Multi-Horizon Forecast Optimality Based on Related Forecast Errors

Christopher G. MacGibbon

Professor Andrew Patton, Faculty Advisor

Professor Grace Kim, Faculty Advisor

Professor Kent Kimbrough, Faculty Advisor

Honors Thesis submitted in partial fulfillment of the requirements for Graduation with Distinction in Economics in Trinity College of Duke University.

Duke University

Durham, North Carolina

2018

Acknowledgements

I would like to thank the three faculty advisors that helped me get to the point of a completed thesis. The advice given to me by Professor Kim and Professor Kimbrough helped keep me on track and organize my paper and points to make them as clear and effective as possible. I am particularly appreciative of the guidance given by Professor Patton. Professor Patton was extremely helpful in all aspects of this process and I am especially grateful of the technical support that he provided.

Abstract

This thesis develops a new Multi-Horizon Moment Conditions test for evaluating multi-horizon forecast optimality. The test is based on the variances, covariances and autocovariances of optimal forecast errors that should have a non-zero relationship for multi-horizon forecasts. A simulation study is conducted to determine the test's size and power properties. Also, the effects of combining the Multi-Horizon Moment Conditions test and the well-known Mincer-Zarnowitz and zero autocorrelation tests into one forecast optimality test are examined. Lastly, an empirical study evaluating forecast optimality for four multi-horizon forecasts made by the Survey of Professional Forecasters is included.

JEL classification: G1; G17; G00

Keywords: Forecast optimality; Forecast errors; Multi-horizon forecast; Squared error loss; Simulation study; Combined test; Survey of Professional Forecasters

1. Introduction

Financial and economic forecasting is an extremely important field in the world today, as it plays a vital role in both policy making and financial planning for individuals, corporations and governments. For example, the Federal Reserve's forecasts for variables such as GDP, CPI and inflation not only give the economic outlook for the United States, but help the members of this organization make policy decisions to try and keep the economy strong and stable. Along the same lines, forecasts of company's earnings carry a lot of weight, as they are important in both guiding management and giving individual investors an indication of the strength of a business they may invest in. Having shown that these forecasts play an important role in society today, it should now be evident that it is crucial to be able to assess whether these forecasts are optimal and doing a good job. For this reason, the field of financial and economic forecast evaluation and the development of tests looking for forecast optimality is essential. It would be problematic to do something, such as set the United States economic policy, based on faulty forecasts. By being able to accurately forecast certain financial metrics, individuals, firms and governments will be able to guide their financial and economic choices to set them up best for success in the future.

There is lot of research that has been done in the field of financial forecast evaluation that focuses on testing for forecast optimality and rationality. This will be discussed in much more detail below, but, at a general level, an optimal forecast is one that minimizes a forecaster's expected loss for the variable that they are forecasting. One of the most widely used theories in this discipline dates all the way back to 1987 when William Nordhaus (1987) introduced the idea of "weak efficiency." Nordhaus (1987) suggested that optimal forecast errors should have no correlation to past forecast errors or forecast information that was available on or before the date of the forecast. In the conclusion of his paper, Nordhaus (1987) writes, "A baboon could

generate a series of weakly efficient forecasts by simply wiring himself to a random-number generator…Hence we should look at weak efficiency as but one attribute of well-constructed forecasts." This quote is extremely meaningful and demonstrates why literature has continued to come through the pipeline regarding testing for forecast rationality. Economists continue to develop conditions that should exist for ideal forecasts and create new tests to complement existing ones. The idea is that an ideal forecast should be able to pass all the different tests looking at forecast optimality and not just one.

Although the field of financial forecast evaluation has been around for a while, it continues to grow and develop. More recently, there has been a strong focus in evaluating multi-horizon forecasts. In a paper by Carlos Capistran (2014) focused on developing a test to evaluate financial forecast optimality, he mentions the importance of multi-horizon forecasts. He goes as far as to say that forecasts for just one horizon are "…in sharp contrast with the way forecasts are produced" (Capistran, 2014). With a multi-horizon forecast, the defining characteristic is that forecasters make predictions at multiple time horizons (represented by h), as opposed to just one. An example of a multi-horizon forecast is a forecast for GDP that is made 1, 2, 3 and 4 quarters before its value is realized. If the date that the actual GDP value is known is 2018 Q1, the 1 quarter out forecast would be made 2017 Q4 ($h = 1$), the 2 quarter out forecast would be made 2017 Q3 ($h = 2$), the 3 quarter out forecast would be made 2017 Q2 ($h = 3$) and the 4 quarter out forecast would be made 2017 Q1 ($h = 4$). Along with this example, Figure 1 below should help give a picture of the different forecasts included within a multi-horizon forecast when the maximum forecast horizon (represented by h_{max}) is 2, 3 and 4. It is important to note that the date that the forecasted variable is realized remains fixed and the horizon defines the date when the forecast was made.

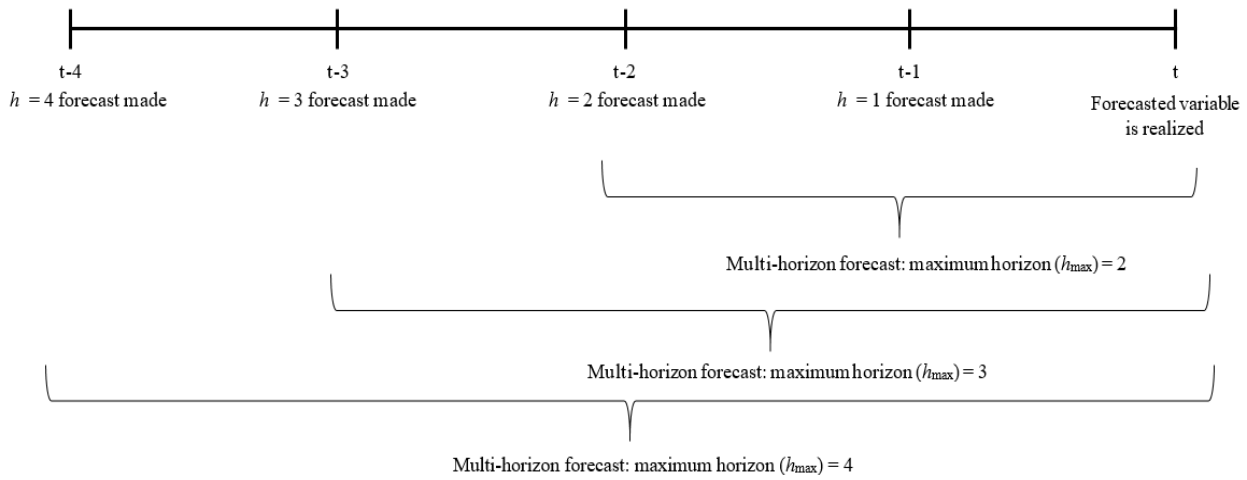


FIG. 1.–Image showing three different multi-horizon forecasts and the horizons they include

Since my thesis focuses on evaluating multi-horizon forecasts, it is not only critical to understand what they are, but also the notation associated with them. A target variable's realized value at time *t*+*h* is represented by Y_{t+h} and the forecast made at time *t* for a value realized at time *t*+*h* is written as $\hat{Y}_{t+h|t}$. The forecast error is then denoted as $e_{t+h|t}$ and defined by the following expression:

$$e_{t+h|t} = Y_{t+h} - \hat{Y}_{t+h|t}. \quad (1)$$

It will be important to remember this notation, particularly during the more technical sections of this piece.

Now that the idea of multi-horizon forecasts should be clear, it is appropriate to talk about how they can be evaluated. The concept of "weak efficiency" described above has been implemented in a regression that is well-known and frequently used to test for the optimality of these types of forecasts. In particular, for a given forecast horizon *h* at time *t* it is common to run a regression of that forecast error on the forecast errors for the forecasts made for time *t* or earlier

to test for forecast optimality. In this regression, the dependent variable is the forecast error for the forecast made at time *t*, *h* periods in the future. On the right-hand side of the forecast is an intercept and several explanatory variables. The explanatory variables are the forecasts made for time *t* or earlier. To give a clearer picture of this regression, it can be represented by the following expression (looking at lags between *h* and K):

$$e_{t+h|t} = \alpha_0 + \alpha_h e_{t|t-h} + \alpha_{h+1} e_{t-1|t-h-1} + \cdots + \alpha_{h+K} e_{t+h-K|t-K} + u_t. \quad (2)$$

The null hypothesis, which says that the forecast is optimal, tests whether $\alpha_i = 0$ for all $i = 0, h, h+1, \ldots h+K$. If there is a relationship between any forecast errors made at or before time *t* and the forecast error for a forecast made at time *t* then the expected value of the forecast error for the forecast made at time *t* is not zero, which is a critical condition of forecast optimality.

Notice in the regression just mentioned, that the forecast error at time *t+h* is regressed on the forecast errors of the forecasts for time *t* and earlier. This says nothing about the relationship between the forecast errors for the forecasts made for 1 and *h-1* periods in the future. The idea of "weak efficiency" implies that there should be no relationship between a forecast error and any variable available at the time the forecast was made. However, this gives no insight into the relationship between forecast errors that should have a non-zero relationship.

For example, imagine that forecasts for the stock price of a restaurant chain are made in 2017 Q2 ($h = 3$) and 2017 Q3 ($h = 2$) for 2018 Q1. Then, in 2017 Q4, a report emerges that numerous people got a foodborne illness from eating at this restaurant, resulting in a large decline in its stock price. The stock price is now much lower than anticipated, impacting the forecast errors for both the $h = 2$ and $h = 3$ forecasts. The news was not available at the time of either forecast, which is why it will affect both forecast errors. This should be reflected in the

relationship between these forecast errors. The relationships between these multi-horizon forecast errors for an optimal forecast have not yet been formally detailed and explored.

In an attempt to add to the pipeline of literature and tests that exist to evaluate multi-horizon financial forecasts, I detail how these forecast errors should relate to each other under squared error loss. I derive what these relationships should be for an optimal forecast, particularly looking at variances, covariances and autocovariances, and then I implement them in a test that looks at whether these relationships hold for a given forecast. I have named this new multi-horizon forecast optimality test the Multi-Horizon Moment Conditions test. I find this name appropriate as the relationships that I derive for the optimal forecast errors are known as "moment conditions" and this optimality test is meant to examine multi-horizon forecasts.

The purpose of the Multi-Horizon Moment Conditions test is to serve as a complement to other existing tests that have already been designed to evaluate forecast optimality. There are many ways that a forecast can be irrational and this test is meant to be one additional checkpoint to make sure that a forecaster is not doing a suboptimal job. Even though different tests evaluate different properties of an optimal or rational forecast, they all have the same objective. Ultimately, a forecaster's goal is to be able to find and develop optimal forecasts for whatever variable is being looked at. These tests are a critical step in the process. If it is identified that a forecast is not optimal, the forecaster should go back and see what they can change in their forecast and how they can improve it to give the best predictions possible. Since these tests have the same goal and serve as complements to each other, that also opens the idea of combining the Multi-Horizon Moment Conditions test with several other existing tests. This should result in a more efficient way to test for forecast optimality and I examine the effectiveness of doing so in this thesis.

The following is the outline of this paper. Section 2 gives an overview of relevant literature, which should both motivate this paper and give background in the field of evaluating financial forecasts. In Section 3, I outline how I derive the properties that I use for the Multi-Horizon Moment Conditions optimality test and discuss how I use Generalized Method of Moments (GMM) to test for those properties. Section 4 provides a simulation study that gives insight into the quality of the Multi-Horizon Moment Conditions test, outlining its size and power properties. In Section 5, I discuss the effectiveness and results from combining several forecast optimality tests. Lastly, Section 6 shows several evaluations of real-world forecasts and Section 7 concludes the paper and provides possible extensions.

2.A. Literature Review

In this section, I review two related, but distinct, subject areas: developing tests to determine the rationality of a forecast and the importance of loss functions in financial forecast evaluation. The main goal of this Literature Review is to put into context where the Multi-Horizon Moment Conditions test fits into the existing field of financial forecast evaluation. I also hope to give background on loss functions and the role they play in testing for financial forecast optimality.

2.B. Development of Tests to Evaluate Forecast Rationality

I will start by detailing the findings and contributions of the relevant literature on developing tests for financial forecast evaluation. I will first discuss the findings of Mincer and Zarnowitz (1969). These authors performed an "Absolute Accuracy Analysis" of economic forecasts and they proposed the following regression to test for forecast optimality:

$$Y_{t+h} = \alpha_0 + \alpha_1 \hat{Y}_{t+h|t} + \hat{v}_t. \quad (3)$$

By testing the joint hypothesis of $\alpha_0 = 0$ and $\alpha_1 = 1$, the authors proposed that one can determine whether the forecasted values are close to the realized values. If the null is rejected, Mincer and Zarnowitz suggested that this test allows for an individual to check if the rejection of rationality is caused by a biased forecast ($\alpha_0 \neq 0$), inefficient forecast ($\alpha_1 \neq 1$) or both. This test is extremely informative, as it not only gives insight into the rationality of a forecast, but also provides intuition for potential reasons why a forecast might not be rational.

Next, I will once again mention the concept of "weak efficiency," which was developed by William Nordhaus (1987). The big idea behind "weak efficiency" is that a forecaster should minimize the expected loss function that is being used, given the forecast information available at the time of the forecast. With almost all optimality properties, this one can be represented in several ways, and it is equivalent to say that forecast errors should be completely uncorrelated to the forecast errors or revisions available at the time of the forecast (Nordhaus, 1987). Michael Clements (1997) presented a method to correct for the issues that Nordhaus's test can have when only a few forecasts are available. He proposed pooling together and simultaneously testing multiple forecast dates and testing for any correlation with past forecast errors. It is important to note that these tests focus on a forecast for a single horizon.

There is also literature focusing on forecast optimality for multi-horizon forecasts. In the paper by Carlos Capistran (2014) that is mentioned in the Introduction, he proposes a new test for multi-horizon forecast optimality based on the "decreasing precision" property. He says that the variance of forecast errors should be greater for longer horizon forecasts. The author mentions that this concept is not new, but that there is no test for it, and is able to develop one. Patton and Timmermann (2012) also noticed a gap in the literature surrounding multi-horizon forecast optimality. They put together general inequality tests for an optimal forecast, as well as

a regression, which were all based on the assumption of mean squared error loss. For example, one of their tests is based off the idea that the covariance of a short-horizon forecast and the realized value should be greater than the covariance of a long-horizon forecast and the realized value. From the work of economists like Capistran, Patton, Timmermann and others, this class of multi-horizon forecasts started to develop a more specific set of optimality tests looking at the internal consistency among the forecasts made at different horizons.

My research, as mentioned earlier, fits in the world of testing for multi-horizon forecast optimality. The work done by Capistran, Patton and Timmermann tests for general conditions that should hold true for multi-horizon forecasts. They look at greater than or less than relationships for different variances, covariances and expected values. To develop the Multi-Horizon Moment Conditions test, I look at exact relationships (meaning equalities instead of inequalities) between certain forecast errors. The relationships that I test are not necessarily a direct extension of the properties derived in these papers, but the literature serves as a great resource for deriving the exact relationships that I evaluate. In particular, I look at variances, covariances and autocovariances of optimal forecast errors. I focus on developing precise relationships that should hold true between certain forecast errors for an optimal multi-horizon forecast.

Additionally, with the current optimality tests looking at exact relationships between forecast errors, they only focus on the relationship of a given forecast error with those that are available on or before the time of the forecast. These conditions do not mention anything about the relationships between forecast errors that should hold for optimal forecasts with overlapping periods of time between when they were made and realized. By looking at the exact relationships

between optimal forecast errors in the time period that is currently being overlooked when testing for forecast optimality, I am able to fill a gap in the existing research in this field.

2.C. Importance of Loss Functions in Evaluating Financial Forecasts

Another important aspect of forecasting are the loss functions used when evaluating financial forecasts. An optimal forecast, denoted as $\hat{Y}^*_{t+h|t}$, is defined by the following equation:

$$\hat{Y}^*_{t+h|t} \equiv \arg \min_{\hat{y} \epsilon Y} E[L(Y_{t+h}, \hat{y})|I_t] \tag{4}$$

(Patton, 2013). In words, this says that the optimal forecast is the one that minimizes the expected loss function, conditional on the information set I_t. This demonstrates that the loss function plays a critical role in the definition of an optimal forecast. The loss function that the Multi-Horizon Moment Conditions (MHMC) test is based on is squared error loss, which is defined in the following way:

$$L(Y_{t+h}, \hat{y}) = (Y_{t+h} - \hat{y})^2 \tag{5}$$

(Patton, 2013). All the moments and conditions that I derive for forecast optimality assume squared error loss.

Having detailed the role of loss functions in defining an optimal forecast, I will now speak to how the chosen loss function can impact the conclusion that is made when determining the optimality of a forecast. In an article by Patton and Timmermann (2007), they mention several properties, such as the non-decreasing variance of the forecast errors with the forecast horizon and the lack of correlation between a forecast error and the forecast errors from on or before the date the forecast was made, that should hold under mean squared error loss. They then go on to say these properties are not too useful "because they do not generally hold under the

other loss functions" (Patton and Timmermann, 2007). Elliot, Kumunjer and Timmermann (2005) echo this idea in a separate piece of literature focused on loss functions. The three authors state in their introduction that although mean squared error loss is commonly used, it is "often difficult to justify on economic ground and is certainly not universally accepted" (Elliot, Kumunjer and Timmermann, 2005). The main takeaway from all of this is that a loss function is crucial to determining the properties of an optimal forecast, and that the conditions that exist for one loss function do not always hold true for other loss functions.

I will now detail the findings from relevant literature focused on asymmetric loss functions. Asymmetric loss functions are loss functions where positive and negative loss are penalized differently. One example where this seems to be the case is with the Federal Reserve's forecast for real GDP. It seems that the Fed penalizes overpredictions of real GDP much more than underpredictions. The intuition behind this is that if the Fed overpredicts real GDP and bases monetary policy on that forecast, then they will be falsely signaling growth in the economy and will not have monetary policy properly aligned to handle the actual GDP growth (Patton and Timmermann, 2007). Another example involves forecasts by the IMF and OECD on government budget deficits. Generally, these forecasts overpredict budget forecasts, as underpredictions are seen as more costly than overpredictions and are penalized harsher in the loss function being used by the forecasters. It is noted in the piece detailing this asymmetric loss that, in some countries, an underprediction is penalized three times more than overpredictions of this variable (Elliot, Kumunjer and Timmermann, 2005).

Additionally, another example of where asymmetric loss can be found is in analyst's forecasts of a company's earnings. One paper finds that, not only is the loss function for the forecast of this variable asymmetric, but that it changes depending on the forecast horizon. For

long horizon forecasts, it is found that analysts penalize underpredictions greater than overpredictions, and that the forecasted value is generally greater than the realized value. An intuitive explanation is that optimistic predictions keep a firm's leaders happy and that the loss from an underprediction, which could lead to a negative reputation for a forecaster, is greater than the loss from an overprediction (Christodoulakis, Stathopoulos and Tessaromatis, 2012). However, for short horizon forecasts, data shows that an analyst's loss function penalizes overpredictions more than underpredictions. The reasoning being that the cost of a negative earnings surprise for a company is worse than the cost of a positive earnings surprise (Christodoulakis, Stathopoulos and Tessaromatis, 2012).

Through these examples, it is shown that there are certainly situations where asymmetric loss functions are used by forecasters. Using conditions that are derived from a separate loss function to determine the optimality of these forecasts is not valid and could result in the rejection of forecast optimality, when the forecasts are actually optimal under the loss function they are using (Patton and Timmermann, 2007). This concept becomes extremely relevant in the Empirical Application section of this thesis. I must justify that the forecasts that I evaluate are created by forecasters with a symmetric loss function, since the MHMC test operates under the assumption of squared error loss.

3. Theoretical Framework

To understand the derivations of the optimal properties that characterize the Multi-Horizon Moment Conditions (MHMC) forecast optimality test, it is important to first understand what defines an optimal forecast and the theory on how it differs from one that is not optimal. As mentioned in the Introduction, an optimal forecast is one that minimizes a forecaster's expected loss function for the data that they are forecasting. Two important variables that are used when

outlining the properties that should hold for an optimal forecast are the optimal forecast itself, as well as the optimal forecast error. The optimal forecast, written as $\hat{Y}^*_{t+h|t}$, was defined in Equation (4) in the Literature Review. This value can be used to get the optimal forecast error, $e^*_{t+h|t}$, by using the following equation:

$$e^*_{t+h|t} = Y_{t+h} - \hat{Y}^*_{t+h|t}. \tag{6}$$

This expression is the same as what is written in Equation (1) above, except the forecasted value is replaced by the optimal forecast.

For the MHMC test, I focus on forecast optimality under squared error loss. Under squared error loss, the following property can be derived:

$$E\left[e^*_{t+h|t}\middle|I_t\right] = 0. \tag{7}$$

This condition says that the expected value of the optimal forecast error for a forecast made at time *t* and realized at time *t+h*, which is represented by $e^*_{t+h|t}$, given the information set I_t, should equal zero. This property must hold for a forecast to be optimal and is used when deriving the relationships that should hold between optimal forecast errors.

I will only assume that the variables studied in this thesis are covariance stationary. A time series is said to be covariance stationary if its expected value, or mean, and variance do not change with time. Hamilton (1994) states that any covariance stationary process can be written as:

$$Y_t = \varepsilon_t + \sum_{j=1}^{\infty} \theta_j \varepsilon_{t-j},\ \varepsilon_t \sim WN(0, \sigma^2). \tag{8}$$

This representation, which is referred to as the "Wold Decomposition" involves an infinite number of parameters, the θ_j values, but I show below that the optimal forecast errors will only depend on a small number of these. As I only have data on the realized values and the forecasts, I do not know the parameters of the Wold Decomposition. However, I demonstrate that I can obtain moment conditions that allow me to estimate these parameters and, more importantly, to test the optimality of the forecasts.

Using the Wold Decomposition, the optimal forecast and optimal forecast error can be derived for each forecast horizon. Under squared error loss, the optimal forecast is the conditional expectation of the Wold Decomposition *h* periods in the future and the optimal forecast error is found by subtracting the optimal forecast from the actual value of the forecast at time *t*+*h*. For example, in the case where the forecast horizon is 1 (*h* = 1),

$$\hat{Y}^*_{t+1|t} = E_t[Y_{t+1}] = \sum_{j=1}^{\infty} \theta_j \varepsilon_{t+1-j}, \tag{9}$$

which means that

$$e^*_{t+1|t} = Y_{t+1-}\hat{Y}^*_{t+1|t} = \varepsilon_{t+1}. \tag{10}$$

By performing this derivation for enough forecast horizons, a general equation can be found for the optimal forecast error. For any forecast horizon *h* greater than or equal to 1,

$$e^*_{t+h|t} = \sum_{j=0}^{h-1} \theta_j \varepsilon_{t+h-j}. \tag{11}$$

The forecast optimality conditions that I derive focus on the variances, covariances and autocovariances of the optimal forecast errors. To give a better sense of what these properties

are, I will walk through how I derive the moment conditions that should hold for an optimal multi-horizon forecast when $h_{max} = 2$. To do this, I start by finding the optimal forecast errors for both $h = 1$ and $h = 2$. In the case where $h = 1$, the optimal forecast error is

$$e^*_{t+1|t} = \varepsilon_{t+1}, \tag{12}$$

and in the case where $h = 2$, the optimal forecast error is

$$e^*_{t+2|t} = \varepsilon_{t+2} + \theta_1 \varepsilon_{t+1}. \tag{13}$$

After deriving these optimal forecast errors, I then look at several variances and covariances that focus on these expressions and determine the values that they should equal if the forecast is truly optimal. The first conditions that I derive are the variance of each individual forecast error:

$$V\left[e^*_{t+1|t}\right] = V[\varepsilon_{t+1}] = \sigma^2 \tag{14}$$

and

$$V\left[e^*_{t+2|t}\right] = V[\varepsilon_{t+2} + \theta_1 \varepsilon_{t+1}] = (1 + \theta_1^2)\sigma^2. \tag{15}$$

These are the optimal variance values for these forecast errors. Next, I derive the first-order autocovariance, which is

$$Cov\left[e^*_{t+2|t}, e^*_{t+1|t-1}\right] = Cov[\varepsilon_{t+2} + \theta_1 \varepsilon_{t+1}, \varepsilon_{t+1} + \theta_1 \varepsilon_t] = \theta_1 \sigma^2. \tag{16}$$

Lastly, I look at the covariances between the different forecast horizons. In the $h = 2$ case, this is the covariances that exist between the $h = 1$ and $h = 2$ forecast errors, which are

$$Cov\left[e^*_{t+2|t+1}, e^*_{t+2|t}\right] = Cov[\varepsilon_{t+2}, \varepsilon_{t+2} + \theta_1\varepsilon_{t+1}] = \sigma^2 \quad (17)$$

and

$$Cov\left[e^*_{t+1|t}, e^*_{t+2|t}\right] = Cov[\varepsilon_{t+1}, \varepsilon_{t+2} + \theta_1\varepsilon_{t+1}] = \theta_1\sigma^2. \quad (18)$$

The derived variance and covariance conditions are a function of σ^2 and θ_1. These parameter values are unknown to whoever is evaluating the forecast and it will always be the case that there are more derived moments than parameters. In the case where $h_{max} = 2$, I derived five properties as a function of σ^2 and θ_1. Since there are more properties or moments than unknown parameters, a tool known as Generalized Method of Moments (GMM) can be used to estimate the parameter values (Hansen, 2017). The parameters estimated are the variables on the right-hand side of the equality in the moment conditions. Thus, the σ^2 and θ_1 terms are estimated using GMM. GMM also gives a value known as the J-statistic, which provides insight into how well the estimated parameter values fit the conditions that should hold for an ideal forecast. The J-statistic is the number that is compared to a critical value and used to reject or fail to reject forecast optimality based on how well the estimated parameter values satisfy the moment conditions that should hold for an optimal forecast.

Above, I derived the moment conditions in the case where $h_{max} = 2$. As part of this study, I also derive the variances, covariances and autocovariances when $h_{max} = 3$ and $h_{max} = 4$. These results are included in the Appendix. The derivations are done the same way, yet there are more variances, covariances and autocovariances that need to be derived as the maximum forecast horizon increases. When $h_{max} = 3$, 15 moment conditions exist and three parameter values must be estimated (σ^2, θ_1 and θ_2). When $h_{max} = 4$, 34 moment conditions exist and four parameter

values must be estimated (σ^2, θ_1, θ_2 and θ_3). I only derive properties for an optimal forecast up to $h_{\max} = 4$ in this thesis. Going to this horizon gives a good sense of the performance of this test for several horizons, and can also be used to evaluate the many real-world multi-horizon forecasts that are done 1, 2, 3 and 4 quarters out.

Although I derive 5 moment conditions when $h_{\max} = 2$, 15 moment conditions when $h_{\max} = 3$ and 34 moment conditions when $h_{\max} = 4$, I do not use all these moments in the GMM test that I implement for each maximum forecast horizon for the MHMC optimality test. All the variances, covariances and autocovariances that I derive for each maximum forecast horizon should hold for an optimal multi-horizon forecast under squared error loss. However, when trying to implement the GMM test in practice, I found that at least one or more moments are redundant for each maximum forecast horizon. This means that they are represented by some linear combination of other moments that should hold for an optimal forecast, and that the moment itself does not need to be included, or else it will result in an unreliable GMM test. By removing the redundant moment, the GMM test is still testing whether that conditions holds and is now reliable. I say that including the redundant moment makes the test unreliable as, when they are included, they blow up the J-statistic and, generally, result in a test statistic that always strongly rejects forecast optimality and does not appear to follow its asymptotic distribution in finite samples. Removing these redundant moments fixes this issue, which is why I do not include them in each GMM test.

As the maximum forecast horizon increases, I remove, at the very least, the same redundant moments that were removed in the GMM tests for forecasts with a shorter maximum forecast horizon. This means, for example, that the $h_{\max} = 4$ GMM test removes the redundant

moments from $h_{max} = 3$ test and any additional moments that were found to be linear combination of other moments for this test. When removing these redundant moment conditions, there are choices of what moment conditions to remove. If I found that moments 1, 2 and 3 were a linear combination of each other, for example, I could have removed moment 1, 2 or 3 and kept the other two. In making these choices, I always strived to remove the redundant moments that best fixed the issue of the J-statistic blowing up and consistently resulting in strong rejection of rationality in all cases. In the case where $h_{max} = 2$, I only use the 4 moments outlined in Equations (14-17) in the GMM test. When $h_{max} = 3$, I use 9 of 15 moments and when $h_{max} = 4$ I use 16 of 34 moments. In the Appendix, I identify the moment conditions that I use in the MHMC GMM test for these two maximum forecast horizons.

To see if the MHMC forecast optimality test that I implement using GMM can be used as a valid way to support or reject forecast optimality, I measure its size and power properties. The size of a test is the percentage of the time it rejects forecast optimality when the forecast is known to be optimal. When testing for size, the goal is to see if the asymptotic distribution of the test statistic is a good approximation to its distribution with a finite sample size. With a GMM test, the calculated J-statistic is χ^2 with the degrees of freedom for each test being equal to the number of moment conditions minus the number of parameters that are being solved for. If the size of a test with these type of distributions is close to the significance level, then it can be said that the asymptotic χ^2 results hold for the sample size being evaluated. If this result does not hold, then the test is not reliable to use at the sample size that the size is being calculated at.

Second, the power of this test is the percentage of the time it rejects forecast optimality when the forecast is known to not be optimal. For a good test, it is expected for the power to

continually increase and eventually reach an asymptote of 100 % as the forecast gets worse. When a forecast is slightly not optimal, a test will not always capture this fact and will not reject optimality for every single forecast. But, as a forecast gets worse, it is expected for the test to capture this a greater percentage of the time and eventually reach a point where it rejects optimality close to, if not exactly, 100 % of the time.

4.A. Simulation Study of Size and Power

To analyze the quality of the Multi-Horizon Moment Conditions (MHMC) forecast optimality test, I created ideal and non-ideal forecasts for a generated data set, and then used that information to determine both its size and power properties. To produce data, I used a first order autoregressive process defined by the following equation:

$$Y_t = \phi Y_{t-1} + u_t \sim WN(0, \sigma_u^2). \tag{19}$$

I chose to use this data generating process as it allows for me to see how the performance of the MHMC test changes as the autocorrelation, or phi (ϕ) value, of the time series changes. By varying this parameter, I gain interesting insight into how it impacts the size, power and ultimately the effectiveness of the MHMC test. I keep the σ_u^2 term constant and set it equal to 1 for each data set that I generate.

When determining the size of the MHMC test, I evaluate it for sample sizes of 50, 100, 250 and 500. Using this range of data points, I can see how the size varies as a function of the sample size. The size of the test should get closer and closer to the significance level as the sample size increases, as it should start to behave more and more like its asymptotic distribution. For this reason, I vary the sample size when evaluating this property to see if this expected behavior holds. When evaluating the power property of the MHMC test, each data set that I

generate has 150 data points. I chose this sample size because it is a value that resembles the sample size that I have when evaluating real-world forecasts, for variables such as the Consumer Price Index (CPI) Inflation Rate and Real GDP Growth, in the Empirical Application section. With power, I want to know the effectiveness of the test near the actual sample size of forecasts that I evaluate, which is why I only analyze this property at one sample size.

After I generated data using Equation (19), the next step was to develop an optimal forecast for the data generating process. In developing an ideal forecast, it is known that the forecast is rational and I used that information to find the size of the MHMC test. To develop an optimal forecast, I set the value of the forecast equal to the expected value of the data generating process conditional on time *t*. This means that for any value of *h* that the optimal forecast is the following:

$$\hat{Y}^{*}_{t|t-h} = E_t[Y_t] = \phi^h Y_{t-h}. \tag{20}$$

After generating the forecast, I input the forecasted and realized values into the MHMC test. With each test, I obtain a J-statistic from GMM, which I then use to determine whether or not the null hypothesis should be rejected, using a significance level of 5 %. By simulating this process 1,000 times and counting how many times I reject the null hypothesis, I obtain the finite-sample size of this test. I first examine the size of the MHMC test to make sure that it is behaving as expected. I then move on and examine the more interesting power properties, that demonstrate how effective the test is at rejecting irrational forecasts.

Next, I will discuss how I analyzed the power of the MHMC test for forecast optimality. To calculate the power of the test in finite samples, I created a non-ideal forecast for the data

generating process mentioned earlier by adding noise to the optimal forecast value. This process is represented by the following equation:

$$\tilde{Y}_{t|t-h} = \hat{Y}^{*}_{t|t-h} + \sigma_n n_t. \tag{21}$$

In this expression, $\tilde{Y}_{t|t-h}$ is the noisy or suboptimal forecast, $\hat{Y}^{*}_{t|t-h}$ is the optimal forecast used for calculating size in Equation (20), n_t is a generated value that is independent and identically distributed that is normal with a mean of 0 and variance of 1, and σ_n is what I refer to as the "noise multiplier." By varying the amount of noise added to the ideal forecast, which is the σ_n value in the equation above, I can see how this parameter can affect the power of the test. The larger the value of σ_n, the greater amount of noise that is added to the optimal forecast. By simulating this process 1,000 times and counting how many times I rejected the null hypothesis of optimality for the non-ideal forecast, I calculated the power by dividing the total number of rejections by the number of simulations.

I developed a MATLAB script that can find the size and power values for the MHMC optimality test in the cases where the maximum forecast horizon is 2, 3 and 4. Not only can this MATLAB script find the size and power properties for the MHMC test, but it also does this for the Mincer-Zarnowitz and zero autocorrelation forecast optimality tests. The reason I generate this information is so that I can compare the same properties of the MHMC test to established, well-known tests used in the field of evaluating financial forecasts. The following, as a reminder, is the Mincer-Zarnowitz regression used to test for forecast optimality:

$$Y_{t+h} = \alpha_0 + \alpha_1 \hat{Y}_{t+h|t} + \hat{v}_t. \tag{22}$$

This test regresses the realized value on the forecasted value and an intercept, with the null hypothesis of forecast optimality testing if $a = 0$ and $b = 1$ (Mincer and Zarnowitz, 1969). This test specifically looks at whether or not the realized value, A_t, is close to the forecasted value, P_t.

Another well-known way to evaluate forecast optimality is with the zero autocorrelation test, which was discussed in the Introduction. This test looks at whether the forecast errors available to the forecaster on their forecast date have any relationship to forecasts made on that date for some date in the future. The following is the regression used for this test when looking at four lagged forecast errors (for any forecast horizon, h):

$$e_{t+h|t} = \beta_0 + \beta_1 e_{t|t-h} + \beta_2 e_{t-1|t-h-1} + \beta_3 e_{t-2|t-h-2} + \beta_4 e_{t-3|t-h-3} + u_t. \quad (23)$$

This test regresses a forecast error on a certain number of forecast errors available to the forecaster at the time of the forecast and an intercept. The null hypothesis of forecast optimality tests if $\beta_0, \beta_1, \dots, \beta_4 = 0$, as there should be no correlation between these forecast errors for an optimal forecast (Nordhaus, 1987).

4.B. Size Discussion

Looking at Panel A in Tables 1, 2 and 3 at the end of this subsection, which are split up by the maximum forecast horizon for the generated multi-horizon forecasts, it is evident that the Multi-Horizon Moment Conditions (MHMC) test has size properties not too far from what is expected. The significance level used when testing was 5 %. If the size is equal or close to this value, then it can be said that the test appears to have good size properties. Although the MHMC test generally has size values close to what is expected, there is an issue for low phi, which is apparent in the column where phi = .1 in all three tables. In this case, particularly when there is a longer maximum forecast horizon, the MHMC test is severely oversized. This means that in the

case where a data set is not very persistent (has a low phi value), that the MHMC test would not be reliable.

I also want to discuss the size values for the Mincer-Zarnowitz and zero autocorrelation tests across these horizons, which can be found in Panel B and C respectively in the three tables below. This is important, as these are two well-known and widely-used forecast optimality tests. They can give a sense of what these values typically are for current forecast optimality tests and also show how well the MHMC test performs compared to them. To start, the zero autocorrelation test appears to be a bit undersized, particularly in the case where $h_{max} = 2$. This means that the test does not reject as many rational tests that would be expected. Next, the Mincer-Zarnowitz regression has great size properties. Across all horizons and all phi, the size values are extremely close to the significance level. Except for the case where phi = .1, the size values are similar for the MHMC test and these existing optimality tests. This is a great result for the MHMC forecast optimality test, as it means that, except for the case of very low phi, it has comparable finite-sample size properties to well-known existing tests.

TABLE 1:

Size values for the three tests when $h_{max} = 2$

Size Values: $h_{max} = 2$					
T	**ϕ**				
	0.1	**0.25**	**0.5**	**0.75**	**0.9**
Panel A: Multi-Horizon Moment Conditions					
50	10.8	6.4	3.2	3.8	4.1
100	9.2	4	3.1	5	4.4
250	5.8	2.7	2.5	3.6	4.4
500	2.7	2.8	2.7	4.7	4.3
Panel B: Mincer-Zarnowitz					
50	3.9	5.1	5.4	6.4	4.5
100	4.7	4.3	5.3	6.4	6.5
250	4.1	4.2	3.6	5.2	5.2
500	3.5	2.6	4	4.7	5.1
Panel C: Zero Autocorrelation					
50	2.5	3.8	4.6	2.9	4.5
100	3	4.9	3.1	3.8	2.4
250	3.8	3.3	4.8	4.2	4.5
500	3.3	1.4	2.7	3.2	3.4

TABLE 2:

Size values for the three tests when $h_{max} = 3$

Size Values: $h_{max} = 3$					
T	**ϕ**				
	0.1	**0.25**	**0.5**	**0.75**	**0.9**
Panel A: Multi-Horizon Moment Conditions					
50	22.5	12.6	5.1	3.9	5.1
100	19.6	8	4.3	5.1	5.3
250	13.7	3.2	3.4	5.9	6.5
500	9.8	2.6	3.5	5.2	6.9
Panel B: Mincer-Zarnowitz					
50	5.4	3.5	5.7	3.5	4.1
100	4.4	4.5	4.3	5.7	6.1
250	4.5	3.3	4.6	6.1	7.1
500	3.2	3.1	2.8	5.7	6.4
Panel C: Zero Autocorrelation					
50	2.9	2.1	2.8	2.9	3.5
100	5.1	3.7	3.6	3.1	4.9
250	4.7	4.4	3.9	4.1	3.9
500	3.7	3.7	3.7	4.2	4.6

TABLE 3:

Size values for the three tests when $h_{max} = 4$

Size Values: $h_{max} = 4$					
T	**ϕ**				
	0.1	**0.25**	**0.5**	**0.75**	**0.9**
Panel A: Multi-Horizon Moment Conditions					
50	38.5	7.6	6.1	4.2	2.5
100	39.7	9.9	4.8	4	3.7
250	39.5	8.5	3.8	4.4	4.2
500	38.7	9.8	4	4	3.1
Panel B: Mincer-Zarnowitz					
50	4.8	3.7	3.2	3.1	4
100	4.2	5.1	4.1	4.9	4.7
250	4.8	4.2	4.9	5.6	6.1
500	3.3	4.4	4	5.4	5.2
Panel C: Zero Autocorrelation					
50	2.5	2.2	2.8	2	2.7
100	4.3	4.2	4.4	4.4	3.1
250	5	4.3	4.5	5.1	3.6
500	4.1	3.5	4.2	3.9	3.5

4.C. Power Discussion

Having established that the Multi-Horizon Moment Conditions (MHMC) test generally has size values close to what is expected for a respectable multi-horizon forecast optimality test when $h_{max} = 2$, 3 and 4, it is appropriate to move on and analyze the power of this test. Based on the size tables from the last subsection, the first value where the MHMC test appears appropriate to use for all maximum forecast horizons is when phi = 0.25. Keeping this in mind, I chose to start my power analysis at this value. I do not analyze the power of the MHMC test for lower phi values since it could appear that the test is extremely powerful, however, that will not be due to the effectiveness of the test. The test statistic was shown to be oversized when phi = 0.1 and the

power from this test would be from the erratic behavior of the test statistic and not its effectiveness. Along with the case when phi = 0.25, I also perform the power analysis when phi = 0.5 and 0.75, to demonstrate how the power of the test changes relative to the Mincer-Zarnowitz and zero autocorrelation test as phi changes.

When plotting the power curves, which is done in Figures 2, 3 and 4 below, I have the power on the y-axis and the noise multiplier on the x-axis. As talked about earlier, the noise multiplier is the variable that controls how poor and noisy the suboptimal forecast is. The larger the value of σ_n in Equation (21), the larger the magnitude of the noise that is added to the optimal forecast and the worse the forecast is. On this graph, the y-axis is showing what percentage of suboptimal forecasts are rejected, for a given level of noise. The forecasts are getting noisier as the noise multiplier increases, which is why the power is expected to increase, eventually reaching a max value of 100 %. The plots also all include a reference line, when the power is equal to 5 %. When the noise multiplier is equal to zero, the power is the same as the size and the reference line can help show how close the size of each test is to the ideal value of 5 %. This line can also help show tests that lack power for this study and always have power of around 5 % for all noise multiplier values.

Now having established how to interpret Figures 2, 3 and 4 below, I would like to discuss the important findings from these plots. To begin, the zero autocorrelation test always has a power value of around 5 % and this value does not increase as the forecast gets worse or come close to 100 %. That means that this test cannot be used to reliably reject optimality for this type of noisy, suboptimal forecast. The property of zero autocorrelation between certain forecasts still holds for both the optimal and non-optimal forecasts in this case.

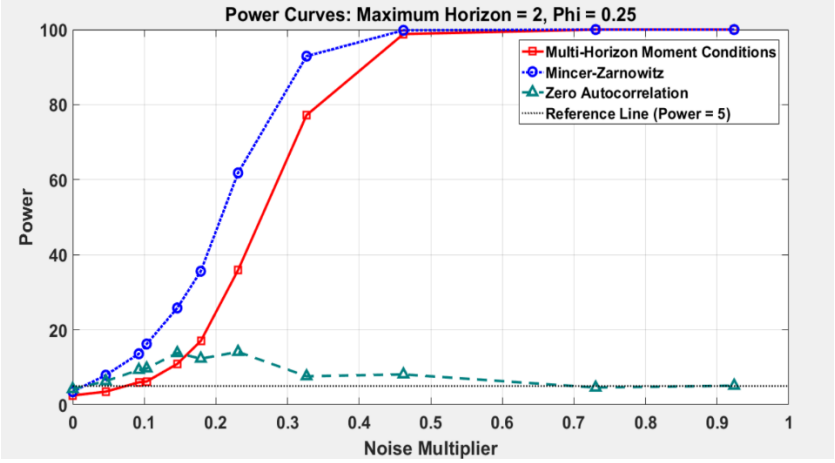
Next, the power curves for the MHMC test have the anticipated shape for all horizons and phi values that are tested. The Mincer-Zarnowitz test also has power curves that look as one would expect. Interestingly enough, the MHMC test has less power than the Mincer-Zarnowitz in certain cases, while in others it has more. Whether or not the MHMC test is more or less powerful is dependent upon the phi value and maximum forecast horizon.

The results across phi for the cases where $h_{max} = 2$ and $h_{max} = 3$ are extremely similar. For these maximum forecast horizons, when phi = 0.25, the Mincer-Zarnowitz's power curve is above that of the MHMC test, which indicates better performance. Next, when phi = 0.5, the Mincer-Zarnowitz's power curve is still above that of the MHMC test, but the two curves are closer together than when phi = 0.25. When phi = 0.75, the result changes and the MHMC test's power curve is above that of the Mincer-Zarnowitz. The conclusion that can be drawn from this is that the MHMC test becomes more powerful relative to the Mincer-Zarnowitz test as the value of phi increases for this type of suboptimal forecast. There is some value between phi = 0.5 and phi = 0.75 where these two tests should have around the same power. For all values of phi greater than this value, the MHMC test is more powerful, while for all values less than this value, the Mincer-Zarnowitz test is more powerful for this type of suboptimal forecast.

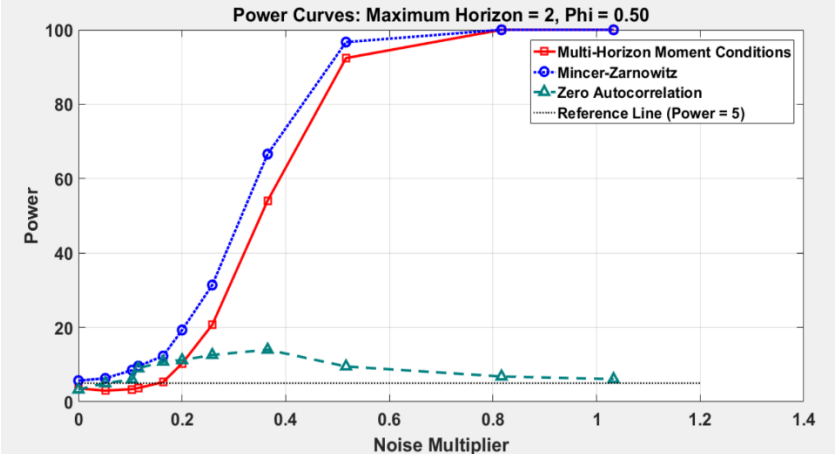
The reason I use the Mincer-Zarnowitz as a reference is because this test is well-known for its strong power. In having similar, or even better performance, the MHMC test can be said to be performing really well. As noted, for certain phi values, when $h_{max} = 2$ and $h_{max} = 3$, the MHMC test performs better. This means that for forecasts with a large enough phi value, it can be argued that the MHMC test is more reliable at rejecting forecast optimality for noisy, non-optimal multi-horizon forecasts than the Mincer-Zarnowitz test.

When $h_{max} = 4$, the MHMC test performs as well as the Mincer-Zarnowitz test when phi = 0.25 and better when phi = 0.5 and 0.75. This is a great result to see for this test, as it does a better job of rejecting this type of suboptimal multi-horizon forecast for a larger range of phi values. The intuition behind this result has to do with the total number of conditions being tested to look at forecast optimality. In the case where $h_{max} = 4$, 16 unique conditions are tested, while this number is 9 when $h_{max} = 3$ and 4 when $h_{max} = 2$. With 16 conditions, the test has more information available to it and more opportunity to determine that a condition is violated and conclude that a forecast is not optimal. With this result, it is evident that the MHMC test appears to perform better than the Mincer-Zarnowitz forecast optimality test for both higher values of phi and longer time horizons for this type of suboptimal multi-horizon forecast.

a



b



c

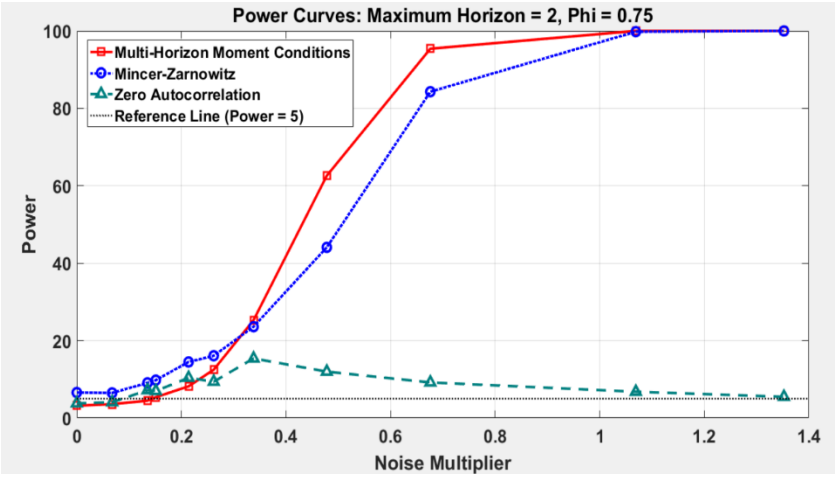
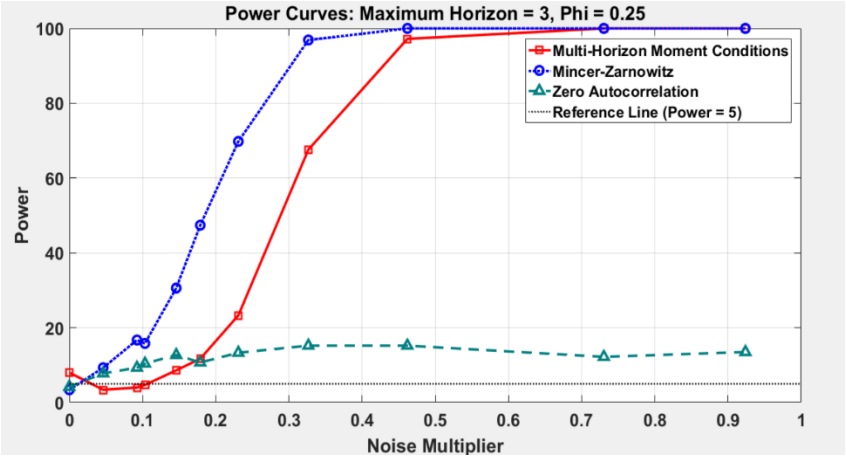
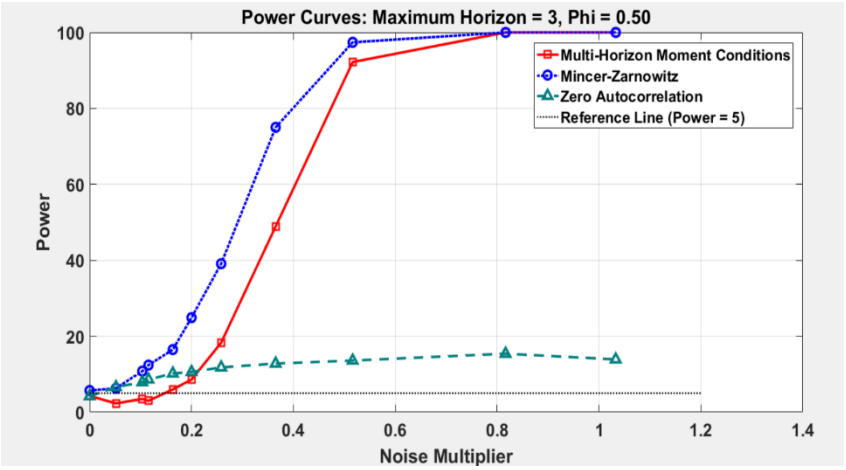


FIG. 2.–Power curves when h_{max} = 2, T = 150 and a. Phi = 0.25, b. Phi = 0.5 and c. Phi = 0.75

a



b



c

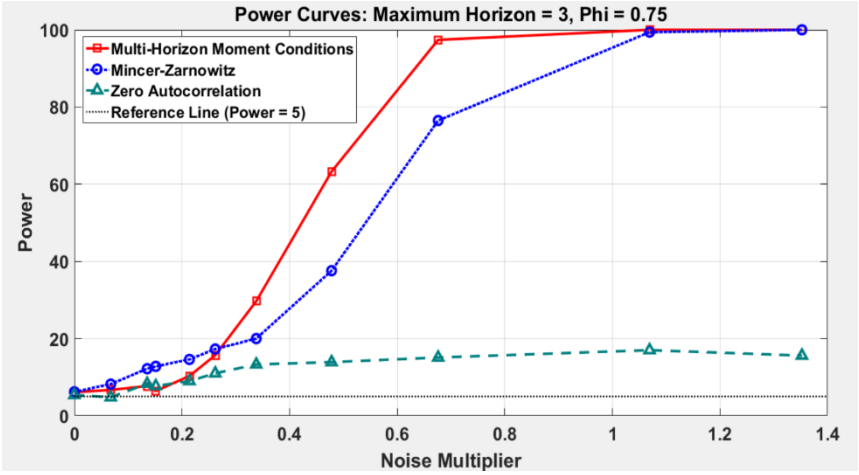
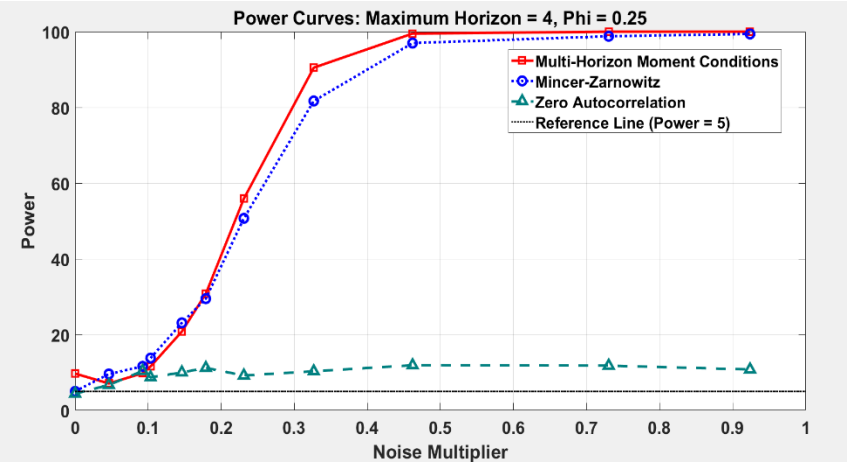
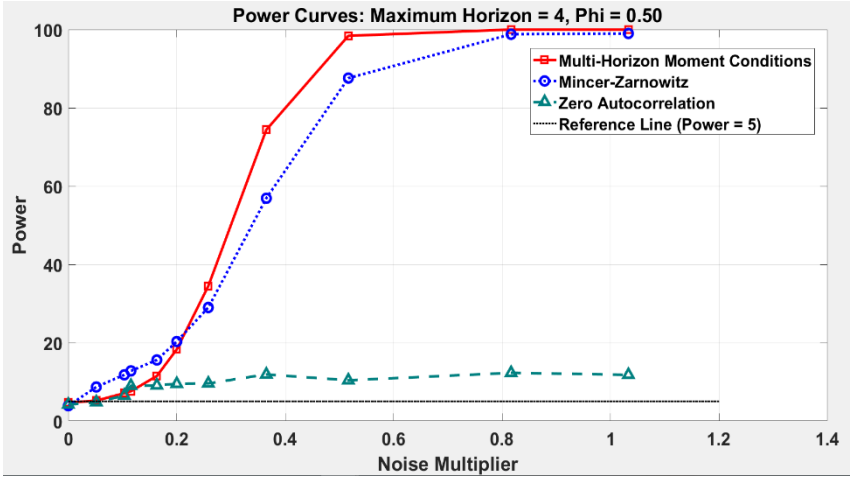


FIG. 3.–Power curves when h_{max} = 3, T = 150 and a. Phi = 0.25, b. Phi = 0.5 and c. Phi = 0.75

a



b



c

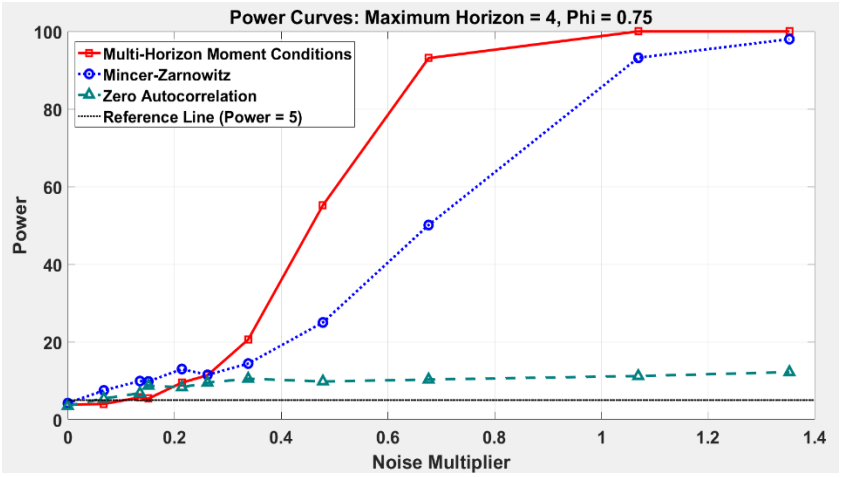


FIG. 4.–Power curves when h_{max} = 4, T = 150 and a. Phi = 0.25, b. Phi = 0.5 and c. Phi = 0.75

Before moving on to general conclusions about these results, I want to get across an extremely important point about the analysis of the power for these tests. In this case, I am specifically looking at the power for the type of non-optimal forecasts that I generated. I generated suboptimal forecasts by adding noise to the optimal forecast, which is detailed in Equation (21) above. There are many ways that a forecast can fail to be optimal, which is why multiple tests exist looking at forecast optimality, and this is just one of those ways. The conclusions that I am drawing about the power properties of these tests applies to how reliably they can reject this type of suboptimal forecast. This is important to understand, as it puts into context some of the results, such as what was found for the zero autocorrelation test. For this type of suboptimal forecast, it has no power, but that does not mean that it has no power for all types of irrational forecasts. Although this power study does not cover the realm of all possible types of irrational forecasts, it still gives strong insight into how the tests perform and compare to each other.

Overall, the MHMC test's power curves, across all maximum forecast horizon and phi values looked at, show the quality of this test. A key component of any test evaluating forecast optimality is that it can reject forecasts that are not optimal and these results demonstrate that the MHMC test has this ability. Also, now that it is evident that the MHMC test can stand on its own and that the moment conditions that it evaluates are legitimate, there is now the possibility to combine this test with the Mincer-Zarnowitz and zero autocorrelation tests. These tests evaluate different properties that should hold for forecast optimality. A combination of these tests would be expected to have both stronger power curves and reject a wider breadth of different types of irrational forecasts. I explore this idea in the next section.

5.A. Combining Forecast Optimality Tests

In the previous section, I claim that a combination of the Mincer-Zarnowitz, zero autocorrelation and Multi-Horizon Moment Conditions (MHMC) forecast optimality tests should result in a test that will have stronger power curves and reject a wider breadth of different types of irrational forecast. I now evaluate the validity of this statement. By combining these three tests in two separate manners, I can use the same suboptimal forecasts that I did in the last section to see how the new power curves compare to the power curves of each individual test.

As mentioned, I combine the three individual tests in two different ways. The first way that I combine these tests is by combining all their moment conditions into one big GMM test. Each test's associated moments are put into one test, from which I obtain a J-statistic and p-value. The second way that I combine these tests is with a method known as Bonferroni bounds. With this method, I perform each test individually with a significance level that is one-third that of the significance level for the overall test. This means that if the significance level that I use for the overall test is 5 %, that I use 1.67 % as the significance level for each individual test to see if it rejects or fails to reject forecast optimality. I then use an "or" statement between all the individual tests that says, if any of them reject optimality, then optimality should be rejected overall by the combined test.

With the combined moments test, I once again have the situation where there are redundant moments. I handle this issue in the same manner that I did earlier, by including only the unique moments in the GMM test. I make sure that I include at least one moment from the Mincer-Zarnowitz, zero autocorrelation and MHMC test so that the combined test includes moments from the three complementary forecast optimality tests. Removing these redundant moments is necessary so that the GMM test is reliable and does not always produce an extremely

large J-statistic that strongly rejects forecast optimality, indicating that the asymptotic distribution does not hold in finite samples.

5.B. Combined Tests Size Discussion

I first evaluated the size of the combined tests, to make sure the asymptotic distribution of the test statistic is a good approximation in a finite sample, before moving on to evaluate the tests' power properties. I perform this size analysis the same way that I did earlier. I look at different sample sizes and values of phi, using a significance level of 5 %, and see how those parameters impact the size value of the combined tests.

The size of the combined tests can be found in Tables 4, 5 and 6. I will first focus on the results from Panel A. Just as with the Multi-Horizon Moment Conditions (MHMC) test, the combined moments test generally has size values close to what is expected for a reliable test. This means that the size is consistently close to the 5 % significance level used when evaluating forecast optimality for most phi and sample sizes. The problematic values are found at low phi, for all maximum forecast horizons, and very high phi when $h_{max} = 2$. In these cases, the combined moments test is oversized and has a size value noticeably greater than the 5 % significance level. It is not surprising that this is the case for low phi values. The MHMC test was oversized for low phi and this issue is not mitigated by combining this test with the Mincer-Zarnowitz and zero-autocorrelation test. It is interesting to see that the combined moments test is now slightly oversized when phi = 0.9 and $h_{max} = 2$. None of these tests appeared to be individually oversized at these values, however, it appears that this phi is impacted when the tests are combined.

Next, I will discuss the size results from combining the individual tests with the Bonferroni bounds method, which can be found in Panel B in the tables below. For each maximum forecast horizon, the Bonferroni bounds combined test is oversized when phi = 0.1 and then has values reasonably close to the 5 % significance level across the other phi values and sample sizes. These results are in agreement with what would be expected based on the size values from each individual test. The values for the higher phi are slightly oversized, particularly in the cases where $h_{max} = 3$ and $h_{max} = 4$, but there are no substantial issues. Just as with the combined moments combined test, these size values are reasonable and it appears valid to move on and evaluate the power of both of these tests.

By combining the three individual tests in two different ways, my goal was to determine the best way to combine these tests into one. In doing this size analysis, I have gained valuable information into solving this problem. By seeing that the Bonferroni bounds test and combined moments test have similar size values, it is apparent that there is no obvious advantage to using one test over the other, in terms of finite-sample size. If there is any difference, it appears that the Bonferroni bounds test generally has more issues with being oversized than the combined moments test. However, this difference is not substantial enough to make any conclusions about what is the better way to combine the Mincer-Zarnowitz and zero autocorrelation test, along with the MHMC test.

TABLE 4:

Size values for the combined tests when $h_{max} = 2$

Size Values: $h_{max} = 2$					
T	**ϕ**				
	0.1	**0.25**	**0.5**	**0.75**	**0.9**
Panel A: Combined Moments					
50	3.2	4.1	3.2	5.2	7.8
100	6.5	6.3	6.5	7.1	11.4
250	6.7	6.1	9	10	9.2
500	5.8	7.6	8.3	8.6	9.7
Panel B: Bonferroni Bounds					
50	13.5	11.4	6.3	8.5	7.4
100	10.9	5.8	6.2	7.4	8.4
250	5.3	4.2	5.5	6.8	7
500	3.1	3.5	2.9	5.2	5.6

TABLE 5:

Size values for the combined tests when $h_{max} = 3$

Size Values: $h_{max} = 3$					
T	**ϕ**				
	0.1	**0.25**	**0.5**	**0.75**	**0.9**
Panel A: Combined Moments					
50	28.9	10.5	4	5.3	4.4
100	30.7	14.9	8.5	8.6	8.3
250	26.5	14.8	7	8.8	9.2
500	18.7	13.2	8.1	8.9	11.1
Panel B: Bonferroni Bounds					
50	25.7	15.6	8.1	7.7	8.1
100	21.8	15.7	8.7	9.7	8.9
250	17.9	8.3	7.4	9.6	11.1
500	13.9	5	5.8	7.6	9.5

TABLE 6:

Size values for the combined tests when $h_{max} = 4$

Size Values: $h_{max} = 4$					
T	**ϕ**				
	0.1	**0.25**	**0.5**	**0.75**	**0.9**
Panel A: Combined Moments					
50	22.4	8.2	6.9	6.1	6
100	34.7	12.4	6.9	6.6	7.8
250	34.6	12.1	7.1	5.9	6.2
500	27.4	13.2	7.6	6.7	6.7
Panel B: Bonferroni Bounds					
50	35.4	10.5	6	5.3	5.3
100	37.3	12.1	8.5	10	8.7
250	38	12.5	6.9	9.1	8.9
500	36.8	11.2	7.5	8.5	8.1

5.C. Combined Tests Power Discussion

When generating results for the combined power curves, I used a sample size of 150, just as I did for the earlier power study. On these graphs, I include the following curves: combined moments, Bonferroni bounds, Mincer-Zarnowitz and Multi-Horizon Moment Conditions (MHMC). I do not include the zero autocorrelation curve in these figures, as it is shown in the last section that this test does not have any power and hovers around 5 % for this type of suboptimal forecast.

I will start by discussing the relevant findings from when phi = 0.25 for all maximum forecast horizons. In looking at the first plot in Figures 5, 6 and 7, both combined tests perform better than the MHMC test alone. This demonstrates that, by combining the three tests, the power of the combined tests is stronger than the weakest link that has power, which is the MHMC test. In two cases, when $h_{max} = 2$ and $h_{max} = 4$, the combined moments power curve is above all other

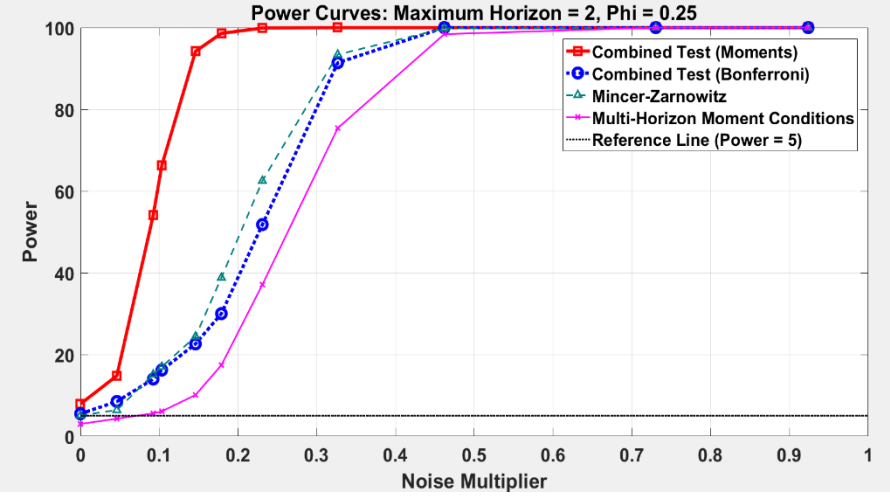
curves, which means that it has more power than the most powerful individual test. Also, these plots demonstrate that the combined tests do reject a wider breadth of suboptimal forecasts. As mentioned, the zero autocorrelation test has no power for this type of irrational forecast. Both combined tests, however, do have power and are able to reject forecast optimality for a greater number of forecasts as they get worse. These initial findings when phi = 0.25 confirm my original hypothesis regarding the benefits of combining the Mincer-Zarnowitz, zero autocorrelation and MHMC test into one.

Looking at the second and third plots in Figures 5, 6 and 7, it is evident that the results hold across phi. The combined moments test is consistently more powerful than the most powerful individual test when $h_{max} = 2$ and $h_{max} = 4$. When $h_{max} = 3$, the combined moments test consistently tracks the most powerful individual test for all values of phi. With the Bonferroni bounds test, it consistently lines up with the most powerful test for all maximum forecast horizons and values of phi. These results demonstrate that the findings when phi = 0.25, hold when phi = 0.5 and phi = 0.75 and that the same benefits from combining the three individual tests into one apply in these cases.

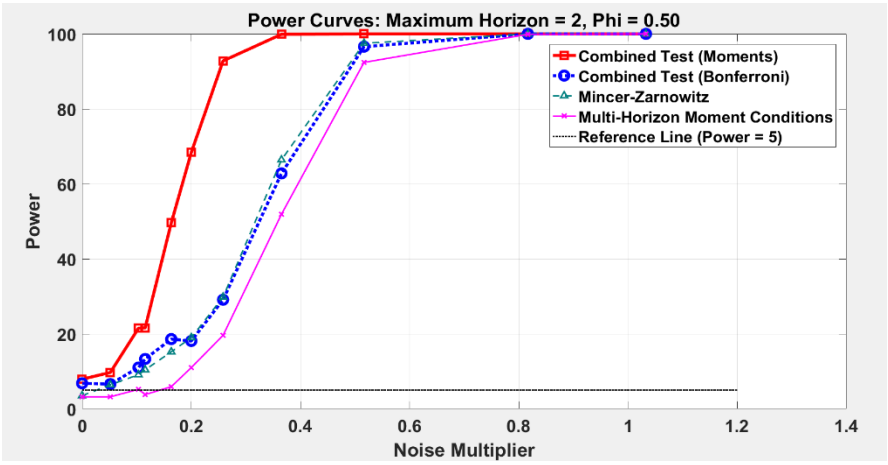
The last piece is to touch on what combined test appears to be doing the best job. There does not appear to be a substantial difference between the size values for the Bonferroni bounds and combined moments tests. However, when looking at each tests' power curves, the combined moments test appears to be doing a better job. For most combinations of phi and maximum forecast horizon, the combined moments test has more power than the Bonferroni bounds combined test. In the very worst case for the combined moments test (when $h_{max} = 3$), it has about the same power as the other combined test. Thus, it appears to be pretty consistently more powerful than the Bonferroni bounds combined test. My conclusion from this is that the

combined moments test is the better way to combine these three forecast optimality tests. I still think the Bonferroni bounds method is a solid way to combine these tests, but that it is not as effective as the combined moments test.

a



b



c

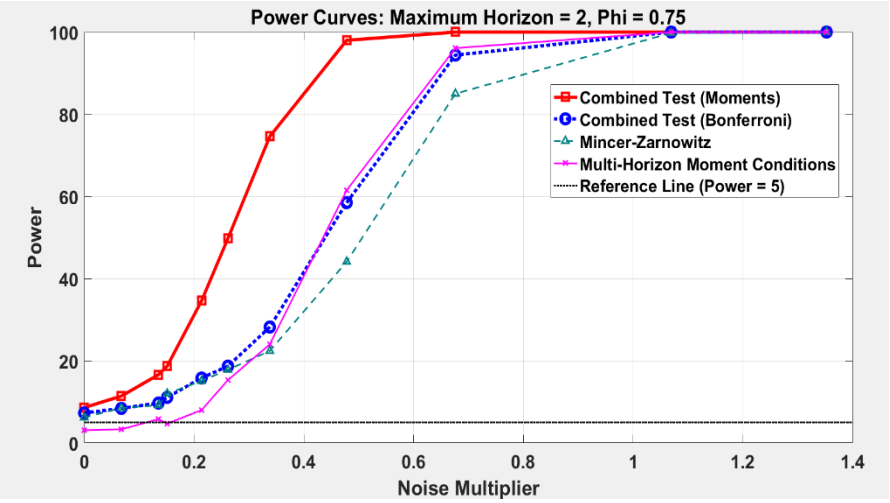
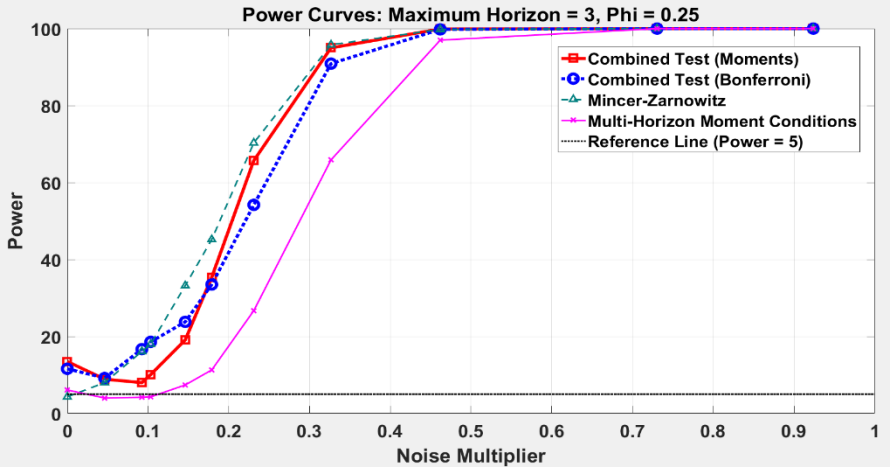
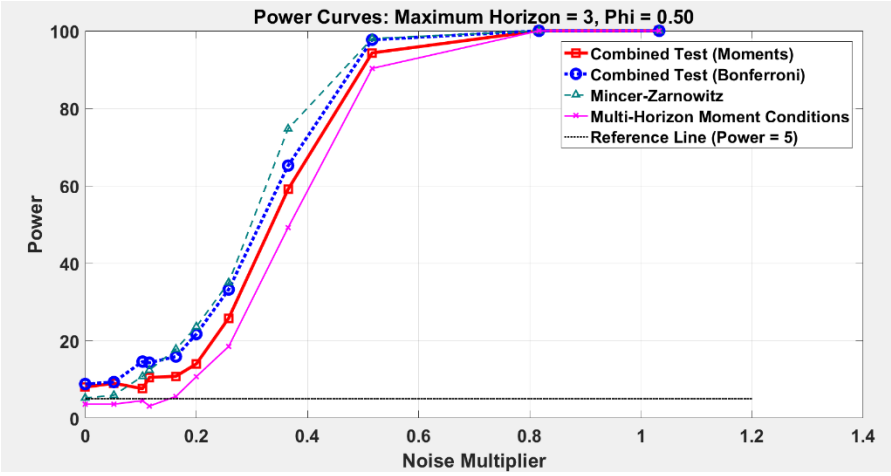


FIG. 5.–Power curves when H =2, T = 150 and a. Phi = 0.25, b. Phi = 0.5 and c. Phi = 0.75

a



b



c

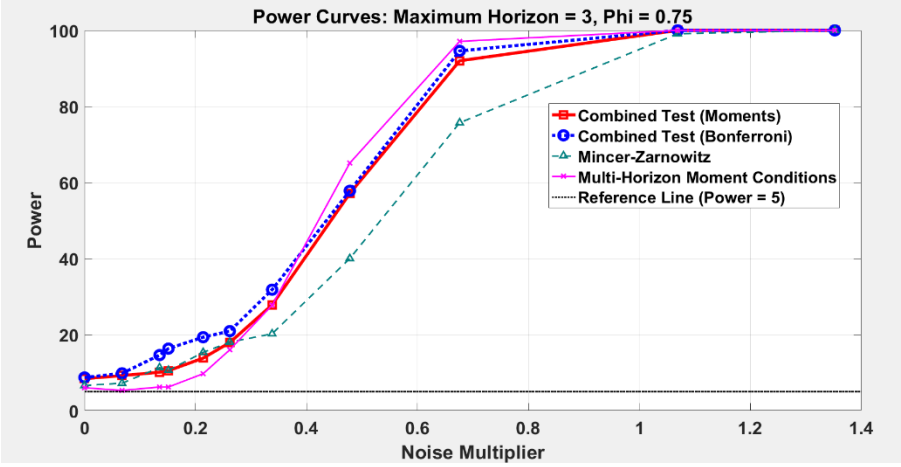
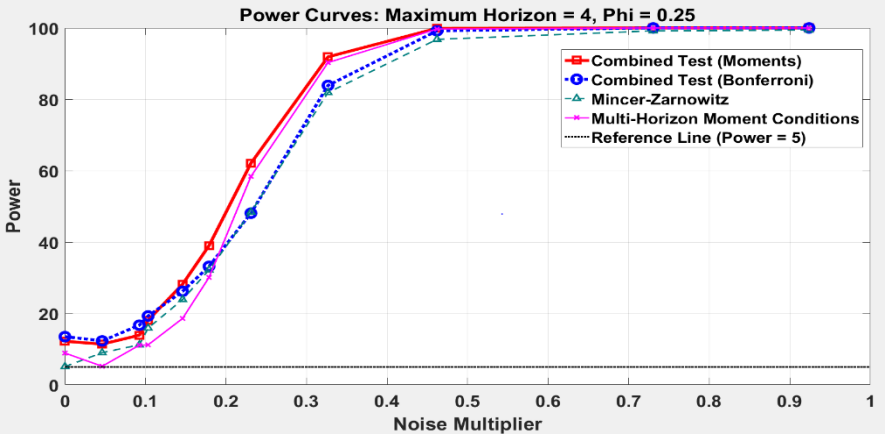
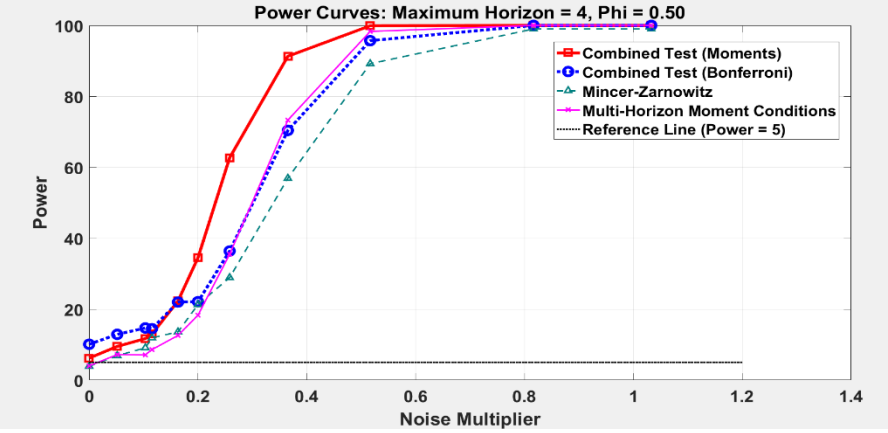


FIG. 6.–Power curves when H =3, T = 150 and a. Phi = 0.25, b. Phi = 0.5 and c. Phi = 0.75

a



b



c

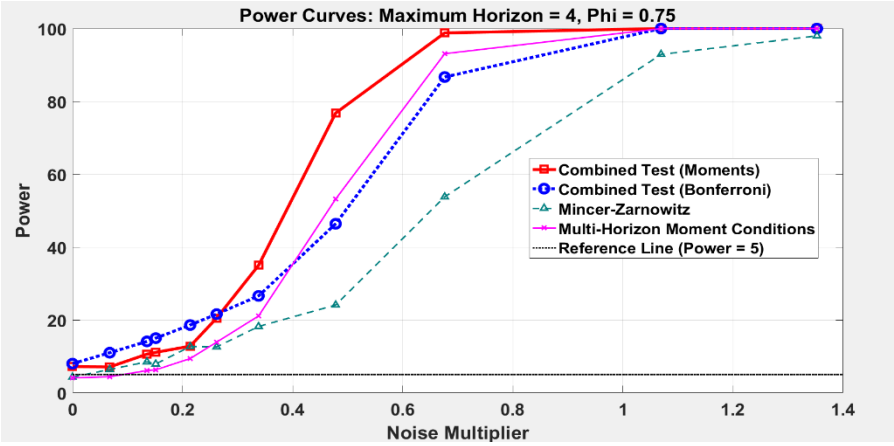


FIG. 7.–Power curves when H =4, T = 150 and a. Phi = 0.25, b. Phi = 0.5 and c. Phi = 0.75

After going through this analysis, I have explained all the tests that I use to evaluate empirical forecasts. I can now see if any of these tests reject forecast optimality of certain forecasts from the Survey of Professional Forecasters. I chose to evaluate forecasts from this group since they are very well-known and develop numerous multi-horizon forecasts that are made quarterly, which means that the MHMC test can be used to test for forecast optimality. The reason forecast optimality tests are developed is to evaluate empirical data, with the intention of seeing if the forecasters are doing an optimal job. If they are not, then the discussion turns into what they could do better and what feature of optimality their forecasts did not pass.

6.A. Empirical Application

Using the Multi-Horizon Moment Conditions (MHMC), Mincer-Zarnowitz, zero autocorrelation and combined tests, I evaluate the optimality of multi-horizon forecasts made quarterly by the Survey of Professional Forecasters for the following variables: Consumer Price Index (CPI) Inflation Rate, Real GDP Growth, Change in the 3-Month Treasury Bill Rate and the Percent Change in the Industrial Production Index ("Survey of Professional Forecasters", 2018). I chose to evaluate these forecasts because they all look at important macroeconomic variables. The Survey of Professional Forecasters provides data for forecasts for each of these variables made by certain individuals. Instead of evaluating one individual forecast, I examine the "consensus" forecast, which is the mean forecasted value across all survey respondents.

For each variable, I evaluate the forecasts that predict how it will change over time. Looking at the change or percent change, as opposed to the absolute level, is important because it makes it much more likely that the data sets are covariance stationary. It is not possible to prove covariance stationarity however, it is possible to reject it for a given time series. For the absolute level of all variables, except for the 3-Month Treasury Bill Rate, there is evidence against them

being covariance stationary. By converting to the change or percent change, it is much more likely that they are covariance stationary, although it cannot be proved. It is essential that these data sets are covariance stationary as this is a required condition for the Wold Decomposition, which is used to derive the moment conditions for the MHMC test. If this assumption does not hold, the Wold Decomposition is not valid, which means that the derived moments for the MHMC test are also no longer valid.

I also want to discuss how I calculate the phi value for each data set. As seen earlier, this value plays an important role in the reliability and effectiveness of the MHMC and combined tests when evaluating the size and power with simulated data and forecasts. To calculate this value, I run the following regression on the realized values of each data set:

$$y_t = \phi_0 + \phi_1 y_{t-1} + e_t. \tag{24}$$

In this regression, the coefficient, ϕ_1, is the estimated phi value for the data set. I also run a 95 % confidence interval on this coefficient, to find the range of values where I can say that I am 95 % certain this phi value lies. I include the results from this regression for each time series in Table 7. In generating these values, it is evident that the CPI Inflation Rate time series has a lower phi value, which leads to poor size properties in the simulation studies. This means that caution is needed when I use the MHMC and combined tests to evaluate this data set. However, the three other times series have relatively large estimated ϕ_1 values, so I feel confident that the MHMC test and combined tests should work well for those.

I include the p-values from each multi-horizon forecast evaluation in Table 8. With each optimality test, the null hypothesis is that the forecast being evaluated is optimal. The p-value can be used to say whether I reject or fail to reject the null hypothesis of optimality. For these

tests, I use a significance level of 5 %, as I did with all the simulations. I highlight the forecasts where the p-value is less than .05, to indicate that forecast optimality is rejected. As mentioned throughout this thesis, the MHMC test, along with the Mincer-Zarnowitz and zero autocorrelation tests are meant to serve as complements to each other. They test for different properties of an optimal forecast under mean squared error loss. This is why I evaluate each multi-horizon forecast with the three individual tests and the combined tests. I can see how many properties, if any, these forecasts appear to be violating for a rational forecast. Also, as shown in the finite-size simulation study, each test individually has around a 5 % chance of falsely rejecting the null hypothesis. This serves as additional motivation for looking at the results with all the individual and combined tests, in order to get the full story behind the optimality of each multi-horizon forecast.

TABLE 7:

Calculation of ϕ_0 and ϕ_1 for all time series

			95 % Confidence Interval	
Variable	**Parameter**	**Estimate**	**Lower**	**Upper**
CPI Inflation Rate	Φ_0	2.409	1.815	3.003
	Φ_1	0.094	-0.074	0.262
Real GDP Growth	Φ_0	1.759	1.171	2.347
	Φ_1	0.358	0.215	0.501
Change in the 3-Month Treasury Bill Rate	Φ_0	-0.044	-0.126	0.037
	Φ_1	0.310	0.172	0.448
Percent Change in the Industrial Production Index	Φ_0	0.908	0.168	1.647
	Φ_1	0.581	0.466	0.697

TABLE 8:

P-values from each test used to evaluate forecasts from the Survey of Professional Forecasters

Survey of Professional Forecasters Forecast Evaluation			
Test	**Maximum Horizon**		
	2	**3**	**4**
Panel A: CPI Inflation Rate (Phi = 0.094, T = 141)			
Mincer-Zarnowitz	**0.021**	**0.046**	0.750
Zero-Autocorrelation	**0.030**	0.235	1.000
Multi-Horizon Moment Conditions	0.451	0.481	0.708
Combined Test: Bonferroni Bounds	0.064	0.137	1.000
Combined Test: Combined Moments	**0.036**	**0.001**	**0.000**
Panel B: Real GDP Growth (Phi = 0.389, T = 168)			
Mincer-Zarnowitz	0.070	0.259	0.739
Zero-Autocorrelation	0.120	0.180	1.000
Multi-Horizon Moment Conditions	0.953	0.924	0.270
Combined Test: Bonferroni Bounds	0.211	0.541	0.810
Combined Test: Combined Moments	0.186	**0.007**	0.067
Panel C: Change in the 3-Month Treasury Bill Rate (Phi = 0.310, T = 144)			
Mincer-Zarnowitz	**0.000**	**0.011**	**0.013**
Zero-Autocorrelation	**0.019**	**0.000**	0.472
Multi-Horizon Moment Conditions	**0.001**	**0.000**	**0.000**
Combined Test: Bonferroni Bounds	**0.001**	**0.000**	**0.000**
Combined Test: Combined Moments	**0.000**	**0.000**	**0.000**
Panel D: Percent Change in the Industrial Production Index (Phi = .581, T = 196)			
Mincer-Zarnowitz	0.252	0.491	0.438
Zero-Autocorrelation	0.125	1.000	0.541
Multi-Horizon Moment Conditions	**0.030**	**0.006**	**0.002**
Combined Test: Bonferroni Bounds	0.089	**0.019**	**0.006**
Combined Test: Combined Moments	**0.009**	**0.003**	**0.033**

Bold = forecast optimality rejected at 5 % significance level

6.B. Consumer Price Index Forecast Evaluation

The first multi-horizon forecast that I will discuss is the one made by the Survey of Professional Forecasters for the CPI Inflation Rate. CPI is defined as the weighted average of the prices for a basket of goods and services and is used to get the CPI Inflation Rate by looking at the percent change in its level over time ("Consumer Price Index – CPI", 2018). Being able to accurately forecast this variable is important for determining what economic policy to set to keep the United States economy healthy and in good standing. For this data set, the estimated phi value is .094 and the sample size is 141. Based on the simulation size study that I conducted, it was shown that, when phi = 0.1, the Multi-Horizon Moment Conditions (MHMC) test and combined tests were oversized when $h_{max} = 3$ and $h_{max} = 4$. This means that the MHMC test, as well as the combined tests, are not reliable ways to evaluate those forecasts. However, when $h_{max} = 2$, this issue no longer exists. I focus on the results from all tests when $h_{max} = 2$ and only the results for the Mincer-Zarnowitz and zero autocorrelation forecast optimality tests when $h_{max} = 3$ and $h_{max} = 4$.

When $h_{max} = 2$, forecast optimality is rejected by both the Mincer-Zarnowitz and zero autocorrelation tests. The combined test using combined moments also rejects optimality, while the combined test using Bonferroni bounds does not. One of the goals of the combined tests is to reject optimality when any of the individual tests do, which shows that the combined moments test appears to be doing a more effective job than the Bonferroni bounds combined test. These results indicate that for the multi-horizon forecast for the CPI Inflation Rate that includes the 1 and 2 quarter out forecasts that the Survey of Professional Forecasters do not appear to be doing an optimal job. The failure of these tests hopefully can give this group insight into why their multi-horizon forecast is not optimal in this case and what they can change to improve it.

As mentioned, when $h_{max} = 3$ and $h_{max} = 4$, the results from the MHMC and combined tests are not reliable, due to the low phi value of the data set. It is interesting to note that optimality is still rejected by the Mincer-Zarnowitz test when $h_{max} = 3$. In this case, the zero autocorrelation test fails to reject forecast optimality. When $h_{max} = 4$, none of the viable tests reject forecast optimality. This demonstrates that the issues in the multi-horizon forecast are more heavily focused in the $h = 1$ and $h = 2$ forecasts and that the Mincer-Zarnowitz and zero autocorrelation criteria for forecast optimality hold more strongly when $h = 3$ and $h = 4$.

There are two key takeaways from the evaluation of the Survey of Professional Forecasters' multi-horizon forecast of the CPI Inflation Rate. First, this analysis shows the effectiveness of the combined moments combined test. One of the goals of this test was to reject a wider breadth of forecasts than any individual test. By evaluating the moment conditions of the Mincer-Zarnowitz, zero autocorrelation and the MHMC test in one test, it rejected forecast optimality when $h_{max} = 2$, even though the MHMC test did not. The Bonferroni bounds test, which combines the three individual tests in a slightly different way, is unsuccessful in doing this. The second takeaway is that the Survey of Professional Forecasters could be doing a better job in their forecasts that are made 1 and 2 quarters out for the CPI Inflation Rate. When $h_{max} = 2$, two different tests reject forecast optimality, meaning two different sets of optimal properties that should hold under squared error loss do not.

6.C. Real GDP Growth Forecast Evaluation

The Survey of Professional Forecasters' multi-horizon forecast of Real GDP Growth in the US is the second forecast that I evaluated. GDP is the value of all goods and services in a country. Real GDP Growth is the percent change in value of GDP over a certain time period, adjusted for inflation ("Gross Domestic Product – GDP", 2018). This is another important

macroeconomic variable to forecast accurately. Knowing what GDP is expected to look like in the future can help shape policy to make sure the US economy stays healthy and stable. For example, if it looks like GDP is going to go down, US policy makers would need to think of a strategy that can help limit the issues that could bring or to turn that trend around.

When $h_{max} = 2$ and $h_{max} = 4$, forecast optimality is never rejected, while when $h_{max} = 3$, there is only one test for which forecast optimality is rejected. This illustrates that, when forecasting Real GDP Growth, the Survey of Professional Forecasters are doing a pretty optimal job. It is reassuring to see that optimality is only rejected for the combined moments combined test when $h_{max} = 3$. It means that, overall, the forecasters are doing a good job of forecasting this important macroeconomic variable. This is insightful information, as it means that these forecasts should be looked at and referenced when setting economic policy, as they are reliable and rational forecasts of Real GDP Growth.

The only result that I would like to discuss further is the rejection of forecast optimality when $h_{max} = 3$ by the combined moments test. There are two possible explanations for this. The first is that the evaluation of all three tests together results in a rejection of forecast optimality. No one test individually rejects forecast optimality, yet, when evaluated together in one test, there is enough there to reject the null hypothesis. This would mean that there are some small issues with each test that, when combined, result in a rejection.

Another possible explanation has to do with the estimation of the phi value. Looking at the 95 % confidence interval in Table 7 for the estimated phi value, the lower bound is .21. If the phi value was actually closer to this value than the estimated .359, there could be issues with the reliability of the Multi-Horizon Moment Conditions (MHMC) and combined tests. As shown, for values of phi that are .25 or lower, there are issues with these tests being oversized when $h_{max} = 3$

and $h_{max} = 4$. Being oversized would result in a higher test statistic and lower p-value from the test. If this was the case with the phi value for this data set, it could explain the low p-value and rejection from this combined test. I find this explanation to be less likely, as this would also affect the MHMC test when $h_{max} = 3$ and $h_{max} = 4$ and the combined tests when $h_{max} = 4$. Nonetheless, whichever explanation is correct, the moral of the story is the same. The Survey of Professional Forecasters do a good job of forecasting Real GDP Growth in the US.

6.D. Change in the 3-Month Treasury Bill Rate Forecast Evaluation

The third multi-horizon forecast that I will discuss is the one made by the Survey of Professional Forecasters for the Change in the 3-Month Treasury Bill Rate. This value is the change in the yield of 3-month treasury bill in the US over time ("Treasury Bill - T-Bill", 2018). Being able to forecast this change is important to investors. The price of a treasury bill is related to its yield and being able to know how the yield will change over time can give insight into how the price will change. This forecast must also take into account variables that impact yield, such as the interest rate in the United States and can give insightful information into those variables.

For these multi-horizon forecasts, optimality is rejected for all tests when $h_{max} = 2$ and $h_{max} = 3$. When $h_{max} = 4$, the zero autocorrelation test is the only test where forecast optimality is not rejected. This demonstrates that for this forecast, the Survey of Professional Forecasters are really doing a suboptimal job. This is the opposite of what was just found for this group's forecasts of Real GDP Growth. In two cases, three different individual tests reject forecast optimality and in one case two different individual tests reject forecast optimality for this multi-horizon forecast. These individual tests all evaluate different properties of optimal forecasts under squared error less, which means that there is a lot that the Survey of Professional

Forecasters need to think about when trying to improve their forecast for the Change in the 3-Month Treasury Bill Rate.

These results indicate that these multi-horizon forecasts for the Change in the 3-Month Treasury Bill Rate are not optimal to use when trying to predict the future value of this variable. The fact that forecast optimality is so strongly rejected in so many cases demonstrate that these forecasts are far from optimal. The Survey of Professional Forecasters need to think about what could cause such strong rejections. They should start by trying to find out if there is one consistent issue that leads to so many tests rejecting forecast optimality or if there are there a plethora of different issues that need to be corrected for. Nonetheless, this analysis should result in one principal conclusion with regard to this forecast. The Survey of Professional Forecasters are not doing an optimal job of forecasting the Change in the 3-Month Treasury Bill Rate and should consider why this is the case.

6.E. Percent Change in the Industrial Production Index Forecast Evaluation

The final multi-horizon forecast that I will discuss is the one made by the Survey of Professional Forecasters for the Percent Change in the Industrial Production Index. The Industrial Production Index level measures real output of all facilities involved with manufacturing, mining and electric and gas utilities in the United States ("Board of Governors of the Federal Reserve System: Industrial Production Index", 2018). By looking at how this level changes over time, the Percent Change in the Industrial Production Index can be computed. Just as with the three other variables that have been evaluated, accurately forecasting this variable is important. Knowing how this value is expected to change over time can give a sense of how this sector is expected to perform, which could help indicate the future strength of the US economy, as well as shape investment decisions.

For this forecast of the Percent Change in the Industrial Production Index, forecast optimality is rejected by the Multi-Horizon Moment Conditions (MHMC) test and both combined tests when $h_{max} = 3$ and $h_{max} = 4$. When $h_{max} = 2$ forecast optimality is rejected by the MHMC test and just the combined moments combined test. This is an extremely meaningful and exciting result for the MHMC test, as it is the only individual test where forecast optimality is rejected. This means that if I had not developed the MHMC test and only evaluated forecast optimality of this multi-horizon forecast with the Mincer-Zarnowitz and zero autocorrelation tests, that I would have failed to reject the null hypothesis of forecast optimality in all cases. The goal of the MHMC test is to serve as a complement to existing tests and to evaluate new features of forecast optimality. In developing this test, I wanted to be able to identify a wider breadth of suboptimal forecasts, with the intent of trying to give forecasters as many tools as possible to make rational and optimal forecasts. In this case, it appears that the Survey of Professional Forecasters are doing a pretty good job of satisfying the properties of an ideal forecast tested by the Mincer-Zarnowitz and zero autocorrelation test, but not as great of a job with the properties that the MHMC test evaluates.

I will now try and give the intuition behind this result and explain why forecast optimality is rejected by the MHMC test, but not by the Mincer-Zarnowitz and zero autocorrelation tests. The Mincer-Zarnowitz test evaluates the regression in Equation (22) above, which looks at whether a forecast is biased or inefficient. This test ultimately is evaluating the absolute accuracy of a forecast and looking at how the forecasted values compare to the realized values of a variable. Since the test fails to reject forecast optimality, it can be said that α_0 is not statistically different from 0 and α_1 is not statistically different from 1 in the Mincer-Zarnowitz regression. The zero autocorrelation test evaluates the regression in Equation (23) above, looking

at specific correlations in forecast errors. There should be no relationship between the forecast errors available at the time a forecast is made and forecast errors from forecasts made on that day. For this forecast, there is no indication that there are correlations in these forecast errors, as forecast optimality is never rejected by the zero autocorrelation test.

Lastly, the MHMC test looks at the variances, covariances and autocovariances between particular forecast errors made for a multi-horizon forecast. Specifically, it looks at the optimal forecast errors for forecasts that have an overlapping period of time between when the forecast is made and realized and should have a non-zero relationship. In each case, these relationships between optimal forecast errors do not appear to hold. A plausible explanation for this is that the information available to forecasters is not being used efficiently across all horizons. The relationship between the forecast errors that I have derived for the MHMC are not holding, which means the forecasts are being differently impacted by news that results in them deviating from the actual value. The main takeaway from this analysis should be that the Survey of Professional Forecasters could do a better job of forecasting the Percent Change in the Industrial Production Index over multiple horizons and that the MHMC test was the tool used to discover this.

6.F. Empirical Application Conclusion

Before concluding this section, I must justify one important assumption that I made for all these forecasts. As a reminder, the moment conditions for the Multi-Horizon Moment Conditions (MHMC) test are all based on squared error loss, which is a symmetric loss function. This means that forecasters do not penalize being above or below the forecasted value differently. If forecasters do not follow this type of loss function, then the tests that I am using to evaluate forecast optimality are not reliable. For these four multi-horizon forecasts, a strong case

can be made that the Survey of Professional Forecasters follow a symmetric loss function. From the Literature Review, it is evident that most times asymmetric loss exists, it is when forecasts are related to developing policy or when companies are trying to give off certain indications about its future performance. With the Survey of Professional Forecasters, these individuals are not pushing policy or trying to influence anyone with their forecasts. They are simply a group of people trying to do their best job to forecast the values of certain variables. There are no clear indications that they would have asymmetric loss for any of the variables that I evaluated and I think it is appropriate to assume squared error loss for these forecasters.

I will conclude this section by summarizing the main points and results. First, it appears that the Survey of Professional Forecasters do an optimal job of forecasting Real GDP Growth, but that there are improvements that can be made with their multi-horizon quarterly forecasts of the CPI Inflation Rate, Change in the 3-Month Treasury Bill Rate and the Percent Change in the Industrial Production Index. Second, the MHMC test demonstrated its usefulness in the forecast evaluation of the Percent Change in the Industrial Production Index forecast, as it was the only individual test to reject forecast optimality. Lastly, the combined moments combined test showed its effectiveness in practice. For the multi-horizon forecast evaluation of the CPI Inflation Rate, it rejected forecast optimality when $h_{\max} = 2$ and only the Mincer-Zarnowitz and zero autocorrelation tests rejected optimality. In this empirical evaluation, this combined test proved that it can reject a wider breadth of forecasts than any individual test, which was shown in the simulation study earlier. This section hopefully provided interesting intuition into how good of a job the Survey of Professional Forecasters are doing forecasting several important macroeconomic variables, while showing the effectiveness and usefulness of the developed MHMC and combined tests in practice.

7. Conclusion

In this thesis, I develop the Multi-Horizon Moment Conditions (MHMC) test for evaluating multi-horizon forecast optimality under squared error loss. The test is meant to serve as a complement to other existing tests that evaluate forecast rationality. It tests for conditions that should hold for an optimal forecast, but are not currently tested for. The MHMC test specifically looks at the variances, covariances and autocovariances of optimal forecast errors that should have a non-zero relationship for multi-horizon forecasts. The zero autocorrelation forecast optimality test says that there should be no correlation between forecast errors available at the time a forecast was made and any forecast errors for forecasts made on that same date for some point in the future. However, there is no test that is currently based on the relationships between optimal forecast errors with overlapping time between when the forecast is made and realized. I derive what these relationships should be when the maximum forecast horizon is 2, 3 and 4 for a multi-horizon forecast.

After deriving the moment conditions for each maximum forecast horizon, I implemented them into a test using Generalized Method of Moments (GMM). I developed a script in MATLAB that can perform the MHMC test and then conducted a simulation study to evaluate its size and power properties. My initial analysis involved looking at the size of this test, along with the well-known Mincer-Zarnowitz and zero autocorrelation tests. When evaluating size, I found that the MHMC test is typically oversized for series with low autocorrelation, but that it generally behaves as anticipated for series with autocorrelation above 0.25. I then performed a power analysis, specifically looking at how powerful the MHMC test is for a noisy, suboptimal forecast. It was found to have power for these types of irrational forecasts and it was shown that there are cases where the MHMC test is more powerful and less powerful than the Mincer-

Zarnowitz test, depending on the phi value of the data generating process and the maximum forecast horizon. The results from the simulation studies that I ran showed that the MHMC test is an effective way to test for forecast optimality and that it would be appropriate to use to evaluate empirical forecasts.

Before moving on to evaluate empirical forecasts using the MHMC test, I developed two different combined forecast optimality tests. Since the Mincer-Zarnowitz, zero autocorrelation and MHMC tests all evaluate different properties of forecast optimality under mean squared error loss, I thought it would be more effective to implement all the properties that they test for into one combined test. The first way that I do this is by combining all the different moments from each test into one big GMM test. Next, the second way that I do this is with a method known as Bonferroni bounds. In this case, I perform each test separately and use a significance level that is one-third of the total significance level chosen for each individual test. After doing this, I evaluate the size and power of these two combined tests. I conclude that the combined moments test is more effective and reliable, as the Bonferroni bounds test is slightly more oversized and typically not as powerful.

I then evaluate the forecast optimality of four different multi-horizon forecasts made by the Survey of Professional Forecasters. These forecasts are all made at 1, 2, 3 and 4 quarter time horizons. I evaluate three multi-horizon forecasts, each with a different maximum forecast horizon, for each test and evaluate each forecast with the three individual forecast optimality tests and two combined tests. The four different forecasts that I evaluate are those for the CPI Inflation Rate, Real GDP Growth, Change in the 3-Month Treasury Bill Rate and the Percent Change in the Industrial Production Index.

There were several interesting results from this analysis. First, for all variables forecasted by the Survey of Professional Forecasters, the only one where I failed to consistently reject optimality was for Real GDP Growth. For all other variables, it appears that the forecasters are violating at least one property of an optimal forecast under squared error loss. Second, I found the combined moments combined test to be more effective in practice when compared to the Bonferroni bounds combined test. It rejected the optimality of forecasts when not all individual tests did, which was specifically the case for the CPI Inflation Rate when $h_{\max} = 2$. Lastly, for the forecast evaluation of the Percent Change in the Industrial Production Index, the only individual test to reject forecast optimality was the MHMC test, demonstrating its usefulness. If only evaluated with the Mincer-Zarnowitz and zero autocorrelation test, forecast optimality would not have been rejected. However, when evaluating this forecast with the MHMC test, rationality was rejected, meaning that it appears that the Survey of Professional Forecasters could be doing a better job.

I believe that I have developed a solid foundation for the MHMC test looking at multi-horizon forecast optimality. However, there are some interesting extensions that exist and further analysis that could be done. To start, I only derive moment conditions under squared error loss for the MHMC test. Not all forecasts can be justified to use this symmetric loss function, which means it is worth seeing if similar moment conditions can be derived from a general loss function. Also, it would be interesting to complete more power studies on different simulated, suboptimal forecasts. I perform one power study where the forecast is suboptimal because it is noisy. There are many ways that a forecast can be irrational and it would be interesting to complete more power studies to see how effective this test is at rejecting different types of irrational forecasts. Lastly, it would be interesting to further examine the issue of the MHMC test

being oversized for series with low autocorrelation. It would be insightful to perform a full analysis of where this test starts to become oversized. With this, it could be worthwhile to see if there is a better way besides GMM to test if the derived moment conditions hold and whether that could mitigate or fix the issue. These are three possible extensions that I have thought about that could continue the development of the MHMC test and help it become as effective as possible in identifying irrational and suboptimal forecasts.

Appendix

A. Optimal forecast errors for all horizons

1. $e^*_{t+1|t} = \varepsilon_{t+1}$

2. $e^*_{t+2|t} = \varepsilon_{t+2} + \theta_1\varepsilon_{t+1}$

3. $e^*_{t+3|t} = \varepsilon_{t+3} + \theta_1\varepsilon_{t+2} + \theta_2\varepsilon_{t+1}$

4. $e^*_{t+4|t} = \varepsilon_{t+4} + \theta_1\varepsilon_{t+3} + \theta_2\varepsilon_{t+2} + \theta_3\varepsilon_{t+1}$

B. 15 moment conditions when $h_{max} = 3$ for the Multi-Horizon Moment Conditions test (A bold number means that the moment condition was included in GMM test)

B.1. Variances

1. $V\left[e^*_{t+1|t}\right] = V[\varepsilon_{t+1}] = \sigma^2$

2. $V\left[e^*_{t+2|t}\right] = V[\varepsilon_{t+2} + \theta_1\varepsilon_{t+1}] = (1 + \theta_1^2)\sigma^2$

3. $V\left[e^*_{t+3|t}\right] = V[\varepsilon_{t+3} + \theta_1\varepsilon_{t+2} + \theta_2\varepsilon_{t+1}] = (1 + \theta_1^2 + \theta_2^2)\sigma^2$

B.2. Autocovariances

4. $Cov\left[e^*_{t+2|t}, e^*_{t+1|t-1}\right] = Cov[\varepsilon_{t+2} + \theta_1\varepsilon_{t+1}, \varepsilon_{t+1} + \theta_1\varepsilon_t] = \theta_1\sigma^2$

5. $Cov\left[e^*_{t+3|t}, e^*_{t+2|t-1}\right] = Cov[\varepsilon_{t+3} + \theta_1\varepsilon_{t+2} + \theta_2\varepsilon_{t+1}, \varepsilon_{t+2} + \theta_1\varepsilon_{t+1} + \theta_2\varepsilon_t] = \theta_1(1 + \theta_2)\sigma^2$

6. $Cov\left[e^*_{t+3|t}, e^*_{t+1|t-2}\right] = Cov[\varepsilon_{t+3} + \theta_1\varepsilon_{t+2} + \theta_2\varepsilon_{t+1}, \varepsilon_{t+1} + \theta_1\varepsilon_t + \theta_2\varepsilon_{t-1}] = \theta_2\sigma^2$

B.3. Covariances

7. $Cov\left[e^*_{t+2|t+1}, e^*_{t+2|t}\right] = Cov[\varepsilon_{t+2}, \varepsilon_{t+2} + \theta_1\varepsilon_{t+1}] = \sigma^2$

8. $Cov\left[e^*_{t+1|t}, e^*_{t+2|t}\right] = Cov[\varepsilon_{t+1}, \varepsilon_{t+2} + \theta_1\varepsilon_{t+1}] = \theta_1\sigma^2$

9. $Cov\left[e^*_{t+3|t+2}, e^*_{t+3|t}\right] = Cov[\varepsilon_{t+3}, \varepsilon_{t+3} + \theta_1\varepsilon_{t+2} + \theta_2\varepsilon_{t+1}] = \sigma^2$

10. $Cov\left[e^*_{t+2|t+1}, e^*_{t+3|t}\right] = Cov[\varepsilon_{t+2}, \varepsilon_{t+3} + \theta_1\varepsilon_{t+2} + \theta_2\varepsilon_{t+1}] = \theta_1\sigma^2$

11. $Cov\left[e^*_{t+1|t}, e^*_{t+3|t}\right] = Cov[\varepsilon_{t+1}, \varepsilon_{t+3} + \theta_1\varepsilon_{t+2} + \theta_2\varepsilon_{t+1}] = \theta_2\sigma^2$

12. $Cov\left[e^*_{t+4|t+2}, e^*_{t+3|t}\right] = Cov[\varepsilon_{t+4} + \theta_1\varepsilon_{t+3}, \varepsilon_{t+3} + \theta_1\varepsilon_{t+2} + \theta_2\varepsilon_{t+1}] = \theta_1\sigma^2$

13. $Cov\left[e^*_{t+3|t+1}, e^*_{t+3|t}\right] = Cov[\varepsilon_{t+3} + \theta_1\varepsilon_{t+2}, \varepsilon_{t+3} + \theta_1\varepsilon_{t+2} + \theta_2\varepsilon_{t+1}] = (1 + \theta_1^2)\sigma^2$

14. $Cov\left[e^*_{t+2|t}, e^*_{t+3|t}\right] = Cov[\varepsilon_{t+2} + \theta_1\varepsilon_{t+1}, \varepsilon_{t+3} + \theta_1\varepsilon_{t+2} + \theta_2\varepsilon_{t+1}] = \theta_1(1 + \theta_2)\sigma^2$

15. $Cov\left[e^*_{t+1|t-1}, e^*_{t+3|t}\right] = Cov[\varepsilon_{t+1} + \theta_1\varepsilon_t, \varepsilon_{t+3} + \theta_1\varepsilon_{t+2} + \theta_2\varepsilon_{t+1}] = \theta_2\sigma^2$

C. 34 moment conditions when $h_{max} = 4$ for the Multi-Horizon Moment Conditions test (A bold number means that the moment condition was included in GMM test)

C.1. Variances

1. $V\left[e^*_{t+1|t}\right] = V[\varepsilon_{t+1}] = \sigma^2$

2. $V\left[e^*_{t+2|t}\right] = V[\varepsilon_{t+2} + \theta_1\varepsilon_{t+1}] = (1 + \theta_1^2)\sigma^2$

3. $V\left[e^*_{t+3|t}\right] = V[\varepsilon_{t+3} + \theta_1\varepsilon_{t+2} + \theta_2\varepsilon_{t+1}] = (1 + \theta_1^2 + \theta_2^2)\sigma^2$

4. $V\left[e^*_{t+4|t}\right] = V[\varepsilon_{t+4} + \theta_1\varepsilon_{t+3} + \theta_2\varepsilon_{t+2} + \theta_3\varepsilon_{t+1}] = (1 + \theta_1^2 + \theta_2^2 + \theta_3^2)\sigma^2$

C.2. Autocovariances

5. $Cov\left[e^*_{t+2|t}, e^*_{t+1|t-1}\right] = Cov[\varepsilon_{t+2} + \theta_1\varepsilon_{t+1}, \varepsilon_{t+1} + \theta_1\varepsilon_t] = \theta_1\sigma^2$

6. $Cov\left[e^*_{t+3|t}, e^*_{t+2|t-1}\right] = Cov[\varepsilon_{t+3} + \theta_1\varepsilon_{t+2} + \theta_2\varepsilon_{t+1}, \varepsilon_{t+2} + \theta_1\varepsilon_{t+1} + \theta_2\varepsilon_t] = \theta_1(1 + \theta_2)\sigma^2$

7. $Cov\left[e^*_{t+3|t}, e^*_{t+1|t-2}\right] = Cov[\varepsilon_{t+3} + \theta_1\varepsilon_{t+2} + \theta_2\varepsilon_{t+1}, \varepsilon_{t+1} + \theta_1\varepsilon_t + \theta_2\varepsilon_{t-1}] = \theta_2\sigma^2$

8. $Cov\left[e^*_{t+4|t}, e^*_{t+3|t-1}\right] = Cov[\varepsilon_{t+4} + \theta_1\varepsilon_{t+3} + \theta_2\varepsilon_{t+2} + \theta_3\varepsilon_{t+1}, \varepsilon_{t+3} + \theta_1\varepsilon_{t+2} + \theta_2\varepsilon_{t+1} + \theta_3\varepsilon_t] = (\theta_1 + \theta_2\theta_1 + \theta_3\theta_2)\sigma^2$

9. $Cov\left[e^*_{t+4|t}, e^*_{t+2|t-2}\right] = Cov[\varepsilon_{t+4} + \theta_1\varepsilon_{t+3} + \theta_2\varepsilon_{t+2} + \theta_3\varepsilon_{t+1}, \varepsilon_{t+2} + \theta_1\varepsilon_{t+1} + \theta_2\varepsilon_t + \theta_3\varepsilon_{t-1}] = (\theta_2 + \theta_3\theta_1)\sigma^2$

10. $Cov\left[e^*_{t+4|t}, e^*_{t+1|t-3}\right] = Cov[\varepsilon_{t+4} + \theta_1\varepsilon_{t+3} + \theta_2\varepsilon_{t+2} + \theta_3\varepsilon_{t+1}, \varepsilon_{t+1} + \theta_1\varepsilon_t + \theta_2\varepsilon_{t-1} + \theta_3\varepsilon_{t-2}] = \theta_3\sigma^2$

C.3. Covariances

11. $Cov\left[e^*_{t+2|t+1}, e^*_{t+2|t}\right] = Cov[\varepsilon_{t+2}, \varepsilon_{t+2} + \theta_1\varepsilon_{t+1}] = \sigma^2$

12. $Cov\left[e^*_{t+1|t}, e^*_{t+2|t}\right] = Cov[\varepsilon_{t+1}, \varepsilon_{t+2} + \theta_1\varepsilon_{t+1}] = \theta_1\sigma^2$

13. $Cov\left[e^*_{t+3|t+2}, e^*_{t+3|t}\right] = Cov[\varepsilon_{t+3}, \varepsilon_{t+3} + \theta_1\varepsilon_{t+2} + \theta_2\varepsilon_{t+1}] = \sigma^2$

14. $Cov\left[e^*_{t+2|t+1}, e^*_{t+3|t}\right] = Cov[\varepsilon_{t+2}, \varepsilon_{t+3} + \theta_1\varepsilon_{t+2} + \theta_2\varepsilon_{t+1}] = \theta_1\sigma^2$

15. $Cov\left[e^*_{t+1|t}, e^*_{t+3|t}\right] = Cov[\varepsilon_{t+1}, \varepsilon_{t+3} + \theta_1\varepsilon_{t+2} + \theta_2\varepsilon_{t+1}] = \theta_2\sigma^2$

16. $Cov\left[e^*_{t+4|t+3}, e^*_{t+4|t}\right] = Cov[\varepsilon_{t+4}, \varepsilon_{t+4} + \theta_1\varepsilon_{t+3} + \theta_2\varepsilon_{t+2} + \theta_3\varepsilon_{t+1}] = \sigma^2$

17. $Cov\left[e^*_{t+3|t+2}, e^*_{t+4|t}\right] = Cov[\varepsilon_{t+3}, \varepsilon_{t+4} + \theta_1\varepsilon_{t+3} + \theta_2\varepsilon_{t+2} + \theta_3\varepsilon_{t+1}] = \theta_1\sigma^2$

18. $Cov\left[e^*_{t+2|t+1}, e^*_{t+4|t}\right] = Cov[\varepsilon_{t+2}, \varepsilon_{t+4} + \theta_1\varepsilon_{t+3} + \theta_2\varepsilon_{t+2} + \theta_3\varepsilon_{t+1}] = \theta_2\sigma^2$

19. $Cov\left[e^*_{t+1|t}, e^*_{t+4|t}\right] = Cov[\varepsilon_{t+1}, \varepsilon_{t+4} + \theta_1\varepsilon_{t+3} + \theta_2\varepsilon_{t+2} + \theta_3\varepsilon_{t+1}] = \theta_3\sigma^2$

20. $Cov\left[e^*_{t+4|t+2}, e^*_{t+3|t}\right] = Cov[\varepsilon_{t+4} + \theta_1\varepsilon_{t+3}, \varepsilon_{t+3} + \theta_1\varepsilon_{t+2} + \theta_2\varepsilon_{t+1}] = \theta_1\sigma^2$

21. $Cov\left[e^*_{t+3|t+1}, e^*_{t+3|t}\right] = Cov[\varepsilon_{t+3} + \theta_1\varepsilon_{t+2}, \varepsilon_{t+3} + \theta_1\varepsilon_{t+2} + \theta_2\varepsilon_{t+1}] = (1 + \theta_1^2)\sigma^2$

22. $Cov\left[e^*_{t+2|t}, e^*_{t+3|t}\right] = Cov[\varepsilon_{t+2} + \theta_1\varepsilon_{t+1}, \varepsilon_{t+3} + \theta_1\varepsilon_{t+2} + \theta_2\varepsilon_{t+1}] = \theta_1(1 + \theta_2)\sigma^2$

23. $Cov\left[e^*_{t+1|t-1}, e^*_{t+3|t}\right] = Cov[\varepsilon_{t+1} + \theta_1\varepsilon_t, \varepsilon_{t+3} + \theta_1\varepsilon_{t+2} + \theta_2\varepsilon_{t+1}] = \theta_2\sigma^2$

24. $Cov\left[e^*_{t+5|t+3}, e^*_{t+4|t}\right] = Cov[\varepsilon_{t+5} + \theta_1\varepsilon_{t+4}, \varepsilon_{t+4} + \theta_1\varepsilon_{t+3} + \theta_2\varepsilon_{t+2} + \theta_3\varepsilon_{t+1}] = \theta_1\sigma^2$

25. $Cov\left[e^*_{t+4|t+2}, e^*_{t+4|t}\right] = Cov[\varepsilon_{t+4} + \theta_1\varepsilon_{t+3}, \varepsilon_{t+4} + \theta_1\varepsilon_{t+3} + \theta_2\varepsilon_{t+2} + \theta_3\varepsilon_{t+1}] = (1 + \theta_1^2)\sigma^2$

26. $Cov\left[e^*_{t+3|t+1}, e^*_{t+4|t}\right] = Cov[\varepsilon_{t+3} + \theta_1\varepsilon_{t+2}, \varepsilon_{t+4} + \theta_1\varepsilon_{t+3} + \theta_2\varepsilon_{t+2} + \theta_3\varepsilon_{t+1}] = \theta_1(1 + \theta_2)\sigma^2$

27. $Cov\left[e^*_{t+2|t}, e^*_{t+4|t}\right] = Cov[\varepsilon_{t+2} + \theta_1\varepsilon_{t+1}, \varepsilon_{t+4} + \theta_1\varepsilon_{t+3} + \theta_2\varepsilon_{t+2} + \theta_3\varepsilon_{t+1}] = (\theta_2 + \theta_1\theta_3)\sigma^2$

28. $Cov\left[e^*_{t+1|t-1}, e^*_{t+4|t}\right] = Cov[\varepsilon_{t+1} + \theta_1\varepsilon_t, \varepsilon_{t+4} + \theta_1\varepsilon_{t+3} + \theta_2\varepsilon_{t+2} + \theta_3\varepsilon_{t+1}] = \theta_3\sigma^2$

29. $Cov\left[e^*_{t+6|t+3}, e^*_{t+4|t}\right] = Cov[\varepsilon_{t+6} + \theta_1\varepsilon_{t+5} + \theta_2\varepsilon_{t+4}, \varepsilon_{t+4} + \theta_1\varepsilon_{t+3} + \theta_2\varepsilon_{t+2} + \theta_3\varepsilon_{t+1}] = \theta_2\sigma^2$

30. $Cov\left[e^*_{t+5|t+2}, e^*_{t+4|t}\right] = Cov[\varepsilon_{t+5} + \theta_1\varepsilon_{t+4} + \theta_2\varepsilon_{t+3}, \varepsilon_{t+4} + \theta_1\varepsilon_{t+3} + \theta_2\varepsilon_{t+2} + \theta_3\varepsilon_{t+1}] = \theta_1(1 + \theta_2)\sigma^2$

31. $Cov\left[e^*_{t+4|t+1}, e^*_{t+4|t}\right] = Cov[\varepsilon_{t+4} + \theta_1\varepsilon_{t+3} + \theta_2\varepsilon_{t+2}, \varepsilon_{t+4} + \theta_1\varepsilon_{t+3} + \theta_2\varepsilon_{t+2} + \theta_3\varepsilon_{t+1}] =$

$(1 + \theta_1^2 + \theta_2^2)\sigma^2$

32. $Cov\left[e^*_{t+3|t}, e^*_{t+4|t}\right] = Cov[\varepsilon_{t+3} + \theta_1\varepsilon_{t+2} + \theta_2\varepsilon_{t+1}, \varepsilon_{t+4} + \theta_1\varepsilon_{t+3} + \theta_2\varepsilon_{t+2} + \theta_3\varepsilon_{t+1}] =$

$(\theta_1 + \theta_1\theta_2 + \theta_2\theta_3)\sigma^2$

33. $Cov\left[e^*_{t+2|t-1}, e^*_{t+4|t}\right] = Cov[\varepsilon_{t+2} + \theta_1\varepsilon_{t+1} + \theta_2\varepsilon_t, \varepsilon_{t+4} + \theta_1\varepsilon_{t+3} + \theta_2\varepsilon_{t+2} + \theta_3\varepsilon_{t+1}] =$

$(\theta_2 + \theta_1\theta_3)\sigma^2$

34. $Cov\left[e^*_{t+1|t-2}, e^*_{t+4|t}\right] = Cov[\varepsilon_{t+1} + \theta_1\varepsilon_t + \theta_2\varepsilon_{t-1}, \varepsilon_{t+4} + \theta_1\varepsilon_{t+3} + \theta_2\varepsilon_{t+2} + \theta_3\varepsilon_{t+1}] =$

$\theta_3\sigma^2$

References

Board of Governors of the Federal Reserve System: Industrial Production Index. (2018). Retrieved March 1, 2018, from https://fred.stlouisfed.org/series/INDPRO

Capistran, C. (2014). Optimality tests for multi-horizon forecasts. *Working Papers - Banco De Mexico*

Christodoulakis, G., Stathopoulos, K., & Tessaromatis, N. (2012). The term structure of loss preference and rationality in analyst earnings forecasts. *Journal of Asset Management, 13*(5), 310-326.

Clements, M. (1997). Evaluating the rationality of fixed-event forecasts. *Journal of Forecasting, 16*, 225-239.

Consumer Price Index - CPI. (2018). Retrieved March 1, 2018, from https://www.investopedia.com/terms/c/consumerpriceindex.asp

Elliot, G., Kumunjer, I., & Timmermann, A. (2005). Estimation and testing of forecast rationality under flexible loss. *The Review of Economic Studies, 72*(4), 1107-1125.

Gross Domestic Product - GDP. (2018). Retrieved March 1, 2018, from https://www.investopedia.com/terms/g/gdp.asp

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press.

Hansen, B. (2017). Econometrics. *University of Wisconsin,* 275-288.

Mincer, J., & Zarnowitz, V. (1969). The evaluation of economic forecasts. *Economic Forecasts and Expectations,* 3-46.

Nordhaus, W. (1987). Forecasting efficiency: Concepts and applications. *The Review of Economics and Statistics, 69*(4), 667-674.

Patton, A. (2013). Properties of Optimal Forecasts. Retrieved March 24, 2018, from http://www.oxford-man.ox.ac.uk/sites/default/files/events/Patton_OMI_lec1_14_white.pdf

Patton, A., & Timmermann, A. (2012). Forecast rationality tests based on multi-horizon bounds. *Journal of Business & Economic Statistics, 30*(1)

Patton, A., & Timmermann, A. (2007). Testing forecast optimality under unknown loss. *Journal of the American Statistical Association, 102*(480), 1172-1184.

Survey of Professional Forecasters (Historical SPF Forecast Data). (2018, February 9). Retrieved February 15, 2018, from https://www.philadelphiafed.org/research-and-data/real-time-center/survey-of-professional-forecasters

Treasury Bill - T-Bill. (2018). Retrieved March 1, 2018, from https://www.investopedia.com/terms/t/treasurybill.asp