Journal of Econometrics 35 (1987) 119–142. North-Holland

LIKELIHOOD AND OTHER APPROACHES TO PREDICTION IN DYNAMIC MODELS

Thomas F. COOLEY

University of Rochester, Rochester, NY 14627, USA
University of California, Santa Barbara, CA 93106, USA

William R. PARKE

University of California, Santa Barbara, CA 93106, USA

In this paper we consider the problem of generating multi-period predictions from two simple dynamic models, an autoregressive model and a geometric random walk. The autoregressive model constitutes a useful paradigm for many of the practical problems of prediction because it possesses a number of features that differentiate it sharply from the standard linear regression model. The geometric random walk model is widely used in macroeconomics and finance and is fundamentally non-normal.

The ideal situation for the prediction problem would be to know the true density of the future observations. Unfortunately, that density depends on parameters that are unknown and must be estimated. We analyze six prediction functions – approximations of the true density – that attempt to circumvent this problem. We contrast the theoretical properties of the likelihood prediction function proposed by Cooley and Parke (1986) with certainty equivalence prediction functions and mean-squared error prediction functions. The results of a Monte Carlo study illustrate the relative performance of the alternative prediction functions for conditional predictions and for the analysis of policy interventions. The results confirm the importance of accounting for parameter uncertainty and approximating the true shape of the future density.

1. Introduction

In this paper we first consider the problem of generating multi-period predictions from simple autoregressive models of the form

$$y_t = \alpha + \gamma y_{t-1} + \varepsilon_t, \qquad t = 1, \ldots, m+n, \tag{1}$$

where the true parameter vector $\theta^0 = (\alpha^0, \gamma^0)$ is unknown. The autoregressive model constitutes a useful paradigm for many practical prediction problems because it possesses a number of features that differentiate it sharply from standard linear regression models. Important characteristics that are highlighted by this model are: (1) the sample period realizations and the unknown future realizations are not independent; (2) the role of parameter uncertainty is compounded in predictions beyond one period; and (3) the distribution of forecast errors will be non-normal even if the underlying sources of uncer-

0304-4076/87/$3.50

tainty are normally distributed. Perhaps equally important, the autoregressive model is one of the few settings in which both Bayesian and mean squared error analyses of predictive distributions have been attempted.[1]

The second model we consider compares various prediction functions in a related, but different setting. Suppose that y_t follows a geometric random walk,

$$y_t/y_{t+1} = \alpha + \varepsilon_t, \qquad t = 1, \ldots, m+n, \tag{2}$$

where the growth rate error terms ε_t are independent. The model is of substantive interest because it occurs naturally in finance and macroeconomics. We also consider an extension of (2) in which these error terms arise from an AR(1) process. The geometric random walk model differs from (1) for our purposes in that the forecast errors are not approximately normally distributed even if the true coefficients are known. Good forecasting performance requires a technique that both accounts for uncertainty in the parameter values and non-normality in the effects of the error terms.

For any model, interest in predictions of the future focuses attention on the probability density generating the vector of future observations $y_f = \{y_t: t = m+1, \ldots, m+n\}$, with forecasts conditioned in some way on the data period observations $y_d = \{y_t: t = 1, \ldots, m\}$. The most informative summary of y_f is given by the density $f(y_f|y_d, \theta^0)$. For model (1) with normally distributed errors this has the form

$$f\left(y_f|y_d, \theta^0\right) \propto \exp\left\{-\tfrac{1}{2}\sum_{t=m+1}^{m+n}\left(y_t - \alpha^0 + \gamma^0 y_{t+1}\right)^2/\sigma^2\right\}. \tag{3}$$

Unfortunately, since θ^0 is unknown, the practical problem is to obtain prediction functions – approximations to $f(y_f|y_d, \theta^0)$ – that have good characteristics, but do not depend on θ^0.

One objective of this paper is to use the context of the autoregressive and geometric random walk models to compare the performances of a predictive likelihood function [Cooley and Parke (1986)] and other prediction functions that approximate $f(y_f|y_d, \theta^0)$. The most important characteristics of a prediction function are that it reveal, as closely as possible, the shape of the true density and that it account for parameter uncertainty. The former is important to obtaining either predictive confidence intervals for y_f, estimates of probabilities of the sort $\Pr(y_f < k)$ for prespecified values of k, or estimates of k such that $\Pr(y_f < k) = p$ for prespecified values of p. The latter ensures that one does not ignore a potentially important source of uncertainty.

[1]See Miller and Thompson (1986) for the former, and Baillie (1979), Box and Jenkins (1976, p. 269), Fuller and Hasza (1981), Yamamoto (1976) and Spitzer and Baillie (1983) for the latter.

We report the results of three numerical prediction experiments for the autoregressive model (1). The first problem is a basic forecast given both parameter and error term uncertainty and a known value for the last observation in the data period. The second problem is an intervention analysis that focuses on forecasting the marginal effects of a change in the intercept α^0 in the presence of both parameter uncertainty and error term uncertainty. The third problem is similar to the second, but highlights parameter uncertainty by setting the error term uncertainty to zero. We view the results of these three experiments not as a comprehensive or conclusive analysis of dynamic models, but rather as providing an indication of the issues that arise in complex forecasting problems. The intent here is to be expository and provide some evidence about the importance of the assumptions about functional form and sources of uncertainty that are commonly made in practical forecasting exercises.

For the geometric random walk model, we focus analytically on the issue of implementation rather than on numerical experiments. We do this for two reasons. First, it will turn out that the predictive likelihood approach yields a well-known optimal solution to the problem. Second, a mean squared analysis not only misses the fundamentally non-normal distribution of y_{m+n} conditional on y_d, but is also difficult to implement.

In the next section we outline the prediction functions we consider. These are shown to be nested in a logical way that facilitates the implementation of the simulation study. Section 3 describes the structure of the numerical prediction experiments for the AR(1) model (1) and the choice of summarizing statistics. Section 4 then analyzes the geometric random walk model (2).

2. Alternative prediction functions

When θ^0 is known, the best summary of the future is the density $f(y_f|y_d, \theta^0)$. In practical situations we have to employ prediction functions that are free of θ^0. There are two formal approaches to removing θ^0 from the predictive density. One can specify a prior density for θ^0 and integrate to obtain a posterior density free of θ^0 [Zellner (1971)]. Alternatively, one can replace θ^0 by sufficient statistics [Lauritzen (1974), Hinkley (1979), Cooley, Parke and Chib (1986)]. Practical implementation of either of the above approaches involves the introduction of sample estimates of θ^0 and there are a variety of less formal approaches based on the use of such estimates. We consider here six prediction functions that can be computed using only sample estimates of θ^0.[2] We illustrate each of them for the AR(1) model.

[2] Our exposition assumes σ^2 is known for convenience. Obviously, this is easily relaxed.

The certainty equivalence prediction function,

$$f(y_f; \hat{\theta}_d) \propto \exp\left\{ -\tfrac{1}{2} \sum_{t=m+1}^{m+n} (y_t - \hat{\alpha}_d - \hat{\gamma}_d y_{t-1})^2 / \sigma^2 \right\}, \tag{4}$$

can be obtained by substituting the least squares parameter estimates $\hat{\theta}_d$ for θ^0 in the multivariate normal density for y_f. The term 'certainty equivalence' is hopeful rather than based on any formal equivalence. Treating $\hat{\theta}_d$ as the true parameter ignores the effects of parameter uncertainty. This prediction function has the advantage that it is simple and should capture the correct shape of the density. Further, it is the prediction function that should result from a 'draw the errors' Monte Carlo approach as described by Brown and Mariano (1983). We now consider four distinct approaches to modifying (4) to incorporate parameter uncertainty: predictive likelihood analysis, mean squared error analysis, Monte Carlo simulation, and Bayesian analysis.

2.1. *Predictive likelihood analysis*

Because the predictive likelihood idea is somewhat new we introduce its definition at some length here. A more complete exposition is contained in Cooley and Parke (1986a, b).

The original definition of predictive likelihood is due to Lauritzen (1974) and Hinkley (1979). The Lauritzen–Hinkley concept recognizes the central importance of $f(y_f; \theta^0)$ for problems of prediction, but uses sufficient statistics to eliminate the unknown parameter θ^0. Let s_d, s_f and s_{d+f} be sufficient reductions of y_d, y_f and their union, respectively. Sufficiency ensures that the density $f(y_d; \theta^0)$ can be factored as

$$f(y_d; \theta^0) = f(y_d | s_d) f(s_d; \theta^0),$$

where $f(y_d|s_d)$ does not depend on θ^0. The Lauritzen–Hinkley definition of predictive likelihood exploits the fact that s_{d+f} is a function of s_f and s_d that does not depend on θ^0.

Definition 1 (Lauritzen–Hinkley). The predictive likelihood function is

$$\text{plik}(y_f|y_d) = f(y_f, s_d | s_{d+f}) = \frac{f(y_f; \theta^0) f(s_d; \theta^0)}{f(s_{d+f}; \theta^0)}.$$

This definition envisions treating $\text{plik}(y_f|y_d)$ as a likelihood function for the future observations y_f. In practical applications the plik could be used to order future values by their plausibility and to obtain confidence intervals for

y_f. This definition has been applied to several econometric problems by Cooley, Parke and Chib (1986), but its applicability is limited. There are some problems for which there is no sufficient reduction of the data and others where minimal sufficient statistics exist but have unworkably complex distributions.

The first step in resolving the shortcomings of Definition 1 exploits the fact that well-behaved maximum likelihood estimates are asymptotically sufficient [Cox and Hinkley (1974)]. This suggests replacing the sufficient statistics s_d and s_{d+f} in Definition 1 by $\hat{\theta}_d$ and $\hat{\theta}_{d+f}$, where the latter represent asymptotically sufficient maximum likelihood estimates. The second step is to use asymptotically valid densities. This leads to the following definition.

Definition 2. The asymptotic predictive likelihood function will be taken to be

$$\text{plik}^{a}(y_f|y_d) = \frac{f(y_f; \theta^0) f^{a}(\hat{\theta}_d; \theta^0)}{f^{a}(\hat{\theta}_{d+f}; \theta^0)},$$

where $f^{a}(\cdot;\cdot)$ denotes an asymptotically valid density.

Definition 2 will be applicable in situations where Definition 1 breaks down and, as the following proposition establishes, it has an easily usable form.

Proposition 1. Definition 2 can be expressed as

$$\text{plik}^{a}(y_f|y_d) = f(y_f; \hat{\theta}_d) \cdot w_1 \cdot w_2 \cdot \exp\{\text{Op}(m^{-3/2})\}, \tag{5}$$

where

$$w_1 = \exp\{-\tfrac{1}{2}\nabla_f(y_f; \hat{\theta}_d) H_{d+f}^{-1}(y_{d+f}; \hat{\theta}_d) \nabla_f'(y_f; \hat{\theta}_d)\},$$

and

$$w_2 = \exp\{\tfrac{1}{2}\text{tr}[H_d^{-1}(y_d; \hat{\theta}_d) H_f(y_f; \hat{\theta}_d)]\}.$$

$\nabla_f(y_f; \hat{\theta}_d)$ is the log gradient function of $f(y_f; \theta)$ evaluated at y_f and $\hat{\theta}_d$, and $H_{d+f}(y_{d+f}; \hat{\theta}_d)$ is the log Hessian of $f(y_{d+f}; \theta)$.

Proof. Cooley and Parke (1986a, b).

For present purposes (5) can be thought of as (4) adjusted for parameter uncertainty. In the models considered in this paper we take w_2 to be constant over variations in y_f. For the model (1) with normally distributed errors, the

predictive likelihood function is

$$\left(y_f - \hat{\theta}_d Z_f\right)' \left[\sigma^2 I_n + Z_f V\left(\hat{\theta}_d\right) Z_f'\right]^{-1} \left(y_f - \hat{\theta}_d Z_f\right), \tag{6}$$

where

$$Z_f = \begin{bmatrix} 1 & y_m \\ \vdots & \vdots \\ 1 & y_{m+n-1} \end{bmatrix},$$

and $V(\hat{\theta}_d)$ is the asymptotic variance–covariance matrix of $\hat{\theta}_d$. This functional form resembles a multivariate normal density, but differs from that density because, for forecasts two or more observations into the future, lagged endogenous values of y_f appear in the covariance matrix $Z_f V(\hat{\theta}_d) Z_f'$.

An important advantage of the predictive likelihood function is that, even if the errors are not taken to be normally distributed, (5) can easily be computed. Although (6) does not then follow from (5), straightforward simulation techniques yield the desired prediction functions. The appendix gives the details of these techniques.

2.2. *Mean squared error analysis*

Mean squared error analysis of the lagged dependent variable model can be characterized as a two-step process. First, replacing θ^0 by $\hat{\theta}_d$ and recursively substituting for unknown lagged values of y_t yields

$$\mathrm{MSE}^e\left(y_f | y_d\right) \propto \exp\left\{-\tfrac{1}{2}\left(y_f - \hat{y}_f\right)' \left(A_f\right)^{-1} \left(y_f - \hat{y}_f\right)\right\}, \tag{7}$$

where $\hat{y}_m$ is the known value y_m, $\hat{y}_t = \hat{\alpha}_d + \hat{\gamma}_d \hat{y}_{t-1}$ recursively defines $(\hat{y}_{m+1}, \ldots, \hat{y}_{m+n})$, and the i, j element of A_f, for $i \geq j$, is given by $(A_f)_{i,j} = \sigma^2 \hat{\gamma}^{i-j}(1 + \hat{\gamma}^2 + \cdots + \hat{\gamma}^{2j})$. This expression is simply another version of the certainty equivalence prediction function (4). The equivalence between the expressions holds only for models with normally distributed, additive error terms.

The second step in deriving the MSE prediction function is to account for the uncertainty in $\hat{\theta}_d$ in the mean (but not the variance) of (7) by the linearization

$$\hat{y}_{m+n} \cong \mathrm{E}\left(y_{m+n}\right) + D_n\left(\hat{\theta}_d - \theta^0\right),$$

where $D_n = \partial \hat{y}_{m+n} / \partial \hat{\theta}_d$. For example, for the third observation in the forecast

period,

$$y_{m+3} = \alpha(1+\gamma+\gamma^2) + \varepsilon_t + \gamma\varepsilon_{t-1} + \gamma^2\varepsilon_{t-2} + \gamma^3 y_m.$$

The vector of derivatives, given by

$$D_{3,1} = 1 + \hat{\gamma} + \hat{\gamma}^2 \quad \text{and} \quad D_{3,2} = \hat{\alpha}(1+2\hat{\gamma}) + 3\hat{\gamma}^2 y_m,$$

is a row in the matrix of derivatives D. This approach yields the mean squared error $A_f + DV(\hat{\theta}_d)D'$, where A_f represents error term uncertainty and $DV(\hat{\theta}_d)D'$ represents parameter uncertainty.

While a mean squared error analysis does not obviously yield a prediction function, we have adopted the normal density,

$$\mathrm{MSE}^{\varepsilon,\theta}(y_f|y_d) \propto \exp\left\{-\tfrac{1}{2}(y_f-\hat{y}_f)'\left[A_f + DV(\hat{\theta}_d)D'\right]^{-1}(y_f-\hat{y}_f)\right\}. \tag{8}$$

This prediction function accounts for parameter uncertainty by adding to the variance–covariance matrix in (7). It intentionally does not account for the non-linear relationship between the estimation errors $\hat{\theta}_d - \theta^0$ and the forecast errors $y_f - \hat{y}_f$.[3]

The relation between predictive likelihood analysis (6) and mean squared error analysis (8) can be understood by noting that (8) could be obtained from (6) by replacing lagged values of y_t in the factors Z_f of $Z_f V(\hat{\theta}_d) Z_f'$ with the plug-in point forecasts $\hat{y}_t$. This substitution eliminates the conditionality of the variance–covariance matrix by ignoring the interaction between error term uncertainty and parameter uncertainty for forecasts more than one period into the future. For example, the forecast error for y_{m+2} depends upon the product of the error in $\hat{\theta}_d$ and (through y_{m+1}) the realization of ε_{m+1}. The predictive likelihood function (6) includes this interaction between ε_{m+1} and $\hat{\theta}_d - \theta^0$, but the mean squared error prediction function takes it to equal zero.

2.3. *The 'draw the error terms, draw the coefficients' Monte Carlo technique*

A common empirical forecasting technique relies on Monte Carlo draws ε_f from the estimated distribution of ε_f. Substituting these draws into the model's equations yields draws from the CEQ prediction function $f(y_f; \hat{\theta}_d)$. This procedure, which we will refer to as $\mathrm{MC}^{\varepsilon}$, has been analyzed extensively by Fair (1980) and Mariano and Brown (1983, 1984). Sometimes these draws are augmented with Monte Carlo draws $\underset{\sim}{\theta}$ from the approximate distribution

[3] Fuller and Hasza (1981) show this interaction term to be of $O(m^{-2})$.

$\underset{\sim}{\theta} \sim N(\hat{\theta}_d, V(\hat{\theta}_d))$. This technique, which we will refer to as $MC^{\varepsilon,\theta}$, is equivalent to taking draws from a mixture $\int f(y_f; \theta) f^a(\theta; \hat{\theta}_d) \partial\theta$. While the latter function may or may not have a theoretical basis as a numerical implementation of some other technique, its position in econometrics is well established. For that reason, we include it here as a practical approach to the prediction problem.

2.4. *Bayesian analysis*

To this point, predictive likelihood analysis has been given a completely frequentist treatment. One virtue of the methodology is that it has a reasonable interpretation from a Bayesian viewpoint as well. Hinkley (1979) shows that the Bayesian predictive posterior $f(y_f|y_d)$ can be factored as

$$f(y_f|y_d) \propto \text{plik}(y_f|y_d) \cdot f(s_{d+f}), \tag{9}$$

where $f(s_{d+f})$ is a prior for the sufficient statistic s_{d+f}. Eq. (9) implies that the predictive likelihood function is a predictive posterior for a uniform prior on s_{d+f}.

In the present case of dependent observations, there arise two complications to this Bayesian view of $\text{plik}(y_f|y_d)$ and to a Bayesian analysis itself. First, the notion of a non-informative prior is deficient because forecasts for this model involve powers of γ and it is impossible to specify uniform priors for γ^0, $(\gamma^0)^2$, and $(\gamma^0)^3$ simultaneously. Equivalently, any prior will be informative for all but at most one future value of $y_m\gamma^n$. Second, the usual uniform prior on θ^0 weights parameter regions of explosive behavior too heavily. When $f(\theta^0)$ is constant the parameter region associated with *non-explosive* behavior has pseudo-measure zero.

Miller and Thompson (1986) implement a Bayesian analysis of an autoregressive model [AR(2) in their case] using a uniform prior for the parameters. This yields a normal-gamma form for the posterior distribution of the parameters. They then draw from this normal-gamma distribution, simulating future draws to build up a predictive posterior. Numerically, their procedure is very similar to a Monte Carlo analysis based on drawings of both error terms and coefficients. Because the principal difference between the two techniques, for this model at least, is in terms of interpreting the numerical results and not in terms of numerical results themselves, we will focus on the Monte Carlo procedure as representative of a Miller and Thompson style Bayesian analysis and do not provide a separate analysis based on such normal-gamma draws.

2.5. *Summary*

It is useful at this point to summarize the strengths and weaknesses of the alternative prediction functions as we see them. The certainty equivalence

prediction function (CEQ) has the functional form of the true density, but ignores parameter uncertainty. The predictive likelihood function (PLIK) is based on an asymptotic approximation to the correct functional form and incorporates parameter uncertainty.

The mean squared error prediction function ($MSE^{\varepsilon,\theta}$) has the virtue that it accounts for parameter uncertainty although it neglects the true shape of the forecast error density by employing a normal approximation. Note also that computing the $MSE^{\varepsilon,\theta}$ function is not to be confused with *minimizing* the mean squared error. Even under quadratic loss, minimizing mean squared error is a formidable task [Chow (1975)] that is not likely to generalize to more complicated models. Finally, note that confidence intervals based on $MSE^{\varepsilon,\theta}$ are strictly appropriate only if they are based on Chebyshev's inequality.

Bayesian posterior predictive densities face conceptual problems only when based on uniform priors as noted above. Bayesian analysis based on informative priors are computationally difficult for problems of modest complexity unless based on a Monte Carlo simulation.

3. Monte Carlo analysis of sample realized prediction functions

The relative merits of these techniques and the practical significance of the theoretical advantages and shortcomings are in part an empirical matter. To shed some light on the significance of the theoretical arguments, we analyze the prediction functions discussed above in three Monte Carlo experiments. Tables 1–3 present the results of these experiments, and figs. 1–4 illustrate typical realized prediction functions for each experiment.

3.1. The three experiments

Experiment 1 is a conditional forecasting problem highlighting the effects of the last known observation y_m. We set $y_m = 5$ for a true model with $\alpha = 0$, $\gamma = 0.9$, and $\varepsilon_t \sim N(0,1)$, $t = 1, \ldots, m+n$. This forecasting problem differs from the linear-normal paradigm in that the conditional expected value,

$$E(y_{m+n}|y_m) = y_m(\gamma^0)^n,$$

is a non-linear function of γ^0. As a consequence the error,

$$y_m(\hat{\gamma}_d)^n - y_m(\gamma^0)^n,$$

is a non-linear function of the estimation error $\hat{\gamma}_d - \gamma^0$. The mean squared error ($MSE^{\varepsilon,\theta}$), Monte Carlo ($MC^{\varepsilon,\theta}$), and predictive likelihood approaches will differ for this experiment to the extent that they treat this non-linearity differently. All three approaches incorporate essentially the same measures of

Table 1

Experiment 1.[a]

	Percentile probabilities					Moments			
	$p = 0.10$	$p = 0.25$	$p = 0.50$	$p = 0.75$	$p = 0.90$	Mean	Std. dev.	Skew.	Kurt.
$n = 1$									
TRUTH	0.10	0.25	0.50	0.75	0.90	4.50	1.00	0.00	3.00
$MSE^{\epsilon}, MC^{\epsilon}, CEQ$	0.08	0.21	0.43	0.68	0.85	4.32	1.00	0.00	3.00
$MSE^{\epsilon,\theta}$	0.08	0.20	0.43	0.69	0.86	4.32	1.04	0.00	3.00
$MC^{\epsilon,\theta}$	0.08	0.20	0.43	0.69	0.87	4.32	1.04	0.00	2.94
PLIK	0.07	0.20	0.43	0.69	0.86	4.32	1.04	0.00	2.93
$n = 5$									
TRUTH	0.10	0.25	0.50	0.75	0.90	2.95	1.85	0.00	3.00
$MSE^{\epsilon}, MC^{\epsilon}, CEQ$	0.09	0.21	0.42	0.64	0.80	2.51	1.76	0.00	3.00
$MSE^{\epsilon,\theta}$	0.07	0.20	0.42	0.66	0.83	2.51	1.94	0.00	3.00
$MC^{\epsilon,\theta}$	0.08	0.20	0.42	0.67	0.84	2.59	1.96	0.21	3.16
PLIK	0.08	0.21	0.43	0.68	0.85	2.63	1.96	0.14	2.93
$n = 10$									
TRUTH	0.10	0.25	0.50	0.75	0.90	1.74	2.15	0.00	3.00
$MSE^{\epsilon}, MC^{\epsilon}, CEQ$	0.10	0.23	0.44	0.65	0.80	1.36	2.00	0.00	3.00
$MSE^{\epsilon,\theta}$	0.08	0.21	0.44	0.67	0.83	1.36	2.21	0.00	3.00
$MC^{\epsilon,\theta}$	0.09	0.22	0.45	0.69	0.86	1.55	2.33	0.38	3.64
PLIK	0.09	0.23	0.46	0.70	0.86	1.62	2.31	0.17	2.87
$n = 15$									
TRUTH	0.10	0.25	0.50	0.75	0.90	1.03	2.25	0.00	3.00
$MSE^{\epsilon}, MC^{\epsilon}, CEQ$	0.11	0.25	0.46	0.67	0.81	0.78	2.07	0.00	3.00
$MSE^{\epsilon,\theta}$	0.09	0.23	0.46	0.69	0.84	0.78	2.29	0.00	3.00
$MC^{\epsilon,\theta}$	0.10	0.24	0.47	0.71	0.87	1.03	2.51	0.49	4.24
PLIK	0.11	0.25	0.49	0.72	0.87	1.07	2.43	0.11	2.85
$n = 20$									
TRUTH	0.10	0.25	0.50	0.75	0.90	0.61	2.28	0.00	3.00
$MSE^{\epsilon}, MC^{\epsilon}, CEQ$	0.12	0.26	0.48	0.69	0.83	0.46	23.10	0.00	3.00
$MSE^{\epsilon,\theta}$	0.10	0.25	0.48	0.71	0.85	0.46	2.31	0.00	3.00
$MC^{\epsilon,\theta}$	0.12	0.26	0.49	0.73	0.88	0.74	2.62	0.63	5.37
PLIK	0.11	0.27	0.51	0.74	0.87	0.79	2.49	0.07	2.83

[a]All figures are averages over 200 data period simulations. For each data period simulation, the numeric integrations use 400 forecast draws.

the error term uncertainty,

$$\sigma^2 \sum_{i=1}^{n} (\gamma^0)^{i-1},$$

which is a linear sum of normal components.

The second and third experiments examine the non-linearity induced by a stylized policy intervention. For experiment 2, we set $y_m = 0$ and $\varepsilon_t \sim N(0,1)$, $t = 1, \ldots, m+n$, but we add 1.0 to the estimated constant term in all future

Table 2

Experiment 2.[a]

	Percentile probabilities					Moments			
	$p=0.10$	$p=0.25$	$p=0.50$	$p=0.75$	$p=0.90$	Mean	Std. dev.	Skew.	Kurt.
$n=1$									
TRUTH	0.10	0.25	0.50	0.75	0.90	1.00	1.00	0.00	3.00
MSE^{ϵ}, MC^{ϵ}, CEQ	0.10	0.25	0.50	0.75	0.90	0.99	1.00	0.00	2.00
$MSE^{\epsilon,\theta}$	0.10	0.25	0.50	0.75	0.90	0.99	1.01	0.00	3.00
$MC^{\epsilon,\theta}$	0.10	0.25	0.50	0.75	0.90	0.99	1.01	0.00	2.95
PLIK	0.10	0.24	0.49	0.75	0.90	0.99	1.01	0.00	2.95
$n=5$									
TRUTH	0.10	0.25	0.50	0.75	0.90	4.10	1.85	0.00	3.00
MSE^{ϵ}, MC^{ϵ}, CEQ	0.10	0.23	0.45	0.68	0.84	3.83	1.76	0.00	3.00
$MSE^{\epsilon,\theta}$	0.09	0.22	0.45	0.69	0.85	3.83	1.85	0.00	3.00
$MC^{\epsilon,\theta}$	0.09	0.22	0.44	0.69	0.86	3.85	1.87	0.11	3.09
PLIK	0.09	0.22	0.45	0.70	0.86	3.89	1.87	0.09	3.00
$n=10$									
TRUTH	0.10	0.25	0.50	0.75	0.90	6.51	2.15	0.00	3.00
MSE^{ϵ}, MC^{ϵ}, CEQ	0.09	0.21	0.40	0.60	0.75	5.83	2.00	0.00	3.00
$MSE^{\epsilon,\theta}$	0.06	0.18	0.40	0.64	0.80	5.83	2.36	0.00	3.00
$MC^{\epsilon,\theta}$	0.07	0.19	0.40	0.65	0.83	5.99	2.47	0.44	3.58
PLIK	0.08	0.20	0.42	0.67	0.83	6.11	2.42	0.19	2.74
$n=15$									
TRUTH	0.10	0.25	0.50	0.75	0.90	7.94	2.25	0.00	3.00
MSE^{ϵ}, MC^{ϵ}, CEQ	0.10	0.21	0.38	0.55	0.69	6.96	2.07	0.00	3.00
$MSE^{\epsilon,\theta}$	0.04	0.16	0.38	0.61	0.77	6.96	2.83	0.00	3.00
$MC^{\epsilon,\theta}$	0.06	0.17	0.38	0.63	0.82	7.33	3.09	0.81	4.60
PLIK	0.07	0.19	0.42	0.66	0.81	7.48	2.79	0.13	2.43
$n=20$									
TRUTH	0.10	0.25	0.50	0.75	0.90	8.78	2.28	0.00	3.00
MSE^{ϵ}, MC^{ϵ}, CEQ	0.11	0.22	0.37	0.52	0.65	7.63	2.10	0.00	3.00
$MSE^{\epsilon,\theta}$	0.03	0.14	0.37	0.60	0.75	7.63	3.23	0.00	3.00
$MC^{\epsilon,\theta}$	0.06	0.16	0.37	0.63	0.81	8.25	3.75	1.24	6.61
PLIK	0.07	0.20	0.41	0.63	0.76	8.24	2.98	0.07	2.33

[a]All figures are averages over 200 data period simulations. For each data period simulation, the numeric integrations use 400 forecast draws.

periods. The conditional expected value,

$$E(y_{m+n}|y_m) = \sum_{i=1}^{n} (\gamma^0)^{i-1},$$

increases toward the limit $1/(1-\gamma^0)$, and the non-linearity becomes more important as the forecast horizon n increases.

The third experiment considers the same intervention in the absence of future error term uncertainty, highlighting the effect of parameter uncertainty

Table 3

Experiment 3.[a]

	Percentile probabilities					Moments			
	$p = 0.10$	$p = 0.25$	$p = 0.50$	$p = 0.75$	$p = 0.90$	Mean	Std. dev.	Skew.	Kurt.
$n = 1$[b]									
$n = 5$									
TRUTH						4.10	0.00	—	—
CEQ	0.28	0.28	0.28	0.28	0.28	3.82	0.00	—	—
$MSE^{\epsilon,\theta}$	0.03	0.10	0.28	0.53	0.72	3.82	0.37	0.00	3.00
$MC^{\epsilon,\theta}$, PLIK	0.03	0.10	0.28	0.53	0.76	3.84	0.38	0.32	3.16
$n = 10$									
TRUTH						6.51	0.00	—	—
CEQ	0.28	0.28	0.28	0.28	0.28	5.77	0.00	—	—
$MSE^{\epsilon,\theta}$	0.02	0.09	0.28	0.51	0.69	5.77	1.08	0.00	3.00
$MC^{\epsilon,\theta}$, PLIK	0.03	0.10	0.28	0.53	0.76	5.92	1.14	0.82	4.22
$n = 15$									
TRUTH						7.94	0.00	—	—
CEQ	0.28	0.28	0.28	0.28	0.28	6.84	0.00	—	—
$MSE^{\epsilon,\theta}$	0.01	0.09	0.28	0.51	0.66	6.84	1.72	0.00	3.00
$MC^{\epsilon,\theta}$, PLIK	0.03	0.10	0.28	0.53	0.76	7.19	1.95	1.30	6.26
$n = 20$									
TRUTH						8.78	0.00	—	—
CEQ	0.28	0.28	0.28	0.28	0.28	7.46	0.00	—	—
$MSE^{\epsilon,\theta}$	0.00	0.09	0.28	0.50	0.64	7.46	2.23	0.00	3.00
$MC^{\epsilon,\theta}$, PLIK	0.03	0.10	0.28	0.53	0.76	8.05	2.75	1.81	9.54

[a]All figures are averages over 1000 data period simulations. For each data period simulation, the numeric integrations use 400 forecast draws.

[b]For $n = 1$, there is no forecast uncertainty for this experiment.

by itself. We again set $y_m = 0$ and add one to the constant term for each future period, but modify the previous experiment by setting $\varepsilon_t = 0$, $t = m + 1, \ldots, m + n$. This experiment focuses attention on the non-linearity in parameters because the forecast problem does not include the normally distributed error term uncertainty. It is the only one of the three for which the unknown future quantity is a non-stochastic point function of the true parameters. Even for experiment 3, a prediction function that reflects parameter uncertainty will yield a non-degenerate predictive density with a non-zero dispersion.[4]

Finally, we note that there are many ways to summarize and describe a realized prediction function, and no one summary will be universally ap-

[4]The limit of (6) as the future term variance goes to zero (but the data period error term variance remains unchanged) is

$$(y_f - \hat{\theta}_d Z_f)' \left[Z_f V(\hat{\theta}_d) Z_f \right]^{-} (y_f - \hat{\theta}_d Z_f).$$

We note that this equation and a 'draw the coefficients' Monte Carlo technique are equivalent in this case and so use the latter as an implementation technique.

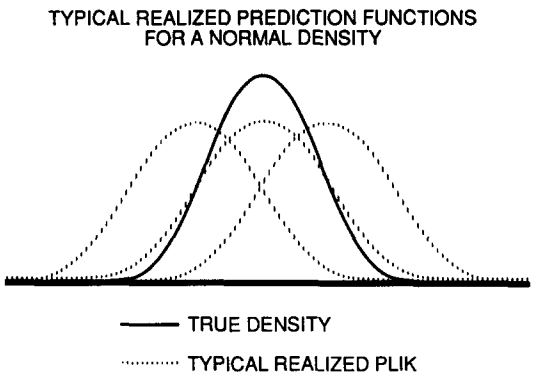

Fig. 1

propriate. [Edwards (1972) makes this point in some detail for parametric likelihood functions.] For a given forecasting problem, it may be appropriate to compute moments, quartiles, modal points, values at specific points, etc. More sophisticated measures might include information measures of differences among probability densities. Here we focus on two basic measures of forecast performance: the accuracy of nominal percentiles as probability forecasts and the accuracy of mean values as forecasts of the future expected value.

3.2. *Percentiles probability*

We compute nominal percentiles for a given prediction function, say $\text{plik}(y_{m+n}|y_d)$, by finding a value $\hat{y}^{\alpha}_{m+n}$ such that

$$\int_{-\infty}^{\hat{y}^{\alpha}_{m+n}} \text{plik}(y_{m+n}|y_d)\,\partial_{ym+n} \Big/ \int_{-\infty}^{+\infty} \text{plik}(y_{m+n}|y_d)\,\partial y_{m+n} = \alpha.$$

The value $\hat{y}^{0.50}_{m+n}$ is then the median, and $(\hat{y}^{0.10}_{m+n}, \hat{y}^{0.90}_{m+n})$ is a natural 80% forecast confidence interval for y_{m+n}. (Figs. 2, 3 and 4 depict typical realized confidence intervals for each prediction function in terms of the 10th, 25th, 50th, 75th and 90th percentiles.) Although it would be possible, using a variant of the Neyman–Pearson lemma, to compute a shorter 80% confidence interval, the interval $(\hat{y}^{0.10}_{m+n}, \hat{y}^{0.90}_{m+n})$ is easier to compute and treats positive and negative forecast errors symmetrically. Furthermore, symmetric probability levels (and one-sided confidence intervals) may be of some importance if, as is common, forecast errors of one sign are taken as bad news, but forecast errors of the opposite sign are actually viewed favorably.

To evaluate these nominal percentiles, we compute $\text{P}(y_{m+n} < \hat{y}^{\alpha}_{m+n}|\hat{\theta}_d)$ for each realized prediction function. The average of these probabilities over the successive data period simulations gives a Monte Carlo estimate of the unconditional probability $\text{P}(y_{m+n} < \hat{y}^{\alpha}_{m+n})$. These latter figures are referred to as percentile probabilities in tables 1, 2, and 3.

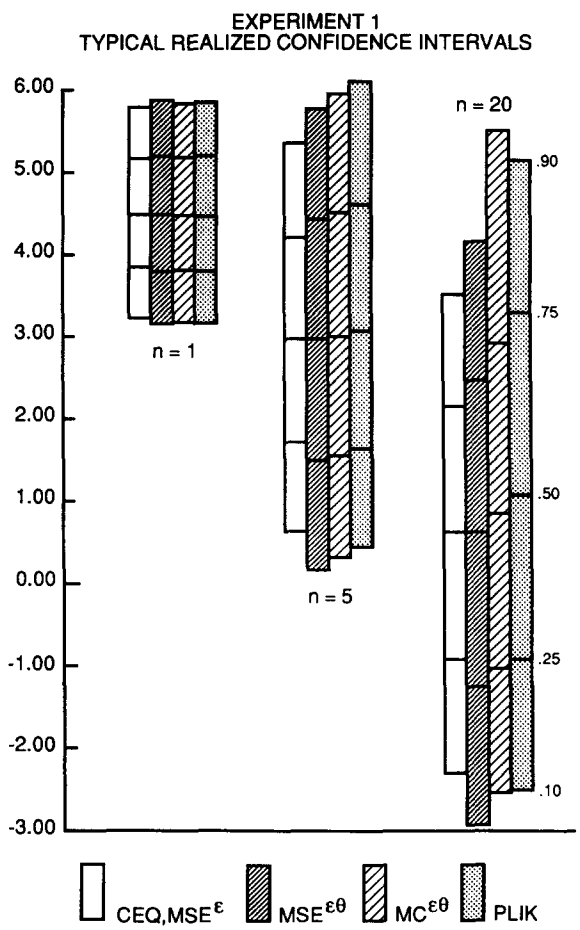

Fig. 2

The results in table 1 for experiment 1 and the typical confidence intervals in fig. 2 show that, for experiment 1, the practical differences among the predictions for the predictive likelihood, Monte Carlo, and mean squared error approaches are not large. For the right-hand tail as reflected in the 0.75 and 0.90 percentiles, the $MC^{\varepsilon,\theta}$ and PLIK probabilities are a bit closer to the desired figures although the differences among techniques are fairly small. For the left-hand tail, the differences among techniques are even smaller.

Experiments 2 and 3 exhibit somewhat greater differences among the percentile probabilities for the various techniques. For experiment 2, the $MC^{\varepsilon,\theta}$ and PLIK figures are noticeably closer to the desired values than the corresponding figures for $MSE^{\varepsilon,\theta}$. This outcome, which is even clearer for experiment 3, reflects the symmetry imposed by the normality assumption underlying the $MSE^{\varepsilon,\theta}$ prediction function. In the absence of the normally

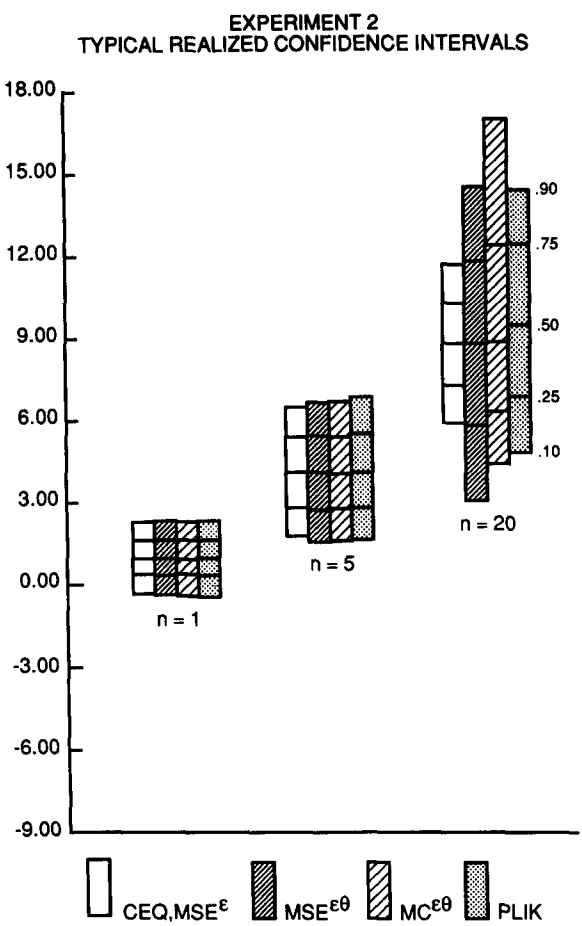

Fig. 3

distributed error term uncertainty in experiment 3, the left-hand tail for the $MSE^{\varepsilon,\theta}$ prediction function is too large to the extent that the nominal 0.10 percentile for $n = 20$ has a true probability of including y_{m+20} that rounds to 0.00. The right-hand tail, on the other hand, is too short in that the 0.90 percentile is to the right of y_{m+20} only with probability 0.64. The results for $MC^{\varepsilon,\theta}$ and PLIK are, of course, not nearly as good as one might hope either, and we find that these prediction experiments reveal several problems common to all the prediction functions.

These results appear to reflect three principal deviations between the actual distribution of $\hat{\theta}_d$ for the sample used here ($m = 100$) and the asymptotic distribution that underlies all the prediction functions:

(1) The estimate of γ^0 is subject to a downward bias. For the sample size, $m = 100$, used here, the bias -0.038 is a substantial fraction of the

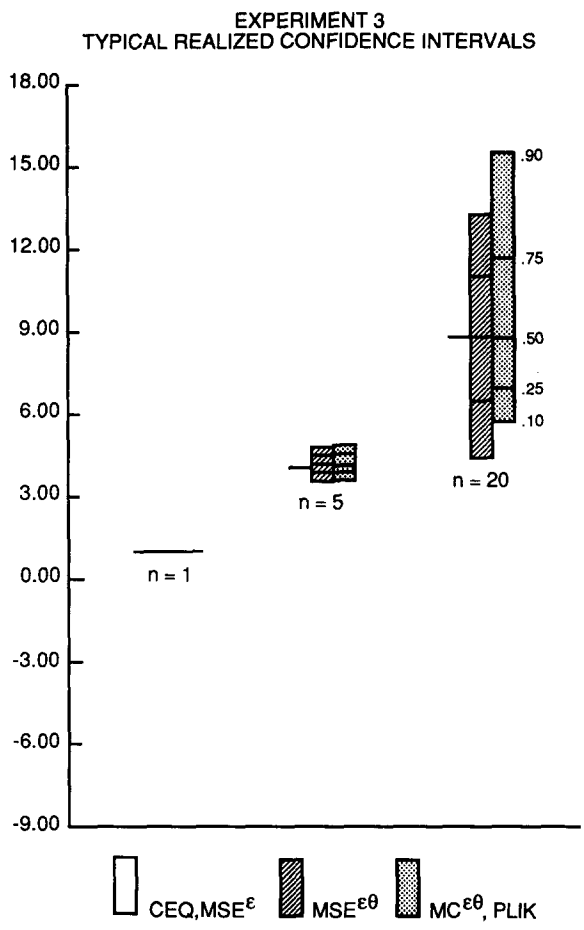

Fig. 4

standard error 0.057. The effect of this bias tends to be magnified for forecasts far into the future.

(2) The distribution of $\hat{\gamma}_d$ is substantially skewed, with a long left-hand tail. The negative skewness in $\hat{\gamma}_d - \gamma^0$ operates in a direction that partially offsets the non-linearities in $(\hat{\gamma}_d)^n - (\gamma^0)^n$ and $1/(1-\hat{\gamma}_d) - 1/(1-\gamma^0)$.

(3) The estimate of $V(\hat{\gamma}_d)$ is not even approximately constant. In particular, the estimated standard error of $\hat{\gamma}_d$ is substantially smaller for overestimates of γ^0 than for underestimates.

Because all these factors combine with the differences among prediction functions, the conclusions drawn must be qualified.

One clear lesson to be drawn from the results of experiment 1 is the importance of accounting for parameter uncertainty. Although the techniques

that ignore parameter uncertainty (CEQ, MSE^{ϵ} and MC^{ϵ}) do well for the short-term forecasts, their percentile probabilities are severely biased as the forecast horizon gets larger. The percentile probability biases for these techniques are substantially greater than those for the methods ($MSE^{\epsilon,\theta}$, $MC^{\epsilon,\theta}$ and PLIK) that do account for parameter uncertainty.

3.3. Point forecasts and moments of prediction functions

The emphasis in this paper is on prediction functions rather than on point forecasts, but it is quite reasonable to use a summary measure of a prediction function as a point forecast of some aspect of the unknown future. For example, the 0.50 percentile computed in tables 1, 2 and 3 is a natural point forecast for the median of the future density. In a similar way, the mean of a prediction function (normalized to integrate to one) is the natural forecast of the future conditional expected value. In practice, this integration involves computing the average of the Monte Carlo draws for $MC^{\epsilon,\theta}$ and computing a weighted average of another set of draws for PLIK. Details of these calculations are given in the appendix.

The expected value forecasts for the CEQ, MSE^{ϵ} and $MSE^{\epsilon,\theta}$ techniques enjoy the advantage of computational simplicity because they all equal $y_m(\hat{\gamma}_d)^n$ for experiment 1 and $\sum_{i=1}^{n}(\hat{\gamma}_d)^{i-1}$ for experiments 2 and 3. There is no claim here that these naive 'plug-in' forecasts are unbiased. The only claim is that $MSE^{\epsilon,\theta}$ uses an asymptotically valid approximation to the mean squared error of the naive 'plug-in' forecast used for a given forecasting problem.

This fact is evident in tables 1, 2 and 3. In nearly every case, the bias of the expected value PLIK and $MC^{\epsilon,\theta}$ forecasts is smaller in absolute value than the bias of the $MSE^{\epsilon,\theta}$ forecast. For example, in experiment 3 at $n = 20$, the bias for $MSE^{\epsilon,\theta}$ is 1.32, but the bias for $MC^{\epsilon,\theta}$ and PLIK is only 0.73. The corresponding figures for experiment 2 at $n = 20$ are -1.15 for $MSE^{\epsilon,\theta}$, -0.53 for $MC^{\epsilon,\theta}$, and -0.54 for PLIK. Both the $MC^{\epsilon,\theta}$ and PLIK point forecasts benefit, in this comparison, from skewness in the two prediction functions that offsets skewness in the distribution of $\hat{\gamma}_d$. The negative bias in $\hat{\gamma}_d$ is also partially offset by the larger right-hand tails in the two prediction functions.

The differences among the various prediction functions are evident in the higher moments given in tables 1, 2 and 3. (These figures are the averages over the data period draws of the realized prediction function moments.) The $MC^{\epsilon,\theta}$ draws exhibit larger variance and much larger skewness and kurtosis than plik$(y_{m+n}|y_d)$ even though the corresponding percentiles do not differ greatly. We can attribute this to the feature of the Monte Carlo technique that the γ draw can and will occasionally be greater than one, producing a forecast draw with the properties of a non-stationary process. The PLIK kurtosis figures are somewhat harder to interpret. Even though the PLIK right-hand tail is also larger than the right-hand tail of a normal density, the left-hand tail

is sufficiently smaller than a normal tail to produce a kurtosis less than 3.00. Overall, these results hint at potential problems with moments as summary measures of non-normal prediction functions and lend support to the use of nominal percentiles in making forecast statements.

4. The geometric random walk model

The numerical results for the AR(1) model suggest, in a setting naturally favorable to mean squared error analysis, two basic qualitative conclusions. First, accounting for parameter uncertainty can be important. Second, for those techniques that do account for parameter uncertainty, the differences do not appear as striking as the similarities. We attribute this to the normal density $f(y_f|y_d; \theta^0)$ that all three techniques reproduce faithfully if θ^0 is known. The probability density $f(y_f|y_d, \theta^0)$ is not a normal density, however, for many important econometric models.

We illustrate the effects of a non-normal true density using the simple geometric random walk of eq. (2),

$$y_t/y_{t-1} = \alpha + \varepsilon_t, \qquad t = 1, \ldots, m+n,$$

where $\varepsilon_t \sim \mathrm{N}(0, \sigma^2)$. For we suppose that the sample is large enough to allow us to focus on uncertainty in α, but not in σ^2. We also note that adding a term $X_t\beta$ to (2) allow for exogenous variables poses no problems.

The difficulties of a mean squared error analysis are apparent for the two-period-ahead forecast error:

$$y_{m+2} - \hat{y}_{m+2} = \alpha^2 - \hat{\alpha}^2 + \alpha(\varepsilon_{m+1} + \varepsilon_{m+2}) + \varepsilon_{m+1}\varepsilon_{m+2}, \tag{10}$$

where $\hat{y}_{m+2} = y_m\hat{\alpha}^2$. A linearization would yield a mean square measure of parameter uncertainty in this case. An asymptotic argument would justify the approximation $m^{1/2}(\alpha^2 - \hat{\alpha}^2) \to^{\mathrm{d}} \mathrm{N}(0, 4\alpha^2 V(\hat{\alpha}_d))$, and the expectation of $\alpha(\varepsilon_{m+1} + \varepsilon_{m+2})$ is zero. The term $\varepsilon_{m+1}\varepsilon_{m+2}$ is not normally distributed, and its mean and variance, which are 0 and σ^4, are hardly a complete description of its shape. For n periods ahead, $\varepsilon_{m+1} \times \cdots \times \varepsilon_{m+n}$ presents even more formidable problems. First one would have to determine the mean squared error of the product of normals which is itself a non-normal density. More importantly, this is the only term that does not diminish with larger data period sample sizes.

The predictive likelihood approach, on the other hand, is even simpler than for the AR(1) model. For (2), the statistic $s_d = \sum_{t=1}^{m} y_t/y_{t-1}$ is sufficient, and Definition 1 can be applied directly using a change of variables between s_f and y_f. We can also obtain an identical predictive likelihood function using Definition 2, avoiding reliance on the existence of sufficient statistics. The true

density is

$$f(y_f|y_d,\theta) \propto \exp\left\{-\tfrac{1}{2}\sum_{t=m+1}^{m+n}(y_t/y_{t-1}-\alpha)^2/\sigma^2\right\}. \tag{11}$$

The first and second log derivatives are then

$$\nabla_f = \sum_{t=m+1}^{m+n}(y_t/y_{t-1}-\alpha)/\sigma^2, \tag{12}$$

and

$$H_{d+f} = -(m+n)/\sigma^2. \tag{13}$$

Applying Definition 2 to matrix expressions for (11), (12) and (13) and using Rao (1973, p. 33) then yields

$$\text{plik}(y_f|y_d) \propto \exp\left\{-\tfrac{1}{2}(r_f-\hat{\alpha}_d\mathbf{1})'[\sigma^2 I + m^{-1}\sigma^2\mathbf{1}\mathbf{1}']^{-1}(r_f-\hat{\alpha}_d\mathbf{1}),\right. \tag{14}$$

where $r_t = y_t/y_{t-1}$ and $\mathbf{1} = (1,\ldots,1)'$. The covariance matrix in this normal density is the sum of the error term variance $\sigma^2 I$ and the parameter variance $m^{-1}\sigma^2\mathbf{1}\mathbf{1}'$.

The predictive likelihood function (14) takes on the familiar form of the probability density for the forecast errors $r_f - \hat{\alpha}_d\mathbf{1}$. Confidence intervals for y_f based on (14), unlike confidence intervals based on mean squared errors, lie by construction entirely within the positive support of $f(y_f|y_d,\theta^0)$ and are asymmetric in accord with the asymmetry inherent in the true density. Further, it is well known that $\hat{\alpha}_d$ is an optimal predictor of r_f.

A slight extension allows us to further distinguish the Lauritzen–Hinkley definition, our definition, and Monte Carlo techniques for a simple extension of the geometric random walk model. Suppose that the growth rate errors ε_t are generated by an AR(1) process,

$$\varepsilon_t = \rho\varepsilon_{t-1} + \eta_t, \qquad \eta_t \sim \text{N}(0,\sigma^2), \tag{15}$$

where η_t, $t = 1,\ldots,m+n$, are independent. Although Definition 1 is not applicable to this model, our extended definition (5) can be applied in a straightforward manner. The true density is

$$f(y_f|y_d,\theta^0) \propto \exp\left\{-\tfrac{1}{2}\sum_{t=m+1}^{m+n}\left(y_t/y_{t-1}-\alpha-\rho(y_{t-1}/y_{t-2}-\alpha)\right)^2/\sigma^2\right\}. \tag{16}$$

Differentiating (16) with respect to α and ρ is straightforward as is applying (5). Weighting draws from $f(y_f|y_d, \hat{\theta}_d)$ by $\exp\{-\frac{1}{2}\nabla_f H_{d+f}^{-1}\nabla_f'\}$ as described in the appendix is an easy Monte Carlo technique. Although the 'draw the error terms, draw the coefficients' Monte Carlo technique does not produce draws from (5) in this case, it might be expected to yield similar numerical results. A mean squared error analysis again faces the severe obstacle that the forecast errors for two or more periods ahead are multiplicative combinations of normally distributed random variables even if α and ρ are known.

6. Summary and conclusions

Our goal here has been twofold. First, we have tried to present the most commonly used approaches to prediction in a way that clarifies their relationship to one another and highlights their theoretical and practical advantages or shortcomings. Our interest in doing this stems from our own work on developing and studying one approach to prediction – predictive likelihood functions. Our second objective was to examine the performance of a variety of prediction functions in the context of a common form of econometric model. This is in contrast to many similar studies that examine only the performance of an advocated technique contrasted occasionally, perhaps, with a strawman candidate.

Our experiments are intended to examine the importance of two features of prediction functions – the way they account for parameter uncertainty (or fail to) and whether they capture the correct form of the density of future observations. One important feature of predictive likelihood analysis is that it focuses on the entire shape of $f(y_f; \theta^0)$ while accounting for parameter uncertainty. A mean squared error approach measures only first and second moments of the forecast errors, and even a Monte Carlo approach is often reduced in practice to computing only two moments. While these simplifications are in some cases algebraically convenient, non-linearities in variables or parameters and dependencies across observations can lead to important asymmetries in forecast error distributions. The results of experiment 3 underscore the importance of capturing the correct shape of the asymmetries introduced by parameter uncertainty.

The predictive likelihood approach is at least as informative as mean squared error analysis in a setting favorable to the latter. In our experiments for the AR(1) model, all error term uncertainty is normally distributed and non-linearities in parameter uncertainty arise only for forecasts two or more periods into the future. Less favorable circumstances such as those involved in the geometric random walk model typically complicate mean squared error analysis and introduce more approximations than are required to implement predictive likelihood analysis. Even so the asymmetries captured by the latter technique are important and evident for the AR(1) model in experiments 2 and 3.

The relatively close agreement of results for predictive likelihood analysis and the 'draw the error terms, draw the coefficients' Monte Carlo technique are not surprising. The two methods are identical for the AR(1) model in the absence of future error term uncertainty. With future error term uncertainty, both procedures deal with the asymmetries in forecast errors introduced by the dependency among observations. While further study will be required to support any generalization, one view of this situation is that coefficient and parameter draws are simply a numerical technique for sampling from a function that closely resembles the predictive likelihood function or the Bayesian posterior for some priors. From that viewpoint, Monte Carlo forecasting is an implementation technique rather than an alternative forecasting theory.

Appendix: Technical details of Monte Carlo methodology

The computational techniques outlined here for the AR(1) model are suitable for a wide range of models. Suppose a generic prediction function for a vector y_f can be written as proportional to

$$\exp\{g(y_f;\hat{\theta}_d)+h(y_f;\hat{\theta}_d)\},$$

where the multivariate normal density $\exp\{g(y_f;\theta^0)\}$ is the true density for y_f and $h(y_f;\hat{\theta}_d)$ is a correction for parameter uncertainty. (For notational simplicity, we are here omitting a constant necessary to ensure that the prediction function integrates to unity.) Our approach to working with predictive likelihood functions numerically incorporates (i) integrating to get marginal likelihoods, (ii) importance sampling, and (iii) antithetic variates.

Although the marginal likelihood of y_{m+n} could be found by integrating $\text{plik}(y_f|\hat{\theta}_d)$ over $y_{m+1},\ldots,y_{m+n-1}$, it is not always necessary to actually perform this integration. Suppose that we are interested in the kth non-central moment A^k of the marginal likelihood for y_{m+n}, which could be obtained as

$$A^k=\int\int y_{m+n}^k\exp\{g(y_f;\hat{\theta}_d)+h(y_f;\hat{\theta}_d)\}\partial y_{m+n}\partial(y_{m+1},\ldots,y_{m+n-1}). \tag{A.1}$$

By numerically generating draws $y_f^{(j)}$, $j=1,\ldots,J$, from $\text{plik}(y_f|\hat{\theta}_d)$, we could form the estimate

$$A_J^k\cong J^{-1}\sum_{j=1}^{J}\left(y_{m+n}^{(j)}\right)^k.$$

For large J, this estimate converges to the desired value A^k because the values $y_{m+n}^{(j)}$, $j=1,\ldots,J$, are drawn from the marginal likelihood for y_{m+n}.

It is, of course, not generally possible to sample randomly from plik$(y_f|\hat{\theta}_d)$, and we must resort to some other sampling strategy. It is often convenient to sample from $\exp\{g(y_f; \hat{\theta}_d)\}$, generating draws $y_{m+n}^{(j)}$, $j = 1, \ldots, J$. The estimate of A^k then becomes

$$A_J^k \cong J^{-1} \sum_{j=1}^{J} \left(y_{m+n}^{(j)}\right)^k \exp\left\{h\left(y_f|\hat{\theta}_d\right)\right\}.$$

This strategy, known as importance sampling, will typically require fewer random draws than a uniform random sampling strategy to yield an accurate estimate of A^k.

More generally, we can efficiently evaluate the integral

$$A = \int_{\Omega} \exp\left\{g\left(y_f; \hat{\theta}_d\right) + h\left(y_f; \hat{\theta}_d\right)\right\} \partial y_f$$

over a region Ω by recognizing that $A = B + C$, where

$$B = \int_{\Omega} \exp\left\{g\left(y_f; \hat{\theta}_d\right)\right\} \partial y_f,$$

and

$$C = \int_{\Omega} \left[\exp\left\{h\left(y_f; \hat{\theta}_d\right)\right\} - 1\right] \exp\left\{g\left(y_f; \hat{\theta}_d\right)\right\} \partial y_f.$$

B can be computed analytically using properties of the conditional normal density $\exp\{g(y_f; \hat{\theta}_d)\}$. C can be approximated by drawing trials $y_f^{(j)}$, $j = 1, \ldots, J$, from the density $\exp\{g(y_f; \hat{\theta}_d)\}$. The average,

$$C_J = \sum_{y_f^{(j)} \in \Omega} \left[\exp\left\{h\left(y_f^{(j)}; \hat{\theta}_d\right)\right\} - 1\right],$$

then converges in probability to C as $J \to \infty$. Effectively, we can compute part B of the integral analytically, leaving only C to deal with via a numerical approximation.

We follow this strategy in computing nominal percentiles for the predictive likelihood function. Let $I(\cdot, k)$ be the indicator function defined by

$$I(y, k) = 1 \quad \text{if} \quad y \leq k,$$
$$= 0 \quad \text{if} \quad y > k.$$

The probability

$$P = \int\int I(y_{m+n}, k) \exp\left\{g\left(y_f; \hat{\theta}_d\right) + h\left(y_f; \hat{\theta}_d\right)\right\}$$
$$\times \partial y_{m+n} \partial (y_{m+1}, \ldots, y_{m+n-1})$$

is computed using

$$P_J = \sum_{j=1}^{J} I\left(y_{m+n}^{(j)}, k\right)\exp\left\{h\left(y_f; \hat{\theta}_d\right)\right\} \Big/ \sum_{j=1}^{J} \exp\left\{h\left(y_f; \hat{\theta}_d\right)\right\}$$

for draws $y_f^{(j)}$, $j = 1, \ldots, J$, from $\exp\{g(y_f; \hat{\theta}_d)\}$.

Antithetic variates may be useful in this context if $h(y_f; \hat{\theta})$ is more or less symmetric about the mode of $g(y_f; \hat{\theta})$. To implement this technique for J evenly divisible by 2, let $\varepsilon_f^{(j)}$, $j = 1, 3, 5, \ldots, J-1$, be independently drawn from $N(0, \sigma^2 I_n)$. The draws $y_f^{(j)}$, $j = 1, 3, 5, \ldots, J-1$, from $\exp\{f(y_f; \hat{\theta}_d)\}$ can then be computed as

$$y_f^{(j)} = \hat{\theta}_d z_f^{(j)} + \varepsilon_f^{(j)}, \qquad j = 1, 3, 5, \ldots, J-1,$$

where $z_t^{(j)} = [1, y_{t-1}^{(j)}]$. The antithetic variates are then

$$y_f^{(j)} = \hat{\theta}_d z_f^{(j)} - \varepsilon_f^{(j-1)}, \qquad j = 2, 4, 6, \ldots, J.$$

Introduction of antithetic variates works well for odd non-central moments because, if k is odd, then

$$\left(y_{m+n}^{(j)}\right)^k \exp\left\{h\left(y_f^{(j)}; \hat{\theta}_d\right)\right\} \quad \text{and} \quad \left(y_{m+n}^{(j-1)}\right)^k \exp\left\{h\left(y_f^{(j-1)}; \hat{\theta}_d\right)\right\}$$

are negatively correlated. Consequently, the variance of a sum involving these draws may be expected to go to zero more quickly than for independent draws.

References

Baillie, R.T., 1979, The asymptotic mean squared error of multistep prediction for the regression model with autoregressive errors, Journal of the American Statistical Association 74, 175–184.

Box, G.E.P. and G.M. Jenkins, 1976, Time series analysis: Forecasting and control (Holden-Day, San Francisco, CA).

Butler, R.W., 1986, Predictive likelihood inference with applications, Journal of the Royal Statistical Society B 48, 1–38.

Chow, G.C., 1975, Multiperiod predictions from stochastic difference equations by Bayesian methods, in: Stephen E. Fienberg and Arnold Zellner, eds., Studies in Bayesian econometrics and statistics (North-Holland, Amsterdam).

Cooley, T.F. and W.R. Parke, 1986a, Asymptotic predictive likelihood and optimal forecasting, Mimeo. (University of California, Santa Barbara, CA).

Cooley, T.F. and W.R. Parke, 1986b, Asymptotic likelihood based prediction functions, Mimeo. (University of California, Santa Barbara, CA).

Cooley, T.F., W.R. Parke and S. Chib, 1986, Prediction functions, Mimeo. (University of California, Santa Barbara, CA).

Cox, D.R. and D.V. Hinkley, 1974, Theoretical statistics (Chapman and Hall, London).
Davison, A.C., 1985, Approximate predictive likelihood, Mimeo. (University of Texas, Austin, TX).
Edwards, A.W.F., 1972, Likelihood (Cambridge University Press, Cambridge).
Fuller, W.A. and D.P. Hasza, 1981, Properties of predictors for autoregressive time series, Journal of the American Statistical Association 76, 155–161.
Hendry, D.F., 1984, Monte Carlo experimentation in econometrics, in: Handbook of econometrics, Vol. 2 (North-Holland, Amsterdam) 939–976.
Hinkley, D.V., 1979, Predictive likelihood, Annals of Statistics 7, 718–728.
Lauritzen, S.L., 1974, Sufficiency, prediction and extreme models, Scandinavian Journal of Statistics 1, 128–134.
Mariano, R. and B. Brown, 1986, Asymptotic behavior of predictors in a nonlinear simultaneous system, International Economic Review 21, 523–536.
Mariano, R. and B. Brown, 1984, Residual based procedures for prediction and estimation of a nonlinear simultaneous system, Econometrica 52, 321–343.
Mathiasen, P.E., 1979, Prediction functions, Scandinavian Journal of Statistics 6, 1–21.
Miller, Robert B. and Patrick A. Thompson, 1986, Sampling the future: A Bayesian approach to forecasting from univariate time series models, Journal of Business and Economic Statistics 4, 427–436.
Rao, C.R., 1973, Linear statistical inference and its applications, 2nd ed. (Wiley, New York).
Rubinstein, Reuven Y., 1981, Simulation and the Monte Carlo method (Wiley, New York).
Spitzer, J.J. and R.T. Baillie, 1983, Small sample properties of predictions from the regression model with autoregressive error, Journal of the American Statistical Association 78, 258–263.
Yamamoto, T., 1976, Asymptotic mean square prediction error for an autoregressive model with estimated coefficients, Journal of the Royal Statistical Society C 25, 123–127.
Zellner, A., 1971, An introduction to Bayesian inference in econometrics (Wiley, New York).