14 Forecasting with VAR Models

14.1 Forecasting with Known Parameters

The discussion of forecasting with VAR models proceeds in two steps. First, we assume that the parameters of the model are known. Although this assumption is unrealistic, it will nevertheless allow us to introduce and analyze important concepts and ideas. In a second step, we then investigate how the results established in the first step have to be amended if the parameters are estimated. The analysis will focus on stationary and causal VAR(1) processes. Processes of higher order can be accommodated by rewriting them in companion form. Thus we have:

$$X_t = \Phi X_{t-1} + Z_t, \qquad Z_t \sim \text{WN}(0, \Sigma),$$

$$X_t = Z_t + \Psi_1 Z_{t-1} + \Psi_2 Z_{t-2} + \ldots = \sum_{j=0}^{\infty} \Psi_j Z_{t-j},$$

where $\Psi_j = \Phi^j$. Consider then the following forecasting problem: Given observations $\{X_T, X_{T-1}, \ldots, X_1\}$, find a linear function, called predictor or forecast function, $\mathbb{P}_T X_{T+h}$, $h \geq 1$, which minimizes the expected quadratic forecast error

$$\mathbb{E}\,(X_{T+h} - \mathbb{P}_T X_{T+h})'\,(X_{T+h} - \mathbb{P}_T X_{T+h}) \\ = \mathbb{E}\,\text{tr}(X_{T+h} - \mathbb{P}_T X_{T+h})(X_{T+h} - \mathbb{P}_T X_{T+h})'.$$

Thereby "tr" denotes the trace operator which takes the sum of the diagonal elements of a matrix. As we rely on linear forecasting functions, $\mathbb{P}_T X_{T+h}$ can be expressed as

$$\mathbb{P}_T X_{T+h} = A_1 X_T + A_2 X_{T-1} + \ldots + A_T X_1 \tag{14.1}$$

K. Neusser, *Time Series Econometrics*, Springer Texts in Business and Economics,
DOI 10.1007/978-3-319-32862-1_14

with matrices $A_1, A_2, \ldots, A_T$ still to be determined. In order to simplify the exposition, we already accounted for the fact that the mean of $\{X_t\}$ is zero.[1] A justification for focusing on linear least-squares forecasts is given in Chap. 3. The first order conditions for the least-squares minimization problem are given by the normal equations:

$$\begin{aligned}\mathbb{E}\left(X_{T+h} - \mathbb{P}_T X_{T+h}\right) X_s' &= \mathbb{E}\left(X_{T+h} - A_1 X_T - \ldots - A_T X_1\right) X_s' \\ &= \mathbb{E} X_{T+h} X_s' - A_1 \mathbb{E} X_T X_s' - \ldots - A_T \mathbb{E} X_1 X_s' = 0, \quad 1 \leq s \leq T.\end{aligned}$$

These equations state that the forecast error $(X_{T+h} - \mathbb{P}_T X_{T+h})$ must be uncorrelated with the available information X_s, $s = 1, 2, \ldots, T$. The normal equations can be written as

$$\begin{aligned}(A_1, A_2, \ldots, A_T) &\begin{pmatrix} \Gamma(0) & \Gamma(1) & \ldots & \Gamma(T-1) \\ \Gamma'(1) & \Gamma(0) & \ldots & \Gamma(T-2) \\ \vdots & \vdots & \ddots & \vdots \\ \Gamma'(T-1) & \Gamma'(T-2) & \ldots & \Gamma(0) \end{pmatrix} \\ &= \left(\Gamma(h)\ \Gamma(h+1)\ \ldots\ \Gamma(T+h-1)\right).\end{aligned}$$

Denoting by $\mathbf{\Gamma}_T$ the matrix

$$\mathbf{\Gamma}_T = \begin{pmatrix} \Gamma(0) & \Gamma(1) & \ldots & \Gamma(T-1) \\ \Gamma'(1) & \Gamma(0) & \ldots & \Gamma(T-2) \\ \vdots & \vdots & \ddots & \vdots \\ \Gamma'(T-1) & \Gamma'(T-2) & \ldots & \Gamma(0) \end{pmatrix},$$

the normal equations can be written more compactly as

$$(A_1, A_2, \ldots, A_T)\, \mathbf{\Gamma}_T = \left(\Gamma(h)\ \Gamma(h+1)\ \ldots\ \Gamma(T+h-1)\right).$$

Using the assumption that $\{X_t\}$ is a VAR(1), $\Gamma(h)$ can be expressed as $\Gamma(h) = \Phi^h \Gamma(0)$ (see Eq. (12.3)) so that the normal equations become

$$\begin{aligned}(A_1, A_2, \ldots, A_T) &\begin{pmatrix} \Gamma(0) & \Phi\Gamma(0) & \ldots & \Phi^{T-1}\Gamma(0) \\ \Gamma(0)\Phi' & \Gamma(0) & \ldots & \Phi^{T-2}\Gamma(0) \\ \vdots & \vdots & \ddots & \vdots \\ \Gamma(0)\Phi'^{T-1} & \Gamma(0)\Phi'^{T-2} & \ldots & \Gamma(0) \end{pmatrix} \\ &= \left(\Phi^h\Gamma(0)\ \Phi^{h+1}\Gamma(0)\ \ldots\ \Phi^{T+h-1}\Gamma(0)\right).\end{aligned}$$

[1] If the mean is non-zero, a constant A_0 must be added to the forecast function.

The easily guessed solution is given by $A_1 = \Phi^h$ and $A_2 = \ldots = A_T = 0$. Thus, the sought-after forecasting function for the VAR(1) process is

$$\mathbb{P}_T X_{T+h} = \Phi^h X_T. \tag{14.2}$$

The forecast error $X_{T+h} - \mathbb{P}_T X_{T+h}$ has expectation zero. Thus, the linear least-squares predictor delivers unbiased forecasts. As

$$X_{T+h} = Z_{T+h} + \Phi Z_{T+h-1} + \ldots + \Phi^{h-1} Z_{T+1} + \Phi^h X_T,$$

the expected squared forecast error (mean squared error) MSE(h) is

$$\begin{aligned}\mathrm{MSE}(h) &= \mathbb{E}\left(X_{T+h} - \Phi^h X_T\right)\left(X_{T+h} - \Phi^h X_T\right)' \\ &= \Sigma + \Phi\Sigma\Phi' + \ldots + \Phi^{h-1}\Sigma\Phi'^{h-1} = \sum_{j=0}^{h-1} \Phi^j \Sigma \Phi'^j. \end{aligned} \tag{14.3}$$

In order to analyze the case of a causal VAR(p) process with $T > p$, we transform the model into the companion form. For $h = 1$, we can apply the result above to get:

$$\mathbb{P}_T Y_{T+1} = \Phi Y_T = \begin{pmatrix} \mathbb{P}_T X_{T+1} \\ X_T \\ X_{T-1} \\ \vdots \\ X_{T-p+2} \end{pmatrix} = \begin{pmatrix} \Phi_1 & \Phi_2 & \ldots & \Phi_{p-1} & \Phi_p \\ I_n & 0 & \ldots & 0 & 0 \\ 0 & I_n & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & I_n & 0 \end{pmatrix} \begin{pmatrix} X_T \\ X_{T-1} \\ X_{T-2} \\ \vdots \\ X_{T-p+1} \end{pmatrix}.$$

This implies that

$$\mathbb{P}_T X_{T+1} = \Phi_1 X_T + \Phi_2 X_{T-1} + \ldots + \Phi_p X_{T-p+1}. \tag{14.4}$$

The forecast error is $X_{T+1} - \mathbb{P}_T X_{T+1} = Z_t$ which has mean zero and covariance variance matrix Σ. In general we have that $\mathbb{P}_T Y_{T+h} = \Phi^h Y_T$ so that $\mathbb{P}_T X_{T+h}$ is equal to

$$\mathbb{P}_T X_{T+h} = \Phi_1^{(h)} X_T + \Phi_2^{(h)} X_{T-1} + \ldots + \Phi_p^{(h)} X_{T-p+1}$$

where $\Phi_i^{(h)}$, $i = 1, \ldots, p$, denote the blocks in the first row of Φ^h. Alternatively, the forecast for $h > 1$ can be computed recursively. For $h = 2$ this leads to:

$$\begin{aligned}\mathbb{P}_T X_{T+2} &= \mathbb{P}_T(\Phi_1 X_{T+1}) + \mathbb{P}_T(\Phi_2 X_T) + \ldots + \mathbb{P}_T\left(\Phi_p X_{T+2-p}\right) + \mathbb{P}_T(Z_{T+2}) \\ &= \Phi_1\left(\Phi_1 X_T + \Phi_2 X_{T-1} + \ldots + \Phi_p X_{T+1-p}\right) \\ &\quad + \Phi_2 X_T + \ldots + \Phi_p X_{T+2-p} \\ &= \left(\Phi_1^2 + \Phi_2\right) X_T + (\Phi_1\Phi_2 + \Phi_3) X_{T-1} + \ldots + \left(\Phi_1\Phi_{p-1} + \Phi_p\right) X_{T+2-p} \\ &\quad + \Phi_1\Phi_p X_{T+1-p}.\end{aligned}$$

For $h > 2$ we proceed analogously. This way of producing forecasts is sometimes called *iterated* forecasts.

In general, the forecast error of a causal VAR(p) process can be expressed as

$$\begin{aligned} X_{T+h} - \mathbb{P}_T X_{T+h} &= Z_{T+h} + \Psi_1 Z_{T+h-1} + \ldots + \Psi_{h-1} Z_{T+1} \\ &= \sum_{j=0}^{h-1} \Phi_j Z_{T+h-j}. \end{aligned}$$

The MSE(h) then is:

$$\mathrm{MSE}(h) = \Sigma + \Psi_1 \Sigma \Psi_1' + \ldots + \Psi_{h-1} \Sigma \Psi_{h-1}' = \sum_{j=0}^{h-1} \Psi_j \Sigma \Psi_j'. \tag{14.5}$$

Example

Consider again the VAR(2) model of Sect. 12.3. The forecast function in this case is then:

$$\begin{aligned} \mathbb{P}_T X_{T+1} &= \Phi_1 X_t + \Phi_2 X_{t-1} \\ &= \begin{pmatrix} 0.8 & -0.5 \\ 0.1 & -0.5 \end{pmatrix} X_t + \begin{pmatrix} -0.3 & -0.3 \\ -0.2 & 0.3 \end{pmatrix} X_{t-1}, \\ \mathbb{P}_T X_{T+2} &= (\Phi_1^2 + \Phi_2) X_t + \Phi_1 \Phi_2 X_{t-1} \\ &= \begin{pmatrix} 0.29 & -0.45 \\ -0.17 & 0.50 \end{pmatrix} X_t + \begin{pmatrix} -0.14 & -0.39 \\ 0.07 & -0.18 \end{pmatrix} X_{t-1}, \\ \mathbb{P}_T X_{T+3} &= (\Phi_1^3 + \Phi_1 \Phi_2 + \Phi_2 \Phi_1) X_t + (\Phi_1^2 \Phi_2 + \Phi_2^2) X_{t-1} \\ &= \begin{pmatrix} 0.047 & -0.310 \\ -0.016 & -0.345 \end{pmatrix} X_t + \begin{pmatrix} 0.003 & -0.222 \\ -0.049 & 0.201 \end{pmatrix} X_{t-1}. \end{aligned}$$

Based on the results computed in Sect. 12.3, we can calculate the corresponding mean squared errors (MSE):

$$\begin{aligned} \mathrm{MSE}(1) &= \Sigma = \begin{pmatrix} 1.0 & 0.4 \\ 0.4 & 2.0 \end{pmatrix}, \\ \mathrm{MSE}(2) &= \Sigma + \Psi_1 \Sigma \Psi_1' = \begin{pmatrix} 1.82 & 0.80 \\ 0.80 & 2.47 \end{pmatrix}, \\ \mathrm{MSE}(3) &= \Sigma + \Psi_1 \Sigma \Psi_1' + \Psi_2 \Sigma \Psi_2' = \begin{pmatrix} 2.2047 & 0.3893 \\ 0.3893 & 2.9309 \end{pmatrix}. \end{aligned}$$

A practical forecasting exercise with additional material is presented in Sect. 14.4.

14.1.1 Wold Decomposition Theorem

At this stage we note that *Wold's theorem* or *Wold's Decomposition* carries over to the multivariate case (see Sect. 3.2 for the univariate case). This Theorem asserts that there exists for each purely non-deterministic stationary process[2] a decomposition, respectively representation, of the form:

$$X_t = \mu + \sum_{j=0}^{\infty} \Psi_j Z_{t-j},$$

where $\Psi_0 = I_n$, $Z_t \sim \text{WN}(0, \Sigma)$ with $\Sigma > 0$ and $\sum_{j=0}^{\infty} \|\Psi_j\|^2 < \infty$. The innovations $\{Z_t\}$ have the property $Z_t = X_t - \widetilde{\mathbb{P}}_{t-1} X_t$ and consequently $Z_t = \widetilde{\mathbb{P}}_t Z_t$. Thereby $\widetilde{\mathbb{P}}_t$ denotes the linear least-squares predictor based on the infinite past $\{X_t, X_{t-1}, \ldots\}$. The interpretation of the multivariate case is analogous to the univariate one.

14.2 Forecasting with Estimated Parameters

In practice the parameters of the VAR model are usually unknown and have therefore to be estimated. In the previous Section we have demonstrated that

$$\mathbb{P}_T X_{T+h} = \Phi_1 \mathbb{P}_T X_{T+h-1} + \ldots + \Phi_p \mathbb{P}_T X_{T+h-p}$$

where $\mathbb{P}_T X_{T+h-j} = Y_{T+h-j}$ if $j \geq h$. Replacing the true parameters by their estimates, we get the forecast function

$$\widehat{\mathbb{P}}_T X_{T+h} = \widehat{\Phi}_1 \widehat{\mathbb{P}}_T X_{T+h-1} + \ldots + \widehat{\Phi}_p \widehat{\mathbb{P}}_T X_{T+h-p}.$$

where a hat indicates the use of estimates. The forecast error can then be decomposed into two components:

$$\begin{aligned} X_{T+h} - \widehat{\mathbb{P}}_T X_{T+h} &= (X_{T+h} - \mathbb{P}_T X_{T+h}) + \left(\mathbb{P}_T X_{T+h} - \widehat{\mathbb{P}}_T X_{T+h}\right) \\ &= \sum_{j=0}^{h-1} \Phi_j Z_{T+h-j} + \left(\mathbb{P}_T X_{T+h} - \widehat{\mathbb{P}}_T X_{T+h}\right). \end{aligned} \tag{14.6}$$

Dufour (1985) has shown that, under the assumption of symmetrically distributed Z_t's (i.e. if Z_t and $-Z_t$ have the same distribution) the expectation of the forecast error is zero even when the parameters are replaced by their least-squares estimates.

[2] A stationary stochastic process is called deterministic if it can be perfectly forecasted from its infinite past. It is called purely non-deterministic if there is no deterministic component (see Sect. 3.2).

This result holds despite the fact that these estimates are biased in small samples. Moreover, the results do not assume that the model is correctly specified in terms of the order p. Thus, under quite general conditions the forecast with estimated coefficients remains unbiased so that $\mathbb{E}\left(X_{T+h} - \widehat{\mathbb{P}}_T X_{T+h}\right) = 0$.

If the estimation is based on a different sample than the one used for forecasting, the two terms in the above expression are uncorrelated so that its mean squared error is by the sum of the two mean squared errors:

$$\begin{aligned}\widehat{\text{MSE}}(h) = &\sum_{j=0}^{h-1} \Psi_j \Sigma \Psi_j' \\ &+\mathbb{E}\left(\mathbb{P}_T X_{T+h} - \widehat{\mathbb{P}}_T X_{T+h}\right)\left(\mathbb{P}_T X_{T+h} - \widehat{\mathbb{P}}_T X_{T+h}\right)'. \end{aligned} \qquad (14.7)$$

The last term can be evaluated by using the asymptotic distribution of the coefficients as an approximation. The corresponding formula turns out to be cumbersome. The technical details can be found in Lütkepohl (2006) and Reinsel (1993). The formula can, however, be simplified considerably if we consider a forecast horizon of only one period. We deduce the formula for a VAR of order one, i.e. taking $X_t = \Phi X_{t-1} + Z_t$, $Z_t \sim \text{WN}(0, \Sigma)$.

$$\mathbb{P}_T X_{T+h} - \widehat{\mathbb{P}}_T X_{T+h} = (\Phi - \widehat{\Phi})X_T = \text{vec}\left((\Phi - \widehat{\Phi})X_T\right) = (X_T' \otimes I_n)\,\text{vec}(\Phi - \widehat{\Phi}).$$

This implies that

$$\begin{aligned}&\mathbb{E}\left(\mathbb{P}_T X_{T+h} - \widehat{\mathbb{P}}_T X_{T+h}\right)\left(\mathbb{P}_T X_{T+h} - \widehat{\mathbb{P}}_T X_{T+h}\right)' \\ &\quad = \mathbb{E}(X_T' \otimes I_n)\,\text{vec}(\Phi - \widehat{\Phi})(\text{vec}(\Phi - \widehat{\Phi}))'(X_T \otimes I_n) \\ &\quad = \mathbb{E}(X_T' \otimes I_n)\frac{\mathbf{\Gamma}_1^{-1} \otimes \Sigma}{T}(X_T \otimes I_n) = \frac{1}{T}\mathbb{E}(X_T' \mathbf{\Gamma}_1^{-1} X_T) \otimes \Sigma \\ &\quad = \frac{1}{T}\mathbb{E}(\text{tr} X_T' \mathbf{\Gamma}_1^{-1} X_T) \otimes \Sigma = \frac{1}{T}\text{tr}(\mathbf{\Gamma}_1^{-1}\mathbb{E}(X_T X_T')) \otimes \Sigma \\ &\quad = \frac{1}{T}(\text{tr}(I_n) \otimes \Sigma) = \frac{n}{T}\Sigma.\end{aligned}$$

Thereby, we have used the asymptotic normality of the least-squares estimator (see Theorem 13.1) and the assumption that forecasting and estimation uses different realizations of the stochastic process. Thus, for $h = 1$ and $p = 1$, we get

$$\widehat{\text{MSE}}(1) = \Sigma + \frac{n}{T}\Sigma = \frac{T+n}{T}\,\Sigma.$$

Higher order models can be treated similarly using the companion form of VAR(p). In this case:

$$\widehat{\text{MSE}}(1) = \Sigma + \frac{np}{T}\Sigma = \frac{T+np}{T}\,\Sigma. \tag{14.8}$$

This is only an approximation as we applied asymptotic results to small sample entities. The expression shows that the effect of the substitution of the coefficients by their least-squares estimates vanishes as the sample becomes large. However, in small sample the factor $\frac{T+np}{T}$ can be sizeable. In the example treated in Sect. 14.4, the covariance matrix Σ, taking the use of a constant into account and assuming 8 lags, has to be inflated by $\frac{T+np+1}{T} = \frac{196+4\times 8+1}{196} = 1.168$. Note also that the precision of the forecast, given Σ, diminishes with the number of parameters.

14.3 Modeling of VAR Models

The previous section treated the estimation of VAR models under the assumption that the order of the VAR, p, is known. In most cases, this assumption is unrealistic as the order p is unknown and must be retrieved from the data. We can proceed analogously as in the univariate case (see Sect. 5.1) and iteratively test the hypothesis that coefficients corresponding to the highest lag, i.e. $\Phi_p = 0$, are simultaneously equal to zero. Starting from a maximal order p_{max}, we test the null hypothesis that $\Phi_{p_{max}} = 0$ in the corresponding VAR(p_{max}) model. If the hypothesis is not rejected, we reduce the order by one to $p_{max} - 1$ and test anew the null hypothesis $\Phi_{p_{max}-1} = 0$ using the smaller VAR($p_{max} - 1$) model. One continues in this way until the null hypothesis is rejected. This gives, then, the appropriate order of the VAR. The different tests can be carried out either as Wald-tests (F-tests) or as likelihood-ratio tests (χ^2-tests) with n^2 degrees of freedom.

An alternative procedure to determine the order of the VAR relies on some information criteria. As in the univariate case, the most popular ones are the Akaike (AIC), the Schwarz or Bayesian (BIC) and the Hannan-Quinn criterion (HQC). The corresponding formula are:

$$\begin{aligned}
\text{AIC(p)}: &\quad \ln\det\widetilde{\Sigma}_p + \frac{2pn^2}{T}, \\
\text{BIC(p)}: &\quad \ln\det\widetilde{\Sigma}_p + \frac{pn^2}{T}\ln T, \\
\text{HQC(p)}: &\quad \ln\det\widetilde{\Sigma}_p + \frac{2pn^2}{T}\ln(\ln T),
\end{aligned}$$

where $\widetilde{\Sigma}_p$ denotes the degree of freedom adjusted estimate of the covariance matrix Σ for a model of order p (see equation(13.5)). n^2p is the number of estimated coefficients. The estimated order is then given as the minimizer of one of these criteria. In practice the Akaike's criterion is the most popular one although it has a tendency to deliver orders which are too high. The BIC and the HQ-criterion on the other hand deliver the correct order on average, but can lead to models which suffer from the omitted variable bias when the estimated order is too low. Examples are discussed in Sects. 14.4 and 15.4.5.

Following Lütkepohl (2006), Akaike's information criterion can be rationalized as follows. Take as a measure of fit the determinant of the one period approximate mean-squared errors $\widehat{\mathrm{MSE}}(1)$ from Eq. (14.8) and take as an estimate of Σ the degrees of freedom corrected version in Eq. (13.5). The resulting criterion is called according to Akaike (1969) the *final prediction error* (FPE):

$$\mathrm{FPE}(p) = \det\left(\frac{T+np}{T} \times \frac{T}{T-np}\widetilde{\Sigma}\right) = \left(\frac{T+np}{T-np}\right)^n \det\widetilde{\Sigma}. \tag{14.9}$$

Taking logs and using the approximations $\frac{T+np}{T-np} \approx 1 + \frac{2np}{T}$ and $\log(1+\frac{2np}{T}) \approx \frac{2np}{T}$, we arrive at

$$\mathrm{AIC}(p) \approx \log FPE(p).$$

14.4 Example: A VAR Model for the U.S. Economy

In this section, we illustrate how to build and use VAR models for forecasting key macroeconomic variables. For this purpose, we consider the following four variables: GDP per capita ($\{Y_t\}$), price level in terms of the consumer price index (CPI) ($\{P_t\}$), real money stock M1 ($\{M_t\}$), and the three month treasury bill rate ($\{R_t\}$). All variables are for the U.S. and are, with the exception of the interest rate, in logged differences.[3] The components of X_t are with the exception of the interest rate stationary.[4] Thus, we aim at modeling $X_t = (\Delta\log Y_t, \Delta\log P_t, \Delta\log M_t, R_t)'$. The sample runs from the first quarter 1959 to the first quarter 2012. We estimate our models, however, only up to the fourth quarter 2008 and reserve the last thirteen quarters, i.e. the period from the first quarter 2009 to first quarter of 2012, for an out-of-sample evaluation of the forecast performance. This forecast assessment has the advantage to account explicitly of the sampling variability in estimated parameter models.

The first step in the modeling process is the determination of the lag-length. Allowing for a maximum of twelve lags, the different information criteria produce the values reported in Table 14.1. Unfortunately, the three criteria deliver different

[3] Thus, $\Delta\log P_t$ equals the inflation rate.

[4] Although the unit root test indicate that R_t is integrated of order one, we do not difference this variable. This specification will not affect the consistency of the estimates nor the choice of the lag-length (Sims et al. 1990), but has the advantage that each component of X_t is expressed in percentage points which facilitates the interpretation.

Table 14.1 Information criteria for the VAR models of different orders

Order	AIC	BIC	HQ
0	−14.498	−14.429	−14.470
1	−17.956	−17.611	−17.817
2	−18.638	**−18.016**	−18.386
3	−18.741	−17.843	−18.377
4	−18.943	−17.768	−18.467
5	−19.081	−17.630	**−18.493**
6	−19.077	−17.349	−18.377
7	−19.076	−17.072	−18.264
8	**−19.120**	−16.839	−18.195
9	−18.988	−16.431	−17.952
10	−18.995	−16.162	−17.847
11	−18.900	−15.789	−17.639
12	−18.884	−15.497	−17.512

Minimum in bold

orders: AIC suggests 8 lags, HQ 5 lags, and BIC 2 lags. In such a situation it is wise to keep all three models and to perform additional diagnostic tests.[5] One such test is to run a horse-race between the three models in terms of their forecasting performance.

We evaluate the forecasts according to the two criteria: the root-mean-squared-error (RMSE) and the mean-absolute-error (MAE)[6]:

$$\text{RMSE}: \quad \sqrt{\frac{1}{h}\sum_{T+1}^{T+h}(\widehat{X}_{it} - X_{it})^2} \tag{14.10}$$

$$\text{MAE}: \quad \frac{1}{h}\sum_{T+1}^{T+h}|\widehat{X}_{it} - X_{it}| \tag{14.11}$$

where $\widehat{X}_{it}$ and X_{it} denote the forecast and the actual value of variable i in period t. Forecasts are computed for a horizon h starting in period T. We can gain further insights by decomposing the mean-squared-error additively into three components:

$$\frac{1}{h}\sum_{T+1}^{T+h}(\widehat{X}_{it} - X_{it})^2 = \left(\left(\frac{1}{h}\sum_{T+1}^{T+h}\widehat{X}_{it}\right) - \overline{X}_i\right)^2$$
$$+(\sigma_{\widehat{X}_i} - \sigma_{X_i})^2 + 2(1-\rho)\sigma_{\widehat{X}_i}\sigma_{X_i}.$$

[5] Such tests would include an analysis of the autocorrelation properties of the residuals and tests of structural breaks.

[6] Alternatively one could use the mean-absolute-percentage-error (MAPE). However, as all variables are already in percentages, the MAE is to be preferred.

The first component measures how far the mean of the forecasts $\frac{1}{h}\sum_{T+1}^{T+h}\widehat{X}_{it}$ is away from the actual mean of the data $\overline{X}_i$. It therefore measures the *bias* of the forecasts. The second one compares the standard deviation of the forecast $\sigma_{\widehat{X}_i}$ to those of the data σ_{X_i}. Finally, the last component is a measure of the unsystematic forecast errors where ρ denotes the correlation between the forecast and the data. Ideally, each of the three components should be close to zero: there should be no bias, the variation of the forecasts should correspond to those of the data, and the forecasts and the data should be highly positively correlated. In order to avoid scaling problems, all three components are usually expressed as a proportion of $\frac{1}{h}\sum_{T+1}^{T+h}(\widehat{X}_{it}-X_{it})^2$:

$$\text{bias proportion:} \qquad \frac{\left(\left(\frac{1}{h}\sum_{T+1}^{T+h}\widehat{X}_{it}\right)-\overline{X}_i\right)^2}{\frac{1}{h}\sum_{T+1}^{T+h}(\widehat{X}_{it}-X_{it})^2} \tag{14.12}$$

$$\text{variance proportion:} \qquad \frac{(\sigma_{\widehat{X}_i}-\sigma_{X_i})^2}{\frac{1}{h}\sum_{T+1}^{T+h}(\widehat{X}_{it}-X_{it})^2} \tag{14.13}$$

$$\text{covariance proportion:} \qquad \frac{2(1-\rho)\sigma_{\widehat{X}_i}\sigma_{X_i}}{\frac{1}{h}\sum_{T+1}^{T+h}(\widehat{X}_{it}-X_{it})^2} \tag{14.14}$$

We use these models to produce dynamic or iterated forecasts $\mathbb{P}_T X_{T+1}, \mathbb{P}_T X_{T+2}, \ldots, \mathbb{P}_T X_{T+h}$. Forecasts for $h \geq 2$ are computed iteratively by inserting for the lagged variables the forecasts obtained in the previous steps. For details see Chap. 14. Alternatively, one may consider a recursive or rolling out-of-sample strategy where the model is reestimated each time a new observation becomes available. Thus, we would evaluate the one-period-ahead forecasts $\mathbb{P}_T X_{T+1}, \mathbb{P}_{T+1} X_{T+2}, \ldots, \mathbb{P}_{T+h-1} X_{T+h}$, the two-period-ahead forecasts $\mathbb{P}_T X_{T+2}, \mathbb{P}_{T+1} X_{T+3}, \ldots, \mathbb{P}_{T+h-2} X_{T+h}$, and so on. The difference between the recursive and the rolling strategy is that in the first case all observations are used for estimation whereas in the second case the sample is rolled over so that its size is kept fixed at T.

Figure 14.1 displays dynamic or iterated forecasts for the four variables expressed in log-levels, respectively in levels for the interest rate. Forecast are evaluated according to the performance measures explained above. The corresponding values are reported in Table 14.2. All models see a quick recovery after the recession in 2008 and are thus much too optimistic. The lowest RMSE for $\log Y_t$ is 5.678 for the VAR(8) model. Thus, GDP per capita is predicted to be on average almost 6 % too high over the forecast period. This overly optimistic forecast is reflected in a large bias proportion which amounts to more than 95 %. The situation looks much better for the price level. All models see an increase in inflation starting in 2009. Especially, the two higher order models fare much better. Their RMSE is just over 1 %. The bias proportion is practically zero for the VAR(8) model. The forecast results of the real money stock are mixed. All models predict a quick recovery. This took indeed place, but first at a more moderate pace. Starting in

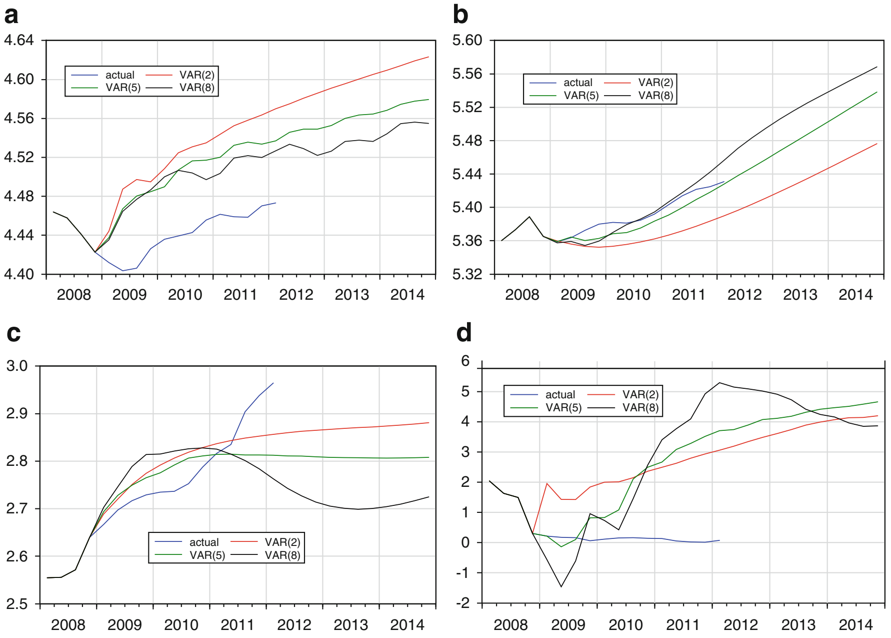

Fig. 14.1 Forecast comparison of alternative models. (**a**) $\log Y_t$. (**b**) $\log P_t$. (**c**) $\log M_t$. (**d**) R_t

mid-2010 the unconventional monetary policy of quantitative easing, however, led to an unforeseen acceleration so that the forecasts turned out to be systematically too low for the later period. Interestingly, the smallest model fared significantly better than the other two. Finally, the results for the interest rates are very diverse. Whereas the VAR(2) model predicts a rise in the interest rate, the other models foresee a decline. The VAR(8) model even predicts a very drastic fall. However, all models miss the continuation of the low interest rate regime and forecasts an increase starting already in 2009. This error can again be attributed to the unforeseen low interest rate monetary policy which was implemented in conjunction with the quantitative easing. This misjudgement resulted in a relatively large bias proportion.

Up to now, we have just been concerned with *point forecasts*. Point forecasts, however, describe only one possible outcome and do not reflect the inherent uncertainty surrounding the prediction problem. It is, thus, a question of scientific integrity to present in addition to the point forecasts also confidence intervals. One straightforward way to construct such intervals is by computing the matrix of mean-squared-errors MSE using Eq. (14.5). The diagonal elements of this matrix can be interpreted as a measure of the forecast error variances for each variable. Under the assumption that the innovations $\{Z_t\}$ are Gaussian, such confidence intervals can be easily computed. However, in practice this assumption is likely to be violated. This problem can be circumvented by using the empirical distribution function of the residuals to implement a bootstrap method similar to the computation of the

Table 14.2 Forecast evaluation of alternative VAR models

	VAR(2)	VAR(5)	VAR(8)
	$\log Y_t$		
RMSE	8.387	6.406	5.678
Bias proportion	0.960	0.961	0.951
Variance proportion	0.020	0.010	0.001
Covariance proportion	0.020	0.029	0.048
MAE	8.217	6.279	5.536
	$\log P_t$		
RMSE	3.126	1.064	1.234
Bias proportion	0.826	0.746	0.001
Variance proportion	0.121	0.001	0.722
Covariance proportion	0.053	0.253	0.278
MAE	2.853	0.934	0.928
	$\log M_t$		
RMSE	5.616	6.780	9.299
Bias proportion	0.036	0.011	0.002
Variance proportion	0.499	0.622	0.352
Covariance proportion	0.466	0.367	0.646
MAE	4.895	5.315	7.762
	R_t		
RMSE	2.195	2.204	2.845
Bias proportion	0.367	0.606	0.404
Variance proportion	0.042	0.337	0.539
Covariance proportion	0.022	0.057	0.057
MAE	2.125	1.772	2.299

RMSE and MAE for $\log Y_t$, $\log P_t$, and $\log M_t$ are multiplied by 100

Value-at-Risk in Sect. 8.4. Figure 14.2 plots the forecasts of the VAR(8) model together with a 80 % confidence interval computed from the bootstrap approach. It shows that, with the exception of the logged price level, the actual realizations fall out of the confidence interval despite the fact that the intervals are already relatively large. This documents the uniqueness of the financial crisis and gives a hard time for any forecasting model.

Instead of computing a confidence interval, one may estimate the probability distribution of possible future outcomes. This provides a complete description of the uncertainty related to the prediction problem (Christoffersen 1998; Diebold et al. 1998; Tay and Wallis 2000; Corradi and Swanson 2006). Finally, one should be aware that the innovation uncertainty is not the only source of uncertainty. As the parameters of the model are themselves estimated, there is also a coefficient uncertainty. In addition, we have to face the possibility that the model is misspecified.

The forecasting performance of the VAR models may seem disappointing at first. However, this was only be a first attempt and further investigations are usually necessary. These may include the search for *structural breaks* (See Bai et al. 1998; Perron 2006). This topic is treated in Sect. 18.1. Another reason for the poor

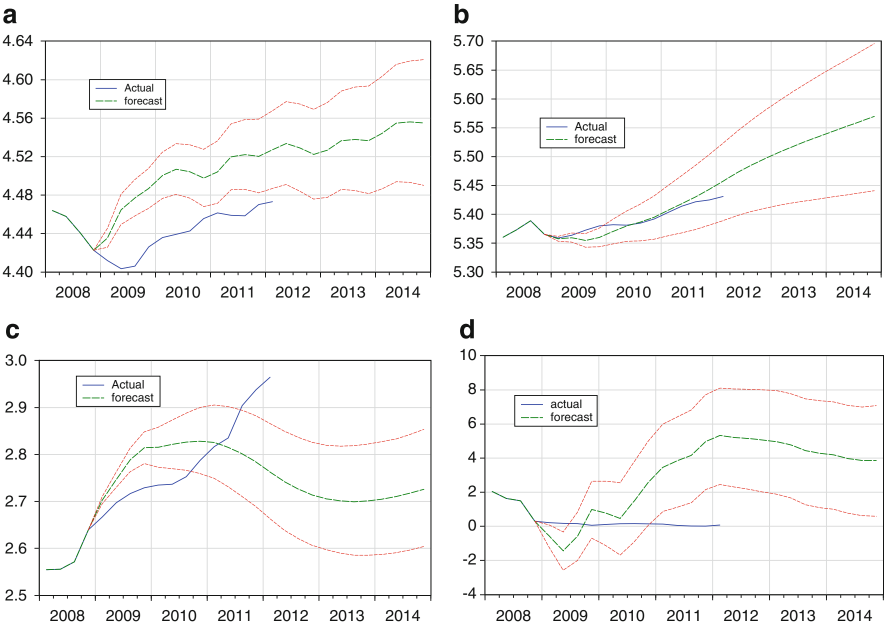

Fig. 14.2 Forecast of VAR(8) model and 80 % confidence intervals (*red dotted lines*). (**a**) log Y_t. (**b**) log P_t. (**c**) log M_t. (**d**) R_t

forecasting may be due to the *over-parametrization* of VAR models. The VAR(8) model, for example, 32 lagged dependent variables plus a constant in each of the four equations which leads to a total 132 parameters. This problem can be dealt with by applying Bayesian shrinkage techniques. This approach, also known as Bayesian VAR (BVAR), was particularly successful when using the so-called Minnesota prior (See Doan et al. 1984; Litterman 1986; Kunst and Neusser 1986; Banbura et al. 2010). This prior is presented in Sect. 18.2.

Besides these more fundamental issues, one may rely on more technical remedies. One such remedy is the use of direct rather iterated forecasts. This difference is best explained in the context of the VAR(1) model $X_t = \Phi X_{t-1} + Z_t, Z_t \sim \mathrm{WN}(0, \Sigma)$. The *iterated* forecast for X_{T+h} uses the OLS-estimate $\widehat{\Phi}$ to compute the forecast $\widehat{\Phi}^h X_T$ (see Chap. 14). Alternatively, one may estimate instead of the VAR(1), the model $X_t = \Upsilon X_{t-h} + Z_t$ and compute the *direct* forecast for X_{T+h} as $\widehat{\Upsilon} X_T$. Although $\widehat{\Upsilon}$ has larger variance than $\widehat{\Phi}$ if the VAR(1) is correctly specified, it is robust to misspecification (see Bhansali 1999; Schorfheide 2005; Marcellino et al. 2006).

Another interesting and common device is intercept correction or residual adjustment. Thereby the constant terms are adjusted in such a way that the residuals of the most recent observation become zero. The model is thereby set back on track. In this way the forecaster can guard himself against possible structural breaks. Residual adjustment can also serve as a device to incorporate anticipated events, like announced policies, which are not yet incorporated into the model. See Clements and Hendry (1996, 2006) for further details and additional forecasting devices.