Evaluation and Combination of Conditional Quantile Forecasts

Raffaella Giacomini* and Ivana Komunjer†,‡

This version: June 2003

Abstract

This paper proposes a method for comparing and combining conditional quantile forecasts in an out-of-sample framework. We construct a Conditional Quantile Forecast Encompassing (CQFE) test as a Wald-type test of superior predictive ability. Rejection of CQFE provides a basis for combination of conditional quantile forecasts. Two central features of our implementation of the principle of encompassing are, first, the use of the 'tick' loss function and, second, a *conditional*, rather than *unconditional* approach to out-of-sample evaluation. Some of the advantages of the conditional approach are that it allows the forecasts to be generated by using general estimation procedures and that it is applicable when the forecasts are based on both nested and non-nested models. The test is also relatively easy to implement using standard GMM techniques. An empirical application to Value-at-Risk evaluation illustrates the usefulness of our method.

Keywords: Encompassing, Loss Function, GMM, Value at Risk

J.E.L. Codes: C12, C22, C52, C53

*University of California, San Diego and Boston College (*rgiacomini@ucsd.edu*).

†Corresponding Author: Division of the Humanities and Social Sciences 228-77, California Institute of Technology, Pasadena CA 91125 (*komunjer@hss.caltech.edu*).

‡Acknowledgements: We are indebted to Graham Elliott, Clive Granger, Jose Lopez, Andrew Patton, Kevin Sheppard and the participants to the 2003 ASSA meeting in Washington, D.C. for their valuable comments and suggestions and to Peter Christoffersen and Eric Ghysels for providing us with their data. We would also like to thank the Editor, Alastair Hall, the Associate Editor and two anonymous referees for useful comments, which led to a considerably improved version of the paper.

1 Introduction

The vast majority of the economic forecasting literature has traditionally focused on producing and evaluating point forecasts for the conditional mean of some variable of interest. More recently, increasing attention has been devoted to other characteristics of the unknown forecast distribution, besides its conditional mean, such as a particular conditional quantile.

A primary example of the growing interest for conditional quantile forecasts is in the context of risk management, as witnessed by the literature on Value at Risk (e.g., Duffie and Pan 1997).[1] There are a variety of approaches to estimating conditional quantiles in general and Value at Risk (VaR) in particular. They range from fully parametric (e.g., Danielsson and de Vries 1997, Barone-Adesi, Bourgoin, and Giannopoulos 1998, Diebold, Schuermann and Stroughair 1998, Embrechts, Resnick and Samorodnitsky 1999, McNeil and Frey 2000), to semi-parametric (e.g., Koenker and Zhao 1996, Taylor 1999, Engle and Manganelli 1999, Chernozhukov and Umanstev 2001, Christoffersen, Hahn and Inoue 2001), and non-parametric (e.g., Battacharya and Gangopadhyay 1990, White 1992).

Given the range of techniques available for producing conditional quantile forecasts, it is necessary to develop adequate tools for their evaluation. A number of authors have focused on *absolute* evaluation, that is, on testing whether a given forecasting model is correctly specified or whether a sequence of forecasts satisfies certain optimality properties. For example, Zheng (1998) and Bierens and Ginther (2001) propose specification tests for evaluating a given model against a generic alternative. Christoffersen (1998), instead, proposes a 'correct conditional coverage' criterion for evaluating a sequence of interval forecasts which does not require knowledge of the underlying model. A potential problem with absolute evaluation is that if different models are rejected as being misspecified, or if they are all accepted, then we are left without any guidance as to which one to choose. In this case, it may be more appropriate to consider *relative* evaluation, which involves comparing the performance of alternative, possibly misspecified, models or sequences of

[1] Ever since August 1996, when US bank regulators adopted a 'market risk' supplement to the Basle Accord (1988), the regulatory capital requirements of commercial banks with trading activities are based on Value at Risk (VaR) estimates. This important measure of market risk is defined as the opposite of a prespecified quantile of the conditional distribution of portfolio returns, and its estimates are routinely generated by the banks' internal models.

forecasts for the same variable and choosing the one that performs the best. This approach is taken by Christoffersen *et al.* (2001), who propose a method for comparing alternative, non-nested VaR estimates. The authors assume that the VaR is a linear function of the volatility and propose estimating the parameters by the information theoretic alternative to GMM due to Kitamura and Stutzer (1997). The evaluation of Christoffersen *et al.* (2001) is conducted in-sample and is only valid if the returns belong to a location-scale family (which implies that the VaR is a linear function of the volatility). Further, in order to apply their test, all competing VaR forecasts must be obtained by the estimation method proposed by Kitamura and Stutzer (1997).

It is however frequently the case that good in-sample performance does not imply good out-of-sample performance and that the models underlying the forecasts remain partially or completely unknown to the forecast user. Moreover, given the variety of approaches to the estimation of conditional quantile models outlined above, it may be of interest to investigate whether different estimation techniques have an effect on forecast performance. In general, in the situation where several forecasts of the same variable are available, it is desirable to have formal testing procedures for out-of-sample comparison, which do not necessarily require knowledge of the underlying model, or, if the model is known, which do no restrict attention to a specific estimation procedure. The goal of this paper is to provide such a test.

Given an appropriate choice of loss function, one could in principle compare the out-of-sample average loss implied by two alternative quantile forecasts using the tests of equal predictive ability proposed by Diebold and Mariano (1995) and West (1996), but these two approaches are not applicable in several important cases, such as when the forecasts are from nested models or when they depend on semi- or non-parametric estimators.

In this paper, we choose a different approach and construct a test for out-of-sample conditional quantile forecast comparison based on the principle of encompassing. The idea of encompassing (e.g., Hendry and Richard 1982, Mizon and Richard 1986, Diebold 1989, Clements and Hendry 1998) requires that a forecast be able to explain the predictive ability of a rival forecast, and thus it can be seen as a test of superior predictive ability.

Two central features of our implementation of the principle of encompassing are, first, the

choice of the relevant loss function, which we argue to be the 'tick' loss function (also known as asymmetric linear loss function of order α) and, second, the focus on *conditional* expected loss, rather than unconditional expected loss in the formulation of the encompassing test. The focus on conditional evaluation links the approach in this paper to the one of Giacomini and White (2003) who propose a general framework for out-of-sample predictive ability testing. Some of the advantages of the conditional approach over the unconditional approach (e.g., West 1996) are that it allows the forecasts to be generated by (estimated) parametric models as well as by semi- and non-parametric techniques and that it is applicable to both nested and non-nested forecast comparisons. The implementation of the test makes use of fairly standard Generalized Method of Moments (GMM) techniques, appropriately modified to accommodate the non-differentiability of our criterion function. As a by-product, our framework also provides a link to Christoffersen's (1998) 'correct conditional coverage' criterion for the absolute evaluation of interval forecasts.

A final feature of our encompassing approach is that it gives a theoretical basis for quantile forecast combination, in cases when neither forecast has superior predictive ability. From a theoretical viewpoint, forecast combination can be seen as a way to pool the information contained in the individual forecasts, and its benefits have been widely advocated since the early work of Bates and Granger (1969).[2] Recent empirical work by Stock and Watson (1999, 2001) has further confirmed the accuracy gains induced by forecast combination for a large number of macroeconomic and financial time series. Surprisingly little empirical work has been done in the context of conditional quantile forecasting. Yet, the benefits of expanding the information set through combination might be particularly evident for quantiles with small nominal coverage - as is usually the case for VaR. Extreme quantiles are very sensitive to the few observations in the tails of the empirical distribution of the sample, and combining forecasts based on different information sets can thus be seen as a way to make the forecast performance more robust to the effects of sample-specific factors.

[2] From a theoretical point of view there are, according to Granger (1989), two situations when it is useful to combine forecasts for the same variable. If the forecasts are based on the same information set, then a forecast combination can only be useful if the original forecasts are sub-optimal according to the relevant loss function. When the forecasts are instead based on different information sets, combining the forecasts can potentially improve the forecasting performance by pooling the information contained in the two sets.

We illustrate the usefulness of the CQFE test by applying it to the problem of VaR evaluation using S&P500 daily return data. We consider a number of popular models for producing 1% and 5% VaR forecasts and generally conclude that the forecast combination outperforms the individual forecasts.

The remainder of the paper is organized as follows: Section 2 describes the environment and gives an overview of our encompassing approach to comparing and combining competing conditional quantile forecasts. Section 3 introduces the test for conditional quantile forecast encompassing and discusses the estimation problem underlying the implementation of the test. Section 4 analyzes the small-sample size and power properties of the proposed test and Section 5 applies the test to the problem of VaR forecast evaluation and combination. Section 6 concludes the paper. All proofs are in the Appendix.

2 Overview

2.1 Description of the Environment

Consider a stochastic process $X \equiv \{X_t : \Omega \longrightarrow \mathbb{R}^{k+1}, k \in \mathbb{N}, t = 1, \ldots, T\}$ defined on a complete probability space $(\Omega, \mathcal{F}, P)$ where $\mathcal{F} \equiv \{\mathcal{F}_t, t = 1, \ldots, T\}$ and $\mathcal{F}_t$ is the σ-field $\mathcal{F}_t \equiv \sigma\{X_s, s \leqslant t\}$. In what follows we partition the observed vector X_t as $X_t \equiv (Y_t, Z_t')'$, where $Y_t : \Omega \to \mathbb{R}$ is a continuous random variable of interest and $Z_t : \Omega \to \mathbb{R}^k$ a vector of explanatory variables. More specifically, we are interested in the α-quantile of the distribution of Y_{t+1} conditional on the information set $\mathcal{F}_t$, $Q_{t,\alpha}$, defined as

$$P_t(Y_{t+1} \leqslant Q_{t,\alpha}) = \alpha, \text{ or} \tag{1}$$

$$Q_{t,\alpha} \equiv F_t^{-1}(\alpha), \tag{2}$$

where $\alpha \in (0, 1)$, F_t is the conditional distribution function of Y_{t+1} and F_t^{-1} its inverse. Using the standard convention, the subscript t under the probability $P(\cdot)$, distribution function $F(\cdot)$, density $f(\cdot)$, expectation $E[\cdot]$ or α-quantile Q_t denotes conditioning on the information set $\mathcal{F}_t$. To further simplify the notation, we hereafter drop the reference to the index α and simply denote the time t

conditional α-quantile as Q_t. As a general rule, a lowercase letter is used to denote observations of the corresponding random variable (e.g. x_t and X_t).

In this paper we propose a test for comparing alternative sequences of one-step-ahead forecasts of Q_t. The evaluation is conducted in an out-of-sample fashion. This consists in dividing the available sample of size T into an in-sample part of size m and an out-of-sample part of size n, so that $T = m + n$. The in-sample portion is used to produce the first set of forecasts and the evaluation is performed over the remaining out-of-sample portion. We impose fairly few restrictions on the way the forecasts are produced. In particular, they may be based on parametric models or be generated by use of semi-parametric or non-parametric techniques. We allow the forecasts to be generated using either: (1) a fixed forecasting scheme, or (2) a rolling window forecasting scheme. For example, in the case of a parametric model, a fixed forecasting scheme involves estimating the parameters only once on the first m observations and using these estimates to produce all the forecasts for the out-of-sample period $t = m + 1, ..., T$. A rolling window forecasting scheme, instead, implies re-estimating the parameters at each out-of-sample point t using an estimation sample containing the m most recent observations, i.e. the observations from date $t - m + 1$ to date t.

Let $\hat{\beta}_{t,m}$ denote the $k \times 1$ vector collecting the time-t estimated parameters from the two models (for parametric forecasting) or whatever semi-parametric or non-parametric estimator used in the construction of the forecasts. In the following, we will use the common notation $\hat{\beta}_{t,m}$ for either forecasting scheme, with the understanding that a fixed forecasting scheme corresponds to the case where $\hat{\beta}_{t,m} = \hat{\beta}_{m,m}$ for all t, $m \leqslant t \leqslant T - 1$, while for the rolling window forecasting scheme $\hat{\beta}_{t,m}$ changes with t but only depends on the previous m observations.

For simplicity, we restrict attention to pairwise comparison and combination, but all the techniques can be readily applied to the case of multiple forecasts. For each time t, $m \leqslant t \leqslant T - 1$, the one-step ahead forecasts of the conditional quantile Q_t formulated at time t are denoted by $\hat{q}_{1m,t} \equiv q_1(x_t, x_{t-1}, ..; \hat{\beta}_{t,m})$ and $\hat{q}_{2m,t} \equiv q_2(x_t, x_{t-1}, ...; \hat{\beta}_{t,m})$, where q_1 and q_2 are $\mathcal{F}_t$-measurable functions. Note that our notation implies, in particular, that the forecasts are formed using the actual realizations of the variables over the out-of-sample period (i.e., the forecasts are not truly

ex-ante).

The crucial requirement that we impose on the functions q_1 and q_2 is that they remain constant over time. This implies, in particular, that use of an expanding estimation window (recursive forecasting scheme) is not allowed, whereas forecasting schemes using (1) a fixed or (2) a rolling window of constant length satisfy the requirement. In the remainder of the paper, we assume that the in-sample size m is a finite constant, chosen by the user a priori. As a consequence, all of our results should be interpreted as being conditional on the given choice of m, but for ease of notation we choose not to make this dependence explicit (except for $\hat{q}_{1m,t}$ and $\hat{q}_{2m,t}$).

2.2 Principles of Forecast Encompassing

Our approach to comparing conditional quantile forecasts is based on the principle of encompassing. Following, for example, Hendry and Richard (1982), Mizon and Richard (1986) and Diebold (1989), encompassing arises when one of two competing forecasts is able to explain the predictive ability of its rival. According to Clements and Hendry (1998, p. 228), a test for forecast encompassing can be generally defined as follows:

> 'A test for forecast encompassing is a test of the conditional efficiency of a forecast, where a forecast is said to be conditionally efficient if the expected loss of a combination of that forecast and a rival forecast is not significantly less than the expected loss of the original forecast alone.'

In the general definition of forecast encompassing proposed by Clements and Hendry (1998), the two key ingredients of a forecast encompassing test are: (1) the loss function involved in the computation of the expected loss and (2) the weights of the forecast combination. Before proceeding with the implementation of such a test we therefore need to discuss each of the two points in more detail. First, note that the choice of the loss function is closely related to which characteristic of the unknown future distribution of the variable one wants to forecast. Let $\hat{f}_t$ be a forecast of some characteristic of interest of the random variable Y_{t+1}, conditional on the information set at time t. The forecast $\hat{f}_t$ is said to be optimal if it minimizes $E_t[L(Y_{t+1} - \hat{f}_t)]$, where L is some loss

function, $L : \mathbb{R} \to \mathbb{R}^+$. Note that the optimal forecast minimizes the expected loss *conditional* on $\mathcal{F}_t$. As discussed in detail below, the focus on conditional, rather than unconditional, expected loss is a central feature of our treatment of both evaluation and combination of forecasts, and one that distinguishes our approach from related literature (e.g., Granger 1989, Taylor and Bunn 1998, Elliott and Timmermann 2002).

Different loss functions L correspond to different optimal forecasts. For example, letting $e_{t+1} \equiv y_{t+1} - \hat{f}_t$, if a quadratic loss function $L(e_{t+1}) = e_{t+1}^2$ is used, then the optimal forecast is the conditional mean of the distribution of Y_{t+1}. If, on the other hand, an absolute value loss function $L(e_{t+1}) = |e_{t+1}|$ is used, the optimal forecast corresponds to the conditional median of the distribution of Y_{t+1}. In the particular case of this paper, the object of interest is Q_t, the conditional α-quantile of the distribution of Y_{t+1}. The corresponding loss function L is the asymmetric linear loss function of order α, $\mathcal{T}_\alpha$, defined as

$$\mathcal{T}_\alpha(e_{t+1}) \equiv (\alpha - 1(e_{t+1} < 0))e_{t+1}, \tag{3}$$

which is also known as 'tick' or 'check' loss function in the literature. We can thus argue that the 'tick' function $\mathcal{T}$ is the implicit loss function whenever the object of interest is a forecast of a particular quantile of the conditional distribution of Y_{t+1}.

This brings us to the second key ingredient of a forecast encompassing test - the choice of weights in the forecast combination. In this paper, we focus on linear combinations of forecasts for the conditional α-quantile of Y_{t+1}, $(\theta_1 \hat{q}_{1m,t} + \theta_2 \hat{q}_{2m,t})$, obtained by choosing a set of weights $(\theta_1, \theta_2) \in \Theta$, with Θ being a compact subset of $\mathbb{R}^2$. The combination weights θ_1 and θ_2 can further be constrained to lie in $(0, 1)$, with $\theta_1 + \theta_2 = 1$, but we choose not to impose this restriction in the paper (for a discussion of restrictions on the combination weights see e.g., Granger and Ramanathan 1984). In general, there are numerous possibilities for the choice of θ_1 and θ_2, each leading to different properties of the forecast combination, as shown by Elliott and Timmermann (2002). In the context of encompassing, however, special role is played by an 'optimal' set of weights (θ_1^*, θ_2^*), which we define next.

2.3 Encompassing for Conditional Quantiles

Based on the general idea by Clements and Hendry (1998), the concept of encompassing for conditional quantile forecasts can then be formalized as follows. A Conditional Quantile Forecast Encompassing (CQFE) test for $\hat{q}_{1m,t}$ with respect to $\hat{q}_{2m,t}$ is a test for conditional efficiency of the forecast $\hat{q}_{1m,t}$, where $\hat{q}_{1m,t}$ is said to be conditionally efficient with respect to $\hat{q}_{2m,t}$ if

$$E_t[\mathcal{T}_\alpha(Y_{t+1} - \hat{q}_{1m,t}\ )] \leqslant E_t[\mathcal{T}_\alpha(Y_{t+1} - (\theta_1\hat{q}_{1m,t}\ + \theta_2\hat{q}_{2m,t})], \text{ for all } (\theta_1, \theta_2) \in \Theta. \tag{4}$$

In practice, testing the inequality (4) is not feasible, since it would involve computing the expected loss for all $(\theta_1, \theta_2) \in \Theta$. Instead, let (θ_1^*, θ_2^*) denote the optimal set of weights, defined as a solution to the minimization problem: $\min_{(\theta_1,\theta_2)\in\Theta} E_t[\mathcal{T}_\alpha(Y_{t+1} - (\theta_1\hat{q}_{1m,t}\ + \theta_2\hat{q}_{2m,t}))]$. We then have that $E_t[\mathcal{T}_\alpha(Y_{t+1} - (\theta_1^*\hat{q}_{1m,t} + \theta_2^*\hat{q}_{2m,t}))] \leqslant E_t[\mathcal{T}_\alpha(Y_{t+1} - (\theta_1\hat{q}_{1m,t} + \theta_2\hat{q}_{2m,t})]$, for every $(\theta_1, \theta_2) \in \Theta$. This in particular implies that

$$E_t[\mathcal{T}_\alpha(Y_{t+1} - (\theta_1^*\hat{q}_{1m,t} + \theta_2^*\hat{q}_{2m,t}))] \leqslant E_t[\mathcal{T}_\alpha(Y_{t+1} - \hat{q}_{1m,t})]. \tag{5}$$

Hence, we obtain the following condition for the conditional efficiency of $\hat{q}_{1m,t}$ with respect to $\hat{q}_{2m,t}$, which is equivalent to that in (4).

Definition 1 (Conditional Quantile Forecast Encompassing) *Let $\hat{q}_{1m,t}$ and $\hat{q}_{2m,t}$ be two alternative forecasts for Q_t. The forecast $\hat{q}_{1m,t}$ is said to encompass $\hat{q}_{2m,t}$ if and only if*

$$E_t[\mathcal{T}_\alpha(Y_{t+1} - \hat{q}_{1m,t}\ )] = E_t[\mathcal{T}_\alpha(Y_{t+1} - (\theta_1^*\hat{q}_{1m,t}\ + \theta_2^*\hat{q}_{2m,t}\ )], a.s. - P, \text{ for t=1,2,...}, \tag{6}$$

where $\mathcal{T}_\alpha$ is the 'tick' loss function defined in (3) *and (θ_1^*, θ_2^*) is a solution to the problem* $\min_{(\theta_1,\theta_2)\in\Theta} E_t[\mathcal{T}_\alpha(Y_{t+1} - (\theta_1\hat{q}_{1m,t} + \theta_2\hat{q}_{2m,t}))]$.

Comments:

1. Interpreting a conditional expectation as a prediction, the equality (6) can be viewed as saying that $\hat{q}_{1m,t}$ encompasses $\hat{q}_{2m,t}$ if the forecaster cannot predict whether the optimal combination of the two forecasts will outperform the original forecast in the future, given what is known today.

This focus on prediction of future performance (conditional expectation), rather than on assessment of average performance (unconditional expectation) in the definition of encompassing distinguishes our approach from the classic encompassing literature (e.g., Hendry and Richard 1982, Mizon and Richard 1986) and establishes a link with the general framework for predictive ability testing proposed by Giacomini and White (2003).

2. Similarly to Giacomini and White (2003), the forecasts $\hat{q}_{1m,t}$ and $\hat{q}_{2m,t}$ in our definition of encompassing explicitly depend on the parameter estimates at time t, rather than on population values of the parameters as in, e.g., West (2001). The implicit premise here is that the forecast user is interested in real-time forecast selection or combination, that is, in selecting at time t the best forecast for time $t+1$ or in combining the available forecasts if neither is found to be superior. The relevant objects of the evaluation are in this case the actual forecasts, depending on parameter estimates, rather than forecasts which depend on population values only achieved in the limit.

3. Focusing on the actual forecasts, rather than the underlying models, in the definition of encompassing means that we do not impose the restriction that the forecasts are estimated using the same 'tick' loss function used for the estimation of the combination weights. As a result, we provide a unified framework for comparing forecasts which may be obtained by utilizing different estimation techniques.

Since the right hand side of equation (6) is the minimum of the conditional expected loss over Θ, the equality in equation (6) will only hold if $\theta_1^* = 1$ and $\theta_2^* = 0$. A CQFE test for $\hat{q}_{1m,t}$ with respect to $\hat{q}_{2m,t}$ can thus be formulated as a test of the null hypothesis $H_{10} : (\theta_1^*, \theta_2^*) = (1, 0)$ against $H_{1a} : (\theta_1^*, \theta_2^*) \neq (1, 0)$. Similarly, to test whether $\hat{q}_{2m,t}$ encompasses $\hat{q}_{1m,t}$, the relevant null and alternative hypotheses would be $H_{20} : (\theta_1^*, \theta_2^*) = (0, 1)$ and $H_{2a} : (\theta_1^*, \theta_2^*) \neq (0, 1)$. If one of the two forecasts represents a natural benchmark, or in cases where economic theory may suggest which of the two hypotheses H_{10} or H_{20} is more relevant, only one of the two encompassing tests will be performed. If, instead, no a priori ordering is available, one would in practice perform a sequential test of H_{10} and H_{20}. If both null hypotheses H_{10} and H_{20} are rejected, the conclusion would be that neither forecast can fully explain the predictive ability of the other, and thus both contain relevant information about Y_{t+1}. In this last case, the forecast combination defined as $\hat{q}_{m,t}$

$\equiv \theta_1^* \hat{q}_{1m,t} + \theta_2^* \hat{q}_{2m,t}$ might outperform both of the original forecasts.

To ease the notation, we hereafter let $\theta \equiv (\theta_1, \theta_2)'$, and $\hat{q}_{m,t} \equiv (\hat{q}_{1m,t}, \hat{q}_{2m,t})'$, so that $\theta_1 \hat{q}_{1m,t} + \theta_2 \hat{q}_{2m,t} = \theta' \hat{q}_{m,t}$. The set of optimal weights $\theta^* \equiv (\theta_1^*, \theta_2^*)'$ introduced in Definition 1 solves

$$\theta^* = \arg\min_{\theta \in \Theta} E_t[(\alpha - 1(Y_{t+1} - \theta' \hat{q}_{m,t} < 0))(Y_{t+1} - \theta' \hat{q}_{m,t})], \tag{7}$$

and the first order condition corresponding to (7) is given in the following Proposition.

Proposition 2 (Correct conditional coverage criterion) *Let θ^* be a solution to the minimization problem* (7)*. Then θ^* satisfies the following first order condition*

$$E_t[\alpha - 1(Y_{t+1} - \theta^{*\prime} \hat{q}_{m,t} < 0)] = 0, a.s. - P. \tag{8}$$

It is interesting to note that the first order condition (8) verified by a solution to the initial minimization problem (7) corresponds exactly to the Christoffersen's (1998) 'correct conditional coverage criterion' for evaluating the performance of interval forecasts. The key appealing property of Christoffersen's correct conditional coverage condition is its intuitive interpretation in terms of information content of a given sequence of interval forecasts. What Proposition 2 shows is that the correct conditional coverage condition also corresponds to a particular choice of loss function. We can thus say that any conditional quantile forecast satisfying the first order condition (8) is optimal in the sense that it minimizes the expected 'tick' loss.

In practice, the optimal vector of weights θ^* in (7) is unknown and needs to be estimated. In the following section we discuss the estimation problem in the conditional framework used to define forecast encompassing.

3 Conditional Quantile Forecast Encompassing (CQFE) Test

We separately discuss the estimation of the combination weights and the implementation of the test.

3.1 GMM Estimation of Optimal Combination Weights

According to the results of Proposition 2, the magnitude in (8) should be uncorrelated with any information available at time t. It should therefore be the case that $E[(\alpha - 1(Y_{t+1} - \theta^{*\prime}\hat{q}_{m,t} < 0))W_t] = 0$, for all $\mathcal{F}_t$-measurable functions W_t. Let W_t^* be an h-vector of variables that are observed at time t and that contain all the relevant information from $\mathcal{F}_t$. We refer to W_t^* as the 'information vector'. Further, denote by g an h-vector-valued function $g : \Theta \times \mathbb{R} \times \mathbb{R}^h \to \mathbb{R}^h$ such that

$$g(\theta; y_{t+1}, w_t^*) \equiv (\alpha - 1(y_{t+1} - \theta'\hat{q}_{m,t} < 0))w_t^*. \tag{9}$$

When the information vector W_t^* contains all the relevant information from $\mathcal{F}_t$, that is, when it includes all elements of the time-t information set potentially correlated with the variable $\alpha - 1(Y_{t+1} - \theta^{*\prime}\hat{q}_{m,t} < 0)$, the solution θ^* to the first order condition (8) coincides with the solution θ^{**} to $E[g(\theta^{**}; Y_{t+1}, W_t^*)] = 0$. In the remainder of the paper, we assume that W_t^* satisfies such requirement and accordingly redefine θ^* to be a solution to

$$g_0(\theta^*) \equiv E[g(\theta^*; Y_{t+1}, W_t^*)] = 0. \tag{10}$$

In practice, the choice of W_t^* depends on the nature of the application considered. Typically, W_t^* consists of different functions of explanatory variables and/or lags of Y_{t+1}, but one may also include previous forecasts or measures of past forecast performance.[3] Some discussion on how to choose the information vector in practical applications is contained in Section 5.

An estimate of θ^* based on condition (10) can be obtained by using Hansen's (1982) GMM approach, appropriately modified to accommodate non-differentiable criterion functions. We propose estimating θ^* over the out-of-sample portion of size $n = T - m$, consisting of the sequence of observations $(w_m^{*\prime}, y_{m+1}, ..., w_{T-1}^{*\prime}, y_T)'$. The GMM estimator of θ^*, denoted $\hat{\theta}_n$, is defined as a local

[3] As stated in Proposition 4, the general requirement on $\{W_t^*\}$ is that it is a strictly stationary and mixing series. In practical applications, one should therefore verify that this assumption if satisfied. One implication is that W_t^* could include previous forecasts provided, as in our assumptions, that the forecast are produced by either a fixed or a rolling window forecasting scheme. The reason is that in these two cases the forecasts are constant measurable functions of a finite window of data and thus inherit the properties of stationarity and mixing from the underlying series.

solution to the minimization problem

$$\min_{\theta\in\Theta}[g_n(\theta)]'\hat{S}^{-1}[g_n(\theta)], \tag{11}$$

where $g_n(\cdot)$ is the sample moment function, $g_n(\theta) \equiv n^{-1}\sum_{t=m}^{T-1} g(\theta; y_{t+1}, w_t^*)$, and $\hat{S}$ a consistent estimator of the asymptotic variance matrix S,

$$S \equiv \lim_{n\to\infty} n \cdot E[g_n(\theta^*)g_n(\theta^*)'] \tag{12}$$

Using the fact that the first order condition (8) implies that $\{g(\theta^*; Y_{t+1}, W_t^*), \mathcal{F}_t\}$ is a martingale difference sequence, a consistent estimator of S is given by

$$\hat{S}(\hat{\theta}_n) \equiv n^{-1}\sum_{t=m}^{T-1} g(\hat{\theta}_n; y_{t+1}, w_t^*)g(\hat{\theta}_n; y_{t+1}, w_t^*)', \tag{13}$$

where $\hat{\theta}_n$ is some consistent initial estimate of θ^*.[4]

The fact that the weighting matrix $\hat{S}^{-1}(\hat{\theta}_n)$ depends on the estimator $\hat{\theta}_n$ itself, calls for a recursive approach. The computation of the GMM estimator $\hat{\theta}_n$ is typically carried out by first choosing an $r \times r$ identity weighting matrix $I_{r\times r}$ and using (11) to compute the corresponding $\hat{\theta}_{n,1}$. The resulting new weighting matrix $\hat{S}^{-1}(\hat{\theta}_{n,1})$ is more efficient than the previous one, and solving (11) leads to a new estimator $\hat{\theta}_{n,2}$. The last two steps can then be repeated until $\hat{\theta}_{n,j}$ equals its previous value $\hat{\theta}_{n,j-1}$.

We now focus on the asymptotic properties of the GMM estimator $\hat{\theta}_n$ obtained as a local solution to the minimization problem (11). In principle, we expect $\hat{\theta}_n$ to converge to any solution to the first order condition (8), which might turn out not to be the best optimal one. Indeed, the expected 'tick' loss in Definition 1 can possibly have multiple local minima, all of which satisfy the first order condition (8). Hence, we need to make sure that the value θ^* which solves $g_0(\theta^*) = 0$ does attain the lowest possible expected 'tick' loss. In other words, θ^* must be a global minimum in equation (7). A sufficient condition granting θ^* to be the best optimum weight is to impose

[4] In cases when the information vector fails to incorporate all the relevant information, condition $g_0(\theta^*) = 0$ is no longer equivalent to the first order condition (8) and $\{g(\theta^*; Y_{t+1}, W_t^*)\}$ is no longer a martingale difference sequence. However, S can still be consistently estimated by using some heteroskedasticity and autocorrelation robust estimator, like the Newey and West's (1987) estimator.

uniqueness, so that any local minimum is also a global one. This is established in the following lemma.

Lemma 3 (Uniqueness) *Let θ^* be a solution to the first order condition* (8)*. Suppose:*
(i) the conditional density of Y_{t+1}, $f_t(\cdot)$, is continuous and strictly positive;
(ii) for $i = 1, 2$: $\hat{q}_{im,t} \neq 0$, a.s. $- P$, and $Corr(\hat{q}_{1m,t}, \hat{q}_{2m,t}) \neq \pm 1$.
Then θ^* is unique.

Assumption (*ii*) is a fairly mild condition which rules out the possibility that the two sequences of forecasts are perfectly correlated, which would happen, e.g., if the two models were proportional or if they only differed by a constant. The consistency result for $\hat{\theta}_n$ is as follows.

Proposition 4 (Consistency) *Let the Assumptions of Proposition 3 hold and further assume:*
(iii) the sequence $\{(W_t^{*\prime}, X_t')'\}$ is strictly stationary and α-mixing with α of size $-r/(r-2)$, with $r > 2$;
(iv) the matrix $E[W_t^* W_t^{*\prime}]$ is nonsingular;
(v) there exist some $\delta > 0$ such that $E||W_t^*||^{2r+\delta} < \infty$;
Then $\hat{\theta}_n \xrightarrow{p} \theta^*$, as $n \to \infty$.

The restriction on the amount of heterogeneity and dependence in the data implied by assumption (*iii*) is in principle stronger than necessary, but we adopt it for convenience. Conditions (*iv*) and (*v*) are fairly standard and imply in particular that all the components of the information vector are not linearly dependent.

We now turn to the asymptotic distribution of $\hat{\theta}_n$. The standard asymptotic normality results for GMM require that the moment function $g_n(\theta)$ be once differentiable, which is not the case here. There are however asymptotic normality results for non-smooth functions and we will hereafter use the one proposed by Newey and McFadden (1994). The basic insight of their approach is that a smoothness condition on the moment function $g_n(\theta)$ can be replaced by the smoothness of its limit, which in the standard GMM case corresponds to the expectation $g_0(\theta)$, with the requirement that

certain remainder terms are small. The asymptotic normality of the GMM estimator $\hat{\theta}_n$ is given in the next proposition.

Proposition 5 (Asymptotic normality) *Let the Assumptions of Proposition 4 hold and further assume:*

(vi) $E||\hat{q}_{m,t}||^4 < \infty$*;*

(vii) the conditional density of Y_{t+1}*,* $f_t(\cdot)$*, is bounded;*

(viii) θ^* *is an interior point of* Θ*.*

Then the GMM estimator $\hat{\theta}_n$ *is asymptotically normal,* $\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{d} \mathcal{N}(0, (\gamma' S^{-1} \gamma)^{-1})$*, with*

$$\gamma \equiv -E[f_t(\theta^{*\prime} \hat{q}_{m,t}) W_t^* \hat{q}'_{m,t}], \tag{14}$$

and S *as defined in* (12)*.*

Comments:

1. Unlike the corresponding result in, e.g., West (2001), the asymptotic matrix is not affected by the presence of estimation uncertainty in the forecasts. From a technical point of view, this is a result of adopting a conditional approach (in particular, of writing the null hypothesis of encompassing in terms of forecasts depending on parameter estimates rather than on probability limits) and of considering asymptotics where only the out-of-sample size grows, whereas the in-sample size remains finite (see also the comments to Theorem 6 below).

2. Note that the expression of γ, which depends on the value of the conditional density f_t evaluated at the optimal combination of quantiles is similar to the one observed in the quantile regression literature (e.g., Koenker and Bassett, 1978).

3.2 Implementation of the CQFE Test

The asymptotic normality result derived above allows us to propose a conditional test of encompassing, which is a test on the coefficients of the optimal combination of quantile forecasts.

We consider conducting the two separate tests: $H_{10} : (\theta_1^*, \theta_2^*) = (1, 0)$ against $H_{1a} : (\theta_1^*, \theta_2^*) \neq (1, 0)$, and $H_{20} : (\theta_1^*, \theta_2^*) = (0, 1)$ against $H_{2a} : (\theta_1^*, \theta_2^*) \neq (0, 1)$, which respectively correspond to

testing whether forecast $\hat{q}_{1m,t}$ encompasses $\hat{q}_{2m,t}$ or whether $\hat{q}_{2m,t}$ encompasses $\hat{q}_{1m,t}$. We propose a Wald test of hypotheses H_{10} and H_{20} in the following theorem, which is the main result of this paper.

Theorem 6 (CQFE Test) *Let the Assumptions of Proposition 5 hold. Consider the test statistics*

$$ENC1_n = n((\hat{\theta}_{1n}, \hat{\theta}_{2n}) - (1,0))\hat{\Omega}^{-1}((\hat{\theta}_{1n}, \hat{\theta}_{2n}) - (1,0))' \quad (15)$$

$$ENC2_n = n((\hat{\theta}_{1n}, \hat{\theta}_{2n}) - (0,1))\hat{\Omega}^{-1}((\hat{\theta}_{1n}, \hat{\theta}_{2n}) - (0,1))', \quad (16)$$

where $(\hat{\theta}_{1n}, \hat{\theta}_{2n})$ *are defined in* (11) *and* $\hat{\Omega}$ *is a consistent estimator of the asymptotic variance* $\Omega \equiv (\gamma' S^{-1}\gamma)^{-1}$ *in* (5).

We then have:

(a) under H_{i0} : $EN\,\mathrm{Ci}_n \xrightarrow{d} \chi^2_2$ *, as* $n \to \infty$, $i = 1, 2$;

(b) under H_{ia} : $EN\,\mathrm{Ci}_n \to +\infty$ *, as* $n \to \infty$, $i = 1, 2$.

Comments:

1. Similarly to Giacomini and White (2003), the asymptotic distribution here is obtained for the number of out-of-sample observations going to infinity, whereas the in-sample size m remains finite. This is in contrast with the approach taken in the existing predictive ability literature (e.g., West 1996, McCracken 2000 etc.), which assumes that both the in-sample and the out-of-sample sizes grow.

2. A result of not letting the in-sample size grow is that the estimation uncertainty embedded in the forecasts does not vanish asymptotically. This is desirable for two main reasons. First, it allows us to focus on forecast, rather than model, evaluation; this means, for example, that our test could be used for assessing the impact on forecast accuracy of using different estimation methods for the same model. Second, it avoids the problems associated with comparison of predictive ability involving nested models. To see why, suppose that we were comparing nested models and that the smaller model were correctly specified. Letting m go to infinity would cause the parameter estimates to converge to their probability limits, which would render the forecasts from the two models asymptotically perfectly correlated, thereby invalidating assumption (ii).

The CQFE test is then implemented as follows. For a desired level of confidence, one first chooses the corresponding critical value c from the χ^2_2 distribution.[5] If $ENC1_n \leqslant c$ we conclude that $\hat{q}_{1m,t}$ encompasses $\hat{q}_{2m,t}$. If $ENC2_n \leqslant c$, we infer that $\hat{q}_{2m,t}$ encompasses $\hat{q}_{1m,t}$. If instead both $ENC1_n$ and $ENC2_n \geqslant c$, the final conclusion is that neither $\hat{q}_{1m,t}$ encompasses $\hat{q}_{2m,t}$, nor $\hat{q}_{2m,t}$ encompasses $\hat{q}_{1m,t}$, in which case the combination quantile $\hat{q}^c_{m,t} \equiv \hat{\theta}_{1n}\hat{q}_{1m,t} + \hat{\theta}_{2n}\hat{q}_{2m,t}$ significantly outperforms its components. The tests proposed have correct asymptotic size, as reflected by the fact that the test statistics have a distribution that is free of nuisance parameters under the null hypotheses and they are consistent, since the test statistics diverge under the alternative.[6]

In the computation of the encompassing test statistics $ENC1_n$ and $ENC2_n$, we propose letting $\hat{\Omega}^{-1} = \hat{\gamma}'\hat{S}^{-1}\hat{\gamma}$, where $\hat{S} \equiv \hat{S}(\hat{\theta}_n)$ is defined in equation (13) and $\hat{\gamma}$ is a numerical derivative estimator of γ in (14) whose jth column is given by

$$\hat{\gamma}_j \equiv [g_n(\hat{\theta}_n + e_j\varepsilon_n) - g_n(\hat{\theta}_n - e_j\varepsilon_n)]/2\varepsilon_n, \tag{17}$$

where e_j is the jth unit vector and ε_n is a small positive constant that depends on the sample size. Note that this particular estimator for γ, suggested by Komunjer (2002), does not require previous estimation of the conditional density f_t, which facilitates its practical computation. The magnitude of ε_n can be chosen so that $\hat{\gamma}$ is consistent: it suffices to consider $\varepsilon_n \to 0$ and $\sqrt{n}\varepsilon_n \to \infty$ (see e.g., Theorem 7.4 in Newey and McFadden 1994, p.2190). In applications one can set $\varepsilon_n = n^{-\delta}$, where δ should in principle be the largest possible, positive and less than 0.5 (as required for consistency of $\hat{\gamma}$), and such that the corresponding standard errors are sufficiently smooth. Choosing a value for δ is in practice a difficult problem and useful guidelines can be found in Newey and McFadden (1994, p. 2190). In the next section, we analyze the robustness of the small sample CQFE test's properties to different choices of δ.

Because of the non-differentiability of the GMM objective function, the maximization problem in (11) requires special attention. In principle, the optimization methods used to solve problems such

[5] The conditional encompassing test for quantile forecasts can be easily generalized to the comparison of k forecasts (or, more generally, k weights). In this case, the limiting distribution of the test statistic will be χ^2_k.

[6] Even though each encompassing test is correctly sized, the overall size might be overstated due to the sequential nature of the testing approach.

as the one in (11) can be sorted into two groups. The first group consists of gradient-based search methods (e.g. Newton-Raphson), which require that the objective function be sufficiently smooth. In the GMM case studied here, none of these methods is applicable.[7] The non-differentiability of the objective function is no longer a problem if one uses an optimization method which is not gradient-based. This second group of search methods regroups algorithms such as simulated annealing, used here, or genetic algorithm, used by Engle and Manganelli (1999). More details on the properties of the simulated annealing algorithm can be found in Goffe, Ferrier and Rogers (1994).

4 Monte Carlo evidence

In this section, we investigate the performance of our CQFE test in finite samples of sizes typically available to financial economists. The proposed evaluation is done along three dimensions: the size of the test, its power and its sensitivity to the choice of ε_n in the construction of $\hat{\gamma}$ above (17). We design our Monte Carlo experiment to match as closely as possible the problem of VaR evaluation and combination, which is the object of the empirical application in the following section. For simplicity, we restrict attention to the Conditional Autoregressive Value at Risk (CAViAR) family of VaR models proposed by Engle and Manganelli (1999). Our choices of models within the CAViAR family, as well as the parameter values used for the simulation are driven by the empirical application.

4.1 Size properties

We consider forecasts generated by the Asymmetric Absolute Value (AAV) CAViAR model,

$$VaR_{AAV,t+1} = \beta_0 + \beta_1 VaR_{AAV,t} + \beta_2 |r_t - \beta_3|, \tag{18}$$

and by the Symmetric Absolute Value (SAV) model,

$$VaR_{SAV,t+1} = \tilde{\beta}_0 + \tilde{\beta}_1 VaR_{SAV,t} + \tilde{\beta}_2 |r_t|, \tag{19}$$

[7] Note that the optimization approach taken in the case of Koenker and Bassett's (1978) quantile regression is based on a linear programming representation of the objective function. Here, however, the GMM objective function is quadratic and the duality theorem traditionally employed no longer holds.

where $VaR_{AAV,t+1}$ and $VaR_{SAV,t+1}$ are forecasts of the conditional α-quantile of $-r_{t+1}$. Our null hypothesis is that the AAV model encompasses the SAV model. To generate data that supports the null hypothesis, we proceed as follows: We first fix the values of the true parameters $(\beta_0, \beta_1, \beta_2, \beta_3)$ and $(\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\beta}_2)$ in (18) and (19), respectively, and replicate $(VaR_{AAV,1}, \ldots, VaR_{AAV,n})$ and $(VaR_{SAV,1}, \ldots, VaR_{SAV,n})$ by assuming that $r_t \sim i.i.d.\mathcal{N}(0, \sigma^2)$ with $\sigma = 0.1$. In this particular case, the in-sample size m is zero and $T = n$. Accordingly, all inference is done conditional on the set of true parameter values $(\beta_0, \beta_1, \beta_2, \beta_3)$ and $(\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\beta}_2)$. Next, we constrain $VaR_{AAV,t+1}$ to be the conditional α-quantile of $-r_{t+1}$ by redefining the original series. For every t, $t = 0, \ldots, n-1$, we let the Data-Generating Process (DGP) be

$$r_{t+1} = -VaR_{AAV,t+1} + u_{t+1}, \tag{20}$$

with $u_{t+1} \sim i.i.d.\mathcal{N}(-\sigma\Phi^{-1}(\alpha), \sigma^2)$, $\sigma = 0.1$, where Φ is the distribution function of a standard normal random variable. By restricting u_{t+1} to have the α-quantile of zero we ensure that the AAV model in (18) produces forecasts of the true conditional α-quantile of $-r_{t+1}$.

The parameter values $(\beta_0, \beta_1, \beta_2, \beta_3) = (0, 0.8, 0.3, 1)$ in (18) and $(\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\beta}_2) = (0, 0.9, 0.2)$ in (19) are chosen so as to match the estimates obtained in the empirical application for $\alpha = 5\%$.

We consider a range of values for the out-of-sample size n and the step size ε_n in (17): $n = (1000, 2500, 5000)$ and $\varepsilon_n = n^{-\delta}$ where δ ranges from 0.4 to 0.5 with increments of 0.01.[8] For each sample size n we generate 500 Monte Carlo replications of the time series $\{r_t\}_{t=1}^n$, $\{VaR_{AAV,t}\}_{t=1}^n$ and $\{VaR_{SAV,t}\}_{t=1}^n$ each of length n.[9] We then consider the forecast combination $(\theta_0 + \theta_{AAV} \cdot VaR_{AAV,t} + \theta_{SAV} \cdot VaR_{SAV,t})$ and construct the GMM estimator $(\hat{\theta}_{0n}, \hat{\theta}_{AAVn}, \hat{\theta}_{SAVn})'$ of the optimal weight vector $(\theta_0^*, \theta_{AAV}^*, \theta_{SAV}^*)'$ according to the procedure described in Section 3. Note that we include a constant term in the forecast combination, thus allowing the empirical coverage of the original forecasts to be different from the 5% nominal value. In our particular case, the AAV forecasts will display correct empirical coverage by construction, whereas the forecasts from the mis-

[8] For smaller sample sizes n the simulated annealing has low convergence rate, hence we chose not to report the corresponding results.

[9] In reality, we generate series of length $n + 50$ and discard the first 50 data points in order to avoid problems linked to the choice of starting values.

specified SAV model will in general be biased. Finally, we compute the proportion of rejections, at the 5% nominal level, of the null hypothesis $H_{10} : (\theta^*_{AAV}, \theta^*_{SAV}) = (1, 0)$. The test statistic $ENC1_n$ is similar to the one in Theorem 6, where we have modified $\hat{\Omega}$ into $R \cdot \hat{\Omega} \cdot R'$ so as to reflect the appropriate parameter restrictions.[10] The information vector W^*_t is $W^*_t \equiv (1, r_t, VaR_{AAV,t}, VaR_{SAV,t})'$. The results are collected in Table 1.

Table 1 : Empirical size of nominal .05 test

	δ										
n	.40	.41	.42	.43	.44	.45	.46	.47	.48	.49	.50
1000	.075	.092	.068	.088	.092	.078	.078	.092	.095	.105	.139
2500	.051	.046	.032	.037	.060	.046	.051	.069	.051	.074	.088
5000	.057	.050	.028	.050	.043	.050	.043	.090	.086	.107	.107

NOTE: Empirical size of the CQFE test for a nominal size of .05. Entries represent the rejection frequencies over 500 Monte Carlo replications of the null hypothesis that forecasts from the AAV CAViAR model encompass forecasts from the SAV CAViAR model when the DGP is the AAV CAViAR. *n* is the sample size and δ is a user-defined constant required in the computation of the numerical derivative estimator in equation (17).

The test appears to be well-sized, with a moderate tendency to over-reject for the smallest sample size $n = 1000$ or for δ close to 0.5.

4.2 Power properties

In order to generate data under the alternative hypothesis of no encompassing of AAV forecasts with respect to SAV forecasts, we first replicate $(VaR_{AAV,1}, \ldots, VaR_{AAV,n})$ and $(VaR_{SAV,1}, \ldots, VaR_{SAV,n})$ for parameter values $(\beta_0, \beta_1, \beta_2, \beta_3) = (0, 0.8, 0.3, 1)$ and $(\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\beta}_2) = (0, 0.9, 0.2)$, respectively, following the procedure described in the previous section and then let the DGP be

$$r_{t+1} = -[\rho VaR_{SAV,t+1} + (1 - \rho) VaR_{AAV,t+1}] + u_{t+1}, \tag{21}$$

where $0 < \rho < 1$ and $u_{t+1} \sim i.i.d.\mathcal{N}(-\sigma\Phi^{-1}(\alpha), \sigma^2)$, $\sigma = 0.1$, as in the previous section. Note that the size study is obtained when the data are generated according to (21) with $\rho = 0$. Accordingly, increasing ρ towards 0.5 allows us to obtain the power curve for the CQFE test. We consider a number of different values for ρ, ranging from $\rho = 0.1$ to $\rho = 0.5$, at increments of 0.1. We keep the step size fixed at $\varepsilon_n = n^{-.45}$. For each parameterization, we generate 500 Monte Carlo replications of

[10] In this particular case, we have $R = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$.

the time series $\{r_t\}_{t=1}^n$, $\{VaR_{AAV,t}\}_{t=1}^n$ and $\{VaR_{SAV,t}\}_{t=1}^n$ and proceed as previously by computing the proportion of rejections of the null hypothesis that $VaR_{AAV,t+1}$ encompasses $VaR_{SAV,t+1}$ at the 5% nominal level. Figure 1 plots the power curves for $n = (1000, 2500, 5000)$.

The test displays fairly good power properties. For example, when the true VaR is an equal-weighted average of AAV and SAV VaRs, the test rejects the null of encompassing at least 85% of the time. As expected, the power increases (although by small amounts) as the size n of the out-of-sample evaluation data set increases.

5 Empirical Evaluation and Combination of VaR forecasts

We illustrate the potential usefulness of our CQFE test by applying it to the problem of VaR evaluation. The importance of VaR has become institutional in August 1996, when US bank regulators adopted a 'market risk' supplement to the Basle Accord of 1988. VaR has thus become a risk-measure for setting capital-adequacy standards of US commercial banks. The data used in our empirical application consist of 16 years of daily returns on the S&P500 index (source: Datastream), from September 1985 to September 2001 ($T = 4176$ observations). The first third of the sample, corresponding to the period from September 1985 to January 1991 ($m = 1392$ observations) is used as the in-sample period, while the remaining two thirds ($n = 2784$ observations) are reserved to evaluate the out-of-sample performance. We adopt a fixed forecasting scheme, which means that all forecasts depend on the same set of parameters estimated over the first m observations. We consider a portfolio consisting of a long position in the index, with an investment horizon of 1 day.

5.1 VaR models

For the purposes of this empirical application we consider the 5% and 1% VaR forecasts originated from four different models: $VaR_{1,t+1}$ and $VaR_{2,t+1}$ are VaR forecasts based on conditional heteroskedasticity models, $r_{t+1}|\mathcal{F}_t \sim \mathcal{D}(0, \sigma_{t+1}^2)$ with $\mathcal{D}$ belonging to a location-scale family of distributions. In this case, VaR is a linear function of the conditional volatility of the returns σ_{t+1} and different VaR models correspond to different specifications for the conditional variance σ_{t+1}^2. Two such specifications are the commonly used GARCH(1,1) model in which $\sigma_{1,t+1}^2 = \omega_0 + \omega_1 \sigma_{1,t}^2 + \omega_2 r_t^2$,

and the JP Morgan's (1996) RiskMetrics model where the variance is obtained as an exponential filter $\sigma^2_{2,t+1} = \lambda\sigma^2_{2,t} + (1-\lambda)r^2_t$, with $\lambda = 0.94$ for daily data. In both cases, the corresponding VaR model is

$$VaR_{i,t+1} = \beta_0 + \beta_1\sigma_{i,t+1}, i = 1, 2. \tag{22}$$

Models such as (22) above have been studied by Christoffersen *et al.* (2001), among others. Hereafter, we refer to $VaR_{1,t+1}$ as GARCH VaR and to $VaR_{2,t+1}$ as RiskMetrics VaR.

A different approach to VaR modeling and estimation is taken by Engle and Manganelli (1999). Here we consider two examples of the CAViAR model proposed by these authors: $VaR_{3,t+1}$ is a forecast based on an Asymmetric Absolute Value (AAV) model

$$VaR_{3,t+1} = \beta_0 + \beta_1 VaR_{3,t} + \beta_2|r_t - \beta_3|, \tag{23}$$

while $VaR_{4,t+1}$ is based on an Asymmetric Slope (AS) model,

$$VaR_{4,t+1} = \beta_0 + \beta_1 VaR_{4,t} + \beta_2 r^+_t + \beta_3 r^-_t, \tag{24}$$

where r^+_t and r^-_t correspond to the positive and the negative part of r_t respectively.[11] Figures 2 and 3 show the out-of-sample sequences of VaR forecasts generated by the above models, together with the sequences of VaR violations.

For each of the four VaR models (22)-(24) we first construct an estimator $\hat{\beta}_m \equiv \hat{\beta}_{m,m}$ of the unknown parameter vector β in sample, i.e. by using the first $m = 1392$ observations. This estimator is then used to form out-of-sample VaR forecasts according to a fixed forecasting scheme. In other words, at each out-of-sample date t, $m \leqslant t \leqslant T-1$, we compute one-step-ahead VaR forecasts $VaR_{i,t+1}$, $i = 1, 2, 3, 4$, based on the four models (22)-(24). The computation is done

[11] The three models $VaR_{1,t+1}$, $VaR_{3,t+1}$ and $VaR_{4,t+1}$ are chosen on the basis of their individual performance in modeling the VaR for the S&P500 index. As shown by Christoffersen, Hahn and Inoue (2001), the GARCH VaR $VaR_{1,t+1}$ is the only VaR measure, among several alternatives considered by the authors, which passes the Christoffersen's (1998) 'conditional coverage test' for both 5% and 1% coverage rates. Similarly, Engle and Manganelli (1999) show that the Asymmetric Absolute Value model $VaR_{3,t+1}$ and the Asymmetric Slope model $VaR_{4,t+1}$ are the best CAViAR specifications for the S&P500 according to a criterion they propose (see Engle and Manganelli, 1999, for details). Finally, the JP Morgan's (1996) RiskMetrics model $VaR_{2,t+1}$ is chosen as a benchmark model commonly used by practitioners.

recursively, meaning that for each $i = 1, 2, 3, 4$, the value of $VaR_{i,t+1}$ depends on the past forecast $VaR_{i,t}$ ($\sigma^2_{i,t}$ in the case of models (22)) and on the out-of-sample realization r_t (respectively r^2_t). For illustration, we report the parameter estimates $\hat{\beta}_m$ in Table 2. Alternatively, one can consider sequences of VaR forecasts provided by different groups of outside researchers/analysts, without knowing the underlying forecasting models, as long as the latter satisfy the assumptions discussed in previous sections.

Table 2 : VaR parameter estimates

$\alpha = 0.01$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\alpha = 0.05$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
Model									
GARCH	0.982	1.597	-	-	GARCH	0.055	1.446	-	-
	(0.048)	(0.033)				(0.075)	(0.095)		
RiskMetrics	0.959	1.698	-	-	RiskMetrics	0.500	1.039	-	-
	(0.104)	(0.125)				(0.104)	(0.137)		
AAV	0.213	0.714	0.761	0.422	AAV	-0.074	0.804	0.328	1.070
	(0.020)	(0.040)	(0.068)	(0.026)		(0.008)	(0.016)	(0.015)	(0.065)
AS	0.460	0.716	0.110	-0.796	AS	0.120	0.834	0.025	-0.404
	(0.029)	(0.061)	(0.011)	(0.081)		(0.048)	(0.058)	(0.006)	(0.123)

NOTE: Parameter estimates for different VaR models. Data: Datastream daily returns on S&P500 from September 1985 to January 1991 ($m = 1392$ observations). The estimation is carried out by GMM in the GARCH and RiskMetrics VaR models and by QML in the CAViAR models. For VaR models where $r_{t+1}|F_t \sim N(0, \sigma^2_{t+1})$ with (1) GARCH volatility $\sigma^2_{t+1} = \omega_0 + \omega_1 \sigma^2_t + \omega_2 r^2_t$ we have $\omega_0 = 0.117$, $\omega_1 = 0.763$ and $\omega_2 = 0.150$, and (2) RiskMetrics volatility $\sigma^2_{t+1} = \lambda \sigma^2_t + (1-\lambda) r^2_t$ we take $\lambda = 0.94$.

As a quick check of the out-of-sample performance of individual VaR models and their equally weighted pairwise combinations ($0.5 \cdot VaR_{i,t+1} + 0.5 \cdot VaR_{j,t+1}$), we compute the empirical coverage a of the corresponding sequence of forecasts, $a \equiv n^{-1} \sum_{t=1}^{n} I_{t+1}$, where I_{t+1} denotes the 'hit' variable $I_{t+1} \equiv 1(Y_{t+1} - VaR_{t+1} < 0)$. If the VaR model under consideration performs well then we expect that it will display correct unconditional coverage, attained when the empirical coverage a equals the nominal coverage α.[12] The out-of-sample empirical coverages are reported in Table 3.

[12] Note that one could devise a simple likelihood ratio test of the null hypothesis that I_{t+1} is Bernoulli(α), which is the main principle of the so-called 'unconditional coverage' test, discussed, among others, by Christoffersen (1998). This test, however, assumes away parameter estimation uncertainty, and thus we decided not to report its results here.

Table 3 : Out-of-sample empirical coverage

$\alpha = 1\%$					
Models		GARCH	RiskMetrics	AAV	AS
	GARCH	0.853%	0.742%	0.705%	0.742%
	RiskMetrics	-	0.705%	0.705%	0.631%
	AAV	-	-	0.742%	0.668%
	AS	-	-	-	0.631%

$\alpha = 5\%$					
Models		GARCH	RiskMetrics	AAV	AS
	GARCH	4.674%	4.711%	4.191%	4.191%
	RiskMetrics	-	4.970%	4.228%	4.191%
	AAV	-	-	4.303%	4.303%
	AS	-	-	-	4.228%

NOTE: Empirical coverage $a = n^{-1}\sum I_{t+1}$ for individual VaR models (diagonal elements) and their equally weighted pairwise combinations (off-diagonal elements). Data: Datastream daily returns on S&P500 from January 1991 to September 2001 ($n = 2784$ observations).

Based on the results from Table 3, we can compare VaR models in terms of the difference between their out-of-sample empirical coverage a and the nominal coverage α. For $\alpha = 1\%$, the best model is GARCH(1,1) with empirical coverage 0.853%, followed by three equally performing models with coverage 0.742%: AAV and equally weighted combinations of GARCH with RiskMetrics and AS. For $\alpha = 5\%$, the best empirical coverage (4.970%) is that of RiskMetrics, followed by an equally weighted combination of RiskMetrics and GARCH (4.711%), and GARCH alone (4.674%). It is interesting to note that, in general, the unconditional coverage of equally weighted combinations $(0.5 \cdot VaR_{i,t+1} + 0.5 \cdot VaR_{j,t+1})$ never outperforms both individual series $VaR_{i,t+1}$ and $VaR_{j,t+1}$.

In order to assess the relative performance of the two models with best empirical coverages, as identified above, we perform our CQFE test. Specifically, we test (1) whether at $\alpha = 1\%$ level, GARCH encompasses AAV, and (2) if at $\alpha = 5\%$ level, RiskMetrics encompasses GARCH. [13]

[13] Before applying the CQFE test, we verified that the sequences of forecasts are not perfectly correlated. The out-of-sample correlation coefficients were .91 for case (1) and .86 for case (2), which made us conclude that assumption (*ii*) is not violated.

5.2 CQFE Test Results

We estimate the optimal combination weights $(\theta_0^*, \theta_i^*, \theta_j^*)'$ in the forecast combination $\theta_0 + \theta_i VaR_{i,t} + \theta_j VaR_{j,t}$ by using the GMM approach described in Section 3. The estimation of the combination weights crucially depends on the choice of the information vector W_t^*. As a general rule, W_t^* should include variables that are part of the time-t information set which are thought to help forecast the conditional quantile of Y_{t+1}. Examples of variables that might be relevant are, e.g., lags of the variable, lags of the quantile forecasts and/or non-linear functions of the above. Also, lags of 'hit' variables $I_{i,t+1}$ can be included, to take into account possible persistence in the variable that indicates whether a 'violation' of the quantile forecast occurred at the given time. The appropriate information vector may also vary with the type of quantile considered. For example, for values of α that approach .5 (corresponding to the conditional median), it is plausible to think that the quantile forecast would be affected by variables that are typically used in conditional mean forecasting. On the other hand, for small values of α, as in the case of our application, the dynamics of the conditional quantile more likely mimic those of the conditional volatility. The vast literature on volatility forecasting can in this case help guide the choice of the relevant variables - e.g., lagged squared returns - to be included in the instrument set. For the purposes of this empirical application, we let $W_t^* \equiv (1, r_t, VaR_{i,t}, VaR_{j,t})'$.

We report the estimated combination weights $\hat{\theta}_{0n}$, $\hat{\theta}_{in}$ and $\hat{\theta}_{jn}$ together with their standard errors in Table 4. It is important to note that the computation of standard errors is based on the numerical derivative $\hat{\gamma}$ given in equation (17), which in turn depends on the size of the step ε_n. In our application we set $\varepsilon_n = n^{-\delta}$ with $\delta = 0.45$ for all weights. Table 4 also contains the values of the test statistics ENC_{in} and ENC_{jn}.

Table 4 : Conditional Quantile Forecast Encompassing Test for VaR measures

Model	$\hat{\theta}_{0n}$	$\hat{\theta}_{in}$	$\hat{\theta}_{jn}$	J	ENC_{in}	ENC_{jn}
$\alpha = 0.01$						
GARCH *(i)* vs AAV *(j)*	-1.506 (0.186)	1.048 (0.193)	0.382 (0.094)	8.636	26.484*	46.982*
$\alpha = 0.05$						
RiskMetrics *(i)* vs GARCH *(j)*	-0.118 (0.097)	0.005 (0.073)	0.565 (0.057)	10.545	188.232*	116.277*

NOTE: Out of sample CQFE test for VaR measures for a portfolio composed of a long position in S&P500 index with an investment horizon of 1 day. Data: Datastream daily returns on S&P500 from January 1991 to September 2001 (n =*2784* observations). The consistent standard errors of the GMM estimator (θ_{0n}, θ_{in}, θ_{jn})' were computed with δ = *0.45* and are reported in parentheses. J is the value of the J-test statistics: $J = g_n(\theta_n)'S^{-1}g_n(\theta_n)$. The marked (*) values of the CQFE test statistics ENC_{in} and ENC_{jn} are significant at the 1% level.

As can be seen from Table 4, neither forecast encompasses its competitor, for both levels of α. This implies that the forecast combination in both cases outperforms the individual forecasts. However, note that for $\alpha = 5\%$ the weight on the RiskMetrics forecast is not significantly different from zero (t-stat $= 0.068$), which suggests that the optimal combination is in this case simply the bias-corrected GARCH forecast.

6 Conclusion

In this paper we propose a Conditional Quantile Forecast Encompassing (CQFE) test for comparing alternative conditional quantile forecasts in an out-of-sample framework. We base our evaluation on the concept of encompassing, which requires that a forecast be able to explain the predictive ability of a rival forecast. The CQFE test can thus be viewed as a test for superior predictive ability. The setup proposed in this paper also allows us to discuss the benefit of forecast combination for quantile forecasts, which becomes relevant when the encompassing tests indicate that neither forecast outperforms its competitor.

The key features of our approach are: (1) the use of the 'tick' loss function rather than the quadratic loss function in the definition of encompassing and (2) a conditional, rather than unconditional, approach to out-of-sample evaluation. Some of the benefits of the conditional approach are that it allows comparison of forecasts based on both nested and non-nested models and of forecasts produced by general estimation procedures.

The implementation of the CQFE test is done by using a fairly standard GMM estimation technique, with optimization procedure appropriately modified to accommodate our non-differentiable criterion function. The CQFE test displays good size and power properties for samples of sizes typically available in financial applications.

We apply the CQFE test to the problem of conditional VaR forecast evaluation using S&P500 daily index returns. At 1% level, we show that a forecast combination (with intercept) of GARCH and AAV CAViAR forecasts outperforms both individual components. A similar result holds at 5% level, where we compare VaR forecasts generated from RiskMetrics and GARCH models. In the latter case, however, we find that the combination weight on the RiskMetrics forecast is not significantly different from zero, indicating that bias-corrected GARCH forecasts for the 5% VaR encompass RiskMetrics forecasts.

Appendix: Proofs

Notation:

if V is a real n-vector, $V \equiv (V_1, \ldots, V_n)'$, then $||V||$ denotes the L_2-norm of V, i.e. $||V||^2 \equiv V'V = \sum_{i=1}^n V_i^2$. If M is a real $n \times n$-matrix, $M \equiv (M_{ij})_{1 \leqslant i,j \leqslant n}$, then $||M||$ denotes the L_∞-norm of M, i.e. $||M|| \equiv \max_{1 \leqslant i,j \leqslant n} |M_{ij}|$.

Proof of Proposition 2. We derive the set of first order conditions corresponding to the minimization problem (7)

$$\theta^* \equiv \arg\min_{\theta \in \Theta} E_t[(\alpha - 1(Y_{t+1} - \theta' \hat{q}_{m,t} < 0))(Y_{t+1} - \theta' \hat{q}_{m,t})] = \arg\min_{\theta \in \Theta} \Sigma(\theta)$$

where $\theta \equiv (\theta_1, \theta_2)' \in \Theta$, Θ being a compact subset of $\mathbb{R}^2$ and $\hat{q}_{m,t} \equiv (\hat{q}_{1m,t}, \hat{q}_{2m,t})'$. We consider

$$\begin{aligned}
\Sigma(\theta) &= E_t[(\alpha - 1(Y_{t+1} - \theta' \hat{q}_{m,t} < 0))(Y_{t+1} - \theta' \hat{q}_{m,t})] \\
&= \int_{\mathbb{R}} (\alpha - 1(Y_{t+1} - \theta' \hat{q}_{m,t} < 0))(y_{t+1} - \theta' \hat{q}_{m,t}) dF_t(y_{t+1}) \\
&= \int_{\mathbb{R}} \alpha (y_{t+1} - \theta' \hat{q}_{m,t}) dF_t(y_{t+1}) - \int_{\mathbb{R}} 1(Y_{t+1} - \theta' \hat{q}_{m,t} < 0)(y_{t+1} - \theta' \hat{q}_{m,t}) dF_t(y_{t+1}) \\
&= \int_{-\infty}^{+\infty} \alpha (y_{t+1} - \theta' \hat{q}_{m,t}) dF_t(y_{t+1}) - \int_{-\infty}^{\theta' \hat{q}_{m,t}} (y_{t+1} - \theta' \hat{q}_{m,t}) dF_t(y_{t+1}) \\
&= \int_{-\infty}^{+\infty} \alpha (y_{t+1} - \theta' \hat{q}_{m,t}) dF_t(y_{t+1}) - \int_{-\infty}^{0} x_{t+1} dF_t(x_{t+1} + \theta' \hat{q}_{m,t}),
\end{aligned}$$

where we have defined $x_{t+1} \equiv y_{t+1} - \theta_t' \hat{q}_{m,t}$. Thus

$$\nabla \Sigma(\theta) = -\alpha \hat{q}_{m,t} - \int_{-\infty}^{0} \hat{q}_{m,t} x_{t+1} f_t(x_{t+1} + \theta' \hat{q}_{m,t}) dx_{t+1},$$

since we assume that the random variable Y_{t+1} has a continuously differentiable conditional density f_t, i.e. $dF_t(y_{t+1}) = f_t(y_{t+1}) dy_{t+1}$ and f_t continuous. By arranging the previous equality we obtain

$$\begin{aligned}
\nabla \Sigma(\theta) &= -\alpha \hat{q}_{m,t} \\
&\quad -[\hat{q}_{m,t} x_{t+1} f_t(x_{t+1} + \theta' \hat{q}_{m,t})]_{-\infty}^{0} + \int_{-\infty}^{0} \hat{q}_{m,t} f_t(x_{t+1} + \theta' \hat{q}_{m,t}) dx_{t+1},
\end{aligned}$$

so that

$$\nabla \Sigma(\theta) = -\alpha \hat{q}_{m,t} + \hat{q}_{m,t} \int_{-\infty}^{\theta' \hat{q}_{m,t}} f_t(y_{t+1}) dy_{t+1}.$$

We can then write

$$\nabla\Sigma(\theta) = -E_t[(\alpha - 1(Y_{t+1} - \theta'\hat{q}_{m,t} < 0))\hat{q}_{m,t}].$$

If θ^* is a solution to the initial minimization problem then $\nabla\Sigma(\theta)|_{\theta^*} = 0$, $a.s. - P$, i.e.

$$E_t[(\alpha - 1(Y_{t+1} - \theta'\hat{q}_{m,t} < 0))\hat{q}_{m,t}] = 0, a.s. - P.$$

The variable $\hat{q}_{m,t}$ being measurable with respect to the information set $\mathcal{F}_t$, we can rewrite the previous equation as

$$E_t[\alpha - 1(Y_{t+1} - \theta'\hat{q}_{m,t} < 0)] = 0, a.s. - P,$$

which completes the proof of Proposition 2. ■

Lemma 7 *For all t, if $Corr(\hat{q}_{1m,t}, \hat{q}_{2m,t}) \neq \pm 1$ and $\hat{q}_{im,t} \neq 0$, $a.s. - P$ for $i = 1, 2$ then $\hat{q}_{1m,t}$ and $\hat{q}_{2m,t}$ are linearly independent, i.e., $\gamma_1\hat{q}_{1m,t} + \gamma_2\hat{q}_{2m,t} = 0, a.s. - P$ implies $\gamma_1 = \gamma_2 = 0$.*

Proof of Lemma 7. By contradiction, suppose there exist $(\gamma_1, \gamma_2) \neq (0, 0)$ such that $\gamma_1\hat{q}_{1m,t} + \gamma_2\hat{q}_{2m,t} = 0$, $a.s. - P$ Without loss of generality, suppose $\gamma_1 \neq 0$. Then $\hat{q}_{1m,t} = -(\gamma_2/\gamma_1)\hat{q}_{2m,t}$, $a.s. - P$, from which it follows that either (1) $\gamma_2 = 0$, which implies that $\hat{q}_{1m,t} = 0$, $a.s. - P$ or (2) $\gamma_2 \neq 0$, which implies that $Corr(\hat{q}_{1m,t}, \hat{q}_{2m,t}) = sgn(-(\gamma_2/\gamma_1)) = \pm 1$. This completes the proof of Lemma 7. ■

Proof of Proposition 3. We show that if θ^* and $\overline{\theta}$ both satisfy the first order condition in (8), then $\theta^* = \overline{\theta}$, i.e.

$$0 = E_t[\alpha - 1(Y_{t+1} - \theta^{*\prime}\hat{q}_{m,t} < 0)] = E_t[\alpha - 1(Y_{t+1} - \overline{\theta}'\hat{q}_{m,t} < 0)],\ a.s. - P \Rightarrow \theta^* = \overline{\theta}.$$

Let W_t denote an element of the information set $\mathcal{F}_t$. Then, the previous statement is equivalent to

$$(\forall W_t \in \mathcal{F}_t, 0 = E[(\alpha - 1(Y_{t+1} - \theta^{*\prime}\hat{q}_{m,t} < 0))W_t] = E[(\alpha - 1(Y_{t+1} - \overline{\theta}'\hat{q}_{m,t} < 0))W_t]) \Rightarrow \theta^* = \overline{\theta}.$$

Consider the difference $\Delta(W_t)$, defined by

$$\Delta(W_t) \equiv E[(\alpha - 1(Y_{t+1} - \theta^{*\prime}\hat{q}_{m,t} < 0))W_t] - E[(\alpha - 1(Y_{t+1} - \overline{\theta}'\hat{q}_{m,t} < 0))W_t].$$

We have

$$\begin{aligned}\Delta(W_t) &= E[W_t(1(Y_{t+1} - \overline{\theta}'\hat{q}_{m,t} < 0) - 1(Y_{t+1} - \theta^{*\prime}\hat{q}_{m,t} < 0))] \\ &= E[W_t(1(\theta^{*\prime}\hat{q}_{m,t} < Y_{t+1} < \overline{\theta}'\hat{q}_{m,t}) - 1(\overline{\theta}'\hat{q}_{m,t} < Y_{t+1} < \theta^{*\prime}\hat{q}_{m,t}))] \\ &= E[W_t E_t[1(\theta^{*\prime}\hat{q}_{m,t} < Y_{t+1} < \overline{\theta}'\hat{q}_{m,t}) - 1(\overline{\theta}'\hat{q}_{m,t} < Y_{t+1} < \theta^{*\prime}\hat{q}_{m,t})]],\end{aligned}$$

since W_t is $\mathcal{F}_t$-measurable. The conditional expectation on the right hand side of the previous equality is in turn equal to

$$\int_{\theta^{*\prime}\hat{q}_{m,t}}^{\overline{\theta}'\hat{q}_{m,t}} f_t(y_{t+1})dy_{t+1} \equiv D_t(\overline{\theta}, \theta^*),$$

where $f_t(\cdot)$ is the conditional density of Y_{t+1}. Thus $\Delta(W_t) = E[W_t D_t(\overline{\theta}, \theta^*)]$ and we have

$$(\forall W_t \in \mathcal{F}_t, \Delta(W_t) = 0) \Rightarrow D_t(\overline{\theta}, \theta^*) = 0, \ a.s. - P.$$

By assumption (i), the conditional density of Y_{t+1}, $f_t(\cdot)$, is continuous and strictly positive on $\mathbb{R}$, so that $D_t(\overline{\theta}, \theta^*)$ can only be almost surely zero when $\overline{\theta}'\hat{q}_{m,t} = \theta^{*\prime}\hat{q}_{m,t}$, $a.s. - P$, i.e., $(\overline{\theta} - \theta^*)'\hat{q}_{m,t} = 0$, $a.s. - P$. From Lemma 7, this implies that $(\overline{\theta} - \theta^*) = 0$. In conclusion, we have that

$$(\forall W_t \in \mathcal{F}_t, \Delta(W_t) = 0) \Rightarrow \theta^* = \overline{\theta},$$

which completes the proof of Proposition 3. ■

Proof of Proposition 4. We first discuss the nature of the sequence $\{g(\theta; Y_{t+1}, W_t^*)\}$. The moment function $g(\theta; Y_{t+1}, W_t^*)$ depends on the data through Y_{t+1}, W_t^* and $\hat{q}_{m,t}$. Let us consider separately the two cases of (1) fixed forecasting scheme and (2) rolling window forecasting scheme. (1) If a fixed forecasting scheme is used, the forecasts $\hat{q}_{m,t}$, $t = m, ..., T-1$ depend, on the one hand, on pre-determined parameter estimates $\hat{\beta}_{m,m}$ hence on the variables $(X_1, \ldots, X_m)$, and on the other hand, on some set of right hand variables of the forecasting model which are observed at time t. Typically, those variables are going to be included in the vector W_t^*. Therefore, by letting $V_{t+1} \equiv (Y_{t+1}, W_t^*, X_1, \ldots, X_m)'$, for every t, $t = m, ..., T-1$, we can rewrite $g(\theta; Y_{t+1}, W_t^*)$ as $g(\theta; V_{t+1})$.

(2) If instead a rolling window forecasting scheme is used, the vector of forecasts $\hat{q}_{m,t}$, $t = m, ..., T-1$

is a constant measurable function of the estimation window which consists of the m most recent observations of X_t. In that case, we can again let $V_{t+1} \equiv (Y_{t+1}, W_t^*, X_t, \ldots, X_{t-m+1})'$, for every t, $t = m, ..., T-1$, and rewrite $g(\theta; Y_{t+1}, W_t^*)$ as $g(\theta; V_{t+1})$.

Since for every t, $t = m, ..., T-1$, V_{t+1} is function of a finite number $(m+2)$ of variables which, by assumption (iii), are strictly stationary and α-mixing, the sequence $\{V_t\}$ is strictly stationary and α-mixing of same size (see, e.g., Theorem 3.49 of White 2001). Note that strict stationarity and α-mixing of $\{V_t\}$ imply ergodicity (see, e.g., Theorem 3.44 in White 2001), so that we can use one of the standard results on the consistency of GMM estimators for stationary and ergodic sequences. Specifically, we verify that the conditions of Theorem 2.6 of Newey and McFadden (1994, pp. 2132-2133) are satisfied in our case (note that the results of Theorem 2.6 hold if the iid assumption is replaced with the condition that $\{V_t\}$ is strictly stationary and ergodic).

First, we need to show that $\hat{S}(\hat{\theta}_n) \xrightarrow{p} S$ where S is the asymptotic covariance matrix defined in equation (12). Recall from equation (13) that $\hat{S}(\theta) \equiv n^{-1} \sum_{t=m}^{T-1} g(\theta; v_{t+1}) g(\theta; v_{t+1})'$, where v_{t+1} is a realization of V_{t+1} defined above. Note that the moment function g is an $\mathcal{F}_{t+1}$-measurable function of $\{V_{t+1}\}$ which is strictly stationary and α-mixing. By using, once again., Theorem 3.49 of White (2001), we can then say that $\{g(\theta; V_{t+1})\}$ and $\{g(\theta; V_{t+1}) g(\theta; V_{t+1})'\}$ are strictly stationary and α-mixing of same size. Hence, we can apply a law of large numbers (LLN) for α-mixing sequences to show that for every $\theta \in \Theta$, $\hat{S}(\theta)$ converges to $\tilde{S}(\theta) \equiv E[g(\theta; V_{t+1}) g(\theta; V_{t+1})']$. Specifically, we check that all the assumptions of Corollary 3.48 in White (2001) hold: first, note that for $r > 2$, we have $-r/(r-1) > -r/(r-2)$ so that the sequence $\{g(\theta; V_{t+1}) g(\theta; V_{t+1})'\}$ is moreover α-mixing with α of size $-r/(r-1)$. We now need to show that for some $\tilde{\delta} > 0$ we have $E||g(\theta; V_{t+1}) g(\theta; V_{t+1})'||^{r+\tilde{\delta}} < \infty$: recall from equation (9) that we have

$$\begin{aligned} ||g(\theta; V_{t+1}) g(\theta; V_{t+1})'|| &= [\alpha - 1(Y_{t+1} - \theta' \hat{q}_{m,t} < 0)]^2 ||W_t^* W_t^{*\prime}|| \\ &\leqslant ||W_t^* W_t^{*\prime}||, a.s. - P. \end{aligned}$$

Moreover, we know, by norm equivalence, that there exist some positive constant c such that

$$||W_t^* W_t^{*\prime}|| = |W_{t,i_0}^* \cdot W_{t,j_0}^*| \leqslant |W_{t,i_0}^*| \cdot |W_{t,j_0}^*| \leqslant c^2 \cdot ||W_t^*||^2, a.s. - P,$$

where i_0 and j_0, $1 \leqslant i_0, j_0 \leqslant h = \dim(W_t^*)$, are such that $||W_t^* W_t^{*\prime}|| = \max_{1 \leqslant i,j \leqslant h} |W_{t,i}^* \cdot W_{t,j}^*| =$

$|W^*_{t,i_0} \cdot W^*_{t,j_0}|$. Hence, $E||g(\theta;V_{t+1})g(\theta;V_{t+1})'||^{r+\tilde{\delta}} \leqslant c^2 \cdot \max\{1, E||W^*_t||^{2r+2\tilde{\delta}}\}$, and so by letting $2\tilde{\delta} = \delta$ and using assumption (v), we get $E||g(\theta;V_{t+1})g(\theta;V_{t+1})'||^{r+\tilde{\delta}} < \infty$. Together, the strict stationarity of $\{g(\theta;V_{t+1})g(\theta;V_{t+1})'\}$ and Corollary 3.48 in White (2001) than ensure that $\hat{S}(\theta) \xrightarrow{p} \tilde{S}(\theta) = E[g(\theta;V_{t+1})g(\theta;V_{t+1})']$. In particular, if $\hat{\theta}_n$ is some previously obtained consistent estimate of θ^*, then $\hat{S}(\hat{\theta}_n) \xrightarrow{p} \tilde{S}(\theta^*) = E[g(\theta^*;V_{t+1})g(\theta^*;V_{t+1})']$ which, due to the fact that $\{g(\theta^*;V_{t+1}), \mathcal{F}_t\}$ is a martingale difference sequence and that $\{g(\theta^*;V_{t+1})g(\theta^*;V_{t+1})'\}$ is strictly stationary, equals the asymptotic covariance matrix S in (12).

We now check that all the other conditions of Theorem 2.6 in Newey and McFadden (1994) are satisfied: in particular, we have $S = E[g(\theta^*;Y_{t+1},W^*_t)g(\theta^*;Y_{t+1},W^*_t)'] = E\{[\alpha - 1(Y_{t+1} - \theta^{*\prime}\hat{q}_{m,t} < 0)]^2 W^*_t W^{*\prime}_t\}$ so that for any $\zeta \in \mathbb{R}^h$, we have $\zeta' \cdot S \cdot \zeta = 0$ if and only if

$$\zeta'[\alpha - 1(Y_{t+1} - \theta^{*\prime}\hat{q}_{m,t} < 0)]^2 W^*_t W^{*\prime}_t \zeta = [\alpha - 1(Y_{t+1} - \theta^{*\prime}\hat{q}_{m,t} < 0)]^2 [W^{*\prime}_t \zeta]^2 = 0, a.s. - P,$$

which is equivalent to $W^{*\prime}_t \zeta = 0, a.s. - P$. Since we know from assumption (iv) that $E[W^*_t W^{*\prime}_t]$ is nonsingular, this last equality implies that ζ needs to be equal to an h-vector of zeros. Hence, the matrices S and its inverse S^{-1} are positive definite, therefore nonsingular. In particular, this implies that $S^{-1}E[g(\theta;V_{t+1})] = 0$ only if $E[g(\theta;V_{t+1})] = 0$ which by using the unicity result in Proposition 3 in turn implies $\theta = \theta^*$. This verifies the condition (i) of Theorem 2.6.

Condition (ii) of Theorem 2.6 is the standard compactness condition on the parameter space Θ which we impose here. The continuity condition (iii) of Theorem 2.6 holds since $g(\theta;V_{t+1})$ is $a.s.$ continuous on Θ. Indeed, note that the only discontinuity point occurs when $Y_{t+1} = \theta^{*\prime}\hat{q}_{m,t}, a.s. - P$ which due to the continuity of Y_{t+1} produces with probability zero.

Finally, condition (iv) of Theorem 2.6 is verified by imposing assumption (v) since for all $\theta \in \Theta$ we have $||g(\theta;V_{t+1})|| \leqslant ||W^*_t||, a.s. - P$, so that $E[\sup_{\theta \in \boldsymbol{\Theta}} ||g(\theta;V_{t+1})||] \leqslant E||W^*_t|| < \max\{1, E||W^*_t||^{2r+\delta}\} < \infty$. We can now safely apply the results of Newey and McFadden's (1994) Theorem 2.6 to show that $\hat{\theta}_n \xrightarrow{p} \theta^*$, which completes the proof of Proposition 4. ■

Lemma 8 (Asymptotic First Order Condition) *Let the Assumptions of Proposition 4 hold. We then have* $\sqrt{n}||g_n(\hat{\theta}_n)|| \xrightarrow{p} 0$.

Proof of Lemma 8. Recall from (11) that $\hat{\theta}_n$ is defined as a local minimum of $[g_n(\theta)]'\hat{S}^{-1}[g_n(\theta)]$ on Θ, where $g_n(\theta) = n^{-1}\sum_{t=m}^{T-1}(\alpha - 1(y_{t+1} - \theta'\hat{q}_{m,t} < 0))w_t^*$. Note that this implies that $\hat{\theta}_n$ is also a local minimum of $||g_n(\theta)||^2 = [g_n(\theta)]' \cdot [g_n(\theta)]$. For $i = 1, 2$ and $j = 1, \ldots, h = \dim(W_t^*)$, let

$$\hat{g}_{n,i,j}(a) \equiv n^{-1}\sum_{t=m}^{T-1}(\alpha - 1(y_{t+1} - (\hat{\theta}_n + \varepsilon e_i)'\hat{q}_{m,t} < 0))w_{t,j}^*,$$

where $\{e_1, e_2\}$ is the standard basis of $\mathbb{R}^2$, and $a \in \mathbb{R}$ is such that for $i = 1, 2$, $\hat{\theta} + ae_i \in \Theta$. Note that $\hat{g}_{n,i,j}(0) = g_{n,j}(\hat{\theta}_n)$ where $g_{n,j}$ corresponds to the jth-component of g_n. For $i = 1, 2$ and for every $j = 1, \ldots, h$ the function $a \mapsto [\hat{g}_{n,i,j}(a)]^2$ is convex , so that for every $\varepsilon > 0$, we have

$$[\hat{g}_{n,i,j}(0)]^2 - [\hat{g}_{n,i,j}(-\varepsilon)]^2 \leqslant \{[\hat{g}_{n,i,j}(\varepsilon)]^2 - [\hat{g}_{n,i,j}(-\varepsilon)]^2\}/2 \leqslant [\hat{g}_{n,i,j}(\varepsilon)]^2 - [\hat{g}_{n,i,j}(0)]^2. \tag{25}$$

Now, note that

$$\begin{aligned}
[\hat{g}_{n,i,j}(\varepsilon)]^2 - [\hat{g}_{n,i,j}(-\varepsilon)]^2 = \; & [\hat{g}_{n,i,j}(\varepsilon) + \hat{g}_{n,i,j}(-\varepsilon)] \cdot [\hat{g}_{n,i,j}(\varepsilon) - \hat{g}_{n,i,j}(-\varepsilon)] \\
= \; & [\hat{g}_{n,i,j}(\varepsilon) + \hat{g}_{n,i,j}(-\varepsilon)] \cdot \\
& [n^{-1}\textstyle\sum_{t=m}^{T-1}(1(y_{t+1} - (\hat{\theta}_n - \varepsilon e_i)'\hat{q}_{m,t} < 0) - \\
& 1(y_{t+1} - (\hat{\theta}_n + \varepsilon e_i)'\hat{q}_{m,t} < 0))w_{t,j}^*],
\end{aligned}$$

so that when $\varepsilon \to 0$, $[\hat{g}_{n,i,j}(\varepsilon)]^2 - [\hat{g}_{n,i,j}(-\varepsilon)]^2 \to 2\hat{g}_{n,i,j}(0)[n^{-1}\sum_{t=m}^{T-1} 1(y_{t+1} = \hat{\theta}_n'\hat{q}_{m,t})w_{t,j}^*]$. By using the inequality (25) it must therefore be the case that

$$P(\hat{g}_{n,i,j}(0)[n^{-1}\sum_{t=m}^{T-1} 1(Y_{t+1} = \hat{\theta}_n'\hat{q}_{m,t})W_{t,j}^*] = 0) = 1. \tag{26}$$

Hence

$$\begin{aligned}
P(\sqrt{n}||g_n(\hat{\theta}_n)|| \; > \; & \epsilon) \leqslant P(\max_{1 \leqslant j \leqslant h} |g_{n,i,j}(\hat{\theta}_n)| > \epsilon/\sqrt{n}) \\
\leqslant \; & P(\max_{1 \leqslant j \leqslant h} |\hat{g}_{n,i,j}(0)| > \epsilon/\sqrt{n}) \\
\leqslant \; & P(\max_{1 \leqslant j \leqslant h} |\hat{g}_{n,i,j}(0)|[n^{-1}\sum_{t=m}^{T-1} 1(Y_{t+1} = \hat{\theta}_n'\hat{q}_{m,t})W_{t,j}^*] > \epsilon/\sqrt{n}),
\end{aligned}$$

where we have used the fact that Y_{t+1} is a continuous random variable, so that $n^{-1}\sum_{t=m}^{T-1} 1(Y_{t+1} = \hat{\theta}_n'\hat{q}_{m,t})W_{t,j}^* = o_p(1)$. Using the condition (26), last the inequality above implies $\sqrt{n}||g_n(\hat{\theta}_n)|| \xrightarrow{p} 0$, which completes the proof of Lemma 8. ■

Proof of Proposition 5. In order to show the asymptotic normality of the GMM estimator $\hat{\theta}_n$ we use the result by Newey and McFadden (1994) and check that all the conditions of their

Theorem 7.2 (p. 2186) are verified. We first need to check that $g_n(\hat{\theta}_n)$ verifies an 'asymptotic first order condition': $g_n(\hat{\theta}_n)'\hat{S}^{-1}g_n(\hat{\theta}_n) \leqslant \inf_{\theta\in\Theta} g_n(\theta)'\hat{S}^{-1}g_n(\theta) + o_p(n^{-1})$. For this it suffices to have $\sqrt{n}||g_n(\hat{\theta}_n)|| \xrightarrow{p} 0$, which is what we have shown in the previous Lemma 8. Note that we also have $\hat{S} \xrightarrow{p} S$, with S nonsingular so that $\hat{S}^{-1} \xrightarrow{p} S^{-1}$. Moreover, S^{-1} is positive definite. We now proceed with checking that conditions (i) to (v) Theorem 7.2 hold.

Recall from our previous discussion that the solution θ^* to the first order condition (8) coincides with the solution θ^{**} to $E[g(\theta^{**}; Y_{t+1}, W_t^*)] = 0$ whenever the information vector W_t^* contains all the relevant information from $\mathcal{F}_t$, that is, when it includes all elements of the time-t information set potentially correlated with the variable $\alpha - 1(Y_{t+1} - \theta^{*\prime}\hat{q}_{m,t} \ < 0)$. Hence, θ^* is a solution to $g_0(\theta^*) = 0$ which shows that condition (i) of Theorem 7.2 hold.

In order to show that the condition (ii) hold, note that g can be written as

$$g(\theta;Y_{t+1}, W_t^*) = [\alpha - H(\theta'\hat{q}_{m,t} - Y_{t+1})]W_t^*,$$

where $H(\cdot)$ is the Heaviside function, i.e. $H(x) = 1$ if $x > 0$ and 0 if $x < 0$. The 'gradient' of $g(\cdot)$ is the function $\Delta : \mathbb{R}^2 \times \mathbb{R}^h \times \Theta \to \mathbb{R}^2 \times \mathbb{R}^h$ such that $\Delta : (\theta;y_{t+1}, w_t^*) \mapsto \Delta(\theta;y_{t+1}, w_t^*)$ with

$$\Delta(\theta;Y_{t+1}, W_t^*) \equiv -\delta(\theta'\hat{q}_{m,t} - Y_{t+1})W_t^*\hat{q}_{m,t}', \tag{27}$$

where $\delta(\cdot)$ represents the Dirac function, i.e. $\delta(x) = 0$ if $x \neq 0$ and $\int_{\mathbb{R}} \delta(x)dx = 1$. Note that $\delta(\cdot)$ is the derivative of $H(\cdot)$, so that we have $|H(x + \varepsilon) - H(x) - \varepsilon\delta(x)| = o(|\varepsilon|)$ for all $x \in \mathbb{R}$. We now show that Δ is indeed a 'gradient' of g in a neighborhood of θ^*, in the sense that $||g(\theta;Y_{t+1}, W_t^*) - g(\theta^*;Y_{t+1}, W_t^*) - \Delta(\theta^*;Y_{t+1}, W_t^*)(\theta - \theta^*)|| = o_p(||\theta - \theta^*||)$. Let

$$r(\theta^*;Y_{t+1}, W_t^*) \equiv ||g(\theta;Y_{t+1}, W_t^*) - g(\theta^*;Y_{t+1}, W_t^*) - \Delta(\theta^*;Y_{t+1}, W_t^*)(\theta - \theta^*)||/||\theta - \theta^*||.$$

In order to simplify the notation we drop the reference to t and let $X \equiv \theta^{*\prime}\hat{q}_{m,t} - Y_{t+1}$ and $\varepsilon \equiv (\theta - \theta^*)'\hat{q}_{m,t}$. Thus

$$\begin{aligned} r(\theta^*;Y_{t+1}, W_t^*) &= ||W_t^*|| \cdot |H(X + \varepsilon) - H(X) - \varepsilon \cdot \delta(X)|/||\theta - \theta^*|| \\ &\leqslant ||W_t^*|| \cdot ||\hat{q}_{m,t}|| \cdot |H(X + \varepsilon) - H(X) - \varepsilon \cdot \delta(X)|/|\varepsilon|, \ a.s. - P, \end{aligned}$$

where we used the fact that $|\varepsilon| \leqslant ||\theta - \theta^*|| \cdot ||\hat{q}_{m,t}||$. Let $A_t \equiv ||W_t^*|| \cdot ||\hat{q}_{m,t}||$. By Cauchy-Schwartz inequality, we have

$$E(A_t^2) \leqslant [E||W_t^*||^4]^{1/2}[E||\hat{q}_{m,t}||^4]^{1/2},$$

so that assumptions (v) and (vi) imply that $E(A_t^2) < \infty$. We now use the finiteness of the second moment of A_t to construct an upper bound for $P(r(\theta^*; Y_{t+1}, W_t^*) > \epsilon)$. For any $\eta > 0$ and any $\epsilon > 0$ let $A \equiv [2E(A_t^2)/\eta]^{1/2} < \infty$ and $\tilde{\epsilon} \equiv \epsilon/A > 0$: we then have

$$\begin{aligned} P(r(\theta^*;Y_{t+1}, W_t^*) > \epsilon) \leqslant & \; P(A_t \cdot |H(X+\varepsilon) - H(X) - \varepsilon \cdot \delta(X)|/|\varepsilon| > \epsilon) \\ \leqslant & \; P(A_t \cdot |H(X+\varepsilon) - H(X) - \varepsilon \cdot \delta(X)|/|\varepsilon| > \epsilon | A_t \leqslant A) \cdot P(A_t \leqslant A) \\ & + P(A_t \cdot |H(X+\varepsilon) - H(X) - \varepsilon \cdot \delta(X)|/|\varepsilon| > \epsilon | A_t > A) \cdot P(A_t > A) \\ \leqslant & \; P(|H(X+\varepsilon) - H(X) - \varepsilon \cdot \delta(X)|/|\varepsilon| > \epsilon/A) + P(A_t > A), \end{aligned}$$

so that by Chebyshev's inequality

$$\begin{aligned} P(r(\theta^*;Y_{t+1}, W_t^*) > \epsilon) \leqslant & \; P(|H(X+\varepsilon) - H(X) - \varepsilon \cdot \delta(X)|/|\varepsilon| > \tilde{\epsilon}) + 1/A^2 \cdot E(A_t^2) \\ \leqslant & \; P(|H(X+\varepsilon) - H(X) - \varepsilon \cdot \delta(X)|/|\varepsilon| > \tilde{\epsilon}) + \eta/2. \end{aligned}$$

Since Dirac delta function is the derivative of Heaviside function we know that given $\tilde{\epsilon} > 0$ and $\eta' \equiv \eta/3 > 0$, there exist some $e > 0$ such that $|\varepsilon| < e$ implies $P(|H(X+\varepsilon) - H(X) - \varepsilon \cdot \delta(X)|/|\varepsilon| > \tilde{\epsilon}) < \eta/3$. Further, recall that $\varepsilon \equiv (\theta - \theta^*)'\hat{q}_{m,t}$, so that for any $e > 0$, there exist some $\rho > 0$ such that $||\theta - \theta^*|| < \rho$ implies $|\varepsilon| < e$, and therefore implies $P(|H(X+\varepsilon) - H(X) - \varepsilon \cdot \delta(X)|/|\varepsilon| > \tilde{\epsilon}) < \eta/3$. For any $\eta > 0$ and any $\epsilon > 0$ we have found $\rho > 0$ such that $||\theta - \theta^*|| < \rho$ implies $P(r(\theta^*; Y_{t+1}, W_t^*) > \epsilon) < \eta$, i.e. we have shown that $P(\lim_{\theta \to \theta^*} r(\theta^*; Y_{t+1}, W_t^*) = 0) = 1$. Therefore, we can say that $g_0(\theta) = E[g(\theta; Y_{t+1}, W_t^*)]$ is differentiable at θ^* with derivative $\gamma \equiv E[\Delta(\theta^*; Y_{t+1}, W_t^*)]$. Using the expression in (27), note that

$$\begin{aligned} \gamma &= E[-\delta(\theta^{*\prime}\hat{q}_{m,t} - Y_{t+1}) W_t^* \hat{q}_{m,t}'] \\ &= E[E_t[-\delta(\theta^{*\prime}\hat{q}_{m,t} - Y_{t+1})] W_t^* \hat{q}_{m,t}'] \\ &= -E[f_t(\theta^{*\prime}\hat{q}_{m,t}) W_t^* \hat{q}_{m,t}'], \end{aligned}$$

where $f_t(\cdot)$ is the density of Y_{t+1} conditional on the information set $\mathcal{F}_t$. We now show that $\gamma' S^{-1} \gamma$ is nonsingular: as previously, consider the quadratic form $\zeta' \gamma' S^{-1} \gamma \zeta$, where $\zeta \in \mathbb{R}^2$. We have

$\zeta'\gamma' S^{-1}\gamma\zeta = 0$ if and only if $\gamma\zeta = 0 \in \mathbb{R}^h$ since, as shown in Proposition 4, S^{-1} is positive definite. On the other hand $\gamma\zeta = 0$ if and only if $E[f_t(\theta^{*\prime}\hat{q}_{m,t})W_t^*\hat{q}'_{m,t}\zeta] = 0$. Given that f_t is assumed to be strictly positive, this last equality hold only if $\hat{q}'_{m,t}\zeta = 0, a.s. - P$. Since by assumption (ii), $\hat{q}_{i,m,t} \neq 0, a.s. - P$, the previous condition can only hold if $\zeta = 0$. Hence, $\gamma' S^{-1}\gamma$ is positive definite, therefore nonsingular.

Condition (iii) of Theorem 7.2 is trivially satisfied by imposing Θ to be compact. In order to show that condition (iv) of Theorem 7.2 holds, i.e. that $\sqrt{n}g_n(\theta^*) \xrightarrow{d} \mathcal{N}(0, S)$, we use a central limit theorem (CLT) for martingale difference sequences (e.g., Corollary 5.26 in White, 2001, p 135): recall from (8) that $\{g(\theta^*; Y_{t+1}, W_t^*), \mathcal{F}_t\}$ is a martingale difference sequence. Also, recall from our previous proof of Proposition 4 that $\hat{S}(\theta^*) = n^{-1}\sum_{t=m}^{T-1} g(\theta^*; y_{t+1}, w_t^*)g(\theta^*; y_{t+1}, w_t^*)'$ is a consistent estimator of the asymptotic covariance matrix S in (12), i.e. $\hat{S}(\theta^*) \xrightarrow{p} S$. In order to apply the CLT provided in Corollary 5.26 of White (2001), we need to show that the following moment condition hold: $E||g(\theta^*; Y_{t+1}, W_t^*)||^{2+\delta} < \infty$ for some $\delta > 0$. We have: $E||g(\theta; Y_{t+1}, W_t^*)||^{2+\delta} \leqslant E||W_t^*||^{2+\delta} \leqslant \max\{1, E||W_t^*||^{2r+\delta}\}$, where $r > 2$, so that by assumption (v), $E||g(\theta; Y_{t+1}, W_t^*)||^{2+\delta} < \infty$ for some $\delta > 0$. We can therefore use Corollary 5.26 of White (2001) to show that $\sqrt{n}g_n(\theta^*) \xrightarrow{d} \mathcal{N}(0, S)$. Finally, Andrews (1994) has shown that the stochastic equicontinuity condition (v) of Theorem 7.2 holds for moment functions such as $g(\theta; Y_{t+1}, W_t^*)$. We can now safely apply the results of Theorem 7.2 in Newey and McFadden (1994) to show that $\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{d} \mathcal{N}(0, (\gamma' S^{-1}\gamma)^{-1}\gamma' S^{-1}\gamma(\gamma' S^{-1}\gamma)^{-1})$, i.e. $\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{d} \mathcal{N}(0, (\gamma' S^{-1}\gamma)^{-1})$, which completes the proof of Proposition 5. ■

Proof of Theorem 6. From Theorem 5, it follows that $\Omega^{-1/2}\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{d} \mathcal{N}(0, I_2)$, where $\Omega^{-1/2}$ is such that $(\Omega^{-1/2})'(\Omega^{-1/2}) = \Omega$ and I_2 indicates the identity matrix of order 2. Since $\hat{\Omega} \xrightarrow{p} \Omega$ as $n \to \infty$, Theorem 4.30 of White (2001) implies that $n(\hat{\theta}_n - \theta^*)'\hat{\Omega}^{-1}(\hat{\theta}_n - \theta^*) \xrightarrow{d} \chi_2^2$, from which (a) and (b) follow. ■

References

[1] Andrews, D.W.K., (1994): 'Empirical Process Methods in Econometrics', in *Handbook of Econometrics*, 4, 2247-2294.

[2] Barone-Adesi, G., Bourgoin, F., and Giannopoulos, K. (1998), 'Don't look back', *Risk*, 11.

[3] Bates J. M., and Granger, C. W. J., (1969): 'The Combination of Forecasts', *Operational Research Quarterly*, 20, 451-468.

[4] Battacharya, P.K. and Gangopadhyay, A.K. (1990), 'Kernel and a nearest-neighbor estimation of a conditional quantile', *Annals of Statistics*, 18, 1400-1415.

[5] Bierens, H.J., and Ginther, D. (2001): 'Integrated Conditional Moment Testing of Quantile Regression Models', *Empirical Economics*, 26, 307- 324.

[6] Chernozhukov, V. and Umantsev, L., (2001): 'Conditional Value-at-Risk: Aspects of Modeling and Estimation', MIT Department of Economics Working Paper, 01-19.

[7] Christoffersen, P., (1998): 'Evaluating Interval Forecasts', *International Economic Review*, 39, 841-862.

[8] Christoffersen, P., Hahn, J., Inoue, A., (2001): 'Testing and Comparing Value-at-Risk Measures', *Journal of Empirical Finance*, 8, 325-342.

[9] Clemen, R. T., (1989): 'Combining Forecasts: a Review and Annotated Bibliography', *International Journal of Forecasting*, 5, 559-583.

[10] Clements, M. P., Hendry, D. F. (1998): *Forecasting Economic Time Series,* Cambridge University Press

[11] Danielsson, J., and de Vries, C. (1997), 'Tail index and quantile estimation with very high frequency data', *Journal of Empirical Finance*, 4, 241-257.

[12] Diebold, F. X., (1989): 'Forecast Combination and Encompassing: Reconciling Two Divergent Literatures', *International Journal of Forecasting*, 5, 589-592.

[13] Diebold, F. X., Mariano, R. S. (1995): 'Comparing Predictive Accuracy', *Journal of Business and Economic Statistics,* 13, 253-263.

[14] Diebold, F.X., Schuermann, T. and Stroughair, J. (1998), 'Pitfalls and Opportunities in the Use of Extreme Value Theory in Risk Management,' in A.-P. N. Refenes, J.D. Moody and A.N. Burgess (eds.), *Advances in Computational Finance*, 3-12, Amsterdam: Kluwer Academic Publishers. Reprinted in *Journal of Risk Finance*, 1 (Winter 2000), 30-36.

[15] Duffie, D. and Pan, J., (1997), 'An Overview of Value at Risk', *Journal of Derivatives*, 4, 7-49.

[16] Elliott, G. and Timmermann, A., (2002): 'Optimal Forecast Combinations under General Loss Functions and Forecast Error Distributions', University of California, San Diego Discussion Paper 2002-08.

[17] Embrechts, P., Resnick, and Samorodnitsky, G. (1999), 'Extreme value theory as a risk management tool', *North American Actuarial Journal*, 3, 30-41.

[18] Engle, R.F., and Manganelli, S. (1999), 'CAViaR: Conditional Autoregressive Value at Risk by Regression Quantiles', UCSD Department of Economics Discussion Paper, 1999-20.

[19] Giacomini, R., White, H. (2003): 'Tests of Conditional Predictive Ability', University of California, San Diego manuscript.

[20] Goffe, W., Ferrier, G.D. and Rogers, J. (1994), 'Global Optimization of Statistical Functions with Simulated Annealing', *Journal of Econometrics*, 60, 65-100.

[21] Granger, C. W. J., (1969): 'Prediction with a Generalized Cost of Error Function', *Operational Research Quarterly*, 20, 199-207.

[22] Granger, C. W. J., (1989): 'Combining Forecasts - Twenty Years Later', *Journal of Forecasting*, 8, 167-173.

[23] Granger, C. W. J. and Ramanathan, R., (1984): 'Improved Methods of Combining Forecasts', *Journal of Forecasting*, 3, 197-204.

[24] Hansen, L.P., (1982): 'Large Sample Properties of Generalized Method of Moments Estimators', *Econometrica*, 50, 1029-1054.

[25] Hendry, D. F., Richard, J. F. (1982): 'On the Formulation of Empirical Models in Dynamic Econometrics', *Journal of Econometrics,* 20, 3-33.

[26] Kitamura, Y., and Stutzer, M., (1997): 'An Information-theoretic Alternative to Generalized Method of Moments Estimation', *Econometrica*, 65, 861-874.

[27] Koenker, R. W., Bassett, G. W. (1978): 'Regression Quantiles', *Econometrica,* 46, 33-50.

[28] Koenker, R. and Zhao, Q. (1996), 'Conditional quantile estimation and inference for ARCH models', *Econometric Theory*, 12, 793-813.

[29] Komunjer, I. (2002), 'Quasi-Maximum Likelihood Estimation for Conditional Quantiles', *Cal-Tech Social Science Working Paper*, 1139.

[30] McNeil, A.J. and Frey, R. (2000), 'Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach', *Journal of Empirical Finance*, 7, 271-300.

[31] Mizon, G. E., Richard, J. F., (1986): 'The Encompassing Principle and its Application to Testing Non-nested Hypotheses', *Econometrica,* 54, 657-678.

[32] Morgan, JP, (1996), *RiskMetrics*, Technical Document, 4th edition, New York.

[33] Newey, W. K. and West, K. D. (1987): 'A Simple, Positive Semidefinite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix', *Econometrica,* 55, 703-708.

[34] Stock, J. H. and Watson, M. W., (1999): 'A Comparison of Linear and Nonlinear Univariate Models for Forecasting Macroeconomic Time Series', Engle, R. F. and White, H. (eds.), *Cointegration, Causality and Forecasting*, Oxford University Press.

[35] Stock, J. H. and Watson, M. W., (2001): 'Forecasting Output and Inflation: the Role of Asset Prices', *Harvard University Department of Economics Working Paper.*

[36] Taylor, J. W., Bunn, D. W. (1998): 'Combining Forecast Quantiles Using Quantile Regression: Investigating the Derived Weights, Estimator Bias and Imposing Constraints', *Journal of Applied Statistics,* 25, 193-206.

[37] Taylor, J. (1999), 'A Quantile Regression Approach to Estimating the Distribution of Multi-Period Returns', *Journal of Derivatives*, Fall, 64-78.

[38] West, K. D. (1996): 'Asymptotic Inference about Predictive Ability', *Econometrica*, 64, 1067-1084.

[39] West, K. D. (2001): 'Encompassing Tests When No Model is Encompassing', *Journal of Econometrics,* 105, 287-308.

[40] White, H. (1992), 'Nonparametric estimation of conditional quantiles using neural networks', in H. White (eds.), *Artificial Neural Networks: Approximation and Learning Theory*, 191-205, Oxford: Blackwell.

[41] White, H. (2001): *Asymptotic Theory for Econometricians*, Academic Press, San Diego.

[42] Zheng, J. X. (1998): 'A Consistent Nonparametric Test of Parametric Regression Models under Conditional Quantile Restrictions', *Econometric Theory,* 14, 123-138.

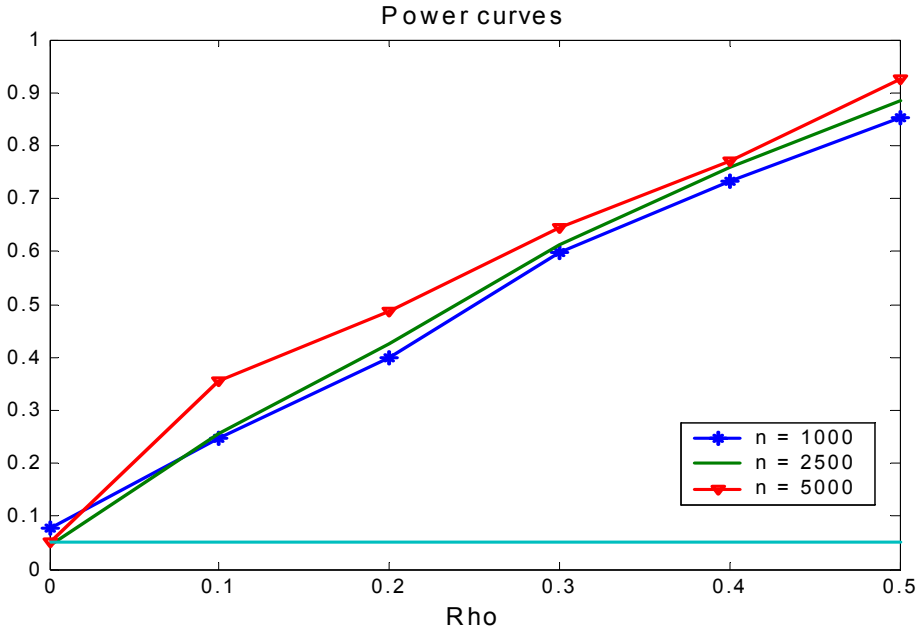

Figure 1: Power curves of the CQFE test in the Monte Carlo experiment discussed in Section 4.2. Each curve represents the rejection frequencies over 500 Monte Carlo replications of the null hypothesis that $VaR_{AAV,t+1}$ encompasses $VaR_{SAV,t+1}$ at the 5% nominal level when the DGP is $r_{t+1} = -[\rho VaR_{SAV,t+1} + (1-\rho)VaR_{AAV,t+1}] + u_{t+1}$. The horizontal axis represents increasing values of ρ.

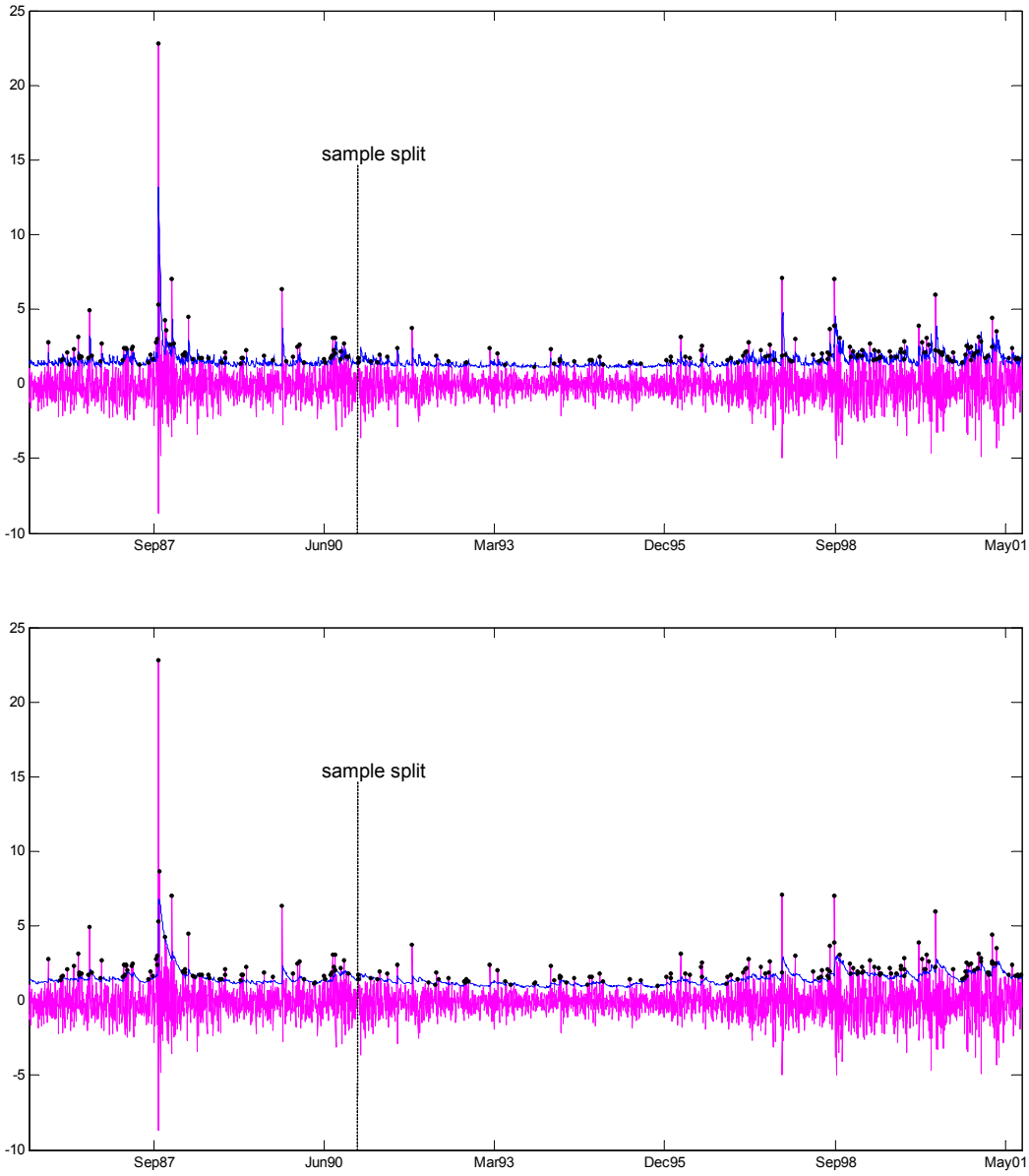

Figure 2: In and out-of-sample daily series of percentage losses on S&P500 index with 5% VaR from the GARCH VAR (top) and the RiskMetrics (bottom) models. VaR violations (or 'hits') are represented by dots.

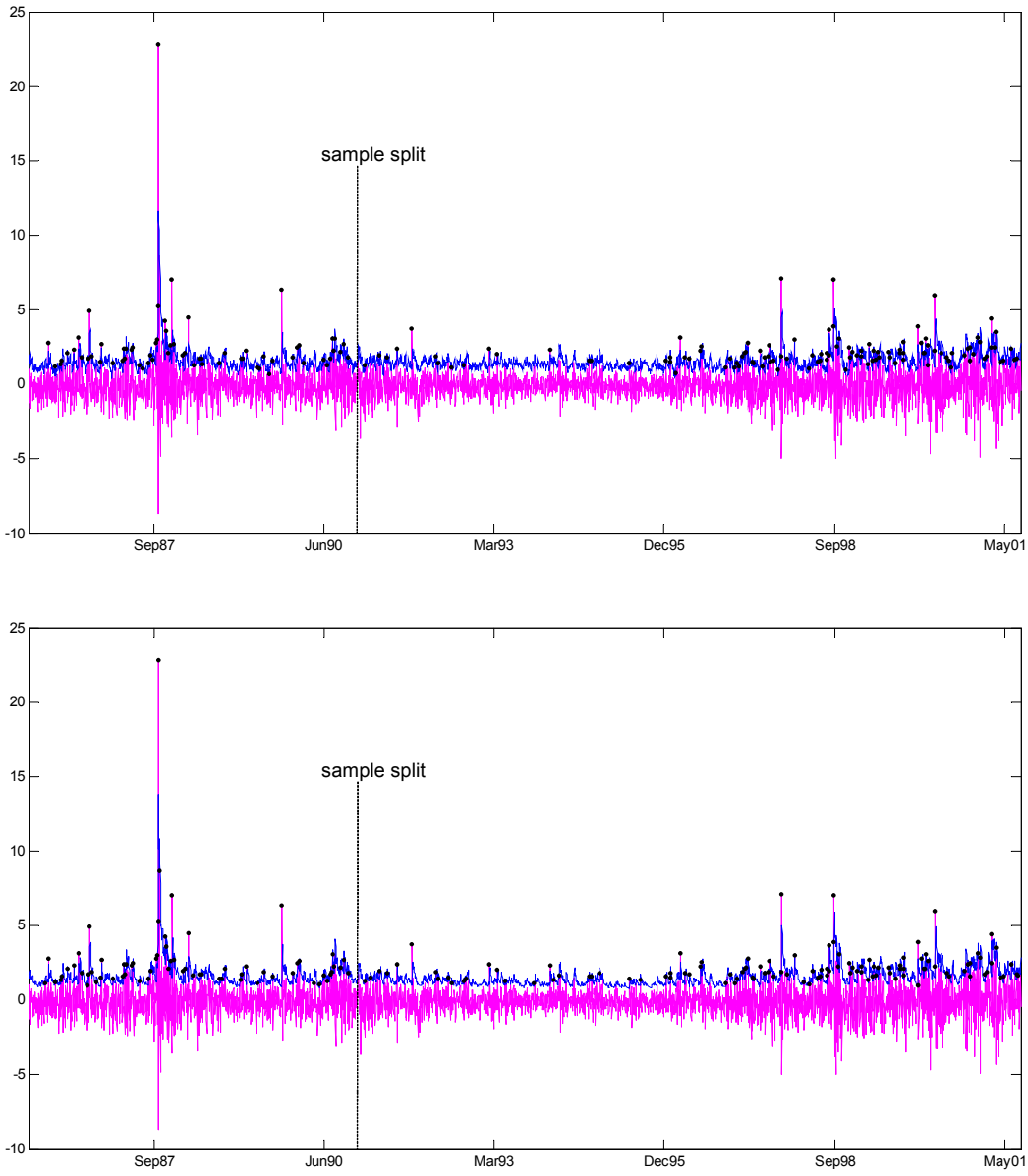

Figure 3: In and out-of-sample daily series of percentage losses on S&P500 index with 5% VaR from the Asymmetric Absolute Value (top) and Asymmetric Slope (bottom) models. VaR violations (or 'hits') are represented by dots.