UNIVERSITY OF CALIFORNIA
RIVERSIDE

Forecast Encompassing in High Dimensional Predictive Models

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Economics

by

Yaojue Xu

September 2023

Dissertation Committee:

Dr. Tae-Hwy Lee, Chairperson
Dr. Gloria Gonzalez-Rivera
Dr. Aman Ullah

The Dissertation of Yaojue Xu is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

First and foremost, I would like to express my profound gratitude to my advisor, Dr. Tae-Hwy Lee, whose unwavering support, invaluable guidance, and constant encouragement have been absolutely pivotal in my journey to completing this thesis and earning a Ph.D. degree in Economics. Under his expert guidance, I delved into the intricacies of econometrics, forecasting, and high-dimensional predictive models, and found myself not only growing in knowledge but in curiosity and determination. The complexities of these subjects were untangled with his patient explanations and insightful perspectives. His relentless passion for the subject became my inspiration, pushing me to strive for excellence and innovation in my research work.

Furthermore, I am deeply appreciative of my thesis committee members, Dr. Gloria Gonzalez-Rivera and Dr. Aman Ullah. Their insightful feedback and constructive criticism on my thesis have been of immense value.

Lastly, I express my deepest appreciation to my parents, Chi Xu and Wenjuan Sun. Their unflagging support, recognition of my achievements, and their resolute protection during my periods of doubt and stress have been my constant source of strength and resilience. They have believed in me, stood by my side, and encouraged me to reach for the stars. Their love and support have been my anchor and my guiding light, and for that, I am eternally grateful.

ABSTRACT OF THE DISSERTATION

Forecast Encompassing in High Dimensional Predictive Models

by

Yaojue Xu

Doctor of Philosophy, Graduate Program in Economics
University of California, Riverside, September 2023
Dr. Tae-Hwy Lee, Chairperson

This thesis advances the field of econometric forecasting and predictor selection by developing a series of novel methods based on the forecast-encompassing principle and the higher-order elicitability concept. These methods enhance predictive regression models by improving predictor selection in the context of various financial and macroeconomic scenarios. Chapter 2 introduces two methods for predictor selection based on out-of-sample forecast encompassing principle in high-dimensional predictive regression models. This principle is applied to forecast U.S. output growth, inflation, and the probability of U.S. recession, demonstrating a lower forecast error loss.

Chapter 3 addresses volatility forecast model comparison by proposing strictly consistent scoring functions based on the Bregman divergence. This new scoring function elicits the mean and variance without the zero mean assumption, facilitating more accurate predictive ability assessment. Chapter 4 delves into the International Monetary Fund's Growth-at-Risk (GaR) and Growth Shortfall (GS) measures, and proposes new tests for predictor selection in GS and GaR using the Fissler-Ziegel (FZ) scoring function.

Expectiles and their applications in financial risk measurement are explored in Chapter 5. The thesis proposes a framework for model selection and model averaging in predictive expectile regression models, providing a significant improvement over model selection. Chapter 6 constructs a new testing framework for Granger Causality in Expected Shortfalls, which previously didn't exist due to the lack of an objective function to evaluate Expected Shortfalls.

Chapter 7 proposes a Granger-Causality test in the predictive regression for the conditional mode and investigates the predictability of equity premium in mode regressions. Chapter 8 studies skewness and kurtosis in stock return distributions and introduces a novel Granger Causality test for these higher moments. The thesis thus significantly enhances the econometric toolbox for financial and macroeconomic forecasting and risk management.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In my thesis, I have established a comprehensive framework for testing Granger Causality across a spectrum of predictive measures within the conditional distribution. This includes, but is not limited to, conditional means, quantiles, expectiles, modes, partial moments, variances, densities, and durations. This framework is fundamentally based on the principle of forecast encompassing (ENC).

Key to the construction of the forecast encompassing test is the existence of a scoring function that is "strictly consistent" for a measure, be it mean, quantile, expectile, and others. A scoring function can be understood as "strictly consistent" if the measure in question is the unique minimizer of the expected score. Additionally, a measure is referred to as "elicitable" if a "strictly consistent" scoring function is found for it (Gneiting, 2011).

"Elicitability" holds vital importance in point forecasts, as it allows us to compare and rank various competing forecasts based on their expected score. As a result, we can choose models with "elicitable" measures using "strictly consistent" scoring functions. For

instance, if the expected score of a competing forecast A is lower than that of forecast B, we would favor A over B.

The encompassing principle possesses the capacity to handle a high-dimensional predictor vector. To this end, I propose two methodologies. The first is derived from the One Covariate at a Time Multiple Testing (OCMT) approach by Chudik, Kapetanios, and Pesaran (2018). The second is founded upon the Lasso-type regularization approach by Tibshirani (1996) and Zou (2006). Respectively, these methodologies are known as ENC-OCMT and ENC-LASSO.

My methodologies offer a broad range of compatibility. They can be applied to linear models, nonparametric additive models, and generalized linear models (GLM) for the conditional mean and more. Furthermore, they can address any measure beyond the conditional mean, like conditional quantiles (Value-at-Risk, Growth-at-Risk), conditional expectiles, and more. This broad applicability is possible because of the structure of ENC-OCMT and ENC-Lasso, which is based on the "generalized" forecast errors from any forecast models for any elicitable measures.

Chapter 2

Encompassing Approaches to Predictor Selection in High-dimensional Predictive Regression Models Using Multiple Testing and LASSO

2.1 Introduction

Forecast evaluation is crucial for selecting models and predictors, which are essential in economics and finance. The forecast encompassing approach with nested models is a widely used method for evaluating forecasts and selecting predictors. The null model

has no predictor, while the alternative model includes a predictor. Consequently, forecast encompassing tests with two nested models can be viewed as Granger causality tests, primarily determining whether a predictor has predictive ability for a variable of interest. Ashley et al. (1980) recommend using out-of-sample forecast comparisons instead of in-sample tests to assess Granger causality. The out-of-sample forecast encompassing approach divides the entire dataset into two parts: in-sample observations and out-of-sample observations. The test estimates parameters and constructs forecasts for the two models using in-sample observations, then evaluates the forecasts using out-of-sample observations. If the alternative model's forecast is more accurate than the null model, the predictor in the alternative model has predictive power for the variable of interest, indicating that the predictor Granger causes the variable of interest. However, if the forecasts in both models exhibit equal accuracy, the predictor does not possess predictive ability for the variable of interest, implying that the predictor does not Granger cause the variable of interest. To achieve accurate out-of-sample forecasts, the number of out-of-sample observations must increase significantly faster than the number of in-sample observations (Inoue et al., 2017). In this paper, we demonstrate the asymptotic normality of the forecast encompassing (ENC) statistic when the number of out-of-sample observations increases arbitrarily faster than the number of in-sample observations.

In macroeconomics, numerous macroeconomic variables exist, and most variables have quarterly data, so the number of variables available for testing predictive ability is much larger than the number of observations. Moreover, in most cases, the number of predictors with predictive ability for the variable of interest increases as the number of

variables used in the test increases. Thus, high-dimensional predictive regression models are widely utilized in various fields. This paper focuses on high-dimensional predictive regression models.

Predictor selection often considers only the marginal effect of a predictor. A particular predictor can be selected when the marginal effect is non-zero. However, this method may not identify all relevant predictors in the model. Correlations might exist among predictors, and the probability of correlations among predictors increases as the number of predictors increases. Some predictors with zero marginal effect on the variable of interest (forecast target) but correlated with other predictors may have an indirect impact on the variable of interest. Therefore, predictor selection should consider both the direct marginal effect and the indirect net impact, especially when there are numerous predictors. We extend the multiple testing approach called One Covariate at a Time Multiple Testing (OCMT) proposed by Chudik et al. (2018) for predictor selection. The OCMT approach focuses on variables' marginal effect and the net impact on the target variable for variable selection using total sample observations. To select predictors in high-dimensional predictive models, we propose a new encompassing approach incorporating the OCMT method, referred to as "ENC-OCMT." Our ENC-OCMT approach considers both marginal effect and net impact for predictor selection. The OCMT procedure selects variables using the in-sample test in the regression model, while our ENC-OCMT approach selects predictors using the out-of-sample test in the predictive regression model.

The OCMT approach by Chudik et al. (2018) can be applied exclusively to linear models for mean regression. Su et al. (2022) demonstrate that OCMT can be applied

to nonparametric additive models. Notably, our new approach, ENC-OCMT, can handle linear models, additive models, and generalized linear models. In this paper, we use three models to select predictors using our ENC-OCMT approach in high-dimensional predictive regression models. The three models are as follows:

Linear Predictive Model:

$$y_{t+1} = \sum_{i=1}^{n} \beta_i x_{i,t} + u_{t+1}, \tag{2.1}$$

Nonparametric Additive Predictive Model:

$$y_{t+1} = \sum_{i=1}^{n} \beta_i f_i\left(x_{i,t}\right) + u_{t+1}, \tag{2.2}$$

Probability Predictive Model:

$$\begin{aligned} &y_{t+1} \sim \text{Bernoulli}\left(\Pr(y_{t+1}=1|x_{i,t})\right), \\ &\Pr(y_{t+1}=1|x_{i,t}) = G\left(\sum_{i=1}^{n} \beta_i x_{i,t}\right) \equiv 1/\left[1+\exp\left(-\sum_{i=1}^{n} \beta_i x_{i,t}\right)\right], \end{aligned} \tag{2.3}$$

where $x_{i,t}$ are the n predictors, for $i = 1, \ldots, n$, with n being the number of predictors.

A widely used variable selection method is the least absolute shrinkage and selection operation (LASSO) by Tibshirani (1996). However, most variable selections using LASSO are based on full sample observations (Fan and Li, 2001; Efron et al., 2004; Zou and Hastie, 2005; Candes and Tao, 2007; Bickel et al., 2009; Zhang, 2010). We propose a new approach for selecting predictors using the LASSO method, based on the out-of-sample forecast encompassing principle. This new approach is called "ENC-LASSO." Compared to the LASSO method, our ENC-LASSO can be applied to out-of-sample predictor selection.

This paper is organized as follows. In Section 2, we discuss the forecast encompassing principle and prove the asymptotic normality of our ENC statistic. In Section 3,

we explain why we propose the ENC-OCMT and ENC-LASSO procedures and illustrate the procedures of ENC-OCMT and ENC-LASSO for out-of-sample predictive selection in high-dimensional linear models. Section 4 demonstrates that the ENC-OCMT and ENC-LASSO approaches can be applied to nonparametric additive predictive models. Section 5 discusses the ENC-OCMT and ENC-LASSO approaches in generalized linear models. We conduct the ENC-OCMT and ENC-LASSO to forecast the probability of a U.S. recession using the probability predictive model.

2.2 Forecast Encompassing Approach

In this section, we introduce the models, describe the encompassing test for the mean predictive regression, and demonstrate that the encompassing statistic used in this paper is asymptotically normal under certain assumptions.

2.2.1 Model

Chudik et al. (2018) proposed a multiple testing approach called One Covariate at a time Multiple Testing (OCMT) for variable selection. The OCMT procedure can be applied only to linear models. In this paper, we propose a new approach for the out-of-sample predictive ability test to select predictors, and our approach can be applied to both linear and nonlinear models. Since the new approach is based on the encompassing principle and OCMT procedure, we first develop the framework of the forecast encompassing approach. To select predictors in high-dimensional predictive regression, we consider the linear model in equation (2.1).

As the OCMT procedure is a multiple testing approach and selects predictors one at a time, we use low-dimensional predictive regression models for the forecast encompassing test to apply the OCMT procedure based on the encompassing principle. We have n + 1 linear regressions: one model without conditioning on any predictors, and the other models conditioned on predictors. We consider n different predictors x_i, $i = 1, \ldots, n$ for n models,, examining one predictor at a time. These nested mean models are

$$\text{Model } 0 : y_{t+1} = a_0 + u_{0,t+1} \tag{2.4}$$

$$\text{Model } i : y_{t+1} = a_i + b_i x_{i,t} + u_{i,t+1}, \quad i = 1, \ldots, n \tag{2.5}$$

where unconditional mean $\mu_{0,t+1} = a_0$ and conditional mean $\mu_{i,t+1} = a_i + b_i x_{i,t}$. The dependent variable y_{t+1} is a scalar. We estimate $\hat{\mu}_{0,t+1}$ and $\hat{\mu}_{i,t+1}$ by following $\hat{\mu}_{0,t+1} = \hat{a}_{0,t}$ and $\hat{\mu}_{i,t+1} = \hat{a}_{i,t} + \hat{b}_{i,t} x_{i,t}$, where $\hat{a}_{0,t}, \hat{a}_{i,t}$ and $\hat{b}_{i,t}$ are estimated based on in-sample observations using the ordinary least squares.

We use a large number of predictors $x_{i,t}$ in our regression models. The number of predictors is larger than the number of observations. In order to use the out-of-sample forecast encompassing approach for the multiple testing methods in high-dimensional models, we use one predictor $x_{i,t}$ at a time for Model i. We check if the predictor $x_{i,t}$ has the predictive ability for the dependent variable y_{t+1}. If the predictor $x_{i,t}$ has the predictive ability for the dependent variable y_{t+1}, we select this predictor and use this predictor in the regression to forecast y_{t+1}. We use nested regression models with one predictor $x_{i,t}$ at a time in the large models, so our out-of-sample forecast encompassing approach is an out-of-sample Granger Causality test. If the predictor $x_{i,t}$ has the predictive ability for the dependent variable y_{t+1}, this predictor Granger causes y_{t+1} in mean regression.

2.2.2 Out-of-sample forecast encompassing test

We now introduce the out-of-sample forecast encompassing test for the predictive ability of a predictor. The models in the equations (7.16) and (7.17) are nested. Due to the asymptotic non-normality of DM statistic for the nested models, we may not use the Diebold Mariano (DM) statistic by Diebold and Mariano (1995) to test the predictive ability of $x_{i,t}$ for y_{t+1}.

Thus, we develop the encompassing statistic and prove that the encompassing statistic is asymptotically normal. In order to find out the encompassing statistic for the mean, we combine two models (Model 0 and Model i) with weights $(1 - \lambda_i)$ and λ_i. We define the score function for the combined model i as a derivative of the expectation of loss function for the combined model with respect to the weight λ_i. The score function is the encompassing statistic we developed using the encompassing principle.

Based on the out-of-sample forecast encompassing principle, we combine Model 0 and Model i with weight $(1 - \lambda_i)$ and λ_i, respectively. Then, the null and alternative hypotheses are

$$\mathbb{H}_0 : \lambda_i = 0 \quad \text{and} \quad \mathbb{H}_1 : \lambda_i \neq 0.$$

Under the null hypothesis $\lambda_i = 0$, the combined model is Model 0, which means that $x_{i,t}$ does not Granger cause y_{t+1}. Under the alternative $\lambda_i \neq 0$, the combined model combines Model 0 and Model i with weights $(1 - \lambda_i)$ and λ_i, which implies that x_t Granger causes y_{t+1}. As mentioned in the previous subsection, we consider the out-of-sample encompassing test, so x_t has the predictive ability for y_{t+1} under the alternative hypothesis.

To find out the optimal weight λ_i and get the encompassing statistic under the null hypothesis, we estimate the weight λ_i minimizing the expectation of least squares loss function with combined mean $\mu_{c,t+1}$. Thus, we have

$$\lambda_i = \arg\min_{\lambda_i} \mathbb{E}L(\mu_{c,i,t+1}, y_{t+1}),$$

where $\mu_{c,i,t+1} = (1-\lambda_i)\,\mu_{0,t+1} + \lambda_i \mu_{i,t+1}$ and $L(\mu_{c,i,t+1}, y_{t+1}) = (y_{t+1} - \mu_{t+1})^2$ is least squares loss function. The encompassing principle is based on the score function with respect to λ_i. We define C_i as the score function that takes a derivative of the expectation of loss function for a combined model with respect to the weight λ. Thus, C_i is the score function for Model i, which is defined as

$$C_i \equiv \frac{\partial \mathbb{E}\left[L(\mu_{c,i,t+1}, y_{t+1})\right]}{\partial \lambda_i} = 0.$$

This score function can be separated into two parts. The first is a derivative of the loss function with respect to the conditional mean for the combined model $\mu_{c,i,t+1}$. Then we refer to the first derivative as the generalized residual function of the conditional mean for the combined model. We use $V_{i,t+1}$ to denote the generalized residual function for Model i. The second is a derivative of the conditional mean for the combined model $\mu_{c,i,t+1}$ with respect to weight λ_i. We call the second derivative to be the test function of the conditional mean. The test function is denoted as $\xi_{i,t+1}$ for the Model i. Thus, the score function C_i for Model i is the expectation of the product of identification function $V_{i,t+1}$ and test function $\xi_{i,t+1}$. Therefore, the score function C_i can be derived as follows:

$$C_i \equiv \frac{\partial \mathbb{E}\left[L(\mu_{c,i,t+1}, Y_{t+1})\right]}{\partial \lambda_i} = \mathbb{E}\left(\frac{\partial L(\mu_{c,i,t+1}, y_{t+1})}{\partial \mu_{t+1}^{(c)}}\right)\left(\frac{\partial \mu_{c,i,t+1}}{\partial \lambda_i}\right)$$

Let $V_{c,i,t+1} = \frac{\partial L(\mu_{c,i,t+1}, y_{t+1})}{\partial \mu_{c,i,t+1}}$ as the generalized residual and $\xi_{i,t+1} = \frac{\partial \mu_{c,i,t+1}}{\partial \lambda_i}$ as the test function. Note that $V_{c,i,t+1} = -2(y_{t+1} - \mu_{c,i,t+1})$ and $\xi_{i,t+1} = \mu_{i,t+1} - \mu_{0,t+1}$. If $\lambda_i = 0$, $\mu_{c,i,t+1} = \mu_{0,t+1}$ and $V_{c,i,t+1} = V_{0,t+1}$. Thus, under the null hypothesis, we have the score function C_i for Model i to be

$$C_i = \mathbb{E}\left[-2(y_{t+1} - \mu_{0,t+1})\left(\mu_{i,t+1} - \mu_{0,t+1}\right)\right] = \mathbb{E}\left(V_{0,t+1}\right)\left(\xi_{i,t+1}\right) = 0.$$

Based on the out-of-sample forecast encompassing principle, we separate the observations into two parts: in-sample observations and out-of-sample observations. The number of in-sample observations is defined as R, and the number of out-of-sample observations is defined as P. The total number of observations is $R + P = T + 1$. We estimate $\hat{a}_0$, $\hat{a}_i$ and $\hat{b}_i$ by using the in-sample observations from 1 to R. And then, we estimate $\hat{\mu}_{t+1} = \hat{a}_0$ and $\hat{\mu}_{i,t+1} = \hat{a}_i + \hat{b}_i x_{i,t}$ by using out-of-sample observations from $R + 1$ to $T + 1$. Then, we use one step ahead method to estimate C_i by $\hat{C}_{i,R,P}$ using the out-of-sample observations from $R + 1$ to $T + 1$. The $\hat{C}_{i,R,P}$ is defined as

$$\hat{C}_{i,R,P} \equiv P^{-1} \sum_{t=R}^{T} \left(\hat{V}_{0,t+1}\hat{\xi}_{i,t+1}\right) \xrightarrow{p} C_i = 0 \tag{2.6}$$

under $\mathbb{H}_0$, as $R, P \to \infty$ and $P/R \to \infty$. Under the null hypothesis, $\lambda_i = 0$, we obtain $\mu_{c,i,t+1} = \mu_{0,t+1}$ and $V_{c,i,t+1} = V_{0,t+1}$. Thus, $\mathbb{E}(\hat{C}_{i,R,P}) = 0$, so $\hat{C}_{i,R,P} \xrightarrow{p} 0$ as $R, P \to \infty$.

In order to get the encompassing statistic for the mean, we consider endowing Model 0 and Model i with the weight $1 - \lambda_i$ and λ_i, respectively, and find out the property of optimal weight λ_i. Theorem 1 gives the relationship between λ_i and C_i. This theorem shows that $\lambda_i = 0$ and $C_i = 0$ are equivalent.

Lemma 1: $\lambda_i = 0$ *if and only if* $C_i = 0$.

In order to test if $x_{i,t}$ has the predictive ability on y_{t+1}, we standardize the $\hat{C}_{i,R,P}$ to the ENC statistic. The ENC statistic is $\text{ENC}_{i,R,P} \equiv \hat{Q}_{i,R,P}^{-0.5}\sqrt{P}\hat{C}_{i,R,P}$, where $Q_{i,R,P} = \text{var}\left(\sqrt{P}\hat{C}_{i,R,P}\right)$ and $Q_{i,R,P} - \hat{Q}_{i,R,P} \xrightarrow{p} 0$.

2.2.3 Asymptotic normality of ENC

Assumptions:

1. $\{x_{i,t}\}$ are weakly stationary processes and $q_{i,t} = x_{i,t}x_{i,t}'$ is bounded for all t and $i = 1, 2$. We define $B_i = (\mathbb{E}q_{i,t})^{-1}$.

2. Let $\pi = \lim_{P,R\to\infty} P/R$. $\pi = \infty$.

In order to get accurate out-of-sample forecasts, selecting an optimal in-sample window is essential. Inoue et al. (2017) find the optimal window size by minimizing the conditional mean squared forecast error in a time-varying predictive regression model. They show that the optimal window size is $R = O(T^{2/3})$. In our setting, $R + P = T + 1$, so $P = O(T)$ when we choose the optimal in-sample window size pointed out by Inoue et al. (2017). Thus, P increases much faster than R. We want to find out the asymptotic normality of $\text{ENC}_{i,R,P}$ when $P/R = \infty$.

Theorem 1: *Under* $\mathbb{H}_0$*, Assumptions 1 and 2,* $\text{ENC}_{i,R,P}$ *is asymptotically standard normal.*

As discussed above, Assumption 2 is important for both theoretical and empirical aspects. We see that $\mathrm{ENC}_{i,R,P}$ follows a standard normal distribution for $\lim P/R \to \infty$, as the case in which $\lim P/R \to 0$ proved by Clark and McCracken (2001).

2.3 Predictor Selection in High-dimensional Linear Models

In most cases, the number of predictors that have the predictive ability on the variable of interest increases as the number of variables we used in the test increases. Thus, the high-dimensional regression model is highly used in many fields. This section proposes a new approach using the OCMT procedure under the encompassing principle. We refer to this approach as ENC-OCMT. This approach extends the OCMT procedure from full-sample variable selection to out-of-sample predictive selection and the ENC statistic from the low-dimensional models (one predictor in a model) to high-dimensional models (many predictors in the model). Thus, the ENC-OCMT can be used for the out-of-sample predictive ability test to select predictors in high-dimensional predictive models.

2.3.1 OCMT Procedure for Out-of-sample Forecast Encompassing Test

Chudik et al. (2018) proposed a new approach for variable selection, which is called OCMT. The OCMT procedure can be applied to only the linear models. In this section, we propose a new approach for the out-of-sample predictive ability test to select predictors, and our approach can be applied to both linear and nonlinear models. Since the new approach is based on the encompassing principle and OCMT procedure, so this approach is referred to as ENC-OCMT. The ENC-OCMT procedure focuses on both the

marginal effect and net impact of variables on the target variable, which is similar to the OCMT procedure. We consider the high-dimensional linear predictive model,

$$y_{t+1} = \sum_{i=1}^{n} \beta_i x_{i,t} + u_{t+1},$$

where $x_{1,t}, \ldots, x_{n,t}$ are the n predictors, for $i = 1, \ldots, n$, and u_{t+1} is an error term. When the number of predictors is small, we can select the predictors directly depending on the marginal effect coefficient β_i. When the marginal effect is not zero, the particular predictor $x_{i,t}$ can be selected. However, for time series data, there might exist correlations among predictors, and the probability of the correlations among predictors increases as the number of predictors in the model increases. Thus, considering the indirect impact among predictors is essential when n is large.

The ENC-OCMT approach attempt to select the predictors $x_{i,t}$ that have marginal effect β_i or net impact θ_i on y_{t+1}, which is different from other multiple testing procedures. The net impact coefficient θ_i from Pesaran and Smith (2014) catches the indirect impact of predictors that have zero marginal effect on the outcome variable but have a correlation with other predictors.

According to the encompassing principle, for the low-dimensional models, we combine 2 models in equations (7.16) and (7.17) with weights $(1 - \lambda_i)$ for Model 0 and λ_i for each Model i. There is no predictor in Model 0, while there is one predictor in each Model i, where $i = 1, \ldots, n$. We combine Model 0 and one Model i at a time to construct the combined model for Model i, so we have n combined models for different Model i. For the high-dimensional models, we combine these $n + 1$ models with weights λ_i for each Model i

and $(1 - \sum_{i=1}^{n} \lambda_i)$ for Model 0, so we get one combined model as follows:

$$\hat{\mu}_{c,t+1} = \left(1 - \sum_{i=1}^{n} \lambda_i\right) \hat{\mu}_{0,t+1} + \sum_{i=1}^{n} \lambda_i \hat{\mu}_{i,t+1}, \tag{2.7}$$

where $\hat{\mu}_{i,t+1}$ is the conditional mean for Model i that is estimated using the out-of-sample observations. Rearrange the equation above, the regression equation is

$$\hat{\mu}_{0,t+1} = \sum_{i=1}^{n} \lambda_i \left(\hat{\mu}_{0,t+1} - \hat{\mu}_{i,t+1}\right) + \hat{\mu}_{c,t+1}, . \tag{2.8}$$

We denote $\hat{u}_{i,t+1}$ as the forecast errors for Model i, so we can calculate $\hat{u}_{i,t+1} = y_{t+1} - \hat{\mu}_{i,t+1}$. Rewriting the equation above with $\hat{u}_{i,t+1}$, we have the regression equation

$$\hat{u}_{0,t+1} = \sum_{i=1}^{n} \lambda_i \left(\hat{u}_{0,t+1} - \hat{u}_{i,t+1}\right) + \hat{u}_{c,t+1}. \tag{2.9}$$

We use the same information to estimate different models, so the residuals of different models are highly correlated with each other. Thus, we should consider the marginal effect $\lambda_i \neq 0$ and net impact $\theta_i \neq 0$ to select predictors. The net impact θ_i for the out-of-sample forecast encompassing approach based on the regression (2.9) is described as

$$\theta_i = \sum_{j=1}^{n} \lambda_j \sigma_{i,j}, \tag{2.10}$$

where $\sigma_{i,j} = E\left(\hat{\boldsymbol{d}}_i' \hat{\boldsymbol{d}}_j\right)$, $\hat{\boldsymbol{d}}_i$ denotes the $P \times 1$ vector of $(\hat{u}_{0,t+1} - \hat{u}_{i,t+1})$ for $t = R, \ldots, T$.

Due to whether the predictors $x_{i,t}$ have marginal effect and net impact on y_{t+1}, there are following four possible cases:

	$\theta_i \neq 0$	$\theta_i = 0$
$\lambda_i \neq 0$	(1). Signals with nonzero net effect	(2). Hidden signals
$\lambda_i = 0$	(3). Pseudo-signals	(4). Noise variables

The first case $\lambda_i \neq 0$ and $\theta_i \neq 0$ implies that predictors, in this case, have both marginal and net impact, whereas predictors in the last case, where $\lambda_i = 0$ and $\theta_i = 0$, are called noise

variables since they have neither marginal effect nor net impact. The third case $\lambda_i = 0$ and $\theta_i \neq 0$ is possible and cannot be neglected. Predictors in the third case are referred to as Pseudo-signals. Obviously, the second case $\lambda_i \neq 0$ and $\theta_i = 0$ is almost impossible to occur when the number of predictors is large. The ENC-OCMT approach selects the predictors in Cases (1), (2), and (3). In other words, we do not select the predictors when they are Noise variables.

Similar to OCMT, the ENC-OCMT procedure is also a multi-stage procedure. The difference is that OCMT checks the covariance between $x_{i,t}$ and y_{t+1}, and ENC-OCMT tests the covariance between the residual $\hat{u}_{0,t+1}$ in Model 0 and the difference of residuals $(\hat{u}_{0,t+1} - \hat{u}_{i,t+1})$ in Model 0 and Model i. If there is covariance between the residual $\hat{u}_{0,t+1}$ in Model 0 and the difference of residuals $(\hat{u}_{0,t+1} - \hat{u}_{i,t+1})$ in Model i, we select the predictor $x_{i,t}$. In the first stage, we consider the n bivariate regressions of $\hat{u}_{0,t+1}$ on $(\hat{u}_{0,t+1} - \hat{u}_{i,t+1})$, for $i = 1, \ldots, n$, which is as follows:

$$\hat{u}_{0,t+1} = \phi_{i,(1)} \left(\hat{u}_{0,t+1} - \hat{u}_{i,t+1}\right) + \hat{u}_{c,i,t+1}, \text{ for } i = 1, \ldots, n, \tag{2.11}$$

where $\hat{u}_{c,i,t+1}$ is an error term of the regression for Model i. Thus, $\phi_{i,(1)}$ can be estimated by regressing $u_{t+1}^{(0)}$ on $u_{t+1}^{(0)} - u_{t+1}^{(i)}$.

In the first stage of ENC-OCMT, we regress $\hat{u}_{0,t+1}$ on $\hat{u}_{0,t+1} - \hat{u}_{i,t+1}$, for $i = 1, 2, \ldots, n$, one predictor at a time. Then, according to the definition of θ in equation (4.26), $\phi_{i,(1)} = \theta_i / \sigma_{ii}$. Denote $t_{\hat{\phi}_{i,(1)}}$ to be the t-ratio of $\hat{\phi}_{i,(1)}$, we have

$$t_{\hat{\phi}_{i,(1)}} = \frac{\hat{\phi}_{i,(1)}}{\text{s.e.}\left(\hat{\phi}_i\right)} = \frac{\hat{\boldsymbol{d}}_i' \hat{\boldsymbol{u}}}{\hat{\sigma}_i \sqrt{\hat{\boldsymbol{d}}_i' \hat{\boldsymbol{d}}_i}} \tag{2.12}$$

where $\hat{\phi}_{i,(1)} = \left(\hat{\boldsymbol{d}}_i'\hat{\boldsymbol{d}}_i\right)^{-1}\hat{\boldsymbol{d}}_i'\hat{\boldsymbol{u}}$ is estimated by minimizing the least squares loss function. $\hat{\boldsymbol{d}}_i$ denotes the $P \times 1$ vector of $(\hat{u}_{0,t+1} - \hat{u}_{i,t+1})$ and $\hat{\boldsymbol{u}}$ denotes the $P \times 1$ vector of $\hat{u}_{0,t+1}$ for $t = R, \ldots, T$. $\hat{\boldsymbol{e}}_{i,(1)}$ denotes the $P \times 1$ vector of residual of the regression of $\hat{\boldsymbol{u}}$ on $\hat{\boldsymbol{d}}_i$ in the first stage, and $\hat{\sigma}_i^2 = \hat{\boldsymbol{e}}_{i,(1)}'\hat{\boldsymbol{e}}_{i,(1)}/P$. The subscript (1) represents the first stage.

It is worth mentioning that we can use our ENC statistic as the t statistic of the regression. In section 2, we show that our ENC statistic for mean regression is the covariance between $\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{d}}_i$. Thus, the ENC statistic is proportional to the t-ratio of the regression. We also prove that the ENC statistic is asymptotically normal under the null hypothesis with no Granger Causality. Thus, we can use ENC statistic to replace t statistic in equation (4.29). Both ENC and t statistics can be used to select predictors compared to the critical value from a standard normal distribution.

This is a two-sided test with alternative hypothesis $\phi_i \neq 0$, so the predictor can be selected when the absolute value of the statistic is greater than the critical value. The first-stage ENC-OCMT selection indicator is given by

$$\widehat{\mathcal{J}}_{i,(1)} = I\left[\left|t_{\hat{\phi}_{i,(1)}}\right| > c_p(n,\delta)\right] \quad \text{for } i = 1, \ldots, n \tag{2.13}$$

where $c_p(n,\delta)$ is a critical value function defined by

$$c_p(n,\delta) = \Phi^{-1}\left(1 - \frac{p}{2f(n,\delta)}\right). \tag{2.14}$$

$\Phi^{-1}(\cdot)$ is the inverse of the standard normal distribution function, $f(n,\delta) = cn^{\delta}$ for some positive constants δ and c, and p $(0 < p < 1)$ is the nominal size of the individual tests to be set by the investigator. According to Chudik et al. (2018), the δ is defined as the critical value exponent. δ is used for the first stage, while δ^* is used in subsequent stages

of OCMT, which δ^* must be greater than δ. We select predictors in the first stage when $\widehat{\mathcal{J}}_{i,(1)} = 1$. According to Chudik et al. (2018), denote the number of predictors selected in the first stage by $\hat{k}_{(1)}$, the index set of the selected predictors in the first stage is denoted by $\hat{\mathcal{S}}_{(1)}$, and the active set is defined as $\mathcal{A}_{(1)} = \{1, \ldots, n\}$, where (1) represents the first stage.

In subsequent stage $j = 2, 3, \ldots$, we consider the $n - k_{(j-1)}$ regressions of $\hat{u}_{0,t+1}$ on the predictors in $\boldsymbol{D}_{(j-1)}$ and, one at a time, d_{it} for i belonging in the active set, $\mathcal{A}_{(j)}$,

$$u_{0,t+1} = \alpha_i + \phi_{i,(j)} \left(u_{0,t+1} - u_{i,t+1}\right) + \sum_{m=1}^{k_{(j-1)}} \gamma_m \left(u_{0,t+1} - u_{m,t+1}\right) + u_{c,t+1},$$

$$\text{for } i \in \mathcal{A}_{(j)}, \quad j = 2, 3, \ldots,$$

where the number of predictors selected in the stage $(j-1)$ is denoted by $\hat{k}_{(j-1)}$, $k_{(j-1)}$ and the number of all predictors selected up to and including stage $(j-1)$ is denoted $k_{(j-1)}$, so $k_{(j-1)} = \sum_{q=1}^{(j-1)} \hat{k}_{(j-1)}$. The matrix $\boldsymbol{D}_{(j-1)}$ denotes the $k_{(j-1)} \times 1$ vector of $(u_{0,t+1} - u_{m,t+1})$ for $t = R, \ldots, T$, and $u_{m,t+1}$ is the residual of the regression for the predictors has been selected in the previous stages, for $m = 1, \ldots, k_{(j-1)}$ and $t = R, \ldots, T$. We then compute the following t-ratios:

$$t_{\hat{\phi}_{i,(j)}} = \frac{\hat{\phi}_{i,(j)}}{\text{s.e.}\left(\hat{\phi}_{i,(j)}\right)} = \frac{\hat{\boldsymbol{d}}_i' \boldsymbol{M}_{(j-1)} \hat{\boldsymbol{u}}}{\hat{\sigma}_{i,(j)} \sqrt{\hat{\boldsymbol{d}}_i' \boldsymbol{M}_{(j-1)} \hat{\boldsymbol{d}}_i}} \quad \text{for } i \in \mathcal{A}_{(j)}, \quad j = 2, 3, \ldots,$$

where $\hat{\phi}_{i,(j)} = \left(\boldsymbol{d}_i' \boldsymbol{M}_{(j-1)} \boldsymbol{d}_i\right)^{-1} \boldsymbol{d}_i' \boldsymbol{M}_{(j-1)} \boldsymbol{u}$ is the least squared estimator of the conditional net effect of $\hat{d}_{it+1}$ on $\hat{u}_{0,t+1}$ in stage j. The estimated variance $\hat{\sigma}^2_{i,(j)} = \boldsymbol{e}_{i,(j)}' \boldsymbol{e}_{i,(j)} / P$, where $\boldsymbol{e}_{i,(j)}$ denotes the $P \times 1$ vector of residual of the regression of $\hat{\boldsymbol{u}}$ on $\hat{\boldsymbol{d}}_i$ and $\boldsymbol{D}_{(j-1)}$ in the stage j. The matrix $\boldsymbol{M}_{(j-1)} = \boldsymbol{I}_P - \boldsymbol{D}_{(j-1)} \left(\boldsymbol{D}_{(j-1)}' \boldsymbol{D}_{(j-1)}\right)^{-1} \boldsymbol{D}_{(j-1)}'$. the subscript (j) represents the stage j.

Predictors are then selected and added to the set of selections when $\widehat{\mathcal{J}}_{i,(j)} = 1$, where $\widehat{\mathcal{J}}_{i,(j)} = I\left[\left|t_{\hat{\phi}_{i,(j)}}\right| > c_p\left(n, \delta^*\right)\right]$. The index set of the selected predictors in stage j is denoted by $\hat{\mathcal{S}}_{(j)}$ and the index set in stage j of all the selected predictors up to and including stage j is denoted by $\mathcal{S}_{(j)}$, so $\mathcal{S}_{(j)} = \mathcal{S}_{(j-1)} \cup \hat{\mathcal{S}}_{(j)}$. The active set in the stage $j+1$ is defined as $\mathcal{A}_{(j+1)} = \{1, \ldots, n\} \backslash \mathcal{S}_{(j)}$. The procedure stops when there is no predictor can be selected in the current stage.

2.3.2 LASSO Procedure for Out-of-sample Forecast Encompassing Test

In this subsection, we propose a new approach for the out-of-sample forecast encompassing tests in the high-dimensional predictive regression models. This approach tests predictive ability using the LASSO under the encompassing principle. Thus, this approach is named ENC-LASSO. The ENC-LASSO procedure still shrinks some coefficients and makes others zero. The predictors with non-zero coefficients are selected through ENC-OCMT.

We use the same combination in the ENC-OCMT and ENC-LASSO. We combine all $n+1$ models, so we have the equation (2.9)

$$\hat{u}_{0,t+1} = \sum_{i=1}^{n} \lambda_i \left(\hat{u}_{0,t+1} - \hat{u}_{i,t+1}\right) + \hat{u}_{c,t+1}.$$

Based on the LASSO method, the ENC-LASSO estimate $\hat{\boldsymbol{\lambda}}_{\text{LASSO}}$ is defined as

$$\hat{\boldsymbol{\lambda}}_{\text{LASSO}} = \arg\min_{\boldsymbol{\lambda}} \sum_{t=R}^{T} \hat{u}_{c,t+1}^2 + \kappa \sum_{i=1}^{n} \left|\lambda_i\right|, \tag{2.15}$$

where $\hat{\boldsymbol{\lambda}}_{\text{LASSO}} = (\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_n)'$ is a vector of estimates λ_i for Model i, κ is a non-negative tuning parameter and the second term is the penalty term.

However, since the LASSO method produces biased through shrinking when the number of variables is large (Fan and Li, 2001), Zou (2006) proposes a weighted LASSO pro-

cedure adding weights to different coefficients in the penalty term, which is called Adaptive LASSO. Zou proves that the Adaptive LASSO with data-dependent weights is consistent. Thus, this paper also tests the predictive ability of predictors using the Adaptive LASSO under the encompassing principle. This procedure is called ENC-ALASSO. Similar to the Adaptive LASSO, we add weights $\hat{w}_i$ to the penalty term in the equation (3.22). The ENC-ALASSO estimate $\hat{\boldsymbol{\lambda}_{\text{ALASSO}}}$ is defined as

$$\hat{\boldsymbol{\lambda}}_{\text{ALASSO}} = \arg\min_{\boldsymbol{\lambda}} \sum_{t=R}^{T} \hat{u}_{c,t+1}^2 + \kappa \sum_{i=1}^{n} \hat{w}_i \left|\lambda_i\right|,$$

where $\hat{\boldsymbol{\lambda}}_{\text{ALASSO}} = (\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_n)'$ is a vector of estimates λ_i for Model i, $\hat{w}_i = 1/|\hat{\lambda}_{i,\text{OLS}}|$, and $\hat{\lambda}_{i,\text{OLS}}$ are estimated minimizing least squares loss function in equation (2.9).

2.3.3 Monte Carlo Simulation

In this subsection, following Chudik et al. (2018) and Su et al. (2022), we use ten different data-generating processes (DGPs), which always include signals and noise variables but with and without pseudo-signals and hidden signals. These simulations are used to confirm that our approaches work perfectly in the finite sample size. The sample combinations $n = (100, 200, 300)$. In our simulation, we set the number of in-sample observations $R = 100$ and the number of out-of-sample observations $P = (200, 300, 500)$. We use rolling windows to estimate parameters by OLS. All experiments are carried out using $2,000$ replications.

The design of DGPs 1-5 follows the setting from Chudik et al. (2018). For all five DGPs, y_{t+1} is generated as

$$y_{t+1} = \beta_1 x_{1,t} + \beta_2 x_{2,t} + \beta_3 x_{3,t} + \beta_4 x_{4,t} + s u_t,$$

where $u_t \sim$ i.i.d. N$(0,1)$ for $t = 1, 2, \ldots, T$. The parameter s is set to make three different goodness of fit $R^2 = 30\%, 50\%$, or 70% (on average). The three R^2 measures correspond to the three choices of s as a low, medium, and high fit.

DGP 1 (No Hidden Signals and No Pseudo-Signals):

There are four settings in DGP 1, which includes signals and noise variables but no hidden signals or pseudo-signals in all four settings. We set $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 1$ in these four settings of this DGP.

1(a). In this setting, the predictors $x_{i,t} = (\varepsilon_{i,t} + \nu g_t) / \sqrt{1+\nu}$, for $i = 1, 2, 3, 4$, are signal variables and the predictors $x_{5,t} = \varepsilon_{5,t}, x_{i,t} = (\varepsilon_{i-1,t} + \varepsilon_{i,t}) / \sqrt{2}$, for $i > 5$, are noise variables, where g_t and ε_{it} are independent draws from N$(0,1)$, for $t = 1, 2, \ldots, T$, and $i = 1, 2, \ldots, n$. We set $\nu = 1$, which implies 50% pairwise correlation among the signal variables.

1(b). In this setting, the predictors $x_{i,t}$ are generated as in DGP 1(a), but with a different setting of $\varepsilon_{i,t}$. The predictors $x_{i,t} = (\varepsilon_{i,t} + g_t) / \sqrt{2}$, for $i = 1, 2, 3, 4$, are signal variables and the predictors $x_{5,t} = \varepsilon_{5,t}, x_{it} = (\varepsilon_{i-1,t} + \varepsilon_{i,t}) / \sqrt{2}$, for $i > 5$, are noise variables, where g_t independently draws from N$(0,1)$, for $t = 1, 2, \ldots, T$, and $i = 1, 2, \ldots, n$, and $\varepsilon_{it} = 0.5\varepsilon_{i,t-1} + \sqrt{0.75} e_{i,t}$ with $e_{i,t} \sim i.i.d.$ N$(0,1)$.

1(c). In this setting, the predictors $x_{i,t} = (\varepsilon_{i,t} + g_t) / \sqrt{2}$, for $i = 1, 2, 3, 4$ are signal variables and the predictors $x_{5,t} = (\varepsilon_{5,t} + b_i f_t) / \sqrt{3}$ and $x_{i,t} = \left[(\varepsilon_{i-1,t} + \varepsilon_{i,t}) / \sqrt{2} + b_i f_t \right] / \sqrt{3}$, for $i > 5$ are noise variables, where $b_i \sim$ i.i.d. N$(1,1)$, $f_t = 0.95 f_{t-1} + \sqrt{1 - 0.95^2} v_t$, and v_t, g_t, and $\varepsilon_{i,t}$ are independent draws from $N(0,1)$.

1(d). In this setting, the predictors $x_{i,t}$ are generated as in DGP 1(a), but with a different setting of ν. The predictors $x_{i,t} = (\varepsilon_{it} + \nu g_t)/\sqrt{1+\nu}$, for $i = 1, 2, 3, 4$, are signal variables and the predictors $x_{5,t} = \varepsilon_{5t}$, $x_{i,t} = (\varepsilon_{i-1,t} + \varepsilon_{i,t})/\sqrt{2}$, for $i > 5$, are noise variables, where g_t and $\varepsilon_{i,t}$ are independent draws from $\mathrm{N}(0,1)$, for $t = 1, 2, \ldots, T$, and $i = 1, 2, \ldots, n$. We set $\nu = \sqrt{\omega/(1-\omega)}$ with two values of ω. Signals have low pairwise correlation when $\omega = 0.2$, while signals have high pairwise correlation when $\omega = 0.8$.

DGP 2 (Pseudo-Signals):

There are two settings in DGP 2, which includes signals, pseudo-signals, and noise variables but no hidden signals in these two settings. We set $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 1$ in these two settings of this DGP.

2(a). In this setting, the predictors $x_{i,t} = (\varepsilon_{i,t} + g_t)/\sqrt{2}$, for $i = 1, 2, 3, 4$, are signal variables, $x_{5t} = \varepsilon_{5t} + \kappa x_{1,t}$, $x_{6,t} = \varepsilon_{6,t} + \kappa x_{2,t}$ are pseudo-signal variables, and $x_{i,t} = (\varepsilon_{i-1,t} + \varepsilon_{i,t})/\sqrt{2}$, for $i > 6$ are noise variables, where g_t and $\varepsilon_{i,t}$ are independent draws from $\mathrm{N}(0,1)$. We set $\kappa = 1.33$ to achieve 80% correlation between the signals $x_{1,t}$ and the pseudo-signal variables $x_{5,t}$ and between the signals $x_{2,t}$ and the pseudo-signal variables $x_{6,t}$.

2(b). In this setting, the predictor matrix $\boldsymbol{x}_{n,t} \sim$ i.i.d. $(\mathbf{0}, \boldsymbol{\Sigma}_x)$ with the elements of variance-covariance matrix $\boldsymbol{\Sigma}_x$ given by $0.5^{|i-j|}$, $1 \leq i, j \leq n$. We define $\boldsymbol{x}_{n,t} = \boldsymbol{\Sigma}_x^{1/2} \boldsymbol{\varepsilon}_t$, where $\boldsymbol{\varepsilon}_t = (\varepsilon_{1,t}, \ldots, \varepsilon_{n,t})'$, and $\varepsilon_{i,t} \sim i.i.d.\mathrm{N}(0,1)$. All predictors in the predictor matrix have correlations with each other. Thus, $x_{i,t}$, for $i = 1, 2, 3, 4$, are signals. We consider $x_{i,t}$, for $5 \leq i \leq 19$, to be pseudo-signals, and $x_{i,t}$, for $i \geq 20$, to be noise variables.

DGP 3 (Hidden Signals):

DGP 3 includes signals, hidden signals, and noise variables but no pseudo-signals. We set $\beta_1 = \beta_2 = \beta_3 = 1$ and $\beta_4 = -1.5$. This implies $\theta_i \neq 0$ for $i = 1, 2, 3$ and $\theta_i = 0$ for $i \geq 4$. $\boldsymbol{x}_{n,t}$ is generated as in DGP 1(a). The predictor matrix $\boldsymbol{x}_{n,t} \sim i.i.d.\,(\mathbf{0}, \boldsymbol{\Sigma}_x)$ with the elements of variance-covariance matrix $\boldsymbol{\Sigma}_x$ given by $0.5^{|i-j|}$, $1 \leq i,\, j \leq n$. We define $\boldsymbol{x}_{n,t} = \boldsymbol{\Sigma}_x^{1/2} \boldsymbol{\varepsilon}_t$, where $\boldsymbol{\varepsilon}_t = (\varepsilon_{1,t}, \ldots, \varepsilon_{n,t})'$, and $\varepsilon_{i,t} \sim i.i.d.\mathrm{N}(0, 1)$.

DGP 4 (Both Hidden Signals and Pseudo-Signals):

There are two settings in DGP 4, which includes all four types of predictors: signals hidden signals, pseudo-signals, and noise variables.

4(a). We generate $\boldsymbol{x}_{n,t}$ in the same way as in DGP 2(a), which features two pseudo-signal variables. We set $\beta_1 = \beta_2 = \beta_3 = 1$ and $\beta_4 = -1.5$. Thus, we have $\theta_i \neq 0$ for $i = 1, 2, 3, 5, 6$, and $\theta_i = 0$ for $i = 4$ and $i \geq 6$. The predictors $x_{1,t}$, $x_{2,t}$ and $x_{3,t}$ are signals, $x_{4,t}$ is hidden variables, $x_{5,t}$ and $x_{6,t}$ are pseudo-signals, and $x_{i,t}$ for $i \geq 7$ are noise variables.

4(b). We generate $\boldsymbol{x}_{n,t}$ in the same way as in DGP 2(b), where all predictors are collinear with signals. We set $\beta_1 = -0.875$ and $\beta_2 = \beta_3 = \beta_4 = 1$. This implies $\theta_i = 0$ for $i = 1$ and $\theta_i > 0$ for all $i > 1$. Thus, $x_{1,t}$ is hidden variable, $x_{2,t}$, $x_{3,t}$, and $x_{4,t}$ are signals. We consider $x_{i,t}$ for $5 \leq i \leq 19$ to be pseudo-signals, and $x_{i,t}$ for $i \geq 20$ to be noise variables.

DGP 5 (Many Signals):

For this DGP, y_{t+1} is generated as

$$y_{t+1} = \sum_{i=1}^{n} i^{-2} x_{i,t} + s u_t,$$

where $\boldsymbol{x}_{n,t}$ are generated as in design DGP 2(b), and $u_t \sim$ i.i.d. $\mathrm{N}(0,1)$. The predictor matrix $\boldsymbol{x}_{n,t} \sim$ i.i.d. $(\mathbf{0}, \boldsymbol{\Sigma}_x)$ with the elements of variance-covariance matrix $\boldsymbol{\Sigma}_x$ given by $0.5^{|i-j|}$, $1 \leq i,\, j \leq n$. We define $\boldsymbol{x}_{n,t} = \boldsymbol{\Sigma}_x^{1/2} \boldsymbol{\varepsilon}_t$, where $\boldsymbol{\varepsilon}_t = (\varepsilon_{1,t}, \ldots, \varepsilon_{n,t})'$, and $\varepsilon_{i,t} \sim$ i.i.d. $\mathrm{N}(0,1)$. We consider the predictors $x_{i,t}$ for $i = 1, \ldots, 11$ to be signals and other predictors to be noise variables.

We focus on three out-of-sample forecast selection methods for the simulation results: ENC-OCMT, ENC-LASSO, and ENC-ALASSO. In our forecast evaluations, we use the most used criteria in the penalized regression methods, which are the following: the true positive rate (TPR) defined by (4.6.1), the false positive rate (FPR) defined by (4.6.1), the false discovery of the approximating model (FDR) defined by (4.6.1) and the false discovery rate of the true model (FDR*) denoted by (4.6.1).

True positive rate:

$$\mathrm{TPR}_{n,T} = \frac{\sum_{i=1}^{n} I\left(\widehat{\mathcal{J}}_i = 1 \text{ and } \beta_i \neq 0\right)}{\sum_{i=1}^{n} I\left(\beta_i \neq 0\right)} \xrightarrow{p} 1.$$

False positive rate:

$$\mathrm{FPR}_{n,T} = \frac{\sum_{i=1}^{n} I\left(\widehat{\mathcal{J}}_i = 1 \text{ and } \beta_i = 0\right)}{\sum_{i=1}^{n} I\left(\beta_i = 0\right)} \xrightarrow{p} 0.$$

False discovery rate:

$$\mathrm{FDR}_{n,T} = \frac{\sum_{i=1}^{n} I\left(\widehat{\mathcal{J}}_i = 1, \text{ and } \beta_i = \theta_i = 0\right)}{\sum_{i=1}^{n} \widehat{\mathcal{J}}_i + 1} \xrightarrow{p} 0.$$

The false discovery rate of the true model:

$$\text{FDR}^*_{n,T} = \frac{\sum_{i=1}^n I\left(\widehat{\mathcal{J}}_i = 1 \text{ and } \beta_i = 0\right)}{\sum_{i=1}^n \widehat{\mathcal{J}}_i + 1} \xrightarrow{p} c.$$

The asymptotic theory of OCMT carries over for ENC-OCMT and ENC-LASSO in terms of the true and false positive rates, and false discovery rate of the approximate and true model (TPR, FPR, FDR, and FDR*). We also report the root of mean squared forecast error (RMSFE) for three out-of-sample forecast selection methods.

We have five DGPs for the Monte Carlo simulation. DGPs 1-5 are linear models from Chudik et al. (2018). We set the sample size n to be 100, 200, and 300, the number of in-sample observations R to be 100, and the number of out-of-sample observations to be 200, 300, and 500 for all ten DGPs. For ENC-OCMT method, we set nominal level $p = 0.1, 0.05, 0.01$, $\delta = 1$ in the first stage, and $\delta^* \in \{1.5, 2\}$ in the subsequent stages.

Different settings lead to the different performance of methods. We report DGP 1(a) in this paper and move other simulation results to Supplemental Appendix. These results show that ENC-OCMT outperforms ENC-LASSO and ENC-ALASSO in most experiments. ENC-OCMT gets the largest TPR, lowest FPR, FDR, FDR* and lowest RMSFE in most experiments among the three methods, which implies that ENC-OCMT can successfully select signals, Hidden Signals, and Pseudo-signals. However, ENC-LASSO has a larger TPR than ENC-OCMT in some experiments, and ENC-ALASSO has a lower FDR and FDR* compared to ENC-LASSO. This means that ENC-LASSO tends to overestimate the number of signals. Lasso also has the same problem in the literature. Overall, no method universally dominates other methods.

2.3.4 Applications to Forecasting Output Growth and Inflation

In this subsection, in order to show the usefulness of ENC-LASSO, ENC-ALASSO, ENC-OCMT, and POST-ENC-OCMT, we use quarterly data from Stock and Watson (2012) to present a macroeconomic forecasting application for U.S. GDP growth and CPI inflation using a large set of macroeconomic variables. The data set contains 109 series from 1960Q3 to 2008Q4. The U.S. GDP growth or differenced log CPI inflation is the target variable. The in-sample periods are from 1960Q3 to 1980Q2 (80 observations) and from 1960Q3 to 1990Q2 (120 observations). The corresponding out-of-sample periods are from 1980Q3 to 2008Q4 (114 observations) and from 1990Q3 to 2008Q4 (74 observations). In this paper, we report 120 in-sample observations and 74 out-of-sample observations and move other empirical results to Supplemental Appendix.

We produce one-step-ahead forecasts using four different procedures: (a), (b) ENC-LASSO and ENC-ALASSO regressions of the target variable. (c), (d) ENC-OCMT and POST-ENC-OCMT procedures applied to regressions of the dependent variable. We select δ using 10-fold cross validation in the first stage of OCMT and ENC-OCMT and $\delta^* = 2\delta$ in the subsequent stages for (d). We consider $p = 0.05$ and generate forecasts using the rolling window.

Table 4 presents the results of forecasting real GDP growth and inflation using various predictive mean regression models, including ENC-LASSO, ENC-ALASSO, ENC-OCMT, and POST-ENC-OCMT. The models are evaluated based on their Mean Squared Forecasting Errors (MSFE) for 74 out-of-sample observations with a rolling window estimation of 120 in-sample observations. Model 0 represents the AR benchmark with factors,

Model S is the AR model with factors and all selected variables using these four methods. The numbers in the parentheses represent how many variables selected using the method. The reported values for MSFE are multiplied by 10^5.

For real output growth, the AR benchmark model inclusive of factors exhibits an MSFE of 3.2732. However, with the integration of ten selected variables through the ENC-LASSO method, the MSFE escalates to 3.8805. This might suggest that the introduction of additional variables does not necessarily enhance the forecasting accuracy of real output growth. Utilizing the ENC-Adaptive LASSO method for variable selection, the incorporation of a single selected variable reduces the MSFE to 3.4736 in the revised model. Intriguingly, no variables were chosen via the ENC-OCMT and POST-ENC-OCMT procedures, resulting in MSFEs equivalent to the AR benchmark with factors, in the absence of any additional variables.

In a similar way, for inflation forecasts, the AR benchmark model inclusive of factors indicates an MSFE of 3.6000. However, when the model includes three variables chosen via the ENC-LASSO method, the MSFE marginally increases to 3.9007. This trend suggests again that integrating additional selected variables may not bolster prediction accuracy. With the inclusion of a single variable chosen using the ENC-Adaptive LASSO method, the MSFE descends to 3.5245, which is the lowest MSFE observed in our study. Interestingly, only one variable is identified when applying the POST-ENC-OCMT method. This variable is distinct from the one chosen via the ENC-Adaptive LASSO method. The MSFE of the model incorporating this POST-ENC-OCMT selected variable is slightly higher at 3.5576, compared to the one selected by the ENC-Adaptive LASSO.

2.4 Predictor Selection in High-dimensional Nonparametric Additive Models

The OCMT procedure can be used for the linear models in mean regression (Chudik et al., 2018). Su et al. (2022) extends the OCMT procedure from linear models to nonparametric additive models. This section shows that our two approaches can also be applied to nonparametric additive models.

2.4.1 Predictive Additive Model

In the previous section, we test the predictive ability in the linear predictive model. In this section, we show hoe to check the predictive ability of predictors $x_{i,t}$ in the predictive additive model. According to Su et al. (2022), we have the non parametric additive model in equation (2.2)

$$y_{t+1} = \sum_{i=1}^{n} \beta_i f_i \left(x_{i,t} \right) + u_{t+1}$$

where y_{t+1} is the dependent variable, $x_{i,t}$ are n predictors, for $i = 1, \ldots, n$, f_i are smooth functions, for $i = 1, \ldots, n$.

2.4.2 Out-of-sample Forecast Encompassing Approach

We test the predictive ability of predictors $x_{i,t}$ on the dependent variable y_{t+1} using the out-of-sample forecast encompassing test. Similar to the regression models in

section 2, we still have $n+1$ nested mean regressions, $i = 1, \cdots, n$.

$$\text{Model } 0 : y_{t+1} = a_0 + u_{0,t+1} \equiv \boldsymbol{f}_0(\boldsymbol{x}_{0,t})'\boldsymbol{\beta}_0 + u_{0,t+1}, \tag{2.16}$$

$$\text{Model } i : y_{t+1} = a_i + b_i f_i(x_{i,t}) + u_{i,t+1} \equiv \boldsymbol{f}_i(\boldsymbol{x}_{i,t})'\boldsymbol{\beta}_i + u_{i,t+1}, \quad i = 1, \ldots, n \tag{2.17}$$

where $\mu_{0,t+1} = \boldsymbol{f}_0(\boldsymbol{x}_{0,t})'\boldsymbol{\beta}_0$ and $\mu_{i,t+1} = \boldsymbol{f}_i(\boldsymbol{x}_{i,t})'\boldsymbol{\beta}_i$ are the unconditional mean and conditional mean of the predictive additive models for Model 0 and Model i respectively. $\boldsymbol{x}_{0,t}$ is a strict subset of $\boldsymbol{x}_{i,t}$. $\boldsymbol{f}_0(\boldsymbol{x}_{0,t})$ is a strict subset of $\boldsymbol{f}_i(\boldsymbol{x}_{i,t})$. $\boldsymbol{x}'_{0,t} = 1, \boldsymbol{x}'_{i,t} = (1, x_{i,t})$, $\boldsymbol{f}_0(\boldsymbol{x}_{0,t})' = 1$, $\boldsymbol{f}_i(\boldsymbol{x}_{i,t})' = (1, f_i(x_{i,t}))$, $\beta_0 = a_0, \beta_i = (a_i, b_i)'$, for $i = 1, \ldots, n$. Thus, we estimate $\hat{\mu}_{0,t+1}$ and $\hat{\mu}_{i,t+1}$ by following $\hat{\mu}_{0,t+1} = \hat{a}_{0,t}$ and $\hat{\mu}_{i,t+1} = \hat{a}_{i,t} - \hat{b}_{i,t} f_i(x_{i,t})$, where $\hat{a}_{0,t}, \hat{a}_{i,t}$ and $\hat{b}_{i,t}$ are estimated minimizing the least squares loss function $L = (y_{t+1} - \mu_{i,t+1})^2 = (u_{i,t+1})^2$. There is no predictor in Model 0 and one predictor in Model i. Under the encompassing principle, we combine Model 0 and Model i with weight $1 - \lambda_i$ and λ_i, respectively, then we get the conditional mean $\hat{\mu}_{c,i,t+1}$ for combined model i, $i = 1, \cdots, n$.

$$\hat{\mu}_{c,i,t+1} = (1 - \lambda_i)\, \hat{\mu}_{0,t+1} + \lambda_i \hat{\mu}_{i,t+1} \tag{2.18}$$

Since we use different weights λ_i to combine Model 0 and different Model i, there is a conditional mean $\hat{\mu}_{c,i,t+1}$ for each combined model i.

We know that the null and alternative hypothesis for the encompassing test is $\mathbb{H}_0 : \lambda_i = 0$, and $\mathbb{H}_1 : \lambda_i \neq 0$. Under the null hypothesis $\lambda_i = 0$, the combined model is Model 0, so that $\mu_{c,i,t+1} = \mu_{0,t+1}$, which means that $x_{i,t}$ does not have the predictive ability on y_{t+1} in mean. Under the alternative $\lambda_i \neq 0$, the combined model combines Model 0 and Model i with weights $1 - \lambda_i$ and λ_i, which implies that $x_{i,t}$ has the predictive ability on y_{t+1} in mean.

In order to construct the encompassing statistic under the null hypothesis, we estimate the weight λ_i using least squared estimation with combined mean $\mu_{c,i,t+1}$. The encompassing principle is based on the score function with respect to λ_i. We define C_i as the score function that takes a derivative of the loss function of the combined model with respect to the weight λ. Thus, C_i is the score function for Model i, which is defined as

$$C_i \equiv \frac{\partial \mathbb{E}L(\mu_{c,i,t+1}, y_{t+1})}{\partial \lambda_i} = 0. \tag{2.19}$$

This score function can be separated into two parts. The first is a derivative of the loss function with respect to the conditional mean for the combined model $\mu_{c,i,t+1}$. Then we define the first derivative as the identification function of the conditional mean for the combined model. We denote $V_{i,t+1}$ as the generalized residual function for Model i. The second is a derivative of the conditional mean for the combined model $\mu_{c,i,t+1}$ with respect to weight λ_i. We call the second derivative to be the test function of the conditional mean. We denote the test function as $\xi_{i,t+1}$ for the Model i. Thus, the score function C_i for Model i is the expectation of the product of identification function $V_{i,t+1}$ and test function $\xi_{i,t+1}$. Therefore, the score function C_i can be derived as follows:

$$\begin{aligned} C_i &\equiv \frac{\partial \mathbb{E}L(\mu_{c,i,t+1}, y_{t+1})}{\partial \lambda_i} \\ &= \mathbb{E}\left(\frac{\partial L(\mu_{c,i,t+1}, y_{t+1})}{\partial \mu_{c,i,t+1}}\right)\left(\frac{\partial \mu_{c,i,t+1}}{\partial \lambda_i}\right) \\ &= \mathbb{E}\left[-2\left(y_{t+1} - \mu_{c,i,t+1}\right)\left(\mu_{i,t+1} - \mu_{0,t+1}\right)\right] \\ &\equiv \mathbb{E}\left(V_{c,i,t+1}\right)\left(\xi_{i,t+1}\right) = 0 \end{aligned} \tag{2.20}$$

where the loss function for the combined model is the least squared loss function $L\left(\mu_{c,i,t+1}, y_{t+1}\right) = \left(y_{t+1} - \mu_{c,i,t+1}\right)^2$ with conditional mean for the combined model $\mu_{c,i,t+1}$. The generalized residual function $V_{c,i,t+1} = \frac{\partial L(\mu_{c,i,t+1}, y_{t+1})}{\partial \mu_{c,i,t+1}} = \mu_{c,i,t+1}$, and the test function $\xi_{i,t+1} =$

$\frac{\partial \mu_{c,i,t+1}}{\partial \lambda_i} \equiv (\mu_{i,t+1} - \mu_{0,t+1})$. Under the null hypothesis $\lambda_i = 0$, since the conditional mean for combined model i is equal to that for Model 0, $\mu_{c,i,t+1} = \mu_{0,t+1}$ and $V_{c,i,t+1} = V_{0,t+1}$. Thus, under the null hypothesis, we have the score function C_i for Model i to be

$$C_i = \mathbb{E}\left[-2\left(y_{t+1} - \mu_{0,t+1}\right)\left(\mu_{i,t+1} - \mu_{0,t+1}\right)\right] = \mathbb{E}\left(V_{0,t+1}\right)\left(\xi_{i,t+1}\right) = 0.$$

We want to determine the predictive ability of predictor $x_{i,t}$ on y_{t+1} using an out-of-sample forecast encompassing test, so we separate the observations into in-sample observations and out-of-sample observations. We denote R to be the number of in-sample observations and P to be the number of out-of-sample observations. We estimate $\hat{a}_0$, $\hat{a}_i$ and $\hat{b}_i$ by using the in-sample observations from 1 to R. And then, we estimate $\hat{\mu}_{0,t+1} = \hat{a}_0$) and $\hat{\mu}_{i,t+1} = \hat{a}_i + \hat{b}_i f_i(x_{i,t})$ by using out-of-sample observations from $R+1$ to $T+1$. After that, we also use the out-of-sample observations to estimate the score function C_i by $\hat{C}_{i,R,P}$. The $\hat{C}_{i,R,P}$ is defined as

$$\hat{C}_{i,R,P} \equiv P^{-1} \sum_{t=R}^{T} \left(\hat{V}_{0,t+1}\hat{\xi}_{i,t+1}\right) \xrightarrow{p} C_i = 0$$

under $\mathbb{H}_0$, as $R, P \to \infty$. Thus, $\mathbb{E}\left(\hat{C}_{i,R,P}\right) = 0$, so $\hat{C}_{i,R,P} \xrightarrow{p} 0$ as $R, P \to \infty$. The ENC statistic is denoted as normalization of $\hat{C}_{i,R,P}$, so the ENC statistic for the nonparametric additive models is asymptotically normal as $R, P \to \infty$ and $P/R \to \infty$.

2.4.3 OCMT Procedure for Out-of-sample Forecast Encompassing Test

According to the encompassing principle, for the low-dimensional models, we combine 2 models in equations (2.16) and (2.17) with weights $(1 - \lambda_i)$ for Model 0 and λ_i for each Model i. There is no predictor in Model 0, while there is one predictor in each Model i, where $i = 1, \ldots, n$. We combine Model 0 and one Model i at a time to construct the

combined model for Model i, so we have n combined models for different Model i. For the high-dimensional models, we combine these $n+1$ models with weights λ_i for each Model i and $(1-\sum_{i=1}^{n}\lambda_i)$ for Model 0, so we get one combined model, which is the same as equation (3.15) in section 3 for the linear predictive model,

$$\hat{\mu}_{c,t+1} = \left(1 - \sum_{i=1}^{n} \lambda_i\right) \hat{\mu}_{0,t+1} + \sum_{i=1}^{n} \lambda_i \hat{\mu}_{i,t+1},$$

where $\hat{\mu}_{i,t+1}$ is the conditional mean for Model i that is estimated using the out-of-sample observations. Rearrange the regression above, we get the new regression equation is the same as equation (3.16),

$$\hat{\mu}_{0,t+1} = \sum_{i=1}^{n} \lambda_i \left(\hat{\mu}_{0,t+1} - \hat{\mu}_{i,t+1}\right) + \hat{\mu}_{c,t+1}, .$$

We denote $\hat{u}_{i,t+1}$ as the forecast errors for Model i, so we can calculate $\hat{u}_{i,t+1} = y_{t+1} - \hat{\mu}_{i,t+1}$. Rewriting the equation above with $\hat{u}_{i,t+1}$, we have the regression equation, which is the same as equation (2.9) in section 3 for the linear predictive model,

$$\hat{u}_{0,t+1} = \sum_{i=1}^{n} \lambda_i \left(\hat{u}_{0,t+1} - \hat{u}_{i,t+1}\right) + \hat{u}_{c,t+1}.$$

Then, we apply the ENC-OCMT approach we described in section 3 to select predictors in the high-dimensional nonparametric additive model. The ENC-OCMT approach in the additive model has the same procedure as that in the linear model. We select predictors in multiple stages when predictors are signals, pseudo-signals, and hidden signals. We stop selecting when no predictor is selected in the current stage.

2.4.4 Monte Carlo Simulation

In this subsection, we want to show that our ENC and ENC-OCMT approaches work well for the nonparametric additive models in finite data. The design of DGPs 6-10 follows the setting from Su et al. (2022). For all five DGPs, we set

$$f_1(x) = x; \qquad f_2(x) = (2x-1)^2; \qquad f_3(x) = \frac{\sin(2\pi x)}{2-\sin(2\pi x)}; \qquad \text{and}$$

$$f_4(x) = 0.1\sin(2\pi x) + 0.2\cos(2\pi x) + 0.3\sin(2\pi x)^2 + 0.4\cos(2\pi x)^3 + 0.5\sin(2\pi x)^3,$$

where $\varepsilon \sim$ i.i.d. $\mathrm{N}(0,1)$.

DGP 6: (No Hidden Signals and No Pseudo-Signals).

DGP 6 includes four independent signals and noise variables but no hidden signals or pseudo-signals. y_{t+1} is generated as follows:

$$y_{t+1} = 2.55f_1(x_{1,t}) + 2.57f_2(x_{2,t}) + 1.68f_3(x_{3,t}) + f_4(x_{4,t}) + \varepsilon_{t+1} \tag{2.21}$$

According to Su et al. (2022), the coefficients before f_i's are set to make each signal have the same strength in terms of variance for independent uniform $x_{1,t}, \ldots, x_{4,t}$. The predictors are generated as follows:

$$x_{i,t} = w_{i,t},\ i = 1, \ldots, 4, \qquad \text{and} \qquad x_{i,t} = \frac{w_{i,t} + u_{1,t}}{2},\ i \geq 5,$$

where $w_{i,t}, i = 1, \ldots, n$, and $u_{1,t}$ are all independent draws from $\mathrm{U}(0,1)$. Define the Signal-to-noise ratio to be $r = \frac{\mathrm{sd}(f)}{\mathrm{sd}(\varepsilon)}$, where $\mathrm{sd}(f)$ is the standard deviation of f and $\mathrm{sd}(\varepsilon)$ is the standard deviation of the error term ε. In this DGP, we set the Signal-to-noise ratio $r = 1.5$.

DGP 7: (Pseudo-Signals)

DGP 7 includes four independent signals, two pseudo-signals, and noise variables but no hidden signals. y_{t+1} is generated from equation (2.21). The predictors are generated as follows:

$$x_{i,t} = w_{i,t},\ i = 1, \ldots, 4, \quad x_{5,t} = \frac{4x_{1,t} + u_{1,t}}{5}, \quad x_{6,t} = \frac{4x_{2,t} + u_{2,t}}{5},$$
$$\text{and} \quad x_{i,t} = \frac{w_{i-2,t} + u_{3,t}}{2},\ i \geq 7,$$

where $w_{i,t}, i = 1, \ldots, (n-2)$, $u_{1,t}$, and $u_{2,t}$ are all independent draws from U$(0,1)$. This setting implies that $x_{i,t}$ for $i = 1, \ldots, 4$ are four independent signals, $x_{5,t}$ and $x_{6,t}$ are two pseudo-signals, and other predictors are noise variables.

DGP 8: (Hidden Signals)

DGP 8 includes four signals, one hidden signal, and noise variables but no pseudo-signals. y_{t+1} is generated from

$$y_{t+1} = 2.55 f_1(x_{1,t}) + 2.57 f_2(x_{2,t}) + 1.68 f_3(x_{3,t}) + f_4(x_{4,t}) + f_5(x_{5,t}) + \varepsilon_{t+1}, \qquad (2.22)$$

where $f_5(x_{5,t}) = -\mathbb{E}[2.55 f_1(x_{1,t}) + 2.57 f_2(x_{2,t}) + 1.68 f_3(x_{3,t}) + f_4(x_{4,t}) \mid x_{5,t}]$. The predictors are generated as follows:

$$x_{i,t} = w_{i,t},\ i = 1, 2, \quad x_{i,t} = \frac{w_{i,t} + u_{1,t}}{2},\ i = 3, 4, \quad x_{5,t} = u_{1,t},$$
$$\text{and} \quad x_{i,t} = \frac{w_{i-1,t} + u_{2,t}}{2},\ i \geq 6$$

where $w_{i,t}, i = 1, \ldots, (n-1)$, $u_{1,t}$, and $u_{2,t}$ are independent draws from U$(0,1)$. This setting implies that $x_{i,t}$ for $i = 1, \ldots, 4$ are four signals, $x_{5,t}$ is one hidden signal, and other predictors are noise variables.

DGP 9: (Both Hidden signals and Pseudo-Signals)

DGP 9 includes four signals, two pseudo-signals, one hidden signal, and noise variables. y_{t+1} is generated from equation (2.22). The predictors are generated as follows:

$$x_j = w_{j,t},\ i = 1, 2, \quad x_j = \frac{w_{i,t} + u_{1,t}}{2},\ i = 3, 4, \quad x_{5,t} = u_{1,t}, \quad x_{6,t} = \frac{4x_{1,t} + u_{2,t}}{5},$$
$$x_{7,t} = \frac{4x_{2,t} + u_{3,t}}{5}, \quad \text{and} \quad x_{i,t} = \frac{w_{i-3,t} + u_{3,t}}{2},\ i \geq 8$$

where $w_{i,t}, i = 1, \ldots, (n-3)$, $u_{1,t}$, $u_{2,t}$, and $u_{3,t}$ are independent draws from U$(0, 1)$. This setting implies that $x_{i,t}$ for $i = 1, \ldots, 4$ are four signals, $x_{5,t}$ is one hidden signal, $x_{6,t}$ and $x_{7,t}$ are two pseudo-signals, and other predictors are noise variables.

DGP 10: (Correlated Signals)

DGP 9 includes four signals, two pseudo-signals, one hidden signal, and noise variables. y_{t+1} is generated from equation (2.21). The predictors are generated as follows:

$$x_{i,t} = \frac{w_{i,t} + u_{1,t}}{2},\ i = 1, \ldots, 4, \quad \text{and} \quad x_{i,t} = \frac{w_{i,t} + u_{2,t}}{2},\ i \geq 5,$$

where $w_{i,t}, i = 1, \ldots, n$, $u_{1,t}$, and $u_{2,t}$ are independent draws from U$(0, 1)$. Thus, there are four signals in this setting and these four signals are correlated with each other. Other predictors are noise variables.

We have five DGPs in this subsection. DGPs 6-10 are nonparametric additive models from Su et al. (2022). We still set the sample size n to be $100, 200$, and 300, the number of in-sample observations R to be 100, and the number of out-of-sample observations to be $200, 300$, and 500 for all five DGPs. For ENC-OCMT method, we set nominal level $p = 0.1, 0.05, 0.01$, $\delta = 1$ in the first stage, and $\delta^* \in \{1.5, 2\}$ in the subsequent stages.

We report DGP 8 in this paper and move other simulation results to Supplemental Appendix. Table 5 shows that ENC-OCMT outperforms ENC-LASSO and ENC-ALASSO.

ENC-OCMT gets the largest TPR, lowest FPR, FDR, FDR* and lowest RMSFE in all experiments among the three methods, which implies that ENC-OCMT can successfully select signals, Hidden Signals, and Pseudo-signals. ENC-LASSO for the nonparametric additive model has the same problem as that for the linear model. ENC-LASSO still has a larger TPR than ENC-OCMT in most experiments. However, ENC-ALASSO has a lower FDR and FDR* compared to ENC-LASSO. Thus, ENC-LASSO tends to overestimate the number of signals, whereas ENC-ALASSO does not have this drawback. Therefore, ENC-OCMT outperforms ENC-LASSO and ENC-ALASSO for the nonparametric additive models.

2.5 Predictor Selection in High-dimensional Generalized Linear Models

In the previous sections, we demonstrated that our two approaches can be applied to linear models and nonparametric additive models. In this section, we extend the ENC-OCMT and ENC-LASSO approaches from linear models to generalized linear models.

2.5.1 Predictive Probability Model

In the previous sections, we consider the linear predictive model and set the dependent variable to be a continuous random variable. In this section, we consider the predictive probability model and set the dependent variable as the discrete random variable. We have

the probability model in equation (2.3),

$$y_{t+1} \sim \text{Bernoulli}\left(\Pr(y_{t+1}=1|\boldsymbol{x}_t)\right)$$

$$\Pr(y_{t+1}=1|\boldsymbol{x}_t) = G\left(\boldsymbol{x}_t'\boldsymbol{\beta}\right) \equiv 1/\left[1+\exp\left(-\boldsymbol{x}_t'\boldsymbol{\beta}\right)\right].$$

where $\boldsymbol{x}_t$ is a vector of $x_{i,t}$, for $i = 1, \ldots, n$, $\boldsymbol{\beta}$ is a vector of parameters β_i. The dependent variable y_{t+1} follows the Bernoulli distribution with the conditional probability $\Pr(y_{t+1} = 1|\boldsymbol{x}_t)$, and the conditional probability is set as binary response models $\Pr(y_{t+1}=1|\boldsymbol{x}_t) = G\left(\boldsymbol{x}_t'\boldsymbol{\beta}\right)$. We set the conditional probability to be the logistic function, so we have $G\left(\boldsymbol{x}_t'\boldsymbol{\beta}\right) \equiv 1/\left[1+\exp\left(-\boldsymbol{x}_t'\boldsymbol{\beta}\right)\right]$.

2.5.2 Out-of-sample Forecast Encompassing Approach

We still determine the predictive ability of predictors on the dependent variable using the out-of-sample forecast encompassing test. Similar to the regression models in section 2, we have $n+1$ nested mean regressions, $i = 1, \cdots, n$.

$$\text{Model } 0 : \Pr(y_{t+1}=1|\boldsymbol{x}_{0,t}) = G(a_0) = G(\boldsymbol{x}_{0,t}'\beta_0) \equiv G_{0,t+1}, \tag{2.23}$$

$$\text{Model } i : \Pr(y_{t+1}=1|\boldsymbol{x}_{i,t}) = G(a_i + b_i x_{i,t}) = G(\boldsymbol{x}_{i,t}'\beta_i) \equiv G_{i,t+1}, \tag{2.24}$$

where $G_{0,t+1}$ and $G_{i,t+1}$ are conditional mean of the predictive probability models for Model 0 and Model i respectively. $\boldsymbol{x}_{0,t}$ is a strict subset of $\boldsymbol{x}_{i,t}$. $\boldsymbol{x}_{0,t}' = 1, \boldsymbol{x}_{i,t}' = (1, x_{i,t}), \beta_0 = a_0, \beta_i = (a_i, b_i)'$. Thus, we estimate $\hat{G}_{0,t+1}$ and $\hat{G}_{i,t+1}$ by following $\hat{G}_{0,t+1} = \hat{a}_{0,t}$ and $\hat{G}_{i,t+1} = \hat{a}_{i,t} - \hat{b}_{i,t}x_{i,t}$, where $\hat{a}_{0,t}, \hat{a}_{i,t}$ and $\hat{b}_{i,t}$ are estimated maximizing the log-likelihood function $L = y_{t+1}\log G_{i,t+1} + (1-y_{t+1})\log(1-G_{i,t+1})$. There is no predictor in Model 0 and one predictor in Model i. Under the encompassing principle, we combine Model 0 and Model i with weight $1-\lambda_i$ and λ_i, respectively, then we get the conditional mean $\hat{G}_{c,i,t+1}$

for combined model i, $i = 1, \cdots, n$.

$$\hat{G}_{c,i,t+1} = (1 - \lambda_i)\, \hat{G}_{0,t+1} + \lambda_i \hat{G}_{i,t+1} \tag{2.25}$$

Since we use different weights λ_i to combine Model 0 and different Model i, there is a conditional mean $\hat{G}_{c,i,t+1}$ for each combined model i.

We know that the null and alternative hypothesis for the encompassing test are $\mathbb{H}_0 : \lambda_i = 0$, and $\mathbb{H}_1 : \lambda_i \neq 0$. Under the null hypothesis $\lambda_i = 0$, the combined model is Model 0, which means that $x_{i,t}$ does not have the predictive ability on y_{t+1} in mean. Under the alternative $\lambda_i \neq 0$, the combined model combines Model 0 and Model i with weights $1 - \lambda_i$ and λ_i, which implies that $x_{i,t}$ has the predictive ability on y_{t+1} in mean.

To find out the encompassing statistic under the null hypothesis, we estimate the weight λ_i maximizing the expectation of log-likelihood function with combined mean $G_{c,i}$. The encompassing principle is based on the score function with respect to λ_i. We define C_i as the score function that takes a derivative of the loss function of the combined model with respect to the weight λ. Thus, C_i is the score function for Model i, which is defined as

$$C_i \equiv \frac{\partial \mathbb{E} L(G_{c,i,t+1}, y_{t+1})}{\partial \lambda_i} = 0. \tag{2.26}$$

This score function can be separated into two parts. The first is a derivative of the loss function with respect to the conditional mean for the combined model $G_{c,i,t+1}$. Then we give a name to the first derivative as the identification function of the conditional mean for the combined model. We use $V_{i,t+1}$ to denote the identification function for Model i. The second is a derivative of the conditional mean for the combined model $G_{c,i,t+1}$ with respect to weight λ_i. The second derivative is called to be the test function of conditional mean

in this paper. We denote $\xi_{i,t+1}$ to be the test function for the Model i. Thus, the score function C_i for Model i is the expectation of the product of identification function $V_{i,t+1}$ and test function $\xi_{i,t+1}$. Therefore, the score function C_i can be derived as follows:

$$\begin{aligned} C_i &\equiv \frac{\partial \mathbb{E} L(G_{c,i,t+1}, y_{t+1})}{\partial \lambda_i} \\ &= \mathbb{E}\left(\frac{\partial L(G_{c,i,t+1}, y_{t+1})}{\partial G_{c,i,t+1}}\right)\left(\frac{\partial G_{c,i,t+1}}{\partial \lambda_i}\right) \\ &= \mathbb{E}\left(\frac{y_{t+1}}{G_{c,i,t+1}} - \frac{1-y_{t+1}}{1-G_{c,i,t+1}}\right)(G_{i,t+1} - G_{0,t+1}) \\ &\equiv \mathbb{E}\left(V_{c,i,t+1}\right)\left(\xi_{i,t+1}\right) = 0 \end{aligned} \tag{2.27}$$

where the loss function for the combined model is the log-likelihood function $L\left(y_{t+1}, G_{c,i,t+1}\right)$ $= -\left[y_{t+1}\log G_{c,i,t+1} + (1-y_{t+1})\log\left(1-G_{c,i,t+1}\right)\right]$ with conditional mean for the combined model $G_{c,i,t+1}$. Then, the identification function $V_{c,i,t+1}$ is defined as $\frac{\partial L(y_{t+1}, G_{c,i,t+1})}{\partial G_{c,i,t+1}} = \left(\frac{y_{t+1}}{G_{c,i,t+1}} - \frac{1-y_{t+1}}{1-G_{c,i,t+1}}\right)$, and the test function $\xi_{i,t+1}$ is defined as $\frac{\partial G_{c,i,t+1}}{\partial \lambda_i} \equiv (G_{i,t+1} - G_{0,t+1})$. Under the null hypothesis $\lambda_i = 0$, since the conditional mean for combined model i is equal to that for Model 0, $G_{c,i,t+1} = G_{0,t+1}$ and $V_{c,i,t+1} = V_{0,t+1}$. Thus, under the null hypothesis, we have the score function C_i for Model i to be

$$C_i = \mathbb{E}\left(\frac{y_{t+1}}{G_{0,t+1}} - \frac{1-y_{t+1}}{1-G_{0,t+1}}\right)(G_{i,t+1} - G_{0,t+1}) = \mathbb{E}\left(V_{0,t+1}\right)\left(\xi_{i,t+1}\right) = 0.$$

For an out-of-sample forecast encompassing test, we separate the observations into two parts: in-sample observations and out-of-sample observations. The number of in-sample observations is defined as R and the number of out-of-sample observations is defined as P. The number of observations is $T+1$. We estimate $\hat{a}_0$, $\hat{a}_i$ and $\hat{b}_i$ by using the in-sample observations from 1 to R. And then, we estimate $\hat{G}_{0,t+1} = 1/(1+\exp(\hat{a}_0))$ and $\hat{G}_{i,t+1} = 1/(1+\exp(\hat{a}_i + \hat{b}_i x_{i,t}))$ by using out-of-sample observations from $R+1$ to $T+1$.

After that, we also use the out-of-sample observations to estimate the score function C_i by $\hat{C}_{i,R,P}$. The $\hat{C}_{i,R,P}$ is defined as

$$\hat{C}_{i,R,P} \equiv P^{-1} \sum_{t=R}^{T} \left[\left(\frac{y_{t+1}}{\hat{G}_{0,t+1}} - \frac{1 - y_{t+1}}{1 - \hat{G}_{0,t+1}} \right) \left(\hat{G}_{i,t+1} - \hat{G}_{0,t+1} \right) \right] \xrightarrow{p} C_i = 0$$

under $\mathbb{H}_0$, as $R, P \to \infty$. Thus, $\mathbb{E}\left(\hat{C}_{i,R,P}\right) = 0$, so $\hat{C}_{i,R,P} \xrightarrow{p} 0$ as $R, P \to \infty$. From section 2, the ENC statistic is denoted as normalization of $\hat{C}_{i,R,P}$, so the ENC statistic is asymptotically normal as $R, P \to \infty$ and $P/R \to \infty$.

2.5.3 OCMT Procedure for Out-of-sample Forecast Encompassing Test

According to the encompassing principle, for the low-dimensional models, we combine 2 models in equations (2.23) and (2.24) with weights $(1 - \lambda_i)$ for Model 0 and λ_i for each Model i. There is no predictor in Model 0, while there is one predictor in each Model i, where $i = 1, \ldots, n$. We combine Model 0 and one Model i at a time to construct the combined model for Model i, so we have n combined models for different Model i. For the high-dimensional models, we combine these $n + 1$ models with weights λ_i for each Model i and $(1 - \sum_{i=1}^{n} \lambda_i)$ for Model 0, so we get one combined model as follows:

$$\hat{G}_{c,t+1} = \left(1 - \sum_{i=1}^{n} \lambda_i \right) \hat{G}_{0,t+1} + \sum_{i=1}^{n} \lambda_i \hat{G}_{i,t+1}, \tag{2.28}$$

where $\hat{G}_{i,t+1}$ is the conditional mean for Model i that is estimated using the out-of-sample observations. Rewriting equation (2.28), we have

$$\hat{G}_{0,t+1} = \sum_{i=1}^{n} \lambda_i \left(\hat{G}_{0,t+1} - \hat{G}_{i,t+1} \right) + \hat{G}_{c,t+1}. \tag{2.29}$$

In the first stage of the ENC-OCMT, we consider n bivariate regressions of $\hat{G}_{0,t+1}$ on $\left(\hat{G}_{0,t+1} - \hat{G}_{i,t+1}\right)$, for $i = 1, \ldots, n$,

$$\hat{G}_{0,t+1} = \psi_{i,(i)} \left(\hat{G}_{0,t+1} - \hat{G}_{i,t+1}\right) + \hat{G}_{c,i,t+1}, \tag{2.30}$$

where $\hat{G}_{c,i,t+1}$ is an error term of the regression for Model i. We denote $\hat{u}_{i,t+1}$ as the residuals for Model i, so we can calculate $\hat{u}_{i,t+1} = y_{t+1} - \hat{G}_{i,t+1}$. Rewriting the equation above with $\hat{u}_{i,t+1}$, we have the regression equation, which is the same as equation (2.9) in section 3 for the linear predictive model,

$$\hat{u}_{0,t+1} = \sum_{i=1}^{n} \lambda_i \left(\hat{u}_{0,t+1} - \hat{u}_{i,t+1}\right) + \hat{u}_{c,t+1}.$$

Then, we apply the ENC-OCMT approach we described in section 3 to select predictors in the high-dimensional nonparametric additive model. The ENC-OCMT approach in the generalized linear model has the same procedure as that in the linear model. We select predictors in multiple stages when predictors are signals, pseudo-signals, and hidden signals. We stop selecting when no predictor is selected in the current stage.

2.5.4 Monte Carlo Simulation

In this subsection, we want to show the good performance of our ENC-OCMT approach for the generalized linear models in finite data. We use equation (2.3) to be DGP 11 for high-dimensional logistic regression models. The setting of predictors is the same as DGP 1(a). The predictors $x_{i,t} = \left(\varepsilon_{i,t} + \nu g_t\right) / \sqrt{1+\nu}$, for $i = 1, 2, 3, 4$, and $x_{5,t} = \varepsilon_{5,t}, x_{i,t} = \left(\varepsilon_{i-1,t} + \varepsilon_{i,t}\right) / \sqrt{2}$, for $i > 5$, where g_t and ε_{it} are independent draws from N$(0, 1)$, for $t = 1, 2, \ldots, T$, and $i = 1, 2, \ldots, n$. We set $\beta_1 = \beta_2 = \beta_3 = 1$, $\beta_4 = -1.5$, $\beta_i = 0$, for $i = 5, \ldots, n$, $\nu = 1$ and $n = 300$. This DGP implies $\theta_i \neq 0$ for $i = 1, 2, 3, 5, 6$, and $\theta_i = 0$

for $i = 4$ and $i > 6$. Thus, the predictors $x_{i,t}$ for $i = 1, 2, 3$ are signals, $x_{4,t}$ is hidden signal, $x_{5,t}$ and $x_{6,t}$ are pseudo-signals, and others are noise variables.

Table 6 shows that the TPR for ENC-OCMT, ENC-LASSO, and ENC-ALASSO tends to be 1 when the number of covariates is much larger than the number of in-sample observations. ENC-LASSO's FPR and FDR are larger than ENC-OCMT's and ENC-ALASSO's. ENC-LASSO in the logistic function has the same problem as in the linear model: ENC-LASSO tends to overestimate the number of signals. From Table 8, we can see that ENC-OCMT can get the lowest forecast loss in the logistic function compared to ENC-LASSO and ENC-ALASSO.

2.5.5 Application to Forecasting probability of Recession

In this subsection, we still use quarterly data from Stock and Watson (2012) to present a macroeconomic forecasting application for U.S. GDP growth using a large set of macroeconomic variables. The data set contains 109 series from $1960Q3$ to $2008Q4$. The in-sample period is from $1960Q3$ to $1980Q2$ (80 observations) and from $1960Q3$ to $1990Q2$ (120 observations) while the out-of-sample period is from $1980Q3$ to $2008Q4$ (114 observations) and from $1990Q3$ to $2008Q4$ (74 observations). We define z_{t+1} as U.S. GDP growth, and then we set the economics status to be recession $y_{t+1} = 1$ if the real GDP Growth is negative $z_{t+1} < 0$ and the economics status to be non-recession $y_{t+1} = 0$ if the real GDP Growth is nonnegative $z_{t+1} \geq 0$. The conditional probability $\Pr(y_{t+1} = 1|\boldsymbol{x}_t) = G(\boldsymbol{x}_t'\boldsymbol{\beta})$ is set to be the logistic function $G(\boldsymbol{x}_t'\boldsymbol{\beta}) = 1/[1 + \exp(-\boldsymbol{x}_t'\boldsymbol{\beta})]$. We use the rolling window in estimation and a one-step-ahead forecast.

Table 7 demonstrates the empirical results of the U.S. recession forecast. From the table, we can see that ENC-OCMT can get the lowest out-of-sample forecast error to forecast the U.S. recession compared to LASSO, OCMT, and ENC-LASSO for two different numbers of out-of-sample observations. In this application of the U.S. recession forecast, OCMT has the highest forecast error, which implies that OCMT cannot work well in the logistic function.

2.6 Conclusion

In this paper, we develop a general forecast encompassing (ENC) framework for testing Granger-Causality in conditional mean regression. We prove that the ENC statistic is asymptotically normal under the null hypothesis of no Granger-Causality. Additionally, we propose two novel predictor selection methods based on the ENC framework. ENC-LASSO employs the LASSO method by Tibshirani (1996), which can simultaneously select predictors based on their out-of-sample predictive ability. ENC-OCMT integrates the OCMT method of Chudik, Kapetanios, and Pesaran (2018) into the forecast encompassing regression.

ENC-OCMT has three main advantages compared to the original OCMT from CKP (2018). First, ENC-OCMT can be applied to nonlinear regression models, making it a versatile and easily adaptable framework for more complex models. Second, ENC-OCMT can be applied not only to the predictive mean regression model but also to any elicitable predictive functionals of the conditional distribution, such as predictive quantile regression models, predictive expectile regression models, predictive mode regression models, predic-

tive expected shortfalls regression models, and predictive density forecast models, among others. Third, ENC-OCMT can be applied to predictive regression models to select predictors based on out-of-sample predictive ability, making it suitable for identifying predictors with Granger-causality.

Simulation results show that ENC-OCMT can outperform OCMT and LASSO in out-of-sample forecasting. Empirical applications demonstrate the method's merits in practice. The empirical results indicate that ENC-OCMT performs well in terms of out-of-sample prediction when forecasting U.S. real GDP growth, U.S. CPI inflation, and U.S. recession.

Table 2.1: DGP 1(a) - Linear Model (Signals Only), $R^2 = 70\%$, $n = 100$

	P	Loss	TPR	FPR	FDR	FDR*
$p = 0.1$,	200	2.1132	1.0000	0.0012	0.0178	0.0178
$\delta = 1$, $\delta^* = 1.5$	300	2.1137	1.0000	0.0012	0.0190	0.0190
	500	2.1131	1.0000	0.0011	0.0177	0.0177
$p = 0.05$,	200	2.1127	1.0000	0.0006	0.0100	0.0100
$\delta = 1$, $\delta^* = 1.5$	300	2.1132	1.0000	0.0006	0.0095	0.0095
	500	2.1126	1.0000	0.0005	0.0083	0.0083
p=0.01,	200	2.1122	1.0000	0.0001	0.0017	0.0017
$\delta = 1$, $\delta^* = 1.5$	300	2.1128	1.0000	0.0001	0.0016	0.0016
	500	2.1121	1.0000	0.0001	0.0018	0.0018
$p = 0.1$,	200	2.1131	1.0000	0.0011	0.0166	0.0166
$\delta = 1$, $\delta^* = 2$	300	2.1137	1.0000	0.0011	0.0174	0.0174
	500	2.1131	1.0000	0.0010	0.0162	0.0162
$p = 0.05$,	200	2.1127	1.0000	0.0006	0.0092	0.0092
$\delta = 1$, $\delta^* = 2$	300	2.1132	1.0000	0.0005	0.0086	0.0086
	500	2.1126	1.0000	0.0005	0.0079	0.0079
$p = 0.01$,	200	2.1122	1.0000	0.0001	0.0016	0.0016
$\delta = 1$, $\delta^* = 2$	300	2.1128	1.0000	0.0001	0.0013	0.0013
	500	2.1121	1.0000	0.0001	0.0017	0.0017
ENC-LASS	200	2.1574	1.0000	0.0514	0.4044	0.4044
	300	2.1631	1.0000	0.0543	0.4161	0.4161
	500	2.1634	1.0000	0.0511	0.3975	0.3975
ENC-ALASSO	200	2.1145	0.9955	0.0036	0.0373	0.0373
	300	2.1148	0.9996	0.0027	0.0285	0.0285
	500	2.1142	1.0000	0.0027	0.0202	0.0202

Table 2.2: DGP 1(a) - Linear Model (Signals Only), $R^2 = 70\%$, $n = 200$

	P	Loss	TPR	FPR	FDR	FDR*
$p = 0.1$,	200	2.1115	1.0000	0.0006	0.0188	0.0188
$\delta = 1$, $\delta^* = 1.5$	300	2.1140	1.0000	0.0006	0.0188	0.0188
	500	2.1131	1.0000	0.0005	0.0167	0.0167
$p = 0.05$,	200	2.1108	1.0000	0.0003	0.0086	0.0086
$\delta = 1$, $\delta^* = 1.5$	300	2.1133	1.0000	0.0003	0.0095	0.0095
	500	2.1113	1.0000	0.0003	0.0085	0.0085
p=0.01,	200	2.1103	1.0000	0.0001	0.0023	0.0023
$\delta = 1$, $\delta^* = 1.5$	300	2.1128	1.0000	0.0000	0.0015	0.0015
	500	2.1109	1.0000	0.0001	0.0019	0.0019
$p = 0.1$,	200	2.1115	1.0000	0.0005	0.0173	0.0173
$\delta = 1$, $\delta^* = 2$	300	2.1139	1.0000	0.0006	0.0179	0.0179
	500	2.1118	1.0000	0.0005	0.0156	0.0156
$p = 0.05$,	200	2.1108	1.0000	0.0002	0.0080	0.0080
$\delta = 1$, $\delta^* = 2$	300	2.1133	1.0000	0.0003	0.0090	0.0090
	500	2.1113	1.0000	0.0002	0.0078	0.0078
$p = 0.01$,	200	2.1103	1.0000	0.0001	0.0021	0.0021
$\delta = 1$, $\delta^* = 2$	300	2.1128	1.0000	0.0000	0.0015	0.0015
	500	2.1109	1.0000	0.0001	0.0018	0.0018
ENC-LASS	200	2.1686	1.0000	0.0334	0.4639	0.4639
	300	2.1755	1.0000	0.0334	0.4545	0.4545
	500	2.1742	1.0000	0.0322	0.4465	0.4465
ENC-ALASSO	200	2.1154	0.9953	0.0034	0.0625	0.0625
	300	2.1178	0.9996	0.0033	0.0486	0.0486
	500	2.1177	1.0000	0.0038	0.0487	0.0487

Table 2.3: DGP 1(a) - Linear Model (Signals Only), $R^2 = 70\%$, $n = 300$

	P	Loss	TPR	FPR	FDR	FDR*
$p = 0.1$,	200	2.1136	1.0000	0.0004	0.0200	0.0200
$\delta = 1$, $\delta^* = 1.5$	300	2.1129	1.0000	0.0003	0.0153	0.0153
	500	2.1133	1.0000	0.0004	0.0183	0.0183
$p = 0.05$,	200	2.1128	1.0000	0.0002	0.0099	0.0099
$\delta = 1$, $\delta^* = 1.5$	300	2.1123	1.0000	0.0002	0.0081	0.0081
	500	2.1127	1.0000	0.0002	0.0102	0.0102
p=0.01,	200	2.1124	1.0000	0.0000	0.0017	0.0017
$\delta = 1$, $\delta^* = 1.5$	300	2.1118	1.0000	0.0000	0.0012	0.0012
	500	2.1124	1.0000	0.0000	0.0021	0.0021
$p = 0.1$,	200	2.1136	1.0000	0.0004	0.0189	0.0189
$\delta = 1$, $\delta^* = 2$	300	2.1129	1.0000	0.0003	0.0146	0.0146
	500	2.1133	1.0000	0.0004	0.0175	0.0175
$p = 0.05$,	200	2.1128	1.0000	0.0002	0.0095	0.0095
$\delta = 1$, $\delta^* = 2$	300	2.1123	1.0000	0.0002	0.0077	0.0077
	500	2.1127	1.0000	0.0002	0.0100	0.0100
$p = 0.01$,	200	2.1123	1.0000	0.0000	0.0016	0.0016
$\delta = 1$, $\delta^* = 2$	300	2.1118	1.0000	0.0000	0.0011	0.0011
	500	2.1124	1.0000	0.0000	0.0020	0.0020
ENC-LASS	200	2.1849	1.0000	0.0268	0.5019	0.5019
	300	2.1837	1.0000	0.0259	0.4955	0.4955
	500	2.1889	1.0000	0.0237	0.4588	0.4588
ENC-ALASSO	200	2.1206	0.9960	0.0033	0.0809	0.0809
	300	2.1197	0.9996	0.0033	0.0644	0.0644
	500	2.1260	1.0000	0.0036	0.0597	0.0597

Table 2.4: DGP 1(a) - Linear Model (Signals Only), $R^2 = 50\%$, $n = 100$

	P	Loss	TPR	FPR	FDR	FDR*
$p = 0.1$,	200	3.2293	1.0000	0.0012	0.0191	0.0191
$\delta = 1$, $\delta^* = 1.5$	300	3.2254	1.0000	0.0013	0.0201	0.0201
	500	3.2322	1.0000	0.0011	0.0169	0.0169
$p = 0.05$,	200	3.2287	1.0000	0.0007	0.0110	0.0110
$\delta = 1$, $\delta^* = 1.5$	300	3.2246	1.0000	0.0007	0.0105	0.0105
	500	3.2313	1.0000	0.0005	0.0086	0.0086
$p = 0.01$,	200	3.2278	1.0000	0.0002	0.0026	0.0026
$\delta = 1$, $\delta^* = 1.5$	300	3.2239	1.0000	0.0001	0.0021	0.0021
	500	3.2306	1.0000	0.0001	0.0019	0.0019
$p = 0.1$,	200	3.2292	1.0000	0.0011	0.0176	0.0176
$\delta = 1$, $\delta^* = 2$	300	3.2252	1.0000	0.0012	0.0186	0.0186
	500	3.2321	1.0000	0.0010	0.0163	0.0163
$p = 0.05$,	200	3.2286	1.0000	0.0006	0.0100	0.0100
$\delta = 1$, $\delta^* = 2$	300	3.2244	1.0000	0.0006	0.0100	0.0100
	500	3.2312	1.0000	0.0005	0.0080	0.0080
$p = 0.01$,	200	3.2278	1.0000	0.0002	0.0024	0.0024
$\delta = 1$, $\delta^* = 2$	300	3.2237	1.0000	0.0001	0.0019	0.0019
	500	3.2306	1.0000	0.0001	0.0019	0.0019
ENC-LASSO	200	3.3102	0.9965	0.0534	0.4134	0.4134
	300	3.3093	0.9999	0.0563	0.4324	0.4324
	500	3.3232	1.0000	0.0562	0.4276	0.4276
ENC-ALASSO	200	3.2353	0.9221	0.0059	0.0731	0.0731
	300	3.2293	0.9761	0.0047	0.0567	0.0567
	500	3.2337	0.9985	0.0029	0.0361	0.0361

Table 2.5: DGP 1(a) - Linear Model (Signals Only), $R^2 = 50\%$, $n = 200$

	P	Loss	TPR	FPR	FDR	FDR*
$p = 0.1$,	200	3.2333	1.0000	0.0006	0.0201	0.0201
$\delta = 1$, $\delta^* = 1.5$	300	3.2246	1.0000	0.0005	0.0168	0.0168
	500	3.2298	1.0000	0.0006	0.0186	0.0186
$p = 0.05$,	200	3.2325	1.0000	0.0003	0.0104	0.0104
$\delta = 1$, $\delta^* = 1.5$	300	3.2242	1.0000	0.0003	0.0091	0.0091
	500	3.2290	1.0000	0.0003	0.0092	0.0092
$p = 0.01$,	200	3.2317	1.0000	0.0001	0.0018	0.0018
$\delta = 1$, $\delta^* = 1.5$	300	3.2239	1.0000	0.0001	0.0020	0.0020
	500	3.2283	1.0000	0.0000	0.0016	0.0016
$p = 0.1$,	200	3.2331	1.0000	0.0006	0.0185	0.0185
$\delta = 1$, $\delta^* = 2$	300	3.2246	1.0000	0.0005	0.0154	0.0154
	500	3.2297	1.0000	0.0006	0.0178	0.0178
$p = 0.05$,	200	3.2324	1.0000	0.0003	0.0095	0.0095
$\delta = 1$, $\delta^* = 2$	300	3.2241	1.0000	0.0003	0.0084	0.0084
	500	3.2289	1.0000	0.0003	0.0089	0.0089
$p = 0.01$,	200	3.2317	1.0000	0.0001	0.0018	0.0018
$\delta = 1$, $\delta^* = 2$	300	3.2238	1.0000	0.0001	0.0019	0.0019
	500	3.2282	1.0000	0.0000	0.0016	0.0016
ENC-LASSO	200	3.3408	0.9973	0.0356	0.4778	0.4778
	300	3.3370	0.9996	0.0357	0.4789	0.4789
	500	3.3482	1.0000	0.0354	0.4778	0.4778
ENC-ALASSO	200	3.2449	0.9281	0.0056	0.1172	0.1172
	300	3.2355	0.9773	0.0043	0.0922	0.0922
	500	3.2371	0.9978	0.0029	0.0587	0.0587

Table 2.6: DGP 1(a) - Linear Model (Signals Only), $R^2 = 50\%$, $n = 300$

	P	Loss	TPR	FPR	FDR	FDR*
$p = 0.1$,	200	3.2290	1.0000	0.0004	0.0200	0.0200
$\delta = 1$, $\delta^* = 1.5$	300	3.2307	1.0000	0.0004	0.0174	0.0174
	500	3.2301	1.0000	0.0004	0.0186	0.0186
$p = 0.05$,	200	3.2280	1.0000	0.0002	0.0097	0.0097
$\delta = 1$, $\delta^* = 1.5$	300	3.2300	1.0000	0.0002	0.0088	0.0088
	500	3.2292	1.0000	0.0002	0.0094	0.0094
$p = 0.01$,	200	3.2270	1.0000	0.0000	0.0020	0.0020
$\delta = 1$, $\delta^* = 1.5$	300	3.2293	1.0000	0.0000	0.0018	0.0018
	500	3.2284	1.0000	0.0000	0.0023	0.0023
$p = 0.1$,	200	3.2287	1.0000	0.0004	0.0188	0.0188
$\delta = 1$, $\delta^* = 2$	300	3.2305	1.0000	0.0003	0.0170	0.0170
	500	3.2301	1.0000	0.0004	0.0179	0.0179
$p = 0.05$,	200	3.2279	1.0000	0.0002	0.0091	0.0091
$\delta = 1$, $\delta^* = 2$	300	3.2299	1.0000	0.0002	0.0087	0.0087
	500	3.2292	1.0000	0.0002	0.0088	0.0088
$p = 0.01$,	200	3.2270	1.0000	0.0000	0.0020	0.0020
$\delta = 1$, $\delta^* = 2$	300	3.2293	1.0000	0.0000	0.0017	0.0017
	500	3.2284	1.0000	0.0000	0.0022	0.0022
ENC-LASSO	200	3.3643	0.9950	0.0282	0.5124	0.5124
	300	3.3631	0.9994	0.0277	0.5128	0.5128
	500	3.3713	1.0000	0.0260	0.4939	0.4939
ENC-ALASSO	200	3.2476	0.9303	0.0053	0.1489	0.1489
	300	3.2450	0.9778	0.0037	0.1116	0.1116
	500	3.2421	0.9986	0.0024	0.0714	0.0714

Table 2.7: DGP 1(a) - Linear Model (Signals Only), $R^2 = 50\%$, $n = 300$

	P	Loss	TPR	FPR	FDR	FDR*
$p = 0.1$,	200	4.9336	0.9985	0.0012	0.0182	0.0182
$\delta = 1$, $\delta^* = 1.5$	300	4.9218	1.0000	0.0011	0.0181	0.0181
	500	4.9268	1.0000	0.0010	0.0159	0.0159
$p = 0.05$,	200	4.9319	0.9969	0.0006	0.0101	0.0101
$\delta = 1$, $\delta^* = 1.5$	300	4.9205	0.9999	0.0005	0.0082	0.0082
	500	4.9253	1.0000	0.0005	0.0076	0.0076
$p = 0.01$,	200	4.9302	0.9906	0.0002	0.0026	0.0026
$\delta = 1$, $\delta^* = 1.5$	300	4.9196	0.9999	0.0001	0.0020	0.0020
	500	4.9247	1.0000	0.0001	0.0022	0.0022
$p = 0.1$,	200	4.9338	0.9985	0.0011	0.0172	0.0172
$\delta = 1$, $\delta^* = 2$	300	4.9217	1.0000	0.0011	0.0172	0.0172
	500	4.9268	1.0000	0.0010	0.0156	0.0156
$p = 0.05$,	200	4.9320	0.9969	0.0006	0.0096	0.0096
$\delta = 1$, $\delta^* = 2$	300	4.9205	0.9999	0.0005	0.0078	0.0078
	500	4.9253	1.0000	0.0005	0.0073	0.0073
$p = 0.01$,	200	4.9303	0.9906	0.0002	0.0024	0.0024
$\delta = 1$, $\delta^* = 2$	300	4.9196	0.9999	0.0001	0.0019	0.0019
	500	4.9247	1.0000	0.0001	0.0022	0.0022
ENC-LASSO	200	5.0461	0.9373	0.0500	0.4010	0.4010
	300	5.0531	0.9800	0.0557	0.4311	0.4311
	500	5.0799	0.9985	0.0613	0.4498	0.4498
ENC-ALASSO	200	4.9364	0.7269	0.0101	0.1291	0.1291
	300	4.9280	0.8378	0.0079	0.1007	0.1007
	500	4.9367	0.9426	0.0067	0.0820	0.0820

Table 2.8: DGP 1(a) - Linear Model (Signals Only), $R^2 = 50\%$, $n = 300$

	P	Loss	TPR	FPR	FDR	FDR*
$p = 0.1$,	200	4.9304	0.9970	0.0006	0.0191	0.0191
$\delta = 1$, $\delta^* = 1.5$	300	4.9286	1.0000	0.0006	0.0196	0.0196
	500	4.9311	1.0000	0.0006	0.0185	0.0185
$p = 0.05$,	200	4.9287	0.9956	0.0003	0.0099	0.0099
$\delta = 1$, $\delta^* = 1.5$	300	4.9270	1.0000	0.0003	0.0106	0.0106
	500	4.9302	1.0000	0.0003	0.0100	0.0100
$p = 0.01$,	200	4.9259	0.9879	0.0001	0.0024	0.0024
$\delta = 1$, $\delta^* = 1.5$	300	4.9264	0.9998	0.0001	0.0020	0.0020
	500	4.9292	1.0000	0.0001	0.0018	0.0018
$p = 0.1$,	200	4.9286	0.9970	0.0006	0.0185	0.0185
$\delta = 1$, $\delta^* = 2$	300	4.9271	1.0000	0.0006	0.0193	0.0193
	500	4.9301	1.0000	0.0006	0.0181	0.0181
$p = 0.05$,	200	4.9304	0.9956	0.0003	0.0096	0.0096
$\delta = 1$, $\delta^* = 2$	300	4.9283	1.0000	0.0003	0.0102	0.0102
	500	4.9311	1.0000	0.0003	0.0097	0.0097
$p = 0.01$,	200	4.9259	0.9879	0.0001	0.0023	0.0023
$\delta = 1$, $\delta^* = 2$	300	4.9263	0.9998	0.0001	0.0019	0.0019
	500	4.9292	1.0000	0.0001	0.0018	0.0018
ENC-LASSO	200	5.0871	0.9265	0.0331	0.4640	0.4640
	300	5.1057	0.9740	0.0352	0.4722	0.4722
	500	5.1308	0.9968	0.0376	0.4938	0.4938
ENC-ALASSO	200	4.9458	0.7254	0.0076	0.1766	0.1766
	300	4.9450	0.8423	0.0069	0.1530	0.1530
	500	4.9523	0.9454	0.0053	0.1189	0.1189

Table 2.9: DGP 1(a) - Linear Model (Signals Only), $R^2 = 50\%$, $n = 300$

	P	Loss	TPR	FPR	FDR	FDR*
$p = 0.1$,	200	4.9281	0.9959	0.0004	0.0191	0.0191
$\delta = 1$, $\delta^* = 1.5$	300	4.9261	0.9998	0.0004	0.0173	0.0173
	500	4.9345	1.0000	0.0004	0.0178	0.0178
$p = 0.05$,	200	4.9269	0.9925	0.0002	0.0110	0.0110
$\delta = 1$, $\delta^* = 1.5$	300	4.9249	0.9998	0.0002	0.0081	0.0081
	500	4.9332	1.0000	0.0002	0.0083	0.0083
$p = 0.01$,	200	4.9247	0.9824	0.0000	0.0023	0.0023
$\delta = 1$, $\delta^* = 1.5$	300	4.9239	0.9991	0.0000	0.0014	0.0014
	500	4.9321	1.0000	0.0000	0.0023	0.0023
$p = 0.1$,	200	4.9281	0.9959	0.0004	0.0186	0.0186
$\delta = 1$, $\delta^* = 2$	300	4.9260	0.9998	0.0003	0.0169	0.0169
	500	4.9345	1.0000	0.0004	0.0171	0.0171
$p = 0.05$,	200	4.9270	0.9925	0.0002	0.0109	0.0109
$\delta = 1$, $\delta^* = 2$	300	4.9248	0.9998	0.0002	0.0078	0.0078
	500	4.9333	1.0000	0.0002	0.0080	0.0080
$p = 0.01$,	200	4.9247	0.9824	0.0000	0.0022	0.0022
$\delta = 1$, $\delta^* = 2$	300	4.9239	0.9991	0.0000	0.0013	0.0013
	500	4.9321	1.0000	0.0000	0.0023	0.0023
ENC-LASSO	200	5.1315	0.9186	0.0262	0.5033	0.5033
	300	5.1469	0.9720	0.0257	0.4988	0.4988
	500	5.1644	0.9978	0.0283	0.5170	0.5170
ENC-ALASSO	200	4.9624	0.7340	0.0072	0.2130	0.2130
	300	4.9555	0.8413	0.0056	0.1775	0.1775
	500	4.9636	0.9486	0.0045	0.1403	0.1403

Table 2.10: Real DGP Growth and Inflation Forecasting Using Predictive Mean Regressions

Real Output Growth		ΔInflation	
Models	MSFE	Models	MSFE
Model 0		**Model 0**	
AR+Factors	3.2732	AR+Factors	3.6000
ENC-LASSO		**ENC-LASSO**	
Model S		**Model S**	
AR + Factors + All selected variables (10)	3.8805	AR + Factors + All selected variables (3)	3.9007
ENC-Adaptive LASSO		**ENC-Adaptive LASSO**	
Model S		**Model S**	
AR + Factors + All selected variables (1)	3.4736	AR + Factors + All selected variables (1)	3.5245
ENC-OCMT		**ENC-OCMT**	
Model S		**Model S**	
AR + Factors + All selected variables (0)	3.2732	Factors + All selected variables (5)	3.6458
POST-ENC-OCMT		**POST-ENC-OCMT**	
Model S		**Model S**	
AR + Factors + All selected variables (0)	3.2732	Factors + All selected variables (1)	3.5576

Table 2.11: DGP 8 - Nonparametric Additive Models (Signals, Pseudo-Signals and Hidden Signals), $n = 100$

	P	Loss	TPR	FPR	FDR	FDR*
$p = 0.1$,	200	2.2301	0.7980	0.0011	0.0158	0.0158
$\delta = 1$, $\delta^* = 1.5$	300	2.2249	0.8004	0.0009	0.0142	0.0142
	500	2.2329	0.8002	0.0012	0.0170	0.0170
$p = 0.05$,	200	2.2353	0.7964	0.0005	0.0077	0.0077
$\delta = 1$, $\delta^* = 1.5$	300	2.2294	0.8002	0.0004	0.0065	0.0065
	500	2.2385	0.8002	0.0006	0.0097	0.0097
$p = 0.01$,	200	2.2455	0.7884	0.0001	0.0019	0.0019
$\delta = 1$, $\delta^* = 1.5$	300	2.2352	0.8002	0.0001	0.0013	0.0013
	500	2.2438	0.8000	0.0001	0.0023	0.0023
$p = 0.1$,	200	2.2340	0.7956	0.0010	0.0149	0.0149
$\delta = 1$, $\delta^* = 2$	300	2.2267	0.8004	0.0009	0.0137	0.0137
	500	2.2332	0.8000	0.0011	0.0158	0.0158
$p = 0.05$,	200	2.2385	0.7912	0.0005	0.0074	0.0074
$\delta = 1$, $\delta^* = 2$	300	2.2305	0.8000	0.0004	0.0065	0.0065
	500	2.2385	0.8000	0.0006	0.0092	0.0092
$p = 0.01$,	200	2.2523	0.7744	0.0001	0.0018	0.0018
$\delta = 1$, $\delta^* = 2$	300	2.2352	0.7996	0.0001	0.0013	0.0013
	500	2.2438	0.8000	0.0001	0.0020	0.0020
ENC-LASSO	200	2.0284	0.5680	0.0715	0.5591	0.5591
	300	2.0306	0.6238	0.0922	0.6166	0.6166
	500	2.0524	0.6662	0.1162	0.6629	0.6629
ENC-ALASSO	200	2.3416	0.4582	0.0049	0.0474	0.0474
	300	2.3041	0.5192	0.0091	0.0690	0.0690
	500	2.2097	0.5868	0.0214	0.1660	0.1660

Table 2.12: DGP 8 - Nonparametric Additive Models (Signals, Pseudo-Signals and Hidden Signals), $n = 100$

	P	Loss	TPR	FPR	FDR	FDR*
$p = 0.1$,	200	2.2264	0.7942	0.0006	0.0156	0.0156
$\delta = 1$, $\delta^* = 1.5$	300	2.2286	0.7996	0.0004	0.0139	0.0139
	500	2.2303	0.8000	0.0006	0.0177	0.0177
$p = 0.05$,	200	2.2325	0.7910	0.0003	0.0079	0.0079
$\delta = 1$, $\delta^* = 1.5$	300	2.2340	0.7994	0.0002	0.0065	0.0065
	500	2.2348	0.8000	0.0003	0.0101	0.0101
$p = 0.01$,	200	2.2402	0.7778	0.0001	0.0018	0.0018
$\delta = 1$, $\delta^* = 1.5$	300	2.2408	0.7994	0.0000	0.0012	0.0012
	500	2.2390	0.8000	0.0001	0.0018	0.0018
$p = 0.1$,	200	2.2304	0.7876	0.0006	0.0151	0.0151
$\delta = 1$, $\delta^* = 2$	300	2.2291	0.7994	0.0004	0.0135	0.0135
	500	2.2306	0.8000	0.0006	0.0167	0.0167
$p = 0.05$,	200	2.2377	0.7828	0.0003	0.0075	0.0075
$\delta = 1$, $\delta^* = 2$	300	2.2344	0.7992	0.0002	0.0063	0.0063
	500	2.2355	0.8000	0.0003	0.0094	0.0094
$p = 0.01$,	200	2.2553	0.7552	0.0001	0.0018	0.0018
$\delta = 1$, $\delta^* = 2$	300	2.2408	0.7974	0.0000	0.0010	0.0010
	500	2.2390	0.8000	0.0001	0.0017	0.0017
ENC-LASSO	200	2.0014	0.5292	0.0396	0.6011	0.6011
	300	2.0378	0.5836	0.0489	0.6340	0.6340
	500	2.0573	0.6350	0.0664	0.7012	0.7012
ENC-ALASSO	200	2.3107	0.4472	0.0031	0.0568	0.0568
	300	2.2700	0.4958	0.0047	0.0731	0.0731
	500	2.1763	0.5686	0.0136	0.1991	0.1991

Table 2.13: DGP 8 - Nonparametric Additive Models (Signals, Pseudo-Signals and Hidden Signals), $n = 100$

	P	Loss	TPR	FPR	FDR	FDR*
$p = 0.1$,	200	2.2475	0.7936	0.0005	0.0199	0.0199
$\delta = 1$, $\delta^* = 1.5$	300	2.2263	0.7998	0.0005	0.0201	0.0201
	500	2.2328	0.8000	0.0003	0.0148	0.0148
$p = 0.05$,	200	2.2400	0.7906	0.0003	0.0108	0.0108
$\delta = 1$, $\delta^* = 1.5$	300	2.2316	0.7992	0.0002	0.0109	0.0109
	500	2.2365	0.8000	0.0002	0.0081	0.0081
$p = 0.01$,	200	2.2466	0.7742	0.0000	0.0024	0.0024
$\delta = 1$, $\delta^* = 1.5$	300	2.2360	0.7984	0.0001	0.0025	0.0025
	500	2.2416	0.8000	0.0001	0.0024	0.0024
$p = 0.1$,	200	2.2517	0.7832	0.0005	0.0191	0.0191
$\delta = 1$, $\delta^* = 2$	300	2.2272	0.7984	0.0004	0.0197	0.0197
	500	2.2335	0.8000	0.0003	0.0139	0.0139
$p = 0.05$,	200	2.2483	0.7766	0.0002	0.0103	0.0103
$\delta = 1$, $\delta^* = 2$	300	2.2328	0.7978	0.0002	0.0105	0.0105
	500	2.2370	0.8000	0.0002	0.0075	0.0075
$p = 0.01$,	200	2.2602	0.7454	0.0000	0.0024	0.0024
$\delta = 1$, $\delta^* = 2$	300	2.2374	0.7960	0.0001	0.0025	0.0025
	500	2.2416	0.8000	0.0001	0.0024	0.0024
ENC-LASSO	200	2.0152	0.5140	0.0299	0.6274	0.6274
	300	2.0265	0.5704	0.0370	0.6603	0.6603
	500	2.0628	0.6234	0.0478	0.7185	0.7185
ENC-ALASSO	200	2.3015	0.4388	0.0026	0.0722	0.0722
	300	2.2605	0.4864	0.0056	0.1024	0.1024
	500	2.1438	0.5658	0.0105	0.2163	0.2163

Table 2.14: Generalized Linear Models (Signals, Pseudo-Signals and Hidden Signals)

	P	Loss	TPR	FPR	FDR	FDR*
ENC-OCMT	200	2.2131	0.6910	0.0003	0.0171	0.3154
$p = 0.05$	300	2.2254	0.8826	0.0002	0.0142	0.2978
$\delta = 1$, $\delta^* = 2$	500	2.2172	0.9031	0.0002	0.0117	0.2961
ENC-LASSO	200	2.2325	0.7221	0.0324	0.4149	0.4413
	300	2.2346	0.9014	0.0308	0.4026	0.4292
	500	2.2351	0.9272	0.0289	0.3798	0.4016
ENC-ALASSO	200	2.2278	0.6233	0.0085	0.1975	0.2037
	300	2.2285	0.8652	0.0054	0.1086	0.1221
	500	2.2291	0.8869	0.0031	0.0552	0.0715

$n = 300$, $R = 100$. Note that $n \gg R$.

Table 2.15: Recession Forecasting Using Logistic Function

	OOS Forecast Loss	
	$R = 80$, $P = 114$	$R = 120$, $P = 74$
Lasso	0.3224	0.3128
OCMT ($p = 0.01$)	0.5917	0.5580
OCMT ($p = 0.05$)	0.5913	0.4706
OCMT ($p = 0.1$)	0.5947	0.4726
ENC-Lasso	0.3165	0.3160
ENC-OCMT ($p = 0.01$)	0.2912	0.2873
ENC-OCMT ($p = 0.05$)	0.2681	0.3057
ENC-OCMT ($p = 0.1$)	0.3118	0.2807

Chapter 3

Higher Order Elicitability and Forecast Encompassing for Volatility Forecasts by Bregman Functions

3.1 Introduction

Over the past thirty years, the academic discourse surrounding volatility has been notably expansive. In the contemporary scholarly landscape, there is a discernible surge in investigations specifically aimed at forecasting volatility. Most papers in the literature on volatility forecast comparison, e.g., Hansen and Lunde (2006 JoE) and Patton (2011 JoE), use a loss function that assumes that the conditional mean is zero and use the realized variance as

a proxy for the true volatility. However, the rankings of volatility models can change depending on the validity of the zero mean assumption and the finite sample quality of the realized variance as a proxy.

In statistical analysis, a functional, such as the mean or the quantile, is referred to as elicitable if it possesses the ability to be uniquely derived from the scoring function. When this occurs, the scoring function in question is termed a strictly consistent scoring function for this functional. Gneiting (2011) points out that the mean or expectation is elicitable, but the variance does not hold this characteristic. This implies that if volatility lacks elicitability, we cannot compare or rank the forecasts of volatility. In other words, we cannot do backtesting for the forecast of volatility. Lambert et al. (2008) propose that even when an univariate functional lacks elicitability, there exists the potential for a multivariate vector to be higher-order elicitable. Given that the construct of variance hinges on the mean, and both the mean and expectation are deemed elicitable, it logically follows that a vector incorporating the mean and variance is higher-order elicitable. Thus, it becomes feasible to identify a strictly consistent scoring function that simultaneously addresses the mean and variance. In this paper, we use the strictly consistent scoring function for mean and variance jointly, which is derived from the Bregman divergence function.

Moreover, we found that the commonly used Gaussian predictive likelihood function is a strictly consistent scoring function since it elicits the same conditional mean and variance as the proposed scoring functions from Bregman divergence. This is because the first-order conditions of the Gaussian predictive likelihood function can be shown to be a nonsingular linear transformation of the first-order conditions from the Bregman scoring

functions. The Gaussian log-likelihood function is therefore a strictly consistent scoring function for mean and variance jointly. In addition, we note that the MSE and QLIKE functions of Patton (2011) are special cases of our proposed scoring functions when the mean return is zero and the variance proxy is the squared return.

Out-of-sample forecast comparison is widely used in many fields because it is suggested to test Granger causality, which is used to determine whether some independent variables can predict the dependent variable (Ashley et al., 1980; Diebold and Mariano, 1995). Many papers focus on out-of-sample tests for equal predictive accuracy and encompassing (Diebold and Mariano, 1995; Clark and McCracken, 2001; Clark and West, 2006, 2007; Ge and Lee, 2020). Diebold and Mariano (1995) introduce Diebold-Mariano (DM) statistics for comparing predictive accuracy when two models are not nested. Clark and McCracken (2001) and Clark and West (2006, 2007) prove that DM statistics have a downward bias for mean regression when two models are nested. Ge and Lee (2020) point out that the DM statistics for quantile regression also have a downside bias for two nested models. In this paper, we extend Clark and McCracken (2001), Clark and West (2006, 2007), and Ge and Lee (2020) from a mean or quantile regression to volatility regression. We also find out DM statistic has a bias for two nested mode models. We will show that the ENC statistic has a zero mean, correct the size, and increase the power.

In macroeconomics and finance, numerous macroeconomic and financial variables exist, and most variables have quarterly data, so the number of variables available for testing predictive ability is much larger than the number of observations. Moreover, in most cases, the number of predictors with predictive ability for the variable of interest

increases as the number of variables used in the test increases. Thus, high-dimensional predictive regression models are widely utilized in various fields. This paper focuses on high-dimensional predictive regression models.

Predictor selection often considers only the marginal effect of a predictor. A particular predictor can be selected when the marginal effect is non-zero. However, this method may not identify all relevant predictors in the model. Correlations might exist among predictors, and the probability of correlations among predictors increases as the number of predictors increases. Some predictors with zero marginal effect on the variable of interest (forecast target) but correlated with other predictors may have an indirect impact on the variable of interest. Therefore, predictor selection should consider both the direct marginal effect and the indirect net impact, especially when there are numerous predictors. We extend the multiple testing approach called One Covariate at a Time Multiple Testing (OCMT) proposed by Chudik et al. (2018) for predictor selection. The OCMT approach focuses on the variables' marginal effect and the net impact on the target variable for variable selection using total sample observations. To select predictors in high-dimensional predictive models, we propose a new encompassing approach incorporating the OCMT method, referred to as "ENC-OCMT." Our ENC-OCMT approach considers both marginal effect and net impact for predictor selection. The OCMT procedure selects variables using the in-sample test in the regression model, while our ENC-OCMT approach selects predictors using the out-of-sample test in the predictive regression model.

A widely used variable selection method is the least absolute shrinkage and selection operation (LASSO) by Tibshirani (1996). However, most variable selections using

LASSO are based on full sample observations (Fan and Li, 2001; Efron et al., 2004; Zou and Hastie, 2005; Candes and Tao, 2007; Bickel et al., 2009; Zhang, 2010). We propose a new approach for selecting predictors using the LASSO method, based on the out-of-sample forecast encompassing principle. This new approach is called "ENC-LASSO." Compared to the LASSO method, our ENC-LASSO can be applied to out-of-sample predictor selection.

The primary aim of this study is to forecast stock market volatility. However, there is a discrepancy between the frequency of stock market returns, which are daily, and that of U.S. Macroeconomic and Financial indicators, which are quarterly. To address this issue within the volatility regressions, we utilize the GARCH-MIDAS model, as suggested by Engle et al. (2013). Consequently, our empirical approach involves the application of the ENC-OCMT and ENC-LASSO methods in conjunction with the GARCH-MIDAS model to forecast the volatility of stock market returns.

Our contributions in this paper are fourfold. First, we propose scoring functions that do not require the zero mean assumption or use a proxy for the unknown true variance. We derive a class of strictly consistent scoring functions from which both the mean and the variance are jointly elicitable. This higher order elicitability as referred to as the Osband principle has been further studied in Fissler et al. (2016) that derived a scoring function that jointly elicits a conditional partial moment (expected shortfalls) together with the corresponding conditional quantiles (value-at-risk). Similarly, we adopt the Osband principle to construct a strictly consistent scoring function for the pair of mean and variance using the Bregman divergence measures. We use the proposed strictly consistent scoring functions to jointly evaluate the mean and volatility forecasts, incorporating the unknown conditional

mean (which may be non-zero and time-varying) and without using a proxy for the unknown conditional variance. Second, we also show that the commonly used Gaussian predictive likelihood function is a strictly consistent scoring function since it elicits the same conditional mean and variance as the proposed scoring functions from the Bregman divergence. This is because the first-order conditions of the Gaussian predictive likelihood function can be shown to be a nonsingular linear transformation of the first-order conditions from the Bregman scoring functions. The Gaussian log-likelihood function is therefore a strictly consistent scoring function for mean and variance jointly. In addition, we note that the MSE and QLIKE functions of Patton (2011) are special cases of our proposed scoring functions when the mean return is zero and the variance proxy is the squared return. Third, in order to compare nested volatility forecast models, we develop a new test for Granger Causality based on the forecast encompassing and forecast combination of the competing models, using the proposed Bregman-based strictly scoring functions for a vector of the conditional mean and variance. We derive the asymptotic distribution of the test statistic. Fourth, we propose two approaches to select predictors in high-dimensional predictive regression models. These two approaches are based on the OCMT and LASSO methods under the forecasting encompassing principle. All these theoretical properties of the proposed scoring functions and inference methods are thoroughly examined by Monte Carlo simulation and empirical applications, for which we also consider the GARCH-MIDAS models to compare various predictors in different frequencies. The simulation results demonstrate that the Bregman-based scoring functions produce consistent and robust rankings of the volatility forecast models without having to assume zero mean and without needing to use a proxy.

The simulation results demonstrate that the forecast encompassing statistic to compare the nested volatility forecast models has good size and power in finite samples. In addition, the simulation results also confirm that our two approaches work perfectly in the finite sample size. The empirical results highlight and contrast the merits of the proposed scoring functions and the inference method based on them.

The structure of this paper unfolds as follows: Section 2 revisits the foundational aspects of elicitability, including its definitions, related lemmas, and theorems. The Gaussian log-likelihood function is showcased in Section 3 as a strictly consistent scoring function for both mean and variance. In Section 4, we delve into the forecast encompassing test within volatility regression and reveal a bias within the DM statistic, whilst establishing that the ENC for volatility converges towards asymptotic standard normality. Section 5 proposes two approaches for predictor selection in high-dimensional predictive models. Monte Carlo simulations are implemented in Section 6, showing the encompassing statistic's ability to rectify bias within DM statistics and improve statistical power. Furthermore, the simulations validate the efficacy of our two proposed approaches in finite sample sizes. In section 7, we present empirical analysis. Section 8 is the conclusion.

3.2 Elicitability

We denote an observation domain $\mathrm{O} \subseteq \mathbb{R}^d$, the cumulative distribution of measure F. Let $\mathcal{F}$ be a class of distribution function on the observation domain O and A be an action domain. We define $\Gamma : \mathcal{F} \to \mathrm{A}$ to be a functional.

Definition 1: (Gneiting (2011), Fissler et al. (2016)) A scoring function is an $\mathcal{F}$ intergrable function $S : \mathrm{A} \times \mathrm{O} \to \mathbb{R}$. It is strictly $\mathcal{F}-$consistent for some function Γ if

$$\mathbb{E}_F S(\Gamma(F), Y) = \mathbb{E}_F S(\gamma, Y) \tag{3.1}$$

for all $F \in \mathcal{F}$ and for all $\gamma \in \mathrm{A}$.

Definition 2: (Gneiting (2011), Fissler et al. (2016)) A functional $\Gamma : \mathcal{F} \to \mathrm{A}$ is called elicitable, if there exists a scoring function S that is strictly $\mathcal{F}-$consistent for Γ.

Definition 3: (Gneiting (2011), Fissler et al. (2016)) A identification function is an $\mathcal{F}$ intergrable function $V : \mathrm{A} \times \mathrm{O} \to \mathbb{R}$. It is a strict $\mathcal{F}-$ identificantion function for Γ if $\mathbb{E}V(\gamma, F) = 0$ holds if and only if $x = \Gamma(F)$.

There is a relation between strictly consistent scoring function S and strict identification function V. There is a nonnegative function $h : \mathbb{R} \to \mathbb{R}$ such that

$$\frac{\mathrm{d}}{\mathrm{d}\gamma} S(\gamma, Y) = h(\gamma) V(\gamma, Y) \tag{3.2}$$

Thus, the strict identification function can be obtained by the derivatives of the strictly consistent scoring function. A statistical functional is elicitable if there exists a scoring function that the correct forecast of the functional is the unique minimizer of the expected score (Fissler et al., 2016). Many statistical functionals are elicitable such as expectation, ratios of expectations, quantiles (Value-at-Risk), and expectiles. However, some functional are not elicitable such as variance, mode, or Expected Shortfall (Conditional Value-at-Risk)

(Gneiting, 2011). Those functionals are not elicitable because we cannot find out a correct forecast for each functional that is the unique minimizer of its expected score. Therefore, we cannot do backtest to compare the forecasts of those functionals with their realized scores. Osband (1985) point out that a non-elicitable functional can be a component of an elicitable functional. In other words, for example, variance is not elicitable, but there exists a 2-elicitable functional of mean and variance. Moreover, Fissler et al. (2016) propose a strictly consistent scoring function for joint Value-at-Risk (VaR) and Expected Shortfall (ES).

Definition 4: (Fissler et al. (2016)) A functional $\Gamma : \mathcal{F} \to A \subseteq \mathbb{R}^k$ is called k-elicitable if there exists a strictly $\mathcal{F}-$consistent scoring function for Γ. Let $k_1, \ldots, k_l \geq 1$ and let $\Gamma_m : \mathcal{F} \to A_m \subseteq \mathbb{R}^{k_m}$ be a k_m -elicitable functional, $m \in \{1, \ldots, l\}$. Then the functional $\Gamma = (\Gamma_1, \ldots, \Gamma_l) : \mathcal{F} \to \mathrm{A}$ is k elicitable where $k = k_1 + \cdots + k_l$ and $A = A_1 \times \cdots \times A_l \subseteq \mathbb{R}^k$.

Theorem 1: (Osband's principle)(Fissler et al. (2016)) Let $\Gamma : \mathcal{F} \to \mathrm{A} \subseteq \mathbb{R}^k$ be a surjective, elicitable and identifiable functional with a strict $\mathcal{F}$-identification function $V : \mathrm{A} \times \mathrm{O} \to \mathbb{R}^k$ and a strictly $\mathcal{F}$ -consistent scoring function $S : \mathrm{A} \times \mathrm{O} \to \mathbb{R}$. If the $\mathbb{E}S(\gamma, F)$ is continuously differentiable, there exists a matrix-valued function $h : \mathrm{int}(\mathrm{A}) \to \mathbb{R}^{k \times k}$ such that for $l \in \{1, \ldots, k\}$

$$\partial_l \mathbb{E}S(\gamma, F) = \sum_{m=1}^{k} h_{lm}(\gamma) \mathbb{E}V_m(\gamma, F) \tag{3.3}$$

for all $\gamma \in \mathrm{int}(\mathrm{A})$ and $F \in \mathcal{F}$. Theorem 6 shows the relationship between a strictly consistent scoring function and strict identification function for many γ.

Strictly consistent scoring functions for a given functional Γ are not unique. In particular, the following result generalizes directly from the one-dimensional case. Let $S : A \times O \to \mathbb{R}$ be a strictly $\mathcal{F}$ -consistent scoring function for a functional $\Gamma : \mathcal{F} \to \mathrm{A}$. Then, for any $\lambda > 0$ and any $\mathcal{F}$ -integrable function $a : \mathrm{O} \to \mathbb{R}$, the scoring function

$$\widetilde{S}(\gamma, y) := \lambda S(\gamma, y) + a(y)$$

is again strictly $\mathcal{F}$-consistent for Γ. Thus, $a(y)$ or any function of y is not necessary for the strictly consistent function.

Osband (1985) points out that a non-elicitable functional can be a component of an elicitable functional. Gneiting (2011) states that variance alone is not elicitable, but there exists a 2-elicitable functional of joint mean and variance. Fissler et al. (2016); Brehmer (2017) explains how to get the strictly consistent scoring function of mean and variance. In order to find out the strictly consistent scoring function of joint mean and variance, we need to use the bijection function and the strictly consistent scoring function of the ratios of the expectation. To get the strictly consistent scoring function of the ratios of the expectation, we need to use Theorem 2.

Theorem 2: (Gneiting (2011)) Consider a measurable weight function $w : D \to [0, \infty)$. Let $\mathcal{F}^{(w)} \subseteq \mathcal{F}$ denote the subclass of the probability distributions in $\mathcal{F}$ which are such that $w(y)f(y)$ has finite integral over D, and the probability measure $F^{(w)}$ with density proportional to $w(y)f(y)$ belongs to $\mathcal{F}$. Define the functional $\Gamma^{(w)} : \mathcal{F}^{(w)} \longrightarrow \mathcal{P}(\mathrm{I}), \quad F \longmapsto \Gamma^{(w)}(F) = \Gamma\left(F^{(w)}\right)$ on this subclass $\mathcal{F}^{(w)}$. Then the following holds:

(a) If Γ is elicitable, then $\Gamma^{(w)}$ is elicitable.

(b) If S is consistent for Γ relative to $\mathcal{F}$, then

$$S^{(w)}(\gamma, y) = w(y)S(\gamma, y)$$

is consistent for $\Gamma^{(w)}$ relative to $\mathcal{F}^{(w)}$

(c) If S is strictly consistent for Γ relative to $\mathcal{F}$, then $S^{(w)}$ is strictly consistent for $\Gamma^{(w)}$ relative to $\mathcal{F}^{(w)}$

Gneiting (2011) show that the functional of the ratios of expectations $\Gamma(F) = \frac{\mathbb{E}_F[r(Y)]}{\mathbb{E}_F[s(Y)]}$ is elicitable. Let $s(y) = w(y)$ and $r(y) = yw(y)$. The strictly consistent scoring function is of the form

$$S(x, y) = s(y)(f(y) - f(\gamma)) - f'(\gamma)(r(y) - \gamma s(y)) + f'(y)(r(y) - ys(y)), \tag{3.4}$$

where f is a strictly convex and differential function.

For variance, $\mathrm{Var}(Y) = \mathbb{E}Y^2 - (\mathbb{E}Y)^2$. Let Γ_1 be the mean functional and Γ_2 be the second moment functional, which we can get from the ratios of expectations with $r(y) = y^2$ and $s(y) = 1$. Then, we have $S(\gamma, y) = f_2(y) - f_2(\gamma) - f_2'(x)(y^2 - \gamma)$.

Proposition 1: (Fissler et al. (2016)) Let λ_m be positive real numbers, $m \in \{1, \ldots, l\}$. The scoring function S is strictly $\mathcal{F}-$consistent for Γ if and only if it is of the form

$$S(\gamma_1, \ldots, \gamma_k, y) := \sum_{m=1}^{l} \lambda_k S_m(\gamma_m, y), \tag{3.5}$$

where $S_m : A_m \times O \to \mathbb{R}$ are strictly $\mathcal{F}$ -consistent scoring function for Γ_m. Proposition 5 indicates that the summation of strictly $\mathcal{F}$ -consistent scoring function is a strictly $\mathcal{F}$ -consistent scoring function.

We can get the strictly consistent scoring function of $(\Gamma_1, \Gamma_2)^\top$ is

$$S(\gamma_1, \gamma_2, y) := -f_1(\gamma_1) - f_1'(\gamma_1)(y - x_1) - f_2(\gamma_2) - f_2'(\gamma_2)\left(y^2 - \gamma_2\right) + a(y),$$

where f_1 and f_2 are strictly convex and differentiable functions.

Theorem 3: (Osband (1985)). Suppose that the class $\mathcal{F}$ is concentrated on the domain A, and let $g : A \to A'$ be a one-to-one mapping. Then the following holds: (a) If Γ is elicitable, then $\Gamma_g = g \circ \Gamma$ is elicitable. (b) If S is consistent for Γ, then the scoring function

$$S_g(x, y) = S\left(g^{-1}(\gamma), y\right)$$

is consistent for Γ_g (c) If S is strictly consistent for Γ, then S_g is strictly consistent for Γ_g

Define the sets $A := \left\{(\gamma_1, \gamma_2) \mid \gamma_2 \geq \gamma_1^2\right\}$ and $A' := \mathbb{R} \times [0, \infty) \subset \mathbb{R}^2$ with a bijection $g : A \to A'$ given by $(\gamma_1, \gamma_2) \mapsto \left(\gamma_1, \gamma_2 - \gamma_1^2\right)$. According to Theorem 3, the inverse of g is given by $g^{-1} : A' \to \text{A}$, $(\gamma_1, \gamma_2) \mapsto \left(\gamma_1, \gamma_2 + \gamma_1^2\right)$. Define Γ_3 be the variance functional. The functional we want to get is $(\Gamma_1, \Gamma_3)^\top$ can now be written as $g((\Gamma_1, \Gamma_2))$ and it is elicitable. Then, we transfer from $(\Gamma_1, \Gamma_2)^\top$ to $(\Gamma_1, \Gamma_{\text{var}})^\top$, and get the strictly consistent scoring function of $(\Gamma_1, \Gamma_{\text{var}})^\top$ as follows:

$$\begin{aligned} S_g(\gamma_1, \gamma_2, y) &= S\left(g_1^{-1}(\gamma_1, \gamma_2), g_2^{-1}(\gamma_1, \gamma_2) y\right) \\ &= -f_1(\gamma_1) - f_1'(\gamma_1)(y - \gamma_1) - f_2\left(\gamma_2 + \gamma_1^2\right) - f_2'\left(\gamma_2 + \gamma_1^2\right)\left[y^2 - \left(\gamma_2 + \gamma_1^2\right)\right] + a(y), \end{aligned} \tag{3.6}$$

where f_1 and f_2 are strictly convex and differential functions.

Proposition 2: Denote S_g as the loss function when $f_1(z) = z$ and $f_2(z) = \log(z)$. For easy to read, let μ replace γ_1 and σ^2 replace γ_2. The "S_g" loss function is as follows:

$$S_g(y, \mu, \sigma^2) = (y - \mu)^2 - \log(y) + \log(\mu^2 + \sigma^2) + \frac{y}{\mu^2 + \sigma^2} - 1. \tag{3.7}$$

We will use this scoring function to test Granger Causality in volatility.

Strict identification functions are obtained from the first-order conditions of the strictly consistent scoring function with respect to the functionals,

$$\mathbb{E}\mathbf{m} \equiv \frac{\partial \mathbb{E}S(\boldsymbol{\gamma}, Y)}{\partial \boldsymbol{\gamma}} = \mathbf{h}(\boldsymbol{\gamma})\mathbb{E}\mathbf{V}(\boldsymbol{\gamma}, Y) = 0,$$

$$\mathbb{E}m_1 \equiv \frac{\partial \mathbb{E}S(\gamma_1, \gamma_3, Y)}{\partial \gamma_1} = h_{11}(\gamma_1, \gamma_3)\mathbb{E}V_1(\gamma_1, \gamma_3, Y) + h_{12}(\gamma_1, \gamma_3)\mathbb{E}V_2(\gamma_1, \gamma_3, Y) = 0,$$

$$\mathbb{E}m_2 \equiv \frac{\partial \mathbb{E}S(\gamma_1, \gamma_3, Y)}{\partial \gamma_3} = h_{21}(\gamma_1, \gamma_3)\mathbb{E}V_1(\gamma_1, \gamma_3, Y) + h_{22}(\gamma_1, \gamma_3)\mathbb{E}V_2(\gamma_1, \gamma_3, Y) = 0,$$

where

$$\mathbf{m} = [m_1, m_2]', \qquad \boldsymbol{\gamma} = [\gamma_1, \gamma_3]',$$

$$\mathbf{V}(\gamma_1, \gamma_3, Y) = \begin{bmatrix} V_1(\gamma_1, \gamma_3, Y) \\ \\ V_2(\gamma_1, \gamma_3, Y) \end{bmatrix},$$

$$\mathbf{h}(\gamma_1, \gamma_3) = \begin{bmatrix} h_{11}(\gamma_1, \gamma_3) & h_{12}(\gamma_1, \gamma_3) \\ \\ h_{21}(\gamma_1, \gamma_3) & h_{22}(\gamma_1, \gamma_3) \end{bmatrix}.$$

The $\mathbf{V}(\gamma_1, \gamma_3, Y)$ is the vector of strict identification functions for mean and variance. The $\mathbf{h}(\gamma_1, \gamma_3)$ is the matrix of the non-negative functions of functionals. The $\mathbf{m}$ is the martingale difference, which is the product of the strict identification functions $\mathbf{V}$ and the non-negative functions $\mathbf{h}$.

The loss function for mean and variance from Bregman divergence is

$$S_g\left(\gamma_1, \gamma_3, y\right) = \left(y - \gamma_1\right)^2 - \log(y^2) + \log\left(\gamma_3 + \gamma_1^2\right) + \frac{y^2}{\gamma_3 + \gamma_1^2} - 1.$$

The martingale difference is the first-order condition of the strictly consistent scoring function with respect to the functional.

$$\mathbb{E}m_{g1} = \frac{\partial \mathbb{E}S_g}{\partial \gamma_1} = -2\left[\mathbb{E}Y - \gamma_1\right] + \frac{2\gamma_1}{\gamma_3 + \gamma_1^2} - \frac{2\gamma_1 \mathbb{E}\left(Y^2\right)}{\left(\gamma_3 + \gamma_1^2\right)^2} = 0$$

$$\mathbb{E}m_{g2} = \frac{\partial \mathbb{E}S_g}{\partial \gamma_3} = \frac{1}{\gamma_3 + \gamma_1^2} - \frac{\mathbb{E}\left(Y^2\right)}{\left(\gamma_3 + \gamma_1^2\right)^2} = 0$$

The martingale difference is the product of the identification function and the function of functionals.

$$\begin{aligned}
\mathbb{E}\mathbf{m} &= \mathbf{h}_g(\gamma_1, \gamma_3)\mathbb{E}\mathbf{V}_g(\gamma_1, \gamma_3, y) \\
&= \begin{bmatrix} h_{g11}(\gamma_1, \gamma_3) & h_{g12}(\gamma_1, \gamma_3) \\ h_{g21}(\gamma_1, \gamma_3) & h_{g22}(\gamma_1, \gamma_3) \end{bmatrix} \begin{bmatrix} \mathbb{E}V_{g1}(\gamma_1, \gamma_3, Y) \\ \mathbb{E}V_{g2}(\gamma_1, \gamma_3, Y) \end{bmatrix} \\
&= \begin{bmatrix} -2 & \frac{2\gamma_1}{(\gamma_3+\gamma_1)^2} \\ 0 & \frac{1}{(\gamma_3+\gamma_1)^2} \end{bmatrix} \begin{bmatrix} \mathbb{E}Y - \gamma_1 \\ \gamma_3 + \gamma_1^2 - \mathbb{E}Y^2 \end{bmatrix} = 0 \qquad \Rightarrow \qquad \begin{bmatrix} \gamma_1 \\ \gamma_3 \end{bmatrix} = \begin{bmatrix} \mathbb{E}Y \\ \mathbb{E}Y^2 - (\mathbb{E}Y)^2 \end{bmatrix}
\end{aligned}$$

3.3 Gaussian Log-likelihood Function

The Gaussian log-likelihood function is also a strictly consistent scoring function for mean and variance. Both the Bregman loss function and Gaussian log-likelihood function for mean and variable can elicit the same strict identification functions of mean and variable.

$$S_L\left(\gamma_1, \gamma_3, y\right) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\left(\gamma_3\right) - \frac{\left(y - \gamma_1\right)^2}{2\gamma_3}.$$

$$\mathbb{E}m_{L1} = \frac{\partial \mathbb{E}S_L}{\partial \gamma_1} = \frac{\mathbb{E}Y - \gamma_1}{\gamma_3} = 0$$
$$\mathbb{E}m_{L2} = \frac{\partial \mathbb{E}S_L}{\partial \gamma_3} = -\frac{1}{2\gamma_3} + \frac{\mathbb{E}\left(Y - \gamma_1\right)^2}{2\gamma_3^2} = 0$$

$$\begin{aligned}
\mathbb{E}\mathbf{m} &= \mathbf{h}_L(\gamma_1, \gamma_3)\mathbb{E}\mathbf{V}_L(\gamma_1, \gamma_3, y) \\
&= \begin{bmatrix} h_{L11}(\gamma_1, \gamma_3) & h_{L12}(\gamma_1, \gamma_3) \\ h_{L21}(\gamma_1, \gamma_3) & h_{L22}(\gamma_1, \gamma_3) \end{bmatrix} \begin{bmatrix} \mathbb{E}V_{L1}(\gamma_1, \gamma_3, Y) \\ \mathbb{E}V_{L2}(\gamma_1, \gamma_3, Y) \end{bmatrix} \\
&= \begin{bmatrix} \frac{1}{\gamma_3} & 0 \\ 0 & -\frac{1}{2\gamma_3^2} \end{bmatrix} \begin{bmatrix} \mathbb{E}Y - \gamma_1 \\ \gamma_3 - \mathbb{E}\left(Y - \gamma_1\right)^2 \end{bmatrix} = 0 \qquad \Rightarrow \qquad \begin{bmatrix} \gamma_1 \\ \gamma_3 \end{bmatrix} = \begin{bmatrix} \mathbb{E}Y \\ \mathbb{E}\left(Y - \mathbb{E}Y\right)^2 \end{bmatrix}
\end{aligned}$$

3.4 Forecast Encompassing Approach

In this section, we introduce the models, describe the encompassing test for the volatility predictive regression, and demonstrate that the encompassing statistic used in this paper is asymptotically normal under certain assumptions.

3.4.1 Model

Chudik et al. (2018) proposed a multiple testing approach called One Covariate at a time Multiple Testing (OCMT) for variable selection. The OCMT procedure can be applied only to linear models. In this paper, we propose a new approach for the out-of-sample predictive ability test to select predictors. Since the new approach is based on the encompassing principle and OCMT procedure, we first develop the framework of the forecast encompassing approach. To select predictors in high-dimensional predictive regression, we

consider the linear model in the below equation:

$$y_{t+1} = \mu_{t+1} + u_{t+1} \tag{3.8}$$

$$\sigma^2_{t+1} = \omega + \alpha u_t^2 + \beta \sigma_t^2 + \sum_{i=1}^{n} \delta_i x_{i,t}^2 \tag{3.9}$$

where $x_{i,t}$ are the n predictors, for $i = 1, \ldots, n$, with n being the number of predictors.

As the OCMT procedure is a multiple testing approach and selects predictors one at a time, we use low-dimensional predictive regression models for the forecast encompassing test to apply the OCMT procedure based on the encompassing principle. We have n + 1 linear regressions: one model without conditioning on any predictors, and the other models conditioned on predictors. We consider n different predictors x_i, $i = 1, \ldots, n$ for n models,, examining one predictor at a time. These nested volatility models are

$$\text{Model } 0: \quad y_{t+1} = \mu_{0,t+1} + u_{0,t+1} \tag{3.10}$$

$$\sigma^2_{0,t+1} = \omega_0 + \alpha_0 u^2_{0,t} + \beta_0 \sigma^2_{1,t} \tag{3.11}$$

$$\text{Model } i: \quad y_{t+1} = \mu_{i,t+1} + u_{i,t+1} \tag{3.12}$$

$$\sigma^2_{i,t+1} = \omega_i + \alpha_2 u^2_{i,t} + \beta_i \sigma^2_{i,t} + \delta_i x^2_{i,t}, \qquad i = 1, \ldots, n \tag{3.13}$$

where unconditional volatility is $\sigma^2_{0,t+1}$ and conditional volatilities are $\sigma^2_{i,t+1}$. The dependent variable y_{t+1} is a scalar. We estimate $\hat{\sigma}^2_{0,t+1}$ and $\hat{\sigma}^2_{i,t+1}$ by following $\hat{\sigma}^2_{0,t+1} = \hat{\omega}_{0,t} + \hat{\alpha}_{0,t} u^2_{0,t} + \hat{\beta}_{0,t} \sigma^2_{0,t}$ and $\hat{\sigma}^2_{i,t+1} = \hat{\omega}_{i,t} + \hat{\alpha}_{i,t} u^2_{i,t} + \hat{\beta}_{i,t} \sigma^2_{i,t} + \hat{\delta}_i x^2_{i,t}$, where $\hat{\omega}_{0,t}, \hat{\alpha}_{0,t}, \hat{\beta}_{0,t}, \hat{\omega}_{i,t}, \hat{\alpha}_{i,t}, \hat{\beta}_{i,t}$ and $\hat{\delta}_{i,t}$ are estimated based on in-sample observations minimizing the loss function.

We use a large number of predictors $x_{i,t}$ in our regression models. The number of predictors is larger than the number of observations. In order to use the out-of-sample forecast encompassing approach for the multiple testing methods in high-dimensional mod-

els, we use one predictor $x_{i,t}$ at a time for Model i. We check if the predictor $x_{i,t}$ has the predictive ability for the dependent variable y_{t+1}. If the predictor $x_{i,t}$ has the predictive ability for the dependent variable y_{t+1}, we select this predictor and use this predictor in the regression to forecast y_{t+1}. We use nested regression models with one predictor $x_{i,t}$ at a time in the large models, so our out-of-sample forecast encompassing approach is an out-of-sample Granger Causality test. If the predictor $x_{i,t}$ has the predictive ability for the dependent variable y_{t+1}, this predictor Granger causes y_{t+1} in volatility regression.

3.4.2 Out-of-sample forecast encompassing test

We now introduce the out-of-sample forecast encompassing test for the predictive ability of a predictor. The models in the equations (3.11) and (3.13) are nested. Due to the asymptotic non-normality of DM statistic for the nested models, we may not use the Diebold Mariano (DM) statistic by Diebold and Mariano (1995) to test the predictive ability of $x_{i,t}$ for y_{t+1}.

Thus, we develop the encompassing statistic and prove that the encompassing statistic is asymptotically normal. In order to find out the encompassing statistic for the volatility, we combine two models (Model 0 and Model i) with weights $(1-\lambda_i)$ and λ_i. We define the score function for the combined model i as a derivative of the expectation of loss function for the combined model with respect to the weight λ_i. The score function is the encompassing statistic we developed using the encompassing principle.

Based on the out-of-sample forecast encompassing principle, we combine Model 0 and Model i with weight $(1-\lambda_i)$ and λ_i, respectively. Then, the null and alternative

hypotheses are

$$\mathbb{H}_0 : \lambda_i = 0 \quad \text{and} \quad \mathbb{H}_1 : \lambda_i \neq 0.$$

Under the null hypothesis $\lambda_i = 0$, the combined model is Model 0, which means that $x_{i,t}$ does not Granger cause y_{t+1}. Under the alternative $\lambda_i \neq 0$, the combined model combines Model 0 and Model i with weights $(1 - \lambda_i)$ and λ_i, which implies that x_t Granger causes y_{t+1}. As mentioned in the previous subsection, we consider the out-of-sample encompassing test, so x_t has the predictive ability for y_{t+1} under the alternative hypothesis.

To find out the optimal weight λ_i and get the encompassing statistic under the null hypothesis, we estimate the weight λ_i minimizing the expectation of least squares loss function with combined mean $\mu_{c,t+1}$. Thus, we have

$$\lambda_i = \arg\min_{\lambda_i} \mathbb{E}S(\mu_{c,i,t+1}, \sigma^2_{c,i,t+1}, Y_{t+1}),$$

where $\mu_{c,i,t+1} = \mu_{0,t+1} = \mu_{i,t+1}$, $\sigma^2_{c,i,t+1} = (1 - \lambda_i)\,\sigma^2_{0,t+1} + \lambda_i \sigma^2_{i,t+1}$, $S(\mu_{c,i,t+1}, \sigma^2_{c,i,t+1}, y_{t+1})$ $= (y_{t+1} - \mu_{c,i,t+1})^2 - \log(y^2_{t+1}) + \log\left(\mu^2_{c,i,t+1} + \sigma^2_{c,i,t+1}\right) + \frac{y^2}{\mu^2_{c,i,t+1} + \sigma^2_{c,i,t+1}} - 1$ is the Bregman loss function. The encompassing principle is based on the score function with respect to λ_i. We define C_i as the score function that takes a derivative of the expectation of loss function for a combined model with respect to the weight λ. Thus, C_i is the score function for Model i, which is defined as

$$C_i \equiv \frac{\partial \mathbb{E}\left[S(\mu_{c,i,t+1}, \sigma^2_{c,i,t+1}, Y_{t+1})\right]}{\partial \lambda_i} = 0.$$

This score function can be separated into four parts. The first part is a derivative of the loss function with respect to the conditional mean for the combined model $\mu_{c,i,t+1}$. Then we refer to the first derivative as the generalized residual function of the conditional mean for

the combined model. We use $h_{1,i,t+1}V_{1,i,t+1}$ to denote the generalized residual function of the conditional mean for Model i. The second part is a derivative of the conditional mean for the combined model $\mu_{c,i,t+1}$ with respect to weight λ_i. We call the second derivative to be the test function of the conditional mean. The test function of the conditional mean is denoted as $\xi_{1,i,t+1}$ for the Model i. The third part is a derivative of the loss function with respect to the conditional volatility for the combined model $\sigma^2_{c,i,t+1}$. Then we refer to the third derivative as the generalized residual function of the conditional volatility for the combined model. We use $h_{2,i,t+1}V_{2,i,t+1}$ to denote the generalized residual function of the conditional volatility for Model i. The fourth part is a derivative of the conditional mean for the combined model $\sigma^2_{c,i,t+1}$ with respect to weight λ_i. We call the fourth derivative to be the test function of the conditional volatility. The test function of the conditional volatility is denoted as $\xi_{2,i,t+1}$ for the Model i.Thus, the score function C_i for Model i is the expectation of the sum of the product of identification function $h_{1,i,t+1}V_{1,i,t+1}$ and test function $\xi_{1,i,t+1}$ and the product of identification function $h_{2,i,t+1}V_{2,i,t+1}$ and test function $\xi_{2,i,t+1}$. Therefore, the score function C_i can be derived as follows:

$$
\begin{aligned}
C_i &\equiv \frac{\partial \mathbb{E}\left[S(\mu_{c,i,t+1}, \sigma^2_{c,i,t+1}, Y_{t+1})\right]}{\partial \lambda_i} \\
&= \mathbb{E}\left[\frac{\partial S(\mu_{c,i,t+1}, \sigma^2_{c,i,t+1}, Y_{t+1})}{\partial \mu_{c,i,t+1}} \frac{\partial \mu_{c,i,t+1}}{\partial \lambda_i} + \frac{\partial S(\mu_{c,i,t+1}, \sigma^2_{c,i,t+1}, Y_{t+1})}{\partial \sigma^2_{c,i,t+1}} \frac{\partial \sigma^2_{c,i,t+1}}{\partial \lambda_i}\right]
\end{aligned}
$$

Let $h_{1,i,t+1}V_{1,i,t+1} = \frac{\partial S(\mu_{c,i,t+1}, \sigma^2_{c,i,t+1}, Y_{t+1})}{\partial \mu_{c,i,t+1}}$ as the generalized residual of the conditional mean, $\nu_{1,i,t+1} = \frac{\partial \mu_{c,i,t+1}}{\partial \lambda_i}$ as the test function of the conditional mean, $h_{2,i,t+1}V_{2,i,t+1} = \frac{\partial S(\mu_{c,i,t+1}, \sigma^2_{c,i,t+1}, Y_{t+1})}{\partial \sigma^2_{c,i,t+1}}$ as the generalized residual of the conditional volatility, $\nu_{2,i,t+1} = \frac{\partial \sigma^2_{c,i,t+1}}{\partial \lambda_i}$ as the test function of the conditional volatility. Note that $h_{1,i,t+1}V_{1,i,t+1} = -2(y_{t+1} - \mu_{c,i,t+1}) + \frac{2\mu_{c,i,t+1}}{\mu^2_{c,i,t+1} + \sigma^2_{c,i,t+1}} - \frac{2\mu_{c,i,t+1} y^2_{t+1}}{(\mu^2_{c,i,t+1} + \sigma^2_{c,i,t+1})^2}$, $\nu_{i,t+1} = \mu_{i,t+1} - \mu_{0,t+1}$, $h_{2,i,t+1}V_{2,i,t+1} =$

$\frac{1}{\mu_{c,i,t+1}^2+\sigma_{c,i,t+1}^2} - \frac{y_{t+1}^2}{(\mu_{c,i,t+1}^2+\sigma_{c,i,t+1}^2)^2}$, and $\nu_{i,t+1} = \sigma_{i,t+1}^2 - \sigma_{0,t+1}^2$. We assume that $\mu_{c,i,t+1} = \mu_{i,t+1} = \mu_{0,t+1}$, so we have $\mu_{i,t+1} - \mu_{0,t+1} = 0$. If $\lambda_i = 0$, $\sigma_{c,i,t+1}^2 = \sigma_{0,t+1}^2$, $h_{2,i,t+1} = h_{2,0,t+1}$ and $V_{2,i,t+1} = V_{2,0,t+1}$. Thus, under the null hypothesis, we have the score function C_i for Model i to be

$$
\begin{aligned}
C_i = \mathbb{E}\Bigg[& \left(-2(y_{t+1} - \mu_{c,i,t+1}) + \frac{2\mu_{c,i,t+1}}{\mu_{c,i,t+1}^2 + \sigma_{c,i,t+1}^2} - \frac{2\mu_{c,i,t+1} y_{t+1}^2}{(\mu_{c,i,t+1}^2 + \sigma_{c,i,t+1}^2)^2} \right) (\mu_{i,t+1} - \mu_{0,t+1}) \\
& \left(\frac{1}{\mu_{c,i,t+1}^2 + \sigma_{c,i,t+1}^2} - \frac{y_{t+1}^2}{(\mu_{c,i,t+1}^2 + \sigma_{c,i,t+1}^2)^2} \right) (\sigma_{i,t+1}^2 - \sigma_{0,t+1}^2) \Bigg] \\
= \mathbb{E} & \left(\frac{1}{\mu_{0,t+1}^2 + \sigma_{0,t+1}^2} - \frac{y_{t+1}^2}{(\mu_{0,t+1}^2 + \sigma_{0,t+1}^2)^2} \right) (\sigma_{i,t+1}^2 - \sigma_{0,t+1}^2) \\
= \mathbb{E} & \left(h_{2,0,t+1} V_{2,0,t+1} \right) (\nu_{2,i,t+1}) = 0.
\end{aligned}
$$

Based on the out-of-sample forecast encompassing principle, we separate the observations into two parts: in-sample observations and out-of-sample observations. The number of in-sample observations is defined as R, and the number of out-of-sample observations is defined as P. The total number of observations is $R + P = T + 1$. We estimate $\hat{\omega}_0, \hat{\alpha}_0, \hat{\beta}_0, \hat{\omega}_i, \hat{\alpha}_i, \hat{\beta}_i$ and $\hat{\delta}_i$ by using the in-sample observations from 1 to R. And then, we estimate $\hat{\sigma}_{0,t+1}^2 = \hat{\omega}_0 + \hat{\alpha}_0 u_{0,t}^2 + \hat{\beta}_0 \sigma_{0,t}^2$ and $\hat{\sigma}_{i,t+1}^2 = \hat{\omega}_i + \hat{\alpha}_i u_{i,t}^2 + \hat{\beta}_0 \sigma_{i,t}^2 + \hat{\delta}_i x_{i,t}$ by using out-of-sample observations from $R + 1$ to $T + 1$. Then, we use one step ahead method to estimate C_i by $\hat{C}_{i,R,P}$ using the out-of-sample observations from $R + 1$ to $T + 1$. The $\hat{C}_{i,R,P}$ is defined as

$$
\hat{C}_{i,R,P} \equiv P^{-1} \sum_{t=R}^{T} \left(\hat{h}_{2,0,t+1} \hat{V}_{2,0,t+1} \hat{\xi}_{2,i,t+1} \right) \xrightarrow{p} C_i = 0 \tag{3.14}
$$

under $\mathbb{H}_0$, as $R, P \to \infty$ and $P/R \to \infty$. Under the null hypothesis, $\lambda_i = 0$, we obtain $\sigma_{c,i,t+1}^2 = \sigma_{0,t+1}^2$, $h_{2,i,t+1} = h_{2,0,t+1}$ and $V_{2,i,t+1} = V_{2,0,t+1}$. Thus, $\mathbb{E}(\hat{C}_{i,R,P}) = 0$, so $\hat{C}_{i,R,P} \xrightarrow{p} 0$ as $R, P \to \infty$.

In order to get the encompassing statistic for the volatility, we consider endowing Model 0 and Model i with the weight $1-\lambda_i$ and λ_i, respectively, and find out the property of optimal weight λ_i. In order to test if $x_{i,t}$ has the predictive ability on y_{t+1}, we standardize the $\hat{C}_{i,R,P}$ to the ENC statistic. The ENC statistic is $\text{ENC}_{i,R,P} \equiv \hat{Q}_{i,R,P}^{-0.5}\sqrt{P}\hat{C}_{i,R,P}$, where $Q_{i,R,P} = \text{var}\left(\sqrt{P}\hat{C}_{i,R,P}\right)$ and $Q_{i,R,P} - \hat{Q}_{i,R,P} \xrightarrow{p} 0$.

3.4.3 Asymptotic Normality of ENC

We prove the asymptotic normality property of the ENC statistic. The following assumptions are used to obtain the limiting distribution in Proposition 2 and its asymptotic normality property in Proposition 3. These assumptions are only sufficient but not necessary and sufficient.

Assumption 1: The parameter estimates $\hat{\beta}_{i,t}, i = 1, 2, t = R, \ldots, T$, satisfy $\hat{\beta}_{i,t} - \beta_i = B_i(t)H_i(t) = \left(R^{-1}\sum_{j=t-R+1}^{t} q_{i,j}\right)^{-1}\left(R^{-1}\sum_{j=t-R+1}^{t} h_{i,j}\right)$.

Assumption 2: Let $U_t = \left[h_{2,t}V_{2,t}, x'_{2,t} - \mathbb{E}x'_{2,t}, h'_{2,t}, \text{vec}\left(h_{2,t}h'_{2,t} - \mathbb{E}h_{2,t}h'_{2,t}\right), \text{vec}\left(q_{2,t} - \mathbb{E}q_{2,t}\right)'\right]$ (a) $\mathbb{E}U_t = 0$, (b) $\mathbb{E}q_{2,t} < \infty$ is p.d., (c) $\mathbb{E}u_t^2 = \sigma^2$. (d) Define $\tilde{U}_t$ the nonredundant elements of U_t, then $R^{-1}\mathbb{E}\left(\sum_{j=t-R+1}^{t}\tilde{U}_j\right)\left(\sum_{j=t-R+1}^{t}\tilde{U}_j\right)' = \Omega < \infty$ is positive definite.

Assumption 3: (a) $\mathbb{E}h_{2,t}h'_{2,t} = \sigma^2\mathbb{E}q_{2,t}$, (b) $\mathbb{E}\left(h_{2,t} \mid h_{2,t-j}, q_{2,t-j}, j = 1, 2, \ldots\right) = 0$.

Assumptions 2 and 3 allow the application of an invariance principle and are sufficient for joint weak convergence of partial sums and averages of these partial sums to Brownian motion and integrals of Brownian motion.

Assumption 4: $\lim_{P,R\to\infty} P/R \equiv \pi \to \infty$.

In order to get accurate out-of-sample forecasts, it is essential to select an optimal in-sample window. However, no consensus opinion has been reached as it depends on the characteristic of the model. Inoue et al. (2017) find out the optimal window size by minimizing the conditional mean squared forecast error in a time-varying predictive regression model. They show that the optimal window size satisfies $R = O(T^{2/3})$. Hong et al. (2020) consider minimizing various out-of-sample forecast errors, including the unconditional, conditional, and global mean squared forecast errors, respectively. The optimal window size they found is $R = O(T^{4/5})$. The out-of-sample window P has a faster divergent rate than in sample window R. Thus, in our model, we set $P/R \to +\infty$.

We have the following main propositions:

Proposition 2: Suppose Assumptions 1-3 hold. Then

$$\begin{aligned}\mathrm{ENC}_P =& \frac{\sum_{t=R}^{T} c_{t+1}}{\sqrt{\sum_{t=R}^{T} c_{t+1}^2 - P\bar{C}^2}} \\ \xrightarrow{d}& \frac{\int_{\xi}^{1} \left[W(s) - W(s-\xi)\right] \mathrm{d}W(s)}{\sqrt{\int_{\xi}^{1} \left[W(s) - W(s-\xi)\right]^2 \mathrm{d}s}}\end{aligned}$$

under $\mathbb{H}_0$, where $\xi = R/T$ and $c_{t+1} = \left(\hat{h}_{2,0,t+1}\hat{V}_{2,0,t+1}\hat{\nu}_{2,i,t+1}\right)$.

Proof: See Appendix A.

Proposition 3: Suppose Assumptions 1-4 hold. Then $\mathrm{ENC}_P \xrightarrow{d} N(0,1)$ under $\mathbb{H}_0$.

Proof: See Appendix A

$$\mathrm{ENC}_P = \frac{\sqrt{P}\hat{C}_P}{\sqrt{\mathrm{avar}\left(\sqrt{P}\hat{C}_P\right)}} \xrightarrow{d} N(0,1) \text{ under } \mathbb{H}_0 \text{ as } R, P \to \infty.$$

3.5 Predictor Selection in High-dimensional Predictive Models

In most cases, the number of predictors that have the predictive ability on the variable of interest increases as the number of variables we used in the test increases. Thus, the high-dimensional regression model is highly used in many fields. This section proposes a new approach using the OCMT procedure under the encompassing principle. We refer to this approach as ENC-OCMT. This approach extends the OCMT procedure from full-sample variable selection to out-of-sample predictive selection and the ENC statistic from the low-dimensional models (one predictor in a model) to high-dimensional models (many predictors in the model). Thus, the ENC-OCMT can be used for the out-of-sample predictive ability test to select predictors in high-dimensional predictive models.

3.5.1 OCMT Procedure for Out-of-sample Forecast Encompassing Test

Chudik et al. (2018) proposed a new approach for variable selection, which is called OCMT. The OCMT procedure can be applied to in-sample predictor selection. In this section, we propose a new approach for the out-of-sample predictive ability test to select predictors. Since the new approach is based on the encompassing principle and OCMT procedure, this approach is referred to as ENC-OCMT. The ENC-OCMT procedure focuses on both the marginal effect and net impact of variables on the target variable, which is

similar to the OCMT procedure. We consider the high-dimensional linear predictive model,

$$y_{t+1} = \mu_{t+1} + u_{t+1}$$
$$\sigma^2_{t+1} = \omega + \alpha u_t^2 + \beta \sigma_t^2 + \sum_{i=1}^{n} \delta_i x_{i,t}^2$$

where $x_{1,t}, \ldots, x_{n,t}$ are the n predictors, for $i = 1, \ldots, n$. When the number of predictors is small, we can select the predictors directly depending on the marginal effect coefficient β_i. When the marginal effect is not zero, the particular predictor $x_{i,t}$ can be selected. However, for time series data, there might exist correlations among predictors, and the probability of the correlations among predictors increases as the number of predictors in the model increases. Thus, considering the indirect impact among predictors is essential when n is large.

The ENC-OCMT approach attempt to select the predictors $x_{i,t}$ that have marginal effect β_i or net impact θ_i on y_{t+1}, which is different from other multiple testing procedures. The net impact coefficient θ_i from Pesaran and Smith (2014) catches the indirect impact of predictors that have zero marginal effect on the outcome variable but have a correlation with other predictors.

According to the encompassing principle, for the low-dimensional models, we combine 2 models in equations (3.11) and (3.13) with weights $(1 - \lambda_i)$ for Model 0 and λ_i for each Model i. There is no predictor in Model 0, while there is one predictor in each Model i, where $i = 1, \ldots, n$. We combine Model 0 and one Model i at a time to construct the combined model for Model i, so we have n combined models for different Model i. For the high-dimensional models, we combine these $n + 1$ models with weights λ_i for each Model i

and $(1 - \sum_{i=1}^{n} \lambda_i)$ for Model 0, so we get one combined model as follows:

$$\hat{\sigma}^2_{c,t+1} = \left(1 - \sum_{i=1}^{n} \lambda_i\right) \hat{\sigma}^2_{0,t+1} + \sum_{i=1}^{n} \lambda_i \hat{\sigma}^2_{i,t+1}, \tag{3.15}$$

where $\hat{\sigma}^2_{i,t+1}$ is the conditional volatility for Model i that is estimated using the out-of-sample observations. Rearrange the equation above, the regression equation is

$$\hat{\sigma}^2_{0,t+1} = \sum_{i=1}^{n} \lambda_i \left(\hat{\sigma}^2_{0,t+1} - \hat{\sigma}^2_{i,t+1}\right) + \hat{\sigma}^2_{c,t+1}, . \tag{3.16}$$

We use the same information to estimate different models, so the residuals of different models are highly correlated with each other. Thus, we should consider the marginal effect $\lambda_i \neq 0$ and net impact $\theta_i \neq 0$ to select predictors. The net impact θ_i for the out-of-sample forecast encompassing approach based on the regression (3.16) is described as

$$\theta_i = \sum_{j=1}^{n} \lambda_j \sigma_{i,j}, \tag{3.17}$$

where $\sigma_{i,j} = E\left(\hat{\boldsymbol{d}}_i' \hat{\boldsymbol{d}}_j\right)$, $\hat{\boldsymbol{d}}_i$ denotes the $P \times 1$ vector of $(\hat{\sigma}^2_{0,t+1} - \hat{\sigma}^2_{i,t+1})$ for $t = R, \ldots, T$.

Due to whether the predictors $x_{i,t}$ have marginal effect and net impact on y_{t+1}, there are following four possible cases:

	$\theta_i \neq 0$	$\theta_i = 0$
$\lambda_i \neq 0$	(1). Signals with nonzero net effect	(2). Hidden signals
$\lambda_i = 0$	(3). Pseudo-signals	(4). Noise variables

The first case $\lambda_i \neq 0$ and $\theta_i \neq 0$ implies that predictors, in this case, have both marginal and net impact, whereas predictors in the last case, where $\lambda_i = 0$ and $\theta_i = 0$, are called noise variables since they have neither marginal effect nor net impact. The third case $\lambda_i = 0$ and $\theta_i \neq 0$ is possible and cannot be neglected. Predictors in the third case are referred to as Pseudo-signals. Obviously, the second case $\lambda_i \neq 0$ and $\theta_i = 0$ is almost impossible to occur

when the number of predictors is large. The ENC-OCMT approach selects the predictors in Cases (1), (2), and (3). In other words, we do not select the predictors when they are Noise variables.

Similar to OCMT, the ENC-OCMT procedure is also a multi-stage procedure. The difference is that OCMT checks the covariance between $x_{i,t}$ and y_{t+1}, and ENC-OCMT tests the covariance between the estimated volatility $\hat{\sigma}^2_{0,t+1}$ in Model 0 and the difference of estimated volatilities $(\hat{\sigma}^2_{0,t+1} - \hat{\sigma}^2_{i,t+1})$ in Model 0 and Model i. If there is covariance between $\hat{sigma}^2_{0,t+1}$ in Model 0 and the difference $(\hat{\sigma}^2_{0,t+1} - \hat{\sigma}^2_{i,t+1})$ in Model i, we select the predictor $x_{i,t}$. In the first stage, we consider the n bivariate regressions of $\hat{\sigma}^2_{0,t+1}$ on $(\hat{\sigma}^2_{0,t+1} - \hat{\sigma}^2_{i,t+1})$, for $i = 1, \ldots, n$, which is as follows:

$$\hat{\sigma}^2_{0,t+1} = \phi_{i,(1)} \left(\hat{\sigma}^2_{0,t+1} - \hat{\sigma}^2_{i,t+1} \right) + \hat{\sigma}^2_{c,i,t+1}, \text{ for } i = 1, \ldots, n, \tag{3.18}$$

where $\hat{\sigma}^2_{c,i,t+1}$ is an error term of the regression for Model i. Thus, $\phi_{i,(1)}$ can be estimated by regressing $\hat{\sigma}^2_{0,t+1}$ on $\hat{\sigma}^2_{0,t+1} - \hat{\sigma}^2_{i,t+1}$ by minimizing the Bregman loss function or negative log-likelihood function.

In the first stage of ENC-OCMT, we regress $\hat{\sigma}^2_{0,t+1}$ on $\hat{\sigma}^2_{0,t+1} - \hat{\sigma}^2_{i,t+1}$, for $i = 1, 2, \ldots, n$, one predictor at a time. Denote $t_{\hat{\phi}_{i,(1)}}$ to be the t-ratio of $\hat{\phi}_{i,(1)}$, we have

$$t_{\hat{\phi}_{i,(1)}} = \frac{\hat{\phi}_{i,(1)}}{\text{s.e.} \left(\hat{\phi}_i \right)} \tag{3.19}$$

The subscript (1) represents the first stage. It is worth mentioning that we can use our ENC statistic as the t statistic of the regression. Since both t statistic $t_{\hat{\phi}_i}$ and our $\text{ENC}_{i,R,P}$ are the statistics to test the covariance between $x_{i,t}$ and y_{t+1}. We also prove that the ENC statistic is asymptotically normal under the null hypothesis with no Granger Causality.

Both ENC and t statistics can be used to select predictors compared to the critical value from a standard normal distribution. Thus, we can use ENC statistic to replace t statistic in equation (19).

This is a two-sided test with alternative hypothesis $\phi_i \neq 0$, so the predictor can be selected when the absolute value of the statistic is greater than the critical value. The first-stage ENC-OCMT selection indicator is given by

$$\widehat{\mathcal{J}}_{i,(1)} = I\left[\left|\mathrm{ENC}_{i,R,P,(1)}\right| > c_p(n,\delta)\right] \quad \text{for } i = 1,\ldots,n \tag{3.20}$$

where $\mathrm{ENC}_{i,R,P,(1)}$ is the ENC statistic in the first stage and $c_p(n,\delta)$ is a critical value function defined by

$$c_p(n,\delta) = \Phi^{-1}\left(1 - \frac{p}{2f(n,\delta)}\right). \tag{3.21}$$

$\Phi^{-1}(\cdot)$ is the inverse of the standard normal distribution function, $f(n,\delta) = cn^{\delta}$ for some positive constants δ and c, and p $(0 < p < 1)$ is the nominal size of the individual tests to be set by the investigator. According to Chudik et al. (2018), the δ is defined as the critical value exponent. δ is used for the first stage, while δ^* is used in subsequent stages of OCMT, which δ^* must be greater than δ. We select predictors in the first stage when $\widehat{\mathcal{J}}_{i,(1)} = 1$. According to Chudik et al. (2018), denote the number of predictors selected in the first stage by $\hat{k}_{(1)}$, the index set of the selected predictors in the first stage is denoted by $\hat{\mathcal{S}}_{(1)}$, and the active set is defined as $\mathcal{A}_{(1)} = \{1,\ldots,n\}$, where the subscript (1) represents the first stage.

In subsequent stage $j = 2, 3, \ldots$, we consider the $n - k_{(j-1)}$ regressions of $\hat{\sigma}^2_{0,t+1}$ on the predictors in $\boldsymbol{D}_{(j-1)}$ and, one at a time, d_{it} for i belonging in the active set, $\mathcal{A}_{(j)}$,

$$\hat{\sigma}^2_{0,t+1} = \alpha_i + \phi_{i,(j)} \left(\hat{\sigma}^2_{0,t+1} - \hat{\sigma}^2_{i,t+1}\right) + \sum_{m=1}^{k_{(j-1)}} \gamma_m \left(\hat{\sigma}^2_{0,t+1} - \hat{\sigma}^2_{m,t+1}\right) + u_{(j),t+1},$$

$$\text{for } i \in \mathcal{A}_{(j)}, \quad j = 2, 3, \ldots.$$

where the number of predictors selected in the stage $(j-1)$ is denoted by $\hat{k}_{(j-1)}$, $k_{(j-1)}$ and the number of all predictors selected up to and including stage $(j-1)$ is denoted $k_{(j-1)}$, so $k_{(j-1)} = \sum_{q=1}^{(j-1)} \hat{k}_{(j-1)}$. The matrix $\boldsymbol{D}_{(j-1)}$ denotes the $k_{(j-1)} \times 1$ vector of $(u_{0,t+1} - u_{m,t+1})$ for $t = R, \ldots, T$, and $\hat{\sigma}^2_{m,t+1}$ is the estimated volatility for the predictors has been selected in the previous stages, for $m = 1, \ldots, k_{(j-1)}$ and $t = R, \ldots, T$. We then compute the $\text{ENC}_{i,R,P,(j)}$ as the t statistic here and the subscript (j) represents the stage j.

Predictors are then selected and added to the set of selections when $\widehat{\mathcal{J}}_{i,(j)} = 1$, where $\widehat{\mathcal{J}}_{i,(j)} = I\left[\left|\text{ENC}_{i,R,P,(j)}\right| > c_p\left(n, \delta^*\right)\right]$. The index set of the selected predictors in stage j is denoted by $\hat{\mathcal{S}}_{(j)}$ and the index set in stage j of all the selected predictors up to and including stage j is denoted by $\mathcal{S}_{(j)}$, so $\mathcal{S}_{(j)} = \mathcal{S}_{(j-1)} \cup \hat{\mathcal{S}}_{(j)}$. The active set in the stage $j+1$ is defined as $\mathcal{A}_{(j+1)} = \{1, \ldots, n\} \backslash \mathcal{S}_{(j)}$. The procedure stops when there is no predictor can be selected in the current stage.

3.5.2 LASSO Procedure for Out-of-sample Forecast Encompassing Test

In this subsection, we propose a new approach for the out-of-sample forecast encompassing tests in the high-dimensional predictive regression models. This approach tests predictive ability using the LASSO under the encompassing principle. Thus, this approach

is named ENC-LASSO. The ENC-LASSO procedure shrinks some coefficients and makes others zero. The predictors with non-zero coefficients are selected through ENC-LASSO.

We use the same combination in the ENC-OCMT and ENC-LASSO. We combine all $n+1$ models, so we have the equation (3.16)

$$\hat{\sigma}^2_{0,t+1} = \sum_{i=1}^{n} \lambda_i \left(\hat{\sigma}^2_{0,t+1} - \hat{\sigma}^2_{i,t+1} \right) + \hat{\sigma}^2_{c,t+1}.$$

Based on the LASSO method, the ENC-LASSO estimate $\hat{\boldsymbol{\lambda}}_{\text{LASSO}}$ is defined as

$$\hat{\boldsymbol{\lambda}}_{\text{LASSO}} = \arg\min_{\boldsymbol{\lambda}} \sum_{t=R}^{T} \left(\hat{\sigma}^2_{c,t+1} \right)^2 + \kappa \sum_{i=1}^{n} |\lambda_i| , \tag{3.22}$$

where $\hat{\boldsymbol{\lambda}}_{\text{LASSO}} = (\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_n)'$ is a vector of estimates λ_i for Model i, κ is a non-negative tuning parameter and the second term is the penalty term.

However, since the LASSO method produces biased through shrinking when the number of variables is large (Fan and Li, 2001), Zou (2006) proposes a weighted LASSO procedure adding weights to different coefficients in the penalty term, which is called Adaptive LASSO. Zou proves that the Adaptive LASSO with data-dependent weights is consistent. Thus, this paper also tests the predictive ability of predictors using the Adaptive LASSO under the encompassing principle. This procedure is called ENC-ALASSO. Similar to the Adaptive LASSO, we add weights $\hat{w}_i$ to the penalty term in equation (3.22). The ENC-ALASSO estimate $\hat{\boldsymbol{\lambda}}_{\text{ALASSO}}$ is defined as

$$\hat{\boldsymbol{\lambda}}_{\text{ALASSO}} = \arg\min_{\boldsymbol{\lambda}} \sum_{t=R}^{T} \left(\hat{\sigma}^2_{c,t+1} \right)^2 + \kappa \sum_{i=1}^{n} \hat{w}_i |\lambda_i| ,$$

where $\hat{\boldsymbol{\lambda}}_{\text{ALASSO}} = (\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_n)'$ is a vector of estimates λ_i for Model i, $\hat{w}_i = 1/|\hat{\lambda}_{i,\text{OLS}}|$, and $\hat{\lambda}_{i,\text{OLS}}$ are estimated minimizing least squares loss function in equation (3.16).

3.6 Monte Carlo Simulations

3.6.1 Simulation Design

In order to show the asymptotic distribution of the forecast encompassing test, we simulate data from the following two DGPs.

In DGP 1, we generate the variable x_t to be an AR(1) process. We set

$$x_t = \rho x_{t-1} + v_t,$$

where $v_t \overset{iid}{\sim} N(0, \sigma_v^2)$, $\rho = 0.5$, $\sigma_v = 1$, $\gamma_{1,t+1} = 0.1$.

$$y_{t+1} = \gamma_{1,t+1} + u_{t+1}$$

$$\gamma_{3,t+1} = \omega + \alpha u_t^2 + \beta \gamma_{3,t} + \delta \cdot x_t^2$$

We consider $\alpha = 0.05$, $\beta = 0.75$, then we have $\omega = (1 - \alpha - \beta)\sigma_u^2 - \delta\sigma_v^2/(1 - \rho)$. We set $\delta = 0$ for the size of the test and $\delta = 0.5$ for the power of the test. In the simulation, we set the number of in-sample observations $R \in \{60, 120, 240\}$ and the number of out-of-sample forecasts $P \in \{48, 240, 1200\}$.

DGP 2 is the GARCH-MIDAS model. This model is a component model of volatility, which is proposed by Engle et al. (2013). In DGP 2, we also generate the variable x_t to be an AR(1) process. We set

$$x_t = \rho x_{t-1} + v_t,$$

where $v_t \overset{iid}{\sim} N(0, \sigma_v^2)$, $\rho = 0.5$, $\sigma_v = 1$. We generate y_t as follows,

$$y_{i,t} - E_{i-1,t}\left(y_{i,t}\right) = \sqrt{g_{i,t}\tau_t}\varepsilon_{i,t}, \tag{3.23}$$

where $\varepsilon_{i,t} \sim N(0,1)$ and $y_{i,t}$ is at day $i = 1, \ldots, N_t$ in a period $t = 1, \ldots, T$. Daily expected returns $E_{i-1,t}(y_{i,t})$ are set to be constant μ. The total number of daily observations is denoted as $N_0 = \sum_{t=1}^{T} N_t$. Equation (23) shows two components: the short-term volatility component $g_{i,t}$, and the long-term volatility component τ_t.

The short-term volatility component is a mean-reverting GARCH(1,1) process:

$$g_{i,t} = (1 - \alpha - \beta) + \alpha \frac{(y_{i-1,t} - \mu)^2}{\tau_t} + \beta g_{i-1,t},$$

where $\alpha > 0, \beta > 0$ and $\alpha + \beta < 1$.

The long-term volatility component is given by:

$$\log(\tau_t) = m + \theta \sum_{k=1}^{K} \varphi_k(\omega_1, \omega_2) x_{t-k},$$

where $\varphi_k(\omega_1, \omega_2)$ is the Beta weighting scheme:

$$\varphi_k(\omega_1, \omega_2) = \frac{(k/(K+1))^{\omega_1 - 1} \cdot (1 - k/(K+1))^{\omega_2 - 1}}{\sum_{l=1}^{K} (l/(K+1))^{\omega_1 - 1} \cdot (1 - l/(K+1))^{\omega_2 - 1}}.$$

We consider $\alpha = 0.05$, $\beta = 0.8$, $m = 0.01$, $\omega_1 = \omega_2 = 5$ and $\mu = 0.5$. We set $\theta = 0$ for the size of the test and $\theta = 0.1$ for the power of the test. In the simulation, we set the number of in-sample observations $R \in \{720, 1080, 1440\}$ and the number of out-of-sample forecasts $P \in \{1440, 2160, 2880\}$.

We compare the ENC statistic with the DM and CCS statistics. The ENC statistic is $ENC_P \equiv \hat{Q}_P^{-0.5} \sqrt{P} \hat{C}_P$, where $\hat{Q}_P = \text{var}\left(\sqrt{P}\hat{C}_P\right)$. The DM statistic is $DM_P \equiv \hat{S}_P^{-0.5} \sqrt{P} \hat{D}_P$, where $\hat{S}_P = \text{var}\left(\sqrt{P}\hat{D}_P\right)$. The CCS statistic is $CCS_P \equiv \hat{W}_P^{-0.5} \sqrt{P} \hat{M}_P$, where $\hat{M}_P = P^{-1} \sum_{t=R}^{T} g\left(\hat{u}_{t+1}^{(1)}\right) \hat{u}_{t+1}^{(1)} x_t$ and $\hat{W}_P = \text{var}\left(\sqrt{P}\hat{M}_P\right)$. $\hat{S}_P, \hat{Q}_P$ and $\hat{W}_P$ are consistent estimators of $S_P = \text{var}\left(\sqrt{P}D_P\right), Q_P = \text{var}\left(\sqrt{P}C_P\right)$ and $W_P = \text{var}\left(\sqrt{P}M_P\right) = \text{var}\left(P^{-0.5} \sum_{t=R}^{T} g\left(u_{t+1}^{(1)}\right) u_{t+1}^{(1)} x_t\right)$, respectively. We use rolling windows to estimate $\hat{\theta} =$

$[\hat{\omega}_t^{(i)}, \hat{\alpha}_t^{(i)}, \hat{\beta}_t^{(i)}, \hat{\delta}_t]$, where $i = 1, 2$, and then predict one step ahead dependent variable $\hat{y}_{t+1}$ and calculate the forecast mean and variance $\hat{\gamma}_{1,t+1}$ and $\hat{\gamma}_{3,t+1}$, then obtain DM_P, ENC_P, and CCS_P statistics. We repeat this procedure 2000 times and get the asymptotic distribution of DM_P, ENC_P, and CCS_P, and the size and power.

In order to check the asymptotic properties of ENC-OCMT and ENC-LASSO, we consider DGP 3. DGP 3 is following DGP 1(a) from Chudik et al. (2018). These simulations are used to confirm that our approaches work perfectly in the finite sample size. The sample combinations $n = (100, 200, 300)$. In our simulation, we set the number of in-sample observations $R = 100$ and the number of out-of-sample observations $P = (200, 300, 500)$. We use rolling windows to estimate parameters by minimizing the negative log-likelihood function. All experiments are carried out using $2,000$ replications.

DGP 3 includes signals and noise variables but no hidden signals or pseudo-signals. y_{t+1} is generated as

$$y_{t+1} = \mu_{t+1} + u_{t+1}$$

$$\sigma_{t+1}^2 = \omega + \alpha u_t^2 + \beta \sigma_t^2 + \delta_1 x_{1,t}^2 + \delta_2 x_{2,t}^2 + \delta_3 x_{3,t}^2 + \delta_4 x_{4,t}^2$$

where $u_t \sim$ i.i.d. $\mathrm{N}(0, 1)$ for $t = 1, 2, \ldots, T$. We set $\delta_1 = \delta_2 = \delta_3 = \delta_4 = 1$ in this DGP. In this setting, the predictors $x_{i,t} = (\varepsilon_{i,t} + \nu g_t) / \sqrt{1 + \nu}$, for $i = 1, 2, 3, 4$, are signal variables and the predictors $x_{5,t} = \varepsilon_{5,t}, x_{i,t} = (\varepsilon_{i-1,t} + \varepsilon_{i,t}) / \sqrt{2}$, for $i > 5$, are noise variables, where g_t and ε_{it} are independent draws from $\mathrm{N}(0, 1)$, for $t = 1, 2, \ldots, T$, and $i = 1, 2, \ldots, n$. We set $\nu = 1$, which implies 50% pairwise correlation among the signal variables.

DGP 4 is used to confirm the asymptotic properties of ENC-OCMT and ENC-LASSO for the GARCH-MIDAS model. We generate y_t to be the same as DGP 2,

$$y_{i,t} - E_{i-1,t}\left(y_{i,t}\right) = \sqrt{g_{i,t}\tau_t}\varepsilon_{i,t},$$

where $\varepsilon_{i,t} \sim N(0,1)$ and $y_{i,t}$ is at day $i = 1, \ldots, N_t$ in a period $t = 1, \ldots, T$. Daily expected returns $E_{i-1,t}\left(y_{i,t}\right)$ are set to be constant μ. The total number of daily observations is denoted as $N_0 = \sum_{t=1}^{T} N_t$. The short-term volatility component $g_{i,t}$ is generated the same as DGP 2.

The short-term volatility component is a mean-reverting GARCH(1,1) process:

$$g_{i,t} = (1 - \alpha - \beta) + \alpha\frac{\left(y_{i-1,t} - \mu\right)^2}{\tau_t} + \beta g_{i-1,t},$$

where $\alpha > 0, \beta > 0$ and $\alpha + \beta < 1$.

The long-term volatility component τ_t is given by:

$$\begin{aligned}\log\left(\tau_t\right) = m &+ \theta_1 \sum_{k=1}^{K} \varphi_k\left(\omega_1, \omega_2\right) x_{1,t-k} + \theta_2 \sum_{k=1}^{K} \varphi_k\left(\omega_1, \omega_2\right) x_{2,t-k} \\ &+ \theta_3 \sum_{k=1}^{K} \varphi_k\left(\omega_1, \omega_2\right) x_{3,t-k} + \theta_4 \sum_{k=1}^{K} \varphi_k\left(\omega_1, \omega_2\right) x_{4,t-k}.\end{aligned}$$

where $\varphi_k\left(\omega_1, \omega_2\right)$ is the Beta weighting scheme. We set $\theta_1 = \theta_2 = \theta_3 = \theta_4 = 1$ in this DGP. The setting of x_t is the same as DSGP 3. We consider $\alpha = 0.05$, $\beta = 0.8$, $m = 0.01$, $\omega_1 = \omega_2 = 5$ and $\mu = 0.5$.

We focus on three out-of-sample forecast selection methods for the simulation results: ENC-OCMT, ENC-LASSO, and ENC-ALASSO. In our forecast evaluations, we use the most used criteria in the penalized regression methods, which are the following: the true positive rate (TPR) defined by (4.6.1), the false positive rate (FPR) defined by

(4.6.1), the false discovery of the approximating model (FDR) defined by (4.6.1) and the false discovery rate of the true model (FDR*) denoted by (4.6.1).

True positive rate:

$$\text{TPR}_{n,T} = \frac{\sum_{i=1}^{n} I\left(\widehat{\mathcal{J}}_i = 1 \text{ and } \beta_i \neq 0\right)}{\sum_{i=1}^{n} I\left(\beta_i \neq 0\right)} \xrightarrow{p} 1. \tag{3.24}$$

False positive rate:

$$\text{FPR}_{n,T} = \frac{\sum_{i=1}^{n} I\left(\widehat{\mathcal{J}}_i = 1 \text{ and } \beta_i = 0\right)}{\sum_{i=1}^{n} I\left(\beta_i = 0\right)} \xrightarrow{p} 0. \tag{3.25}$$

False discovery rate:

$$\text{FDR}_{n,T} = \frac{\sum_{i=1}^{n} I\left(\widehat{\mathcal{J}}_i = 1, \text{ and } \beta_i = \theta_i = 0\right)}{\sum_{i=1}^{n} \widehat{\mathcal{J}}_i + 1} \xrightarrow{p} 0. \tag{3.26}$$

The false discovery rate of the true model:

$$\text{FDR}^*_{n,T} = \frac{\sum_{i=1}^{n} I\left(\widehat{\mathcal{J}}_i = 1 \text{ and } \beta_i = 0\right)}{\sum_{i=1}^{n} \widehat{\mathcal{J}}_i + 1} \xrightarrow{p} c. \tag{3.27}$$

The asymptotic theory of OCMT carries over for ENC-OCMT and ENC-LASSO in terms of the true and false positive rates, and false discovery rate of the approximate and true model (TPR, FPR, FDR, and FDR*). We also report the root of mean squared forecast error (RMSFE) for three out-of-sample forecast selection methods.

For DGPs 3 and 4, we set the sample size n to be $100, 200$, and 300, the number of in-sample observations R to be 100, and the number of out-of-sample observations to be $200, 300$, and 500. For ENC-OCMT method, we set nominal level $p = 0.1, 0.05, 0.01$, $\delta = 1$ in the first stage, and $\delta^* \in \{1.5, 2\}$ in the subsequent stages.

3.6.2 Simulation Results

Figures 1-2 show that ENC statistics for GARCH model have proper sizes and good powers. The distributions of ENC for GARCH models are asymptotically normal under the null hypothesis.

Different settings lead to the different performance of methods. These results show that ENC-OCMT outperforms ENC-LASSO and ENC-ALASSO in most experiments. ENC-OCMT gets the largest TPR, lowest FPR, FDR, FDR* and lowest RMSFE in most experiments among the three methods, which implies that ENC-OCMT can successfully select signals, Hidden Signals, and Pseudo-signals. However, ENC-LASSO has a larger TPR than ENC-OCMT in some experiments, and ENC-ALASSO has a lower FDR and FDR* compared to ENC-LASSO. This means that ENC-LASSO tends to overestimate the number of signals. Lasso also has the same problem in the literature. Overall, no method universally dominates other methods.

3.7 Empirical Analysis

The dataset utilized in our analysis is sourced from Fang et al. (2020), including the S&P500, and U.S. macroeconomic & financial indicators from 1969Q1 to 2018Q4. The dataset includes the natural logarithm of daily S&P500 index prices, representing stock market returns, alongside quarterly macroeconomic and financial data. This dataset comprises sixteen macroeconomic variables, such as the real GDP growth rate, industrial production

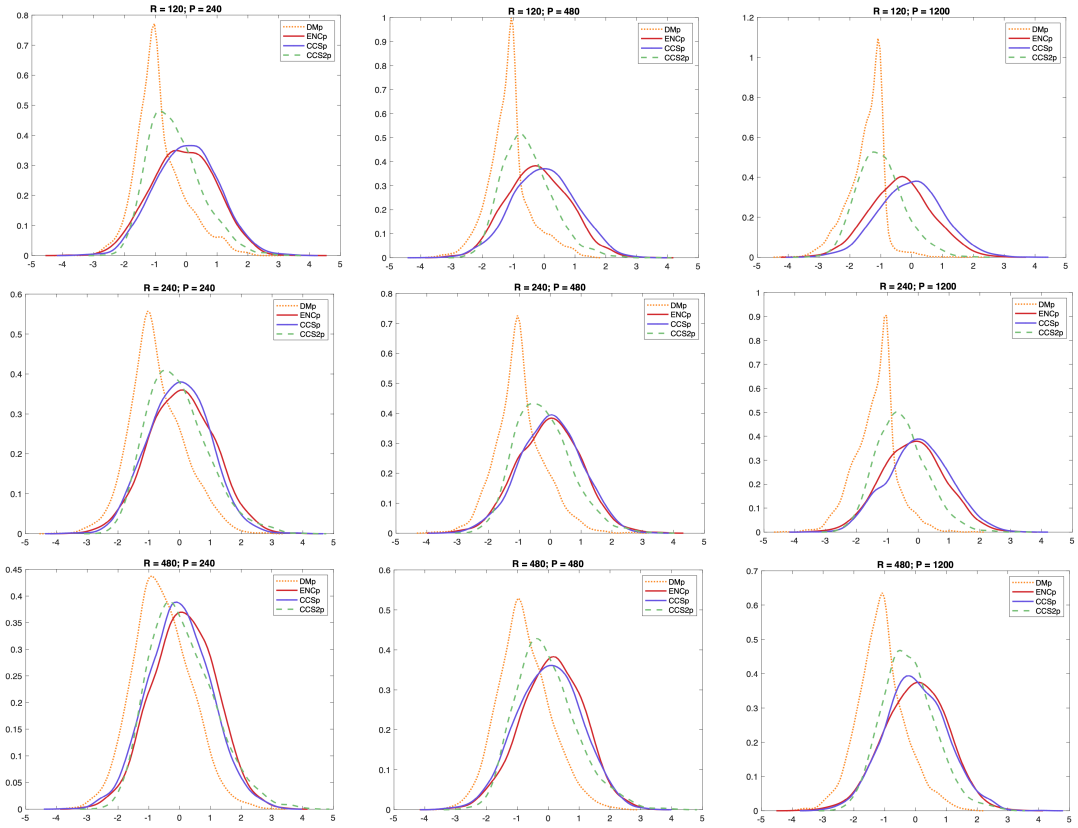

Figure 3.1: Simulation Results of DGP 1: Size of test

$b = 0$, $\rho = 0.5$ and 2000 repeats

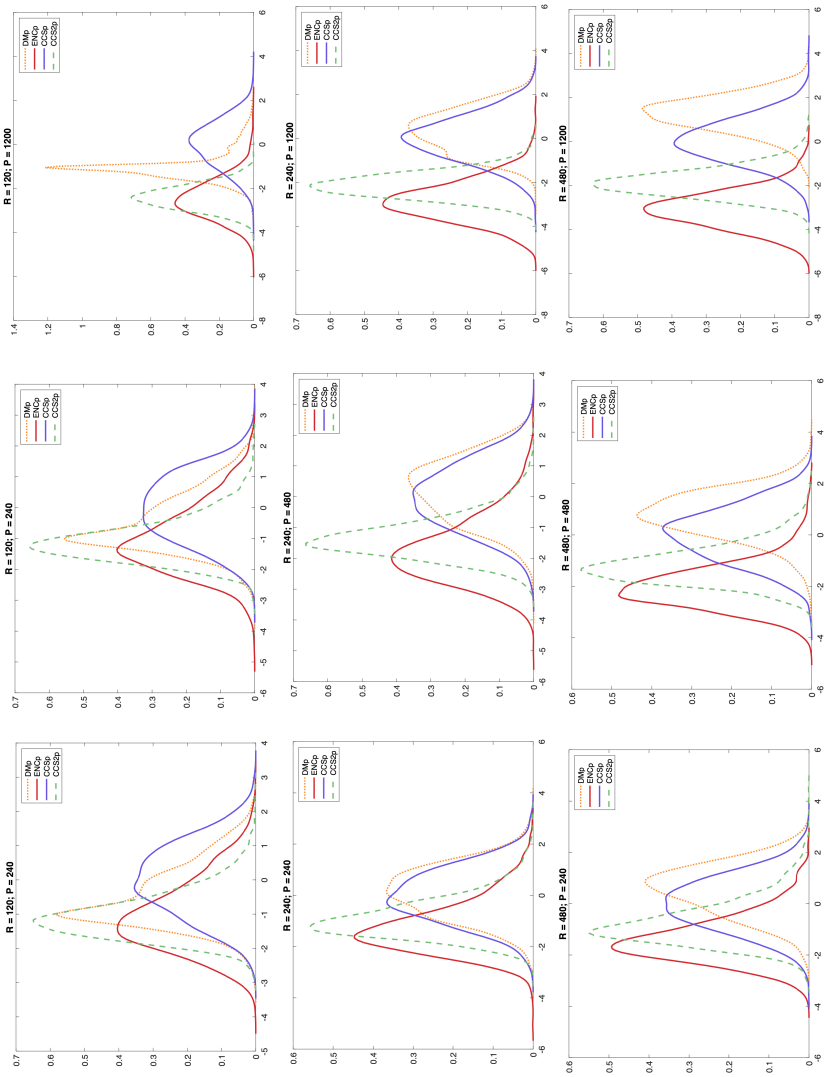

Figure 3.2: Simulation Results of DGP 1: Power of test

$b = 0.05$, $\rho = 0.5$ and 2000 repeats

growth rate, unemployment rate, housing starts, after-tax nominal corporate profits, real personal consumption, Consumer Price Index (CPI), Producer Price Index (PPI), Chicago Fed National Activity Index (CFNAI), new orders index of the Institute of Supply Management (ISM), monetary base, and the University of Michigan consumer sentiment index. Additionally, six financial variables are taken into account, including term spread, default spread, equity market returns (MKT), short-term reversal factor (STR), implied volatility (IV), and realized volatility (RV).

In this section, we want to select predictors for forecasting stock market volatility. While the stock market returns are recorded on a daily basis, the potential predictors, the U.S. macroeconomic and financial indicators, are documented quarterly. Consequently, we intend to employ our ENC-OCMT and ENC-LASSO techniques of predictor selection, in conjunction with the GARCH-MIDAS model. It's worth noting that in equation (23), $\sigma_{i,t}^2 = g_{i,t}\tau_t$. We divide the full sample into two subsets: in-sample data ranging from 1969Q1 to 2006Q4 (152 observations) and out-of-sample data spanning from 2007Q1 to 2018Q4 (48 observations). The parameters $\mu, \alpha, \beta, m, \theta, \omega_1, \omega_2$ are estimated by minimizing the Bregman loss function or negative log-likelihood function, utilizing the in-sample observations. Subsequently, we evaluate the forecast errors and forecast losses using out-of-sample observations. We employ the rolling window approach in estimation and implement a one-step-ahead forecast.

3.8 Conclusion

The research conducted throughout this paper has greatly enriched our understanding of volatility forecasting, shedding light on new possibilities for further advancements in the field. By proposing scoring functions that do not require the zero mean assumption or a proxy for the unknown true variance, this work extends the current state of knowledge in this area, and offers innovative tools to evaluate the mean and volatility forecasts jointly. The rigorously derived strictly consistent scoring functions can bridge the gap left by non-elicitable variance, thus contributing to a more accurate and robust comparison and ranking of volatility forecasts. Furthermore, these functions are underpinned by the Bregman divergence measures, providing a solid mathematical base for their application.

Moreover, the development of new tests for Granger Causality and the novel approaches proposed for predictor selection in high-dimensional predictive regression models, namely ENC-OCMT and ENC-LASSO, are significant advancements in the field. Our simulation results show the effectiveness and robustness of these methods even in finite sample sizes, making them valuable tools for practical forecasting. The methods we propose in this paper prove to be not only theoretically sound but also practically effective, especially in high-frequency and high-dimensional settings, typically encountered in financial and macroeconomic forecasting. The integration of these methods with the GARCH-MIDAS model further illustrates their potential for extensive applications. However, future research is warranted to explore additional applications and potential modifications of the proposed

methods, with a view towards refining and broadening their utility in different forecasting contexts.

3.9 Appendix A

Denote $\theta_i = [\omega_i, \alpha_i, \beta_i, \delta_i]', i = 1, 2.$

Score:

$$k_{i,t} = \frac{\partial S(\mu, \sigma^2, y)}{\partial \sigma^2} \frac{\partial \sigma^2}{\partial \theta} = \left[\frac{1}{\mu^2 + \sigma^2(\theta)} - \frac{y^2}{(\mu^2 + \sigma^2(\theta))^2} \right] \nabla' \sigma^2(\theta) = 0$$

$$K_i(t) = \frac{1}{R} \sum_{t=1}^{R} k_{i,t}$$

$$h_{i,t} = -k_{i,t}$$

$$H_i(t) = \frac{1}{R} \sum_{t=1}^{R} h_{i,t}(\theta)$$

By using Taylor Expansion, we have

$$K_{i,t} + \Lambda_i(t, \theta^*)(\hat{\theta} - \theta_0) = 0$$

$$\hat{\theta} - \theta_0 = -\Lambda_i(t, \theta^*)^{-1} K_{i,t} = \Lambda_i(t, \theta^*)^{-1} H_i(t) = B_i(t, \theta^*) H_i(t) = B_i H_i(t) + o_P(1)$$

where

$$\begin{aligned} \Lambda(\theta) &= \frac{\partial K_i(t)}{\partial \theta} \\ &= \frac{1}{R} \sum_{t=1}^{R} \left\{ \left[-\frac{1}{(\mu^2 + \sigma^2(\theta))^2} + \frac{2y^2}{(\mu^2 + \sigma^2(\theta))^3} \right] \nabla' \sigma^2(\theta) \nabla \sigma^2(\theta) \right. \\ &\quad \left. + \left[\frac{1}{\mu^2 + \sigma^2(\theta)} - \frac{y^2}{(\mu^2 + \sigma^2(\theta))^2} \right] \nabla^2 \sigma^2(\theta) \right\} \end{aligned}$$

$$B_i(t, \theta^*) = \Lambda_i(t, \theta^*)^{-1} = B_i + o_p(1)$$

where B_i equals to $B_i(t)$ evaluated at θ_0 and θ^* lies between $\hat{\theta}$ and θ_0. By the continuous mapping theorem, $\theta \xrightarrow{p} \theta^0 \Rightarrow B_i(t) \xrightarrow{p} B_i$.

Therefore, $B_1 = 1$ and $B_2 = \begin{pmatrix} 1 & 0 \\ 0 & \Lambda(\theta_2^0)^{-1} \end{pmatrix}^{-1}$. Let J denote the selection matrix

(1,0) such that $Jh_{2,t} = h_{1,t}$. Let $\sup_t$ denote $\sup_{t-R+1\leq\cdot\leq t}$. And matrix A and C will defined in Lemma 3 and $\tilde{h}_{2,t} = Z^{-1}A'CB_2^{1/2}h_{2,t}, \tilde{H}_2(t) = Z^{-1}A'CB_2^{1/2}H_2(t)$. U_t and $\tilde{U}_t$ will defined in Assumption 2. By the central limit theorem, $\sqrt{R}\left[R^{-1}\sum_{s=t-R+1}^{t} h_{2,s}\right] \sim N\left(0, Z^2B_2^{-1}\right)$, where $Z^2B_2^{-1} = \mathbb{E}[k_t(\theta)k_t(\theta)']$. We use $\sum_t$ instead of $\sum_{s=t-R+1}^{t}$ and use $\hat{\sigma}_{t+1}$ instead of $\sigma_{t+1}(\hat{\theta})$ for simplicity.

Lemma 1: (Clark & McCracken, 2001) Let Assumptions 1-4 hold. For i=1,2,

$$\sum_t H_i'(t)B_ih_{i,t+1}h_{i,t+1}'B_jH_j'(t) = \sum_t H_i'(t)B_i\mathbb{E}(h_{i,t+1}h_{i,t+1}')B_jH_j'(t) + o_p(t).$$

Proof: Adding and subtracting $\mathbb{E}(h_{i,t+1}h_{i,t+1}')$,

$$\begin{aligned}
&\sum_t H_i'(t)B_ih_{i,t+1}h_{i,t+1}'B_jH_j(t) \\
&= \sum_t H_i'(t)B_i\mathbb{E}(h_{i,t+1}h_{i,t+1}')B_jH_j(t) + \sum_t H_i'(t)B_i\left(h_{i,t+1}h_{i,t+1}' - \mathbb{E}(h_{i,t+1}h_{i,t+1}')\right)B_jH_j(t).
\end{aligned}$$

Now, we need to show the second term is $o_p(1)$.

$$\begin{aligned}
&\sum_t H_i'(t)B_i\left(h_{i,t+1}h_{i,t+1}' - \mathbb{E}(h_{i,t+1}h_{i,t+1}')\right)B_jH_j(t) \\
&= T^{-1/2}\sum_t(\frac{T}{t})^2\left[T^{-1/2}\sum_{s=1}^{t}h_{j,s}'B_j \otimes T^{-1/2}\sum_{s=1}^{t}h_{i,s}'B_i\right] \\
&\times \text{vec}\left[T^{-1/2}\left(h_{i,t+1}h_{i,t+1}' - \mathbb{E}(h_{i,t+1}h_{i,t+1}')\right)\right]
\end{aligned}$$

$\sum_t(\frac{T}{t})^2\left[\frac{1}{\sqrt{T}}\sum_{s=1}^{t}h_{j,s}'B_j \otimes \frac{1}{\sqrt{T}}\sum_{s=1}^{t}h_{i,s}'B_i\right]\text{vec}\left[\frac{1}{\sqrt{T}}\left(h_{i,t+1}h_{i,t+1}' - \mathbb{E}(h_{i,t+1}h_{i,t+1}')\right)\right]$ is $Op(1)$ follows from Assumption 2, Corollary 29.19 of Davidson (1994) and Theorem 3.1 of Hansen (1992). The proof is complete.

Lemma 2: Let Assumptions 1-3 hold. (a) Let $-J'B_1J + B_2 = M$ and $B_2^{-1/2}MB_2^{-1/2} = Q$, then Q is idempotent. (b). Define matrix $A = (0,1)'$ and $C = I_{2\times 2}$ Then $Q = CAA'C$.

Proof: (a).

$$\begin{aligned} Q &\equiv B_2^{-1/2}MB_2^{-1/2} \\ &= \begin{pmatrix} 1 & 0 \\ 0 & \Lambda(\theta_2^0)^{-1} \end{pmatrix}^{1/2} \begin{pmatrix} 1 & 0 \\ 0 & \Lambda(\theta_2^0)^{-1} \end{pmatrix}^{-1} \begin{pmatrix} 1 & 0 \\ 0 & \Lambda(\theta_2^0)^{-1} \end{pmatrix}^{1/2} \\ &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \end{aligned}$$

so Q is an idempotent matrix with rank 1.

(b). Lemma A4(b) of Clark and McCracken (2001).

Lemma 3: (Clark & McCracken, 2001) Let Assumptions 1-4 hold.

$$\sum_t \tilde{H}_2'(t)\tilde{h}_{2,t+1} \xrightarrow{d} \xi^{-1} \int_\xi^1 \left[W(s) - W(s-\xi)\right] \mathrm{d}W(s).$$

Proof:

$$\begin{aligned} \sum_t \tilde{H}_2'(t)\tilde{h}_{2,t+1} &= \sum_t \left(\frac{1}{R} \sum_{j=t-R+1}^{t} \tilde{h}_{2,j}' \right) \tilde{h}_{2,t+1} \\ &= \sum_t \left(\frac{1}{R} \sum_{j=1}^{t} \tilde{h}_{2,j}' - \frac{1}{R} \sum_{j=1}^{t-R} \tilde{h}_{2,j}' \right) \tilde{h}_{2,t+1} \\ &= \frac{T}{R} \left[\sum_t \left(T^{-1/2} \sum_{j=1}^{t} \tilde{h}_{2,j}' - T^{-1/2} \sum_{j=1}^{t-R} \tilde{h}_{2,j}' \right) \left(T^{-1/2} \tilde{h}_{2,t+1} \right) \right] \end{aligned}$$

According to the continues mapping theorem and Corollary 29.19 of Davidson (1994), $T^{-1/2}\sum_{j=1}^{t}\tilde{h}'_{2,j} \Longrightarrow W(s), T^{-1/2}\sum_{j=1}^{t-R}\tilde{h}'_{2,j} \Longrightarrow W(s-\xi)$, $T^{-1/2}\tilde{h}_{2,t+1} \Longrightarrow \mathrm{d}W(s)$. Since $T/R = \xi^{-1}$, the proof is complete.

Lemma 4: (Clark & McCracken, 2001) Let Assumptions 1-4 hold.

$$\sum_{t}\tilde{H}'_2(t)\tilde{H}_2(t) \xrightarrow{d} \xi^{-2}\int_{\xi}^{1}\left[W(s)-W(s-\xi)\right]^2 \mathrm{d}s.$$

Proof:

$$\begin{aligned}
\sum_{t}\tilde{H}'_2(t)\tilde{H}_2(t) &= \sum_{t}\left(\frac{1}{R}\sum_{j=t-R+1}^{t}\tilde{h}'_{2,j}\right)'\left(\frac{1}{R}\sum_{j=t-R+1}^{t}\tilde{h}'_{2,j}\right)\\
&= \sum_{t}\left(\frac{1}{R}\sum_{j=1}^{t}\tilde{h}'_{2,j}-\frac{1}{R}\sum_{j=1}^{t-R}\tilde{h}'_{2,j}\right)'\left(\frac{1}{R}\sum_{j=1}^{t}\tilde{h}'_{2,j}-\frac{1}{R}\sum_{j=1}^{t-R}\tilde{h}'_{2,j}\right)\\
&= \frac{1}{T}\left(\frac{T}{R}\right)^2\sum_{t}\left[\left(T^{-1/2}\sum_{j=1}^{t}\tilde{h}'_{2,j}-T^{-1/2}\sum_{j=1}^{t-R}\tilde{h}'_{2,j}\right)'\right.\\
&\quad\left.\times\left(T^{-1/2}\sum_{j=1}^{t}\tilde{h}'_{2,j}-T^{-1/2}\sum_{j=1}^{t-R}\tilde{h}'_{2,j}\right)\right]
\end{aligned}$$

According to the continues mapping theorem and Corollary 29.19 of Davidson (1994), $T^{-1/2}\sum_{j=1}^{t}\tilde{h}'_{2,j} \Longrightarrow W(s), T^{-1/2}\sum_{j=1}^{t-R}\tilde{h}'_{2,j} \Longrightarrow W(s-\xi)$, $T^{-1/2}\tilde{h}_{2,t+1} \Longrightarrow \mathrm{d}W(s)$ and $1/T \Longrightarrow ds$. Since $T/R = \xi^{-1}$, the proof is complete.

Lemma 5: Let Assumptions 1-4 hold.

$$\begin{aligned}
&\sum_t \left[V_2(\hat{\mu}_t^{(1)}, (\hat{\sigma}_t^{(1)})^2, y_{t+1}) \left((\hat{\sigma}_t^{(1)})^2 - (\hat{\sigma}_t^{(2)})^2 \right) \right] \\
&= \sum_t \left[\left(\frac{1}{(\hat{\mu}_t^{(1)})^2 + (\hat{\sigma}_t^{(1)})^2} - \frac{y^2}{\left((\hat{\mu}_t^{(1)})^2 + (\hat{\sigma}_t^{(1)})^2 \right)^2} \right) \left((\hat{\sigma}_t^{(1)})^2 - (\hat{\sigma}_t^{(2)})^2 \right) \right] \\
&= Z^2 \sum_t \tilde{H}_2'(t) \tilde{h}_{2,t+1} + o_p(1)
\end{aligned}$$

Proof: Taking Taylor Expansion, we can get every term in the last equation as follows:

$$\frac{1}{(\hat{\mu}_t^{(1)})^2 + (\hat{\sigma}_t^{(1)})^2} = \frac{1}{(\mu_t^{(1)})^2 + (\sigma_t^{(1)})^2} - \frac{\nabla'(\sigma_t^{(1)})^2}{\left((\hat{\mu}_t^{(1)})^2 + (\hat{\sigma}_t^{(1)})^2 \right)^2} \left(\hat{\theta}_1 - \theta_1 \right) + O_p \left((\hat{\theta}_1 - \theta_1)^2 \right)$$

$$\frac{y^2}{\left((\hat{\mu}_t^{(1)})^2 + (\hat{\sigma}_t^{(1)})^2 \right)^2} = \frac{y^2}{\left((\mu_t^{(1)})^2 + (\sigma_t^{(1)})^2 \right)^2} + \frac{2y^2 \nabla'(\sigma_t^{(1)})^2}{\left((\hat{\mu}_t^{(1)})^2 + (\hat{\sigma}_t^{(1)})^2 \right)^3} \left(\hat{\theta}_1 - \theta_1 \right) + O_p \left((\hat{\theta}_1 - \theta_1)^2 \right)$$

$$(\hat{\sigma}_t^{(1)})^2 - (\hat{\sigma}_t^{(2)})^2 = \nabla'(\sigma_t^{(1)})^2 (\hat{\theta}_1 - \theta_1) - \nabla'(\sigma_t^{(2)})^2 (\hat{\theta}_2 - \theta_2) + O_p \left((\hat{\theta}_1 - \theta_1)^2 \right) - O_p \left((\hat{\theta}_2 - \theta_2)^2 \right)$$

Substituting those terms into the equation, we can get

$$\begin{aligned}
&\sum_t \left[\left(\frac{1}{(\hat{\mu}_t^{(1)})^2 + (\hat{\sigma}_t^{(1)})^2} - \frac{y^2}{\left((\hat{\mu}_t^{(1)})^2 + (\hat{\sigma}_t^{(1)})^2 \right)^2} \right) \left((\hat{\sigma}_t^{(1)})^2 - (\hat{\sigma}_t^{(2)})^2 \right) \right] \\
&= \sum_t \Bigg\{ \Bigg[\underbrace{\left(\frac{1}{(\mu_t^{(1)})^2 + (\sigma_t^{(1)})^2} - \frac{y^2}{\left((\mu_t^{(1)})^2 + (\sigma_t^{(1)})^2 \right)^2} \right)}_{A^{(1)}} \\
&\quad + \underbrace{\left(-\frac{\nabla'(\sigma_t^{(1)})^2}{\left((\hat{\mu}_t^{(1)})^2 + (\hat{\sigma}_t^{(1)})^2 \right)^2} - \frac{2y^2 \nabla'(\sigma_t^{(1)})^2}{\left((\hat{\mu}_t^{(1)})^2 + (\hat{\sigma}_t^{(1)})^2 \right)^3} \right) \left(\hat{\theta}_1 - \theta_1 \right) + 2O_p \left((\hat{\theta}_1 - \theta_1)^2 \right)}_{B^{(1)}} \Bigg] \\
&\quad \times \left[\underbrace{\nabla'(\sigma_t^{(1)})^2 (\hat{\theta}_1 - \theta_1) - \nabla'(\sigma_t^{(2)})^2 (\hat{\theta}_2 - \theta_2)}_{C^{(1)}} + \underbrace{O_p \left((\hat{\theta}_1 - \theta_1)^2 \right) - O_p \left((\hat{\theta}_2 - \theta_2)^2 \right)}_{C^{(2)}} \right] \Bigg\}
\end{aligned}$$

$$
\begin{aligned}
D_1 &= A^{(1)}C^{(1)} \\
&= \sum_t \left[\left(\frac{1}{(\mu_t^{(1)})^2 + (\sigma_t^{(1)})^2} - \frac{y^2}{\left((\mu_t^{(1)})^2 + (\sigma_t^{(1)})^2\right)^2} \right) \right. \\
&\quad \left. \times \left(\nabla'(\sigma_t^{(1)})^2(\hat{\theta}_1 - \theta_1) - \nabla'(\sigma_t^{(2)})^2(\hat{\theta}_2 - \theta_2) \right) \right] \\
&= \sum_t \left[\left(\frac{1}{(\mu_t^{(1)})^2 + (\sigma_t^{(1)})^2} - \frac{y^2}{\left((\mu_t^{(1)})^2 + (\sigma_t^{(1)})^2\right)^2} \right) \nabla'(\sigma_t^{(1)})^2(\hat{\theta}_1 - \theta_1) \right. \\
&\quad \left. - \left(\frac{1}{(\mu_t^{(1)})^2 + (\sigma_t^{(1)})^2} - \frac{y^2}{\left((\mu_t^{(1)})^2 + (\sigma_t^{(1)})^2\right)^2} \right) \nabla'(\sigma_t^{(2)})^2(\hat{\theta}_2 - \theta_2) \right] \\
&= \sum_t \left[-h'_{1,t+1} B_1 H_1(t) + h'_{2,t+1} B_2 H_2(t) \right] + o_p(1) \\
&= \sum_t \left[-h'_{2,t+1} J' B_1 J H_2(t) + h'_{2,t+1} B_2 H_2(t) \right] + o_p(1) \\
&= \sum_t \left[h'_{2,t+1} M H_2(t) \right] + o_p(1) \\
&= \sum_t \left[h'_{2,t+1} B_2^{1/2} B_2^{-1/2} M B_2^{-1/2} B_2^{1/2} H_2(t) \right] + o_p(1) \\
&= \sum_t \left[h'_{2,t+1} B_2^{1/2} Q B_2^{1/2} H_2(t) \right] + o_p(1) \\
&= \sum_t \left[h'_{2,t+1} B_2^{1/2} C A A' C B_2^{1/2} H_2(t) \right] + o_p(1) \\
&= Z^2 \sum_t \tilde{H}'_2(t) \tilde{h}_{2,t+1} + o_p(1)
\end{aligned}
$$

Now we need to show that D_2, D_3, D_4 are lower order than D_1. The order of each element is showed below.

$$
A^{(1)} = O_p(1), \qquad B^{(1)} = O_p\left(\frac{1}{\sqrt{R}}\right), \qquad C^{(1)} = O_p\left(\frac{1}{\sqrt{R}}\right), \qquad C_2 = O_p\left(\frac{1}{R}\right)
$$

$$D_1 = \sum_t A^{(1)}C^{(1)} = \sum_t \left[O_p\left(1\right) O_p\left(\frac{1}{\sqrt{R}}\right)\right] = O_p\left(\frac{P}{\sqrt{R}}\right))$$

$$D_2 = \sum_t A^{(1)}C^{(2)} = \sum_t \left[O_p\left(1\right) O_p\left(\frac{1}{R}\right)\right] = O_p\left(\frac{P}{R}\right)$$

$$D_3 = \sum_t B^{(1)}C^{(1)} = \sum_t \left[O_p\left(\frac{1}{\sqrt{R}}\right) O_p\left(\frac{1}{\sqrt{R}}\right)\right] = O_p\left(\frac{P}{R}\right)$$

$$D_4 = \sum_t B^{(1)}C^{(2)} = \sum_t \left[O_p\left(\frac{1}{\sqrt{R}}\right) O_p\left(\frac{1}{R}\right)\right] = O_p\left(\frac{P}{R^{3/2}}\right)$$

Therefore, all D_2, D_3, D_4 are lower order than D_1. The proof is complete.

Lemma 6: Let Assumptions 1-4 hold.

$$\sum_t \left[V_2(\hat{\mu}_t^{(1)}, (\hat{\sigma}_t^{(1)})^2, y_{t+1}) \left(\hat{\sigma}_t^{2(1)} - \hat{\sigma}_t^{2(2)} \right) \right]^2 - P\bar{C}^2$$

$$= \sum_t \left[\left(\frac{1}{(\hat{\mu}_t^{(1)})^2 + (\hat{\sigma}_t^{(1)})^2} - \frac{y^2}{\left((\hat{\mu}_t^{(1)})^2 + (\hat{\sigma}_t^{(1)})^2 \right)^2} \right) \left((\hat{\sigma}_t^{(1)})^2 - (\hat{\sigma}_t^{(2)})^2 \right) \right]^2 - P\bar{C}^2$$

$$= Z^4 \sum_t \tilde{H}_2'(t)\tilde{H}_2(t) + o_p(1)$$

Proof:

$$\sum_t \left[V_2(\hat{\mu}_t^{(1)}, (\hat{\sigma}_t^{(1)})^2, y_{t+1}) \left((\hat{\sigma}_t^{(1)})^2 - (\hat{\sigma}_t^{(2)})^2 \right) \right]^2$$

$$= \sum_t \left[\left(\frac{1}{(\hat{\mu}_t^{(1)})^2 + (\hat{\sigma}_t^{(1)})^2} - \frac{y^2}{\left((\hat{\mu}_t^{(1)})^2 + (\hat{\sigma}_t^{(1)})^2 \right)^2} \right) \left((\hat{\sigma}_t^{(1)})^2 - (\hat{\sigma}_t^{(2)})^2 \right) \right]^2$$

$$= \sum_t \left[-h_{1,t+1}' B_1 H_1(t) + h_{2,t+1}' B_2 H_2(t) \right]^2 + o_p(1).$$

$$= \sum_t H_1'(t) B_1' h_{1,t+1} h_{1,t+1}' B_1 H_1(t) + \sum_t H_2'(t) B_2' h_{2,t+1} h_{2,t+1}' B_2 H_2(t)$$

$$- 2 \sum_t H_1'(t) B_1' h_{1,t+1} h_{2,t+1} B_2 H_2(t) + o_p(1)$$

$$= \sum_t H_1'(t) B_1' \mathbb{E}\left(h_{1,t+1} h_{1,t+1}'\right) B_1 H_1(t) + \sum_t H_2'(t) B_2' \mathbb{E}\left(h_{2,t+1} h_{2,t+1}'\right) B_2 H_2(t)$$

$$- 2 \sum_t H_1'(t) B_1' \mathbb{E}\left(h_{1,t+1} h_{2,t+1}'\right) B_2 H_2(t) + o_p(1)$$

$$= Z^2 \sum_t H_1'(t) B_1 H_1(t) + Z^2 \sum_t H_2'(t) B_2 H_2(t) - 2Z^2 \sum_t H_1'(t) B_2 H_2(t) + o_p(1)$$

$$= Z^2 \sum_t H_2'(t) J' B_1 J H_2(t) + Z^2 \sum_t H_2'(t) B_2 H_2(t) - 2Z^2 \sum_t H_2'(t) J' B_1 J H_2(t) + o_p(1)$$

$$= Z^2 \sum_t H_2'(t) \left[-J' B_1 J + B_2\right] H_2(t) + o_p(1)$$

$$= Z^2 \sum_t H_2'(t) M H_2(t) + o_p(1)$$

$$= Z^2 \sum_t H_2'(t) B_2^{1/2} B_2^{-1/2} M B_2^{-1/2} B_2^{1/2} H_2(t) + o_p(1)$$

$$= Z^2 \sum_t H_2'(t) B_2^{1/2} C A A' C B_2^{1/2} H_2(t) + o_p(1)$$

$$= Z^4 \sum_t \tilde{H}_2'(t) \tilde{H}_2(t) + o_p(1)$$

Lemma 3 implies that $\bar{C} = O_p(P^{-1})$, so $P\bar{C}^2 = o_p(1)$. Thus, the proof is completed.

Table 3.1: Simulation Results of DGP 1: Size of Test and Power of Test

	Size of Test											
	$P = 240$				$P = 480$				$P = 1200$			
Repeat=2000	DM_P	ENC_P	$CCS1_P$	$CCS2_P$	DM_P	ENC_P	$CCS2_P$		$CCS2_P$ DM_P	ENC_P	$CCS1_P$	$CCS2_P$
$R = 120$	0.006	0.054	0.045	0.023	0.000	0.050	0.052	0.022	0.000	0.062	0.053	0.087
$R = 240$	0.008	0.049	0.046	0.033	0.004	0.046	0.047	0.028	0.000	0.052	0.046	0.028
$R = 480$	0.012	0.052	0.054	0.056	0.006	0.051	0.052	0.047	0.000	0.053	0.045	0.022
	Power of Test											
	$P = 240$				$P = 480$				$P = 1200$			
Repeat=2000	DM_P	ENC_P	$CCS1_P$	$CCS2_P$	DM_P	ENC_P	$CCS2_P$		$CCS2_P$ DM_P	ENC_P	$CCS1_P$	$CCS2_P$
$R = 120$	0.024	0.201	0.044	0.070	0.013	0.373	0.053	0.241	0.004	0.737	0.049	0.795
$R = 240$	0.051	0.245	0.046	0.057	0.069	0.457	0.047	0.154	0.088	0.833	0.054	0.614
$R = 480$	0.120	0.274	0.049	0.050	0.177	0.551	0.049	0.118	0.321	0.915	0.044	0.465

The table above shows the size of DM_P, ENC_P, and CCS_P test in DGP under 5% nominal level, $b = 0$, $\phi = 0.5$.

The table below shows the size of DM_P, ENC_P, and CCS_P test in DGP under 5% nominal level, $b = 0.05$, $\phi = 0.5$.

Table 3.2: Combining Volatility Models

	$\delta = 0$		
	$R = 240, P = 240$	$R = 240, P = 480$	$R = 240, P = 1200$
$\hat{\lambda}_{R,P}$	-0.1683	0.0056	0.0098
$\mathbb{E}[S_g(\gamma_1, \gamma_3^{(1)}, y)]$	2.2420	2.2719	2.2766
$\mathbb{E}[S_g(\gamma_1, \gamma_3^{(2)}, y)]$	2.2618	2.3167	2.4678
$\mathbb{E}[S_g(\gamma_1, \gamma_3^{(c)}, y)]$	2.2373	2.2689	2.2756
	$\delta = 0.1$		
	$R = 240, P = 240$	$R = 240, P = 480$	$R = 240, P = 1200$
$\hat{\lambda}_{R,P}$	1.0723	0.8672	0.8142
$\mathbb{E}[S_g(\gamma_1, \gamma_3^{(1)}, y)]$	2.3178	2.4677	2.3387
$\mathbb{E}[S_g(\gamma_1, \gamma_3^{(2)}, y)]$	2.2617	2.4189	2.3291
$\mathbb{E}[S_g(\gamma_1, \gamma_3^{(c)}, y)]$	2.2614	2.4169	2.2796

Table above: Repeat=2000, $\delta = 0$. Since $\delta = 0$, Model 1 is the "oracle model".

Table below: Repeat=2000, $\delta = 0.1$. Since $\delta \neq 0$, Model 2 is the "oracle model".

The combined model is better than the "oracle model".

Table 3.3: Simulation Results of DGP 3: ENC-OCMT for GARCH Model

	P	TPR	FPR	FDR	FDR*
ENC-OCMT	100	1.0000	0.0052	0.0121	0.0121
p=0.05,	300	1.0000	0.0045	0.0090	0.0090
$\delta = 1, \delta^* = 2$	500	1.0000	0.0041	0.0092	0.0092
ENC-LASSO	200	1.0000	0.0641	0.3167	0.3167
	300	1.000	0.0603	0.3210	0.3210
	500	1.000	0.0585	0.3063	0.3063
ENC-Adaptive LASSO	200	0.9972	0.0101	0.0283	0.0283
	300	0.9991	0.0093	0.0275	0.0275
	500	1.0000	0.0089	0.0218	0.0218

$n = 100, R = 100$. Note that $n \gg R$.

Table 3.4: Simulation Results of DGP 4: ENC-OCMT for GARCH-MIDAS Model

	P	TPR	FPR	FDR	FDR*
ENC-OCMT	100	1.0000	0.0092	0.0143	0.0143
p=0.05,	300	1.0000	0.0070	0.0107	0.0107
$\delta = 1, \delta^* = 2$	500	1.0000	0.0089	0.0094	0.0094
ENC-LASSO	200	1.0000	0.0752	0.3823	0.3823
	300	1.000	0.0736	0.3986	0.3986
	500	1.000	0.0703	0.3701	0.3701
ENC-Adaptive LASSO	200	0.9946	0.0124	0.0315	0.0315
	300	0.9991	0.0102	0.0298	0.0298
	500	1.0000	0.0098	0.0235	0.0235

$n = 100, R = 100$. Note that $n \gg R$.

Table 3.5: Granger-Causality from Goyal-Welch Study

$R:P=0.25:0.75$				$R:P=0.5:0.5$				$R:P=0.75:0.25$			
INFL											
DM_P	ENC_P	CCS_P	$CCS2_P$	DM_P	ENC_P	CCS_P	$CCS2_P$	DM_P	ENC_P	CCS_P	$CCS2_P$
2.2691	-2.7857	-2.8673	-2.3104	4.1221	-4.5142	-2.4501	-2.4048	1.7046	-2.0271	1.4468	2.0197
(0.0116)	(0.0053)	(0.0041)	(0.0209)	(0.0000)	(0.0000)	(0.0143)	(0.0162)	(0.0441)	(0.0426)	(0.1480)	(0.0434)
BM											
DM_P	ENC_P	CCS_P	$CCS2_P$	DM_P	ENC_P	CCS_P	$CCS2_P$	DM_P	ENC_P	CCS_P	$CCS2_P$
1.9844	-3.0256	-2.5288	-3.0701	-0.1468	-0.1931	-2.4776	-3.3241	1.0969	-1.4005	2.6215	3.6848
(0.0236)	(0.0025)	(0.0114)	(0.0021)	(0.5584)	(0.8469)	(0.0132)	(0.0009)	(0.1364)	(0.1614)	(0.0088)	(0.0002)
SVAR											
DM_P	ENC_P	CCS_P	$CCS2_P$	DM_P	ENC_P	CCS_P	$CCS2_P$	DM_P	ENC_P	CCS_P	$CCS2_P$
1.7538	-2.6440	-1.3079	-0.9979	-3.2338	1.8413	-0.9013	-0.8715	0.2757	-0.4601	-0.6998	-0.8552
(0.0397)	(0.0082)	(0.1909)	(0.3183)	(0.9994)	(0.0656)	(0.3674)	(0.3835)	(0.3914)	(0.6454)	(0.4840)	(0.3924)
LTR											
DM_P	ENC_P	CCS_P	$CCS2_P$	DM_P	ENC_P	CCS_P	$CCS2_P$	DM_P	ENC_P	CCS_P	$CCS2_P$
0.9517	-1.9756	1.7878	-1.6995	2.2124	-2.5411	1.3675	-0.5446	-2.4760	2.3733	-0.4868	0.9037
(0.1706)	(0.0482)	(0.0738)	(0.0892)	(0.0135)	(0.0111)	(0.1715)	(0.5860)	(0.9934)	(0.0176)	(0.6264)	(0.3661)

Table 3.6: Model combination result from Welch-Goyal Study

$R:P=25\%:75\%$				$R:P=50\%:50\%$				$R:P=75\%:25\%$			
INFL		BM		INFL		BM		INFL		BM	
λ_P	Loss	$\hat{\lambda}_P$	Loss	$\hat{\lambda}_P$	Loss	$\hat{\lambda}_P$	Loss	$\hat{\lambda}_P$	Loss	$\hat{\lambda}_P$	Loss
	1.4242[b]		1.4242		1.4936		1.4936		1.7193		1.7193
1.6406[a]	1.4058[c]	1.2746	1.4097	2.1277	1.4766	0.2893	1.4955	3.6474	1.7079	2.3833	1.7131
	1.4034[d]		1.4091		1.4591		1.4933		1.6922		1.7095
$R:P=25\%:75\%$				$R:P=50\%:50\%$				$R:P=75\%:25\%$			
SVAR		LTR		SVAR		LTR		SVAR		LTR	
λ_P	Loss	$\hat{\lambda}_P$	Loss	$\hat{\lambda}_P$	Loss	$\hat{\lambda}_P$	Loss	$\hat{\lambda}_P$	Loss	$\hat{\lambda}_P$	Loss
	1.4242		1.4242		1.4936		1.4936		1.7193		1.7193
1.1084	1.4156	0.8448	1.4158	-0.1708	1.5063	2.8924	1.4834	1.4808	1.7185	-7.9864	1.7224
	1.4155		1.4156		1.4927		1.4754		1.7185		1.7028

[a]: The optimal weight on the second model $\hat{\lambda}$..

[b]: The average of the loss of volatility from forecasting of the first model $\phi^{(1)}$.

[c]: The average of the loss of volatility from forecasting of the second model $\phi^{(2)}$.

[d]: The average of the loss of volatility from forecasting of the combined model $\phi^{(c)}$.

Table 3.7: Predictor Selections for Stock Market Volatility Model

Models	OOS Forecast Error Loss
Model 0 (AR)	2.6237
ENC-LASSO	2.9733
ENC-Adaptive LASSO	2.7349
ENC-OCMT	2.5915
POST-ENC-OCMT	2.5915

The reported value is OOS Forecast Error Loss $\times 10^5$.

Chapter 4

Forecast Encompassing Principle in High Dimensional Predictive Models of Growth-at-Risk and Growth-Shortfalls

4.1 Introduction

The necessity for reliable downside risk metrics in macroeconomic risk management has been increasingly recognized. The past decade has seen a surge in policymakers focusing on downside risk. Recently, the International Monetary Fund (IMF) popularized a risk measure for GDP growth named Growth-at-Risk (GaR), defined by Adrian et al. (2019) as the worst conditional GDP growth distribution at a specific coverage level (5th percentile)

conditional on financial states. However, GaR carries two major drawbacks. Firstly, GaR lacks subadditivity, rendering it incapable of satisfying the coherence criteria introduced by Artzner et al. (1999). The absence of subadditivity. Secondly, GaR concentrates on the probability rather than the magnitude of losses, leading to potential discrepancies in extreme losses between investments with identical GaR. Consequently, GaR may not be a fitting risk measure in certain contexts. In response to these limitations, Chavleishvili et al. (2021) introduced a measure of adverse real economic impact, Growth Shortfall (GS), defined as the expected GDP growth when it falls below GaR. Hence, GaR and GS parallel the definitions of VaR and ES, respectively, with the former pair assessing financial returns risk and the latter evaluating GDP growth risk.

However, GS risk measures are not "elicitable" according to Gneiting (2011), signifying that there is no "strictly consistent" scoring function for these risk measures. This hinders the comparison of competing forecasts of these risk measures concerning the consistent score (Fissler & Ziegel, 2016). Consequently, evaluating and comparing GS forecasts prove challenging due to this shortcoming. Gneiting (2011) observed that while the mean is elicitable, the variance is not. Yet, a pair of mean and variance proves elicitable. Fissler and Ziegel (2016) furthered this by introducing a strictly consistent scoring function for GaR and GS, rendering the pair of GaR and GS elicitable. This availability of a strictly consistent scoring function for GaR and GS (the FZ scoring function) has expedited the progress of GS forecasting. For example, Patton et al. (2019) proposed dynamic models for VaR and ES forecasting utilizing the FZ scoring function, and Taylor (2019) leveraged the FZ scoring function to forecast VaR and ES based on the asymmetric Laplace distribution.

In the current paper, the FZ scoring function is used for in-sample estimation and out-of-sample forecasting to test Granger causality.

The technique of out-of-sample forecast comparison is extensively utilized across various disciplines since it provides a method to test Granger causality, which determines the predictability of a dependent variable based on independent variables (Ashley et al., 1980). There is a wide array of literature focusing on out-of-sample tests for equality of accuracy and encompassing (Diebold & Mariano, 1995; Clark & West, 2006, 2007). This requires two models. A pair of models are considered "nested" when a larger model encompasses all variables in a smaller model, plus one or more other variables absent in the smaller model. On the other hand, two models are "non-nested" when they share the same variables, with each model also containing one or more unique variables. These non-nested models overlap without one nesting the other.

In situations where non-nested models are compared, Diebold and Mariano (1995) presented the DM statistic for predictive accuracy comparison. Under the null hypothesis of no difference in predictive accuracy, the DM test statistic asymptotically approaches a N(0,1) distribution. However, Harvey et al. (1997) indicated that while the DM test performs adequately for large samples, it could be oversized for moderate samples. Therefore, they amended the DM statistic by correcting the bias (HNL statistic), employing a critical value from the student's t-distribution instead of the standard normal distribution used in the DM test. But, for nested model comparisons, the DM statistic for equal accuracy and the HNL statistic for encompassing are invalid due to their inability to converge to the standard normal distribution (Clark and McCracken, 2001). Further, Clark and West

(2006, 2007) demonstrated that the DM statistic for mean regression bears a downside bias, which can be rectified by adding a non-negative adjustment term. In the present paper, the mean regression from Clark and West (2006, 2007) is extended to GS, and it's shown that the DM statistic for GS using the FZ scoring function also has a bias under the null hypothesis. The paper then develops the encompassing statistic for Granger causality in GS, demonstrating satisfactory performance in size and power tests.

In the fields of macroeconomics and finance, we encounter a profusion of macroeconomic and financial variables, most of which provide quarterly data. Consequently, when evaluating predictive abilities, we often encounter a larger number of variables than observations. Additionally, as we increase the pool of variables in a test, we typically find a rise in the number of predictors that show a predictive ability for the variable of interest. This prevalence has led to the widespread use of high-dimensional predictive regression models across various disciplines, which is the focus of this paper.

Typically, predictor selection is conducted by evaluating the marginal effect of a predictor, selecting it if this effect is non-zero. However, this approach may overlook important variables within the model. As the number of predictors increases, so does the likelihood of correlations existing amongst them. Even if certain predictors show zero marginal effect on the variable of interest (the forecast target), they may still indirectly impact it through their correlation with other predictors. Therefore, it is vital to consider both the direct marginal effects and indirect net impacts when selecting predictors, especially in the context of numerous predictors.

To address this issue, we have expanded upon the multiple testing approach, termed One Covariate at a Time Multiple Testing (OCMT), as proposed by Chudik et al. (2018). The OCMT approach concentrates on the marginal effect and net impact of the variables on the target variable for variable selection, utilizing total sample observations. To select predictors in high-dimensional predictive models, we introduce a novel encompassing approach that incorporates the OCMT method, termed as "ENC-OCMT." Our ENC-OCMT approach takes into account both marginal effect and net impact for predictor selection. While the OCMT procedure selects variables based on the in-sample test in the regression model, our ENC-OCMT approach selects predictors using an out-of-sample test in the predictive regression model.

A popular method for variable selection is the least absolute shrinkage and selection operation (LASSO), as introduced by Tibshirani (1996). Yet, most LASSO-based variable selections rely on full sample observations (Fan & Li, 2001; Efron et al., 2004; Zou & Hastie, 2005; Candes & Tao, 2007; Bickel et al., 2009; Zhang, 2010). We propose a fresh approach for selecting predictors utilizing the LASSO method, which is grounded on the out-of-sample forecast encompassing principle. This innovative approach, which we term "ENC-LASSO," presents a comparative advantage to the traditional LASSO method as it facilitates out-of-sample predictor selection.

4.2 Objective function for GaR and GS

Consider an observation domain, O, $y, x \in \mathrm{O} \subseteq \mathbb{R}^{d_1+d_2}$, $d_1 = \dim(y)$ and $d_2 = \dim(x)$, the conditional distribution $F \equiv F(Y_{t+1}|\mathcal{I}_t)$ for Y_{t+1} given X_t. Let $\mathcal{F}$ be a class of

distribution function on the observation domain O, and let A be an action domain, $\gamma \in$ A. We define $\Gamma : \mathcal{F} \to$ A to be functional. For example, $\Gamma(F(y|x))$ may be $\mathbb{E}(Y|X)$, $\mathrm{Q}(Y|X)$, $\mathrm{Var}(Y|X)$, $\mathrm{Mode}(Y|X)$ or $\mathrm{GS}(Y|X)$, where $\mathbb{E}(Y|X)$ is the conditional mean, $\mathrm{Q}(Y|X)$ is the conditional quantile, $\mathrm{Var}(Y|X)$ is the conditional variance, $\mathrm{Mode}(Y|X)$ is the conditional mode and $\mathrm{GS}(Y|X)$ is the conditional Growth Shortfall. Note that Γ can be a vector of several of these.

Definition 1: (Gneiting, 2011; Fissler et al., 2016) A scoring function is an $\mathcal{F}$-integrable function $S : A \times O \to \mathbb{R}$. S is said to be *$\mathcal{F}$-consistent* for a functional $\Gamma : \mathcal{F} \to$ A if $\mathbb{E}_F S(\Gamma(F), Y) \leq \mathbb{E}_F S(\gamma, Y)$ for all $F \in \mathcal{F}$ and for all $\gamma \in$ A. Furthermore, S is *strictly $\mathcal{F}$-consistent* for Γ if it is $\mathcal{F}$-consistent for Γ and if $\mathbb{E}_F S(\Gamma(F), Y) = \mathbb{E}_F S(\gamma, Y)$ implies that $\gamma = \Gamma(F)$ for all $F \in \mathcal{F}$ and for all $\gamma \in$ A.

Definition 2: (Gneiting, 2011; Fissler et al., 2016) An identification function is an $\mathcal{F}$-integrable function $V : A \times O \to \mathbb{R}^k$. V is said to be an *$\mathcal{F}$-identification function* for a functional $\Gamma : \mathcal{F} \to \mathrm{A} \subseteq \mathbb{R}^k$ if $\mathbb{E}_F V(\Gamma(F), Y) = 0$ for all $F \in \mathcal{F}$. Furthermore, V is a *strict $\mathcal{F}$-identification function* for Γ if $\mathbb{E}_F V(\gamma, Y) = 0$ holds if and only if $\gamma = \Gamma(F)$ for all $F \in \mathcal{F}$ and for all $\gamma \in$ A.

According to Fissler et al. (2016), there is a relation among strictly consistent scoring function S, strict identification function V, and a matrix-valued function h. We define a vector-valued $\mathbf{m}$ to be a vector of m_1 and m_2, and a vector-valued $\boldsymbol{\gamma}$ to be a vector of γ_1

and γ_2. Taking the first order conditions of the expected loss function with respect to γ, γ_1 and γ_2 respectively, we can get $\mathbb{E}_F m$, $\mathbb{E}_F m_1$, and $\mathbb{E}_F m_2$ to be

$$\mathbb{E}_F \mathbf{m} \equiv \frac{\partial \mathbb{E}_F S(\boldsymbol{\gamma}, Y)}{\partial \boldsymbol{\gamma}} = \mathbb{E}_F \mathbf{h}(\boldsymbol{\gamma}) \mathbf{V}(\boldsymbol{\gamma}, Y) = 0, \tag{4.1}$$

$$\begin{aligned}
\mathbb{E}_F m_1 \equiv & \frac{\partial \mathbb{E}_F S(\gamma_1, \gamma_2, Y)}{\partial \gamma_1} = \mathbb{E}_F h_{11}(\gamma_1, \gamma_2) V_1(\gamma_1, \gamma_2, Y) + \mathbb{E}_F h_{12}(\gamma_1, \gamma_2) V_2(\gamma_1, \gamma_2, Y) = 0, \\
\mathbb{E}_F m_2 \equiv & \frac{\partial \mathbb{E}_F S(\gamma_1, \gamma_2, Y)}{\partial \gamma_2} = \mathbb{E}_F h_{21}(\gamma_1, \gamma_2) V_1(\gamma_1, \gamma_2, Y) + \mathbb{E}_F h_{22}(\gamma_1, \gamma_2) V_2(\gamma_1, \gamma_2, Y) = 0,
\end{aligned}$$

where

$$\mathbf{m} = [m_1, m_2]', \qquad \boldsymbol{\gamma} = [\gamma_1, \gamma_2]',$$

$$\mathbf{V}(\gamma_1, \gamma_2, Y) = \begin{bmatrix} V_1(\gamma_1, \gamma_2, Y) \\ V_2(\gamma_1, \gamma_2, Y) \end{bmatrix},$$

$$\mathbf{h}(\gamma_1, \gamma_2) = \begin{bmatrix} h_{11}(\gamma_1, \gamma_2) & h_{12}(\gamma_1, \gamma_2) \\ h_{21}(\gamma_1, \gamma_2) & h_{22}(\gamma_1, \gamma_2) \end{bmatrix}.$$

Concerning equation (7.1), the strict identification function V is a function of variable y and functional γ, and the function h is a function of γ only. There is no y in the function of h. Functional forms of V1 and V2 are shown in Eq (6.3).

Despite the GaR (quantile) being elicitable, GS is not (Gneiting, 2011). However, Fissler et al. (2016) propose a scoring function strictly consistent for joint GaR and GS. Consequently, a pair of GaR and GS becomes 2-elicitable.

$$\begin{aligned}
S(\gamma_1, \gamma_2, y) = & (1\{y \leq \gamma_1\} - \alpha) G_1(\gamma_1) - 1\{y \leq \gamma_1\} G_1(y) \\
& + G_2(\gamma_2) \left(\gamma_2 - \gamma_1 + \frac{1}{\alpha} 1\{y \leq \gamma_1\} (\gamma_1 - y) \right) - \mathcal{G}(\gamma_2) + a(y),
\end{aligned} \tag{4.2}$$

where γ_1 represents GaR and γ_2 represents GS, $\mathcal{G}' = G_2$, G_1 is increasing function, and G_2 is increasing and convex function. Therefore, a pair of VaR and ES is 2-elicitable. With

regard to Fissler et al. (2016), taking the derivatives of the scoring function with respect to γ_1 and γ_2, we have

$$\mathbb{E}_F m_1 = \frac{\partial \mathbb{E}_F S(\gamma_1, \gamma_2, Y)}{\partial \gamma_1} = \mathbb{E}_F \left[G_1'(\gamma_1) + \frac{1}{\alpha} G_2(\gamma_2) \right] (1\{y \leq \gamma_1\} - \alpha) = 0,$$

$$\begin{aligned}\mathbb{E}_F m_2 = \frac{\partial \mathbb{E}_F S(\gamma_1, \gamma_2, Y)}{\partial \gamma_2} &= \mathbb{E}_F \frac{\gamma_1 G_2'(\gamma_2)}{\alpha} (1\{y \leq \gamma_1\} - \alpha) + \mathbb{E}_F G_2'(\gamma_2) \left(\gamma_2 - \frac{1}{\alpha} y 1\{y \leq \gamma_1\} \right) \\ &= 0,\end{aligned}$$

where the strict identification function is

$$\mathbf{V}(\gamma_1, \gamma_2, y) = \begin{bmatrix} V_1(\gamma_1, \gamma_2, Y) \\ \\ V_2(\gamma_1, \gamma_2, Y) \end{bmatrix} = \begin{bmatrix} 1\{y \leq \gamma_1\} - \alpha \\ \\ \gamma_2 - \frac{1}{\alpha} y 1\{y \leq \gamma_1\} \end{bmatrix}, \tag{4.3}$$

and the matrix-valued function h is

$$\mathbf{h}(\gamma_1, \gamma_2) = \begin{bmatrix} G_1'(\gamma_1) + \frac{1}{\alpha} G_2(\gamma_2) & 0 \\ \\ \frac{\gamma_1 G_2'(\gamma_2)}{\alpha} & G_2'(\gamma_2) \end{bmatrix}. \tag{4.4}$$

For the first component of the identification function V, $\mathbb{E}_F (1\{y \leq \gamma_1\} - \alpha) = 0$ if and only if $\gamma_1 = \text{VaR}_\alpha(y)$. For the second component of V, $\mathbb{E}_F \left(\gamma_2 - \frac{1}{\alpha} y 1\{y \leq \gamma_1\} \right)$ if and only if $\gamma_2 = \text{ES}_\alpha(y)$. Therefore, $\mathbb{E}_F V(\gamma_1, \gamma_2, y) = 0$ if and only if $(\gamma_1; \gamma_2)$ equals the true GaR and GS of y, so the joint functional of VaR (Quantile) and ES is 2-elicitible. It indicates that this scoring function is strictly consistent for a pair of GaR and GS, which means that it can be used for forecasts comparison of GaR and GS.

The importance of the positive homogeneity of the scoring function for forecast comparison is noted by Patton (2011b) states that positive homogeneity of the scoring function is important for forecast comparison. Furthermore, Patton and Sheppard (2009) demonstrates that zero degree homogeneity results in greater power for the Diebold-Mariano tests in volatility forecasting.

Patton et al. (2019) denote S_{FZ0} as the scoring function with homogeneity of degree zero. S_{FZ0} is homogeneous of degree zero if and only if $G_1(\gamma) = 0$ and $G_2(\gamma) = -1/\gamma$. For easy to read, let q replace γ_1 and e replace γ_2. Then, the "FZ0" scoring function is

$$S_{FZ0}\left(q_\alpha, e_\alpha, y\right) = S_\alpha\left(q_\alpha, e_\alpha, y\right) = -\frac{1}{\alpha e_\alpha}\mathbf{1}\{y \leq q_\alpha\}(q_\alpha - y) + \frac{q_\alpha}{e_\alpha} + \log(-e_\alpha) - 1. \quad (4.5)$$

4.3 Forecast Encompassing Test for Granger Causality in GaR and GS

In this section, we leverage this FZ scoring function S_{FZ0} to test forecast encompassing in Growth Shortfall. We subsequently develop a forecast encompassing test in the GS, where the conditional distribution $F \equiv F(Y_{t+1}|\mathcal{I}_t)$ for Y_{t+1} given X_t. We have two models: one unconditioned on x_t, and the other conditioned on x_t. The unconditional GS of the unconditional distribution $F_1 = F_Y(Y)$ is $\Gamma(F_1)$. The conditional GS of the conditional distribution $F_2 = F_{Y|X}(Y|X)$ is the functional $\Gamma(F_2)$. The two nested models for GaR q_{t+1} are the conditional distribution $F_2 = F_{Y|X}(Y|X)$ is the functional $\Gamma(F_2)$. The $n+1$ nested models are

$$\text{Model } 0: \quad y_{t+1} = \sigma_{0,t+1}\eta_{0,t+1} \quad (4.6)$$

$$\sigma^2_{0,t+1} = \omega + \beta\sigma^2_{0,t} + \gamma y_t^2 \quad (4.7)$$

$$\text{Model } i: \quad y_{t+1} = \sigma_{i,t+1}\eta_{i,t+1} \quad (4.8)$$

$$\sigma^2_{i,t+1} = \omega + \beta\sigma^2_{i,t} + \gamma y_t^2 + \pi x_t^2 \quad (4.9)$$

where $\eta_t \sim iid\ N(0,1)$. Under this setting, the GaR and GS are proportional to σ_{t+1}, with

$$\left(\text{VaR}^\alpha_{t+1}, \text{ES}^\alpha_{t+1}\right) = \left(q^\alpha_{t+1}, e^\alpha_{t+1}\right) = \left(a_\alpha, b_\alpha\right)\sigma_{t+1}$$

For a standard Normal distribution, with CDF and PDF denoted Φ and ϕ, we have:

$$a_\alpha = \Phi^{-1}(\alpha)$$

$$b_\alpha = -\phi\left(\Phi^{-1}(\alpha)\right)/\alpha$$

The FZ scoring function is the scoring function S_{FZ0} in Patton et al. (2019), that is

$$S_\alpha(q, e, y) = -\frac{1}{\alpha e}\mathbf{1}\{y \leq q\}(q - y) + \frac{q}{e} + \log(-e) - 1.$$

According to Patton et al. (2019), the derivative of FZ scoring function with respect to GaR q_{t+1} and GS e_{t+1} are

$$\mathbb{E}_t\left[-\frac{1}{\alpha e_{t+1}}\mathbf{1}\{Y_{t+1} \leq q_{t+1}\} + \frac{1}{e_{t+1}}\right] = 0, \tag{4.10}$$

$$\mathbb{E}_t\left[\frac{1}{\alpha\left(e_{t+1}\right)^2}\mathbf{1}\left\{Y_{t+1} \leq q_{t+1}\right\}\left(q_{t+1} - Y\right) - \frac{q_{t+1}}{e_{t+1}^2} + \frac{1}{e_{t+1}}\right] = 0, \tag{4.11}$$

which means that $m_1 = -\frac{1}{\alpha e_{t+1}}\mathbf{1}\{y_{t+1} \leq q_{t+1}\} + \frac{1}{e_{t+1}}$ and $m_2 = \frac{1}{\alpha e_{t+1}^2}\mathbf{1}\{y \leq q_{t+1}\}(q_{t+1} - y_{t+1}) - \frac{q_{t+1}}{e_{t+1}^2} + \frac{1}{e_{t+1}}$ are martingale difference sequences. In the last section, we have already defined $m = (m_1, m_2)'$. So, rewriting equations (6.8) and (6.9), we have

$$\mathbb{E}_t\mathbf{m} = \mathbb{E}_t\mathbf{h}(q_{t+1}, e_{t+1})\mathbf{V}(q_{t+1}, e_{t+1}, y_{t+1}) = 0, \tag{4.12}$$

where the strict identification functions in equation (7.1) are

$$\mathbf{V}(q_{t+1}, e_{t+1}, y_{t+1}) = \begin{bmatrix} V_1(q_{t+1}, e_{t+1}, y_{t+1}) \\ V_2(q_{t+1}, e_{t+1}, y_{t+1}) \end{bmatrix} = \begin{bmatrix} 1\{y_{t+1} \leq q_{t+1}\} - \alpha \\ e_{t+1} - \frac{1}{\alpha}y_{t+1}1\{y_{t+1} \leq q_{t+1}\} \end{bmatrix}, \tag{4.13}$$

and the matrix-valued function h is

$$\mathbf{h}(q_{t+1}, e_{t+1}) = \begin{bmatrix} -\frac{1}{\alpha e_{t+1}} & 0 \\ -\frac{q_{t+1}}{\alpha e_{t+1}^2} & \frac{1}{e_{t+1}^2} \end{bmatrix}. \tag{4.14}$$

In equation (6.11), V_1 shows how we define VaR (quantile). We have $\mathbb{E}_t V_1 = 0$, so $\mathbb{E}_t(1\{y_{t+1} \leq q_{t+1}\} - \alpha) = 0$ implies $\mathbb{E}_t(1\{y_{t+1} \leq q_{t+1}\}) = \alpha$. V_2 shows how we define ES. We know that $\mathbb{E}_t V_2 = 0$, so $\mathbb{E}_t(e_{t+1} - \frac{1}{\alpha} y_{t+1} 1\{y_{t+1} \leq q_{t+1}\})$ implies $e_{t+1} = \frac{1}{\alpha}\mathbb{E}_t(y_{t+1} 1\{y_{t+1} \leq q_{t+1}\})$.

In order to test the equal predictive accuracy of the two nested models, we set the null and alternative hypotheses are

$$\mathbb{H}_0 : \mathbb{E}_t \left[S_\alpha \left(q_{0,t+1}, e_{0,t+1}, Y_{t+1} \right) - S_\alpha \left(q_{i,t+1}, e_{i,t+1}, Y_{t+1} \right) \right] = 0 \tag{4.15}$$

$$\mathbb{H}_1 : \mathbb{E}_t \left[S_\alpha \left(q_{0,t+1}, e_{0,t+1}, Y_{t+1} \right) - S_\alpha \left(q_{i,t+1}, e_{i,t+1}, Y_{t+1} \right) \right] > 0 \tag{4.16}$$

We set the alternative hypothesis is one side because we expect that the forecasts from Model 2 are better than those from Model 1, as Clark and West (2006) did. When coefficient $b = 0$, x does not Granger cause y. When coefficient $b \neq 0$, x Granger causes y. Define R as the in-sample observations. Define P as the out-of-sample forecasts. The DM statistics based on the FZ loss differential is

$$D = \mathbb{E}_t \left[S_\alpha \left(q_{0,t+1}, e_{0,t+1}, Y_{t+1} \right) - S_\alpha \left(q_{i,t+1}, e_{i,t+1}, Y_{t+1} \right) \right]$$

$$\hat{D}_{R,P} = P^{-1} \sum_{t=R}^{T} \left[S_\alpha \left(\hat{q}_{0,t+1}, \hat{e}_{0,t+1}, Y_{t+1} \right) - S_\alpha \left(\hat{q}_{i,t+1}, \hat{e}_{i,t+1}, Y_{t+1} \right) \right]$$

where $T+1 = R+P$. For mean regression with nested models, Clark and West (2006, 2007) prove that the DM statistic has downward bias and show that adjusted DM statistic (DM statistics plus the adjusted term) obtains zero mean expectation under the null hypothesis, corrects the size, and increases the power. Thus, at first, we attempt to show that the DM statistic based on the FZ loss differential $\hat{D}_{R,P}$ has a non-zero mean under the null

hypothesis.

Thereom 1: the DM statistic of the FZ loss differential $\hat{D}_{R,P}$ has a non-zero mean. Thus, $\mathbb{E}\hat{D}_{R,P} \neq 0$ under $\mathbb{H}_0$.

Proof: In order to show that $\hat{D}_P$ is non-zero, we consider eight cases. (1).$Y_{t+1} > \hat{q}_{0,t+1}$, $Y_{t+1} > \hat{q}_{i,t+1}$, and $\hat{e}_{0,t+1} > \hat{e}_{i,t+1}$. (2).$Y_{t+1} > \hat{q}_{0,t+1}$, $Y_{t+1} > \hat{q}_{i,t+1}$ and $\hat{e}_{0,t+1} < \hat{e}_{i,t+1}$. (3).$Y_{t+1} \leq \hat{q}_{0,t+1}$, $Y_{t+1} > \hat{q}_{i,t+1}$ and $\hat{e}_{0,t+1} > \hat{e}_{i,t+1}$. (4).$Y_{t+1} \leq \hat{q}_{0,t+1}$, $Y_{t+1} > \hat{q}_{i,t+1}$, and $\hat{e}_{0,t+1} < \hat{e}_{i,t+1}$. (5).$Y_{t+1} > \hat{q}_{0,t+1}$, $Y_{t+1} \leq \hat{q}_{i,t+1}$ and $\hat{e}_{0,t+1} > \hat{e}_{i,t+1}$. (6).$Y_{t+1} > \hat{q}_{0,t+1}$, $Y_{t+1} \leq \hat{q}_{i,t+1}$, and $\hat{e}_{0,t+1} < \hat{e}_{i,t+1}$. (7).$Y_{t+1} \leq \hat{q}_{0,t+1}$, $Y_{t+1} \leq \hat{q}_{i,t+1}$ and $\hat{e}_{0,t+1} > \hat{e}_{i,t+1}$. (8).$Y_{t+1} \leq \hat{q}_{0,t+1}$, $Y_{t+1} \leq \hat{q}_{i,t+1}$ and $\hat{e}_{0,t+1} < \hat{e}_{i,t+1}$. In these eight cases, we can show that $\hat{D}_{R,P}$ has a non-zero mean. See Appendix A2 for details.

Due to the bias of DM statistic, we want to develop the encompassing test to show that the encompassing statistic has zero mean. In order to find out the encompassing statistic for the ES, we build a combined FZ scoring function by combining two models, VaR q and ES e, and take the derivative of the expectation of the combined FZ scoring function with respect to the weight λ.

Theorem 2: Under $\mathbb{H}_0$,

$$C_i \equiv \mathbb{E}_t \left[m_{1,0} \left(\hat{q}_{i,t+1} - \hat{q}_{0,t+1} \right) + m_{2,0} \left(\hat{e}_{i,t+1} - \hat{e}_{0,t+1} \right) \right] = 0. \tag{4.17}$$

$$\hat{C}_{i,R,P} \equiv P^{-1} \sum_{t=R}^{T} \left[\hat{m}_{1,0,t+1} \left(\hat{q}_{i,t+1} - \hat{q}_{0,t+1} \right) + \hat{m}_{2,0,t+1} \left(\hat{e}_{i,t+1} - \hat{e}_{0,t+1} \right) \right] \xrightarrow{p} C_i = 0 \tag{4.18}$$

as $R, P \to \infty$ and $P/R \to \infty$.

Proof: We combine model 1 and model 2 with weight $1-\lambda$ and λ respectively. To estimate the expectation of the FZ scoring function with combined GaR $q_{t+1}^{(c)}$ and GS $e_{t+1}^{(c)}$ to find out the encompassing statistic under the null hypothesis, we have

$$\lambda = \arg\min_{\lambda} \mathbb{E}_t S_\alpha(q_{c,t+1}, e_{c,t+1}, Y_{t+1}),$$

where $q_{c,t+1} = (1-\lambda)\, q_{0,t+1} + \lambda q_{i,t+1}$ and $e_{c,t+1} = (1-\lambda)\, e_{0,t+1} + \lambda e_{i,t+1}$. We define C to be the derivative of the expected FZ scoring function with respect to λ, then we have

$$\begin{aligned}
C_i &\equiv \frac{\partial \mathbb{E}_t \left[S_\alpha(q_{c,t+1}, e_{c,t+1}, Y_{t+1}) \right]}{\partial \lambda} \\
&= \frac{\partial \mathbb{E}_t S_\alpha(q_{c,t+1}, e_{c,t+1}, Y_{t+1})}{\partial q_{c,t+1}} \frac{\partial q_{c,t+1}}{\partial \lambda} + \frac{\partial \mathbb{E}_t S_\alpha(q_{c,t+1}, e_{c,t+1}, Y_{t+1})}{\partial e_{c,t+1}} \frac{\partial e_{c,t+1}}{\partial \lambda} \\
&= \mathbb{E}_t \left[m_{1,c} \left(q_{i,t+1} - q_{0,t+1} \right) + m_{2,c} \left(e_{i,t+1} - e_{0,t+1} \right) \right] \\
&= \mathbb{E}_t \left[m_{1,0} \left(q_{i,t+1} - q_{0,t+1} \right) + m_{2,0} \left(e_{i,t+1} - e_{0,t+1} \right) \right] = 0
\end{aligned}$$

under $\mathbb{H}_0$. Due to the derivative, C_i should be 0. We estimate C by $\hat{C}_{i,R,P}$, so $\hat{C}_{R,P}$ is defined as

$$\hat{C}_{i,R,P} \equiv P^{-1} \sum_{t=R}^{T} \left[\hat{m}_{1,0,t+1} \left(\hat{q}_{i,t+1} - \hat{q}_{0,t+1} \right) + \hat{m}_{2,0,t+1} \left(\hat{e}_{i,t+1} - \hat{e}_{0,t+1} \right) \right] \xrightarrow{p} C_i = 0 \qquad (4.19)$$

under $\mathbb{H}_0$, as $R, P \to \infty$ and $P/R \to \infty$. Under the null hypothesis, $\lambda = 0$, we obtain $q_{c,t+1} = q_{0,t+1}$ and $e_{c,t+1} = e_{0,t+1}$. Thus, we obtain $\mathbb{E}(\hat{C}_{i,R,P}) = 0$, so $\hat{C}_{i,R,P} \xrightarrow{p} 0$ as $R, P \to \infty$. See Appendix A2 for details.

In order to get the encompassing statistic for ES, we consider endowing Model

1 and Model 2 with the weight $1-\lambda$ and λ respectively, and find out the property of optimal weight $\hat{\lambda}$. Theorem 3 gives the relationship between $\hat{\lambda}$ and $\hat{C}_P$.

Theorem 3: We denote $\lambda = \arg\min_\lambda \mathbb{E}_t S\left(q_{c,t+1}, e_{c,t+1}, Y_{t+1}\right)$. Under $\mathbb{H}_0, \hat{\lambda} \to 0$, $\hat{C}_{i,R,P} \xrightarrow{p} 0$ as $R, P \to \infty$.

Proof: See Appendix B.

This theorem shows that $\hat{\lambda} \to 0$ and $\hat{C}_{R,P} \xrightarrow{p} 0$ as $R, P \to \infty$ are equivalent.

We compare the DM statistic and the encompassing statistic with the CCS statistic, the test statistic by Chao et al. (2001). Chao et al. (2001) show that the CCS statistic is $\hat{M}_{i,R,P} = P^{-1}\sum_{t=R}^{T} \hat{u}_{0,t+1} x_t$. In order to compare the equal predictive accuracy of two nested expectile models, we standardize these three statistics. The DM statistic is $\mathrm{DM}_{i,R,P} \equiv \hat{S}_{i,R,P}^{-0.5}\sqrt{P}\hat{D}_{i,R,P}$, where $S_{i,R,P} = \mathrm{var}\left(\sqrt{P}\hat{D}_{i,R,P}\right)$ and $S_{i,R,P} - \hat{S}_{R,P} \xrightarrow{p} 0$. The encompassing statistic is $\mathrm{ENC}_P \equiv \hat{Q}_{i,R,P}^{-0.5}\sqrt{P}\hat{C}_{i,R,P}$, where $Q_{i,R,P} = \mathrm{var}\left(\sqrt{P}\hat{C}_{i,R,P}\right)$ and $Q_{i,R,P} - \hat{Q}_{i,R,P} \xrightarrow{p} 0$. The CCS statistic is $\mathrm{CCS}_P \equiv \hat{W}_{i,R,P}^{-0.5}\sqrt{P}\hat{M}_{i,R,P}$, where $W_{i,R,P} = \mathrm{var}\left(\sqrt{P}\hat{M}_{i,R,P}\right)$ and $W_{i,R,P} - \hat{W}_{i,R,P} \xrightarrow{p} 0$.

4.4 Asymptotic Normality of the ENC Statistic for GaR and GS

In this section, we compare two nested models using encompassing test and model combination, then explore the properties of the two methodologies.

4.4.1 Encompassing Test for Equal Predictive Accuracy

We derive the asymptotic distribution of ENC_P. To show ENC_P to be asymptotical standard normality, we simplify the ENC_P statistic.

$$\text{ENC}_{i,R,P} = Q_P^{-1/2}\sqrt{P}\hat{C}_{i,R,P} = \frac{\sum_t c_t}{\sqrt{\sum_t c_t^2 - P\hat{C}_{i,R,P}^2}},$$

where

$$\hat{m}_{1,0,t+1} = -\frac{1}{\alpha\hat{e}_{0,t+1}\mathbf{1}\{Y_{t+1} \leq \hat{q}_{0,t+1}} + \frac{1}{\hat{e}_{0,t+1}}$$

$$\hat{m}_{2,0,t+1} = \frac{1}{\alpha\left(\hat{e}_{0,t+1}\right)^2}\mathbf{1}\left\{Y_{t+1} \leq \hat{q}_{0,t+1}\right\}\left(\hat{q}_{0,t+1} - Y_{t+1}\right) - \frac{\hat{q}_{0,t+1}}{\left(\hat{e}_{0,t+1}\right)^2} + \frac{1}{\hat{e}_{0,t+1}}.$$

$$\hat{C}_{i,R,P} = P^{-1}\sum_{t=R}^{T} c_t = P^{-1}\sum_{t=R}^{T}\hat{m}_{1,0,t+1}\left(\hat{q}_{i,t+1} - \hat{q}_{0,t+1}\right) - \hat{m}_{2,0,t+1}\left(\hat{e}_{i,t+1} - \hat{e}_{0,t+1}\right).$$

In order to obtain the limiting distribution in Theorem 4, we need some notations. We define $g\left(u_{i,t+1}\right) = [\alpha - 1\left(u_{i,t+1} < 0\right)]$. The score $k_{i,t}(\beta_{i,t})$ is the derivative of the expected FZ scoring function with respect to $\beta_{i,t}$. Let $h_{i,t+1} = -k_{i,t+1}(\beta_{i,t})$ and $H_i(t+1) - \frac{1}{R}\sum_{t=1}^{R} h_{i,t+1}$. We denote the Hessian to be $B_i^{-1} = \mathbb{E}_F\Lambda(\beta_{i,t})$, which is the second order condition of expected FZ scoring function with respect to $\beta_{i,t}$. Therefore, $B_1 = 1$ and $B_2 = \begin{pmatrix} 1 & 0 \\ 0 & \mathbb{E}_F\Lambda(\beta_{2,t}) \end{pmatrix}^{-1}$. By the central limit theorem, $\sqrt{R}\left[R^{-1}\sum_{s=t-R+1}^{t} h_{2,s}\right] \sim N\left(0, Z^2B_2^{-1}\right)$, where $Z^2B_2^{-1} = \mathbb{E}[k_{i,t}(\beta_{i,t})k_{i,t}(\beta_{i,t})']$.

The following assumptions can be used to obtain the limiting distribution in Theorem 4. These assumptions are only sufficient but not necessary and sufficient.

Assumption 1: The parameter estimates $\hat{\beta}_{i,t}, i = 1, 2, t = R, \ldots, T$, satisfy $\hat{\beta}_{i,t} - \beta_i = B_i(t)H_i(t) = \left(R^{-1}\sum_{j=t-R+1}^{t}\Lambda_{i,j}\right)^{-1}\left(R^{-1}\sum_{j=t-R+1}^{t}h_{i,j}\right)$.

Assumption 2: Let $U_t = \left[\mathbb{E}k_t(\theta), x'_{2,t} - \mathbb{E}x'_{2,t}, h'_{2,t}, \mathrm{vec}\left(h_{2,t}h'_{2,t} - \mathbb{E}h_{2,t}h'_{2,t}\right),\right]$. (a) $\mathbb{E}U_t = 0$. (b) Denote $\tilde{U}_t$ to be the vector of U_t. $R^{-1}\mathrm{E}\left(\sum_{j=t-R+1}^{t}\tilde{U}_j\right)\left(\sum_{j=t-R+1}^{t}\tilde{U}_j\right)'$ $= \Omega < \infty$ is p.d.

Assumption 3: (a) $\mathbb{E}h_{2,t}h'_{2,t} = Z^2B_2^{-1}$, (b) $\mathbb{E}\left(h_{2,t} \mid h_{2,t-j}, j = 1, 2, \ldots\right) = 0$.

Assumptions 2 and 3 allow the application of an invariance principle and are sufficient for joint weak convergence of partial sums and averages of these partial sums to Brownian motion and integrals of Brownian motion.

Assumption 4: $\lim_{P,R\to\infty} P/R = \pi = \infty$.

In order to get accurate out-of-sample forecasts, it is essential to select an optimal in-sample window. However, no consensus opinion has been reached as it depends on the characteristics of the model. Inoue et al. (2017) find the optimal window size by minimizing the conditional mean squared forecast error in a time-varying predictive regression

model. They show that the optimal window size satisfies $R = O(T^{2/3})$. Hong et al. (2020) consider minimizing various out-of-sample forecast errors, including the unconditional, conditional, and global mean squared forecast errors, respectively. The optimal window size they found is $R = O(T^{4/5})$. The out-of-sample window P has a faster divergent rate than the in-sample window R. Thus, in our model, we set $P/R \to +\infty$.

Theorem 4: Let Assumptions 1-4 hold. $\mathrm{ENC}_{\mathrm{i,R,P}} \xrightarrow{d} N(0,1)$, as $P, R \to \infty$, under $\mathbb{H}_0$: $\lambda = 0$.

West (1996) proves that ENC statistic is asymptotically normal under $\pi \geq 0$ for mean regression when two models are non-nested. Clark and McCracken (2001) show that $\mathrm{ENC}_{\mathrm{i,R,P}}$ statistic is asymptotically standard normal under $\pi = 0$, and $\mathrm{ENC}_{\mathrm{i,R,P}}$ statistic is not standard normal under $\pi > 0$ when two mean models are nested models. In this theorem, we show that $\mathrm{ENC}_{\mathrm{i,R,P}}$ statistic is asymptotically standard normality under π tends to ∞. See Appendix C for proof of Theorem 4.

4.5 Predictor Selection in High-dimensional Predictive Models

The potential for the number of predictive variables affecting the outcome of interest often amplifies as the quantity of utilized variables in the test escalates. As such, the high-dimensional regression model has found widespread application across various disciplines. This section introduces a novel approach, christened as ENC-OCMT, which leverages

the Out-of-sample Comparison of Model Tests (OCMT) procedure under the overarching encompassing principle. This methodology broadens the scope of the OCMT procedure, from full-sample variable selection to out-of-sample predictive selection, while extending the Encompassing (ENC) statistic from low-dimensional models (comprising one predictor) to high-dimensional models (featuring numerous predictors). Therefore, ENC-OCMT is designed for out-of-sample predictive ability testing, facilitating the selection of predictors in high-dimensional predictive models.

4.5.1 OCMT Procedure for Out-of-sample Forecast Encompassing Test for GaR and GS

Chudik et al. (2018) introduced a fresh strategy for variable selection, named OCMT (Out-of-sample Comparison of Model Tests), primarily suited for in-sample predictor selection. In this section, we put forth a novel methodology, specifically designed for out-of-sample predictive ability tests to assist in predictor selection. Drawing from the encompassing principle and the OCMT procedure, we term this new approach ENC-OCMT. This procedure evaluates both the marginal effect and the cumulative impact of variables on the outcome variable, mirroring the functionality of the OCMT procedure. We apply ENC-OCMT in the context of the high-dimensional linear predictive model,

$$y_{t+1} = \sigma_{i,t+1}\eta_{i,t+1} \tag{4.20}$$

$$\sigma^2_{i,t+1} = \omega + \beta\sigma^2_{i,t} + \gamma y_t^2 + \sum_{i=1}^{n} \pi_i x^2_{i,t} \tag{4.21}$$

$$\left(q^{\alpha}_{t+1}, e^{\alpha}_{t+1}\right) = (a_\alpha, b_\alpha)\, \sigma_{t+1} \tag{4.22}$$

In the given model, $x_{1,t}, \ldots, x_{n,t}$ denote the n predictors for $i = 1, \ldots, n$. In scenarios with a small number of predictors, it is possible to select predictors based directly on the marginal effect coefficient, β_i. When this marginal effect is non-zero, the specific predictor $x_{i,t}$ can be selected. However, for time series data, correlations among predictors may exist, and the likelihood of such interdependencies rises with an increase in the number of predictors in the model. Thus, when dealing with a large n, it is critical to consider the indirect impact among predictors.

The ENC-OCMT methodology strives to select the predictors $x_{i,t}$ that have either a marginal effect (β_i) or a net impact (θ_i) on y_{t+1}. This sets it apart from other multiple testing procedures. The net impact coefficient θ_i, as described by Pesaran and Smith (2014), accounts for the indirect impact of predictors which, despite having zero marginal effect on the outcome variable, maintain a correlation with other predictors.

In accordance with the encompassing principle, for low-dimensional models, we amalgamate two models in equations (6) and (7), assigning weights of $(1 - \lambda_i)$ to Model 0 and λ_i to each Model i. While Model 0 lacks predictors, each Model i has one, where $i = 1, \ldots, n$. By combining Model 0 with one Model i at a time, we establish the combined model for each Model i, resulting in n combined models, each corresponding to a different Model i.

For high-dimensional models, these $n + 1$ models are combined, with each Model i receiving a weight of λ_i and Model 0 receiving a weight of $(1 - \sum_{i=1}^{n} \lambda_i)$. This results in a single composite model, represented as follows:

$$\hat{\sigma}_{c,t+1} = \left(1 - \sum_{i=1}^{n} \lambda_i\right) \hat{\sigma}_{0,t+1} + \sum_{i=1}^{n} \lambda_i \hat{\sigma}_{i,t+1}, \tag{4.23}$$

where $\hat{sigma}_{i,t+1}$ is the conditional standard deviation for Model i that is estimated using the out-of-sample observations. Since $\left(q_{t+1}^{\alpha}, e_{t+1}^{\alpha}\right) = (a_{\alpha}, b_{\alpha})$, we can combine GaR and GS using the same weights λ_i as follows:

$$\begin{aligned} \hat{q}_{c,t+1} &= \left(1 - \sum_{i=1}^{n} \lambda_i\right) \hat{q}_{0,t+1} + \sum_{i=1}^{n} \lambda_i \hat{q}_{i,t+1} \\ \hat{e}_{c,t+1} &= \left(1 - \sum_{i=1}^{n} \lambda_i\right) \hat{e}_{0,t+1} + \sum_{i=1}^{n} \lambda_i \hat{e}_{i,t+1} \end{aligned} \tag{4.24}$$

Rearrange the equation above, the regression equation is

$$\hat{\sigma}_{0,t+1} = \sum_{i=1}^{n} \lambda_i \left(\hat{\sigma}_{0,t+1} - \hat{q}_{i,t+1}\right) + \hat{\sigma}_{c,t+1}, . \tag{4.25}$$

Given that we utilize identical information to estimate diverse models, the residuals across different models are significantly correlated. Consequently, we need to consider both the marginal effect $\lambda_i \neq 0$ and net impact $\theta_i \neq 0$ while selecting predictors. The net impact θ_i, applicable to the out-of-sample forecast encompassing approach, is predicated on the regression (4.25) and can be characterized as follows:

$$\theta_i = \sum_{j=1}^{n} \lambda_j \sigma_{i,j}, \tag{4.26}$$

where $\sigma_{i,j} = E\left(\hat{\boldsymbol{d}}_i' \hat{\boldsymbol{d}}_j\right)$, $\hat{\boldsymbol{d}}_i$ denotes the $P \times 1$ vector of $(\hat{\sigma}_{0,t+1} - \hat{\sigma}_{i,t+1})$ for $t = R, \ldots, T$.

Due to whether the predictors $x_{i,t}$ have marginal effect and net impact on y_{t+1}, there are following four possible cases:

	$\theta_i \neq 0$	$\theta_i = 0$
$\lambda_i \neq 0$	(1). Signals with nonzero net effect	(2). Hidden signals
$\lambda_i = 0$	(3). Pseudo-signals	(4). Noise variables

In the first case, where $\lambda_i \neq 0$ and $\theta_i \neq 0$, predictors exert both a marginal and net impact. On the contrary, in the last case where $\lambda_i = 0$ and $\theta_i = 0$, predictors are termed as noise

variables since they don't have either marginal effect or net impact. The third case $\lambda_i = 0$ and $\theta_i \neq 0$ is a plausible scenario that cannot be ignored, where predictors are recognized as pseudo-signals. Evidently, the second case $\lambda_i \neq 0$ and $\theta_i = 0$ is highly unlikely when the number of predictors is large. The ENC-OCMT approach targets predictors in the first, second, and third cases, specifically excluding noise variables.

Like OCMT, the ENC-OCMT procedure is also a multi-stage operation, albeit with one key difference: while OCMT examines the covariance between $x_{i,t}$ and y_{t+1}, ENC-OCMT tests the covariance between the estimated standard deviation $\hat{\sigma}_{0,t+1}$ in Model 0 and the difference in estimated standard deviation $(\hat{\sigma}_{0,t+1} - \hat{\sigma}_{i,t+1})$ between Model 0 and Model i. If a covariance exists between $\hat{\sigma}_{0,t+1}$ in Model 0 and the difference $(\hat{\sigma}_{0,t+1} - \hat{\sigma}_{i,t+1})$ in Model i, we choose the predictor $x_{i,t}$. In the initial stage, we contemplate the n bivariate regressions of $\hat{\sigma}_{0,t+1}$ on $(\hat{\sigma}_{0,t+1} - \hat{\sigma}_{i,t+1})$, for $i = 1, \ldots, n$. This process can be detailed as follows:

$$\hat{\sigma}_{0,t+1} = \phi_{i,(1)} \left(\hat{\sigma}_{0,t+1} - \hat{\sigma}_{i,t+1}\right) + \hat{\sigma}_{c,i,t+1}, \text{ for } i = 1, \ldots, n, \tag{4.27}$$

where $\hat{\sigma}_{c,i,t+1}$ is an error term of the regression for Model i. We know that $\left(q^{\alpha}_{t+1}, e^{\alpha}_{t+1}\right) = (a_\alpha, b_\alpha)\, \sigma_{t+1}$. Thus, $\phi_{i,(1)}$ can be estimated by regressing $\hat{\sigma}_{0,t+1}$ on $\hat{\sigma}_{0,t+1} - \hat{\sigma}_{i,t+1}$ by minimizing the FZ loss function as follows:

$$S\left(y_{t+1}, q_{i,t+1}, e_{i,t+1}\right) = -\frac{1}{\alpha \hat{e}_{c,t+1}} 1 y_{t+1} \leq \hat{q}_{t+1} \left(y_{t+1} - \hat{q}_{t+1}\right) + \frac{\hat{q}_{t+1}}{\hat{e}_{c,t+1}} + \log\left(-\hat{e}_{c,t+1}\right) - 1 \tag{4.28}$$

In the first stage of ENC-OCMT, we regress $\hat{\sigma}_{0,t+1}$ on $\hat{\sigma}_{0,t+1} - \hat{\sigma}_{i,t+1}$, for $i = 1, 2, \ldots, n$, one predictor at a time. Denote $t_{\hat{\phi}_{i,(1)}}$ to be the t-ratio of $\hat{\phi}_{i,(1)}$, we have

$$t_{\hat{\phi}_{i,(1)}} = \frac{\hat{\phi}_{i,(1)}}{\text{s.e.}\ \left(\hat{\phi}_i\right)} \tag{4.29}$$

The "(1)" subscript denotes the first stage. It's crucial to note that our ENC statistic can be utilized as the t statistic of the regression since both t statistic $t_{\hat{\phi}i}$ and our $\text{ENC}_{i,R,P}$ are measured testing the covariance between $x_{i,t}$ and y_{t+1}. We have also demonstrated that under the null hypothesis with no Granger Causality, the ENC statistic is asymptotically normal. Both the ENC and t statistics, when compared to the critical value from a standard normal distribution, can be used for predictor selection. As such, the ENC statistic can be employed to substitute the t statistic in equation (4.29).

This is a two-sided test with alternative hypothesis $\phi_i \neq 0$, so the predictor can be selected when the absolute value of the statistic is greater than the critical value. The first-stage ENC-OCMT selection indicator is given by

$$\widehat{\mathcal{J}}_{i,(1)} = I\left[\left|\text{ENC}_{i,R,P,(1)}\right| > c_p(n,\delta)\right] \quad \text{for } i = 1,\ldots,n \tag{4.30}$$

where $\text{ENC}_{i,R,P,(1)}$ is the ENC statistic in the first stage and $c_p(n,\delta)$ is a critical value function defined by

$$c_p(n,\delta) = \Phi^{-1}\left(1 - \frac{p}{2f(n,\delta)}\right). \tag{4.31}$$

$\Phi^{-1}(\cdot)$ signifies the inverse of the standard normal distribution function, while $f(n,\delta) = cn^{\delta}$ is defined for some positive constants δ and c. p $(0 < p < 1)$ is the nominal size of the individual tests to be determined by the researcher. Following Chudik et al. (2018), δ is coined as the critical value exponent. The first stage uses δ, and subsequent stages of OCMT employ δ, where δ must exceed δ. Predictors are selected in the first stage when $\widehat{\mathcal{J}}_{i,(1)} = 1$. In line with Chudik et al. (2018), let's define $\hat{k}_{(1)}$ as the number of predictors selected in the first stage, $\hat{\mathcal{S}}_{(1)}$ as the index set of the chosen predictors in the first stage, and $\mathcal{A}_{(1)} = 1,\ldots,n$ as the active set. Here, the subscript "(1)" signifies the first stage.

In subsequent stage $j = 2, 3, \ldots$, we consider the $n - k_{(j-1)}$ regressions of $\hat{\sigma}_{0,\alpha.t+1}$ on the predictors in $\boldsymbol{D}_{(j-1)}$ and, one at a time, d_{it} for i belonging in the active set, $\mathcal{A}_{(j)}$,

$$\hat{\sigma}_{0,t+1} = \alpha_i + \phi_{i,(j)} \left(\hat{\sigma}_{0,t+1} - \hat{\sigma}_{i,t+1}\right) + \sum_{m=1}^{k_{(j-1)}} \gamma_m \left(\hat{\sigma}_{0,t+1} - \hat{\sigma}_{m,t+1}\right) + u_{(j),t+1},$$

$$\text{for } i \in \mathcal{A}_{(j)}, \quad j = 2, 3, \ldots.$$

In this framework, $\hat{k}_{(j-1)}$ represents the number of predictors selected in stage $(j-1)$, while $k_{(j-1)}$ signifies the total number of predictors chosen up to, and inclusive of, stage $(j-1)$. Therefore, $k_{(j-1)} = \sum_{q=1}^{(j-1)} \hat{k}_{(j-1)}$. The term $\boldsymbol{D}_{(j-1)}$ designates a $k_{(j-1)} \times 1$ vector consisting of the differences $(u_{0,t+1} - u_{m,t+1})$ for $t = R, \ldots, T$. Meanwhile, $\hat{q\sigma} m, t+1$ refers to the estimated GaR for the predictors selected in previous stages, applicable for $m = 1, \ldots, k(j-1)$ and $t = R, \ldots, T$. The term $\text{ENC}_{i,R,P,(j)}$ is computed as the t-statistic in this context, with the subscript (j) indicating the stage j.

In the proposed methodology, predictors are identified and incorporated into the selection set when $\widehat{\mathcal{J}}_{i,(j)} = 1$, where $\widehat{\mathcal{J}}_{i,(j)} = I\left[\left|\text{ENC}_{i,R,P,(j)}\right| > c_p\left(n, \delta^*\right)\right]$. The set of predictors chosen in stage j is represented by $\hat{\mathcal{S}}_{(j)}$, while $\mathcal{S}_{(j)}$ signifies the index set of all predictors chosen up to and including stage j. Thus, $\mathcal{S}_{(j)} = \mathcal{S}_{(j-1)} \cup \hat{\mathcal{S}}_{(j)}$. In addition, the active set for stage $j+1$ is denoted as $\mathcal{A}_{(j+1)} = 1, \ldots, n \backslash \mathcal{S}(j)$. The procedure concludes when no further predictor can be selected in the current stage.

4.5.2 LASSO Procedure for Out-of-sample Forecast Encompassing Test for GaR and GS

In this subsection, we introduce a novel methodology for performing out-of-sample forecast encompassing tests in high-dimensional predictive regression models. This tech-

nique leverages the Least Absolute Shrinkage and Selection Operator (LASSO) under the encompassing principle to test predictive ability and is thus referred to as ENC-LASSO. The ENC-LASSO procedure serves to both shrink certain coefficients and eliminate others by setting them to zero. Consequently, predictors with non-zero coefficients are selected using the ENC-LASSO methodology.

We use the same combination in the ENC-OCMT and ENC-LASSO. We combine all $n+1$ models, so we have the equation (4.25)

$$\hat{\sigma}_{0,t+1} = \sum_{i=1}^{n} \lambda_i \left(\hat{\sigma}_{0,t+1} - \hat{\sigma}_{i,t+1} \right) + \hat{\sigma}_{c,t+1}.$$

Based on the LASSO method, the ENC-LASSO estimate $\hat{\boldsymbol{\lambda}}_{\text{LASSO}}$ is defined as

$$\begin{aligned} \hat{\boldsymbol{\lambda}}_{\text{LASSO}} = \arg\min_{\boldsymbol{\lambda}} \sum_{t=R}^{T} \left(-\frac{1}{\alpha \hat{e}_{c,t+1}} 1 y_{t+1} \leq \hat{q}_{t+1} \left(y_{t+1} - \hat{q}_{t+1} \right) + \frac{\hat{q}_{t+1}}{\hat{e}_{c,t+1}} + \log\left(-\hat{e}_{c,t+1} \right) - 1 \right) \\ + \kappa \sum_{i=1}^{n} \left| \lambda_i \right|, \end{aligned} \tag{4.32}$$

where $\hat{\boldsymbol{\lambda}}_{\text{LASSO}} = (\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_n)'$ is a vector of estimates λ_i for Model i, κ is a non-negative tuning parameter and the second term is the penalty term.

However, as the LASSO method is known to generate bias via shrinking when the number of variables is large (Fan and Li, 2001), Zou (2006) introduced a variant known as Adaptive LASSO. This procedure applies distinct weights to different coefficients in the penalty term, thus moderating the bias associated with the traditional LASSO. Zou demonstrated that Adaptive LASSO with data-dependent weights maintains consistency. Accordingly, this paper also applies the Adaptive LASSO under the encompassing principle to test predictors' predictive ability. This methodology is denoted as ENC-ALASSO. Analogous to the Adaptive LASSO, we incorporate weights $\hat{w}_i$ into the penalty term in equation

(**??**). The ENC-ALASSO estimate $\hat{\boldsymbol{\lambda}}$. ALASSO is defined as follows:

$$\hat{\boldsymbol{\lambda}}_{\text{ALASSO}} = \arg\min_{\boldsymbol{\lambda}} \sum_{t=R}^{T} \left(-\frac{1}{\alpha \hat{e}_{c,t+1}} 1 y_{t+1} \leq \hat{q}_{t+1} \left(y_{t+1} - \hat{q}_{t+1} \right) + \frac{\hat{q}_{t+1}}{\hat{e}_{c,t+1}} + \log\left(-\hat{e}_{c,t+1} \right) - 1 \right) + \kappa \sum_{i=1}^{n} \hat{w}_i \left| \lambda_i \right|,$$

where $\hat{\boldsymbol{\lambda}}_{\text{ALASSO}} = (\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_n)'$ is a vector of estimates λ_i for Model i, $\hat{w}_i = 1/|\hat{\lambda}_{i,\text{check}}|$, and $\hat{\lambda}_{i,\text{check}}$ are estimated minimizing check loss function in equation (4.25).

4.6 Monte Carlo Simulations

4.6.1 Simulation Design

In order to check the asymptotic properties of ENC-OCMT and ENC-LASSO for GaR and GS, we have one DGP. This DGP includes signals and noise variables. There are nether hidden signals nor pseudo-signals. In this DGP, y_{t+1} is generated as

$$y_{t+1} = \sigma_{i,t+1} \eta_{i,t+1}$$

$$\sigma^2_{i,t+1} = \omega + \beta \sigma^2_{i,t} + \gamma y^2_t + \pi_1 x^2_{1,t} + \pi_2 x^2_{2,t} + \pi_3 x^2_{3,t} + \pi_4 x^2_{4,t}$$

where $\eta_t \sim iid\ N(0,1)$, $t = 1, 2, \ldots, T$. We set $\pi_1 = \pi_2 = \pi_3 = \pi_4 = 1$ in this DGP. In this setting, the predictors $x_{i,t} = (\varepsilon_{i,t} + \nu g_t) / \sqrt{1+\nu}$, for $i = 1, 2, 3, 4$, are signal variables and the predictors $x_{5,t} = \varepsilon_{5,t}, x_{i,t} = (\varepsilon_{i-1,t} + \varepsilon_{i,t}) / \sqrt{2}$, for $i > 5$, are noise variables, where g_t and ε_{it} are independent draws from N$(0,1)$, for $t = 1, 2, \ldots, T$, and $i = 1, 2, \ldots, n$. We set $\nu = 1$, which implies 50% pairwise correlation among the signal variables. We minimize the FZ loss function to estimate the conditional GaR and GS jointly. We estimate GaR and

GS for Model 0 and Model i to be defined as

$$q_{0,t+1} = a_\alpha \sigma_{0,t+1}, \quad e_{0,t+1} = b_\alpha \sigma_{0,t+1}$$

$$q_{i,t+1} = a_\alpha \sigma_{i,t+1}, \quad e_{i,t+1} = b_\alpha \sigma_{i,t+1}$$

where $a_\alpha = \Phi^{-1}(\alpha)$ and $b_\alpha = -\phi\left(\Phi^{-1}(\alpha)\right)/\alpha$.

We focus on three out-of-sample forecast selection methods for the simulation results: ENC-OCMT, ENC-LASSO, and ENC-ALASSO. In our forecast evaluations, we use the most used criteria in the penalized regression methods, which are the following: the true positive rate (TPR) defined by (4.6.1), the false positive rate (FPR) defined by (4.6.1), the false discovery of the approximating model (FDR) defined by (4.6.1) and the false discovery rate of the true model (FDR*) denoted by (4.6.1).

True positive rate:

$$\text{TPR}_{n,T} = \frac{\sum_{i=1}^{n} I\left(\widehat{\mathcal{J}}_i = 1 \text{ and } \beta_i \neq 0\right)}{\sum_{i=1}^{n} I\left(\beta_i \neq 0\right)} \xrightarrow{p} 1.$$

False positive rate:

$$\text{FPR}_{n,T} = \frac{\sum_{i=1}^{n} I\left(\widehat{\mathcal{J}}_i = 1 \text{ and } \beta_i = 0\right)}{\sum_{i=1}^{n} I\left(\beta_i = 0\right)} \xrightarrow{p} 0.$$

False discovery rate:

$$\text{FDR}_{n,T} = \frac{\sum_{i=1}^{n} I\left(\widehat{\mathcal{J}}_i = 1, \text{ and } \beta_i = \theta_i = 0\right)}{\sum_{i=1}^{n} \widehat{\mathcal{J}}_i + 1} \xrightarrow{p} 0.$$

The false discovery rate of the true model:

$$\text{FDR}^*_{n,T} = \frac{\sum_{i=1}^{n} I\left(\widehat{\mathcal{J}}_i = 1 \text{ and } \beta_i = 0\right)}{\sum_{i=1}^{n} \widehat{\mathcal{J}}_i + 1} \xrightarrow{p} c.$$

The asymptotic theory of OCMT carries over for ENC-OCMT and ENC-LASSO in terms of the true and false positive rates, and false discovery rate of the approximate and true

model (TPR, FPR, FDR, and FDR*). We report the FZ loss for three out-of-sample forecast selection methods.

We set the sample size n to be $100, 200$, and 300, the number of in-sample observations R to be 100, and the number of out-of-sample observations to be $200, 300$, and 500. For ENC-OCMT method, we set nominal level $p = 0.1, 0.05, 0.01$, $\delta = 1$ in the first stage, and $\delta^* \in \{2\}$ in the subsequent stages.

4.6.2 Simulation Results

The simulation results for GaR and GS, as indicated by Table 1, draw attention to certain parallels and distinctions among the methodologies. Once again, ENC-OCMT shines with the lowest Loss values (1.8810, 1.8760, and 1.8720 for P=200, 300, and 500 respectively), implying an optimized predictive capacity. Its TPR, like the GaR simulation, remains high and improves with an increase in P. The false discovery rates (FDR and FDR*), while higher than in the GaR case, still compare favorably against those for ENC-LASSO. ENC-LASSO, though showing high TPR values, suffers from considerably high FDR rates, signaling a higher probability of falsely identified predictors. The ENC-Adaptive LASSO, like in the GaR case, strikes a balance with comparatively controlled FDR and high TPR values, although it doesn't surpass the ENC-OCMT in overall performance. These results collectively underline the superior performance of the ENC-OCMT in both simulation scenarios.

4.7 Empirical Analysis

Our empirical examination utilizes the quarterly data from Stock and Watson (2012), focusing on macroeconomic forecasting for U.S. GDP growth and CPI inflation against a wide set of macroeconomic variables. The dataset consists of 109 series that cover the period from the third quarter of 1960 through to the fourth quarter of 2008. For the purposes of our study, U.S. GDP growth is used as the target variable. Two in-sample periods are defined: the first from the third quarter of 1960 to the second quarter of 1980 (80 observations), and the second from the third quarter of 1960 to the second quarter of 1990 (120 observations). Correspondingly, the out-of-sample periods are from the third quarter of 1980 to the fourth quarter of 2008 (114 observations) and from the third quarter of 1990 to the fourth quarter of 2008 (74 observations).

When assessing the GaR model, the ENC-OCMT outperforms other methods, yielding the smallest OOS Forecast Loss of 0.0877 (Table 3). In terms of predictor selection (Table 3), ENC-Lasso opts for "Real private domestic investment," "Purchasing managers' index," "Price Index: Others," "US effective exchange rate," and "Money stock: m2." On the other hand, ENC-OCMT selects "First lag of the dependent variable," "Residential price index," "Unemploy: by duration: average," and "Employees, nonfarm" as predictors. DGP 1 results for GaR (Table 1) further illustrate the superiority of ENC-OCMT, showing the lowest Loss across varying P values and consistently high TPR with minimal FPR and FDR.

For the GS model, the ENC-OCMT again produces the most favorable results, demonstrating the smallest OOS Forecast Loss of 0.1008 (Table 4). Regarding predic-

tor selection (Table 4), ENC-Lasso picks "Real private domestic investment," "Purchasing managers' index," "Employees, nonfarm," "US effective exchange rate," and "Money stock: m2," plus "Consumer credit outstanding." Conversely, ENC-OCMT selects "First lag of the dependent variable," "Third lag of the dependent variable," "MZM money stock (FRB St. Louis)," and "Employees, nonfarm." The DGP 2 results for GS (Table 2) further underscore the dominance of ENC-OCMT with the lowest Loss across different P values, high TPR, and minimized FPR and FDR.

4.8 Conclusion

In this paper, we propose a novel predictor selection method based on the ENC framework. The goal is to select the predictors with non-zero marginal effects or non-zero net impacts in terms of their OOS predictive ability in high dimensional predictive regression models. ENC-OCMT is based on the multiple testing approach of CKP's (2018) OCMT. Compared to CKP-OCMT which can be applied only to the conditional mean, ENC-OCMT can be applied to any elicitable functionals. In this paper, we propose the ENC-OCMT and ENC-LASSO methods for GaR and GS. The simulation results show that the asymptotic theory of CKP-OCMT carries over for ENC-OCMT in terms of TPR, FPR, and FDR, in selection of the predictors for predictive models in GaR and GS. The empirical results show that ENC-OCMT can get lower losses in terms of OOS prediction.

Table 4.1: Simulation Result for GaR and GS

	P	Loss	TPR	FPR	FDR	FDR*
ENC-OCMT	200	1.8810	0.9730	0.0340	0.0380	0.0380
$p = 0.05$	300	1.8760	0.9840	0.0320	0.0370	0.0370
$\delta = 1$, δ^*=2	500	1.8720	0.9920	0.0280	0.0330	0.0330
ENC-LASSO	200	1.9120	0.9810	0.0730	0.6120	0.6120
	300	1.8990	0.9850	0.0680	0.5840	0.5840
	500	1.8930	0.9970	0.0610	0.5370	0.5370
ENC-Adaptive LASSO	200	1.8810	0.9710	0.0520	0.1970	0.1970
	300	1.8770	0.9790	0.0500	0.1930	0.1930
	500	1.8740	0.9900	0.0480	0.1890	0.1890

$\alpha = 0.1, n = 100, R = 100$

Table 4.2: Predictor Selection for the GaR and GS Model

	LASSO	OCMT	ENC-LASSO	ENC-OCMT
OOS Forecast Loss	0.1811	0.5734	0.1054	0.1008

Predictor Selection

ENC-LASSO	ENC-OCMT
Real private domestic investment	First lag of the dependent variable
Purchasing managers' index	Third lag of the dependent variable
Employees, nonfarm	MZM money stock (FRB St. Louis)
US effective exchange rate	Employees, nonfarm
Money stock: m2	
Consumer credit outstanding	

Chapter 5

Forecast Encompassing and Granger Causality in Predictive Expectile Regression Models

5.1 Introduction

Granger Causality, whether one time series is useful in forecasting another, has been developed by Granger (1969, 1980). It aims to investigate the predictive ability of the additional variable (covariate) by comparing two nested models, the ideas and methods derive from non-nested model comparisons (Diebold and Mariano, 1995; Harvey et al., 1998) and associated statistics are typically asymptotically standard normal. For two nested models (West, 1996; Chao et al., 2001; Corradi and Swanson, 2002; McCracken, 2000; Clark and McCracken, 2001; Clark and West, 2006, 2007), the asymptotic distribution of the associated

test statistic based on loss-differential may degenerate or have a non-standard distribution (West, 1996; Clark and McCracken, 2001, 2005, 2009). The DM test, tends to under-reject the null hypothesis of no Granger Causality of the covariate as shown by Clark and McCracken (2001), Clark and West (2006, 2007), they show that the DM statistic for mean regression has negative mean so they propose using encompassing test (Harvey et al., 1998) to examine Granger Causality. Clark and McCracken (2001); Clark and West (2007) point out that the ENC statistic is asymptotically normal under some restrictions on the number of in-sample and out-of-sample observations. Corradi and Swanson (2002) and Chao et al. (2001) propose another two conditional moment tests, namely CS and CCS test, under regularity conditions, the tests are asymptotically standard normal.

Besides the conditional mean model, Granger-causality in the conditional quantile model has been studied by Lee and Yang (2012) for the out-of-sample predictive ability and by Chuang et al. (2009) for in-sample variable selection tests. Quantile model is related with Value at Risk (VaR hereinafter), which measures the maximum potential loss of a given portfolio over a certain period at a given confidence level, it has been considered as the standard measure of financial market risk. VaR is the quantile depending on the given tail probability. However, VaR has two main drawbacks. First, it can not satisfy the "coherent" condition because it lacks subadditivity (Artzner et al., 1999), which implies that the total financial risk of a portfolio should not be larger than the sum of financial risk of individuals in the portfolio due to diversity reduces risk. Second, VaR focuses on the probability rather than the magnitude of the losses and there might have different losses depending on different tail distributions with the same VaRs (Chuang et al., 2009).

Due to these two weaknesses of VaR, Basel III in 2013 proposed a new risk measure, Expected Shortfall (ES hereinafter). Increasing number of financial institutions choose ES as their main risk measure. ES is defined as the conditional expectation of the return which are less than the VaR, so ES is more sensitive to the magnitude of extreme losses. ES is also known as conditional VaR, or CVaR (Artzner et al., 1999; Rockafellar et al., 2000; Fissler et al., 2016). Artzner et al. (1999) prove that ES is "coherent". However, the measure of ES is not a "elicitable" measure (Gneiting, 2011; Fissler et al., 2016), which means that is no strictly consistent loss function for the ES risk measure. Thus, how to estimate ES becomes a main issue.

Newey and Powell (1987) introduce expectiles that are estimated by minimizing the asymmetric least squares (ALS hereinafter) loss function. Expetile is "coherent" (Artzner et al., 1999) and "elicitable". (Gneiting, 2011). Kuan et al. (2009) propose an expectile-based VaR (EVaR hereinafter) to compare with the conventional quantile-based VaR (QVaR hereinafter) and find out EVaR is more sensitive to capture the magnitude of extreme losses than QVaR. Taylor (2008) shows that there is a link between expectile and ES. Moreover, if the distribution of returns is known, and there is a simple expression of expectile as a function of ES. It is convenient to get the ES from using expectiles.

In this paper, we use the ALS loss function of Newey and Powell (1987) to construct the encompassing test to test Granger causality in expectile regressions. We show that under $\mathbb{H}_0$, the statistic of ENC test is asymptotically standard normal when the ratio number of out-of-sample observation to in-sample observations tends to infinity, hence it has zero mean and correct size, whereas the DM test is serious undersized. In model averaging, we find out

that under $\mathbb{H}_0$, the optimal weight of model averaging is in favor of model without covariate (Model 1), endowing whole weight on Model 1 and 0 weight on model with covariate (Model 2), we show the relationship between ENC test and model averaging. We also find out how the optimal weight switch from Model 1 to Model 2 when the signal-to-noise ratio (SNR hereinafter) increases in Model 2.

This paper is organized as follows. section 2 reviews the definitions and theorem of the elicitability. Section 3 reviews the encompassing test for Granger causality in mean by Clark and West (2006, 2007). Section 4 introduces the encompassing test for Granger causality in expectiles. Section 5 introduces model averaging and its properties. Section 6 conducts Monte Carlo simulation of ENC test and model averaging. Section 7 presents empirical analysis of Welch and Goyal (2008) study for the expectile regressions. Section 8 concludes. The proofs are collected in Appendix.

5.2 Elicitability

We denote an observation domain O for $y, x \in \mathrm{O} \subseteq \mathbb{R}^{d_1+d_2}$, $d_1 = \dim(y)$ and $d_2 = \dim(x)$, the conditional distribution $F \equiv F_{Y|X}$ for Y given X. Let $\mathcal{F}$ be a class of distribution function on the observation domain O, and let A be an action domain, $\gamma \in \mathrm{A}$. We define $\Gamma : \mathcal{F} \to \mathrm{A}$ to be functional. For example, $\Gamma(F(y|x))$ may be $\mathbb{E}(Y|X)$, $Q(Y|X)$, $V(Y|X)$, Mode$(Y|X)$ or ES$(Y|X)$, where $\mathbb{E}(Y|X)$ is the conditional mean, $Q(Y|X)$ is the conditional quantile, $V(Y|X)$ is the conditional variance, Mode$(Y|X)$ is the conditional mode and ES$(Y|X)$ is the conditional Expected Shortfall. Note that Γ can be a vector of several of these.

Definition 1: (Gneiting, 2011; Fissler et al., 2016) A scoring function is an $\mathcal{F}$-integrable function $S : A \times O \to \mathbb{R}$. S is said to be *$\mathcal{F}$-consistent* for a functional $\Gamma : \mathcal{F} \to \mathrm{A}$ if $\mathbb{E}_F S(\Gamma(F), Y) \leq \mathbb{E}_F S(\gamma, Y)$ for all $F \in \mathcal{F}$ and for all $\gamma \in \mathrm{A}$. Furthermore, S is *strictly $\mathcal{F}$-consistent* for Γ if it is $\mathcal{F}$-consistent for Γ and if $\mathbb{E}_F S(\Gamma(F), Y) = \mathbb{E}_F S(\gamma, Y)$ implies that $\gamma = \Gamma(F)$ for all $F \in \mathcal{F}$ and for all $\gamma \in \mathrm{A}$.

Definition 2: (Gneiting, 2011; Fissler et al., 2016) A functional $\Gamma : \mathcal{F} \to \mathrm{A} \subseteq \mathbb{R}^k$ is called *k-elicitable*, if there exists a strictly $\mathcal{F}$-consistent scoring function for Γ.

Definition 3: (Gneiting, 2011; Fissler et al., 2016) An identification function is an $\mathcal{F}$-integrable function $V : A \times O \to \mathbb{R}^k$. V is said to be an *$\mathcal{F}$-identification function* for a functional $\Gamma : \mathcal{F} \to \mathrm{A} \subseteq \mathbb{R}^k$ if $\mathbb{E}_F V(\Gamma(F), Y) = 0$ for all $F \in \mathcal{F}$. Furthermore, V is a *strict $\mathcal{F}$-identification function* for Γ if $\mathbb{E}_F V(\gamma, Y) = 0$ holds if and only if $\gamma = \Gamma(F)$ for all $F \in \mathcal{F}$ and for all $\gamma \in \mathrm{A}$.

Osband (1985) point out that there is a relation between strictly consistent scoring function S and strict identification function V. There is a nonnegative function $h : \mathbb{R} \to \mathbb{R}$ such that

$$\frac{\mathrm{d}}{\mathrm{d}\gamma} S(\gamma, y) = h(\gamma) V(\gamma, y). \tag{5.1}$$

The strict identification function can be obtained by the partial derivative of the strictly consistent scoring function. A statistical functional is elicitable if there exists a

scoring function that the correct forecast of the functional is the unique minimizer of the expected score. We can compare or rank the forecasts of the elicitable functional with their realized scores (Fissler et al., 2016). Many statistical functionals are 1-elicitable such as expectation, ratios of expectations, quantiles (Value-at-Risk), and expectiles. However, some functional are not 1-elicitable such as variance, mode, or Expected Shortfall (Gneiting, 2011).

Proposition 1: (Savage, 1971) The class of the strictly consistent scoring functions of mean or expectation functional are Bregman divergence. The Bregman divergence is

$$S(\gamma, y) = f(y) - f(\gamma) - f'(\gamma)(y - \gamma), \tag{5.2}$$

where f is a strictly convex and differential function.

Then, taking the derivative, we can get $\frac{\mathrm{d}}{\mathrm{d}\mu} S(\gamma, y) = f''(\gamma)(\gamma - y)$, where $h(\gamma) = f''(\gamma)$ and the strictly consistent identification function $V(\gamma, y) = \gamma - y$ in equation (7.1). The mean functional $\mathbb{E}_F(y) = \gamma$ when $\mathbb{E}_F V = 0$. Therefore, mean or expectations are elicitable. If the function $f(y) = y^2$, we have the squared error scoring function

$$S(\gamma, y) = (y - \gamma)^2, \tag{5.3}$$

which is strictly consistent with the mean functional.

For the expectile, the asymmetric squares scoring function is

$$\mathrm{S}_\tau(\gamma_\tau, y) = \begin{cases} \tau\,(y - \gamma_\tau)^2 & \text{if } \gamma_\tau \leq y, \\ (1 - \tau)\,(y - \gamma_\tau)^2 & \text{if } \gamma_\tau \geq y. \end{cases} \tag{5.4}$$

By considering the characteristics of the Bregman divergence, Gneiting (2011) states that the scoring function is consistent for the $\tau-$expectile if and only if the form of the scoring function is

$$S_\tau(\gamma_\tau, y) = \text{abs}\left(\tau - 1\{\gamma_\tau \geq y\}\right)\left[f(y) - f(\gamma_\tau) - f'\left(\gamma_\tau\right)\left(y - \gamma_\tau\right)\right], \tag{5.5}$$

where f is convex and differential function. Taking the derivative of $S_\tau(\gamma_\tau, Y)$, we can get $\frac{\mathrm{d}}{\mathrm{d}\gamma_\tau} S_\tau\left(\gamma_\tau, Y\right) = h\left(\gamma_\tau\right) V\left(\gamma_\tau, Y\right) = \text{abs}\left(1\{\gamma_\tau \geq y\} - \tau\right) f''(\gamma_\tau)(\gamma_\tau - y)$, where $h(\gamma_\tau) = f''(\gamma_\tau)$ and the strictly consistent identification function $V(\gamma_\tau, Y) = \text{abs}\left(1\{\gamma_\tau \geq y\} - \tau\right)\left(\gamma_\tau - y\right)$ in equation (7.1). When the function $f(y) = y^2$, we can get the well-known scoring function for expectile functional from Newey and Powell (1987)

$$S_\tau(\gamma_\tau, y) = \text{abs}\left(\tau - 1\{\gamma_\tau \geq y\}\right)\left(y - \gamma_\tau\right)^2, \tag{5.6}$$

which is strictly consistent. Therefore, expectiles are 1-elicitable.

When $\tau = 0.5$, the strictly consistent scoring function for expectile in equation (5.16) becomes $S_{0.5}(\gamma, Y) = (y - \gamma)^2$, which is the same strictly consistent scoring function for mean in equation (5.3).

5.3 Granger Causality in Mean

This section reviews testing the Granger Causlity in conditional mean because the mean regression is the special case for the expectile regression. We have two nested models. One model is without covariate X, denoted as Model 1, and the other model is

with conditioning on X, denoted as Model 2. The two nested models are

$$\text{Model 1} \quad : \quad y_{t+1} = c_1 + u_{t+1}^{(1)} \equiv x_{1,t}'\beta_1 + u_{t+1}^{(1)}, \tag{5.7}$$

$$\text{Model 2} \quad : \quad y_{t+1} = c_2 + bx_t + u_{t+1}^{(2)} \equiv x_{2,t}'\beta_2 + u_{t+1}^{(2)}, \tag{5.8}$$

where the conditional mean of y for model i are $\mu_t^{(i)} = x_{i,t}'\beta_i$, $u_{t+1}^{(i)} = y_{t+1} - \mu_t^{(i)}$, $i = 1, 2$. The dependent variable y_{t+1} is a scalar random variable. The independent variable x_t is a stationary variable. $x_{1,t}$ is a strict subset of $x_{2,t}$. In the mean models, $x_{1,t}' = 1, \beta_1 = c_1, x_{2,t}' = (1, x_t), \beta_2 = (c_2, b)$. The forecast errors $\hat{u}_{t+1}^{(1)}$ and $\hat{u}_{t+1}^{(2)}$ are estimated by $\hat{u}_{t+1}^{(1)} = y_{t+1} - \hat{c}_{1,t}$ and $\hat{u}_{t+1}^{(2)} = y_{t+1} - \hat{c}_{2,t} - \hat{b}_t x_t$. In the mean regression, we use the squared error loss function: $S(\mu, y) = \frac{1}{2}(y - \mu)^2 = \frac{1}{2}u^2$, with $u = y - \mu$. For Model 2, the derivative of the squared error loss function with respect to the mean μ is

$$\mathbb{E}\left(u^{(2)} \middle| X\right) = 0, \tag{5.9}$$

so μ is the martingale difference series.

In order to test the Granger Causality of the covariate, the null and alternative hypotheses are

$$\mathbb{H}_0 : \mathbb{E}\left[S(\mu^{(1)}, Y) - S(\mu^{(2)}, Y)\right] = 0, \tag{5.10}$$

$$\mathbb{H}_1 : \mathbb{E}\left[S(\mu^{(1)}, Y) - S(\mu^{(2)}, Y)\right] > 0. \tag{5.11}$$

Under $\mathbb{H}_0$, X does not Granger-cause Y in mean. Under $\mathbb{H}_1$, X Granger-causes Y in mean. According to Clark and McCracken (2001) and Clark and West (2006), the test is one-sided because if X has predictive power, the forecast from Model 2 is superior to that from Model 1. The DM test statistic (Diebold and Mariano, 1995) based on the mean squared predictive

error (MSPE) loss differential is

$$D = \mathbb{E}\left[\left(u^{(1)}\right)^2 - \left(u^{(2)}\right)^2\right], \tag{5.12}$$

which can be estimated by

$$\hat{D}_P = P^{-1}\sum_{t=R}^{T}\left[\left(\hat{u}_{t+1}^{(1)}\right)^2 - \left(\hat{u}_{t+1}^{(2)}\right)^2\right] \xrightarrow{p} D \quad \text{as } R, P \to \infty, \tag{5.13}$$

where R is the number of in-sample observations, P is the out-of-sample forecasts and $T + 1 = R + P$. Clark and McCracken (2001) and Clark and West (2007) rewrite the sample MSPE loss differential as

$$\begin{aligned}\hat{D}_P &= P^{-1}\sum_{t=R}^{T}\frac{1}{2}\left[\left(\hat{u}_{t+1}^{(1)}\right)^2 - \left(\hat{u}_{t+1}^{(2)}\right)^2\right] \\ &= P^{-1}\sum_{t=R}^{T}\hat{u}_{t+1}^{(1)}\left(\hat{u}_{t+1}^{(1)} - \hat{u}_{t+1}^{(2)}\right) - \frac{1}{2}P^{-1}\sum_{t=R}^{T}\left(\hat{u}_{t+1}^{(1)} - \hat{u}_{t+1}^{(2)}\right)^2.\end{aligned}$$

We denote $\hat{C}_P \equiv P^{-1}\sum_{t=R}^{T}\hat{u}_{t+1}^{(1)}\left(\hat{u}_{t+1}^{(1)} - \hat{u}_{t+1}^{(2)}\right)$ and $\hat{V}_P \equiv P^{-1}\sum_{t=R}^{T}\left(\hat{u}_{t+1}^{(1)} - \hat{u}_{t+1}^{(2)}\right)^2$. Clark and McCracken (2001) show that

$$\hat{C}_P \equiv P^{-1}\sum_{t=R}^{T}\hat{u}_{t+1}^{(1)}\left(\hat{u}_{t+1}^{(1)} - \hat{u}_{t+1}^{(2)}\right) \xrightarrow{p} 0, \tag{5.14}$$

under $\mathbb{H}_0$ as $R, P \to \infty$. $\hat{V}_P$ is non-negative. Thus, $\hat{D}_P$ is expected to be negative under $\mathbb{H}_0$, so we have $\mathbb{E}\hat{D}_P \leq 0$.

Remark: Clark and McCracken (2001) use Monte Carlo simulation to show that $\hat{D}_P$ is downward biased and under-sized under $\mathbb{H}_0$. They also show that $\hat{C}_P$ has zero expectation and correct size under $\mathbb{H}_0$.

5.4 Granger Causality in Expectiles

This section introduces the Granger Causality in expectile. We define τ-expectile loss function as

$$\mathrm{S}_\tau(\mu_\tau, y) = \begin{cases} \tau\,(y-\mu_\tau)^2 & \text{if } \mu_\tau \le y, \\ (1-\tau)\,(y-\mu_\tau)^2 & \text{if } \mu_\tau \ge y. \end{cases} \tag{5.15}$$

where μ_τ is the τ-expectile of y, $\tau \in (0,1)$. Equation (5.15) can be written as

$$\mathrm{S}_\tau(\mu_\tau, y) = \mathrm{abs}\,(\tau - 1(\mu_\tau \ge y)) \cdot (y-\mu_\tau)^2, \tag{5.16}$$

where abs represents the absolute value. Equation (5.16) is the ALS loss function from Newey and Powell (1987). Gneiting (2011) proves that the ALS loss function is a strictly consistent scoring function for expectile, and so expectile is elicitable. The mean regression is the special case of the expectile regression. When $\tau = 0.5$, Equation (5.16) is the squared loss function in the previous section.

We develop the Granger Causality test for conditional expectile models. Consider two nested models, Model 1 is without conditioning on X and Model 2 is with conditioning on X,

$$\text{Model 1}: y_{t+1} = c_{1,\tau} + u^{(1)}_{\tau,t+1} \equiv x'_{1,t}\beta_{1,\tau} + u^{(1)}_{\tau,t+1}, \tag{5.17}$$

$$\text{Model 2}: y_{t+1} = c_{2,\tau} + b_\tau x_t + u^{(2)}_{\tau,t+1} \equiv x'_{2,t}\beta_{2,\tau} + u^{(2)}_{\tau,t+1}, \tag{5.18}$$

The expectile of Model i is $\mu^{(i)}_{\tau,t}$, where $\mu^{(1)}_{\tau,t} = x'_{1,t}\beta_{1,\tau}$ and $\mu^{(2)}_{\tau,t} = x'_{2,t}\beta_{2,\tau}$. The dependent variable y_{t+1} is a scalar random variable. The independent variable x_t is a stationary variable. $x_{1,t}$ is a strict subset of $x_{2,t}$. $x'_{1,t} = 1, \beta_{1,\tau} = a_{1,\tau}, x'_{2,t} = (1, x_t)\,, \beta_{2,\tau} = (c_{2,\tau}, b_\tau)'$. Thus, we estimate $\hat{u}^{(1)}_{t+1,\tau}$ and $\hat{u}^{(2)}_{t+1,\tau}$ by $\hat{u}^{(1)}_{t+1,\tau} = y_{t+1} - \hat{c}_{1,t,\tau}$ and $\hat{u}^{(2)}_{t+1,\tau} = y_{t+1} - \hat{c}_{2,t,\tau} - \hat{b}_{t,\tau}x_t$.

In the expectile regression, Equation (5.16) can be written as:

$$S_\tau(\mu_\tau, y) \equiv \frac{1}{2} g_\tau\left(u_\tau\right) \cdot u_\tau^2, \tag{5.19}$$

where $g_\tau(u_\tau) = \text{abs}(\tau - 1(u_\tau < 0))$ and $u_\tau = y - \mu_\tau$. Taking a derivative of the ALS loss function with respect to the expectile, we get the martingale difference series. Below the subscript τ is omitted for simplicity.

Newey and Powell (1987, p. 844) point out that $S_\tau(\mu_\tau, y)$ is differentiable and convex in u_τ, so that $S_\tau(\mu_\tau, y)$ is differentiable and convex in μ. For Model 2, the derivative of ALS loss function with respect to the expectile μ is

$$\frac{\partial \mathbb{E}\left[S(\mu^{(2)}(\tau), Y) \middle| x_t\right]}{\partial \mu(\tau)} = \mathbb{E}\left[g\left(u^{(2)}\right) u^{(2)} \middle| X\right] = 0, \tag{5.20}$$

hence $g\left(u\right) u$ is a martingale difference series.

Under $\mathbb{H}_0$, X does not Granger-cause Y in the τ-expectile. We have the null hypothesis

$$\mathbb{H}_0 : \mathbb{E}\left[S(\mu^{(1)}, Y) - S(\mu^{(2)}, Y)\right] = 0. \tag{5.21}$$

Under $\mathbb{H}_1$, X Granger-causes Y in the τ-expectile, the alternative hypothesis is

$$\mathbb{H}_1 : \mathbb{E}\left[S(\mu^{(1)}, Y) - S(\mu^{(2)}, Y)\right] > 0. \tag{5.22}$$

Under the alternative hypothesis, X Granger-causes Y in τ-expectile, hence we expect forecasts from Model 2 to be superior to those from Model 1. The DM test based on the ALS loss-differential is

$$D = \mathbb{E}\left[\frac{1}{2} g\left(u^{(1)}\right)\left(u^{(1)}\right)^2 - \frac{1}{2} g\left(u^{(2)}\right)\left(u^{(2)}\right)^2\right], \tag{5.23}$$

which can be estimated by

$$\hat{D}_P = P^{-1}\sum_{t=R}^{T}\left[\frac{1}{2}g\left(\hat{u}_{t+1}^{(1)}\right)\left(\hat{u}_{t+1}^{(1)}\right)^2 - \frac{1}{2}g\left(\hat{u}_{t+1}^{(2)}\right)\left(\hat{u}_{t+1}^{(2)}\right)^2\right] \xrightarrow{p} D \quad as \quad R, P \to \infty, \quad (5.24)$$

In section 2, Clark and McCracken (2001), Clark and West (2006, 2007) prove that the DM test has downward bias while the ENC test has zero means in mean regression under $\mathbb{H}_0$. The ENC test corrects the size and increases the power. In this section, we show that comparing the two nested expectile models, under $\mathbb{H}_0$, the ENC test has the correct size whereas the DM test is undersized.

To prove $\hat{D}_P$ is biased, we rewrite the DM test based on the ALS loss-differential as

$$\begin{aligned}\hat{D}_P &= P^{-1}\sum_{t=R}^{T}\left[\frac{1}{2}g\left(\hat{u}_{t+1}^{(1)}\right)\left(\hat{u}_{t+1}^{(1)}\right)^2 - \frac{1}{2}g\left(\hat{u}_{t+1}^{(2)}\right)\left(\hat{u}_{t+1}^{(2)}\right)^2\right] \\ &= P^{-1}\sum_{t=R}^{T} g\left(\hat{u}_{t+1}^{(1)}\right)\hat{u}_{t+1}^{(1)}\left(\hat{u}_{t+1}^{(1)} - \hat{u}_{t+1}^{(2)}\right) \\ &\quad - P^{-1}\sum_{t=R}^{T}\frac{1}{2}\left[g\left(\hat{u}_{t+1}^{(1)}\right)\left(\hat{u}_{t+1}^{(1)} - \hat{u}_{t+1}^{(2)}\right)^2 - g\left(\hat{u}_{t+1}^{(1)}\right)\left(\hat{u}_{t+1}^{(2)}\right)^2 + g\left(\hat{u}_{t+1}^{(2)}\right)\left(\hat{u}_{t+1}^{(2)}\right)^2\right],\end{aligned}$$

where we denote the first term as $\hat{C}_P$ and the second term as $\hat{V}_P$. Then we have $\hat{D}_P = \hat{C}_P - \hat{V}_P$. Now we show that $\hat{C}_P$ tends to 0 as $P\to\infty$ and that $\hat{V}_P$ is non-negative, which implies that $\hat{D}_P$ is expected to be negative under the null hypothesis.

Proposition 2: Under $\mathbb{H}_0$,

$$\hat{C}_P \equiv P^{-1}\sum_{t=R}^{T} g\left(\hat{u}_{t+1}^{(1)}\right)\hat{u}_{t+1}^{(1)}\left(\hat{u}_{t+1}^{(1)} - \hat{u}_{t+1}^{(2)}\right) \xrightarrow{p} 0 \quad (5.25)$$

as $R, P\to\infty$.

Proof: From our models, we have $u_{t+1}^{(1)} = y_{t+1} - x_{1,t}'\beta_1$ and $u_{t+1}^{(2)} = y_{t+1} - x_{2,t}'\beta_2$, so

we can obtain $(u_{t+1}^{(1)} - u_{t+1}^{(2)}) = y_{t+1} - x_{1,t}'\beta_1 - y_{t+1} + x_{2,t}'\beta_2 = x_{2,t}'\beta_2 - x_{1,t}'\beta_1$. Thus, we have $\mathbb{E}[g\left(u^{(1)}\right)u^{(1)}(u^{(1)} - u^{(2)})] = \mathbb{E}[g\left(u(1)\right)u^{(1)}(x_{2,t}'\beta_{2,\tau} - x_{1,t}'\beta_{1,\tau})]$. We know that $g\left(u_{t+1}^{(1)}\right)u_{t+1}^{(1)}$ is a martingale difference series, so $g\left(u_{t+1}^{(1)}\right)u_{t+1}^{(1)}$ is uncorrected with both $x_{1,t}$ and $x_{2,t}$. Thus, we prove that $\mathbb{E}[g\left(u(1)\right)u^{(1)}(x_{2,t}'\beta_{2,\tau} - x_{1,t}'\beta_{1,\tau})] = 0$ and we have $\hat{C}_P \xrightarrow{p} 0$ as $R, P \to \infty$.

Next, we prove that $\hat{V}_P$ is non-negative.

Proposition 3: Let

$$\hat{V}_P \equiv P^{-1}\sum_{t=R}^{T}\frac{1}{2}\left[g\left(\hat{u}_{t+1}^{(1)}\right)\left(\hat{u}_{t+1}^{(1)} - \hat{u}_{t+1}^{(2)}\right)^2 - g\left(\hat{u}_{t+1}^{(1)}\right)\left(\hat{u}_{t+1}^{(2)}\right)^2 + g\left(\hat{u}_{t+1}^{(2)}\right)\left(\hat{u}_{t+1}^{(2)}\right)^2\right]. \tag{5.26}$$

Then, $\hat{V}_P \geq 0$.

Proof: See Appendix A.

Summarizing Propositions 2 and 3, we have $\mathbb{E}\hat{D}_P < 0$ under $\mathbb{H}_0$. DM has a downward bias with two nested models under H_0.

Now, we want to propose an ENC test for Granger Causality with two nested models. We build a combined ALS scoring function $S_\tau(\mu^{(c)}, y)$ by combing Model 1 and Model 2 with the weight $1 - \lambda$ and λ respectively, so $\mu^{(c)} = (1 - \lambda)\mu^{(1)} + \lambda\mu^{(2)}$. Then, taking a derivative of the expectation of combined ALS scoring function with respect to λ, we get the first order condition $\mathbb{E}[g\left(u^{(c)}\right)u^{(c)}\left(u^{(1)} - u^{(2)}\right)] = 0$. Under $\mathbb{H}_0$, when $\lambda = 0$, $\mu^{(c)} = \mu^{(1)}$ and $u^{(c)} = u^{(1)}$, then we obtain $\mathbb{E}[g\left(u^{(1)}\right)u^{(1)}\left(u^{(1)} - u^{(2)}\right)] = 0$.

Proposition 4: Denote the ENC statistic as $\hat{C}_P$. Under $\mathbb{H}_0$,

$$C \equiv \mathbb{E}_F \left[g\left(u^{(1)}\right) u^{(1)} \left(u^{(1)} - u^{(2)}\right)\right] = 0, \tag{5.27}$$

$$\hat{C}_P \equiv P^{-1} \sum_{t=R}^{T} g\left(\hat{u}_{t+1}^{(1)}\right) \hat{u}_{t+1}^{(1)} \left(\hat{u}_{t+1}^{(1)} - \hat{u}_{t+1}^{(2)}\right) \xrightarrow{p} C = 0 \tag{5.28}$$

as $R, P \to \infty$.

Proof: See Appendix A.

Proposition 2 and Proposition 4 state that the ENC statistic $\hat{C}_P$ tends to zero when the numbers of in-sample observation R and out-of-sample observation P tend to infinity. We prove the asymptotic normality property of the ENC statistic. The following assumptions are used to obtain the limiting distribution in Proposition 5 and its asymptotic normality property in Proposition 6. These assumptions are only sufficient but not necessary and sufficient.

Assumption 1: The parameter estimates $\hat{\beta}_{i,t}, i = 1, 2, t = R, \ldots, T$, satisfy $\hat{\beta}_{i,t} - \beta_i = B_i(t)H_i(t) = \left(R^{-1} \sum_{j=t-R+1}^{t} q_{i,j}\right)^{-1} \left(R^{-1} \sum_{j=t-R+1}^{t} h_{i,j}\right)$, where $q_{i,j} = g\left(u_j^{(i)}\right) x_{i,j-1} x_{i,t-1}' / \mathbb{E}\left[g\left(u_j^{(i)}\right)\right]$ and $h_{i,j} = g\left(u_j^{(i)}\right) u_j^{(i)} x_{i,j-1} \mathbb{E}\left[g\left(u_j^{(i)}\right)\right]$.

Assumption 2: Let $U_t = \left[g\left(u_t\right) u_t, x_{2,t}' - \mathbb{E}x_{2,t}', h_{2,t}', \mathrm{vec}\left(h_{2,t}h_{2,t}' - \mathbb{E}h_{2,t}h_{2,t}'\right), \mathrm{vec}\left(q_{2,t} - \mathbb{E}q_{2,t}\right)'\right]$ (a) $\mathbb{E}U_t = 0$, (b) $\mathbb{E}q_{2,t} < \infty$ is p.d., (c) $\mathbb{E}u_t^2 = \sigma^2$. (d) Define $\tilde{U}_t$ the nonredundant elements of U_t, then $R^{-1}\mathbb{E}\left(\sum_{j=t-R+1}^{t} \tilde{U}_j\right)\left(\sum_{j=t-R+1}^{t} \tilde{U}_j\right)' = \Omega < \infty$ is positive definite.

Assumption 3: (a) $\mathbb{E}h_{2,t}h_{2,t}' = \sigma^2 \mathbb{E}q_{2,t}$, (b) $\mathbb{E}\left(h_{2,t} \mid h_{2,t-j}, q_{2,t-j}, j = 1, 2, \ldots\right) = 0$.

Assumptions 2 and 3 allow the application of an invariance principle and are sufficient for joint weak convergence of partial sums and averages of these partial sums to Brownian motion and integrals of Brownian motion.

Assumption 4: $\lim_{P,R\to\infty} P/R \equiv \pi \to \infty$.

In order to get accurate out-of-sample forecasts, it is essential to select an optimal in-sample window. However, no consensus opinion has been reached as it depends on the characteristics of the model. Inoue et al. (2017) find out the optimal window size by minimizing the conditional mean squared forecast error in a time-varying predictive regression model. They show that the optimal window size satisfies $R = O(T^{2/3})$. Hong et al. (2020) consider minimizing various out-of-sample forecast errors, including the unconditional, conditional, and global mean squared forecast errors, respectively. The optimal window size they found is $R = O(T^{4/5})$. The out-of-sample window P has a faster divergent rate than the in-sample window R. Thus, in our model, we set $P/R \to +\infty$.

We have the following main theorems:

Proposition 5: Suppose Assumptions 1-3 hold. Then

$$\begin{aligned}\mathrm{ENC}_P =& \frac{\sum_{t=R}^{T} c_{t+1}}{\sqrt{\sum_{t=R}^{T} c_{t+1}^2 - P\bar{C}^2}} \\ \xrightarrow{d}& \frac{\int_{\xi}^{1} \left[W(s) - W(s-\xi)\right] \mathrm{d}W(s)}{\sqrt{\int_{\xi}^{1} \left[W(s) - W(s-\xi)\right]^2 \mathrm{d}s}}\end{aligned}$$

under $\mathbb{H}_0$, where $\xi = R/T$ and $c_{t+1} = g\left(\hat{u}_{t+1}^{(1)}\right)\hat{u}_{t+1}^{(1)}\left(\hat{u}_{t+1}^{(1)} - \hat{u}_{t+1}^{(2)}\right)$.

Proof: See Appendix B.

Proposition 6: Suppose Assumptions 1-4 hold. Then $\text{ENC}_P \xrightarrow{d} N(0,1)$ under $\mathbb{H}_0$.

Proof: See Appendix B

Remarks: Diebold and Mariano (1995), West (1996) prove that ENC statistic is asymptotically normal under $\pi \geq 0$ when two models are non-nested. Clark and McCracken (2001) show that ENC statistic is asymptotically standard normal under $\pi = 0$ and but not standard normal under $\pi > 0$ when two models are nested. In Propositions 5 and 6, we show that ENC statistic is asymptotically standard normality under π tends to $+\infty$.

5.5 Model Averaging in Expectile Regressions

In this section, we consider combining two nested expectile regressions and find out the optimal combination weights in order to get the minimum loss. In the mean model, Clark and McCracken (2009) show that the weight endowed on Model 1 when combining the two nested model is negatively related to the SNR of Model 2, especially, when the covariate in Model 2 has no predictive power, the weight endowed on Model one is expected to be 1. Diebold (1989) discusses the use of model selection and model averaging. Peng and Yang (2021) compare model selection with model averaging and show that model averaging reaches the minimum loss, which is superior to the minimum loss obtained from model selection. We consider endowing Model 1 and Model 2 with the weight $1-\lambda$ and λ respectively and find out the property of optimal weight $\hat{\lambda}$. Proposition 7 gives the relationship

between $\hat{\lambda}$ and $\hat{C}_P$.

Proposition 7: We denote $\lambda = \arg\min_\lambda \mathbb{E}_F S(\mu^{(c)}, Y)$, where $\mu^{(c)} = (1-\lambda)\,\mu^{(1)} + \lambda\mu^{(2)}$. Under $\mathbb{H}_0$, $\hat{\lambda} \to 0$ if and only if $\hat{C}_P \xrightarrow{p} 0$ as $R, P \to \infty$.

Proof: See Appendix C.

Clark and McCracken (2009) show that in two nested mean models, the optimal weight endowing on Model 2 is SNR/ $(1 + \text{SNR})$, under the null hypothesis of no Granger Causality, the SNR in model 2 is 0, hence the optimal weight endowing on Model 1 and 2 are 0 and 1 respectively. From the equation of the optimal weight, if the SNR of Model 2 increases, e.g., b^2, σ_x^2 increases or σ_u^2 decreases in Equation (5.8) will lead to higher weight on Model 2. Proposition 6 shows that the optimal weight for the nested expectile model has the same property as nested mean model.

5.6 Monte Carlo Simulation

5.6.1 Simulation design

We set two DGPs to show the asymptotic distribution of encompassing test. In DGP1, we generate $u_{t+1}^{(2)}$ following uniform distribution on $[a_1, a_2]$, where $a_1 < 0 < a_2$. The error term $u_{t+1}^{(2)}$ needs to satisfy $\mathbb{E}\left[g\left(u^{(2)}\right) u^{(2)} \middle| X\right] = 0$. Thus, the relationship between

a_1 and a_2 can be expressed as $a_1 = -\sqrt{\frac{\tau}{1-\tau}}a_2$.[1] We set $u_{t+1}^{(2)}$ to have variance 1, hence $a_2 = \frac{\sqrt{12}}{1+\sqrt{\frac{\tau}{1-\tau}}}$ and $a_1 = -\frac{\sqrt{12}}{1+\sqrt{\frac{1-\tau}{\tau}}}$[2]. E.g. if we set $\tau = 0.1$, then $a_1 = -\frac{\sqrt{3}}{2}$ and $a_2 = \frac{3\sqrt{3}}{2}$, so $u_{t+1}^{(2)} \sim U[-\frac{\sqrt{3}}{2}, \frac{8\sqrt{3}}{3}]$. The mean-standard deviation ratio of the error term $u_{t+1}^{(2)}$ should be

$$\frac{\mathbb{E}(u_{t+1}^{(2)})}{\sqrt{Var(u_{t+1}^{(2)})}} = \frac{\frac{a_1+a_2}{2}}{\frac{a_1-a_2}{\sqrt{12}}} = \frac{\sqrt{3}(a_1+a_2)}{a_1-a_2}.$$

Next we generate the error term of $u_{t+1}^{(1)}$ as follows:

$$\mathbb{E}(u_{t+1}^{(1)}) = \mathbb{E}(y_{t+1} - c_1) = \mathbb{E}(c_2 + bx_t + u_{t+1}^{(2)} - c_1) = c_2 + 0 + \frac{a_1+a_2}{2} - c_1,$$

$$Var(u_{t+1}^{(1)}) = Var(y_{t+1} - c_1) = Var(c_2 + bx_t + u_{t+1}^{(2)}) - c_1) = b^2\sigma_x^2 + 1.$$

Thus, in order to ensure $\mathbb{E}\left[g\left(u^{(1)}\right)u^{(1)}\middle|X\right] = 0$, under uniform distribution of our DGP, the error term of $u_{t+1}^{(1)}$ should also satisfy

$$\frac{\mathbb{E}(u_{t+1}^{(1)})}{\sqrt{Var(u_{t+1}^{(1)})}} = \frac{c_2 + \frac{a_1+a_2}{2} - c_1}{\sqrt{b^2\sigma_x^2+1}} = \frac{\sqrt{3}(a_1+a_2)}{a_1-a_2}, \tag{5.29}$$

which implies that

$$c_2 - c_1 = -\frac{a_1+a_2}{2} + \frac{\sqrt{3}(a_1+a_2)}{a_1-a_2}\sqrt{b^2\sigma_x^2+1}. \tag{5.30}$$

Without loss of generality, we set $c_1 = 1$. In DGP 1, we set stationary process $x_t = \phi x_{t-1} + v_t$ be AR(1) stationary process, where $v_t \overset{iid}{\sim} N(0, \sigma_v^2)$, thus, $\{x_t\}$ has zero mean and variance $\sigma_v^2/(1-\phi)$. we consider $\tau \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, $\phi \in \{0, 0.95\}$, $\sigma_v \in \{1\}$, $\sigma_u \in \{0.1\}$ for two DGPs. $b = 0$ for the size of the test and $b = 0.1$ for the power of the test.

[1] $u_{t+1}^{(2)}$ follows uniform distribution on $[a_1, a_2]$, so we have $\tau\int_0^{a_2} u_{t+1}^{(2)}\frac{1}{a_2-a_1}du_{t+1}^{(2)} + (1-\tau)\int_{a_1}^{0} u_{t+1}^{(2)}\frac{1}{a_2-a_1}du_{t+1}^{(2)} = \frac{\tau}{a_2-a_1}(u_{t+1}^{(2)})^2|_0^{a_2} + \frac{(1-\tau)}{a_2-a_1}(u_{t+1}^{(2)})^2|_{a_1}^0 = \frac{\tau a_2^2}{a_2-a_1} - \frac{(1-\tau)a_1^2}{a_2-a_1} = 0$. Simplifying this equation, we obtain that the relationship between a_1 and a_2 is $a_1 = -\sqrt{\frac{\tau}{1-\tau}}a_2$.

[2] $u_{t+1}^{(2)}$ follows uniform distribution on $[a_1, a_2]$, so we have the variance $Var = \frac{(a_2-a_1)^2}{12} = \frac{(a_2+\sqrt{\frac{\tau}{1-\tau}}a_1)^2}{12} = \frac{(1+\sqrt{\frac{\tau}{1-\tau}})^2a_2^2}{12} = 1$. Simplifying this equation, we obtain a_1 and a_2 are functions of τ.

In DGP2, we generate $u_{t+1}^{(2)}$ following normal distribution. All the Results about DGP 2 are almost the same as DGP 1. See Supplemental Appendix for details.

We use MATLAB to estimate the models. For Model 1, we use rolling scheme to regress $\{y_s\}_{s=t-R+1}^{t}$ on $\{x_{1,s-1}\}_{s=t-R+1}^{t}$ to obtain $\hat{c}_{1,t}$. Since $\{x_{1,s-1}\}_{s=t-R+1}^{t}$ is a vector of 1, the forecast dependent variable is $\hat{y}_{1,t+1} = \hat{c}_{1,t}$, which is also called "historical expectile", the forecast error is $\hat{u}_{t+1}^{(1)} = y_{t+1} - \hat{c}_t$, where $t = R, ..., T$. For Model 2, we use rolling scheme to regress $\{y_s\}_{s=t-R+1}^{t}$ on $\{1, x_{s-1}\}_{s=t-R+1}^{t}$ to obtain $\left(\hat{c}_{2,t}, \hat{b}_t\right)$, so we can get the forecast dependent variable is $\hat{y}_{2,t+1} = \hat{c}_{2,t} + \hat{b}_t x_t$ and the forecast error is $\hat{u}_{t+1}^{(2)} = y_{t+1} - \hat{y}_{2,t+1}$.

In the simulation, we set the number of in-sample observations $R \in \{60, 120, 240\}$ and the number of the out-of-sample forecasts $P \in \{48, 240, 1200\}$. We compare the ENC statistics with the DM and CCS statistics. The ENC statistic is $ENC_P \equiv \hat{Q}_P^{-0.5}\sqrt{P}\hat{C}_P$, where $\hat{Q}_P = \text{var}\left(\sqrt{P}\hat{C}_P\right)$. The DM statistic is $DM_P \equiv \hat{S}_P^{-0.5}\sqrt{P}\hat{D}_P$, where $\hat{S}_P = \text{var}\left(\sqrt{P}\hat{D}_P\right)$. The CCS statistic is $CCS_P \equiv \hat{W}_P^{-0.5}\sqrt{P}\hat{M}_P$, where $\hat{M}_P = P^{-1}\sum_{t=R}^{T} g\left(\hat{u}_{t+1}^{(1)}\right)\hat{u}_{t+1}^{(1)}x_t$ and $\hat{W}_P = \text{var}\left(\sqrt{P}\hat{M}_P\right)$. $\hat{S}_P$, $\hat{Q}_P$ and $\hat{W}_P$ are consistent estimators of $S_P = \text{var}\left(\sqrt{P}D_P\right)$, $Q_P = \text{var}\left(\sqrt{P}C_P\right)$ and $W_P = \text{var}\left(\sqrt{P}M_P\right) = \text{var}\left(P^{-0.5}\sum_{t=R}^{T} g\left(u_{t+1}^{(1)}\right)u_{t+1}^{(1)}x_t\right)$, respectively. We use rolling windows to estimate $\hat{c}_{1,t}$, $\hat{c}_{2,t}$, and $\hat{b}_t$, and then predict one step ahead dependent variable $\hat{y}_{t+1}$ and calculate the forecast error $\hat{u}_{t+1}$, then obtain DM_P, ENC_P, and CCS_P statistics. We repeat this procedure 2000 times and get the asymptotic distribution of DM_P, ENC_P, and CCS_P, and the size and power. For model averaging, we estimate $\hat{\lambda}$ by regressing $\left\{\hat{u}_{t+1}^{(1)}\right\}_{t=R}^{T}$ on $\left\{\left(\hat{u}_{t+1}^{(1)} - \hat{u}_{t+1}^{(2)}\right)\right\}_{t=R}^{T}$ and use the optimal weight $\hat{\lambda}$ to calculate the forecast error $\hat{u}_{t+1}^{(c)}$ from combined model. Then, we can obtain the out-of-sample loss for Model 1, Model 2, and the combined model. We repeat this procedure 2000

times and get the out-of-sample average ALS loss for Model 1, Model 2 and the combined model using the following model:

$$\hat{S}_{\tau}^{(i)} \equiv \frac{1}{P}\sum_{t=R}^{T} \text{abs}\left(\tau - 1\left(\hat{u}_{t+1}^{(i)} < 0\right)\right)\left(\hat{u}_{t+1}^{(i)}\right)^2, \quad i = 1, 2, c, \tag{5.31}$$

where superscripts (1), (2), and (c) represent Model 1, Model 2 and the combined model. Comparing these three losses, we can find out which forecast is the best one in expectile regression. We report the results of DGP1.

5.6.2 Simulation result on encompassing test

For space, we report results only for DGP 1. The results for DGP 2 are reported in the supplemental appendix as the results are similar.

Tables 1-4 and Figures 1-4 show the Monte Carlo simulation asymptotic distribution and the size and power of DM_P, ENC_P and CCS_P statistics for DGP 1. Tables 1-2 show the size of the test with $b = 0$, $\phi = 0, 0.95$, and $\sigma_u = 0.1$. The tables demonstrate that the sizes of ENC_P and CCS_P are good under 5% nominal level, whereas the size of test for DM_P is much less than 5% under 5% nominal level, which implies that DM_P has a downward bias under $\mathbb{H}_0$. Tables 3-4 show the power of the test with $b = 0.1$, $\phi = 0, 0.95$, and $\sigma_u = 0.1$ in DGP 1. These tables show that the power of tests for these three statistics DM_P, ENC_P and CCS_P are more significant when the signal-noise ratio (SNR) is higher. Figures 1-2 and 3-4 of supplemental material show the asymptotic distribution of DM_P, ENC_P and CCS_P statistics under $\mathbb{H}_0$ in DGP 1. From Figures 1-2, we can see that DM_P has the negative mean, which implies that DM_P has a downward bias under $\mathbb{H}_0$. However, the distributions of ENC_P, and CCS_P are close to the standard normal distribution. When

$R = 60$ and $P = 1200$, the distributions of ENC_P, and CCS_P are bell-shaped with highest value approximately equaling to $1/\sqrt{2\pi}$. Figures 3-4 show the asymptotic distribution of DM_P, ENC_P and CCS_P statistics under $\mathbb{H}_1$ in DGP 1. We can see that ENC_P has the highest power under $\mathbb{H}_1$, which provides the same results as Tables 3-4.

5.6.3 Simulation result on combining forecasts

We report results for DGP 1. We use the out-of-sample average ALS loss to show the forecast of the combined model outperforms any other individual models. Tables 5-8 show out-of-sample average of $\hat{S}_\tau^{(1)}$ of Model 1, $\hat{S}_\tau^{(2)}$ of Model 2, $\hat{S}_\tau^{(c)}$ of combined model from 2000 Monte Carlo simulations. For each R, P, and τ, we have a block containing 6 numbers in 2 columns. The first column of each block shows the mean, skewness, and kurtosis of $\hat{\lambda}_P$. The second column of each block shows the out-of-sample average of $\hat{S}_\tau^{(1)}$, $\hat{S}_\tau^{(2)}$, $\hat{S}_\tau^{(c)}$.

Tables 5-6 show the estimate of $\hat{\lambda}_P$ and the out-of-sample average of Model 1, Model 2, and combined model under $\mathbb{H}_0$. From these two tables, for different R, P, and τ, we can see that the combined model has the lowest out-of-sample average ALS loss, so the combined model outperforms Model 1 and Model 2. In other words, model averaging is superior to model selection. Moreover, the optimal weight $\hat{\lambda}$ is decreasing when the out-of-sample observations P increases. Furthermore, all optimal weights $\hat{\lambda}$ are negative under different R and P, but the optimal weight $\hat{\lambda}$ are close to zero when P is large. In addition, all three out-of-sample average ALS losses are decreasing as R increases and are increasing

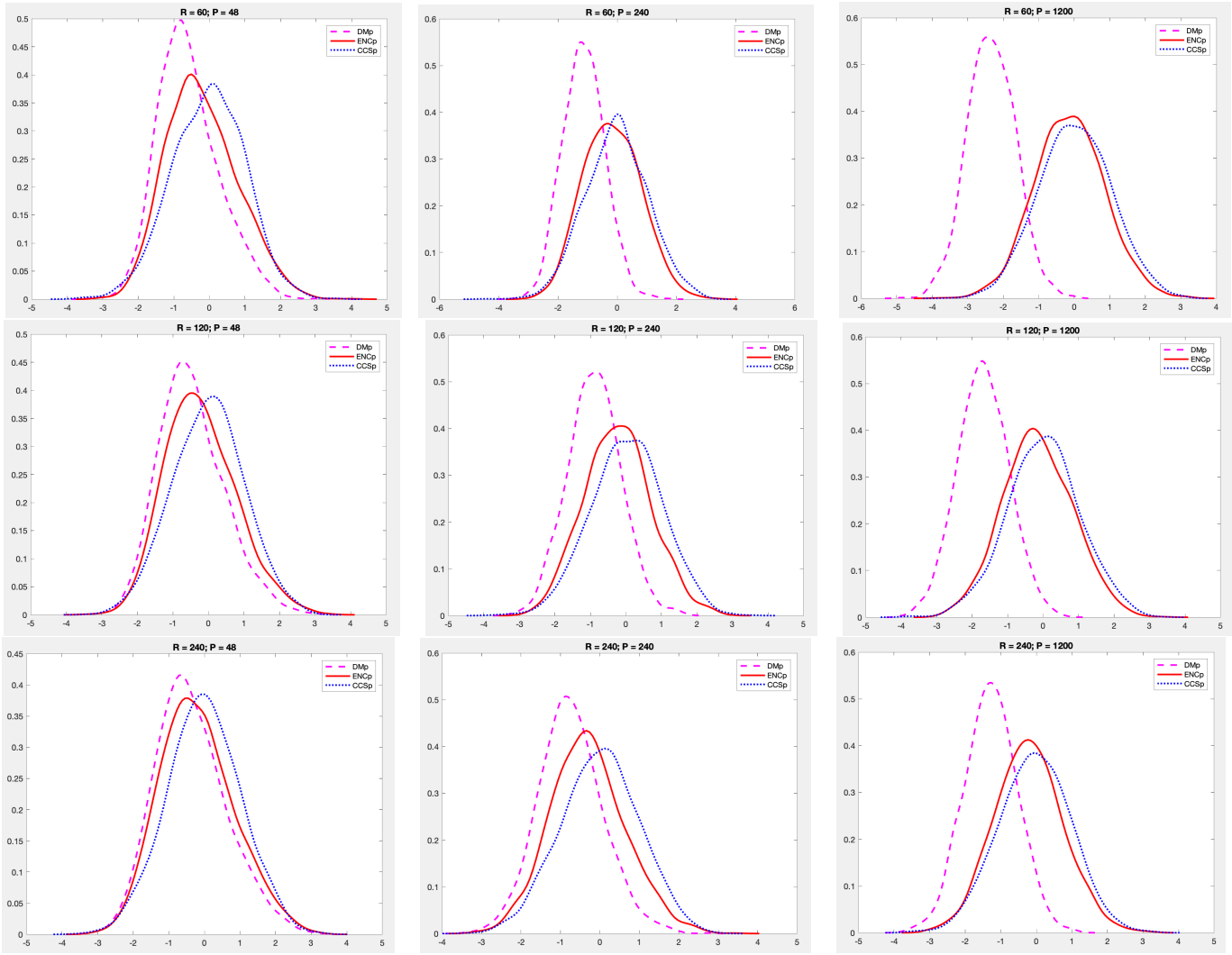

Figure 5.1: Simulation Results of DGP 1: Size of Test, $= 0.1$, $\phi = 0$

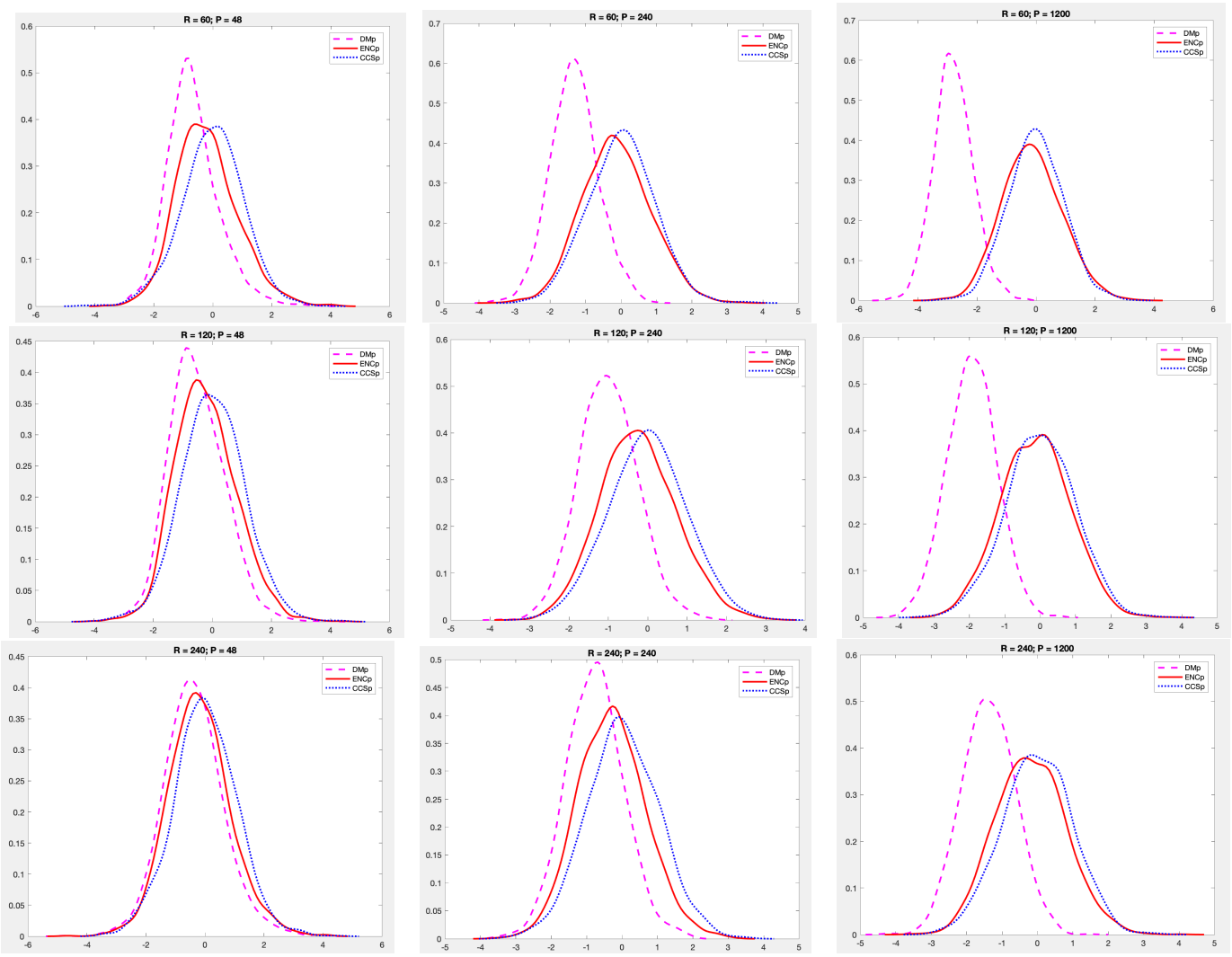

Figure 5.2: Simulation Results of DGP 1: Size of Test, $= 0.1$, $\phi = 0.95$

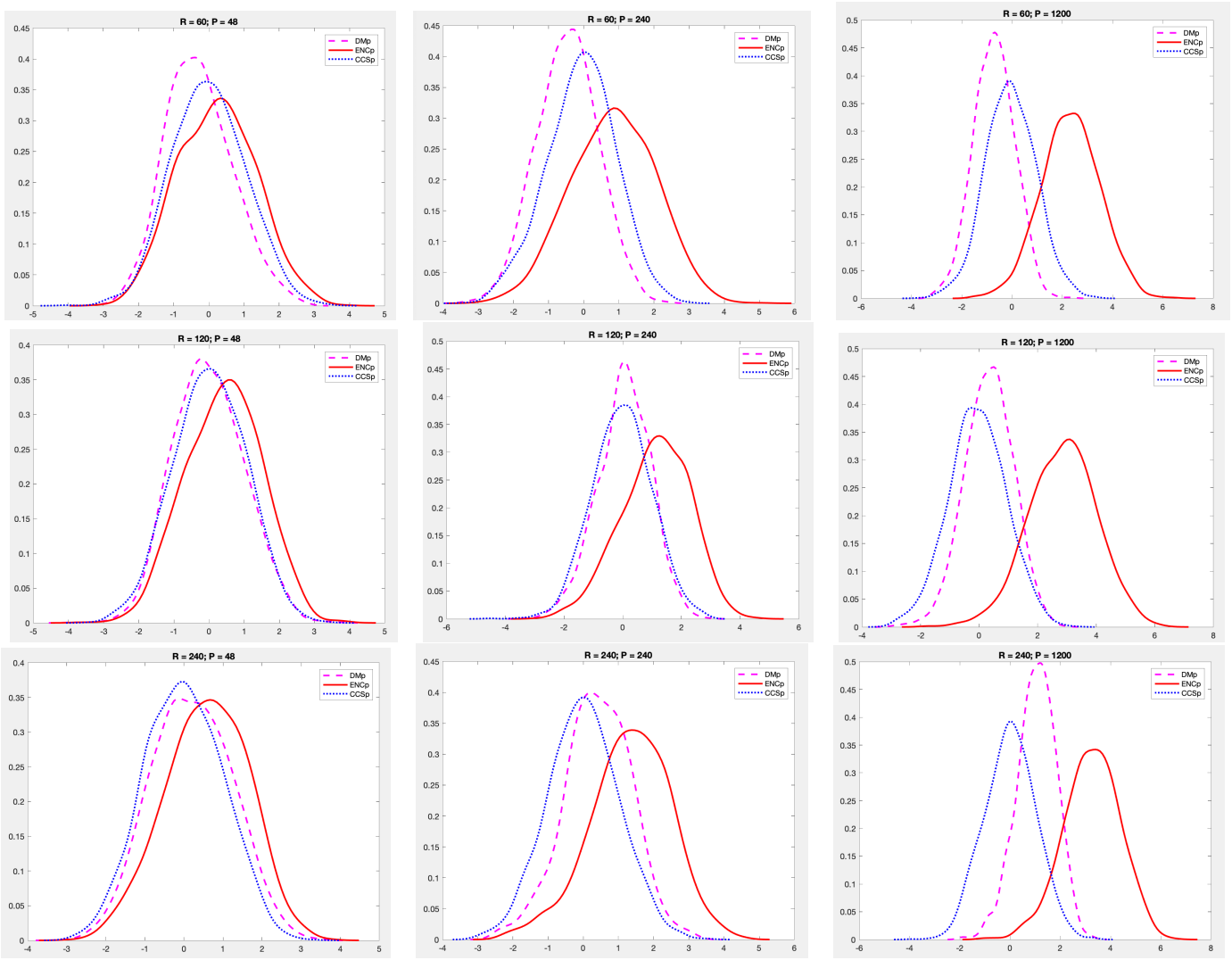

Figure 5.3: Simulation Results of DGP 1: Size of Test, $= 0.1$, $\phi = 0$

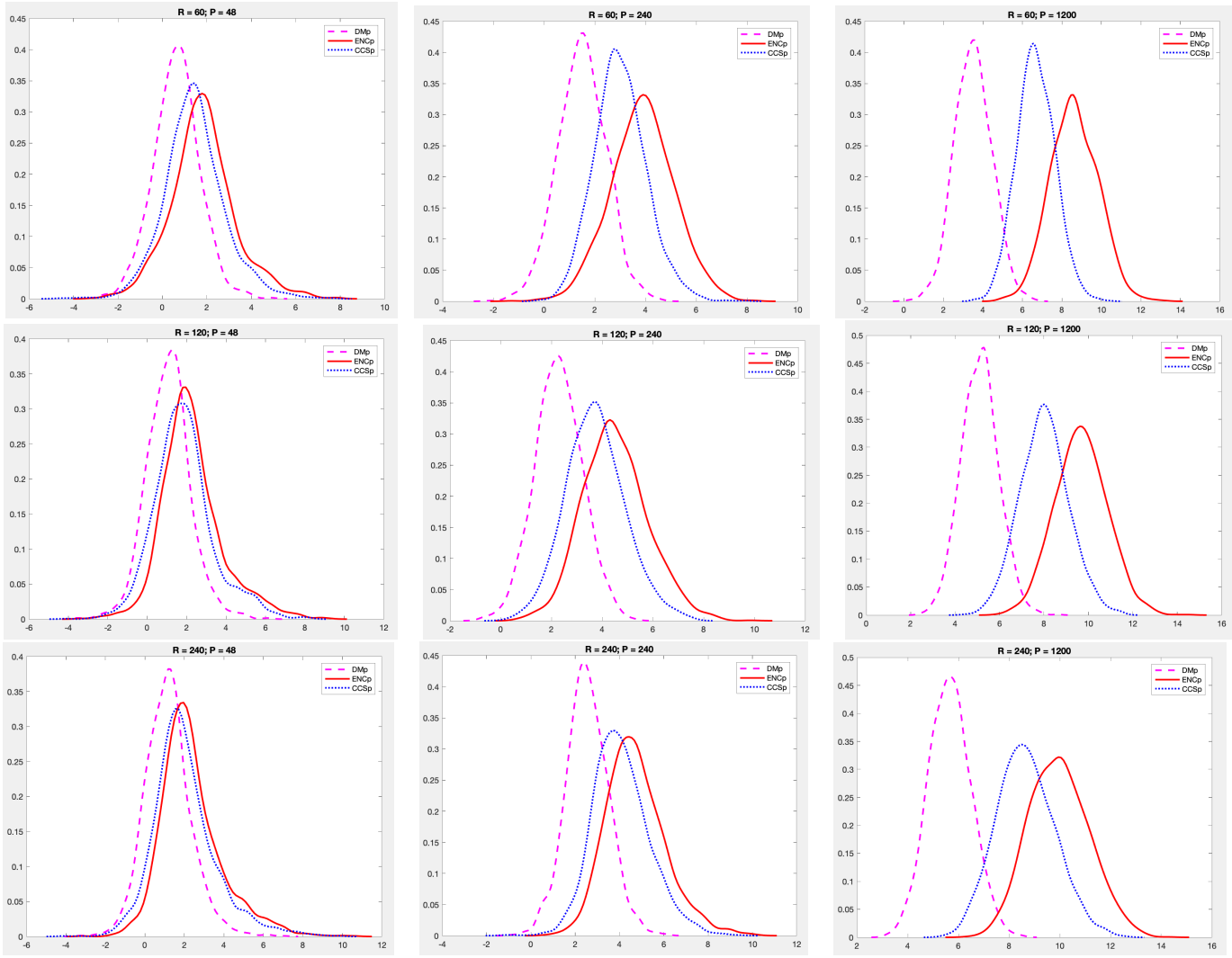

Figure 5.4: Simulation Results of DGP 1: Size of Test, $= 0.1$, $\phi = 0$

as P increases. This is because we can predict y_{t+1} accurately by using more information y in the previous periods when R increases. However, if P increases, the forecast periods are increasing, and the forecast of y_{t+1} is becoming difficult.

Tables 7-8 show the estimate of $\hat{\lambda}_P$ and the out-of-sample average of Model 1, Model 2, and combined model under $\mathbb{H}_1$. From the two tables, for different R, P, and τ, we can see that the combined model also has the lowest out-of-sample average ALS loss. Model averaging is superior to model selection. Moreover, the optimal weight $\hat{\lambda}$ is increases when the out-of-sample observations P increase. Moreover, when $b = 0.1$ and $\phi = 0$, optimal weight $\hat{\lambda}$ are negative under $R = 60, 120$ or $P = 48$ and are positive under $R = 240$ or $P = 240, 1200$. The reason is that we can predict y_{t+1} by using the information y in the previous periods when R is small or P is small, so we favor Model 1. However, we can use dependent variable $\{x_t\}$ to improve the prediction accuracy when R is large or P is large. When $b = 0.1$ and $\phi = 0.95$, all optimal weights $\hat{\lambda}$ are positive and are most likely in the range from 0.5 to 1. This is because we need to use the dependent variables $\{x_t\}$ and the information y to predict y_{t+1} when ϕ is large. Furthermore, all three out-of-sample average ALS losses are decreasing as R increases and are increasing as P increases.

5.7 Predictive Expectile Regressions for the Equity Premium

In this section, we use the data from Welch and Goyal (2008). We chose to use annual, quarterly, and monthly data from January 1926 to December 2019. We consider 18 different macroeconomic and financial variables that are similar to the variables Welch and

Goyal (2008) used. We build two nested expectile models to test if these variables Granger cause the equity premium. For Model 1, we find out the expectile by using the previous R in-sample observations of equity premium and then use the estimated mode to predict the future equity premium at time $R+1$. For Model 2, we estimate the constant term and covariance term by using the previous R in-sample observations of equity premium and independent variable and then use these estimated coefficients and independent variable

5.7.1 Granger Causality in the predictive expectile regressions for equity premium

We set that the forecasts begin 20 years after the sample and the forecasts begin in 1965. We use a rolling window to estimate and set $\tau \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. Tables 9, 11, and 13 show the results of three statistics DM_P, ENC_P and CCS_P and their p-values by using different additional variables x_t based on annual, quarterly and monthly data when the out-of-sample forecasts begin 20 years after data are available. Tables 10, 12, and 14 demonstrate the results when the out-of-sample forecasts begin in 1965.

From Table 9, when out-of-sample forecasts begin 20 years after data are available, based on the annual data, dividend price ratio (d/p), dividend yield (d/y), earning price ratio (e/p), percent equity issuing (eqis), investment capital ratio (i/k) and consumption-based macroeconomic ratios (cay) Granger-cause equity premium respectively with different τ in expectile by using ENC test. Moreover, for the mean case ($\tau = 0.5$), d/p, d/y, and i/k can predict the equity premium by using the ENC test. Thus, we can see that e/p, eqis, and cay can predict the equity premium with different expectiles but not with $\tau = 0.5$ for the mean. Furthermore, by using the DM test, there are no macroeconomic and financial

variables that can predict the equity premium in mean regression. These results of the DM test in mean are consistent with those in Welch and Goyal (2008). Thus, for the mean case, d/p, d/y, and i/k can Granger-cause the equity premium by using the ENC test but not the DM test.

Table 10 shows the results of out-of-sample forecasts beginning in 1965. By using the annual data, it demonstrates the same results for the mean from the DM test as in Table 9 and in Welch and Goyal (2008). In Table 10, based on ENC statistic, there are six more macroeconomic and financial variables that Granger-cause the equity premium with different τ in expectile, but only one variable i/k can predict the equity premium in mean. Hence, although there is no x_t Grange-cause the equity premium in the mean by using DM statistic, i/k can Granger-cause the equity premium in the mean by using the ENC test.

5.7.2 Model averaging in the predictive expectile regressions for equity premium

In the empirical analysis, we still set $\tau \in \{0.01, 0.05, 0.1, 0.3, 0.5, 0.7, 0.9, 0.95, 0.99\}$. Table 15 shows the results of mean of optimal weight $\hat{\lambda}$ and out-of-sample average ALS loss for these three models $\hat{S}_\tau^{(1)}$, $\hat{S}_\tau^{(2)}$ and $\hat{S}_\tau^{(c)}$ by using different independent variables x_t under different τ. From Table 15, we can see that the combined model has the lowest out-of-sample average ALS loss compared with Model 1 and Model 2 under different τ for all 15 variables. It implies that the model averaging is superior to model selection. This result is consistent with the simulation result in Section 5.3 and the result from a recent paper by Peng and Yang (2021). However, the optimal weights differ with independent variables x_t and τ. The optimal weights for default return spread (dfr) under different τ are negative,

while the optimal weights for other variables under different τ are positive.

5.8 Conclusions

We point out that the ENC test has a good size under the null hypothesis of no Granger Causality of the covariate, whereas the DM test seriously rejects the null hypothesis. Furthermore, by showing that the asymptotic distribution of ENC statistic is standard normal under $\mathbb{H}_0$, we explain the reason why the ENC test has a good size whereas the DM test is undersized. Under the alternative hypothesis of Granger Causality of the covariate, we point out that the ENC test has good power. In the Monte-Carlo simulation, we demonstrate that the ENC statistic, compared with the DM statistic, has the correct size and good power under $\mathbb{H}_0$. We show the relationship between the ENC test and model averaging. We use Welch-Goyal data to test if macroeconomic and financial variables Granger Cause the equity premium under different expectiles. We demonstrate that the ENC test significantly improves the predictive accuracy regardless of the expectiles, which can not be found in the DM test. In this paper, we focus on linear predictive expectile models with one additional predictor at a time just for simplicity, which can be generalized to nonlinear expectile regression models with many additional candidate predictors in high dimensions, which need to be regularized. We leave this for future work.

5.9 Appendix A. Proofs of Propositions 3-4

Proof of Proposition 3.

We only need to show that the each term of $\hat{V}_P$ is non-negative, i.e.,

$$g\left(\hat{u}_{t+1}^{(1)}\right)\left(\hat{u}_{t+1}^{(1)}-\hat{u}_{t+1}^{(2)}\right)^2-g\left(\hat{u}_{t+1}^{(1)}\right)\left(\hat{u}_{t+1}^{(2)}\right)^2+g\left(\hat{u}_{t+1}^{(2)}\right)\left(\hat{u}_{t+1}^{(2)}\right)^2$$

is non-negative at each t. We consider 2 cases.

Case 1: If $\hat{u}_{t+1}^{(1)}\times\hat{u}_{t+1}^{(2)}>0$, then $g\left(\hat{u}_{t+1}^{(1)}\right)=g\left(\hat{u}_{t+1}^{(2)}\right)$ and

$$\begin{aligned}&g\left(\hat{u}_{t+1}^{(1)}\right)\left(\hat{u}_{t+1}^{(1)}-\hat{u}_{t+1}^{(2)}\right)^2-g\left(\hat{u}_{t+1}^{(1)}\right)\left(\hat{u}_{t+1}^{(2)}\right)^2+g\left(\hat{u}_{t+1}^{(2)}\right)\left(\hat{u}_{t+1}^{(2)}\right)^2\\&=g\left(\hat{u}_{t+1}^{(1)}\right)\left(\hat{u}_{t+1}^{(1)}-\hat{u}_{t+1}^{(2)}\right)^2\geq 0.\end{aligned}$$

Case 2: If $\hat{u}_{t+1}^{(1)}\times\hat{u}_{t+1}^{(2)}<0$, then

$$\begin{aligned}&g\left(\hat{u}_{t+1}^{(1)}\right)\left(\hat{u}_{t+1}^{(1)}-\hat{u}_{t+1}^{(2)}\right)^2-g\left(\hat{u}_{t+1}^{(1)}\right)\left(\hat{u}_{t+1}^{(2)}\right)^2+g\left(\hat{u}_{t+1}^{(2)}\right)\left(\hat{u}_{t+1}^{(2)}\right)^2\\&=g\left(\hat{u}_{t+1}^{(1)}\right)\left(\hat{u}_{t+1}^{(1)}\right)^2-2g\left(\hat{u}_{t+1}^{(1)}\right)\hat{u}_{t+1}^{(1)}\hat{u}_{t+1}^{(2)}+g\left(\hat{u}_{t+1}^{(2)}\right)\left(\hat{u}_{t+1}^{(2)}\right)^2>0.\end{aligned}$$

Thus, $\hat{V}_P$ is non-negative.

Proof of Proposition 4.

We estimate the ALS scoring function with combined μ to find out the adjusted ALS loss-differential. We have

$$\lambda=\arg\min_{\lambda}\mathbb{E}_F S(\mu^{(c)}(\tau),Y),$$

where $\mu_{c,t}(\tau)=(1-\lambda)\,\mu_{1,t}(\tau)+\lambda\mu_{2,t}(\tau)$, $u_{t+1}^{(1)}=y_{t+1}-\hat{\mu_{1,t}}(\tau)$ and $u_{t+1}^{(2)}=y_{t+1}-\hat{\mu_{2,t}}(\tau)$, so we have $u_{t+1}^{(c)}=(1-\lambda)\,u_{t+1}^{(1)}+\lambda u_{t+1}^{(2)}$. First, we rewrite the ALS scoring function with

combined error term as follow:

$$\begin{aligned}
\mathbb{E}S(\mu^{(c)}(\tau), Y) &= \mathbb{E}\left[\frac{1}{2}\,\text{abs}\left(\tau - 1(u_{t+1}^{(c)} < 0)\right)\left(u_{t+1}^{(c)}\right)^2\right] \\
&= \frac{1}{2}\int_{(1-\lambda)x_{1,t}'\beta_1+\lambda x_{2,t}'\beta_2}^{+\infty} \tau\left[y_{t+1} - (1-\lambda)\,x_{1,t}'\beta_1 - \lambda x_{2,t}'\beta_2\right]^2 f_t\left(y_{t+1}\right) dy_{t+1} \\
&\quad + \frac{1}{2}\int_{-\infty}^{(1-\lambda)x_{1,t}'\beta_1+\lambda x_{2,t}'\hat{\beta}_2} (1-\tau)\left[y_{t+1} - (1-\lambda)\,x_{1,t}'\beta_1 - \lambda x_{2,t}'\beta_2\right]^2 \\
&\quad \times f_t\left(y_{t+1}\right) dy_{t+1}.
\end{aligned}$$

Then, taking a derivative of the expected ALS scoring function with respect to λ, we have

$$\begin{aligned}
C &\equiv \nabla_\lambda \mathbb{E}_F S(\mu^{(c)}(\tau), Y) \\
&= \frac{1}{2}\int_{(1-\lambda)x_{1,t}'\beta_1+\lambda x_{2,t}'\beta_2}^{+\infty} \frac{\partial\tau\left[y_{t+1} - (1-\lambda)\,x_{1,t}'\beta_1 - \lambda x_{2,t}'\beta_2\right]^2 f_t\left(y_{t+1}\right)}{\partial\lambda} dy_{t+1} \\
&\quad - \frac{1}{2}\tau\left[(1-\lambda)\,x_{1,t}'\beta_1 + \lambda x_{2,t}'\beta_2 - (1-\lambda)\,x_{1,t}'\beta_1 - \lambda x_{2,t}'\beta_2\right]^2 f_t\left(y_{t+1}\right) \\
&\quad \times \frac{\partial\left(1-\lambda\right)x_{1,t}'\beta_1 + \lambda x_{2,t}'\beta_2}{\partial\lambda} \\
&\quad + \frac{1}{2}\int_{-\infty}^{(1-\lambda)x_{1,t}'\beta_1+\lambda x_{2,t}'\beta_2} \frac{\partial\left(1-\tau\right)\left[y_{t+1} - (1-\lambda)\,x_{1,t}'\beta_1 - \lambda x_{2,t}'\beta_2\right]^2 f_t\left(y_{t+1}\right)}{\partial\lambda} dy_{t+1} \\
&\quad + \frac{1}{2}\tau\left[(1-\lambda)\,x_{1,t}'\beta_1 + \lambda x_{2,t}'\beta_2 - (1-\lambda)\,x_{1,t}'\beta_1 - \lambda x_{2,t}'\beta_2\right]^2 f_t\left(y_{t+1}\right) \\
&\quad \times \frac{\partial\left(1-\lambda\right)x_{1,t}'\beta_1 + \lambda x_{2,t}'\beta_2}{\partial\lambda} \\
&= -2\int_{(1-\lambda)x_{1,t}'\beta_1+\lambda x_{2,t}'\beta_2}^{+\infty} \tau\left[y_{t+1} - (1-\lambda)\,x_{1,t}'\beta_1 - \lambda x_{2,t}'\beta_2\right] f_t\left(y_{t+1}\right) \\
&\quad \times \left(-x_{1,t}'\beta_1 + x_{2,t}'\beta_2\right) dy_{t+1} \\
&\quad - 2\int_{-\infty}^{(1-\lambda)x_{1,t}'\beta_1+\lambda x_{2,t}'\beta_2} (1-\tau)\left[y_{t+1} - (1-\lambda)\,x_{1,t}'\beta_1 - \lambda x_{2,t}'\beta_2\right] f_t\left(y_{t+1}\right) \\
&\quad \times \left(-x_{1,t}'\beta_1 + x_{2,t}'\beta_2\right) dy_{t+1}
\end{aligned}$$

$$
\begin{aligned}
&= -\int_{(1-\lambda)x'_{1,t}\beta_1+\lambda x'_{2,t}\beta_2}^{+\infty} \tau\left[y_{t+1}-(1-\lambda)\,x'_{1,t}\beta_1-\lambda x'_{2,t}\beta_2\right] \\
&\quad\times f_t\left(y_{t+1}\right)\left(y_{t+1}-x'_{1,t}\beta_1-y_{t+1}+x'_{2,t}\beta_2\right)dy_{t+1} \\
&\quad-\int_{-\infty}^{(1-\lambda)x'_{1,t}\beta_1+\lambda x'_{2,t}\beta_2}(1-\tau)\left[y_{t+1}-(1-\lambda)\,x'_{1,t}\beta_1-\lambda x'_{2,t}\beta_2\right] \\
&\quad\times f_t\left(y_{t+1}\right)\left(y_{t+1}-x'_{1,t}\beta_1-y_{t+1}+x'_{2,t}\beta_2\right)dy_{t+1} \\
&= -\int_{(1-\lambda)x'_{1,t}\beta_1+\lambda x'_{2,t}\beta_2}^{+\infty} \tau\left[y_{t+1}-(1-\lambda)\,x'_{1,t}\beta_1-\lambda x'_{2,t}\beta_2\right] f_t\left(y_{t+1}\right) \\
&\quad\times\left(u_{t+1}^{(1)}-u_{t+1}^{(2)}\right)dy_{t+1} \\
&\quad-\int_{-\infty}^{(1-\lambda)x'_{1,t}\beta_1+\lambda x'_{2,t}\beta_2}(1-\tau)\left[y_{t+1}-(1-\lambda)\,x'_{1,t}\beta_1-\lambda x'_{2,t}\beta_2\right] f_t\left(y_{t+1}\right) \\
&\quad\times\left(u_{t+1}^{(1)}-u_{t+1}^{(2)}\right)dy_{t+1} \\
&= -\left(u_{t+1}^{(1)}-u_{t+1}^{(2)}\right)\int_{(1-\lambda)x'_{1,t}\beta_1+\lambda x'_{2,t}\beta_2}^{+\infty} \tau\left[y_{t+1}-(1-\lambda)\,x'_{1,t}\beta_1-\lambda x'_{2,t}\beta_2\right] dF_t\left(y_{t+1}\right) \\
&\quad-\left(u_{t+1}^{(1)}-u_{t+1}^{(2)}\right)\int_{-\infty}^{(1-\lambda)x'_{1,t}\beta_1+\lambda x'_{2,t}\beta_2}(1-\tau)\left[y_{t+1}-(1-\lambda)\,x'_{1,t}\beta_1-\lambda x'_{2,t}\beta_2\right] dF_t\left(y_{t+1}\right) \\
&= -\mathbb{E}_F\left[g\left(u_{t+1}^{(c)}\right)u_{t+1}^{(c)}\left(u_{t+1}^{(1)}-u_{t+1}^{(2)}\right)\right] \\
&= -\mathbb{E}_F\left[g\left(u_{t+1}^{(1)}\right)u_{t+1}^{(1)}\left(u_{t+1}^{(1)}-u_{t+1}^{(2)}\right)\right]=0 \quad \text{ under } \mathbb{H}_0.
\end{aligned}
$$

$$
\hat{C}_P = P^{-1}\sum_{t=R}^{T} g\left(\hat{u}_{t+1}^{(1)}\right)\hat{u}_{t+1}^{(1)}\left(\hat{u}_{t+1}^{(1)}-\hat{u}_{t+1}^{(2)}\right) \xrightarrow{p} \mathbb{E}_F\left[g\left(u_{t+1}^{(1)}\right)u_{t+1}^{(1)}\left(u_{t+1}^{(1)}-u_{t+1}^{(2)}\right)\right]=0
$$

as $R, P \to \infty$ under $\mathbb{H}_0$. Under $\mathbb{H}_0$, $\lambda = 0$, then we have $\mu^{(c)} = \mu^{(1)}$, $u_{t+1}^{(c)} = u_{t+1}^{(1)}$ and $g\left(u_{t+1}^{(c)}\right) = g\left(u_{t+1}^{(1)}\right)$. Thus, $C = \mathbb{E}\left[g\left(u_{t+1}^{(1)}\right)u_{t+1}^{(c)}\left(u_{t+1}^{(1)}-u_{t+1}^{(2)}\right)\right] = 0$, so $\hat{C}_P \xrightarrow{p} 0$ as $R, P \to \infty$.

5.10 Appendix B. Proofs of Propositions 5 and 6

This section proves Propositions 5 and 6. Since $ENC_P \equiv \hat{Q}_P^{-0.5}\sqrt{P}\hat{C}_P$, to show $\hat{C}_P$ has zero mean in Proposition 4, we only need to show ENC_P has zero mean, it is also the sufficient condition of Theorem2. Under $\mathbb{H}_0$, $u_{t+1}^{(1)} = u_{t+1}^{(2)} \equiv u_{t+1}$. We have defined $\hat{u}_{t+1}^{(1)} = y_{t+1} - \hat{c}_{1,t} \equiv y_{t+1} - x_{1,t}'\hat{\beta}_{1,t}$ and $\hat{u}_{t+1}^{(2)} = y_{t+1} - \hat{c}_{2,t} - \hat{b}_t x_t \equiv y_{t+1} - x_{2,t}'\hat{\beta}_{2,t}$, where $x_{1,t}' = 1$ and $x_{2,t}' = (1, x_t)$. Let $q_{i,t} = g\left(u_t^{(i)}\right) x_{i,t-1}x_{i,t-1}'/\mathbb{E}\left[g\left(u_t^{(i)}\right)\right]$ and $B_i = \mathbb{E}\left[q_{i,t}\right]^{-1}$, $h_{i,t} = g\left(u_t^{(i)}\right) \times u_t^{(i)} \times x_{i,t-1}' \times \mathbb{E}\left[g\left(u_t^{(i)}\right)\right]$, $i = 1, 2$. Therefore $B_1 = 1$ and $B_2 = \begin{pmatrix} 1 & 0 \\ 0 & \mathbb{E}\left(x_t^2\right) \end{pmatrix}^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & \sigma_v^2/\left(1-\phi^2\right) \end{pmatrix}^{-1}$. Let J denote the selection matrix (1,0) such that $Jh_{2,t} = h_{1,t}$, $\sup_t$ denote $\sup_{t-R+1\leq\cdot\leq t}$, $\sum_t$ denote $\sum_{t=R}^{T}$, and matrix A and C will defined in Lemma 3 and $\tilde{h}_{2,t} = \sigma^{-1}A'CB_2^{1/2}h_{2,t}$, $\tilde{H}_2(t) = \sigma^{-1}A'CB_2^{1/2}H_2(t)$. U_t and $\tilde{U}_t$ are defined in Assumption 2, $\xi = R/T$. By the central limit theorem, $\sqrt{R}\left[R^{-1}\sum_{s=t-R+1}^{t} h_{2,s}\right] \sim N\left(0, \sigma^2 B_2^{-1}\right)$, where $\sigma^2 = \mathbb{E}g\left(u_t^2\right)\sigma^2$ from Newey and Powell (1987).

Proof of Proposition 5.

To prove Theorem 2, we need the following seven lemmas.

Lemma 1: (Clark and McCracken, 2001) Let Assumptions 1-3 hold. For $i = 1, 2$, $\sum_t h_{i,t}'B_i(t)H_i(t) = \sum_t h_{i,t+1}'B_iH_i(t) + o_p(1)$.

Proof: Adding and subtracting B_i, we can get

$$\sum\nolimits_t h_{i,t+1}'B_i(t)H_i(t) = \sum\nolimits_t h_{i,t}'B_iH_i(t) + \sum\nolimits_t h_{i,t}'\left(B_i(t) - B_i\right)H_i(t).$$

Now, we need to show the second term on the right-hand side is $o_p(1)$.

$$\begin{aligned}&\sum\nolimits_t h'_{i,t}\left(B_i(t)-B_i\right)H_i(t)\\ =&T^{-1/2}\sum_t\left(\frac{T}{R}\right)\text{vec}\left[T^{1/2}\left(B_i(t)-B_i\right)\right]\left[T^{-1/2}h_{i,t}\otimes\left(T^{-1/2}\sum_{j=1}^{t}h_{ij}\right)\right].\end{aligned}$$

where $\sum_t\left(\frac{T}{R}\right)\text{vec}\left[T^{1/2}\left(B_i(t)-B_i\right)\right]\left[T^{-1/2}h_{i,t}\otimes\left(T^{-1/2}\sum_{j=1}^{t}h_{ij}\right)\right]$ is $O_p(1)$ following Theorem 3.1 of Hansen (1992). Term $T^{-1/2}$ is $o(1)$.

Lemma 2: (Clark and McCracken, 2001) Let Assumptions 1-3 hold. For $i=1,2$, we have

$$\sum\nolimits_t H'_i(t)B_ih_{i,t+1}h'_{i,t+1}B_jH'_j(t)=\sum\nolimits_t H'_i(t)B_i\mathbb{E}(h_{i,t+1}h'_{i,t+1})B_jH'_j(t)+o_p(t).$$

Proof: Adding and subtracting $\mathbb{E}(h_{i,t+1}h'_{i,t+1})$,

$$\begin{aligned}&\sum\nolimits_t H'_i(t)B_ih_{i,t+1}h'_{i,t+1}B_jH_j(t)\\ &=\sum\nolimits_t H'_i(t)B_i\mathbb{E}(h_{i,t+1}h'_{i,t+1})B_jH_j(t)+\sum H'_i(t)B_i\left(h_{i,t+1}h'_{i,t+1}-\mathbb{E}(h_{i,t+1}h'_{i,t+1})\right)B_jH_j(t).\end{aligned}$$

Now, we need to show the second term is $o_p(1)$.

$$\begin{aligned}&\sum\nolimits_t H'_i(t)B_i\left(h_{i,t+1}h'_{i,t+1}-\mathbb{E}(h_{i,t+1}h'_{i,t+1})\right)B_jH_j(t)\\ =&T^{-1/2}\sum_t\left(\frac{T}{R}\right)^2\left[T^{-1/2}\sum_{s=1}^{t}h'_{j,s}B_j\otimes T^{-1/2}\sum_{s=1}^{t}h'_{i,s}B_i\right]\\ &\text{vec}\left[T^{-1/2}\left(h_{i,t+1}h'_{i,t+1}-\mathbb{E}(h_{i,t+1}h'_{i,t+1})\right)\right].\end{aligned}$$

$\sum_t\left(\frac{T}{R}\right)^2\left[\frac{1}{\sqrt{T}}\sum_{s=1}^{t}h'_{j,s}B_j\otimes\frac{1}{\sqrt{T}}\sum_{s=1}^{t}h'_{i,s}B_i\right]\text{vec}\left[\frac{1}{\sqrt{T}}\left(h_{i,t+1}h'_{i,t+1}-\mathbb{E}(h_{i,t+1}h'_{i,t+1})\right)\right]$ is $O_p\left(1\right)$, which follows from Assumption 2, Corollary 29.19 of Davidson (1994) and Theorem 3.1 of Hansen (1992). $T^{-1/2}$ is $o(1)$. The proof is complete.

Lemma 3: Let $M = -JB_1J + B_2$. We can verify that $Q \equiv B_2^{-1/2} M B_2^{-1/2} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$, which is an idempotent matrix with rank 1. Define matrix $A = (0,1)'$ and $C{=}I_{2\times 2}$ Then $Q = CAA'C$. We have the following result

$$\textbf{(a).}\ T^{-1/2} \sum_{j=t-R+1}^{t} \tilde{h}_{2,j}\tilde{h}'_{2,j} \to I_{1\times 1}$$

$$\textbf{(b).}\ T^{-1/2} \sum_{j=t-R+1}^{t} \tilde{h}_{2,j} \to \xi^{-1}\left[W(s) - W(s-\xi)\right]$$

Proof: Note that $\mathbb{E}\left(\tilde{h}_{2,j}\right) = \mathbb{E}\left(\sigma^{-1}A'CB_2^{1/2}h_{2,t}\right) = 0$ then

$$\begin{aligned}
&Var\left(T^{-1/2}\sum_{j=t-R+1}^{t}\tilde{h}_{2,j}\tilde{h}'_{2,j}\right)\\
=&T^{-1}\left(\sigma^{-1}A'CB_2^{1/2}\right)\left(\sum_{j=t-R+1}^{t} Var\left(h_{2,j}h'_{2,j}\right)\right)\left(\sigma^{-1}A'CB_2^{1/2}\right)'\\
=&\sigma^{-2}\left(A'CB_2^{1/2}\right)\left(\sigma^2 B_2^{-1}\right)\left(A'CB_2^{1/2}\right)'\\
=&A'CB_2^{1/2}B_2^{-1}B_2^{1/2}C'A\\
=&A'CC'A = Tr\left(C'AA'C\right) = Tr\left(Q\right) = I_{1\times 1}.
\end{aligned}$$

and

$$\begin{aligned}
& T^{-1/2}\sum_{j=t-R+1}^{t}\tilde{h}_{2,j}\\
= \quad & T^{-1/2}\sum_{j=1}^{t}\tilde{h}_{2,j} - T^{-1/2}\sum\nolimits_{j=1}^{t-R}\tilde{h}_{2,j}\\
\Longrightarrow \quad & W(s) - W(s-\xi).
\end{aligned}$$

Lemma 4: (Clark and McCracken, 2001) Let Assumptions 1-4 hold, then

$$\sum\nolimits_t \tilde{H}_2'(t)\tilde{h}_{2,t+1} \xrightarrow{d} \xi^{-1} \int_{\xi}^{1} \left[W(s) - W(s-\xi)\right] \mathrm{d}W(s).$$

Proof:

$$\begin{aligned}
\sum\nolimits_t \tilde{H}_2'(t)\tilde{h}_{2,t+1} &= \sum\nolimits_t \left(\frac{1}{R} \sum_{j=t-R+1}^{t} \tilde{h}_{2,j}' \right) \tilde{h}_{2,t+1} \\
&= \sum\nolimits_t \left(\frac{1}{R} \sum_{j=1}^{t} \tilde{h}_{2,j}' - \frac{1}{R} \sum_{j=1}^{t-R} \tilde{h}_{2,j}' \right) \tilde{h}_{2,t+1} \\
&= \frac{T}{R} \left[\sum\nolimits_t \left(T^{-1/2} \sum_{j=1}^{t} \tilde{h}_{2,j}' - T^{-1/2} \sum_{j=1}^{t-R} \tilde{h}_{2,j}' \right) \left(T^{-1/2} \tilde{h}_{2,t+1} \right) \right].
\end{aligned}$$

According to the continuous mapping theorem and Corollary 29.19 of Davidson (1994),

$T^{-1/2} \sum_{j=1}^{t} \tilde{h}_{2,j}' \Longrightarrow W(s)$, $\quad T^{-1/2} \sum_{j=1}^{t-R} \tilde{h}_{2,j}' \Longrightarrow W(s-\xi)$ and $T^{-1/2} \tilde{h}_{2,t+1} \Longrightarrow \mathrm{d}W(s)$.

Since $\left(\frac{T}{R}\right) = \xi^{-1}$, the proof is complete.

Lemma 5: (Clark and McCracken, 2001) Let Assumptions 1-4 hold, then

$$\sum\nolimits_t \tilde{H}_2'(t)\tilde{H}_2(t) \xrightarrow{d} \xi^{-2} \int_{\xi}^{1} \left[W(s) - W(s-\xi)\right]^2 \mathrm{d}s$$

Proof:

$$\begin{aligned}\sum\nolimits_t \tilde{H}_2'(t)\tilde{H}_2(t) &= \sum\nolimits_t \left(\frac{1}{R}\sum_{j=t-R+1}^{t}\tilde{h}_{2,j}'\right)'\left(\frac{1}{R}\sum_{j=t-R+1}^{t}\tilde{h}_{2,j}'\right)\\ &= \sum\nolimits_t \left(\frac{1}{R}\sum_{j=1}^{t}\tilde{h}_{2,j}' - \frac{1}{R}\sum_{j=1}^{t-R}\tilde{h}_{2,j}'\right)'\left(\frac{1}{R}\sum_{j=1}^{t}\tilde{h}_{2,j}' - \frac{1}{R}\sum_{j=1}^{t-R}\tilde{h}_{2,j}'\right)\\ &= \frac{1}{T}\left(\frac{T}{R}\right)^2\sum\nolimits_t \left(T^{-1/2}\sum_{j=1}^{t}\tilde{h}_{2,j}' - T^{-1/2}\sum_{j=1}^{t-R}\tilde{h}_{2,j}'\right)'\\ &\times\left(T^{-1/2}\sum_{j=1}^{t}\tilde{h}_{2,j}' - T^{-1/2}\sum_{j=1}^{t-R}\tilde{h}_{2,j}'\right).\end{aligned}$$

According to the continuous mapping theorem and Corollary 29.19 of Davidson (1994), and Lemma 4, the proof is complete.

Lemma 6: Let Assumptions 1-4 hold. $\sum_t g\left(\hat{u}_{t+1}^{(1)}\right)\hat{u}_{t+1}^{(1)}\left(\hat{u}_{t+1}^{(1)} - \hat{u}_{t+1}^{(2)}\right) = \sigma^2\sum_t \tilde{H}_2'(t)\tilde{h}_{2,t+1} + o_p(1)$.

Proof: We have $\hat{u}_{t+1}^{(1)} = u_{t+1} - x_{1,t}'\left(\hat{\beta}_{1,t} - \beta_1\right)$ and $\hat{u}_{t+1}^{(1)} = u_{t+1} - x_{2,t+1}'\left(\hat{\beta}_{2,t} - \beta_2\right)$, hence $\hat{u}_{t+1}^{(1)} - \hat{u}_{t+1}^{(2)} = x_{2,t}'\left(\hat{\beta}_{2,t} - \beta_2\right) - x_{1,t}'\left(\hat{\beta}_{1,t} - \beta_1\right)$. According to Newey and Powell (1987), ALS is differentiable and convey in u, with $\partial[g(u)\cdot u^2]/\partial u = 2g(u)\cdot u$. Then, by using the

Taylor expansion, we can obtain

$$\begin{aligned}
&\textstyle\sum_t g\left(\hat{u}_{t+1}^{(1)}\right)\hat{u}_{t+1}^{(1)}\left(\hat{u}_{t+1}^{(1)}-\hat{u}_{t+1}^{(2)}\right)\\
&=\textstyle\sum_t g\left(\hat{u}_{t+1}^{(1)}\right)\hat{u}_{t+1}^{(1)}\left[x_{2,t}'\left(\hat{\beta}_{2,t}-\beta_2\right)-x_{1,t}'\left(\hat{\beta}_{1,t}-\beta_1\right)\right]\\
&=\underbrace{\textstyle\sum_t g\left(u_{t+1}^{(1)}\right)u_{t+1}^{(1)}\left[x_{2,t}'\left(\hat{\beta}_{2,t}-\beta_2\right)-x_{1,t}'\left(\hat{\beta}_{1,t}-\beta_1\right)\right]}_{G_1}\\
&\quad+\underbrace{\textstyle\sum_t g\left(u_{t+1}^{(1)}\right)\left(\hat{u}_{t+1}^{(1)}-u_{t+1}^{(1)}\right)\left[x_{2,t}'\left(\hat{\beta}_{2,t}-\beta_2\right)-x_{1,t}'\left(\hat{\beta}_{1,t}-\beta_1\right)\right]}_{G_2}\\
&\quad+o_p\left(1\right).
\end{aligned}$$

Now, we want to show that terms G_2 have lower order of magnitude than term G_1.

$$\begin{aligned}
G_1&=\sum_t g\left(u_{t+1}^{(1)}\right)u_{t+1}^{(1)}\left[x_{2,t}'\left(\hat{\beta}_{2,t}-\beta_{2,t}\right)-x_{1,t}'\left(\hat{\beta}_{1,t}-\beta_{1,t}\right)\right]=PO_p\left(1\right)O_p\left(\frac{1}{\sqrt{R}}\right)\\
&=O_p\left(\frac{P}{\sqrt{R}}\right).
\end{aligned}$$

$$\begin{aligned}
G_2&=\sum_t g\left(u_{t+1}^{(1)}\right)\left(\hat{u}_{t+1}^{(1)}-u_{t+1}^{(1)}\right)\left[x_{2,t}'\left(\hat{\beta}_{2,t}-\beta_{2,t}\right)-x_{1,t}'\left(\hat{\beta}_{1,t}-\beta_{1,t}\right)\right]\\
&=PO_p\left(1\right)O_p\left(\frac{1}{\sqrt{R}}\right)O_p\left(\frac{1}{\sqrt{R}}\right)\\
&=O_p\left(\frac{P}{R}\right).
\end{aligned}$$

Therefore, G_2 has lower order than G_1.

G_1 can be shown that

$$\begin{aligned}
G_1 &= \sum\nolimits_t g\left(u_{t+1}^{(1)}\right) u_{t+1}^{(1)}\left[x_{2,t}'\left(\hat{\beta}_{2,t}-\beta_2\right)-x_{1,t}'\left(\hat{\beta}_{1,t}-\beta_1\right)\right] \\
&= \sum\nolimits_t g\left(u_{t+1}^{(1)}\right) u_{t+1}^{(1)} x_{2,t}'\left(\hat{\beta}_{2,t}-\beta_2\right)-g\left(u_{t+1}^{(1)}\right) u_{t+1}^{(1)} x_{1,t}'\left(\hat{\beta}_{1,t}-\beta_1\right) \\
&= \sum\nolimits_t \left[-h_{1,t+1}' B_1(t) H_1(t)+h_{2,t+1}' B_2(t) H_2(t)\right] \\
&= \sum\nolimits_t \left[-h_{1,t+1}' B_1 H_1(t)+h_{2,t+1}' B_2 H_2(t)\right]+o_p(1) \\
&= \sum\nolimits_t \left[-h_{2,t+1}' J' B_1 J H_2(t)+h_{2,t+1}' B_2 H_2(t)\right]+o_p(1) \\
&= \sum\nolimits_t \left[h_{2,t+1}' M H_2(t)\right]+o_p(1) \\
&= \sum\nolimits_t \left[h_{2,t+1}' B_2^{1/2} B_2^{-1/2} M B_2^{-1/2} B_2^{1/2} H_2(t)\right]+o_p(1) \\
&= \sum\nolimits_t \left[h_{2,t+1}' B_2^{1/2} Q B_2^{1/2} H_2(t)\right]+o_p(1) \\
&= \sum\nolimits_t \left[h_{2,t+1}' B_2^{1/2} C A A' C B_2^{1/2} H_2(t)\right]+o_p(1) \\
&= \sigma^2 \sum\nolimits_t \tilde{H}_2'(t) \tilde{h}_{2,t+1}+o_p(1).
\end{aligned}$$

Thus, $\sum_t g\left(\hat{u}_{t+1}^{(1)}\right) \hat{u}_{t+1}^{(1)}\left(\hat{u}_{t+1}^{(1)}-\hat{u}_{t+1}^{(2)}\right)=G_1+G_2+o_p(1)$. G_2 has lower order than G_1 and $G_1=\sigma^2 \sum_t \tilde{H}_2'(t) \tilde{h}_{2,t+1}+o_p(1)$. $\sum_t g\left(\hat{u}_{t+1}^{(1)}\right) \hat{u}_{t+1}^{(1)}\left(\hat{u}_{t+1}^{(1)}-\hat{u}_{t+1}^{(2)}\right)=\sigma^2 \sum_t \tilde{H}_2'(t) \tilde{h}_{2,t+1}+o_p(1)$. The proof is completed.

Lemma 7: Let Assumptions 1-4 hold. Then $\sum_t\left[g\left(\hat{u}_{t+1}^{(1)}\right)\left(\hat{u}_{t+1}^{(1)}-\hat{u}_{t+1}^{(2)}\right)\right]^2-P\bar{C}^2=\sigma^4 \sum_t \tilde{H}_2'(t) \tilde{H}_2(t)+o_p(1)$.

Proof:

$$
\begin{aligned}
&\sum_t \left[g\left(\hat{u}_{t+1}^{(1)}\right) \hat{u}_{t+1}^{(1)} \left(\hat{u}_{t+1}^{(1)} - \hat{u}_{t+1}^{(2)}\right)\right]^2 \\
&= \sum_t \left\{ g\left(\hat{u}_{t+1}^{(1)}\right) \hat{u}_{t+1}^{(1)} \left[x_{2,t}' \left(\hat{\beta}_{2,t} - \beta_{2,t}\right) - x_{1,t}' \left(\hat{\beta}_{1,t} - \beta_{1,t}\right)\right]\right\}^2 \\
&= \sum_t \left\{ \left[g\left(u_{t+1}'\right) u_{t+1}^{(1)} + g\left(\hat{u}_{t+1}^{(1)}\right) \hat{u}_{t+1}^{(1)} - g\left(u_{t+1}^{(1)}\right) u_{t+1}^{(1)}\right]\right. \\
&\quad \left. \times \left[x_{2,t}' \left(\hat{\beta}_{2,t} - \beta_{2,t}\right) - x_{1,t}' \left(\hat{\beta}_{1,t} - \beta_{1,t}\right)\right]\right\}^2 \\
&= \sum_t \left[-h_{1,t+1}' B_1(t) H_1(t) + h_{2,t+1}' B_2(t) H_2(t)\right]^2 + o_p(1). \\
&= \sum_t H_1'(t) B_1'(t) h_{1,t+1} h_{1,t+1}' B_1(t) H_1(t) + \sum_t H_2'(t) B_2'(t) h_{2,t+1} h_{2,t+1}' B_2(t) H_2(t) \\
&\quad - 2 \sum_t H_1'(t) B_1'(t) h_{1,t+1} h_{2,t+1} B_2(t) H_2(t) + o_p(1) \\
&= \sum_t H_1'(t) B_1'(t) \mathbb{E}\left(h_{1,t+1} h_{1,t+1}'\right) B_1(t) H_1(t) + \sum_t H_2'(t) B_2'(t) \mathbb{E}\left(h_{2,t+1} h_{2,t+1}'\right) B_2(t) H_2(t) \\
&\quad - 2 \sum_t H_1'(t) B_1'(t) \mathbb{E}\left(h_{1,t+1} h_{2,t+1}'\right) B_2(t) H_2(t) + o_p(1) \\
&= \sigma^2 \sum_t H_2'(t) J' B_1 J H_2(t) + \sigma^2 \sum_t H_2'(t) B_2 H_2(t) - 2\sigma^2 \sum_t H_2'(t) J' B_1 J H_2(t) + o_p(1) \\
&= \sigma^2 \sum_t H_2'(t) \left[-J' B_1 J + B_2\right] H_2(t) + o_p(1) \\
&= \sigma^2 \sum_t H_2'(t) M H_2(t) + o_p(1) \\
&= \sigma^2 \sum_t H_2'(t) B_2^{1/2} B_2^{-1/2} M B_2^{-1/2} B_2^{1/2} H_2(t) + o_p(1) \\
&= \sigma^2 \sum_t H_2'(t) B_2^{1/2} C A A' C B_2^{1/2} H_2(t) + o_p(1) \\
&= \sigma^4 \sum_t \tilde{H}_2'(t) \tilde{H}_2(t) + o_p(1).
\end{aligned}
$$

since $\bar{C} = O_p(P^{-1})$, therefore $P\bar{C}^2 = o_p(1)$. The proof is complete.

Combining Lemmas 1-7,

$$\begin{aligned}
\mathrm{ENC}_P &= \frac{\sum_{t=R}^{T} g\left(\hat{u}_{t+1}^{(1)}\right) \hat{u}_{t+1}^{(1)} \left(\hat{u}_{t+1}^{(1)} - \hat{u}_{t+1}^{(2)}\right)}{\sqrt{\sum_{t=R}^{T} \left[g\left(\hat{u}_{t+1}^{(1)}\right) \hat{u}_{t+1}^{(1)} \left(\hat{u}_{t+1}^{(1)} - \hat{u}_{t+1}^{(2)}\right)\right]^2 - P\bar{C}^2}} \\
&= \frac{\sigma^2 \sum_t \tilde{H}_2'(t)\tilde{h}_{2,t+1}}{\sqrt{\sigma^4 \sum_t \tilde{H}_2'(t)\tilde{H}_2(t)}} \\
&\xrightarrow{d} \frac{\int_\xi^1 \left[W(s) - W(s-\xi)\right] \mathrm{d}W(s)}{\sqrt{\int_\xi^1 \left[W(s) - W(s-\xi)\right]^2 \mathrm{d}s}}
\end{aligned}$$

under $\mathbb{H}_0$. Therefore, the proof of Proposition 5 is completed.

Proof of Proposition 6.

Proof: We divide $[0,1]$ to $n = \left[\frac{1}{\xi}\right] = \left[\frac{T}{R}\right]$ equal segments and let $t = [sn]$, where $[\cdot]$ denote the integer part and $s \in [0,1]$. Let $\{\nu_i\}_{i=1}^n$ is mixing sequence with $\mathbb{E}(\nu) = 0$ and $Var(\nu) = 1$. We denote $V_t = \sum_{i=1}^t \nu_i$ to be the partial sum, then $V_t = \sum_{i=1}^t \nu_i \sim N(0,t)$. Thus,

$$\frac{V_t}{\sqrt{n}} = \frac{\sum_{i=1}^t \nu_i}{\sqrt{n}} = V_n(s) \Rightarrow W(s),$$

where $V_n(s)$ is a CADLAG and $V_n(s)$ is a Wiener process.

$$\begin{aligned} n^{-1}\sum_{t=1}^n \nu_{t-1}\nu_t &= n^{-1}\sum_{t=1}^n V_{t-1}\nu_t - n^{-1}\sum_{t=1}^n V_{t-2}\nu_t \\ &\Rightarrow \int_\xi^1 W(s)\,\mathrm{d}W(s) - \int_\xi^1 W(s-\xi)\,dW(s) \\ &= \int_\xi^1 [W(s) - W(s-\xi)]\,\mathrm{d}W(s) \\ n^{-2}\sum_{t=1}^n \nu_{t-1}^2 &= n^{-2}\sum_{t=1}^n (V_{t-1} - V_{t-2})^2 \\ &\Rightarrow \int_\xi^1 [W(s) - W(s-\xi)]^2\,\mathrm{d}s \end{aligned}$$

We usd DGP to generate i.i.d standard normal sequence $\{\nu_t\}$ and design the following model

$$\nu_{t+1} = \zeta\nu_t + \epsilon_t.$$

The true parameter $\zeta = 0$, The estimator $\hat{\zeta} = \sum_{t=1}^n \nu_{t-1}\nu_t / \sum_{t=1}^n \nu_{t-1}^2$ and $Var(\hat{\zeta}) = (\sum_{t=1}^n \nu_{t-1}^2)^{-1} Var(\nu) = (\sum_{t=1}^n \nu_{t-1}^2)^{-1}$. By using CLT,

$$\frac{\sum_{t=1}^n u_{t-1}u_t}{\sqrt{\sum_{t=1}^n u_{t-1}^2}} \Rightarrow \frac{\int_\xi^1 [W(s) - W(s-\xi)]\,\mathrm{d}W(s)}{\sqrt{\int_\xi^1 [W(s) - W(s-\xi)]^2\,\mathrm{d}s}} \xrightarrow{d} N(0,1).$$

The proof is complete.

5.11 Appendix C: Proof of Proposition 7

The error terms $u_{t+1}^{(1)} = y_{t+1} - \mu_{t+1}^{(1)}$ and $u_{t+1}^{(2)} = y_{t+1} - \mu_{t+1}^{(2)}$, so we have the combined error term $u_{t+1}^{(c)} = (1-\lambda)\, u_{t+1}^{(1)} + \lambda u_{t+1}^{(2)}$. The expected loss function from combining forecasts is

$$\mathbb{E}S\left(\mu^{(c)}, Y\right) = \mathbb{E}\left[\text{abs}\left(\tau - \mathbf{1}(u^{(c)} < 0)\right)\left(u^{(c)}\right)^2\right]. \tag{5.32}$$

Taking the first order condition of Equation (5.32) with respect to λ, we have

$$\begin{aligned} C^{(c)} &\equiv \nabla_\lambda \mathbb{E}S(\mu^{(c)}, Y) \\ &= -\mathbb{E}\left[g\left(u^{(c)}\right) u^{(c)}\left(u^{(1)} - u^{(2)}\right)\right] = 0 \end{aligned}$$

To show the necessary condition, if $\lambda = 0$, then $\hat{u}_{t+1}^{(c)} = \hat{u}_{t+1}^{(1)}$. we have

$$\hat{C}_P^{(c)} \equiv P^{-1} \sum_{t=R}^{T} \left[g\left(\hat{u}_{t+1}^{(1)}\right) \hat{u}_{t+1}^{(1)}\left(\hat{u}_{t+1}^{(1)} - \hat{u}_{t+1}^{(2)}\right)\right] \xrightarrow{p} 0$$

as shown in Proposition 3 under $\mathbb{H}_0$. To show the sufficient condition, since

$$C^{(c)} = -\mathbb{E}_F\left[g\left(u^{(c)}\right)\left((1-\lambda)\, u^{(1)} + \lambda u^{(2)}\right)\left(u^{(1)} - u^{(2)}\right)\right]$$

taking the second order condition of Equation (5.32) with respect to λ, we have

$$\frac{\partial^2 \mathbb{E}S(\mu^{(c)}, Y)}{\partial \lambda^2} = \frac{\partial C^{(c)}}{\partial \lambda} = \mathbb{E}[g\left(u^{(c)}\right)\left(u^{(1)} - u^{(2)}\right)^2 > 0$$

Therefore there is a unique value $\hat{\lambda} \to 0$ such that $\hat{C}_P^{(c)} \to 0$.

Table 5.1: Simulation Results of DGP 1: Size of Test, $\phi = 0$

	$P = 48$			$P = 240$			$P = 1200$		
	DM_P	ENC_P	CCS_P	DM_P	ENC_P	CCS_P	DM_P	ENC_P	CCS_P
$\tau = 0.1$									
$R = 60$	0.011	0.040	0.065	0.001	0.033	0.059	0.000	0.040	0.060
$R = 120$	0.023	0.045	0.048	0.002	0.035	0.056	0.000	0.038	0.061
$R = 240$	0.033	0.047	0.057	0.003	0.034	0.053	0.007	0.030	0.062
$\tau = 0.3$									
$R = 60$	0.001	0.041	0.059	0.000	0.033	0.050	0.000	0.039	0.047
$R = 120$	0.016	0.033	0.052	0.002	0.030	0.056	0.000	0.027	0.047
$R = 240$	0.024	0.040	0.059	0.006	0.030	0.060	0.000	0.032	0.052
$\tau = 0.5$									
$R = 60$	0.010	0.036	0.050	0.002	0.030	0.049	0.000	0.048	0.047
$R = 120$	0.012	0.031	0.041	0.001	0.023	0.052	0.000	0.033	0.048
$R = 240$	0.023	0.039	0.056	0.004	0.031	0.054	0.000	0.031	0.041
$\tau = 0.7$									
$R = 60$	0.011	0.039	0.066	0.000	0.029	0.054	0.000	0.042	0.047
$R = 120$	0.022	0.030	0.056	0.001	0.028	0.044	0.000	0.033	0.052
$R = 240$	0.022	0.039	0.053	0.005	0.027	0.056	0.000	0.033	0.049
$\tau = 0.9$									
$R = 60$	0.011	0.049	0.057	0.001	0.033	0.049	0.000	0.046	0.048
$R = 120$	0.024	0.047	0.047	0.001	0.040	0.046	0.000	0.032	0.047
$R = 240$	0.027	0.043	0.052	0.005	0.033	0.042	0.000	0.034	0.047

Table 5.2: Simulation Results of DGP 1: Size of Test, $\phi = 0.95$

	$P = 48$			$P = 240$			$P = 1200$		
	DM_P	ENC_P	CCS_P	DM_P	ENC_P	CCS_P	DM_P	ENC_P	CCS_P
$\tau = 0.1$									
$R = 60$	0.009	0.042	0.056	0.000	0.036	0.039	0.000	0.045	0.034
$R = 120$	0.017	0.042	0.068	0.000	0.029	0.049	0.000	0.032	0.046
$R = 240$	0.031	0.049	0.059	0.006	0.034	0.053	0.000	0.034	0.043
$\tau = 0.3$									
$R = 60$	0.006	0.032	0.056	0.004	0.028	0.055	0.000	0.044	0.048
$R = 120$	0.016	0.039	0.056	0.002	0.031	0.045	0.000	0.037	0.052
$R = 240$	0.033	0.047	0.062	0.004	0.030	0.049	0.000	0.029	0.040
$\tau = 0.5$									
$R = 60$	0.008	0.033	0.058	0.000	0.035	0.041	0.000	0.042	0.039
$R = 120$	0.013	0.032	0.047	0.001	0.027	0.051	0.000	0.041	0.048
$R = 240$	0.023	0.035	0.049	0.002	0.032	0.049	0.000	0.034	0.046
$\tau = 0.7$									
$R = 60$	0.010	0.030	0.046	0.000	0.036	0.036	0.000	0.046	0.035
$R = 120$	0.016	0.040	0.049	0.001	0.032	0.044	0.000	0.039	0.039
$R = 240$	0.026	0.044	0.053	0.002	0.019	0.048	0.000	0.029	0.052
$\tau = 0.9$									
$R = 60$	0.009	0.042	0.055	0.000	0.039	0.043	0.000	0.043	0.040
$R = 120$	0.019	0.038	0.051	0.002	0.032	0.050	0.000	0.043	0.049
$R = 240$	0.030	0.050	0.071	0.004	0.029	0.041	0.000	0.044	0.052

Table 5.3: Simulation Results of DGP 1: Power of Test, $\phi = 0$

	$P = 48$			$P = 240$			$P = 1200$		
	DM_P	ENC_P	CCS_P	DM_P	ENC_P	CCS_P	DM_P	ENC_P	CCS_P
$\tau = 0.1$									
$R = 60$	0.035	0.114	0.065	0.011	0.260	0.051	0.002	0.727	0.048
$R = 120$	0.057	0.139	0.044	0.032	0.346	0.048	0.053	0.850	0.043
$R = 240$	0.084	0.163	0.048	0.090	0.412	0.053	0.230	0.920	0.052
$\tau = 0.3$									
$R = 60$	0.018	0.091	0.051	0.009	0.234	0.049	0.001	0.640	0.043
$R = 120$	0.047	0.104	0.049	0.021	0.271	0.043	0.021	0.756	0.037
$R = 240$	0.058	0.122	0.045	0.060	0.317	0.044	0.127	0.849	0.053
$\tau = 0.5$									
$R = 60$	0.020	0.078	0.057	0.008	0.213	0.043	0.001	0.599	0.051
$R = 120$	0.045	0.111	0.049	0.021	0.267	0.051	0.025	0.754	0.052
$R = 240$	0.053	0.119	0.049	0.060	0.304	0.054	0.124	0.837	0.050
$\tau = 0.7$									
$R = 60$	0.018	0.092	0.052	0.005	0.220	0.048	0.002	0.630	0.043
$R = 120$	0.047	0.114	0.053	0.020	0.271	0.062	0.024	0.762	0.054
$R = 240$	0.068	0.137	0.055	0.062	0.339	0.051	0.138	0.854	0.045
$\tau = 0.9$									
$R = 60$	0.036	0.110	0.043	0.010	0.286	0.041	0.002	0.721	0.047
$R = 120$	0.064	0.132	0.058	0.033	0.351	0.047	0.066	0.860	0.054
$R = 240$	0.079	0.150	0.045	0.086	0.402	0.057	0.221	0.929	0.041

Table 5.4: Simulation Results of DGP 1: Power of Test, $\phi = 0.95$

	$P = 48$			$P = 240$			$P = 1200$		
	DM_P	ENC_P	CCS_P	DM_P	ENC_P	CCS_P	DM_P	ENC_P	CCS_P
$\tau = 0.1$									
$R = 60$	0.154	0.545	0.406	0.470	0.968	0.923	0.976	1.000	1.000
$R = 120$	0.280	0.652	0.532	0.756	0.988	0.967	1.000	1.000	1.000
$R = 240$	0.304	0.657	0.560	0.843	0.996	0.983	1.000	1.000	1.000
$\tau = 0.3$									
$R = 60$	0.120	0.449	0.387	0.352	0.931	0.903	0.933	1.000	1.000
$R = 120$	0.222	0.559	0.474	0.652	0.980	0.950	0.999	1.000	1.000
$R = 240$	0.248	0.592	0.514	0.744	0.989	0.968	1.000	1.000	1.000
$\tau = 0.5$									
$R = 60$	0.095	0.451	0.398	0.320	0.920	0.883	0.916	1.000	1.000
$R = 120$	0.195	0.549	0.474	0.623	0.973	0.956	0.999	1.000	1.000
$R = 240$	0.237	0.576	0.506	0.726	0.989	0.962	1.000	1.000	1.000
$\tau = 0.7$									
$R = 60$	0.108	0.456	0.394	0.346	0.926	0.894	0.930	1.000	1.000
$R = 120$	0.217	0.552	0.470	0.653	0.981	0.959	0.999	1.000	1.000
$R = 240$	0.237	0.581	0.507	0.751	0.991	0.974	1.000	1.000	1.000
$\tau = 0.9$									
$R = 60$	0.162	0.527	0.416	0.445	0.970	0.932	0.971	1.000	1.000
$R = 120$	0.261	0.632	0.520	0.744	0.986	0.962	1.000	1.000	1.000
$R = 240$	0.308	0.647	0.543	0.837	0.995	0.981	1.000	1.000	1.000

Table 5.5: Simulation Results of DGP 1: Model averaging, $b = 0$, $\phi = 0$, $R = 60$

		$P = 48$		$P = 240$		$P = 1200$	
Repeat=2000		$\hat{\lambda}$	Loss	$\hat{\lambda}$	Loss	$\hat{\lambda}$	Loss
			0.2284[b]		0.2286		0.2290
$\tau = 0.1$	$R = 60$	-0.6525[a]	0.2335[c]	-0.1873	0.2333	-0.0445	0.2338
			0.2239[d]		0.2277		0.2288
			0.4462		0.4462		0.4456
$\tau = 0.3$	$R = 60$	-0.7400	0.4543	-0.2107	0.4541	-0.0508	0.4536
			0.4373		0.4443		0.4452
			0.5067		0.5076		0.5084
$\tau = 0.5$	$R = 60$	-0.7528	0.5156	-0.2433	0.5168	-0.0464	0.5173
			0.4968		0.5055		0.5080
			0.4451		0.4466		0.4459
$\tau = 0.7$	$R = 60$	-0.8013	0.4528	-0.2275	0.4546	-0.0526	0.4540
			0.4354		0.4447		0.4456
			0.2280		0.2288		0.2289
$\tau = 0.9$	$R = 60$	-0.7897	0.2327	-0.1919	0.2335	-0.0523	0.2336
			0.2230		0.2278		0.2287

[a]: The Monte Carlo average of the combination weight $\hat{\lambda}$

[b]: The Monte Carlo average of the ALS loss $\hat{S}_\tau^{(1)}$ from forecasting Model 1.

[c]: The Monte Carlo average of the ALS loss $\hat{S}_\tau^{(2)}$ from forecasting Model 2.

[d]: The Monte Carlo average of the ALS loss $\hat{S}_\tau^{(c)}$ from forecasting combined model.

Table 5.6: Simulation Results of DGP 1: Model averaging, $b = 0$, $\phi = 0$, $R = 120$

		$P = 48$		$P = 240$		$P = 1200$	
Repeat=2000		$\hat{\lambda}$	Loss	$\hat{\lambda}$	Loss	$\hat{\lambda}$	Loss
			0.2265		0.2267		0.2271
$\tau = 0.1$	$R = 120$	-1.1518	0.2286	-0.4080	0.2289	-0.0983	0.2293
			0.2218		0.2258		0.2270
			0.4440		0.4425		0.4415
$\tau = 0.3$	$R = 120$	-1.3720	0.4481	-0.4134	0.4463	-0.0894	0.4453
			0.4344		0.4406		0.4412
			0.5021		0.5038		0.5045
$\tau = 0.5$	$R = 120$	-1.3190	0.5069	-0.4294	0.5082	-0.0925	0.5088
			0.4916		0.5018		0.5041
			0.4429		0.4419		0.4418
$\tau = 0.7$	$R = 120$	-1.2740	0.4466	-0.4236	0.4457	-0.0854	0.4456
			0.4332		0.4400		0.4415
			0.2264		0.2266		0.2272
$\tau = 0.9$	$R = 120$	-1.2162	0.2286	-0.4160	0.2287	-0.1069	0.2293
			0.2215		0.2256		0.2270

Table 5.7: Simulation Results of DGP 1: Model averaging, $b = 0$, $\phi = 0$, $R = 240$

		$P = 48$		$P = 240$		$P = 1200$	
Repeat=2000		$\hat{\lambda}$	Loss	$\hat{\lambda}$	Loss	$\hat{\lambda}$	Loss
			0.2253		0.2260		0.2259
$\tau = 0.1$	$R = 240$	-1.6764	0.2262	-0.6372	0.2270	-0.1991	0.2269
			0.2205		0.2252		0.2257
			0.4421		0.4392		0.4403
$\tau = 0.3$	$R = 240$	-2.0478	0.4439	-0.7630	0.4411	-0.1833	0.4421
			0.4326		0.4374		0.4399
			0.5028		0.5032		0.5020
$\tau = 0.5$	$R = 240$	-2.0518	0.5051	-0.6825	0.5052	-0.1868	0.5041
			0.4917		0.5012		0.5016
			0.4412		0.4407		0.4399
$\tau = 0.7$	$R = 240$	-2.1069	0.4432	-0.7623	0.4426	-0.1927	0.4418
			0.4319		0.4389		0.4395
			0.2258		0.2258		0.2261
$\tau = 0.9$	$R = 240$	-1.8714	0.2270	-0.7180	0.2268	-0.1674	0.2271
			0.2210		0.2248		0.2259

Table 5.8: Simulation Results of DGP 1: Model averaging, $b = 0$, $\phi = 0.95$, $R = 60$

		$P = 48$		$P = 240$		$P = 1200$	
Repeat=2000		$\hat{\lambda}$	Loss	$\hat{\lambda}$	Loss	$\hat{\lambda}$	Loss
			0.2284		0.2285		0.2291
$\tau = 0.1$	$R = 60$	-0.6725	0.2357	-0.1552	0.2358	-0.0407	0.2365
			0.2235		0.2275		0.2289
			0.4463		0.4455		0.4453
$\tau = 0.3$	$R = 60$	-0.6920	0.4594	-0.1613	0.4578	-0.0314	0.4577
			0.4370		0.4436		0.4449
			0.5111		0.5077		0.5081
$\tau = 0.5$	$R = 60$	-0.6280	0.5249	-0.1648	0.5213	-0.0356	0.5218
			0.5009		0.5055		0.5076
			0.4498		0.4457		0.4456
$\tau = 0.7$	$R = 60$	-0.6714	0.4615	-0.1631	0.4580	-0.0345	0.4580
			0.4406		0.4438		0.4452
			0.2291		0.2287		0.2289
$\tau = 0.9$	$R = 60$	-0.5181	0.2362	-0.1482	0.2360	-0.0464	0.2362
			0.2244		0.2277		0.2287

Table 5.9: Simulation Results of DGP 1: Model averaging, $b = 0$, $\phi = 0.95$, $R = 120$

		$P = 48$		$P = 240$		$P = 1200$	
Repeat=2000		$\hat{\lambda}$	Loss	$\hat{\lambda}$	Loss	$\hat{\lambda}$	Loss
			0.2275		0.2267		0.2272
$\tau = 0.1$	$R = 120$	-1.1864	0.2302	-0.3609	0.2296	-0.0794	0.2301
			0.2226		0.2257		0.2270
			0.4417		0.4425		0.4420
$\tau = 0.3$	$R = 120$	-1.1866	0.4465	-0.3475	0.4475	-0.0816	0.4473
			0.4322		0.4407		0.4417
			0.5058		0.5031		0.5042
$\tau = 0.5$	$R = 120$	-1.3805	0.5124	-0.3811	0.5091	-0.0811	0.5102
			0.4946		0.5010		0.5038
			0.4419		0.4424		0.4417
$\tau = 0.7$	$R = 120$	-1.3875	0.4474	-0.3378	0.4476	-0.0764	0.4470
			0.4325		0.4407		0.4414
			0.2274		0.2270		0.2271
$\tau = 0.9$	$R = 120$	-1.1522	0.2301	-0.3270	0.2299	-0.0785	0.2300
			0.2224		0.2261		0.2269

Table 5.10: Simulation Results of DGP 1: Model averaging, $b = 0$, $\phi = 0.95$, $R = 240$

Repeat=2000		$P = 48$ $\hat{\lambda}$	$P = 48$ Loss	$P = 240$ $\hat{\lambda}$	$P = 240$ Loss	$P = 1200$ $\hat{\lambda}$	$P = 1200$ Loss
			0.2256		0.2258		0.2261
$\tau = 0.1$	$R = 240$	-2.2661	0.2267	-0.6639	0.2270	-0.1721	0.2273
			0.2205		0.2248		0.2259
			0.4417		0.4400		0.4399
$\tau = 0.3$	$R = 240$	-2.1057	0.4439	-0.6642	0.4423	-0.1638	0.4422
			0.4325		0.4383		0.4396
			0.5011		0.5026		0.5024
$\tau = 0.5$	$R = 240$	-2.3348	0.5036	-0.6886	0.5052	-0.1659	0.5050
			0.4905		0.5006		0.5020
			0.4414		0.4402		0.4400
$\tau = 0.7$	$R = 240$	-2.3768	0.4435	-0.6757	0.4424	-0.1644	0.4422
			0.4317		0.4384		0.4396
			0.2258		0.2260		0.2257
$\tau = 0.9$	$R = 240$	-2.4199	0.2270	-0.7637	0.2272	-0.1599	0.2269
			0.2206		0.2250		0.2255

Table 5.11: Simulation Results of DGP 1: Model averaging, $b = 0.1$, $\phi = 0$, $R = 60$

Repeat=2000		$P = 48$ $\hat{\lambda}$	$P = 48$ Loss	$P = 240$ $\hat{\lambda}$	$P = 240$ Loss	$P = 1200$ $\hat{\lambda}$	$P = 1200$ Loss
			0.2325		0.2319		0.2319
$\tau = 0.1$	$R = 60$	-0.1921	0.2345	0.2255	0.2337	0.3677	0.2335
			0.2263		0.2299		0.2305
			0.4505		0.4502		0.4505
$\tau = 0.3$	$R = 60$	-0.3014	0.4535	0.1579	0.4539	0.3328	0.4538
			0.4392		0.4468		0.4484
			0.5136		0.5129		0.5136
$\tau = 0.5$	$R = 60$	-0.2543	0.5171	0.1532	0.5172	0.3244	0.5174
			0.5010		0.5093		0.5114
			0.4510		0.4494		0.4504
$\tau = 0.7$	$R = 60$	-0.2811	0.4544	0.1646	0.4530	0.3287	0.4538
			0.4398		0.4460		0.4483
			0.2321		0.2318		0.2319
$\tau = 0.9$	$R = 60$	-0.2637	0.2337	0.2237	0.2336	0.3613	0.2336
			0.2257		0.2298		0.2305

Table 5.12: Simulation Results of DGP 1: Model averaging, $b = 0.1$, $\phi = 0$, $R = 120$

		$P = 48$		$P = 240$		$P = 1200$	
Repeat=2000		$\hat{\lambda}$	Loss	$\hat{\lambda}$	Loss	$\hat{\lambda}$	Loss
			0.2303		0.2299		0.2300
$\tau = 0.1$	$R = 120$	-0.0564	0.2295	0.4101	0.2289	0.5469	0.2292
			0.2235		0.2274		0.2281
			0.4474		0.4474		0.4463
$\tau = 0.3$	$R = 120$	-0.1905	0.4469	0.3015	0.4464	0.4946	0.4456
			0.4359		0.4432		0.4436
			0.5099		0.5097		0.5092
$\tau = 0.5$	$R = 120$	-0.2676	0.5094	0.2700	0.5092	0.4839	0.5085
			0.4965		0.5055		0.5062
			0.4470		0.4470		0.4470
$\tau = 0.7$	$R = 120$	-0.3539	0.4467	0.2906	0.4461	0.4929	0.4462
			0.4355		0.4429		0.4442
			0.2309		0.2303		0.2298
$\tau = 0.9$	$R = 120$	-0.0686	0.2298	0.3716	0.2294	0.5476	0.2289
			0.2239		0.2278		0.2279

Table 5.13: Simulation Results of DGP 1: Model averaging, $b = 0.1$, $\phi = 0$, $R = 240$

		$P = 48$		$P = 240$		$P = 1200$	
Repeat=2000		$\hat{\lambda}$	Loss	$\hat{\lambda}$	Loss	$\hat{\lambda}$	Loss
			0.2292		0.2289		0.2289
$\tau = 0.1$	$R = 240$	0.5100	0.2271	0.5971	0.2269	0.7230	0.2268
			0.2219		0.2259		0.2265
			0.4454		0.4442		0.4448
$\tau = 0.3$	$R = 240$	0.0719	0.4427	0.4888	0.4413	0.6585	0.4421
			0.4332		0.4394		0.4414
			0.5071		0.5070		0.5075
$\tau = 0.5$	$R = 240$	-0.0061	0.5038	0.4555	0.5041	0.6474	0.5046
			0.4927		0.5018		0.5038
			0.4436		0.4458		0.4450
$\tau = 0.7$	$R = 240$	0.2021	0.4411	0.4972	0.4429	0.6552	0.4423
			0.4315		0.4409		0.4416
			0.2294		0.2291		0.2293
$\tau = 0.9$	$R = 240$	0.4854	0.2274	0.5395	0.2271	0.7141	0.2272
			0.2225		0.2261		0.2269

Table 5.14: Simulation Results of DGP 1: Model averaging, $b = 0.1$, $\phi = 0.95$, $R = 60$

		$P = 48$		$P = 240$		$P = 1200$	
Repeat=2000		$\hat{\lambda}$	Loss	$\hat{\lambda}$	Loss	$\hat{\lambda}$	Loss
			0.2539		0.2547		0.2550
$\tau = 0.1$	$R = 60$	0.6624	0.2359	0.7586	0.2360	0.7779	0.2364
			0.2297		0.2337		0.2347
			0.4849		0.4857		0.4852
$\tau = 0.3$	$R = 60$	0.5710	0.4597	0.7275	0.4582	0.7548	0.4580
			0.4495		0.4543		0.4548
			0.5499		0.5522		0.5510
$\tau = 0.5$	$R = 60$	0.5708	0.5219	0.7178	0.5223	0.7521	0.5216
			0.5096		0.5179		0.5181
			0.4832		0.4865		0.4858
$\tau = 0.7$	$R = 60$	0.5644	0.4576	0.7241	0.4594	0.7579	0.4583
			0.4464		0.4554		0.4552
			0.2550		0.2544		0.2546
$\tau = 0.9$	$R = 60$	0.6728	0.2363	0.7595	0.2358	0.7771	0.2361
			0.2305		0.2336		0.2343

Table 5.15: Simulation Results of DGP 1: Model averaging, $b = 0.1$, $\phi = 0.95$, $R = 120$

		$P = 48$		$P = 240$		$P = 1200$	
Repeat=2000		$\hat{\lambda}$	Loss	$\hat{\lambda}$	Loss	$\hat{\lambda}$	Loss
			0.2554		0.2557		0.2562
$\tau = 0.1$	$R = 120$	0.9327	0.2303	0.9072	0.2297	0.9077	0.2300
			0.2252		0.2290		0.2296
			0.4858		0.4874		0.4870
$\tau = 0.3$	$R = 120$	0.8898	0.4476	0.8890	0.4472	0.8930	0.4474
			0.4384		0.4459		0.4467
			0.5530		0.5536		0.5541
$\tau = 0.5$	$R = 120$	0.8914	0.5121	0.8794	0.5105	0.8897	0.5102
			0.5024		0.5089		0.5095
			0.4867		0.4859		0.4869
$\tau = 0.7$	$R = 120$	0.9188	0.4462	0.8852	0.4463	0.8935	0.4471
			0.4373		0.4449		0.4464
			0.2572		0.2563		0.2561
$\tau = 0.9$	$R = 120$	0.9386	0.2305	0.9045	0.2302	0.9085	0.2300
			0.2257		0.2295		0.2297

Table 5.16: Simulation Results of DGP 1: Model averaging, $b = 0.1$, $\phi = 0.95$, $R = 120$

		$P = 48$		$P = 240$		$P = 1200$	
Repeat=2000		$\hat{\lambda}$	Loss	$\hat{\lambda}$	Loss	$\hat{\lambda}$	Loss
			0.2565		0.2559		0.2560
$\tau = 0.1$	$R = 240$	1.0065	0.2278	0.9744	0.2267	0.9614	0.2272
			0.2230		0.2260		0.2271
			0.4858		0.4870		0.4863
$\tau = 0.3$	$R = 240$	0.9707	0.4440	0.9664	0.4429	0.9525	0.4423
			0.4344		0.4415		0.4421
			0.5523		0.5527		0.5528
$\tau = 0.5$	$R = 240$	1.0004	0.5044	0.9527	0.5051	0.9523	0.5048
			0.4946		0.5037		0.5046
			0.4840		0.4868		0.4861
$\tau = 0.7$	$R = 240$	0.9891	0.4398	0.9656	0.4423	0.9533	0.4420
			0.4308		0.4411		0.4418
			0.2532		0.2571		0.2560
$\tau = 0.9$	$R = 240$	1.0038	0.2254	0.9622	0.2273	0.9608	0.2271
			0.2206		0.2266		0.2270

Table 5.17: Encompassing result from Welch-Goyal Study (Annual Data, Up to 2019) - I

	Forecasts begin 20 years after sample							
τ	0.01	0.05	0.1	0.3	0.5	0.7	0.9	0.95
d/p								
DM_p	-1.9120	-1.5970	-1.0028	-0.1077	0.0303	0.0139	-0.0558	-0.2326
p-value	0.9721	0.9449	0.8420	0.5429	0.4879	0.4944	0.5222	0.5920
ENC_p	-0.3404	-0.3176	0.6706	1.9949	2.2134	2.1946	1.8392	1.6677
p-value	0.7335	0.7508	0.5025	0.0461	0.0269	0.0282	0.0659	0.0954
CCS_p	2.0655	1.4429	0.9331	0.2299	-0.0220	-0.2612	-0.6827	-0.9482
p-value	0.0389	0.1490	0.3508	0.8181	0.9825	0.7940	0.4948	0.3430
d/y								
DM_p	-1.5520	-1.0934	-0.4092	0.2700	0.0256	-0.3747	-1.0061	-1.0854
p-value	0.9397	0.8629	0.6588	0.3936	0.4898	0.6461	0.8428	0.8611
ENC_p	2.4899	1.9611	2.4605	2.5579	2.1918	1.4900	0.2818	0.0156
p-value	0.0128	0.0499	0.0139	0.0105	0.0284	0.1362	0.7781	0.9876
CCS_p	2.0726	1.4657	0.9594	0.2309	-0.0535	-0.3274	-0.7932	-1.0665
p-value	0.0382	0.1427	0.3373	0.8174	0.9574	0.7434	0.4277	0.2862
e/p								
DM_p	-1.5677	-0.9948	-0.5494	-0.6802	-1.1550	-1.2935	-0.6998	-0.2196
p-value	0.9415	0.8401	0.7086	0.7518	0.8760	0.9021	0.7580	0.5869
ENC_p	-0.4023	0.7720	0.9560	0.5602	0.0736	0.1318	1.2892	1.8470
p-value	0.6875	0.4401	0.3391	0.5754	0.9413	0.8952	0.1973	0.0647
CCS_p	2.0511	1.4196	0.9018	0.1918	-0.0641	-0.3254	-0.7845	-1.0236
p-value	0.0403	0.1557	0.3672	0.8479	0.9489	0.7449	0.4328	0.3060

Table 5.18: Encompassing result from Welch-Goyal Study (Annual Data, Up to 2019) - II

	Forecasts begin 20 years after sample							
τ	0.01	0.05	0.1	0.3	0.5	0.7	0.9	0.95
d/e								
DM_p	-1.3756	-1.0404	-0.8809	-0.9078	-1.2490	-1.3390	-1.1824	-1.2478
p-value	0.9155	0.8509	0.8108	0.8180	0.8942	0.9097	0.8815	0.8940
ENC_p	-0.4359	0.3874	0.5182	0.5839	0.4041	0.3304	0.5193	0.0413
p-value	0.6629	0.6985	0.6043	0.5593	0.6861	0.7411	0.6036	0.9671
CCS_p	1.7686	1.3303	0.9335	0.3604	0.1654	0.0768	-0.0195	-0.2732
p-value	0.0770	0.1834	0.3506	0.7185	0.8686	0.9388	0.9844	0.7847
svar								
DM_p	-1.6303	-1.6518	-1.5999	-1.4428	-1.2904	-1.1761	-1.1609	-1.1644
p-value	0.9485	0.9507	0.9452	0.9255	0.9015	0.8802	0.8772	0.8779
ENC_p	-0.6297	-0.2792	-0.5570	-0.5228	-0.3700	-0.3929	-0.5931	-0.6561
p-value	0.5289	0.7801	0.5775	0.6011	0.7114	0.6944	0.5531	0.5118
CCS_p	-1.4252	-0.7901	-0.2589	0.4962	0.6580	0.8308	1.0340	1.0934
p-value	0.1541	0.4295	0.7957	0.6197	0.5106	0.4061	0.3011	0.2742
infl								
DM_p	0.2274	0.2449	0.0073	-0.4288	-0.7306	-0.8375	-0.8615	-0.9065
p-value	0.4100	0.4033	0.4971	0.6660	0.7675	0.7988	0.8055	0.8177
ENC_p	0.9211	1.5327	1.5182	0.8051	0.0885	-0.3472	-0.3425	-0.2784
p-value	0.3096	0.2011	0.2497	0.8619	0.4709	0.7594	0.5633	0.9341
CCS_p	-1.2492	-0.5386	-0.1153	0.4351	0.5036	0.2817	-0.0572	0.0537
p-value	0.2116	0.5902	0.9082	0.6635	0.6145	0.7782	0.9544	0.9572

Table 5.19: Encompassing result from Welch-Goyal Study (Annual Data, Up to 2019) - III

	Forecasts begin 20 years after sample							
τ	0.01	0.05	0.1	0.3	0.5	0.7	0.9	0.95
dfy								
DM_p	-1.4504	-1.4581	-1.4880	-1.2765	-1.3347	-1.5078	-1.7460	-1.7431
p-value	0.9265	0.9276	0.9316	0.8991	0.9090	0.9342	0.9596	0.9593
ENC_p	-0.2405	-0.5578	-0.8435	-0.5306	-0.2892	-0.0324	0.1917	0.1160
p-value	0.8100	0.5769	0.3990	0.5957	0.7724	0.9741	0.8480	0.9077
CCS_p	-1.2683	-0.1995	0.4105	0.9680	0.9683	0.8302	0.5400	0.7125
p-value	0.2047	0.8419	0.6815	0.3330	0.3329	0.4064	0.5892	0.4762
lty								
DM_p	0.8602	0.2356	-0.0606	-0.6675	-0.9121	-0.9914	-1.0745	-1.1135
p-value	0.1948	0.4069	0.5242	0.7478	0.8191	0.8392	0.8587	0.8673
ENC_p	1.4694	1.2331	1.2065	0.9629	0.7887	0.7179	0.6731	0.8651
p-value	0.1417	0.2175	0.2276	0.3356	0.4303	0.4728	0.5009	0.3870
CCS_p	-1.6215	-0.8601	-0.4160	0.0934	0.1525	0.1478	0.1676	0.4204
p-value	0.1049	0.3898	0.6774	0.9256	0.8788	0.8825	0.8669	0.6742
tbl								
DM_p	-1.0789	-1.0923	-1.2348	-1.2698	-1.4106	-1.5171	-1.4774	-1.3693
p-value	0.8597	0.8627	0.8915	0.8979	0.9208	0.9354	0.9302	0.9146
ENC_p	1.5844	1.1829	0.7462	0.4249	0.2179	0.1272	-0.0272	-0.0569
p-value	0.1131	0.2368	0.4555	0.6709	0.8275	0.8988	0.9783	0.9547
CCS_p	-1.6029	-1.1145	-0.8595	-0.5279	-0.4044	-0.2643	-0.0636	0.1967
p-value	0.1089	0.2650	0.3901	0.5975	0.6859	0.7916	0.9493	0.8441

Table 5.20: Encompassing result from Welch-Goyal Study (Annual Data, Up to 2019) - IV

	Forecasts begin 20 years after sample							
τ	0.01	0.05	0.1	0.3	0.5	0.7	0.9	0.95
tms								
DM_p	-2.2556	-1.9033	-1.7227	-1.4715	-1.6908	-1.7492	-1.6048	-1.5425
p-value	0.9880	0.9715	0.9575	0.9294	0.9546	0.9599	0.9457	0.9385
ENC_p	-0.7807	-0.2088	-0.3460	-0.2133	-0.4888	-0.7917	-0.8642	-0.6155
p-value	0.4350	0.8346	0.7293	0.8311	0.6250	0.4285	0.3875	0.5382
CCS_p	-0.8591	0.1592	0.8730	1.7726	1.6841	1.3021	0.8190	0.9382
p-value	0.3903	0.8735	0.3827	0.0763	0.0922	0.1929	0.4128	0.3482
b/m								
DM_p	-0.7265	-0.7796	-0.7093	-0.6437	-0.5337	-0.4589	-0.6548	-0.8013
p-value	0.7662	0.7822	0.7609	0.7401	0.7032	0.6768	0.7437	0.7885
ENC_p	0.2715	-0.0305	0.1964	0.5160	0.8239	1.1000	1.2308	1.2297
p-value	0.7860	0.9756	0.8443	0.6058	0.4100	0.2713	0.2184	0.2188
CCS_p	-1.0371	0.2054	0.8308	1.3976	1.2877	0.9224	0.3351	0.4455
p-value	0.2997	0.8372	0.4061	0.1622	0.1978	0.3563	0.7375	0.6559
dfr								
DM_p	-1.1114	-1.1116	-1.1916	-1.5952	-1.9224	-1.8454	-1.2981	-1.4566
p-value	0.8668	0.8668	0.8833	0.9447	0.9727	0.9675	0.9029	0.9274
ENC_p	0.3848	0.9152	-0.3583	-0.3294	0.0699	0.5366	0.5552	0.3465
p-value	0.7004	0.3601	0.7201	0.7419	0.9443	0.5916	0.5788	0.7290
CCS_p	0.7625	1.3666	1.8232	1.9120	1.6760	1.3815	0.4648	0.2683
p-value	0.4457	0.1717	0.0683	0.0559	0.0937	0.1671	0.6421	0.7885

Table 5.21: Encompassing result from Welch-Goyal Study (Annual Data, Up to 2019) - V

	Forecasts begin 20 years after sample							
τ	0.01	0.05	0.1	0.3	0.5	0.7	0.9	0.95
ltr								
DM_p	-1.4491	-1.2919	-1.3573	-1.4840	-1.4680	-1.0265	-0.1435	-0.0357
p-value	0.9263	0.9018	0.9127	0.9311	0.9289	0.8477	0.5570	0.5142
ENC_p	-0.1969	0.8327	0.0510	0.0004	0.5620	1.3869	1.8830	1.8542
p-value	0.8439	0.4050	0.9593	0.9996	0.5741	0.1655	0.0597	0.0637
CCS_p	-0.5993	0.2187	0.7822	1.3238	1.4074	1.4214	0.8144	0.7182
p-value	0.5490	0.8269	0.4341	0.1856	0.1593	0.1552	0.4154	0.4726
ntis								
DM_p	-1.3518	-1.3023	-1.1471	-1.4668	-1.6569	-1.8689	-1.3832	-1.2378
p-value	0.9118	0.9036	0.8743	0.9288	0.9512	0.9692	0.9167	0.8921
ENC_p	-1.2322	-1.0225	-0.5671	0.0932	0.4507	0.5459	0.6505	0.7233
p-value	0.2179	0.3065	0.5707	0.9258	0.6522	0.5851	0.5154	0.4695
CCS_p	0.2030	0.2939	0.2359	-0.0514	-0.4676	-0.9465	-1.1972	-0.8999
p-value	0.8391	0.7689	0.8135	0.9590	0.6401	0.3439	0.2312	0.3682
eqis								
DM_p	-0.1805	-0.1462	-0.2089	-0.2086	-0.2161	-0.3677	-0.5637	-0.8387
p-value	0.5716	0.5581	0.5827	0.5826	0.5855	0.6435	0.7135	0.7992
ENC_p	2.5796	2.2614	1.8226	1.4047	1.3324	1.2814	1.5777	1.3287
p-value	0.0099	0.0237	0.0684	0.1601	0.1827	0.2000	0.1146	0.1839
CCS_p	-1.4907	-0.9251	-0.6474	-0.2841	-0.2784	-0.3606	-0.4352	-0.0817
p-value	0.1360	0.3549	0.5174	0.7764	0.7807	0.7184	0.6634	0.9349

Table 5.22: Encompassing result from Welch-Goyal Study (Annual Data, Up to 2019) - VI

	Forecasts begin 20 years after sample							
τ	0.01	0.05	0.1	0.3	0.5	0.7	0.9	0.95
csp								
DM_p	-0.6856	-0.1514	0.3044	0.2741	-0.0956	-0.0625	0.5918	0.8124
p-value	0.7535	0.5602	0.3804	0.3920	0.5381	0.5249	0.2770	0.2083
ENC_p	0.0752	0.4409	1.0119	0.8977	0.6093	0.7210	1.0328	1.3515
p-value	0.9400	0.6593	0.3116	0.3694	0.5423	0.4709	0.3017	0.1765
CCS_p	0.6727	0.1472	-0.2561	-0.7449	-0.8476	-0.8320	-0.7851	-0.8156
p-value	0.5011	0.8830	0.7979	0.4564	0.3967	0.4054	0.4324	0.4147
i/k								
DM_p	2.4588	3.0034	2.3678	1.6148	1.3176	1.1269	1.2319	1.3055
p-value	0.0070	0.0013	0.0089	0.0532	0.0938	0.1299	0.1090	0.0959
ENC_p	2.9851	3.6328	3.3909	2.6660	2.2266	1.9042	1.7628	1.6957
p-value	0.0028	0.0003	0.0007	0.0077	0.0260	0.0569	0.0779	0.0899
CCS_p	-1.7229	-1.2220	-0.8729	-0.3555	-0.1869	-0.0142	0.2661	0.5442
p-value	0.0849	0.2217	0.3827	0.7222	0.8517	0.9887	0.7902	0.5863
cay								
DM_p	0.7687	0.3182	0.0745	0.0222	-0.0553	-0.2519	-1.1082	-2.0044
p-value	0.2210	0.3752	0.4703	0.4912	0.5221	0.5995	0.8661	0.9775
ENC_p	1.7586	1.9596	2.0321	1.7214	1.3934	1.0067	-0.0951	-1.0945
p-value	0.0786	0.0500	0.0421	0.0852	0.1635	0.3141	0.9242	0.2737
CCS_p	0.8761	0.7546	0.7592	0.9045	1.0639	1.2125	1.2726	1.2171
p-value	0.3810	0.4505	0.4477	0.3657	0.2874	0.2253	0.2032	0.2236

Chapter 6

Higher Order Elicitability and Forecast Encompassing in Predictive Models of Expected Shortfalls

6.1 Introduction

Risk management is important in financial institutions, and choosing a proper risk measure is crucial in financial risk management. Two risk measures are most widely used in financial markets: Value at Risk (VaR) and Expected Shortfall (ES). Basel II in 1996 proposed VaR as a proper risk measure, so VaR has become the standard measure of financial market risk. VaR measures the maximum potential loss of a given portfolio over a certain period

at a given confidence level, so VaR is the quantile with a given tail probability. However, A quantile-based VaR has two main drawbacks. First, VaR cannot satisfy the "coherent" since VaR lacks subadditivity. Artzner et al. (1999) introduce coherent risk measures. They define a risk measure to be coherent if the risk measure satisfies translation invariance ($\rho(X + a) = \rho(X) + a$ for all $a \in \mathbb{R}$), subadditivity ($\rho(X + Y) = \rho(X) + \rho(Y)$), positive homogeneity ($\rho(\lambda X) = \lambda\rho(X)$ for all $\rho \geq 0$) and monotonicity ($Y \leq X$ implies $\rho(Y) \leq \rho(X)$). Thus, lack of subadditivity means that the VaR of a portfolio can be larger than the sum of the individual VaRs, which violates the conventional concept that diversity reduces risk. Second, VaR focuses on the probability of the losses but not the magnitude of the losses. Two investments may have the same VaR, but the extreme losses may be different. Thus, VaR might not be an appropriate risk measure in some situations. Basel III in 2013 proposed another risk measure, which is the Expected Shortfall. ES is defined as the conditional expectation of the return given that it exceeds the VaR, so more sensitive to the magnitude of extreme losses.

Measures of the downside risk are essential in risk management for Macroeconomics. The increasing number of policymakers has focused on the downside risk in the last decade. The International Monetary Fund (IMF) has recently popularized a risk measure for GDP growth called Growth-at-Risk (GaR). GaR is the worst conditional GDP growth distribution at a given coverage level (5th percentile) depending on financial conditions (Adrian et al., 2019). Moreover, Adrian et al. (2019); Chavleishvili et al. (2021) define a measure of adverse real economic impact to be Growth Shortfall (GS). GS is the expectation of the GDP growth when it is less than GaR. Thus, GaR and GS have the

same definition as the VaR and ES respectively. VaR and ES measure the risk of financial returns, while GaR and GS measure the risk of GDP growth.

However, the measure of ES and GS is not "elicitable" (Gneiting, 2011). Elicitable risk measure means there is a "strictly consistent" scoring function for this risk measure. Then I can compare competing forecasts of this risk measure with respect to the consistent score (Fissler et al., 2016). There is no strictly consistent scoring function for ES and GS. Therefore, it is difficult to evaluate and compare the ES and GS forecasts due to this drawback.

Gneiting (2011) points out that the mean is elicitable, but the variance is not. However, a pair of mean and variance is elicitable. Fissler et al. (2016) introduce a strictly consistent scoring function for VaR and ES, so the pair of VaR and ES is elicitable. The existence of a strictly consistent scoring function for VaR and ES (FZ scoring function) accelerates the development of ES forecast. Patton et al. (2019) propose new dynamic models to forecast VaR and ES by using the FZ scoring function. Taylor (2019) uses the FZ scoring function to forecast VaR and ES based on the asymmetric Laplace distribution. In this paper, in order to test Granger causality, I also use the strictly consistent scoring function for VaR and ES to conduct in-sample estimation and out-of-sample forecasting.

Out-of-sample forecast comparison is widely used in many fields because it is suggested to test Granger Causality, which determines whether some independent variables can predict the dependent variable (Ashley et al., 1980). Many papers in the literature focus on out-of-sample tests for equal accuracy and encompassing (Diebold and Mariano, 1995; Clark and West, 2006, 2007). Out-of-sample tests for equal accuracy and encompassing

need two models. When the big model includes all the variables in the small model and has one or more other variables that are not in the small model, these two models are called nested models because the big model nests the small model. When these two models include the same variables and each model also has one or more unique variables, these two models are called non-nested models. These two models overlap each other, and no one can nest another.

When two non-nested models are to be compared, Diebold and Mariano (1995) introduce the DM statistic for comparing predictive accuracy. Under the null hypothesis of no difference in predictive accuracy, the test statistic DM is asymptotically N(0,1) distributed. However, Harvey et al. (1997) points out that the DM test performs well for large samples but could oversize for moderate samples. Hence, they modified the DM statistic by correcting the bias (HNL statistic). They use a critical value from the student's t-distribution instead of the standard normal distribution that the DM test used. However, when there are two nested models to be compared, according to Clark and McCracken (2001), the DM statistic for equal accuracy and the HNL statistic for encompassing are invalid because these statistics cannot converge to the standard normal distribution. Moreover, If there are two nested models, Clark and West (2006) and Clark and West (2007) prove that the DM statistic for mean regression has a downside bias, and this downward bias can be corrected if they add a non-negative adjustment term on it. In this paper, I extend the mean regression from Clark and West (2006, 2007) to ES and show that DM statistics for ES using the FZ scoring function also have a bias under the null hypothesis. I developed the encompassing

statistic for Granger Causality in ES. The encompassing statistic performs well in the size and power tests.

I consider two applications in this paper. The first application focuses on the financial market. I want to examine if the Macroeconomic and financial variables Granger Cause the equity premium of S&P 500 in VaR and ES. The second application is devoted to Macroeconomics. I want to check if the GDP growth can be predicted by the financial conditions in GaR and GS by using the FZ scoring function.

The paper is organized as follows. In section 2, I review the definitions, lemmas, and theorems of elicitability. In section 3, I show that the DM statistic has a bias with nested models, and introduce the encompassing test for Granger-Causality in forecasting ES or GS by using the FZ scoring function. In section 4, I prove the ENC for the FZ scoring function is asymptotically standard normal as the number of out-of-sample forecasts tends to infinity. In section 5, I conduct Monte Carlo simulations to show that the encompassing statistic has good size and power. In section 6, I present an empirical analysis. Proofs are presented in the appendix.

6.2 Objective function for ES and GS

We denote an observation domain O for $y, x \in \mathrm{O} \subseteq \mathbb{R}^{d_1+d_2}$, $d_1 = \dim(y)$ and $d_2 = \dim(x)$, the conditional distribution $F \equiv F(Y_{t+1}|\mathcal{I}_t)$ for Y_{t+1} given X_t. Let $\mathcal{F}$ be a class of distribution function on the observation domain O, and let A be an action domain, $\gamma \in \mathrm{A}$. We define $\Gamma : \mathcal{F} \to \mathrm{A}$ to be functional. For example, $\Gamma(F(y|x))$ may be $\mathbb{E}(Y|X)$, $\mathrm{Q}(Y|X)$, $\mathrm{Var}(Y|X)$, $\mathrm{Mode}(Y|X)$ or $\mathrm{ES}(Y|X)$, where $\mathbb{E}(Y|X)$ is the conditional mean, $\mathrm{Q}(Y|X)$ is the

conditional quantile, $\text{Var}(Y|X)$ is the conditional variance, $\text{Mode}(Y|X)$ is the conditional mode and $\text{ES}(Y|X)$ is the conditional Expected Shortfall. Note that Γ can be a vector of several of these.

Definition 1: (Gneiting, 2011; Fissler et al., 2016) A scoring function is an $\mathcal{F}$-integrable function $S : A \times O \to \mathbb{R}$. S is said to be *$\mathcal{F}$-consistent* for a functional $\Gamma : \mathcal{F} \to \text{A}$ if $\mathbb{E}_F S(\Gamma(F), Y) \leq \mathbb{E}_F S(\gamma, Y)$ for all $F \in \mathcal{F}$ and for all $\gamma \in \text{A}$. Furthermore, S is *strictly $\mathcal{F}$-consistent* for Γ if it is $\mathcal{F}$-consistent for Γ and if $\mathbb{E}_F S(\Gamma(F), Y) = \mathbb{E}_F S(\gamma, Y)$ implies that $\gamma = \Gamma(F)$ for all $F \in \mathcal{F}$ and for all $\gamma \in \text{A}$.

Definition 2: (Gneiting, 2011; Fissler et al., 2016) An identification function is an $\mathcal{F}$-integrable function $V : A \times O \to \mathbb{R}^k$. V is said to be an *$\mathcal{F}$-identification function* for a functional $\Gamma : \mathcal{F} \to \text{A} \subseteq \mathbb{R}^k$ if $\mathbb{E}_F V(\Gamma(F), Y) = 0$ for all $F \in \mathcal{F}$. Furthermore, V is a *strict $\mathcal{F}$-identification function* for Γ if $\mathbb{E}_F V(\gamma, Y) = 0$ holds if and only if $\gamma = \Gamma(F)$ for all $F \in \mathcal{F}$ and for all $\gamma \in \text{A}$.

A statistical functional is elicitable if a scoring function exists that the correct forecast of the functional is a unique minimizer of the expected score. We can compare or rank the forecasts of the elicitable functional with their realized scores (Fissler et al., 2016). Many statistical functionals are 1-elicitable, such as expectation, ratios of expectations, quantiles (Value-at-Risk), and expectiles. However, some functional are not 1-elicitable, such as variance, mode, and Expected Shortfall (Gneiting, 2011). Osband (1985) points

out that a non-elicitable functional can be a component of an elicitable functional. For example, variance is not elicitable, but there exists a 2-elicitable functional of mean and variance.

Definition 3: (Fissler et al., 2016) A functional $\Gamma : \mathcal{F} \to A \subseteq \mathbb{R}^2$ is called 2-elicitable if there exists a strictly $\mathcal{F}-$consistent scoring function for Γ. Then the functional $\Gamma = (\Gamma_1, \Gamma_2) : \mathcal{F} \to$ A is 2 elicitable.

According to Fissler et al. (2016), there is a relation among strictly consistent scoring function S, strict identification function V, and a matrix-valued function h. We define a vector-valued $\mathbf{m}$ to be a vector of m_1 and m_2, and a vector-valued $\boldsymbol{\gamma}$ to be a vector of γ_1 and γ_2. Taking the first order conditions of the expected loss function with respect to γ, γ_1 and γ_2 respectively, we can get $\mathbb{E}_F m$, $\mathbb{E}_F m_1$, and $\mathbb{E}_F m_2$ to be

$$\mathbb{E}_F \mathbf{m} \equiv \frac{\partial \mathbb{E}_F S(\boldsymbol{\gamma}, Y)}{\partial \boldsymbol{\gamma}} = \mathbf{h}(\boldsymbol{\gamma}) \mathbb{E}_F \mathbf{V}(\boldsymbol{\gamma}, Y) = 0, \tag{6.1}$$

$$\mathbb{E}_F m_1 \equiv \frac{\partial \mathbb{E}_F S(\gamma_1, \gamma_2, Y)}{\partial \gamma_1} = h_{11}(\gamma_1, \gamma_2) \mathbb{E}_F V_1(\gamma_1, \gamma_2, Y) + h_{12}(\gamma_1, \gamma_2) \mathbb{E}_F V_2(\gamma_1, \gamma_2, Y) = 0,$$

$$\mathbb{E}_F m_2 \equiv \frac{\partial \mathbb{E}_F S(\gamma_1, \gamma_2, Y)}{\partial \gamma_2} = h_{21}(\gamma_1, \gamma_2) \mathbb{E}_F V_1(\gamma_1, \gamma_2, Y) + h_{22}(\gamma_1, \gamma_2) \mathbb{E}_F V_2(\gamma_1, \gamma_2, Y) = 0,$$

where

$$\mathbf{m} = [m_1, m_2]', \qquad \boldsymbol{\gamma} = [\gamma_1, \gamma_2]',$$

$$\mathbf{V}(\gamma_1, \gamma_2, Y) = \begin{bmatrix} V_1(\gamma_1, \gamma_2, Y) \\ V_2(\gamma_1, \gamma_2, Y) \end{bmatrix},$$

$$\mathbf{h}(\gamma_1, \gamma_2) = \begin{bmatrix} h_{11}(\gamma_1, \gamma_2) & h_{12}(\gamma_1, \gamma_2) \\ h_{21}(\gamma_1, \gamma_2) & h_{22}(\gamma_1, \gamma_2) \end{bmatrix}.$$

Concerning equation (7.1), the strict identification function V is a function of variable y and functional γ, and the function h is a function of γ only. There is no y in the function of h. Functional forms of V1 and V2 are shown in Eq (6.3).

Quantile is elicitable, but Expected Shortfall is not elicitable (Gneiting, 2011). Fissler et al. (2016) propose a strictly consistent scoring function of joint Value-at-risk (VaR) and Expected Shortfall (ES) as

$$\begin{aligned} S(\gamma_1, \gamma_2, y) = &(1\{y \leq \gamma_1\} - \alpha) G_1(\gamma_1) - 1\{y \leq \gamma_1\} G_1(y) \\ &+ G_2(\gamma_2) \left(\gamma_2 - \gamma_1 + \frac{1}{\alpha} 1\{y \leq \gamma_1\} (\gamma_1 - y) \right) - \mathcal{G}(\gamma_2) + a(y), \end{aligned} \tag{6.2}$$

where γ_1 represents VaR and γ_2 represents ES, $\mathcal{G}' = G_2$, G_1 is increasing function, and G_2 is increasing and convex function. Therefore, a pair of VaR and ES is 2-elicitable. With regard to Fissler et al. (2016), taking the derivatives of the scoring function with respect to γ_1 and γ_2, we have

$$\mathbb{E}_F m_1 = \frac{\partial \mathbb{E}_F S(\gamma_1, \gamma_2, Y)}{\partial \gamma_1} = \left[G_1'(\gamma_1) + \frac{1}{\alpha} G_2(\gamma_2) \right] \mathbb{E}_F (1\{y \leq \gamma_1\} - \alpha) = 0,$$

$$\begin{aligned} \mathbb{E}_F m_2 = \frac{\partial \mathbb{E}_F S(\gamma_1, \gamma_2, Y)}{\partial \gamma_2} &= \frac{\gamma_1 G_2'(\gamma_2)}{\alpha} \mathbb{E}_F (1\{y \leq \gamma_1\} - \alpha) + G_2'(\gamma_2) \mathbb{E}_F \left(\gamma_2 - \frac{1}{\alpha} y 1\{y \leq \gamma_1\} \right) \\ &= 0, \end{aligned}$$

where the strict identification function is

$$\mathbf{V}(\gamma_1, \gamma_2, y) = \begin{bmatrix} V_1(\gamma_1, \gamma_2, Y) \\ V_2(\gamma_1, \gamma_2, Y) \end{bmatrix} = \begin{bmatrix} 1\{y \leq \gamma_1\} - \alpha \\ \gamma_2 - \frac{1}{\alpha} y 1\{y \leq \gamma_1\} \end{bmatrix}, \tag{6.3}$$

and the matrix-valued function h is

$$\mathbf{h}(\gamma_1, \gamma_2) = \begin{bmatrix} G_1'(\gamma_1) + \frac{1}{\alpha} G_2(\gamma_2) & 0 \\ \frac{\gamma_1 G_2'(\gamma_2)}{\alpha} & G_2'(\gamma_2) \end{bmatrix}. \tag{6.4}$$

For the first component of the identification function V, $\mathbb{E}_F\left(1\{y \leq \gamma_1\} - \alpha\right) = 0$ if and only if $\gamma_1 = \text{VaR}_\alpha(y)$. For the second component of V, $\mathbb{E}_F\left(\gamma_2 - \frac{1}{\alpha} y 1\{y \leq \gamma_1\}\right)$ if and only if $\gamma_2 = \text{ES}_\alpha(y)$. Therefore, $\mathbb{E}_F V(\gamma_1, \gamma_2, y) = 0$ if and only if $(\gamma_1; \gamma_2)$ equals the true VaR and ES of y, so the joint functional of VaR (Quantile) and ES is 2-elicitible. It indicates that this scoring function is strictly consistent for a pair of VaR and ES, which means that it can be used for forecasts comparison of VaR and ES.

Patton (2011b) states that positive homogeneity of the scoring function is important for forecast comparison. Moreover, Patton and Sheppard (2009) find out that homogeneity of degree zero leads to the higher power of Diebold-Mariano tests in volatility forecast. Patton et al. (2019) denote S_{FZ0} as the scoring function with homogeneity of degree zero. S_{FZ0} is homogeneous of degree zero if and only if $G_1(\gamma) = 0$ and $G_2(\gamma) = -1/\gamma$. For easy to read, let q replace γ_1 and e replace γ_2. Then, the "FZ0" scoring function is

$$S_{FZ0}\left(q_\alpha, e_\alpha, y\right) = S_\alpha\left(q_\alpha, e_\alpha, y\right) = -\frac{1}{\alpha e_\alpha} \mathbf{1}\{y \leq q_\alpha\}(q_\alpha - y) + \frac{q_\alpha}{e_\alpha} + \log(-e_\alpha) - 1. \tag{6.5}$$

In this paper, we use this FZ scoring function S_{FZ0} to test the forecast encompassing in Expected Shortfall.

6.3 Forecast Encompassing Test for Granger Causality in Expected Shortfall

In this section, we develop the forecast encompassing test in the ES. The conditional distribution $F \equiv F(Y_{t+1}|\mathcal{I}_t)$ for Y_{t+1} given X_t. Define $\gamma_1 = q$ for VaR. Define $\gamma_2 = e$ for ES. We have two models. One model is without conditioning on x_t, and the other model is with conditioning on x_t. The unconditional ES of the unconditional distribution $F_1 = F_Y(Y)$ is $\Gamma(F_1)$. The conditional ES of the conditional distribution $F_2 = F_{Y|X}(Y|X)$ is the functional $\Gamma(F_2)$. The two nested models for VaR q_{t+1} are

$$\text{Model 1}: \quad y_{t+1} = c_{1,\alpha} + u^{(1)}_{t+1,\alpha} \equiv x'_{1,t}\beta_{1,\alpha} + u^{(1)}_{\alpha,t+1} \tag{6.6}$$

$$\text{Model 2}: \quad y_{t+1} = c_{2,\alpha} + b_{1,\alpha}x_t + u^{(2)}_{\alpha,t+1} \equiv x'_{2,t}\beta_{2,\alpha} + u^{(2)}_{\alpha,t+1} \tag{6.7}$$

where $q^{(1)}_{\alpha,t+1} = c_{1,\alpha} = x'_{1,t}\beta_{1,\alpha}$ and $q^{(2)}_{\alpha,t+1} = c_{2,\alpha} = x'_{2,t}\beta_{2,\alpha}$. ES for Model 1 and Model 2 are defined as $e^{(1)}_{t+1} = \mathbb{E}_F\left[Y_{t+1} \mid Y_{t+1} \leq q^{(1)}_{\alpha,t+1}\right]$ and $e^{(2)}_{t+1} = \mathbb{E}_F\left[Y_{t+1} \mid Y_{t+1} \leq q^{(2)}_{\alpha,t+1}\right]$. The dependent variable y_{t+1} is a scalar random variable. $x_{1,t}$ is a strict subset of $x_{2,t}$. $x'_{1,t} = 1, \beta_{1,\alpha} = c_{1,\alpha}, x'_{2,t} = (1, x_t), \beta_{2,\alpha} = (c_{2,\alpha}, b_\alpha)$. Below the subscript α is omitted for simplicity.

The FZ scoring function is the scoring function S_{FZ0} in Patton et al. (2019), that is

$$S_\alpha(q, e, y) = -\frac{1}{\alpha e}\mathbf{1}\{y \leq q\}(q - y) + \frac{q}{e} + \log(-e) - 1.$$

According to Patton et al. (2019), the first oder conditions of FZ scoring function with respect to VaR q_{t+1} and ES e_{t+1} are

$$\mathbb{E}_F\left[-\frac{1}{\alpha e_{t+1}}\mathbf{1}\{Y_{t+1}\leq q_{t+1}\}+\frac{1}{e_{t+1}}\right]=0, \tag{6.8}$$

$$\mathbb{E}_F\left[\frac{1}{\alpha\left(e_{t+1}\right)^2}\mathbf{1}\left\{Y_{t+1}\leq q_{t+1}\right\}\left(q_{t+1}-Y\right)-\frac{q_{t+1}}{e_{t+1}^2}+\frac{1}{e_{t+1}}\right]=0, \tag{6.9}$$

which means that $m_1=-\frac{1}{\alpha e_{t+1}}\mathbf{1}\{y_{t+1}\leq q_{t+1}\}+\frac{1}{e_{t+1}}$ and $m_2=\frac{1}{\alpha e_{t+1}^2}\mathbf{1}\{y\leq q_{t+1}\}(q_{t+1}-y_{t+1})-\frac{q_{t+1}}{e_{t+1}^2}+\frac{1}{e_{t+1}}$ are martingale difference sequences. In the last section, we have already defined $m=(m_1,m_2)'$. So, rewriting equations (6.8) and (6.9), we have

$$\mathbb{E}_F\mathbf{m}=\mathbf{h}(q_{t+1},e_{t+1})\mathbb{E}_F\mathbf{V}(q_{t+1},e_{t+1},y_{t+1})=0, \tag{6.10}$$

where the strict identification functions in equation (7.1) are

$$\mathbf{V}(q_{t+1},e_{t+1},y_{t+1})=\begin{bmatrix}V_1(q_{t+1},e_{t+1},y_{t+1})\\ \\ V_2(q_{t+1},e_{t+1},y_{t+1})\end{bmatrix}=\begin{bmatrix}1\{y_{t+1}\leq q_{t+1}\}-\alpha\\ \\ e_{t+1}-\frac{1}{\alpha}y_{t+1}1\{y_{t+1}\leq q_{t+1}\}\end{bmatrix}, \tag{6.11}$$

and the matrix-valued function h is

$$\mathbf{h}(q_{t+1},e_{t+1})=\begin{bmatrix}-\frac{1}{\alpha e_{t+1}} & 0\\ \\ -\frac{q_{t+1}}{\alpha e_{t+1}^2} & \frac{1}{e_{t+1}^2}\end{bmatrix}. \tag{6.12}$$

In equation (6.11), V_1 shows how we define VaR (quantile). We have $\mathbb{E}_F V_1=0$, so $\mathbb{E}_F(1\{y_{t+1}\leq q_{t+1}\}-\alpha)=0$ implies $E_F(1\{y_{t+1}\leq q_{t+1}\})=\alpha$. V_2 shows how we define ES. We know that $\mathbb{E}_F V_2=0$, so $\mathbb{E}_F(e_{t+1}-\frac{1}{\alpha}y_{t+1}1\{y_{t+1}\leq q_{t+1}\})$ implies $e_{t+1}=\frac{1}{\alpha}\mathbb{E}_F(y_{t+1}1\{y_{t+1}\leq q_{t+1}\})$.

In order to test the equal predictive accuracy of the two nested models, we set the null and alternative hypotheses are

$$\mathbb{H}_0 : \mathbb{E}_F \left[S_\alpha \left(q_{t+1}^{(1)}, e_{t+1}^{(1)}, Y_{t+1} \right) - S_\alpha \left(q_{t+1}^{(2)}, e_{t+1}^{(2)}, Y_{t+1} \right) \right] = 0 \tag{6.13}$$

$$\mathbb{H}_1 : \mathbb{E}_F \left[S_\alpha \left(q_{t+1}^{(1)}, e_{t+1}^{(1)}, Y_{t+1} \right) - S_\alpha \left(q_{t+1}^{(2)}, e_{t+1}^{(2)}, Y_{t+1} \right) \right] > 0 \tag{6.14}$$

We set the alternative hypothesis is one side because we expect that the forecasts from Model 2 are better than those from Model 1, as Clark and West (2006) did. When coefficient $b = 0$, x does not Granger cause y. When coefficient $b \neq 0$, x Granger causes y. Define R as the in-sample observations. Define P as the out-of-sample forecasts. The DM statistics based on the FZ loss differential is

$$D = \mathbb{E}_F \left[S_\alpha \left(q_{t+1}^{(1)}, e_{t+1}^{(1)}, Y_{t+1} \right) - S_\alpha \left(q_{t+1}^{(2)}, e_{t+1}^{(2)}, Y_{t+1} \right) \right]$$

$$\hat{D}_{R,P} = P^{-1} \sum_{t=R}^{T} \left[S_\alpha \left(\hat{q}_{t+1}^{(1)}, \hat{e}_{t+1}^{(1)}, Y_{t+1} \right) - S_\alpha \left(\hat{q}_{t+1}^{(2)}, \hat{e}_{t+1}^{(2)}, Y_{t+1} \right) \right]$$

where $T + 1 = R + P$. For mean regression with nested models, Clark and West (2006, 2007) prove that the DM statistic has a downward bias and show that adjusted DM statistics (DM statistics plus the adjusted term) obtain zero mean expectation under the null hypothesis, corrects the size, and increases the power. Thus, at first, we attempt to show that the DM statistic based on the FZ loss differential $\hat{D}_{R,P}$ has a non-zero mean under the null hypothesis.

Thereom 1: the DM statistic of the FZ loss differential $\hat{D}_{R,P}$ has a non-zero mean. Thus, $\mathbb{E}\hat{D}_{R,P} \neq 0$ under $\mathbb{H}_0$.

Proof: In order to show that $\hat{D}_P$ is non-zero, we consider eight cases. (1).$Y_{t+1} > \hat{q}_{t+1}^{(1)}, Y_{t+1} >$

$\hat{q}_{t+1}^{(2)}$ and $\hat{e}_{t+1}^{(1)} > \hat{e}_{t+1}^{(2)}$.(2).$Y_{t+1} > \hat{q}_{t+1}^{(1)}, Y_{t+1} > \hat{q}_{t+1}^{(2)}$ and $\hat{e}_{t+1}^{(1)} < \hat{e}_{t+1}^{(2)}$.(3).$Y_{t+1} \leq \hat{q}_{t+1}^{(1)}, Y_{t+1} > \hat{q}_{t+1}^{(2)}$ and $\hat{e}_{t+1}^{(1)} > \hat{e}_{t+1}^{(2)}$.(4).$Y_{t+1} \leq \hat{q}_{t+1}^{(1)}, Y_{t+1} > \hat{q}_{t+1}^{(2)}$, and $\hat{e}_{t+1}^{(1)} < \hat{e}_{t+1}^{(2)}$.(5).$Y_{t+1} > \hat{q}_{t+1}^{(1)}, Y_{t+1} \leq \hat{q}_{t+1}^{(2)}$ and $\hat{e}_{t+1}^{(1)} > \hat{e}_{t+1}^{(2)}$.(6).$Y_{t+1} > \hat{q}_{t+1}^{(1)}, Y_{t+1} \leq \hat{q}_{t+1}^{(2)}$, and $\hat{e}_{t+1}^{(1)} < \hat{e}_{t+1}^{(2)}$. (7).$Y_{t+1} \leq \hat{q}_{t+1}^{(1)}, Y_{t+1} \leq \hat{q}_{t+1}^{(2)}$ and $\hat{e}_{t+1}^{(1)} > \hat{e}_{t+1}^{(2)}$.(8).$Y_{t+1} \leq \hat{q}_{t+1}^{(1)}, Y_{t+1} \leq \hat{q}_{t+1}^{(2)}$ and $\hat{e}_{t+1}^{(1)} < \hat{e}_{t+1}^{(2)}$. In these eight cases, we can show that $\hat{D}_{R,P}$ has a non-zero mean. See Appendix A2 for details.

Due to the bias of DM statistic, we want to develop the encompassing test to show that the encompassing statistic has zero mean. In order to find out the encompassing statistic for the ES, we build a combined FZ scoring function by combining two models, VaR q and ES e, and take the derivative of the expectation of the combined FZ scoring function with respect to the weight λ.

Theorem 2: Under $\mathbb{H}_0$,

$$C \equiv \mathbb{E}_F \left[m_1^{(1)} \left(\hat{q}_{t+1}^{(2)} - \hat{q}_{t+1}^{(1)} \right) + m_2^{(1)} \left(\hat{e}_{t+1}^{(2)} - \hat{e}_{t+1}^{(1)} \right) \right] = 0. \tag{6.15}$$

$$\hat{C}_{R,P} \equiv P^{-1} \sum_{t=R}^{T} \left[\hat{m}_{1,t+1}^{(1)} \left(\hat{q}_{t+1}^{(2)} - \hat{q}_{t+1}^{(1)} \right) + \hat{m}_{2,t+1}^{(1)} \left(\hat{e}_{t+1}^{(1)} - \hat{e}_{t+1}^{(2)} \right) \right] \xrightarrow{p} C = 0 \tag{6.16}$$

as $R, P \to \infty$ and $P/R \to \infty$.

Proof: We combine model 1 and model 2 with weight $1-\lambda$ and λ respectively. To estimate the expectation of the FZ scoring function with combined VaR $q_{t+1}^{(c)}$ and ES $e_{t+1}^{(c)}$ to find out the encompassing statistic under the null hypothesis, we have

$$\lambda = \arg\min_{\lambda} \mathbb{E}_F S_\alpha(q_{t+1}^{(c)}, e_{t+1}^{(c)}, Y_{t+1}),$$

where $q_{t+1}^{(c)} = (1-\lambda)\, q_{t+1}^{(1)} + \lambda q_{t+1}^{(2)}$ and $e_{t+1}^{(c)} = (1-\lambda)\, e_{t+1}^{(1)} + \lambda e_{t+1}^{(2)}$. We define C to be the first order condition of the expected FZ scoring function with respect to λ, then we have

$$\begin{aligned} C &\equiv \frac{\partial \mathbb{E}_F \left[S_\alpha(q_{t+1}^{(c)}, e_{t+1}^{(c)}, Y_{t+1}) \right]}{\partial \lambda} \\ &= \frac{\partial \mathbb{E}_F S_\alpha(q_{t+1}^{(c)}, e_{t+1}^{(c)}, Y_{t+1})}{\partial q^{(c)}} \frac{\partial q_{t+1}^{(c)}}{\partial \lambda} + \frac{\partial \mathbb{E}_F S_\alpha(q_{t+1}^{(c)}, e_{t+1}^{(c)}, Y_{t+1})}{\partial e_{t+1}^{(c)}} \frac{\partial e_{t+1}^{(c)}}{\partial \lambda} \\ &= \mathbb{E}\left[m_1^{(c)} \left(q_{t+1}^{(2)} - q_{t+1}^{(1)} \right) + m_2^{(c)} \left(e_{t+1}^{(2)} - e_{t+1}^{(1)} \right) \right] \\ &= \mathbb{E}\left[m_1^{(1)} \left(q_{t+1}^{(2)} - q_{t+1}^{(1)} \right) + m_2^{(1)} \left(e_{t+1}^{(2)} - e_{t+1}^{(1)} \right) \right] = 0 \end{aligned}$$

under $\mathbb{H}_0$.

Due to the first order condition, C should be 0. We estimate C by $\hat{C}_{R,P}$, so $\hat{C}_{R,P}$ is defined as

$$\hat{C}_{R,P} \equiv P^{-1} \sum_{t=R}^{T} \left[\hat{m}_{1,t+1}^{(1)} \left(\hat{q}_{t+1}^{(2)} - \hat{q}_{t+1}^{(1)} \right) + \hat{m}_{2,t+1}^{(1)} \left(\hat{e}_{t+1}^{(2)} - \hat{e}_{t+1}^{(1)} \right) \right] \xrightarrow{p} C = 0 \tag{6.17}$$

under $\mathbb{H}_0$, as $R, P \to \infty$ and $P/R \to \infty$

Under the null hypothesis, $\lambda = 0$, we obtain $q_{t+1}^{(c)} = q_{t+1}^{(1)}$ and $e_{t+1}^{(c)} = e_{t+1}^{(1)}$. Thus, we obtain $\mathbb{E}(\hat{C}_{R,P}) = 0$, so $\hat{C}_{R,P} \xrightarrow{p} 0$ as $R, P \to \infty$. See Appendix A2 for details.

In order to get the encompassing statistic for ES, we consider endowing Model 1 and Model 2 with the weight $1-\lambda$ and λ respectively, and find out the property of optimal weight $\hat{\lambda}$. Theorem 3 gives the relationship between $\hat{\lambda}$ and $\hat{C}_P$.

Theorem 3: We denote $\lambda = \arg\min_\lambda \mathbb{E}_F S\left(q_{t+1}^{(c)}, e_{t+1}^{(c)}, Y_{t+1} \right)$. Under $\mathbb{H}_0, \hat{\lambda} \to 0$, $\hat{C}_{R,P} \xrightarrow{p} 0$ as $R, P \to \infty$.

Proof: See Appendix B.

This theorem shows that $\hat{\lambda} \to 0$ and $\hat{C}_{R,P} \xrightarrow{p} 0$ as $R, P \to \infty$ are equivalent.

We compare the DM statistic and the encompassing statistic with the CCS statistic, the test statistic by Chao et al. (2001). Chao et al. (2001) show that the CCS statistic is $\hat{M}_{R,P} = P^{-1} \sum_{t=R}^{T} \hat{u}_{t+1}^{(1)} x_t$. In order to compare the equal predictive accuracy of two nested expectile models, we standardize these three statistics. The DM statistic is $\text{DM}_P \equiv \hat{S}_{R,P}^{-0.5} \sqrt{P} \hat{D}_{R,P}$, where $S_{R,P} = \text{var}\left(\sqrt{P}\hat{D}_{R,P}\right)$ and $S_{R,P} - \hat{S}_{R,P} \xrightarrow{p} 0$. The encompassing statistic is $\text{ENC}_P \equiv \hat{Q}_{R,P}^{-0.5} \sqrt{P} \hat{C}_{R,P}$, where $Q_{R,P} = \text{var}\left(\sqrt{P}\hat{C}_{R,P}\right)$ and $Q_{R,P} - \hat{Q}_{R,P} \xrightarrow{p} 0$. The CCS statistic is $\text{CCS}_P \equiv \hat{W}_{R,P}^{-0.5} \sqrt{P} \hat{M}_{R,P}$, where $W_{R,P} = \text{var}\left(\sqrt{P}\hat{M}_{R,P}\right)$ and $W_{R,P} - \hat{W}_{R,P} \xrightarrow{p} 0$.

6.4 Asymptotic Normality of the ENC Statistic for ES

In this section, we compare two nested models using encompassing test and model combination, then explore the properties of the two methodologies.

6.4.1 Encompassing Test for Equal Predictive Accuracy

We derive the asymptotic distribution of ENC_P. To show ENC_P to be asymptotical standard normality, we simplify the ENC_P statistic.

$$\text{ENC}_P = Q_P^{-1/2} \sqrt{P} \hat{C}_{R,P} = \frac{\sum_t c_t}{\sqrt{\sum_t c_t^2 - P\hat{C}_{R,P}^2}},$$

where

$$\hat{m}_{1,t+1} = -\frac{1}{\alpha \hat{e}_{t+1}^{(1)}} \mathbf{1}\{Y_{t+1} \leq \hat{q}_{t+1}^{(1)}\} + \frac{1}{\hat{e}_{t+1}^{(1)}}$$

$$\hat{m}_{2,t+1} = \frac{1}{\alpha \left(\hat{e}_{t+1}^{(1)}\right)^2} \mathbf{1}\left\{Y_{t+1} \leq \hat{q}_{t+1}^{(1)}\right\} \left(\hat{q}_{t+1}^{(1)} - Y_{t+1}\right) - \frac{\hat{q}_{t+1}^{(1)}}{\left(\hat{e}_{t+1}^{(1)}\right)^2} + \frac{1}{\hat{e}_{t+1}^{(1)}}.$$

$$\hat{C}_P = P^{-1} \sum_{t=R}^{T} c_t = P^{-1} \sum_{t=R}^{T} \hat{m}_{1,t+1} \left(\hat{q}_{t+1}^{(2)} - \hat{q}_{t+1}^{(1)}\right) - \hat{m}_{2,t+1} \left(\hat{e}_{t+1}^{(2)} - \hat{e}_{t+1}^{(1)}\right).$$

In order to obtain the limiting distribution in Theorem 4, we need some notations. We define $g\left(u_{t+1}^{(i)}\right) = \left[\alpha - 1\left(u_{t+1}^{(i)} < 0\right)\right]$. The score $k_{i,t}(\beta_{i,t})$ is the derivative of the expected FZ scoring function with respect to $\beta_{i,t}$. Let $h_{i,t+1} = -k_{i,t+1}(\beta_{i,t})$ and $H_i(t+1) - \frac{1}{R}\sum_{t=1}^{R} h_{i,t+1}$. We denote the Hessian to be $B_i^{-1} = \mathbb{E}_F \Lambda(\beta_{i,t})$, which is the second order condition of expected FZ scoring function with respect to $\beta_{i,t}$. Therefore, $B_1 = 1$ and $B_2 = \begin{pmatrix} 1 & 0 \\ 0 & \mathbb{E}_F \Lambda(\beta_{2,t}) \end{pmatrix}^{-1}$. By the central limit theorem, $\sqrt{R}\left[R^{-1}\sum_{s=t-R+1}^{t} h_{2,s}\right] \sim N\left(0, Z^2 B_2^{-1}\right)$, where $Z^2 B_2^{-1} = \mathbb{E}[k_{i,t}(\beta_{i,t}) k_{i,t}(\beta_{i,t})']$.

The following assumptions can be used to obtain the limiting distribution in Theorem 4. These assumptions are only sufficient but not necessary and sufficient.

Assumption 1: The parameter estimates $\hat{\beta}_{i,t}, i = 1, 2, t = R, \ldots, T$, satisfy $\hat{\beta}_{i,t} - \beta_i = B_i(t) H_i(t) = \left(R^{-1}\sum_{j=t-R+1}^{t} \Lambda_{i,j}\right)^{-1} \left(R^{-1}\sum_{j=t-R+1}^{t} h_{i,j}\right)$.

Assumption 2: Let $U_t = \left[\mathbb{E}k_t(\theta), x_{2,t}' - \mathbb{E}x_{2,t}', h_{2,t}', \mathrm{vec}\left(h_{2,t}h_{2,t}' - \mathbb{E}h_{2,t}h_{2,t}'\right), \right]$. (a) $\mathbb{E}U_t = 0$. (b) Denote $\tilde{U}_t$ to be the vector of U_t. $R^{-1}\mathrm{E}\left(\sum_{j=t-R+1}^{t} \tilde{U}_j\right)\left(\sum_{j=t-R+1}^{t} \tilde{U}_j\right)'$ $= \Omega < \infty$ is p.d.

Assumption 3: (a) $\mathbb{E}h_{2,t}h_{2,t}' = Z^2 B_2^{-1}$, (b) $\mathbb{E}\left(h_{2,t} \mid h_{2,t-j}, j = 1, 2, \ldots\right) = 0$.

Assumptions 2 and 3 allow the application of an invariance principle and are sufficient for joint weak convergence of partial sums and averages of these partial sums to Brownian motion and integrals of Brownian motion.

Assumption 4: $\lim_{P,R\to\infty} P/R = \pi = \infty$.

In order to get accurate out-of-sample forecasts, it is essential to select an optimal in-sample window. However, no consensus opinion has been reached as it depends on the characteristics of the model. Inoue et al. (2017) find the optimal window size by minimizing the conditional mean squared forecast error in a time-varying predictive regression model. They show that the optimal window size satisfies $R = O(T^{2/3})$. Hong et al. (2020) consider minimizing various out-of-sample forecast errors, including the unconditional, conditional, and global mean squared forecast errors, respectively. The optimal window size they found is $R = O(T^{4/5})$. The out-of-sample window P has a faster divergent rate than the in-sample window R. Thus, in our model, we set $P/R \to +\infty$.

Theorem 4: Let Assumptions 1-4 hold. $\mathrm{ENC_P} \xrightarrow{d} N(0,1)$, as $P, R \to \infty$, under $\mathbb{H}_0$: $\lambda = 0$.

West (1996) proves that ENC statistic is asymptotically normal under $\pi \geq 0$ for mean

regression when two models are non-nested. Clark and McCracken (2001) show that ENC_P statistic is asymptotically standard normal under $\pi = 0$, and ENC_P statistic is not standard normal under $\pi > 0$ when two mean models are nested models. In this theorem, we show that ENC_P statistic is asymptotically standard normality under π tends to ∞. See Appendix C for proof of Theorem 4.

6.5 Monte Carlo Simulation

6.5.1 Simulation Design

In order to show the asymptotic distribution of encompassing test for FZ scoring function, we have three DGPs.

In both two DGPs, we set x_t in Model 2 as an AR(1) stationary process, so $x_t = \phi x_{t-1} + v_t$, where $v_t \overset{iid}{\sim} N(0, \sigma_v^2)$. Thus, x has zero mean and variance $\sigma_v^2/(1-\phi)$. The error term $u_{t+1}^{(2)}$ of Model 2 for Value-at-risk (quantile) satisfies

$$\mathbb{E}\left(\alpha - \mathbf{1}\left(u_{t+1}^{(2)} < 0\right) \mid x_t\right) = 0 \tag{6.18}$$

which means that $\alpha - \mathbf{1}(u_{t+1}^{(2)} < 0)$ is a martingale difference series and the conditional quantile of $v_{t+1}^{(2)}$ given x_t is zero.

In DGP 1, we generate the u_{t+1}^2 following normal distribution. The mean and variance of u_{t+1}^2 satisfy

$$\frac{\mathbb{E}\left(u_{t+1}^{(2)}\right)}{\sqrt{\text{Var}\left(u_{t+1}^{(2)}\right)}} = \frac{-\Phi^{-1}(\alpha)\sigma_u}{\sigma_u} = -\Phi^{-1}(\alpha) \tag{6.19}$$

We set the $\text{Var}\left(u_{t+1}^{(2)}\right) = \sigma_u^2$, then we can get $\mathbb{E}\left(u_{t+1}^{(2)}\right) = -\Phi^{-1}(\alpha)\sigma_u$. Second, we generate $\{y_{t+1}\}$ from $y_{t+1} = c_2 + bx_t + e_{t+1}^{(2)}$, then we has the mean and variance as follows:

$$\mathbb{E}\left(y_{t+1}\right) = \mathbb{E}\left(c_{12} + b_1 x_t + u_{t+1}^{(2)}\right) = c_{12} + 0 - \Phi^{-1}(\alpha)\sigma_u$$

$$\mathrm{Var}\left(y_{t+1}\right) = \mathrm{Var}\left(c_{12} + b_1 x_t + e_{t+1}^{(2)}\right) = b_1^2\sigma_x^2 + \sigma_u^2$$

In order to make sure all $\hat{v}_{t+1}$ and $\hat{e}_{t+1}$ to be negative, we set $\mathbb{E}(Y) = -3$, which implies that $c_{12} = -3 + \Phi^{-1}(\alpha)\sigma_u$. Third, we find out mean and variance of $u_{t+1}^{(1)}$ as follows:

$$\mathbb{E}\left(u_{t+1}^{(1)}\right) = \mathbb{E}\left(y_{t+1} - c_{11}\right) = \mathbb{E}\left(c_{12} + b_1 x_t + u_{t+1}^{(2)} - c_{11}\right) = c_{12} + 0 - \Phi^{-1}(\alpha)\sigma_u - c_{11}$$

$$\mathrm{Var}\left(u_{t+1}^{(1)}\right) = \mathrm{Var}\left(y_{t+1} - c_{11}\right) = \mathrm{Var}\left(c_{12} + b_1 x_t + e_{t+1}^{(2)} - c_{11}\right) = b_1^2\sigma_x^2 + \sigma_u^2$$

We set $u_{t+1}^{(1)}$ following normal distribution and should satisfy

$$\frac{\mathbb{E}\left(u_{t+1}^{(1)}\right)}{\sqrt{\mathrm{Var}\left(u_{t+1}^{(1)}\right)}} = \frac{c_{12} - c_{11} - \Phi^{-1}(\alpha)\sigma_u}{\sqrt{b_1^2\sigma_x^2 + \sigma_u^2}} = -\Phi^{-1}(\alpha) \tag{6.20}$$

Simplifying the above equation, $c_{12} - c_{11} = \Phi^{-1}(\alpha)\sigma_u - \Phi^{-1}(\alpha)\sqrt{b_1^2\sigma_x^2 + \sigma_e^2}$. Since $c_{12} = -3 + \Phi^{-1}(\alpha)\sigma_u$, $c_{11} = -3 + \Phi^{-1}(\alpha)\sqrt{b_1^2\sigma_x^2 + \sigma_e^2}$. Under the null, when $b_1 = 0$, $c_{11} = c_{12}$. We consider $\alpha \in \{0.01, 0.025, 0.05, 0.1\}, \phi \in \{0, 0.95\}, \sigma_u \in \{1\}, \sigma_v \in \{1\}$, and $b \in \{0, 0.1\}$.

In DGP2, we generate y_{t+1} by using GARCH(1,1).

$$y_{t+1} = \sigma_{t+1}u_{t+1}$$

$$\sigma_{t+1} = \omega + \zeta u_t^2 + \beta\sigma_t^2 + bx_t^2$$

$$u_t \sim N(0,1)$$

We set the parameters to be $\zeta = 0.05$, $\beta = 0.85$, $b \in \{0, 0.1\}$, and $\omega = 1 - \zeta - \beta - \delta$. We make x_t follow $N(0,1)$. According to Patton et al. (2019), Under DGP2, by using standard normal distribution for u_{t+1},

$$\mathrm{VaR}_{\alpha,t+1} = a_\alpha\sigma_{t+1} \qquad a_\alpha = \Phi^{-1}(\alpha)$$

$$\mathrm{ES}_{\alpha,t+1} = b_\alpha\sigma_{t+1} \qquad b_\alpha = -\phi\left(\Phi^{-1}(\alpha)\right)/\alpha$$

In DGP3, following Patton et al. (2019), we generate y_{t+1} by using GAS models.

$$y_{t+1} = \mu_t + \sigma_{t+1} u_{t+1},$$

$$q_{t+1} = w_1 + d_1 q_t + a_1 V_{1,t},$$

$$e_{t+1} = w_2 + d_2 e_t + a_2 V_{2,t} + b x_t.$$

Recall that $V_{1,t} = 1\{y_t \leq q_t\} - \alpha$, and $V_{2,t} = e_t - \frac{1}{\alpha} 1\{y_t \leq q_t\} y_t$. We still generate u_{t+1} to follow standard normal distribution, so we have

$$q_{t+1} = \mu_{t+1} + a_\alpha \sigma_{t+1},$$

$$e_{t+1} = \mu_{t+1} + b_\alpha \sigma_{t+1}.$$

We set the parameters to be $b_{11} = b_{22} = 0.95$, $a_{11} = a_{22} = 0.05$, $b \in \{0, 0.1\}$, $w_1 = (1 - b_{11}) a_\alpha$, and $w_2 = (1 - b_{22}) b_\alpha$. We make x_t follow $N(0, 1)$.

In our simulation, we set the number of in-sample observations $R \in \{120, 240, 480\}$ and out-of-sample forecasts $P \in \{240, 480, 1200\}$. We use rolling windows to estimate θ by minimizing S_{FZ0}. In DGP1, $\theta_1 = c_1$ for Model 1 and $\theta_2 = (c_2, b)'$ for Model 2. Then, we can obtain the one step ahead forecast $q_{t+1}^{(1)} = c_1$ and $e_{t+1}^{(1)} = \mathbb{E}_F(Y_{t+1} | Y_{t+1} \leq q_{t+1}^{(1)})$ for Model 1 and $q_{t+1}^{(2)} = c_2 + b x_t$ and $e_{t+1}^{(2)} = \mathbb{E}_F(Y_{t+1} | Y_{t+1} \leq q_{t+1}^{(2)}$ for Model 2. In DGP2, $\theta_1 = (\omega_1, \zeta_1, \beta_1, a_{1,\alpha}, b_{1,\alpha})$ for Model 1 and $\theta_2 = (\omega_2, \zeta_2, \beta_2, a_{2,\alpha}, b_{2,\alpha}, \delta)$ for Model 2. Then, we can obtain the one step ahead forecast $q_{t+1}^{(1)} = a_{1,\alpha} \sigma_{t+1}^{(1)}$ and $e_{t+1}^{(1)} = b_{1,\alpha} \sigma_{t+1}^{(1)}$ for Model 1 and $q_{t+1}^{(2)} = a_{1=2,\alpha} \sigma_{t+1}^{(2)}$ and $e_{t+1}^{(2)} = b_{2,\alpha} \sigma_{t+1}^{(2)}$ for Model 2. In DGP3, $\theta_1 = (w_{11}, d_{11}, a_{11}, w_{21}, d_{21}, a_{21})$ for Model 1 and $\theta_2 = (w_{12}, d_{12}, a_{12}, w_{22}, d_{22}, a_{22}, b)$ for Model 2. Then, we can obtain the one step ahead forecast $q_{t+1}^{(1)} = w_{11} + d_{11} q_t + a_{11} V_{1,t}$ and $e_{t+1}^{(1)} = w_{21} + d_{21} e_t + a_{21} V_{2,t}$ for Model 1 and $q_{t+1}^{(2)} = w_{12} + d_{12} q_t + a_{12} V_{1,t}$ and $e_{t+1}^{(2)} = w_{22} + d_{22} e_t + a_{22} V_{2,t}$ for Model 2. By using those forecast variables, we can obtain $\mathrm{DM}_{R,P}$, $\mathrm{ENC}_{R,P}$, and $\mathrm{CCS}_{R,P}$

statistics. Repeating this procedure 2000 times, we get the asymptotic distributions of $\mathrm{DM}_{R,P}$, $\mathrm{ENC}_{R,P}$, and $\mathrm{CCS}_{R,P}$, and the size and power of these three statistics.

6.5.2 Simulation result on encompassing test

Figures 1-2 and tables 1-2 show the Monte Carlo simulation in DGP1 for asymptotic distribution and the size and powers of $\mathrm{DM}_{R,P}$, $\mathrm{ENC}_{R,P}$, and $\mathrm{CCS}_{R,P}$ statistics under different α, b, and ϕ. Table 1 shows the result of the size of the test under $b = 0$ and $\phi = 0$. For different in-sample observations R, out-of-sample forecasts P, and α, the result demonstrates that $\mathrm{DM}_{R,P}$ is much less than 5% under the 5% nominal level, implying that DM_P has a downward bias under $\mathbb{H}_0$. However, the sizes of $\mathrm{ENC}_{R,P}$ and $\mathrm{CCS}_{R,P}$ are good under the 5% nominal level. Table 2 shows the result of the power of the test under $b = 0.1$ and $\phi = 0.95$. It illustrates that $\mathrm{ENC}_{R,P}$ has the highest power and $\mathrm{DM}_{R,P}$ has the higher power than $\mathrm{CCS}_{R,P}$ for different in-sample observations R, out-of-sample forecasts P and α.

Figure 1 shows the asymptotic distribution of $\mathrm{DM}_{R,P}$, $\mathrm{ENC}_{R,P}$, and $\mathrm{CCS}_{R,P}$ statistics under $\mathbb{H}_0$ when $\alpha = 0.05$. These figures demonstrate that $\mathrm{DM}_{R,P}$ has a negative mean and high kurtosis, which implies that $\mathrm{DM}_{R,P}$ has a downward bias under $\mathbb{H}_0$. However, the asymptotic distributions of $\mathrm{ENC}_{R,P}$ and $\mathrm{CCS}_{R,P}$ are close to the standard normal distribution. Figure 2 shows the asymptotic distribution of $\mathrm{DM}_{R,P}$, $\mathrm{ENC}_{R,P}$ and $\mathrm{CCS}_{R,P}$ statistic under $\alpha = 0.05$, $b = 0.1$ and $\phi = 0, 0.95$. From these figures, we can see that the asymptotic distribution of $\mathrm{CCS}_{R,P}$ move to the left, so CCS_P has the lowest mean compared to $\mathrm{DM}_{R,P}$

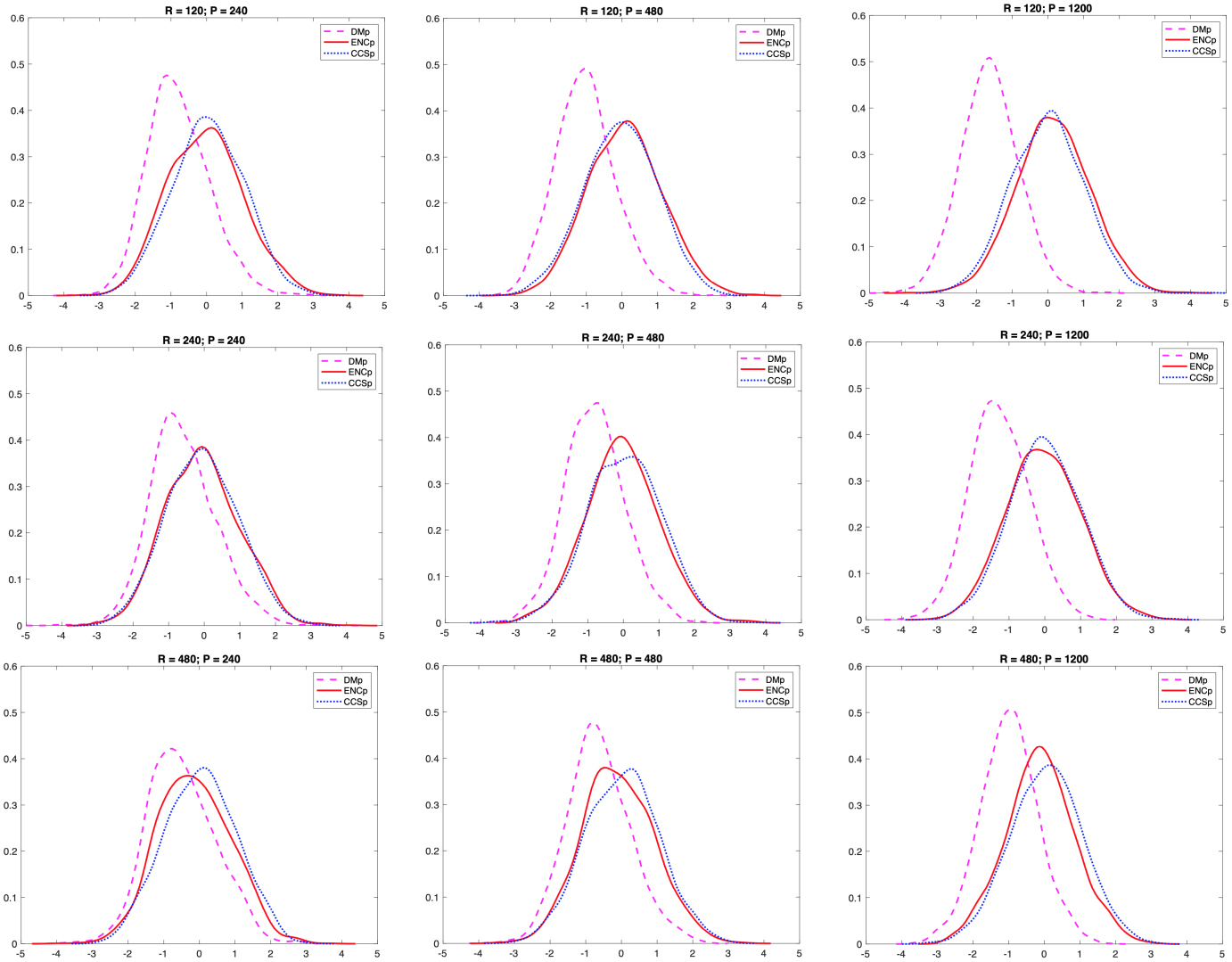

Figure 6.1: Simulation Results of DGP 1: Size of Test, $\alpha = 0.05$

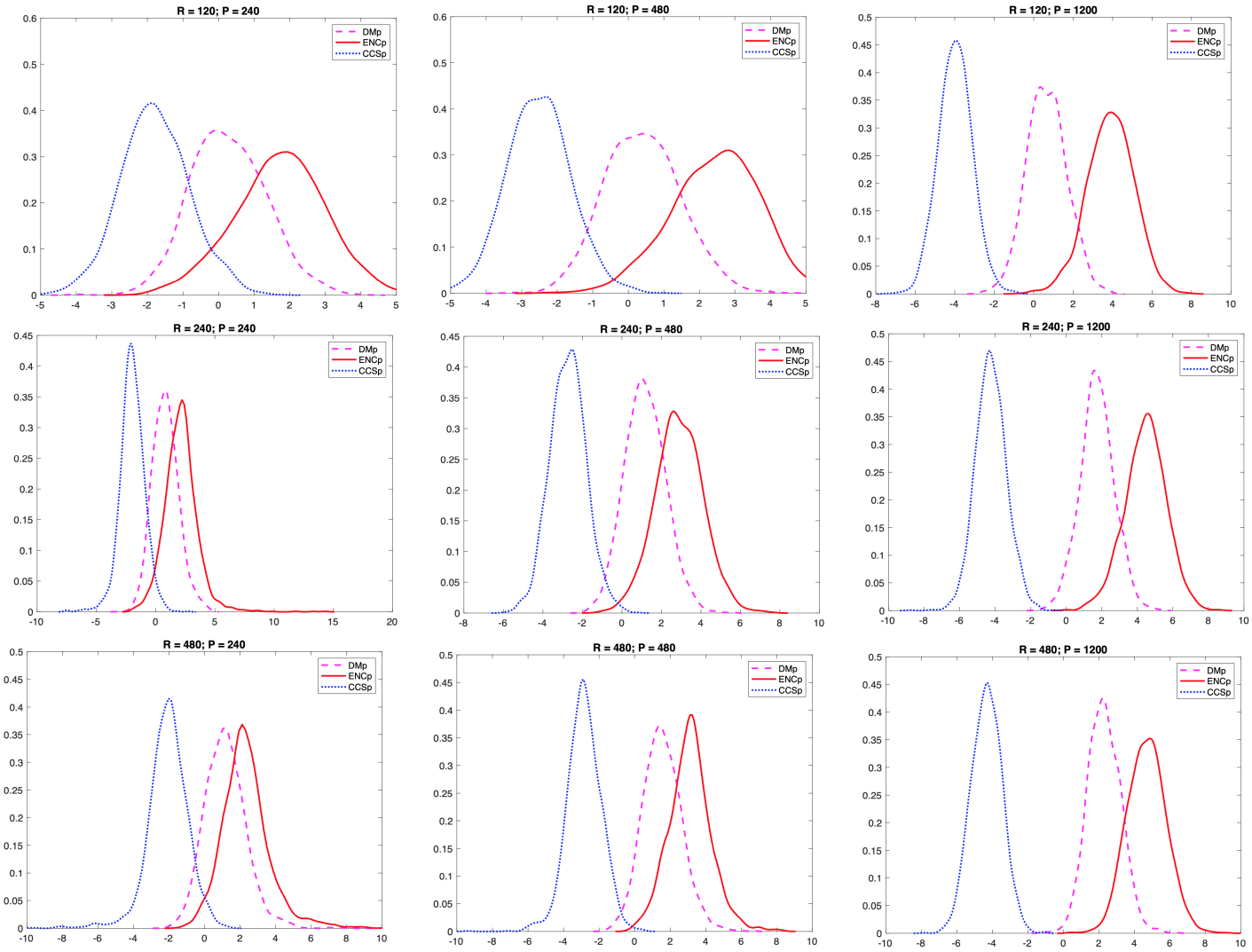

Figure 6.2: Simulation Results of DGP 1: Power of Test, $\alpha = 0.05$

and $\mathrm{ENC}_{R,P}$. Moreover, the asymptotic distribution of $\mathrm{DM}_{R,P}$ and $\mathrm{ENC}_{R,P}$ move to the right, and the means of $\mathrm{DM}_{R,P}$ are lower than means of $\mathrm{ENC}_{R,P}$.

6.6 Empirical Analysis

6.6.1 Empirical Analysis of VaR and ES for Monthly Equity Premium

In this subsection, we use the data from Welch and Goyal (2008) to check the predictive variable in VaR and ES for Equity Premium. We chose to use monthly data from January 1926 to December 2019, which contains 1128 observations. There is a total of 15 Macroeconomic and financial variables.

We build two nested models for VaR and ES to test if these 15 variables Granger cause the equity premium respectively. For Model 1, we find out the VaR and ES by using the previous R in-sample observations of the equity premium and then use estimated VaR and ES to predict the future equity premium at time $R+1$. For Model 2, we estimate the constant term and covariance term using the previous R in-sample observations of equity premium and an independent variable and then use these estimated coefficients and the independent variable to forecast the future equity premium. In this empirical work, the forecasts begin 40 years after the data are available, and the out-of-sample forecasts begin in 1965. We choose $\alpha \in \{0.01, 0.025, 0.05, 0.1\}$.

Tables 5 - 6 show the ENC statistics for 15 Macroeconomic and financial variables under four different α. In Table 5, forecasts begin 40 years after the data are available, which means the number of in-sample observations $R = 480$. From this table, when $\alpha = 0.01$, we can see that the dividend payout ratio is the only variable that Granger Causes the equity

premium in VaR and ES at a 10% significance level. When we choose $\alpha = 0.025$, the earning price ratio, stock variance, and default yield spread Granger Cause the equity premium on VaR and ES respectively at a 5% significance level. Moreover, the equity premium can be predicted by the dividend yield, treasury-bill rate, default return spread, and consumption, wealth, and income ratio in VaR and ES at $\alpha = 0.05$, respectively. When $\alpha = 0.1$, seven variables Granger Cause the equity premium in VaR and ES: dividend price ratio, dividend yield, earning price ratio, dividend payout ratio, default return spread, long-term return, and consumption, wealth, income ratio.

Table 6 demonstrates the ENC statistics for 15 variables when out-of-sample forecasts begin in 1965. From this table, when $\alpha = 0.01$, we can see that the term spread is the only variable that Granger Causes the equity premium in VaR and ES at a 10% significance level. The long-term return and earning price ratio Granger Cause the equity premium in VaR and ES at a 1% significance level when $\alpha = 0.025$ and $\alpha = 0.05$ respectively. Moreover, when $\alpha = 0.1$, the equity premium can be predicted by the dividend price ratio, dividend yield, earning price ratio, default return spread, and consumption, wealth, and income ratio. From Tables 5 - 6, we can see that the decreasing number of variables can Granger Cause the equity premium when we decrease α. Thus, fewer variables have the predictive ability of the equity premium in VaR and ES when we focus on extreme losses. Compared to forecasts beginning in 1965, the equity premium can be predicted by more variables at different α when forecasts begin 40 years after the data are available. Thus, more variables can Granger Cause the equity premium when the number of in-sample observations becomes small.

6.6.2 Empirical Analysis of GaR and GS for Quarterly GDP Growth

Measures of the downside risk are essential in risk management. The increasing number of policymakers has focused on the downside risk in the last decade. The International Monetary Fund (IMF) has recently popularized a risk measure for GDP growth called Growth-at-Risk (GaR). GaR is the worst conditional GDP growth distribution at a given coverage level (5th percentile) depending on financial conditions (Adrian et al., 2019). Moreover, Adrian et al. (2019); Chavleishvili et al. (2021) define a measure of adverse real economic impact to be Growth Shortfall (GS). GS is the expectation of the GDP growth when it is less than GaR. In order to check if financial conditions have the predictive ability of the macroeconomic risk, we choose to use the National Financial Conditions Index (NFCI), which is computed by the Federal Reserve Bank of Chicago. The NFCI is calculated from a dynamic factor model with 105 financial variables to estimate the weekly US financial conditions. Financial conditions are tighter than average when NFCI is positive, while financial conditions are looser than average when NFCI is negative.

In this subsection, we use the data from Adrian et al. (2019) to check if the NFCI can predict GDP growth on GaR and GS. The data are Quarterly from 1973Q1 to 2019Q4. There is a total of 188 observations. We consider GDP growth as our y_{t+h} and NFCI as our x_t. The test is an h-step out-of-sample Granger Causality test. We choose $h \in \{0, 1, 2, 3, 4, 5, 6, 7, 8\}$. Forecasts of GaR and GS begin 20 years after data are available, which means $R = 80$. The rolling window estimation is used in this test. We also use the FZ scoring function for a pair of GaR and GS to examine the forecast encompassing test for Granger Causality on GaR and GS.

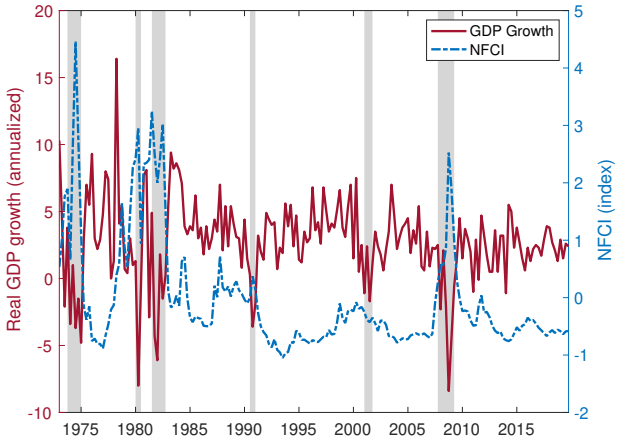

Figure 6.3: The times series of the GDP growth and the NFCI

Figure 3 shows the time series of the real GDP growth and the NFCI. This figure is the same as Figure 2 in Adrian et al. (2019). Due to the properties of the GDP growth and the NFCI, there is a negative relationship between these two variables. In the regular periods, when GDP growth is positive, GDP growth is more volatile than the NFCI. During the National Bureau of Economic Research (NBER) recession dates, the GDP growth reaches negative outcomes, and the NFCI reaches positive outcomes. Figure 3 demonstrates the negative relationship between GDP growth and NFCI. Thus, deteriorations in financial conditions coincide with the decreases in GDP growth. Then, we want to use the forecast encompassing test to examine if the NFCI Granger Causes the GDP growth on GaR and GS by using the FZ loss function.

Table 7 displays the ENC statistics and p-values of the Granger Causality test for various h when $\alpha = 0.05$. This table shows that the NFCI Granger Causes the four quarters ahead (one year) GDP growth on GaR and GS at the 10% significance level. It implies that the downside risk will increase in four quarters when the financial conditions

become tighter in the current quarter. However, the NFCI cannot Granger Cause GDP growth in the other quarters ahead.

Moreover, we want to explore if there are significant changes in $\hat{C}_{R,t}$ in different periods. Section 2 states that the encompassing statistic $\text{ENC}_{R,P}$ can be obtained by standardizing the $\hat{C}_{R,t}$. Thus, we plot $\hat{C}_{R,t}$ for various of h to check the change in $\hat{C}_{R,t}$ in the out-of-sample period when $\alpha = 0.05$. Figure 6 shows the time series of $\hat{C}_{R,t}$, $\hat{C}_{1,R,t}$, and $\hat{C}_{2,R,t}$, where $\hat{C}_{1,R,t} = \hat{m}_{1,t}\left(G\hat{a}R_{2,t} - G\hat{a}R_{1,t}\right)$ is the GaR term in the $\hat{C}_{R,t}$, and $\hat{C}_{2,R,t} = \hat{m}_{2,t}\left(\hat{GS}_{2,t} - \hat{GS}_{1,t}\right)$ is the GS term in the $\hat{C}_{R,t}$. From Figure 6, we can see that there are spikes of $\hat{C}_{R,t}$, $\hat{C}_{1,R,t}$, and $\hat{C}_{2,R,t}$ for various h during the NBER recession dates. The spike becomes smaller when the quarter ahead h becomes larger. When h is zero, current quarter GDP growth and NFCI are used. Then the spike is more considerable. The spike becomes negligible as the number of quarters ahead h goes up.

Furthermore, we check if there are differences between the forecast GaR without the NFCI and with the NFCI. Figure 4 demonstrates the changes of the $G\hat{a}R_{1,R,t}$, $G\hat{a}R_{2,R,t}$, $\hat{GS}_{1,R,t}$, and $\hat{GS}_{2,R,t}$ in the out-of-sample period, where $G\hat{a}R_{1,R,t}$ and $G\hat{a}R_{2,R,t}$ are the forecast GaR without the NFCI and with the NFCI, and $\hat{GS}_{1,R,t}$ and $\hat{GS}_{2,R,t}$ are the forecast GS without the NFCI and with the NFCI. This figure displays that the difference between $G\hat{a}R_{1,R,t}$ and $G\hat{a}R_{2,R,t}$ and the difference between $\hat{GS}_{1,R,t}$ and $\hat{GS}_{2,R,t}$ become small during the NBER recession dates when the number of quarters ahead h increases from 0 to 8. For the NBER recession dates from 2007Q4 to 2009Q2, the difference between $G\hat{a}R_{1,R,t}$ and $G\hat{a}R_{2,R,t}$ and the difference between $\hat{GS}_{1,R,t}$ and $\hat{GS}_{2,R,t}$ are large when $h = 0, 1, 2, 3, 4$.

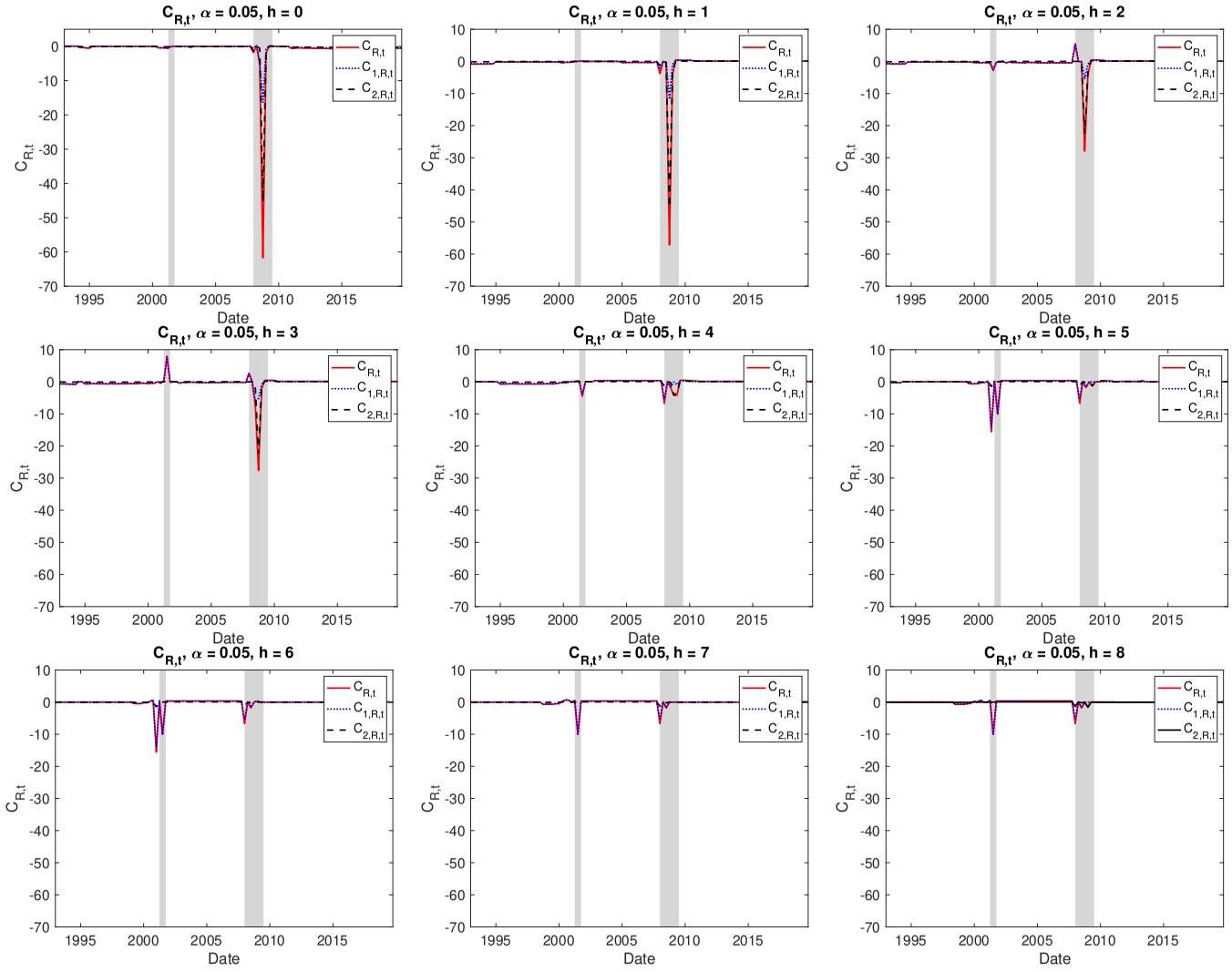

Figure 6.4: Empirical Results: The Time series of the $\hat{C}_{R,t}$ and $\hat{C}_{1,R,t}$

However, the difference between $\hat{GaR}_{1,R,t}$ and $\hat{GaR}_{2,R,t}$ is very small, and the difference between $\hat{GS}_{1,R,t}$ and $\hat{GS}_{2,R,t}$ is almost zero when $h = 5, 6, 7, 8$.

6.7 Conclusion

In this paper, I develop a new forecast encompassing test for Granger Causality for VaR and ES. The strictly consistent scoring function for VaR and ES is from Fissler et al. (2016) and Patton et al. (2019). I show that the DM statistic has a bias for the FZ scoring function with two nested models under the null hypothesis. I exhibit that the encompassing statistic (ENC) has zero mean under the null hypothesis with no Granger Causality. I prove that the ENC statistic is asymptotically standard normal. Monte Carlo simulation demonstrates that the ENC statistic has good size and has the highest power under the alternative hypothesis $\mathbb{H}_1$ in finite samples compared to DM and CCS statistics. I demonstrate that ENC statistic improves the predictive accuracy compared with the DM test. I am considering two applications. The first empirical analysis of VaR and ES illustrates some Macroeconomic and financial variables Granger Cause the equity premium in VaR and ES at different α. The second empirical analysis of GaR and GS shows that the NFCI Granger Causes the four quarters ahead of GDP growth on GaR and GS when α is 0.05.

6.8 Appendix A

A1: Proof of Theorm 1.

DM statistics of FZ loss-differential can be rewrite as follows:

$$\hat{D}_P = P^{-1}\sum_{t=R}^{T}\left[-\frac{1}{\alpha\hat{e}_{t+1}^{(1)}}\mathbf{1}\left\{Y_{t+1}\leq\hat{q}_{t+1}^{(1)}\right\}\left(\hat{q}_{t+1}^{(1)}-Y_{t+1}\right)+\frac{\hat{q}_{t+1}^{(1)}}{\hat{e}_{t+1}^{(1)}}+\log\left(-\hat{e}_{t+1}^{(1)}\right)\right.$$
$$\left.+\frac{1}{\alpha\hat{e}_{t+1}^{(2)}}\mathbf{1}\left\{Y_{t+1}\leq\hat{q}_{t+1}^{(2)}\right\}\left(\hat{q}_{t+1}^{(2)}-Y_{t+1}\right)-\frac{\hat{q}_{t+1}^{(2)}}{\hat{e}_{t+1}^{(2)}}-\log\left(-\hat{e}_{t+1}^{(2)}\right)\right]$$

In order to prove that $\hat{D}_P$ is non-zero at each t, we denote $\hat{D}_P = P^{-1}\sum_{t=R}^{T} d_t$ and show d_t is non-zero at each t.

(1). $Y_{t+1} > \hat{q}_{t+1}^{(1)}, \quad Y_{t+1} > \hat{q}_{t+1}^{(2)}, \quad \text{and} \quad \hat{e}_{t+1}^{(1)} > \hat{e}_{t+1}^{(2)}$

$$d_t = \underbrace{\frac{\hat{q}_{t+1}^{(1)}}{\hat{e}_{t+1}^{(1)}}}_{>0}\underbrace{-\frac{\hat{q}_{t+1}^{(2)}}{\hat{e}_{t+1}^{(2)}}}_{<0}+\underbrace{\left[\log\left(-\hat{e}_{t+1}^{(1)}\right)-\log\left(-\hat{e}_{t+1}^{(2)}\right)\right]}_{<0}$$

(2). $Y_{t+1} > \hat{q}_{t+1}^{(1)}, \quad Y_{t+1} > \hat{q}_{t+1}^{(2)}, \quad \text{and} \quad \hat{e}_{t+1}^{(1)} < \hat{e}_{t+1}^{(2)}$

$$d_t = \underbrace{\frac{\hat{q}_{t+1}^{(1)}}{\hat{e}_{t+1}^{(1)}}}_{>0}\underbrace{-\frac{\hat{q}_{t+1}^{(2)}}{\hat{e}_{t+1}^{(2)}}}_{<0}+\underbrace{\left[\log\left(-\hat{e}_{t+1}^{(1)}\right)-\log\left(-\hat{e}_{t+1}^{(2)}\right)\right]}_{>0}$$

(3). $Y_{t+1} \leq \hat{q}_{t+1}^{(1)}, \quad Y_{t+1} > \hat{q}_{t+1}^{(2)}, \quad \text{and} \quad \hat{e}_{t+1}^{(1)} > \hat{e}_{t+1}^{(2)}$

$$d_t = P^{-1}\sum_{t=R}^{T}\left[-\frac{\hat{q}_{t+1}^{(1)}-Y_{t+1}}{\alpha\hat{e}_{t+1}^{(1)}}+\frac{\hat{q}_{t+1}^{(1)}}{\hat{e}_{t+1}^{(1)}}+\log\left(-\hat{e}_{t+1}^{(1)}\right)-\frac{\hat{q}_{t+1}^{(2)}}{\hat{e}_{t+1}^{(2)}}-\log\left(-\hat{e}_{t+1}^{(2)}\right)\right]$$
$$= \underbrace{\frac{(\alpha-1)\,\hat{q}_{t+1}^{(1)}+Y_{t+1}}{\alpha\hat{e}_{t+1}^{(1)}}}_{<0}\underbrace{-\frac{\hat{q}_{t+1}^{(2)}}{\hat{e}_{t+1}^{(2)}}}_{<0}+\underbrace{\left[\log\left(-\hat{e}_{t+1}^{(1)}\right)-\log\left(-\hat{e}_{t+1}^{(2)}\right)\right]}_{<0}<0$$

(4). $Y_{t+1} \leq \hat{q}_{t+1}^{(1)}, \quad Y_{t+1} > \hat{q}_{t+1}^{(2)}, \quad \text{and} \quad \hat{e}_{t+1}^{(1)} < \hat{e}_{t+1}^{(2)}$

$$d_t = P^{-1}\sum_{t=R}^{T}\left[-\frac{\hat{q}_{t+1}^{(1)}-Y_{t+1}}{\alpha\hat{e}_{t+1}^{(1)}}+\frac{\hat{q}_{t+1}^{(1)}}{\hat{e}_{t+1}^{(1)}}+\log\left(-\hat{e}_{t+1}^{(1)}\right)-\frac{\hat{q}_{t+1}^{(2)}}{\hat{e}_{t+1}^{(2)}}-\log\left(-\hat{e}_{t+1}^{(2)}\right)\right]$$

$$= \underbrace{\frac{(\alpha-1)\,\hat{q}_{t+1}^{(1)} + Y_{t+1}}{\alpha \hat{e}_{t+1}^{(1)}}}_{<0} \underbrace{-\frac{\hat{q}_{t+1}^{(2)}}{\hat{e}_{t+1}^{(2)}}}_{<0} + \underbrace{\left[\log\left(-\hat{e}_{t+1}^{(1)}\right) - \log\left(-\hat{e}_{t+1}^{(2)}\right)\right]}_{>0}$$

(5). $Y_{t+1} > \hat{q}_{t+1}^{(1)}, \quad Y_{t+1} \le \hat{q}_{t+1}^{(2)}, \quad \text{and} \quad \hat{e}_{t+1}^{(1)} > \hat{e}_{t+1}^{(2)}$

$$d_t = P^{-1}\sum_{t=R}^{T}\left[\frac{\hat{q}_{t+1}^{(1)}}{\hat{e}_{t+1}^{(1)}} + \log\left(-\hat{e}_{t+1}^{(1)}\right) + \frac{\hat{q}_{t+1}^{(2)} - Y_{t+1}}{\alpha \hat{e}_{t+1}^{(2)}} - \frac{\hat{q}_{t+1}^{(2)}}{\hat{e}_{t+1}^{(2)}} - \log\left(-\hat{e}_{t+1}^{(2)}\right)\right]$$

$$= \underbrace{\frac{\hat{q}_{t+1}^{(1)}}{\hat{e}_{t+1}^{(1)}}}_{>0} + \underbrace{\frac{(1-\alpha)\,\hat{q}_{t+1}^{(2)} - Y_{t+1}}{\alpha \hat{e}_{t+1}^{(2)}}}_{>0} + \underbrace{\left[\log\left(-\hat{e}_{t+1}^{(1)}\right) - \log\left(-\hat{e}_{t+1}^{(2)}\right)\right]}_{<0}$$

(6). $Y_{t+1} > \hat{q}_{t+1}^{(1)}, \quad Y_{t+1} \le \hat{q}_{t+1}^{(2)}, \quad \text{and} \quad \hat{e}_{t+1}^{(1)} < \hat{e}_{t+1}^{(2)}$

$$d_t = P^{-1}\sum_{t=R}^{T}\left[\frac{\hat{q}_{t+1}^{(1)}}{\hat{e}_{t+1}^{(1)}} + \log\left(-\hat{e}_{t+1}^{(1)}\right) + \frac{\hat{q}_{t+1}^{(2)} - Y_{t+1}}{\alpha \hat{e}_{t+1}^{(2)}} - \frac{\hat{q}_{t+1}^{(2)}}{\hat{e}_{t+1}^{(2)}} - \log\left(-\hat{e}_{t+1}^{(2)}\right)\right]$$

$$= \underbrace{\frac{\hat{q}_{t+1}^{(1)}}{\hat{e}_{t+1}^{(1)}}}_{>0} + \underbrace{\frac{(1-\alpha)\,\hat{q}_{t+1}^{(2)} - Y_{t+1}}{\alpha \hat{e}_{t+1}^{(2)}}}_{>0} + \underbrace{\left[\log\left(-\hat{e}_{t+1}^{(1)}\right) - \log\left(-\hat{e}_{t+1}^{(2)}\right)\right]}_{>0} > 0$$

(7). $Y_{t+1} \le \hat{q}_{t+1}^{(1)}, \quad Y_{t+1} \le \hat{q}_{t+1}^{(2)}, \quad \text{and} \quad \hat{e}_{t+1}^{(1)} > \hat{e}_{t+1}^{(2)}$

$$d_t = P^{-1}\sum_{t=R}^{T}\left[-\frac{\hat{q}_{t+1}^{(1)} - Y_{t+1}}{\alpha \hat{e}_{t+1}^{(1)}} + \frac{\hat{q}_{t+1}^{(1)}}{\hat{e}_{t+1}^{(1)}} + \log\left(-\hat{e}_{t+1}^{(1)}\right) + \frac{\hat{q}_{t+1}^{(2)} - Y_{t+1}}{\alpha \hat{e}_{t+1}^{(2)}} - \frac{\hat{q}_{t+1}^{(2)}}{\hat{e}_{t+1}^{(2)}} - \log\left(-\hat{e}_{t+1}^{(2)}\right)\right]$$

$$= \underbrace{\frac{(\alpha-1)\,\hat{q}_{t+1}^{(1)} + Y_{t+1}}{\alpha \hat{e}_{t+1}^{(1)}}}_{<0} + \underbrace{\frac{(1-\alpha)\,\hat{q}_{t+1}^{(2)} - Y_{t+1}}{\alpha \hat{e}_{t+1}^{(2)}}}_{>0} + \underbrace{\left[\log\left(-\hat{e}_{t+1}^{(1)}\right) - \log\left(-\hat{e}_{t+1}^{(2)}\right)\right]}_{<0}$$

(8). $Y_{t+1} \le \hat{q}_{t+1}^{(1)}, \quad Y_{t+1} \le \hat{q}_{t+1}^{(2)}, \quad \text{and} \quad \hat{e}_{t+1}^{(1)} < \hat{e}_{t+1}^{(2)}$

$$d_t = P^{-1}\sum_{t=R}^{T}\left[-\frac{\hat{q}_{t+1}^{(1)} - Y_{t+1}}{\alpha \hat{e}_{t+1}^{(1)}} + \frac{\hat{q}_{t+1}^{(1)}}{\hat{e}_{t+1}^{(1)}} + \log\left(-\hat{e}_{t+1}^{(1)}\right) + \frac{\hat{q}_{t+1}^{(2)} - Y_{t+1}}{\alpha \hat{e}_{t+1}^{(2)}} - \frac{\hat{q}_{t+1}^{(2)}}{\hat{e}_{t+1}^{(2)}} - \log\left(-\hat{e}_{t+1}^{(2)}\right)\right]$$

$$= \underbrace{\frac{\hat{q}_{t+1}^{(1)}}{\hat{e}_{t+1}^{(1)}}}_{>0} + \underbrace{\frac{(1-\alpha)\,\hat{q}_{t+1}^{(2)} - Y_{t+1}}{\alpha \hat{e}_{t+1}^{(2)}}}_{>0} + \underbrace{\left[\log\left(-\hat{e}_{t+1}^{(1)}\right) - \log\left(-\hat{e}_{t+1}^{(2)}\right)\right]}_{>0}$$

Thus, we find out$\mathbb{E}\hat{D}_P \neq 0$, so the DM statistics has a bias.

A2: Proof of Theorem 2.

We estimate the FZ scoring function with combined error term to find out the the encompassing statistic. We have

$$\lambda = \arg\min_{\lambda} \mathbb{E} S_\alpha \left(q_{t+1}^{(c)}, e_{t+1}^{(c)}, Y_{t+1} \right)$$

where $q_{t+1}^{(c)} = (1-\lambda)\, q_{t+1}^{(1)} + \lambda q_{t+1}^{(2)}$ and $e_{t+1}^{(c)} = (1-\lambda)\, e_{t+1}^{(1)} + \lambda \hat{e}_{t+1}^{(2)}$. First, we rewrite the expectation of the FZ scoring function with combined functionals $q_{t+1}^{(c)}$ and $e_{t+1}^{(c)}$ as follows:

$$\begin{aligned}
\mathbb{E}_F S_\alpha \left(q_{t+1}^{(c)}, e_{t+1}^{(c)}, Y_{t+1} \right) &= \int_{-\infty}^{+\infty} -\frac{1}{\alpha e_{t+1}^{(c)}} \mathbf{1}\{Y_{t+1} \le q_{t+1}^{(c)}\}(q_{t+1}^{(c)} - Y_{t+1}) \mathrm{d}F_t(Y_{t+1}) + \frac{q_{t+1}^{(c)}}{e_{t+1}^{(c)}} \\
&\quad + \log(-e_{t+1}^{(c)}) - 1 \\
&= \int_{-\infty}^{q_{t+1}^{(c)}} -\frac{1}{\alpha e_{t+1}^{(c)}} (q_{t+1}^{(c)} - Y_{t+1}) \mathrm{d}F_t(Y_{t+1}) + \frac{q_{t+1}^{(c)}}{e_{t+1}^{(c)}} + \log(-e_{t+1}^{(c)}) - 1 \\
&= \int_{-\infty}^{0} \frac{1}{\alpha e_{t+1}^{(c)}} Z_{t+1} \mathrm{d}F_t(Z_{t+1} + q_{t+1}^{(c)}) + \frac{q_{t+1}^{(c)}}{e_{t+1}^{(c)}} + \log(-e_{t+1}^{(c)}) - 1 \\
&= Z_{t+1} F_t(Z_{t+1} + q_{t+1}^{(c)})|_{-\infty}^{0} - \int_{-\infty}^{0} \frac{1}{\alpha e_{t+1}^{(c)}} F_t(Z_{t+1} + q_{t+1}^{(c)}) \mathrm{d}Z_{t+1} \\
&\quad + \frac{q_{t+1}^{(c)}}{e_{t+1}^{(c)}} + \log(-e_{t+1}^{(c)}) - 1 \\
&= -\frac{1}{\alpha e_{t+1}^{(c)}} \int_{-\infty}^{0} F_t(Z_{t+1} + q_{t+1}^{(c)}) \mathrm{d}Z_{t+1} + \frac{q_{t+1}^{(c)}}{e_{t+1}^{(c)}} + \log(-e_{t+1}^{(c)}) - 1
\end{aligned}$$

Taking derivative of expectation of the FZ scoring function with combined functionals $q^{(c)}$ and $e^{(c)}$ with respect to combined functionals $q^{(c)}$ and $e^{(c)}$, we have

$$
\begin{aligned}
m_{1,t+1}^{(c)} &= \frac{\partial \mathbb{E}_F\left[S_\alpha\left(q_{t+1}^{(c)}, e_{t+1}^{(c)}, Y_{t+1}\right)\right]}{\partial q_{t+1}^{(c)}} \\
&= -\frac{1}{\alpha e_{t+1}^{(c)}} \int_{-\infty}^{0} f_t(Z_{t+1} + q_{t+1}^{(c)}) \mathrm{d}Z_{t+1} + \frac{1}{e_{t+1}^{(c)}} \\
&= -\frac{1}{\alpha e_{t+1}^{(c)}} \int_{-\infty}^{q_{t+1}^{(c)}} f_t(Y_{t+1}) \mathrm{d}Y_{t+1} + \frac{1}{e_{t+1}^{(c)}} \\
&= -\frac{1}{\alpha e_{t+1}^{(c)}} F_t(q_{t+1}^{(c)}) + \frac{1}{e_{t+1}^{(c)}} \\
&= \mathbb{E}_F\left[-\frac{1}{\alpha e_{t+1}^{(c)}} \mathbf{1}\{Y_t \leq q_{t+1}^{(c)}\} + \frac{1}{e_{t+1}^{(c)}}\right] \\
m_{2,t+1}^{(c)} &= \frac{\partial \mathbb{E}_F\left[S_\alpha\left(q_{t+1}^{(c)}, e_{t+1}^{(c)}, Y_{t+1}\right)\right]}{\partial e_{t+1}^{(c)}} \\
&= \frac{1}{\alpha\left(e_{t+1}^{(c)}\right)^2} \int_{-\infty}^{+\infty} \mathbf{1}\{Y_{t+1} \leq q_{t+1}^{(c)}\}(q_{t+1}^{(c)} - Y_{t+1}) \mathrm{d}F_t(Y_{t+1}) - \frac{q_{t+1}^{(c)}}{\left(e_{t+1}^{(c)}\right)^2} + \frac{1}{e_{t+1}^{(c)}} \\
&= \mathbb{E}_F\left[\frac{1}{\alpha\left(e_{t+1}^{(c)}\right)^2} \mathbf{1}\left\{Y_{t+1} \leq \hat{q}_{t+1}^{(c)}\right\}\left(q_{t+1}^{(c)} - Y_{t+1}\right) - \frac{q_{t+1}^{(c)}}{\left(e_{t+1}^{(c)}\right)^2} + \frac{1}{\hat{e}_{t+1}^{(c)}}\right]
\end{aligned}
$$

Taking derivative of combined functionals $q^{(c)}$ and $e^{(c)}$ with respect to weight λ, we obtain

$$
\begin{aligned}
\frac{\partial q_{t+1}^{(c)}}{\partial \lambda} &= \left(q_{t+1}^{(2)} - q_{t+1}^{(1)}\right) \\
\frac{\partial e_{t+1}^{(c)}}{\partial \lambda} &= \left(e_{t+1}^{(2)} - e_{t+1}^{(1)}\right)
\end{aligned}
$$

In order to find the optimal weight λ^*, we take the first order condition of the expected FZ scoring function with respect to λ.

F.O.C w.r.t λ:

$$
\begin{aligned}
C &\equiv \frac{\partial \mathbb{E}_F \left[S_\alpha \left(q_{t+1}^{(c)}, e_{t+1}^{(c)}, Y_{t+1} \right) \right]}{\partial \lambda} \\
&= \frac{\partial \mathbb{E}_F S_\alpha \left(q_{t+1}^{(c)}, e_{t+1}^{(c)}, Y_{t+1} \right)}{\partial q_{t+1}^{(c)}} \frac{\partial q_{t+1}^{(c)}}{\partial \lambda} + \frac{\partial \mathbb{E}_F S_\alpha \left(q_{t+1}^{(c)}, e_{t+1}^{(c)}, Y_{t+1} \right)}{\partial e_{t+1}^{(c)}} \frac{\partial e_{t+1}^{(c)}}{\partial \lambda} \\
&= \mathbb{E}_F \left[m_{1,t+1}^{(1)} \left(q_{t+1}^{(2)} - q_{t+1}^{(1)} \right) + m_{2,t+1}^{(c)} \left(e_{t+1}^{(2)} - e_{t+1}^{(1)} \right) \right] \\
&= \mathbb{E}_F \left[m_{1,t+1}^{(1)} \left(q_{t+1}^{(2)} - q_{t+1}^{(1)} \right) + m_{2,t+1}^{(2)} \left(e_{t+1}^{(2)} - e_{t+1}^{(1)} \right) \right] = 0 \qquad \text{under } \mathbb{H}_0
\end{aligned}
$$

$$
\begin{aligned}
\hat{C}_P &= P^{-1} \sum_{t=R}^{T} \left[\hat{m}_{1,t+1}^{(c)} \left(\hat{q}_{t+1}^{(1)} - \hat{q}_{t+1}^{(2)} \right) + \hat{m}_{2,t+1}^{(c0} \left(\hat{e}_{t+1}^{(1)} - \hat{e}_{t+1}^{(2)} \right) \right] \\
&= P^{-1} \sum_{t=R}^{T} \left[\hat{m}_{1,t+1}^{(1)} \left(\hat{q}_{t+1}^{(2)} - \hat{q}_{t+1}^{(1)} \right) + \hat{m}_{2,t+1}^{(2)} \left(\hat{e}_{t+1}^{(2)} - \hat{e}_{t+1}^{(1)} \right) \right] \\
&\xrightarrow{p} C = 0 \quad \text{under } \mathbb{H}_0 \quad \text{as } R, P \to \infty \text{ and } P/R \to \infty.
\end{aligned}
$$

Under $\mathbb{H}_0, \lambda = 0$, then we have $q_{t+1}^{(c)} = q_{t+1}^{(1)}$ and $e_{t+1}^{(c)} = e_{t+1}^{(1)}$. Thus, $\mathbb{E}(\hat{C}_P) = 0$, so $\hat{C}_P \xrightarrow{p} 0$, as $R, P \to \infty$ and $R, P \to \infty$.

Appendix B: Proof of Theorem 3

We have the combined VaR $q_{t+1}^{(c)} = (1-\lambda)q_{t+1}^{(1)} + \lambda q_{t+1}^{(2)}$ and ES $e_{t+1}^{(c)} = (1-\lambda)e_{t+1}^{(1)} + \lambda e_{t+1}^{(2)}$. The expected loss function from combining forecast is

$$\begin{aligned} &\mathbb{E}_F S\left(q_{t+1}^{(c)}, e_{t+1}^{(c)}, Y_{t+1}\right) \\ =&\mathbb{E}_F\left[-\frac{1}{\alpha e_{t+1}^{(c)}}\mathbf{1}\{Y_{t+1} \le q_{t+1}^{(c)}\}\left(q_{t+1}^{(c)} - Y_{t+1}\right) + \frac{q_{t+1}^{(c)}}{e_{t+1}^{(c)}} + \log\left(-e_{t+1}^{(c)}\right) - 1\right] \end{aligned} \tag{6.21}$$

Taking the first order condition of Equation (6.21) with respect to λ, we have

$$C^{(c)} \equiv \nabla_\lambda \mathbb{E} S\left(q_{t+1}^{(c)}, e_{t+1}^{(c)}, Y_{t+1}\right) = \mathbb{E}\left[m_{1,t+1}^{(c)}\left(q_{t+1}^{(2)} - q_{t+1}^{(1)}\right) + m_{2,t+1}^{(c)}\left(e_{t+1}^{(2)} - e_{t+1}^{(1)}\right)\right] = 0$$

To show the necessary condition, if $\lambda = 0$, then $\hat{q}_{t+1}^{(c)} = \hat{q}_{t+1}^{(1)}$ and $\hat{e}_{t+1}^{(c)} = \hat{e}_{t+1}^{(1)}$, so that $\hat{m}_{1,t+1}^{(c)} = \hat{m}_{1,t+1}^{(1)}$, $\hat{m}_{2,t+1}^{(c)} = \hat{m}_{2,t+1}^{(1)}$. we have

$$\hat{C}_P^{(c)} \equiv P^{-1}\sum_{t=R}^{T}\left[\hat{m}_{1,t+1}^{(1)}\left(q_{t+1}^{(2)} - q_{t+1}^{(1)}\right) + \hat{m}_{2,t+1}^{(1)}\left(e_{t+1}^{(2)} - e_{t+1}^{(1)}\right)\right] \xrightarrow{p} 0$$

as shown in Theorem 2 under $\mathbb{H}_0$. To show the sufficient condition, taking the second order condition of Equation (6.21) with respect to λ, we have

$$\begin{aligned} &\frac{\partial^2 \mathbb{E}_F S\left(q_{t+1}^{(c)}, e_{t+1}^{(c)}, Y_{t+1}\right)}{\partial \lambda^2} = \frac{\partial C^{(c)}}{\partial \lambda} \\ =&\mathbb{E}_F\left[\frac{1}{\alpha\left(e_{t+1}^{(c)}\right)^3}\left[\alpha - 1\{Y_{t+1} \le q_{t+1}^{(c)}\}\right]\left[e_{t+1}^{(c)}\left(q_{t+1}^{(2)} - q_{t+1}^{(1)}\right) + 2\left(e_{t+1}^{(2)} - e_{t+1}^{(1)}\right)\right]\left(e_{t+1}^{(2)} - e_{t+1}^{(1)}\right)\right. \\ &\left. -\frac{2}{\left(e_{t+1}^{(c)}\right)^3}\left[e_{t+1}^{(c)} - \frac{1}{\alpha}Y_{t+1}1\{Y_{t+1} \le q_{t+1}^{(c)}\}\right]\left(e_{t+1}^{(2)} - e_{t+1}^{(1)}\right)^2 + \frac{1}{\left(e_{t+1}^{(c)}\right)^2}\left(e_{t+1}^{(2)} - e_{t+1}^{(1)}\right)^2\right] \\ =&\mathbb{E}_F\left[\frac{1}{\left(e_{t+1}^{(c)}\right)^2}\left(e_{t+1}^{(2)} - e_{t+1}^{(1)}\right)^2\right] > 0, \end{aligned}$$

where $\mathbb{E}_F\left[\alpha - 1\{Y_{t+1} \le q_{t+1}^{(c)}\}\right] = 0$ and $\mathbb{E}_F\left[e_{t+1}^{(c)} - \frac{1}{\alpha}Y_{t+1}1\{Y_{t+1} \le q_{t+1}^{(c)}\}\right] = 0$. Therefore there is a unique value $\hat{\lambda} \to 0$ such that $\hat{C}_P^{(c)} \to 0$.

Appendix C: Proof of Theorem 4

Under $\mathbb{H}_0, u_{t+1}^{(1)} = u_{t+1}^{(2)} = u_{t+1}$. We define $g\left(u_{t+1}^{(i)}\right) = \left[\alpha - 1\left(u_{t+1}^{(i)} < 0\right)\right]$ and the score $k_{i,t}(\beta_{i,t})$ as follows:

$$\begin{aligned}
k_{i,t+1}(\beta_{i,t}) =& \frac{\partial L\left(Y_{t+1}, q_{t+1}(\beta_{i,t}), e_{t+1}(\beta_{i,t}); \alpha\right)}{\partial \beta_{i,t}} \\
=& \nabla q_{t+1}(\beta_{i,t})' \frac{1}{\alpha e_{t+1}(\beta_{i,t})} g\left(u_{t+1}^{(i)}\right) \\
& - \nabla e_{t+1}(\beta_{i,t})' \frac{1}{\alpha e_{t+1}(\beta_{i,t})^2}\left[g\left(u_{t+1}^{(i)}\right) q_{t+1}(\beta_{i,t}) + (\alpha - g\left(u_t^{(i)}\right))Y_{t+1} + \alpha e_{t+1}(\beta_{i,t})\right]
\end{aligned}$$

Then, let $h_{i,t+1}$ and $H_i(t+1)$ to be following.

$$\begin{aligned}
h_{i,t} =& - k_{i,t+1}\left(\beta_{i,t}\right) \\
=& \nabla q_{t+1}(\beta_{i,t})' \frac{1}{-\alpha e_{t+1}(\beta_{i,t})} g\left(u_{t+1}^{(i)}\right) \\
& + \nabla e_{t+1}(\beta_{i,t})' \frac{1}{\alpha e_{t+1}(\beta_{i,t})^2}\left[g\left(u_{t+1}^{(i)}\right) q_{t+1}(\beta_{i,t}) + \left(\alpha - g\left(u_{t+1}^{(i)}\right)\right) Y_{t+1} + \alpha e_{t+1}(\beta_{i,t})\right] \\
H_i(t+1) =& - \frac{1}{R}\sum_{t=1}^{R} k_{i,t+1}\left(\beta\right) \\
=& \frac{1}{R}\sum_{t=1}^{R}\Bigg\{\nabla q_{t+1}(\beta_{i,t})' \frac{1}{-\alpha e_{t+1}(\beta_{i,t})} g\left(u_{t+1}^{(i)}\right) \\
& + \nabla e_{t+1}(\beta_{i,t})' \frac{1}{\alpha e_{t+1}(\beta_{i,t})^2}\left[g\left(u_{t+1}^{(i)}\right) q_{t+1}(\beta_{i,t}) + (\alpha - g\left(u_{t+1}^{(i)}\right))Y_{t+1}\right. \\
& \left. + \alpha e_{t+1}(\beta_{i,t})\right]\Bigg\}
\end{aligned}$$

We denote the Hessian to be

$$\begin{aligned}
B_i^{-1} =& \mathbb{E}_F \Lambda(\beta_{i,t}) \\
=& \frac{\partial \mathbb{E}_F\left[k_{i,t}(\beta_{i,t})\right]}{\partial \beta_{i,t}} \\
=& \mathbb{E}\left[\frac{f_{t+1}\left(q_{t+1}(\beta_{i,t})\right)}{-\alpha e_{t+1}(\beta_{i,t})} \nabla q_{t+1}(\beta_{i,t})' \nabla q_t(\beta_{i,t}) + \frac{1}{e_{t+1}(\beta_{i,t})^2} \nabla e_{t+1}(\beta_{i,t})' \nabla e_{t+1}(\beta_{i,t})\right]
\end{aligned}$$

Therefore, $B_1 = 1$ and $B_2 = \begin{pmatrix} 1 & 0 \\ 0 & \mathbb{E}_F\Lambda(\beta_{2,t}) \end{pmatrix}^{-1}$. Let J denote the selection matrix (1,0) such that $Jh_{2,t} = h_{1,t}$. Let $\sup_t$ denote $\sup_{t-R+1\leq\cdot\leq t}$. And matrix A and C will defined in Lemma 3 and $\tilde{h}_{2,t} = Z^{-1}A'CB_2^{1/2}h_{2,t}, \tilde{H}_2(t) = Z^{-1}A'CB_2^{1/2}H_2(t)$. U_t and $\tilde{U}_t$ will defined in Assumption 2. By the central limit theorem, $\sqrt{R}\left[R^{-1}\sum_{s=t-R+1}^{t} h_{2,s}\right] \sim N\left(0, Z^2B_2^{-1}\right)$, where $Z^2B_2^{-1} = \mathbb{E}[k_t(\theta)k_t(\theta)']$. We use $\sum_t$ instead of $\sum_{j=t-R+1}^{t}$ and use $\hat{q}_{t+1}^{(i)}$ and $\hat{e}_{t+1}^{(i)}$ instead of $q_{t+1}(\hat{\beta}_{i,t})$ and $e_{t+1}(\hat{\beta}_{i,t})$ for simplicity.

Lemma 1: (Clark and McCracken, 2001) Let Assumptions 1-4 hold.

For i=1,2, $\sum_t H_i'(t)B_i(t)h_{i,t+1}h_{i,t+1}'B_j(t)H_j'(t) = \sum_t H_i'(t)B_i\mathbb{E}(h_{i,t+1}h_{i,t+1}')B_jH_j'(t) + o_p(t)$.

Proof: Adding and subtracting $\mathbb{E}(h_{i,t+1}h_{i,t+1}')$,

$$\sum_t H_i'(t)B_i(t)h_{i,t+1}h_{i,t+1}'B_j(t)H_j(t)$$

$$= \sum_t H_i'(t)B_i\mathbb{E}(h_{i,t+1}h_{i,t+1}')B_jH_j(t) + \sum_t H_i'(t)B_i\left(h_{i,t+1}h_{i,t+1}' - \mathbb{E}(h_{i,t+1}h_{i,t+1}')\right)B_jH_j(t).$$

Now, we need to show the second term is $o_p(1)$.

$$\sum_t H_i'(t)B_i\left(h_{i,t+1}h_{i,t+1}' - \mathbb{E}(h_{i,t+1}h_{i,t+1}')\right)B_jH_j(t)$$

$$= T^{-1/2}\sum_t (\frac{T}{t})^2\left[T^{-1/2}\sum_{s=1}^{t} h_{j,s}'B_j \otimes T^{-1/2}\sum_{s=1}^{t} h_{i,s}'B_i\right]$$

$$\times \text{vec}\left[T^{-1/2}\left(h_{i,t+1}h_{i,t+1}' - \mathbb{E}(h_{i,t+1}h_{i,t+1}')\right)\right]$$

$\sum_t(\frac{T}{t})^2\left[\frac{1}{\sqrt{T}}\sum_{s=1}^{t} h_{j,s}'B_j \otimes \frac{1}{\sqrt{T}}\sum_{s=1}^{t} h_{i,s}'B_i\right]\text{vec}\left[\frac{1}{\sqrt{T}}\left(h_{i,t+1}h_{i,t+1}' - \mathbb{E}(h_{i,t+1}h_{i,t+1}')\right)\right]$ is $Op(1)$ follows from Assumption 2, Corollary 29.19 of Davidson (1994) and Theorem 3.1 of

Hansen (1992). The proof is complete.

Lemma 2: Let Assumptions 1-3 hold. (a) Let $-J'B_1J+B_2=M$ and $B_2^{-1/2}MB_2^{-1/2}=Q$, then Q is idempotent. (b). Define matrix $A=(0,1)'$ and $C=I_{2\times 2}$ Then $Q=CAA'C$.

Proof: (a).

$$\begin{aligned} Q &\equiv B_2^{-1/2}MB_2^{-1/2} \\ &= \begin{pmatrix} 1 & 0 \\ 0 & \mathbb{E}_F\Lambda(\beta_{2,t}) \end{pmatrix}^{1/2} \begin{pmatrix} 1 & 0 \\ 0 & \mathbb{E}_F\Lambda(\beta_{2,t}) \end{pmatrix}^{-1} \begin{pmatrix} 1 & 0 \\ 0 & \mathbb{E}_F\Lambda(\beta_{2,t}) \end{pmatrix}^{1/2} \\ &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \end{aligned}$$

so Q is an idempotent matrix with rank 1.

(b). Lemma A4(b) of Clark and McCracken (2001).

Lemma 3: (Clark and McCracken, 2001) Let Assumptions 1-4 hold.

$$\sum_t \tilde{H}_2'(t)\tilde{h}_{2,t+1} \xrightarrow{d} \xi^{-1}\int_\xi^1 [W(s)-W(s-\xi)]\,\mathrm{d}W(s),$$

where $\xi=R/T$ and $\sum_t$ denotes $\sum_{t=R+1}^T$ (similarly hereinafter).

Proof:

$$\begin{aligned}\sum_t \tilde{H}_2'(t)\tilde{h}_{2,t+1} &= \sum_t \left(\frac{1}{R}\sum_{j=t-R+1}^{t}\tilde{h}_{2,j}'\right)\tilde{h}_{2,t+1} \\ &= \sum_t \left(\frac{1}{R}\sum_{j=1}^{t}\tilde{h}_{2,j}' - \frac{1}{R}\sum_{j=1}^{t-R}\tilde{h}_{2,j}'\right)\tilde{h}_{2,t+1} \\ &= \frac{T}{R}\left[\sum_t \left(T^{-1/2}\sum_{j=1}^{t}\tilde{h}_{2,j}' - T^{-1/2}\sum_{j=1}^{t-R}\tilde{h}_{2,j}'\right)\left(T^{-1/2}\tilde{h}_{2,t+1}\right)\right]\end{aligned}$$

According to the continues mapping theorem and Corollary 29.19 of Davidson (1994),

$T^{-1/2}\sum_{j=1}^{t}\tilde{h}_{2,j}' \Longrightarrow W(s)$, $\quad T^{-1/2}\sum_{j=1}^{t-R}\tilde{h}_{2,j}' \Longrightarrow W(s-\xi)$ and $T^{-1/2}\tilde{h}_{2,t+1} \Longrightarrow \mathrm{d}W(s)$.

Since $T/R = \xi^{-1}$, the proof is complete.

Lemma 4: (Clark and McCracken, 2001) Let Assumptions 1-4 hold.

$$\sum_t \tilde{H}_2'(t)\tilde{H}_2(t) \xrightarrow{d} \xi^{-2}\int_{\xi}^{1}\left[W(s) - W(s-\xi)\right]^2 \mathrm{d}s.$$

Proof:

$$\begin{aligned}\sum_t \tilde{H}_2'(t)\tilde{H}_2(t) &= \sum_t \left(\frac{1}{R}\sum_{j=t-R+1}^{t}\tilde{h}_{2,j}'\right)'\left(\frac{1}{R}\sum_{j=t-R+1}^{t}\tilde{h}_{2,j}'\right) \\ &= \sum_t \left(\frac{1}{R}\sum_{j=1}^{t}\tilde{h}_{2,j}' - \frac{1}{R}\sum_{j=1}^{t-R}\tilde{h}_{2,j}'\right)'\left(\frac{1}{R}\sum_{j=1}^{t}\tilde{h}_{2,j}' - \frac{1}{R}\sum_{j=1}^{t-R}\tilde{h}_{2,j}'\right) \\ &= \frac{1}{T}\left(\frac{T}{R}\right)^2\sum_t \left(T^{-1/2}\sum_{j=1}^{t}\tilde{h}_{2,j}' - T^{-1/2}\sum_{j=1}^{t-R}\tilde{h}_{2,j}'\right)' \\ &\quad \times \left(T^{-1/2}\sum_{j=1}^{t}\tilde{h}_{2,j}' - T^{-1/2}\sum_{j=1}^{t-R}\tilde{h}_{2,j}'\right)\end{aligned}$$

According to the continuous mapping theorem and Corollary 29.19 of Davidson (1994),

$T^{-1/2}\sum_{j=1}^{t}\tilde{h}_{2,j}' \Longrightarrow W(s), T^{-1/2}\sum_{j=1}^{t-R}\tilde{h}_{2,j}' \Longrightarrow W(s-\xi), T^{-1/2}\tilde{h}_{2,t+1} \Longrightarrow \mathrm{d}W(s)$ and

$1/T \Longrightarrow ds$. Since $T/R = \xi^{-1}$, the proof is complete.

Lemma 5: Let Assumptions 1-4 hold.

$$
\begin{aligned}
&\sum_t \left[\hat{m}_{1,t+1}\left(\hat{q}^{(2)}_{t+1} - \hat{q}^{(1)}_{t+1}\right) + \hat{m}_{2,t+1}\left(\hat{e}^{(2)}_{t+1} - \hat{e}^{(1)}_{t+1}\right)\right] \\
&= \sum_t \left[\left(-\frac{1}{\alpha \hat{e}^{(1)}_{t+1}}\mathbf{1}\{Y_{t+1} \le \hat{v}^{(1)}_{t+1}\} + \frac{1}{\hat{e}^{(1)}_{t+1}}\right)\left(\hat{q}^{(2)}_{t+1} - \hat{q}^{(1)}_{t+1}\right)\right. \\
&\quad \left. + \left(\frac{1}{\alpha\left(\hat{e}^{(1)}_{t+1}\right)^2}\mathbf{1}\left\{Y_{t+1} \le \hat{q}^{(1)}_{t+1}\right\}\left(\hat{q}^{(1)}_{t+1} - Y_{t+1}\right) - \frac{\hat{q}^{(1)}_{t+1}}{\left(\hat{e}^{(1)}_{t+1}\right)^2} + \frac{1}{\hat{e}^{(1)}_{t+1}}\right)\left(\hat{e}^{(2)}_{t+1} - \hat{e}^{(1)}_{t+1}\right)\right] \\
&= Z^2 \sum_t \tilde{H}_2'(t)\tilde{h}_{2,t+1} + o_p(1)
\end{aligned}
$$

Proof:

$$
\begin{aligned}
&\sum_t \left[\left(-\frac{1}{\alpha \hat{e}^{(1)}_{t+1}}\mathbf{1}\{Y_{t+1} \le \hat{q}^{(1)}_{t+1}\} + \frac{1}{\hat{e}^{(1)}_{t+1}}\right)\left(\hat{q}^{(2)}_{t+1} - \hat{q}^{(1)}_{t+1}\right)\right. \\
&+ \left.\left(\frac{1}{\alpha\left(\hat{e}^{(1)}_{t+1}\right)^2}\mathbf{1}\left\{Y_{t+1} \le \hat{q}^{(1)}_{t+1}\right\}\left(\hat{q}^{(1)}_{t+1} - Y_{t+1}\right) - \frac{\hat{q}^{(1)}_{t+1}}{\left(\hat{e}^{(1)}_{t+1}\right)^2} + \frac{1}{\hat{e}^{(1)}_{t+1}}\right)\left(\hat{e}^{(2)}_{t+1} - \hat{e}^{(1)}_{t+1}\right)\right] \\
&= \sum_t \left\{\frac{1}{\alpha \hat{e}^{(1)}_{t+1}}\left(\alpha - \mathbf{1}\{Y_{t+1} \le \hat{q}^{(1)}_{t+1}\}\right)\left(\hat{q}^{(2)}_{t+1} - \hat{q}^{(1)}_{t+1}\right)\right. \\
&- \left.\frac{1}{\alpha\left(\hat{e}^{(1)}_{t+1}\right)^2}\left[\left(\alpha - \mathbf{1}\left\{Y_{t+1} \le \hat{q}^{(1)}_{t+1}\right\}\right)\hat{q}^{(1)}_{t+1} + \mathbf{1}\left\{Y_{t+1} \le \hat{q}^{(1)}_{t+1}\right\}Y_{t+1} - \alpha \hat{e}^{(1)}_{t+1}\right]\left(\hat{e}^{(2)}_{t+1} - \hat{e}^{(1)}_{t+1}\right)\right\} \\
&= \sum_t \left\{\frac{1}{\alpha \hat{e}^{(1)}_{t+1}} g\left(\hat{u}^{(1)}_{t+1}\right)\left(\hat{q}^{(2)}_{t+1} - \hat{q}^{(1)}_{t+1}\right)\right. \\
&- \left.\frac{1}{\alpha\left(\hat{e}^{(1)}_{t+1}\right)^2}\left[g\left(\hat{u}^{(1)}_{t+1}\right)\hat{q}^{(1)}_{t+1} + \left(\alpha - g\left(\hat{u}^{(1)}_{t+1}\right)\right)Y_{t+1} - \alpha \hat{e}^{(1)}_{t+1}\right]\left(\hat{e}^{(2)}_{t+1} - \hat{e}^{(1)}_{t+1}\right)\right\}
\end{aligned}
$$

Taking Taylor Expansion and using generalized function by Galfand and Shilov, every term in the last equation becomes

$$\frac{1}{\alpha\hat{e}_{t+1}^{(1)}} = \frac{1}{\alpha e_{t+1}^{(1)}} - \frac{\nabla e_{t+1}^{(1)\prime}}{\alpha\left(e_{t+1}^{(1)}\right)^2}\left(\hat{\beta}_1 - \beta_1\right) + O_p\left(\left(\hat{\beta}_{1,t} - \beta_1\right)^2\right)$$

$$g\left(\hat{u}_{t+1}^{(1)}\right) = g\left(u_{t+1}^{(1)}\right) + \delta\left(u_{t+1}^{(1)}\right)\left(\hat{u}_{t+1}^{(1)} - u_{t+1}^{(1)}\right) + O_p\left(\left(\hat{u}_{t+1}^{(1)} - u_{t+1}^{(1)}\right)^2\right)$$

$$\begin{aligned}\hat{q}_{t+1}^{(2)} - \hat{q}_{t+1}^{(1)} =& q_{t+1}^{(2)} + \nabla q_{t+1}^{(2)}\left(\hat{\beta}_{2,t} + \beta_2\right) \\ &+ O_p\left(\left(\hat{\beta}_{2,t} - \beta_2\right)^2\right) - q_{t+1}^{(1)} - \nabla q_{t+1}^{(1)}\left(\hat{\beta}_{1,t} - \beta_1\right) + O_p\left(\left(\hat{\beta}_{1,t} - \beta_1\right)^2\right)\end{aligned}$$

$$\frac{1}{\alpha\left(\hat{e}_{t+1}^{(1)}\right)^2} = \frac{1}{\alpha\left(e_{t+1}^{(1)}\right)^2} + \frac{-2\nabla e_{t+1}^{(1)\prime}}{\alpha\left(e_{t+1}^{(1)}\right)^3}\left(\hat{\beta}_{1,t} - \beta_1\right) + O_p\left(\left(\hat{\beta}_{1,t} - \beta_1\right)^2\right)$$

$$\begin{aligned}g\left(\hat{u}_{t+1}^{(1)}\right)\hat{q}_{t+1}^{(1)} &= g\left(u_{t+1}^{(1)}\right)q_{t+1}^{(1)} + g\left(\hat{u}_{t+1}^{(1)}\right)\hat{q}_{t+1}^{(1)} - g\left(u_{t+1}^{(1)}\right)q_{t+1}^{(1)} \\ &= g\left(u_{t+1}^{(1)}\right)q_{t+1}^{(1)} + g\left(\hat{u}_{t+1}^{(1)}\right)\left(\hat{q}_{t+1}^{(1)} - q_{t+1}^{(1)}\right) + \left(g\left(\hat{u}_{t+1}^{(1)}\right) - g\left(u_{t+1}^{(1)}\right)\right)q_{t+1}^{(1)} \\ &= g\left(u_{t+1}^{(1)}\right)q_{t+1}^{(1)} + g\left(\hat{u}_{t+1}^{(1)}\right)\left(\hat{q}_{t+1}^{(1)} - q_{t+1}^{(1)}\right) + \delta\left(u_{t+1}^{(1)}\right)\left(\hat{u}_{t+1}^{(1)} - u_{t+1}^{(1)}\right)q_{t+1}^{(1)} \\ &\quad + O_p\left(\left(\hat{u}_{t+1}^{(1)} - u_{t+1}^{(1)}\right)^2\right)q_{t+1}^{(1)}\end{aligned}$$

$$\begin{aligned}\left(\alpha - g\left(\hat{u}_{t+1}^{(1)}\right)\right) &= \left(\alpha - g(u_{t+1}^{(1)}) - g(\hat{u}_{t+1}^{(1)}) + g(u_{t+1}^{(1)})\right) \\ &= \left(\alpha - g\left(u_{t+1}^{(1)}\right)\right) - \left(g\left(\hat{u}_{t+1}^{(1)}\right) - g\left(u_{t+1}^{(1)}\right)\right) \\ &= \left(\alpha - g\left(u_{t+1}^{(1)}\right)\right) - \left[\delta\left(u_{t+1}^{(1)}\right)\left(\hat{u}_{t+1}^{(1)} - u_{t+1}^{(1)}\right) + O_p\left(\left(\hat{u}_{t+1}^{(1)} - u_{t+1}^{(1)}\right)^2\right)\right]\end{aligned}$$

$$\begin{aligned}\alpha\hat{e}_{t+1}^{(1)} &= \alpha\left(e_{t+1}^{(1)} + \nabla e_{t+1}^{(1)}\left(\hat{\beta}_{1,t} - \beta_1\right) + O_p\left(\left(\hat{\beta}_{1,t} - \beta_1\right)^2\right)\right) \\ &= \alpha e_{t+1}^{(1)} + \alpha\left(\nabla e_{t+1}^{(1)}\left(\hat{\beta}_{1,t} - \beta_1\right) + O_p\left(\left(\hat{\beta}_{1,t} - \beta_1\right)^2\right)\right)\end{aligned}$$

$$\begin{aligned}&\sum_t \left[\left(-\frac{1}{\alpha\hat{e}_{t+1}^{(1)}}\mathbf{1}\{Y_{t+1} \le \hat{q}_{t+1}^{(1)}\} + \frac{1}{\hat{e}_{t+1}^{(1)}}\right)\left(\hat{q}_{t+1}^{(2)} - \hat{q}_{t+1}^{(1)}\right)\right. \\ &\left. + \left(\frac{1}{\alpha\left(\hat{e}_{t+1}^{(1)}\right)^2}\mathbf{1}\left\{Y_{t+1} \le \hat{q}_{t+1}^{(1)}\right\}\left(\hat{q}_{t+1}^{(1)} - Y_{t+1}\right) - \frac{\hat{q}_{t+1}^{(1)}}{\left(\hat{e}_{t+1}^{(1)}\right)^2} + \frac{1}{\hat{e}_{t+1}^{(2)}}\right)\left(\hat{e}_{t+1}^{(1)} - \hat{e}_{t+1}^{(1)}\right)\right]\end{aligned}$$

$$
\begin{aligned}
&= \sum_t \Bigg\{ \Bigg[\underbrace{\frac{1}{\alpha e_{t+1}^{(1)}}}_{A_1} - \underbrace{\frac{\nabla e_{t+1}^{(1)\prime}}{\alpha \left(e_{t+1}^{(1)}\right)^2} \left(\hat{\beta}_{1,t} - \beta_1\right) + O_p\left(\left(\hat{\beta}_{1,t} - \beta_1\right)^2\right)}_{A_2} \Bigg] \times \\
&\Bigg[\underbrace{g\left(u_{t+1}^{(1)}\right)}_{B1} + \underbrace{\delta\left(u_{t+1}^{(1)}\right)\left(\hat{u}_{t+1}^{(1)} - u_{t+1}^{(1)}\right) + O_p\left(\left(\hat{u}_{t+1}^{(1)} - u_{t+1}^{(1)}\right)^2\right)}_{B2} \Bigg] \times \\
&\Bigg[\underbrace{\nabla q_{t+1}^{(2)}\left(\hat{\beta}_{2,t} - \beta_2\right) - \nabla q_{t+1}^{(1)}\left(\hat{\beta}_{1,t} - \beta_1\right)}_{C_1} + \underbrace{O_p\left(\left(\hat{\beta}_{1,t} - \beta_1\right)^2\right) + O_p\left(\left(\hat{\beta}_{2,t} - \beta_1\right)^2\right)}_{C_2} \Bigg] \\
&+ \Bigg[\underbrace{\frac{1}{-\alpha\left(e_{t+1}^{(1)}\right)^2}}_{D_1} + \underbrace{\frac{2\nabla e_{t+1}^{(1)\prime}}{\alpha (e_{t+1}^{(1)})^3}\left(\hat{\beta}_{1,t} - \beta_1\right) - O_p\left(\left(\hat{\beta}_{1,t} - \beta_1\right)^2\right)}_{D_2} \Bigg] \times \\
&\Bigg[\underbrace{g\left(u_{t+1}^{(1)}\right) q_{t+1}^{(1)} + \left(\alpha - g\left(u_{t+1}^{(1)}\right)\right) Y_{t+1} - \alpha e_{t+1}^{(1)}}_{E_1} \\
&\underbrace{+ g\left(\hat{u}_{t+1}^{(1)}\right)\left(\hat{q}_{t+1}^{(1)} - q_{t+1}^{(1)}\right) + \delta\left(u_{t+1}^{(1)}\right)\left(\hat{u}_{t+1}^{(1)} - u_{t+1}^{(1)}\right) q_{t+1}^{(1)} + O_p\left(\left(\hat{u}_{t+1}^{(1)} - u_{t+1}^{(1)}\right)^2\right) q_{t+1}^{(1)}}_{E_2} \\
&\underbrace{- \left(\delta\left(u_{t+1}^{(1)}\right)\left(\hat{u}_{t+1}^{(1)} - u_{t+1}^{(1)}\right) + O_p\left(\left(\hat{u}_{t+1}^{(1)} - u_{t+1}^{(1)}\right)^2\right)\right) Y_{t+1}}_{E_2} \\
&\underbrace{-\alpha\left(\nabla e_{t+1}^{(1)}\left(\hat{\beta}_{1,t} - \beta_1\right) + O_p\left(\left(\hat{\beta}_{1,t} - \beta_1\right)^2\right)\right)\Bigg]}_{E_2} \times \\
&\Bigg[\underbrace{\nabla e_{t+1}^{(2)}\left(\hat{\beta}_{2,t} - \beta_2\right) - \nabla e_{t+1}^{(1)}\left(\hat{\beta}_{1,t} - \beta_1\right)}_{F_1} + \underbrace{O_p\left(\left(\hat{\beta}_{1,t} - \beta_1\right)^2\right) + O_p\left(\left(\hat{\beta}_{2,t} - \beta_2\right)^2\right)}_{F_2} \Bigg] \Bigg\}
\end{aligned}
$$

$$
\begin{aligned}
AA &= A_1B_1C_1 + D_1E_1F_1 \\
&= \sum_t \frac{1}{\alpha e_{t+1}^{(1)}} g\left(u_{t+1}^{(1)}\right)\left(\nabla q_{t+1}^{(2)}\left(\hat{\beta}_{2,t}-\beta_2\right) - \nabla q_{t+1}^{(1)}\left(\hat{\beta}_{1,t}-\beta_1\right)\right) \\
&\quad - \frac{1}{\alpha\left(e_{t+1}^{(1)}\right)^2}\left(g\left(u_{t+1}^{(1)}\right)q_{t+1}^{(1)} + \left(\alpha - g\left(u_{t+1}^{(1)}\right)\right)Y_{t+1} - \alpha e_{t+1}^{(1)}\right) \\
&\quad \times \left(\nabla e_{t+1}^{(2)}\left(\hat{\beta}_{2,t}-\beta_2\right) - \nabla e_{t+1}^{(1)}\left(\hat{\beta}_{1,t}-\beta_1\right)\right) \\
&= \sum_t \left[\nabla q_{t+1}^{(2)\prime}\frac{1}{\alpha e_{t+1}^{(1)}} g(u_{t+1}^{(1)})(\hat{\beta}_{2,t}-\beta_2) - \nabla q_{t+1}^{(1)\prime}\frac{1}{\alpha e_{t+1}^{(1)}} g(u_{t+1}^{(1)})(\hat{\beta}_{1,t}-\beta_1)\right. \\
&\quad - \nabla e_{t+1}^{(2)\prime}\frac{1}{\alpha(e_{t+1}^{(1)})^2}\left(g(u_{t+1}^{(1)})q_{t+1}^{(1)} + \left(\alpha - g(u_{t+1}^{(1)})\right)Y_{t+1} - \alpha e_{t+1}^{(1)}\right)(\hat{\beta}_{2,t}-\beta_2) \\
&\quad \left. + \nabla e_{t+1}^{(1)}\frac{1}{\alpha(e_{t+1}^{(1)})^2}\left(g(u_{t+1}^{(1)})q_{t+1}^{(1)} + \left(\alpha - g(u_{t+1}^{(1)})\right)Y_{t+1} - \alpha e_{t+1}^{(1)}\right)(\hat{\beta}_{1,t}-\beta_1)\right] \\
&= \sum_t \left\{\left[\nabla q_{t+1}^{(2)\prime}\frac{1}{\alpha e_{t+1}^{(1)}} g(u_{t+1}^{(1)})\right.\right. \\
&\quad \left.\left. - \nabla e_{t+1}^{(2)\prime}\frac{1}{\alpha(e_{t+1}^{(1)})^2}\left(g(u_{t+1}^{(1)})q_{t+1}^{(1)} + \left(\alpha - g(u_{t+1}^{(1)})\right)Y_{t+1} - \alpha e_{t+1}^{(1)}\right)\right](\hat{\beta}_{2,t}-\beta_2) \\
&\quad - \left[\nabla q_{t+1}^{(1)\prime}\frac{1}{\alpha e_{t+1}^{(1)}} g(u_{t+1}^{(1)})\right. \\
&\quad \left. - \nabla e_{t+1}^{(1)}\frac{1}{\alpha(e_{t+1}^{(1)})^2}\left(g(u_{t+1}^{(1)})q_{t+1}^{(1)} + \left(\alpha - g(u_{t+1}^{(1)})\right)Y_{t+1} - \alpha e_{t+1}^{(1)}\right)\right](\hat{\beta}_{1,t}-\beta_1) \\
&= \sum_t \left[-h_{1,t+1}' B_1(t) H_1(t) + h_{2,t+1}' B_2(t) H_2(t)\right] \\
&= \sum_t \left[-h_{1,t+1}' B_1 H_1(t) + h_{2,t+1}' B_2 H_2(t)\right] + o_p(1) \\
&= \sum_t \left[-h_{2,t+1}' J' B_1 J H_2(t) + h_{2,t+1}' B_2 H_2(t)\right] + o_p(1) = \sum_t \left[h_{2,t+1}' M H_2(t)\right] + o_p(1) \\
&= \sum_t \left[h_{2,t+1}' B_2^{1/2} B_2^{-1/2} M B_2^{-1/2} B_2^{1/2} H_2(t)\right] + o_p(1) = \sum_t \left[h_{2,t+1}' B_2^{1/2} Q B_2^{1/2} H_2(t)\right] \\
&= \sum_t \left[h_{2,t+1}' B_2^{1/2} C A A' C B_2^{1/2} H_2(t)\right] + o_p(1) = Z^2 \sum_t \tilde{H}_2'(t)\tilde{h}_{2,t+1} + o_p(1)
\end{aligned}
$$

Now we need to show that AB, AC, AD, AE, AF, AG, AH are lower order than AA. The order of each element is showed below.

$$A_1 = O_p(1), \qquad A_2 = O_p\left(\frac{1}{\sqrt{R}}\right), \qquad B_1 = O_p(1), \qquad B_2 = O_p\left(\frac{1}{R}\right),$$
$$C_1 = O_p\left(\frac{1}{\sqrt{R}}\right), \qquad C_2 = O_p\left(\frac{1}{R}\right), \qquad D_1 = O_p(1), \qquad D2 = O_p\left(\frac{1}{\sqrt{R}}\right),$$
$$E_1 = O_p(1), \qquad E_2 = O_p\left(\frac{1}{\sqrt{R}}\right), \qquad F_1 = O_p\left(\frac{1}{\sqrt{R}}\right), \qquad F_2 = O_p\left(\frac{1}{R}\right),$$

Besides $\sum_t \delta(u_{t+1}^{(1)})(\hat{u}_{t+1}^{(1)} - u_{t+1}^{(1)})q_{t+1}^{(1)}$, the order of other elements are straight forwards. For the order of $\sum_t \delta(u_{t+1}^{(1)})(\hat{u}_{t+1}^{(1)} - u_{t+1}^{(1)})q_{t+1}^{(1)}$, following Phillips (1991), we can get

$$P^{-1/2}\sum_t \left(\hat{u}_{t+1}^{(1)} - u_{t+1}^{(1)}\right) q_{t+1}^{(1)} \xrightarrow{d} N(0, Q)$$

and

$$P^{-1}\sum_t \delta\left(u_{t+1}^{(1)}\right)\left(\hat{u}_{t+1}^{(1)} - u_{t+1}^{(1)}\right) q_{t+1}^{(1)}$$
$$\xrightarrow{p} \lim_{P\to\infty} P^{-1}\sum_t E\left(\delta\left(u_{t+1}^{(1)}\right)\right)\left(\hat{u}_{t+1}^{(1)} - u_{t+1}^{(1)}\right) q_{t+1}^{(1)} = f(0)Q$$

Thus, we have

$$\sum_t \delta\left(u_{t+1}^{(1)}\right)\left(\hat{u}_{t+1}^{(1)} - u_{t+1}^{(1)}\right) q_{t+1}^{(1)} = Pf(0)O_p(\frac{1}{R}) = O_p\left(\frac{P}{R}\right)$$

By using the order of each element, the order of each term in the $\sum_t(\hat{m}_{1,t}(\hat{q}_{t+1}^{(1)} - \hat{q}_{t+1}^{(2)}) + \hat{m}_{2,t}(\hat{e}_{t+1}^{(1)} - \hat{e}_{t+1}^{(2)})]$ can be proved as following.

$$AA = \sum_t A_1 B_1 C_1 + D_1 E_1 F_1$$

$$= \sum_t \left[O_p(1) O_p(1) O_p\left(\frac{1}{\sqrt{R}}\right) + O_p(1) O_p(1) O_p\left(\frac{1}{\sqrt{R}}\right) \right] = O_p\left(\frac{P}{\sqrt{R}}\right))$$

$$AB = \sum_t A_1 B_1 C_2 + D_1 E_1 F_2$$

$$= \sum_t \left[O_p(1) O_p(1) O_p\left(\frac{1}{R}\right) + O_p(1) O_p(1) O_p\left(\frac{1}{R}\right) \right] = O_p\left(\frac{P}{R}\right)$$

$$AC = \sum_t A_1 B_2 C_1 + D_1 E_2 F_1$$

$$= \sum_t \left[O_p(1) O_p\left(\frac{1}{R}\right) O_p\left(\frac{1}{\sqrt{R}}\right) + O_p(1) O_p\left(\frac{1}{\sqrt{R}}\right) O_p\left(\frac{1}{\sqrt{R}}\right) \right] = O_p\left(\frac{P}{R}\right)$$

$$AD = \sum_t A_1 B_2 C_2 + D_1 E_2 F_2$$

$$= \sum_t \left[O_p(1) O_p\left(\frac{1}{R}\right) O_p\left(\frac{1}{R}\right) + O_p(1) O_p\left(\frac{1}{\sqrt{R}}\right) O_p\left(\frac{1}{R}\right) \right] = O_p\left(\frac{P}{R^{3/2}}\right)$$

$$AE = \sum_t A_2 B_1 C_1 + D_2 E_1 F_1$$

$$= \sum_t \left[O_p\left(\frac{1}{\sqrt{R}}\right) O_p(1) O_p\left(\frac{1}{\sqrt{R}}\right) + O_p\left(\frac{1}{\sqrt{R}}\right) O_p(1) O_p\left(\frac{1}{\sqrt{R}}\right) \right] = O_p\left(\frac{P}{R}\right)$$

$$AF = \sum_t A_2 B_1 C_2 + D_2 E_1 F_2$$

$$= \sum_t \left[O_p\left(\frac{1}{\sqrt{R}}\right) O_p(1) O_p\left(\frac{1}{R}\right) + O_p\left(\frac{1}{\sqrt{R}}\right) O_p(1) O_p\left(\frac{1}{R}\right) \right] = O_p\left(\frac{P}{R^{3/2}}\right)$$

$$AG = \sum_t A_2 B_2 C_1 + D_2 E_2 F_1$$

$$= \sum_t O_p\left(\frac{1}{\sqrt{R}}\right) O_p\left(\frac{1}{R}\right) O_p\left(\frac{1}{\sqrt{R}}\right) + O_p\left(\frac{1}{\sqrt{R}}\right) O_p\left(\frac{1}{\sqrt{R}}\right) O_p\left(\frac{1}{\sqrt{R}}\right) = O\left(\frac{P}{R^{3/2}}\right)$$

$$AH = \sum_t A_2 B_2 C_2 + D_2 E_2 F_2$$

$$= \sum_t \left[O_p\left(\frac{1}{\sqrt{R}}\right) O_p\left(\frac{1}{R}\right) O_p\left(\frac{1}{R}\right) + O_p\left(\frac{1}{\sqrt{R}}\right) O_p\left(\frac{1}{\sqrt{R}}\right) O_p\left(\frac{1}{R}\right) \right] = O_p\left(\frac{P}{R^2}\right)$$

Therefore, all AB, AC, AD, AE, AF, AG, AH are lower order than AA. The proof is complete.

Lemma 6: Let Assumptions 1-4 hold.

$$
\begin{aligned}
&\sum_t \left[\hat{m}_{1,t+1}\left(\hat{q}^{(2)}_{t+1}-\hat{q}^{(1)}_{t+1}\right)+\hat{m}_{2,t+1}\left(\hat{e}^{(2)}_{t+1}-\hat{e}^{(1)}_{t+1}\right)\right]^2 - P\bar{C}^2 \\
=&\sum_t \left[\left(-\frac{1}{\alpha\hat{e}^{(1)}_{t+1}}\mathbf{1}\{Y_{t+1}\le \hat{q}^{(1)}_{t+1}\}+\frac{1}{\hat{e}^{(1)}_{t+1}}\right)\left(\hat{q}^{(2)}_{t+1}-\hat{q}^{(1)}_{t+1}\right)\right. \\
+&\left.\left(\frac{1}{\alpha\left(\hat{e}^{(1)}_{t+1}\right)^2}\mathbf{1}\left\{Y_{t+1}\le \hat{q}^{(1)}_{t+1}\right\}\left(\hat{q}^{(1)}_{t+1}-Y_{t+1}\right)-\frac{\hat{q}^{(1)}_{t+1}}{\left(\hat{e}^{(1)}_{t+1}\right)^2}+\frac{1}{\hat{e}^{(1)}_{t+1}}\right)\left(\hat{e}^{(2)}_{t+1}-\hat{e}^{(1)}_{t+1}\right)\right]^2 - P\bar{C}^2 \\
=&Z^4\sum_t \tilde{H}_2'(t)\tilde{H}_2(t)+o_p(1)
\end{aligned}
$$

Proof:

$$
\begin{aligned}
&\sum_t \left[\hat{m}_{1,t+1}\left(\hat{q}^{(2)}_{t+1}-\hat{q}^{(1)}_{t+1}\right)+\hat{m}_{2,t+1}\left(\hat{e}^{(2)}_{t+1}-\hat{e}^{(1)}_{t+1}\right)\right]^2 \\
=&\sum_t \left[\left(-\frac{1}{\alpha\hat{e}^{(1)}_{t+1}}\mathbf{1}\{Y_{t+1}\le \hat{q}^{(1)}_{t+1}\}+\frac{1}{\hat{e}^{(1)}_{t+1}}\right)\left(\hat{q}^{(2)}_{t+1}-\hat{q}^{(1)}_{t+1}\right)\right. \\
&\left.+\left(\frac{1}{\alpha\left(\hat{e}^{(1)}_{t+1}\right)^2}\mathbf{1}\left\{Y_{t+1}\le \hat{q}^{(1)}_{t+1}\right\}\left(\hat{q}^{(1)}_{t+1}-Y_{t+1}\right)-\frac{\hat{q}^{(1)}_{t+1}}{\left(\hat{e}^{(1)}_{t+1}\right)^2}+\frac{1}{\hat{e}^{(1)}_{t+1}}\right)\left(\hat{e}^{(2)}_{t+1}-\hat{e}^{(1)}_{t+1}\right)\right]^2 \\
=&\sum_t \left[-h'_{1,t+1}B_1(t)H_1(t)+h'_{2,t+1}B_2(t)H_2(t)\right]^2 \\
=&\sum_t \left[-h'_{1,t+1}B_1H_1(t)+h'_{2,t+1}B_2H_2(t)\right]^2+o_p(1) \\
=&\sum_t H_1'(t)B_1'h_{1,t+1}h'_{1,t+1}B_1H_1(t)+\sum_t H_2'(t)B_2'h_{2,t+1}h'_{2,t+1}B_2H_2(t) \\
&-2\sum_t H_1'(t)B_1'h_{1,t+1}h_{2,t+1}B_2H_2(t)+o_p(1)
\end{aligned}
$$

$$= \sum_t H_1'(t) B_1' \mathbb{E}\left(h_{1,t+1} h_{1,t+1}'\right) B_1 H_1(t) + \sum_t H_2'(t) B_2' \mathbb{E}\left(h_{2,t+1} h_{2,t+1}'\right) B_2 H_2(t)$$

$$- 2 \sum_t H_1'(t) B_1' \mathbb{E}\left(h_{1,t+1} h_{2,t+1}'\right) B_2 H_2(t) + o_p(1)$$

$$= Z^2 \sum_t H_1'(t) B_1 H_1(t) + Z^2 \sum_t H_2'(t) B_2 H_2(t) - 2Z^2 \sum_t H_1'(t) B_2 H_2(t) + o_p(1)$$

$$= Z^2 \sum_t H_2'(t) J' B_1 J H_2(t) + Z^2 \sum_t H_2'(t) B_2 H_2(t) - 2Z^2 \sum_t H_2'(t) J' B_1 J H_2(t) + o_p(1)$$

$$= Z^2 \sum_t H_2'(t) \left[-J' B_1 J + B_2\right] H_2(t) + o_p(1)$$

$$= Z^2 \sum_t H_2'(t) M H_2(t) + o_p(1)$$

$$= Z^2 \sum_t H_2'(t) B_2^{1/2} B_2^{-1/2} M B_2^{-1/2} B_2^{1/2} H_2(t) + o_p(1)$$

$$= Z^2 \sum_t H_2'(t) B_2^{1/2} C A A' C B_2^{1/2} H_2(t) + o_p(1)$$

$$= Z^4 \sum_t \tilde{H}_2'(t) \tilde{H}_2(t) + o_p(1)$$

Lemma 3 implies that $\bar{C} = O_p(P^{-1})$, so $P\bar{C}^2 = o_p(1)$. Thus, the proof is completed.

Theorem 2: Let Assumptions 1-4 hold. $\mathrm{ENC}_P \xrightarrow{d} N(0,1)$, as $P, R \to \infty$, under $\mathbb{H}_0$: $\lambda = 0$.

Proof: In order to show that ENC_P is asymptotically standard normal, we need to show

$$\begin{aligned}
\mathrm{ENC}_P &= Q_P^{-1/2}\sqrt{P}\hat{C}_P \\
&= \frac{\sum_t \hat{m}_{1,t+1}\left(\hat{q}_{t+1}^{(2)} - \hat{q}_{t+1}^{(1)}\right) + \hat{m}_{2,t+1}\left(\hat{e}_{t+1}^{(2)} - \hat{e}_{t+1}^{(1)}\right)}{\sqrt{\sum_t \left[\hat{m}_{1,t+1}\left(\hat{q}_{t+1}^{(2)} - \hat{q}_{t+1}^{(1)}\right) + \hat{m}_{2,t+1}\left(\hat{e}_{t+1}^{(2)} - \hat{e}_{t+1}^{(1)}\right)\right]^2 - P\bar{C}^2}} \\
&\Longrightarrow \frac{\xi^{-1}\int_\xi^1 \left[W(s) - W(s-\xi)\right] \mathrm{d}W(s)}{\sqrt{\xi^{-2}\int_\xi^1 \left[W(s) - W(s-\xi)\right]^2 \mathrm{d}s}} \xrightarrow{d} N(0,1)
\end{aligned}$$

We divided $[0,1]$ to $n = [1/\xi] = [T/R]$ equal segments and let $t = [sn]$, where $[\cdot]$ denote the integer part and $s \in [0,1]$. Let $\{\nu_i\}_{i=1}^n$ is mixing sequence with $\mathbb{E}(\nu) = 0$ and $Var(\nu) = 1$. We denote $V_t = \sum_{i=1}^t \nu_i$ to be the partial sum, then $V_t = \sum_{i=1}^t \nu_i \sim N(0,t)$. Thus,

$$\frac{V_t}{\sqrt{n}} = \frac{\sum_{i=1}^t \nu_i}{\sqrt{n}} = V_n(s) \Rightarrow W(s),$$

where $V_n(s)$ is a CADLAG process and $V_n(s)$ is a Wiener process.

$$\begin{aligned}
n^{-1}\sum_{t=1}^n \nu_{t-1}\nu_t &= n^{-1}\sum_{t=1}^n V_{t-1}\nu_t - n^{-1}\sum_{t=1}^n V_{t-2}\nu_t \\
&\Rightarrow \int_\xi^1 W(s)\,\mathrm{d}W(s) - \int_\xi^1 W(s-\xi)\,\mathrm{d}W(s) \\
&= \int_\xi^1 \left[W(s) - W(s-\xi)\right]\mathrm{d}W(s) \\
n^{-2}\sum_{t=1}^n \nu_{t-1}^2 &= n^{-2}\sum_{t=1}^n \left(V_{t-1} - V_{t-2}\right)^2 \\
&\Rightarrow \int_\xi^1 \left[W(s) - W(s-\xi)\right]^2 \mathrm{d}s
\end{aligned}$$

We generate an AR(1) regression to show the asymptotical standard normality of ENC_P.

$$\nu_{t+1} = \zeta\nu_t + \epsilon_t$$

$\hat{\zeta} = \sum_{t=1}^{n} \nu_{t-1}\nu_t / \sum_{t=1}^{n} \nu_{t-1}^2$ and $Var(\hat{\zeta}) = (\sum_{t=1}^{n} \nu_{t-1}^2)^{-1} Var(\nu) = (\sum_{t=1}^{n} \nu_{t-1}^2)^{-1}$.

By using CLT,

$$\frac{\int_{\xi}^{1} [W(s) - W(s-\xi)]\, dW(s)}{\sqrt{\int_{\xi}^{1} [W(s) - W(s-\xi)]^2 \, \mathrm{d}s}} \Rightarrow \frac{\sum_{t=1}^{n} u_{t-1}u_t}{\sqrt{\sum_{t=1}^{n} u_{t-1}^2}} \xrightarrow{d} N(0,1)$$

Table 6.1: Simulation Results of DGP 1: Size of Test

	$P = 240$			$P = 480$			$P = 1200$		
	DM_P	ENC_P	CCS_P	DM_P	ENC_P	CCS_P	DM_P	ENC_P	CCS_P
$\alpha = 0.01$									
$R = 120$	0.005	0.091	0.040	0.003	0.098	0.044	0.001	0.089	0.047
$R = 240$	0.053	0.123	0.035	0.019	0.093	0.056	0.004	0.086	0.040
$R = 480$	0.070	0.105	0.034	0.026	0.075	0.048	0.009	0.062	0.054
$\alpha = 0.025$									
$R = 120$	0.006	0.078	0.049	0.003	0.079	0.052	0.001	0.088	0.056
$R = 240$	0.026	0.074	0.055	0.008	0.060	0.056	0.001	0.066	0.053
$R = 480$	0.030	0.064	0.049	0.013	0.058	0.043	0.004	0.054	0.057
$\alpha = 0.05$									
$R = 120$	0.008	0.057	0.054	0.003	0.065	0.045	0.001	0.068	0.052
$R = 240$	0.013	0.055	0.048	0.002	0.054	0.054	0.000	0.050	0.048
$R = 480$	0.015	0.043	0.059	0.009	0.049	0.059	0.000	0.044	0.054
$\alpha = 0.1$									
$R = 120$	0.003	0.053	0.043	0.002	0.056	0.055	0.000	0.064	0.047
$R = 240$	0.009	0.047	0.052	0.002	0.043	0.054	0.000	0.045	0.054
$R = 480$	0.014	0.041	0.050	0.007	0.044	0.054	0.001	0.046	0.063

This table show the size of DM_P, ENC_P, and CCS_P test under 5% significance level, $b = 0$, $\phi = 0$.

Table 6.2: Simulation Results of DGP 2: Size of Test

	$P = 240$			$P = 480$			$P = 1200$		
	DM_P	ENC_P	CCS_P	DM_P	ENC_P	CCS_P	DM_P	ENC_P	CCS_P
$\alpha = 0.05$									
$R = 120$	0.029	0.064	0.048	0.011	0.055	0.035	0.003	0.072	0.041
$R = 240$	0.032	0.059	0.096	0.015	0.056	0.053	0.004	0.066	0.040
$R = 480$	0.040	0.077	0.121	0.025	0.057	0.076	0.009	0.055	0.038

This table show the size of DM_P, ENC_P, and CCS_P test under 5% significance level, $b = 0$, $\phi = 0$.

Table 6.3: Simulation Results of DGP 1: Power of Test

	$P = 240$			$P = 480$			$P = 1200$		
	DM_P	ENC_P	CCS_P	DM_P	ENC_P	CCS_P	DM_P	ENC_P	CCS_P
$\alpha = 0.01$									
$R = 120$	0.035	0.238	0.007	0.018	0.349	0.002	0.008	0.673	0.000
$R = 240$	0.119	0.348	0.009	0.112	0.493	0.002	0.076	0.721	0.000
$R = 480$	0.240	0.460	0.028	0.225	0.571	0.003	0.232	0.790	0.000
$\alpha = 0.025$									
$R = 120$	0.060	0.406	0.002	0.153	0.500	0.003	0.025	0.840	0.000
$R = 240$	0.153	0.497	0.003	0.162	0.687	0.000	0.220	0.931	0.000
$R = 480$	0.251	0.568	0.004	0.288	0.758	0.000	0.496	0.954	0.000
$\alpha = 0.05$									
$R = 120$	0.104	0.543	0.001	0.125	0.749	0.000	0.165	0.964	0.000
$R = 240$	0.221	0.664	0.001	0.321	0.854	0.000	0.535	0.986	0.000
$R = 480$	0.313	0.713	0.000	0.445	0.891	0.000	0.758	0.997	0.000
$\alpha = 0.1$									
$R = 120$	0.162	0.707	0.001	0.241	0.894	0.000	0.470	0.999	0.000
$R = 240$	0.333	0.793	0.000	0.480	0.956	0.000	0.854	1.000	0.000
$R = 480$	0.420	0.853	0.000	0.623	0.975	0.000	0.938	1.000	0.000

This table show the power of DM_P, ENC_P, and CCS_P test under 5% significance level, $b = 0.1$, $\phi = 0.95$.

Table 6.4: Simulation Results of DGP 2: Power of Test

	$P = 240$			$P = 480$			$P = 1200$		
	DM_P	ENC_P	CCS_P	DM_P	ENC_P	CCS_P	DM_P	ENC_P	CCS_P
$\alpha = 0.05$									
$R = 120$	0.022	0.098	0.045	0.012	0.100	0.136	0.000	0.177	0.657
$R = 240$	0.071	0.144	0.058	0.066	0.213	0.077	0.060	0.319	0.359
$R = 480$	0.155	0.201	0.064	0.167	0.363	0.068	0.272	0.654	0.247

This table show the power of DM_P, ENC_P, and CCS_P test under 5% significance level, $b = 0.1$, $\phi = 0.0$.

Table 6.5: Predicting $\boldsymbol{\gamma}_\alpha = (\text{VaR}_\alpha, \text{ES}_\alpha)'$ of the Conditional Distribution of the Equity Premium - I

			Forecasts begin 40 years			
	Variable	Data	$\alpha = 0.010$	$\alpha = 0.025$	$\alpha = 0.050$	$\alpha = 0.100$
d/p	Dividend price ratio	1872 - 2019	0.8456	0.0217	-0.9950	-1.7111*
d/y	Dividend yield	1872 - 2019	0.9034	0.2156	-2.0096**	-2.0647**
e/p	Earning price ratio	1872 - 2019	0.1563	-2.1621**	-1.1869	-2.0079**
d/e	Dividend payout ratio	1872 - 2019	-1.8649*	-1.4992	-1.5904	-2.5146**
svar	Stock variance	1885 - 2019	-0.8076	-2.1099**	-1.3688	-1.6135
infl	Inflation	1913 - 2019	0.5453	1.3445	1.6212	-1.2970
dfy	Default yield spread	1919 - 2019	-0.2556	-2.1410**	-0.5588	0.4735
lty	Long term yield	1919 - 2019	0.4369	0.8794	1.3936	-0.96409
tbl	Treasury-bill rate	1920 - 2019	-0.6512	1.7758*	2.6561***	-0.2662
tms	Team Spread	1920 - 2019	1.2146	-0.6777	-1.3864	0.1449
b/m	Book to market	1921 - 2019	-0.4917	0.1936	0.2225	1.0035
dfr	Default return spread	1926 - 2019	-0.6867	0.9984	-1.8607*	-2,9087***
ltr	Long term return	1926 - 2019	-0.7686	-1.3939	-1.3442	-2.5768***
ntis	Net equity expansion	1927 - 2019	-0.6009	-0.0934	0.4599	0.8047
csp	Consumption, wealth, income ratio	1937 - 2002	-0.7636	-1.5200	-2.1509**	-1.8635*

Table 6.6: Predicting $\boldsymbol{\gamma}_\alpha = (\text{VaR}_\alpha, \text{ES}_\alpha)'$ of the Conditional Distribution of the Equity Premium - II

			Forecasts begin 1965			
	Variable	Data	$\alpha = 0.010$	$\alpha = 0.025$	$\alpha = 0.050$	$\alpha = 0.100$
d/p	Dividend price ratio	1872 - 2019	1.1461	-0.0489	-0.9532	-2.7234***
d/y	Dividend yield	1872 - 2019	0.7004	0.1788	-1.1910	-2.4216**
e/p	Earning price ratio	1872 - 2019	-1.0353	-2.1651**	-2.7181***	-3.3621***
d/e	Dividend payout ratio	1872 - 2019	-0.9068	-0.5818	0.1831	1.3407
svar	Stock variance	1885 - 2019	-0.4051	-1.6676*	-1.1233	-1.2407
infl	Inflation	1913 - 2019	0.2271	0.4527	-0.9735	-0.6127
dfy	Default yield spread	1919 - 2019	-0.3626	-1.0582	-0.4422	0.5596
lty	Long term yield	1919 - 2019	-0.6788	0.1752	-0.5766	-0.6432
tbl	Treasury-bill rate	1920 - 2019	0.4081	-0.9483	0.7650	1.0696
tms	Team Spread	1920 - 2019	-1.9046*	-0.8820	-1.7015*	-0.7368
b/m	Book to market	1921 - 2019	0.4758	-0.9507	-0.7258	0.8496
dfr	Default return spread	1926 - 2019	-0.9792	-0.5778	-1.6071	-2.1071**
ltr	Long term return	1926 - 2019	-0.7699	-2.9069***	-1.0749	-1.5730
ntis	Net equity expansion	1927 - 2019	0.4822	-1.8989*	0.3957	0.7531
csp	Consumption, wealth, income ratio	1937 - 2002	-0.6788	-1.6012	-1.3447	-3.2380***

Table 6.7: Predicting $\boldsymbol{\gamma}_\alpha = (\mathrm{GaR}_\alpha, \mathrm{GS}_\alpha)'$ of the Conditional Distribution of the GDP Growth

	$h=0$	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=7$	$h=8$
$\mathrm{ENC}_{R,P}$	-1.4929	-1.2786	-1.5109	-1.4508	-1.8355	-1.4053	-1.2638	-0.8704	-1.1025
P-value	0.1355	0.2010	0.1308	0.1468	0.0664	0.1599	0.2063	0.3841	0.2702

Chapter 7

Forecast Encompassing and Granger Causality in Predictive Modal Regressions

7.1 Introduction

In recent years, an increasing number of researchers have focused on modal regression. Kemp and Silva (2012) introduce a semi-parametric conditional mode regression estimator when the dependent variable has a continuous conditional density. They also show that the proposed estimator is consistent and has an asymptotic distribution. Kemp et al. (2020) propose a semiparametric estimation of a random vector's conditional mode with a continuous conditional joint density. However, the mode is not 1-elicitable because there does not exist a strictly consistent scoring function for the mode. Thus, there is no possibility

of comparing and ranking the forecasts of mode based on the realized score. Dimitriadis et al. (2019) formalize those ideas from Kemp and Silva (2012) and Kemp et al. (2020) in the theoretical framework. They define the generalized modal midpoint as the minimizer of the expected loss function for mode and prove that the generalized modal midpoint tends to mode when the bandwidth tends to zero. Thus, they define mode as asymptotically elicitable when the bandwidth tends to zero. In this paper, we want to test the forecast encompassing for the conditional mode.

Out-of-sample forecast comparison is widely used in many fields because it is suggested to test Granger causality, which is used to determine whether some independent variables can predict the dependent variable (Ashley et al., 1980; Diebold and Mariano, 1995). Many papers focus on out-of-sample tests for equal predictive accuracy (Diebold and Mariano, 1995; Clark and McCracken, 2001; Clark and West, 2006, 2007). Clark and McCracken (2001) and Clark and West (2006, 2007) prove that the DM statistic has a downward bias for mean regression when two models are nested. Moreover, they show that the forecast encompassing (ENC) statistic has a zero mean under the null hypothesis. In this paper, we show that the DM statistic has a bias for two nested mode models. Furthermore, we show that the ENC statistic of mode regression is asymptotically standard normality with zero mean under the null hypothesis.

The paper is organized as follows. In section 2, we review the definitions and theorems of the elicitability. In section 3, we discuss the elicitability of the mode. Section 4 develops the forecast encompassing test in the conditional mode regression. We show that the DM statistic has a bias, and the ENC statistic has zero mean under the null hypothe-

sis. In section 5, we prove the ENC for the conditional mode regression is asymptotically standard normality as the number of out-of-sample observations tends to infinity. Section 6 shows that the ENC statistic has good size and standard normal distribution in the finite sample by conducting the Monte Carlo simulation. In section 7, we present empirical analysis to find that some macroeconomic and financial variables Granger cause the equity premium.

7.2 Elicitability

We denote an observation domain O for $y, x \in \mathrm{O} \subseteq \mathbb{R}^{d_1+d_2}$, $d_1 = \dim(y)$ and $d_2 = \dim(x)$, the conditional distribution $F \equiv F_{Y|X}$ for Y given X. Let $\mathcal{F}$ be a class of distribution function on the observation domain O, and let A be an action domain, $\gamma \in \mathrm{A}$. We define $\Gamma : \mathcal{F} \to \mathrm{A}$ to be functional. For example, $\Gamma(F(y|x))$ may be $\mathbb{E}(Y|X)$, $Q(Y|X)$, $V(Y|X)$, $\mathrm{Mode}(Y|X)$ or $\mathrm{ES}(Y|X)$, where $\mathbb{E}(Y|X)$ is the conditional mean, $Q(Y|X)$ is the conditional quantile, $V(Y|X)$ is the conditional variance, $\mathrm{Mode}(Y|X)$ is the conditional mode and $\mathrm{ES}(Y|X)$ is the conditional Expected Shortfall. Note that Γ can be a vector of several of these.

Definition 1: (Gneiting, 2011; Fissler et al., 2016) A scoring function is an $\mathcal{F}$-integrable function $S : A \times O \to \mathbb{R}$. S is said to be *$\mathcal{F}$-consistent* for a functional $\Gamma : \mathcal{F} \to \mathrm{A}$ if $\mathbb{E}_F S(\Gamma(F), Y) \leq \mathbb{E}_F S(\gamma, Y)$ for all $F \in \mathcal{F}$ and for all $\gamma \in \mathrm{A}$. Furthermore, S is *strictly $\mathcal{F}$-consistent* for Γ if it is $\mathcal{F}$-consistent for Γ and if $\mathbb{E}_F S(\Gamma(F), Y) = \mathbb{E}_F S(\gamma, Y)$ implies that $\gamma = \Gamma(F)$ for all $F \in \mathcal{F}$ and for all $\gamma \in \mathrm{A}$.

Definition 2: (Gneiting, 2011; Fissler et al., 2016) A functional $\Gamma : \mathcal{F} \to \mathrm{A} \subseteq \mathbb{R}^k$ is called *k-elicitable*, if there exists a strictly $\mathcal{F}$-consistent scoring function for Γ.

A statistical functional is elicitable if there exists a scoring function that the correct forecast of the functional is the unique minimizer of the expected score. We can compare or rank the forecasts of the elicitable functional with their realized scores (Fissler et al., 2016). Many statistical functionals are 1-elicitable, such as expectation, ratios of expectations, quantiles (Value-at-Risk), and expectiles. However, some functionals are not 1-elicitable, such as variance, mode, and Expected Shortfall (Gneiting, 2011).

7.2.1 Identification Function for Moment Conditions

Taking the derivative of the consistent scoring function, then we can get the identification function.

Definition 3: (Gneiting, 2011; Fissler et al., 2016) An identification function is an $\mathcal{F}$-integrable function $V : A \times O \to \mathbb{R}^k$. V is said to be an *$\mathcal{F}$-identification function* for a functional $\Gamma : \mathcal{F} \to \mathrm{A} \subseteq \mathbb{R}^k$ if $\mathbb{E}_F V(\Gamma(F), Y) = 0$ for all $F \in \mathcal{F}$. Furthermore, V is a *strict $\mathcal{F}$-identification function* for Γ if $\mathbb{E}_F V(\gamma, Y) = 0$ holds if and only if $\gamma = \Gamma(F)$ for all $F \in \mathcal{F}$ and for all $\gamma \in \mathrm{A}$.

Osband (1985) pointed out the strict identification function can be obtained by the partial derivative of the strictly consistent scoring function. There is a nonnegative

function $h : \mathbb{R} \to \mathbb{R}$ such that

$$\frac{\mathrm{d}}{\mathrm{d}\gamma} S\left(\gamma, Y\right) = \frac{\mathrm{d}}{\mathrm{d}\gamma} S\left(\gamma, Y\right) = h\left(\gamma\right) V\left(\gamma, Y\right). \tag{7.1}$$

7.2.2 Examples

We show some examples of functionals that are elicitable. The consistent scoring functions for the mean and expectiles are the Bregman divergence, but the consistent scoring functions for quantiles and mode are not.

First, for the mean, Savage (1971) shows that the strictly scoring function S is consistent for the mean functional if and only if it is of the form

$$S(\gamma, y) = f(y) - f(\gamma) - f'(\gamma)(y - \gamma), \tag{7.2}$$

where f is a strictly convex and differentiable function. Taking the derivative of the strictly consistent scoring function for the mean, we get

$$\frac{\mathrm{d}}{\mathrm{d}\gamma} S\left(\gamma, y\right) = f''(\gamma)(\gamma - y), \tag{7.3}$$

where $h(\gamma) = f''(\gamma)$ and the strict identification function $V(\gamma, y) = \gamma - y$. If we choose $f(y) = y^2$, the strictly consistent scoring function becomes the squared error scoring function $S(\gamma, y) = (y - \gamma)^2$.

Second, for the $\tau-$expectile, Gneiting (2011) shows that the strictly scoring function S is consistent for the $\tau-$expectile if and if it is the form

$$S(\gamma, y) = \mathrm{abs}\left(\tau - 1(\gamma \geq y)\right) \left[f(y) - f(\gamma) - f'(\gamma)(y - \gamma)\right], \tag{7.4}$$

where f is a strictly convex and differentiable function. Taking the derivative of the strictly consistent scoring function for the $\tau-$expectile, we have

$$\frac{\mathrm{d}}{\mathrm{d}\gamma} S\left(\gamma, y\right) = \mathrm{abs}\left(\tau - 1(\gamma \geq y)\right) f''(\gamma)(\gamma - y), \tag{7.5}$$

where $h(\gamma) = f''(\gamma)$ and the strict identification function $V(\gamma, y) = \mathrm{abs}\left(\tau - 1(\gamma \geq y)\right)(\gamma - y)$. If we choose the $f(y) = y^2$, the strictly consistent scoring function becomes the asymmetric least squares (ALS) scoring function from Newey and Powell (1987), that is $S(\gamma, y) = \mathrm{abs}(\tau - 1(\gamma \geq y))(y - \gamma)^2$.

Third, for the α-quantile, Thomson (1979) shows that the strictly scoring function S is consistent for the α-quantile functional if and only if it is of the form

$$S(\gamma, y) = \left(\alpha - 1(\gamma \geq y)\right)\left(g(y) - g(\gamma)\right), \tag{7.6}$$

where g is a strictly increasing function. Taking the derivative of the strictly consistent scoring function for the α-quantile, we get

$$\frac{\mathrm{d}}{\mathrm{d}\gamma} S\left(\gamma, y\right) = \left(1(\gamma \geq y) - \alpha\right) g'(\gamma), \tag{7.7}$$

where $h(\gamma) = g'(\gamma)$ and the strictly identification function $V(\gamma, y) = 1(\gamma \geq y) - \alpha$. If we choose the $g(y) = y$, then the strictly consistent scoring function becomes well-known $\alpha-$quantile scoring function from Koenker and Bassett Jr (1978), that is $S(\gamma, y) = \left(\alpha - 1(\gamma \geq y)\right)\left(y - \gamma\right)$. Hence, the strictly consistent scoring function for the $\alpha-$quantile is not a Bregman divergence.

Next, we discuss the mode regression in Section 3, where we define the scoring rule for the mode functional, which is not a Bregman divergence.

7.3 Elicitability of Mode

In this section, we discuss the scoring rule, asymptotic electability, and identification function for the mode.

7.3.1 Scoring Rule for the Mode and the Modal Midpoint

The scoring function of the mode can be written as one minus kernel or minus kernel. See Lee (1989, 1993), Kemp and Silva (2012) and Dimitriadis et al. (2019).

Definition 4: (Lee, 1989, 1993; Kemp and Silva, 2012) The scoring function for the mode is

$$S_{\delta}^{K}(\gamma, y) = -\frac{1}{\delta} K\left(\frac{y-\gamma}{\delta}\right), \tag{7.8}$$

where δ is the strictly positive bandwidth and $K(\cdot)$ denotes a smooth kernel function.

Definition 5: (Dimitriadis et al., 2019) $\Gamma_{\delta}(F)$ is the *conditional modal midpoint* that is the minimizer of the (expected) scoring function in equation (7.8)

$$\Gamma_{\delta}(F) = \underset{\gamma \in \mathbb{R}}{\arg\min}\ \mathbb{E}_{F}\left[S_{\delta}^{K}(\gamma, Y)\right]. \tag{7.9}$$

Assumption 1: F is a unimodal distribution, so that $S_{\delta}^{K}(\gamma, y)$ in equation (7.8) is strictly $\mathcal{F}$-consistent scoring function.

Therefore, the conditional modal midpoint $\Gamma_\delta(F)$ is 1-elicitable. However, note that the conditional mode $\Gamma(F)$ is not elicitable.

7.3.2 Asymptotic Elicitability

Gneiting (2011) mentions informally that the mode is an optimal forecast under the zero-one scoring function $S_\delta(\gamma, y) = 1(|\gamma - y| > \delta)$. $\Gamma_\delta(F)$ is defined as the modal midpoint functional. The scoring function $S_\delta(\gamma, y)$ is consistent for $\Gamma_\delta(F)$. Thus, the mode functional can be defined as $\Gamma(F) = \lim_{\delta \to 0} \Gamma_\delta(F)$ (Gneiting, 2011). Dimitriadis et al. (2019) formalize the idea from Gneiting (2011), Kemp and Silva (2012) and Kemp et al. (2020) and introduce the concept of a "asymptotic elicitability".

Definition 6: (Gneiting, 2011; Dimitriadis et al., 2019) The functional $\Gamma : \mathcal{F} \to A \subseteq \mathbb{R}$ is *asymptotically elicitable* if there exists a sequence of elicitable functional $\Gamma_\delta : \mathcal{F} \to A \subseteq \mathbb{R}$ such that $\lim_{\delta \to 0} \Gamma_\delta(F) = \Gamma(F)$ for all $F \in \mathcal{F}$.

The following theorem shows that the mode is asymptotically elicitable when the bandwidth δ tends to 0.

Theorem 1: (Dimitriadis et al., 2019) For the class of distributions $\mathcal{F}$ consists of absolutely continuous unimodal distributions with bounded density and for any kernel function K which is positive, smooth, $\int K(u)\mathrm{d}u = 1$ and $\log(K(u))$ is a concave function, the func-

tional Γ_δ induced by the scoring function (7.8) is well-defined for all $\delta > 0$, it holds that

$$\lim_{\delta \to 0} \Gamma_\delta(F) = \Gamma(F), \tag{7.10}$$

for all $F \in \mathcal{F}$, where $\Gamma(F)$ is the conditional mode of the conditional distribution $F \equiv F_{Y|X}$.

Proof: Dimitriadis et al. (2019) prove that $\lim_{\delta \to 0} \Gamma_\delta(F) = \Gamma(F)$. Let $Y|X \sim F \in \mathcal{F}$

$$\begin{aligned} \mathbb{E}_F \left[S_\delta^K(\gamma, Y) \right] &= -\int \frac{1}{\delta} K\left(\frac{y-\gamma}{\delta} \right) f(y|x) \mathrm{d}y \\ &= -\int K(v) f(\gamma + v\delta|x) \mathrm{d}v \\ &= -\int f(\gamma + v\delta|x) \mathrm{d}\mathcal{K}(v), \end{aligned} \tag{7.11}$$

where $f(y|x)$ is the density of $F \equiv F_{Y|X}$, $v = (y-\gamma)/\delta$ and the kernel $K(\cdot)$ is the density of the probability measure $\mathcal{K}'(\cdot)$. Then, assume the density f is bounded, we have

$$\lim_{\delta \to 0} \mathbb{E}_F \left[S_\delta^K(\gamma, Y) \right] = -\int \lim_{\delta \to 0} f(\gamma + v\delta|x) \mathrm{d}\mathcal{K}(v) = -\int f(\gamma|x) \mathrm{d}\mathcal{K}(v) = -f(\gamma|x) \tag{7.12}$$

for all $\gamma \in \mathbb{R}$ as $\int \mathrm{d}\mathcal{K}(v) = 1$. It holds that there exits some $x \in \mathrm{supp}(F)$ such that $-f(\gamma|x) < -f(\tilde{\gamma}|x)$ for all $\tilde{\gamma} \in \mathbb{R}$, where $\gamma = \Gamma(F)$ is the mode of the distribution F by definition. As $S_\delta^K(\gamma, F)$ is a continuous function in γ and δ, we get that

$$\Gamma(F) = \lim_{\delta \to 0} \Gamma_\delta(F) = \lim_{\delta \to 0} \left(\arg\min_\gamma S_\delta^K(\gamma, Y) \right) = \arg\min_x \left(\lim_{\delta \to 0} S_\delta^K(\gamma, Y) \right). \tag{7.13}$$

The proof is complete.

The mode is not 1-elicitable but 1-elicitable when the bandwidth δ tends to 0, i.e., the mode is asymptotically 1-elicitable.

7.3.3 Identification Function for the Modal Midpoint

Equation (7.1) shows that strict identification can be obtained by taking the derivative of the strictly consistent scoring function. Taking the derivative of the strictly consistent scoring function of the mode, we get the strict identification function and martingale difference series for the modal midpoint.

Theorem 2: (Dimitriadis et al., 2019) The first order condition of the loss function with respect to the mode γ is

$$\mathbb{E}_F \left[\frac{1}{\delta^2} K' \left(\frac{u}{\delta} \right) \middle| X \right] = 0, \quad i = 1, 2, \tag{7.14}$$

where $u = Y - \gamma$, which means that $\frac{1}{\delta^2} K' \left(\frac{u}{\delta} \right)$ is a martingale difference series. The strictly identification function V_δ^K in equation (7.1) is

$$V_\delta^K(\gamma, y) = \frac{1}{\delta^2} K' \left(\frac{u}{\delta} \right), \tag{7.15}$$

and the function h_δ^K in equation (7.1) is $h_\delta^K(\gamma) = 1$ for the mode regression.

Proof: Recall the strictly consistent scoring function $S_\delta^K(\gamma, y) = -\frac{1}{\delta} K \left(\frac{y - \gamma}{\delta} \right)$ in equation (7.8). We take the derivative of the expected scoring function with respect to γ. The

first order condition (F.O.C.) with respect to (w.r.t.) γ gives

$$\begin{aligned}\frac{\partial \mathbb{E}_F\left[S_\delta^K(\gamma,Y)\middle|X\right]}{\partial\gamma} &= \mathbb{E}_F\left[\frac{\partial S_\delta^K(\gamma,Y)}{\partial\gamma}\middle|X\right]\\ &= \mathbb{E}_F\left[-\frac{1}{\delta}K'\left(\frac{Y-\gamma}{\delta}\right)\left(-\frac{1}{\delta}\right)\middle|X\right]\\ &= \mathbb{E}_F\left[\frac{1}{\delta^2}K'\left(\frac{u}{\delta}\right)\middle|X\right]\\ &= h_\delta^K(\gamma)\mathbb{E}_F\left[V_\delta^K(\gamma,Y)\middle|X\right]=0.\end{aligned}$$

We obtain $\mathbb{E}_F\left[\frac{1}{\delta^2}K'\left(\frac{u}{\delta}\right)\middle|X\right]=0$, which implies that $\frac{1}{\delta^2}K'\left(\frac{u}{\delta}\right)$ is a martingale difference series. Thus, the strictly identification function V_δ^K in equation (7.1) is $V_\delta^K(\gamma,y)=\frac{1}{\delta^2}K'\left(\frac{u}{\delta}\right)$ and the function h_δ^K in equation (7.1) is $h_\delta^K(\gamma)=1$ for the mode regression.

7.4 Forecast Encompassing Test for Granger Causality in Predictive Mode Regression

In this section, we discuss the estimation of the modal regression and develop a new test for Granger Causality in the conditional mode regression.

7.4.1 Models

We have two models. One model is without conditioning on x and the other model is with conditioning on x. The unconditional mode of the unconditional distribution $F_1 = F_Y(Y)$ is $\Gamma(F_1)$. The conditional mode of the conditional distribution $F_2 = F_{Y|X}(Y|X)$ is

the functional $\Gamma(F_2)$. The two nested mode models are

$$\text{Model 1} : y_{t+1} = a_1 + u_{t+1}^{(1)} \equiv x_{1,t}'\beta_1 + u_{t+1}^{(1)}, \tag{7.16}$$

$$\text{Model 2} : y_{t+1} = a_{2,} + bx_t + u_{t+1}^{(2)} \equiv x_{2,t}'\beta_2 + u_{t+1}^{(2)}, \tag{7.17}$$

where $\gamma_t^{(1)} = x_{1,t}'\beta_1$ and $\gamma_t^{(2)} = x_{2,t}'\beta_2$.The dependent variable y_{t+1} is a scalar random variable. The independent variable x_t is a stationary variable. $x_{1,t}$ is a strict subset of $x_{2,t}$. $x_{1,t}' = 1, \beta_1 = a_1, x_{2,t}' = (1, x_t), \beta_2 = (a_2, b)$. Thus, we estimate $\hat{u}_{t+1}^{(1)}$ and $\hat{u}_{t+1}^{(2)}$ by following $\hat{u}_{t+1}^{(1)} = y_{t+1} - \hat{a}_{1,t}$ and $\hat{u}_{t+1}^{(2)} = y_{t+1} - \hat{a}_{2,t} - \hat{b}_t x_t$.

7.4.2 Estimation of Modal Regression

The scoring function of mode includes the Kernel function, so we need to choose the bandwidth δ_R. Following Kemp and Silva (2012), we define

$$\delta_n = k \,\text{MAD}\, n^{-1/7}. \tag{7.18}$$

For the value of k, we choose $k = 0.8, 1.6$. According to Kemp and Silva (2012), $k = 1.6$ is inspired by Silverman (1986) rule-of-thumb. $k = 0.8$ is chosen to show under-smoothing. MAD denotes the median of the absolute deviation from the median ordinary least squares residual,

$$\text{MAD} = \text{med}_t \left[\text{abs} \left(\left(y_{t+1} - x_{i,t}'\beta_i \right) - \text{med}_t \left(y_{t+1} - x_{i,t}'\beta_i \right) \right) \right], \tag{7.19}$$

where β is the ordinary least squares (OLS) estimator.

In this paper, we choose the standard normal density kernel, that is

$$K(\frac{y-\gamma}{\delta}) = \frac{1}{\sqrt{2\pi}\delta} \exp\left(-\frac{(y-\gamma)^2}{2\delta^2} \right). \tag{7.20}$$

In order to find out the moment condition, taking the derivative of the expectation of equation (7.8) with respect to β_i, we get

$$\begin{aligned}\frac{\partial \mathbb{E}_F S_\delta^K(\gamma, Y)}{\partial \beta_i} &= \mathbb{E}_F\left[\frac{\partial S_\delta^K(\gamma, Y)}{\partial \beta_i}\right] \\ &= \mathbb{E}_F\left[\sum_{t=1}^{n} -\frac{1}{\delta_n^4\sqrt{2\pi}}\exp\left(-\frac{\left(y_{t+1}-x_{i,t}'\beta_i\right)^2}{2\delta_n^2}\right)\left(y_{t+1}-x_{i,t}'\beta_i\right)x_{i,t}'\right] \\ &= \mathbb{E}_F\left[\sum_{t=1}^{n} \exp\left(-\frac{\left(y_{t+1}-x_{i,t}'\beta_i\right)^2}{2\delta_n^2}\right)\left(y_{t+1}-x_{i,t}'\beta_i\right)x_{i,t}'\right] = 0.\end{aligned}$$

To find out the estimators, we solve the following moment condition

$$\mathbb{E}_F\left[\sum_{t=1}^{n} w_t\left(\beta_i\right)\left(y_{t+1}-x_{i,t}'\beta_i\right)x_{i,t}'\right] = 0, \tag{7.21}$$

where the weight is

$$w_t\left(\beta_i\right) = \exp\left(-\frac{\left(y_{t+1}-x_{i,t}'\beta_i\right)^2}{2\delta_n^2}\right). \tag{7.22}$$

Solving the moment equation (7.21) gives the estimator of the mode regression when the δ_n tends to zero. Solving the moment equation (7.21) gives the estimator of the mean regression when the δ_n tends to ∞. The estimator $\hat{\beta}_i$ can be computed as iterated weighted least squares estimators, i.e., as the solution to the equation:

$$\hat{\beta}_i = \left[\sum_{t=1}^{n} w_t\left(\hat{\beta}_i\right)x_{i,t}x_{i,t}'\right]^{-1}\sum_{t=1}^{n} w_t\left(\hat{\beta}_i\right)x_{i,t}y_{t+1}, \tag{7.23}$$

where $w_t\left(\hat{\beta}_i\right) = \exp\left(-\frac{\left(y_{t+1}-x_{i,t}'\hat{\beta}_i\right)^2}{2\delta_n^2}\right)$.

Taking the derivative of scoring function with respect to $\beta_{i,t}$, we get the score [1]

$$h_{i,t} = \frac{\partial S_\delta^K(\gamma, y)}{\partial \beta_{i,t}} = -\frac{1}{\delta_n^2}K'\left(\frac{u_t^{(i)}}{\delta_n}\right)x_{i,t-1}. \tag{7.24}$$

[1]We note that $h_{i,t}$ is called the score (the first order condition of $S(\gamma, Y)$), while $S(\gamma, Y)$ is called the scoring function or scoring rule.

Then, let $H_i(t)$ as follows:

$$H_i(t) = \frac{1}{R}\sum_{j=t-R+1}^{t} h_{i,t} = -\frac{1}{R}\sum_{j=t-R+1}^{t} \frac{1}{\delta_R^2} K'\left(\frac{u_t^{(i)}}{\delta}\right) x_{i,t-1}. \tag{7.25}$$

Taking the derivative of $h_{i,t}$ with respect to $\beta_{i,t}$, the Hessian is obtained as follows:

$$\Lambda_{i,t} = \frac{\partial h_{i,t}}{\partial \beta} = \frac{1}{\delta_n^3} K''\left(\frac{u_t^{(i)}}{\delta_n}\right) x_{i,t-1} x'_{i,t-1}. \tag{7.26}$$

According to Kemp and Silva (2012), we get that

$$\frac{1}{n}\sum_{t=1}^{n} \Lambda_{i,t} = \frac{1}{n}\sum_{t=1}^{n} \frac{1}{\delta_n^3} K''\left(\frac{\hat{u}_t^{(i)}}{\delta_n}\right) x_{i,t-1} x'_{i,t-1} = B_i^{-1} + o_p(1), \tag{7.27}$$

where

$$B_i^{-1} = \mathbb{E}_F\left[f''_{u_t^{(i)}|X}\left(0|x_{i,t}\right) x_{i,t} x'_{i,t}\right] = \lim_{n\to\infty} \mathbb{E}_F\left[\frac{1}{\delta_n^3}\left\{K''\left(\frac{u_t^{(i)}}{\delta_n}\right)\right\} x_{i,t} x'_{i,t}\right], \tag{7.28}$$

where $f''_{u^{(i)}|X}\left(0|x_t\right) = \frac{\partial^2 L}{\partial\beta\partial\beta}\Big|_{\beta_0}$ and $\mathbb{E}_F f''_{u^{(i)}|X}\left(0|x_t\right)$ is negative definite. Kemp and Silva (2012) shows that the asymptotic distribution of $\hat{\beta}_{i,t}$ is

$$\sqrt{n\delta_n^3}\left(\hat{\beta}_{i,t} - \beta_i\right) = -B_i^{-1}\left[\frac{1}{\sqrt{n\delta_n}}\sum_{t=1}^{n} K'\left(\frac{u_{i,t}}{\delta_n}\right) x_{i,t}\right] + o_p(1) \xrightarrow{d} N\left(0, B_i^{-1} A_i B_i^{-1}\right), \tag{7.29}$$

where $A_i = \lim_{n\to\infty} \mathbb{E}_F\left[\frac{1}{\delta_n}\left\{K'\left(\frac{u_i}{\delta_n}\right)\right\}^2 x_i x'_i\right]$. Equation (7.29) shows that the mode regression estimator converges to a normal distribution.

7.4.3 Testing for Equal Predictive Ability

Diebold and Mariano (1995) introduce a test for comparing two models in out-of-sample prediction. The null and alternative hypotheses are

$$\mathbb{H}_0 : \mathbb{E}_F\left[S_\delta^K(\gamma_1, Y) - S_\delta^K(\gamma_2, Y)\right] = 0, \tag{7.30}$$

$$\mathbb{H}_1 : \mathbb{E}_F\left[S_\delta^K(\gamma_1, Y) - S_\delta^K(\gamma_2, Y)\right] > 0. \tag{7.31}$$

Under $\mathbb{H}_0$, X does not Granger-cause Y in the mode. Under $\mathbb{H}_1$, X Granger-causes Y in the mode. The alternative hypothesis is one-sided. The reason is that we expect forecasts from Model 2 to be superior to those from Model 1 as Clark and McCracken (2001) and Clark and West (2006) did. Define R as the number of in-sample observations. Define P as the number of out-of-sample forecasts. We get the DM test statistic based on the loss differential is

$$D = \mathbb{E}_F \left[-\frac{1}{\delta_R} K\left(\frac{u_{t+1}^{(1)}}{\delta_R}\right) + \frac{1}{\delta_R} K\left(\frac{u_{t+1}^{(2)}}{\delta_R}\right) \right], \tag{7.32}$$

$$\hat{D}_P = P^{-1} \sum_{t=R}^{T} \left[-\frac{1}{\delta_R} K\left(\frac{\hat{u}_{t+1}^{(1)}}{\delta_R}\right) + \frac{1}{\delta_R} K\left(\frac{\hat{u}_{t+1}^{(2)}}{\delta_R}\right) \right] \xrightarrow{p} D \tag{7.33}$$

as $P \to \infty$, where $u_{t+1}^{(i)} = y_{t+1} - \gamma_t^{(i)}$ and $T + 1 = R + P$. We show that the DM statistic based on the mode loss differential $\hat{D}_P$ has a non-zero mean under null hypothesis. Proposition 3 shows that $\hat{D}_P$ is biased.

Proposition 3: The DM statistic of the mode loss differential $\hat{D}_P$ has a non-zero mean. Thus, $\mathbb{E}\hat{D}_P \neq 0$ under $\mathbb{H}_0$.

Proof: We use the standard normal density kernel $K\left(\frac{\hat{u}_{t+1}}{\delta_R}\right) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\hat{u}_{t+1}}{\delta_R}\right)^2\right)$. In order to prove that $\hat{D}_P$ is non-zero, we need to get $d_t = -\frac{1}{\delta_R} K\left(\frac{\hat{u}_{t+1}^{(1)}}{\delta_R}\right) + \frac{1}{\delta_R} K\left(\frac{\hat{u}_{t+1}^{(2)}}{\delta_R}\right)$ is non-zero. When $|\hat{u}_{t+1}^{(1)}| > |\hat{u}_{t+1}^{(2)}|$, $K\left(\frac{\hat{u}_{t+1}^{(2)}}{\delta_R}\right) > K\left(\frac{\hat{u}_{t+1}^{(1)}}{\delta_R}\right) > 0$, so $d_t > 0$. When $|\hat{u}_{t+1}^{(1)}| < |\hat{u}_{t+1}^{(2)}|$, $K\left(\frac{\hat{u}_{t+1}^{(1)}}{\delta_R}\right) > K\left(\frac{\hat{u}_{t+1}^{(2)}}{\delta_R}\right) > 0$, so $d_t < 0$. Thus, we get that $\hat{D}_P$ is non-zero at each t.

Thus, the DM statistic has bias under the null hypothesis.

7.4.4 A New Test for Granger Causality in Modal Regression

In this paper, we develop a new test for Grander Causality based on the combined mode

$$\gamma^{(c)} = (1-\lambda)\gamma^{(1)} + \lambda\gamma^{(2)}. \tag{7.34}$$

The null and alternative hypotheses are

$$\mathbb{H}_0 : \lambda = 0, \text{ and } \mathbb{H}_1 : \lambda \neq 0. \tag{7.35}$$

Under $\mathbb{H}_0 : \lambda = 0$, there is no Granger Causality between X and Y. Under $\mathbb{H}_1 : \lambda \neq 0$, there exists Granger Causality between X and Y.

Equation (7.1) shows $\mathbb{E}_F\left[V_\delta^K(\gamma, Y)|X\right] = 0$, which implies $\mathbb{E}_F\left[V_\delta^K(\gamma, Y)\xi(X)\right] = 0$, $\forall\ \xi(X)$. Equation (7.15) shows that the identification function for the mode is $V_\delta^K = \frac{1}{\delta_R^2}K'\left(\frac{u}{\delta_R}\right)$. In this new test, we choose the test function $\xi(X) = u^{(1)} - u^{(2)}$. Proposition 4 gives the property of the encompassing statistic and shows why we want to choose the test function as $\xi(X) = u^{(1)} - u^{(2)}$.

Proposition 4: Denote the forecast encompassing statistic as $\hat{C}_P$. Under $\mathbb{H}_0$,

$$C \equiv \mathbb{E}_F\left[\frac{1}{\delta_R^2}K'\left(\frac{u_{t+1}^{(1)}}{\delta_R}\right)\left(u_{t+1}^{(1)} - u_{t+1}^{(2)}\right)\right] = 0, \tag{7.36}$$

$$\hat{C}_P \equiv P^{-1}\sum_{t=R}^{T}\left[\frac{1}{\delta_R^2}K'\left(\frac{\hat{u}_{t+1}^{(1)}}{\delta_R}\right)\left(\hat{u}_{t+1}^{(1)} - \hat{u}_{t+1}^{(2)}\right)\right] \xrightarrow{p} C = 0 \tag{7.37}$$

as $P \to \infty$. Therefore, we have $C = \mathbb{E}_F\left[V_{\delta_R}^K(\gamma, Y)\xi(X)\right] = 0$ when we choose the test function $\xi(X) = u_{t+1}^{(1)} - u_{t+1}^{(2)}$.

Proof: We combined Model 1 and Model 2 with weight $1-\lambda$ and λ respectively. In order to find out the forecast encompassing statistic of the mode regression under the null hypothesis, we estimate the expectation of the scoring function with combined $\gamma^{(1)}$ and $\gamma^{(2)}$, so that we have

$$\lambda = \arg\min_{\lambda} \mathbb{E}_F S_\delta^K(\gamma^{(c)}, Y), \tag{7.38}$$

where $\gamma_t^{(c)} = (1-\lambda)\,\gamma_t^{(1)} + \lambda\gamma_t^{(2)}$. Taking the first order condition on the expectation of mode scoring function with respect to λ, we get

$$C = \frac{\partial \mathbb{E}_F S_\delta^K(\gamma^{(c)}, Y)}{\partial \lambda} = \mathbb{E}_F \frac{\partial S_\delta^K(\gamma^{(c)}, Y)}{\partial \lambda} = \mathbb{E}_F \left[\frac{1}{\delta_R^2} K' \left(\frac{u_{t+1}^{(1)}}{\delta_R} \right) \left(u_{t+1}^{(1)} - u_{t+1}^{(2)} \right) \right] = 0,$$

$$\hat{C}_P = P^{-1} \sum_{t=R}^{T} \left[\frac{1}{\delta_R^2} K' \left(\frac{\hat{u}_{t+1}^{(1)}}{\delta_R} \right) \left(\hat{u}_{t+1}^{(1)} - \hat{u}_{t+1}^{(2)} \right) \right] \xrightarrow{p} C = 0 \quad \text{as } P \to \infty,$$

under $\mathbb{H}_0$ where $u_{t+1}^{(i)} = y_{t+1} - \gamma_t^{(i)}$. Under the null hypothesis, $\lambda = 0$, then we have $u_{t+1}^{(c)} = u_{t+1}^{(1)}$. Thus, we obtain $\mathbb{E}_F(\hat{C}_P) = 0$, so $\hat{C}_P \xrightarrow{p} 0$ as $P \to \infty$. Therefore, we have $C = \mathbb{E}_F \left[V_{\delta_R}^K(\gamma, Y)\xi(X) \right] = 0$ when we choose the test function $\xi(X) = u_{t+1}^{(1)} - u_{t+1}^{(2)}$.

We compare the DM statistic and the ENC statistic with the CCS statistic which is the test statistic by Chao et al. (2001). Chao et al. (2001) show that the CCS statistic is $\hat{M}_P = P^{-1} \sum_{t=R}^{T} \hat{e}_{t+1}^{(1)} x_t$. In order to compare the equal predictive accuracy of two nested mode models, we standardize these three statistics. The DM statistic is $DM_P \equiv \hat{S}_P^{-0.5} \sqrt{P} \hat{D}_P$, where $S_P = \text{var}\left(\sqrt{P}\hat{D}_P \right)$ and $S_P - \hat{S}_P \xrightarrow{p} 0$. The ENC statistic is $ENC_P \equiv \hat{Q}_P^{-0.5} \sqrt{P} \hat{C}_P$, where $Q_P = \text{var}\left(\sqrt{P}\hat{C}_P \right)$ and $Q_P - \hat{Q}_P \xrightarrow{p} 0$. The CCS statistic is $CCS_P \equiv \hat{W}_P^{-0.5} \sqrt{P} \hat{M}_P$, where $W_P = \text{var}\left(\sqrt{P}\hat{M}_P \right)$ and $W_P - \hat{W}_P \xrightarrow{p} 0$.

7.5 Asymptotic Normality of the ENC Statistic for Mode

In this section, we derive the asymptotic distribution of ENC_P. To show ENC_P to be asymptotical standard normality, we denote c_t as

$$\hat{C}_P \equiv P^{-1}\sum_{t=R}^{T} c_t \equiv P^{-1}\sum_{t=R}^{T} \frac{1}{\delta^2} K'\left(\frac{\hat{u}_{t+1}^{(1)}}{\delta}\right)\left(\hat{u}_{t+1}^{(1)} - \hat{u}_{t+1}^{(2)}\right).$$

We simplify the ENC_P statistic as follows:

$$\mathrm{ENC}_P = Q_P^{-1/2}\sqrt{P}\hat{C}_P = \frac{\sum_{t=R}^{T} c_t}{\sqrt{\sum_t c_t^2 - P\hat{C}_P^2}} = \frac{\sum_{t=R}^{T} \frac{1}{\delta^2} K'\left(\frac{\hat{u}_{t+1}^{(1)}}{\delta}\right)\left(\hat{u}_{t+1}^{(1)} - \hat{u}_{t+1}^{(2)}\right)}{\sqrt{\sum_{t=R}^{T}\left[\frac{1}{\delta^2} K'\left(\frac{\hat{u}_{t+1}^{(1)}}{\delta}\right)\left(\hat{u}_{t+1}^{(1)} - \hat{u}_{t+1}^{(2)}\right)\right]^2 - P\hat{C}_P^2}}.$$

The following assumptions with Assumption 1 in section 3.1 can be used to obtain the limiting distribution in Theorem 3. These assumptions are only sufficient but not necessary and sufficient.

Assumption 2: The parameter estimates $\hat{\beta}_{i,t}, i = 1, 2, t = R, \ldots, T$, satisfy $\hat{\beta}_{i,t} - \beta_i = B_i H_i(t) + o_p(1) = B_i\left(R^{-1}\sum_{j=t-R+1}^{t} h_{i,j}\right) + o_p(1)$, where B_i is defined in equation (7.28), $H_i(t)$ is defined in equation (7.25), and h_{it} is defined in equation (7.24).

Assumption 3: Let $U_t = \left[\frac{1}{\delta_R^2} K'\left(\frac{u_t}{\delta_R}\right), x_{2,t}' - \mathbb{E}x_{2,t}', h_{2,t}', \mathrm{vec}\left(h_{2,t}h_{2,t}' - \mathbb{E}h_{2,t}h_{2,t}'\right)\right]'$ (a) $\mathbb{E}U_t = 0$, (b) Define $\tilde{U}_t$ the nonredundant elements of U_t, then $\frac{1}{R}\mathbb{E}(\sum_{j=t-R+1}^{t}\tilde{U}_j)(\sum_{j=t-R+1}^{t}\tilde{U}_j)' = \Omega < \infty$ is p.d.

Assumption 4: (a) $\mathbb{E}h_{2,t}h_{2,t}' = \sigma^2 B_2^{-1}$, (b) $\mathbb{E}\left(h_{2,t}|h_{2,t-j}, j = 1, 2, \ldots\right) = 0$.

Assumptions 3 and 4 allow the application of an invariance principle and are sufficient for joint weak convergence of partial sums and averages of these partial sums to

Brownian motion and integrals of Brownian motion.

Assumption 5: $f_{u|X}(u|x)$ is three times differentiable with respect to u for all x such that: $(a) f^{(j)}_{u|X}(u|x) = \partial^j f_{u|X}(u|x)/\partial u^j$ is uniformly bounded for $j = 1, 2, 3; (b) E\left[f^{(2)}_{u|X}(0|x_i)\, x_i x_i'\right]$ is negative definite.

Assumption 6: $K : \mathbb{R} \to \mathbb{R}, K(u)$ is a continuously differentiable kernel function such that: $(a) \int_{-\infty}^{\infty} K(u)\mathrm{d}u = 1$; $(b) \sup_{u\in R} |K(u)| = c_0 < \infty$; $(c) \sup_{u\in R} |K'(u)| = c_1 < \infty$; $(d) \int_{-\infty}^{\infty} uK(u)\mathrm{d}u = 0$; $(e) \lim_{u\to\pm\infty} K(u) = 0$; $(f) \int_{-\infty}^{\infty} u^2|K(u)|\mathrm{d}u = M_0 < \infty$; $(g) \int_{-\infty}^{\infty} |K'(u)|^2\mathrm{d}u = M_1 < \infty$; $(h) \sup_{u\in R} |K''(u)| = M_2 < \infty$; $(i) \sup_{u\in R} |K'''(u)| = M_3 < \infty$; $(j) \int_{-\infty}^{\infty} |K''(u)|^2\ \mathrm{d}u = M_4 < \infty$.

Assumption 6 includes standard conditions on kernels in the nonparametric literature (Kemp and Silva, 2012).

Assumption 7: δ_R is a strictly positive bandwidth such that for $R \to \infty, (a)\delta_R \to 0; (b)\frac{R\delta_R}{\ln(R)} \to \infty$; $(c) R\delta_R^7 = o(1)$; $(d)\frac{R\delta_R^5}{\ln(R)} \to \infty$.

Assumption 7 is the bandwidth assumption, which implies that shrinks to zero at an appropriate rate (Kemp and Silva, 2012).

Assumption 8: $\lim_{P,R\to\infty} P/R = \infty$.

In order to get accurate out-of-sample forecasts, selecting an optimal in-sample window is essential. Inoue et al. (2017) find the optimal window size by minimizing the conditional mean squared forecast error in a time-varying predictive regression model. They show that the optimal window size is $R = O_p(T^{2/3})$. Moreover, Hong et al. (2020) point

out that the optimal window size minimizes various out-of-sample forecast criteria, which correspond to the unconditional, conditional, and global mean squared forecast error. They show that the optimal window size is $R = O_p(T^{4/5})$. In our setting, $R + P = T$, so $P = O_p(T)$. Whatever the optimal window size is $R = O_p(T^{2/3})$ or $R = O_p(T^{4/5})$, P has a faster convergence rate than R. Thus, we want to find out the asymptotic normality of ENC_P when $P/R = \infty$.

Theorem 3: Let Assumptions 1-8 hold. $\text{ENC}_P \xrightarrow{d} N(0,1)$, as $P, R \to \infty$, under $\mathbb{H}_0$: $\lambda = 0$.

See Appendix A for proof of Theorem 3.

7.6 Monte Carlo Simulation

In this section, we generate data by two data generating processes. We show that the ENC_P has good size and the distributions of ENC_P are standard normal.

7.6.1 Simulation Design

In order to show the distribution of forecast encompassing test, we simulate data from the following two DGPs.

In DGP 1, we generate the additional variable x_t in Model 2 to be an AR(1) process. We set

$$x_t = \phi x_{t-1} + v_t, \tag{7.39}$$

where $|\phi| < 1$, $v \sim N(0, \sigma_v^2)$. Following Dimitriadis et al. (2019), we generate the error term $u_{t+1} \overset{iid}{\sim} SN(0, \sigma_u^2, \eta)$, where $SN(0, \sigma_u^2, \eta)$ is a skewed normal distribution and η is the skewness of the error term u.

DGP 2 is from Kemp and Silva (2012), where x_t is generated from the $\chi^2_{(3)}$ distribution, and scaled to have variance equal to one. The error term u_{t+1} from equation (7.17) is generated from a rescaled log-gamma random variable

$$u_{t+1} = -\lambda \ln\left(Z_{t+1}\right), \quad \lambda > 0, \tag{7.40}$$

where Z_{t+1} has a gamma distribution with mean α/κ and variance α/κ^2, for $\alpha, \kappa > 0$

In this setting, we can see that the mode of u_{t+1} is given by $\lambda \ln(\kappa/\alpha)$. Thus, when we set $\kappa = \alpha$, we get that u_{t+1} has zero mode. According to Kemp and Silva (2012), the mean of u_{t+1} is $\mu_u = \lambda\left[\ln(\kappa) - \psi_0(\alpha)\right]$, where $\psi_0(\cdot)$ denotes the digamma function. The variance of u_{t+1} is given by $\lambda^2\psi_1(\alpha)$, where $\psi_1(\cdot)$ is the trigamma function. When we set the $\lambda = \psi_1(\alpha)^{-0.5}$, the variance of error term is one. Thus, we can get u_{t+1} is positively skewed, with coefficient of skewness $-\psi_2(\alpha)\psi_1(\alpha)^{-3/2}$, where $\psi_2(\cdot)$ is the quadrigamma function. For example, if $\alpha = 5$ or $\alpha = 0.05$, the coefficient of skewness of u_{t+1} is approximately 0.5 or 2 respectively.

For the bandwidth δ_R, the setting are the same in both DGPs. Following Kemp and Silva (2012), we define

$$\delta_R = k\,\mathrm{MAD}\,R^{-1/7}. \tag{7.41}$$

For the value of k, we choose $k = 0.8, 1.6$. According to Kemp and Silva (2012), $k = 1.6$ is inspired by Silverman (1986) rule-of-thumb. $k = 0.8$ is chosen to show under-smoothing. MAD denotes the median of the absolute deviation from the median ordinary least squares

residual,

$$\text{MAD} = \text{med}_t \left[\text{abs} \left(\left(y_{t+1} - x'_{i,t}\beta_i \right) - \text{med}_t \left(y_{t+1} - x'_{i,t}\beta_i \right) \right) \right], \tag{7.42}$$

where β is the OLS estimator.

We consider $\alpha \in \{5, 0.05\}$, $\eta \in \{-0.5, -0.25, 0, 0.25, 0.5\}$, $\phi \in \{0\}$, $\sigma_v \in \{1\}$, $\sigma_u \in \{1\}$, $c_2 \in \{0.5\}$, and $b \in \{0, 0.1\}$.

We use the MATLAB to estimate and forecast the models. We choose to use a standard normal density Kernel. To find out the estimators, Kemp and Silva (2012) point out that we can solve the following moment condition

$$\mathbb{E}_F \left[\sum_{t=1}^{R} w_t \left(\beta_i \right) \left(y_{t+1} - x'_{i,t}\beta_i \right) x'_{i,t} \right] = 0, \tag{7.43}$$

where the weight $w_t \left(\beta_i \right) = \exp \left(- \frac{\left(y_{t+1} - x'_{i,t}\beta_i \right)^2}{2\delta_R^2} \right)$. The equation (7.43) is for the mode regression when δ_R tends to zero. The equation (7.43) is for the mean regression when δ_R tends to ∞. The estimator $\hat{\beta}_i$ can be computed as iterated weighted least squares estimators, i.e, as the solution to the equation:

$$\hat{\beta}_i = \left[\sum_{t=1}^{R} w_t \left(\hat{\beta}_i \right) x_{i,t} x'_{i,t} \right]^{-1} \sum_{t=1}^{R} w_t \left(\hat{\beta}_i \right) x_{i,t} y_{t+1}, \tag{7.44}$$

where $w_t \left(\hat{\beta}_i \right) = \exp \left(- \frac{\left(y_{t+1} - x'_{i,t}\hat{\beta}_i \right)^2}{2\delta_R^2} \right)$.

For Model 1, we regress $\{y_s\}_{s=t-R+1}^{t}$ on constant term to get $\hat{c}_{1,t}$, where $t = R, \cdots, T$. For Model 2, we regress $\{y_s\}_{s=t-R+1}^{t}$ on $\{1, x_{s-1}\}_{s=t-R+1}^{t}$ to get $\hat{a}_{2,t}$ and $\hat{b}_t$. The forecast errors are $\hat{u}_{t+1}^{(1)} = y_{t+1} - \hat{a}_{1,t}$ for Model 1 and $\hat{u}_{t+1}^{(1)} = y_{t+1} - \hat{a}_{2,t} - \hat{b}_t x_t$, where $t = R, \cdots, T$. We choose to use rolling window for estimation. The in-sample observations $R \in \{120, 240, 480\}$ and the out-of-sample observations $P \in \{240, 480, 1200\}$. We repeat 2000 times to find out the distributions of DM_P, ENC_P and CCS_P.

7.6.2 Simulation Results

Figures 1-4 and Tables 1-4 show the Monte Carlo simulation distribution and the size and power of the DM_P, ENC_P, and CCS_P statistics in DGP 1 and DGP 2 with different skewness. Table 1 shows the size of the test under $\mathbb{H}_0$ with different skewness η in DGP 1. This table demonstrates that DM_P is much less than 5% under the 5% nominal level, which means the DM_P has a bias under the null hypothesis. However, compared to the DM_P, the size of the ENC_P and CCS_P is good under the 5% nominal level. Table 2 shows the size of the test under $\mathbb{H}_0$ with different skewness in DGP 2. This table shows that the size of the ENC_P is good under the 5% nominal level, but the size of the DM_P and CCS_P is not good. Table 3 shows the power of the test under $\mathbb{H}_1$ with different skewness η in DGP 1. This table demonstrates that the ENC_P has the highest power for different in-sample observations R, out-of-sample forecasts P, and skewness η in DGP 1. Moreover, the power of the CCS_P is lower than the $\mathrm{ENC_P}$ but is higher than the DM_P.

Figure 1 shows the distributions of the DM_P, ENC_P, and CCS_P statistics with the skewness $\eta = 0.5$ under $\mathbb{H}_0$ in DGP 1. These figures in Figure 1 demonstrate that the $\mathrm{DM_P}$ has a negative mean and high kurtosis, implying that the DM_P has a downward bias under $\mathbb{H}_0$. The distributions of the ENC_P and $\mathrm{CCS_P}$ are close to the standard normal distribution under the null hypothesis. Figure 2 shows the distributions of the DM_P, ENC_P, and CCS_P statistics with the skewness $\eta = 2$ under $\mathbb{H}_0$ in DGP 2. The distribution of ENC_P is close to the standard normal distribution under the null hypothesis, but the distributions of DM_P and CCS_P are not standard normal. Figure 3 shows the asymptotic distribution of the DM_P, $\mathrm{ENC_P}$, and CCS_P statistics with the skewness $\eta = 0.5$ under $\mathbb{H}_1$ in DGP 1. From

these figures in Figure 3, we can see that the mean of the DM_P is lower than the means of the ENC_P and the CCS_P.

7.7 Empirical Analysis

In this section, we use the annual data up to 2019 from Welch and Goyal (2008). Using Augmented The dickey-Fuller test statistic, stock variance (svar), inflation rate (infl), book-to-market ratio (b/m), and long-term return (ltr) are stationary processes under 1% level, book-to-market ratio (b/m) is stationary under 5% level. Thus, we build two nested mode models to test if these four stationary variables Granger cause the equity premium. For Model 1, we find the mode using the previous R in-sample observations of the equity premium. Then, we use the estimated mode to predict the future equity premium at time $R+1$. For Model 2, we estimate the constant and covariance terms using the previous R in-sample observations of equity premium and independent variables. We then use these estimated coefficients and independent variables at time $R+1$ to predict the equity premium at time $R+1$.

We set that the forecast begins 20 years after the sample and the forecast after 1965. We use rolling windows to estimate. Table 5 shows the results of three statistics DM_P, ENC_P, and CCS_P and their p-values using different additional variables x_t with different R and P. Results in the first three columns are obtained when the forecast begins 20 years after the sample and results in columns 4-6 show the DM_P, ENC_P, and CCS_P and p-values for these three statistics when the forecast begins 1965.

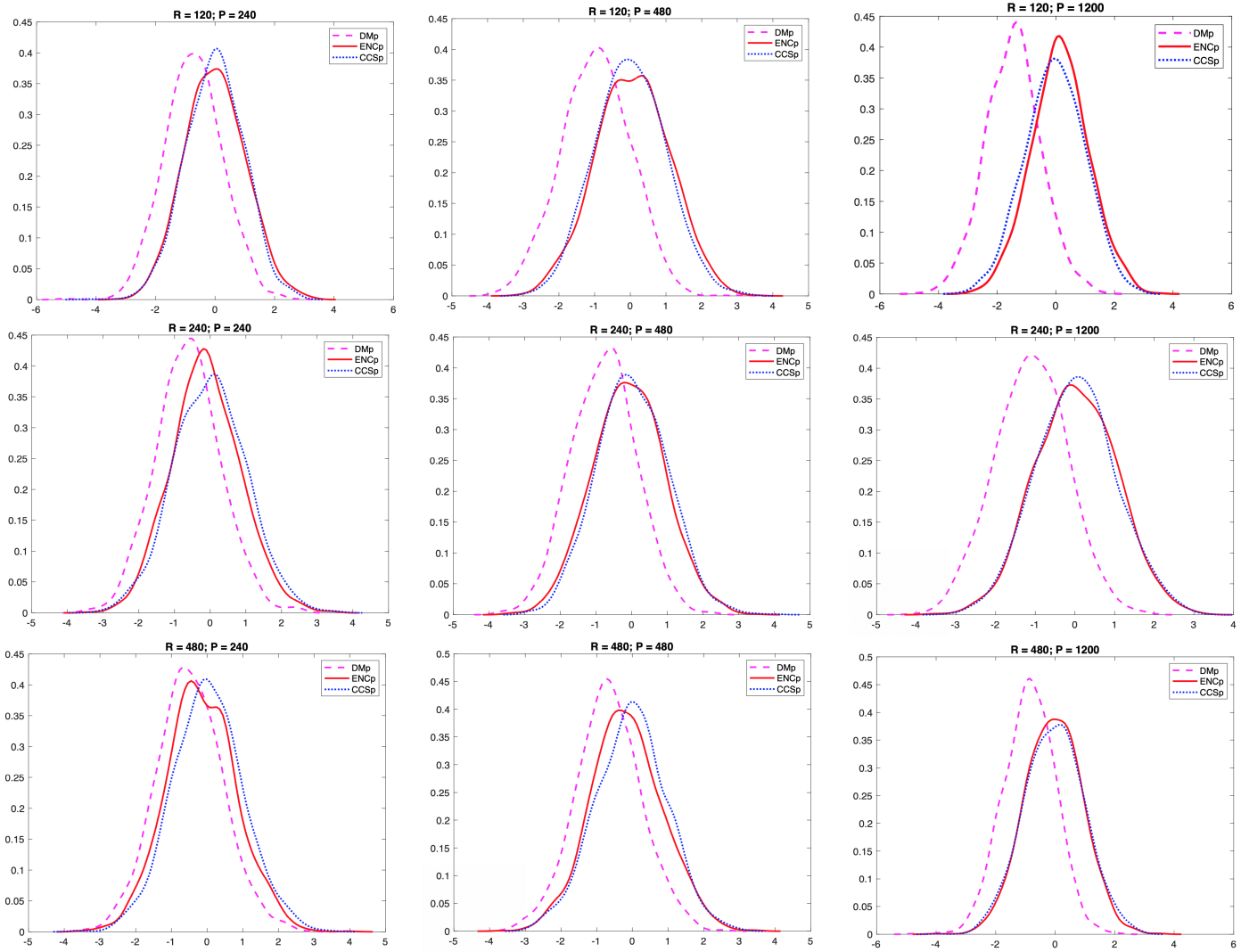

Figure 7.1: Simulation Results of DGP 1: Size of Test, $\eta = 0.5$

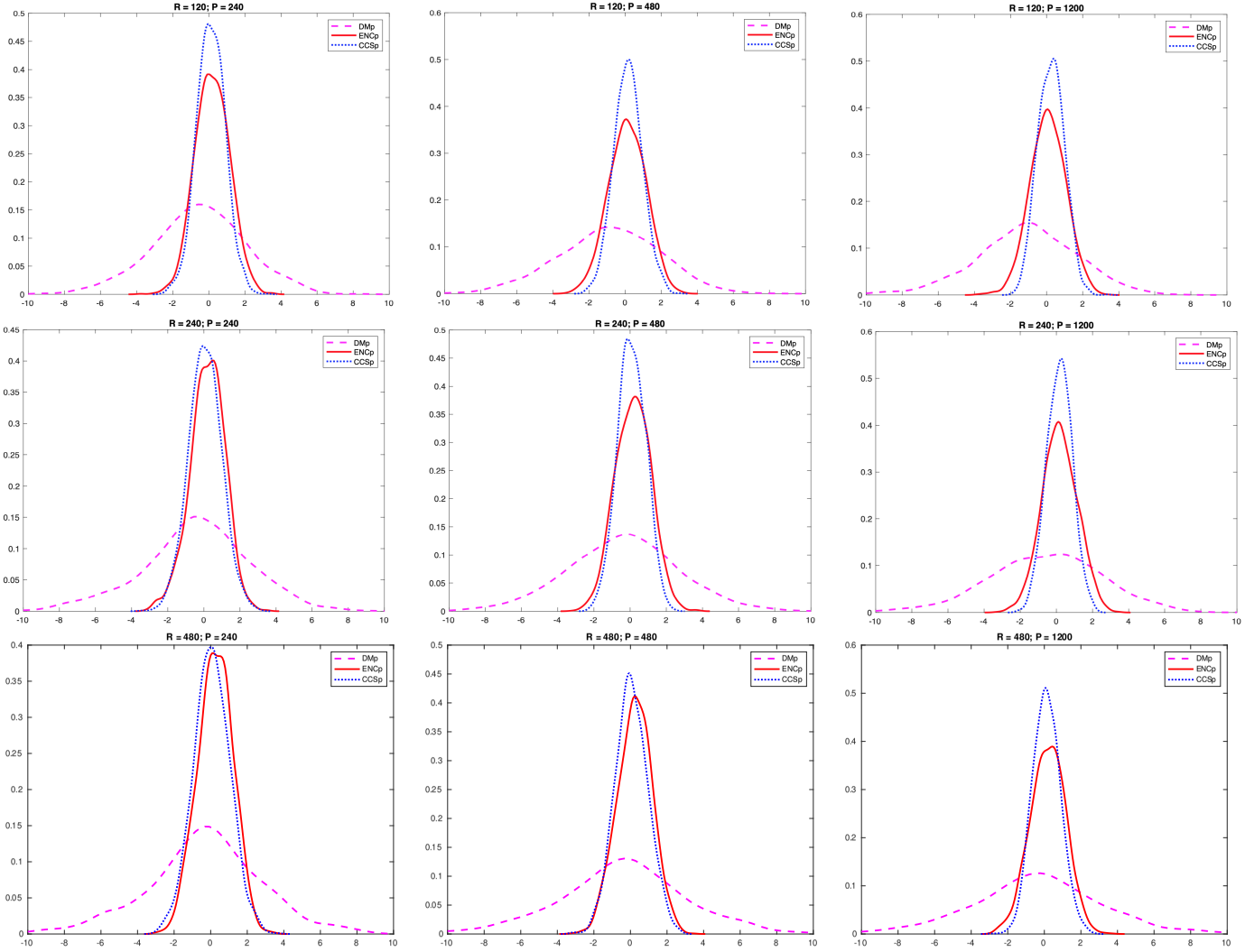

Figure 7.2: Simulation Results of DGP 2: Size of Test, $\alpha = 0.05$

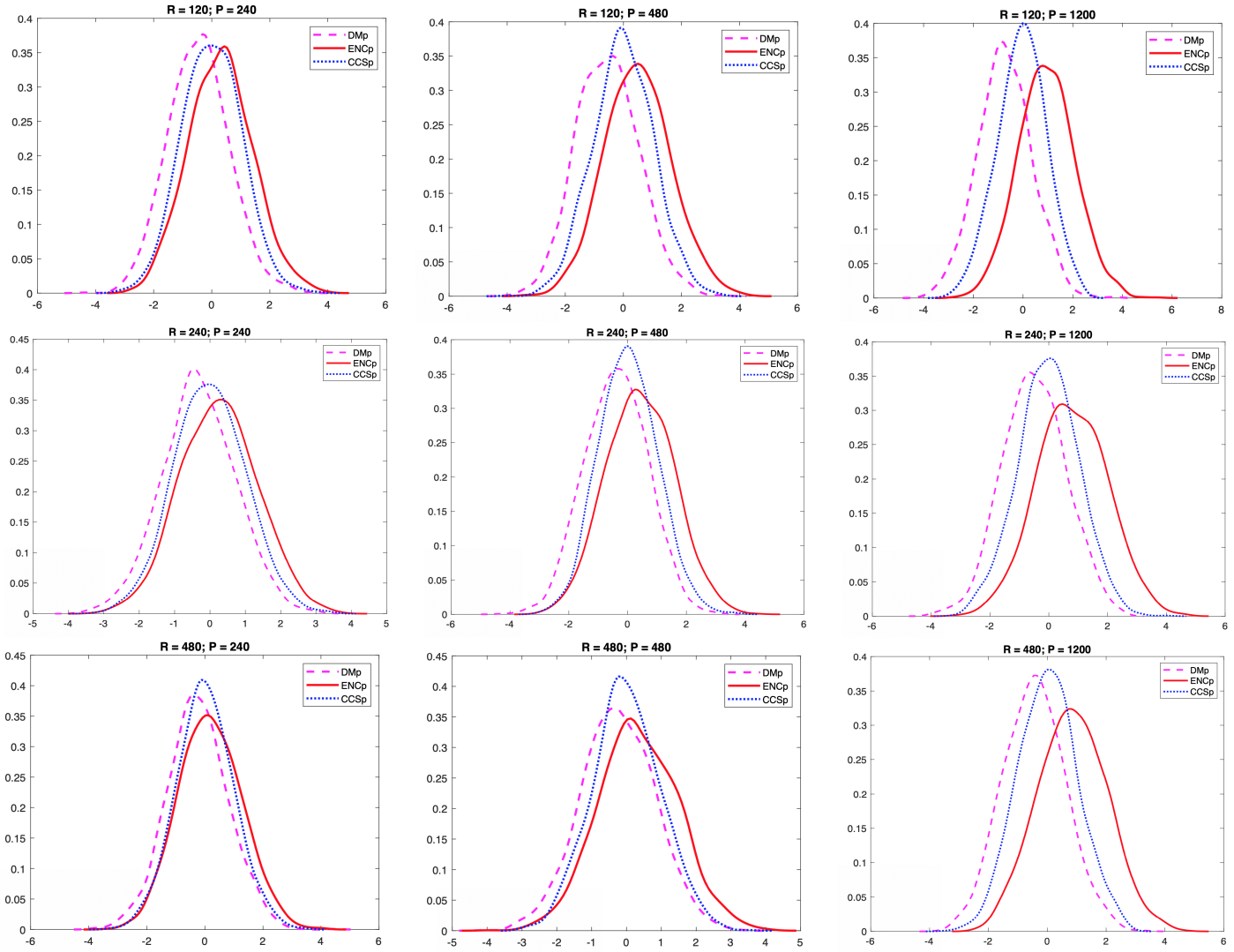

Figure 7.3: Simulation Results of DGP 1: Power of Test, $\eta = 0.5$

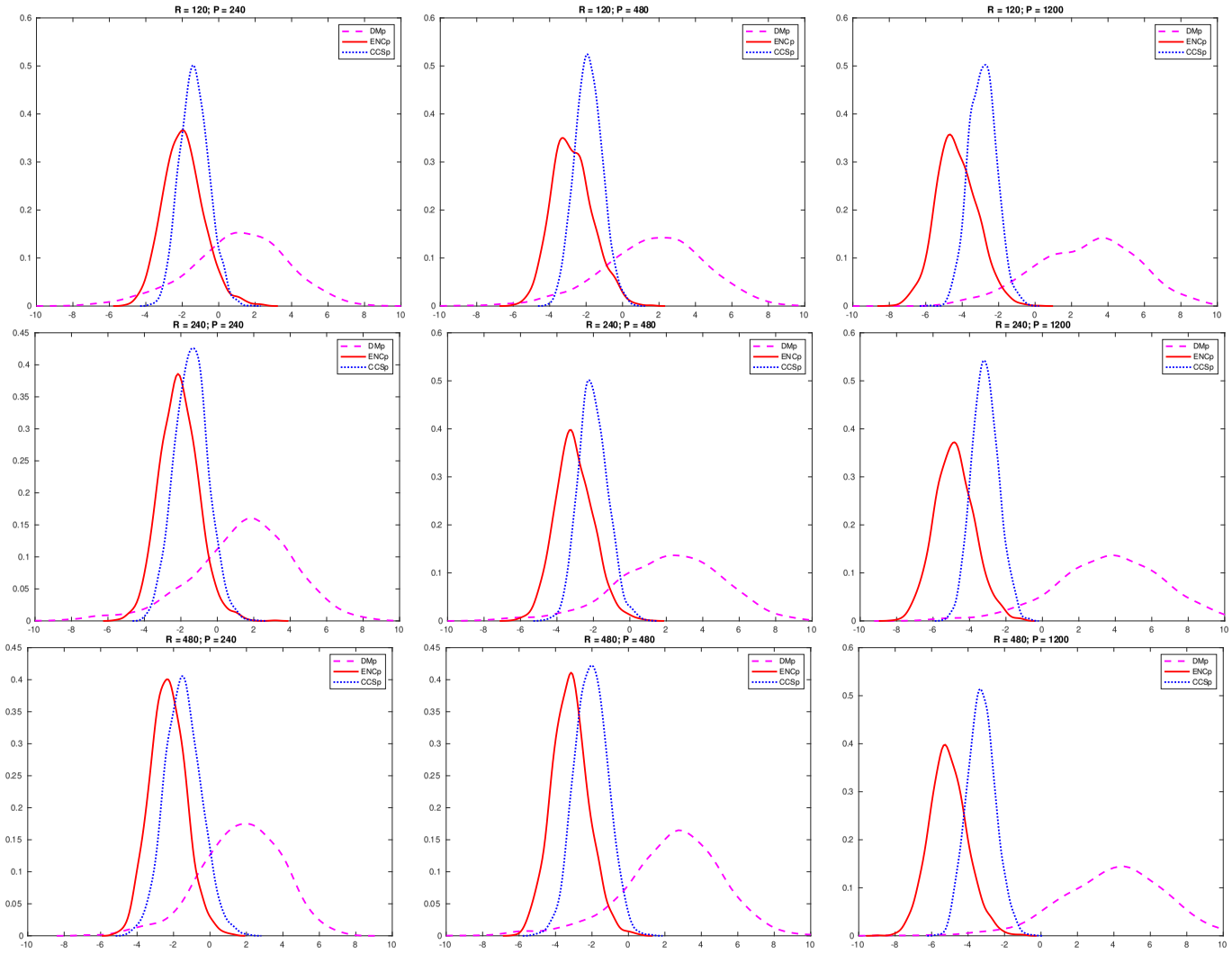

Figure 7.4: Simulation Results of DGP 2: Power of Test, $\alpha = 0.05$

From Table 5, we can see that the long-term return (ltr) Granger causes the equity premium at 5% significance level based on the ENC_P statistic when the forecast begins 20 years after the sample. However, the equity premium can be predicted by the inflation rate (infl) at 5% significance level based on the CCS_P statistic when the forecast begins 20 years after the sample and the forecast after 1965. Therefore, our prediction shows that the long-term return and inflation rate Granger-cause the equity premium in the mode using the Welch and Goyal (2008) data.

7.8 Conclusion

In this paper, we develop a new forecast encompassing test of Granger Causality in the conditional mode regression. The scoring function for the mode is based on the papers by Kemp and Silva (2012) and Dimitriadis et al. (2019). We show that the ENC statistic has zero mean under the null hypothesis of no GC in the mode. We also prove that the ENC statistic of the mode is asymptotically standard normal under the null hypothesis. Monte Carlo simulation demonstrates that the ENC statistic has good size and power and has a standard normal distribution in finite samples. Using the macroeconomic and financial predictors, we find out that the long-term return and inflation rate Granger-cause the equity premium in the conditional mode regression.

Appendix A: Proof of Theorem 3

Under $\mathbb{H}_0, u_{t+1}^{(1)} = u_{t+1}^{(2)} = u_{t+1}$. We have defined $\hat{u}_{t+1}^{(1)} = y_{t+1} - \hat{c}_{1,t} \equiv y_{t+1} - x_{1,t}'\hat{\beta}_{1,t}$ and $\hat{u}_{t+1}^{(2)} = y_{t+1} - \hat{c}_{2,t} - \hat{b}_t x_t \equiv y_{t+1} - x_{2,t}'\hat{\beta}_{2,t}$, where $x_{1,t}' = 1$ and $x_{2,t}' = (1, x_t)$. Taking the derivative of loss function with respect to β, we can get the score as follows:

$$h_{i,t} = \frac{\partial S_\delta^K(\gamma, y)}{\partial \beta} = -\frac{1}{\delta_R^2} K' \left(\frac{u_t^{(i)}}{\delta_R} \right) x_{i,t-1}. \tag{7.45}$$

Then, let $H_i(t)$ as follows:

$$H_i(t) = \frac{1}{R} \sum_{j=t-R+1}^{t} h_{i,t} = -\frac{1}{R} \sum_{j=t-R+1}^{t} \frac{1}{\delta_R^2} K' \left(\frac{u_t^{(i)}}{\delta} \right) x_{i,t-1}. \tag{7.46}$$

Taking the derivative of $h_{i,t}$ with respect to β, we can get the Hessian as follows:

$$\Lambda_{i,t} = \frac{\partial h_{i,t}}{\partial \beta} = \frac{1}{\delta_R^3} K'' \left(\frac{u_t^{(i)}}{\delta_R} \right) x_{i,t-1} x_{i,t-1}'. \tag{7.47}$$

According to Kemp and Silva (2012), we can obtain

$$\frac{1}{R} \sum_{j=t-R+1}^{t} \Lambda_{i,t} = \frac{1}{R} \sum_{j=t-R+1}^{t} \frac{1}{\delta_R^3} K'' \left(\frac{\hat{u}_t^{(i)}}{\delta_R} \right) x_{i,t-1} x_{i,t-1}' = B_i^{-1} + o_p(1), \tag{7.48}$$

where

$$B_i^{-1} = \mathbb{E}_F \left[f''_{u_t^{(i)}|X} \left(0|x_{i,t}\right) x_{i,t} x_{i,t}' \right] = \lim_{R\to\infty} \mathbb{E}_F \left[\frac{1}{\delta_R^3} \left\{ K'' \left(\frac{u_t^{(i)}}{\delta_R} \right) \right\} x_{i,t} x_{i,t}' \right], \tag{7.49}$$

where $f''_{u^{(i)}|X}\left(0|x_t\right) = \frac{\partial^2 L}{\partial\beta\partial\beta}\Big|_{\beta_0}$ and $\mathbb{E}_F f''_{u^{(i)}|X}\left(0|x_t\right)$ is negative definite. Therefore, $B_1 = 1$ and $B_2 = \begin{pmatrix} 1 & 0 \\ 0 & -E\left[f''_{u^{(i)}|X}(0|x_t) x_t x_t' \right] \end{pmatrix}^{-1}$. Let $k_1 = 1$ the number of parameters to be estimated in Model 1 and k_2 the additional numbers of parameters to be estimated in Model 2.Let J denote the selection matrix (1, 0) such that $Jh_{2,t} = h_{1,t}$. Let sup $_t$ denote sup $_{t-R+1\le\cdot\le t}$. And matrix A and C will defined in Lemma 2 and $\tilde{h}_{2,t} =$

$Z^{-1}A'CB_2^{1/2}h_{2,t}, \tilde{H}_2(t) = Z^{-1}A'CB_2^{1/2}H_2(t)$. U_t and $\tilde{U}_t$ will defined in Assumption 2. According to Kemp and Silva (2012), the asymptotic distribution of $\hat{\beta}_{i,t}$ is

$$\sqrt{R\delta_R^3}\left(\hat{\beta}_{i,t} - \beta_i\right) = -B_i^{-1}\left[\frac{1}{\sqrt{R\delta_R}}\sum_{j=t-R+1}^{t} K'\left(\frac{u_{i,t}}{\delta_R}\right)x_{i,t}\right] + o_p(1) \xrightarrow{d} N\left(0, B_i^{-1}A_iB_i^{-1}\right), \tag{7.50}$$

where $A_i = \lim_{R\to\infty}\mathbb{E}_F\left[\frac{1}{\delta_R}\left\{K'\left(\frac{u_i}{\delta_R}\right)\right\}^2 x_i x_i'\right]$.

By the central limit theorem,

$$\sqrt{R\delta_R^3}\left[R^{-1}\sum_{s=t-R+1}^{t} h_{2,s}\right] \xrightarrow{d} N\left(0, Z^2B_2^{-1}\right), \tag{7.51}$$

where

$$Z^2B_2^{-1} = A_2 = \lim_{R\to\infty}\mathbb{E}_F\left[\frac{1}{\delta_R}\left\{K'\left(\frac{u_t^{(2)}}{\delta_R}\right)\right\}^2 x_{2,t}x_{2,t}'\right]. \tag{7.52}$$

For simplificity, we denote $\sum_{s=t-R+1}^{t}$ as $\sum_t$. To complete the proof, we need the following lemmas.

Lemma 1: (Clark and McCracken, 2001) (Not-for-Publication Appendix Lemma A3) Let Assumptions 1-8 hold. For i=1,2, $\sum_t H_i'(t)B_ih_{i,t+1}h_{i,t+1}'B_jH_j'(t) = \sum_t H_i'(t)B_i\mathbb{E}(h_{i,t+1}h_{i,t+1}')$ $B_jH_j'(t) + o_p(t)$.

Proof: Adding and subtracting $\mathbb{E}(h_{i,t+1}h_{i,t+1}')$,

$$\sum_t H_i'(t)B_ih_{i,t+1}h_{i,t+1}'B_jH_j(t)$$

$$= \sum_t H_i'(t)B_i\mathbb{E}(h_{i,t+1}h_{i,t+1}')B_jH_j(t) + \sum_t H_i'(t)B_i\left(h_{i,t+1}h_{i,t+1}' - \mathbb{E}(h_{i,t+1}h_{i,t+1}')\right)B_jH_j(t).$$

Now, we need to show the second term is $o_p(1)$. We have

$$\begin{aligned}
&\sum_t H_i'(t)B_i\left(h_{i,t+1}h_{i,t+1}' - \mathbb{E}(h_{i,t+1}h_{i,t+1}')\right)B_jH_j(t) \\
&= T^{-1/2}\sum_t (\frac{T}{t})^2\left[T^{-1/2}\sum_{s=1}^{t} h_{j,s}'B_j \otimes T^{-1/2}\sum_{s=1}^{t} h_{i,s}'B_i\right] \\
&\quad \times \text{vec}\left[T^{-1/2}\left(h_{i,t+1}h_{i,t+1}' - \mathbb{E}(h_{i,t+1}h_{i,t+1}')\right)\right].
\end{aligned}$$

$\sum_t (\frac{T}{t})^2\left[\frac{1}{\sqrt{T}}\sum_{s=1}^{t} h_{j,s}'B_j \otimes \frac{1}{\sqrt{T}}\sum_{s=1}^{t} h_{i,s}'B_i\right]\text{vec}\left[\frac{1}{\sqrt{T}}\left(h_{i,t+1}h_{i,t+1}' - \mathbb{E}(h_{i,t+1}h_{i,t+1}')\right)\right]$ is $Op(1)$, which follows from Assumption 2, Corollary 29.19 of Davidson (1994) and Theorem 3.1 of (Hansen, 1992). The proof is complete.

Lemma 2: Let Assumptions 1-7 hold. Let $-J'B_1J + B_2 = M$ and $B_2^{-1/2}MB_2^{-1/2} = Q$, then Q is idempotent. Define matrix $A = (0,1)'$ and $C = I_{2\times 2}$ Then $Q = CAA'C$.

Proof:

$$\begin{aligned}
Q &\equiv B_2^{-1/2}MB_2^{-1/2} \\
&= \begin{pmatrix} 1 & 0 \\ 0 & -E\left[f''_{u_t^{(i)}|X}(0 \mid x_t)x_tx_t'\right] \end{pmatrix}^{1/2} \begin{pmatrix} 1 & 0 \\ 0 & -E\left[f''_{u_t^{(i)}|X}(0 \mid x_t)x_{i,t}x_t'\right] \end{pmatrix}^{-1} \times \\
&\quad \begin{pmatrix} 1 & 0 \\ 0 & -E\left[f''_{u_t^{(i)}|X}(0 \mid x_t)x_tx_t'\right] \end{pmatrix}^{1/2} \\
&= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},
\end{aligned}$$

so Q is an idempotent matrix with rank 1.

Lemma 3: (Clark and McCracken, 2001) (Not-for-Publication Appendix Lemma A5 and A6) Let Assumptions 1-7 hold. Define $\zeta = R/T$.

$\sum_t \tilde{H}_2'(t)\tilde{h}_{2,t+1} \xrightarrow{d} \zeta^{-1} \int_\zeta^1 [W(s) - W(s-\zeta)]\, \mathrm{d}W(s)$.

Proof:

$$\begin{aligned}
\sum_t \tilde{H}_2'(t)\tilde{h}_{2,t+1} &= \sum_t \left(\frac{1}{R} \sum_{j=t-R+1}^{t} \tilde{h}_{2,j}' \right) \tilde{h}_{2,t+1} \\
&= \sum_t \left(\frac{1}{R} \sum_{j=1}^{t} \tilde{h}_{2,j}' - \frac{1}{R} \sum_{j=1}^{t-R} \tilde{h}_{2,j}' \right) \tilde{h}_{2,t+1} \\
&= \frac{T}{R} \left[\sum_t \left(T^{-1/2} \sum_{j=1}^{t} \tilde{h}_{2,j}' - T^{-1/2} \sum_{j=1}^{t-R} \tilde{h}_{2,j}' \right) \left(T^{-1/2} \tilde{h}_{2,t+1} \right) \right].
\end{aligned}$$

According to the continuous mapping theorem and Corollary 29.19 of Davidson (1994), $T^{-1/2} \sum_{j=1}^{t} \tilde{h}_{2,j}' \Longrightarrow W(s)$, $\quad T^{-1/2} \sum_{j=1}^{t-R} \tilde{h}_{2,j}' \Longrightarrow W(s-\zeta)$ and $T^{-1/2}\tilde{h}_{2,t+1} \Longrightarrow \mathrm{d}W(s)$. Since $T/R = \zeta^{-1}$, the proof is complete.

Lemma 4: (Clark and McCracken (2001) (Not-for-Publication Appendix Lemma A5 and A7) Let Assumptions 1-7 hold. $\sum_t \tilde{H}_2'(t)\tilde{H}_2(t) \xrightarrow{d} \zeta^{-2} \int_\zeta^1 [W(s) - W(s-\zeta)]^2\, \mathrm{d}s$.

Proof:

$$\begin{aligned}
\sum_t \tilde{H}_2'(t)\tilde{H}_2(t) &= \sum_t \left(\frac{1}{R} \sum_{j=t-R+1}^{t} \tilde{h}_{2,j}' \right)' \left(\frac{1}{R} \sum_{j=t-R+1}^{t} \tilde{h}_{2,j}' \right) \\
&= \sum_t \left(\frac{1}{R} \sum_{j=1}^{t} \tilde{h}_{2,j}' - \frac{1}{R} \sum_{j=1}^{t-R} \tilde{h}_{2,j}' \right)' \left(\frac{1}{R} \sum_{j=1}^{t} \tilde{h}_{2,j}' - \frac{1}{R} \sum_{j=1}^{t-R} \tilde{h}_{2,j}' \right) \\
&= \frac{1}{T} \left(\frac{T}{R} \right)^2 \sum_t \left(T^{-1/2} \sum_{j=1}^{t} \tilde{h}_{2,j}' - T^{-1/2} \sum_{j=1}^{t-R} \tilde{h}_{2,j}' \right)' \\
&\quad \times \left(T^{-1/2} \sum_{j=1}^{t} \tilde{h}_{2,j}' - T^{-1/2} \sum_{j=1}^{t-R} \tilde{h}_{2,j}' \right).
\end{aligned}$$

According to the continuous mapping theorem and Corollary 29.19 of Davidson (1994), $T^{-1/2}\sum_{j=1}^{t}\tilde{h}'_{2,j} \Longrightarrow W(s), T^{-1/2}\sum_{j=1}^{t-R}\tilde{h}'_{2,j} \Longrightarrow W(s-\zeta), T^{-1/2}\tilde{h}_{2,t+1} \Longrightarrow \mathrm{d}W(s)$ and $1/T \Longrightarrow \mathrm{d}s$. Since $T/R = \zeta^{-1}$, the proof is complete.

Lemma 5: Let Assumptions 1-8 hold.

$$\sum_t \frac{1}{\delta_R^2} K'\left(\frac{\hat{u}_{t+1}^{(1)}}{\delta_R}\right)\left(\hat{u}_{t+1}^{(1)} - \hat{u}_{t+1}^{(2)}\right) = Z^2 \sum_t \tilde{H}_2'(t)\tilde{h}_{2,t+1} + o_p(1). \tag{7.53}$$

Proof: We have $\hat{u}_{1,t+1} = u_{t+1} - x'_{1,t}\left(\hat{\beta}_{1,t} - \beta_{1,t}\right)$ and $\hat{u}_{2,t+1} = u_{t+1} - x'_{2,t+1}\left(\hat{\beta}_{2,t} - \beta_{2,t}\right)$, so $\hat{u}_{1,t+1} - \hat{u}_{2,t+1} = x'_{2,t}\left(\hat{\beta}_{2,t} - \beta_{2,t}\right) - x'_{1,t}\left(\hat{\beta}_{1,t} - \beta_{1,t}\right)$.

$$\begin{aligned}
&\sum_t \frac{1}{\delta_R^2} K'\left(\frac{\hat{u}_{t+1}^{(1)}}{\delta_R}\right)\left(\hat{u}_{t+1}^{(1)} - \hat{u}_{t+1}^{(2)}\right) \\
&= \sum_t \frac{1}{\delta_R^2} K'\left(\frac{\hat{u}_{t+1}^{(1)}}{\delta_R}\right)\left[x'_{2,t}\left(\hat{\beta}_{2,t} - \beta_{2,t}\right) - x'_{1,t}\left(\hat{\beta}_{1,t} - \beta_{1,t}\right)\right] \\
&= \sum_t \frac{1}{\delta_R^2}\left[K'\left(\frac{u_{t+1}^{(1)}}{\delta_R}\right) + \frac{1}{\delta_R}K''\left(\hat{u}_{t+1}^{(1)} - u_{t+1}^{(1)}\right) + O_p\left(\left(\hat{u}_{t+1}^{(1)} - u_{t+1}^{(1)}\right)^2\right)\right] \times \\
&\quad \left[x'_{2,t}\left(\hat{\beta}_{2,t} - \beta_{2,t}\right) - x'_{1,t}\left(\hat{\beta}_{1,t} - \beta_{1,t}\right)\right] \\
&= \underbrace{\sum_t \frac{1}{\delta_R^2} K'\left(\frac{u_{t+1}^{(1)}}{\delta_R}\right)\left[x'_{2,t}\left(\hat{\beta}_{2,t} - \beta_{2,t}\right) - x'_{1,t}\left(\hat{\beta}_{1,t} - \beta_{1,t}\right)\right]}_{A} \\
&\quad + \underbrace{\sum_t \frac{1}{\delta_R^3} K''\left(\frac{u_{t+1}^{(1)}}{\delta_R}\right)\left(\hat{u}_{t+1}^{(1)} - u_{t+1}^{(1)}\right)\left[x'_{2,t}\left(\hat{\beta}_{2,t} - \beta_{2,t}\right) - x'_{1,t}\left(\hat{\beta}_{1,t} - \beta_{1,t}\right)\right]}_{B_1} \\
&\quad + \underbrace{\sum_t \frac{1}{\delta_R^2} O_p\left(\left(\hat{u}_{t+1}^{(1)} - u_{t+1}^{(1)}\right)^2\right)\left[x'_{2,t}\left(\hat{\beta}_{2,t} - \beta_{2,t}\right) - x'_{1,t}\left(\hat{\beta}_{1,t} - \beta_{1,t}\right)\right]}_{B_2}.
\end{aligned}$$

$$
\begin{aligned}
A &= \sum_t \frac{1}{\delta_R^2} K'\left(\frac{u_{t+1}^{(1)}}{\delta_R}\right)\left[x_{2,t}'\left(\hat{\beta}_{2,t}-\beta_{2,t}\right)-x_{1,t}'\left(\hat{\beta}_{1,t}-\beta_{1,t}\right)\right] \\
&= \sum_t \left[-\frac{1}{\delta_R^2} K'\left(\frac{u_{t+1}^{(1)}}{\delta_R}\right) x_{1,t}'\left(\hat{\beta}_{1,t}-\beta_{1,t}\right)+\frac{1}{\delta_R^2} K'\left(\frac{u_{t+1}^{(1)}}{\delta_R}\right) x_{2,t}'\left(\hat{\beta}_{2,t}-\beta_{2,t}\right)\right] \\
&= \sum_t \left[-h_{1,t+1}' B_1 H_1(t)+h_{2,t+1}' B_2 H_2(t)\right]+o_p(1) \\
&= \sum_t \left[-h_{2,t+1}' J' B_1 J H_2(t)+h_{2,t+1}' B_2 H_2(t)\right]+o_p(1) \\
&= \sum_t \left[h_{2,t+1}' M H_2(t)\right]+o_p(1) \\
&= \sum_t \left[h_{2,t+1}' B_2^{1/2} B_2^{-1/2} M B_2^{-1/2} B_2^{1/2} H_2(t)\right]+o_p(1) \\
&= \sum_t \left[h_{2,t+1}' B_2^{1/2} Q B_2^{1/2} H_2(t)\right]+o_p(1) \\
&= \sum_t \left[h_{2,t+1}' B_2^{1/2} C A A' C B_2^{1/2} H_2(t)\right]+o_p(1) \\
&= Z^2 \sum_t \tilde{H}_2'(t)\tilde{h}_{2,t+1}+o_p(1).
\end{aligned}
$$

Now, we need to show that B1 and B2 have lower order than A. Kemp and Silva (2012) show that $\hat{\beta}-\beta = O_p(R^{-1/2}\delta_R^{-3/2})$. The limitations of condition in Assumption 5 imply that $\delta_R \propto R^{-\kappa}$ where $\kappa \in (1/7, 1)$. Thus, the fastest convergence of $\hat{\beta}-\beta$ is obtained by choosing $\delta_R \approx R^{-1/7}$ resulting in a convergence rates close to $R^{2/7}$.

$$
\begin{aligned}
A &= \sum_t \frac{1}{\delta_R^2} K'\left(\frac{u_{t+1}^{(1)}}{\delta_R}\right)\left[x_{2,t}'\left(\hat{\beta}_{2,t}-\beta_{2,t}\right)-x_{1,t}'\left(\hat{\beta}_{1,t}-\beta_{1,t}\right)\right] = PO_p(1)O_p\left(\frac{1}{R^{2/7}}\right) \\
&= O_p\left(\frac{P}{R^{2/7}}\right). \\
B1 &= \sum_t \frac{1}{\delta_R^3} K''\left(\frac{u_{t+1}^{(1)}}{\delta_R}\right)\left(\hat{u}_{t+1}^{(1)}-u_{t+1}^{(1)}\right)\left[x_{2,t}'\left(\hat{\beta}_{2,t}-\beta_{2,t}\right)-x_{1,t}'\left(\hat{\beta}_{1,t}-\beta_{1,t}\right)\right] \\
&= PO_p(\frac{1}{R^{2/7}})O_p\left(\frac{1}{R^{2/7}}\right) = O_p\left(\frac{P}{R^{4/7}}\right).
\end{aligned}
$$

$$B2 = \sum_t \frac{1}{\delta_R^2} O_p\left(\left(\hat{u}_{t+1}^{(1)} - u_{t+1}^{(1)}\right)^2\right)\left[x_{2,t}'\left(\hat{\beta}_{2,t} - \beta_{2,t}\right) - x_{1,t}'\left(\hat{\beta}_{1,t} - \beta_{1,t}\right)\right]$$
$$= PO_p\left(\frac{1}{R^{4/7}}\right) O_p\left(\frac{1}{R^{2/7}}\right) = O_p\left(\frac{P}{R^{6/7}}\right).$$

Therefore, B1 and B2 have a lower order than A. The proof is complete.

Lemma 6: Let Assumptions 1-8 hold.

$$\sum_t \left[\frac{1}{\delta_R^2} K'\left(\frac{\hat{u}_{t+1}^{(1)}}{\delta_R}\right)\left(\hat{u}_{t+1}^{(1)} - \hat{u}_{t+1}^{(2)}\right)\right]^2 - P\bar{C}^2 = Z^4 \sum_t \tilde{H}_2'(t)\tilde{H}_2(t) + o_p(1). \tag{7.54}$$

Proof:

$$\sum_t \left[\frac{1}{\delta_R^2} K'\left(\frac{\hat{u}_{t+1}^{(1)}}{\delta_R}\right)\left(\hat{u}_{t+1}^{(1)} - \hat{u}_{t+1}^{(2)}\right)\right]^2$$
$$= \sum_t \left[-h_{1,t+1}' B_1(t) H_1(t) + h_{2,t+1}' B_2(t) H_2(t)\right]^2 + o_p(1).$$
$$= \sum_t H_1'(t) B_1'(t) h_{1,t+1} h_{1,t+1}' B_1(t) H_1(t) + \sum_t H_2'(t) B_2'(t) h_{2,t+1} h_{2,t+1}' B_2(t) H_2(t)$$
$$- 2\sum_t H_1'(t) B_1'(t) h_{1,t+1} h_{2,t+1} B_2(t) H_2(t) + o_p(1)$$
$$= \sum_t H_1'(t) B_1'(t) \mathbb{E}\left(h_{1,t+1} h_{1,t+1}'\right) B_1(t) H_1(t) + \sum_t H_2'(t) B_2'(t) \mathbb{E}\left(h_{2,t+1} h_{2,t+1}'\right) B_2(t) H_2(t)$$
$$- 2\sum_t H_1'(t) B_1'(t) \mathbb{E}\left(h_{1,t+1} h_{2,t+1}'\right) B_2(t) H_2(t) + o_p(1)$$
$$= Z^2 \sum_t H_2'(t) J' B_1 J H_2(t) + Z^2 \sum_t H_2'(t) B_2 H_2(t) - 2Z^2 \sum_t H_2'(t) J' B_1 J H_2(t) + o_p(1)$$
$$= Z^2 \sum_t H_2'(t)\left[-J' B_1 J + B_2\right] H_2(t) + o_p(1)$$
$$= Z^2 \sum_t H_2'(t) M H_2(t) + o_p(1)$$
$$= Z^2 \sum_t H_2'(t) B_2^{1/2} B_2^{-1/2} M B_2^{-1/2} B_2^{1/2} H_2(t) + o_p(1)$$

$$= Z^2 \sum_t H_2'(t) B_2^{1/2} CAA'CB_2^{1/2} H_2(t) + o_p(1)$$

$$= Z^4 \sum_t \tilde{H}_2'(t) \tilde{H}_2(t) + o_p(1).$$

Lemma 4 implies that $\bar{C} = O_p(P^{-1})$, so $P\bar{C}^2 = o_p(1)$. Thus, the proof is completed.

Table 7.1: Simulation Results of DGP 1: Size of Test

	$P = 240$			$P = 480$			$P = 1200$		
	DM_P	ENC_P	CCS_P	DM_P	ENC_P	CCS_P	DM_P	ENC_P	CCS_P
$\eta = -0.5$									
$R = 120$	0.003	0.045	0.055	0.001	0.049	0.046	0.001	0.060	0.056
$R = 240$	0.010	0.040	0.052	0.003	0.044	0.046	0.002	0.054	0.060
$R = 480$	0.009	0.036	0.050	0.002	0.045	0.049	0.003	0.044	0.048
$\eta = -0.25$									
$R = 120$	0.005	0.046	0.053	0.002	0.051	0.047	0.001	0.054	0.053
$R = 240$	0.007	0.048	0.053	0.003	0.047	0.048	0.002	0.051	0.050
$R = 480$	0.008	0.044	0.054	0.005	0.043	0.048	0.003	0.052	0.052
$\eta = 0.25$									
$R = 120$	0.003	0.049	0.051	0.001	0.053	0.045	0.001	0.057	0.052
$R = 240$	0.005	0.046	0.050	0.003	0.049	0.049	0.002	0.053	0.051
$R = 480$	0.008	0.044	0.053	0.004	0.041	0.047	0.003	0.053	0.052
$\eta = 0.5$									
$R = 120$	0.007	0.053	0.053	0.001	0.055	0.055	0.001	0.062	0.052
$R = 240$	0.010	0.042	0.055	0.004	0.049	0.054	0.002	0.055	0.053
$R = 480$	0.014	0.043	0.056	0.005	0.040	0.048	0.002	0.043	0.044

This table shows the size of DM_P, ENC_P, and CCS_P test in DGP1 under $\mathbb{H}_0$ under 5% nominal level, $k = 1.6$, $c_2 = 0$, $b = 0$, $\phi = 0$, $\sigma_u = 1$.

Table 7.2: Simulation Results of DGP 2: Size of Test

	$P = 240$			$P = 480$			$P = 1200$		
	DM_P	ENC_P	CCS_P	DM_P	ENC_P	CCS_P	DM_P	ENC_P	CCS_P
$\alpha = 5$									
$R = 120$	0.181	0.047	0.022	0.173	0.052	0.019	0.164	0.053	0.012
$R = 240$	0.219	0.048	0.040	0.230	0.043	0.021	0.237	0.051	0.009
$R = 480$	0.248	0.045	0.055	0.257	0.047	0.027	0.261	0.047	0.015
$\alpha = 0.05$									
$R = 120$	0.197	0.044	0.018	0.182	0.053	0.016	0.162	0.051	0.009
$R = 240$	0.229	0.046	0.037	0.244	0.046	0.011	0.240	0.050	0.005
$R = 480$	0.250	0.043	0.052	0.261	0.043	0.029	0.264	0.046	0.013

This table shows the size of DM_P, ENC_P, and CCS_P test in DGP 2 under $\mathbb{H}_0$ under the 5% nominal level, $k = 1.6$, $b_1 = 0$, $b_0 = 0$. $\alpha = 5$, the coefficient of skewness of u_{t+1} is approximately 0.5. $\alpha = 0.05$, the coefficient of skewness of u_{t+1} is approximately 2.

Table 7.3: Simulation Results of DGP 1: Power of Test

	$P = 240$			$P = 480$			$P = 1200$		
	DM_P	ENC_P	CCS_P	DM_P	ENC_P	CCS_P	DM_P	ENC_P	CCS_P
$\eta = -0.5$									
$R = 120$	0.020	0.129	0.046	0.021	0.169	0.056	0.016	0.277	0.053
$R = 240$	0.031	0.120	0.049	0.027	0.158	0.052	0.022	0.272	0.056
$R = 480$	0.055	0.130	0.052	0.031	0.141	0.053	0.024	0.281	0.061
$\eta = -0.25$									
$R = 120$	0.017	0.136	0.051	0.030	0.269	0.051	0.016	0.243	0.053
$R = 240$	0.029	0.131	0.051	0.034	0.211	0.053	0.032	0.248	0.054
$R = 480$	0.041	0.127	0.053	0.041	0.235	0.054	0.040	0.254	0.056
$\eta = 0.25$									
$R = 120$	0.018	0.133	0.049	0.023	0.263	0.051	0.019	0.240	0.051
$R = 240$	0.031	0.125	0.051	0.037	0.197	0.053	0.030	0.245	0.052
$R = 480$	0.039	0.122	0.056	0.043	0.235	0.052	0.041	0.252	0.054
$\eta = 0.5$									
$R = 120$	0.028	0.121	0.049	0.023	0.166	0.055	0.013	0.271	0.048
$R = 240$	0.029	0.115	0.052	0.022	0.150	0.050	0.020	0.268	0.055
$R = 480$	0.046	0.103	0.056	0.030	0.134	0.053	0.022	0.280	0.062

This table shows the power of DM_P, ENC_P, and CCS_P test in DGP1 under $\mathbb{H}_1$ under the 5% nominal level, $k = 1.6$, $b_1 = 0.1$, $b_0 = 0$, $\phi = 0$, $\sigma_u = 1$.

Table 7.4: Simulation Results of DGP 2: Size of Test

	$P = 240$			$P = 480$			$P = 1200$		
	DM_P	ENC_P	CCS_P	DM_P	ENC_P	CCS_P	DM_P	ENC_P	CCS_P
$\alpha = 5$									
$R = 120$	0.354	0.497	0.189	0.445	0.682	0.401	0.572	0.853	0.803
$R = 240$	0.477	0.526	0.225	0.542	0.798	0.469	0.704	0.952	0.916
$R = 480$	0.512	0.632	0.297	0.623	0.843	0.537	0.796	0.987	0.934
$\alpha = 0.05$									
$R = 120$	0.429	0.512	0.219	0.527	0.773	0.461	0.657	0.975	0.889
$R = 240$	0.503	0.569	0.284	0.588	0.845	0.550	0.762	0.994	0.940
$R = 480$	0.528	0.647	0.318	0.660	0.891	0.542	0.801	0.997	0.950

This table shows the power of DM_P, ENC_P, and CCS_P test in DGP 2 under $\mathbb{H}_0$ under the 5% nominal level, $k = 1.6$, $b_1 = 0.1$, $b_0 = 0$. $\alpha = 5$, the coefficient of skewness of u_{t+1} is approximately 0.5. $\alpha = 0.05$, the coefficient of skewness of u_{t+1} is approximately 2.

Table 7.5: Predictive Modal Regression (Monthly Data)

	Forecast begin 20 years			Forecast begin 1965		
	DM_P	ENC_P	CCS_P	DM_P	ENC_P	CCS_P
svar	-1.2738	1.4273	-1.406	-1.3089	-0.4438	1.4015
p-value	0.8986	0.1535	0.1597	0.9047	0.6572	0.1611
infl	1.1131	-1.8062	2.2323	0.8413	-1.6808	2.2040
p-value	0.1328	0.0709	0.0256	0.2001	0.0928	0.0275
b/m	-0.8322	0.5138	1.5236	-1.5673	0.3858	1.8843
p-value	0.7974	0.6074	0.1276	0.9415	0.6997	0.0595
ltr	-2.1092	2.2888	-0.0992	-4.2166	1.1018	-0.5410
p-value	0.9825	0.0221	0.9209	0.9999	0.2706	0.5885

Chapter 8

Forecast Encompassing Test for Granger Causality in Predictive Model of Skewness and Kurtosis

8.1 Introduction

The landmark work Markowitz (1952) suggested a framework for asset allocation that set the fundamentals of modern portfolio theory. Markowitz formalizes the intuition that investors try to maximize their expected return while minimizing risk with a theory for investors' decision-making. Based on the two-parameter (mean-variance) model of portfolio selection, Markowitz's normative theory led to the development of the theory of the valuation of risk assets by Sharpe (1964) and Lintner (1965), whose Capital Asset Pricing Model (CAPM) is ubiquitous in financial applications. Markowitz's mean-variance framework and the CAPM

have come under much scrutiny over the last 60 years. By defining risk as the variance of future returns, two underpinnings of the CAPM are well known. The first is that asset returns are assumed to be normal, or more precisely, elliptically distributed. The second is that investors have mean-variance preferences when investors have quadratic utility functions that are only defined by their means and variances. However, we may ask if both assumptions hold in practice. It is reasonable to expect that, when skewness is present, it may account for misleading results. Studies have shown that approaches that model the returns using a normal distribution, whether explicitly or implicitly, may not perform satisfactorily when applied to real data. In fact, the role of skewness in return distributions has been relatively neglected before, and become a vibrant topic of discussion nowadays. It is widely recognized in the literature that financial returns do not always follow elliptically symmetric distributions. They generally exhibit non-normal features such as skewness, heavy tails, and kurtosis.

Numerous studies in the empirical finance literature have reported evidence of skewness or asymmetries in the distribution of financial returns, over the last four decades. See Friend and Westerfield (1980), Singleton and Wingender (1986), Lim (1989), Richardson and Smith (1993), Harvey and Siddique (1999, 2000), Aït-sahali and Brandt (2001), Chen et al. (2001), Cont (2001), Patton (2004), Hueng and McDonald (2005), Engle and Patton (2007), and He et al. (2008), among others. The assumed or declared significant asymmetries in return distributions have been discovered. It is particularly true of certain types of assets, described specifically by Low et al. (2012), such as hedge funds, emerging markets, or assets with credit risk, and is even more evident during financial crises. Hansen (1994)

conjectured that the reason why most applications have ignored higher-order features of the return distributions may be because only the mean and variance generate significant excitement. But this does not imply that higher-order features should be completely ignored. Since the presence of asymmetries violates the assumption of elliptically symmetric distributed financial returns in the traditional mean-variance analysis, it implies that the elliptical distributions cannot deal with returns in the presence of skewness, which is an empirical feature of returns. The CAPM's failure to correctly assess performance results from the fact that the CAPM-based beta does not capture skewness and other higher-order moments of the return distribution that investors value. For portfolio or asset returns that are highly skewed, the correct beta differs substantially from the CAPM beta. Investors typically do distinguish between upside and downside risks. Early studies such as Rubinstein (1973), Arditti (1967), Scott and Horvath (1980), and Kimball (1989) suggest that most investors have skewness preference. This preference for skewness implies that prices in equilibrium reflect more than mean and variance.

From a financial perspective, skewness is crucial since it may itself be considered as a measure of risk. Kim and White (2004) state that if investors prefer right-skewed portfolios, then a left-skewed portfolio should contain a "skew premium" which is appealing to investors. In the context of hedge funds, and options, Fung and Hsieh (2001) and Mitchell and Pulvino (2001) showed that the hedge funds returns have significant left-tail risk, Corrado and Su (1997) attributed the anomaly known as "volatility skew" in option pricing to the skewness and kurtosis of the return distribution. With respect to optimal portfolio allocation, Chunhachinda et al. (1997) showed that portfolio allocation can change

considerably if higher than second moments are considered in selection. The role played by skewness in risk management is described by Rosenberg and Schuermann (2006), Grigoletto and Lisi (2009), and Lee and McLachlan (2018). Measures such as Value-at-Risk (VaR) and expected shortfall (ES) are widely used in the finance industry to evaluate market risk exposure. Typically, the estimation of risk measures relies on the assumption that the returns follow some parametric distribution. Approaches used skewed distributions for this purpose may lead to more precise estimates of these risk measures. Skewness is important not only from a financial point of view but also from a statistical perspective. Some studies found significant conditional skewness, see Jensen and Lunde (2001), León et al. (2005), and Lanne and Pentti (2007), among others. Bai and Ng (2001, 2005) proposed a test to investigate the constant skewness of the marginal and conditional distribution of returns. In addition, some economic theories have been offered as an explanation of the mechanism generating the asymmetry, including leverage effects (Christie, 1982), the volatility feedback mechanism (Campbell and Hentschel, 1992), stochastic bubbles models (Blanchard and Watson, 1982) and investor heterogeneity (Hong and Stein, 2003).

In consequence, the ability to incorporate skewness is an important practical consideration. Kraus and Litzenberger (1976) presented a three-parameter CAPM which extended the Sharpe-Lintner standard CAPM framework, incorporating the effect of skewness in return distributions, making the assumption that investors have a preference for positive return skewness in their portfolios. Hwang and Satchell (1999) developed a higher-order moment CAPM, and study its performance on emerging markets data. Leland (1999) presents a simple risk measure to implement CAPM and correctly captures all elements of

risk, including skewness, kurtosis, and other characteristics that further describe the shape of the return distribution. Harvey and Siddique (2000) suggested an asset pricing model incorporates conditional skewness. Dittmar (2002) provided empirical evidence for the role of co-skewness in equity pricing. Harvey et al. (2004) proposed a method for optimal portfolio selection with higher moments using a Bayesian decision framework. Brandt et al. (2009) suggested incorporating higher-order moments for returns in the portfolio allocation process. Atilgan et al. (2020) derived a skewness-aware beta, and study whether it is consistent with returns observed in the market.

Motivated by the empirical and theoretical evidence that skewness and kurtosis in the distribution of financial time series and are time-varying, we want to examine the predictive ability of forecast models of financial returns to incorporate skewness and kurtosis. In this article, we proposed a test to investigate the Granger-causality in predictive models of skewness and kurtosis. In addition to financial motivation, the other reason for doing so is statically motivated. Since two competing forecast models for the conditional mean may be combined to produce a forecast that is better than the two forecasts (Bates and Granger, 1969). It is worth investigating how to test the predictive ability (Granger-causality) of a predictor for a forecast target of interest. Early studies such as Mincer and Zarnowitz (1969), Nelson (1972), and Granger and Newbold (1973) suggest an approach to test if two models are different in forecastability. The predictive ability of two forecast models has been compared in non-nested models, by Diebold and Mariano (1995), West (1996), White (2000) and Hansen and Lunde (2005). Under the null hypothesis that two non-nested models have no difference in predictive ability, the DM test statistics (Diebold and Mariano,

1995), which is the expected loss-differential of the two models, is asymptotically normal. The loss differential may vary with the loss function. For example, the DM test statistic is constructed by the loss-differential if the objective of interest is the one-step ahead mean squared prediction error (MSPE). A loss function is said to be strictly consistent for a functional, for example, the loss-differential, of the conditional distribution function if the optimal functional can be obtained as a unique solution from minimizing the expected loss. A functional is said to be elicitable if there exists a strictly consistent loss function, as defined by Gneiting (2011). Besides the well-known squared error loss, there are some commonly used strictly consistent loss functions such as loss function for quantile forecasts, binary forecasts, probability forecasts, and density forecasts. The comparison in non-nested models has been extended to nested models by works such as Clark and McCracken (2001, 2005), Clark and West (2006, 2007), Giacomini and White (2006). Out-of-sample forecast comparison is widely used in many fields because it is suggested to test for Granger-causality, determining whether one-time series is useful in forecasting another, proposed by Granger (1969). See Ashley et al. (1980). When there are two nested models to be compared, the DM test statistics lost their asymptotic normality, since it has a negative bias. The bias corrected MSPE loss-differential has zero means, and exactly constructs the encompassing (ENC) test statistic. From the perspective of the framework for forecast encompassing principle, Clark and McCracken (2001, 2005, 2009), discussed how the optimal forecast combination can be used to construct the test statistic for Granger-causality of a predictor.

In this article, we consider the Granger-causality test in two competing forecast models beyond the conditional mean framework. We focus on distributions with skewness

and kurtosis, which capture the distributional shape of financial time series. Time-varying skewness and kurtosis were introduced by Hansen (1994), who extended the ARCH framework by adopting a conditional generalized Student's t-distribution. The dynamics on skewness and kurtosis are introduced by modeling shape parameters. This skewed t-distribution is a generalization of the Student's t-distribution, with additional parameters to accommodate skewness, rendering it suitable for handling the asymmetric distributional shape of financial data. More models with dynamic conditional skewness and/or kurtosis have been studied in the literature, such as Harvey and Siddique (1999), Jondeau and Rockinger (2003), Brooks et al. (2005), and Yan (2005). Lee and McLachlan (2013) have considered the use of flexible asymmetric distributions for modeling portfolio returns and observed that they provided a closer fit to the real data set in their example. Others, extend the univariate skew normal distribution (Azzalini, 1985) with additional parameters to regulate skewness and other features. In this paper, we follow Hansen (1994) in the use of skewed t-distribution, generate the competing forecast models that can capture the empirically observed time-varying skewness and kurtosis and focus on testing forecast encompassing in these features.

This paper is organized as follows. In sections 2 and 3, we introduce the encompassing test for Granger-causality in skewness and kurtosis by using the log-score scoring rule. In section 4, we prove the ENC test statistic is asymptotically standard normal as the number of out-of-sample forecasts tends to infinity. In section 5, we conduct Monte Carlo simulations to show that the ENC test statistic has better size and power. In section 6, we present empirical analysis.

8.2 Forecast Encompassing Test for Granger Causality in Skewness

In this section, we develop the forecast encompassing test in the skewness regression. In order to test the out-of-sample predictive ability of x_t on y_{t+1}, we consider two skewness nested models:

$$\begin{aligned}
\text{Model 1: } y_{t+1} &= \sqrt{\frac{\nu-2}{\nu}}\varepsilon_{1,t+1}, \qquad & \varepsilon_{1,t+1} &\sim t(\nu, \eta_{1,t+1}), \\
\eta_{1,t+1} &= 2\left(\frac{e^{\theta_1}}{1+e^{\theta_1}} - 0.5\right), & & \\
\text{Model 2: } y_{t+1} &= \sqrt{\frac{\nu-2}{\nu}}\varepsilon_{2,t+1}, \qquad & \varepsilon^{(2)}_{t+1} &\sim t(\nu, \eta_{2,t+1}), \\
\eta_{2,t+1} &= 2\left(\frac{e^{\theta_2+f(x_t)}}{1+e^{\theta_2+f(x_t)}} - 0.5\right), & &
\end{aligned}$$

where $f(x_t)$ is linear function βx_t in this paper. Consider the skewed Student's t-distribution, where $-1 < \eta < 1$ is the skewness parameter, and $2 < \nu < \infty$ is the degree of freedom, introduced by Hansen (1994). It can capture the time-varying conditional skewness and kurtosis. We consider a combined functional of the forecast skewness parameter $\eta_{c,t+1}$ to be a weighted average of the forecast skewness parameters $\eta_{1,t+1}$ in Model 1 and $\eta_{2,t+1}$ in Model 2,

$$\eta_{c,t+1} = (1-\lambda)\,\eta_{1,t+1} + \lambda\eta_{2,t+1}.$$

To test the predictive ability of a predictor for a forecast target of interest, one method is to consider out-of-sample tests for Granger-Causality in a framework based on the forecast encompassing principle. The forecast encompassing is to test $\mathrm{H}_0 : \lambda = 0$ versus $\mathrm{H}_1 : \lambda \neq 0$. Under the null hypothesis, when $\lambda = 0$, x_t does not Grange-cause y_{t+1} in skewness. According to Hansen (1994), the density function of the skewed Student's t-distribution is

as follows, with omitted subscripts for simplicity,

$$g(y,\eta) = \begin{cases} bc\left(1+\frac{1}{\nu-2}\left(\frac{by+a}{1-\eta}\right)^2\right)^{-(\nu+1)/2} & y < -a/b, \\ bc\left(1+\frac{1}{\nu-2}\left(\frac{by+a}{1+\eta}\right)^2\right)^{-(\nu+1)/2} & y \geq -a/b, \end{cases}$$

where

$$a = 4\eta c\left(\frac{\nu-2}{\nu-1}\right) \qquad b^2 = 1+3\eta^2-a^2 \qquad c = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi(\nu-2)}\Gamma\left(\frac{\nu}{2}\right)}$$

In this paper, we consider the negative log-score function as a strictly consistent loss function for skewness models, which is

$$L = -\ln g\left(y_{t+1}, \eta_{t+1}\right).$$

We estimate the optimal forecast combination λ^* by minimizing the negative log-score function for the combined model,

$$\lambda^* = \arg\min_{\lambda} \mathbb{E}\left[-\ln g\left(Y_{t+1}, \eta_{c,t+1}\right)\right].$$

The optimal forecast combination λ^* can be used to construct the test statistic for Granger-causality of a predictor, and λ^* would satisfy the first-order condition,

$$\frac{\partial}{\partial\lambda}\mathbb{E}\left[\ln g\left(Y_{t+1}, \eta_{c,t+1}\right)\right] = 0.$$

If we may interchange the operations of differentiation and expectation, the first-order condition is

$$\mathbb{E}\frac{\partial}{\partial\lambda}\left[\ln g\left(Y_{t+1}, \eta_{c,t+1}\right)\right] = \mathbb{E}\left[\frac{\partial}{\partial\eta_{c,t+1}}\ln g\left(Y_{t+1}, \eta_{c,t+1}\right)\frac{\partial}{\partial\lambda}\eta_{c,t+1}\right] = 0$$

Taking the derivative of the log-score function with respect to $\eta_{c,t+1}$, we obtain the identification function $V_{c,t+1}$, with omitted subscripts for simplicity,

$$V = \frac{\partial L}{\partial \eta} = \begin{cases} -\frac{1}{b} * \frac{\partial b}{\partial \eta} + \frac{\nu+1}{\nu-2} * \frac{1}{d_1} * \left(\frac{by+a}{1-\eta}\right) * e_1 & y < -a/b, \\ -\frac{1}{b} * \frac{\partial b}{\partial \eta} + \frac{\nu+1}{\nu-2} * \frac{1}{d_2} * \left(\frac{by+a}{1-\eta}\right) * e_2 & y \geq -a/b, \end{cases}$$

where

$$\frac{\partial a}{\partial \eta} = 4c\left(\frac{\nu-2}{\nu-1}\right) \qquad \frac{\partial b}{\partial \eta} = \frac{1}{2b}\left(6\eta - 2a\frac{\partial a}{\partial \eta}\right)$$

$$d_1 = 1 + \frac{1}{\nu-2}\left(\frac{by+a}{1-\eta}\right)^2 \qquad d_2 = 1 + \frac{1}{\nu-2}\left(\frac{by+a}{1+\eta}\right)^2$$

$$e_1 = \frac{\frac{\partial b}{\partial \eta}y + \frac{\partial a}{\partial \eta}}{1-\eta} + \frac{by+a}{(1-\eta)^2} \qquad e_2 = \frac{\frac{\partial b}{\partial \eta}y + \frac{\partial a}{\partial \eta}}{1+\eta} - \frac{by+a}{(1+\eta)^2}$$

Taking the derivative of $\eta_{c,t+1}$ with respect to λ, we obtain $\eta_{2,t+1} - \eta_{1,t+1}$. Therefore, the first-order condition can be written as $\mathbb{E}\left[V_{c,t+1}\left(\eta_{2,t+1} - \eta_{1,t+1}\right)\right] = 0$. The out-of-sample test for Granger-Causality can be done in the approach of forecast encompassing test, which is equivalent to test $\mathrm{H}_0 : \mathbb{E}\left[V_{1,t+1}\left(\eta_{2,t+1} - \eta_{1,t+1}\right)\right] = 0$ versus $\mathrm{H}_1 : \mathbb{E}\left[V_{1,t+1}\left(\eta_{2,t+1} - \eta_{1,t+1}\right)\right] \neq 0$.

Considering our objective to examine out-of-sample predictive abilities, we partition the entire sample dataset $T+1$ into two subsets: the in-sample observations (denoted as R) and the out-of-sample observations (represented as P). We use the in-sample observations to estimate the parameters $\hat{\theta}_{1,t}, \hat{\theta_{2,t}}$, and $\hat{\beta}_t$. Then, we employ the out-of-sample observations to compute the forecast for the skewness parameters, denoted as $\hat{\eta}_{t+1}$. Thus, under the null hypothesis, we have the score for the out-of-sample test to be

$$\hat{C}_{R,P} \equiv \frac{1}{P}\sum_{t=R}^{T}\left[V_{1,t+1}\left(\hat{\eta}_{2,t+1} - \hat{\eta}_{1,t+1}\right)\right] \xrightarrow{p} E\left(\hat{C}_{R,P}\right) \text{ as } R, P \to \infty \text{ under } \mathbb{H}_0.$$

We compare the forecast encompassing test we proposed with the other two standard tests, the DM and CCS tests. The DM test is popular for evaluating the predictive ability of two forecasting models proposed by Diebold and Mariano (1995). The DM test statistic $\hat{D}_{R,P}$ is formed as

$$\hat{D}_{R,P} = \frac{1}{P}\sum_{t=R}^{T}\left[\ln\hat{g}\left(y_{t+1},\eta_{2,t+1}\right) - \ln\hat{g}\left(y_{t+1},\eta_{2,t+1}\right)\right].$$

The CCS test is the out-of-sample test of causality that resembles an encompassing test applied to forecasts from nested models proposed by Chao et al. (2001). The CCS test statistic $\hat{M}_{R,P}$ is formed as

$$\hat{M}_{R,P} = \frac{1}{P}\sum_{t=R}^{T} V_{1,t+1}x_t.$$

In order to compare the predictive accuracy of two nested skewness models, we standardize the three statistics. The standardized encompassing statistic is $\mathrm{ENC}_{R,P} \equiv \hat{Q}_{R,P}^{-0.5}\sqrt{P}\hat{C}_{R,P}$, where $Q_{R,P} = \mathrm{var}\left(\sqrt{P}\hat{C}_{R,P}\right)$ and $Q_{R,P} - \hat{Q}_{R,P} \xrightarrow{p} 0$. The standardized DM statistic is $\mathrm{DM}_{R,P} \equiv \hat{S}_{R,P}^{-0.5}\sqrt{P}\hat{D}_{R,P}$, where $S_{R,P} = \mathrm{var}\left(\sqrt{P}\hat{D}_{R,P}\right)$ and $S_{R,P} - \hat{S}_{R,P} \xrightarrow{p} 0$. The standardized CCS statistic is $\mathrm{CCS}_{R,P} \equiv \hat{W}_{R,P}^{-0.5}\sqrt{P}\hat{M}_{R,P}$, where $W_{R,P} = \mathrm{var}\left(\sqrt{P}\hat{M}_{R,P}\right)$ and $W_{R,P} - \hat{W}_{R,P} \xrightarrow{p} 0$.

8.3 Forecast Encompassing Test for Granger Causality in Kurtosis

In this section, we develop the forecast encompassing test in the kurtosis. We have two models. One model is without conditioning on x_t and the other model is with conditioning

on x_t. The two nested kurtosis models are

$$\text{Model 1: } y_{t+1} = \sqrt{\frac{\nu_{1,t+1}-2}{\nu_{1,t+1}}}\varepsilon_{1,t+1}, \qquad \varepsilon_{1,t+1} \sim t(\nu_{1,t+1}),$$

$$\nu_{1,t+1} = \theta_1,$$

$$\text{Model 2: } y_{t+1} = \sqrt{\frac{\nu_{2,t+1}-2}{\nu_{2,t+1}}}\varepsilon_{2,t+1}, \qquad \varepsilon_{2,t+1} \sim t(\nu_{2,t+1}),$$

$$\nu_{2,t+1} = \theta_2 + f(x_t),$$

where ν is the degree of freedom, and $f(x_t)$ is linear function βx_t. Consider a combined functional of the forecast degree of freedom $\nu_{c,t+1}$, a weighted average of the two forecast degree of freedom

$$\nu_{c,t+1} = (1-\lambda)\,\nu_{1,t+1} + \lambda\nu_{2,t+1}$$

In this paper, we consider the negative log-score function as a strictly consistent scoring function for kurtosis models, we have

$$L = -\ln g\left(y_{t+1}, \nu_{t+1}\right),$$

where the density of the student t-distribution is

$$g(y,\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{(\nu-2)\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{y^2}{\nu-2}\right)^{-\frac{\nu+1}{2}}.$$

We estimate the optimal weight λ^* by minimizing the log score scoring function for the combined model,

$$\lambda^* = \arg\min_{\lambda} \mathbb{E}\left[-\ln g\left(Y_{t+1}, \nu_{c,t+1}\right)\right].$$

In order to get the optimal forecast weight λ^*, we take the first-order condition of the expectation of the loss function with respect to the weight,

$$\mathbb{E}\frac{\partial}{\partial\lambda}\left[\ln g\left(Y_{t+1}, \nu_{c,t+1}\right)\right] = \mathbb{E}\left[\frac{\partial}{\partial\nu_{c,t+1}}\ln g\left(Y_{t+1}, \nu_{c,t+1}\right)\frac{\partial}{\partial\lambda}\nu_{c,t+1}\right] = 0$$

Taking the derivative of the loss function with respect to ν, we obtain the identification function $V_{c,t+1}$, with omitted subscripts for simplicity,

$$V = \frac{\partial L}{\partial \nu} = -\frac{1}{2}\left[\frac{\gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu+1}{2})} - \frac{1}{(\nu-2)\pi} - \frac{\gamma(\frac{\nu}{2})}{\Gamma(\frac{\nu}{2})} - \ln\left(1+\frac{y^2}{\nu-2}\right) + \frac{(\nu+1)y^2}{(\nu-2)(1+y^2)}\right],$$

where $\gamma(\nu)$ is the derivative of Gamma function $\Gamma(\nu)$ with respect to ν. Taking the derivative of $\nu_{c,t+1}$ with respect to λ, we obtain $\nu_{2,t+1} - \nu_{1,t+1}$. Therefore, the first-order condition can be written as $\mathbb{E}\left[V_{c,t+1}\left(\nu_{2,t+1} - \nu_{1,t+1}\right)\right] = 0$. The out-of-sample test for Granger-Causality can be done in the approach of forecast encompassing test, which is equivalent to test $\mathrm{H}_0 : \mathbb{E}\left[V_{1,t+1}\left(\nu_{2,t+1} - \nu_{1,t+1}\right)\right] = 0$ versus $\mathrm{H}_1 : \mathbb{E}\left[V_{1,t+1}\left(\nu_{2,t+1} - \nu_{1,t+1}\right)\right] \neq 0$.

Considering our objective to examine out-of-sample predictive abilities, we partition the entire sample dataset $T+1$ into two subsets: the in-sample observations (denoted as R) and the out-of-sample observations (represented as P). We use the in-sample observations to estimate the parameters $\hat{\theta}_{1,t}, \hat{\theta_{2,t}}$, and $\hat{\beta}_t$. Then, we employ the out-of-sample observations to compute the forecast for the skewness parameters, denoted as $\hat{\eta}_{t+1}$. Thus, under the null hypothesis, we have the score for the out-of-sample test to be

$$\hat{C}_{R,P} \equiv \frac{1}{P}\sum_{t=R}^{T}\left[V_{1,t+1}\left(\hat{\nu}_{2,t+1} - \hat{\nu}_{1,t+1}\right)\right] \xrightarrow{p} E(\hat{C}_{R,P}) \text{ as } R, P \to \infty \text{ under } \mathbb{H}_0$$

We will compare the ENC test with the DM test by Diebold and Mariano (1995) and the CCS test by Chao et al. (2001). The DM statistic is

$$\hat{D}_{R,P} = \frac{1}{P}\sum_{t=R}^{T}\left[\ln\hat{g}\left(y_{t+1}, \nu_{t+1}^{(2)}\right) - \ln\hat{g}\left(y_{t+1}, \nu_{t+1}^{(1)}\right)\right]$$

The CCS statistic is

$$\hat{M}_{R,P} = \frac{1}{P}\sum_{t=R}^{T} V_{t+1}^{(1)} x_t.$$

In order to compare the equal predictive accuracy of two nested kurtosis models, we standardize these three statistics. The DM statistic is $\mathrm{DM}_{R,P} \equiv \hat{S}_{R,P}^{-0.5}\sqrt{P}\hat{D}_{R,P}$, where $S_{R,P} = \mathrm{var}\left(\sqrt{P}\hat{D}_{R,P}\right)$ and $S_{R,P} - \hat{S}_{R,P} \xrightarrow{p} 0$. The encompassing statistic is $\mathrm{ENC}_{R,P} \equiv \hat{Q}_{R,P}^{-0.5}\sqrt{P}\hat{C}_{R,P}$, where $Q_{R,P} = \mathrm{var}\left(\sqrt{P}\hat{C}_{R,P}\right)$ and $Q_{R,P} - \hat{Q}_{R,P} \xrightarrow{p} 0$. The CCS statistic is $\mathrm{CCS}_{R,P} \equiv \hat{W}_{R,P}^{-0.5}\sqrt{P}\hat{M}_{R,P}$, where $W_{R,P} = \mathrm{var}\left(\sqrt{P}\hat{M}_{R,P}\right)$ and $W_{R,P} - \hat{W}_{R,P} \xrightarrow{p} 0$.

8.4 Monte Carlo Simulation

In our simulation, we compare the predictive ability of Model 1 with that from Model 2. We set the number of in-sample observations $R \in \{100, 200, 500\}$ and the number of out-of-sample forecasts $P \in \{200, 500, 1000\}$. In order to show the asymptotic distribution of encompassing test for the log-score scoring rule, we set x_t in Model 2 as an AR(1) stationary process, $x_t = \phi x_{t-1} + v_t$, where $v_t \overset{iid}{\sim} N(0, \sigma_v^2)$. Thus, x_t has zero mean and variance $\sigma_v^2/(1-\phi)$. We generate y_{t+1} following the skewed Student's t-distribution where η_{t+1} is the skewness of y_{t+1}. If $\eta_{t+1} > 0$, the mode of the density is to the left of zero and y_{t+1} is skewed to the right, and vice-versa when $\eta_{t+1} < 0$. As suggested by Hansen (1994), we generate the skewness parameter η_{t+1} from the modified logistic transformation, $\eta_{t+1} = \Lambda\left(x_t'\beta\right)$, where $\Lambda(z) = (1 - e^{-z})/(1 + e^{-z})$, to ensure $-1 < \eta_{t+1} < 1$ at all times.

We present the evaluation of the finite-sample size and power of tests of forecast encompassing.

The size of a test is the probability of rejecting H_0 given H_0 is true. To evaluate the size of tests, the coefficient β is set at 0. For a given degree of freedom ν, in Model 1 under H_0 the coefficient β set at 0 implies the skewness of y_{t+1} is zero. We use rolling windows to

obtain the one-step ahead log-likelihood estimation $\hat{\eta}_{1,t+1} = \hat{\theta}_1$, using the Matlab Toolbox by Patton (2002). Then, we obtain $\hat{\eta}_{2,t+1} = \Lambda\left(x_t'\hat{\theta}_1\right)$ to construct Model 2. In the size experiments, Model 1 forecasts encompass Model 2 forecasts.

The power of a test is the probability of reject $\mathbb{H}_0$ given $\mathbb{H}_0$ is false. To evaluate the power of tests, β is set at 0.1 and 0.2. For a given degree of freedom ν, in Model 1 under $\mathbb{H}_1$ the coefficient β set other than 0 implies y_{t+1} is skewed. We use rolling windows to obtain the one-step ahead log-likelihood estimation $\hat{\eta}_{1,t+1}$ and $\hat{\eta}_{2,t+1}$. In the power experiments, Model 1 forecasts do not encompass Model 2 forecasts.

We use rolling windows to estimate the forecast combination λ by minimizing the negative log-score function. Using those forecast variables, we then obtain $\mathrm{DM}_{R,P}$, $\mathrm{ENC}_{R,P}$, and $\mathrm{CCS}_{R,P}$ statistics. Repeating this procedure 2000 times, we obtain the asymptotic distribution of $\mathrm{DM}_{R,P}$, $\mathrm{ENC}_{R,P}$, and $\mathrm{CCS}_{R,P}$.

8.5 Empirical Analysis

There are some variables that have been shown to be predictors of financial asset returns in various studies. In this section, an empirical application is conducted to check if these variables Granger-cause the asset returns in skewness.

8.5.1 Data

To measure financial asset returns, we consider the return on the Center for Research in Security Prices (CRSP) value-weighted stock market portfolio. Industry portfolios sorting

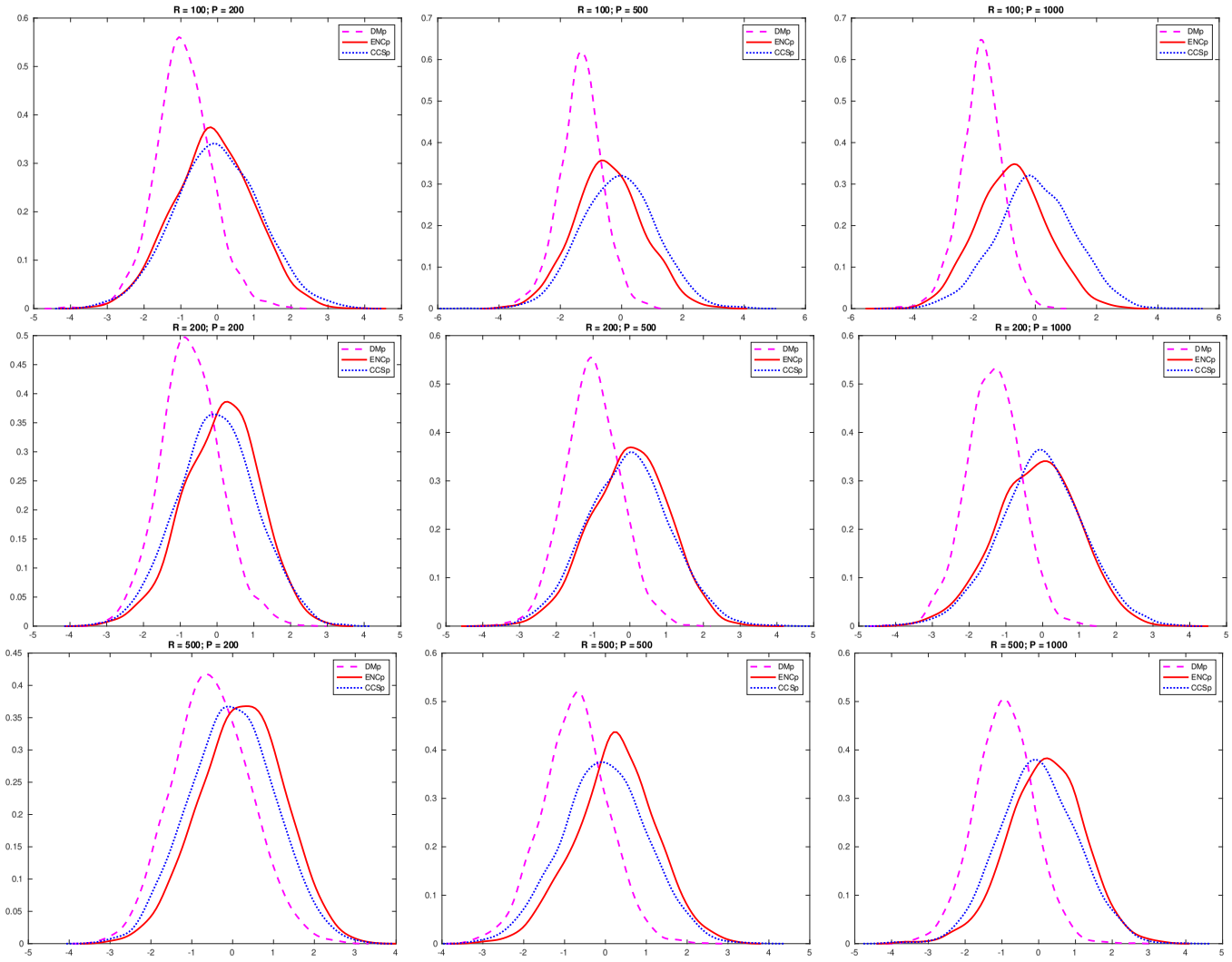

Figure 8.1: Simulation Results of Skewness: Size of Test

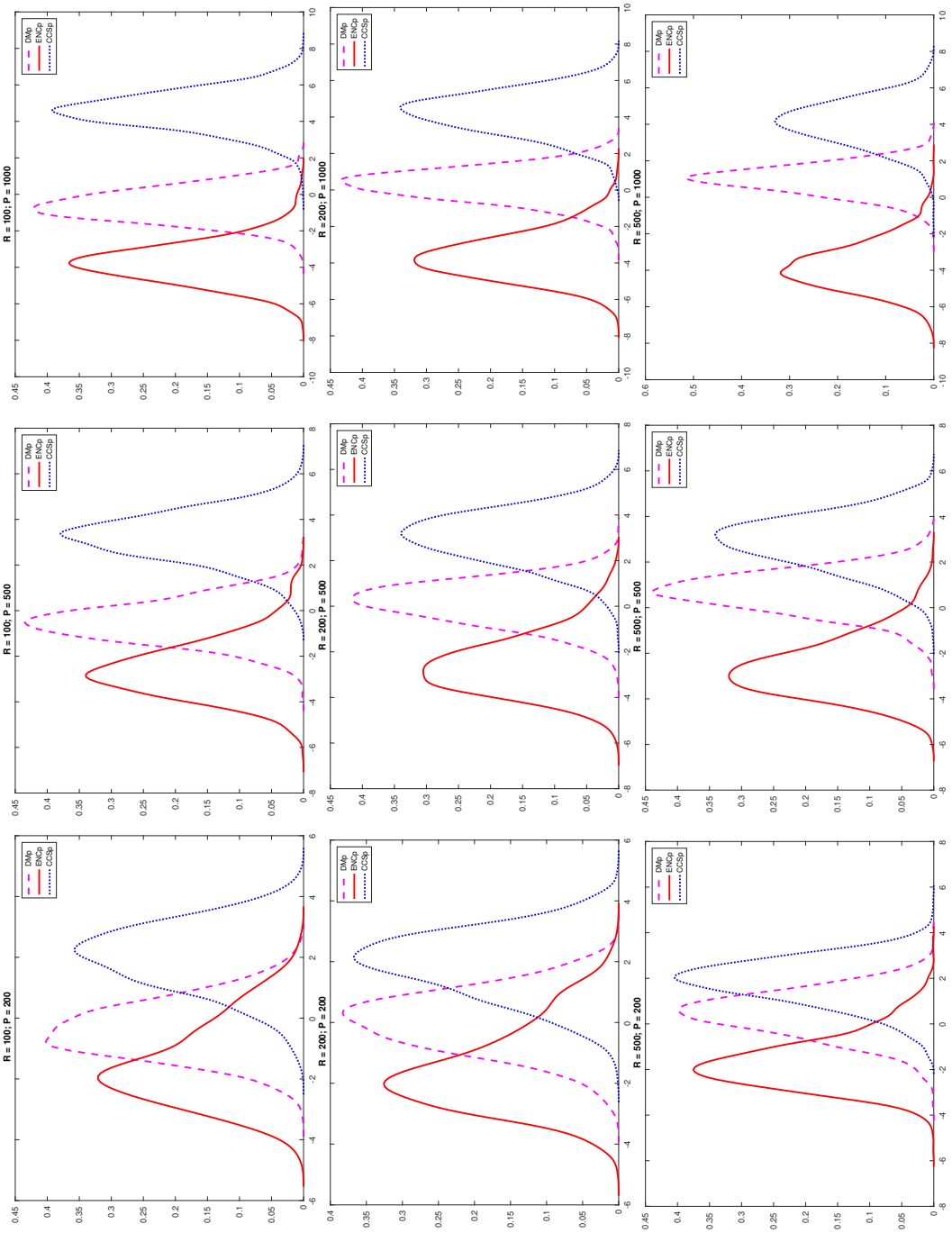

Figure 8.2: Simulation Results of Skewness: Power of Test

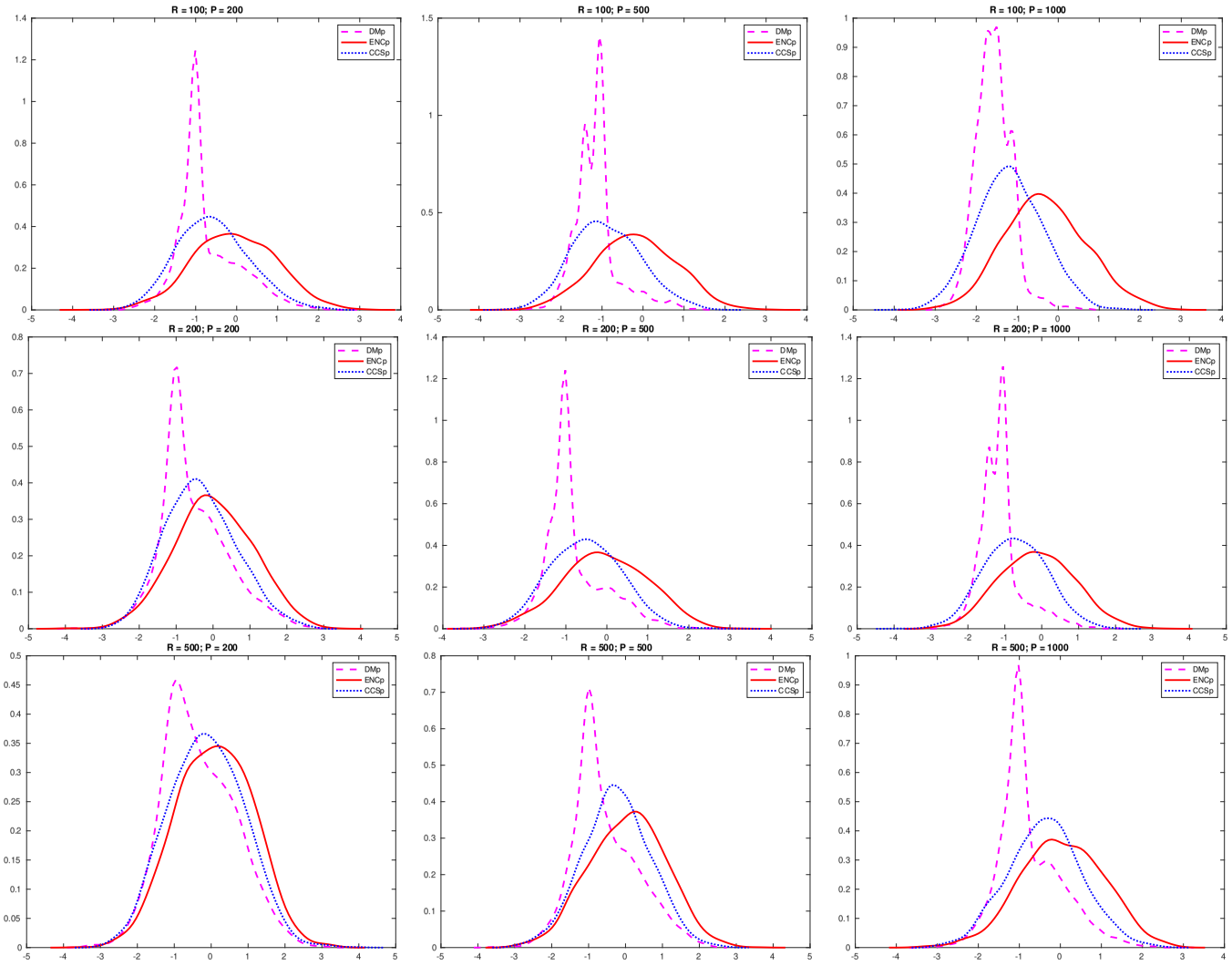

Figure 8.3: Simulation Results of Kurtosis: Size of Test

by some criteria is often motivated by previous empirical evidence. Firms are sorted into industry portfolios according to their two-digit Standard Industry Classification (SIC) code. Two-digit SIC grouping industry portfolios are studied in many studies, such as Moskowitz and Grinblatt (1999), and Dittmar (2002). We obtain the 10 industry-sorted portfolio return data for from Kenneth R. French's data library. The 10 industries include consumer non-durables (NoDur), consumer durables (Durbl), manufacturing (Manuf), energy (Energy), business equipment (HiTec), telephone and television transmission (Telcm), wholesale, retail and some services (Shops), healthcare, medical equipment and drugs (Hlth), utilities (Utils), and others (Other).

A set of variables that are commonly used to predict asset returns in the return predictability literature. See Fama and French (1988, 1989), Welch and Goyal (2008), Huang et al. (2015), and Jiang et al. (2019), Rapach et al. (2016). The predictors we focused on are dividend yield, book-to-market ratio, and T-bill return. A huge number of studies investigate the predictive power of the dividend yield. The dividends data we use are twelve-month moving sums of dividends paid on the S&P 500 index, from Amit Goyal's website. The dividend yield is the difference between the log of dividends and the log of lagged prices. Kothari and Shanken (1997), Pontiff and Schall (1998), explore the book-to-market ratio. The book-to-market ratio is the ratio of book value to market value for the Dow Jones Industrial Average, from Amit Goyal's website. Campbell (1987) shows that term premia in Treasury bill returns can predict stock returns. Fama and Schwert (1977), Ferson (1989), and Shanken (1990) examine the T-bill return. We obtain the yield on the three-month T-bill in excess of the yield on the one-month T-bill, from CRSP.

The monthly data cover the period from January 1927 to December 2021, totaling 1147 observations. By using Augmented Dickey-Fuller test statistic, dividend yield, book-to-market ratio, and T-bill return are stationary processes under the 5% level. We build two nested models to test if these stationary predictor variables Granger-cause the 10 industry portfolios returns in skewness. For Model 1, we find out the empirical skewness by P rolling window of the R in-sample observations of portfolio returns. For Model 2, we use OLS to estimate the future skewness of the portfolio returns by the previous predictor variables. We then obtain $\mathrm{DM}_{R,P}$, $\mathrm{ENC}_{R,P}$, and $\mathrm{CCS}_{R,P}$ statistics for $R : P = 2 : 3$, $1 : 1$, and $3 : 2$.

8.5.2 Results

Table 5 presents the test statistics for the Granger-causality test in skewness. The purpose of the test is to check if the three predictor variables Granger-cause the asset returns in skewness. Note the DM test is a one-side test, where the alternative hypothesis states that the expected loss of Model 1 is greater than Model 2. Therefore, we reject the null if the test statistics of DM are greater than the critical value (eg. 1.645 in 5% significance level). While the ENC test and CCS test are two-sided tests, and we reject the null if the test statistics are greater than 1.95 or less than -1.95 in the 5% significance level. Rejecting the null indicates there is not enough evidence to show the predictor variables not Granger-cause the asset returns in skewness.

From Table 5, first, we find that the DM statistics tend not to reject the null, compared to the other two statistics, which confirms that the DM statistic is biased. Secondly, comparing the results of different $R\,P$ ratios, we find that the less in-sample observations, the more rejections, especially for the ENC test of dividend yield and book-to-market ra-

tio. Thirdly, among the 10-industry sorted portfolios, more rejections appear in consumer nondurables (NoDur) and manufacturing (Manuf), indicating that the predictor variables tend to Granger-cause the returns of nondurables and manufacturing industry portfolios in skewness. While fewer rejections appear in consumer durables (Durbl) and healthcare, medical equipment, and drugs (Hlth), providing not enough evidence the predictor variables Granger-cause the returns of durables and health industry portfolios in skewness. Lastly, we observe that the three predictor variables have similar power in predicting the skewness of portfolio returns.

8.6 Conclusion

To test the predictive ability of a predictor for a forecast target of interest, we consider the Granger-causality test in two competing forecast models, based on the forecast encompassing principle. In order to handle the asymmetric and fat-tailed financial data, we focus on the forecast encompassing test in skewness and kurtosis. Using the conditional generalized Student's t-distribution, we construct the encompassing (ENC) test statistic. From the Monte Carlo simulation, we find that the ENC test statistic has better size and power, compared to the other two statistics. The empirical analysis shows that based on the ENC test, among the 10-industry sorted portfolios, the predictor variables such as dividend yield, book-to-market ratio, and T-bill return tend to have significant predictive power on the returns of nondurables and manufacturing industry portfolios, but not on durables and health industry portfolios in skewness.

Several extensions can be considered. First, if we consider the nested forecast models for both skewness and Kurtosis together, it would improve the model to fit the real distribution shape of financial data. Second, we focus on the conditional skewed Student's t-distribution. If we go for more generalized distributions, such as the generalized linear model (GLM), it would improve the range of the application further.

Table 8.1: Simulation Results: Size of Test

Panel A: ν									
	$P = 200$			$P = 500$			$P = 1000$		
Repeat=2000	DM_P	ENC_P	CCS_P	DM_P	ENC_P	CCS_P	DM_P	ENC_P	CCS_P
$R = 100$	0.006	0.054	0.052	0.000	0.047	0.083	0.000	0.054	0.166
$R = 200$	0.024	0.062	0.043	0.003	0.058	0.044	0.000	0.053	0.070
$R = 500$	0.024	0.047	0.043	0.011	0.052	0.038	0.006	0.049	0.044
Panel B: η									
	$P = 200$			$P = 500$			$P = 1000$		
Repeat=2000	DM_P	ENC_P	CCS_P	DM_P	ENC_P	CCS_P	DM_P	ENC_P	CCS_P
$R = 100$	0.002	0.069	0.090	0.000	0.102	0.098	0.000	0.158	0.121
$R = 200$	0.005	0.054	0.069	0.000	0.057	0.080	0.000	0.082	0.093
$R = 500$	0.011	0.057	0.055	0.004	0.050	0.061	0.002	0.059	0.062

This table shows the size of DM_P, ENC_P, and CCS_P test under 5% significance level, $\beta = 0$, $\phi = 0$.

Table 8.2: Simulation Results: Power of Test

Panel A: ν									
	$P = 200$			$P = 500$			$P = 1000$		
Repeat=2000	DM_P	ENC_P	CCS_P	DM_P	ENC_P	CCS_P	DM_P	ENC_P	CCS_P
$R = 100$	0.032	0.149	0.225	0.097	0.662	0.720	0.049	0.857	0.870
$R = 200$	0.051	0.377	0.341	0.240	0.718	0.729	0.335	0.924	0.932
$R = 500$	0.212	0.503	0.623	0.318	0.785	0.780	0.410	0.917	0.973
Panel B: η									
	$P = 200$			$P = 500$			$P = 1000$		
Repeat=2000	DM_P	ENC_P	CCS_P	DM_P	ENC_P	CCS_P	DM_P	ENC_P	CCS_P
$R = 100$	0.017	0.409	0.540	0.009	0.707	0.874	0.005	0.933	0.991
$R = 200$	0.047	0.429	0.507	0.039	0.699	0.818	0.058	0.901	0.967
$R = 500$	0.104	0.441	0.480	0.128	0.727	0.785	0.194	0.916	0.957

This table shows the size of DM_P, ENC_P, and CCS_P test under 5% significance level, $\phi = 0$, $\beta = 0.1$ for DGP 1 and $b = 0.2$ for DGP 2.

Table 8.3: Simulation Results: Model Averaging, $b = 0$, Panel A

Panel A: ν						
	$P = 200$		$P = 500$		$P = 1000$	
Repeat=2000	$\hat{\lambda}$	Loss	$\hat{\lambda}$	Loss	$\hat{\lambda}$	Loss
$R = 100$	0.2269[a]	1.3947[b]	0.2577	1.3937	0.2435	1.3952
		1.4323[c]		1.4319		1.4337
		1.3928[d]		1.3930		1.3948
$R = 200$	0.0636	1.3931	0.1884	1.3932	0.2282	1.3922
		1.4114		1.4106		1.4098
		1.3906		1.3922		1.3918
$R = 500$	-0.3288	1.3906	-0.0505	1.3909	0.0659	1.3909
		1.3964		1.3970		1.3966
		1.3878		1.3898		1.3903

[a]: The Monte Carlo average of the combination weight $\hat{\lambda}$.

[b]: The Monte Carlo average of the log score loss $\hat{S}_\tau^{(1)}$ from forecasting Model 1.

[c]: The Monte Carlo average of the log score loss $\hat{S}_\tau^{(2)}$ from forecasting Model 2.

[d]: The Monte Carlo average of the log score loss $\hat{S}_\tau^{(c)}$ from forecasting combined model.

Table 8.4: Simulation Results: Model Averaging, $b = 0$, Panel B

Panel B: η						
	$P = 200$		$P = 500$		$P = 1000$	
Repeat=2000	$\hat{\lambda}$	Loss	$\hat{\lambda}$	Loss	$\hat{\lambda}$	Loss
		1.3719		1.3716		1.3724
$R = 100$	-0.3013	1.3798	-0.1164	1.3796	-0.581	1.3804
		1.3695		1.3707		1.3719
		1.3717		1.3730		1.3722
$R = 200$	-0.6520	1.3748	-0.3042	1.3761	-0.1709	1.3753
		1.3693		1.3720		1.3717
		1.3709		1.3721		1.3719
$R = 500$	-1.3739	1.3719	-0.7439	1.3732	-0.4523	1.3729
		1.3681		1.3712		1.3714

[a]: The Monte Carlo average of the combination weight $\hat{\lambda}$.

[b]: The Monte Carlo average of the log score loss $\hat{S}_\tau^{(1)}$ from forecasting Model 1.

[c]: The Monte Carlo average of the log score loss $\hat{S}_\tau^{(2)}$ from forecasting Model 2.

[d]: The Monte Carlo average of the log score loss $\hat{S}_\tau^{(c)}$ from forecasting combined model.

Table 8.5: Test statistics of Granger-causality test in skewness - Dividend yield

Panel A: Dividend yield										
$R:P=2:3$										
	NoDur	Durbl	Manuf	Enrgy	HiTec	Telcm	Shops	Hlth	Utils	Other
DM_P	-2.81	-5.21	-7.98	2.64	-1.53	-3.60	-0.79	2.22	4.66	-4.41
ENC_P	-2.04	2.25	7.24	-3.39	-1.96	-2.32	1.39	0.14	-5.50	4.70
CCS_P	11.05	1.61	14.40	-1.52	4.45	3.28	2.68	1.55	3.73	14.64
$R:P=1:1$										
DM_P	-3.88	1.00	-7.66	0.33	-4.24	-6.65	-5.93	-0.58	-4.39	-7.24
ENC_P	-0.76	1.46	8.20	-3.06	0.16	0.22	0.86	-2.36	-1.47	6.60
CCS_P	12.63	2.08	15.63	-1.71	5.58	4.56	10.81	-1.51	2.27	15.33
$R:P=3:2$										
DM_P	-5.40	0.83	-2.58	-1.72	-1.22	-6.62	-4.10	-1.81	-6.03	-0.78
ENC_P	1.69	1.88	5.22	-2.36	-2.12	1.70	-0.19	-1.51	0.43	4.14
CCS_P	11.63	1.74	13.42	-0.93	1.03	5.12	10.42	-0.15	3.31	13.52

Table 8.6: Test statistics of Granger-causality test in skewness - Book-to-market ratio

Panel B: Book-to-market ratio										
$R:P=2:3$										
	NoDur	Durbl	Manuf	Enrgy	HiTec	Telcm	Shops	Hlth	Utils	Other
DM_P	4.45	1.20	9.15	2.18	6.08	4.44	2.52	3.26	4.57	7.60
ENC_P	-4.62	2.30	-6.21	-3.18	-3.68	-5.72	1.22	-0.23	-5.69	-7.07
CCS_P	-7.84	-0.87	-12.01	0.42	-2.86	-0.12	-1.98	-0.20	-2.82	-14.40
$R:P=1:1$										
DM_P	3.21	0.92	9.40	1.07	4.79	1.30	1.47	0.54	1.50	8.30
ENC_P	-5.82	1.44	0.21	-2.75	-1.68	-3.50	-3.69	-2.17	-3.34	-1.41
CCS_P	-8.55	-3.94	-11.45	2.03	-1.14	-1.68	-6.25	2.51	-0.80	-12.17
$R:P=3:2$										
DM_P	1.53	-1.45	8.38	0.29	5.68	-0.24	1.97	0.10	-1.29	8.66
ENC_P	-4.86	1.64	1.28	-2.73	-2.76	-3.00	-5.30	-1.57	-2.10	-0.50
CCS_P	-10.40	-2.96	-10.40	1.39	0.79	-4.46	-8.76	0.10	-3.23	-10.37

Table 8.7: Test statistics of Granger-causality test in skewness - Tbill return

Panel C: Tbill return										
$R:P=2:3$										
	NoDur	Durbl	Manuf	Enrgy	HiTec	Telcm	Shops	Hlth	Utils	Other
DM_P	-0.57	0.07	-1.26	1.99	2.66	-1.19	0.85	0.92	0.80	-1.04
ENC_P	-3.31	1.83	-3.64	-3.24	-3.23	-5.21	1.25	-0.52	-5.44	-3.43
CCS_P	-9.92	-1.06	-13.98	1.12	-3.27	-0.66	-2.27	0.02	-1.74	-15.73
$R:P=1:1$										
DM_P	6.98	-7.06	9.51	0.79	6.35	-4.88	6.62	1.12	-2.94	7.58
ENC_P	-9.14	-1.51	-0.75	-2.63	-2.39	-2.75	-8.13	-2.55	-2.73	-0.30
CCS_P	-10.65	-3.59	-13.78	1.53	-3.05	-2.51	-8.81	2.22	-1.15	-14.38
$R:P=3:2$										
DM_P	4.15	0.44	1.68	-4.00	-3.58	0.38	0.06	-0.45	0.34	-0.11
ENC_P	-8.38	-4.78	-9.90	0.34	0.01	-4.02	-7.87	-1.23	-3.74	-10.18
CCS_P	-10.29	-3.93	-10.23	1.40	0.95	-3.51	-8.64	0.72	-2.06	-10.37

Chapter 9

Conclusions

In conclusion, this thesis makes a significant contribution to the field of econometric forecasting and predictor selection by developing a set of innovative methods based on the forecast-encompassing principle and higher-order elicitability concept. These methods offer robust improvements to predictive regression models, specifically in the context of various financial and macroeconomic scenarios. These advancements range from enhancing the forecasting of U.S. output growth, inflation, and recession probability to exploring expectiles and their applications in financial risk measurement. They also provide new frameworks for Granger Causality tests in the predictive regression for the conditional mode, skewness and kurtosis in stock return distributions, and Expected Shortfalls. These diverse applications of the developed methods demonstrate their far-reaching implications in various fields and offer a considerable enhancement to the econometric toolbox for financial and macroeconomic forecasting and risk management.

Beyond their immediate applications, the concepts and methods proposed in this thesis have broader implications for the future of econometric forecasting. They lay the groundwork for more nuanced predictor selection and risk assessment in increasingly complex economic scenarios. The integration of the Granger Causality Test across various predictive functionals of the conditional distribution and the accommodation of linear models, nonparametric additive models, and generalized linear models for the conditional mean and beyond, paves the way for more sophisticated modeling and predictive techniques. By offering a means to compare different competing forecasts and rank them in terms of their expected score, these models make significant strides towards improving the accuracy of economic predictions. The potential of these innovative methodologies and the continued expansion of this field present exciting opportunities for further research and applications.

References

Adrian, T., N. Boyarchenko, and D. Giannone (2019). Vulnerable growth. *American Economic Review 109*(4), 1263–89.

Aït-sahali, Y. and M. W. Brandt (2001). Variable selection for portfolio choice. *The Journal of Finance 56*(4), 1297–1351.

Arditti, F. D. (1967). Risk and the required return on equity. *The Journal of Finance 22*(1), 19–36.

Artzner, P., F. Delbaen, J.-M. Eber, and D. Heath (1999). Coherent measures of risk. *Mathematical Finance 9*(3), 203–228.

Ashley, R., C. Granger, and R. Schmalensee (1980). Advertising and aggregate consumption: An analysis of causality. *Econometrica 48*(5), 1149–1167.

Atilgan, Y., K. O. Demirtas, A. D. Gunaydin, and I. Kirli (2020). Average skewness in global equity markets. *International Review of Finance*.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, 171–178.

Bai, J. and S. Ng (2001). A consistent test for conditional symmetry in time series models. *Journal of Econometrics 103*(1-2), 225–258.

Bai, J. and S. Ng (2005). Tests for skewness, kurtosis, and normality for time series data. *Journal of Business & Economic Statistics 23*(1), 49–60.

Bates, J. M. and C. W. Granger (1969). The combination of forecasts. *Journal of the Operational Research Society 20*(4), 451–468.

Bickel, P., Y. Ritov, and A. Tsybakov (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of statistics 37*(4), 1705–1732.

Blanchard, O. J. and M. W. Watson (1982). Bubbles, rational expectations and financial markets.

Brandt, M. W., P. Santa-Clara, and R. Valkanov (2009). Parametric portfolio policies: Exploiting characteristics in the cross-section of equity returns. *The Review of Financial Studies 22*(9), 3411–3447.

Brehmer, J. (2017). Elicitability and its application in risk management. *arXiv preprint arXiv:1707.09604*.

Brooks, C., S. P. Burke, S. Heravi, and G. Persand (2005). Autoregressive conditional kurtosis. *Journal of Financial Econometrics 3*(3), 399–421.

Campbell, J. Y. (1987). Stock returns and the term structure. *Journal of Financial Economics 18*(2), 373–399.

Campbell, J. Y. and L. Hentschel (1992). No news is good news: An asymmetric model of changing volatility in stock returns. *Journal of Financial Economics 31*(3), 281–318.

Candes, E. and T. Tao (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics 35*(6), 2313–2351.

Chamberlain, G. (1982). The general equivalence of granger and sims causality. *Econometrica: Journal of the Econometric Society*, 569–581.

Chao, J., V. Corradi, and N. Swanson (2001). Out-of-sample tests for Granger causality. *Macroeconomic Dynamics 5*(4), 598–620.

Chavleishvili, S., R. F. Engle, S. Fahr, M. Kremer, S. Manganelli, and B. Schwaab (2021). The risk management approach to macro-prudential policy.

Chen, J., H. Hong, and J. C. Stein (2001). Forecasting crashes: Trading volume, past returns, and conditional skewness in stock prices. *Journal of Financial Economics 61*(3), 345–381.

Christie, A. A. (1982). The stochastic behavior of common stock variances: Value, leverage and interest rate effects. *Journal of Financial Economics 10*(4), 407–432.

Chuang, C.-C., C.-M. Kuan, and H.-Y. Lin (2009). Causality in quantiles and dynamic stock return–volume relations. *Journal of Banking & Finance 33*(7), 1351–1360.

Chudik, A., G. Kapetanios, and H. Pesaran (2018). A one covariate at a time, multiple testing approach to variable selection in high-dimensional linear regression models. *Econometrica 86*(4), 1479–1512.

Chunhachinda, P., K. Dandapani, S. Hamid, and A. J. Prakash (1997). Portfolio selection and skewness: Evidence from international stock markets. *Journal of Banking & Finance 21*(2), 143–167.

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

Clark, T. and K. West (2006). Using out-of-sample mean squared prediction errors to test the martingale difference hypothesis. *Journal of Econometrics 135*(1-2), 155–186.

Clark, T. E. and M. W. McCracken (2005). Evaluating direct multistep forecasts. *Econometric Reviews 24*(4), 369–404.

Clark, T. E. and M. W. McCracken (2009). Combining forecasts from nested models. *Oxford Bulletin of Economics and Statistics 71*(3), 303–329.

Clark, T. E. and K. D. West (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics 138*(1), 291–311.

Cont, R. (2001). Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance 1*(2), 223.

Corradi, V. and N. R. Swanson (2002). A consistent test for nonlinear out of sample predictive accuracy. *Journal of Econometrics 110*(2), 353–381.

Corrado, C. J. and T. Su (1997). Implied volatility skews and stock return skewness and kurtosis implied by stock option prices. *The European Journal of Finance 3*(1), 73–85.

Davidson, J. (1994). *Stochastic limit theory: An introduction for econometricians*. OUP Oxford.

Diebold, F. and R. Mariano (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics 20*(1), 134–144.

Diebold, F. X. (1989). Forecast combination and encompassing: Reconciling two divergent literatures. *International Journal of Forecasting 5*(4), 589–592.

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

Dimitriadis, T., A. J. Patton, and P. Schmidt (2019). Testing forecast rationality for measures of central tendency. *arXiv:1910.12545*.

Dittmar, R. F. (2002). Nonlinear pricing kernels, kurtosis preference, and evidence from the cross section of equity returns. *The Journal of Finance 57*(1), 369–403.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *The Annals of Statistics 32*(2), 407–499.

Elliott, G., A. Timmermann, and I. Komunjer (2005). Estimation and testing of forecast rationality under flexible loss. *The Review of Economic Studies 72*(4), 1107–1125.

Engle, R. F., E. Ghysels, and B. Sohn (2013). Stock market volatility and macroeconomic fundamentals. *Review of Economics and Statistics 95*(3), 776–797.

Engle, R. F. and A. J. Patton (2007). What good is a volatility model? In *Forecasting volatility in the financial markets*, pp. 47–63. Elsevier.

Fama, E. F. and K. R. French (1988). Dividend yields and expected stock returns. *Journal of Financial Economics 22*(1), 3–25.

Fama, E. F. and K. R. French (1989). Business conditions and expected returns on stocks and bonds. *Journal of Financial Economics 25*(1), 23–49.

Fama, E. F. and G. W. Schwert (1977). Asset returns and inflation. *Journal of Financial Economics 5*(2), 115–146.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*(456), 1348–1360.

Fang, T., T.-H. Lee, and Z. Su (2020). Predicting the long-term stock market volatility: A garch-midas model with variable selection. *Journal of Empirical Finance 58*, 36–49.

Feng, X., X. He, and J. Hu (2011). Wild bootstrap for quantile regression. *Biometrika 98*(4), 995–999.

Ferson, W. E. (1989). Changes in expected security returns, risk, and the level of interest rates. *The Journal of Finance 44*(5), 1191–1217.

Fissler, T., J. F. Ziegel, et al. (2016). Higher order elicitability and osband's principle. *The Annals of Statistics 44*(4), 1680–1707.

Friend, I. and R. Westerfield (1980). Co-skewness and capital asset pricing. *The Journal of Finance 35*(4), 897–913.

Fung, W. and D. A. Hsieh (2001). The risk in hedge fund strategies: Theory and evidence from trend followers. *The Review of Financial Studies 14*(2), 313–341.

Ge, Y. and T. H. Lee (2020). Comparing predictive accuracy of nested quantile models using encompassing test.

Gel'fand, I. M. and N. Y. Vilenkin (2014). *Generalized functions: Applications of harmonic analysis*, Volume 4. Academic press.

Giacomini, R. and H. White (2006). Tests of conditional predictive ability. *Econometrica 74*(6), 1545–1578.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association 106*(494), 746–762.

Gneiting, T. and A. E. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association 102*(477), 359–378.

Gneiting, T. and R. Ranjan (2011). Comparing density forecasts using threshold-and quantile-weighted scoring rules. *Journal of Business & Economic Statistics 29*(3), 411–422.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica 37*(3), 424–438.

Granger, C. W. (1980). Testing for causality: A personal viewpoint. *Journal of Economic Dynamics and control 2*, 329–352.

Granger, C. W. and P. Newbold (1973). Some comments on the evaluation of economic forecasts. *Applied Economics 5*(1), 35–47.

Grigoletto, M. and F. Lisi (2009). Looking for skewness in financial time series. *The Econometrics Journal 12*(2), 310–323.

Hansen, B. (1994). Autoregressive conditional density estimation. *International Economic Review*, 705–730.

Hansen, B. E. (1992). Convergence to stochastic integrals for dependent heterogeneous processes. *Econometric Theory*, 489–500.

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

Hansen, P. R. and A. Timmermann (2012). Choice of sample split in out-of-sample forecast evaluation.

Harvey, A. G., E. Watkins, and W. Mansell (2004). *Cognitive behavioural processes across psychological disorders: A transdiagnostic approach to research and treatment*. Oxford University Press, USA.

Harvey, C. R. and A. Siddique (1999). Autoregressive conditional skewness. *Journal of Financial and Quantitative Analysis 34*(4), 465–487.

Harvey, C. R. and A. Siddique (2000). Conditional skewness in asset pricing tests. *The Journal of Finance 55*(3), 1263–1295.

Harvey, D., S. Leybourne, and P. Newbold (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting 13*(2), 281–291.

Harvey, D., S. Leybourne, and P. Newbold (1998). Tests for forecast encompassing. *Journal of Business & Economic Statistics 16*(2), 254–259.

He, C., A. Silvennoinen, and T. Teräsvirta (2008). Parameterizing unconditional skewness in models for financial time series. *Journal of Financial Econometrics 6*(2), 208–230.

Hong, H. and J. C. Stein (2003). Differences of opinion, short-sales constraints, and market crashes. *The Review of Financial Studies 16*(2), 487–525.

Hong, Y., Y. Sun, and S. Wang (2020). Selection of an optimal rolling window in time-varying predictive regression. *Cornell University Working Paper*.

Huang, D., F. Jiang, J. Tu, and G. Zhou (2015). Investor sentiment aligned: A powerful predictor of stock returns. *The Review of Financial Studies 28*(3), 791–837.

Hueng, C. J. and J. B. McDonald (2005). Forecasting asymmetries in aggregate stock market returns: Evidence from conditional skewness. *Journal of Empirical Finance 12*(5), 666–685.

Hwang, S. and S. E. Satchell (1999). Modelling emerging market risk premia using higher moments. *International Journal of Finance & Economics 4*(4), 271–296.

Inoue, A., L. Jin, and B. Rossi (2017). Rolling window selection for out-of-sample forecasting with time-varying parameters. *Journal of Econometrics 196*(1), 55–67.

Inoue, A. and L. Kilian (2005). In-sample or out-of-sample tests of predictability: Which one should we use? *Econometric Reviews 23*(4), 371–402.

Jensen, M. B. and A. Lunde (2001). The nig-s&arch model: a fat-tailed, stochastic, and autoregressive conditional heteroskedastic volatility model. *The Econometrics Journal 4*(2), 319–342.

Jiang, F., J. Lee, X. Martin, and G. Zhou (2019). Manager sentiment and stock returns. *Journal of Financial Economics 132*(1), 126–149.

Jondeau, E. and M. Rockinger (2003). Conditional volatility, skewness, and kurtosis: existence, persistence, and comovements. *Journal of Economic Dynamics and Control 27*(10), 1699–1737.

Kemp, G. C., P. M. Parente, and J. S. Silva (2020). Dynamic vector mode regression. *Journal of Business & Economic Statistics 38*(3), 647–661.

Kemp, G. C. and J. S. Silva (2012). Regression towards the mode. *Journal of Econometrics 170*(1), 92–101.

Kim, T.-H. and H. White (2004). On more robust estimation of skewness and kurtosis. *Finance Research Letters 1*(1), 56–73.

Kimball, M. S. (1989). Precautionary saving in the small and in the large.

Koenker, R. and G. Bassett Jr (1978). Regression quantiles. *Econometrica*, 33–50.

Kothari, S. P. and J. Shanken (1997). Book-to-market, dividend yield, and expected market returns: A time-series analysis. *Journal of Financial Economics 44*(2), 169–203.

Kraus, A. and R. H. Litzenberger (1976). Skewness preference and the valuation of risk assets. *The Journal of Finance 31*(4), 1085–1100.

Kuan, C.-M., J.-H. Yeh, and Y.-C. Hsu (2009). Assessing value at risk with care, the conditional autoregressive expectile models. *Journal of Econometrics 150*(2), 261–270.

Lambert, N. S., D. M. Pennock, and Y. Shoham (2008). Eliciting properties of probability distributions. In *Proceedings of the 9th ACM Conference on Electronic Commerce*, pp. 129–138.

Lanne, M. and S. Pentti (2007). Modeling conditional skewness in stock returns. *The European Journal of Finance 13*(8), 691–704.

Lee, M.-J. (1989). Mode regression. *Journal of Econometrics 42*(3), 337–349.

Lee, M.-J. (1993). Quadratic mode regression. *Journal of Econometrics 57*(1-3), 1–19.

Lee, S. X. and G. J. McLachlan (2013). Modelling asset return using multivariate asymmetric mixture models with applications to estimation of value-at-risk. In *20Th International Congress On Modelling and Simulation (Modsim2013)*, pp. 1128–1234. Modelling and Simulation Society of Australia and New Zealand.

Lee, S. X. and G. J. McLachlan (2018). Risk measures based on multivariate skew normal and skew t-mixture models. *Asymmetric Dependence in Finance: diversification, correlation and portfolio management in market downturns*, 152–168.

Lee, T.-H. and W. Yang (2012). Money-income granger-causality in quantiles. *Advances in Econometrics 30*, 385–409.

Leland, H. E. (1999). Beyond mean–variance: Performance measurement in a nonsymmetrical world (corrected). *Financial Analysts Journal 55*(1), 27–36.

León, Á., G. Rubio, and G. Serna (2005). Autoregresive conditional volatility, skewness and kurtosis. *The Quarterly Review of Economics and Finance 45*(4-5), 599–618.

Lim, K.-G. (1989). A new test of the three-moment capital asset pricing model. *Journal of Financial and Quantitative Analysis 24*(2), 205–216.

Lintner, J. (1965). Security prices, risk, and maximal gains from diversification. *The Journal of Finance 20*(4), 587–615.

Low, C., D. Pachamanova, and M. Sim (2012). Skewness-aware asset allocation: A new theoretical framework and empirical evidence. *Mathematical Finance: An International Journal of Mathematics, Statistics and Financial Economics 22*(2), 379–410.

Markowitz, H. (1952). The utility of wealth. *Journal of Political Economy 60*(2), 151–158.

McCracken, M. W. (2000). Robust out-of-sample inference. *Journal of Econometrics 99*(2), 195–223.

McCracken, M. W. (2007). Asymptotics for out of sample tests of granger causality. *Journal of Econometrics 140*(2), 719–752.

Mincer, J. A. and V. Zarnowitz (1969). The evaluation of economic forecasts. In *Economic forecasts and expectations: Analysis of forecasting behavior and performance*, pp. 3–46. NBER.

Mitchell, M. and T. Pulvino (2001). Characteristics of risk and return in risk arbitrage. *the Journal of Finance 56*(6), 2135–2175.

Moskowitz, T. J. and M. Grinblatt (1999). Do industries explain momentum? *The Journal of Finance 54*(4), 1249–1290.

Nelson, C. R. (1972). The prediction performance of the frb-mit-penn model of the us economy. *The American Economic Review 62*(5), 902–917.

Newey, W. K. and J. L. Powell (1987). Asymmetric least squares estimation and testing. *Econometrica*, 819–847.

Nolde, N., J. F. Ziegel, et al. (2017). Elicitability and backtesting: Perspectives for banking regulation. *The Annals of Applied Statistics 11*(4), 1833–1874.

Osband, K. (1985). *Providing incentives for better cost forecasting*. Ph. D. thesis, University of California, Berkeley.

Patton, A. J. (2004). On the out-of-sample importance of skewness and asymmetric dependence for asset allocation. *Journal of Financial Econometrics 2*(1), 130–168.

Patton, A. J. (2011a). Data-based ranking of realised volatility estimators. *Journal of Econometrics 161*(2), 284–303.

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

Patton, A. J. and K. Sheppard (2009). Evaluating volatility and correlation forecasts. In *Handbook of financial time series*, pp. 801–838. Springer.

Patton, A. J., J. F. Ziegel, and R. Chen (2019). Dynamic semiparametric models for expected shortfall (and value-at-risk). *Journal of Econometrics 211*(2), 388–413.

Peng, J. and Y. Yang (2021). On improvability of model selection by model averaging. *Journal of Econometrics*.

Pesaran, H. and R. Smith (2014). Signs of impact effects in time series regression models. *Economics Letters 122*(2), 150–153.

Phillips, P. C. B. (1991). A shortcut to lad estimator asymptotics. *Econometric Theory*, 450–463.

Pontiff, J. and L. D. Schall (1998). Book-to-market ratios as predictors of market returns. *Journal of Financial Economics 49*(2), 141–160.

Rapach, D. E., M. C. Ringgenberg, and G. Zhou (2016). Short interest and aggregate stock returns. *Journal of Financial Economics 121*(1), 46–65.

Richardson, M. and T. Smith (1993). A test for multivariate normality in stock returns. *Journal of Business*, 295–321.

Rockafellar, R. T., S. Uryasev, et al. (2000). Optimization of conditional value-at-risk. *Journal of risk 2*, 21–42.

Rosenberg, J. V. and T. Schuermann (2006). A general approach to integrated risk management with skewed, fat-tailed risks. *Journal of Financial Economics 79*(3), 569–614.

Rubinstein, M. E. (1973). A comparative statics analysis of risk premiums. *The Journal of Business 46*(4), 605–615.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association 66*(336), 783–801.

Scott, R. C. and P. A. Horvath (1980). On the direction of preference for moments of higher order than the variance. *The Journal of Finance 35*(4), 915–919.

Shanken, J. (1990). Intertemporal asset pricing: An empirical investigation. *Journal of Econometrics 45*(1-2), 99–120.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance 19*(3), 425–442.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, Volume 26. CRC press.

Sims, C. A. et al. (1986). Are forecasting models usable for policy analysis? *Quarterly Review 10*(Win), 2–16.

Singleton, J. C. and J. Wingender (1986). Skewness persistence in common stock returns. *Journal of Financial and Quantitative Analysis 21*(3), 335–341.

Stock, J. and M. Watson (2012). Disentangling the channels of the 2007-2009 recession. Technical report, National Bureau of Economic Research.

Su, L., T. Yang, Z. Tao, Yonghui, and Q. Zhou (2022). A one-covariate-at-a-time method for nonparametric additive models. *arXiv preprint arXiv:2204.12023*.

Taylor, J. W. (2008). Estimating value at risk and expected shortfall using expectiles. *Journal of Financial Econometrics 6*(2), 231–252.

Taylor, J. W. (2019). Forecasting value at risk and expected shortfall using a semiparametric approach based on the asymmetric laplace distribution. *Journal of Business & Economic Statistics 37*(1), 121–133.

Thomson, W. (1979). Eliciting production possibilities from a well-informed manager. *Journal of Economic Theory 20*, 360–380.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 58*(1), 267–288.

Welch, I. and A. Goyal (2008). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies 21*(4), 1455–1508.

West, K. D. (1996). Asymptotic inference about predictive ability. *Econometrica*, 1067–1084.

White, H. (2000). A reality check for data snooping. *Econometrica 68*(5), 1097–1126.

Xie, S., Y. Zhou, and A. T. Wan (2014). A varying-coefficient expectile model for estimating value at risk. *Journal of Business & Economic Statistics 32*(4), 576–592.

Yan, J. (2005). Asymmetry, fat-tail, and autoregressive conditional density in financial return data with systems of frequency curves. *University of Iowa: Department of Statistics and Actuarial Science*.

Yao, Q. and H. Tong (1996). Asymmetric least squares regression estimation: a nonparametric approach. *Journal of Nonparametric Statistics 6*(2-3), 273–292.

Zhang, C. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics 38*(2), 894–942.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association 101*(476), 1418–1429.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*(2), 301–320.