7

Probability forecasting: Concepts and analysis

Having considered the econometric issues involved in the estimation, testing and interpretation of a long-run structural VAR model, in this chapter we turn attention to the use of the model in probability forecasting. Much of the material would be relevant to forecasts based on any type of model. However, the material is particularly relevant here since VARs are frequently employed in forecasting. Moreover, given the size and simplicity of the structure of most VAR models, these models are particularly well-suited to an investigation of the various types of uncertainty that influence forecasts, and their use in decision-making.

7.1 Probability forecasting

In much of what follows, we are concerned with the notion of *probability forecasting*, arguing that these convey the uncertainties surrounding forecasts from a macroeconomic model in a very straightforward way and one that is most useful in decision-making. A probability forecast is a statement of the likelihood of a specified event taking place conditional on the available information and can be estimated on the basis of any macroeconomic model. The event can be defined with respect to the values of a single variable or a set of variables, measured at a particular time, or at a sequence of times, or over a particular interval of time in the future.

For example, in a macroeconomic context, suppose that the focus of interest is inflation, Δp_t, and output growth, Δy_t. Then events that might

be of interest include

$$\begin{aligned}
\pi_{1t} &= \Pr\left(\Delta p_{t+1} < a_1 \mid \Im_t\right), \\
\pi_{2t} &= \Pr\left(\Delta p_{t+1} < a_1, \Delta y_{t+1} > a_2 \mid \Im_t\right), \\
\pi_{3t} &= \Pr\left(\Delta p_{t+h} < a_1, \Delta y_{t+h} > a_2 \mid \Im_t\right), \; h \geq 1, \\
\pi_{4t} &= \Pr\left(\Delta p_{t+1} < a_1, \Delta p_{t+2} < a_1, \Delta p_{t+3} < a_1, \Delta p_{t+4} < a_1 \mid \Im_t\right),
\end{aligned} \tag{7.1}$$

where $\Im_t$ denotes a non-decreasing information set up to time t. The first example illustrates a *single* event while the others relate to *joint* events involving either more than one variable or a variable considered at more than one time horizon. Examples one and two are concerned with the *one-step ahead* forecast horizon, example three is concerned with an *h-step ahead* forecast horizon, and the fourth example relates to a *multiple-step ahead* forecast horizon. The probability of all events are conditional on $\Im_t$.

The calculation of probability forecasts remains relatively unusual, however. Macroeconomic forecasts are typically presented in the form of point forecasts and their uncertainty is characterised (if at all) by forecast confidence intervals. Focusing on point forecasts is justified when the underlying decision problems faced by agents and the government are linear in constraints and quadratic in the loss function; the so-called LQ problem. But for most decision problems, reliance on point forecasts will not be sufficient and probability forecasts will be needed (see, for example Granger and Pesaran, 2000a,b).

The need for probability forecasts is also acknowledged by a variety of researchers and institutions. In the statistics literature, for example, Dawid (1984) has been advocating the use of probability forecasting in a sequential approach to the statistical analysis of data; the so-called 'prequential approach'. In the macroeconometric modelling literature, Fair (1980) was one of the first to compute probability forecasts using a macroeconometric model of the US economy. For example, in a macroeconomic context, the motivation for the current monetary policy arrangements in the UK is that it provides for transparency in policy-making and an economic environment in which firms and individuals are better able to make investment and consumption decisions. The range of possible decisions that a firm can make regarding an investment plan represents the firm's action space. The 'states of nature' in this case are defined by all of the possible future out-turns for the macroeconomy. For example, referring to the illustrative events above, the investment decision might rely on inflation in the next period, or the average rate of inflation over some longer period, remaining below a target level; or interest might focus on the future path

of inflation and output growth considered together. In making a decision, the firm should define a loss function which evaluates the profits or losses associated with each point in the action space and given any 'state of nature'. Except for LQ decision problems, decision rules by individual households and firms will generally require probability forecasts with respect to different threshold values reflecting their specific cost–benefit ratios. For this purpose, we need to provide estimates of the whole probability distribution function of the events of interest, rather than point forecasts or particular forecast intervals which are likely to be relevant only to the decision problem of a few. Probability event forecasts can also convey important information on the properties of a model. For example, long-run neutrality of output growth to inflation (or *vice versa*) would imply that

$$\begin{aligned}
&\lim_{h\to\infty} \Pr\left(\Delta p_{t+h} < a_1, \Delta y_{t+h} > a_2 \mid \Im_t\right) \\
&\quad = \left[\lim_{h\to\infty} \Pr\left(\Delta p_{t+h} < a_1 \mid \Im_t\right)\right] \times \left[\lim_{h\to\infty} \Pr\left(\Delta y_{t+h} > a_2 \mid \Im_t\right)\right].
\end{aligned} \tag{7.2}$$

7.1.1 *Probability forecasts in a simple univariate AR(1) model*

As an illustration we first consider probability forecasts in the case of a simple univariate AR(1) model. This serves to illustrate the use of the concept in a simple context, but also demonstrates some of the (perhaps surprising) features of probability forecasts and highlights the problems involved in calculating probability forecasts analytically (as opposed to the use of the simulation methods described below).

Consider the following AR(1) model for the log of real output y_t:

$$y_t = \mu + (1-\rho)\gamma t + \rho y_{t-1} + u_t, \quad t = 1, 2, \ldots, T, T+1, \ldots, T+H, \tag{7.3}$$

where u_t's are independently and identically distributed random variables with a zero mean and variance σ^2. In the case where output can be assumed to be trend stationary (*i.e.* $|\rho| < 1$), the trend growth rate of y_t is given by γ. In the case where y_t is difference stationary (*i.e.* $\rho = 1$), the average growth rate will be given by μ. The restricted specification of the trend coefficient in (7.3) ensures that irrespective of whether y_t is trend stationary or first difference stationary its deterministic trend component is linear.

Defining the lag polynomial

$$\rho_h(L) = 1 + \rho L + \rho^2 L^2 + \cdots + \rho^{h-1} L^{h-1},$$

then, by successive substitution in (7.3), we can obtain

$$y_{T+h} = \rho^h y_T + \rho_h(L)\left[\mu + (1-\rho)\gamma(T+h) + u_{T+h}\right], \; h = 1, 2, \ldots H,$$

which after some algebra yields

$$y_{T+h} = \rho^h y_T + \delta(h, T) + h\gamma + v_{T+h}, \tag{7.4}$$

where

$$\delta(h, T) = \left(\frac{1-\rho^h}{1-\rho}\right)\mu - \left(\frac{1-\rho^h}{1-\rho}\right)\rho\gamma + T(1-\rho^h)\gamma,$$

$$v_{T+h} = \sum_{i=0}^{h-1} \rho^i u_{T+h-i}.$$

For a given initial value and the sample size, T, the sum of the terms $\rho^h y_T$ and $\delta(h, T)$ is of $O(1)$ in h and will be dominated by $h\gamma$ as the forecast horizon, h, is extended. Note that

$$\lim_{h\to\infty} \delta(h, T) = \frac{\mu - \rho\gamma}{1-\rho} + T\gamma, \quad \text{if } |\rho| < 1,$$

and $\delta(h, T) = h(\mu - \gamma)$ if $\rho = 1$. Therefore, for sufficiently large h, the deterministic component of y_{T+h} will be given by $h\gamma + T\gamma + (\mu - \rho\gamma)/(1-\rho)$ if $|\rho| < 1$, and by $y_T + h\mu$ if $\rho = 1$. It is interesting to note that irrespective of whether y_t has a unit root or not, the mean of the h-step ahead forecast will be of the same order of magnitude.

Also, for reasonably long forecast horizons, the composite error term v_{T+h} will be approximately distributed as a normal variate even if the underlying errors, u_t, were not normally distributed. In particular, for sufficiently large h we have

$$v_{T+h} \backsimeq N\left[0, \sigma^2\left(\sum_{j=1}^{h} \rho^{2(j-1)}\right)\right]. \tag{7.5}$$

Unlike the point forecasts, the orders of the variance of the h-step ahead forecasts differ depending on whether $|\rho| < 1$ or $\rho = 1$. Under the former $V(y_{T+h}|\Im_t) = V(v_{T+h}) = O(1)$, whilst under the latter $V(v_{T+h}) = O(h)$. But as we shall see, the probability forecasts have similar limit properties under $|\rho| < 1$ or $\rho = 1$, when y_t contains deterministic trends.

FORECASTING GROWTH PROBABILITIES: AN ANALYTIC SOLUTION

To illustrate the nature of probability forecasts in the univariate AR(1) model, we present below expressions for forecasts of output growth over different horizons. Specifically, we consider the four-period average growth rate over the period $T+h-4$ to $T+h$, for any arbitrary horizon h, and also the average growth rate in y_t over the period T to $T+h$, for horizon h. The four-period average growth rate is given by

$$\frac{y_{T+h} - y_{T+h-4}}{4} = f_1(\rho, y_T, \delta(4, T)) + \gamma + \frac{v_{T+h} - v_{T+h-1}}{4}, \quad h = 4, 5, \ldots H, \tag{7.6}$$

where $f_1(\rho, y_T, \delta(4, T)) = \rho^{h-4}\{-(1-\rho^4)y_T + \delta(4, T)\}/4$, while the average growth rate of y_t over the period T to $T+h$ is given by

$$\frac{y_{T+h} - y_T}{h} = f_2(\rho, y_T, \delta(h, T)) + \gamma + h^{-1} v_{T+h}, \quad h = 1, 2, \ldots H, \tag{7.7}$$

where $f_2(\rho, y_T, \delta(h, T)) = \{-(1-\rho^h)y_T + \delta(h, T)\}/h$. The four-period average given by (7.6) provides a good example of a typical event of interest; setting $h = 4$, for example, we would have the annual growth rate over the coming year if quarterly data were used. Given the trended nature of the y_t process, the 'long average' in (7.7) provides useful insights on the long-run properties of the probability forecasts. In what follows, we examine probability forecasts of $(y_{T+h} - y_T)/h$ in both the stationary and unit root cases, but we focus on the case where parameters are known, so that the only source of uncertainty relates to the future shocks.

Case 1: y_t is trend stationary ($|\rho| < 1$)

If y_t is trend stationary, then (7.5) provides

$$\frac{v_{T+h} - v_{T+h-4}}{4} \sim N\left[0, \frac{\sigma^2}{4^2}\left(1+\rho^2\right)\left(2 - (1-\rho^4)\rho^{2(h-4)}\right)\right], \tag{7.8}$$

while

$$\frac{1}{h} v_{T+h} \sim N\left[0, \frac{\sigma^2}{h^2}\left(\frac{1-\rho^{2h}}{1-\rho^2}\right)\right]. \tag{7.9}$$

From (7.6) and using (7.8), we have

$$\Pr\left(\frac{y_{T+h}-y_{T+h-4}}{4} < a \mid \mathfrak{I}_T\right)$$
$$= \Pr\left\{\frac{v_{T+h}-v_{T+h-4}}{4} < [a-\gamma-f_1(\rho,y_T,\delta(4,T))] \mid \mathfrak{I}_T\right\}$$
$$= \Phi\left\{\frac{4[a-\gamma-f_1(\rho,y_T,\delta(4,T))]}{\sigma\sqrt{(1+\rho^2)(2-(1-\rho^4)\rho^{2(h-4)})}}\right\},$$

while from (7.7)

$$\Pr\left(\frac{y_{T+h}-y_T}{h} < a \mid \mathfrak{I}_T\right)$$
$$= \Pr\left\{h^{-1}v_{T+h} < [a-\gamma-f_2(\rho,y_T,\delta(h,T))] \mid \mathfrak{I}_T\right\},$$
$$= \Phi\left\{\frac{h\sqrt{1-\rho^2}[a-\gamma-f_2(\rho,y_T,\delta(h,T))]}{\sigma\sqrt{1-\rho^{2h}}}\right\}, \tag{7.10}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of a standard normal variate. For sufficiently large h we now have

$$\lim_{h\to\infty}\left[\Pr\left(\frac{y_{T+h}-y_{T+h-4}}{4} < a \mid \mathfrak{I}_T\right)\right] = \Phi\left(\frac{4(a-\gamma)}{\sigma\sqrt{2(1+\rho^2)}}\right),$$

and the probability of the four-period average falling below a given threshold converges to a constant.

For the long average, as $h \to \infty$ we have

$$\lim_{h\to\infty}\left[\Pr\left(\frac{y_{T+h}-y_T}{h} < a \mid \mathfrak{I}_T\right) - \Phi\left(\frac{h\sqrt{1-\rho^2}(a-\gamma)}{\sigma}\right)\right] = 0, \tag{7.11}$$

and hence

$$\lim_{h\to\infty}\Pr\left(\frac{y_{T+h}-y_T}{h} < a \mid \mathfrak{I}_t\right) = \left\{\begin{array}{ll} 1 & \text{if } a>\gamma \\ 0.5 & \text{if } a=\gamma \\ 0 & \text{if } a<\gamma \end{array}\right\}.$$

This shows that, at the infinite horizon, the probability of events relating to the long average will typically degenerate to values of zero or one, depending on the value of the trend growth rate, γ, relative to the selected threshold value. This property follows directly from the fact that y_t tends

(mean-reverts) to its deterministic trend path as $h \to \infty$. In addition this result also explains why the long-run forecasts of trend stationary models are not affected by intercept adjustments.

Case 2: y_t has a unit root ($\rho = 1$).

In this case, the analysis simplifies considerably and we have

$$\Pr\left(\frac{y_{T+h}-y_{T+h-4}}{4} < a \mid \mathfrak{I}_T\right) = \Phi\left[\frac{\sqrt{4}(a-\mu)}{\sigma}\right], \tag{7.12}$$

and

$$\Pr\left(\frac{y_{T+h}-y_T}{h} < a \mid \mathfrak{I}_T\right) = \Phi\left[\frac{\sqrt{h}(a-\mu)}{\sigma}\right]. \tag{7.13}$$

For the long average case

$$\lim_{h\to\infty}\Pr\left(\frac{y_{T+h}-y_T}{h} < a \mid \mathfrak{I}_T\right) = \left\{\begin{array}{ll} 1 & \text{if } a>\mu \\ 0.5 & \text{if } a=\mu \\ 0 & \text{if } a<\mu \end{array}\right\}, \tag{7.14}$$

which is the same as the result obtained for the trend stationary case.

The above discussion highlights an extremely important property of the probability forecasts, showing that the probability of the long-run average growth rate, $(y_{T+h}-y_T)/h$, will take a value of zero or one at the infinite horizon whether or not there exists a unit root in the series. Comparison of (7.10) and (7.13) shows that the speeds with which the probability forecasts degenerate are given by $h\sqrt{(1-\rho^2)}$ and $\sqrt{h}$ for the trend stationary and the unit root processes, respectively. Thus, the main distinction between the stationary and unit root case is the speed with which the zero/unity boundary is reached.

Consider now the effect of parameter uncertainty on the probability forecasts, and for simplicity assume that $\rho = 1, \sigma^2$ is given and that the unknown mean growth rate, μ, is estimated by the sample mean, $\widehat{\mu} = T^{-1}\sum_{t=1}^{T}\Delta y_t$. To allow for parameter uncertainty, we first write (7.7) for $\rho = 1$ as

$$\frac{y_{T+h}-y_T}{h} = \widehat{\mu} + (\mu-\widehat{\mu}) + h^{-1}v_{T+h}, \tag{7.15}$$

and let μ to be unknown conditional on the past observations given by the information set, $\mathfrak{I}_T = \{y_1, y_2, \ldots, y_T\}$. The uncertainty associated with

μ can be characterised by[1]

$$\mu - \widehat{\mu} | \Im_T \sim N\left(0, \frac{\sigma^2}{T}\right). \tag{7.16}$$

This result can be viewed as the posterior distribution of μ with respect to diffuse priors for μ. Using (7.16) in conjunction with (7.15), we have

$$h^{-1}\left(y_{T+h} - y_T\right) \sim N\left(\widehat{\mu}, \frac{\sigma^2}{T} + \frac{\rho^2}{h}\right),$$

and therefore,

$$\Pr\left(\frac{y_{T+h} - y_T}{h} < a \mid \Im_T\right) = \Phi\left(\frac{a - \widehat{\mu}}{\sigma\sqrt{\frac{1}{T} + \frac{1}{h}}}\right).$$

The result in this case depends on the relative size of T and h. For a fixed T and as $h \to \infty$,

$$\lim_{h\to\infty} \Pr\left(\frac{y_{T+h} - y_T}{h} < a \mid \Im_T\right) = \Phi\left(\frac{\sqrt{T}[a - \hat{\mu}]}{\sigma}\right),$$

which differ from the limit result given by (7.14) when μ is known. Clearly, result (7.14) follows if T and $h \to \infty$, jointly. In this case the uncertainty surrounding the value of μ vanishes as $T \to \infty$ and we return to the case of known μ. In the case where h is relatively small, the effect of parameter uncertainty on the probability estimates is of order T^{-1}. To establish this result we first write $\pi_t = \Pr[h^{-1}\left(y_{T+h} - y_T\right) < a \mid \Im_T]$ as

$$\pi_t(x) = \Phi\left[\theta\,(1+x)^{-1/2}\right],$$

where $\theta = \sqrt{h}\,(a - \widehat{\mu})/\sigma$ and $x = h/T$. Expanding $\pi_t(x)$ around $x = 0$, we have[2]

$$\pi_t(x) = \pi_t(0) - \left[\frac{\theta}{2}\phi(\theta)\right]x + O(x^2),$$

where $\pi_t(0)$ corresponds to the probability estimate that ignores parameter uncertainty. Hence, for finite h we have

$$\pi_t(x) = \pi_t(0) + O\left(\frac{h}{T}\right),$$

[1] It is also assumed that conditional on $\Im_T$, μ and v_{T+h} are i.i.d. normal variables.
[2] Such an expansion is sensible since h is assumed to be small relative to T.

as required. This result holds more generally and in practice the effect of parameter uncertainty on probability forecasts would be of second-order importance when h is small and T relatively large.

7.2 Modelling forecast uncertainties

Returning to a more general setting, model-based forecasts are subject to five different types of uncertainties:

- future uncertainty
- parameter uncertainty (for a given model)
- model uncertainty
- policy uncertainty
- measurement uncertainty (data inadequacies and measurement errors).

Here, we focus on the first three and consider how to allow for them in the computation of probability forecasts. Policy and measurement uncertainties pose special problems of their own and will not be addressed here. Future uncertainty refers to the effects of unobserved future shocks on forecasts, while parameter and model uncertainties are concerned with the robustness of forecasts to the choice of parameter values (for a given model) and more generally the alternative models under consideration.[3]

7.2.1 *Future and parameter uncertainties*

The standard textbook approach to taking account of future and parameter uncertainties is through the use of confidence intervals around point forecasts. Instead of a point forecast, an interval forecast is provided. Although such forecast intervals may contain important information about probability forecasts of interest to a particular decision-maker, they do not allow for a full recovery of the forecast probability distribution function which is needed in decision-making contexts where the decision problem is not of the LQ type. The relationships between forecast intervals and probability forecasts become even more tenuous when forecasts of *joint* events or forecasts from multiple models are considered. For example, it would be impossible to infer the probability of the joint event of a positive output growth and an inflation rate falling within a pre-specified range from

[3] For a discussion on the problem of model uncertainty, see Draper (1995) and Chatfield (1995).

given variable-specific forecast intervals. In fact, even if the primary object of interest is a point forecast, as we shall see below, consideration of probability forecasts can help clarify how best to pool point mean and volatility forecasts in the presence of model uncertainty.

For the purpose of exposition, initially we abstract from parameter uncertainty and consider the following simple linear regression model:

$$y_t = \mathbf{x}'_{t-1}\boldsymbol{\beta} + u_t, \qquad t = 1, 2, \ldots, T,$$

where $\mathbf{x}_{t-1}$ is a $k \times 1$ vector of predetermined regressors, $\boldsymbol{\beta}$ is a $k \times 1$ vector of fixed but unknown coefficients, and $u_t \sim N(0, \sigma^2)$. The optimal forecast of y_{T+1} at time T (in the mean squared error sense) is given by $\mathbf{x}'_T\boldsymbol{\beta}$. In the absence of parameter uncertainty, the calculation of a probability forecast for a specified event is closely related to the more familiar concept of forecast confidence interval. For example, suppose that we are interested in the probability that the value of y_{T+1} lies below a specified threshold, say a, conditional on $\mathfrak{I}_T = (y_T, \mathbf{x}_T, y_{T-1}, \mathbf{x}_{T-1}, \ldots)$, the information available at time T. For given values of $\boldsymbol{\beta}$ and σ^2, we have

$$\Pr\left(y_{T+1} < a \mid \mathfrak{I}_T\right) = \Phi\left(\frac{a - \mathbf{x}'_T\boldsymbol{\beta}}{\sigma}\right),$$

where as before $\Phi(\cdot)$ is the standard Normal cumulative distribution function while the $(1-\alpha)$% forecast interval for y_{T+1} (conditional on $\mathfrak{I}_T$) is given by $\mathbf{x}'_T\boldsymbol{\beta} \pm \sigma\Phi^{-1}(1 - (\alpha/2))$.

The two approaches, although related, are motivated by different considerations. The point forecast provides the threshold value $a = \mathbf{x}'_T\boldsymbol{\beta}$ for which $\Pr\left(y_{T+1} < a \mid \mathfrak{I}_T\right) = 0.5$, while the forecast interval provides the threshold values $c_L = \mathbf{x}'_T\boldsymbol{\beta} - \sigma\Phi^{-1}(1 - (\alpha/2))$, and $c_U = \mathbf{x}'_T\boldsymbol{\beta} + \sigma\Phi^{-1}(1 - (\alpha/2))$ for which $\Pr\left(y_{T+1} < c_L \mid \mathfrak{I}_T\right) = \alpha/2$ and $\Pr\left(y_{T+1} < c_U \mid \mathfrak{I}_T\right) = 1 - (\alpha/2)$. Clearly, the threshold values, c_L and c_U, associated with the $(1-\alpha)$% forecast interval may or may not be of interest.[4] Only by chance will the forecast interval calculations provide information in a way which is directly useful in specific decision-making contexts.

The relationship between probability forecasts and interval forecasts becomes even more obscure when parameter uncertainty is also taken into account. In the context of the above regression model, the point estimate

[4] The association between probability forecasts and interval forecasts is even weaker when one considers *joint* events. Many different such intervals will be needed for the purpose of characterising the probability forecasts of joint events.

of the forecast is given by $\widehat{y}_{T+1} = \mathbf{x}'_T\widehat{\boldsymbol{\beta}}_T$, where

$$\widehat{\boldsymbol{\beta}}_T = \mathbf{Q}_{T-1}^{-1}\mathbf{q}_T,$$

is the Ordinary Least Squares (OLS) estimate of $\boldsymbol{\beta}$, with

$$\mathbf{Q}_{T-1} = \sum_{t=1}^{T}\mathbf{x}_{t-1}\mathbf{x}'_{t-1}, \quad \text{and} \quad \mathbf{q}_T = \sum_{t=1}^{T}\mathbf{x}_{t-1}y_t.$$

The relationship between the actual value of y_{T+1} and its time T predictor can be written as

$$\begin{aligned} y_{T+1} &= \mathbf{x}'_T\boldsymbol{\beta} + u_{T+1} \\ &= \mathbf{x}'_T\widehat{\boldsymbol{\beta}}_T + \mathbf{x}'_T(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_T) + u_{T+1}, \end{aligned} \tag{7.17}$$

so that the forecast error, ξ_{T+1}, is given by

$$\xi_{T+1} = y_{T+1} - \widehat{y}_{T+1} = \mathbf{x}'_T(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_T) + u_{T+1}.$$

This example shows that the point forecasts, $\mathbf{x}'_T\widehat{\boldsymbol{\beta}}_T$, are subject to two types of uncertainties, namely that relating to $\boldsymbol{\beta}$ and that relating to the distribution of u_{T+1}. For any given sample of data, $\mathfrak{I}_T$, $\widehat{\boldsymbol{\beta}}_T$ is known and can be treated as fixed. On the other hand, although $\boldsymbol{\beta}$ is assumed fixed at the estimation stage, it is unknown to the forecaster and, from this perspective, it is best viewed as a random variable at the forecasting stage. Hence, in order to compute probability forecasts which account for future as well as parameter uncertainties, we need to specify the joint probability distribution of $\boldsymbol{\beta}$ and u_{T+1}, conditional on $\mathfrak{I}_T$. As far as u_{T+1} is concerned, we continue to assume that

$$u_{T+1} | \mathfrak{I}_T \sim N(0, \sigma^2),$$

and to keep the exposition simple, for the time being we shall assume that σ^2 is known and that u_{T+1} is distributed independently of $\boldsymbol{\beta}$. For $\boldsymbol{\beta}$, noting that

$$\left(\widehat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}\right) | \mathfrak{I}_T \sim N\left(\mathbf{0}, \sigma^2\mathbf{Q}_{T-1}^{-1}\right), \tag{7.18}$$

we assume that

$$\boldsymbol{\beta} \mid \mathfrak{I}_T \sim N\left(\widehat{\boldsymbol{\beta}}_T, \sigma^2\mathbf{Q}_{T-1}^{-1}\right), \tag{7.19}$$

which is akin to a Bayesian approach with non-informative priors for $\boldsymbol{\beta}$. Hence

$$\xi_{T+1} \mid \mathfrak{I}_T \sim N\left[0, \sigma^2\left(1 + \mathbf{x}'_T\mathbf{Q}_{T-1}^{-1}\mathbf{x}_T\right)\right].$$

The $(1-\alpha)\%$ forecast interval in this case is given by

$$c_{LT} = \mathbf{x}_T'\widehat{\boldsymbol{\beta}}_T - \sigma\left\{1+\mathbf{x}_T'\mathbf{Q}_{T-1}^{-1}\mathbf{x}_T\right\}^{1/2}\Phi^{-1}\left(1-\frac{\alpha}{2}\right), \tag{7.20}$$

and

$$c_{UT} = \mathbf{x}_T'\widehat{\boldsymbol{\beta}}_T + \sigma\left\{1+\mathbf{x}_T'\mathbf{Q}_{T-1}^{-1}\mathbf{x}_T\right\}^{1/2}\Phi^{-1}\left(1-\frac{\alpha}{2}\right). \tag{7.21}$$

When σ^2 is unknown, under the standard non-informative Bayesian priors on $(\boldsymbol{\beta},\sigma^2)$, the appropriate forecast interval can be obtained by replacing σ^2 by its unbiased estimator, $\hat{\sigma}_T^2 = (T-k)^{-1}\sum_{t=1}^{T}(y_t-\mathbf{x}_{t-1}'\widehat{\boldsymbol{\beta}}_T)'(y_t-\mathbf{x}_{t-1}'\widehat{\boldsymbol{\beta}}_T)$, and $\Phi^{-1}(1-(\alpha/2))$ by the $(1-(\alpha/2))\%$ critical value of the standard t-distribution with $T-k$ degrees of freedom. Although such interval forecasts have been discussed in the econometrics literature, the particular assumptions that underlie them are not often fully recognised.

Using this interpretation, the effect of parameter uncertainty on forecasts can also be obtained via stochastic simulations, by generating alternative forecasts of y_{T+1} for different values of $\boldsymbol{\beta}$ (and σ^2) drawn from the conditional probability distribution of $\boldsymbol{\beta}$ given by (7.19). Alternatively, one could estimate probability forecasts by focusing directly on the probability distribution of y_{T+1} for a given value of $\mathbf{x}_T$, simultaneously taking into account both parameter and future uncertainties. For example, in the simple case where σ^2 is known, this can be achieved by simulating $y_{T+1}^{(j,s)}$, where

$$y_{T+1}^{(j,s)} = \mathbf{x}_T'\widehat{\boldsymbol{\beta}}^{(j)} + u_{T+1}^{(s)}, \qquad j=1,2,\ldots,J,\ \ s=1,2,\ldots,S,$$

$\widehat{\boldsymbol{\beta}}^{(j)}$ is the jth random draw from $N\left(\widehat{\boldsymbol{\beta}}_T,\sigma^2\mathbf{Q}_{T-1}^{-1}\right)$, and $u_{T+1}^{(s)}$ is the sth random draw from $N\left(0,\sigma^2\right)$, which is independant of the drawing $\widehat{\boldsymbol{\beta}}^{(j)}$.[5] This is an example of the parametric 'bootstrap predictive density' discussed in Harris (1989). In large samples, the stochastic simulation approach will be equivalent to the analytical methods discussed above, as J and $S\to\infty$. However, as argued below, it is more generally applicable and will be used in our empirical application.

An alternative approach to allowing for the effects of future and parameter uncertainties on prediction of y_{T+1} would be to follow the literature on 'predictive likelihoods', where a predictive density for y_{T+1} conditional on $\mathfrak{I}_T$ is derived directly.[6] In the case of the regression example, the problem

[5] In the realistic case where σ^2 is unknown it is replaced by $\hat{\sigma}_T^2$.

[6] A large number of different predictive likelihoods have been suggested in the statistics literature. Bjørnstad (1990) provides a review.

has been studied by Levy and Perng (1986) who show that the optimal prediction density for y_{T+1}, in the Kullback–Leibler information-theoretic sense, is the Student t-distribution with $T-k$ degrees of freedom, having the location $\hat{y}_{T+1} = \mathbf{x}_T'\widehat{\boldsymbol{\beta}}_T$ and the dispersion $\hat{\sigma}_T^2\left(1+\mathbf{x}_T'\mathbf{Q}_{T-1}^{-1}\mathbf{x}_T\right)$. This is the same as the Bayes predictive density of $y_{T+1}\mid\mathfrak{I}_T$ with a non-informative prior on $(\boldsymbol{\beta},\sigma^2)$. In this way Levy and Perng provide a non-Bayesian interpretation of Bayes predictive density in the context of linear regression models.

7.2.2 *Model uncertainty: Combining probability forecasts*

Suppose we are interested in a decision problem that requires probability forecasts of an event defined in terms of one or more elements of $\mathbf{z}_t$, over the period $t=T+1,T+2,\ldots,T+h$, where $\mathbf{z}_t=(z_{1t},z_{2t},\ldots,z_{nt})'$ is an $n\times 1$ vector of the variables of interest and h is the forecast (decision) horizon. Assume also that the data generating process (DGP) is unknown and the forecasts are made considering m different models indexed by i (that could be nested or non-nested). Each model, M_i, $i=1,2,\ldots,m$, is characterised by a probability density function of $\mathbf{z}_t$ defined over the estimation period $t=1,2,\ldots,T$, as well as the forecast period $t=T+1,T+2,\ldots,T+h$, in terms of a $k_i\times 1$ vector of unknown parameters, $\boldsymbol{\theta}_i$, assumed to lie in the compact parameter space, $\boldsymbol{\Theta}_i$. Model M_i is then defined by

$$M_i:\left\{f_i\left(\mathbf{z}_1,\mathbf{z}_2,\ldots,\mathbf{z}_T,\mathbf{z}_{T+1},\mathbf{z}_{T+2},\ldots,\mathbf{z}_{T+h};\boldsymbol{\theta}_i\right),\boldsymbol{\theta}_i\in\boldsymbol{\Theta}_i\right\}, \tag{7.22}$$

where $f_i(\cdot)$ is the joint probability density function of past and future values of $\mathbf{z}_t$. Conditional on each model, M_i, being true we shall assume that the true value of $\boldsymbol{\theta}_i$, which we denote by $\boldsymbol{\theta}_{i0}$, is fixed and remains constant across the estimation and the prediction periods and lies in the interior of $\boldsymbol{\Theta}_i$. We denote the maximum likelihood estimator of $\boldsymbol{\theta}_{i0}$ by $\widehat{\boldsymbol{\theta}}_{iT}$, and assume that it satisfies the usual regularity conditions so that

$$\sqrt{T}\left(\widehat{\boldsymbol{\theta}}_{iT}-\boldsymbol{\theta}_{i0}\right)|M_i \overset{a}{\sim} N\left(\mathbf{0},\mathbf{V}_{\theta_i}\right),$$

where $\overset{a}{\sim}$ stands for 'asymptotically distributed as', $\mathbf{V}_{\theta_i}$ is a positive definite matrix, and $T^{-1}\mathbf{V}_{\theta_i}$ is the asymptotic covariance matrix of $\widehat{\boldsymbol{\theta}}_{iT}$ conditional on M_i, with $\mathbf{V}_{\theta_i}$ being a positive definite matrix.[7] Under these assumptions, parameter uncertainty only arises when T is finite and $\widehat{\boldsymbol{\theta}}_{iT}\xrightarrow{a}\boldsymbol{\theta}_{i0}$ as $T\to\infty$.

[7] In the case of cointegrating VAR models, a more general version of this result is needed. This is because the cointegrating coefficients converge to their asymptotic distribution at a faster rate than the other parameters in the model. However, the general results of this section are not affected by this complication.

The case where $\boldsymbol{\theta}_{i0}$ could differ across the estimation and forecast periods poses new difficulties and can be resolved in a satisfactory manner if one is prepared to formalise how $\boldsymbol{\theta}_{i0}$ changes over time. See, for example, Pesaran, Timmermann and Pettenuzzo (2004).[8]

7.2.3 *Bayesian model averaging*

The object of interest is the probability density function of $\mathbf{Z}_{T+1,h} = (\mathbf{z}_{T+1}, \ldots, \mathbf{z}_{T+h})$ conditional on the available observations at the end of period T, $\mathbf{Z}_T = (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_T)$, denoted by $\Pr(\mathbf{Z}_{T+1,h} | \mathbf{Z}_T)$. For this purpose, models and their parameters serve as intermediate inputs in the process of characterisation and estimation of $\Pr(\mathbf{Z}_{T+1,h} | \mathbf{Z}_T)$. The Bayesian approach provides an elegant and logically coherent solution to this problem, with a full solution given by the so-called 'Bayesian model averaging' formula (*e.g.* Draper (1995) and Hoeting *et al.* (1999)):

$$\Pr(\mathbf{Z}_{T+1,h} | \mathbf{Z}_T) = \sum_{i=1}^{m} \Pr(M_i | \mathbf{Z}_T) \Pr(\mathbf{Z}_{T+1,h} | \mathbf{Z}_T, M_i), \tag{7.23}$$

where $\Pr(M_i | \mathbf{Z}_T)$ is the posterior probability of model M_i,

$$\Pr(M_i | \mathbf{Z}_T) = \frac{\Pr(M_i) \Pr(\mathbf{Z}_T | M_i)}{\sum_{j=1}^{m} \Pr(M_j) \Pr(\mathbf{Z}_T | M_j)}. \tag{7.24}$$

$\Pr(M_i)$ is the prior probability of model M_i, $\Pr(\mathbf{Z}_T | M_i)$ is the integrated likelihood,

$$\Pr(\mathbf{Z}_T | M_i) = \int_{\boldsymbol{\theta}_i} \Pr(\boldsymbol{\theta}_i | M_i) \Pr(\mathbf{Z}_T | M_i, \boldsymbol{\theta}_i) \, d\boldsymbol{\theta}_i. \tag{7.25}$$

$\Pr(\boldsymbol{\theta}_i | M_i)$ is the prior on $\boldsymbol{\theta}_i$ conditional on M_i, $\Pr(\mathbf{Z}_T | M_i, \boldsymbol{\theta}_i)$ is the likelihood function of model M_i, and $\Pr(\mathbf{Z}_{T+1,h} | \mathbf{Z}_T, M_i)$ is the posterior predictive density of model M_i defined by

$$\Pr(\mathbf{Z}_{T+1,h} | \mathbf{Z}_T, M_i) = \int_{\boldsymbol{\theta}_i} \Pr(\boldsymbol{\theta}_i | \mathbf{Z}_T, M_i) \Pr(\mathbf{Z}_{T+1,h} | \mathbf{Z}_T, M_i, \boldsymbol{\theta}_i) \, d\boldsymbol{\theta}_i, \tag{7.26}$$

in which $\Pr(\boldsymbol{\theta}_i | \mathbf{Z}_T, M_i)$ is the posterior probability of $\boldsymbol{\theta}_i$ given model M_i:

$$\Pr(\boldsymbol{\theta}_i | \mathbf{Z}_T, M_i) = \frac{\Pr(\boldsymbol{\theta}_i | M_i) \Pr(\mathbf{Z}_T | M_i, \boldsymbol{\theta}_i)}{\sum_{j=1}^{m} \Pr(M_j) \Pr(\mathbf{Z}_T | M_j)}. \tag{7.27}$$

[8] Pesaran, Timmermann and Pettenuzzo (2004) propose a Bayesian procedure that allows for the possibility of new breaks over the forecast horizon, taking account of the size and duration of past breaks (if any) by means of a hierarchical hidden Markov chain model. Predictions are formed by integrating over the hyper parameters from the meta distributions that characterise the stochastic break point process.

The Bayesian approach requires *a priori* specifications of $\Pr(M_i)$ and $\Pr(\boldsymbol{\theta}_i | M_i)$ for $i = 1, 2, \ldots, m$, and further assumes that one of the m models being considered is the DGP so that $\Pr(\mathbf{Z}_{T+1,h} | \mathbf{Z}_T)$ defined by (7.23) is proper.

7.2.4 *Pooling of forecasts*

The Bayesian model averaging formula also provides a simple 'optimal' solution to the problem of *pooling* of the point and volatility forecasts. In the context of the above set-up the point forecasts are given by $E(\mathbf{Z}_{T+1,h} | \mathbf{Z}_T, M_i)$, $i = 1, 2, \ldots, m$, and can be combined in a variety of ways as discussed extensively in the literature. For reviews of the forecast combination literature see Clemen (1989), Granger (1989), Diebold and Lopez (1996) and Newbold and Harvey (2002).

In general the combined or pooled point forecasts can be written as

$$E_w(\mathbf{Z}_{T+1,h} | \mathbf{Z}_T) = \sum_{i=1}^{m} w_{iT} E(\mathbf{Z}_{T+1,h} | \mathbf{Z}_T, M_i),$$

where w_{iT}, $i = 1, 2, \ldots, m$ are the weights attached to the individual point forecasts. The main issues are: Should the weights be non-negative and add up to unity? Should they be based on past relative performance of the alternative models and hence be time varying? How should the relative performance of the various models be measured, namely should we be using in-sample criteria of fit and parsimony or out-of-sample realised performance?

In situations where the models under consideration are thought to be exhaustive (and hence the true data generating process is thought to lie in the set of models under consideration), the Bayesian approach can be used to provide a coherent answer to these questions. Under Bayesian model averaging (BMA) the weights, w_{iT}, are set to the posterior probability of model M_i and hence are non-negative and satisfy the additivity condition, $\sum_{i=1}^{m} w_{iT} = 1$. Using the Bayesain weights the combined point forecast is given by

$$E(\mathbf{Z}_{T+1,h} | \mathbf{Z}_T) = \sum_{i=1}^{m} \Pr(M_i | \mathbf{Z}_T) E(\mathbf{Z}_{T+1,h} | \mathbf{Z}_T, M_i).$$

In practice the derivation of the model-specific probability weights pose a number of conceptual and computations issues that will be briefly addressed below.

In cases where the models under consideration are not exhaustive and the underlying data generation process could be time varying, non-Bayesian weights might be more appropriate. Many alternatives have been proposed in the literature. Amongst these the simple average rule where equal weights are attached to the alternative forecasts tends to perform surprisingly well, as noted originally by Clemen (1989).[9] Recently, Granger and Jeon (2004) have proposed a modification of this procedure where the average rule is applied to a subset of best performing models. This modification, referred to as 'thick' modelling, is particularly relevant when there are many forecasts under consideration.

Pooling of forecast variances can also be considered. Under BMA we have (*e.g.* Draper, 1995)

$$\begin{aligned}\text{Var}\left(\mathbf{Z}_{T+1,h}|\mathbf{Z}_T\right) = &\sum_{i=1}^{m}\Pr\left(M_i|\mathbf{Z}_T\right)\text{Var}\left(\mathbf{Z}_{T+1,h}|\mathbf{Z}_T,M_i\right)\\ &+\sum_{i=1}^{m}\Pr\left(M_i|\mathbf{Z}_T\right)\left[E\left(\mathbf{Z}_{T+1,h}|\mathbf{Z}_T,M_i\right)\right.\\ &\left.-E\left(\mathbf{Z}_{T+1,h}|\mathbf{Z}_T\right)\right]^2,\end{aligned}$$

Once again, more generally, we could have

$$\begin{aligned}\text{Var}_w\left(\mathbf{Z}_{T+1,h}|\mathbf{Z}_T\right) = &\sum_{i=1}^{m}w_{iT}\text{Var}\left(\mathbf{Z}_{T+1,h}|\mathbf{Z}_T,M_i\right)\\ &+\sum_{i=1}^{m}w_{iT}\left[E\left(\mathbf{Z}_{T+1,h}|\mathbf{Z}_T,M_i\right)-E\left(\mathbf{Z}_{T+1,h}|\mathbf{Z}_T\right)\right]^2,\end{aligned}$$

where the weights w_{iT} could be obtained using Bayesian or non-Bayesian procedures. The first term in the above expression accounts for within model variability and the second term for between model variability. Clearly, a procedure that only combines the forecast variances will not be correct unless all models have the same point forecasts. Pooling of predictive densities clearly does not imply using averages of the moments of the underlying distributions except for the first moments.

There is no doubt that the Bayesian model averaging provides an attractive solution to the problem of accounting for model uncertainty. But its strict application can be problematic particularly in the case of high-dimensional models such as the vector error correction model of the UK economy considered in our empirical work. The major difficulties lie in

[9] Recent Monte Carlo evidence that attempts to explain this empirical finding is provided by Hendry and Clements (2004) and Smith and Wallis (2005).

the choice of the space of models to be considered, the model priors $\Pr(M_i)$, and the specification of meaningful priors for the unknown parameters, $\Pr(\boldsymbol{\theta}_i|M_i)$. The computational issues, while still considerable, are partly overcome by Monte Carlo integration techniques. For an excellent overview of the issues involved in the application of BMA approach to forecasting, see Hoeting *et al.* (1999). See also Fernandez *et al.* (2001a,b) and Pesaran and Zaffaroni (2005) for specific applications.

Putting the problem of model specification to one side, the two important components of BMA formula are the posterior probability of the models, $\Pr(M_i|\mathbf{Z}_T)$, and the posterior density functions of the parameters, $\Pr(\boldsymbol{\theta}_i|\mathbf{Z}_T,M_i)$, for $i=1,\ldots,m$. In what follows we therefore consider different approximations of $\Pr(M_i|\mathbf{Z}_T)$ and $\Pr(\boldsymbol{\theta}_i|\mathbf{Z}_T,M_i)$, assuming that T is sufficiently large that the sample observations dominate the choice of the priors; in essence adopting a classical stance within an otherwise Bayesian framework. See also Garratt *et al.* (2003b).

7.3 Computation of probability forecasts: Some practical issues

Suppose the *joint event* of interest is defined by $\boldsymbol{\varphi}\left(\mathbf{Z}_{T+1,h}\right) < \mathbf{a}$, where $\boldsymbol{\varphi}(\cdot)$ and $\mathbf{a}$ are the $L\times 1$ vectors $\boldsymbol{\varphi}(\cdot)=(\varphi_1(\cdot),\varphi_2(\cdot),\ldots,\varphi_L(\cdot))'$, $\mathbf{a}=(a_1,a_2,\ldots,a_L)'$, $\varphi_j(\mathbf{Z}_{T+1,h})$ is a scalar function of the variables over the forecast horizon $T+1,\ldots,T+h$, and a_j is the 'threshold' value associated with $\varphi_j(\cdot)$. To simplify the exposition, we denote this joint event by $\mathfrak{A}_\varphi$. The (conditional) probability forecast associated with this event assuming that model M_i holds is given by

$$\pi_i\left(\mathbf{a},h;\boldsymbol{\varphi}(\cdot),\boldsymbol{\theta}_i\right)=\Pr\left[\boldsymbol{\varphi}\left(\mathbf{Z}_{T+1,h}\right)<\mathbf{a}\,|\mathbf{Z}_T,M_i,\boldsymbol{\theta}_i\right]. \tag{7.28}$$

In practice, we might be interested in computing probability forecasts for a number of alternative threshold values over the range $a_j\in[a_{\min},a_{\max}]$.

With future uncertainty only

If the model is known to be M_i defined by (7.22) but the value of $\boldsymbol{\theta}_i$ is not known, a *point estimate* of $\pi_i\left(\mathbf{a},h;\boldsymbol{\varphi}(\cdot),\boldsymbol{\theta}_i\right)$ can be obtained by

$$\pi_i\left(\mathbf{a},h;\boldsymbol{\varphi}(\cdot),\widehat{\boldsymbol{\theta}}_{iT}\right)=\int_{\mathfrak{A}_\varphi}f_i\left(\mathbf{Z}_{T+1,h}|\mathbf{Z}_T,M_i,\widehat{\boldsymbol{\theta}}_{iT}\right)d\mathbf{Z}_{T+1,h}. \tag{7.29}$$

This probability distribution function only takes account of future uncertainties that arise from the model's stochastic structure, as it is computed

for a given density function, M_i, and a given value of $\boldsymbol{\theta}_i$, namely $\widehat{\boldsymbol{\theta}}_{iT}$. It is also known as the 'profile predictive likelihood'. See, for example, Bjørnstad (1990).

With future and parameter uncertainty

To allow for parameter uncertainty, we assume that conditional on $\mathbf{Z}_T$, the probability distribution function of $\boldsymbol{\theta}_i$ is given by $g(\boldsymbol{\theta}_i|\mathbf{Z}_T, M_i)$. Then,

$$\tilde{\pi}_i(\mathbf{a}, h; \boldsymbol{\varphi}(\cdot)) = \int_{\boldsymbol{\theta}_i \in \boldsymbol{\Theta}_i} \pi_i(\mathbf{a}, h; \boldsymbol{\varphi}(\cdot), \boldsymbol{\theta}_i)\, g(\boldsymbol{\theta}_i|\mathbf{Z}_T, M_i)\, d\boldsymbol{\theta}_i, \tag{7.30}$$

or equivalently,

$$\tilde{\pi}_i(\mathbf{a}, h; \boldsymbol{\varphi}(\cdot)) = \int_{\boldsymbol{\theta}_i \in \boldsymbol{\Theta}_i} \int_{\mathfrak{A}_\varphi} f_i(\mathbf{Z}_{T+1,h}|\mathbf{Z}_T, M_i, \boldsymbol{\theta}_i)\, g(\boldsymbol{\theta}_i|\mathbf{Z}_T, M_i)\, d\mathbf{Z}_{T+1,h} d\boldsymbol{\theta}_i. \tag{7.31}$$

Computation of (7.31) requires the knowledge of $g(\boldsymbol{\theta}_i|\mathbf{Z}_T, M_i)$. In the absence of model priors $\Pr(M_i)$ or priors for the unknown parameters, $\Pr(\boldsymbol{\theta}_i|M_i)$, we might assume

$$\boldsymbol{\theta}_i|\mathbf{Z}_T, M_i \overset{a}{\sim} N\left(\widehat{\boldsymbol{\theta}}_{iT}, T^{-1}\widehat{\mathbf{V}}_{\boldsymbol{\theta}_i}\right). \tag{7.32}$$

In this case, the point estimate of the probability forecast, $\pi_i(\mathbf{a}, h; \boldsymbol{\varphi}(\cdot), \widehat{\boldsymbol{\theta}}_{iT})$, and the alternative estimate, $\tilde{\pi}_i(\mathbf{a}, h; \boldsymbol{\varphi}(\cdot))$, that allows for parameter uncertainty are asymptotically equivalent as $T \to \infty$. The latter is the 'bootstrap predictive density' described in Harris (1989), who demonstrates that it performs well in a number of important cases. Also, both of these estimates under M_i tend to $\pi_i(\mathbf{a}, h; \boldsymbol{\varphi}(\cdot), \boldsymbol{\theta}_{i0})$, which is the profile predictive likelihood evaluated at the true value $\boldsymbol{\theta}_{i0}$. In practice, computations of $\pi_i(\mathbf{a}, h; \boldsymbol{\varphi}(\cdot), \widehat{\boldsymbol{\theta}}_{iT})$ and $\tilde{\pi}_i(\mathbf{a}, h; \boldsymbol{\varphi}(\cdot))$ are typically carried out by stochastic simulations (see Section 7.3.2 below), and the two estimates will differ by terms that are $O(h/T)$ and will be very close when h is small and T large.[10]

With future and model uncertainty

The probability estimates that allow for model uncertainty can now be obtained using the Bayesian averaging procedure. Abstracting from parameter uncertainty we have

$$\pi(\mathbf{a}, h; \boldsymbol{\varphi}(\cdot), \widehat{\boldsymbol{\theta}}_T) = \sum_{i=1}^{m} w_{iT} \pi_i(\mathbf{a}, h; \boldsymbol{\varphi}(\cdot), \widehat{\boldsymbol{\theta}}_{iT}), \tag{7.33}$$

[10] See Bjørnstad (1990, 1998) for reviews of the literature on predictive likelihood analysis.

where $\widehat{\boldsymbol{\theta}}_T = \left(\widehat{\boldsymbol{\theta}}'_{1T}, \ldots, \widehat{\boldsymbol{\theta}}'_{mT}\right)'$, and the weights, $w_{iT} \geq 0$ can be derived by approximating the posterior probability of model M_i by[11]

$$\ln \Pr(M_i|\mathbf{Z}_T) = LL_{iT} - \left(\frac{k_i}{2}\right) \ln(T) + O(1), \tag{7.34}$$

where LL_{iT} is the maximised value of the log-likelihood function for model M_i. This is the familiar Schwarz (1978) Bayesian information criterion for model selection. The use of this approximation leads to the following choice for w_{iT}:

$$w_{iT} = \frac{\exp(\Delta_{iT})}{\sum_{j=1}^{m} \exp(\Delta_{jT})}, \tag{7.35}$$

where $\Delta_{iT} = SBC_{iT} - \max_j(SBC_{jT})$ and $SBC_{iT} = LL_{iT} - \left(\frac{k_i}{2}\right)\ln(T)$. Alternatively, following Burnham and Anderson (1998), one could use Akaike weights defined by $\Delta_{iT} = AIC_{iT} - \max_j(AIC_{jT})$, $AIC_{iT} = LL_{iT} - k_i$. While the Schwarz weights are asymptotically optimal if the DGP lies in the set of models under consideration, the Akaike weights are likely to perform better when the true model does not lie in the set of models under consideration, that are viewed as approximations to a complex and (possibly) unknown DGP.

With future, parameter and model uncertainty

When parameter uncertainty is also taken into account, we have

$$\tilde{\pi}(\mathbf{a}, h; \boldsymbol{\varphi}(\cdot)) = \sum_{i=1}^{m} w_{iT} \tilde{\pi}_i(\mathbf{a}, h; \boldsymbol{\varphi}(\cdot)), \tag{7.36}$$

where $\tilde{\pi}_i(\mathbf{a}, h; \boldsymbol{\varphi}(\cdot))$ is the bootstrap predictive density defined by (7.31) that makes use of the normal approximation given by (7.32). Again, in practice, computations of $\pi_i(\mathbf{a}, h; \boldsymbol{\varphi}(\cdot))$and $\tilde{\pi}_i(\mathbf{a}, h; \boldsymbol{\varphi}(\cdot))$ are typically carried out by stochastic simulations (see Section 7.3.2 below).

7.3.1 *Computation of probability forecasts using analytic methods*

In this subsection, we outline the computational difficulties that typically will be encountered in the calculation of probability forecasts. We illustrate this using the simpler case in which it is assumed that the parameters of

[11] See also Draper (1995) for approximate posterior probability forecasts, conditional on the model M_i being true.

the model are known, so that only stochastic uncertainty is considered, and the probability forecast is evaluated according to (7.29).

In this case there is generally no conceptual difficulty in evaluating the probability of an event taking place using (7.29) for known $\widehat{\boldsymbol{\theta}}$. However, the computation can become complicated because of the form of the functions $\boldsymbol{\varphi}$ or due to the difficulties arising from the selection of appropriate limits of integration for the expression, or because of the complexity of the event to be forecast even if the functions $\boldsymbol{\varphi}$ are reasonably simple.

Consider, for example, the linear case in which the joint event of interest $\boldsymbol{\varphi}\left(\mathbf{z}_{T+1}, \ldots, \mathbf{z}_{T+h}\right)$ can be expressed by

$$\boldsymbol{\varphi}\left(\mathbf{z}_{T+1}, \ldots, \mathbf{z}_{T+h}\right) = \boldsymbol{\varphi}\left(\hat{\mathbf{z}}_{T+1}, \ldots, \hat{\mathbf{z}}_{T+h}\right) + \boldsymbol{\nu}_{T+h}, \tag{7.37}$$

where $\boldsymbol{\varphi}\left(\hat{\mathbf{z}}_{T+1}, \ldots, \hat{\mathbf{z}}_{T+h}\right)$ represents a (consistent) estimate of $\boldsymbol{\varphi}\left(\mathbf{z}_{T+1}, \ldots, \mathbf{z}_{T+h}\right)$, based on estimated model parameter values $\widehat{\boldsymbol{\theta}}_T$, and the stochastic uncertainty surrounding the estimate is captured by an $L \times 1$ vector of the corresponding forecast errors, $\boldsymbol{\nu}_{T+h}$, which is assumed to be normally distributed with zero means and an $L \times L$ positive covariance matrix, $\boldsymbol{\Sigma}_{\nu}$. In this case, the probability forecast defined by (7.29) is given by

$$\begin{aligned}
&\widehat{\pi}\left(\mathbf{a}, h; \boldsymbol{\varphi}(.), \widehat{\boldsymbol{\theta}}_T\right) \\
&\quad = \Pr\left(\boldsymbol{\varphi}\left(\mathbf{z}_{T+1}, \ldots, \mathbf{z}_{T+h}\right) < \mathbf{a}\right) = \Pr\left(\boldsymbol{\nu}_{T+h} < \mathbf{a} - \boldsymbol{\varphi}\left(\hat{\mathbf{z}}_{T+1}, \ldots, \hat{\mathbf{z}}_{T+h}\right)\right) \\
&\quad = \int_{-\infty}^{a_L^*} \cdots \int_{-\infty}^{a_1^*} \left[(2\pi)^{-\frac{1}{2}L} |\boldsymbol{\Sigma}_{\nu}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} \boldsymbol{\nu}_{T+h}' \boldsymbol{\Sigma}_{\nu}^{-1} \boldsymbol{\nu}_{T+h}\right)\right] d\nu_{T+h,1} \cdots d\nu_{T+h,L},
\end{aligned}$$

where

$$a_j^* = a_j - \varphi_j\left(\hat{\mathbf{z}}_{T+1}, \ldots, \hat{\mathbf{z}}_{T+h}\right), \quad j = 1, 2, \ldots, L.$$

Even in this relatively simple case, the evaluation of the probability involves L multiple integrals and, unless L is small (1 or 2), its computation would be quite demanding.

7.3.2 *Computation of probability forecasts based on VAR models by stochastic simulation*

In this subsection, we describe the steps involved in the calculation of probability forecasts based on a vector error correction model described in Section 6.3, using stochastic simulation techniques. Consider the underlying vector error correction model, (6.86), which can be

rewritten as

$$\mathbf{z}_t = \sum_{i=1}^{p} \boldsymbol{\Phi}_i \mathbf{z}_{t-i} + \mathbf{a}_0 + \mathbf{a}_1 t + \mathbf{H}\boldsymbol{\zeta}_t, \; t = 1, 2, \ldots, T, \tag{7.38}$$

where $\boldsymbol{\Phi}_1 = \mathbf{I}_m - \boldsymbol{\alpha}\boldsymbol{\beta}' + \boldsymbol{\Gamma}_1$, $\boldsymbol{\Phi}_i = \boldsymbol{\Gamma}_i - \boldsymbol{\Gamma}_{i-1}$, $i = 2, \ldots, p-1$, $\boldsymbol{\Phi}_p = -\boldsymbol{\Gamma}_{p-1}$, and $\boldsymbol{\zeta}_t$ is assumed to be a serially uncorrelated *i.i.d.* vector of shocks with zero means and a positive definite covariance matrix, $\boldsymbol{\Sigma}_{\zeta\zeta}$ given by (6.88). In what follows, we consider the calculation of probability forecasts first for given values of the parameters, and then taking into account parameter uncertainty.

FORECASTS IN THE ABSENCE OF PARAMETER UNCERTAINTY

Suppose that the ML estimators of $\boldsymbol{\Phi}_i$, $i = 1, \ldots, p$, $\mathbf{a}_0$, $\mathbf{a}_1$, $\mathbf{H}$ and $\boldsymbol{\Sigma}_{\zeta\zeta}$ are given and denoted by $\hat{\boldsymbol{\Phi}}_i$, $i = 1, \ldots, p$, $\hat{\mathbf{a}}_0$, $\hat{\mathbf{a}}_1$, $\hat{\mathbf{H}}$ and $\hat{\boldsymbol{\Sigma}}_{\zeta\zeta}$, respectively. Then, the point estimates of the h-step ahead forecasts of $\mathbf{z}_{T+h}$ conditional on $\mathfrak{I}_T$, denoted by $\hat{\mathbf{z}}_{T+h}$, can be obtained recursively as

$$\hat{\mathbf{z}}_{T+h} = \sum_{i=1}^{p} \hat{\boldsymbol{\Phi}}_i \hat{\mathbf{z}}_{T+h-i} + \hat{\mathbf{a}}_0 + \hat{\mathbf{a}}_1 (t+h), \quad h = 1, 2, \ldots, \tag{7.39}$$

where the initial values, $\mathbf{z}_T, \mathbf{z}_{T-1}, \ldots, \mathbf{z}_{T-p+1}$, are given. To obtain probability forecasts by stochastic simulation, we simulate the values of $\mathbf{z}_{T+h}$ by

$$\begin{aligned}
\mathbf{z}_{T+h}^{(r)} &= \sum_{i=1}^{p} \hat{\boldsymbol{\Phi}}_i \mathbf{z}_{T+h-i}^{(r)} + \hat{\mathbf{a}}_0 + \hat{\mathbf{a}}_1 (t+h) + \hat{\mathbf{H}}\boldsymbol{\zeta}_{T+h}^{(r)}, \\
&\quad h = 1, 2, \ldots; \; r = 1, 2, \ldots, R,
\end{aligned} \tag{7.40}$$

where superscript '(r)' refers to the rth replication of the simulation algorithm, and $\mathbf{z}_T^{(r)} = \mathbf{z}_T$, $\mathbf{z}_{T-1}^{(r)} = \mathbf{z}_{T-1}, \ldots, \mathbf{z}_{T-p+1}^{(r)} = \mathbf{z}_{T-p+1}$ for all r. The $\boldsymbol{\zeta}_{T+h}^{(r)}$'s can be drawn either by parametric or non-parametric methods as described in Section 7.3.3 below. The probability that $\boldsymbol{\varphi}_\ell\left(\mathbf{z}_{T+1}^{(r)}, \ldots, \mathbf{z}_{T+h}^{(r)}\right) < \mathbf{a}_\ell$, is computed as

$$\pi_R\left(\mathbf{a}_\ell, h; \boldsymbol{\varphi}_\ell(\cdot), \widehat{\boldsymbol{\theta}}\right) = \frac{1}{R} \sum_{r=1}^{R} I\left(\mathbf{a}_\ell - \boldsymbol{\varphi}_\ell\left(\mathbf{z}_{T+1}^{(r)}, \ldots, \mathbf{z}_{T+h}^{(r)}\right)\right),$$

where $\widehat{\boldsymbol{\theta}}$ is a vector containing estimates of all the parameters, and $I(A)$ is an indicator function which takes the value of unity if $A > 0$, and zero otherwise. To simplify the notation we denote $\pi_R\left(\mathbf{a}_\ell, h; \boldsymbol{\varphi}_\ell(\cdot), \widehat{\boldsymbol{\theta}}\right)$ by $\pi_R\left(\mathbf{a}_\ell\right)$. The predictive probability distribution function is now given by $\pi_R\left(\mathbf{a}_\ell\right)$ as the threshold values, $\mathbf{a}_\ell$, are varied over the relevant regions.

FORECASTS IN THE PRESENCE OF PARAMETER UNCERTAINTY

To allow for parameter uncertainty, we use the bootstrap procedure and first simulate S *(in-sample)* values of $\mathbf{z}_t$, $t = 1, 2, \ldots, T$, denoted by $\mathbf{z}_t^{(s)}$, $s = 1, \ldots, S$, where

$$\mathbf{z}_t^{(s)} = \sum_{i=1}^{p} \hat{\mathbf{\Phi}}_i \mathbf{z}_{t-i}^{(s)} + \hat{\mathbf{a}}_0 + \hat{\mathbf{a}}_1 t + \mathbf{H}\zeta_t^{(s)}, \quad t = 1, 2, \ldots, T, \tag{7.41}$$

realisations are used for the initial values, $\mathbf{z}_{-1}, \ldots, \mathbf{z}_{-p}$, and $\zeta_t^{(s)}$'s can be drawn either by parametric or non-parametric methods (see Section 7.3.3 below). Having obtained the S set of simulated in-sample values, $\left(\mathbf{z}_1^{(s)}, \mathbf{z}_2^{(s)}, \ldots, \mathbf{z}_T^{(s)}\right)$, the VAR($p$) model (7.38) is estimated S times to obtain the ML estimates, $\hat{\mathbf{\Phi}}_i^{(s)}$, $\hat{\mathbf{a}}_0^{(s)}$, $\hat{\mathbf{a}}_1^{(s)}$, $\hat{\mathbf{H}}^{(s)}$ and $\hat{\mathbf{\Sigma}}_{\zeta\zeta}^{(s)}$, for $i = 1, 2, \ldots, p$, and $s = 1, 2, \ldots, S$.

For each of these bootstrap replications, R replications of the h-step ahead point forecasts are computed as

$$\mathbf{z}_{T+h}^{(r,s)} = \sum_{i=1}^{p} \hat{\mathbf{\Phi}}_i^{(s)} \mathbf{z}_{T+h-i}^{(r,s)} + \hat{\mathbf{a}}_0^{(s)} + \hat{\mathbf{a}}_1^{(s)}(t+h) + \hat{\mathbf{H}}^{(s)} \zeta_{T+h}^{(r,s)}, \tag{7.42}$$

for $h = 1, 2, \ldots H$; $r = 1, 2, \ldots, R$ and $s = 1, 2, \ldots, S$, and the predictive distribution function is then computed as

$$\pi_{R,S}(\mathbf{a}_\ell) = \frac{1}{SR} \sum_{r=1}^{R} \sum_{s=1}^{S} I\left[\mathbf{a}_\ell - \boldsymbol{\varphi}_\ell\left(\mathbf{z}_{T+1}^{(r,s)}, \ldots, \mathbf{z}_{T+h}^{(r,s)}\right)\right].$$

Bootstrapping cointegrating models can be done either for a fixed number of cointegrating relations (obtained from estimates based on the actual time series), or the cointegrating relations could be re-estimated for each bootstrap replication. In our empirical applications we follow the former, but allow for the uncertainty surrounding the number of cointegrating vectors by means of model averaging techniques; namely different choices of the number of cointegrating relations are regarded as different models.

7.3.3 *Generating simulated errors*

We now provide more details on the mechanism by which shocks are generated in stochastic simulation methods described above. There are two basic ways that the in-sample and future errors, $\zeta_t^{(s)}$ and $\zeta_{T+h}^{(r,s)}$ respectively, can be simulated so that the contemporaneous correlations that

exist across the errors in the different equations of the VAR model are taken into account and maintained. The first is a *parametric* method where the errors are drawn from an assumed probability distribution function. Alternatively, one could employ a *non-parametric* procedure. The latter is slightly more complicated and is based on re-sampling techniques in which the simulated errors are obtained by a random draw from the in-sample estimated residuals (*e.g.* Hall, 1992).

Parametric approach

Under this approach the errors are drawn for example, from a multivariate distribution with zero means and the covariance matrix, $\hat{\mathbf{\Sigma}}_{\zeta\zeta}^{(s)}$. To obtain the simulated errors for m variables over h periods we first generate mh draws from an assumed *i.i.d.* distribution which we denote by $\epsilon_{T+i}^{(r,s)}$, $i = 1, 2, \ldots, h$. These are then used to obtain $\left\{\zeta_{T+i}^{(r,s)}, i = 1, 2, \ldots h\right\}$ computed as $\zeta_{T+h}^{(r,s)} = \hat{\mathbf{P}}^{(s)} \epsilon_{T+h}^{(r,s)}$ for $r = 1, 2, \ldots, R$ and $s = 1, 2, \ldots, S$, where $\hat{\mathbf{P}}^{(s)}$ is the lower triangular Choleski factor of $\hat{\mathbf{\Sigma}}_{\zeta\zeta}^{(s)}$ such that $\hat{\mathbf{\Sigma}}_{\zeta\zeta}^{(s)} = \hat{\mathbf{P}}^{(s)} \hat{\mathbf{P}}^{(s)\prime}$, and $\hat{\mathbf{\Sigma}}_{\zeta\zeta}^{(s)}$ is the estimate of $\mathbf{\Sigma}_{\zeta\zeta}$ in the sth replication of the bootstrap procedure set out above. In the absence of parameter uncertainty, we obtain $\zeta_{T+h}^{(r)} = \hat{\mathbf{P}} \epsilon_{T+h}^{(r)}$ with $\hat{\mathbf{P}}$ being the lower triangular Choleski factor of $\hat{\mathbf{\Sigma}}_{\zeta\zeta}$. In our applications, reported in Chapter 11, for each r and s, we generate $\epsilon_{T+i}^{(r,s)}$ as *i.i.d.*$N(0, \mathbf{I}_m)$, although other parametric distributions such as the multivariate Student t-distribution can also be used.

Non-parametric approaches

The most obvious non-parametric approach to generating the simulated errors, $\zeta_{T+h}^{(r,s)}$, which we denote 'Method 1', is simply to take h random draws with replacements from the in-sample residual vectors $\left\{\hat{\zeta}_1^{(s)}, \ldots, \hat{\zeta}_T^{(s)}\right\}$. The simulated errors thus obtained clearly have the same distribution and covariance structure as that observed in the original sample. However, this procedure is subject to the criticism that it could introduce serial dependence at longer forecast horizons since the pseudo-random draws are made from the same set of relatively small T vector of residuals.

An alternative non-parametric method for generating simulated errors, 'Method 2', makes use of the Choleski decomposition of the estimated covariance employed in the parametric approach. For a given choice of $\hat{\mathbf{P}}^{(s)}$ a set of mT transformed error terms $\left\{\hat{\epsilon}_1^{(s)}, \ldots, \hat{\epsilon}_T^{(s)}\right\}$ are computed such that $\hat{\epsilon}_t^{(s)} = \hat{\mathbf{P}}^{(s)-1} \hat{\zeta}_t^{(s)}$, $t = 1, 2, \ldots, T$. The mT individual error terms

are uncorrelated with each other, but retain the distributional information contained in the original observed errors. A set of *mh* simulated errors are then obtained by drawing with replacement from these transformed residuals, denoted by $\left\{\epsilon_{T+1}^{(r,s)}, \ldots, \epsilon_{T+h}^{(r,s)}\right\}$. These are then used to obtain $\left\{\zeta_{T+1}^{(r,s)}, \ldots, \zeta_{T+h}^{(r,s)}\right\}$, using the transformations $\zeta_{T+h}^{(r,s)} = \hat{\mathbf{P}}^{(s)}\epsilon_{T+h}^{(r,s)}$ for $r = 1, 2, \ldots, R$ and $s = 1, 2, \ldots, S$. Given that the $\hat{\mathbf{P}}^{(s)}$ matrix is used to generate the simulated errors, it is clear that $\zeta_{T+h}^{(r,s)}$ again has the same covariance structure as the original estimated errors. And being based on errors drawn at random from the transformed residuals, these simulated errors will also display the same distributional features. Further, given that the re-sampling occurs from the mT transformed error terms, Method 2 also has the advantage over Method 1 that the serial dependence introduced through sampling with replacement is likely to be less problematic.

Choice of approach

The two non-parametric approaches described above have the advantage over the parametric approach that they make no distributional assumptions on the error terms, and are better able to capture the uncertainties arising from (possibly rare) extreme observations. However, they suffer from the fact that they require random sampling *with replacement*. Replacement is essential as otherwise the draws at longer forecast horizons are effectively 'truncated' and unrepresentative. On the other hand, for a given sample size, it is clear that re-sampling from the observed errors with replacement inevitably introduces serial dependence in the simulated forecast errors at longer horizons as the same residuals are drawn repeatedly. When generating simulated errors over forecast horizons, therefore, this provides an argument for the use of non-parametric methods over shorter forecast horizons, but suggests that a greater reliance might be placed on the parametric approach for the generation of probability forecasts at longer time horizons.

7.4 Estimation and forecasting with conditional models

The density function $f_i(\cdot)$ given in (7.22) can be decomposed in two ways. First, a sequential conditioning decomposition can be employed to write $f_i(\cdot)$ as the product of the conditional distributions on successive

observations on the $\mathbf{z}_t$,

$$f_i(\mathbf{Z}_t; \mathbf{z}_0, \boldsymbol{\theta}) = \prod_{s=1}^{t} f_i(\mathbf{z}_s \mid \mathbf{Z}_{s-1}; \mathbf{z}_0, \boldsymbol{\theta}_i),$$

where $\mathbf{Z}_s = (\mathbf{z}_0, \mathbf{z}_1, \ldots, \mathbf{z}_s)$ for given initial values $\mathbf{z}_0$. Second, since we frequently wish to distinguish between variables which are endogenous, denoted by $\mathbf{y}_t$, and those which are exogenous, denoted by $\mathbf{x}_t$, we can write $\mathbf{z}_t = (\mathbf{y}_t', \mathbf{x}_t')'$ and use the factorisation:

$$f_i(\mathbf{z}_t \mid \mathbf{Z}_{t-1}; \mathbf{z}_0, \boldsymbol{\theta}) = f_{iy}(\mathbf{y}_t \mid \mathbf{x}_t, \mathbf{Z}_{t-1}; \mathbf{z}_0, \boldsymbol{\theta}_{iy}) \times f_{ix}(\mathbf{x}_t \mid \mathbf{Z}_{t-1}; \mathbf{z}_0, \boldsymbol{\theta}_{ix}), \quad (7.43)$$

where $f_{iy}(\mathbf{y}_t \mid \mathbf{x}_t, \mathbf{Z}_{t-1}; \mathbf{z}_0, \boldsymbol{\theta}_{iy})$ is the conditional distribution of $\mathbf{y}_t$ given $\mathbf{x}_t$ under model M_i and the information available at time $t-1$, $\mathbf{Z}_{t-1}$, and $f_{ix}(\mathbf{x}_t \mid \mathbf{Z}_{t-1}; \mathbf{z}_0, \boldsymbol{\theta}_{ix})$ is the marginal density of $\mathbf{x}_t$ conditional on $\mathbf{Z}_{t-1}$. Note that the unknown parameters $\boldsymbol{\theta}_i$ are decomposed into the parameters of interest, $\boldsymbol{\theta}_{iy}$, and the parameters of the marginal density of the exogenous variables, $\boldsymbol{\theta}_{ix}$. In the case where $\mathbf{x}_t$ is strictly exogenous, knowledge of the marginal distribution of $\mathbf{x}_t$ does not help with the estimation of $\boldsymbol{\theta}_{iy}$, and estimation of these parameters can therefore be based entirely on the conditional distribution, $f_{iy}(\mathbf{y}_t \mid \mathbf{x}_t, \mathbf{Z}_{t-1}; \mathbf{z}_0, \boldsymbol{\theta}_{iy})$.

Despite this, parameter uncertainty relating to $\boldsymbol{\theta}_{ix}$ can continue to be relevant for probability forecasts of the endogenous variables, $\mathbf{y}_t$, and forecast uncertainty surrounding the endogenous variables is affected by the way the uncertainty associated with the future path of the exogenous variables is resolved. In practice, the future values of $\mathbf{x}_t$ are often treated as known and fixed at pre-specified values. The resultant forecasts for $\mathbf{y}_t$ are then referred to as *scenario (or conditional) forecasts*, with each scenario representing a different set of assumed future values of the exogenous variables. This approach underestimates the degree of forecast uncertainties. A more plausible approach would be to treat $\mathbf{x}_t$ as strongly or weakly exogenous (as appropriate) at the estimation stage, but to allow for the forecast uncertainties of the endogenous and the exogenous variables jointly. The exogeneity assumption will simplify the estimation process but does not eliminate the need for a joint treatment of future and model uncertainties associated with the exogenous variables and the endogenous variables.