INTERNATIONAL ECONOMIC REVIEW
Vol. 39, No. 4, November 1998

EVALUATING INTERVAL FORECASTS*

BY PETER F. CHRISTOFFERSEN[†1]

McGill University, Canada

A complete theory for evaluating interval forecasts has not been worked out to date. Most of the literature implicitly assumes homoskedastic errors even when this is clearly violated, and proceed by merely testing for correct *unconditional* coverage. Consequently, I set out to build a consistent framework for *conditional* interval forecast evaluation, which is crucial when higher-order moment dynamics are present. The new methodology is demonstrated in an application to the exchange rate forecasting procedures advocated in risk management.

1. INTRODUCTION

The vast majority of research in economic forecasting centers around producing and evaluating *point* forecasts. Point forecasts are clearly of first-order importance. They are relatively easy to compute, very easy to understand, and they typically guide the immediate action taken by the forecast user. For example, the production manager in a firm wants a forecast of sales in order to decide on production, the chief financial officer wants a forecast of portfolio returns in order to decide on rebalancing, and the central bank governor wants a forecast of inflation in order to carry out monetary policy.

Quickly the question arises: should the user be content with a point forecast? Quite clearly she should not. By nature, point forecasts are of limited value since they only describe one (albeit important) possible outcome. Interval forecasts are equally important—and often neglected—aids. They indicate the likely range of outcomes, thereby allowing for thorough contingency planning. Thus, the production manager can check the hypothetical inventory holdings under various probable sales conditions, the chief financial officer can assess the effects on the solvency of the firm of a range of possible portfolio returns, and the central bank governor can plan policy actions contingent on likely inflationary developments.[2]

* Manuscript received December 1995.
† E-mail: *christop@management.mcgill.ca*
[1] I would like to thank Albert Ando, Anil Bera, Jeremy Berkowitz, Tim Bollerslev, Valentina Corradi, Frank Diebold, Lorenzo Giorgianni, Jin Hahn, Jose Lopez, Roberto Mariano, and two referees for their useful comments. All remaining inadequacies are my own. Special thanks are due to Barbara and Edward Netter for their generous financial support.

[2] A survey of actual interval forecasts, provided by professional forecasters of macroeconomic time series, can be found in Croushore (1993).

Faced with the task of calculating interval predictions, the applied forecaster can build on a large literature, which is summarized in Chatfield (1993). However, when the forecast user wants to *evaluate* a set of interval forecasts produced by the forecaster, not many tools are available. This paper is intended to address the deficiency by clearly defining what is meant by a 'good' interval forecast, and describing how to test if a given interval forecast deserves the label 'good.'

One of the motivations of Engle's (1982) classic paper was to form *dynamic* interval forecasts around point predictions. The insight was that the intervals should be narrow in tranquil times and wide in volatile times, so that the occurrences of observations outside the interval forecast would be spread out over the sample and not come in clusters. An interval forecast that fails to account for higher-order dynamics may be correct on average (have correct unconditional coverage), but in any given period it will have incorrect *conditional* coverage characterized by clustered outliers. These concepts will be defined precisely below, and tests for correct conditional coverage are suggested.

Chatfield (1993) emphasizes that model misspecification is a much more important source of poor interval forecasting than is simple estimation error. Thus, my testing criterion and the tests of this criterion are model free. In this regard, the approach taken here is similar to the one taken by Diebold and Mariano (1995). This paper can also be seen as establishing a formal framework for the ideas suggested in Granger et al., (1989).

Recently, financial market participants have shown increasing interest in interval forecasts as measures of uncertainty. Thus, I apply my methods to the interval forecasts provided by J.P. Morgan (1995). Furthermore, the so-called 'Value-at-Risk' measures suggested for risk measurement correspond to tail forecasts, that is, one-sided interval forecasts of portfolio returns. Lopez (1996) evaluates these types of forecasts applying the procedures developed in this paper.

The remainder of the paper is structured as follows. Section 2 establishes a general efficiency criterion along with a more narrowly defined but more easily applied conditional coverage criterion for interval forecasts. Section 3 establishes some simple tests of the univariate conditional coverage criterion. Section 4 introduces various extensions to the benchmark univariate case. Finally, Section 5 analyzes, in an application to daily exchange rate returns, the interval forecast from J.P. Morgan (1995), along with two competing forecasts.

2. THE FRAMEWORK FOR CONDITIONAL COVERAGE TESTING

2.1. *Defining A Testing Criterion.* The objective of this section is to define a general criterion of goodness for an out-of-sample interval forecast of a given time series. In order to acknowledge the finding in the current literature that model misspecification is the most important source of poor interval forecasts, this paper makes no assumptions on the underlying data generating process. The aim is to develop tests of the *forecasting methodology* being applied—regardless of what it might be—not of any hypothesized underlying true conditional distribution.

The primitives of the analysis are the following: Observe a sample path, $\{y_t\}_{t=1}^T$, of the time series y_t. Also available is a corresponding sequence of *out-of-sample* interval forecasts, $\{(L_{t|t-1}(p), U_{t|t-1}(p))\}_{t=1}^T$, where $L_{t|t-1}(p)$, and $U_{t|t-1}(p)$ are the lower and upper limits of the ex ante interval forecast for time t made at time $t-1$ for the coverage probability, p.

Given the realizations of the time series and the interval forecasts, the indicator variable is defined as,[3]

DEFINITION 1. The indicator variable, I_t, for a given interval forecast, $(L_{t|t-1}(p), U_{t|t-1}(p))$ for time t, made at time $t-1$, is defined as,

$$I_t = \begin{cases} 1, & \text{if } y_t \in \left[L_{t|t-1}(p), U_{t|t-1}(p)\right] \\ 0, & \text{if } y_t \notin \left[L_{t|t-1}(p), U_{t|t-1}(p)\right] \end{cases}.$$

With this definition, I am ready to establish the general testing criterion for interval forecasts as follows:

DEFINITION 2. Say that the sequence of interval forecasts, $\{(L_{t|t-1}(p), U_{t|t-1}(p))\}_{t=1}^T$, is *efficient* with respect to information set Ψ_{t-1}, if $E[I_t|\Psi_{t-1}] = p$, for all t.

In the definition of conditional efficiency the indicator variable is combined with a general conditioning set. This approach enables me to form tests of the interval forecasts without relying on any distributional assumptions on the process being forecasted. This is important in most applications in economics where any kind of distributional assumption is highly questionable. In value-at-risk (VAR) applications, the underlying returns series is nonstationary by construction, since the portfolio is typically changing over time. Furthermore, VAR forecasts are often plagued by misspecification due to time-varying covariances and options risk approximations, so that abstaining from distributional assumptions is crucial.

Notice also that standard evaluation of interval forecasts (e.g., Baillie and Bollerslev 1992, and McNees 1995) proceeds by simply comparing the nominal coverage, $\sum_{t=1}^T I_t/T$ to the true coverage, p. In my framework this corresponds to testing for conditional efficiency with respect to the empty information set, $\Psi_{t-1} = \varnothing$, that is, testing that $E[I_t] = p$, for all t. But, I am not content with this *unconditional* hypothesis. In the presence of higher-order dynamics, testing the *conditional* efficiency of the sequence is important.

VAR estimates have been mentioned as an application of interval forecasting, where the intervals are one-sided.[4] By a one-sided or open interval, I mean that $(L_{t|t-1}(p), U_{t|t-1}(p))$ is equal to either $(L_{t|t-1}(p), +\infty)$, as in VAR, or $(-\infty, U_{t|t-1}(p))$.

[3] In this section attention is restricted to symmetric intervals; the asymmetric case is treated in Section 4.2.

[4] See Kupiec (1995) and Lopez (1996) for the details.

With these terms appropriately defined, the analysis of one-sided intervals corresponds exactly to that of two-sided intervals.

2.2. *An Operational Testing Criterion.* Now I want to make the criterion for out-of-sample interval forecasts operational, and develop easily implementable testing procedures. To construct a readily applied test, consider the information set that consists of past realizations of the indicator sequence, $\Psi_{t-1} = \{I_{t-1}, I_{t-2}, I_{t-3}, \ldots, I_1\}$. The following result is then easily established,

LEMMA 1. *Testing* $E[I_t|\Psi_{t-1}] = E[I_t|I_{t-1}, I_{t-2}, I_{t-3}, \ldots, I_1] = p$, *for all* t, *is equivalent to testing that the sequence* $\{I_t\}$ *is identically and independently distributed Bernoulli with parameter* p. *Write* $\{I_t\} \overset{\text{iid}}{\sim} \text{Bern}(p)$.

PROOF. If $E[I_t|\Psi_{t-1}] = E[I_t|I_{t-1}, I_{t-2}, I_{t-3}, \ldots, I_1] = p$, then, by the definition of the expectation of a binary (0, 1) variable, $\Pr(I_t|I_{t-1}, I_{t-2}, I_{t-3}, \ldots, I_1) = p$ for all t. This implies both independence and that $\Pr(I_t) = p$, for all t. Thus, $\{I_t\} \overset{\text{iid}}{\sim} \text{Bern}(p)$, $\forall t$. The converse is obvious. Q.E.D.

For precision, note the following.

DEFINITION 3. Say that a sequence of interval forecasts, $\{(L_{t|t-1}(p), U_{t|t-1}(p))\}_{t=1}^{T}$, has correct conditional coverage if $\{I_t\} \overset{\text{iid}}{\sim} \text{Bern}(p)$, $\forall t$.

Consider now a given sequence of interval forecasts. The sequence could be constructed from a parametric or nonparametric statistical time series model, or be composed from entirely judgmental forecasts or anything else. The idea is to test the i.i.d. Bern(p) hypothesis for the sequence of interval forecasts in order to get an indication of how close the actual coverage is to the correct *conditional* coverage.[5]

3. A LIKELIHOOD RATIO FRAMEWORK FOR CONDITIONAL COVERAGE TESTING

This section develops an easily applied and unified framework for testing the conditional coverage hypothesis. It can be done conveniently in a likelihood ratio testing framework. The following specifies an LR test of correct unconditional coverage, an LR test of independence, and an LR test that combines the two to form a complete test of the conditional coverage.

3.1. *The LR Test of Unconditional Coverage.* Consider the indicator sequence, $\{I_t\}_{t=1}^{T}$, constructed from a given interval forecast. To test the unconditional coverage, the hypothesis that $E[I_t] = p$ should be tested against the alternative $E[I_t] \neq p$,

[5] Testing for the Bernoulli property has an interesting parallel in the statistics literature on quality control where the fraction of nonconforming articles in a sample is investigated. Refer to Grant and Leavenworth (1988) for the details.

given independence.[6] The likelihood under the null hypothesis is simply

$$L(p; I_1, I_2, \ldots, I_T) = (1-p)^{n_0} p^{n_1},$$

and under the alternative

$$L(\pi; I_1, I_2, \ldots, I_T) = (1-\pi)^{n_0} \pi^{n_1}.$$

Testing for unconditional coverage can be formulated as a standard likelihood ratio test,

$$\mathrm{LR}_{\mathrm{uc}} = -2\log\left[L(p; I_1, I_2, \ldots, I_T)/L(\hat{\pi}; I_1, I_2, \ldots, I_T)\right] \overset{\text{asy}}{\sim} \chi^2(s-1) = \chi^2(1),$$

where $\hat{\pi} = n_1/(n_0 + n_1)$ is the maximum likelihood estimate of π, and $s = 2$ is the number of possible outcomes of the sequence.

This procedure tests the coverage of the interval but it does not have any power against the alternative that the zeros and ones come clustered together in a time-dependent fashion. In the test above, the order of the zeros and ones in the indicator sequence does not matter, only the total number of ones plays a role.

It was stressed above that simply testing for the correct unconditional coverage is insufficient when dynamics are present in the higher-order moments. The two tests presented below make up for this deficiency. The first tests the independence assumption, and the second jointly tests for independence and correct coverage, thus giving a complete test of correct conditional coverage.

3.2. *The LR Test of Independence.* Now I test the independence part of my conditional coverage hypothesis. Independence will be tested against an explicit first-order Markov alternative.[7]

Consider a binary first-order Markov chain, $\{I_t\}$, with transition probability matrix

$$\Pi_1 = \begin{bmatrix} 1-\pi_{01} & \pi_{01} \\ 1-\pi_{11} & \pi_{11} \end{bmatrix},$$

where $\pi_{ij} = \Pr(I_t = j | I_{t-1} = i)$. The approximate likelihood function for this process is

$$L(\Pi_1; I_1, I_2, \ldots, I_T) = (1-\pi_{01})^{n_{00}} \pi_{01}^{n_{01}} (1-\pi_{11})^{n_{10}} \pi_{11}^{n_{11}},$$

where n_{ij} is the number of observations with value i followed by j. As is standard, I condition on the first observation everywhere.[8] It is then easy to maximize the log-likelihood function and solve for the parameters, which are simply ratios of the

[6] Kupiec (1995) and McNees (1995) apply similar tests of unconditional coverage.

[7] Complementary tests based on autocorrelations can be found in Granger et al., (1989), and in tests based on runs in Christoffersen (1996). Notice also the analogy to categorical data analysis, e.g., Andersen (1994).

[8] The exact likelihood ratio tests including the first observation can be found in Christoffersen (1996).

counts of the appropriate cells:

$$\hat{\Pi}_1 = \begin{bmatrix} \dfrac{n_{00}}{n_{00}+n_{01}} & \dfrac{n_{01}}{n_{00}+n_{01}} \\ \dfrac{n_{10}}{n_{10}+n_{11}} & \dfrac{n_{11}}{n_{10}+n_{11}} \end{bmatrix}.$$

Consider now the output sequence, $\{I_t\}$, from an interval model. I estimate a first-order Markov chain model on the sequence, and test the hypothesis that the sequence is independent by noting that

$$\Pi_2 = \begin{bmatrix} 1-\pi_2 & \pi_2 \\ 1-\pi_2 & \pi_2 \end{bmatrix},$$

corresponds to independence. The likelihood under the null becomes

$$L(\Pi_2; I_1, I_2, \ldots, I_T) = (1-\pi_2)^{(n_{00}+n_{10})}\pi_2^{(n_{01}+n_{11})},$$

and the ML estimate is $\hat{\Pi}_2 = \hat{\pi}_2 = (n_{01}+n_{11})/(n_{00}+n_{10}+n_{01}+n_{11})$.

From Hoel (1954) I have the following standard result: the LR test of independence is asymptotically distributed as a χ^2 with $(s-1)^2$ degrees of freedom, that is,

$$\begin{aligned} \mathrm{LR}_{\mathrm{ind}} &= -2\log\left[L\left(\hat{\Pi}_2; I_1, I_2, \ldots, I_T\right)/L\left(\hat{\Pi}_1; I_1, I_2, \ldots, I_T\right)\right] \\ &\overset{\text{asy}}{\sim} \chi^2\left((s-1)^2\right) = \chi^2(1). \end{aligned}$$

Again, I am working with a binary sequence, so $s = 2$. Notice that this test does not depend on the true coverage p, and thus only tests the independence part of my hypothesis. This is, of course, a limitation, but it provides for interesting testing of the dynamics in interval forecast without testing for the true error distribution, as in the example in Section 5. The $\mathrm{LR}_{\mathrm{ind}}$ test is also useful for testing the appropriateness of Bonferroni region forecasts in the multivariate case (see Section 4.1). However, ultimately I would like to test jointly for independence and correct probability parameter, p. This is done below.

3.3. *The Joint Test of Coverage and Independence.* The above tests for unconditional coverage and independence are now combined to form a complete test of conditional coverage. In effect, the null of the unconditional coverage test will be tested against the alternative of the independence test. Consequently, I need to find the distribution of

$$\mathrm{LR}_{\mathrm{cc}} = -2\log\left[L(p; I_1, I_2, \ldots, I_T)/L\left(\hat{\Pi}_1; I_1, I_2, \ldots, I_T\right)\right],$$

and have the following result:

PROPOSITION. *The distribution of the LR test of conditional coverage is asymptotically* χ^2 *with degrees of freedom* $s(s-1)$, *that is,*

$$\mathrm{LR}_{\mathrm{cc}} = -2\log\left[L(p; I_1, I_2, \ldots, I_T)/L\left(\hat{\Pi}_1; I_1, I_2, \ldots, I_T\right)\right] \overset{\text{asy}}{\sim} \chi^2(s(s-1)) = \chi^2(2).$$

PROOF. See the Appendix.

Notice that if I condition on the first observation in the test for unconditional coverage the result is that $\hat{\pi} = \hat{\pi}_2 = \hat{\Pi}_2$. This in turn implies that, when ignoring the first observation, the three LR tests are numerically related by the following identity,

$$\mathrm{LR}_{\mathrm{cc}} = \mathrm{LR}_{\mathrm{uc}} + \mathrm{LR}_{\mathrm{ind}}.$$

This LR framework enables joint testing of randomness and correct coverage while retaining the individual hypotheses as subcomponents. Furthermore, these tests are easy to carry out.

4. EXTENSIONS TO THE BASIC FRAMEWORK

4.1. *The Multivariate Case: Region Forecasts and Bonferroni Bands.* The extension of the testing procedures in Section 3 to evaluation of multivariate forecasts presents no conceptual difficulties. I am given a sample of an m-variate time series, $\{Y_t\}_{t=1}^T$, and a sequence of out-of-sample region forecasts,

$$\left\{R_{t|t-1}(p)\right\}_{t=1}^T,$$

where the desired coverage, p, of the region is prespecified. $R_{t|t-1}(p) \in \Re^m$ is the region forecast for time t made at time $t-1$. Again, define the indicator variable I_t by

$$I_t = \begin{cases} 1, & \text{if } Y_t \in R_{t|t-1}(p) \\ 0, & \text{if } Y_t \notin R_{t|t-1}(p) \end{cases}.$$

It follows that the testing procedure for the multivariate case is identical to the univariate case, so at the theoretical level nothing further needs to be said. As a practical matter, however, the region forecasts can be difficult to compute and interpret, and often the forecaster relies on Bonferroni's method for constructing the joint forecast regions. This method splices together a conservative joint forecast

region from m (one for each time series) individual interval forecasts $\{(L_{i,t|t-1}(1-\tau), U_{i,t|t-1}(1-\tau))\}_{i=1}^{m}$, where $\tau = (1-p)/m$. The resulting joint forecast region has a coverage of *at least* p,

$$\Pr\left(Y_t \in \left(L_{1,t|t-1}(1-\tau), U_{1,t|t-1}(1-\tau)\right) \times \cdots \times \left(L_{m,t|t-1}(1-\tau), U_{m,t|t-1}(1-\tau)\right)\right) \geq 1 - m\tau = p.$$

From an evaluation perspective, these Bonferroni bands are interesting in that my methods allow for separate tests of independence and coverage. Since the coverage will most likely be incorrect, I test independence separately from coverage: the $\mathrm{LR}_{\mathrm{ind}}$ test is useful for this purpose. Rejecting a Bonferroni region forecast in the $\mathrm{LR}_{\mathrm{cc}}$ test should not lead one to the conclusion that the forecast is bad, if what is rejected is that the coverage is too large.

4.2. *Testing for Asymmetries in the Tail Probabilities.* In the previous tests, I did not make explicit whether the realizations that fell outside the predicted interval were in the upper or lower tail of the conditional distribution. If the conditional distribution is symmetric and the predicted intervals are symmetric, this is not critical. On the other hand, if one is concerned about the calibration of each tail individually, or want an asymmetric interval, the framework needs to be generalized as developed below.

Let α_l and α_u be the desired lower and upper tail probabilities, respectively. Then in my previous notation, $1-p = \alpha_1 + \alpha_u$, and the old set-up corresponds to $\alpha_l = \alpha_u = (1-p)/2$.

Now, define

$$S_t = \begin{cases} 1, & \text{if } y_t \leq L_{t|t-1}(\alpha_l) \\ 2, & \text{if } L_{t|t-1}(\alpha_l) < y_t < U_{t|t-1}(\alpha_u) \\ 3, & \text{if } y_t \geq U_{t|t-1}(\alpha_u) \end{cases}$$

Under the null that the interval forecast is correctly calibrated, the transition matrix for S_t is

$$\Pi_0 = \begin{bmatrix} \alpha_l & 1-\alpha_l-\alpha_u & \alpha_u \\ \alpha_l & 1-\alpha_l-\alpha_u & \alpha_u \\ \alpha_l & 1-\alpha_l-\alpha_u & \alpha_u \end{bmatrix},$$

the alternative of independence, but incorrect coverage is

$$\Pi_2 = \begin{bmatrix} \pi_l & 1-\pi_l-\pi_u & \pi_u \\ \pi_l & 1-\pi_l-\pi_u & \pi_u \\ \pi_l & 1-\pi_l-\pi_u & \pi_u \end{bmatrix},$$

and the full alternative allowing for first-order dependence and incorrect coverage is

$$\Pi_1 = \begin{bmatrix} \pi_{ll} & 1-\pi_{ll}-\pi_{lu} & \pi_{lu} \\ \pi_{ml} & 1-\pi_{ml}-\pi_{mu} & \pi_{mu} \\ \pi_{ul} & 1-\pi_{ul}-\pi_{uu} & \pi_{uu} \end{bmatrix}.$$

The three LR tests can then be used again. The test of unconditional coverage is

$$\begin{aligned} \mathrm{LR}_{\mathrm{uc}} &= -2\log\left[L(\Pi_0; S_1, S_2, \ldots, S_T)/L\left(\hat{\Pi}_2; S_1, S_2, \ldots, S_T\right)\right] \\ &\overset{\text{asy}}{\sim} \chi^2(s-1) = \chi^2(2), \end{aligned}$$

where $s = 3$ is the number of states. The test of independence is

$$\begin{aligned} \mathrm{LR}_{\mathrm{ind}} &= -2\log\left[L\left(\hat{\Pi}_2; S_1, S_2, \ldots, S_T\right)/L\left(\hat{\Pi}_1; S_1, S_2, \ldots, S_T\right)\right] \\ &\overset{\text{asy}}{\sim} \chi^2\left((s-1)^2\right) = \chi^2(4). \end{aligned}$$

And the test for conditional coverage is

$$\begin{aligned} \mathrm{LR}_{\mathrm{cc}} &= -2\log\left[L(\Pi_0; S_1, S_2, \ldots, S_T)/L\left(\hat{\Pi}_1; S_1, S_2, \ldots, S_T\right)\right] \\ &\overset{\text{asy}}{\sim} \chi^2(s(s-1)) = \chi^2(6). \end{aligned}$$

4.3. *Expanding the Information Set.* Suppose an interval forecast is rejected using the tests above. One would then like to find out what was causing the rejection. In the tests derived above, the independence of future realizations of the indicator sequence was only tested with respect to the past values of I_t. This restriction makes for easily applied tests of conditional coverage and simple independence. However, one might want to put the interval forecasts through some closer scrutiny and test if a realization outside the predicted interval is associated with certain values of other variables, or combinations of these. Consider the following binary regression framework.

I want to test for the sensitivity of the interval forecast to a q by one vector of observed variables, Z_{t-1}. Z_{t-1} could, of course, include y_{t-1} and I_{t-1}. Then I have

$$\Psi_{t-1} = \{Z_{t-1}, Z_{t-2}, \ldots, Z_1\}.$$

Under the null hypothesis that the current estimates are efficient with respect to this information set, I estimate the relation,

$$I_t = \alpha + \beta' f(Z_{t-1}) + \epsilon_t,$$

where $f(\cdot)$: $\Re^q \to \Re^k$.

LEMMA 2. *Testing for the null hypothesis of interval forecast efficiency,* $E[I_t|\Psi_{t-1}] = p$, *versus the alternative* $E[I_t|\Psi_{t-1}] = \alpha + \beta' f(Z_{t-1})$, *is equivalent to testing,* $[\alpha \;\; \beta'] = [p \;\; \tilde{0}']$, *where* $\tilde{0}$ *is a* k *by one vector of zeros.*

The test of interval forecast efficiency with respect to the information set Ψ_{t-1} can be considered a joint test of independence (slopes equal zero), and correct unconditional coverage (constant term equals p). This framework allows for interesting inference on the interval forecast methodology used. A significantly positive β_i coefficient indicates that the corresponding regressor is not efficiently applied in the current methodology: the probability of getting a realization outside the predicted interval is indeed dependent on $f_i(Z_{t-1})$.

Notice that under the null, the error term will indeed be homoskedastic,

$$\epsilon_t = \begin{cases} 1-p, & \textit{with probability } p \\ -p, & \textit{with probability } 1-p \end{cases},$$

thus, standard inference procedures apply. In closing, note that a natural alternative to the regression approach would be to apply the J-test from Hansen's (1982) GMM framework. My general criterion $E[I_t|\Psi_{t-1}]=p$ implies

$$E\left[(I_t-p)'f(Z_{t-q})\right]=\tilde{0}, \qquad q=1,2,\ldots,t-1$$

which gives me k moment restrictions for each lag of Z_t.

5. APPLICATION: INTERVAL FORECASTING OF DAILY EXCHANGE RATES

Now turn to an empirical application of the methodology developed above. One of the main points of this paper is the important difference between conditional and unconditional interval coverage. Therefore, let us assess one's ability to discern between the two in a realistic finite sample setting. This is done in a simple Monte Carlo experiment, tailored to the subsequent exchange rate application.

In the application, I test a particular real-life interval forecast where the difference between conditional and unconditional coverage is crucial, namely the interval forecast for daily financial time series provided by J.P. Morgan (1995). The performance of this particular interval forecasting methodology is assessed using four daily exchange rate returns.

5.1. *Static Interval Forecasts of Simulated GARCH Processes.* The first step is to get some evidence on how powerful the $\mathrm{LR}_{\mathrm{ind}}$ tests are in rejecting inappropriate interval forecasts under realistic finite-sample conditions. To this end, imagine a univariate time series generated by Bollerslev's (1986) Gaussian GARCH(1,1) model,

$$y_t|\Omega_{t-1} \sim N(0,h_t), \qquad h_t=\omega+\alpha y_{t-1}^2+\beta h_{t-1}.$$

Now, consider a *static* interval forecast of this time series based on the quantiles of the unconditional distribution,

$$[L(p),U(p)]=\left[F^{-1}\left(\frac{1-p}{2}\right),F^{-1}\left(\frac{1+p}{2}\right)\right],$$

where $F(\cdot)$ is the unconditional, time-invariant cdf of y_t.

Figure 1a shows a typical realization of the GARCH process ($\alpha = 0.1$, $\beta = 0.85$, $\omega = 0.05$) along with the static interval forecast. By construction, this interval forecast will have close to perfect unconditional coverage, and thus it will pass the unconditional interval forecast evaluation test, LR_{uc}. However, this is obviously not a good conditional interval forecast, but I want to be able to reject it as such.

Even though the unconditional coverage is correct, in each period the conditional coverage is,

$$F_{t|t-1}(U(p)) - F_{t|t-1}(L(p)) = \Phi\left(U(p)/\sqrt{h_t}\right) - \Phi\left(L(p)/\sqrt{h_t}\right) = 2\Phi\left(U(p)/\sqrt{h_t}\right) - 1 \equiv p_t \neq p.$$

The conditional coverage, p_t, is not constant over time. Since h_t exhibits positive autocorrelation, so too will p_t. For the particular example in Figure 1b, which shows $\{p_t\}$ for the GARCH realization and static interval forecast from Figure 1a, the first-order autocorrelation coefficient is 0.94. As p_t in real applications is unobserved, I hope to detect the misspecification of the static interval forecast by testing for dependence in $\{I_t\}$ over time.

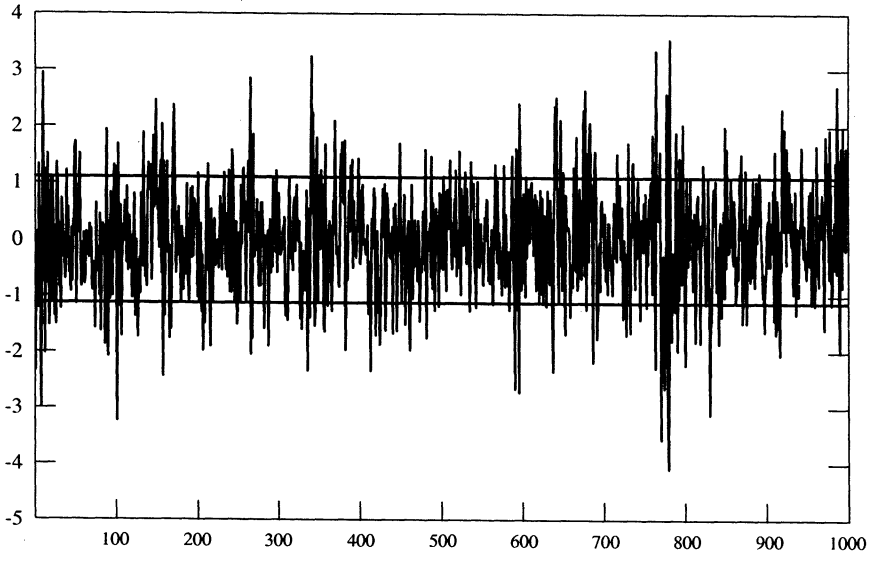

FIGURE 1A

GARCH(1,1) WITH STATIC 75 PER CENT INTERVAL FORECASTS*

* Figure 1a shows a Gaussian GARCH(1,1) realization of length 1000, along with the static 75 per cent interval forecasts, i.e. the 12.5 per cent and 87.5 per cent quantiles of the unconditional distribution. The parameter values of the process are, $\alpha = 0.1$, $\beta = 0.85$, and $\omega = 0.05$.

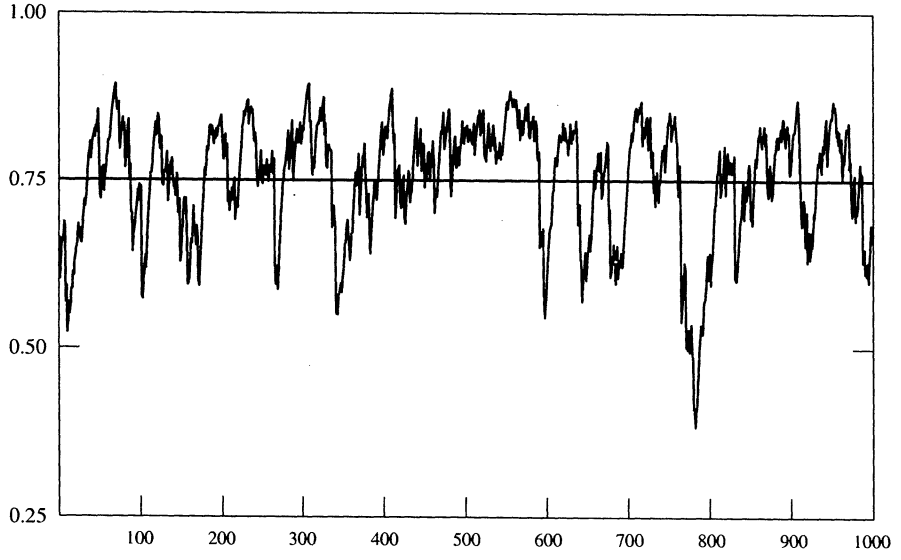

FIGURE 1B

CONDITIONAL COVERAGE OF STATIC 75 PER CENT INTERVAL FORECAST[†]

Figure 2 shows the power of the LR_{ind} test to reject the static interval forecasts when the data is generated by a GARCH process. I have let the actual coverage probability, p, vary between 0.50 and 0.95. The GARCH parameter configuration corresponds to that of Figure 1a. The sample size varies between 250 and 2,000 in increments of 250, and the number of Monte Carlo replications is 1,000. The shape of the power plots illustrates that—given a certain sample size—dependence is the hardest to reject when p is either quite small (close to 0.50) or quite large (close to 0.95). When p is small, the GARCH effects do not have much impact on the true interval forecasts. When p is large, the alternative becomes harder to distinguish from the null, as the number of switches between zero and one, even under the null, is quite small (as n_0/T is small). Figure 2 also illustrates the need for sizable samples. This is quite natural since I am trying to draw inference about the tails of the distribution; a similar point is made for unconditional tests by Kupiec (1995).

5.2. *Interval Forecasting of Daily Exchange Rates.* Keeping in mind the sample sizes required for discerning between conditional and unconditional coverage, let us put the interval evaluation tests to use in a real-life application. I am interested in testing the performance of a forecasting methodology suggested by J.P. Morgan's (1995) RiskMetrics, a new framework for risk measurement. The simple, very tractable methods suggested therein have been found to work well for *variance*

[†] Figure 1b shows the actual *conditional* coverage of the static 75 per cent interval forecasts from the top panel.

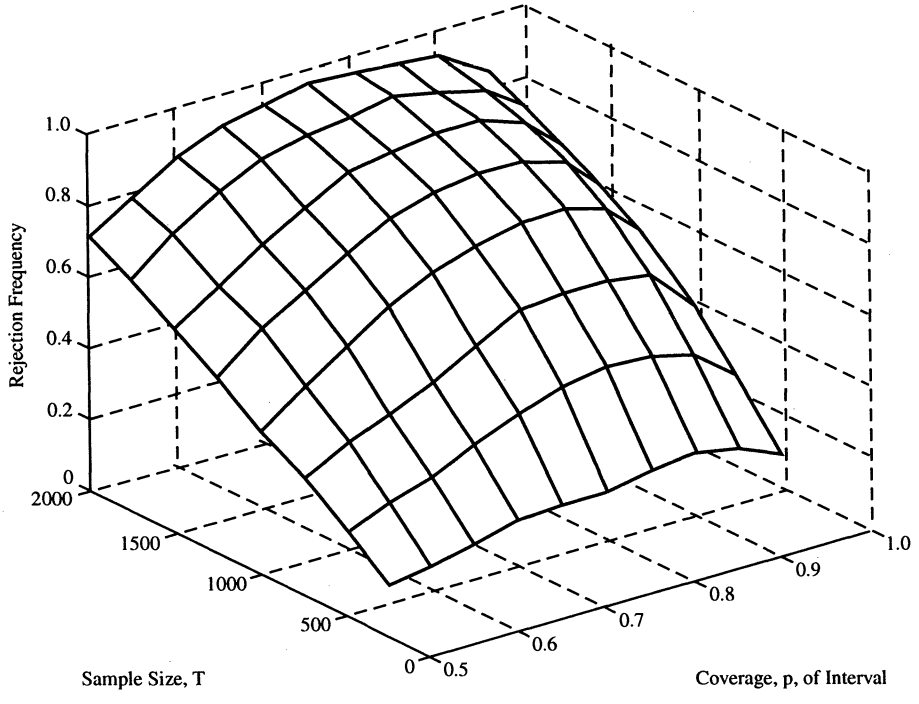

FIGURE 2

POWER OF LR_{IND} TEST AGAINST GARCH DEPENDENCE MONTE CARLO REJECTION FREQUENCY OF STATIC INTERVAL FORECASTS*

forecasting (Boudoukh et al., 1995). But these same models are also used to produce *interval* forecasts, so I want to evaluate them on their own merits.

The interval forecast suggested by J.P. Morgan is,

$$\left(L_{t|t-1}(p), U_{t|t-1}(p)\right) = \left(\Phi^{-1}\left(\frac{1-p}{2}\right)\sigma_t, \Phi^{-1}\left(\frac{1+p}{2}\right)\sigma_t\right)$$

where

$$\sigma_t^2 = (1-\lambda)\sum_{i=0}^{\infty}\lambda^i y_{t-1-i}^2 = \lambda\sigma_{t-1}^2 + (1-\lambda)y_{t-1}^2.$$

Thus, the RiskMetrics forecasts are based on an exponential smoothing model for the squares combined with a standard Gaussian density for the innovations. For

*The figure shows the simulated probabilities of rejecting a static interval forecast using the LR_{ind} test when the true DGP is a Gaussian GARCH(1,1) with parameters $\alpha = 0.1$, $\beta = 0.85$, and $\omega = 0.05$. The significance level is 5 per cent. The coverage, p, of the interval forecasts varies between 50 and 95 per cent. The sample sizes run from 250 to 2,000 in intervals of 250. The number of Monte Carlo replications is 1,000.

daily data, the persistence parameter, λ, is fixed at 0.94. The parameter of the forecasting model is simply *calibrated* at that given value. Notice that the exponential model from RiskMetrics corresponds to a particular IGARCH model without drift, that is, $\alpha + \beta = 1$, $\beta = \lambda$, and $\omega = 0$.

Consistent with most of the literature on exchange rate prediction (see Diebold and Nason 1990), J.P. Morgan does not model any conditional mean dynamics in its forecast. The time varying interval is simply placed around a constant mean.

J.P. Morgan's interval is tested along with two peers. The first is constructed from an estimated GARCH(1,1) model with Student's t innovations (Bollerslev 1987). The second is a simple static forecast such as the one in Figure 1a, based on the in-sample empirical quantiles. I employ daily mid-market (average of bid and offer) log-differences in the British pound, German mark, Japanese yen and Swiss franc vis-a-vis the U.S. dollar, and have a total of 4,000 observations from January 1980 to May 1995.[9] The experiment entails first estimating the parameters necessary (including the empirical quantiles, but *not* λ) to form the three forecasts on the first 2,000 observations. Then I fix the parameters, including the empirical fractiles, and do out-of-sample interval forecasting and forecast evaluation on the last 2,000 observations.

The results of the testing are presented in Figure 3–6. Each figure has three panels. The three lines in each panel give the values of the relevant LR statistics for each of the competing forecasts. The long dashes give the LR value for the RiskMetrics forecast, the solid line for the GARCH-t interval forecasts, and the short dashes for the static forecast. The horizontal, solid line in each panel corresponds to the 5 per cent critical value of the relevant χ^2 distribution. Any LR statistic above this value is statistically significant at the 5 per cent level.

The top panel of Figure 3–6 shows the value of the LR_{cc} statistic, that is, the complete test of conditional coverage. The middle panel shows the LR_{uc} statistic, the test of unconditional coverage. Finally, the bottom panel shows the LR_{ind} statistic, the test of independence. By the decomposition property of these statistics, the values in the middle and bottom panels for each forecast sum to the value in the top panel.

The main results for the three competing interval forecasts can be summarized as follows:

(i) The exponential interval forecast from RiskMetrics passes the independence test, across coverage rates and across exchange rates. It passes the unconditional coverage test for certain coverage rates, typically $p = 0.8 - 0.9$, but fails in most other cases. For the complete test of conditional coverage, the rejections of unconditional coverage lead to rejection of the forecast outside the $0.8 - 0.9$ range of coverages.

(ii) The static interval forecasts fail the independence test in most cases. The tests for unconditional coverage, on the other hand, are passed in general. This indicates that the unconditional distribution does not seem to change over the course of the sample. However, the rejection of independence

[9] The data source is Datastream International.

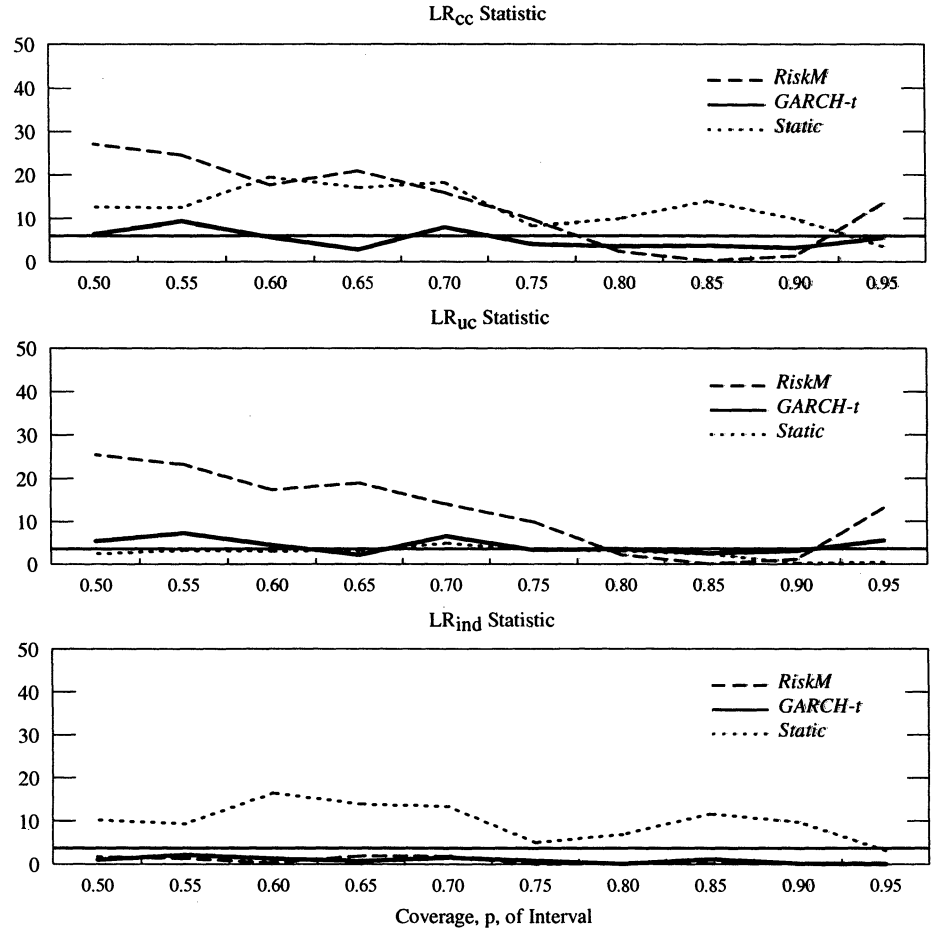

FIGURE 3

BRITISH POUND: LIKELIHOOD RATIO STATISTICS OF CONDITIONAL COVERAGE, UNCONDITIONAL COVERAGE, AND INDEPENDENCE*

* The top panel shows the LR statistics of conditional coverage for three interval forecasts: The long dash is J.P. Morgan's exponential RiskMetrics forecast, the solid line is the GARCH(1,1)-t forecast, and the short dash is the static forecast. The solid horizontal line represents the 5 per cent significance level of the appropriate χ^2 distribution. The test values are plotted for coverages ranging between 50 and 95 per cent. The middle and bottom panels show the corresponding values of the LR tests of unconditional coverage and independence, respectively.

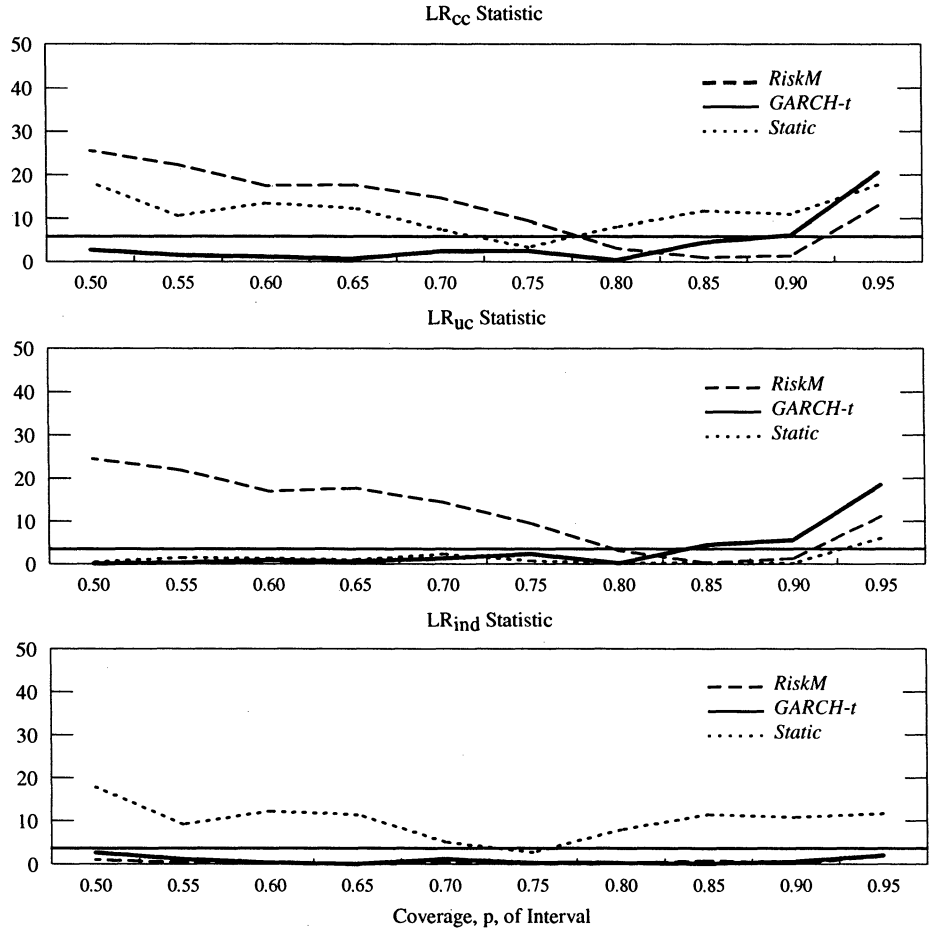

FIGURE 4

GERMAN MARK: LIKELIHOOD RATIO STATISTICS OF CONDITIONAL COVERAGE, UNCONDITIONAL COVERAGE, AND INDEPENDENCE*

* The top panel shows the LR statistics of conditional coverage for three interval forecasts: The long dash is J.P. Morgan's exponential RiskMetrics forecast, the solid line is the GARCH(1,1)-t forecast, and the short dash is the static forecast. The solid horizontal line represents the 5 per cent significance level of the appropriate χ^2 distribution. The test values are plotted for coverages ranging between 50 and 95 per cent. The middle and bottom panels show the corresponding values of the LR tests of unconditional coverage and independence, respectively.

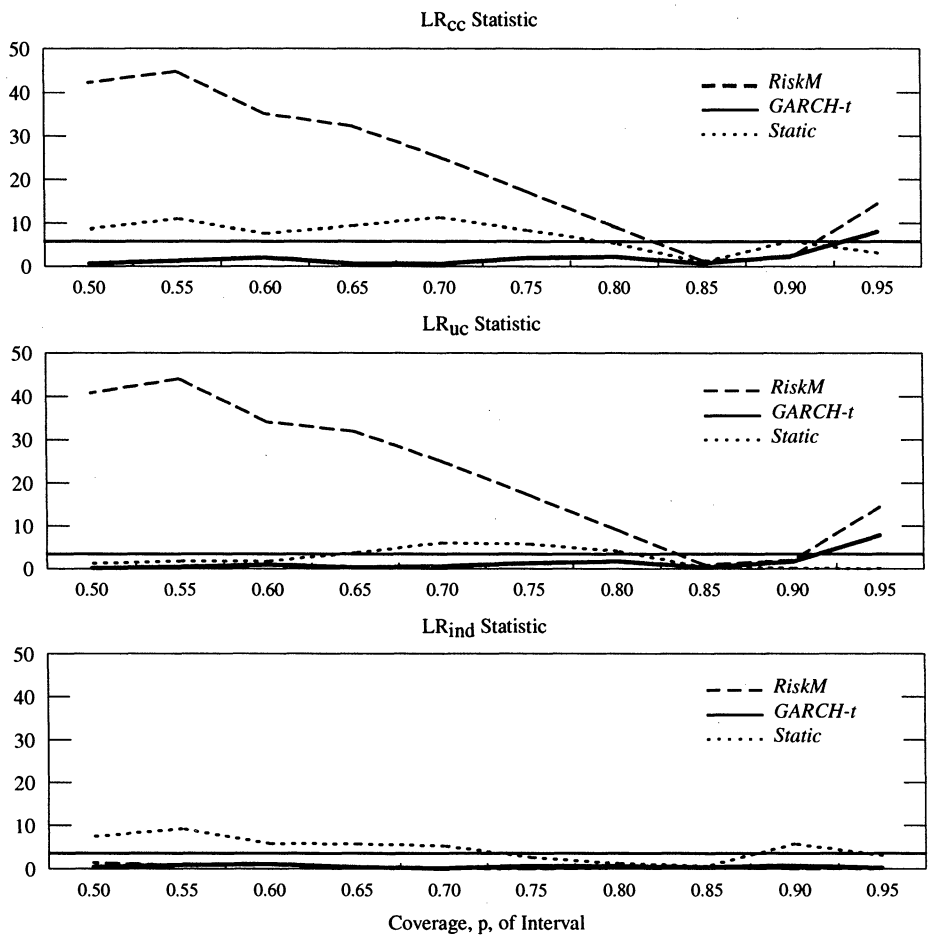

FIGURE 5

JAPANESE YEN: LIKELIHOOD RATIO STATISTICS OF CONDITIONAL COVERAGE, UNCONDITIONAL COVERAGE, AND INDEPENDENCE*

* The top panel shows the LR statistics of conditional coverage for three interval forecasts: The long dash is J.P. Morgan's exponential RiskMetrics forecast, the solid line is the GARCH(1,1)-t forecast, and the short dash is the static forecast. The solid horizontal line represents the 5 per cent significance level of the appropriate χ^2 distribution. The test values are plotted for coverages ranging between 50 and 95 per cent. The middle and bottom panels show the corresponding values of the LR tests of unconditional coverage and independence, respectively.

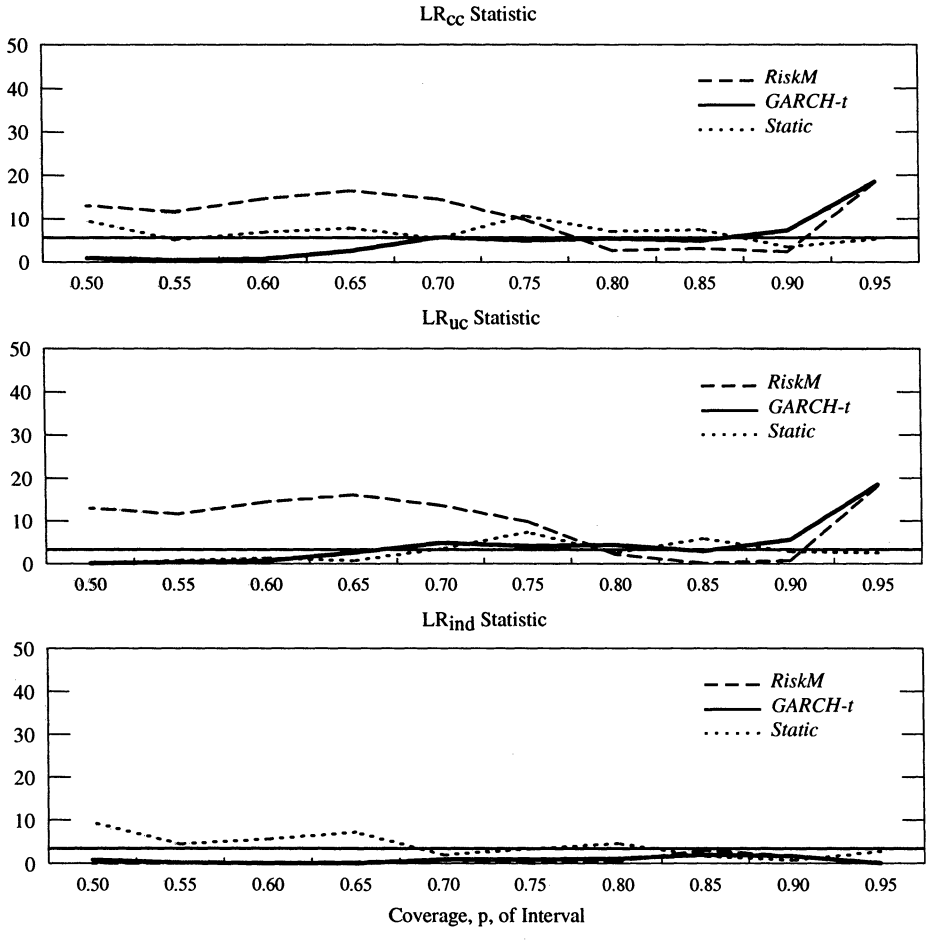

FIGURE 6

SWISS FRANC: LIKELIHOOD RATIO STATISTICS OF CONDITIONAL COVERAGE, UNCONDITIONAL COVERAGE, AND INDEPENDENCE*

* The top panel shows the LR statistics of conditional coverage for three interval forecasts: The long dash is J.P. Morgan's exponential RiskMetrics forecast, the solid line is the GARCH(1,1)-t forecast, and the short dash is the static forecast. The solid horizontal line represents the 5 per cent significance level of the appropriate χ^2 distribution. The test values are plotted for coverages ranging between 50 and 95 per cent. The middle and bottom panels show the corresponding values of the LR tests of unconditional coverage and independence, respectively.

leads to a rejection of conditional coverage for these flat forecasts in most cases. The notable exceptions are for the Swiss franc (across coverage rates), and for $p = 0.95$ (across exchange rates). In the case of the Swiss franc, there is not much evidence of conditional variance dynamics in the prediction sample period.

(iii) Finally, the interval forecasts from the GARCH-t model perform the best overall. The independence test is passed everywhere, and the unconditional coverage is captured in most cases also. The only exceptions to this is for $p = 0.95$, where the unconditional coverage is rejected in some cases.

The performance across interval forecasts for the case where $p = 0.95$ is quite interesting as the static forecast seems to outperform the two dynamic forecasts. Table 1 reports the nominal coverage rates for the various forecasts and exchange rates. It is evident that the GARCH-t forecast is rejected when $p = 0.95$ because it is

TABLE 1
COVERAGE RATES AND AVERAGE WIDTHS OF INTERVAL FORECASTS*

British Pound	50.00	55.00	60.00	65.00	70.00	75.00	80.00	85.00	90.00	95.00
RiskMetrics	55.60	60.30	64.50	69.60	73.80	78.00	81.30	84.90	89.30	93.15
Garch(1,1)-t	52.55	57.95	62.30	66.55	72.60	76.75	81.65	86.25	91.15	96.10
Static	51.75	57.00	61.90	66.90	72.25	76.75	81.55	86.20	90.30	94.70
RiskMetrics	0.90	1.01	1.12	1.24	1.38	1.53	1.71	1.92	2.19	2.61
Garch(1,1)-t	0.84	0.95	1.06	1.18	1.32	1.48	1.66	1.89	2.20	2.71
Static	0.80	0.92	1.04	1.17	1.31	1.48	1.68	1.91	2.27	2.83
German Mark	50.00	55.00	60.00	65.00	70.00	75.00	80.00	85.00	90.00	95.00
RiskMetrics	55.50	60.15	64.45	69.40	73.80	77.90	81.55	85.35	89.25	93.30
Garch(1,1)-t	50.20	55.55	60.95	65.75	71.10	76.40	80.30	86.65	91.55	96.95
Static	50.70	56.30	61.20	66.00	71.50	75.75	80.30	85.45	90.25	96.15
RiskMetrics	0.92	1.03	1.15	1.28	1.41	1.57	1.75	1.96	2.24	2.67
Garch(1,1)-t	0.83	0.94	1.05	1.18	1.32	1.48	1.68	1.92	2.26	2.85
Static	0.79	0.90	1.02	1.16	1.33	1.48	1.67	1.99	2.35	3.12
Japanese Yen	50.00	55.00	60.00	65.00	70.00	75.00	80.00	85.00	90.00	95.00
RiskMetrics	57.10	62.30	66.30	70.90	75.00	78.90	82.65	85.75	89.05	93.05
Garch(1,1)-t	50.40	55.70	61.00	65.50	70.70	76.05	81.15	85.45	90.85	96.30
Static	51.20	56.45	61.40	67.00	72.45	77.25	81.75	85.40	90.25	95.05
RiskMetrics	0.86	0.97	1.08	1.19	1.32	1.47	1.64	1.84	2.10	2.50
Garch(1,1)-t	0.74	0.84	0.94	1.05	1.18	1.33	1.52	1.75	2.09	2.68
Static	0.72	0.82	0.93	1.05	1.18	1.37	1.56	1.79	2.16	2.79
Swiss Franc	50.00	55.00	60.00	65.00	70.00	75.00	80.00	85.00	90.00	95.00
RiskMetrics	54.00	58.75	64.10	69.20	73.70	78.00	81.30	85.25	89.45	92.80
Garch(1,1)-t	50.20	55.50	60.85	66.65	72.20	76.90	81.85	86.35	91.55	96.95
Static	50.40	55.90	61.20	65.85	71.85	77.55	81.40	86.90	91.10	95.75
RiskMetrics	1.01	1.13	1.26	1.40	1.55	1.72	1.92	2.16	2.46	2.94
Garch(1,1)-t	0.94	1.06	1.19	1.32	1.48	1.66	1.87	2.14	2.50	3.12
Static	0.90	1.02	1.16	1.29	1.48	1.69	1.91	2.26	2.66	3.35

* For each exchange rate panel and each true coverage rate, $p = 50$ to 95 per cent, the top half of the panel shows the nominal coverage rate and the bottom half of the panel shows the average width of the interval prediction over the sample.

too cautious. The RiskMetrics forecast, on the other hand, is too confident and gives a nominal coverage below the desired 0.95 probability level. It is also interesting to note that while the nominal coverage for the GARCH-t forecast is everywhere higher than for the static forecast, the average width of the GARCH-t forecast is smaller everywhere.

The observation made in Chatfield (1993) that out-of-sample interval forecasts tend to be too narrow in practice does not hold in this application. If anything, these forecasts tend to be too wide. In almost all instances, the GARCH-t and static forecasts are either correct on average, or a little too wide. The RiskMetrics forecasts are too wide for all coverage rates up to 0.90, but then too narrow for 0.90 and 0.95. Thus, the Gaussian innovation assumption fails at small as well as at large coverage rates.

In conclusion, while the independence test is passed in general by the two dynamic forecasts, there is room for improvement in specifying the innovations distribution where the empirical quantiles forecast performs better. Superior performance might be achieved by combining the parametric, dynamic variance specification, with a (nonparametric) empirical quantiles approach for the innovations, as is done in Engle and Gonzalez-Rivera (1991).

6. SUMMARY AND DIRECTIONS FOR FUTURE RESEARCH

This paper has introduced a general conditional efficiency criterion for evaluating interval forecasts. An easily applied version of this criterion—conditional coverage—is also presented. I suggest a likelihood ratio test of conditional coverage that decomposes into subtests of independence and unconditional coverage, respectively, and is easy to carry out. The separate evaluation of higher-order dynamics and the distributional assumption, which these tests offer, is interesting and useful. It is constructive in that it can indicate whether the dynamics or the innovation distribution (or both) is misspecified. Extensions of the basic set-up, to including general information sets, asymmetric intervals and multivariate time series, are provided.

In an application to daily exchange rates it is shown that the calibrated interval forecast from RiskMetrics, J.P. Morgan's risk measurement methodology, passes the tests for certain coverage rates, but fails for most others. Interval forecasts from an estimated GARCH-t pass the tests in most cases; both in terms of getting the dynamics right, and getting the nominal coverage right. It is interesting that both dynamic, parametric forecasts are often rejected in favor of static interval forecasts when the desired coverage rate is high. In this case, the RiskMetrics forecast turns out to be overly confident, while the GARCH-t forecast is overly cautious. Combining a simply dynamic variance specification with a nonparametric error distribution is likely to present a favorable alternative.

Financial market participants and regulators have recently shown increasing interest in interval forecasting. The so-called Value-at-Risk measures suggested for risk measurement (e.g., Kupiec and O'Brien 1995) correspond to appropriately defined one-sided interval (or tail) forecasts of portfolio returns. Lopez (1996) evaluates these types of forecasts by applying the procedures developed in this paper, along with the methods suggested in Kupiec (1995), and Crnkovic and

Drachman (1996). The latter work is similar in spirit to Diebold et al., (1998), who build on the ideas set forth in this paper to design a framework for density forecast evaluation.

In the Value-at-Risk setup, where attention is confined to the lower tail of the distribution, new challenges face the forecast evaluator who is concerned with testing conditional coverage. The loss of information from having only a one-sided interval forecast can be serious when volatility dynamics are present. This problem, along with investigating the relevance of variance dynamics for risk managers in general, is the topic of current research (Christoffersen and Diebold 1997).

APPENDIX
DISTRIBUTION OF THE LR TEST OF CONDITIONAL COVERAGE

Conditional on the first observation, the likelihood function for a first-order Markov chain with s states is $L = \prod_{i,j}^{s} \pi_{ij}^{n_{ij}}$. Consider testing the null hypothesis that $\pi_{ij} = \pi_j$. The ML estimates under the alternative are $\hat{\pi}_{ij} = n_{ij}/n_i$, with $n_i = \sum_{j=1}^{s} n_{ij}$. I want to find the distribution of $-2\log(\lambda)$, where $\lambda = L_0(\pi_j)/L(\hat{\pi}_{ij})$. Bartlett (1951) shows that the transition counts, n_{ij}, are asymptotically normally distributed so that

$$L \sim c|A|^{1/2}\exp\left(-\tfrac{1}{2}[n-\mu]'A[n-\mu]\right),$$

where $[n-\mu]$ is the vector of the linearly independent variables $n_{ij} - \mu_{ij}$, with μ_{ij} being the expected value of n_{ij}. Using this result in the expression for λ provides:

$$\lambda \sim \frac{c|A^0|^{1/2}\exp\left(-\tfrac{1}{2}[n-\mu^0]'A^0[n-\mu^0]\right)}{c|\hat{A}|^{1/2}\exp\left(-\tfrac{1}{2}[n-\hat{\mu}]'\hat{A}[n-\hat{\mu}]\right)},$$

where the parameters have been replaced by their ML estimates. This can then be written as

$$-2\log(\lambda) \sim \log\left(\frac{|\hat{A}|}{|A^0|}\right) + [n-\mu^0]'A^0[n-\mu^0] + [n-\hat{\mu}]'\hat{A}[n-\hat{\mu}].$$

Under the null, $\hat{\pi}_{ij}$ converges to $\pi_{ij} = \pi_j$, thus, $|\hat{A}|$ converges to $|A^0|$, and I get

$$-2\log(\lambda) \sim [n-\mu^0]'A^0[n-\mu^0] + [n-\hat{\mu}]'\hat{A}[n-\hat{\mu}].$$

It can be shown that $\hat{\mu}_{ij} = n_{ij}$, so that the second term in this expression will vanish, and what is left is $-2\log(\lambda) \sim [n-\mu^0]'A^0[n-\mu^0]$.

The typical element in $[n-\mu^0]$ is $w_{ij} = n_{ij} - \pi_i n_j$. Notice that there are $s-1$ independent restrictions of the form $\sum_{i=1}^{s} n_{ij} = n_j = \sum_{i=1}^{s} n_{ji}$, and in addition, $\sum_{i,j}^{s} n_{ij} = n$. Thus, there are only $s^2 - s = s(s-1)$ independent variables in the quadratic form, and I get

$$-2\log(\lambda) \sim \chi^2(s(s-1)).$$

In the binary case, $s = 2$, and I get a $\chi^2(2)$ distribution.

REFERENCES

ANDERSEN, E.B., *The Statistical Analysis of Categorical Data*, 3rd edition (Berlin: Springer Verlag, 1994).

BARTLETT, M.S., "The Frequency Goodness of Fit Test for Probability Chains," *Proceedings of the Cambridge Philosophical Society* 47 (1951), 86–95.

BAILLIE, R.T. AND T. BOLLERSLEV, "Prediction in Dynamic Models with Time-Dependent Conditional Variances," *Journal of Econometrics* 51 (1992), 91–113.

BOLLERSLEV, T., "Generalized Autoregressive Conditional Heteroskedasticity," *Journal of Econometrics* 31 (1986), 307–327.

———, "A Conditionally Heteroskedastic Time Series Model for Speculative Prices and Rates of Return," *Review of Economics and Statistics* 69 (1987), 542–547.

BOUDOUKH, J., M. RICHARDSON, AND R.F. WHITELAW, "Taking the Pain out of Volatility Estimation," mimeo, Stern School of Business, New York University, 1995.

CHATFIELD, C., "Calculating Interval Forecasts," *Journal of Business and Economics Statistics* 11 (1993), 121–135.

CHRISTOFFERSEN, P.F., "Essays on Forecasting in Economics," Ph.D. Dissertation, University of Pennsylvania, 1996.

——— AND F.X. DIEBOLD, "How Relevant is Volatility Forecasting for Financial Risk Management?" Wharton Financial Institutions Center, Working Paper No. 97–45, 1997.

CRNKOVIC, C. AND J. DRACHMAN, "Quality Control," *Risk* 9 (1996), 138–143.

CROUSHORE, D., "Introducing: The Survey of Professional Forecasters," Federal Reserve Bank of Philadelphia, *Business Review*, November/December (1993), 3–15.

DIEBOLD, F.X. AND R.S. MARIANO, "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics* 13 (1995), 253–265.

——— AND J.A. NASON, "Nonparametric Exchange Rate Prediction?" *Journal of International Economics* 28 (1990), 315–332.

———, A. TAY, AND T. GUNTHER, "Evaluating Density Forecasts," International Economic Review, this issue, 863–883.

ENGLE, R.F., "Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of United Kingdom Inflation," *Econometrica* 50 (1982), 987–1007.

——— AND G. GONZALEZ-RIVERA, "Semiparametric ARCH Models," *Journal of Business and Economic Statistics* 9 (1991), 345–359.

GRANGER, C.W.J., H. WHITE, AND M. KAMSTRA, "Interval Forecasting. An Analysis Based Upon ARCH-Quantile Estimators," *Journal of Econometrics* 40 (1989), 87–96.

GRANT, E.L. AND R.S. LEAVENWORTH, *Statistical Quality Control*, 6th edition (New York: McGraw-Hill, 1988).

HANSEN, L.P., "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica* 50 (1982), 1029–1054.

HOEL, P.G., "A Test for Markov Chains," *Biometrika* 41 (1954), 430–433.

J.P. MORGAN, "RiskMetrics—Technical Document," 3rd edition, Morgan Guaranty Trust Company, New York, 1995.

KUPIEC, P., "Techniques for Verifying the Accuracy of Risk Measurement Models," *Journal of Derivatives* 3 (1995), 73–84.

——— AND J. O'BRIEN, "Internal Affairs," *Risk* 8 (1995), 43–47.

LOPEZ, J.A., "Regulatory Evaluation of Value-at-Risk Models," mimeo, Federal Reserve Bank of New York, 1996.

MCNEES, S., "Forecast Uncertainty: Can It Be Measured?" mimeo, Federal Reserve Bank of Boston, 1995.