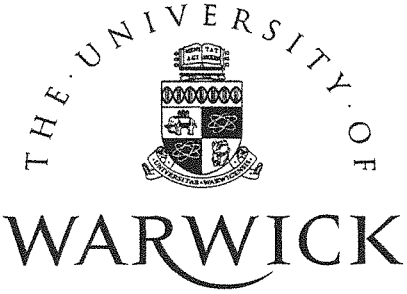
EVALUATING THE FORECAST OF DENSITIES OF LINEAR AND NON-LINEAR MODELS: APPLICATIONS TO OUTPUT GROWTH AND UNEMPLOYMENT

Michael P. Clements and Jeremy Smith

No.509

WARWICK ECONOMIC RESEARCH PAPERS



DEPARTMENT OF ECONOMICS

EVALUATING THE FORECAST DENSITIES OF LINEAR AND NON-LINEAR MODELS: APPLICATIONS TO OUTPUT GROWTH AND UNEMPLOYMENT

Michael P. Clements and Jeremy Smith
Department of Economics
University of Warwick
Coventry
ENGLAND

No.509

March 1998

Evaluating the forecast densities of linear and non-linear models: Applications to output growth and unemployment

Michael P. Clements and Jeremy Smith*
Department of Economics,
University of Warwick,
Coventry, CV4 7AL.

March 12, 1998

Abstract

In economics density forecasts are rarely available, and as a result attention has traditionally focused on point forecasts of the mean and the use of mean square error statistics to represent the loss function. We extend the methods of forecast density evaluation in Diebold, Gunther and Tay (1997) to compare linear and non-linear model-based forecasts of US output growth and changes in the unemployment rate. Of prime concern is whether concentrating solely on the first moment obscures the potential usefulness of non-linear models as forecasting devices.

1 Introduction

Traditionally economic forecasting has been primarily concerned with the production and evaluation of point forecasts (see, e.g., Wallis, 1995). While the point forecast is the 'most likely' outcome, in most practical settings it is hard to imagine that the user of the forecasts will be indifferent to the likely *range* of outcomes. Summary information regarding the degree of uncertainty associated with forecasts, such as standard errors, is sometimes provided. Chatfield (1993) surveys the large literature on calculating interval forecasts, where an interval forecast comprises an upper and lower bound to define the interval, together with the probability that the actual outcome falls within that interval. However, it is only recently that Christoffersen (1997) has analysed how to evaluate such forecasts. Christoffersen (1997) suggests that a 'good' interval forecast should have correct *conditional* coverage – the interval is wider in volatile periods than in those of greater tranquility – so that the occurrences of observations outside the interval are not clustered in the former and completely absent from the latter. Continuing this trend toward providing a more complete description of the uncertainty surrounding forecasts, Diebold *et al.* (1997) propose methods for evaluating density forecasts, and Diebold, Tay and Wallis (1997) apply these ideas to survey-based inflation forecasts.

Evaluating the complete forecast density function could be particularly relevant for comparisons of forecasts from non-linear and linear models since mean squared forecast error (MSFE) statistics of point forecasts may fail to do the non-linear models justice. Assessed on MSFE, the

*The first author acknowledges financial support under ESRC grant L116251015. We are grateful to David Cox, Martin Cripps and Ken Wallis for helpful comments. Email: M.P.Clements@warwick.ac.uk and Jeremy.Smith@warwick.ac.uk, tel.: 01203 523055, FAX: 01203 523032.

forecasting ability of non-linear models is often not clearly better than that of linear alternatives, notwithstanding the apparent ability of such models to characterise business cycle features of the data. Thus, the survey by De Gooijer and Kumar (1992) concludes that 'no uniformity seems to exist in the evidence presented on the forecasting ability of non-linear models'. Diebold and Nason (1990) give four reasons why non-linear models may fail to forecast better than the simplest linear model even when linearity is rejected statistically in-sample. First, there may be non-linearities in the *even*-ordered conditional moments, which cannot be exploited for improved first-moment prediction. Secondly, non-linearities signalled by statistical testing could be due to outliers or structural breaks. Thirdly, the conditional-mean non-linearities, while a feature of the data generating process (DGP), are not 'large' enough to yield forecasting gains. Fourthly, the wrong type of non-linear model is used. Portmanteau tests may reject linearity but will not suggest the appropriate non-linear alternative, and some tests designed to have power against a specific non-linear model may reject against other non-linear models (see Clements and Smith, 1998).

Clements and Smith (1996) control for a number of these factors in a simulation study, and conclude that the third reason is relevant in a number of cases. That is, if we take the DGP to be an empirical non-linear model, once the model has to be specified and its parameters estimated on the data, then the non-linear model rarely performs better (on MSFE) than a (mis-specified) linear model. However, forecast performance is significantly affected by where the process is at the time the forecast is made, and such models sometimes fare better on qualitative, 'direction-of-change' statistics.

The aim of this paper is to compare linear and non-linear models in terms of their density forecasts, using the technology proposed in Diebold *et al.* (1997). Because economic time series tend to be inter-related, improved forecasts can often be obtained by modelling systems of equations. This naturally gives rise to joint forecast densities, and an innovation in this paper is a proposed method of evaluating such densities, which extends Diebold *et al.* (1997).

In section 2 we discuss the results in the literature on comparisons based on traditional MSFE measures, for the series we consider, as well as other ways of evaluating the non-linear model forecasts. We also formally test for the presence of more than one regime in the series of interest, to see if we can reject a linear characterization of the series in favour of the type of non-linear model we are interested in. Section 3 discusses linear and non-linear systems approaches to modelling. In section 4 we apply and extend the forecast density evaluation approach of Diebold *et al.* (1997), the results of which are reported in section 5. Section 6 concludes.

2 Comparing the forecast performance of linear and non-linear models

In this section we consider various ways of comparing the forecast performance of non-linear and linear models. We illustrate with reference to the results in the literature for the two series we consider – the first difference of the logarithm of US GNP, i.e., the growth rate of output (denoted x), and the change in the US unemployment rate (u). These series are ideal for our purposes, since they have been modelled using self-exciting threshold autoregressive (SETAR) models[1], and this is the class of non-linear model we focus on in this paper, and because non-

[1]Such models were first proposed by Tong (1978), Tong and Lim (1980) and Tong (1983) – see also Tong (1995).

linear system models of the two series have appeared in the literature. Section 2.1 contains a brief outline of the SETAR class of model. Section 2.2 considers traditional MSFE-based measures, section 2.3 evaluation in terms of correctly predicting the direction of change of the variable, or the regime, and section 2.4 reports on impulse response analysis for non-linear models. Section 2.5 formally tests for a linear model against a two-regime SETAR model.

2.1 SETAR models

Briefly, the SETAR model class supposes the series can be modelled as a number of distinct regimes, where the regimes are characterised by different conditional distributions of the process, each parameterised by an autoregression. For the SETAR model, the regimes depend on observable lagged values of the process, in contrast to the Hamilton (1989) model, where the regime-switching process is an unobservable discrete first-order Markov process. The SETAR assumption greatly facilitates the estimation of the model.

When there are two regimes, then the process is in regime $i = 1$ at period t when $y_{t-d} \leq r$, and otherwise ($y_{t-d} > r$) in regime $i = 2$:

$$y_t = \phi_0^{\{i\}} + \phi_1^{\{i\}} y_{t-1} + \ldots + \phi_p^{\{i\}} y_{t-p} + \epsilon_t^{\{i\}}, \quad \epsilon_t^{\{i\}} \sim \mathsf{IN}\left(0, \sigma^{2\{i\}}\right), \quad i = 1, 2 \qquad (1)$$

where the parameters super-scripted by $\{i\}$ may vary across regime. The orders of the autoregressions may differ across regimes (so that p is the maximum lag order and some of the $\phi_j^{\{i\}}$ may be zero for some i). We make the assumption that the disturbances are independent normal. This is stronger than the iid assumption sometimes made. It will allow us to simulate the model by drawing normal random variates. An assumption of normality is made for the disturbances of the linear AR model and for the systems models.

Stationarity and ergodicity conditions are discussed in, e.g., Tong (1995). In the following, y_t is the first difference of (log) output (x_t), or the change in the unemployment rate (u_t)[2].

2.2 MSFE measures and regime-dependence

Tiao and Tsay (1994) consider a two-regime SETAR model and a four-regime refinement, where $p = 2$. Potter (1995) estimates a SETAR(2; 5, 5) but with the third and fourth lags restricted to zero under both regimes, and $d = 2$ and $r = 0$. The estimation period in both instances is 1947 – 90. A noteworthy feature of SETAR models of US GNP over this period is a large negative coefficient on the second lag in the lower regime, indicating that the US economy moves swiftly out of recession. The empirical forecast performance of the SETAR model relative to a linear AR model is markedly improved when the comparison is made in terms of how well they forecast when the economy is in recession, illustrating the now well-known path dependence of the outcome of such comparisons. Clements and Smith (1996, 1997) find evidence for this effect in empirical and Monte Carlo analyses of the forecast performance of SETAR and linear models. If forecasts are not evaluated conditional upon the regime, then the gains in the minority regime need to be sufficiently large to ensure that the SETAR performs well on average.

In the SETAR model switching is exogenous. Following on from the work of Beaudry and Koop (1993), Pesaran and Potter (1997) develop a three-regime SETAR model with an

[2]The precise data definitions are as follows: X is seasonally-adjusted real US GNP. We splice together data for 1947 – 90 at 1982 prices (as used by, e.g., Potter, 1995) with a more recent vintage of data at 1992 prices for the period 1959 – 96. The US unemployment rate series is the quarterly series used by Montgomery, Zarnowitz, Tsay and Tiao (1996) for the period 1948 – 93.

endogenously changing floor and ceiling, whereby the ceiling regime is in force when the growth rate is fast, so that the economy is overheating, and the floor when recent output growth has been slow or negative. On the RMSFE for mean prediction, their model fares worse than a random walk with drift. However, as they note, the forecast period of 1993 – 96 was unusually calm by historical standards and contained no recessions. Moreover, the RMSFE for their model for predicting the conditional variance of the process was superior to that of linear models.

2.3 Predicting the direction of change

Regime-switching models may be better suited to predicting movements between regimes, rather than small movements within a regime. Direction of change tests[3] are one way of capturing this idea. The tests are closely related to the standard χ^2 test of independence between actual and predicted directions of change based on the 2×2 contingency table. Clements and Smith (1996) report these tests and related measures for forecasts from linear and non-linear models of a number of economic and financial variables. Pesaran and Potter (1997) consider the number of times the 'floor and ceiling' and linear models of US GNP correctly predict negative output growth (loosely, 'recessions') in-sample, and find the non-linear model offers an improvement on this criterion.

Nevertheless, such criteria count one-for-one a forecast of a very small increase, when in fact a small decline occurred, with a forecast of a large increase when a large decline occurred, (if the regimes are positive and negative growth, say), so that such measures would appear a useful complement to, rather than substitute for, MSFE measures.

2.4 Impulse response analysis

Rather than comparing the forecast performance of linear and non-linear models, some authors have focused on the different implications for the propagation of shocks. Koop, Pesaran and Potter (1996) develop generalized non-linear impulse response functions (GIs) (see also Gallant, Rossi and Tauchen, 1993) to analyse the response of non-linear models to shocks. Their analysis recognises that linear impulse response analysis is inappropriate, since the impact of the shock is dependent upon the sign and size of the shock, and the position of the process when the shock hits. The non-linear models of US GNP of Beaudry and Koop (1993), Potter (1995) and Pesaran and Potter (1997) are shown to imply different degrees of persistence between shocks occurring in recessions and booms.

2.5 Testing for more than 1 regime

Hansen (1996) presents a framework for testing the null of linearity against the alternative of threshold autoregression, that delivers valid inference when the nuisance parameters r and d are unknown. The testing procedure is non-standard because these nuisance parameters are unidentified under the null (of a single regime – linearity). The full-sample results are recorded in table 1, where we report the p-values for the $\sup T_T$, $\text{ave} T_T$ and $\exp T_T$ statistics of the null of linearity (see Hansen, 1996 for details).

[3]These were developed in the context of predicting rates of return on market investments by Henriksson and Merton (1981). Schnader and Stekler (1990) and Stekler (1994) applied the approach to macroeconomic prediction, and Pesaran and Timmermann (1992) suggested a number of refinements and extensions.

Table 1 Asymptotic p-values of linear null versus SETAR model.

Series	Output growth – x	Unemployment rate changes – u
SETAR model	$(2;2,2)$	$(2;2,2)$
Sample	1947 – 94	1948 – 93
	Robust LM Statistics	
SupT_T	0.616	0.013
ExpT_T	0.519	0.007
AveT_T	0.506	0.008
	Standard LM Statistics	
SupT_T	0.079	0.001
ExpT_T	0.158	0.001
AveT_T	0.321	0.001

The results were obtained using Bruce Hansen's Gauss code `tar.prg`. Similar results for the US unemployment rate were obtained if instead the lag order was set at 4.

The null of a single regime is clearly rejected for u_t. But the evidence for SETAR non-linearities in x_t is weak, confirming the finding of Hansen (1996) – on any test the null of linearity is not rejected at the 5% level. These results are similar to those reported by Potter (1995, Table IV, p.115.). However, Potter (1995) (same table) also records Monte Carlo evidence indicating that the tests are too conservative, particularly the heteroscedasticity-robust versions, and that the power at the nominal 5% level is low. Correcting for size, he finds evidence in favour of non-linearity at the 10% level. A similar size correction here might suggest the same conclusion.

3 Multivariate Models

Since there are likely to be important interactions between output and unemployment, we extend the analysis to multivariate models, and consider systems of equations in the change in the unemployment rate and GNP growth. One of the simplest is a bivariate vector-autoregression (VAR):

$$\mathbf{Y}_t = \mathbf{A}(L)\mathbf{Y}_{t-1} + \varepsilon_t \tag{2}$$

where $\mathbf{Y}_t = [u_t \quad x_t]'$. A third-order model appeared adequate, so $\mathbf{A}(L)$ is a second-order matrix lag polynomial and ε_t is assumed to be an independently normally distributed error process with covariance matrix:

$$E\left[\varepsilon_t \varepsilon_t'\right] = \sum = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}. \tag{3}$$

There are few multivariate threshold autoregressive models, but Koop *et al.* (1996) suggest the following model[4]:

$$\mathbf{Y}_t = \mathbf{A}(L)\mathbf{Y}_{t-1} + \mathbf{B}\mathbf{Z}_{t-1} + \mathbf{H}_t\epsilon_t$$

where $\mathbf{A}(L)$ is as before, and $\mathbf{B}$ is a (2×2) matrix containing the coefficients on the non-linear terms $\mathbf{Z}_t = [CDR_t, OH_t]'$, which are defined by the following recursions (see Pesaran and Potter, 1997):

[4]This model is a multivariate extension of the univariate model of Beaudry and Koop (1993) and Pesaran and Potter (1997).

$$
\begin{aligned}
F_t &= \begin{cases} 1(x_t < r_f) & \text{if } F_{t-1} = 0 \\ 1(CDR_{t-1} + x_t < 0) & \text{if } F_{t-1} = 1 \end{cases} \\
CDR_t &= \begin{cases} (x_t - r_f)F_t & \text{if } F_{t-1} = 0 \\ (CDR_{t-1} - x_t)F_t & \text{if } F_{t-1} = 1 \end{cases} \\
C_t &= 1(F_t = 0)1(x_t > r_c)(x_{t-1} > r_c) \\
OH_t &= C_t(OH_{t-1} + x_t - r_c) \\
COR_t &= 1(F_t + C_t = 0)
\end{aligned}
$$

where $1(A)$ is an indicator function taking the value unity when the statement in brackets is true and zero otherwise, r_c and r_f are the ceiling and floor threshold values. The error process $\epsilon_t \sim IN(0, I)$ and the variance-covariance matrix of the error term is such that:

$$H_t = (H_0 COR_{t-1} + H_1 F_{t-1} + H_2 C_{t-1})$$

where:

$$H_j = \begin{bmatrix} \sigma^2_{j11} & \sigma_{j12} \\ \sigma_{j12} & \sigma^2_{j22} \end{bmatrix} \qquad j = 0, 1, 2$$

and:

$$
\begin{aligned}
\sigma^2_{011} &= \tfrac{1}{T_0} \sum_{t=1}^{T} COR_{t-1}(u_t - \mathbf{A}(L)_1 \mathbf{Y}_{t-1})^2 \\
\sigma^2_{022} &= \tfrac{1}{T_0} \sum_{t=1}^{T} COR_{t-1}(x_t - \mathbf{A}(L)_2 \mathbf{Y}_{t-1})^2 \\
\sigma_{012} &= \tfrac{1}{T_0} \sum_{t=1}^{T} COR_{t-1}(u_t - \mathbf{A}(L)_1 \mathbf{Y}_{t-1})(x_t - \mathbf{A}(L)_2 \mathbf{Y}_{t-1}) \\
\sigma^2_{111} &= \tfrac{1}{T_1} \sum_{t=1}^{T} F_{t-1}(u_t - \mathbf{A}(L)_1 \mathbf{Y}_{t-1} - B_{11} CDR_{t-1})^2 \\
\sigma^2_{122} &= \tfrac{1}{T_1} \sum_{t=1}^{T} F_{t-1}(x_t - \mathbf{A}(L)_2 \mathbf{Y}_{t-1} - B_{21} CDR_{t-1})^2 \\
\sigma_{112} &= \tfrac{1}{T_1} \sum_{t=1}^{T} F_{t-1}(u_t - \mathbf{A}(L)_1 \mathbf{Y}_{t-1} - B_{11} CDR_{t-1})(x_t - \mathbf{A}(L)_2 \mathbf{Y}_{t-1} - B_{21} CDR_{t-1}) \\
\sigma^2_{211} &= \tfrac{1}{T_2} \sum_{t=1}^{T} C_{t-1}(u_t - \mathbf{A}(L)_1 \mathbf{Y}_{t-1} - B_{12} OH_{t-1})^2 \\
\sigma^2_{222} &= \tfrac{1}{T_2} \sum_{t=1}^{T} C_{t-1}(x_t - \mathbf{A}(L)_2 \mathbf{Y}_{t-1} - B_{22} OH_{t-1})^2 \\
\sigma_{212} &= \tfrac{1}{T_2} \sum_{t=1}^{T} C_{t-1}(u_t - \mathbf{A}(L)_1 \mathbf{Y}_{t-1} - B_{12} OH_{t-1})(x_t - \mathbf{A}(L)_2 \mathbf{Y}_{t-1} - B_{22} OH_{t-1})
\end{aligned}
$$

$\mathbf{A}(L)_r$, $r = 1, 2$ is the r-th row of $\mathbf{A}(L)$, B_{ij} is the ijth element of $\mathbf{B}$ $(i, j = 1, 2)$, and $T_0 = \sum_{t=1}^{T} COR_{t-1}$, $T_1 = \sum_{t=1}^{T} F_{t-1}$, $T_2 = \sum_{t=1}^{T} C_{t-1}$. To simplify estimation, we follow Koop *et al.* (1996) in taking the floor (r_f) and ceiling (r_c) threshold values as given, from Pesaran and Potter (1997), at -0.8755 and 0.5391, respectively. The model can then be estimated using an iterative GLS procedure, with iterations limited to 10, which appears to be sufficient to ensure

Table 2 NVAR model – full-sample statistics.

	x_t		u_t	
	Coeff.	t-value	Coeff.	t-value
Intercept	0.276	1.636	0.097	1.899
x_{t-1}	0.199	1.747	−0.041	−1.216
x_{t-2}	0.300	2.798	−0.093	−2.992
x_{t-3}	0.050	0.513	−0.001	−0.028
u_{t-1}	−1.043	−3.176	0.510	5.413
u_{t-2}	0.400	1.188	0.004	0.042
u_{t-3}	0.248	0.964	−0.209	−2.984
CDR_{t-1}	−0.640	−2.614	0.072	0.603
OH_{t-1}	−0.044	−0.993	0.009	0.789

The error variances for the VAR equations for x and u were 0.8971 and 0.0883, compared to error variances of 0.8604, 1.3309 and 0.5639 for x, and 0.0988, 0.2176 and 0.0415, for u, in the corridor, floor and ceiling regimes, respectively.

convergence of all parameters. This model is referred to as the NVAR, to distinguish it from the VAR.

Table 2 provides some statistics relating to the estimated NVAR. On the basis of the full-sample estimates, the non-linear terms are not significant in the u_t equation. Even so, there are clear regime-dependencies in the equation's error variances, which may play an important role in constructing the forecast densities. Note also that there are significant interactions between x and u, in both directions, suggesting the systems approaches (VAR and NVAR) may do better than the univariate AR and SETAR models.

The MSFE results reported in section 5 do not favour the non-linear models, and in this respect match those of De Gooijer and Kumar (1992), Pesaran and Potter (1997) and Clements and Smith (1997), *inter alia* – allowing for non-linearity yields little or no gains in terms of MSFE when the implicit aim of the exercise is forecasting the mean of the future distribution of the variable of interest. However, as presaged in the introduction, such a negative conclusion may be unwarranted if the non-linear model performs well in terms of predicting the overall density function, rather than simply the first moment, to which we now turn.

4 Evaluating density forecasts

We evaluate the model-based forecasts using the approach of Diebold *et al.* (1997). This requires calculating the probability integral transforms of the actual realizations of the variables (changes in unemployment or GNP growth) over the forecast period ($\{y_t\}_{t=1}^n$, $t = 1, \ldots, n$) with respect to the forecast densities of the SETAR and AR models, denoted by $\{p_t(y_t)\}_{t=1}^n$. That is, we evaluate:

$$\{z_t\}_{t=1}^n = \left\{ \int_{-\infty}^{y_t} p_t(u)du \right\}_{t=1}^n . \tag{4}$$

When the model forecast density corresponds to the true predictive density (given by the DGP, and denoted by $f_t(y_t)$), i.e., $p_t(y_t) = f_t(y_t)$, then Diebold *et al.* (1997) show that $\{z_t\}_{t=1}^n \sim$

iid$U[0,1]$. The result that $z_t \sim U[0,1]$ can be found in, e.g., Kendall, Stuart and Ord, 1987, sections 1.27 and 30.36. Diebold *et al.* (1997) make the result operational in the time series context by establishing that the z_t sequence are independent when the true densities are used at each t. Hence the idea is to evaluate the forecast density by assessing whether there is statistically significant evidence that the realizations do not come from that density – this amounts to testing whether the $\{z_t\}$ series depart from the iid uniform assumption.

For the AR univariate model, $p_t^h(y_t)$ (where the h-subscript denotes the h-step ahead forecast density) can be calculated analytically. Ignoring parameter estimation uncertainty and assuming the AR model disturbances are independently normally distributed, $p_t^h(y_t)$ is gaussian with mean and variance given by simple functions of the estimated model parameters and equation standard error. While the SETAR model $p_t^h(y_t)$ cannot be found analytically, it can be calculated by simulation[5]. In fact for each t we simulate both univariate model densities by Monte Carlo (setting the number of replications to 500, and drawing gaussian errors), and evaluate (4) using these densities.

Joint densities of the change in unemployment and output growth are simulated by Monte Carlo from the systems models – the VAR and NVAR. We can then calculate the z-values for x and u separately, ignoring the joint nature of the forecast density – these are referred to as 'marginal' z-values. As an extension to the evaluation framework of Diebold *et al.* (1997) we propose the use of conditional z-values, as an additional aid to forecast evaluation. Writing the conditional forecast density of, say, x given u takes on its realized value as $p_{X|U}(x \mid U = u)$, we integrate the realized value of x_t against the model forecast density:

$$\{z_{x,t}\}_{t=1}^n = \left\{ \int_{-\infty}^{x_t} p_{X|U}(w \mid U_t = u_t) dw \right\}_{t=1}^n . \tag{5}$$

When $p_{X|U}(x \mid U_t = u_t)$ corresponds to the actual conditional forecast density ($f_{X|U}(x \mid U_t = u_t)$), it follows immediately that $\{z_{x,t}\}_{t=1}^n \sim$ iid$U[0,1]$. The conditional density is generated as follows. Under the null that the model (i.e., the VAR or NVAR) is correctly specified (and ignoring parameter estimation uncertainty), then the 1-step ahead forecast error vector will be the realized disturbance term. For example, in the VAR given by (2) where $\boldsymbol{\varepsilon}_t = [\varepsilon_{u,t}\ \varepsilon_{x,t}]' \sim IN(\mathbf{0}, \sum)$ adopt the notation:

$$E\left[\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t'\right] = \sum = \begin{bmatrix} \sigma_u^2 & \sigma_{ux} \\ \sigma_{ux} & \sigma_x^2 \end{bmatrix}. \tag{6}$$

Then $p_{X|U}(x \mid U_t = u_t)$ can be simulated by drawing random disturbances for $\varepsilon_{x,t}$ from:

$$\varepsilon_{x,t} \sim N\left(\rho \varepsilon_{u,t} \frac{\sigma_x}{\sigma_u}, \sigma_x^2 \left(1 - \rho^2\right)\right) \tag{7}$$

where $\rho = \sigma_{ux}/\sigma_u \sigma_x$.

For the NVAR, $\sum$ depends on the regime.

This sampling procedure allows us to calculate sequences for both $\{z_{x,t}\}$ and $\{z_{u,t}\}$ for 1-step ahead forecasts. Multi-step conditional z's can be calculated for the linear VAR in this

[5]See, e.g., Tiao and Tsay (1994) and Clements and Smith (1997). Gallant *et al.* (1993) and Koop *et al.* (1996) provide analyses of the construction of conditional densities for non-linear time series models, in the context of impulse response analysis.

way. The 2-step ahead forecast error vector (of period t conditional on $t-2$) is $\boldsymbol{\varepsilon}_t + \mathbf{A}\boldsymbol{\varepsilon}_{t-1}$, with $E[\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t' \mid t-2] = \sum_2 = \sum + \mathbf{A}\sum\mathbf{A}'$. Hence the conditional 2-step ahead density for x given $U_t = u_t$ can be simulated by drawing $\varepsilon_{x,t|t-2}$ from a normal distribution with moments as functions of $(\varepsilon_{u,t|t-2}, \sum_2)$ – *c.f.* (7). The NVAR can not be treated in this way, because even assuming $\boldsymbol{\varepsilon}_t$ is gaussian the 2-step ahead error distribution will not have a simple form.

As a further extension to the evaluation framework of Diebold *et al.* (1997) we consider the use of 'joint' z-values. Writing the joint forecast density as $p_{x,u}(x,u)$, we integrate the realized pair $\{x_t, u_t\}$ against the model forecast density:

$$\{z_{j,t}\}_{t=1}^{n} = \left\{ \int_{-\infty}^{x_t} \int_{-\infty}^{u_t} p_{x,u}(w_1,w_2)dw_2dw_2 \right\}_{t=1}^{n} . \tag{8}$$

The 1-step forecasts of x and u are normal conditional on period $t-1$, given the assumption of normality of the disturbances. If the covariance matrix of the disturbances is diagonal, then the 1-step forecasts are independent and the joint can be written as the product of the two marginals. Then when $p_{x,u}(x,u)$ corresponds to the actual joint forecast density ($f_{x,u}(x,u) = f_x(x) \times f_u(u)$), we can show that:

$$\{z_{j,t}\}_{t=1}^{n} \sim \mathsf{iid}\,(-\ln W) \tag{9}$$

where W is the distribution function for the product of two independent $U[0,1]$ uniform random variables – see Appendix A – which enables a 'test' of the joint forecast density. As for the univariate transforms, the $z_{j,t}$ are calculated by integrating the simulated densities for $p_{x,u}(x,u)$.

Typically, as is the case here, the forecasts will be correlated, and under the null of correct specification the marginal z's will be $U[0,1]$ but not independent, so (9) is inapplicable. We use simulation to obtain a reference distribution for the joint z's under the null of correct specification, against which the distribution of the empirical joint z's can be compared.

Finally, notice that similar problems arise for sequences of multi-step forecasts. For a k-step ahead forecast horizon there will only be n/k independent forecasts, since optimal forecasts will exhibit $k-1$ order dependence. Thus 2-step forecasts of t and $t+1$ will not be independent, and even under correct specification the corresponding z's, although $U[0,1]$, will not be independent. We proceed by dividing the z's for the 2-step forecasts in to two sets, taking the first, third etc. for the first set, and putting the remainder in the second. When the z's are evaluated and plotted, we therefore proceed as if we have two unrelated sets. To save space, in each instance we report results for the set which offers the most evidence against the null. Alternatively, we obtain the theoretical cdf of sequences of 2-step forecasts by simulation, as suggested in the previous paragraph for the joint z's. Appendix B details a method of treating sequences of multi-step forecasts from linear models, by adapting the method of conditioning described above, but this approach is not pursued in the empirical work.

5 Results

The SETAR model used for US GNP growth is similar to Potter (1995), although the process is a second-order autoregression in both regimes (the lag 5 terms are excluded). We set $d = 2$ but r and the autoregressive coefficients are determined from the data. The sample period is 1947:1 – 94:4. The model is first estimated on the available sample up to 1977:2, and forecasts generated for 1 through to 5-steps ahead. The end date is then extended by one observation,

so the model is estimated up to 1977:3, and 1- through to 5-step ahead forecasts are calculated taking 1977:3 as the forecast origin. Continuing in this fashion we obtain seventy 1- through to 5-step ahead forecasts, where the final 1-step forecast is of 1992:2 and the final 5-step forecast is of 1993:2.

Montgomery *et al.* (1996) estimated SETAR models of US unemployment on monthly and quarterly data. Their model was a 'TAR component' model, although from a forecasting perspective the computationally simpler SETAR model fares little worse. Montgomery *et al.* (1996) find the model capable of recording gains relative to a linear model when unemployment is rising rapidly in the contraction phase of the business cycle. We consider a three-regime model on quarterly data, where the process is assumed to be a second-order autoregression in each regime, and we set $d = 2$. The threshold values r_1, r_2 and the autoregressive coefficients are estimated from the data. The model is estimated and forecast over the same periods as that for output, described above, as are the VAR and NVAR models.

Table 3 reports the empirical MSFEs for the linear and non-linear univariate models and systems for 1 to 5-step ahead forecasts. The multivariate approaches fair better than the univariate at short horizons, while the non-linear models are never much better than their linear counterparts (and sometimes much worse).

Table 3 MSFE Performance of SETAR and Linear AR Models.

h	AR(2)	SETAR	VAR(2)	NVAR
		GNP		
1	0.762	0.827	0.599	0.588
2	0.834	0.890	0.806	0.880
3	0.851	0.902	0.826	1.007
4	0.858	0.872	0.846	1.117
5	0.862	0.875	0.881	1.204
		Unemployment		
1	0.071	0.077	0.066	0.064
2	0.103	0.109	0.092	0.088
3	0.114	0.115	0.104	0.106
4	0.116	0.119	0.113	0.122
5	0.114	0.116	0.113	0.129

We now move on to an evaluation of the z-transforms. Tests for the null hypothesis of iid uniformity is a joint hypothesis. Diebold *et al.* (1997) argue that tests of iid$U[0,1]$, such as $\sum_{t=1}^{n} -2\log z_t \sim \chi^2_{2n}$ under the null, may often be of little practical value, since it will not be apparent which part of the null (iid or uniformity) is at odds with data, and instead advocate more informal data analysis. We follow their suggestion, and consider each part of the hypotheses in turn.

5.1 Testing for independence

The iid assumption is tested using LM tests for serial correlation, and since dependence may occur in higher moments, we also consider $(z - \overline{z})^j$ for integer j up to 3. Table 4 reports LM tests for autocorrelation up to fourth and second order for the 1- and 2-step ahead z-values,

respectively. For both the GNP and unemployment series we are unable to reject the null hypothesis of no serial correlation in any of the first 3 moments of the marginal or conditional z's for any of the models at the conventional 5% level. The iid assumption is rejected for the VAR joint z's. At the 10% significance level there is some evidence of serial correlation in the third moment of the AR 1-step unemployment z's.

Table 4 LM Tests for iid.

	AR	SETAR	VAR	NVAR	AR	SETAR	VAR	NVAR	VAR	NVAR
					1-step					
Moment		Unemployment					GNP		Joint	
1	0.434	0.826	0.980	0.513	0.991	0.911	0.860	0.963	**0.010**	0.176
2	0.218	0.859	0.782	0.512	0.928	0.813	0.726	0.944	**0.053**	0.235
3	**0.077**	0.838	0.431	0.571	0.738	0.622	0.640	0.884	0.133	0.244
			Unemp.|GNP				GNP|Unemp.			
1			0.137	0.651			0.513	0.519		
2			0.429	0.437			0.365	0.349		
3			0.530	0.381			0.292	0.337		
					2-step					
		Unemployment					GNP			
1	0.349	0.353	0.356	0.504	0.594	0.899	0.808	0.777		
1	**0.099**	0.173	0.601	0.423	0.986	0.499	0.857	0.839		
2	0.267	0.896	0.553	0.327	0.834	0.632	0.735	0.397		
2	0.137	0.258	0.414	0.263	0.954	0.664	0.941	0.976		
3	0.285	0.928	0.466	0.297	0.853	0.662	0.613	0.330		
3	0.167	0.303	0.364	0.482	0.915	0.756	0.934	0.969		

The table records the p-values for χ^2 LM tests of serial corelation up to fourth-order (1-step) and second-order (2-step).

5.2 Assessing uniformity

We assess the uniformity aspect (conditional on the iid part) by plotting the actual cdfs of the z_t's against the theoretical cdfs (the 45° line). The 95% confidence intervals drawn alongside the 45° lines are based on the critical values tabulated by Miller (1956).[6] Figure 1 plots the cdfs for the univariate models. For 1-step forecasts the AR model empirical cdfs for both x and u hit the 95% confidence intervals, while the SETAR cdfs are clearly interior. However the 2-step ahead SETAR cdf for x touches the boundary.

Figure 2 plots a selection of cdfs for the VAR. The marginals for x and u at 1-step, and for u at 2-step, are borderline, and the cdf for $x \mid u$ clearly crosses the lower boundary. The theoretical cdf and confidence intervals for the joint z are obtained by simulation assuming the model in question is the data generating process.[7] While there is no evidence against the distributional assumption, the confidence intervals are invalidated by the rejection of iid for the empirical z's.

[6] Miller (1956, Table 1) reports exact critical values of Kolmogorov Statistics for small sample sizes, n up to 100. The 95% confidence intervals are the 45° line ± the relevant values.

[7] 1000 sets of 70 1-step joint z's are simulated. Each set of 70 is sorted separately and cumulated. The theoretical cdf is traced out by the median across the 1000 sets for each of the 70 points, and the 95% interval is the 25^{th} and 975^{th} largest of the 1000 sets at each of the 70 points.

Figure 3 plots the same selection of cdfs for the NVAR. These suggest fewer violations of the distributional assumptions. As for the VAR, the cdf for $x \mid u$ crosses the lower boundary. The cdf for the joint z signals problems not evident from the marginals and conditionals – there are too few realized pairs of values with a low probability of occurring under the NVAR. Put another way, the spread of the NVAR density is too wide.

In general the forecast density evaluation criteria appear to favour the non-linear models, notwithstanding the poor performance of these models on MSFE comparisons. This finding is consistent with the results of Pesaran and Potter (1997, p.685), who report that 'The 'worst' model for mean prediction by the RMSFE criterion is the floor and ceiling model', while 'The 'best model' for predicting the conditional variance of the growth rates is clearly the floor and ceiling model'.

6 Conclusions

This paper looks at the forecast performance of linear and non-linear, univariate and systems, models of the change in the US unemployment rate, and the US GNP growth rate, using quarterly data over the period 1948 – 94. On an MSFE evaluation of conditional mean predictions, the non-linear models rarely yield much of an improvement over the linear models. Not unsurprisingly, the multivariate models are better than the univariate.

However, using the results of Diebold *et al.* (1997) we found that the non-linear multivariate model appears to provide a better characterisation of the density of future realizations of the variables over this period than does the VAR. While there was little evidence against the iid assumption for the probability integral transforms of the forecasts produced by any of the models, there appeared to be less evidence against the assumption of uniformity for the NVAR than for the VAR.

Finally, we extended the approach of Diebold *et al.* (1997) to the evaluation of conditional and joint forecast densities, and considered the evaluation of sequences of multi-step forecasts within this framework.

Our results suggest a narrow focus on MSFE criteria may be misleading, and evaluation techniques which consider the entire forecast density may discriminate between models which would otherwise appear very similar. Such an avenue of investigation may be particularly fruitful when comparing forecasts from linear and non-linear models.

One aspect we have not explored in this paper is the evaluation of forecast densities conditional upon a particular regime. As noted above, such a strategy favours non-linear models on MSFE comparisons, and might do the same for forecast density comparisons. It would amount to selecting the subset of z's (marginal or conditional) corresponding to, say, x being in a particular regime at the forecast origin. In the empirical examples considered in this paper we have relatively few observations on the z's, particularly since the subsets of key interest would probably correspond to regimes with a minority of the data points. Such a strategy may be worthwhile when the focus is on the larger data sets typical of financial variables.

References

Beaudry, P., and Koop, G. (1993). Do recessions permanently affect output. *Journal of Monetary Economics*, **31**, 149–163.

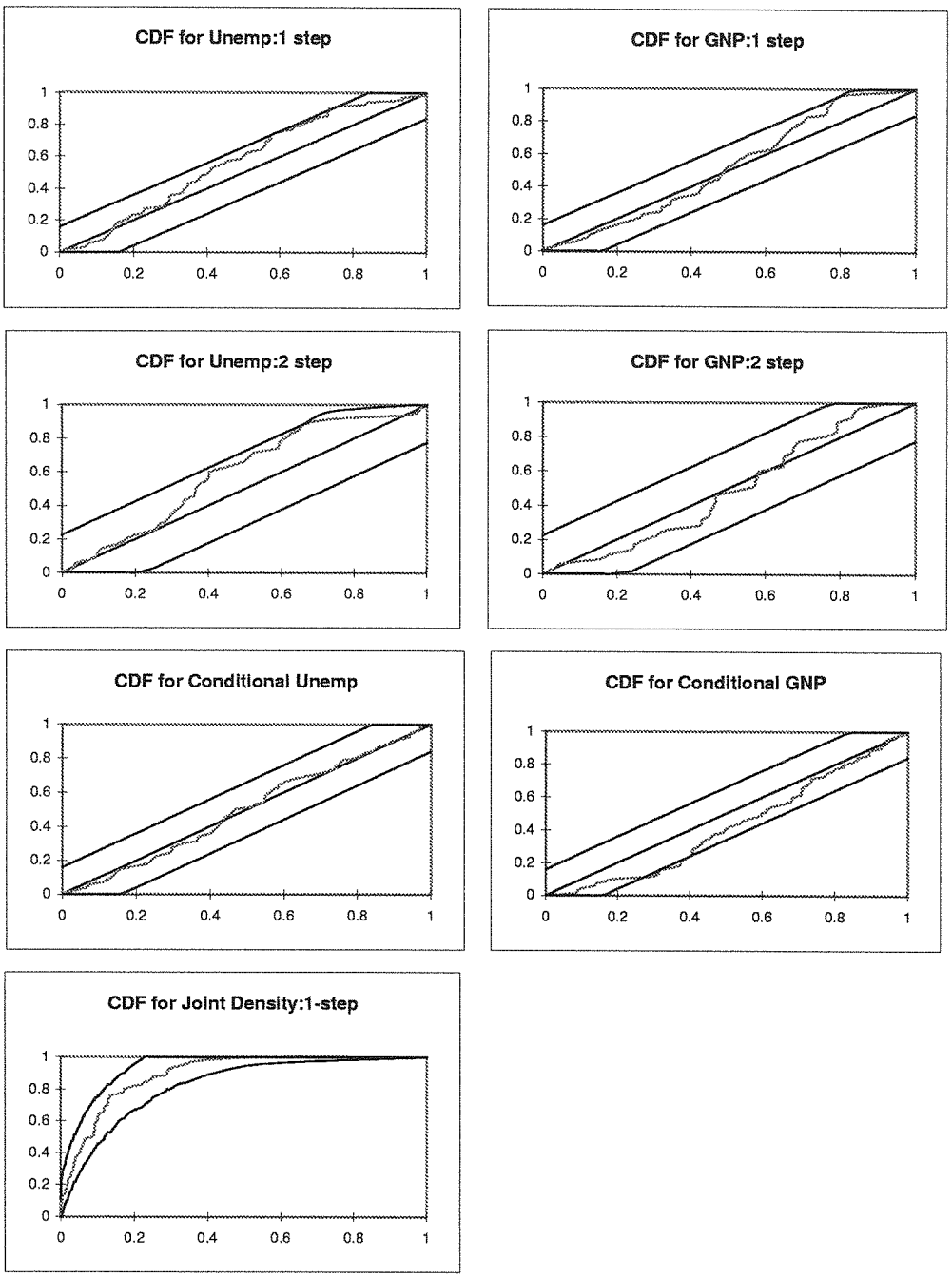


Figure 2 CDFs of VAR marginal, conditional and joint z-values.

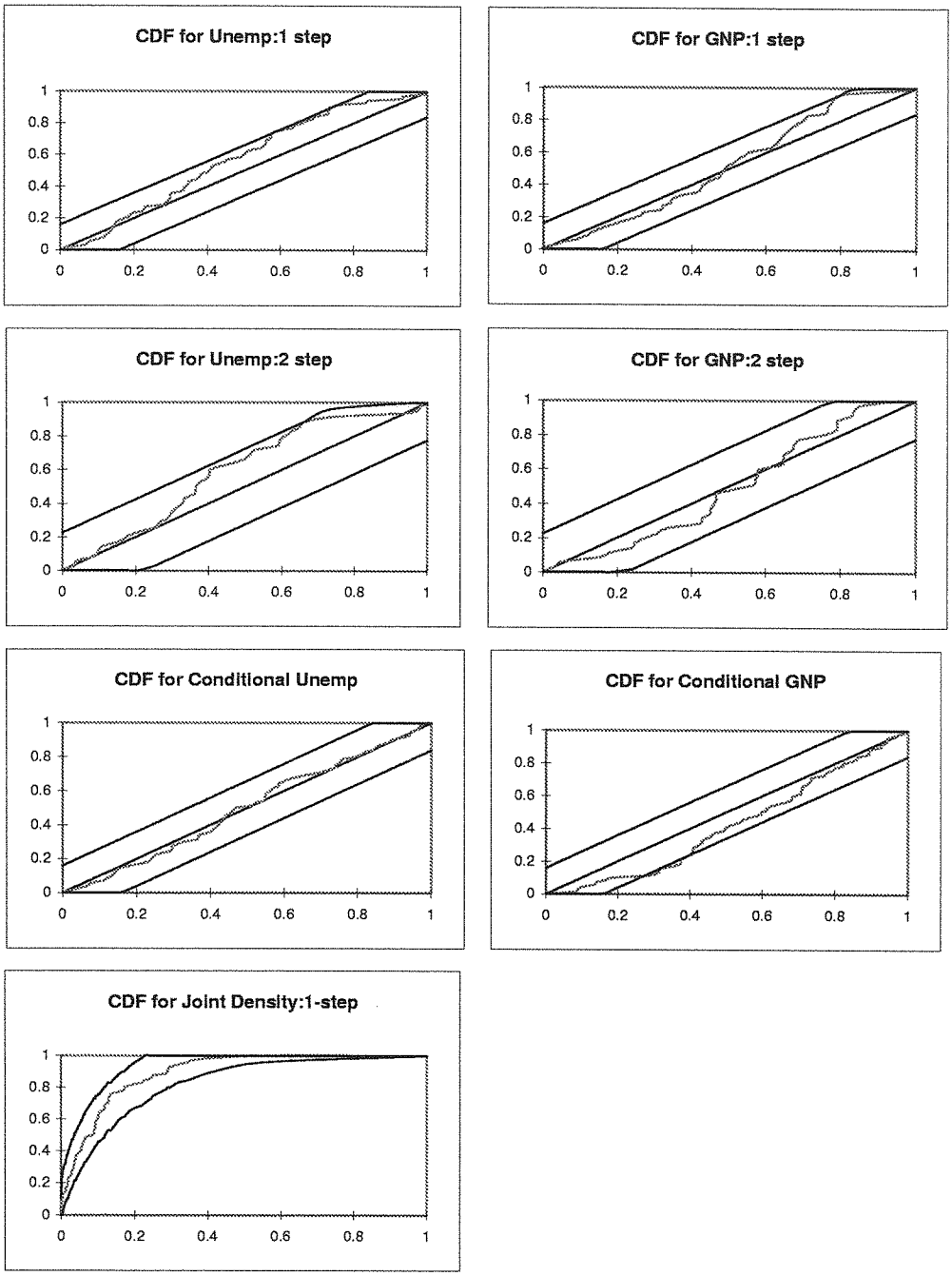


Figure 2 CDFs of VAR marginal, conditional and joint z-values.

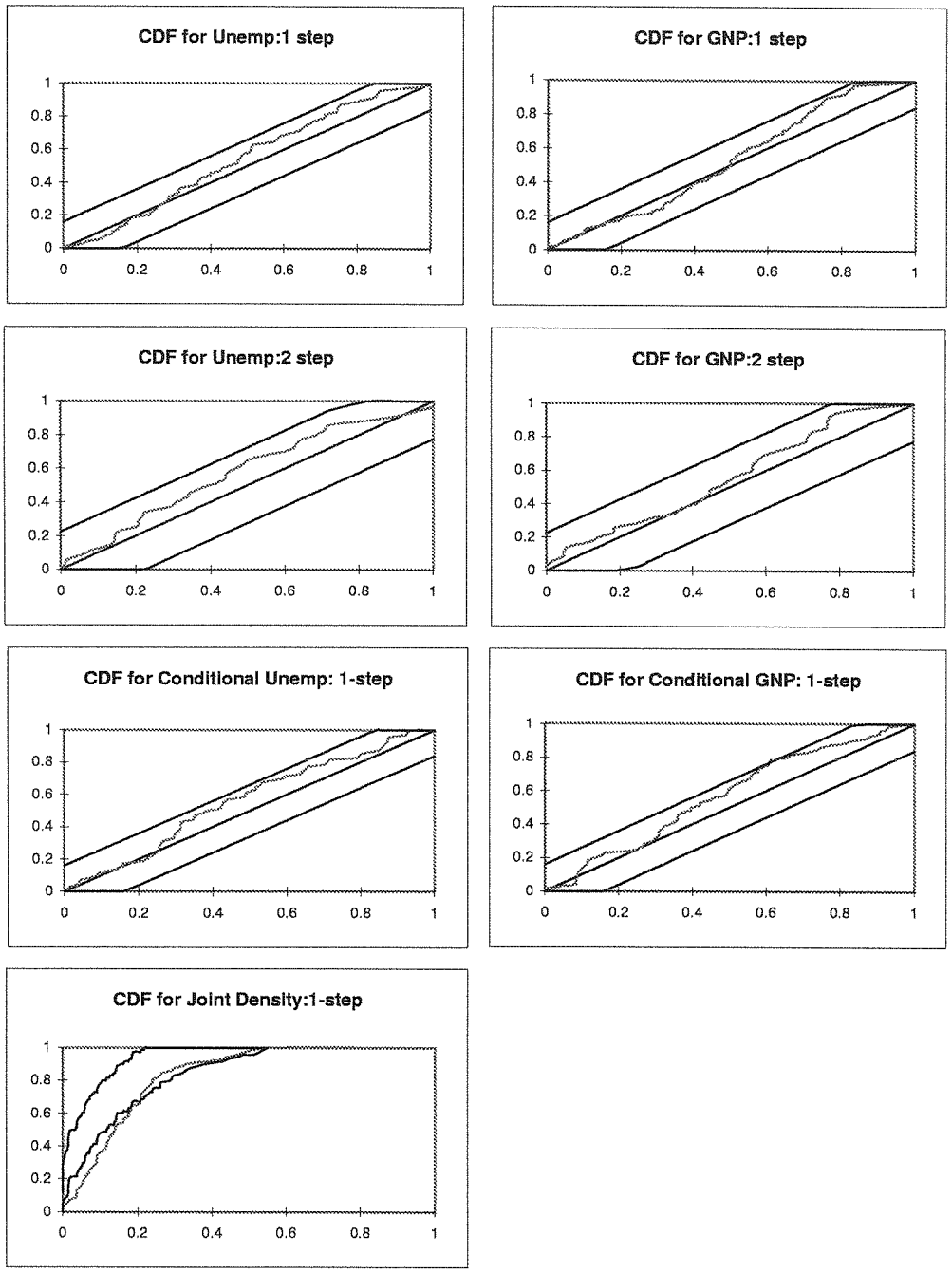


Figure 3 CDFs of NVAR marginal, conditional and joint z-values.

Chatfield, C. (1993). Calculating interval forecasts. *Journal of Business and Economic Statistics*, **11**, 121–135.

Christoffersen, P. F. (1997). Evaluating interval forecasts. *International Economic Review*. Forthcoming.

Clements, M. P., and Smith, J. (1996). A Monte Carlo study of the forecasting performance of empirical SETAR models. Warwick Economic Research Papers No. 464, Department of Economics, University of Warwick.

Clements, M. P., and Smith, J. (1997). The performance of alternative forecasting methods for SETAR models. *International Journal of Forecasting*, **13**, 463–475.

Clements, M. P., and Smith, J. (1998). Testing between self-exciting threshold autorogressive and structural change models. Mimeo, Department of Economics, University of Warwick.

De Gooijer, J. G., and Kumar, K. (1992). Some recent developments in non-linear time series modelling, testing and forecasting. *International Journal of Forecasting*, **8**, 135–156.

Diebold, F. X., Gunther, T. A., and Tay, A. S. (1997). Evaluating density forecasts. Manuscript, Department of Economics, University of Pennsylvania.

Diebold, F. X., and Nason, J. A. (1990). Nonparametric exchange rate prediction. *Journal of International Economics*, **28**, 315–332.

Diebold, F. X., Tay, A. S., and Wallis, K. F. (1997). Evaluating density forecasts of inflation: The survey of professional forecasters. Macromodelling Bureau Discussion Paper No. 38, University of Warwick.

Foster, D., and Vohra, R. V. (1996). Asymptotic calibration. Manuscript, Department of Statistics, the Wharton School, University of Pennsylvania.

Gallant, A. R., Rossi, P. E., and Tauchen, G. (1993). Nonlinear dynamic structures. *Econometrica*, **61**, 871–907.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, **57**, 357–384.

Hansen, B. E. (1996). Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica*, **64**, 413–430.

Henriksson, R. D., and Merton, R. C. (1981). On market timing and investment performance. II Statistical procedures for evaluating forecast skills. *Journal of Business*, **54**, 513–533.

Kendall, M. G., Stuart, A., and Ord, J. K. (1987). *Advanced Theory of Statistics*, 5th edn., Vol. 1 and 2. London: Charles Griffin and Co.

Koop, G., Pesaran, M. H., and Potter, S. M. (1996). Impulse response analysis in nonlinear multivariate models. *Journal of Econometrics*, **74**, 119–147.

Miller, L. H. (1956). Table of percentage points of Kolmogorov statistics. *Journal of the American Statistical Association*, **51**, 111–121.

Montgomery, A. L., Zarnowitz, V., Tsay, R. S., and Tiao, G. C. (1996). Nonlinearity in modeling and forecasting the U.S. unemployment rate. Mimeo, Wharton School, University of Pennsylvania, Philadelphia.

Pesaran, M. H., and Potter, S. M. (1997). A floor and ceiling model of US Output. *Journal of Economic Dynamics and Control*, **21**, 661–695.

Pesaran, M., and Timmermann, A. (1992). A simple nonparametric test of predictive performance. *Journal of Business and Econometic Statistics*, **10**, 461–465.

Potter, S. (1995). A nonlinear approach to U.S. GNP. *Journal of Applied Econometrics*, **10**, 109–125.

Schnader, M. H., and Stekler, H. O. (1990). Evaluating predictions of change. *Journal of Business*, **63**, 99–107.

Spanos, A. (1986). *Statistical Foundations of Econometric Modelling.* Cambridge: Cambridge University Press.

Stekler, H. O. (1994). Are economic forecasts valuable?. *Journal of Forecasting*, **13**, 495–505.

Tiao, G. C., and Tsay, R. S. (1994). Some advances in non-linear and adaptive modelling in time-series. *Journal of Forecasting*, **13**, 109–131.

Tong, H. (1978). On a threshold model. In Chen, C. H. (ed.), *Pattern Recognition and Signal Processing*, pp. 101–141. Amsterdam: Sijhoff and Noordoff.

Tong, H. (1983). *Threshold Models in Non-Linear Time Series Analysis*: Springer-Verlag, New York.

Tong, H. (1995). *Non-linear Time Series. A Dynamical System Approach.* Oxford: Clarendon Press. First published 1990.

Tong, H., and Lim, K. S. (1980). Threshold autoregression, limit cycles and cyclical data. *Journal of The Royal Statistical Society*, **B 42**, 245–292.

Wallis, K. F. (1995). Large-scale macroeconometric modelling. In Pesaran, M. H., and Wickens, M. R. (eds.), *Handbook of Applied Econometrics: Macroeconomics*: Basil Blackwell.

7 Appendix A. Independent forecasts

Proof of equation (9)

Let X_1 and X_2 be the probability integral transforms for x and u based on the marginal densities, obtained from the joint density. Assuming the model density is the actual forecast density, and the joint is the product of the marginals, then $X_1 \sim U(0,1)$, $X_2 \sim U(0,1)$, and X_1 and X_2 are independent. Hence the joint distribution function $F_{X_1X_2}$ is the product of the marginals, $F_{X_1X_2}(x_1, x_2) = x_1x_2$.

Using a change of variables (e.g., Spanos, 1986, pp.105-7):

$$\begin{aligned} Y_1 &= X_1X_2 \\ Y_2 &= X_2 \end{aligned}$$

for which the determinant of the Jacobian for the inverse transformation is:

$$J = \det \frac{\partial(X_1, X_2)}{\partial(Y_1, Y_2)} = \begin{bmatrix} \frac{1}{Y_2} & -\frac{Y_1}{Y_2^2} \\ 0 & 1 \end{bmatrix}$$

then the joint density function of (Y_1, Y_2) is:

$$f_{Y_1Y_2} = \frac{1}{Y_2} \tag{10}$$

where $0 < Y_1 < Y_2$ and $0 < Y_1 < 1$.

Since Y_1 is the random variable of interest, integrating Y_2 out of $f_{Y_1Y_2}$ over the permissible range gives:

$$f_{Y_1} = \int_{Y_1}^{1} Y_2^{-1} dY_2 = [\ln Y_2]_{Y_1}^{1} = -\ln Y_1.$$

The distribution function is:

$$F_{Y_1} = Y_1 - Y_1 \ln Y_1, \qquad 0 < Y_1 < 1$$

so:

$$F_{Y_1} = \Pr(X_1 < x_1, X_2 < x_2).$$

8 Appendix B. Sequences of multi-step forecasts

Consider an AR(1) with gaussian errors:

$$y_t = \alpha y_{t-1} + \varepsilon_t \tag{11}$$

and $\varepsilon_t \sim IN(0, \sigma^2)$.

Consider the 2-step forecast of t given $t-2$:

$$y_{t|t-2} = \alpha^2 y_{t-2}$$

has forecast error $\varepsilon_t + \alpha\varepsilon_{t-1}$ and error variance $(1+\alpha^2)\sigma^2$.

For a 2-step forecast of $t+1$, the error is $\varepsilon_{t+1} + \alpha\varepsilon_t$ with variance $(1+\alpha^2)\sigma^2$. The forecast errors are correlated as:

$$E[(\varepsilon_t + \alpha\varepsilon_{t-1})(\varepsilon_{t+1} + \alpha\varepsilon_t)] = \alpha\sigma^2.$$

The forecast density of the 2-step forecast of t is:

$$y_{t|t-2} \sim N\left(\alpha^2 y_{t-2}, \left(1+\alpha^2\right)\sigma^2\right)$$

and:

$$y_{t+1|t-1} \sim N\left(\alpha^2 y_{t-1}, \left(1+\alpha^2\right)\sigma^2\right)$$

but $y_{t|t-2}$ and $y_{t+1|t-1}$ are not independent:

$$C\left[y_{t|t-2}, y_{t+1|t-1}\right] = \alpha\sigma^2,$$

so that although the corresponding probability integral transforms (pit's), z_t and z_{t+1}, are $U[0,1]$, they will not be independent.

$$\begin{bmatrix} y_{t|t-2} \\ y_{t+1|t-1} \end{bmatrix} \sim N\left(\begin{array}{c} \alpha^2 y_{t-2} \\ \alpha^2 y_{t-1} \end{array}, \begin{bmatrix} \left(1+\alpha^2\right)\sigma^2 & \alpha\sigma^2 \\ \alpha\sigma^2 & \left(1+\alpha^2\right)\sigma^2 \end{bmatrix}\right) \equiv N\left(\begin{array}{c} \sigma_1^2 \\ \sigma_{12} \end{array}, \begin{array}{c} \sigma_{12} \\ \sigma_1^2 \end{array}\right).$$

So consider the 2-step ahead density of $y_{t+1|t-1}$ given that $Y_{t|t-2} = y_t$.

This can be simulated from:

$$y_{t+1} = \alpha^2 y_{t-1} + v_{t+1}$$

by drawing:

$$v_{t+1} \sim N\left(\rho(\varepsilon_t + \alpha\varepsilon_{t-1})\frac{\sigma_1}{\sigma_1}, \sigma_1^2\left(1-\rho^2\right)\right) \equiv N\left(\rho(\varepsilon_t + \alpha\varepsilon_{t-1}), \sigma^2\frac{\left(1+\alpha^2+\alpha^4\right)}{1+\alpha^2}\right) \tag{12}$$

where $\rho = \sigma_{12}/\sigma_1\sigma_1 = \alpha/\left(1+\alpha^2\right)$.

So:

$$p_{Y_{t+1}|Y_t}(y_{t+1} \mid Y_t = y_t) \sim N\left(\alpha^2 y_{t-1} + \rho(\varepsilon_t + \alpha\varepsilon_{t-1}), \sigma^2\frac{\left(1+\alpha^2+\alpha^4\right)}{1+\alpha^2}\right). \tag{13}$$

So the z for the 2-step ahead forecast of y_{t+1} conditional on the 2-step ahead forecast of y_t being correct, is given by calculating the pit for y_{t+1} against (13). This is $U[0,1]$ under correct specification.

Note the ε_j are not observed, but $\varepsilon_t + \alpha\varepsilon_{t-1}$ is the 2-step ahead error in forecasting period t.