Lecture 15
Forecasting

1

Forecasting

• A shock is often used to describe an unexpected change in a variable or in the value of the error terms at a particular time period.

• A shock is defined as the difference between expected (a forecast) and what actually happened.

• One of the most important objectives in time series analysis is to forecast its future values. It is the primary objective of ARIMA modeling:

• Two types of forecasts.

- In sample (prediction): The expected value of the RV (in-sample), given the estimates of the parameters.

- Out of sample (forecasting): The value of a future RV that is not observed by the sample.

Forecasting – Basic Concepts

• Any forecasts needs an information set, I_T. This includes data, models and/or assumptions available at time *T*. The forecasts will be conditional on I_T.

• The variable to forecast Y_{T+l} is a RV. It can be fully characterized by a pdf.

• In general, it is difficult to get the pdf for the forecast. In practice, we get a point estimate (the forecast) and a C.I.

• <u>Notation</u>:
- Forecast for *T*+*l* made at *T*: $\hat{Y}_{T+l}$, $\hat{Y}_{T+l|T}$, $\hat{Y}_T(l)$.

- *T*+*l* forecast error: $e_{T+l} = Y_{T+l} - \hat{Y}_{T+l}$

- Mean squared error (MSE): $MSE(e_{T+l}) = E[Y_{T+l} - \hat{Y}_{T+l}]^2$

Forecasting – Basic Concepts

• To get a point estimate, $\hat{Y}_{T+\ell}$, we need a cost function to judge various alternatives. This cost function is call *loss function*. Since we are working with forecast, we work with a expected loss function.

• A popular loss functions is the MSE, which is quadratic and symmetric. We can use asymmetric functions, for example, functions that penalize positive errors more than negative errors.

• If we use the MSE as the cost function, we look for $\hat{Y}_{T+l}$, which minimizes it. That is

$$\min E[e_{T+\ell}^2] = E[(Y_{T+\ell} - \hat{Y}_{T+\ell})^2] = E[Y_{T+\ell}{}^2 - 2Y_{T+\ell}\hat{Y}_{T+\ell} + \hat{Y}_{T+\ell}{}^2]$$

Then, f.o.c. implies:

$$E[-2Y_{T+\ell} + 2\hat{Y}_{T+\ell}] = 0 \qquad \Rightarrow E[Y_{T+\ell}] = \hat{Y}_{T+\ell}.$$

Forecasting – Basic Concepts

• The optimal point forecast under MSE is the (conditional) mean:

$$\hat{Y}_{T+l} = E[Y_{T+l} \,|\, I_T]$$

• Different loss functions lead to different optimal forecast. For example, for the MAE, the optimal point forecast is the median.

• The computation of $E[Y_{T+l}|\, I_T]$ depends on the distribution of $\{\varepsilon_t\}$. If $\{\varepsilon_t\} \sim$ WN, then $E[\varepsilon_{T+l}|\, I_T] = 0$, which greatly simplifies computations, especially in the linear model.

• Then, for ARMA(*p, q)* stationary process (with a Wold representation), the minimum MSE linear forecast (best linear predictor) of Y_{T+l}, conditioning on I_T is:

$$Y_{T+l} = \theta_0 + \Psi_l \varepsilon_{T+l} + \Psi_{l+1}\varepsilon_{T+l-1} + \cdots$$

Forecasting Steps for ARMA Models

• Process:

- ARIMA model $\quad Y_t = \phi Y_{t-1} + a_t$

$$\Downarrow$$

- Estimation (Evaluation in-sample) $\quad \hat{\phi}$ (Estimate of ϕ)

$$\Downarrow$$

$$\hat{Y}_t = \hat{\phi} Y_{t-1} \text{ (Prediction)}$$

$$\Downarrow$$

- Forecast (Evaluation out-of-sample) $\quad \hat{Y}_{t+1} = \hat{\phi}\hat{Y}_t$ (Forecast)

Forecasting From ARMA Models

• We observe the time series: $I_T = \{Y_1, Y_2, \ldots, Y_T\}$.

- At time T, we want to forecast: $Y_{T+1}, Y_{T+2}, \ldots, Y_{T+l}$.
- T: The forecast origin.
- l: Forecast horizon
- $\hat{Y}_T(l)$: *l-step ahead* forecast = Forecasted value Y_{T+l}

• Use the conditional expectation of $Y_{T+\ell}$, given the observed sample.

$$\hat{Y}_T(\ell) = E\left(Y_{T+\ell} \middle| Y_T, Y_{T-1}, \cdots, Y_1\right)$$

<u>Example</u>: One-step ahead forecast: $\hat{Y}_{T+1} = E[Y_{T+1} \mid Y_T, Y_{T-1}, \ldots, Y_1]$

• Forecast accuracy to be measured by MSE

$\Rightarrow$ conditional expectation, best forecast.

7

Forecasting From ARMA Models

• The stationary ARMA model for Y_t is

$$\phi_p(B)Y_t = \theta_0 + \theta_q(B)a_t$$

or

$$Y_t = \theta_0 + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q}$$

• Assume that we have data $Y_1, Y_2, \ldots, Y_T$. We want to forecast $Y_{T+\ell}$. Then,

$$Y_{T+\ell} = \theta_0 + \phi_1 Y_{T+\ell-1} + \cdots + \phi_p Y_{T+\ell-p} + \varepsilon_{T+\ell} - \theta_1 \varepsilon_{T+\ell-1} - \cdots - \theta_q \varepsilon_{T+\ell-q}$$

• Considering the Wold representation:

$$Y_{T+\ell} = \theta_0 + \Psi(B)\varepsilon_t = \theta_0 + \frac{\theta_q(B)}{\phi_p(B)}\varepsilon_t$$

$$= \theta_0 + \varepsilon_{T+\ell} + \Psi_1 \varepsilon_{T+\ell-1} + \Psi_2 \varepsilon_{T+\ell-2} + \cdots + \Psi_\ell \varepsilon_T + \cdots$$

Forecasting From ARMA Models

• Taking the expectation of $Y_{T+\ell}$, we have

$$\hat{Y}_T(\ell) = E\left(Y_{T+\ell} \middle| Y_T, Y_{T-1}, \cdots, Y_1\right)$$
$$= \Psi_\ell \varepsilon_T + \Psi_{\ell+1}\varepsilon_{T-1} + \cdots$$

where $E\left(\varepsilon_{T+j} \middle| Y_T, \cdots, Y_1\right) = \begin{cases} 0, & j > 0 \\ \varepsilon_{T+j}, & j \le 0 \end{cases}$

• Then, we define the forecast error:

$$e_T(\ell) = Y_{T+\ell} - \hat{Y}_T(\ell)$$
$$= \varepsilon_{T+\ell} + \Psi_1 \varepsilon_{T+\ell-1} + \cdots + \Psi_{\ell-1}\varepsilon_{T+1} = \sum_{i=0}^{\ell-1} \Psi_i \varepsilon_{T+\ell-i}$$

• The expectation of the forecast error: $E\left(e_T(\ell)\right) = 0$

Forecasting From ARMA Models

• The expectation of the forecast error: $E\left(e_T(\ell)\right) = 0$
$\Rightarrow$ the forecast in unbiased.

• The variance of the forecast error:

$$Var\left(e_T(\ell)\right) = Var\left(\sum_{i=0}^{\ell-1} \Psi_i \varepsilon_{T+\ell-i}\right) = \sigma^2 \sum_{i=0}^{\ell-1} \Psi_i^2$$

• <u>Example 1</u>: One-step ahead forecast ($l = 1$).

$$Y_{T+1} = \theta_0 + \varepsilon_{T+1} + \Psi_1 \varepsilon_T + \Psi_2 \varepsilon_{T-1} + \cdots$$
$$\hat{Y}_T(1) = \theta_0 + \Psi_1 \varepsilon_T + \Psi_2 \varepsilon_{T-1} + \cdots$$
$$e_T(1) = Y_{T+1} - \hat{Y}_T(1) = \varepsilon_{T+1}$$
$$Var\left(e_T(1)\right) = \sigma^2$$

Forecasting From ARMA Models

• Example 2: One-step ahead forecast (l = 2).

$$Y_{T+2} = \theta_0 + \varepsilon_{T+2} + \Psi_1 \varepsilon_{T+1} + \Psi_2 \varepsilon_T + \cdots$$
$$\hat{Y}_T(2) = \theta_0 + \Psi_2 \varepsilon_T + \cdots$$
$$e_T(2) = Y_{T+2} - \hat{Y}_T(2) = \varepsilon_{T+2} + \Psi_1 \varepsilon_{T+1}$$
$$Var\left(e_T(2)\right) = \sigma^2\left(1 + \Psi_1^2\right)$$

• Note: $\lim_{\ell \to \infty} \hat{Y}_T(\ell) - \mu = 0$

$$\lim_{\ell \to \infty} Var\left(e_T(\ell)\right) = \gamma_0 < \infty$$

• As we forecast into the future, the forecasts are not very interesting (unconditional forecasts!). That is why ARMA (or ARIMA) forecasting is useful only for short-term forecasting.

Forecasting From ARMA Model: C.I.

• A 100(1- α)% prediction interval for $Y_{T+\ell}$ (l-steps ahead) is

$$\hat{Y}_T(\ell) \pm z_{\alpha/2}\sqrt{Var\left(e_T(\ell)\right)}$$
$$\hat{Y}_T(\ell) \pm z_{\alpha/2}\sigma \sqrt{\sum_{i=0}^{\ell-1} \Psi_i^2}$$

• Example: 95% C.I. for the 2 step-ahead forecast

$$\hat{Y}_T(2) \pm 1.96\,\sigma\sqrt{1 + \Psi_1^2}$$

• When computing prediction intervals from data, we substitute estimates for parameters, giving approximate prediction intervals.

Note: Since $\hat{\Psi}$'s are RV, MSE[ε_{T+l}]=MSE[e_{T+l}]= $\sigma^2 \sum_{i=0}^{\ell-1} \Psi_i^2$

Forecasting From ARMA Model: Updating

• Suppose we have T observations at time $t=T$. We have a good ARMA model for Y_t. We obtain the forecast for Y_{T+1}, Y_{T+2}, etc.

• At $t = T + 1$, we observe Y_{T+1}. Now, we update our forecasts using the original value of Y_{T+1} and the forecasted value of it.

• The forecast error is: $e_T(\ell) = Y_{T+\ell} - \hat{Y}_T(\ell) = \sum_{i=0}^{\ell-1} \Psi_i \varepsilon_{T+\ell-i}$

We can also write this as

$$\begin{aligned} e_{T-1}(\ell+1) &= Y_{T-1+\ell+1} - \hat{Y}_{T-1}(\ell+1) \\ &= \sum_{i=0}^{\ell} \Psi_i \varepsilon_{T-1+\ell+1-i} = \sum_{i=0}^{\ell} \Psi_i \varepsilon_{T+\ell-i} \\ &= \sum_{i=0}^{\ell-1} \Psi_i \varepsilon_{T+\ell-i} + \Psi_\ell \varepsilon_T = e_T(\ell) + \Psi_\ell \varepsilon_T \end{aligned}$$

Forecasting From ARMA Model: Updating

• Then,

$$\begin{aligned} Y_{T+\ell} - \hat{Y}_{T-1}(\ell+1) &= Y_{T+\ell} - \hat{Y}_T(\ell) + \Psi_\ell \varepsilon_T \\ \hat{Y}_T(\ell) &= \hat{Y}_{T-1}(\ell+1) + \Psi_\ell \varepsilon_T \\ \hat{Y}_T(\ell) &= \hat{Y}_{T-1}(\ell+1) + \Psi_\ell \left\{ Y_T - \hat{Y}_{T-1}(1) \right\} \\ \hat{Y}_{T+1}(\ell) &= \hat{Y}_T(\ell+1) + \Psi_\ell \left\{ Y_{T+1} - \hat{Y}_T(1) \right\} \end{aligned}$$

• <u>Example</u>: $l = 1$, $T = 100$.

$$\hat{Y}_{101}(1) = \hat{Y}_{100}(2) + \Psi_1 \left\{ Y_{101} - \hat{Y}_{100}(1) \right\}$$

Forecasting From ARMA Model: Transformations

• If we use variance stabilizing transformation, after the forecasting, we need to convert the forecasts for the original series.

• For example, if we use log-transformation, then,

$$E\left[Y_{T+\ell}\,|Y_1,\cdots,Y_T\right] \geq \exp\left\{E\left[\ln\left(Y_{T+\ell}\right)|\ln\left(Y_1\right),\cdots,\ln\left(Y_T\right)\right]\right\}$$

• If $X \sim N(\mu, \sigma^2)$, then, $E\left[exp\left(X\right)\right] = exp\left(\mu + \frac{\sigma^2}{2}\right)$.

• The MSE forecast for the original series is:

$$\exp\left[\hat{Z}_n(\ell) + \frac{1}{2}Var\left(e_n(\ell)\right)\right] \quad \text{where } Z_{n+\ell} = \ln\left(Y_{n+\ell}\right)$$

$$\mu = E\left(Z_{n+\ell}|Z_1,\cdots,Z_n\right) \qquad \sigma^2 = Var\left(Z_{n+\ell}|Z_1,\cdots,Z_n\right)$$

Forecasting From ARMA Model: Remarks

• In general, we need a large *T*. Better estimates and it is possible to check for model stability and check forecasting ability of model by withholding data.

• Seasonal patterns also need large *T*. Usually, you need 4 to 5 seasons to get reasonable estimates.

• Parsimonious models are very important. Easier to compute and interpret models and forecasts. Forecasts are less sensitive to deviations between parameters and estimates.

Forecasting From Simple Models: ES

• Industrial companies, with a lot of inputs and outputs, want quick and inexpensive forecasts. Easy to fully automate.

• Exponential Smoothing Models (ES) fulfill these requirements.

• In general, these models are limited and not optimal, especially compared with Box-Jenkins methods.

• Goal of these models: Suppress the short-run fluctuation by smoothing the series. For this purpose, a weighted average of all previous values works well.

• There are many ES models. We will go over the Simple Exponential Smoothing (SES) and Holt-Winter's Exponential Smoothing (HW ES).

Forecasting From Simple Models: ES

• Observed time series: $Y_1, Y_2, \ldots, Y_T$

• The equation for the model is $S_t = \alpha Y_{t-1} + (1-\alpha) S_{t-1}$

where

- α: the smoothing parameter, $0 \leq \alpha \leq 1$
- Y_t: the value of the observation at time t
- S_t: the value of the smoothed observation at time t.

• The equation can also be written as

$$S_t = S_{t-1} + \alpha(Y_{t-1} - S_{t-1}) = S_{t-1} + \alpha(\text{forecast error})$$

• Then, the forecast is: $S_{t+1} = \alpha Y_t + (1-\alpha) S_t = S_t + \alpha(Y_t - S_t)$

That is, a simple updating equation.

Forecasting From Simple Models: ES

• Q: Why Exponential?

For the observed time series $Y_1, Y_2, \ldots, Y_n$, Y_{n+1} can be expressed as a weighted sum of previous observations.

$$\hat{Y}_t(1) = c_0 Y_t + c_1 Y_{t-1} + c_2 Y_{t-2} + \cdots$$

where c_i's are the weights.

• Giving more weights to the recent observations, we can use the geometric weights (decreasing by a constant ratio for every unit increase in lag):

$$\Rightarrow c_i = \alpha(1-\alpha)^i; \quad i = 0,1,\ldots; 0 \le \alpha \le 1.$$

• Then, $\hat{Y}_t(1) = \alpha(1-\alpha)^0 Y_t + \alpha(1-\alpha)^1 Y_{t-1} + \alpha(1-\alpha)^2 Y_{t-2} + \cdots$

$$\hat{Y}_t(1) = \alpha Y_t + (1-\alpha)\hat{Y}_{t-1}(1) \quad \Rightarrow S_{t+1} = \alpha Y_t + S_t$$

19

Forecasting From Simple Models: Selecting α

• Choose α between 0 and 1.

- If α = 1, it becomes a naive model; if α is close to 1, more weights are put on recent values. The model fully utilizes forecast errors.
- If α is close to 0, distant values are given weights comparable to recent values. Set α ≈ 0 when there are big random variations in Y_t.
- α is often selected as to minimize the MSE.

• In empirical work, 0.05 ≤ α ≤ 0.3 are used (α ≈ 1 is used rarely).

Numerical Minimization Process:

- Take different α values ranging between 0 and 1.
- Calculate 1-step-ahead forecast errors for each α.
- Calculate MSE for each case.

Choose α which has the min MSE: $e_t = Y_t - S_t \Rightarrow min \sum_{t=1}^{n} e_t^2 \Rightarrow \alpha$

20

Forecasting From Simple Models: Example

Time	Y_t	S_{t+1} (α=0.10)	$(Y_t-S_t)^2$
1	5	-	-
2	7	(0.1)5+(0.9)5=5	4
3	6	(0.1)7+(0.9)5=5.2	0.64
4	3	(0.1)6+(0.9)5.2=5.28	5.1984
5	4	(0.1)3+(0.9)5.28=5.052	1.107
		TOTAL	**10.945**

$$MSE = \frac{SSE}{n-1} = 2.74$$

• Calculate this for α=0.2, 0.3,…, 0.9, 1 and compare the MSEs. Choose α with minimum MSE

21

Forecasting From Simple Models: Remarks

• Some computer programs automatically chooses the optimal α using the search method or non-linear optimization techniques.

• Initial Value Problem

– Set S_2 to Y_1 is one method of initialization.

– Also, take the average of, say first 4 or 5 observations. Use this average as an initial value.

• This model ignores trends or seasonalities. Not very realistic. But, deterministic components, D_t, can be easily incorporated. The model that incorporates both features is called *Holt-Winter's ES*.

22

Forecasting From Simple Models: HW ES

• Now, we introduce trend (T_t) and seasonality (I_t) factors. Both can can be included as additively or multiplicatively factors.

• Details

- We use multiplicative seasonalities –i.e., Y_t/I_t– and additive trend.
- The forecast, S_t, is adjusted by the deterministic trend: $S_t + T_t$.
- The trend, T_t, is a weighted average of T_{t-1} and the change in S_t.
- The seasonality is also a weighted average of I_{t-S} and the Y_t/S_t

• Then, the model has three equations:

$$S_t = \alpha \frac{Y_{t-1}}{I_{t-s}} + (1-\alpha)(S_{t-1} - T_{t-1})$$

$$T_t = \gamma (S_t - S_{t-1}) + (1-\gamma) T_{t-1}$$

$$I_t = \delta \frac{Y_t}{S_t} + (1-\delta) I_{t-s}$$

23

Forecasting From Simple Models: HW ES

• We think of (Y_t/S_t) as capturing *seasonal effects*.

s = # of periods in the seasonal cycles

(s = 4, for quarterly data)

• We have only three parameters :

α = smoothing parameter

γ = trend coefficient

δ = seasonality coefficient

• Q: How do we determine these 3 parameters?

- Ad-hoc method: α ,γ and δ can be chosen as value between $0.02 < \alpha, \gamma, \delta < 0.2$
- Minimization of the MSE, as in SES.

24

Forecasting From Simple Models: HW ES

• h-step ahead forecast $\hat{Y}_t(h) = (S_t + hT_t)I_{t+h-s}$

Note: Seasonal factor is multiplied in the h-step ahead forecast

• Initial values for algorithm

- We need at least one complete season of data to determine the initial estimates of I_{t-s}.

- Initial values:

$$1.S_0 = \sum_{t=1}^{s} Y_t / s$$

$$2.T_0 = \frac{1}{s}\left(\frac{Y_{s+1} - Y_1}{s} + \frac{Y_{s+2} - Y_2}{s} + \cdots + \frac{Y_{s+s} - Y_s}{s}\right)$$

$$or\ T_0 = \left[\left\{\sum_{t=1}^{s} Y_t / s\right\} - \left\{\sum_{t=s+1}^{2s} Y_t / s\right\}\right] / s$$

25

Forecasting From Simple Models: HW ES

• Algorithm to compute initial values for seasonal component I_s.

Assume we have T observation and quarterly seasonality (s=4):

(1) Compute the averages of each of T years.

$$A_t = \sum_{i=1}^{4} Y_{t,i} / 4, \quad t = 1,2,\cdots,6 \quad \text{(yearly averages)}$$

(2) Divide the observations by the appropriate yearly mean: $Y_{t,i}/A_t$.

(3) I_s is formed by computing the average $Y_{t,i}/A_t$ per year:

$$I_s = \sum_{i=1}^{T} Y_{t,s} / A_t 4 \quad s = 1,2,3,4$$

26

Forecasting From Simple Models: HW ES

• Remarks

- Note that, if a computer program selects $\gamma = 0 = \delta$, this does not a lack of trend or seasonality. It implies a constant (deterministic) component.

- In this case, an ARIMA model with deterministic trend may be a more appropriate model.

- For HW ES, a seasonal weight near one implies that a non-seasonal model may be more appropriate.

- We modeled seasonalities as multplicative:

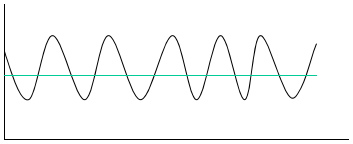
=> Multiplicative seasonality: $\text{Forecast}_t = S_t * I_{t-s}$.

- But, seasonal components can also be additive. For example, during the month of December sales at a particular store increase by $X every year. In this case, we just $X to the December forecast.

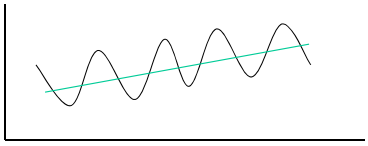
=> Additive seasonality: $\text{Forecast}_t = S_t + I_{t-s}$.

27

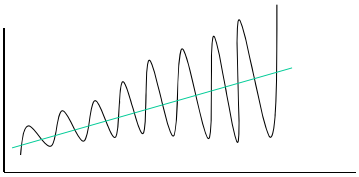
ES Models – Different Types



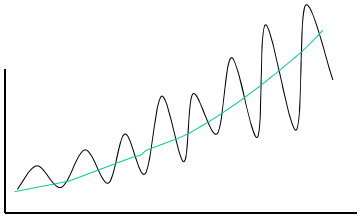
1. No trend and additive seasonal variability (1,0)



2. Additive seasonal variability with an additive trend (1,1)

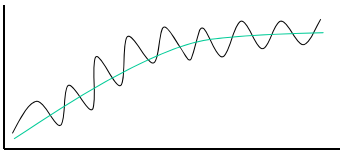


3. Multiplicative seasonal variability with an additive trend (2,1)

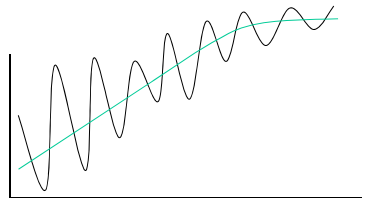


4. Multiplicative seasonal variability with a multiplicative trend (2,2)

ES Models – Different Types



5. Dampened trend with additive seasonal variability (1,1)



6. Multiplicative seasonal variability and dampened trend (2,2)

- Select the type of model to fit based on the presence of
- Trend – additive or multiplicative, dampened or not
- Seasonal variability – additive or multiplicative

Evaluation of forecasts

- Summary of loss functions of out-of-sample forecast accuracy:

Mean Error = $\frac{1}{m}\sum_{i=T+1}^{T+m}(\hat{y}_i - y_i) = \frac{1}{m}\sum_{i=T+1}^{T+m} e_i$

Mean Absolute Error (MAE) = $\frac{1}{m}\sum_{i=T+1}^{T+m}|\hat{y}_i - y_i| = \frac{1}{m}\sum_{i=T+1}^{T+m}|e_i|$

Mean Squared Error (MSE) = $\frac{1}{m}\sum_{i=T+1}^{T+m}(\hat{y}_i - y_i)^2 = \frac{1}{m}\sum_{i=T+1}^{T+m} e_i^2$

Root Mean Square Error (RMSE)= $\sqrt{\frac{1}{m}\sum_{i=T+1}^{T+m}(\hat{y}_i - y_i)^2} = \sqrt{\frac{1}{m}\sum_{i=T+1}^{T+m} e_i^2}$

Theil's U-stat = $U = \frac{\sqrt{\frac{1}{m}\sum_{i=T+1}^{T+m} e_i^2}}{\sqrt{\frac{1}{T}\sum_{i=1}^{T} y_i^2}}$

Evaluation of forecasts – DM Test

• To determine if one model predicts better than another, we define the loss differential between two forecasts:

$$d_t = g(e_t^{M1}) - g(e_t^{M2}).$$

where g(.) is the forecasting loss function. M1 and M2 are two competing sets of forecasts –could be from models or something else.

• We only need $\{e_t^{M1}\}$ & $\{e_t^{M2}\}$, not the structure of M1 or M2. In this sense, this approach is "*model-free*."

• Typical (symmetric) loss functions: $g(e_t) = e_t^2$ & $g(e_t) = |e_t|$.

• But other g(.)'s can be used: $g(e_t) = \exp(\lambda e_t^2) - \lambda e_t^2$ $(\lambda>0)$.

Evaluation of forecasts – DM Test

• Then, we test the null hypotheses of equal predictive accuracy:

H_0: $E[d_t] = 0$ vs.

H_1: $E[d_t] = \mu \neq 0$.

- Diebold and Mariano (1995) assume $\{e_t^{M1}\}$ & $\{e_t^{M2}\}$ is covariance stationarity and other regularity conditions (finite $Var[d_t]$, independence of forecasts after *l* periods) needed to apply CLT. Then,

$$\frac{\bar{d} - \mu}{\sqrt{Var[\bar{d}]/T}} \xrightarrow{d} N(0,1), \qquad \bar{d} = \frac{1}{m}\sum_{i=T+1}^{T+m} d_i$$

• Then, the DM test is a simple *z-test*:

$$DM = \frac{\bar{d}}{\sqrt{\hat{V}ar[\bar{d}]/T}} \xrightarrow{d} N(0,1)$$

Evaluation of forecasts – DM Test

where $\hat{Var}[\bar{d}]$ is a consistent estimator of the variance, usually based on sample autocovariances of d_t:

$$\hat{Var}[\bar{d}] = \gamma(0) + 2\sum_{j=k}^{l} \gamma(j)$$

• Assume the l-step ahead forecast errors have zero autocorrelations at order l. Harvey et al. (1998) propose a small-sample modification of the DM test:

$$DM^* = DM/\{[T + 1 - 2l + l(l-1)/T]/T\}1/2 \sim t_{T-l}.$$

• If ARCH is suspected, replace l with $[0.5\sqrt{(T)}]+l$. ([.]=integer part).

•<u>Note</u>: If $\{e_t^{M1}\}$ & $\{e_t^{M2}\}$ are perfectly correlated, the numerator and denominator of the DM test are both converging to 0 as $T \to \infty$.

$\Rightarrow$ Avoid DM test when this situation is suspected (say, two nested models.) Though, in small samples, it is OK.

Evaluation of forecasts – DM Test

Example: Code in R

```
 dm.test <- function (e1, e2, h = 1, power = 2) {
d <- c(abs(e1))^power - c(abs(e2))^power
  d.cov <- acf(d, na.action = na.omit, lag.max = h - 1, type = "covariance", plot = FALSE)$acf[, , 1]
  d.var <- sum(c(d.cov[1], 2 * d.cov[-1]))/length(d)
  dv <- d.var#max(1e-8,d.var)
  if(dv > 0)
    STATISTIC <- mean(d, na.rm = TRUE) / sqrt(dv)
  else if(h==1)
    stop("Variance of DM statistic is zero")
  else
  {
    warning("Variance is negative, using horizon h=1")
    return(dm.test(e1,e2,alternative,h=1,power))
  }
    n <- length(d)
  k <- ((n + 1 - 2*h + (h/n) * (h-1))/n)^(1/2)
  STATISTIC <- STATISTIC * k
  names(STATISTIC) <- "DM"
}
```

Evaluation of forecasts – DM Test: Remarks

• The DM tests is routinely used. Its "model-free" approach has appeal. There are model-dependent tests, see West (1996), Clark and McCracken (2001), and, more recent, Clark and McCracken (2011), with more complicated asymptotic distributions.

• The loss function does not need to be symmetric (like MSE).

• The DM test is based on the notion of unconditional –i.e., on average over the whole sample- expected loss.

• Following Morgan, Granger and Newbold (1977), the DM statistic can be calculated by regression of d_t, on an intercept, using NW SE. But, we can also condition on variables that may explain d_t. We move from an unconditional to a conditional expected loss perspective.

Evaluation of forecasts – Conditional Test

• Giacomini and White (2006) present a general framework for out-of-sample predictive ability testing, characterised by the formulation of tests (such as tests for equality of forecasts) based on conditional expected loss. Now,

$$E[\hat{d}_t \mid F_{t-1}] = 0 \;\Rightarrow\; E[h_{t-1}\, \hat{d}_t] = 0,$$

where h_{t-1} is a F_{t-1} measurable function of dimension q.

<u>Note</u>: G&W (2006) also differs from the standard approach to testing for predictive ability in that it compares forecasting methods (estimation + model) rather than forecasting models.

• The test becomes a Wald test, with an asymptotic $\chi^2(q)$ distribution.

Combination of Forecasts

• Idea – from Bates & Granger (*Operations Research Quarterly,* 1969):

- We have different forecasts from R models:

$$\hat{Y}_T^{M1}(\ell), \hat{Y}_T^{M2}(\ell), \ldots \hat{Y}_T^{MR}(\ell)$$

• Q: Why not combine them?

$$\hat{Y}_T^{Comb}(\ell) = \omega_{M1}\hat{Y}_T^{M1}(\ell) + \omega_{M2}\hat{Y}_T^{M2}(\ell) + \ldots + \omega_{MR}\hat{Y}_T^{MR}(\ell)$$

• Very common practice in economics, finance and politics, reported by the press as "consensus forecast." Usually, as a simple average.

• Q: Advantage? Lower forecast variance. Diversification argument.

Intuition: Individual forecasts are each based on partial information sets (say, private information) or models.

37

Combination of Forecasts – Optimal Weights

• The variance of the forecasts is:

$$Var[\hat{Y}_T^{Comb}(\ell)] = \sum_{j=1}^{R}(\omega_{Mj})^2 Var[\hat{Y}_T^{Mj}(\ell)] + 2\sum_{j=1}^{R}\sum_{i=j+1}^{R}\omega_{Mj}\omega_{Mi}Co\,\mathrm{var}[\hat{Y}_T^{Mj}(\ell)\hat{Y}_T^{Mi}(\ell)]$$

Note: Ideally, we would like to have negatively correlated forecasts.

• Assuming unbiased forecasts and uncorrelated errors,

$$Var[\hat{Y}_T^{Comb}(\ell)] = \sum_{j=1}^{R}(\omega_{Mj})^2\sigma_j^2$$

Example: Simple average: $\omega_j = 1/R$. Then, $Var[\hat{Y}_T^{Comb}(\ell)] = 1/R^2\sum_{j=1}^{R}\sigma_j^2$.

• We can derived optimal weights –i,e., ω_j's that minimize the variance of the forecast. Under the uncorrelated assumption:

$$\omega_{Mj}^* = \sigma_j^{-2} \Big/ \sum_{j=1}^{R}\sigma_j^{-2}$$

Combination of Forecasts – Optimal Weights

• Under the uncorrelated assumption:

$$\omega_{Mj}^* = \sigma_j^{-2} \Big/ \sum_{j=1}^{R} \sigma_j^{-2}$$

• The ω_j*'s are inversely proportional to their variances.

• In general, forecasts are biased and correlated. The correlations will appear in the above formula for the optimal weights. For the two forecasts case:

$$\omega_{Mj}^* = (\sigma_1^2 - \sigma_{12}) \big/ (\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}) = (\sigma_1^2 - \rho\sigma_1\sigma_2) \big/ (\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2)$$

• We do not observe the forecast variances and covariances, nor the biases. We need a history of forecasts to estimate the optimal weights.

Combination of Forecasts: Regression Weights

• In general, forecasts are biased and correlated. The correlations will appear in the above formula for the optimal weights. Ideally, we would like to have negatively correlated forecasts.

• Granger and Ramanathan(1984) used a regression method to combine forecasts.

- Regress the actual value on the forecasts. The estimated coefficients are the weights.

$$y_{T+\ell} = \beta_1 \hat{Y}_T^{M\,1}(\ell) + \beta_2 \hat{Y}_T^{M\,2}(\ell) + \ldots + \beta_R \hat{Y}_T^{MR}(\ell) + \varepsilon_{T+\ell}$$

• Should use a constrained regression

– Omit the constant

– Enforce non-negative coefficients.

– Constrain coefficients to sum to one

40

Combination of Forecasts: Regression Weights

• Remarks:

- To get weights, we do not include a constant. Here, we are assuming unbiased forecasts. If the forecasts are biased, we include a constant.

- To account for potential correlation of errors, Coulson and Robbins (1993) suggests allowing for ARMA residuals or include y_{T+l-1} in the regression.

- Time varying weights are also possible.

• Should weights matter? Two views:

- Simple averages outperform more complicated combination techniques --Stock and Watson (1999) and Fildes and Ord (2002).

- Sampling variability may affect weight estimates to the extent that the combination has a larger MSFE --Harvey and Newbold (2005).

- Bayesian techniques, using priors, may help in the latter situation.

41

Combination of Forecasts: Bayesian Weights

• In our discussion of model selection, we mentioned that the BIC is consistent. That means, the probability that a model is true, given the data is proportional to BIC:

$$P(M_j \mid \text{data}) \propto \exp(-BIC_j/2).$$

• Based on this, we use the BIC of different models to derive weights. This is a simplified form of *Bayesian model averaging* (BMA).

• Easy calculation of weights. Let BIC* be the smallest BIC among the R models considered. Define $\Delta BIC_{Mj} = BIC_{Mj} - BIC^*$.

Then,

$$\omega_{Mj}^* = \exp(-\Delta BIC_{Mj}/2)$$

$$\omega_{Mj} = \omega_{Mj}^* / \sum_{j=1}^{R} \omega_{Mj}^*$$

42

Combination of Forecasts: Bayesian Weights

• Steps:

(1) Compute BIC for the R different models.

(2) Find best-fitting BIC*.

(3) Compute ΔBIC & exp(-ΔBIC/2).

(4) Add up all values and re-normalize.

• BMA puts the most weight on the model with the smallest BIC.

• Some authors have suggested replacing BIC with AIC in the weight formula –i.e., $\omega_j \propto \exp(-AIC_j/2)$.

- There is no clear theory for this formula. It is simple and works well in practice.

- This method is called weighted AIC (WAIC).

43

Combination of Forecasts: Bayesian Weights

• Q: Does it make a difference the criteria used? Two situations:

(1) The selection criterion (AIC, BIC) are close for competing models. Then, it is difficult to select one over the other.

- WAIC and BMA will produce similar weights.

(2) The selection criterion are different.

- WAIC and BMA will produce different weights.

- They will give zero weight if the difference is large, say, above 10.

Q: Which one to use?

- Not clear. WAIC works well in practice.

<u>General finding</u>: Simple averaging works well, but it is not optimal. A combination beats the lowest criteria used.

44

Forecasting: Final Comments

• Since Bates and Granger (1969) and Granger and Ramanathan (1984), combination weights have generally been chosen to minimize a symmetric, squared-error loss function.

• But, asymmetric loss functions can also be used. Elliot and Timmermann (2004) allow for general loss functions (and distributions). They find that the optimal weights depend on higher order moments, such a skewness.

• It is also possible to forecast quantiles and combine them. Testing of quantile forecasts can be based on the general approach of G&W (2006). Giacomini and Komunjer (2005) present an application.

45