process may be a reasonable representation of the data-genrating process 15 years ago, which may suggest that the market structure for construction hasn't changed severely after the housing bubble.

3. Model Selection and Diagnostics for 2010 to 2015 Time Period

We will choose the number of states for our model on the data from 2010 to 2015 by comparing the log-likelihood fit of our model and the simplicity of our model on the data from this time frame for models with up to 8 fitted hidden states. We wil fit these models using the Baum-Welch algorithm. Prior to training, we will initialize our parameters with a random row-stochastic transition matrix, a random distribution initialization of our initial state vector, and an initialization of our emission distributions as standardized Gaussian distributions. We choose to initialize these emission distributions this way because our emission variable (relative price change) is continuous and a Gaussian distribution is a simple univariate distribution

[1] "CAT Split History." SplitHistory.com. SplitHistory.com, 2016. Web. 05 May 2016.

that has easy to interpret parameters. However, in the later parts of the model selection phase, we will consider other possible emission distributions if necessary.

While it would be more optimal to select the number of states for our model using cross-validation by forward-chaining on a test set, calculating the cross-validated log-likelihood score for a given model would take $O(f(h^{n/f}))$ iterations, where f is the number of folds used in cross-validation, n is the length of our training set, and h is the number of hidden states in our model. Given that this algorithm grows very large in the number of hidden states used to fit our model, it would be reasonable to simply select our model based on how well it fits the data and how interpretable the model can be rather than to select it based on out-of-sample prediction.

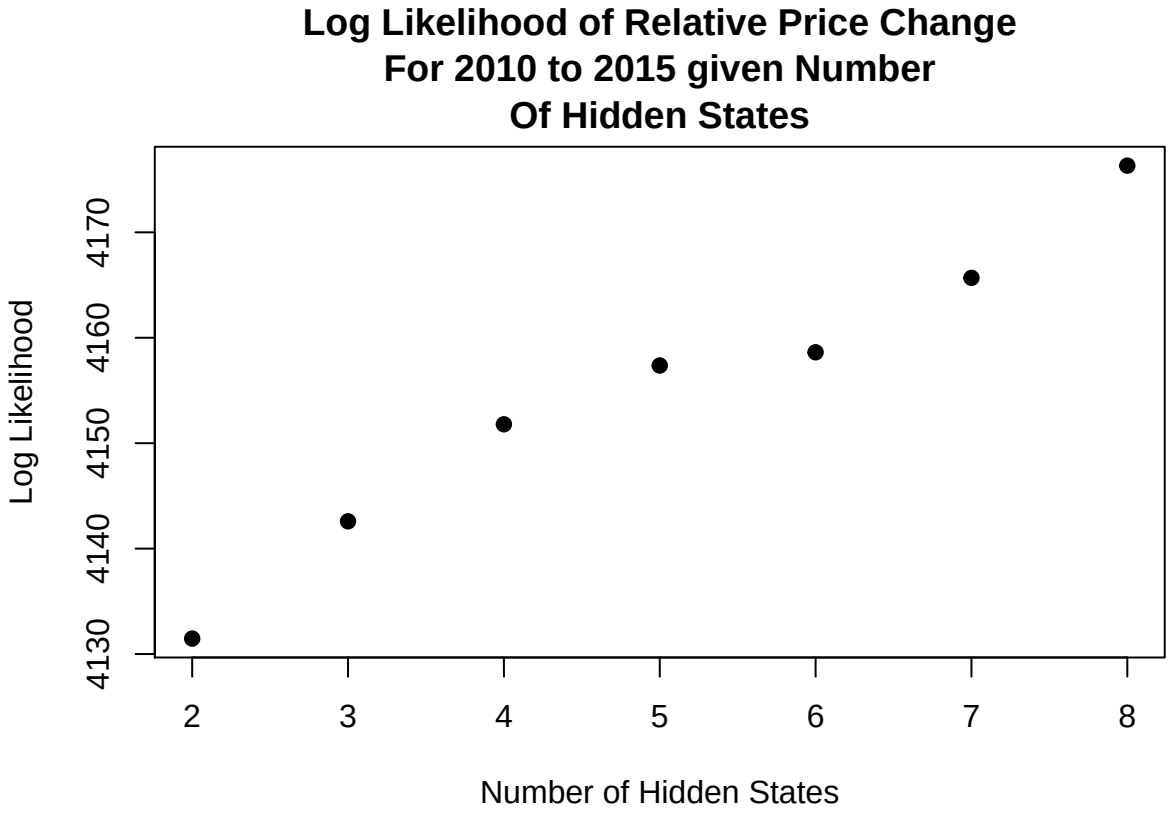

Figure 5: Graph of Log-Likelihood for the data range 2010-2015 for our Fitted Hidden Markov Models given the number of hidden states. All emission distributions are Gaussian.

We see that while the hidden markov model with 7 hidden states seems to be fitting the data the best, models with fewer hidden states are only performing slightly poorer than the 7 state model. Thus, in order to have a more interpretable model for comparison with a selected model from the 2000-2006 time frame, it could be reasonable to sacrifice a small amount of fit to the 2010-2015 data in order to have a model with fewer hidden states. Thus, I will choose a two state hidden markov model to fit this data.

In order to see how to test our assumption of Gaussian-distributed emissions given a particular hidden state, we will view our distributions of relative price change given the hidden state assignment generated by the Viterbi algorithm on this time frame.

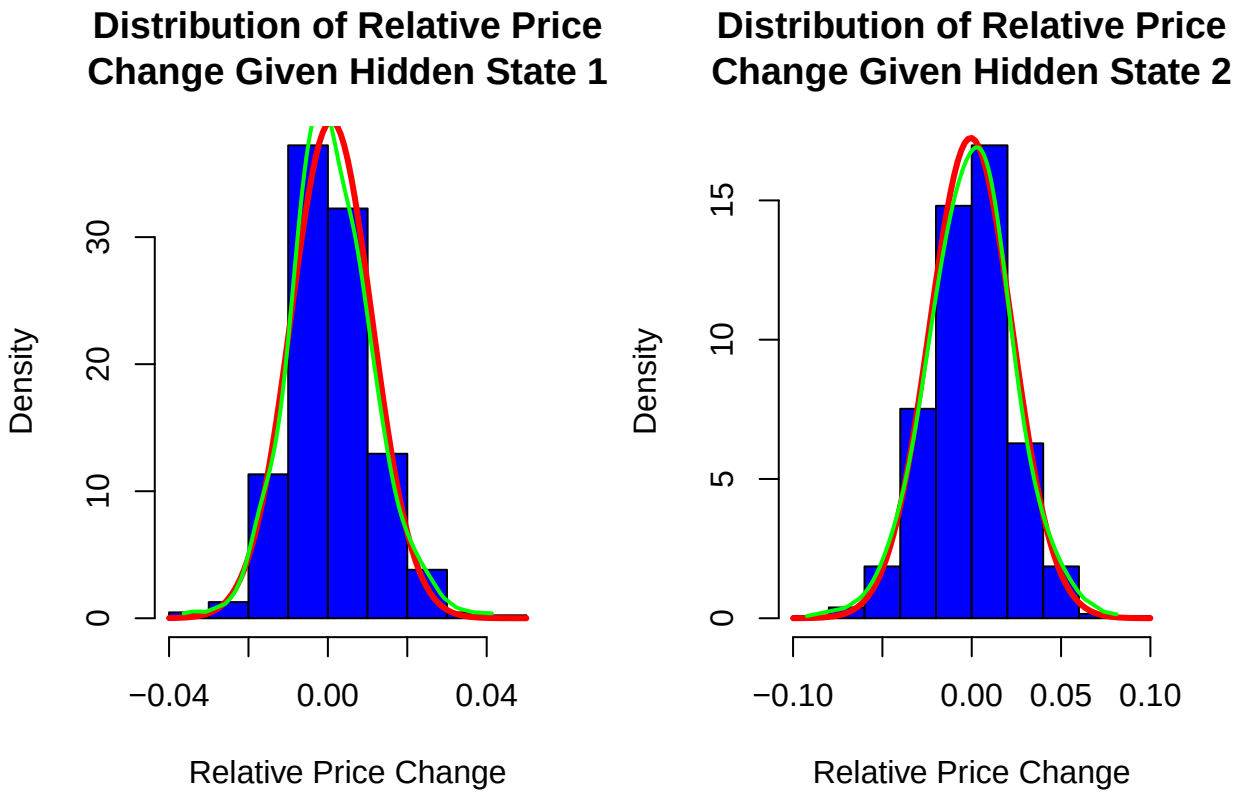

Figure 6: Plot of our distributions of relative price change for our 2010-2015 data given an estimated hidden state (estimated via the Viterbi Algorithm). The red line indicates the estimated Gaussian distribution of the emission probabilities given a particular state, and the green line indicates the kernel density estimate for fitting the distribution.

We see for our distributions of relative price change given hidden state 1 estimated by the Viterbi Algorithm (Figure 6, left), our estimated Gaussian curve for the distribution of emissions (red) is relatively close in structure to the kernel density estimate of the distribution (green), which suggests the chosen distribution is a relatively good fit for the data estimated to be in hidden state 1. For the distribution of relative price change given hidden state 2 (Figure 6, right), our estimated Gaussian curve (red) is slightly off from the kernel density estimate (green) in terms of mean, although it is structurally quite similar in shape to the kernel density estimate; this suggests that our fitted mean might be slightly off for the emission distribution given state 2, but the actual distribution class (in this case, Gaussian Distributions) may be an appropriate set of distributions to consider. Thus, while we may need to consider other possible choices for means and standard deviations for the emission distributions given our estimate hidden states, it seems relatively reasonable to treat the emission distributions given hidden states as Gaussian.

Hence, for the 2010-2015 data, we selected a model with 2 hidden states and Gaussian emission distributions.

4. Inference for the 2010-2015 Model

We see the log-likelihood for the model we fit on the 2010-2015 data is 4131, which is only a slightly poorer fit when compared to models that fit with more hidden states (see Figure 5). Thus, we could argue that this model is predicting the 2010-2015 data about as well as more complex alternatives.

	State 1	State 2
Probability of Starting in State	0	1

Table 1: Our Initial State Probability Distribution for the Model fit for the 2010-2015 data.

While the actual estimated probability for starting in state 1 is non-zero (1.447321×10^{-63} to be exact), it is so small that it is virtually zero. Thus, our model predicts that with almost certainty, we will start the time series in Hidden State 2.

	To State 1	To State 2
From State 1	0.95830	0.04166
From State 2	0.05354	0.94650

Table 2: Our estimated transition probability matrix for the MOdel fit for the 2010-2015 data.

We see that given that we are in state i, there is around a 95% chance I stay in state i and around a 5% chance I switch to state $j, j \neq i$. Essentially, this implies that within the model, there is a very small probability of switching states at any given moment in the realization of the chain, and once I reach a given state, I am likely to keep staying in that state over the subsequent steps in a realization of the chain. This suggests that in this hidden markov model, we stay within particular state regimes for a while given the high probability of staying in the same state during the transition phase, and thus this suggests that these states represent sustained effects throughout many periods in a realization of our data.

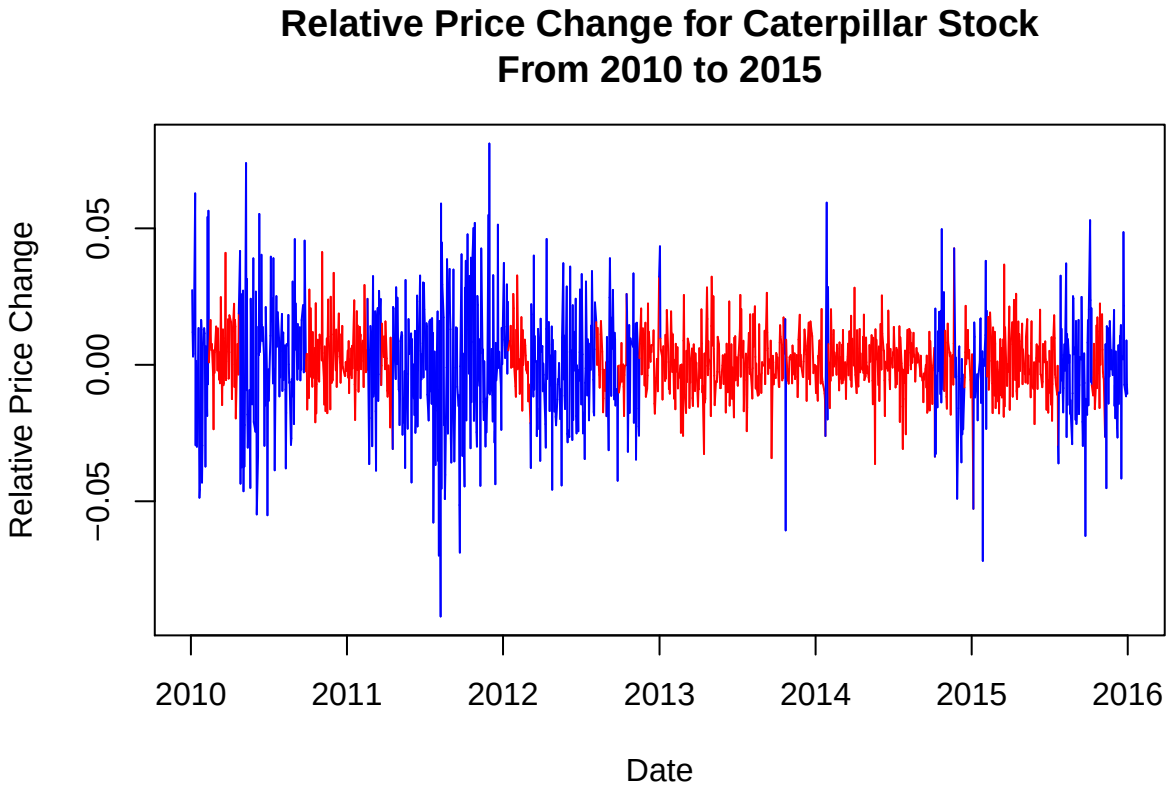

Figure 7: Plot of Relative Price Change for Caterpillar Stock With segments colored with the state the previous event was in. The hidden states were calculated using the Viterbi algorithm on our two state model. The first hidden state is colored in red, and the second hidden state is colored in blue.

We see that when we estimate the most likely hidden state path that our fitted model finds for the 2010-2015 data (Figure 7), our second hidden state (blue) represents instance of large fluctuations within the relative price change of the stock and our first hidden state (red) represents instances of small fluctuation within the relative price change of a stock. Thus, to some degree, the second hidden state represents instances of high volatility within the price of stock and the first hidden state represents instances of low volatility within the price of a stock. This aligns with our hypothesized theory that relative price change goes through two

different regimes: regimes of major fluctuations within the price of a stock and regimes of low fluctuations within the price of a stock. Our suggestion that the hidden states represent sustained effects in the data is apparent when we see the most likely hidden state path for our fitted data, as we often go through sizable periods of staying in either hidden state 1 or hidden state 2, with very few periods in which we rapidly switch between the two hidden states.

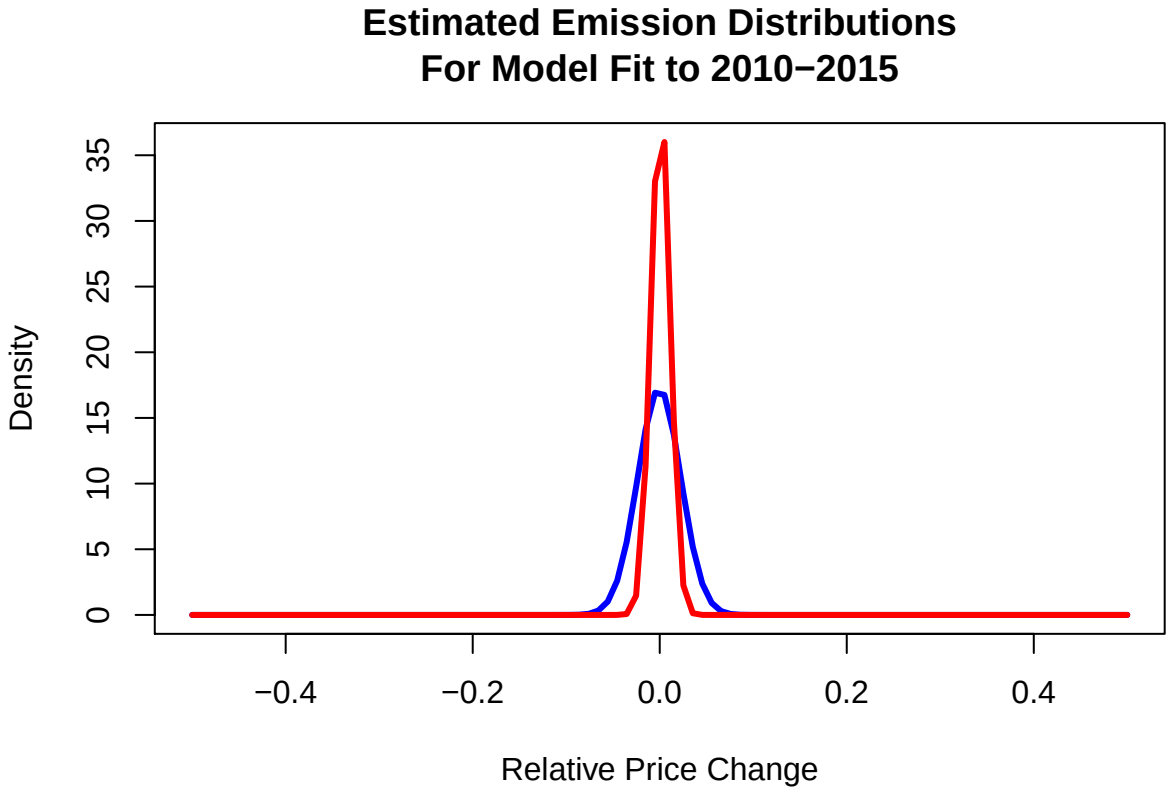

Figure 8: Graph of our Gaussian Emission Distributions given hidden states estimated via the Viterbi algorithm for the 2010-2015 data, Where the red line represents the emission distribution given hidden state 1 and the blue line represents the emission distribution given hidden state 2.

We see that our emission distributions are both centered almost at 0, but the distribution given hidden state 2 (blue) has slightly higher estimated standard error than the distribution given hidden state 1 (red). Thus, it is apparent that our relative price change variable tends to be centering around 0, but the emission distribution under hidden state 2 features greater variance from the mean than the emission distribution under hidden state 1. This provides more evidence that hidden state 2 is predicted to more often generate larger absolute relative price changes than hidden state 1. This further emphasizes the suggestion that hidden state 12is predicted to generate instances on average of high volatility within the price of the Caterpillar Stock while hidden state 1 is predicted to generate instances on average of low volatility within the price of the Caterpillar Stock.

5. Model Selection and Diagnostics for 2000-2006

Just as we did when we selected our model for the 2010-2015 section of the data, we will select our model for the 2000-2006 dataset based on the log-likelihood of seeing the data under our model and the interpretability (i.e. simplicity) of the model. We will consider up to 8 states for our model, and we will train our models using the Baum-Welch algorithm. Prior to training, we will initialize our parameters with a random row-stochastic transition matrix, a random discrete distribution for our initial state vector, and emission distributions modeled as standardized Gaussian distributions. As we did for model selection for the 2010-2015 dataset,

This choice to use Gaussian distributions for our emission distributions will be reviewed in the following sections. The choice to select a model based on how well it fits the data rather than based on minimal generalization error is due to the inefficiency of calculating cross-validated log-likelihood for our sequential dataset (see section 3).

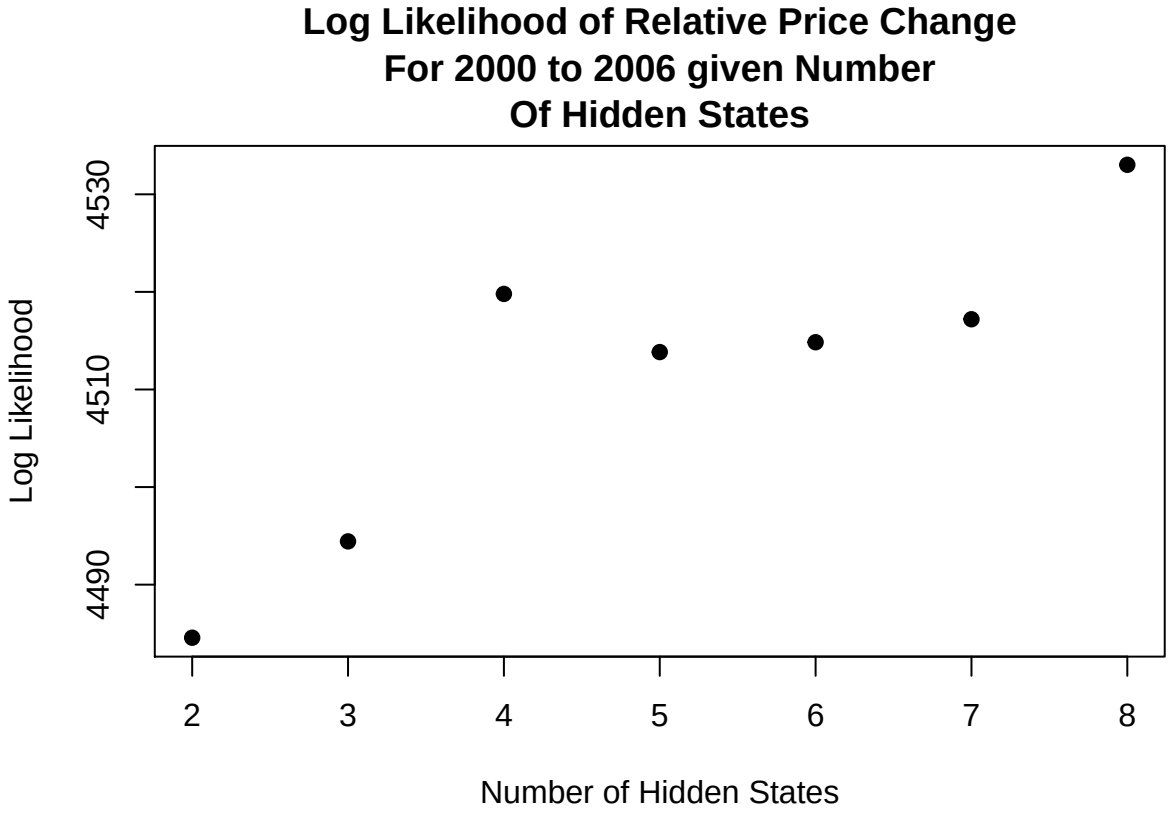

Figure 9: Log Likelihood for the Relative Price Change Data from 2000 to 2006 given the number of hidden states. These log-likelihoods were calculated using a fit generated by the Baum-Welch algorithm.

We see that we are steadily increasing the log-likelihood of seeing the data from 2000 to 2006. However, we notice that the difference between the log-likelihood of the model with 8 levels of hidden states and the log-likelihood of the model with 2 levels of hidden states is 48.46. Given that this is a small difference in terms of the scale of the individual log-likelihoods (this is a 1.081% increase in log-likelihood from 2 levels of hidden states to 8 levels of hidden states), it seems more reasonable to choose a model with simpler interpretability than to choose a model that performs slightly better when it comes to fitting the data. Thus, we will select the model with two levels of hidden states.

We will now check to see if our assumption of Gaussian distributed emissions is a reasonable assumption for the model we fit. We will first assign hidden states to our observations using the Viterbi algorithm, and then see if our observations given hidden states are approximately distributed based on our estimated emission distributions.

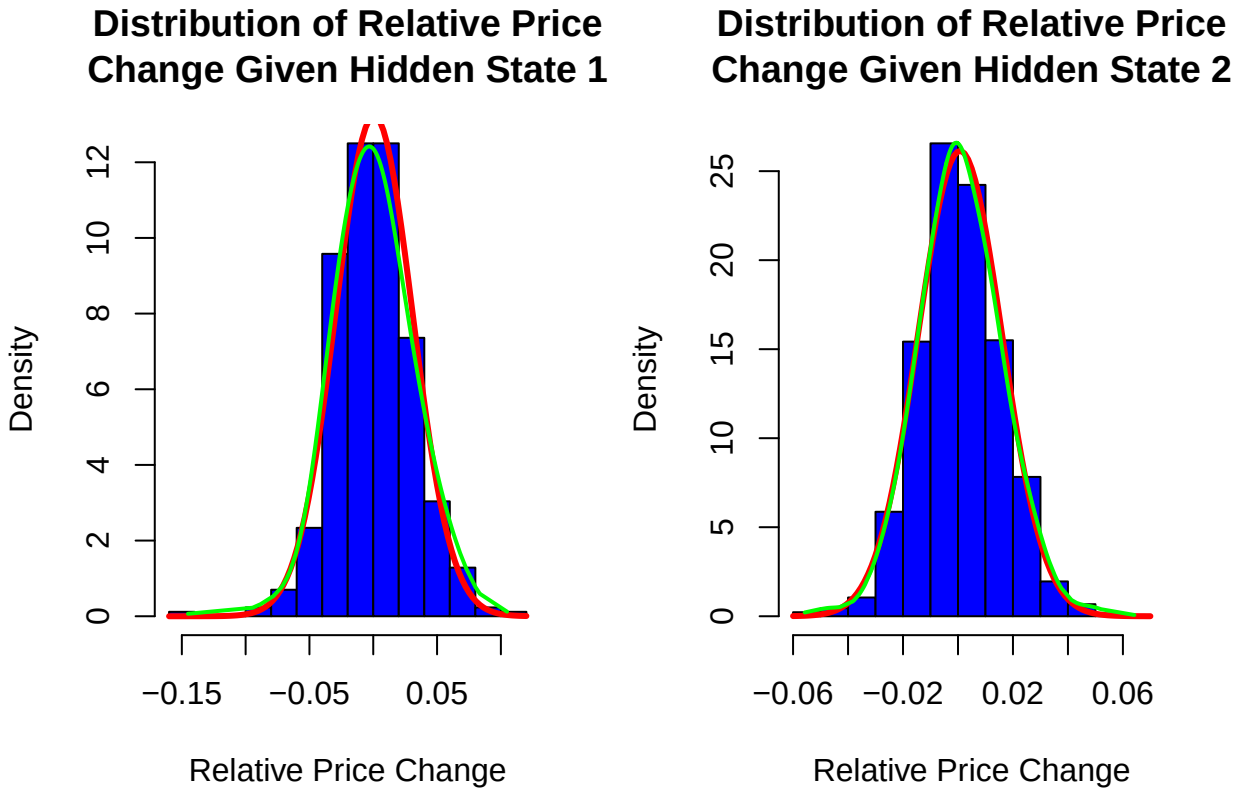

Figure 10: Distributions of Relative Price Change Given Hidden State Levels on the 2000-2006 part of the dataset. These hidden state levels are calculated upon this part of the dataset using the Viterbi Algorithm. The red line indicates the conditional Gaussian distributions estimated by our selected HMM, and the green line indicates the Kernel Density estimate of the given conditional distribution.

We can see that our fitted Gaussian Distribution for the emissions of hidden state 2 is very similar to the Kernel Density Estimate for the distribution of observations labeled to be generated by hidden state 2 (Figure 10, right), which suggests that this fitted Gaussian Distribution is a reasonable estimated distribution for this part of our data. However, when we look at the distribution of observables given hidden state 1 (Figure 10, left), we can see that our fitted Gaussian Curve is slightly underfitting the tails of the distribution observed by our respective Kernel Density Estimate. That being said, in order to keep both our model from 2010-2015 and our model from 2000-2006 comparable, this slight bias in the Gaussian emission distributions is acceptable as long as we take into account the small biases it introduces into our interpretations of the conditional emission distributions. Thus, I will leave these estimates of conditional emission distributions as Gaussian.

Hence, our final selected model to fit for the 2000-2006 part of the dataset is a hidden markov model with two levels of hidden states and Gaussian conditional emission distributions.

6. Inference for the 2000-2006 Model

The log-likelihood of the current model fit to the 2000-2006 data is 4485. As discussed before, this log-likelihood is only slightly smaller than the log-likelihood of alternative models with more hidden state levels fit to the same dataset. Thus, this model is predicting the 2000-2006 about as on par as more complex HMMs.

	State 1	State 2
Probability of Starting in State	1	0

Table 3: Our Initial State Probability Distribution for the model fit for the 2000-2006 data.

Similar to our model fit on the data from 2010 to 2015, we see an initial state distribution that sets one of our states' probability to being near zero, although the actual fitted probability of initially starting in hidden state 2 is $2.4742612 \times 10^{-51}$. This model suggests that we are predicting on average that we will start a given time series generation process in hidden state 1.

	To State 1	To State 2
From State 1	0.94610	0.05389
From State 2	0.01719	0.98280

Table 4: Our estimated transition probability matrix for the model fit for the 2000-2006 data.

We see that by our transition probability matrix, when we are in a given hidden state at some point in the data-generating process, we are predicted to stay in that state with very high probability; thus, it is likely that the hidden states are predicting sustained long-run regimes in our data-generating process rather than short effects within our fitted data, as our model is predicting that we will not often have many consecutive switches between the hidden states in the time series generation process. However, in this instance, we see that the probability of staying in hidden state 2 given that our model is within hidden state 2 is estimated with near-to-one probability, which suggests hidden state 2 is approaching the behavior of an absorbing state. That being said, given that there is estimated to be some non-zero probability of switching from state 2 to state 1, we cannot claim that hidden state 2 is estimated to be exactly an absorbing state.

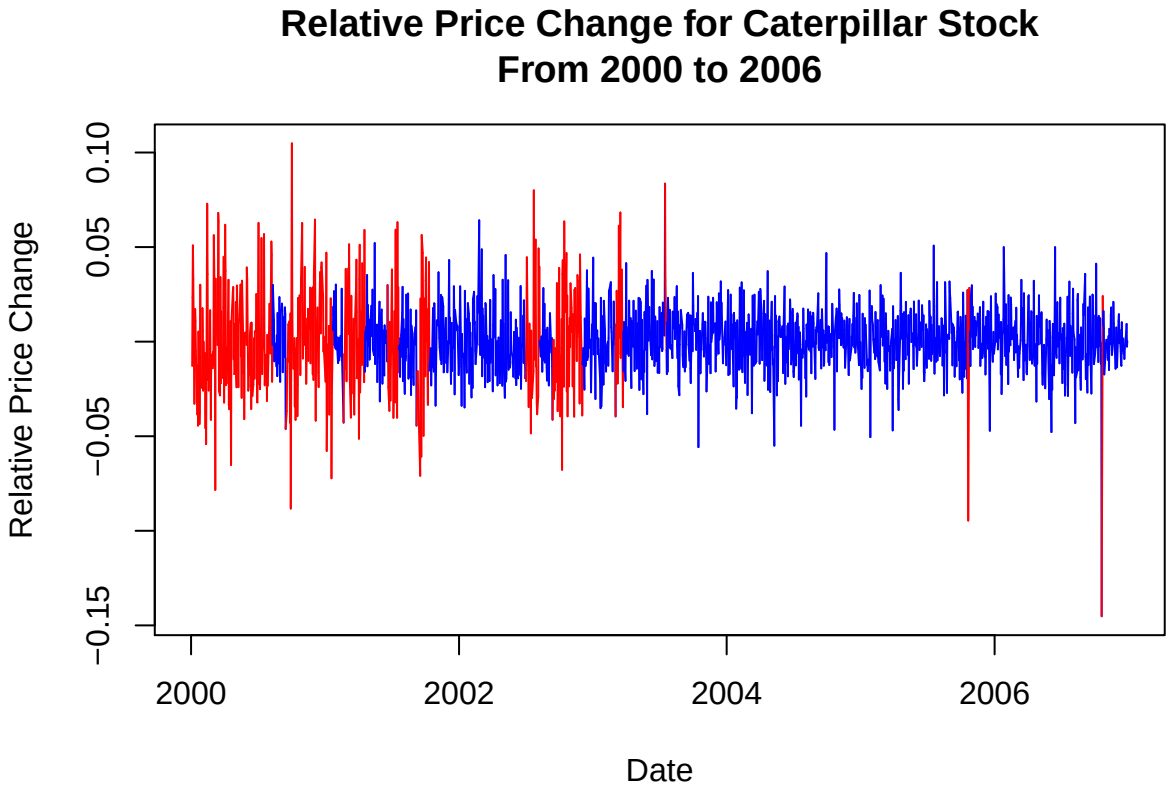

Figure 11: Plot of Relative Price Change for Caterpillar Stock With segments colored with the state the previous event was in. The hidden states were calculated using the Viterbi algorithm on our two state model. The first hidden state is colored in red, and the second hidden state is colored in blue.

Again, we see one state (hidden state 1, red) indicating instances of large relative price changes and one state (hidden state 2, blue) indicating instances of smaller relative price changes. Thus, we see our model fit our hidden states to predict regimes of high volatility within a stock price and low volatility within a stock price.

That being said, we see that we are estimated to more likely stay in the low-volatility hidden state in our model than we are estimated to stay in the high-volatility hidden state, which suggests that the time

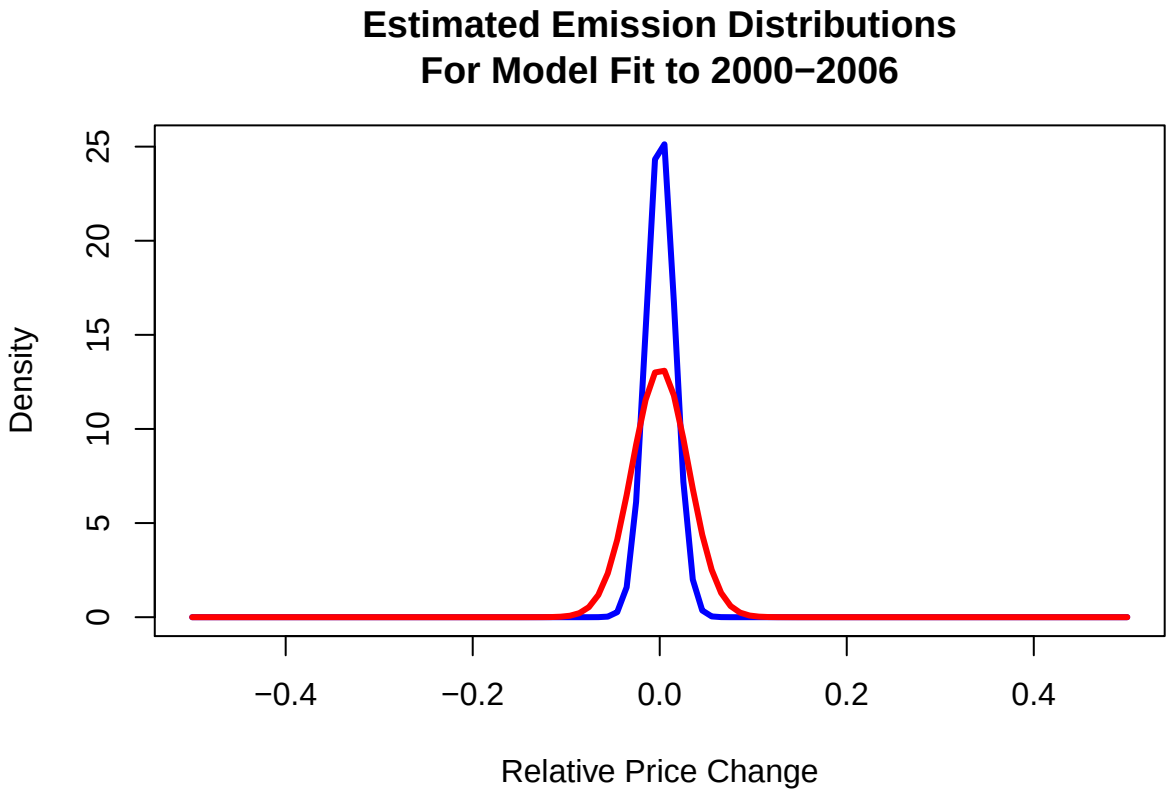

Figure 12: Graph of our Gaussian Emission Distributions Given Hidden States estimated via the Viterbi algorithm for the 2000-2006 data, Where the red line represents the emission distribution given hidden state 1 and the blue line represents the emission distribution given hidden state 2.

As we saw before, the estimated conditional emission distributions feature a similar mean (close to 0), but they feature very different variance. Notably, in this case our conditional emission distribution given hidden state 1 (red) has a much large estimated standard error than the conditional emission distribution given hidden state 2 (blue). This suggests that while both hidden states are expected to on average generate relative price changes around 0, hidden state 1 is predicted more often generate observations with larger absolute deviations from the mean than the observations generated by hidden state 2. This provides more evidence that we have one hidden state that is predicted to generate a regime of high volatility within the relative price of a stock and one hidden state that is predicted to generate a regime of low volatility within the relative price of a stock.

7. Comparisons between the 2000-2006 Model and the 2010-2015 Model

We see that there are many similarities between the two models we fit. Both of the models selected have two fitted hidden states, both models have initial state distributions that are close to degenerate, and both models seem to be fitting one hidden state to represent a high-volatility regime in our observations and one hidden state to represent a low-volatility regime in our observations. However, we see that in the model fit in on the 2000-2006 data, we are slightly more likely stay in the low-volatility regime given that we are in the low-volatility regime than we are in the model fit on the 2010-2015 data, which suggests that the 2000-2006 data may represent a time frame of relatively sustained low volatility within the stock price when compared to the 2010-2015 data.

This suggests that post-financial crisis in 2008, caterpillar's stock was predicted to more likely transition to a high-volatility regime from a low-volatility regime than it was prior to 2008. Whether or not this greater volatility post-financial crisis was caused by the great disruptions within the construction market from the real estate bubble requires further causal analysis of our models with a general model for disruption within the real estate market.

8. Conclusion

We see that the models we fit for our two time frames were rather simple, since we found that models with 2 hidden states fit our data almost as well as models with more hidden states. We saw that both of our models fit a set of hidden states which predicted high-volatility and low-volatility regimes within changes in Caterpillar's stock price, although the model fit on the 2000-2006 data predicted a higher probability of sustaining a low-volatility regime (i.e. not transitioning from a low-volatility to high-volatility regime) than the model fit on the 2010-2015 data. This suggests that the data-generating process on the 2000-2006 data is predicted to represent a lower-volatility environment for the construction stock, which may suggest that post-financial crisis, Caterpillar's stock price is much more likely to face large disruptions within its valuation.

There are several lines of research that would be useful to study after this analysis. Due to the fact that factorial HMMs have the potential to represent long-run and short-run effects within a time series, it may be useful to compare the effectiveness of these models with our simpler models fit in our analysis. Given the fact that it would be interesting to see if the collapse of the real estate market caused greater volatility within the valuation of construction companies, it would also be useful to study this dataset using a causal analysis between the models we fit and covariates that measure disruption within the construction and real estate market.