Dynamically Selecting & Combining

Volatility Forecasts

by

Jason Weitze

An honors thesis submitted in partial fulfillment

of the requirements for the degree of

Bachelor of Science

Undergraduate College

Leonard N. Stern School of Business

New York University

May 2019

Professor Marti G. Subrahmanyam	Professor Rob F. Engle
Faculty Adviser	Thesis Adviser

Dynamically Selecting & Combining Volatility Forecasts*

Jason Weitze

May 2019

Abstract

It is often the case that there are number of different methods for forecasting any given economic time series. For instance, New York University's Volatility Institute has 13 forecasters, which each provide a different forecast of the next day's volatility. Having many choices is often great, but at the end of the day we need a single forecast to make decisions with, so which one should we use? In this paper, we explore several different methods for choosing from and/or combining a collection of forecasts. To do so, we consider unconditional and conditional Diebold-Mariano tests, model confidence sets, and discrete state Markov processes. At the end, we propose a method for modelling the daily best forecast amongst the Volatility Institute's volatility forecasts for the SP500. In particular, we consider the model confidence set of the collection of volatility forecasters, and we model the remaining forecasters dynamically as a discrete state Markov process.

1 Introduction

Generally, there are a number of different methods out there to forecast any given economic time series, whether it be GDP, inflation, volatility, etc. Having good predictions of these values is often of immense value, but if there are many

*Email: jason.weitze@stern.nyu.edu

different forecasting methods and thus many different forecasts available, which one should be used? Is there one forecaster which is the best for all time? Is there a set of forecasters which are each the best during separate periods of time? More generally, is there a systematic way in which I can take a collection of forecasts and arrive at a single forecast to use?

If you are an economic theorist, it may well be that you are in favor of a particular theory and thus would prefer to use a model, which was built on that theory. In this case the problem is solved, you know precisely which forecast you would like to use. For this reason, this paper may not be as relevant to you. Rather, this paper is for those who have been given a number of forecasts or have access to a number of different forecasts of the same economic time series but are left to do decide which forecast to use on their own.

In fact, this was a questions we first asked ourselves when looking at the collection of forecasts used by New York University's Volatility Institute, which maintains 13 different volatility forecasters. While dealing with 13 forecasters is already a lot, since ARCH models were first proposed in Engle (1982), the number of models in existence has continued to grow. For instance, in Lunde and Hansen (2005), the authors managed to compile a list of more than 300 ARCH-type models. While the Volatility Institute certainly doesn't have that many models, we still have no means of determining which forecast we should actually use. For the sake of simplicity, in this paper we focus specifically on the SP500, which has just 11 models that the Volatility Institute maintains.

Thus, in this paper we aim to explore and discuss several different methods for deciding which forecast to use from a collection of forecasts, and we use the 11 models maintained by the Volatility Institute as our case study.

Fortunately, there is already an enormous literature on model and forecast comparisons that we can lean on. In this paper, we pay particular attention to the work done in Diebold and Mariano (2002), which introduced the now famous Diebold-Mariano tests as well as to the work done in Hansen et al. (2003) and Hansen et al. (2011), which introduced the notion of a model confidence set. Nonetheless, there are many other methods that have been proposed for comparing models including Clark and McCracken (2001) and West (1996). Also, while we do not touch upon information theoretic model selection procedures in this paper, as detailed in P. Burnham and R. Anderson (2002), there are a number of such methods as well.

Since a number of methods already exist in the literature for comparing models and forecasts, we refrain from proposing an entirely new method for comparing forecasts and instead build off of pre-existing methods like Diebold-Mariano tests and model confidence sets. In this paper, we consider the scenario in which every day we are given a collection of forecasts for the coming day's volatility and we attempt to use these tools to determine which forecast or set of forecasts are the best. Thus, our contribution to this literature is in exploring the practical value of these tools in the context of a collection of volatility forecasters as well as in determining what to do when these methods tell us that several forecasters are the best.

In the end we propose a single, systematic method from which you can dynamically determine what your daily forecasts should be if you're given a large collection of forecasters. We will demonstrate in Section 3 that we can distill a large set of forecasters down to a small sample of forecasters using a model confidence set. Once we have a smaller set of forecasters to consider, we propose modeling the daily best forecaster as a discrete state Markov process.

In the remaining sections of this paper, we go into more details. Starting in Section 2, we discuss the various methodological choices we make in this paper. Subsequently, we discuss our empirical findings in Section 3 before presenting our conclusions in Section 4.

2 Methodology

In this section, we discuss some of our methodological choices. We start by discussing our notation in Section 2.1. In the following section (Section 2.2), we discuss the general theoretical scenario that we are considering. Afterwards, we discuss our choice of loss function, the QLIKE function, in Section 2.3. Then, in Section 2.4 we discuss the difference between forecasters and models, which is a somewhat arbitrary distinction that we make for the sake of clarity in this paper.

Further, while we use several widely used methods to compare our forecasters like Diebold-Mariano tests and Model Confidence Sets, we also use Markov chains to model the 'best' forecaster as a stochastic process. As this is a less than standard method, we describe it further in the following section (Section 2.5). Afterwards, in Section 2.6 we explain why we can not compare the different approaches we use in this paper, at least not from a statistical perspective.

2.1 Notation

For the sake of simplicity, we attempt to list out most of the notation used in this paper below.

- σ_t^2 is the volatility at time t.
- $\hat{\sigma}_t^2$ is a proxy for volatility at time t (We will use squared returns).
- h_t is a scalar volatility forecast and $h_{t,i}$ is the i^{th} forecast of a collection of forecasts.
- $\mathbb{H}_t$ is a set of volatility forecasts at time t (i.e. $\mathbb{H}_t := \{h_{t,1}, \ldots, h_{t,k}\}$).
- h_i denotes the sequence of volatility forecasts from the i^{th} volatility forecaster (i.e. $h_i := (h_{t,i}) = \{h_{1,i}, h_{2,i}, h_{3,i}, \ldots\}$)
- $\mathbb{H}$ is a set of volatility forecast sequences (i.e. $\mathbb{H} := \{h_1, \ldots, h_k\}$).
- $\mathbf{h}_t$ denotes a vector of the forecasts at time t (i.e. $\mathbf{h}_t := (h_{t,1}, \ldots, h_{t,k})^T$).
- L denotes a loss function (it takes $\hat{\sigma}_t^2$ and h_t as arguments). In general, this will refer to the QLIKE loss function, which we discuss in Section 2.3.
- $\mathcal{F}_t$ is the information set at time t.
- l_t is a loss at time t (i.e. $l_t := L(\hat{\sigma}_t^2, h_t)$) and $l_{t,i}$ is the loss of the i^{th} forecast at time t.
- $d_{t,ij}$ is a loss differential at time t (i.e. $d_{t,ij} := l_{t,i} - l_{t,j}$).
- $\mathcal{M}_{1-\alpha}$ denotes a $(1-\alpha)\%$ model confidence set.
- e_i denotes a vector of all zeroes except for a one in the i^{th} position (i.e. a one-hot vector).

2.2 The General Scenario

As is normally the case, we are hoping to find the best forecast h_t of our variable of interest: conditional variance (σ_t^2). In this particular instance, we unfortunately cannot observe conditional variance, so instead we must settle for the next best thing, a conditionally unbiased volatility proxy ($\hat{\sigma}_t^2$). While using a conditionally unbiased volatility proxy could be problematic, we will explain in Section 2.3 why, based on our specific choice of loss function, we are justified in our use of the proxy. So, for our purposes let us simply consider the overarching goal to be finding optimizing the following minimization problem:

$$\min_{h \in \mathbb{R}_{++}} \mathbb{E}[L(\hat{\sigma}_t^2, h) | \mathcal{F}_{t-1}]$$

However, in our scenario, we are not asked to find some general forecast h. Rather, we are provided with a family of forecasts $\mathbb{H}_t$ for each time t and we are asked to choose or create a forecast from $\mathbb{H}_t$. We start with the assumption that we have no information about the relative quality of the forecasts in $\mathbb{H}_t$, so we start off with a uninformative, uniform prior on each forecast being the best. The hope is that once we throw in some information like their past performance (i.e. historical values of $l_{t,i}$ for each forecaster), we can get a more informative distribution on the k forecasters. The expectation is that gathering more information about our k forecasters will enable us to decide, systematically, what our forecast should be.

In order to determine what our final forecast should be, we could use $\mathbb{H}_t$ in a number of different ways and we will consider several of them in this paper. The first way we will consider (in Section 3.1) is whether or not we can select a single, dominant forecast to use for all time. A little more formally, we are interested in seeing if there is an $h_i \in \mathbb{H}$ such that its sequence of forecasts ($h_{t,i}$) has the lowest expected loss for all time. Afterwards, we will instead consider (in Section 3.2) whether or not some subset of $\mathbb{H}$ are dominant forecasters. We will do so by creating a model confidence set out of $\mathbb{H}$. Using the model confidence set we will be able to use the other methods on a reduced set of forecasters. Alternatively, we could simply consider the average of the forecasts that end up in our model confidence sets.

Up until now we have not considered the fact that the best forecaster in $\mathbb{H}$ may vary. In the following section (Section 3.3), we change that by trying to see if we can pick the best forecast in $\mathbb{H}_t$ using our current information set $\mathcal{F}_{t-1}$ in the following section (Section 3.3). From a more formal perspective, this could be framed as the following minimization problem:

$$h_t^* := \arg\min_{h\in\mathbb{H}_t} \mathbb{E}[L(\hat{\sigma}_t^2, h)|\mathcal{F}_{t-1}]$$

In order to tackle this minimization problem empirically, we consider the value of a conditional Diebold Mariano test as well as the possibility of modeling the best forecast as a discrete state Markov process. Then, in the following section (Section 3.4), we consider the value of a time-varying function f_t that combines the forecasts $\mathbb{H}_t$ in a manner that may depend on the current information set $\mathcal{F}_{t-1}$. More formally:

$$h_t^* := \arg\min_{f_t:\mathbb{H}_t\to\mathbb{R}_{++}} \mathbb{E}[L(\hat{\sigma}_t^2, f_t(h_{t,1},\ldots,h_{t,k}))|\mathcal{F}_{t-1}]$$

To help constrain this optimization problem a little, we will focus our attention on functions f_t, which are linear in the forecasts. In particular, we consider the value of using the transition probabilities from our discrete state Markov process as our linear weights.

2.3 What Is QLIKE & Why Are We Using It?

To start, let us note that QLIKE is a loss function with the following functional form:

$$L(\hat{\sigma}_t^2, h_t) = \log(h_t) + \frac{\hat{\sigma}_t^2}{h_t}$$

We use this loss function as opposed to common alternatives like mean squared error (MSE) or mean absolute error (MAE) because it happens to have 2 desirable properties.

The first reason we use QLIKE is because it is a robust loss function as per the definition provided in Patton (2011). In brief, Patton (2011) considers a loss function to be robust if, when ranking the expected losses of 2 competing volatility forecasters, we get the same ranking whether we use the true conditional variance or a conditionally unbiased proxy like squared returns. This is an important definition as we are primarily interested in comparing competing forecasts against the conditional variance, but we do not have access to it as it is an unobservable, latent variable. For this reason, we are left to compare forecasts based on a conditionally unbiased volatility proxy, so we would hope that the rankings we derive using the proxy are consistent with the rankings we would see if we could observe the true conditional variance, which is precisely why we prefer robust loss functions.

In and of itself, this is not a sufficient rationale for why we specifically opt to use QLIKE, as Patton (2011) notes that there are many other robust loss functions. That said, for our purposes there is really only one other well-known robust loss function worth considering: MSE. So, we could choose to use MSE instead, but we prefer QLIKE because it is less sensitive to extreme events. This is due to the fact that MSE depends, predominately, on the magnitude of the forecast error $(\hat{\sigma}_t^2 - h_t)$ whereas QLIKE depends, predominately, on the standardized forecast error $(\hat{\sigma}_t^2/h_t)$. Thus, while MSE can be dominated by a handful of outliers, QLIKE gives more equitable weight to each forecast. However, it is worth noting that depending on the specific application it may be desirable to use MSE, in which case the methods outlined in this paper will still apply but the results may vary.

2.4 On Forecasters vs. Models

There is a very subtle distinction that needs to be made between models and forecasters, at least as far as this paper is concerned. In this paper, we consider a model to be a theory driven mechanism for producing forecasts. On the other hand, we consider a forecaster to be a black-box that simply spits out forecasts when asked, so for our purposes a forecaster is theory agnostic.

Fundamentally, there need not be any distinction between the two. We draw the line in the sand merely to avoid accidentally suggesting that we are making any claims on economic theory. Generally, whether or not we use a given model is dependent on whether or not we believe in the particular theory that underpins it, so determining that a given model produces better forecasts would suggest that the underlying theory is also better. However, we want to avoid making any claims about the underlying economic theory and instead content ourselves with finding the forecasters which produce the best forecasts.

We are careful to not compare models in part because we intend to use the Diebold Mariano test as proposed in Diebold and Mariano (2002), and as Diebold (2015) stresses, this test only requires us to make assumptions about the forecasts' pairwise loss differentials (d_{12t}). In our case, we only want to make assumptions about the forecasts' pairwise loss differentials, so our conclusions will only be able to make claims about competing forecasters and not about the underlying economic models.

2.5 Markov Chain Models

Later in this paper, in Section 3.3 and Section 3.4, we will explore the value in modeling our best forecaster in a more dynamic manner. Normally, when we are looking for the best forecaster, we are looking for which forecaster has the lowest expected loss over all time. However, it is often the case that our best forecaster is not the best forecaster all of the time, and thus the title of 'best' forecaster might change from day to day. To that end, we could model the best forecaster as a stochastic process over the k forecasters we have available to us.

Let us consider X_t to be our random variable. Then we can define X_t in the following manner:

$$X_t := \{i : i = \arg \min_{j \in \{1,\dots,k\}} L(\hat{\sigma}_t^2, h_{t,j})\}$$

Then we can attempt to model X_t as a discrete state Markov chain, with k states and a transition matrix P'. In this case, the transition matrix P' should have entries of the form:

$$P'_{ij} := \mathbb{P}(X_t = j | X_{t-1} = i).$$

Using our historical data to get empirical transition probabilities, we can generate an empirical version of P', which we will denote P.

Once we have an empirical transition matrix P, we can start using that information to determine what forecast we should use, but first we need to isolate the transition probabilities conditional on today's best model. So note that we can rewrite the i^{th} row of P as $e_i^T P$, which is a $1 \times k$ vector of the following probabilities:

$$e_i^T P := (\mathbb{P}(X_t = 1 | X_{t-1} = i), \ldots, \mathbb{P}(X_t = k | X_{t-1} = i))$$

For convenience, right now let us make $i := X_{t-1}$ so that $e_i^T P$ is the vector of transition probabilities for this next forecast because this notation is significantly more concise than $e_{X_{t-1}}^T P$, which we would otherwise have to write every time.

Now, we can use this vector to pick our forecast. If, for some reason, we want to select a single forecast from $\mathbb{H}_t$, we could simply select the forecast corresponding to the largest element of $e_i^T P$, so if the j^{th} element is maximal, then we would choose to use the forecast $h_{t,j}$. More formally, we would choose to use the forecast $h_{t,j}$ where we get j from the following maximization problem:

$$j := \arg\max_j e_i^T P e_j$$

Now we have a means of selecting a single forecast to use, but note that by selecting a single forecast we are throwing away a lot of information about all of the other forecasts, which, while less likely, could nonetheless be the best forecasts tomorrow. To that end, recall that when we first got our forecasts, we assumed that we had no prior information about them and chose to put an uninformative, uniform prior on them. Well we could now update that prior by using $e_i^T P$ as a, hopefully, more informative distribution over our 'best' forecasters. If we are using $e_i^T P$ as our distribution over which forecaster will be best tomorrow, we could use this to take the empirical expectation of the best forecast tomorrow. In this way, we are attempting to estimate the expected best forecast each day. To do so, we simply take the dot product of $e_i^T P$ with our vector of forecasts at time t given by $\bar{h}_t$. Following this logic, we would make the following forecasts:

$$h_t := e_i^T P \bar{h}_t$$

2.6 Comparing our Different Methods

When all is said and done, we would like to be able to compare the different methods we discuss in this paper. It would be great if we could end this paper by saying that you should simply use Diebold-Mariano tests or if we could say that Markov chain models gave us the best forecasts, but unfortunately we cannot. This is an unfortunate result of our methodology as well as our underlying data.

The methodology is a problem because the second we add one combined model into the final comparison we run into problems with our forecasts being correlated by construction. For this reason, our resultant loss sequences will, necessarily, be dependent, so even simple comparisons like a Diebold-Mariano test would be inadequate. Now we could normally address this problem with most other datasets by simply splitting a hold-out sample and evaluating the average losses of each method on a separate subset of the hold-out sample, but this too is infeasible. We cannot use this strategy because it requires our subsets to consist of i.i.d. samples, but because we are using an economic time series (the SP500), we are unfortunately dealing with a highly dependent set of data rather than the independent data we would need. While there may be some alternate strategy for dealing with these problems, for the reasons listed above, we avoid making the claim that any given method is preferred over any of the alternative methods suggested here.

While we opt to not take a stance one way or another, we attempt to show some of the pros and cons of each method talked about in this paper in the hopes that you can make an informed decision as to how you want to systematically choose which forecast to use. Further, even though every strategy discussed in this paper has its downsides, we submit that each method proposed in this paper is a perfectly valid strategy.

3 Empirical Results

Before jumping into the complexities of dealing with a multitude of forecasters, let us start by considering a smaller collection of forecasters for which we have access to a larger sample period of forecasts. In particular, we start by considering the simplified scenario in which we are given only 3 forecasters: GARCH, EGARCH and GJR-GARCH.[1] Each forecaster is fit to the SP500 where we use squared returns as proxy for true conditional variance.

To start, we will consider a very simple Diebold-Mariano test in Section 3.1 followed by Model Confidence sets in Section 3.2. Then we consider methods for dynamically selecting which forecaster to use each day in Section 3.3 before exploring how we might dynamically combine our forecasts each day. Once we have explored the efficacy of the various tools on these 3 forecasters, we will expand our analysis to 8 of the forecasters maintained by NYU's Volatility Institute in Section 3.5.[2]

3.1 Diebold Mariano Tests

First on our list is to consider traditional Diebold-Mariano tests, which were designed simply to test the null hypothesis two forecasts have the same expected loss sequence, where we use the QLIKE loss function:

$$l_{t,i} = L(h_t, \hat{\sigma}_t^2) = \log(h_t) + \frac{\hat{\sigma}_t^2}{h_t^2}$$

Thus, the Diebold-Mariano test is testing whether or not $\mathbb{E}[l_{t,i}] = \mathbb{E}[l_{t,j}]$ (or equivalently, whether or not $\mathbb{E}[d_{t,ij}] = 0$). For context, look at Figure 1 to see the loss sequence for GARCH model, the loss sequence for the EGARCH model, and their loss differential sequence from top-to-bottom.

However, we need to note that the Diebold-Mariano test is only applicable if certain assumptions hold, at least approximately. Each of the assumptions are made about the random variable $d_{t,ij}$. In particular, we need $d_{t,ij}$ to have a time-invariant mean, variance, and covariance (i.e. covariance stationary). Realistically, the loss differential sequence presented in Figure 1 does exhibit some amount of clustering and there appear to be periods of time in which the loss differential is skewed one way or the other, whereas we would like to see a more consistent skew

[1] We have forecasts from these three forecasters since 1986 whereas we only have forecasts since 1990 for the rest of our forecasters.

[2] If you are interested in learning more about the 8 forecasters we used in this analysis, we have provided more detail on the parametric forms of these models in the appendix (Section 5.1).

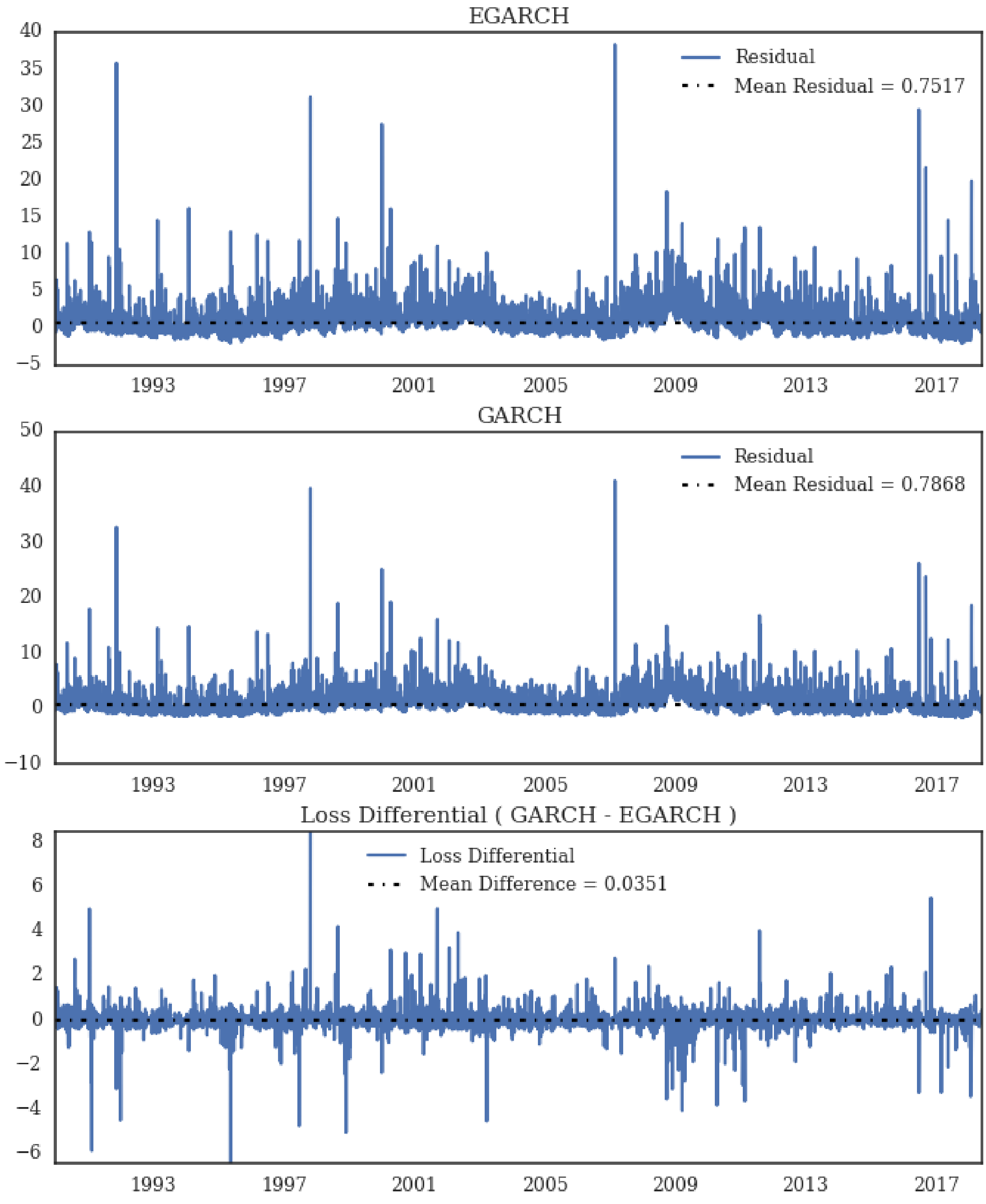

Figure 1: The top graph depicts the loss sequence for the GARCH forecaster. The middle graph depicts the loss sequence for the EGARCH forecaster. The bottom graph depicts the loss differential sequence for the two forecasters. In particular, the bottom graph depicts the loss differential for EGARCH−GARCH.

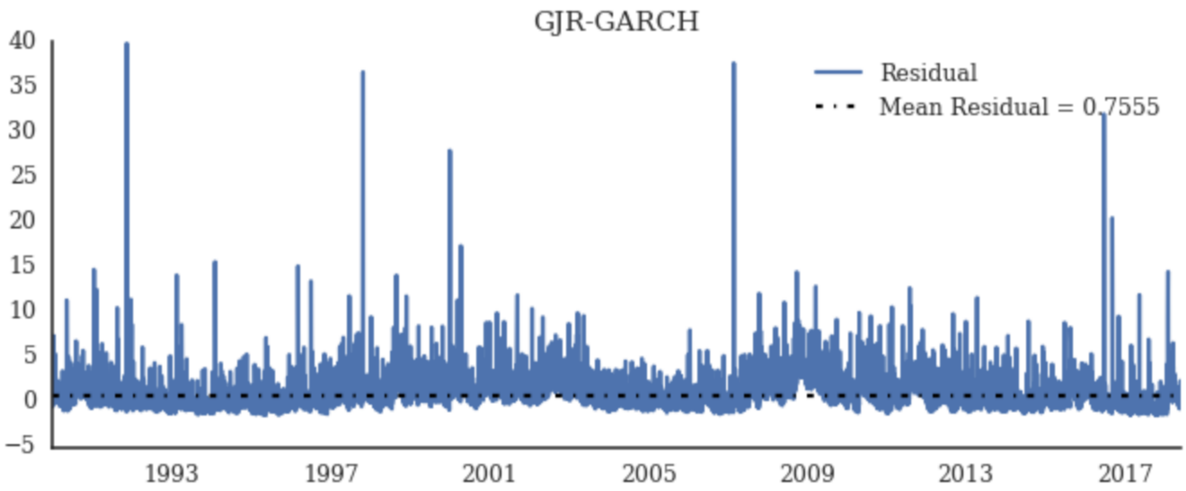

Figure 2: This graph depicts the loss sequence for the GJR-GARCH forecaster.

(if any) if the sequence was indeed stationary. In fact, we will attempt to leverage the fact that the unconditional loss differential is non-stationary in Section 2.5. For these reasons, it would be unreasonable for us to suppose that the assumptions do indeed hold. Nonetheless, let us assume that we our loss differential sequence is approximately stationary so that we can at least examine the results of the Diebold-Mariano tests. But, we need to remember to take any and all results from these tests with a grain of salt.[3]

Before looking at our tests' results, we should note that in our tests, we report several different p-values, which correspond to a different estimator of our loss differential sequences' standard errors. Of late there has been a lot of concern as to which standard errors are the most appropriate in a given scenario, so in this paper, rather than taking a stance one way or another, we consider several different estimators of the standard errors. The first p-value we include is the one corresponding to a conventional standard error. The second p-value corresponds to White standard errors, which were first proposed in White (1980). The third p-value corresponds to Newey-West standard errors, which were first proposed in Newey and West (1987).

Now we are ready to examine the pairwise Diebold-Mariano test results on our three forecasters:

[3] Keep these assumptions in mind for later, as we will need to make simple assumptions on the unconditional loss differential to compute our model confidence sets in Section 3.2.

	Coeffs	P-Vals	White P-Vals	Newey-West P-Vals
GJR-GARCH−GARCH	-0.0383	0.0***	0.0***	0.0***
GJR-GARCH−EGARCH	0.0023	0.5519	0.5518	0.5542
GARCH−EGARCH	0.0406	0.0***	0.0***	0.0***

These results seem to suggest that that the EGARCH and GJR-GARCH are both significantly better than the GARCH model. However, while the EGARCH loss is on average lower than the GJR-GARCH we can not, statistically speaking, differentiate the two forecasters' loss sequences.[4]

Seemingly, this makes the Diebold-Mariano test less than ideal for selecting a single forecaster. In this instance, we can't use the Diebold-Mariano test to argue that a single forecaster is dominant. So what can you conclude? Well, if we fail to show that two of our forecasters, in this case the EGARCH and GJR-GARCH, have distinct loss sequences, then we can conclude that their loss sequences are statistically non-differentiable. Thus, both the EGARCH and GJR-GARCH are effectively tied for the best forecaster.

While we have reached the conclusion of our Diebold-Mariano tests, we have yet to resolve our initial problem: what forecast should we use? If indeed the two forecasters are statistically indistinguishable, we could try a few different options in order to get a single forecast. To start, let us recall that when we first received our set of forecasters $\mathbb{H}_t$, we started by assuming that we have no bias towards one forecaster over another and thus had an uninformative, uniform prior over the various forecasters. We could emulate this logic, and since the two forecasts are statistically indistinguishable, we could assign equal probabilities to each forecaster and randomly select which forecast to use each day. This certainly works, but note that the expectation of the forecast this method would provide is simply the average of the 'best' forecasters. For this reason, instead of randomly picking which forecast to use, the more systematic approach may be as straightforward as averaging the

[4] These results indicate that, for instance, the GJR-GARCH forecaster is better than the GARCH forecaster, because we are modelling loss differential sequences of the form GJR-GARCH− GARCH. Thus, a positive coefficient suggests that GJR-GARCH>GARCH and a negative coefficient suggests that GJR-GARCH<GARCH. Since the coefficient for GJR-GARCH− GARCH has a negative coefficient, we conclude that GJR-GARCH's expected loss sequences is less than that of the GARCH forecaster. Similar logic applies for the following two comparisons.

'best' forecasters, which is precisely the methodology we end up employing.

At this point, we have a method for getting a single forecast, but we have yet to discuss a potential problem with this method. In particular, note that if we have k forecasters to compare, we are potentially running k choose 2 hypothesis tests, and while its not difficult to tell a computer to compute more test statistics, it does potentially kill our tests' significance. This is because the more tests that we run at a fixed significance level α, the more likely it is that at least one of the null hypotheses is incorrectly rejected. In this context, that means we are increasingly likely to mistakenly say that two of our forecasters have statistically distinguishable loss sequences when in fact they do not. Fortunately, problems like this have been well-studied, so we can opt to use a Bonferroni-style correction, akin to the ones proposed in Dunn (1958) and Dunn (1961).[5]

3.2 Model Confidence Sets

Just a few years ago, the idea of a model confidence set was proposed in Hansen et al. (2003) and Hansen et al. (2011), so they are still fairly new to the world of econometrics. Fortunately, they operate much like a confidence intervals making them, at least at a high level, quite accessible. A model confidence set, $\mathcal{M}_{1-\alpha}$, with a significance level α, is a subset of your models $\mathbb{H}$ that contains the best model at least $1 - \alpha\%$ of the time. Thus, if we choose a significance level $\alpha = 0.05$, a model confidence set $\mathcal{M}_{.95} \subset \mathbb{H}$ contains the best forecaster in $\mathbb{H}$ at least 95% of the time.

Before going any further, it is worth noting that the construction of model confidence sets relies on the assumption that $\mu_{ij} := \mathbb{E}[d_{t,ij}]$ is constant (i.e. time-invariant). As we noted in the last section (Section 3.1), this assumption is not valid for our loss differential sequences. However, just like we did in the last section, we can continue on with this analysis by taking our results with a grain of salt and by noting that the assumptions are approximately valid. In any case, we proceed to compute the model confidence set for our set of forecasters.

[5] Note that the Italian mathematician Carlo Emilio Bonferroni did not develop the Bonferroni correction himself.

In order to compute the model confidence sets, we use a method that is detailed in Hansen et al. (2011), but for the sake of clarity, we provide an outline of the algorithm in the appendix (Section 5.2). As part of the process, the algorithm produce p-values for each model, and the models' p-values are used to determine their inclusion in the model confidence set. In particular, if we consider a model confidence set with $\alpha = 0.05$, then a given model is included in $\mathcal{M}_{0.95}$ if it's p-value p_i is greater than or equal to $\alpha = 0.05$.

At this point, lets proceed to compute the model confidence set for our three forecasters:

Forecaster	p-value
GARCH	0.000
EGARCH	0.632
GJR-GARCH	1.000

If we use these p-values and an $\alpha = 0.05$ significance level, we would conclude that the EGARCH and GJR-GARCH are in the 95% model confidence set (i.e. EGARCH, GJR-GARCH $\in \mathcal{M}_{.95}$), but we would exclude the GARCH forecaster.

Now that we have a model confidence set, we have to figure out how to use it. The most basic solution would be to simply average the forecasters in your model confidence set much like we did with the statistically indistinguishable forecasters in our Diebold-Mariano tests. This is certainly a strategy one could employ, but one could also try using one of our other forecast selection strategies on the model confidence set. When you start off with a large number of forecasters, it can be overwhelming, difficult, and often impractical to try to use the other methods on the set of forecasters. For instance, if we have 100 forecasters, in order to use something like a conditional Diebold-Mariano test, which we consider in Section 3.3, you would have to consider 100 choose 2 or 4,950 hypothesis tests. However, if you instead only considered the few forecasters in the model confidence set you will likely have an order of magnitude fewer hypothesis tests to consider (e.g. if the model confidence set consists of 5 forecasters, you would only have 5 choose 2 or 10 hypothesis tests to consider). Note that this also means that while running Diebold-Mariano tests on a large number of forecasters may necessitate the use of

Bonferroni-style corrections, the model confidence set is already designed to work with a multitude of forecasters, so it does not need any corrections. For all of these reasons, model confidence sets are potentially invaluable in dwindling down a large set of forecasters, and we use it for this purpose in Section 3.3.

While this is certainly helpful, the model confidence set is not without its own drawbacks. The main concern is that the model confidence set is looking to include the best forecaster over all time. Thus, if there is a particular forecaster that is rarely the best forecaster, but during certain time periods is regularly the best, we will be less likely to include it in our model confidence set. This only becomes a problem if we are interested in using one of our methods to dynamically select the best forecast because if we use them on the model confidence set we will have prematurely excluded a forecaster which is best during a particular period of time. For instance, it turns out that while the GARCH forecaster is not in our model confidence set, it is actually the best forecaster about a third of the time, so if we tried to predict the best forecaster using the model confidence set, we would necessarily be missing the best forecaster at least a third of the time.

Nonetheless, if you have 100 forecasters in $\mathbb{H}$, modeling their transition probabilities using a Markov chain (as we do in Section 3.4) will be quite difficult whereas modeling the transition probabilities of the forecasters in the model confidence set may be significantly easier. Modeling 100 forecasters with a Markov chain would be impractical because it would require that we estimate 10,000 transition probabilities, so even if we have 10,000 data points our estimates for the transition probabilities will be extremely noisy. However, if you only have 5 forecasters after reducing your sample to the model confidence set, then you will only need to estimate 25 transition probabilities, which will lead to a significantly more reasonable model.

Accordingly, the choice to use a model confidence set to reduce the size of your $\mathbb{H}$ depends on a number of factors. If you simply have too many forecasters in your $\mathbb{H}$, you may need to reduce its size using a model confidence set to make some of the other methods we suggest a little more practical. However, if your $\mathbb{H}$ is already reasonably small, you may prefer to not use a model confidence set in order to avoid prematurely losing potentially good forecasters. It may also depend on the method you are trying to use. If you are trying to dynamically find the best forecaster, you may be less inclined to use a model confidence set on a small set of forecasters as you will inevitably be losing valuable information about forecasters which are occasionally the best, and thus occasionally the forecast you would like to use. However, if you are simply trying to find a single forecast or set of forecasters to average in the same manner over all time you may be more inclined towards a model confidence set even if your set of forecasters is already small. This is because you are not interested in a forecaster which is best merely for a small period of time, rather you want to know which forecaster is the best over all time, so losing some forecasters which are on average not the best is a potentially good thing.

In an effort to depict both, a use-case for the model confidence set $\mathcal{M}_{.95}$ and a use-case for using the entire set of forecasts $\mathbb{H}$, we use the model confidence set in Section 3.3 and we use the entire set of forecasts in 3.4.

3.3 Conditional Diebold Mariano Tests

At this point it is time to start considering whether or not we can use the current information set $\mathcal{F}_{t-1}$ to dynamically select which forecast $h_{t,i}$ we should use from the set $\mathbb{H}_t$. To do so, we consider a conditional Diebold-Mariano tests. A conditional Diebold-Mariano test is one which tests the expected loss differential between two series conditional on some other variable. In particular, we consider the expected loss differential series conditional on the four quarters of the year. Running this test on the loss differential series for EGARCH−GJR-GARCH (i.e. the forecasters from our model confidence set in Section 3.2), gives us the following coefficients and p-values:

	Coefficients	P-Values	White P-Values	Newey-West P-Values
Q1	0.0103	0.1839	0.2733	0.2881
Q2	-0.0017	0.8249	0.7966	0.7937
Q3	-0.0013	0.8719	0.8729	0.8748
Q4	-0.0168	0.0314**	0.0143**	0.0119**

Looking at these results, we see that no matter which standard errors you choose to use, the p-values for the Q4 dummy variable are significant at the 5% level. At the same time, none of the other coefficients are significant, even at a 10% level, which seems to suggest that the two forecasters' loss sequences are not statistically differentiable for the first 3 quarters of the year. Thus, while the EGARCH and GJR-GARCH are unconditionally statistically indistinguishable (as per Section 3.1), it turns out that in the fourth quarter the EGARCH forecasts are significantly better.[6]

Following our logic in Section 3.1, we can continue to use the forecasts' average during the quarters in which the two forecasters have statistically indistinguishable loss sequences. Whereas in the fourth quarter, when the EGARCH's loss sequence is significantly better, we can simply use the EGARCH forecaster.

3.4 Dynamic Forecast Combination: Markov Models

Now we consider the possibility of modeling the daily best forecaster dynamically using the Markov models we described in Section 2.5. But, before we go ahead and compute all of our transition probabilities, we take a moment to look at the data and confirm that a Markov model might indeed be useful.

To that end, we want to see if there is any regime structure in our stochastic process of best forecasters X_t that could be informative. The first thing we might consider is whether there is any structure in how long a given forecaster remains the best. For convenience, we refer to the period of time in which a single forecaster remains as the best forecaster as a regime. Note that if X_t evolves in an uninformative manner we would expect to see regimes that decay very quickly. In particular, we might expect to see the probability that a given regime is of length L to follow a geometric distribution with a parameter of $1/k$. In our case, we are only considering 3 forecasters, so if there is no structure we might expect that the regime lengths decay at a rate of 1/3. However, if we actually look at the data, we see that our regimes lengths decay at a rate closer to 1/2. While, the regime lengths do not quite fit a geometric distribution with parameter 1/2, they come reasonably close,

[6] We are modelling the loss differential sequence EGARCH$-$ GJR-GARCH, so a positive coefficient suggests that EGARCH¿GJR-GARCH and a negative coefficient suggests that EGARCH¡GJR-GARCH. Since Q4 has a negative coefficient, we conclude that EGARCH's expected loss sequences is less than that of the GJR-GARCH in Q4.

and a graph of the regime lengths (Figure 3) certainly looks to be approximately geometric: Thus there seems to be a non-trivial structure in our best forecaster

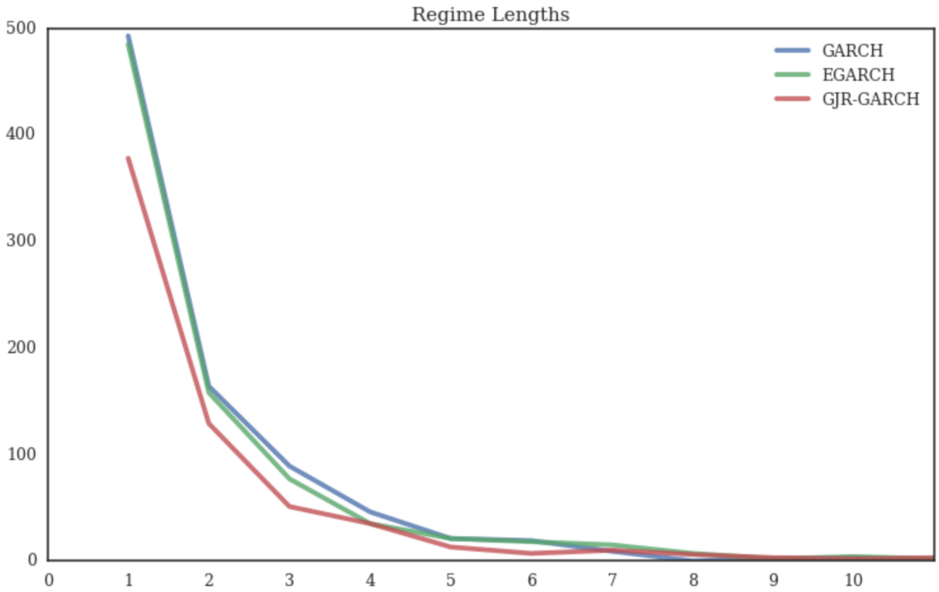

Figure 3: This graph depicts the regime lengths frequencies of our three forecasters. Recall that by regime length, we mean the number of days in a row in which a given forecaster provides the best forecast.

regimes, which means that the regime we are in is informative about what regime we will be in tomorrow. In particular, we should expect that the best forecaster today is the most likely to be the best forecaster tomorrow. Since, information about the current state of the world (i.e. the regime we are currently in), can help inform our guess as to the state of the world tomorrow (i.e. the regime we will be in tomorrow), it seems worthwhile to consider modeling this process as a discrete state Markov chain.

By now we have hopefully provided a compelling reason to model the daily best forecaster X_t as a Markov chain, so we should proceed to see what it looks like and what we can do with it. It turns out that if we model the best forecaster X_t as a Markov Chain we see the following transition probabilities (interestingly each forecaster has about a 50% chance of remaining the 'best' forecaster the following day):

	To	EGARCH	GARCH	GJR-GARCH
From				
EGARCH		0.49	0.31	0.20
GARCH		0.3	0.51	0.19
GJR-GARCH		0.24	0.26	0.50

Now that we have the transition probabilities, we can neatly neatly write them as the empirical transition matrix P for our Markov chain:

$$P = \begin{pmatrix} 0.49 & 0.31 & 0.20 \\ 0.30 & 0.51 & 0.19 \\ 0.24 & 0.26 & 0.50 \end{pmatrix}$$

Before going any further, we should note the non-trivial structure that is present in this Markov chain. In particular, note that the diagonal values (i.e. the likelihood of tomorrow's best forecaster being the same as today's) are fairly consistently about 1/2. If however, the Markov model was uninformative, we would expect that all of the values in this transition matrix would be approximately 1/3. This indicates that modeling these forecasters as a Markov chain is not merely a nice mathematical exercise, but rather that is actually capturing something in the day-to-day evolution of the best forecaster.

Given that the empirical transition probabilities are nontrivial, we can start using this empirical transition matrix P to determine what single forecaster we should use. Recall from Section 2.5 that if you wanted to choose a single forecaster to use tomorrow, we determined that you should do the following. First, suppose that i is yesterday's best forecaster (i.e. $i = X_{t-1}$), then we said that $e_i^T P$ was the vector of empirical probabilities for today's best forecaster X_t. Thus, if you wanted to pick a single best forecaster, we said choose the maximal element of $e_i^T P$ and call it the j^{th} element. Then we would predict that X_t is j and would thus forecast $h_{t,j}$. But note that in ever case, our matrix P of empirical transition probabilities suggests that the most likely best forecaster tomorrow is the best forecaster today. Thus, if $i = X_{t-1}$, then using this method, we would predict that $i = X_t$ and thus that we would forecast $h_{t,i}$.

Alternatively, we also suggested that the transition probabilities could be used as weights to linearly combine our forecasts. Once again suppose that i is yesterday's best forecaster (i.e. $i = X_{t-1}$). Then, recall that if we wanted to combine the forecasts for today we demonstrated that you simply need to compute $e_i^T P\mathbf{h}_t$ to get the linear combination of your forecasts. Note that this strategy is an essential part of the strategy we end up recommending in Section 3.5.

We can also go one step further by computing our Markov chain's steady state probabilities. Moreover, we can compute the empirical unconditional probability of any given forecaster being the best by measuring the proportion of days in which each forecaster was the best. Its a good sanity check, to compare the values of the steady probabilities with those of the empirical unconditional probabilities because if all went well, we should hopefully get the same percentages from each method. If we compute these values, we get the following probabilities, and fortunately, our two sets of probabilities seem to agree, so we have satisfied out sanity check:

	Steady State Probability	Empirical Unconditional
EGARCH	0.35	0.35
GARCH	0.37	0.37
GJR-GARCH	0.28	0.28

At this point, we have successfully constructed a Markov chain for our 3 forecasters to model the transition probabilities of the stochastic process X_t, and in theory we could model any k forecasters using the same methodology. However, this has a few potential downsides. First note that, right now we only have to estimate 9 transition probabilities, but if instead we had k forecasters, we would need to estimate k^2 transition probabilities instead. Logistically, this would not be a problem, but the more transition probabilities we need to estimate, the more noisy our estimates will be. This is because we are considering a fixed sample, so the relative frequency of each transition occurring in our sample would decrease. Thus, there is a modelling problem with using the Markov chain strategy to model k forecasters when k is large. The second problem that we should take note of is that by linearly weighting our k forecasters by transition probabilities, we are potentially diluting the quality of our resultant forecast if their are inferior forecasters. For these reasons, there is an incentive to weed out some of the inferior forecasters before modeling the transition probabilities of a subset of the forecasters with a

Markov chain. Not only will we have fewer transition probabilities to estimate if we consider a subset of the forecasters, but we will also have fewer inferior forecasters weighing down our final forecast.

3.5 Results From Using all Forecasters

Now that we have explored the value of several methods for producing daily volatility forecasts for the SP500 using a small set of only 3 forecaster, it is time to consider what would happen if we instead consider all 8 of our forecasters. Due to the simplicity and dynamism of the Markov chain methods proposed in Section 2.5, we want to use these methods to model our 8 forecasters. However, as pointed out in Section 3.4, considering too many forecasters at once may lead to inferior forecasts from our Markov chain model. Fortunately, we saw in Section 3.2 that computing a model confidence set can systematically provide us with a subset of forecasters. Not only does computing a model confidence set provide us with a subset of forecasters, but it is also systematically excluding the forecasters which are significantly worse than the true best forecaster.

For these reasons we propose the following systematic procedure for computing a single forecast from a collection of forecasters. First, we suggest computing a model confidence set of $k^* < k$ forecasters from your original set of k forecasters. Then, if k^* happens to be 1, we are done, and we simply use the sole forecaster in the model confidence set to produce our forecasts. On the other hand, if $k^* > 1$, we estimate the transition probabilities for the k^* forecasters in our model confidence set and model these forecasters as a Markov chain.[7]

Finally, we have a systematic process by which we can produce a single forecast from a collection of forecasters, so we can finally model our 8 volatility forecasters and see what happens. To start, lets compute the model confidence set:

[7] In the event that the model confidence set still contains too many forecasters (e.g. $k^* = 10$), simply continue to raise the significance level α of your model confidence set until you have a more suitable number of forecasters. The cutoff for what defines 'too-many' is very situation specific so we avoid defining a sharp threshold in this paper.

Model Name	P-Value	In $\mathcal{M}_{0.95}$?
GAS-GARCH-T	0.000	No
GARCH	0.000	No
S0GARCH	0.185	Yes
EGARCH	0.185	Yes
SGARCH	0.205	Yes
APARCH	0.205	Yes
GJR-GARCH	0.205	Yes
AGARCH	1.000	Yes

Thus, there are 6 forecasters remaining in the model confidence set. If we were that this might be too many forecasters to model as a Markov chain, we could go ahead and increase the significance level α of our model confidence sets to $\alpha = 0.2$, in which case we would reduce our model confidence set to only 4 forecasters. In our case, we have enough data that 6 forecasters should be fine, so we continue our process by computing the transition matrix for these 6 forecasters. Doing so gives us the following transition probabilities:

	To:	AGARCH	APARCH	EGARCH	GJR-GARCH	S0GARCH	SGARCH
From:							
AGARCH		.45	.07	.11	.10	.10	.17
APARCH		.11	.41	.05	.115	.135	.18
EGARCH		.128	.037	.428	.125	.115	.167
GJR-GARCH		.12	.1	.09	0.4	0.11	0.18
SOGARCH		.18	.12	.145	.145	.33	.08
SGARCH		.135	.11	.14	.14	.055	.42

Just as was the case when we modeled only 3 forecasters as a Markov chain in Section 3.4, we see that the Markov chain is highly non-trivial. In particular, we would expect every cell in the above table to be 1/6 if there was indeed no structure in the day-to-day evolution of the best forecaster. The easiest way to see this is by looking once again at the diagonals to note that they are each significantly larger than 1/6, which indicates that tomorrow's best forecaster is significantly more likely to be the same as today's best forecaster than random chance. Given that the transition matrix is once again non-trivial, we can go ahead and use it to provide linear weights on our forecasts conditional on yesterday's best forecaster.

Further, given that we have a Markov chain we can compute its steady state probabilities and compare them to the empirical unconditional probabilities of each of the forecasters as one final sanity check. We provide the appropriate probabilities in the following table, and fortunately, both sets of probabilities are identical.

	Steady State Probability	Empirical Unconditional Probability
AGARCH	0.196	0.196
APARCH	0.128	0.128
EGARCH	0.159	0.159
GJR-GARCH	0.171	0.171
S0GARCH	0.130	0.130
SGARCH	0.216	0.216

4 Conclusion

We started off by asking how we could come up with a single volatility forecast in a systematic manner if we were given a collection of volatility forecasts. In order to address this question, we considered several different methods for comparing models and/or forecasts from the literature (e.g. Diebold-Mariano tests and Model Confidence Sets), and tried them out on a sample of volatility forecasters produced by one of three forecasters. In doing so, we observed the pros and cons and of these various methods and ended up a proposing a single method to systematically reduce a collection of forecasts to a single forecast. We proposed that a collection of forecasters could be reduced to a smaller subset of forecasts by only considering those forecasters in a model confidence set. Once we determine which forecasters are in the model confidence set, we suggested that the daily best forecaster among these remaining forecasters be modeled as a discrete state Markov Process. Finally, once we have the transition matrix for the Markov chain, we recommended using the transition probabilities as linear weight for the daily forecasts to combine the volatility forecasts in a dynamic manner. In this way, we have demonstrated a method for systematically reducing a collection of volatility forecasts to get a single forecast, so that the end-user can look at a single forecast instead of being forced to choose from a collection of forecasts.

5 Appendix

5.1 Model List

In this paper we make reference to a number of different models for volatility, which we briefly explain below. In general, we assume that the our underlying asset, the SP500, has returns of the form $r_t = \mu + \varepsilon_t$. To start, we list our model's names as well as the form of their ε_t term:

Model Acronym	Model Name	Form of ε_t.[a]
AGARCH	Asymmetric GARCH	$\varepsilon_t = \sigma_t z_t$
APARCH	Asymmetric Power ARCH	$\varepsilon_t = \sigma_t z_t$
EGARCH	Exponential GARCH	$\varepsilon_t = \sigma_t z_t$
GARCH	GARCH	$\varepsilon_t = \sigma_t z_t$
GAS-GARCH-T	Generalized Autoregressive Score GARCH-T	$\varepsilon_t = \sigma_t \zeta_t$
GJR-GARCH	Glosten-Jagannathan-Runkle GARCH	$\varepsilon_t = \sigma_t z_t$
SGARCH	Spline GARCH	$\varepsilon_t = \sqrt{\sigma_t^2 \tau_t} z_t$
S0GARCH	Zero Slope Spline GARCH	$\varepsilon_t = \sqrt{\sigma_t^2 \tau_t} z_t$

[a] We let z_t be drawn from a standard normal distribution and we let ζ_t be drawn from a student's-t distribution with v degrees of freedom.

On the next page, we list our models' parametric forms.

Model Acronym	Parametric Form[a]	Parameters to Fit
AGARCH	$\sigma_t^2 = \omega + \alpha(\varepsilon_{t-1} - \gamma)^2 + \beta\sigma_{t-1}^2$	$\omega, \alpha, \beta, \gamma$
APARCH	$\sigma_t^\delta = \omega + \alpha(\lvert\varepsilon_{t-1}\rvert - \gamma\varepsilon_{t-1})^\delta + \beta\sigma_{t-1}^\delta$	$\omega, \alpha, \beta, \gamma, \delta$
EGARCH	$\ln(\sigma_t^2) = \omega + \alpha(\lvert z_{t-1}\rvert - \mathbb{E}\lvert z_{t-1}\rvert) + \gamma z_{t-1} + \beta\ln(\sigma_{t-1}^2)$	$\omega, \alpha\beta, \gamma$
GARCH	$\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2$	ω, α, β
GAS-GARCH-T	$\sigma_t^2 = \omega + \alpha\left(\frac{v+3}{v}\right)\left[\left(1 + \frac{\varepsilon_{t-1}^2}{(v-2)\sigma_{t-1}^2}\right)^{-1}\left(\frac{v+1}{v-2}\right)\varepsilon_{t-1}^2 - \sigma_{t-1}^2\right]$ $+\beta\sigma_{t-1}^2$	ω, α, β, v
GJR-GARCH	$\sigma_t^2 = \omega + (\alpha + \gamma\mathbf{1}_{\{r_{t-1} \geq \mu\}})\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2$	$\omega, \alpha, \beta, \gamma$
SGARCH	$\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2$ $\tau_t = \exp\left(\sum_{i=1}^k \phi_i(t - t_i)_+^2\right)$	$\omega, \alpha, \beta, \phi_1, \ldots, \phi_k, k$
S0GARCH	$\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2$ $\tau_t = \exp\left(\sum_{i=1}^k \phi_i(t - t_i)_+^2\right)$ Where we constrain ϕ_i to satisfy: $\sum_{i=1}^k \phi_i(T - t_i) = 0$	$\omega, \alpha, \beta, \phi_1, \ldots, \phi_k, k$

[a] The values for the $t_1 = 0, \ldots, t_k = T$ are chosen to partition the time stamps in our data set into $k - 1$ intervals of equal size.

5.2 Model Confidence Set Algorithm

Here we provide the outline of an algorithm for constructing a model confidence set using the T_{max} test statistic that was suggested in Hansen et al. (2011). We suppose there are N data points and k models that we are considering. Further, we consider B bootstrapped samples (in practice we let $B = 1000$) which are each of size s (in practice we let $s = \sqrt{N}$). Then the algorithm operates in the following manner:

1. Construct the B bootstrapped samples.

2. Start off by considering all k models and let $\mathcal{M} = \{1, \ldots, k\}$

3. Compute the average loss of each model: $\bar{l}_i = N^{-1}\sum_{t=1}^N l_{t,i}$

4. For each bootstrapped sample b, and for each model i, compute the average loss of each model (denoted $\bar{l}_{b,i}$)

5. For each bootstrapped sample b, and for each model i, compute a centered version of the bootstrapped losses given by $c_{b,i} := \bar{l}_{b,i} - \bar{l}_i$.

6. Initialize the dummy variable $p_{last} = 0$.

7. While $|\mathcal{M}| > 1$

   i Compute the average loss of the models in $\mathcal{M}$, which is given by: $\bar{l} = |\mathcal{M}|^{-1} \sum_{i \in \mathcal{M}} \bar{l}_i$

   ii For each bootstrapped sample b, compute the average centered loss of the models in $\mathcal{M}$, which is given by: $c_b = |\mathcal{M}|^{-1} \sum_{i \in \mathcal{M}} c_{b,i}$

   iii For each model i in $\mathcal{M}$, compute the average centered loss of it's bootstrapped losses, which is given by: $c_i = B^{-1} \sum_{b=1}^{B} c_{b,i}$

   iv For each model i in $\mathcal{M}$, compute the variance of the centered bootstrapped losses, which is given by: $v_j = B^{-1} \sum_{b=1}^{B} (c_{b,j} - c_j)^2$

   v Compute the average loss of each bootstrapped sample, which is given by: $\bar{l}_b = |\mathcal{M}|^{-1} \sum_{i \in \mathcal{M}} \bar{l}_{b,i}$

   vi Compute the T_{max} test statistic given by $T_{max} = \max_{j \in \mathcal{M}} \frac{\bar{l}_j}{v_j}$

   vii For each bootstrapped sample b, compute the bootstrapped test statistic given by $T_{b,max} = max_{j \in \mathcal{M}} \frac{\bar{l}_{b,j} - \bar{l}_b}{v_j}$

   viii Determine which model j to eliminate from $\mathcal{M}$, by finding the j that gave you T_{max}. Thus, j is given by $j = \arg\max_{j \in \mathcal{M}} \frac{\bar{l}_j}{v_j}$ and so we can redefine cM as $\mathcal{M} := \mathcal{M} \setminus \{j\}$.

   ix Assign model j a p-value given by $\hat{p} = \max\{B^{-1} \sum_{b=1}^{B} \mathbf{1}_{T_{b,max} \geq T_{max}}, p_{last}\}$

   x Reset p_{last} to be $p_{last} := \hat{p}$

8. Assign remaining model a p-value of 1.0

9. Models with $\hat{p} \geq \alpha$ get added to the $1 - \alpha\%$ model confidence set denoted by $\mathcal{M}_{1-\alpha}$.

References

Clark, T. and McCracken, M. (2001). Tests of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics*, 105(1):85–110.

Diebold, F. (2015). Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of diebold-mariano tests. *Journal of Business Economic Statistics*, 33(1):1–1.

Diebold, F. and Mariano, R. (2002). Comparing predictive accuracy. *Journal of Business Economic Statistics*, 20(1):134–44.

Dunn, O. J. (1958). Estimation of the means of dependent variables. *The Annals of Mathematical Statistics*, 29(4):1095–1111.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64.

Engle, R. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50(4):987–1007.

Hansen, P., Lunde, A., and Nason, J. (2003). Choosing the best volatility models: The model confidence set approach. *Oxford Bulletin of Economics and Statistics*, 65(s1):839–861.

Hansen, P., Lunde, A., and Nason, J. (2011). The model confidence set. *Econometrica*, 79(2):453–497.

Lunde, A. and Hansen, P. (2005). A forecast comparison of volatility models: does anything beat a garch(1,1)? *Journal of Applied Econometrics*, 20(7):873–889.

Newey, W. and West, K. (1987). A simple, positive semi-definite, heteroscedasticity and autocorrelation consistent covariance matrix. *Applied Econometrics*, 55(3):703–708.

P. Burnham, K. and R. Anderson, D. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, volume 67.

Patton, A. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, 160(1):246–256.

West, K. (1996). Asymptotic inference about predictive ability. *Econometrica*, 64(5):1067–84.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–38.