Module 9 - Selection of Models

MATH1307 Forecasting

Prepared by: Dr. Haydar Demirhan based on the textbook by Hyndman et al., Forecasting with Exponential Smoothing: The State Space Approach. Springer, 2008.

Introduction

Selection of the most suitable model from a set of tentative models is important in terms of getting precise forecasts. From the perspective of this module, the set of tentative models can include many specific models within the general innovations state-space models. There are also a bunch of different approaches for model selection in the literature. In this module, we will focus on a subset of these approaches.

We will describe the use of information criteria for selecting among the innovations state-space models. We will consider four commonly recommended information criteria and one relatively new information criterion.

We will use the MASE to develop measures for comparing model selection procedures. These measures will be used to compare the five information criteria with each other, and the commonly applied prediction validation method for model selection using the M3 competition data (Makridakis and Hibon, 2000) and a hospital data set.

We also compare the results with the application of damped trend models for all time series. Finally, some implications of these comparisons will be given.

Information Criteria for Model Selection

In terms of forecasting, our main aim is to select the model with the best predictive ability on average. Finding the model with the smallest within-sample one-step-ahead forecast errors, or even the one with the maximum likelihood does not assure us that the model will be the best one for forecasting.

To deal with this issue, we can use an information criterion which penalizes the likelihood to compensate for the potential overfitting of data. Many different information criteria can be used for model selection under different circumstances. The general structure of information criteria is as the following:

$$IC = -2\log(likelihood) + penalty \tag{1}$$

With such a measure, the model with the minimum likelihood will not be selected as the best one. Instead, the penalty term will increase the value of IC for the models that are not good in terms of the type of penalty considered. The following table summarizes the ICs, we will consider in this module.

Criterion	Penalty	Source
AIC	$2q$	Akaike (1974)
BIC	$q\log(n)$	Schwarz (1978)
HQIC	$2q\log(\log(n))$	Hannan and Quinn (1979)
AICc	$2qn/(n-q-1)$	Sugiura (1978)
LEIC	qc	Billah et al. (2003)

where q is the number of parameters and free states in $\hat{X}_0$, and n is the number of observations. Nearly, all of the penalties are functions of the number of parameters and the sample size. We compute the IC for each model or a bunch if ICs for each model and choose the model with the minimum IC. We do not expect that ICs will agree on one particular model, but in most cases, a subset of considered ICs agree on one of the tentative models. Notice that, this process will not give us the best model ever, but it will indicate a reasonable model for forecasting.

The *Akaike Information Criterion* (AIC) penalize the likelihood by the penalty of twice the number of parameters in the model. This criterion tends to choose more parsimonious models. The AIC has been criticized because it is inconsistent and tends to overfit models.

The *Schwarz Bayesian information criterion* (BIC) uses $q\log(n)$ as the penalty function. So, it considers the number of parameters and the sample size simultaneously. We need to have a sample large enough to feed the parameter estimation process. So, we need more information as the number of parameters increases. BIC considers this debate. BIC provides a Bayesian solution to the problem of model identification in the sense that the BIC is asymptotically minimized at the model order having the highest posterior probability. In our case, the *order* is the number of parameters and free states. As the sample size increases, the BIC is minimized at the true order with a probability that approaches unity. Thus, it is order-consistent.

The *Hannan–Quinn information criterion* (HQIC) uses double logarithm in the penalty function $2q\log(\log(n))$. The purpose of using the logarithm twice in a nested manner is to make the criterion more order consistent. But in this case, it loses efficiency.

The *bias corrected* AIC (AICc) is a corrected version of AIC. While the BIC and HQIC are order-consistent, they are not asymptotically efficient like the AIC. But AIC has a negative bias. The AICc is an asymptotically efficient information criterion that does an approximate correction for this negative bias and has been shown to provide better model order choices for small samples.

The *linear empirical information criterion* (LEIC) is based on the penalty function qc, where c is estimated empirically for an ensemble of N similar time series with M competing models. The procedure for estimating c requires that a specified number of time periods H be withheld from each time series. The forecasting errors for these withheld time periods are used to compare the competing models and determine a value for c.

For the model selection purposes, the function `ets()` utilizes one of AIC, BIC, or AICc in the auto-selection mode over value of argument `ic=c("aicc","aic","bic")`. The HQIC can be computed easily using the output of `ets()` function.

Computation of AIC, BIC, AICc, and HQIC are illustrated using the `ausgdp` data below.

```
data("ausgdp")
fit.ausgdp <- ets(ausgdp,"AAN" , damped=FALSE , upper=rep(1,4))
n = length(ausgdp) # The number of observations
q = (length(fit.ausgdp$par)+1) # The number of parameters and free
  states
AIC = -2*fit.ausgdp$loglik + 2*q
BIC = -2*fit.ausgdp$loglik + q*log(n)
AICc = -2*fit.ausgdp$loglik + 2*q*n/(n-q-1)
HQIC = -2*fit.ausgdp$loglik+ 2*q*log(log(n))
# Put the results into a data frame
IC = data.frame(AIC , BIC , AICc , HQIC)
IC
```

```
##         AIC      BIC     AICc     HQIC
## 1 1201.463 1214.828 1202.057 1206.881
```

```
# Compare with those calculated by the ets() function
IC.ets = data.frame(fit.ausgdp$aic , fit.ausgdp$bic , fit.ausgdp$a
  icc)
colnames(IC.ets) = c("AIC" , "BIC" , "AICc")
IC.ets
```

```
##        AIC      BIC     AICc
## 1 1201.463 1214.828 1202.057
```

The following function computed information criteria along with the MASE measure for the bunch of time series and models.

```
GoFVals = function(data, H, models){
  M = length(models) # The number of competing models
  N = length(data) # The number of considered time series
  fit.models = list()
  series = array(NA, N*M)
  FittedModels = array(NA, N*M)
  AIC = array(NA, N*M)
  AICc = array(NA, N*M)
  BIC = array(NA, N*M)
  HQIC = array(NA, N*M)
  MASE = array(NA, N*M)
  mean.MASE = array(NA, N)
  median.MASE = array(NA, N)
  GoF = data.frame(series, FittedModels, AIC, AICc, BIC, HQIC, MAS
  E)
  count = 0
  for ( j in 1:N){
    sum.MASE = 0
    sample.median = array(NA, M)
    for ( i in 1: M){
      count = count + 1
      fit.models[[count]] = ets(data[[j]], model = models[i])
      GoF$AIC[count] = fit.models[[count]]$aic
      GoF$AICc[count] = fit.models[[count]]$aicc
      GoF$BIC[count] = fit.models[[count]]$bic
      q = length(fit.models[[count]]$par)
      GoF$HQIC[count] = -2*fit.models[[count]]$loglik+ 2*q*log(log
  (length(data[[j]])))
      GoF$MASE[count] = accuracy(fit.models[[count]])[6]
      sum.MASE = sum.MASE + GoF$MASE[count]
      sample.median[i] = GoF$MASE[count]
      GoF$series[count] = j
      GoF$FittedModels[count] = models[i]
    }
    mean.MASE[j] = sum.MASE / N
    median.MASE[j] = median(sample.median)
  }
  return(list(GoF = GoF, mean.MASE = mean.MASE, median.MASE = medi
  an.MASE))
}
```

Linear empirical information criterion

Let's have N time series shown by $\{Y_t^{(j)}\}, j = 1, \ldots, N$, and the number of observations in each series be n_j. To calculate LEIC, each series should be divided into two parts like training and validation parts. Let $n_j^* = n_j - H$ be the number of observations in the training part and the validation segment consists of the last H observations. Let n be the length of the longest series within the considered set of series, i.e. $n = \max\{n_j^*; j = 1, \ldots, N\}$. The following algorithms are used to find the value of LEIC.

Model Estimation

1. For each of the N series, use the first n_j^* observations to estimate the parameters and initial state vector in each of the M competing models by maximum likelihood estimation.
2. Record the maximized log-likelihoods for all estimated models.

Penalty estimation

1. For each trial value of c do the following
   a. For each time series select a model with the minimum LEIC using

$$IC = -2 \log \text{likelihood} + q\zeta(n) \tag{2}$$

   and $\zeta(n) =$ the trial value for c.
   b. For each forecast horizon h, $h = 1, \ldots, H$, and time series compute the absolute scaled error ASE:

$$ASE(h, c, j) = \frac{|y_{n_j^*+h}^{(j)} - \hat{y}_{n_j^*}^{(c,j)}(h)|}{MAE_j}, \tag{3}$$

   where $MAE_j = (n_j^* - 1)^{-1} \sum_{t=2}^{n_j^*} |y_t^{(j)} - y_{t-1}^{(j)}|$, and $\hat{y}_{n_j^*}^{(c,j)}(h)$ is the h-step-ahead forecast using the model selected for the jth series.
2. For each value of c and for each forecast horizon h, calculate the mean absolute scaled error $MASE$ across the N time series to obtain

$$MASE(h, c) = \frac{1}{N} \sum_{j=1}^{N} ASE(h, c, j). \tag{4}$$

3. Select a value of $c^{(h)}$ by minimizing the $MASE(h, c)$ over the grid of c values. Thus, a value of $c^{(h)}$ is selected for each forecast horizon h, $h = 1, \ldots, H$.
4. Compute the final value of c by averaging the H values of $c^{(h)}$:

$$c = \frac{1}{H} \sum_{h=1} H c^{(h)}. \tag{5}$$

Observe that $MASE(H, i, j)$ is an average across H forecast horizons for a specified model i and time series j, while $MASE(h, c)$ is an average across N time series for a fixed forecasting horizon h and model determined by trial value c. Also, it is important

to re-estimate the parameters and initial state vector for the selected model by using all of the n_j values.

This algorithm is implemented with the following function:

```
LEIC = function(data, H, models){
  M = length(models) # The number of competing models
  N = length(data) # The number of considered time series
  n = array(NA, N) # Array to hold the length of each series
  for ( j in 1:N){ # Find the length of each series
    n[j] = length(data[[j]])
  }
  n.max = max(n)
  n.star = n - H

  # Fit the models
  fit.models = list()
  count = 0
  for ( j in 1:N ){ # For each series of length nj*
    for ( i in 1:M){ # Fit all models
      count = count + 1
      fit.models[[count]] = ets(ts(data[[j]][1:n.star],frequency =
  frequency(data[[j]])), model = models[i])
    }
  }

  c = seq(0.15 , 2*log(n.max) , 0.05)
  ASE = array(NA, dim = c(H, length(c), N))
  MASE = array(NA, dim = c(H, length(c)))
  leic = array(NA, count)
  for ( k in 1:length(c)){
    count = 0
    for ( j in 1:N ){
      for ( i in 1:M){
        count = count + 1
        q = length(fit.models[[count]]$par) # the number of parame
  ters + the number of free states
        leic[count] = -2*fit.models[[count]]$loglik + q * c[k]
      }
    }

    best.model = fit.models[[which(leic == min(leic))]]# gives the
  order of the model with miniumum leic
    summ = 0
    for (h in 1:H){
      summ = 0
      for ( j in 1:N){
        summ.2 = 0
        for ( t in 2:n[j] ){
          summ.2 = summ.2 + abs(data[[j]][t] - data[[j]][t-1])
        }
        ASE[h, k , j] = sum( abs(data[[j]][(n.star + 1):(n.star+
  h)] - forecast(best.model, h = h)$mean) / summ.2 )
        summ = summ + ASE[h, k , j]
      }
```

```
      MASE[h, k] = summ / N
    }
  }
  ch = array(NA, H)
  for ( h in 1:H){
    ch[h] = MASE[h, min( which(min(MASE[h,]) == MASE[h, ]) )] # th
  e first min is to take minimum of minimums
  }
  c.opt = mean(ch)

  series = array(NA, N*M)
  FittedModels = array(NA, N*M)
  values = array(NA, N*M)
  leic = data.frame(series, FittedModels, values)
  count = 0
  for ( j in 1:N ){
    for ( i in 1:M){
      count = count + 1
      q = length(fit.models[[count]]$par) # the number of paramete
  rs + the number of free states
      leic$series[count] = j
      leic$FittedModels[count] = models[i]
      leic$values[count] = -2*fit.models[[count]]$loglik + q * c.o
  pt
    }
  }

  return(list(leic = leic, c.opt = c.opt))
}
```

Let's illustrate the use of this function over `ausgdp` and `usgdp` series available in `expsmooth` package.

```
data("ausgdp")
data("usgdp")
# Take the exact same window of usgdp series
usGDP = window(usgdp,start = c(1971,3), end=c(1998,1))

# Create the list of series reqiured by the LEIC function
data = list()
data[[1]] = ts(ausgdp, start = c(1971,3), frequency = 4)
data[[2]] = ts(usGDP, start = c(1971,3), frequency = 4)

# Specify the forecast horizon
H = 5

# Specify the models we will focus on
models = c("ANN", "MNN", "AAN", "AAA")
# This set of models includes the ones we found
# suitable in Module 8.

# Call the GoFVals function
GoFVals(data = data, H = H, models = models)
```

```
## $GoF
##   series FittedModels      AIC     AICc      BIC     HQIC
MASE
## 1      1          ANN 1333.377 1333.610 1341.396 1333.545 0.281
7838
## 2      1          MNN 1333.176 1333.409 1341.194 1333.343 0.281
7836
## 3      1          AAN 1196.733 1197.573 1212.770 1200.151 0.134
8853
## 4      1          AAA 1209.909 1211.765 1233.964 1216.577 0.141
6623
## 5      2          ANN 1408.040 1408.273 1416.059 1408.207 0.281
7374
## 6      2          MNN 1413.435 1413.668 1421.454 1413.602 0.281
7537
## 7      2          AAN 1335.245 1335.839 1348.609 1337.579 0.181
6359
## 8      2          AAA 1341.790 1344.081 1368.518 1349.542 0.178
9328
##
## $mean.MASE
## [1] 0.4200575 0.4620299
##
## $median.MASE
## [1] 0.2117229 0.2316866
```

```
# Call the LEIC function
LEIC(data = data, H = H, models = models)
```

```
## $leic
##   series FittedModels   values
## 1      1          ANN 1255.672
## 2      1          MNN 1257.190
## 3      1          AAN 1128.851
## 4      1          AAA 1136.419
## 5      2          ANN 1327.080
## 6      2          MNN 1334.166
## 7      2          AAN 1260.973
## 8      2          AAA 1261.074
##
## $c.opt
## [1] 0.3592123
```

The optimal value of c is found 0.359, which is close to 0.25 mentioned as optimal by the authors of the Textbook. For AUSGDP series minimum LEIC is found as 1128.249 from the AAN model, and for USGDP series it is 1261.082 from AAA model and 1262.571 from AAN model. For all other measures, mostly AAN model is supported for AUSGDP series and AAA model is supported for USGDP series while AAN model is the second best model.

Choosing a Model Selection Procedure

It is not suitable and useful to fit all models that we can fit to a dataset and then apply a model selection procedure. Instead, we first use descriptive analysis to get some sense about the structure of trend and seasonal components and existence of changing variance in the series and select a subset of available models as the set of tentative models. Then, we fit these models and conduct diagnostic analyses to see if the residuals support the model assumptions or are there any unusual residuals that can introduce a bias in the estimation procedure. This reduces the size of the set of tentative models which we can apply ICc and other criteria to select the most promising model. Then, we need to choose a model selection procedure to move on.

Measures for Comparing Model Section Procedures

In our comparisons, we will include the following procedures for choosing forecasting models for N time series:

- A single model for all time series
- Minimum IC (AIC, BIC, AICc, HQIC, LEIC)
- Prediction validation (VAL)

We will define three measures based on MASE to compare model selection procedures for forecasting.

Let the set of observations for the time series $\{Y_t^{(j)}\}, j = 1, \ldots, N$ be split into two parts: a *fitting set* of the first n_j values and a *forecasting set* of the last H values. All of the measures are based on the mean absolute scaled forecast error $\text{MASE}(H, i, j)$, where $i = 1, \ldots, M$. So, we will have a different MASE value for each set of observations, forecasting set, and model. The models are numbered from 1 to M, and for model selection procedure k, we denote the number of the model selected for time series $\{Y_t^{(j)}\}$ by k_j. The rank $r(H, k_j, j)$ for procedure k and set of observations j is the rank of $\text{MASE}(H, k_j, j)$ among the values of $\text{MASE}(H, i, j), i = 1, \ldots, M$, when they are ranked in ascending order.

For a specified model selection procedure k and number of forecasting horizons H, the following measures will be computed:

$$\begin{aligned} \text{Mean rank MASE}(H, k) &= \frac{1}{N}\sum_{j=1}^{N} r(H, k_j, j), \\ \text{Mean MASE}(H, k) &= \frac{1}{N}\sum_{j=1}^{N} \text{MASE}(H, k_j, j), \\ \text{Median MASE}(H, k) &= \text{median}\{\text{MASE}(H, k_j, j); j = 1, \ldots, N\}. \end{aligned} \tag{6}$$

We can write another bunch of similar measures using MAPE as well. We will use these three measures to select the model selection procedure.

The following `MASEvalues()` function computes Mean rank $\text{MASE}(H, k)$, Mean $\text{MASE}(H, k)$ and Median $\text{MASE}(H, k)$ for a given set of time series datasets and MASE values obtained by a model selection procedure.

```
MASEvalues = function(data, H, model, MASEs){
  # MASEs: All MASE values resulting from a model selection proced
  ure
  N = length(data) # The number of considered time series
  MASEs = sort(MASEs)
  MASE.model = array(NA, N)
  MASE.rank = array(NA, N)
  fit.models = list()
  for ( j in 1:N){
    fit.models[[j]] = ets(data[[j]], model = model)
    MASE.model[j] = accuracy(fit.models[[j]])[6]
    MASE.rank[j] = which(MASE.model[j] == MASEs)
  }
  mean.rank.MASE = mean(MASE.rank)
  # Mean of MASE values over all considered datasets based on the
   best model
  # which is selected by a particular model selection procedure
  mean.MASE = mean(MASE.model)
  median.MASE = median(MASE.model)
  return(list(mean.rank.MASE = mean.rank.MASE, mean.MASE = mean.MA
  SE, median.MASE = median.MASE))
}
```

For illustration, let's consider the AUSGDP and USGDP series. For these series mostly AAN model was supported by IC method. First, we will get MASE values from the output of `GoFVals()` function and feed them into the `MASEvalues()` function.

```
GoFs = GoFVals(data = data, H = H, models = models)
MASEs = GoFs$GoF$MASE
MASEvalues(data = data, H = H, model = "AAN", MASEs = MASEs)
```

```
## $mean.rank.MASE
## [1] 2.5
##
## $mean.MASE
## [1] 0.1582606
##
## $median.MASE
## [1] 0.1582606
```

Then we will compare these results with those obtained by the prediction validation method.

Prediction validation

Prediction validation method is frequently used to select a model from a set of M tnetative models in practice. The algorithm we apply for prediction validation is as the following:

1. Divide the fitting set for time series $\{y_t^{(j)}\}$ of lenght n_j into two segments: the first segment consists of $n_j^* = n_j - H$ observations, and the second segment consists of the last H observations.
2. Using $\{y_1^{(j)}\}$ to $\{y_{n_j^*}^{(j)}\}$, find the maximum likelihood estimates for each model i, $i = 1, \ldots, M$.
3. For each model i, compute forecasts.
4. Compute $MASE(H, i, j)$ defined in the algorithm for LEIC.
5. Choose model k_j, where

$$MASE(H, k_j, j) = \min\{MASE(H, i, j), i = 1, \ldots, M\}. \quad (7)$$

The parameters and initial state vector for the selected model must be re-estimated using all n_j values.

Implementation of this prediction validation approach is very similar to the calculation of LEIC. The following algorithm implements the prediction validation approach:

```
pVal = function(data, H, models){
  M = length(models) # The number of competing models
  N = length(data) # The number of considered time series
  n = array(NA, N) # Array to hold the length of each series
  for ( j in 1:N){ # Find the length of each series
    n[j] = length(data[[j]])
  }
  n.max = max(n)
  n.star = n - H

  # Fit the models
  fit.models = list()
  forecasts = list()
  count = 0
  for ( j in 1:N ){ # For each series of length nj*
    for ( i in 1:M){ # Fit all models
      count = count + 1
      fit.models[[count]] = ets(ts(data[[j]][1:n.star],frequency =
  frequency(data[[j]])), model = models[i])
      forecasts[[count]] = forecast(fit.models[[count]])$mean
    }
  }

  ASE = array(NA, dim = c(H, M, N))
  MASE.1 = array(NA, dim = c(M, N))
  MASE = array(NA, N)
  MASE.model = array(NA, N)
  summ = 0
  for (h in 1:H){
    summ = 0
    count = 0
    for ( j in 1:N){
      MAE = 0
      for ( t in 2:n[j] ){
        MAE = MAE + abs(data[[j]][t] - data[[j]][t-1])
      }
      for ( i in 1:M ){
        count = count + 1
        ASE[h, i , j] = sum( abs(data[[j]][(n.star + 1):(n.star+
  h)] - forecast(fit.models[[count]], h = h)$mean) / MAE )
      }
    }
  }

  for ( j in 1:N){
    for ( i in 1:M ){
      MASE.1[i , j] = 0
      for (h in 1:H){
        MASE.1[i , j] = MASE.1[i , j] + ASE[h, i , j]
```

```
      }
      MASE.1[i , j] = MASE.1[i , j]  / H
    }
    MASE[j] = min(MASE.1[ , j])
    MASE.model[j] = models[which(MASE[j] == MASE.1[ , j] )]
  }

  return(list(MASE = MASE, best.model = MASE.model))
}
```

Let's illustrate the use of this function over `ausgdp` and `usgdp` series available in `expsmooth` package. WE will use the same data and setting as those used to compute LEIC.

```
H = 5
models = c("ANN", "MNN", "AAN", "AAA")
pVal(data = data, H = H, models = models)
```

```
## $MASE
## [1] 0.06111059 0.02707487
##
## $best.model
## [1] "AAA" "AAA"
```

Notice that we worked over two similar series here: AUSGDP and USGDP in the same timeframe. The results of the prediction validation approach suggest that for GDP series AAN model consistently gives the best prediction results with MASE values of 0.06453209 and 0.01531074 for Australian and United States GDP series between 1971 and 1998.

Practical Illustration

Comparing Selection Procedures on the M3 Data

In this section, we will use the M3 competition data (Makridakis and Hibon, 2000) to compare the mentioned model selection procedures. For the 645 annual time series, because there is no seasonality, we apply innovations state-space models without seasonality. The number and percentages of time series with minimum MASE for each of non-seasonal models are shown below:

Model	Count	Percent
ETS(A,N,N)	141	21.86
ETS(A,M,N)	84	13.02
ETS(M,M,N)	74	11.47
ETS(A,M_d,N)	72	11.16
ETS(M,A,N)	56	8.68
ETS(A,A,N)	54	8.37
ETS(M,M_d,N)	52	8.06
ETS(A,A_d,N)	40	6.20
ETS(M,N,N)	37	5.74
ETS(M,A_d,N)	35	5.43

* Taken from: Hyndman et al., Forecasting with Exponential Smoothing: The State Space Approach. Springer, 2008.

The model ETS(A,N,N) is the most frequent model which is selected as the best within 10 candidate models for 141 series according to MASE criterion. However, each model fits best for some of the series according to MASE criterion.

For 1428 monthly M# time series, the number and percentages of time series with minimum MASE for each of seasonal and non-seasonal innovations state space models are shown below:

Model	Count	Percent	Model	Count	Percent
ETS(M,M,N)	92	6.44	ETS(A,M,A)	40	2.80
ETS(M,A,N)	81	5.67	ETS(A,M_d,N)	39	2.73
ETS(M,A,M)	78	5.46	ETS(A,M_d,M)	39	2.73
ETS(A,M,N)	76	5.32	ETS(A,N,A)	38	2.66
ETS(A,N,N)	69	4.83	ETS(M,A_d,M)	37	2.59
ETS(A,A,M)	63	4.41	ETS(M,A_d,N)	36	2.52
ETS(M,M,M)	60	4.20	ETS(M,M_d,M)	35	2.45
ETS(A,N,M)	58	4.06	ETS(A,A_d,N)	34	2.38
ETS(A,A,N)	57	3.99	ETS(M,M_d,A)	33	2.31
ETS(A,M,M)	54	3.78	ETS(M,M_d,N)	33	2.31
ETS(M,N,M)	49	3.43	ETS(A,A_d,M)	32	2.24
ETS(M,N,A)	48	3.36	ETS(M,M,A)	30	2.10
ETS(M,N,N)	47	3.29	ETS(A,A,A)	30	2.10
ETS(M,A,A)	44	3.08	ETS(A,A_d,A)	30	2.10
ETS(M,A_d,A)	43	3.01	ETS(A,M_d,A)	23	1.61

* Taken from: Hyndman et al., Forecasting with Exponential Smoothing: The State Space Approach. Springer, 2008.

For the monthly data, which includes seasonality, we reach the same conclusion that every model is best for some of the time series. Therefore, it might be beneficial to have a procedure for choosing from among all these models.

On the other hand, what happens if we fit the same model to all series? The following display shows the results of mean rank, mean MASE, and median MASE, for each non-seasonal model when fitted to all annual series.

Model	Mean rank	Mean MASE	Median MASE
ETS(A,A_d,N)	4.97	2.92	1.82
ETS(M,A_d,N)	5.23	2.97	1.95
ETS(A,A,N)	5.25	2.99	1.97
ETS(A,M_d,N)	5.29	3.57	1.75
ETS(M,M_d,N)	5.31	3.24	1.89
ETS(M,A,N)	5.37	2.96	2.01
ETS(A,M,N)	5.77	4.18	1.96
ETS(M,M,N)	5.87	3.63	2.05
ETS(A,N,N)	5.93	3.17	2.26
ETS(M,N,N)	6.02	3.19	2.26

* Taken from: Hyndman et al., Forecasting with Exponential Smoothing: The State Space Approach. Springer, 2008.

This time, we see that the ETS(A,Ad,N) model fits better than the others to all datasets in an overall sense. So, different measures can give different results to the model selection problem. This also supports the notion of trying to find a model selection procedure.

Including the quarterly data as well, best methods and models for each measure based on MASE and the IC which gives the minimum value are shown below:

Measure	Data type	Damped trend	AIC	Best method(s)
(a) Comparison using MASE for linear models				
Mean Rank	Annual	1.86/ETS(A,A_d,N)	1.84	1.84/AIC
	Quarterly	2.96/ETS(A,A_d,A)	3.08	3.08/AIC
	Monthly	3.29/ETS(A,A_d,A)	3.07	3.03/AICc
Mean MASE	Annual	2.92/ETS(A,A_d,N)	2.94	2.94/AIC
	Quarterly	2.14/ETS(A,A_d,A)	2.15	2.15/AIC, LEIC
	Monthly	2.09/ETS(A,A_d,A)	2.06	2.05/AICc
Median MASE	Annual	1.82/ETS(A,A_d,N)	1.82	1.82/AIC
	Quarterly	1.46/ETS(A,A_d,A)	1.47	1.47/AIC
	Monthly	1.12/ETS(A,A_d,A)	1.08	1.07/AICc
(b) Comparison using MASE for all models				
Mean Rank	Annual	4.97/ETS(A,A_d,N)	5.42	5.29/BIC
	Quarterly	12.84/ETS(M,A_d,M)	13.97	13.97/AIC
	Monthly	14.09/ETS(A,A_d,M)	13.50	13.29/AICc
Mean MASE	Annual	2.92/ETS(A,A_d,N)	3.30	2.91/LEIC
	Quarterly	2.13/ETS(M,A_d,M)	2.27	2.27/AIC
	Monthly	2.10/ETS(A,A_d,M)	2.07	2.08/AIC, AICc, HQIC
Median MASE	Annual	1.82/ETS(A,A_d,N)	1.98	1.92/LEIC
	Quarterly	1.50/ETS(M,A_d,M)	1.54	1.54/AIC
	Monthly	1.10/ETS(A,A_d,M)	1.10	1.07/HQIC

* Taken from: Hyndman et al., Forecasting with Exponential Smoothing: The State Space Approach. Springer, 2008.

When MAPE is used instead of MASE, the following results are obtained.

(c) Comparison using MAPE for linear models				
Mean Rank	Annual	1.86/ETS(A,A_d,N)	1.83	1.83/AIC
	Quarterly	2.98/ETS(A,A_d,A)	3.07	3.07/AIC
	Monthly	3.22/ETS(A,A_d,A)	3.08	3.06/AICc
Mean MAPE	Annual	22.66/ETS(A,A_d,N)	22.00	21.33/AICc
	Quarterly	12.06/ETS(A,A_d,A)	11.95	11.94/LEIC
	Monthly	22.01/ETS(A,A_d,A)	21.75	21.23/AICc
Median MAPE	Annual	10.92/ETS(A,A_d,N)	11.18	11.16/AICc, LEIC
	Quarterly	5.32/ETS(A,A_d,A)	5.46	5.46/AIC
	Monthly	9.30/ETS(A,A_d,A)	9.29	9.29/AIC, AICc
(d) Comparison using MAPE for all models				
Mean Rank	Annual	4.98/ETS(A,A_d,N)	5.45	5.26/LEIC
	Quarterly	12.86/ETS(M,A_d,M)	13.87	13.87/AIC
	Monthly	13.76/ETS(A,A_d,M)	13.62	13.54/AICc
Mean MAPE	Annual	22.66/ETS(A,A_d,N)	25.42	20.71/LEIC
	Quarterly	11.96/ETS(M,A_d,M)	12.23	12.15/HQIC
	Monthly	20.02/ETS(A,A_d,M)	21.63	21.62/HQIC
Median MAPE	Annual	10.92/ETS(A,A_d,N)	11.54	11.16/LEIC
	Quarterly	5.22/ETS(M,A_d,M)	5.62	5.54/VAL
	Monthly	9.15/ETS(A,A_d,M)	9.03	8.96/VAL

* Taken from: Hyndman et al., Forecasting with Exponential Smoothing: The State Space Approach. Springer, 2008.

There are three potential non-seasonal linear models for the annual time series and six potential linear models for the quarterly and monthly data. The last two columns in the table indicate that, among the model selection methods, the AIC always has the minimum, or nearly the minimum, value for each measure. On the other hand, applying the damped trend model gives satisfactory results. One could decide to choose these models rather than use the AIC model selection method. However, one should be reassured that the AIC will do as well as or better than the encompassing model, and it would lead to the selection of simpler models when possible.

When we include multiplicative nonlinear models, we consider 30 models in total. In this case, other measures like BIC and HQIC give better models for annual and monthly data while AIC is still good for quarterly models. Damped trend models seem to be promising for the multiplicative case but which damped trend model to use is not that clear (ETS(M,Ad,M) or ETS(A,Ad,M)). The mean MASE and median MASE do not decrease when the number of models in the selection process is increased from six to 30. For both monthly and quarterly time series, one should consider using the AIC with an expanded set of linear models, but far fewer than all 30 models.

When we use MAPE instead of MASE, we get different results for multiplicative models.

See the following display for the comparison of individual methods over only linear models and all models:

Measure	Data type	AIC	BIC	HQIC	AICc	LEIC	VAL
(a) Comparison of methods using MASE for linear models							
Mean Rank	Annual	1.84	1.86	1.86	1.88	1.86	1.97
	Quarterly	3.08	3.24	3.14	3.16	3.12	3.26
	Monthly	3.07	3.15	3.05	3.03	3.23	3.20
Mean MASE	Annual	2.94	2.96	2.95	2.96	2.95	3.04
	Quarterly	2.15	2.21	2.16	2.17	2.15	2.19
	Monthly	2.06	2.13	2.09	2.05	2.19	2.17
Median MASE	Annual	1.82	1.85	1.85	1.85	1.85	1.95
	Quarterly	1.47	1.58	1.50	1.49	1.49	1.53
	Monthly	1.08	1.11	1.08	1.07	1.12	1.10
(b) Comparison of methods using MASE for all models							
Mean Rank	Annual	5.42	5.29	5.39	5.33	5.31	5.55
	Quarterly	13.97	14.75	14.20	14.47	15.14	14.87
	Monthly	13.50	13.60	13.33	13.29	14.78	13.92
Mean MASE	Annual	3.30	3.28	3.29	3.26	2.91	3.37
	Quarterly	2.27	2.38	2.29	2.29	2.40	2.29
	Monthly	2.08	2.10	2.08	2.08	2.19	2.20
Median MASE	Annual	1.98	1.95	1.97	1.97	1.92	2.00
	Quarterly	1.54	1.57	1.55	1.56	1.61	1.55
	Monthly	1.10	1.11	1.07	1.09	1.14	1.09

* Taken from: Hyndman et al., Forecasting with Exponential Smoothing: The State Space Approach. Springer, 2008.

This table gives more insight into the use of ICs for model comparison. In most cases for linear models, AIC or AICc seems to be promising. But, the use of other measures should be considered for non-linear models.

Comparing Selection Procedures on a Hospital Data Set

In this hospital data set, each time series comprises a monthly patient count for one of 20 products that are related to medical problems. The total number of series in this dataset is 767. The following display shows the results for the hospital data set using the MASE with only linear models and all models.

Measure	Data type	AIC	BIC	HQIC	AICc	LEIC	VAL
(a) Comparison of methods using MASE for linear models							
Mean Rank	Monthly	3.10	3.07	3.01	3.10	3.07	3.36
Mean MASE	Monthly	0.94	0.91	0.92	0.94	0.91	0.96
Median MASE	Monthly	0.83	0.83	0.83	0.83	0.83	0.84
(b) Comparison of methods using MASE for all models							
Mean Rank	Monthly	13.66	13.25	13.22	13.52	13.53	14.80
Mean MASE	Monthly	0.98	0.96	0.97	0.98	0.95	1.00
Median MASE	Monthly	0.84	0.84	0.83	0.84	0.85	0.86

* Taken from: Hyndman et al., Forecasting with Exponential Smoothing: The State Space Approach. Springer, 2008.

For linear models with monthly data, BIC and HQIC perform better and all the measures are close to each other according to median MASE. The results for all models are similar to those obtained for only linear models. A difference from the findings with the M3 data is that we found that it is not a good idea to use a single damped trend model for forecasting.

Implications for Model Selection Procedures

We can draw the following inferences from the applications with real data:

- The AIC model selection method was shown to be a reasonable choice among the six model selection methods for the three types of data (annual, quarterly, and monthly) in the M3 data and for the monthly time series in the hospital data.
- The number of observations for annual data is always likely to be small (i.e., less than or equal to 40), and thus the IC procedures may not have sufficient data to compete with simply choosing a single model such as the ETS(A,Ad,N) model when all ten non-seasonal models are considered.
- However, using the AIC on the three linear non-seasonal models fared as well as the ETS(A,Ad,N) and would allow the possibility of choosing simpler models, especially when there is a mild trend in the data. Thus, for annual time series, we recommend using the AIC and choosing among the three linear non-seasonal models.
- For the monthly data, the AIC is better than the choice of selecting a single damped trend model in both the M3 data and the hospital data.
- Because it is definitely not clear which single model to use, we suggest using the AIC.
- One might also consider limiting the choice of models to a set that includes the linear models but is smaller than the complete set of 30 models.
- We make the same recommendations for the quarterly time series, with additional emphasis on reducing the number of models from 30.

Notice that, the results shown here are all based on real datasets rather than a Monte Carlo simulation. So, as seen in both applications, it is hard to generalize the results for all possible datasets.

Summary

In this module, we considered the model selection problem in terms of the innovations state-space models. We considered different sets of tentative models for annual and monthly/quarterly series, which includes seasonality. Four criteria used to select models are discussed an illustrated in terms of how they penalize the likelihood considering the number of parameters in the sample and sample size. We considered two datasets that include large numbers of real time series and applied model selection criteria to these series to see their frequency of selecting the best model in relation to the MASE, which is independent of the scale; hence, appropriately used to compare two non-nested models.

References

Akaike, H. (1974) A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, 19, 716–723.

Billah, B., R. J. Hyndman and A. B. Koehler (2003) Empirical information criteria for time series forecasting model selection, Working paper 02/03, Department of Econometrics & Business Statistics, Monash University.

Billah, B., R. J. Hyndman and A. B. Koehler (2005) Empirical information criteria for time series forecasting model selection, *Journal of Statistical Computation & Simulation*, 75(10), 831–840.

Hannan, E. J. and B. Quinn (1979) The determination of the order of an autoregression, *Journal of the Royal Statistical Society*, Series B, 41(2), 190–195.

Hyndman, R.J., Koehler, A.B., Ord, J.K., and Snyder, R.D. (2008) Forecasting with exponential smoothing: the state space approach (http://www.exponentialsmoothing.net), Springer-Verlag.

Makridakis, S. and M. Hibon (2000) The M3-competition: results, conclusions and implications, *International Journal of Forecasting*, 16, 451–476.

Schwarz, G. (1978) Estimating the dimension of a model, *The Annals of Statistics*, 6, 461–464.

Sugiura, N. (1978) Further analysis of the data by Akaike's information criterion and the finite corrections, *Communications in Statistics*, A7, 13–26.