Elicitability and Encompassing for Volatility Forecasts by Bregman Functions

Tae-Hwy Lee[*], Ekaterina Seregina[†], Yaojue Xu[‡]

September 30, 2023

Abstract

In this paper, we construct a class of strictly consistent scoring functions based on the Bregman divergence measure, which jointly elicit the mean and variance. We use the scoring functions to develop a novel out-of-sample forecast encompassing test in volatility predictive models. We show the encompassing test is asymptotically normal. Simulation results demonstrate the merits of the proposed Bregman scoring functions and the forecast encompassing test. The forecast encompassing test exhibits a proper size and good power in finite samples. In an empirical application, we investigate the predictive ability of macroeconomic and financial variables in forecasting the equity premium volatility.

Keywords: strictly consistent scoring function, elicitability, Bregman divergence, Granger-causality, encompassing, model averaging, equity premium.

JEL Classification: C53, E37, E27

[*]Department of Economics, University of California, Riverside, CA 92521. E-mail: taelee@ucr.edu
[†]Department of Economics, Colby College, Waterville, ME 04901. E-mail: eseregin@colby.edu
[‡]Department of Economics, Colby College, Waterville, ME 04901. E-mail: yaojuexu@colby.edu

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) and Patton (2011), use a loss function that assumes that the conditional mean is zero and uses 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 & Mariano, 1995). Many papers focus on out-of-sample tests for equal predictive accuracy and encompassing (Diebold & Mariano, 1995; Clark & McCracken, 2001; Clark & West, 2006, 2007). 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. In this paper, we extend Clark and McCracken (2001) and Clark and West (2006, 2007) from a mean regression to volatility regression. We also find out DM statistic has a bias for two nested mode models. We propose an out-of-sample test based on the forecast encompassing (ENC) principle. We define the test statistic to be ENC. We will show that the ENC statistic has a zero mean, correct the size, and increase the power.

Our contributions in this paper are twofold. 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 and Ziegel (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, 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. 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 statistics to compare the nested volatility forecast models has good size and power in finite samples.

The structure of this paper unfolds as follows: Section 2 revisits the foundational aspects of elicitability, including its definitions, related lemmas, and theorems. The Bregman divergence 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, whilst establishing that the ENC for volatility converges towards asymptotic standard normality. Monte Carlo simulations are implemented in Section 5, showing the encompassing statistic's ability to rectify bias within DM statistics and improve statistical power. In section 6, we present empirical analysis. Section 7 is the conclusion.

2 Elicitability

We denote an observation domain $\mathrm{O} \subseteq \mathbb{R}$, 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 and Ziegel (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

$$\mathbb{E}_F S(\Gamma(F), F) = \mathbb{E}_F S(\gamma, F), \tag{1}$$

for all $F \in \mathcal{F}$ and for all $\gamma \in \mathrm{A}$. In this paper, F is defined to be the conditional distribution of y_{t+1} given x_t. The F is defined as follows,

$$F \equiv F_{t+1}(y) = \mathrm{Pr}\left(y_{t+1} < y | x_t\right).$$

Definition 2: (Gneiting (2011), Fissler and Ziegel (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 and Ziegel (2016)) An identification function is an $\mathcal{F}$ integrable 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 $\gamma = \Gamma(F)$.

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

$$\frac{\mathrm{d}}{\mathrm{d}\gamma} S(\gamma, y) = h(\gamma) V(\gamma, y). \tag{2}$$

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 such that the correct forecast of the functional is the unique minimizer of the expected score (Fissler & Ziegel, 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.

3 Bregman Divergence Scoring Functions

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 and Ziegel (2016) propose a strictly consistent scoring function for joint Value-at-Risk (VaR) and Expected Shortfall (ES).

Definition 4: (Fissler and Ziegel (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 $A = A_1 \times \cdots \times A_l \subseteq \mathbb{R}^k$.

Remark: 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. Let $\Gamma_m : \mathcal{F} \to A_m \subseteq \mathbb{R}$ be a elicitable functional, $m \in \{1, 2\}$. Then the functional $\Gamma = (\Gamma_1, \Gamma_2) : \mathcal{F} \to \mathrm{A}$ is 2 elicitable where $A = A_1 \times A_2 \subseteq \mathbb{R}^2$.

Definition 5: (Fissler and Ziegel (2016)) An identification function is an $\mathcal{F}$-integrable function V. V is a strict $\mathcal{F}$-identification function for Γ if $\partial V(\gamma, F) = 0$ holds if and only if $\gamma = \Gamma(F)$ for all $F \in \mathcal{F}$ and for all $\gamma \in A$.

Theorem 1: (Osband's principle)(Fissler and Ziegel (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 $l \in \{1, \ldots, k\}$

$$\partial_l \mathbb{E}S(\gamma, F) = \sum_{m=1}^{k} h_{lm}(\gamma) \mathbb{E}V_m(\gamma, F), \tag{3}$$

for all $\gamma \in \mathrm{int}(\mathrm{A})$ and $F \in \mathcal{F}$. We denote the l-th partial derivative of the loss function $S(\gamma, F)$ with respect to γ as $\partial_l S$. h_{lm} is a continuous function obtained after taking the l-th partial derivative with respect to γ for $m = 1, \ldots, k$. Theorem 1 shows the relationship between a strictly consistent scoring function and strict identification function for many γ.

Remark: When k is 2 in Theorem 1 and only first partial derivative is needed, then there exists a matrix-valued function $h : \mathrm{int}(\mathrm{A}) \to \mathbb{R}^{2 \times 2}$ such that for $l \in \{1, \ldots, k\}$

$$\partial \mathbb{E}S(\gamma, F) = h_1(\gamma)\mathbb{E}V_1(\gamma, F) + h_2(\gamma)\mathbb{E}V_2(\gamma, F),$$

where ∂S is the first partial derivative of the loss function $S(\gamma, F)$ with the respect to functional γ.

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. 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.

Savage (1971) shows that the mean is elicitable and the strictly consistent scoring function for the mean is

$$S(\gamma, y) = f_1(y) - f_1(\gamma) - f_1'(\gamma)(y - \gamma), \tag{4}$$

where f_1 is a strictly convex and differentiable function. Banerjee et al. (2005) refer to Equation (4) as a Bregman function. Gneiting (2011) shows that the functional of the ratios of expectations $\Gamma(F) = \frac{\mathbb{E}_F[r(y)]}{\mathbb{E}_F[s(y)]}$ is elicitable. The strictly consistent scoring function is of the form

$$S(\gamma, y) = s(y)\left(f_2(y) - f_2(\gamma)\right) - f_2'(\gamma)\left(r(y) - \gamma s(y)\right) + f_2'(y)\left(r(y) - y s(y)\right), \tag{5}$$

where f_2 is a strictly convex and differentiable function. We can get the strictly consistent scoring function for the second moment when we choose $r(y) = y^2$ and $s(y) = 1$ in Equation (5). We can omit all terms that do not depend on γ from the strictly consistent scoring function, hence, ignoring the last term in Equation (5) we get:

$$S(\gamma, y) = f_2(y) - f_2(\gamma) - f_2'(\gamma)(y^2 - \gamma). \tag{6}$$

Theorem 2: (Fissler and Ziegel (2016)) Let λ_m be positive real numbers, $m \in \{1, \ldots, k\}$. The scoring function S is strictly $\mathcal{F}$-consistent for Γ if and only if it is of the form

$$S(\gamma_1, \ldots, \gamma_k, y) := \sum_{m=1}^{k} \lambda_m S_m(\gamma_m, y), \tag{7}$$

where $S_m : A_m \times O \to \mathbb{R}$ are strictly $\mathcal{F}$-consistent scoring function for Γ_m. Theorem 2 indicates that the summation of strictly $\mathcal{F}$-consistent scoring function is a strictly $\mathcal{F}$-consistent scoring function.

Remark: When we have a vector with two elicitable functionals $[\Gamma_1, \Gamma_2]'$, the strictly consistent scoring function for this vector of functionals is

$$S(\gamma_1, \gamma_2, y) = S_1(\gamma_1, y) + S_2(\gamma_2, y), \tag{8}$$

Let Γ_1 be the mean functional and Γ_2 be the second moment functional, the strictly consistent scoring function of the mean and second moment jointly is

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

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

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

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 $\boldsymbol{\gamma} = (\gamma_1, \gamma_2) \mapsto \left(\gamma_1, \gamma_2 - \gamma_1^2\right)$. According to Theorem 3, the inverse of g is given by $g^{-1} : A' \to A$, $\boldsymbol{\gamma}_g = (\gamma_1, \gamma_3) \mapsto \boldsymbol{\gamma} = (\gamma_1, \gamma_2)$. Define Γ_3 be the variance functional. The functional we want to get is $(\Gamma_1, \Gamma_3)'$ can now be written as $g\left((\Gamma_1, \Gamma_2)'\right)$ and it is elicitable. Then, we transfer from $(\Gamma_1, \Gamma_2)'$ to $(\Gamma_1, \Gamma_3)'$, and get the strictly consistent scoring function of $(\Gamma_1, \Gamma_3)'$ as follows:

$$\begin{aligned} S_g(\gamma_1, \gamma_3, y) &= S\left(g_1^{-1}(\gamma_1, \gamma_2), g_2^{-1}(\gamma_1, \gamma_2), y\right) \\ &= f_1(y) - f_1(\gamma_1) - f_1'(\gamma_1)(y - \gamma_1) + f_2(y) - f_2\left(\gamma_2 + \gamma_1^2\right) \\ &\quad - f_2'\left(\gamma_2 + \gamma_1^2\right)\left[y^2 - \left(\gamma_2 + \gamma_1^2\right)\right], \end{aligned} \tag{9}$$

where f_1 and f_2 are strictly convex and differential functions.

Theorem 4: Denote $S_g(\gamma_1, \gamma_3, y)$ as the strictly consistent scoring function of mean and variance with $f_1(z) = z^2$ and $f_2(z) = -\log(z)$,

$$S_g(\gamma_1, \gamma_3, y) = (y - \gamma_1)^2 - \log(y^2) + \log\left(\gamma_3 + \gamma_1^2\right) + \frac{y^2}{\gamma_3 + \gamma_1^2} - 1. \tag{10}$$

We will use the S_g as the strictly consistent scoring function to test predictive ability in predictive volatility models.

From Definition 5 and Theorem 1, we know that strict identification functions can be obtained from the derivatives of the strictly consistent scoring function with respect to the functionals,

$$\begin{aligned}
\mathbb{E}_F\mathbf{m} \equiv& \frac{\partial \mathbb{E}_F S_g(\boldsymbol{\gamma}_g, Y)}{\partial \boldsymbol{\gamma}_g} = \mathbf{h}(\boldsymbol{\gamma}_g)\mathbb{E}_F\mathbf{V}(\boldsymbol{\gamma}_g, Y) = 0,\\
\mathbb{E}_F m_1 \equiv& \frac{\partial \mathbb{E}_F S_g(\gamma_1, \gamma_3, Y)}{\partial \gamma_1}\\
=& h_{11}(\gamma_1, \gamma_3)\mathbb{E}_F V_1(\gamma_1, \gamma_3, Y) + h_{12}(\gamma_1, \gamma_3)\mathbb{E}_F V_2(\gamma_1, \gamma_3, Y)\\
=& 0,\\
\mathbb{E}_F m_2 \equiv& \frac{\partial \mathbb{E}_F S_g(\gamma_1, \gamma_3, Y)}{\partial \gamma_3}\\
=& h_{21}(\gamma_1, \gamma_3)\mathbb{E}_F V_1(\gamma_1, \gamma_3, Y) + h_{22}(\gamma_1, \gamma_3)\mathbb{E}_F V_2(\gamma_1, \gamma_3, Y)\\
=& 0,
\end{aligned}$$

where

$$\mathbf{m} = (m_1, m_2)', \qquad \boldsymbol{\gamma}_g = (\gamma_1, \gamma_3)',$$

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

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

The $\boldsymbol{\gamma}_g$ is a vector of γ_1 and γ_3. The **m** is a vector of two martingale differences m_1 and m_2. The **m** is obtained by the first derivative of the conditional expectation of strictly consistent scoring function with respect to the vector of functionals, so **m** is equal to zero. Moreover, the conditional expectation is conditioning on the past x_t. Thus, the m_1 is the martingale difference for the mean and m_2 is the martingale difference for the volatility. The **m** is the martingale difference for the vector of mean and volatility, and is the product of the strict identification functions **V** and the functions **h**. The $\mathbf{V}(\gamma_1, \gamma_3, Y)$ is a vector of strict identification functions for mean and variance. The $\mathbf{h}(\gamma_1, \gamma_3)$ is a matrix of the functions of functionals.

The scoring function of mean and variance from Bregman divergence S_g is

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

Thus, the martingale differences for the mean and volatility are derivative of the strictly consistent scoring function with respect to the functionals:

$$\mathbb{E}_F m_1 = \frac{\partial \mathbb{E}_F S_g}{\partial \gamma_1} = \mathbb{E}_F \left[-2\left(y-\gamma_1\right) + \frac{2\gamma_1}{\gamma_3+\gamma_1^2} - \frac{2\gamma_1\left(y^2\right)}{\left(\gamma_3+\gamma_1^2\right)^2} \right] = 0,$$
$$\mathbb{E}_F m_2 = \frac{\partial \mathbb{E}_F S_g}{\partial \gamma_3} = \mathbb{E}_F \left[\frac{1}{\gamma_3+\gamma_1^2} - \frac{\left(y^2\right)}{\left(\gamma_3+\gamma_1^2\right)^2} \right] = 0,$$

where m_1 is the martingale difference sequence of the mean and m_2 is the martingale difference sequence of the volatility. The m_1 and m_2 are martingale difference sequences because the expectations of the first-order conditions are taken with respect to the conditional distribution F. The martingale difference **m** can be presented as the product of the identification function **V** and the function of functionals **h** as follows:

$$\begin{aligned}\mathbb{E}_F \mathbf{m} &= \mathbf{h}(\gamma_1,\gamma_3)\mathbb{E}_F \mathbf{V}(\gamma_1,\gamma_3,y) \\ &= \begin{bmatrix} h_{11}(\gamma_1,\gamma_3) & h_{12}(\gamma_1,\gamma_3) \\ h_{21}(\gamma_1,\gamma_3) & h_{22}(\gamma_1,\gamma_3) \end{bmatrix} \mathbb{E}_F \begin{bmatrix} V_1(\gamma_1,\gamma_3,Y) \\ V_2(\gamma_1,\gamma_3,Y) \end{bmatrix} \\ &= \begin{bmatrix} -2 & \frac{2\gamma_1}{(\gamma_3+\gamma_1)^2} \\ 0 & \frac{1}{(\gamma_3+\gamma_1)^2} \end{bmatrix} \mathbb{E}_F \begin{bmatrix} y-\gamma_1 \\ \gamma_3+\gamma_1^2-y^2 \end{bmatrix} = 0.\end{aligned}$$

Then, we can derive the vector of γ_1 and γ_3 as follows:

$$\begin{bmatrix} \gamma_1 \\ \gamma_3 \end{bmatrix} = \begin{bmatrix} \mathbb{E}_F(y) \\ \mathbb{E}_F\left(y-\mathbb{E}_F(y)\right)^2 \end{bmatrix}.$$

Therefore, the scoring function $S_g(\gamma_1,\gamma_3,y)$ elicits γ_1 (mean) and γ_3 (variance) jointly.

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.

4.1 Model

In this paper, we propose a new approach to test the out-of-sample predictive ability of a variable in volatility modelling. Since the new approach is based on the encompassing principle, we first develop the framework of the forecast encompassing approach.

We have two models. One model is without conditioning on x and the other model is with conditioning on x. The volatility of the distribution $F_1 = F_Y(Y)$ of Model 1 is $\Gamma(F_1)$. The conditional volatility of the conditional distribution $F_2 = F_{Y|X}(Y|X)$ of Model 2 is the functional $\Gamma(F_2)$. The two nested models are

$$\text{Model 1}: \quad y_{t+1} = \mu_{t+1} + u_{1,t+1} \tag{11}$$

$$\sigma^2_{1,t+1} = \omega_1 + \alpha_1 u^2_{1,t} + \beta_1 \sigma^2_{1,t} \tag{12}$$

$$\text{Model 2}: \quad y_{t+1} = \mu_{t+1} + u_{2,t+1} \tag{13}$$

$$\sigma^2_{2,t+1} = \omega_2 + \alpha_2 u^2_{2,t} + \beta_2 \sigma^2_{2,t} + \delta_2 x^2_t, \tag{14}$$

where $\sigma^2_{1,t+1}$ and $\sigma^2_{2,t+1}$ are conditional variances of Model 1 and Model 2. We consider the same conditional mean model for μ_{t+1} for both Model 1 and Model 2. The dependent variable y_{t+1} is a scalar. All the parameters are estimated based on in-sample observations minimizing the loss function in Equation (10). Then we can get the volatilities $\hat{\sigma}^2_{1,t+1}$ and $\hat{\sigma}^2_{2,t+1}$ using the estimated parameters. We use μ_{t+1} to represent $\gamma_{1,t+1}$ and σ^2_{t+1} to represent $\gamma_{3,t+1}$, so we have the Bregman loss function

$$S_g\left(\mu_{t+1}, \sigma^2_{t+1}, y_{t+1}\right) = \left(y_{t+1} - \mu_{t+1}\right)^2 - \log(y^2_{t+1}) + \log\left(\sigma^2_{t+1} + \mu^2_{t+1}\right) + \frac{y^2_{t+1}}{\sigma^2_{t+1} + \mu^2_{t+1}} - 1.$$

Using the out-of-sample forecast encompassing approach in predictive models, we check if the predictor x_t has the predictive ability for the dependent variable y_{t+1}. If the predictor x_t has the predictive ability for the dependent variable y_{t+1}, we select it and use this predictor in the regression to forecast y_{t+1}. We use nested regression models with one predictor x_t in the large model, so our out-of-sample forecast encompassing approach is an out-of-sample Granger-causality test. If the predictor x_t has the predictive ability for the dependent variable y_{t+1}, this predictor Granger causes y_{t+1} in volatility regression.

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 (12) and (14) 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_t for y_{t+1}.

Thus, we develop the encompassing statistic and prove that the encompassing statistic is asymptotically normal. In order to find the encompassing statistic for the volatility, we combine two models (Model 1 and Model 2) with weights $(1 - \lambda)$ and λ. We define the moment function for the combined model as a derivative of the expectation of loss function for the combined model with respect to the weight λ. The moment function is the encompassing statistic we developed using the encompassing principle.

Based on the out-of-sample forecast encompassing principle, we combine Model 1 and Model 2 with weight $(1 - \lambda)$ and λ, respectively. Then, the null and alternative hypotheses are

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

Under the null hypothesis $\lambda = 0$, the combined model is Model 1, which means that x_t does not Granger-cause y_{t+1}. Under the alternative $\lambda \neq 0$, the combined model combines Model 1 and Model 2 with weights $(1 - \lambda)$ and λ, 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 and get the encompassing statistic under the null hypothesis, we estimate the weight λ by minimizing the expectation of least squares loss function with combined volatility $\sigma^2_{c,t+1}$. We use the weight λ to combine volatilities in two models so that $\sigma^2_{c,t+1} = (1-\lambda)\,\sigma^2_{1,t+1} + \lambda\sigma^2_{2,t+1}$. Thus, we obtain λ from

$$\lambda = \arg\min_{\lambda} \mathbb{E}_F S_g(\mu_{t+1}, \sigma^2_{c,t+1}, y_{t+1}), \tag{15}$$

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

$$M_c \equiv \frac{\partial \mathbb{E}_F\left[S_g(\mu_{t+1}, \sigma^2_{c,t+1}, Y_{t+1})\right]}{\partial\lambda} = 0. \tag{16}$$

The moment function M_c can be derived as follows:

$$\begin{aligned} M_c \equiv& \frac{\partial \mathbb{E}_F\left[S_g(\mu_{t+1}, \sigma^2_{c,t+1}, y_{t+1})\right]}{\partial\lambda} \\ =& \mathbb{E}_F\left(\frac{\partial S_g(\mu_{t+1}, \sigma^2_{c,t+1}, y_{t+1})}{\partial \sigma^2_{c,t+1}}\right)\left(\frac{\partial \sigma^2_{c,t+1}}{\partial\lambda}\right). \end{aligned} \tag{17}$$

From the equation above, we can see that the moment function M_c can be presented in the product of two parts. The first part of Equation (17) is a derivative of the loss function with respect to the conditional volatility for the combined model $\sigma^2_{c,t+1}$. We refer to the first derivative as the generalized residual function (Gourieroux et al., 1987) of the conditional volatility for the combined model. We define $h_{c,t+1}$ as a function of $\sigma^2_{c,t+1}$:

$$h_{c,t+1} = \frac{1}{(\mu^2_{t+1} + \sigma^2_{c,t+1})^2}.$$

Let $V_{c,t+1}$ be the strict identification function of variance for the combined model $\sigma^2_{c,t+1}$:

$$V_{c,t+1} = (\mu^2_{t+1} + \sigma^2_{c,t+1})^2 - y^2_{t+1}.$$

We use $h_{c,t+1}V_{c,t+1}$ to denote the generalized residual function of the conditional volatility for the combined model. The second part of Equation (17) is a derivative of the conditional volatility for the combined model $\sigma^2_{c,t+1}$ with respect to weight λ. We call the second derivative to be the test function of the conditional volatility. The test function of the conditional volatility is denoted as ν_{t+1}. Thus, the moment function M_c is the expectation of the product of the generalized residual function $h_{c,t+1}V_{c,t+1}$ and the test function ν_{t+1}.

Note that $h_{c,t+1}V_{c,t+1} = \frac{1}{\mu^2_{t+1}+\sigma^2_{c,t+1}} - \frac{y^2_{t+1}}{(\mu^2_{t+1}+\sigma^2_{c,t+1})^2}$, and $\nu_{t+1} = \sigma^2_{2,t+1} - \sigma^2_{1,t+1}$. If $\lambda = 0$, $\sigma^2_{c,t+1} = \sigma^2_{0,t+1}$, $\sigma^2_{c,t+1} = \sigma^2_{1,t+1}$, $h_{c,t+1} = h_{1,t+1}$ and $V_{c,t+1} = V_{1,t+1}$, where

$h_{1,t+1}V_{1,t+1}$ is the generalized residual function of model 1. Thus, under the null hypothesis, we have the moment function M_c to be

$$\begin{aligned} M_c &= \mathbb{E}_F\left(\frac{1}{\mu_{t+1}^2+\sigma_{1,t+1}^2} - \frac{y_{t+1}^2}{(\mu_{t+1}^2+\sigma_{1,t+1}^2)^2}\right)\left(\sigma_{2,t+1}^2 - \sigma_{1,t+1}^2\right) \\ &= \mathbb{E}_F\left(h_{1,t+1}V_{1,t+1}\right)\left(\nu_{t+1}\right) \\ &= M_1 = 0, \end{aligned}$$

where M_1 is the moment function under the null hypothesis (Model 1). The M_c is the moment function for finding the optimal weight λ by minimizing the expectation of the scoring function for the combined model $S_g(\mu_{t+1}, \sigma_{c,t+1}^2, y_{t+1})$, so M_c is the first order condition of the expectation of the scoring function with respect to the weight, which is equal to zero.

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 $\hat{\omega}_1, \hat{\alpha}_1, \hat{\beta}_1, \hat{\omega}_2, \hat{\alpha}_2, \hat{\beta}_2$ and $\hat{\delta}_2$ by using the in-sample observations from 1 to R. And then, we estimate $\hat{\sigma}_{1,t+1}^2 = \hat{\omega}_1 + \hat{\alpha}_1 u_{1,t}^2 + \hat{\beta}_1 \sigma_{1,t}^2$ and $\hat{\sigma}_{2,t+1}^2 = \hat{\omega}_2 + \hat{\alpha}_2 u_{2,t}^2 + \hat{\beta}_2 \sigma_{2,t}^2 + \hat{\delta}_2 x_t$ by using out-of-sample observations from $R+1$ to $T+1$. Then, we use one step ahead method to estimate M_0 by $\hat{M}_{R,P}$ using the out-of-sample observations from $R+1$ to $T+1$. The $\hat{M}_{R,P}$ is defined as

$$\hat{M}_{R,P} \equiv P^{-1}\sum_{t=R}^{T}\left(\hat{h}_{1,t+1}\hat{V}_{1,t+1}\hat{\nu}_{t+1}\right) \xrightarrow{p} M_0 = 0, \tag{18}$$

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

In order to get the encompassing statistic for the volatility, we consider endowing Model 1 and Model 2 with the weight $1-\lambda$ and λ, respectively, and find out the property of optimal weight λ. In order to test if x_t has the predictive ability on y_{t+1}, we standardize the $\hat{M}_{R,P}$ to the ENC statistic. The ENC statistic is $\mathrm{ENC}_{R,P} \equiv \hat{Q}_{R,P}^{-0.5}\sqrt{P}\hat{M}_{R,P}$, where $\hat{Q}_{R,P}$ is a consistent estimator for $Q_{R,P} = \mathrm{var}\left(\sqrt{P}\hat{M}_{R,P}\right)$.

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 $\hat{\theta}_{i,t}, i = 1, 2, t = R, \ldots, T$, satisfy $\hat{\theta}_{i,t} - \theta_i = B_i(t)H_i(t) = \left(R^{-1}\sum_{j=t-R+1}^{t} q_{i,j}\right)^{-1}\left(R^{-1}\sum_{j=t-R+1}^{t} \kappa_{i,j}\right)$, where $q_1 = 1$, $q_2 = \mathbb{E}\left(x_t^2\right)$, and κ_i is the derivative of the loss function with respect to the parameter vector θ, $\theta = (\omega, \alpha, \beta, \delta)'$.

Assumption 2: Let $U_t = [k_{2,t}V_{2,t}, x_t' - \mathbb{E}x_t', \kappa_{2,t}', \text{vec}\left(\kappa_{2,t}\kappa_{2,t}' - \mathbb{E}\kappa_{2,t}k_{2,t}'\right),$ $\text{vec}\left(q_{2,t} - \mathbb{E}q_{2,t}\right)']$. (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) $\mathbb{E}\kappa_{2,t}\kappa_{2,t}' = \sigma^2\mathbb{E}q_{2,t}$, (b) $\mathbb{E}\left(\kappa_{2,t} \mid \kappa_{2,t-j}, q_{2,t-j}, j = 1, 2, \ldots\right) = 0$.

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. 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 $R = O(T^{2/3})$. The out-of-sample window P has a faster divergent rate than the in-sample window R. This result is consistent with Assumption 4.

We can get the ENC statistics by standardizing the estimated moment function $\hat{M}_{R,P}$, then we have the following main theorem:

Theorem 5: Suppose Assumptions 1-3 hold. Then

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

under $\mathbb{H}_0$, where $\xi = R/T$, $c_{t+1} = \hat{h}_{1,t+1}\hat{V}_{1,t+1}\hat{\nu}_{t+1}$, $\bar{C} = 1/P\sum_{t=R}^{T} c_{t+1}$ and $W(s)$ is a standard Wiener process, $s \in (0, 1)$.

Theorem 6: Suppose Assumptions 1-4 hold. Then $\text{ENC}_{R,P} \xrightarrow{d} N(0, 1)$ under $\mathbb{H}_0$.

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

5 Monte Carlo Simulations

We divide the simulation results into two parts. First, we confirm the theoretical properties of $\mathrm{ENC}_{R,P}$ in finite sample: we examine size and power properties, as well as asymptotic distribution of the proposed statistics. Second, we estimate the optimal combination weight, λ, and examine the performance of the combined model.

5.1 Simulation Design

In order to examine the finite sample properties of the asymptotic distribution of the forecast encompassing test, we simulate data from the following DGP. We generate the variable x_t to be an AR(1) process. We generate y_{t+1} as follows:

$$
\begin{aligned}
y_{t+1} &= \mu_{t+1} + u_{t+1}, \\
\sigma_{t+1}^2 &= \omega + \alpha u_t^2 + \beta \sigma_t^2 + \delta x_t^2,
\end{aligned}
$$

where $x_t = \rho x_{t-1} + v_t$, $v_t \overset{iid}{\sim} N(0, \sigma_v^2)$, $\rho = 0.5$, $\sigma_v = 1$, $\mu_{t+1} = 0.1$, $\alpha = 0.05$, $\beta = 0.75$, and $\omega = (1-\alpha-\beta)\sigma_u^2 - \delta\sigma_v^2/(1-\rho)$. 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 $P \in \{48, 240, 1200\}$.

We compare the ENC statistic with the DM, CCS, and CCS2 statistics. The ENC statistic is

$$
\mathrm{ENC}_{R,P} \equiv \hat{Q}_{R,P}^{-0.5} \sqrt{P} \hat{M}_{R,P},
$$

where $Q_{R,P} = \mathrm{var}\left(\sqrt{P}\hat{M}_{R,P}\right)$. The DM statistic is

$$
\mathrm{DM}_{R,P} \equiv \hat{S}_{R,P}^{-0.5} \sqrt{P} \hat{D}_{R,P},
$$

where $\hat{D}_{R,P} = P^{-1}\sum_{t=R}^{T}\left(S_g(\mu_{t+1}, \sigma_{1,t+1}^2, y_{t+1}) - S_g(\mu_{t+1}, \sigma_{2,t+1}^2, y_{t+1})\right)$ and $S_{R,P} = \mathrm{var}\left(\sqrt{P}\hat{D}_{R,P}\right)$. The CCS statistic (Chao et al., 2001) is

$$
\mathrm{CCS}_{R,P} \equiv \hat{W}_{1,R,P}^{-0.5} \sqrt{P} \hat{Z}_{1,R,P},
$$

where $\hat{Z}_{1,R,P} = P^{-1}\sum_{t=R}^{T}\hat{h}_{1,t+1}\hat{V}_{1,t+1}x_t$ and $W_{1,R,P} = \mathrm{var}\left(\sqrt{P}\hat{Z}_{1,R,P}\right)$. The CCS2 statistic is

$$
\mathrm{CCS2}_{R,P} \equiv \hat{W}_{2,R,P}^{-0.5} \sqrt{P} \hat{Z}_{2,R,P},
$$

where $\hat{Z}_{2,R,P} = P^{-1}\sum_{t=R}^{T}\hat{h}_{1,t+1}\hat{V}_{1,t+1}x_t^2$ and $W_{2,R,P} = \mathrm{var}\left(\sqrt{P}\hat{Z}_{2,R,P}\right)$. $\hat{Q}_{R,P}$, $\hat{S}_{R,P}$, $\hat{W}_{1,R,P}$ and $\hat{W}_{2,R,P}$ are consistent estimators of $Q_{R,P} = \mathrm{var}\left(\sqrt{P}M_{R,P}\right)$, $S_{R,P} = \mathrm{var}\left(\sqrt{P}D_{R,P}\right)$, $W_{1,R,P} = \mathrm{var}\left(\sqrt{P}Z_{1,R,P}\right)$, and $W_{2,R,P} = \mathrm{var}\left(\sqrt{P}Z_{2,R,P}\right)$, respectively. We use rolling windows to estimate $\hat{\theta} = [\hat{\omega}_{i,t}, \hat{\alpha}_{i,t}, \hat{\beta}_{i,t}, \hat{\delta}_t]$, where $i = 1, 2$, and then predict one step ahead dependent variable $\hat{y}_{t+1}$ and calculate the forecast mean and variance $\hat{\mu}_{t+1}$ and $\hat{\sigma}_{t+1}^2$, then obtain $\mathrm{DM}_{R,P}$, $\mathrm{ENC}_{R,P}$, $\mathrm{CCS}_{R,P}$, and and $\mathrm{CCS2}_{R,P}$statistics. We repeat this procedure 2000 times and get the asymptotic distribution of $\mathrm{DM}_{R,P}$, $\mathrm{ENC}_{R,P}$, $\mathrm{CCS}_{R,P}$, and $\mathrm{CCS2}_{R,P}$ and the size and power.

5.2 Simulation Results

Table 1 examines the size properties of four statistics of interest ($\text{ENC}_{R,P}$, $\text{DM}_{R,P}$, $\text{CCS}_{R,P}$ and $\text{CCS2}_{R,P}$). We consider the statistic values at the 5% nominal level. For $P = 240$, the $\text{ENC}_{R,P}$ statistic shows good size across all repeats, staying relatively close to the 0.05 mark. Meanwhile, $\text{DM}_{R,P}$ tends to undervalue, and $\text{CCS}_{R,P}$ and $\text{CCS2}_{R,P}$ fluctuate around the ideal mark. At $P = 480$ and $P = 1200$, we notice a similar trend where the $\text{ENC}_{R,P}$ statistic maintains good size, particularly evident with $P = 480$ where it hovers very close to 0.05 across all repeats.

Table 2 examines the power properties of the statistics. We notice that the power increases as the R and P values increase. Particularly, $\text{ENC}_{R,P}$ and $\text{CCS2}_{R,P}$ move towards the value of 1 as P increases, indicating improved power. At $P = 1200$, these statistics portray high power, with $\text{ENC}_{R,P}$ nearing or exceeding 0.9 in the higher R values, showing a very substantial power.

Notably, $\text{ENC}_{R,P}$ performs well in terms of both size and power. In contrast, $\text{CCS}_{R,P}$ can achieve good size properties but exhibits poor power, whereas $\text{CCS2}_{R,P}$ has good power but poor performance in terms of size. $\text{DM}_{R,P}$ has the worst performance for both size and power estimation.

Figures 1-2 verify asymptotic normality of $\text{ENC}_{R,P}$: ENC statistics for the GARCH model have proper size and good power properties. The distributions of ENC for GARCH models are asymptotically normal under the null hypothesis.

Table 3 and Table 4 report the estimated combination weight, $\hat{\lambda}_{R,P}$, and the out-of-sample forecast error loss of the combined model. The latter ($\sigma^2_{c,t+1}$) consistently has the lowest forecast error loss compared to the individual models ($\sigma^2_{1,t+1}$ and $\sigma^2_{2,t+1}$) across different R and P values. This is an indication that the combined model is superior and presents a better alternative for forecasting.

6 Empirical Analysis

In the empirical application, we study predictive ability of financial and macroeconomic variables in forecasting conditional volatility of the equity premium. The data is sourced from Welch and Goyal (2008) and spans a period from January 1926 to December 2018. The dataset comprises a total of 1116 monthly observations. The observations include a variable y_{t+1}, which represents equity premium, and a variable x_t, which represents different financial measures such as the inflation rate (INFL), book-to-market ratio (BM), stock variance (SVAR), or long-term return (LTR). Recall that the number of in-sample observations is defined as R, the number of out-of-sample observations is defined as P. The total number of observations is $R + P = T + 1$. We consider three cases: $\xi = \frac{R}{T} = \{1/4, 2/4, 3/4\}$.

Table 5 illustrates the Granger-causality relationships in the volatility model using Welch-Goyal dataset. Four predictors, x_t (INFL, BM, SVAR, and LTR) are investigated for their predictive ability to forecast the volatility of y_{t+1} (equity premium). The table comprises three different risk-to-premium $\xi = \frac{R}{T}$ ratios and utilizes four statistical measures, namely

DM (one-sided test), and ENC, CCS, and CCS2 (two-sided tests). A p-value less than 0.05 indicates that the null hypothesis is rejected, and the predictor has predictive ability. For instance, for the INFL and LTR predictors all p-values associated with the ENC are less than 0.05. This suggests that these two variables have predictive ability to forecast volatility of the equity premium. For BM and SVAR, the results are significant for $\xi = 1/4$. This suggests varying effectiveness of these two predictors in the context of forecasting volatility of the equity premium.

Table 6 reports the estimated optimal combination weight and explores the performance of the combined models at various ξ ratios. The optimal weights, denoted by $\hat{\lambda}_P$, vary across different ξ ratios. The combined model always obtains the lowest out-of-sample forecast error loss compared to Model 1 and Model 2.

7 Conclusion

In this paper we have developed a novel approach for joint estimation and forecasting of the mean and volatility using scoring functions. In contrast to the existing literature, our framework does not require an assumption of zero mean or a proxy for an unknown true variance. We develop an out-of-sample forecast encompassing test to determine if a predictor has predictive ability to forecast volatility of the target variable. We prove that our ENC statistic is asymptotically normal under the null hypothesis. In the Monte-Carlo simulation, we demonstrate that the ENC statistic, compared with the DM statistic and other statistics, has the correct size under the null hypothesis of no Granger-causality of the covariate, and good power under the alternative hypothesis of Granger-causality of the covariate. We show that the asymptotic distributions of ENC statistics are standard normal under $\mathbb{H}_0$ in a finite sample. We apply our method to the empirical application using the Welch-Goyal dataset that contains macroeconomic and financial variables. We find that inflation and long-term return Granger-cause equity premium in volatility, whereas book-to-market ratio and stock variance have varying predictive effectiveness depending on the length of in-sample and out-of-sample windows.

References

Ashley, R., Granger, C. W., & Schmalensee, R. (1980). Advertising and aggregate consumption: An analysis of causality. *Econometrica*, 1149–1167.

Banerjee, A., Guo, X., & Wang, H. (2005). On the optimality of conditional expectation as a bregman predictor. *IEEE Transactions on Information Theory*, *51*(7), 2664–2669.

Brehmer, J. (2017). Elicitability and its application in risk management. *arXiv preprint arXiv:1707.09604*.

Chao, J., Corradi, V., & Swanson, N. R. (2001). Out-of-sample tests for granger causality. *Macroeconomic Dynamics*, *5*(4), 598–620.

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

Clark, T. E., & West, K. D. (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., & West, K. D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics*, *138*(1), 291–311.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & economic statistics*, *20*(1), 134–144.

Fissler, T., & Ziegel, J. F. (2016). Higher order elicitability and osband's principle. *The Annals of Statistics*, *44*(4), 1680–1707.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, *106*(494), 746–762.

Gourieroux, C., Monfort, A., Renault, E., & Trognon, A. (1987). Generalised residuals. *Journal of econometrics*, *34*(1-2), 5–32.

Hansen, P. R., & Lunde, A. (2006). Consistent ranking of volatility models. *Journal of Econometrics*, *131*(1-2), 97–121.

Inoue, A., Jin, L., & Rossi, B. (2017). Rolling window selection for out-of-sample forecasting with time-varying parameters. *Journal of Econometrics*, *196*(1), 55–67.

Lambert, N. S., Pennock, D. M., & Shoham, Y. (2008). Eliciting properties of probability distributions. In *Proceedings of the 9th acm conference on electronic commerce* (pp. 129–138).

Osband, K. (1985). *Providing incentives for better cost forecasting* (Unpublished doctoral dissertation). University of California, Berkeley.

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

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, *66*(336), 783–801.

Welch, I., & Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, *21*(4), 1455–1508.

Table 1: Size of test

	$P = 240$				$P = 480$				$P = 1200$			
Repeat=2000	DM	ENC	CCS	CCS2	DM	ENC	CCS	CCS2	DM	ENC	CCS	CCS2
$R = 120$	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

This table shows the size of DM, ENC, CCS, and CCS2 test in DGP under 5% nominal level, $\delta = 0$, $\phi = 0.5$.
Note: The test function of CCS, CCS2 and ENC are x, x^2, and $\hat{\sigma}_2^2 - \hat{\sigma}_1^2$ respectively.

Table 2: Power of test

	$P = 240$				$P = 480$				$P = 1200$			
Repeat=2000	DM	ENC	CCS	CCS2	DM	ENC	CCS	CCS2	DM	ENC	CCS	CCS2
$R = 120$	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

This table shows the power of DM, ENC, CCS, and CCS2 test in DGP under 5% nominal level, $\delta = 0.05$, $\phi = 0.5$.
Note: The test function of CCS, CCS2 and ENC are x, x^2, and $\hat{\sigma}_2^2 - \hat{\sigma}_1^2$ respectively.

Table 3: Combining Volatility Models

	$R = 240, P = 240$	$R = 240, P = 480$	$R = 240, P = 1200$
$\hat{\lambda}_{R,P}$	-0.1683	0.0056	0.0098
$\mathbb{E}[S_g(\mu, \sigma_1^2, y)]$	2.2420	2.2719	2.2766
$\mathbb{E}[S_g(\mu, \sigma_2^2, y)]$	2.2618	2.3167	2.4678
$\mathbb{E}[S_g(\mu, \sigma_c^2, y)]$	2.2373	2.2689	2.2756

Repeat=2000, $\delta = 0$.

Table 4: Combining Volatility Models

	$R = 240, P = 240$	$R = 240, P = 480$	$R = 240, P = 1200$
$\hat{\lambda}_{R,P}$	1.0723	0.8672	0.8142
$\mathbb{E}[S_g(\mu, \sigma_1^2, y)]$	2.3178	2.4677	2.3387
$\mathbb{E}[S_g(\mu, \sigma_2^2, y)]$	2.2617	2.4189	2.3291
$\mathbb{E}[S_g(\mu, \sigma_c^2, y)]$	2.2614	2.4169	2.2796

Repeat=2000, $\delta = 0.1$.

Table 5: Granger-causality using Welch-Goyal Dataset

INFL											
$R:T=1:4$				$R:T=2:4$				$R:T=3:4$			
DM	ENC	CCS	CCS2	DM	ENC	CCS	CCS2	DM	ENC	CCS	CCS2
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											
$R:T=1:4$				$R:T=2:4$				$R:T=3:4$			
DM	ENC	CCS	CCS2	DM	ENC	CCS	CCS2	DM	ENC	CCS	CCS2
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											
$R:T=1:4$				$R:T=2:4$				$R:T=3:4$			
DM	ENC	CCS	CCS2	DM	ENC	CCS	CCS2	DM	ENC	CCS	CCS2
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											
$R:T=1:4$				$R:T=2:4$				$R:T=3:4$			
DM	ENC	CCS	CCS2	DM	ENC	CCS	CCS2	DM	ENC	CCS	CCS2
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)

Note: The test function of CCS, CCS2 and ENC are x, x^2, and $\hat{\sigma}_2^2 - \hat{\sigma}_1^2$ respectively.

Table 6: Model combination result using Welch-Goyal Dataset

$R:T=1:4$				$R:T=1:2$				$R:T=3:4$			
INFL		BM		INFL		BM		INFL		BM	
$\hat{\lambda}_{R,P}$	Loss	$\hat{\lambda}_{R,P}$	Loss	$\hat{\lambda}_{R,P}$	Loss	$\hat{\lambda}_{R,P}$	Loss	$\hat{\lambda}_{R,P}$	Loss	$\hat{\lambda}_{R,P}$	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

$R:T=1:4$				$R:T=1:2$				$R:T=3:4$			
SVAR		LTR		SVAR		LTR		SVAR		LTR	
$\hat{\lambda}_{R,P}$	Loss	$\hat{\lambda}_{R,P}$	Loss	$\hat{\lambda}_{R,P}$	Loss	$\hat{\lambda}_{R,P}$	Loss	$\hat{\lambda}_{R,P}$	Loss	$\hat{\lambda}_{R,P}$	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 $\hat{\lambda}_{R,P}$.

[b]: The average of the loss of volatility from forecasting of Model 1.

[c]: The average of the loss of volatility from forecasting of Model 2.

[d]: The average of the loss of volatility from forecasting of the combined model.

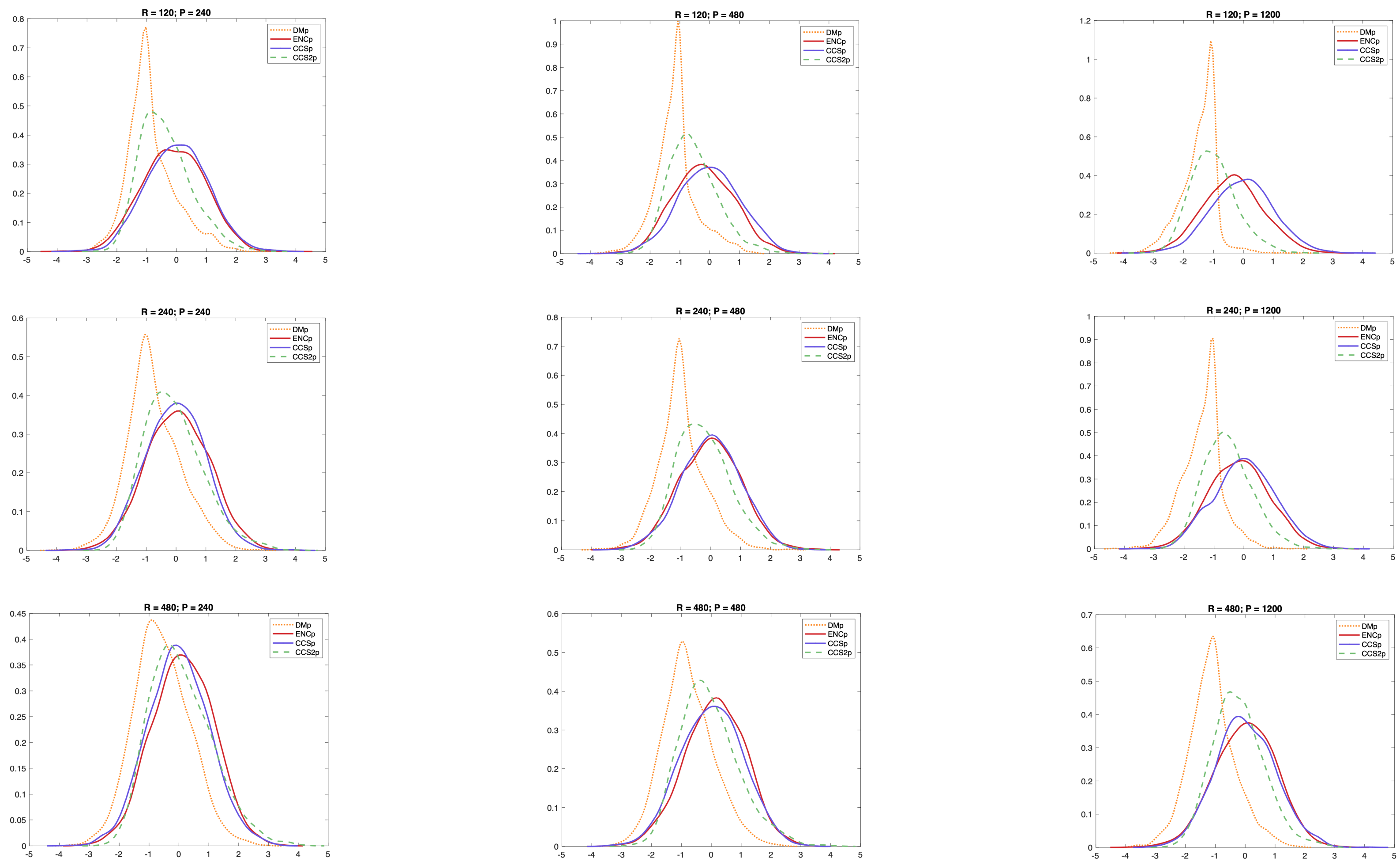

Figure 1: Size of test, $b = 0$, $\rho = 0.5$ and 2000 repeats.

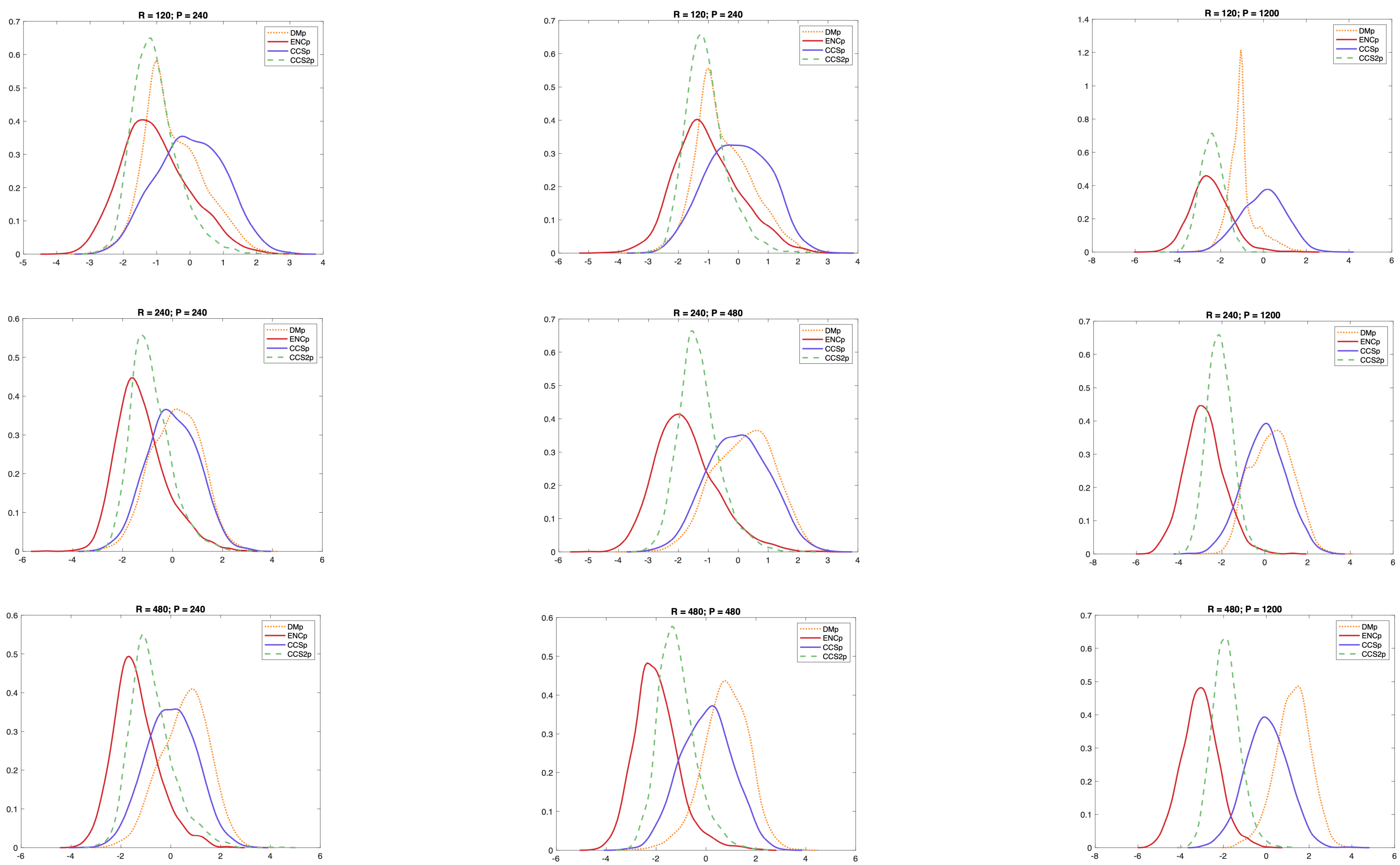

Figure 2: Power of test, $b = 0.05$, $\rho = 0.5$ and 2000 repeats.