Lecture 6: Forecasting

Bus 41910, Time Series Analysis, Mr. R. Tsay

Forecasting is one of the main objectives of time series analysis. Typically, we use the criterion of **minimum mean squared errors** to produce point forecasts. Like other statistical forecasts, there are two main sources of uncertainty involved. The first uncertainty is concerned with "future" variables and the second is the uncertainty about the model used. An example of the second uncertainty is that the parameters of the model used are estimates, not "true" values. For simplicity, we shall first focus on "conditional forecasts" which assume that the fitted model is the *true* model. In other words, we shall begin with methods that ignore the second source of uncertainty. Later we shall discuss methods that take into consideration model uncertainty.

A. Forecast of a general linear model: Recall that there are three representations for a general linear time series Z_t:

- MA representation:
$$Z_t = \mu + \sum_{i=0}^{\infty} \psi_i a_{t-i} \quad \text{with} \quad \psi_0 = 1.$$
- AR representation:
$$Z_t = d + \sum_{i=1}^{\infty} \pi_i Z_{t-i} + a_t,$$
where $d = \pi(1)\mu$ is a constant.
- ARMA (or ARIMA) representation:
$$\phi(B)(1-B)^d Z_t = c + \theta(B) a_t.$$

Here we do not separate seasonal from non-seasonal models in the discussion of ARIMA representation. The argument in effect applies to all ARMA models. Also, for simplicity, we shall let $c = d = \mu = 0$.

These three representations are useful in different aspects of forecasting. Suppose that the *forecast origin* is T and we are interested in forecasting $Z_{T+\ell}$ for $\ell > 0$. Such a forecast is called an ℓ-step ahead forecast at time T, with ℓ being the *forecast horizon*. Mathematically, we like to find a forecast, say $\hat{Z}$, which satisfies

$$E[(Z_{T+\ell} - \hat{Z})^2 | F_T] = \min_f E[(Z_{T+\ell} - f)^2 | F_T],$$

where $\hat{Z}$ and f are functions of $Z_T, Z_{T-1}, \cdots$ and the model and F_T denotes the available information at time T. Obviously, the forecast turns out to be the conditional expectation of $Z_{T+\ell}$ given $Z_T, Z_{T-1}, \cdots$ and the model. [Use $E[(Z_{T+\ell} - \hat{Z}_T(\ell)) + (\hat{Z}_T(\ell) - f)]^2$, where

$\hat{Z}_T(\ell) = E(Z_{T+\ell}|F_T)$ is the conditional expectation of $Z_{T+\ell}$ given the information available at time T.] Denote the ℓ-step ahead forecast by $\hat{Z}_T(\ell)$. Then,

$$\hat{Z}_T(\ell) = E(Z_{T+\ell}|Z_T, Z_{T-1}, \cdots, \text{model}).$$

The ℓ-step ahead forecast error is

$$e_T(\ell) = Z_{T+\ell} - \hat{Z}_T(\ell)$$

and the variance of the forecast error is

$$\text{Var}[e_T(\ell)] = \text{Var}[Z_{T+\ell} - \hat{Z}_T(\ell)].$$

From the MA representation of $Z_{T+\ell}$, we have (assuming $\mu = 0$)

$$\hat{Z}_T(\ell) = \psi_\ell a_T + \psi_{\ell+1} a_{T-1} + \cdots = \sum_{i=0}^{\infty} \psi_{\ell+i} a_{T-i}$$

and

$$e_T(\ell) = \sum_{i=0}^{\ell-1} \psi_i a_{t+\ell-i} \quad \text{and} \quad \text{Var}[e_T(\ell)] = (\sum_{i=0}^{\ell-1} \psi_i^2)\sigma_a^2.$$

In particular, $e_T(1) = a_{T+1}$ and $\text{Var}[e_T(1)] = \sigma_a^2$. Alternatively, from the AR representation of $Z_{T+\ell}$, we obtain

$$\hat{Z}_T(1) = \sum_{i=0}^{\infty} \pi_{1+i} Z_{T-i} = \sum_{i=1}^{\infty} \pi_i Z_{T+1-i} \quad \text{and} \quad e_T(1) = a_{T+1}.$$

For $\ell > 1$, consider

$$Z_{T+\ell} = \pi_1 Z_{T+\ell-1} + \pi_2 Z_{T+\ell-2} + \pi_3 Z_{T+\ell-3} + \cdots + a_{T+\ell}.$$

Taking conditional expectation, we obtain

$$Z_T(\ell) = \pi_1 \hat{Z}_T(\ell-1) + \cdots + \pi_{\ell-1} \hat{Z}_T(1) + \pi_\ell Z_T + \pi_{\ell+1} Z_{T-1} + \cdots.$$

This equation can be used repeatedly to obtain a general formula for forecasting. For instance, for $\ell = 2$,

$$\begin{aligned}
\hat{Z}_T(2) &= \pi_1 \hat{Z}_T(1) + \pi_2 Z_T + \pi_3 Z_{T-1} + \cdots \\
&= \pi_1(\pi_1 Z_T + \pi_2 Z_{T-1} + \cdots) + \pi_2 Z_T + \pi_3 Z_{T-1} + \cdots \\
&= (\pi_1^2 + \pi_2) Z_T + (\pi_1\pi_2 + \pi_3) Z_{T-1} + (\pi_1\pi_3 + \pi_4) Z_{T-2} + \cdots \\
&= \sum_{i=0}^{\infty} \pi_{1+i}^{(2)} Z_{T-i}
\end{aligned}$$

where $\pi_j^{(2)} = \pi_1\pi_j + \pi_{j+1}$ for $j > 0$. In general, we have

$$\hat{Z}_T(\ell) = \sum_{i=0}^{\infty} \pi_{1+i}^{(\ell)} Z_{T-i} = (\pi_1^{(\ell)} + \pi_2^{(\ell)} B + \cdots) Z_T = (\pi_1^{(\ell)} B^\ell + \pi_2^{(\ell)} B^{\ell+1} + \cdots) Z_{T+\ell}$$

where $\pi_j^{(\ell)}$ is a function of π-weights and ψ-weights of Z_t.
To derive the general formula for $\hat{Z}_T(\ell)$, recall that its forecast error is

$$e_T(\ell) = \sum_{i=0}^{\ell-1} \psi_i a_{T+\ell-i} = (\psi_0 + \psi_1 B + \cdots + \psi_{\ell-1} B^{\ell-1}) a_{T+\ell}.$$

From the AR representation, we have

$$\pi(B) Z_{T+\ell} = a_{T+\ell}.$$

Multiplying the above equation by $(\psi_0 + \psi_1 B + \cdots + \psi_{\ell-1} B^{\ell-1})$, we obtain

$$(\psi_0 + \psi_1 B + \cdots + \psi_{\ell-1} B^{\ell-1}) \pi(B) Z_{T+\ell} = (\psi_0 + \psi_1 B + \cdots + \psi_{\ell-1} B^{\ell-1}) a_{T+\ell}.$$

Since $\psi(B)\pi(B) = 1$ for all B, it is easily seen that the coefficients of B^i in the left hand side are zero for $i = 1, 2, \cdots, \ell - 1$. Therefore, in conjunction with the fact that the right hand side is the ℓ-step ahead forecast error, the above equation in effect is in the form

$$Z_{T+\ell} - \hat{Z}_T(\ell) = \text{forecast error.}$$

Consequently, we have

$$(\psi_0 + \psi_1 B + \cdots + \psi_{\ell-1} B^{\ell-1})(1 - \pi_1 B - \pi_2 B^2 - \cdots) = 1 - \pi_1^{(\ell)} B^\ell - \pi_2^{(\ell)} B^{\ell+1} + \cdots.$$

By equating coefficients, we obtain that, for $j > 0$,

$$\pi_j^{(\ell)} = \sum_{i=0}^{\ell-1} \pi_{j+i} \psi_{\ell-i-1}$$

In particular, for $j = 1$, we have

$$\pi_1^{(\ell)} = \psi_\ell.$$

It is easy to show that $\pi_j^{(\ell)} = \pi_{j+1}^{(\ell-1)} + \psi_{\ell-1}\pi_j$ where, of course, $\pi_j^{(1)} = \pi_j$.

Eventual Forecasting Function: In many applications, we are interested in obtaining ℓ-step ahead forecasts for $\ell = 1, \cdots, k$. In this situation, the ARMA representation is useful. For convenience, we extend the forecasting notation by defining

$$\hat{Z}_T(\ell) = Z_{T+\ell} \quad \text{for} \quad \ell \leq 0.$$

Also, define

$$\hat{a}_T(\ell) = \begin{cases} 0 & \text{for } \ell > 0 \\ a_{T+\ell} & \text{for } \ell \leq 0 \end{cases}$$

Then, taking conditional expectation of the equation

$$Z_{T+\ell} - \phi_1 Z_{T+\ell-1} - \cdots - \phi_p Z_{T+\ell-p} = a_{T+\ell} - \theta_1 a_{T+\ell-1} - \cdots - \theta_q a_{T+\ell-q}$$

we obtain

$$\hat{Z}_T(\ell) - \phi_1 \hat{Z}_T(\ell-1) - \cdots - \phi_p \hat{Z}_T(\ell-p) = \hat{a}_T(\ell) - \theta_1 \hat{a}_T(\ell-1) - \cdots - \theta_q \hat{a}_T(\ell-q).$$

This is a recursive formula which can be used to compute the forecasts. In particular, for $\ell > q$, we have

$$\hat{Z}_T(\ell) - \phi_1 \hat{Z}_T(\ell-1) - \cdots - \phi_p \hat{Z}_T(\ell-p) = 0.$$

By letting the backshift operator "B" operate on ℓ, we have

$$\phi(B)\hat{Z}_T(\ell) = 0 \quad \text{for} \quad \ell > q.$$

This is called the "eventual forecasting function" of Z_t. It describes the forecasting pattern of Z_t. Once again, the forecasts satisfy the AR difference equation for $\ell > q$. Similar results hold for the ARIMA models.

Updating Formula: In some applications, we may wish to update the forecast in light of newly available information. For instance, we may wish to update $\hat{Z}_T(2)$ when Z_{T+1} becomes available. A simple example is that the US government may revise its October's forecast of December unemployment rate when the November unemployment rate is available. The general situation is as follows:
Original forecast: $\hat{Z}_T(\ell)$.
Updated forecast: $\hat{Z}_{T+1}(\ell-1)$.
What is the relation between these two forecasts?
To answer this question, write the model as

$$Z_{T+\ell} = \hat{Z}_T(\ell) + a_{T+\ell} + \psi_1 a_{T+\ell-1} + \cdots + \psi_{\ell-1} a_{T+1}.$$

On the other hand, we can also write it as

$$Z_{T+\ell} = \hat{Z}_{T+1}(\ell-1) + a_{T+\ell} + \psi_1 a_{T+\ell-1} + \cdots + \psi_{\ell-2} a_{T+2}.$$

Therefore, an updating formula is

$$\hat{Z}_{T+1}(\ell-1) = \hat{Z}_T(\ell) + \psi_{\ell-1} a_{T+1}.$$

Thus, $\psi_{\ell-1}$ is the only quantity needed to revise the forecast when a_{T+1} is available. This is natural in light of the fact that $\psi_{\ell-1}$ is the effect of the innovation a_{T+1} on $Z_{T+\ell}$.

Remark: For models with some deterministic component such as a time trend, one can use the above method to forecast the stochastic components, then adjust the forecasts by adding the values of the deterministic component. Of course, only stochastic component has forecast error, conditional on the given model.

B. Some simple models: We next consider the forecasts of some simple ARIMA models. Results of these simple models are informative.

a. AR(1) Model. $Z_t - \phi Z_{t-1} = a_t$. Here the π-weights are $\pi_1 = \phi$ and $\pi_j = 0$ for $j > 1$, and the ψ-weights are $\psi_i = \phi^i$. From the model and the above result, we have

$$\hat{Z}_T(\ell) = \phi^\ell Z_T, \quad e_T(\ell) = a_{T+\ell} + \phi a_{T+\ell-1} + \cdots + \phi^{\ell-1} a_{T+1}, \quad \mathrm{Var}[e_T(\ell)] = (1 + \phi^2 + \cdots + \phi^{2(\ell-1)})\sigma_a^2.$$

Clearly, the forecast $\hat{Z}_T(\ell) \to 0$ as $\ell \to \infty$. In general, $\hat{Z}_T(\ell)$ goes to the mean of Z_t as ℓ goes to infinity. Since $|\phi| < 1$, the variance of forecast error converges to $\frac{1}{1-\phi^2}\sigma_a^2$, which is the variance of Z_t. In summary, for stationary AR(1) series, the serial correlation decays exponentially to zero, implying that the current value Z_T has essentially no information about the remote future observation $Z_{T+\ell}$ for large ℓ. Therefore, the long-term forecast is the marginal distribution of Z_t. Of course, the short-term forecast is different from the marginal distribution of Z_t.

b. MA(1) Model: $Z_t = a_t - \theta a_{t-1}$ with $|\theta| < 1$. It is easy to see that for this simple model

$$\hat{Z}_T(\ell) = \begin{cases} -\theta a_T & \text{for } \ell = 1 \\ 0 & \text{for } \ell > 1, \end{cases} \qquad e_T(\ell) = \begin{cases} a_{T+1} & \text{for } \ell = 1 \\ a_{T+\ell} - \theta a_{T+\ell-1} & \text{for } \ell > 1, \end{cases}$$

$$\mathrm{Var}[e_T(\ell)] = \begin{cases} \sigma_a^2 & \text{for } \ell = 1 \\ (1+\theta^2)\sigma_a^2 & \text{for } \ell > 1. \end{cases}$$

Thus, only the 1-step ahead forecast is different from the marginal distribution of Z_t. This is obvious as the memory of an MA(1) process vanishes after 1 time period.

c. ARMA(2,1) Model: $Z_t - \phi_1 Z_{t-1} - \phi_2 Z_{t-2} = a_t - \theta a_{t-1}$. For simplicity, I shall only give the forecasts of this example. You can easily obtain the forecast errors and their variance via the ψ-weights of Z_t. The forecasts are

$$\hat{Z}_T(1) = \phi_1 Z_T + \phi_2 Z_{T-1} - \theta a_T, \quad \hat{Z}_T(2) = \phi_1 \hat{Z}_T(1) + \phi_2 Z_T$$

$$\hat{Z}_T(\ell) = \phi_1 \hat{Z}_T(\ell - 1) + \phi_2 \hat{Z}_T(\ell - 2), \quad \text{for} \quad \ell \geq 3.$$

From the stationarity, the forecasts go to the mean of Z_t exponentially. For the short-term forecasts, the pattern depends on the roots of $(1 - \phi_1 B - \phi_2 B^2) = 0$. For instance, the forecasts will have a damped sine-cosine pattern for complex roots.

d. Random Walk: $Z_t = Z_{t-1} + a_t$. Here the ψ-weights are $\psi_i = 1$ for all i. Therefore,

$$\hat{Z}_T(\ell) = Z_t, \quad e_T(\ell) = a_{T+\ell} + \cdots + a_{T+1}, \quad \mathrm{Var}[e_T(\ell)] = \ell\sigma_a^2.$$

Thus, the forecasts form a *horizontal line* with value Z_T. The variance of forecast error diverges to infinity as ℓ increases. This makes sense as no one would trust the long-term forecasts of a random walk.

e. ARIMA(0,1,1) Model: $Z_t = Z_{t-1} + a_t - \theta a_{t-1}$. Recall that the ψ-weights of this exponential smoothing model are $\psi_i = (1-\theta)$ for all i. Therefore,

$$\hat{Z}_T(\ell) = \begin{cases} Z_T - \theta a_T & \text{for } \ell = 1 \\ \hat{Z}_T(1) & \text{for } \ell > 1, \end{cases} \qquad e_T(\ell) = \begin{cases} a_{T+1} & \text{for } \ell = 1 \\ a_{T+\ell} + (1-\theta)(a_{T+\ell-1} + \cdots + a_{T+1}) & \text{for } \ell > 1 \end{cases}$$

$$\mathrm{Var}[e_T(1)] = \sigma_a^2, \quad \mathrm{Var}[e_T(\ell)] = [1 + (\ell-1)(1-\theta)^2]\sigma_a^2, \quad \text{for} \quad \ell > 1.$$

f. ARIMA(0,2,2) Model: $Z_t = 2Z_{t-1} - Z_{t-2} + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2}$. Again, for simplicity, I shall only give the results of forecasts.

$$\hat{Z}_T(\ell) = \begin{cases} 2Z_T - Z_{T-1} - \theta_1 a_T - \theta_2 a_{T-1} & \text{for } \ell = 1 \\ 2\hat{Z}_T(1) - Z_T - \theta_2 a_T & \text{for } \ell = 2 \\ 2\hat{Z}_T(\ell-1) - \hat{Z}_T(\ell-2) & \text{for } \ell > 2. \end{cases}$$

g. Seasonal Model: $Z_t = \Phi Z_{t-4} + a_t$. Here the ψ-weights are $\psi_{4i} = \Phi^i$ and $\psi_i = 0$, otherwise. The forecasts are

$$\hat{Z}_T(\ell) = \begin{cases} \Phi^{v+1} Z_{T-3} & \text{if } \ell = 4v+1 \\ \Phi^{v+1} Z_{T-2} & \text{if } \ell = 4v+2 \\ \Phi^{v+1} Z_{T-1} & \text{if } \ell = 4v+3 \\ \Phi^{v} Z_T & \text{if } \ell = 4v \end{cases}$$

which has a damped seasonal pattern. The forecast errors and their variances are easy to compute. (Exercise!)

C. Simple Exponential Smoothing

Consider the 1-step ahead prediction at origin T. Given data $\{Z_T, Z_{T-1}, \cdots, Z_1\}$, what is the (linear) prediction of Z_{T+1}? Common sense implies that in general the most recent observations should be more relevant in such a prediction. Mathematically, we can formulate the idea as follows: (i) Suppose that the weight for Z_t is ω, i.e. initial weight, and (ii) the weight is discounted by a constant rate δ, i.e. the weight for Z_{T-1} is $\delta\omega$, that for Z_{T-2} is $\delta^2\omega$, etc., where $0 < \delta < 1$. Thus, the prediction is

$$\tilde{Z}_T(1) = \omega Z_T + \delta\omega Z_{T-1} + \delta^2\omega Z_{T-2} + \cdots + \delta^{T-1}\omega Z_1.$$

However, any decent prediction *should* not change the scale of the measurement meaning that the weights should sum to 1, i.e.

$$\omega + \delta\omega + \delta^2\omega + \cdots + \delta^{T-1}\omega = 1.$$

In other words,

$$\omega(1 + \delta + \cdots + \delta^{T-1}) = \omega\frac{1-\delta^T}{1-\delta} = 1.$$

Therefore, $\omega = (1-\delta)/(1-\delta^T)$. Consequently, $\omega \to 1-\delta$ as $T \to \infty$.
For simplicity, we assume that T is sufficiently large so that $\omega = 1-\delta$. In this case, the 1-step prediction is

$$\tilde{Z}_T(1) = (1-\delta)\sum_{i=0}^{\infty}\delta^i Z_{T-i}, \tag{1}$$

which depends only on a single parameter δ, the discount rate.

Updating

Assume that δ is known. The 1-step ahead prediction at time $T+1$ is

$$\begin{aligned} \tilde{Z}_{T+1}(1) &= (1-\delta)[Z_{T+1} + \delta Z_T + \delta^2 Z_{T-1} + \cdots] \\ &= (1-\delta)Z_{T+1} + \delta(1-\delta)[Z_T + \delta Z_{T-1} + \delta^2 Z_{T-2} + \cdots] \\ &= (1-\delta)Z_{T+1} + \delta\tilde{Z}_T(1). \end{aligned} \tag{2}$$

This result says that given the "old" prediction $Z_T(1)$ and the new observation Z_{T+1}, the new prediction is $(1-\delta)Z_{T+1} + \delta\tilde{Z}_T(1)$, which is a weighted average of the old prediction and the new data point. The original forecast is discounted by δ.

Estimation

The prediction error at the forecast origin T is $e_T(1) = Z_{T+1} - \tilde{Z}_T(1)$, which is a function of δ. Consider an estimation period for T from t_0 to t_1. The parameter δ can be obtained by minimizing the sum of squared errors of prediction, i.e.

$$\hat{\delta} = \arg\min_{0<\delta<1}\sum_{T=t_0}^{t_1} e_T^2(1).$$

In this class, we use maximum likelihood estimate via ARIMA models.

Relation to ARIMA models

The time series Z_t can be written as

$$Z_t = \tilde{Z}_{t-1}(1) + a_t, \tag{3}$$

where $\tilde{Z}_{t-1}(1)$ is the forecast of simple exponential smoothing at origin $t-1$ and a_t is the forecast error. Similarly,

$$Z_{t-1} = \tilde{Z}_{t-2}(1) + a_{t-1}. \tag{4}$$

Using the updating formula, $\tilde{Z}_{t-1}(1) = (1-\delta)Z_{t-1} + \delta\tilde{Z}_{t-2}(1)$, we have $\tilde{Z}_{t-1}(1) - \delta\tilde{Z}_{t-2}(1) = (1-\delta)Z_{t-1}$. Multiplying Eq.(4) by δ, subtracting it from Eq. (3), and using the above identity, we have

$$Z_t - \delta Z_{t-1} = (1-\delta)Z_{t-1} + a_t - \delta a_{t-1}.$$

Consequently,

$$Z_t - Z_{t-1} = a_t - \delta a_{t-1},$$

which is an ARIMA(0,1,1) model with $\theta = \delta$. Thus, simple exponential smoothing model is an ARIMA(0,1,1) model with the constraint that $0 < \theta < 1$.

D. Combining Forecasts

Again, consider the 1-step ahead forecast of Z_{T+1} at the origin T. Suppose that there are m forecasting methods available and they produce *unbiased* forecasts $Z_{T,j}(1)$ for $j = 1, \cdots, m$. By unbiased forecast we mean that the expectation of the associated forecast error is zero. Empirical experience shows that a linear combination of these m forecasts often performs better in the mean squared error sense than the individual forecast. For instance, the simple average $Z_T(1) = \frac{1}{m}\sum_{j=1}^{M} Z_{T,j}(1)$ is often used in practice; see the book by Granger and Newbold for reference. Other combined forecasts include median forecast or weighted averages. However, there is no single combined forecast that systematically outperforms the others.

Methods of combining forecasts

Method 1: Consider the case of two unbiased forecasts $Z_{T,1}(1)$ and $Z_{T,2}(1)$. Then,

$$Z_{T+1} = Z_{T,1}(1) + a_{1,T+1} = Z_{T,2}(1) + a_{2,T+1},$$

where $a_{i,T+1}$ denotes the forecast error of $Z_{T,i}(1)$ and $E(a_{i,T+1}) = 0$. Assume that the variance of $a_{i,T+1}$ is σ_i^2 and the covariance between $a_{1,T+1}$ and $a_{2,T+1}$ is σ_{12}. Let $Z_T(1) = \alpha Z_{T,1}(1) + (1-\alpha)Z_{T,2}(1)$ be an arbitrary combined forecast. The forecast error of this combined forecast is $e_{T+1} = \alpha a_{1,T+1} + (1-\alpha)a_{2,T+1}$. The variance of forecast error is

$$\mathrm{Var}(e_{T+1}) = \alpha^2\sigma_1^2 + (1-\alpha)^2\sigma_2^2 + 2\alpha(1-\alpha)\sigma_{12}.$$

Theoretically, we can find α that minimizes $\mathrm{Var}(e_{T+1})$. To this end, we consider two cases.
Case A: Assume that the two forecasts are independent so that $\sigma_{12} = 0$. In this case, $\mathrm{Var}(e_{T+1}) = \alpha^2\sigma_1^2 + (1-\alpha)^2\sigma_2^2$. Therefore, $\alpha = \frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2}$, and we have

$$\begin{aligned} Z_T(1) &= \frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2} Z_{T,1}(1) + \frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2} Z_{T,2}(1) \\ &= \frac{1/\sigma_1^2}{1/\sigma_1^2 + 1/\sigma_2^2} Z_{T,1}(1) + \frac{1/\sigma_2^2}{1/\sigma_1^2 + 1/\sigma_2^2} Z_{T,2}(1) \\ &= \frac{p_1}{p_1+p_2} Z_{T,1}(1) + \frac{p_2}{p_1+p_2} Z_{T,2}(1), \end{aligned}$$

where $p_i = \sigma_i^{-2}$ is called the precision of the forecast $Z_{T,i}(1)$. In other words, the weight is determined by the relative precision of the individual forecast. This is appealing. If p_1 is much larger than p_2, i.e. σ_1^2 is much smaller than σ_2^2, then $Z_{T,1}(1)$ should be more reliable and this is shown by the heavier weight $p_1/(p_1+p_2)$. In general, if there are m unbiased forecasts available and they are all uncorrelated with precisions p_i, then the optimal forecast is

$$Z_T(1) = \sum_{i=1}^{m} \frac{p_i}{\sum_{j=1}^{m} p_j} Z_{T,i}(1).$$

Case B: $\sigma_{12} \neq 0$. In this case, the optimal α is

$$\alpha = \frac{\sigma_2^2 + \sigma_{12}}{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}}.$$

Theoretically, the variance of this combined forecast is not greater than $\min\{\sigma_1^2, \sigma_2^2\}$. In this sense, one should combine the forecasts. However, σ_1^2, σ_2^2 and σ_{12} are unknown. Using their sample estimates may affect the performance of combined forecast.

Method 2: In practice, the forecasts are often correlated because they are typically based on similar information. The combined forecast of Case B of Method 1 becomes much more involved. A practice procedure is to use multiple linear regression. Again, consider the case of two forecast methods and 1-step ahead forecasts. Suppose that there is a *forecasting* period available for T from t_0 to t_1. Let $Z_{T,i}(1)$ be the 1-step ahead forecast of method i at forecast origin T. Consider the multiple linear regression

$$Z_{T+1} = \beta_0 + \beta_1 Z_{T,1}(1) + \beta_2 Z_{T,2}(1) + e_t, \quad T = t_0, \cdots, t_1,$$

that relates the forecasts to their target Z_{T+1}. Estimate the above multiple linear regression by the ordinary least squares method. The estimate $\hat{\beta}_i$ for $i = 0, 1, 2$ are then used to obtain the 1-step ahead combined forecast for origin $T > t_1$. That is,

$$\widehat{Z}_T(1) = \hat{\beta}_0 + \hat{\beta}_1 Z_{T,1}(1) + \hat{\beta}_2 Z_{T,2}, \quad T > t_1.$$

The constant term β_0 is used to handle any bias in the individual forecasts. Obviously, adopting this regression approach means that we assume that the weights continue to hold for $T > t_1$. This may not be true in real application. Consequently, optimal combined forecast may not exist. This regression approach applies to more than two individual forecasts.

In practice, the estimates $\hat{\beta}_i$ may depend of the forecast horizon ℓ. In addition, the weights $\hat{\beta}_i$ may also depend on other available information. For instance, to forecast unemployment rate Z_t, the weights $\hat{\beta}_i$ of the individual forecasts may be a function of the growth rate of Gross National Product.

E. Forecast Evaluation

In general, it is hard to evaluate the performance of different forecasting methods. A common practice is to use out-of-sample forecasts. One divides the data into estimation subsample and forecasting subsample. Competing models are estimated using data in the estimation subsample. The fitted models are then applied to obtain forecasts of the observations in the forecasting subsample. In most studies, various forecast horizons are used to compare performance of different forecasting models.

Two general criteria are often used to evaluate the preformance of a fitted model in the forecasting subsample. The first criterion is to calculate mean squared of forecast errors (MSFE). A model with smaller MSFE is judged to be a *better* forecasting model. This criterion has its share of weakness, however. It may depend on the choice of forecasting subsample, and different forecast horizons may result in different conclusions. Furthermore, the criterion is sensitive to outliers in the evaluation period.

Another criterion is based on density function of the forecast errors. Consider the 1-step ahead forecasts. If the entertained model is adequate for the time series under study, then its 1-step ahead forecast errors should follow the distribution of a_{T+1}. Let $F(x)$ be the cumulative distribution function of a_{T+1} and let e_{T+i} be the 1-step ahead forecast error, where $i = 1, \ldots, h$ with h is the number of data points in the forecasting subsample. Consider the transformation

$$\epsilon_i = F(e_{T+i}), \quad i = 1, \ldots, h.$$

Under the assumption that the fitted model is the true model, then $\{\epsilon_i\}_{i=1}^{h}$ should form a random sample from a Uniform [0,1] distribution. One can make use of this property to test the validity of the fitted model. See, for instance, Diebold, Gunther, and Tay (1998, *International Economic Review*) and tay and Wallis (2000, *Journal of Forecasting*) and the references therein.

F. Model Uncertainty

Finally, in application, the uncertainty in the model used and the uncertainty in parameter estimates of a specified model are also important. For a given forecast model, parameter uncertainty can be handled by using the predictive distribution of a forecast. One approach is to use MCMC method in estimation and forecasting. The approach automatically takes care of parameter uncertainty. Model uncertainty is much harder to handle. Model averaging and combined forecasts are methods that can be used to deal with model uncertainty. Consult the literature on Bayesian model averaging to obtain further details.

In what follows, we shall briefly mention two forecasting methods that can handle model uncertainty. The first method is the Bayesian forecasting through simulation. The second is an adaptive forecasting method.

F1. Bayesian forecast: A good reference of this approach is Thompson and Miller (1986, JBES, pp. 427-436). For simplicity, this approach often focuses on AR models. Implementation of this approach to ARMA and MA models can be quite involved, because of

the non-linear nature of the MA parameters. However, recent developments in Markov Chain Monte Carlo (MCMC) methods, e.g. the Gibbs sampler, have mitigated some of the difficulties in using this approach.
Consider the AR model

$$Z_t = \phi_0 + \phi_1 Z_{t-1} + \cdots + \phi_p Z_{t-p} + a_t$$

Here the unknown parameters are $\mathbf{\Phi} = (\phi_0, \phi_1, \cdots, \phi_p)'$ and σ_a^2. Often it is more convenient to reparametrize σ_a^2 in term of the "precision" parameter τ such that $\tau = \sigma_a^{-2}$. A typical Bayesian analysis is then as follows:

- Specify a joint prior distribution of the parameters $\mathbf{\Phi}$ and τ.
- Assume the process Z_t is normally distributed.
- Obtain the posterior distribution of the parameters $\mathbf{\Phi}$ and τ.
- Consider the predictive distribution of the future observations on which the forecasts are based.

Let $\boldsymbol{Z} = (Z_1, Z_2, \cdots, Z_n)'$ be the observed data. Under the diffuse prior

$$f(\mathbf{\Phi}, \tau) \approx \frac{1}{\tau}$$

Zellner (1971, p. 195) showed that the posterior distribution of τ is gamma and the conditional distribution of $\mathbf{\Phi}$ given $\boldsymbol{Z}$ and τ is normal. That is,

$$f(\tau|\boldsymbol{Z}) \sim \Gamma(v/2, 2/S(\hat{\mathbf{\Phi}}))$$

and

$$f(\mathbf{\Phi}|\boldsymbol{Z}, \tau) \sim N(\hat{\mathbf{\Phi}}, [\tau \boldsymbol{X}'\boldsymbol{X}]^{-1})$$

where $v = n - p - 1$ and $\boldsymbol{X}'\boldsymbol{X}$ is the design-matrix of the AR(p) model, $\hat{\mathbf{\Phi}}$ is the least squares estimate of $\mathbf{\Phi}$ and $S(\hat{\mathbf{\Phi}})$ denotes the residual sum of squares. [These results are standard in Bayesian analysis. They are closely related to the case of *iid* normal with unknown mean and unknown variance and Jeffery's prior. See any textbook of Bayesian inference, e.g. DeGroot (1970) or Box and Tiao (1973).]
Turn to prediction. A joint ℓ-step ahead forecast requires the predictive distribution

$$f(Z_{n+1}, \cdots, Z_{n+\ell}|\boldsymbol{Z}) = \int f(Z_{n+1}, \cdots, Z_{n+\ell}, \tau, \mathbf{\Phi}|\boldsymbol{Z}) d\mathbf{\Phi} d\tau$$

$$= \int f(Z_{n+1}, \cdots, Z_{n+\ell}|\mathbf{\Phi}, \tau, \boldsymbol{Z}) f(\mathbf{\Phi}|\tau, \boldsymbol{Z}) f(\tau|\boldsymbol{Z}) d\mathbf{\Phi} d\tau \qquad (5)$$

where the integration is over the region $-\infty < \phi_i < \infty$ and $\tau > 0$.

For an AR(p) model with $\ell = 1$, the corresponding predictive distribution is a Student t with $n - p - 1$ degrees of freedom. For general ℓ, we have

$$f(Z_{n+1}, \cdots, Z_{n+\ell}|\boldsymbol{Z}) = \prod_{i=1}^{\ell} f(Z_{n+i}|\boldsymbol{Z}, Z_{n+1}, \cdots, Z_{n+\ell-1}).$$

Unfortunately, products of t-distribution do not corresponding to a closed form distribution. Thus, to use the above predictive distribution, one often tries to match certain moments in order to produce point forecasts and forecast intervals.

On the other hand, consider the equation in (5). The last two term of the integrand are gamma and normal, respectively, whereas the first term of the integrand is the density of a sequence of ℓ random variables of an AR(p) model. It is, therefore, easy to "simulate" the predictive distribution.

A Simulation Procedure:

1. Choose a value of τ from the gamma distribution $\Gamma(v/2, 2/S(\hat{\boldsymbol{\Phi}}))$.
2. Choose a set of parameter $\boldsymbol{\Phi}$ from the conditional distribution $N(\hat{\boldsymbol{\Phi}}, [\tau \boldsymbol{X}'\boldsymbol{X}]^{-1})$.
3. Simulate a path of $Z_{n+1}, \cdots, Z_{n+\ell}$ by using the AR(p) model, i.e., draw $a_{n+1}, \cdots, a_{n+\ell}$ from $N(0, \tau^{-1})$ and use the parameters chosen in Steps 1 and 2.
4. Repeat Steps 1-3 for many times and use the collection of the paths to make forecasts and forecast intervals.

For illustration, see the analysis of U.S. quarterly unemployment rate in Thompson and Miller (1986).

F2. Adaptive Forecasting: This approach is motivated by the idea that **ALL** statistical models are wrong. Thus, the principal of maximum likelihood is not applicable (strictly speaking). Thus, instead of selecting a most plausible model for a given data set, one simply entertains a "simple" model such as the exponential smoothing ARIMA(0,1,1) model or an ARMA(1,1) model for a non-seasonal time series. In practice, the selection of a model can be governed by theoretical as well as practical consideration.

Denote the unknown parameters of the entertained simple model by $\boldsymbol{\theta}$. Suppose that one is interested in the ℓ-step ahead forecasts. Then the parameter $\boldsymbol{\theta}$ is estimated by minimizing the sum of squares of the ℓ-step ahead forecast errors. Of course, for 1-step ahead forecast, this approach reduces to the traditional least squares method. However, for multi-step ahead forecasts, it is different from the least squares or maximum likelihood method.

For discussion, suppose that the entertained model is ARMA(1,1) model

$$(1 - \beta B)Z_t = (1 - \eta B)b_t$$

where b_t might not be a white noise series, as we do not believe that the model is the "true" model. For this ARMA(1,1) model, the ψ-weights are $\psi_i = \beta^{i-1}(\beta - \eta)$. Therefore, the forecast errors are

$$e_t(\ell) = \begin{cases} b_{t+1} & \text{for } \ell = 1 \\ b_{t+2} + (\beta - \eta) b_{t+1} & \text{for } \ell = 2 \\ b_{t+3} + (\beta - \eta) b_{t+2} + \beta(\beta - \eta) b_{t+1} & \text{for } \ell = 3 \\ \vdots & \vdots \end{cases}$$

The parameters β and η are then estimated, for ℓ-step ahead forecasts, by minimizing

$$S(\ell, \beta, \eta) = \sum_{t=\ell+1}^{n-\ell} [e_t(\ell)]^2.$$

This is a non-linear optimization problem which can be solved by any package, e.g. IMSL or NAG subroutine.
For application of this adaptive approach in forecasting long-memory time series, see Tiao and Tsay (1994, JoF) and the references therein.

Finally, there are many other forecasting methods. For instance, Litterman (1980, 86, JBES) discussed Bayesian forecasting of vector AR models and Garcia-Ferrer, et al (1987, JBES) discussed pooled forecasts (or shrinkage estimates).