Journal of Forecasting
J. Forecast. **23**, 541–557 (2004)
Published online in Wiley InterScience (www.interscience.wiley.com). **DOI**: 10.1002/for.930

Comparing the Accuracy of Density Forecasts from Competing Models

LUCIO SARNO[1]* AND GIORGIO VALENTE[2]
[1] *University of Warwick, UK, and Centre for Economic Policy Research, UK*
[2] *University of Warwick, UK*

ABSTRACT

A rapidly growing literature emphasizes the importance of evaluating the forecast accuracy of empirical models on the basis of density (as opposed to point) forecasting performance. We propose a test statistic for the null hypothesis that two competing models have equal density forecast accuracy. Monte Carlo simulations suggest that the test, which has a known limiting distribution, displays satisfactory size and power properties. The use of the test is illustrated with an application to exchange rate forecasting. Copyright © 2004 John Wiley & Sons, Ltd.

KEY WORDS forecasting; forecast evaluation; density forecast; exchange rates

INTRODUCTION

A large body of literature in econometrics and applied economics has focused on evaluating the forecast accuracy of economic models (e.g. see the survey of Diebold and Lopez, 1996 and references cited therein). Although this literature has traditionally focused on accuracy evaluations based on point forecasts, several authors have recently emphasized the importance of evaluating the forecast accuracy of economic models on the basis of density, as opposed to point, forecasting performance (see, *inter alia*, Diebold *et al*., 1998; Granger and Pesaran, 1999, 2000; Timmermann, 2000; Pesaran and Skouras, 2001; Wallis, 2003).

In a decision-theoretical context, the need to consider the predictive density of a time series—as opposed to considering only its conditional mean and variance—seems fairly accepted in the light of the argument that economic agents may not have loss functions that depend symmetrically on the realizations of future values of potentially non-Gaussian variables. In this case, agents are interested in knowing not only the mean and variance of the variables in question, but their full predictive densities. In various contexts in economics and finance—among which the recent boom in financial risk management represents an obvious case—there is an increasingly strong need to provide and evaluate density forecasts.[1]

* Correspondence to: Lucio Sarno, Finance Group, Warwick Business School, University of Warwick, Coventry, CV4 7AL, UK. E-mail: lucio.sarno@warwick.ac.uk

[1] For a recent survey of the literature on density forecasting and a discussion of its applications in macroeconomics and finance, see Tay and Wallis (2000); see also Corker *et al*. (1986), De Gooijer and Zerom (2000), Weigend and Shi (2000) and Lopez (2001).

Several researchers have recently proposed methods for evaluating density forecasts. For example, Diebold *et al.* (1998) extend previous work—cited in the next section—on the probability integral transform and show how it is possible to evaluate a model-based predictive density and to test formally the hypothesis that the predictive density implied by a particular model corresponds to the true predictive density. Similar ideas have been developed in a different fashion by, *inter alia*, Anderson *et al*. (1994), Li (1996), Granger and Pesaran (1999), Berkowitz (2001) and Li and Tkacz (2001). In general, this line of research has produced several methods either to measure the closeness of two density functions or to test the hypothesis that the predictive density generated by a particular model corresponds to the true predictive density.

However, several gaps still remain in the literature on density forecasting. In particular, one would like to have a statistical procedure for comparing the accuracy of density forecasts produced by competing models. Econometric methods currently available allow the researcher to compare a model-based density forecast to the true predictive density, but they do not allow us to test, for example, the hypothesis that two competing model-based predictive densities are equally close to the true predictive density that the researcher wishes to forecast. Put differently, no testing procedure is available—to the best of our knowledge—that allows the researcher to formally discriminate between alternative models in terms of density forecasting performance.[2] In some sense, it would be desirable to have, in the context of density forecasting, an analogous procedure to the test developed in the context of point forecasting by Diebold and Mariano (1995), who derived a test statistic for the null hypothesis that two models have equal (point) forecast accuracy.

This paper contributes to the relevant literature in that we propose a test statistic for the null hypothesis that two competing models have equal density forecast accuracy. This test statistic, which is based on the concept of integrated square difference rather than the probability integral transform, may be seen as the analogue of the test of Diebold and Mariano (1995) in the context of density forecasting. The proposed test has several attractive properties. In particular, it has a known limiting standard normal distribution and it is easy to implement in practice. Unlike most related tests in this context, our proposed test statistic does not involve testing a joint hypothesis of i.i.d. and uniformity (or normality), rendering the interpretation of the test results straightforward. Also, the test is fairly general and could be applied to density forecasts provided by virtually any econometric model, regardless of the functional form and of the estimation method employed. In addition, Monte Carlo simulations, designed to investigate the size and power properties of this test statistic, suggest that the test has satisfactory empirical size and power properties in a finite sample.

The remainder of the paper is set out as follows. In the next section, we present our test statistic and provide a brief discussion of how this test is linked to the literature on density forecasting. The third section presents the results from carrying out a battery of Monte Carlo simulations designed to examine the empirical size and power properties of the test. In the fourth section we provide an illustrative application of the proposed test statistic in the context of exchange rate forecasting. A final section briefly summarizes and concludes.

[2] To date, researchers have compared model-based density forecasts from competing models without directly testing the hypothesis of equidistance of the competing density forecasts from the true predictive density, using more informal methods (e.g. see Clements and Smith, 2000).

COMPARING THE ACCURACY OF DENSITY FORECASTS FROM COMPETING MODELS

A test statistic

Let $f(y)$, $g_1(y)$ and $g_2(y)$ be three probability density functions with distribution functions F, G_1 and G_2, respectively; F, G_1 and G_2 are absolutely continuous with respect to the Lesbegue measure in $\Re^p$. Let $f(y)$ be the probability density function of the variable y_t over the period $t = 1, \ldots, T$, whereas $g_1(y)$ and $g_2(y)$ are the probability density functions implied by two competing forecasting models, say M_1 and M_2.

We are interested in testing the null hypothesis of equidistance of the probability densities $g_1(y)$ and $g_2(y)$ from $f(y)$, that is

$$H_0 \quad : \quad \text{dist}[f(y), g_1(y)] = \text{dist}[f(y), g_2(y)] \tag{1}$$

where the operator dist denotes a generic measure of distance.

A conventional measure of global closeness between two functions is the integrated square difference (ISD) (e.g. see Pagan and Ullah, 1999):

$$\text{ISD} = \int [\phi(x) - \gamma(x)]^2 dx \tag{2}$$

where $\phi(\cdot)$ and $\gamma(\cdot)$ denote probability density functions; given the definition of ISD in equation (2), ISD ≥ 0, and ISD = 0 only if $\phi(x) = \gamma(x)$. Using (2) we can rewrite the null hypothesis H_0 in (1) as follows:

$$\begin{aligned} H_0 \quad &: \quad \int [f(y) - g_1(y)]^2 dy = \int [f(y) - g_2(y)]^2 dy \\ &: \quad \text{ISD}_1 - \text{ISD}_2 = 0 \end{aligned} \tag{3}$$

In (3), the null hypothesis of equal density forecast accuracy of models M_1 and M_2 is written as the null hypothesis of equality of two integrated square differences or, equivalently, as the null hypothesis that the difference between two integrated square differences is zero.[3]

Consider three series of realizations from $f(y)$, $g_1(y)$ and $g_2(y)$, say $\{y_t\}_{t=1}^{T}$, $\{\hat{y}_{1t}\}_{t=1}^{T_1}$ and $\{\hat{y}_{2t}\}_{t=1}^{T_2}$, respectively.[4] With observations $\{y_t\}_{t=1}^{T}$, $\{\hat{y}_{1t}\}_{t=1}^{T}$ and $\{\hat{y}_{2t}\}_{t=1}^{T}$ we can consistently estimate the unknown functions $f(y)$, $g_1(y)$ and $g_2(y)$ using kernel estimation, obtaining:

$$\hat{f}(y) = \frac{1}{Th} \sum_{i=1}^{T} K\left(\frac{y_i - y}{h} \right) \tag{4}$$

[3] Li and Tkacz (2001) derive a test for evaluating conditional density functions that also uses the notion of integrated square difference. It is important to note, however, that the null hypothesis tested by the Li–Tkacz (2001) test statistic is different from the null hypothesis of the test statistic proposed in the present paper. Indeed, Li and Tkacz suggest that their test 'can be used to determine whether a sample of random variables originates from a given conditional distribution' (p. 2), therefore allowing us to test the null of equality of the conditional density function implied by a model and a nonparametric estimate of the true conditional density function. Hence, their test statistic focuses on the same null hypothesis studied by the growing literature in this context (see Diebold *et al*., 1998) and it is not designed to test the null of equal density forecast accuracy between competing models.

[4] For simplicity and for clarity of exposition, throughout this section, we consider the case where $T_1 = T_2 = T$, although the results derived below can easily be extended to the more general case where $T_1 \neq T_2$.

$$\hat{g}_1(y) = \frac{1}{Th}\sum_{i=1}^{T} K\left(\frac{y_{1i} - y}{h}\right) \tag{5}$$

$$\hat{g}_2(y) = \frac{1}{Th}\sum_{i=1}^{T} K\left(\frac{y_{2i} - y}{h}\right) \tag{6}$$

where $K(\cdot)$ is the kernel function and h is the smoothing parameter.[5]

Using (4)–(6) we can then obtain a consistent estimate of the integrated square differences ISD_1 and ISD_2, say $\widehat{\mathrm{ISD}}_1$ and $\widehat{\mathrm{ISD}}_2$. Define $d = \widehat{\mathrm{ISD}}_1 - \widehat{\mathrm{ISD}}_2$ as the estimated relative distance between the probability density functions. In order to test for the statistical significance of d, the next step is to calculate a confidence interval for d.

In the spirit of the analysis of Hall (1992), define $\{y_i^j\}_{i=1}^T$, $\{\hat{y}_{1i}^j\}_{i=1}^T$, $\{\hat{y}_{2i}^j\}_{i=1}^T$ as the jth resample of the original data $\{y_t\}_{t=1}^T$, $\{\hat{y}_{1t}\}_{t=1}^T$, $\{\hat{y}_{2t}\}_{t=1}^T$, drawn randomly with replacement. From these resamples it is possible to obtain consistent bootstrap estimates of the density functions $\hat{f}^j(y)$, $\widehat{g_1^j}(y)$, $\widehat{g_2^j}(y)$ and, consequently, of $d^j = \widehat{\mathrm{ISD}}_1^{\,j} - \widehat{\mathrm{ISD}}_2^{\,j}$.[6]

Consider a sample path $\{d^j\}_{j=1}^B$, where B is the number of bootstrap replications. Under general conditions,[7] we have:

$$\sqrt{B}(\bar{d} - \mu) \xrightarrow{d} N(0, \sigma^2) \tag{7}$$

where

$$\bar{d} = \frac{1}{B}\sum_{j=1}^{B} d^j = \frac{1}{B}\sum_{j=1}^{B}\left(\widehat{\mathrm{ISD}}_1^{\,j} - \widehat{\mathrm{ISD}}_2^{\,j}\right) \tag{8}$$

is the average difference of the estimated relative distances over B bootstrap replications. Because in large samples the average difference $\bar{d}$ is approximately normally distributed with mean μ and variance σ^2/B, the large-sample statistic for testing the null hypothesis that models M_1 and M_2 have equal density forecast accuracy is:

$$\eta = \frac{\bar{d}}{\sqrt{\hat{\sigma}^2/B}} \xrightarrow{d} N(0,1) \tag{9}$$

where $\hat{\sigma}^2$ is a consistent estimate of σ^2.[8]

[5] In practice, for several econometric models $\hat{g}_1(y)$ and $\hat{g}_2(y)$ are known analytically. However, for more complex, nonlinear models we may not know the probability density function and therefore need to estimate it. In this paper we consider non-parametric estimation of density functions as a general procedure to implement the test statistic discussed below, but it should be clear that the test is directly applicable also when the probability density function is known analytically. Nevertheless, note that the kernel density estimates (4)–(6) will not converge to the true densities unless they are time-invariant, which is a disadvantage of using kernel estimation in this context relative to analytical forms of the model-based density functions or testing procedures based on the probability integral transform.

[6] Note that the data $\{y_i^j\}_{i=1}^T$ can only be resampled with replacement if it is independently and identically distributed (i.i.d.). If there is time dependence, the bootstrap procedure needs to be modified to accommodate it.

[7] See Kendall and Stuart (1976, ch. 11).

[8] On the consistency of the bootstrap estimates of σ^2 in this context see Hall (1992) and Mammen (1992).

Discussion

The literature on evaluating density forecasts is largely based on the idea of the probability integral transform, which goes back at least as far as Rosenblatt (1952). The essence of this methodology is to calculate the probability integral transform, say z_t, of the realizations of the variable of interest over the forecast period with respect to the density forecast produced by a model. If this density forecast corresponds to the true predictive density then the sequence $\{z_t\}_{t=1}^{T}$ is i.i.d. $U[0, 1]$ (Kendall *et al*., 1987, sections 1.27 and 30.36). Diebold *et al*. (1998) illustrate formally how it is possible to evaluate the model forecast density by testing whether there is statistically significant evidence that $\{z_t\}_{t=1}^{T}$ does not depart from the i.i.d. uniform assumption. Hence a test of the null hypothesis that $\{z_t\}_{t=1}^{T}$ is i.i.d. $U[0, 1]$ is tantamount to a test that the model density forecast corresponds to the true predictive density. Obviously, the null of i.i.d. uniformity is a joint hypothesis. Diebold *et al*. (1998) argue that tests of i.i.d. uniformity may often be of little practical use since, when the null hypothesis is rejected, it may not be apparent which leg of the joint hypothesis—i.i.d. or uniformity—is violated. Diebold *et al.* (1998) therefore propose a more 'informal' data analysis. Berkowitz (2001) suggests that a more powerful test may be obtained by working with the inverse normal cumulative distribution function transformation of the $\{z_t\}_{t=1}^{T}$ sequence, which becomes a standard normal variate under the null hypothesis that the model density forecast equals the true predictive density. In this case, one would test whether the transformed realizations are i.i.d. $N(0, 1)$.

Another related paper is due to Wallis (2003), who recasts recently proposed likelihood ratio tests of goodness-of-fit and independence of interval forecasts in the framework of Pearson chi-squared statistics, also considering their extensions to density forecasts and giving special attention to the calculation of small-sample distributions. The proposed test statistics are particularly useful in macroeconomic applications where researchers do not have a large number of observations. Indeed, Wallis (2003) illustrates the potential use of these tests in an application to two series of density forecasts of inflation, namely the US Survey of Professional Forecasters and the Bank of England fan charts.[9]

Using this type of test one could, for example, test whether $f(y) = g_1(y)$ or $f(y) = g_2(y)$, using the notation of the previous section. However, it would not be possible to test whether $g_1(y)$ or $g_2(y)$ is closer to the true predictive density, i.e. it would not be possible to establish whether model M_1 or model M_2 performs better in density forecasting.[10]

In some sense, therefore, the η test statistic fills a gap in the relevant literature in that it is—to the best of our knowledge—the first test statistic designed for testing the null hypothesis that two competing model-based predictive densities are equally close to the density that the researcher wishes to forecast. The η test may also be seen in some ways as a test of relative model adequacy as opposed to a test designed to evaluate density forecasts, which has been the main focus of the relevant literature to date (e.g. Wallis, 2003 and references cited therein). Also, the η test has several virtues. First, under the assumptions described above, the test has a known limiting distribution, which could be derived from basic results in asymptotic theory. Second, the η test can be applied to density forecasts provided by virtually any model, regardless of the functional form and estimation method

[9] Other tests focusing on the closeness between two distribution functions include, for example, the tests due to Anderson *et al*. (1994), Li (1996) and Li and Tkacz (2001).

[10] Presumably, however, it is possible to extend the ideas underlying the work using the probability integral transform to derive a test statistic for comparing the predictive ability of competing models, essentially testing the same null hypothesis tested by the η test statistic in (9). Indeed, this is an immediate avenue for future research. To date, however, researchers have compared model-based density forecasts without directly testing the hypothesis of equidistance of the competing density forecasts from the true predictive density (e.g. see Clements and Smith, 2000).

employed and regardless of whether one is interested in one-step-ahead forecasting or multi-step-ahead forecasting.[11] Third, the η test statistic does not involve testing a joint null hypothesis (i.i.d. and either $U(0, 1)$ or $N(0, 1)$) since it is not based on the probability integral transform. Fourth, as shown in the next section, the η test appears to perform very satisfactorily in terms of size and power properties. Fifth, as illustrated in the final part of the paper, the η test is quite easy to implement in practice.[12]

SIZE AND POWER PROPERTIES OF THE η TEST: MONTE CARLO EVIDENCE

This section reports the results from carrying out Monte Carlo simulations designed to investigate the empirical size and power properties of the η test.

Size

In order to evaluate the empirical size properties of the η test statistic we set up the following experiment. The data generating process (DGP) consists of three probability density functions $f(y)$, $g_1(y)$, $g_2(y)$, each with distribution $N(0, 1)$. We assume that $T_1 = T_2 = T$ for simplicity and investigate $B \in \{10, 25, 50, 100, 500, 1000\}$ and $T \in \{50, 75, 100, 250, 500\}$.[13] In estimation we use the Gaussian kernel function, and the smoothing parameter is calculated according to Silverman's (1986, p. 45) rule: $h = 1.06\sigma T^{-\frac{1}{5}}$.[14] All of the Monte Carlo results discussed in this paper were constructed using 5000 replications in each experiment, with identical random numbers across experiments.[15]

Table I reports the estimated mean value, the estimated variance and the estimated test sizes at the 10%, 5% and 1% significance levels respectively for each value of B and T examined. The simulation results suggest satisfactory size properties. As expected, the performance of the test improves with the number of bootstrap replications; also, the results suggest that with 100 bootstrap replications the empirical size is quite close to the theoretical value. Note that the size of the test is virtually independent of the sample size T, so that even for a small sample size (e.g. $T = 50$) the empirical size of the η test displays properties that are qualitatively identical to the properties obtained for large sample sizes (e.g. $T = 500$). This is not surprising given the definition of the η test statistic (9), which indicates that T does not affect directly the convergence of the test.[16]

---

[11] For linear models it is possible to calculate both one- and multi-step-ahead forecasts analytically. For nonlinear models, while one-step-ahead forecasts can be obtained analytically, multi-step-ahead forecasts usually require bootstrap or Monte Carlo integration methods—see Granger and Terasvirta (1993, ch. 8), Franses and van Dijk (2000, ch. 3–4) and references cited therein.

[12] However, it should be noted that the η test statistic relies upon somewhat stronger assumptions than the test proposed by, for example, Diebold *et al.* (1998) or Berkowitz (2001) using the probability integral transform. See footnotes 5 and 6.

[13] At each replication, for each value of T considered, we generated a sample size of $T + 100$ and discarded the first 100 observations, leaving a sample size T for the analysis; this should reduce the dependence of the results on the initialization.

[14] On alternative choices of the smoothing parameter, see, for example, Pagan and Ullah (1999, ch. 2) and references cited therein.

[15] The computation was carried out using the binned kernel density estimator (Silverman, 1982; Scott and Sheater, 1985; Hardle and Scott, 1992). This procedure allows us to obtain a computationally efficient approximation of the probability functions $\hat{f}(y)$, $\hat{g}_1(y)$, $\hat{g}_2(y)$. The approximation we used has a discrete convolution structure which can be computed using the fast Fourier transform. The number of grid points used is 25, which should provide an estimate of the probability functions virtually indistinguishable from the exact estimate (see Wand and Jones, 1995, appendix D).

[16] In order to assess the robustness of these Monte Carlo results, we performed a number of safeguards. For example, we re-executed the battery of experiments discussed above using the Epanechnikov kernel instead of the Gaussian kernel. We considered several alternative different rules governing the smoothing parameter on the basis of the relationship $h = cT^{(-\frac{1}{5}+k)}$, which is more general than Silverman's rule and collapses to Silverman's rule if $c = 1.06\sigma$ and $k = 0$. We also investigated

Table I. Empirical size of the η test

	Mean	Var	10%	5%	1%
		$T = 50$			
$B = 10$	0.012	1.377	0.158	0.093	0.030
$B = 25$	0.005	1.150	0.121	0.067	0.017
$B = 50$	0.004	1.054	0.109	0.056	0.014
$B = 100$	0.003	1.046	0.108	0.054	0.012
$B = 500$	–0.002	1.031	0.106	0.052	0.011
$B = 1000$	–0.001	1.025	0.104	0.051	0.011
		$T = 75$			
$B = 10$	–0.022	1.410	0.155	0.091	0.032
$B = 25$	0.018	1.118	0.113	0.065	0.017
$B = 50$	–0.016	1.066	0.110	0.058	0.014
$B = 100$	0.010	1.035	0.104	0.056	0.012
$B = 500$	0.008	0.979	0.098	0.054	0.009
$B = 1000$	–0.004	0.985	0.099	0.051	0.009
		$T = 100$			
$B = 10$	0.016	1.371	0.152	0.090	0.029
$B = 25$	–0.014	1.111	0.118	0.066	0.015
$B = 50$	–0.012	1.056	0.110	0.056	0.012
$B = 100$	–0.011	1.024	0.105	0.055	0.012
$B = 500$	–0.008	0.098	0.096	0.047	0.009
$B = 1000$	–0.006	1.001	0.103	0.052	0.010
		$T = 250$			
$B = 10$	0.015	1.390	0.155	0.090	0.030
$B = 25$	0.011	1.103	0.115	0.062	0.016
$B = 50$	–0.008	1.046	0.109	0.054	0.012
$B = 100$	0.007	1.030	0.107	0.053	0.009
$B = 500$	0.003	0.996	0.096	0.047	0.010
$B = 1000$	–0.002	0.997	0.094	0.049	0.010
		$T = 500$			
$B = 10$	–0.022	1.321	0.149	0.086	0.026
$B = 25$	–0.018	1.113	0.116	0.066	0.018
$B = 50$	0.011	1.060	0.111	0.056	0.013
$B = 100$	–0.009	1.037	0.108	0.053	0.011
$B = 500$	–0.007	1.012	0.104	0.052	0.010
$B = 1000$	–0.005	0.986	0.098	0.049	0.010

Notes: Mean and var denote the sample mean and the sample variance of the η test statistic calculated by Monte Carlo methods, as described in the text. 10%, 5% and 1% are the estimated empirical rejection rates.

Power

We investigated the power of the η test against several specific alternatives. In particular, setting the true predictive density $f(y)$ consistent with a $N(0, 1)$ distribution, we calculated the percentage of rejections of the (false) null hypothesis of equal density forecast accuracy under six alternatives represented by the following DGPs:

the size properties of the η test in the presence of undersmoothing ($k < 0$) and oversmoothing ($k > 0$), executing Monte Carlo experiments for values of $c \in \{0.8, 1.0, 1.2\}$ and $k = \{-0.2, -0.1, 0.1, 0.2\}$. These Monte Carlo simulations yielded results that are qualitatively identical to the ones given in Table I. In turn, these results (not reported to conserve space but available upon request) indicate that our Monte Carlo evidence is not particularly subject to a problem of specificity (Hendry, 1984).

1. $g_1(y) = N(0, 1)$, $g_2(y) = N(2, 10)$;
2. $g_1(y) = N(2, 5)$, $g_2(y) = N(3, 8)$;
3. $g_1(y) = N(0, 1)$, $g_2(y) = \chi^2(5)$;
4. $g_1(y) = N(0, 1)$, $g_2(y) = t(5)$;
5. $g_1(y) = t(5)$, $g_2(y) = \chi^2(5)$;
6. $g_1(y) = N(0, 1)$, $g_2(y) = 0.2 \cdot N(1, 1) + 0.8 \cdot N(10, 0.01)$.

Given our results in the previous subsection, we used a fixed value of $T = 100$ in all experiments, and considered $B = \{10, 25, 50, 100\}$. All other aspects of the DGP design are identical to experiments discussed in the previous subsection. The set of different alternatives and competing densities considered is fairly broad in order to explore how excess skewness (e.g. the χ^2 distribution in DGPs 3 and 5), excess kurtosis (e.g. the t distribution in DGP 4) and the presence of multimodality (e.g. the mixture of normal distributions in DGP 6) influence the performance of the η test. Also, the examination of the power properties of the η test under DGPs 1 and 2 is interesting for assessing the power of the test when the competing densities are associated with the same form of distribution but with different moments.

In Figure 1 we have plotted the percentage of rejections of the null hypothesis of equidistance of $g_1(y)$ and $g_2(y)$ from $f(y)$ under each of the DGPs 1 to 6. As Figure 1 clearly reveals, the η test is quite powerful, detecting the false alternatives and rejecting the null of equidistance with a high probability even when the number of bootstrap replications B is fairly small. In fact, with $B = 100$ the rejection rate of the η test is very close to unity for all DGPs (departures from the null hypothesis) examined.

In order to investigate the robustness of these power results, we then re-executed the same battery of simulations under several different underlying distributions for the true predictive density $f(y)$. For example, we set up an experiment where the true predictive density $f(y)$ is set as $t(5)$, all other settings of the DGP design being the same as in the previous experiments.[17] The set of alternative DGPs investigated is the following:

1. $g_1(y) = t(5)$, $g_2(y) = N(2, 10)$;
2. $g_1(y) = t(6)$, $g_2(y) = N(2, 10)$;
3. $g_1(y) = t(5)$, $g_2(y) = \chi^2(5)$;
4. $g_1(y) = t(5)$, $g_2(y) = N(0, 1)$;
5. $g_1(y) = N(0, 1)$, $g_2(y) = \chi^2(5)$;
6. $g_1(y) = t(5)$, $g_2(y) = 0.2 \cdot N(1, 1) + 0.8 \cdot N(10, 0.01)$.

The estimated power functions, plotted in Figure 2, again indicate that the power performance of the test is satisfactory. However, in this case the number of bootstrap replications appears to play a somewhat more important role since the performance of the test improves more substantially when the number of bootstrap replications increases relative to the case where $f(y)$ is $N(0, 1)$.

AN ILLUSTRATIVE EXAMPLE: FORECASTING EXCHANGE RATES

We shall illustrate the practical use of the η test with a simple application to out-of sample exchange rate forecasting. Consider two multivariate models of nominal exchange rate determination, based

[17] We also considered a $\chi^2(5)$ and a $N(2, 5)$, obtaining similar results (not reported but available upon request).

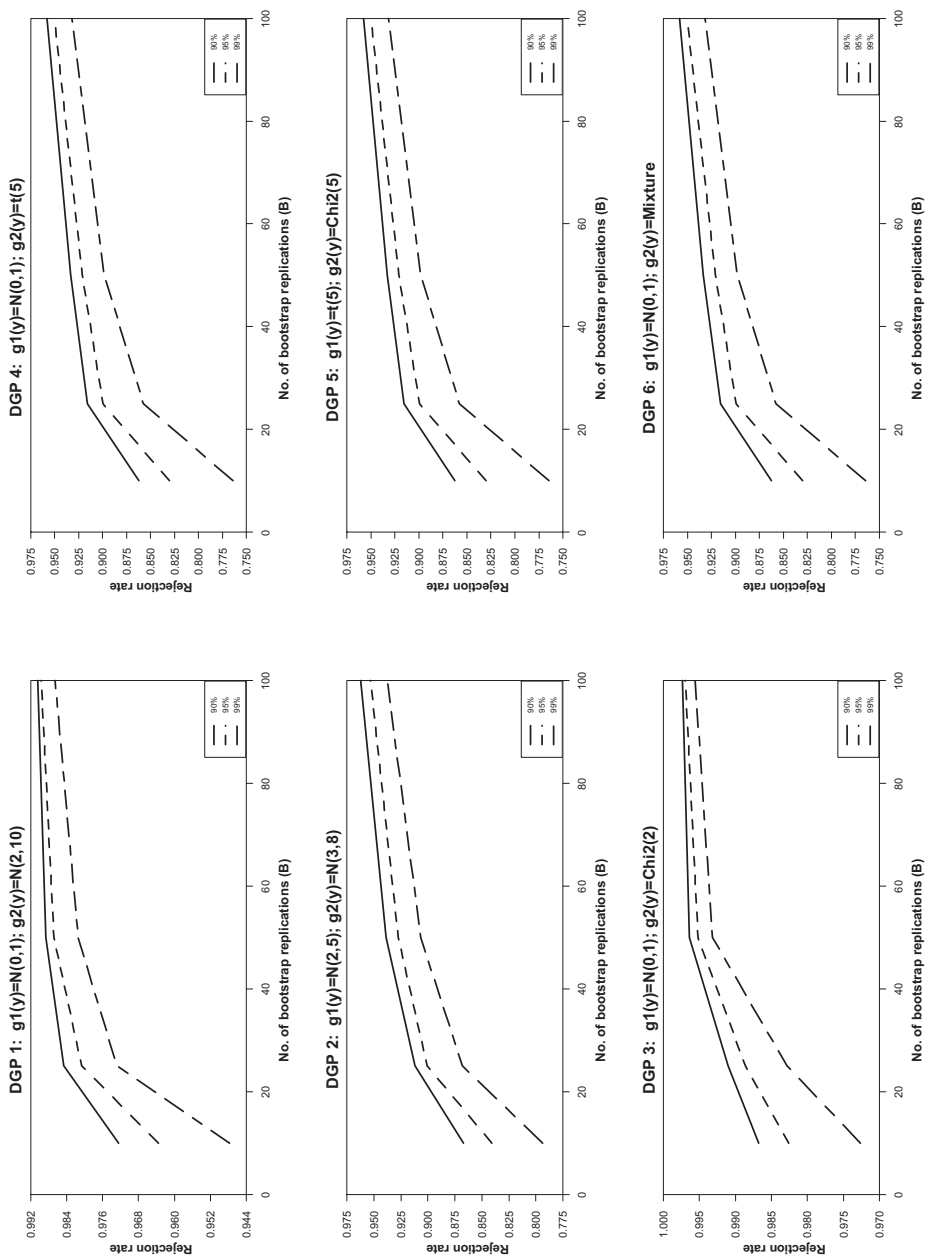

Figure 1. Estimated power functions: $f(y) = N(0, 1)$

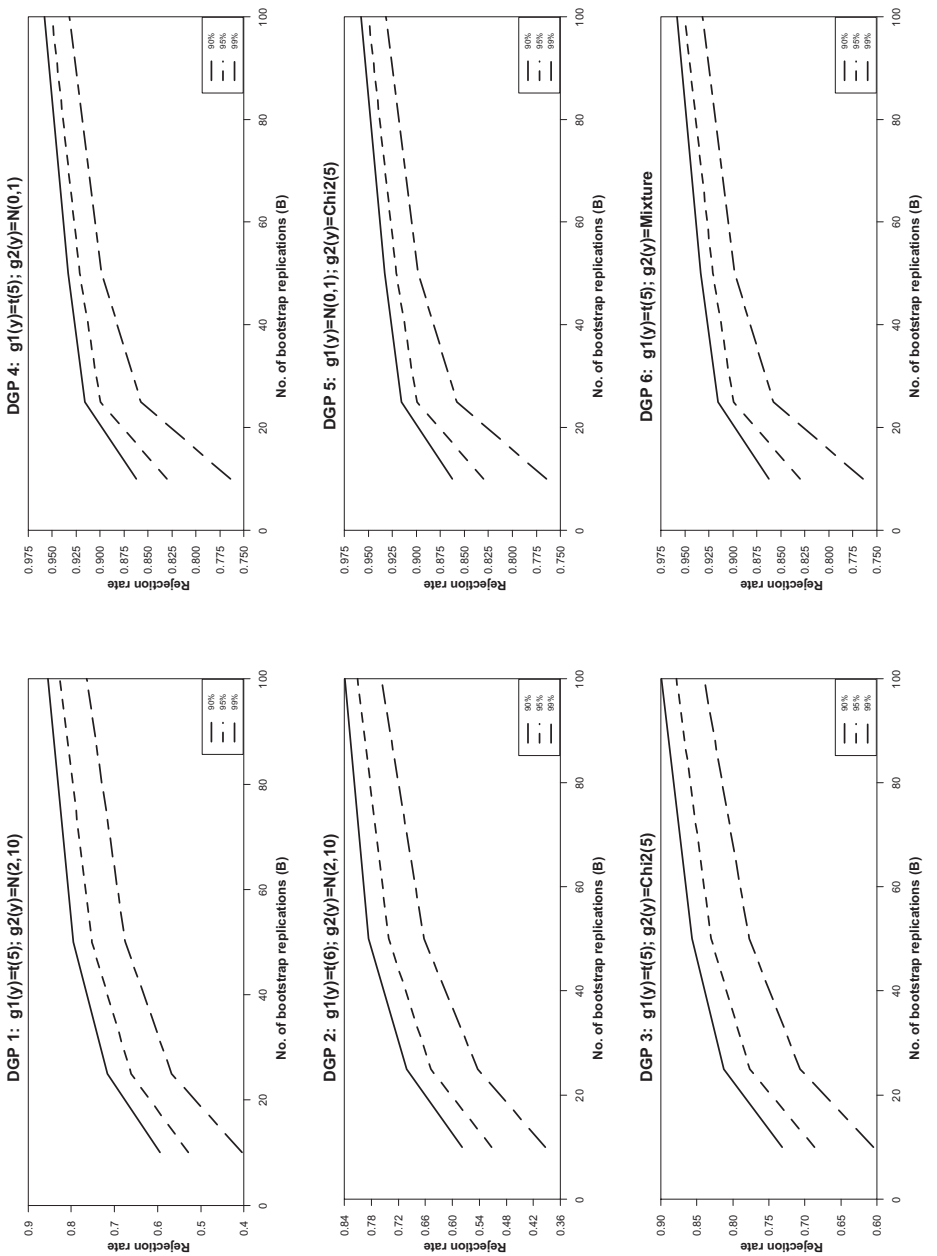

Figure 2. Estimated power functions: $f(y) = t(5)$

on the spot–forward relationship (say model M_1) and the long-run purchasing power parity (PPP) hypothesis (say model M_2).

Model M_1 is related to the literature on foreign exchange market efficiency which tests whether the forward rate is an optimal predictor of the future spot exchange rate, as it should be under the risk-neutral efficient market hypothesis (RNEMH) (e.g. see Hodrick, 1987). Although a large empirical literature has provided evidence that rejects the optimality of the forward rate as optimal predictor of the future spot exchange rate and therefore the validity of the RNEMH, some recent contributions suggest that (the term structure of) forward premia contain valuable information about future exchange rate movements that can be exploited for forecasting exchange rates (e.g. Clarida and Taylor, 1997; Clarida *et al*., 2003). Our model M_1 is a bivariate version of the vector equilibrium correction model (VECM) proposed by Clarida and Taylor (1997). Define s_t and f_t as the logarithm of the spot nominal bilateral exchange rate and the logarithm of the one-month forward exchange rate, respectively. Assuming that both the spot exchange rate and the forward rate are nonstationary and that they have a common stochastic trend (cointegrate), as recorded by a large empirical literature (see Clarida and Taylor, 1997 and references cited therein), then it is possible to characterize the spot–forward relationship using a VECM representation where the long-run equilibrium condition is the forward premium $s_t - f_t$ (Engle and Granger, 1987):

$$\begin{bmatrix}\Delta s_t \\ \Delta f_t\end{bmatrix} = \begin{bmatrix}\mu_1 \\ \mu_2\end{bmatrix} + \sum_{i=1}^{p-1} \Gamma_i \begin{bmatrix}\Delta s_{t-i} \\ \Delta f_{t-i}\end{bmatrix} + \Pi_{M_1} \begin{bmatrix}s_{t-1} \\ f_{t-1}\end{bmatrix} + \begin{bmatrix}\varepsilon_{1t} \\ \varepsilon_{2t}\end{bmatrix} \tag{10}$$

where $\Pi_{M_1} = \alpha_{M_1}\beta'_{M_1}$ is the long-run impact matrix whose rank r determines the number of cointegrating vectors (e.g. Johansen, 1995) and $[\varepsilon_{1t}\ \varepsilon_{2t}]'$ is a vector of disturbances.

Model M_2 in this example is based on an international parity condition, the long-run PPP hypothesis, which is often viewed as a long-run equilibrium condition holding through arbitrage in international goods markets and is assumed in much open-economy modelling (e.g. Rogoff, 1996; Sarno and Taylor, 2002). PPP states that the nominal bilateral exchange rate is equal to the ratio of the relevant national price levels of the two countries considered. A number of researchers have tested for cointegration between the nominal spot exchange rate and relative prices as a way of testing the validity of long-run PPP, with mixed results. However, while very few contemporary economists would hold that PPP holds continuously in the real world, 'most instinctively believe in some variant of purchasing power parity as an anchor for long-run real exchange rates' (Rogoff, 1996, p. 647), and indeed the implication or assumption of much reasoning in international macroeconomics is that some form of PPP holds at least as a long-run relationship.

Define $z_t \equiv p_t - p_t^*$ as the relative price, where p_t and p_t^* denote the logarithm of the domestic and foreign price levels, respectively. If long-run PPP holds, we can express the dynamic relationship between the nominal spot exchange rate and relative prices in a VECM of the form:

$$\begin{bmatrix}\Delta s_t \\ \Delta z_t\end{bmatrix} = \begin{bmatrix}v_1 \\ v_2\end{bmatrix} + \sum_{i=1}^{p-1} \Xi_i \begin{bmatrix}\Delta s_{t-i} \\ \Delta z_{t-i}\end{bmatrix} + \Pi_{M_2} \begin{bmatrix}s_{t-1} \\ z_{t-1}\end{bmatrix} + \begin{bmatrix}e_{1t} \\ e_{2t}\end{bmatrix} \tag{11}$$

where the long-run impact matrix $\Pi_{M_2} = \alpha_{M_2}\beta'_{M_2}$ and $[e_{1t}\ e_{2t}]'$ is a vector of disturbances.

In order to calculate the η test we first estimated the competing models (10) and (11) using monthly bilateral dollar exchange rate data (domestic price of the foreign currency) vis-à-vis the Japanese yen and the UK pound. Time series for bilateral dollar exchange rates and one-month forward rates

Table II. Forecasting results

	MAE	MSE
	US dollar–Japanese yen	
Model M_1	0.010863	0.000239
Model M_2	0.011742	0.000256
DM test	0.18740 [0.826]	0.0988 [0.922]
η test	6.8669 [0.0]	
	US dollar–UK sterling	
Model M_1	0.00778	0.000136
Model M_2	0.00843	0.000157
DM test	0.26112 [0.794]	0.16212 [0.838]
η test	4.2032 [2.6×10^{-5}]	

Notes: Models M_1 and M_2 denote the VECM based on the spot–forward relationship (10) and the VECM based on purchasing power parity (11), respectively. MAE and MSE are the mean absolute error and the mean square error respectively, calculated using the one-step-ahead forecast series (108 data points) as described in the text. DM is the Diebold and Mariano (1995) test statistic for the null hypothesis that models M_1 and M_2 have equal point forecast accuracy; η test is the test statistic for the null hypothesis that models M_1 and M_2 have equal density forecast accuracy, constructed according to equation (9) using 100 bootstrap replications. Figures in brackets denote *p*-values; *p*-values equal to zero up to the eighth decimal point are recorded as [0.0].

over the sample period from January 1979 to December 2000 were obtained from *Datastream*, whereas time series for the consumer price index for Japan, the UK and the USA were obtained from the International Monetary Fund's *International Financial Statistics* CD. We estimated models (10) and (11) using data from January 1979 to December 1991, leaving the data from January 1992 to the end of the sample period for calculating out-of-sample dynamic one-step-ahead forecasts.[18]

The results of the forecasting exercise are given in Table II, where we report both conventional measures of predictive accuracy, such as the mean absolute error (MAE) and the mean square error (MSE), as well as the η test statistic. Although for each exchange rate the spot–forward VECM yields lower MAEs and MSEs than the PPP VECM, the results suggest that the forecasting performance of the two competing models M_1 and M_2 is very similar in that the MAEs and MSEs produced by the two models are very close. This similarity is formally confirmed by carrying out the Diebold and Mariano (1995) test, which tests the null hypothesis that the competing models perform equally well in terms of point forecasting performance. In fact, we are unable to reject, at conventional significance levels, the null hypothesis of equal predictive accuracy under the Diebold–Mariano test for both exchange rates considered.[19] These results would imply that, under the specific measures of predictive accuracy examined, the out-of-sample forecasting performances of models based on the

[18] The lag length, *p*, was set equal to unity for both VECMs, consistent with conventional information criteria. Note, however, that for each model, we did not employ a general-to-specific procedure to achieve a parsimonious specification of the VECM, although a more parsimonious specification may lead to better forecasting results for both models (10) and (11). Nevertheless, we thought this was unnecessary given the merely illustrative nature of the present empirical exercise.

[19] We also calculated Diebold–Mariano tests by bootstrap in order to take into account the impact of smallsample bias and parameter uncertainty on the distribution of the tests. The results were, however, qualitatively identical to the ones reported in Table II.

spot–forward relationship (model M_1) and PPP (model M_2) are not statistically different at conventional nominal levels of significance. Put differently, given that the two competing models have the same functional form and lag structure and the only difference between them is the variable used for forecasting exchange rates (the forward rate in model M_1 and the relative price in model M_2), these results may be viewed as implying that the information content of the forward rate is statistically equivalent to the information content embedded in relative prices for the purpose of forecasting the exchange rate.

However, inspecting Figure 3, which displays the one-step-ahead forecasts and the density forecasts from the competing models together with the actual realizations of the corresponding series and the true predictive density for each exchange rate examined, a different result arises. Although none of the forecast densities implied by the competing models M_1 and M_2 appears particularly close to the true predictive density,[20] the forecasts produced by the PPP VECM (11) are more leptokurtic than the ones obtained from the spot–forward VECM (10). Simple visual inspection of the graphs in Figure 3 suggests that the distance between the forecast density of the spot–forward VECM (10) from the true predictive density is shorter than the distance between the forecast density of the PPP VECM (11) and the true predictive density. This visual evidence is, in fact, supported by the results of the η test, reported in the last row of Table II. For both exchange rates examined, the η test, calculated using 100 bootstrap replications, strongly rejects the null hypothesis of equidistance of the competing predictive densities from the true predictive density. In turn, these results imply that, in contrast with the implications of the MAEs and MSEs discussed earlier, the spot–forward model (model M_1) is superior to the PPP model (model M_2) in terms of out-of-sample forecasting performance, suggesting that the information content of the forward rate is more valuable than the information content of relative prices for the purpose of forecasting the exchange rate.

CONCLUSION

This paper contributes to the recent line of research that emphasizes the need to evaluate the forecasting ability of empirical models on the basis of density forecast accuracy. The recent relevant literature has proposed several methods either to measure the closeness of two density functions or to test the hypothesis that the predictive density generated by a particular model corresponds to the true predictive density. The specific contribution of this paper is that it provides a test statistic for comparing the accuracy of density forecasts produced by competing models and formally testing the hypothesis that two competing model-based density forecasts are equally close to the density that the researcher wishes to forecast. This test is, in the context of density forecasting, the analogue of the test statistic developed by Diebold and Mariano (1995) for testing the null hypothesis that two models have equal forecast accuracy in the context of point forecasting.

Our proposed test statistic displays several attractive properties in that it has a known limiting standard normal distribution and—unlike available testing procedures—does not involve testing a

[20] Note, for example, that the true predictive density has fatter tails than the predictive densities from either model M_1 or M_2, suggesting that none of the two simple models considered in this application is particularly good at capturing the higher moments in the exchange rate data examined. Logical extensions of the linear VECMs used here which might achieve a more accurate description of these fat tails include VECMs that allow for autoregressive conditional heteroskedasticity or for regime switching (see Clarida *et al*., 2003; Sarno and Valente, 2004).

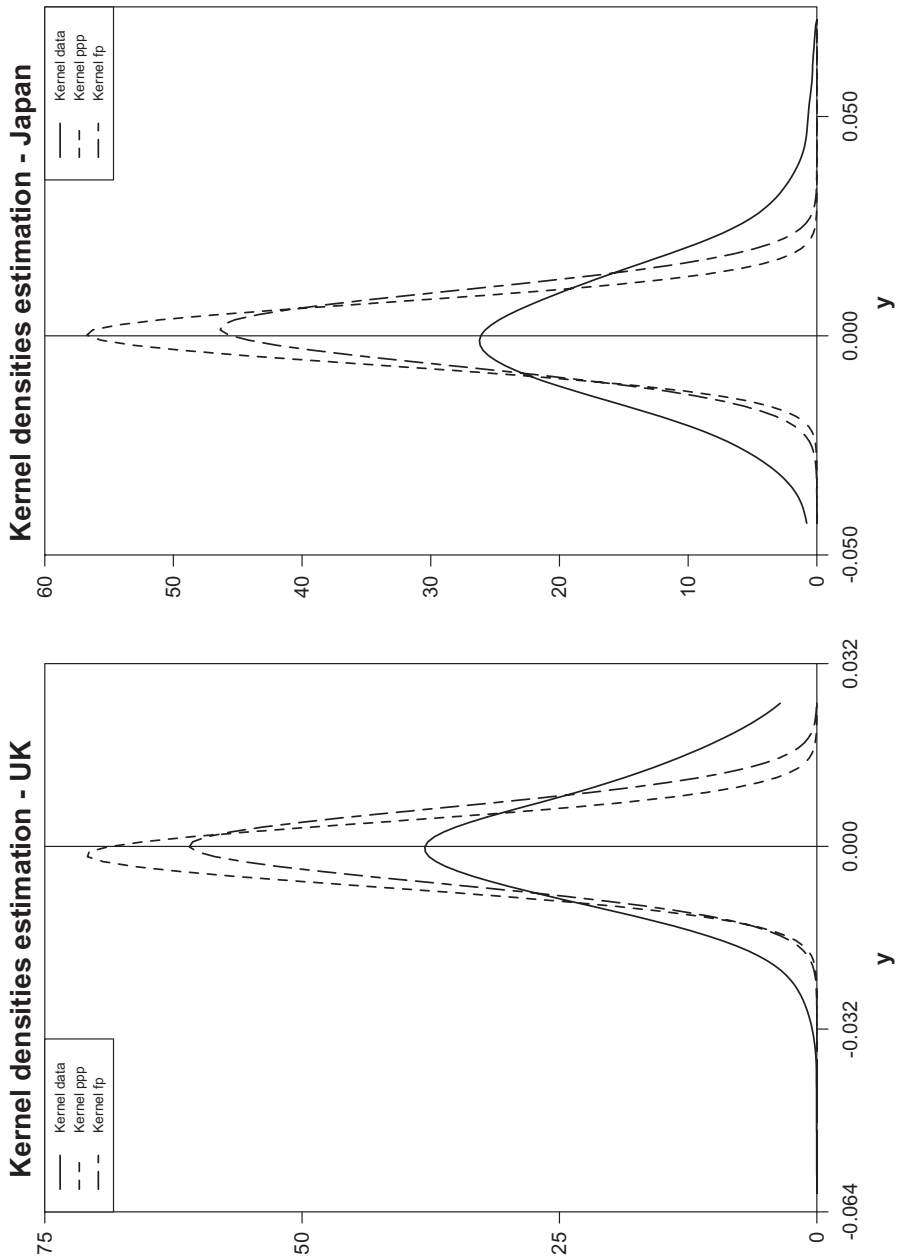

Figure 3. Forecast densities estimation

joint hypothesis. The test is easy to implement in practice, as illustrated in an application to exchange rate forecasting. Also, the test is found to have satisfactory empirical size and power properties in a simulation exercise. Nevertheless, this test circumvents the problem of testing a joint hypothesis by relying on somewhat stronger assumptions than other methods proposed in the literature that are based on the probability integral transform. Relaxation of these assumptions is an immediate avenue for future research. In particular, the assumption of time-invariance of the densities over the forecast horizon could be relaxed by using recursive kernel estimation, which would allow us to test the null hypothesis of equal density forecast accuracy on time-varying densities period by period (Yamato, 1971; Nobel *et al*., 1998).

ACKNOWLEDGEMENTS

This paper was partly written while Lucio Sarno was a Visiting Scholar at the International Monetary Fund, the Federal Reserve Bank of St. Louis and the Central Bank of Norway. The Economic and Social Research Council (ESRC) provided financial support (Grant Ref. RES-000-22-0404). The authors are grateful for comments to an anonymous referee, Kit Baum, Yin-Wong Cheung, Mike Clements, Mike Dueker, Jerry Coakley, Dick van Dijk, Ana-Maria Fuertes, Ken Wallis, Mark Wohar in addition to participants at the 2002 Society for Computational Economics Annual Conference in Aix-en-Provence. The authors alone are responsible for any errors that may remain.

REFERENCES

Anderson NH, Hall P, Titterington DM. 1994. Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density functions. *Journal of Multivariate Analysis* **50**: 41–54.

Berkowitz J. 2001. Testing density forecasts, with applications to risk management. *Journal of Business and Economic Statistics* **19**: 465–474.

Clarida RH, Taylor MP. 1997. The term structure of forward exchange premiums and the forecastability of spot exchange rates: correcting the errors. *Review of Economics and Statistics* **89**: 353–361.

Clarida RH, Sarno L, Taylor MP, Valente G. 2003. The out-of-sample success of term structure models as exchange rate predictors: one step beyond. *Journal of International Economics* **60**: 61–83.

Clements MP, Smith J. 2000. Evaluating the linear densities of linear and non-linear models: applications to output growth and unemployment. *Journal of Forecasting* **19**: 255–276.

Corker RJ, Holly S, Ellis RG. 1986. Uncertainty and forecast precision. *International Journal of Forecasting* **2**: 53–69.

De Gooijer JG, Zerom D. 2000. Kernel-based multistep-ahead predictions of the US short-term interest rate. *Journal of Forecasting* **19**: 335–353.

Diebold FX, Lopez JA. 1996. Forecast evaluation and combination. In *Handbook of Statistics 14*, Maddala GS, Rao CA (eds). Elsevier: Amsterdam; 241–268.

Diebold FX, Mariano RS. 1995. Comparing predictive accuracy. *Journal of Business and Economic Statistics* **13**: 253–263.

Diebold FX, Gunther TA, Tay AS. 1998. Evaluating density forecasts with applications to financial risk management. *International Economic Review* **39**: 863–883.

Engle RE, Granger CWJ. 1987. Co-integration and equilibrium correction representation, estimation and testing. *Econometrica* **55**: 251–276.

Franses PH, van Dijk D. 2000. *Non-linear Time Series Models in Empirical Finance*. Cambridge University Press: Cambridge.

Granger CWJ, Pesaran MH. 1999. A decision theoretic approach to forecast evaluation. In *Statistics and Finance: An Interface*, Chan WS, Lin WK, Tong, H (eds). Imperial College Press: London.
Granger CWJ, Pesaran MH. 2000. Economic and statistical measures of forecast accuracy. *Journal of Forecasting* **19**: 537–560.
Granger CWJ, Terasvirta T. 1993. *Modelling Nonlinear Economic Relationships*. Oxford University Press: Oxford.
Hall P. 1992. Effect of bias estimation on coverage accuracy of bootstrap confidence intervals for a probability density. *Annals of Statistics* **20**: 675–694.
Hardle WK, Scott DW. 1992. Smoothing by weighted averaging of rounded points. *Computational Statistics* **7**: 97–128.
Hendry DF. 1984. Monte Carlo experimentation in econometrics. In *Handbook of Econometrics*, Griliches Z, Intrilligator MD (eds). North-Holland: Amsterdam.
Hodrick RJ. 1987. *The Empirical Evidence on the Efficiency of Forward and Futures Foreign Exchange Markets*. Harwood: London.
Johansen S. 1995. *Likelihood-based Inference in Cointegrated VAR Models*. Oxford University Press: Oxford.
Kendall MG, Stuart A. 1976. *The Advanced Theory of Statistics*, Vol. 1, 4th edn. Charles Griffin and Co: London.
Kendall MG, Stuart A, Ord JK. 1987. *The Advanced Theory of Statistics*, Vols 1–2, 5th edn. Charles Griffin and Co: London.
Li Q. 1996. Nonparametric testing of closeness between two unknown distribution functions. *Econometric Reviews* **15**: 261–274.
Li F, Tkacz G. 2001. A consistent bootstrap test for conditional density functions with time-dependent data. Bank of Canada, Working Paper No. 2001–21.
Lopez JA. 2001. Evaluating the predictive accuracy of volatility models. *Journal of Forecasting* **20**: 87–109.
Mammen E. 1992. *When Does Bootstrap Work: Asymptotic Results and Simulations*. Lecture Notes in Statistics, 77. Springer-Verlag: Berlin.
Nobel AB, Morvai G, Kulkarni SR. 1998. Density estimation from an individual numerical sequence. *IEEE Transactions on Information Theory* **44**: 537–541.
Pagan A, Ullah A. 1999. *Nonparametric Econometrics*. Cambridge University Press: Cambridge.
Pesaran MH, Skouras S. 2001. Decision-based methods for forecast evaluation. In *A Companion to Economic Forecasting*, Clements MP, Hendry DF (eds). Blackwell: Oxford.
Rogoff K. 1996. The purchasing power parity puzzle. *Journal of Economic Literature* **34**: 647–668.
Rosenblatt M. 1952. Remarks on a multivariate transformation. *Annals of Mathematical Statistics* **23**: 470–472.
Sarno L, Taylor MP. 2002. Purchasing power parity and the real exchange rate. *International Monetary Fund Staff Papers* **49**: 65–105.
Sarno L, Valente G. 2004. Empirical exchange rate models and currency risk: some evidence from density forecasts. University of Warwick, mimeo.
Scott DW, Sheater SJ. 1985. Kernel density estimation with binned data. *Communication in Statistics* **14**: 1353–1359.
Silverman BW. 1982. Kernel density estimation using the fast Fourier transformation. *Applied Statistics* **31**: 93–97.
Silverman BW. 1986. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall: New York.
Tay AS, Wallis KF. 2000. Density forecasting: a survey. *Journal of Forecasting* **19**: 235–254.
Timmermann A. 2000. Density forecasting in economics and finance: Editorial. *Journal of Forecasting* **19**: 231–234.
Wallis KF. 2003. Chi-squared tests of interval and density forecasts, and the Bank of England's fan charts. *International Journal of Forecasting* **19**: 165–175.
Wand MP, Jones MC. 1995. *Kernel Smoothing*. Chapman and Hall: New York.
Weigend AS, Shi S. 2000. Predicting daily probability distributions of S&P500 returns. *Journal of Forecasting* **19**: 375–392.
Yamato H. 1971. Sequential estimation of a continuous probability function and mode. *Bulletin of Mathematical Statistics* **14**: 1–12.

Authors' biographies:
Lucio Sarno is Professor of Finance and Chairman of the Accounting and Finance Group, Warwick Business School, University of Warwick and a Research Affiliate of the CEPR in London. Professor Sarno

has published over 40 papers in refereed economics and finance journals and several books. Consultancy work includes projects for the IMF, World Bank, US Federal Reserve, Italian Ministry of Finance and the Central Bank of Norway.

Giorgio Valente is Lecturer in Finance, Warwick Business School, University of Warwick. His publications include the *Journal of International Economics, Journal of Applied Econometrics, Journal of Futures Markets*.

Authors' address:
Lucio Sarno and Giorgio Valente, Warwick Business School, University of Warwick, Coventry CV4 7AL, UK.

