ESTIMATING THE DIMENSION OF A MODEL

BY GIDEON SCHWARZ

Hebrew University

The problem of selecting one of a number of models of different dimensions is treated by finding its Bayes solution, and evaluating the leading terms of its asymptotic expansion. These terms are a valid large-sample criterion beyond the Bayesian context, since they do not depend on the a priori distribution.

Qualitatively both our procedure and Akaike’s give “a mathematical formulation of the principle of parsimony in model building.” Quantitatively, since our procedure differs from Akaike’s only in that the dimension is multiplied by $\frac{1}{2} \log n$, our procedure leans more than Akaike’s towards lower-dimensional models (when there are 8 or more observations). For large numbers of observations the procedures differ markedly from each other. If the assumptions we made in Section 2 are accepted, Akaike’s criterion cannot be asymptotically optimal. This would contradict any proof of its optimality, but no such proof seems to have been published, and the heuristics of Akaike [1] and of Tong [4] do not seem to lead to any such proof.

Outline

1. Forecasting
2. Model Selection
Forecasting Goal: Minimize Mean Square Error

We wish to choose a prediction that is \( m \) time points ahead by \( \hat{Z}_n(m) \) which minimizes the mean square error (MSE), i.e. \( \hat{Z}_n(m) \) is the minimizer of

\[
E \left[ (Z_{n+m} - g(Z))^2 \right]
\]

(where \( Z = (Z_1, \ldots, Z_n) \))

Note that

\[
E \left[ (y - g(Z))^2 \right] = E \left[ E \left[ (y - g(Z))^2 | Z \right] \right]
\]

\[
= E \left[ \frac{E[y^2|Z] - 2g(Z)E[y^2|Z] + g(Z)^2}{\frac{\partial}{\partial g(Z)}} \right]
\]

\[
\Rightarrow g(Z) = E(y|Z)
\]

Therefore

\[
\hat{Z}_n(m) = E(Z_{n+m}|Z_n, Z_{n-1}, \ldots, Z_1)
\]
Best Linear Predictor

We will focus on linear predictors of the form

\[ \hat{Z}_n(m) = \alpha_0 + \sum_{k=1}^{n} \alpha_k Z_k \]  

(\star)

**Theorem (Best Linear Prediction for Stationary Processes)**

*Given data \(Z_1, Z_2, \ldots, Z_n\), the best linear predictor of \(Z_{n+m}\) for \(m \geq 1\) is found by solving*

\[ E \left[ (Z_{n+m} - \hat{Z}_n(m))Z_k \right] = 0, \quad k = 0, 1, \ldots, n \]

*where \(x_0 = 1\).*

Equations (\star) are called the prediction equations, and they provide a useful tool in computing the best linear predictor \(\hat{Z}_n(m)\).
\[ \mathbb{E} \left[ (Z_{n+m} - \hat{Z}_n(m))Z_k \right] = 0, \quad k = 0, 1, \ldots, n \]
Outline

1. Forecasting

2. Model Selection
Examples in R

```r
> ar2 = arima.sim(list(order=c(2,0,0), ar=c(1.5,-.75)), n = 100)
> ts.plot(ar2)

> ar2.acf = ARMAacf(ar=c(1.5,-.75), ma=0, 24)
> ar2.pacf = ARMAacf(ar=c(1.5,-.75), ma=0, 24, pacf=T)
> par(mfrow=c(2,1))
> acf(ar2, lwd=2)
> lines(0:24, ar2.acf, type="p", lwd=3, col="red")
> pacf(ar2, lwd=2)
> lines(ar2.pacf, type="p", lwd=3, col="red")
```
Examples in R (cont.)

Series ar2

ACF
-0.5  0.0  0.5  1.0

Partial ACF
-0.5  0.0  0.5
> ma2 = arima.sim(list(order=c(0,0,2), ma=c(1.5,-.75)), n = 100)
> ts.plot(ma2)

> ma2.acf = ARMAacf(ma=c(1,1), ar=0, 24)
> ma2.pacf = ARMAacf(ma=c(1,1), ar=0, 24, pacf=T)
> par(mfrow=c(2,1))
> acf(ma2, lwd=2)
> lines(0:24,ma2.acf, type="p",lwd=3, col="red")
> pacf(ma2,lwd=2)
> lines(ma2.pacf, type="p",lwd=3, col="red")
Examples in R (cont.)

Series ma2

- ACF values are plotted for various lags.
- The graph shows the autocorrelation function for the series.
- The significance levels are indicated by dashed lines.

- Partial ACF values are also plotted for the same lags.
- The partial autocorrelation function helps in identifying the order of the MA component.
- Similar significance levels are indicated by dashed lines.

Lag values range from 0 to 20.

Arthur Berg
Forecasting Model Selection

Examples in R (cont.)

```r
> arma22 = arima.sim(list(order=c(2,0,2),ar=c(1.5,-.75) ,ma=c(1.5,-.75)), n = 100)
> ts.plot(arma22)

> arma22.acf = ARMAacf(ma=c(1,1), ar=c(1.5,-.75), 24)
> arma22.pacf = ARMAacf(ma=c(1,1), ar=c(1.5,-.75), 24, pacf=T)
> par(mfrow=c(2,1))
> acf(arma22, lwd=2)
> lines(0:24,arma22.acf, type="p",lwd=3, col="red")
> pacf(arma22,lwd=2,ylim=c(-1,1))
> lines(arma22.pacf, type="p",lwd=3, col="red")
```
Examples in R (cont.)

Series arma22

ACF

Partial ACF

Lag
### Table: ACF and PACF for Causal and Invertible ARMA Models

<table>
<thead>
<tr>
<th></th>
<th>AR$(p)$</th>
<th>MA$(q)$</th>
<th>ARMA$(p,q)$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ACF</strong></td>
<td>Tails off</td>
<td>Cuts off after lag $q$</td>
<td>Tails off</td>
</tr>
<tr>
<td><strong>PACF</strong></td>
<td>Cuts off after lag $p$</td>
<td>Tails off</td>
<td>Tails off</td>
</tr>
</tbody>
</table>
Forecasting Model Selection

Recruitment Series

```r
> setwd("C:/Users/berg.UFAD/Documents/sta 6857/R")
> rec = ts(scan("mydata/recruit.dat"), start=1950, frequency=12)
```
Recruitment Series (cont.)

```r
> par(mfrow=c(2,1))
> acf(rec, 48)
> pacf(rec, 48)
```
> (fit = ar.ols(rec, order=2, demean=F, intercept=T))

Call:
ar.ols(x = rec, order.max = 2, demean = F, intercept = T)

Coefficients:
   1     2
1.3541 -0.4632

Intercept: 6.737 (1.111)

Order selected 2  sigma^2 estimated as 89.72

> fit$asy.se

$x.mean
[1] 1.110599

$ar
[1] 0.04178901 0.04187942
> rec.pr = predict(rec.yw, n.ahead=24)
> U = rec.pr$pred + rec.pr$se
> L = rec.pr$pred - rec.pr$se
> minx = min(rec,L)
> maxx = max(rec,U)
> ts.plot(rec, rec.pr$pred, xlim=c(1980,1990), ylim=c(minx,maxx))
> lines(rec.pr$pred, col="red", type="o")
> lines(U, col="blue", lty="dashed")
> lines(L, col="blue", lty="dashed")
Model Selection with the F-Statistic

- **Forward Step** — Start with a minimal model. Increase the model by adding in the variable which produces the largest F-statistic. Repeat until no additional variable added produces a significant F-statistic.

- **Backward Step** — Start with a full model. Decrease the model by removing the variable which gives the smallest contribution. Repeat until all remaining variables are significant.

- **Forward/Backward Step** — Increase the model by adding in the variable which produces the largest F-statistic. Decrease the model by removing weakest variable as long as its contribution is not significant. Repeat until no additional variables added produces a significant F-statistic and all variables in the model are significant.

---

**Example**

**Minimal Model:** \( x_t = \beta_1 + w_t \)

**Complete Model:** \( x_t = \beta_1 + \beta_2 t + \beta_3 t^2 + \beta_4 \sin t + w_t \)

Use the above procedures to find a model in between the minimal and complete models like:

\( x_t = \beta_1 + \beta_4 \sin t + w_t \).
You may wish to view my slides on Linear Models with R (click).

```r
> x <- seq(0, 10, .1)
> y <- 1 + x + cos(x) + rnorm(x)
> yy <- 1 + x + cos(x)
> plot(x, y, lwd = 3)
> lines(x, yy, lwd = 3)
```
Example in R (cont.)

```r
> f1 <- lm(y ~ x + I(x^2) + I(x^3) + sin(x) + cos(x) + exp(x))
> anova(f1)

Analysis of Variance Table

Response: y

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>1</td>
<td>692.94</td>
<td>692.94</td>
<td>515.5497 &lt; 2.2e-16 ***</td>
</tr>
<tr>
<td>I(x^2)</td>
<td>1 0.03</td>
<td>0.03</td>
<td>0.0214</td>
<td>0.8841297</td>
</tr>
<tr>
<td>I(x^3)</td>
<td>1 19.69</td>
<td>19.69</td>
<td>14.6468</td>
<td>0.0002335 ***</td>
</tr>
<tr>
<td>sin(x)</td>
<td>1 1.42</td>
<td>1.42</td>
<td>1.0552</td>
<td>0.3069485</td>
</tr>
<tr>
<td>cos(x)</td>
<td>1 11.01</td>
<td>11.01</td>
<td>8.1934</td>
<td>0.0051832 **</td>
</tr>
<tr>
<td>exp(x)</td>
<td>1 0.03</td>
<td>0.03</td>
<td>0.0192</td>
<td>0.8900193</td>
</tr>
<tr>
<td>Residuals</td>
<td>94 126.34</td>
<td>1.34</td>
<td>---</td>
<td></td>
</tr>
</tbody>
</table>
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 1

> f2 <- lm(y ~ x + I(x^2) + I(x^3) + sin(x) + cos(x))
> anova(f2)

Analysis of Variance Table

Response: y

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>1</td>
<td>692.94</td>
<td>692.94</td>
<td>520.9277 &lt; 2.2e-16 ***</td>
</tr>
<tr>
<td>I(x^2)</td>
<td>1 0.03</td>
<td>0.03</td>
<td>0.0216</td>
<td>0.8835280</td>
</tr>
<tr>
<td>I(x^3)</td>
<td>1 19.69</td>
<td>19.69</td>
<td>14.7996</td>
<td>0.0002164 ***</td>
</tr>
<tr>
<td>sin(x)</td>
<td>1 1.42</td>
<td>1.42</td>
<td>1.0662</td>
<td>0.3044251</td>
</tr>
<tr>
<td>cos(x)</td>
<td>1 11.01</td>
<td>11.01</td>
<td>8.2789</td>
<td>0.0049533 **</td>
</tr>
<tr>
<td>Residuals</td>
<td>95 126.37</td>
<td>1.33</td>
<td>---</td>
<td></td>
</tr>
</tbody>
</table>
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```
Example in R (cont.)

```r
> f3 <- lm(y ~ x + I(x^3) + sin(x) + cos(x))
> anova(f3)

Analysis of Variance Table

Response: y

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>1</td>
<td>692.94</td>
<td>692.94</td>
<td>&lt; 2.2e-16 ***</td>
</tr>
<tr>
<td>I(x^3)</td>
<td>1</td>
<td>0.34</td>
<td>0.34</td>
<td>0.2549 0.6148</td>
</tr>
<tr>
<td>sin(x)</td>
<td>1</td>
<td>0.16</td>
<td>0.16</td>
<td>0.1231 0.7265</td>
</tr>
<tr>
<td>cos(x)</td>
<td>1</td>
<td>31.62</td>
<td>31.62</td>
<td>24.0135 3.88e-06 ***</td>
</tr>
<tr>
<td>Residuals</td>
<td>96</td>
<td>126.40</td>
<td>1.32</td>
<td></td>
</tr>
</tbody>
</table>

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```r
> f4 <- lm(y ~ x + I(x^3) + cos(x))
> anova(f4)

Analysis of Variance Table

Response: y

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>1</td>
<td>692.94</td>
<td>692.94</td>
<td>&lt; 2.2e-16 ***</td>
</tr>
<tr>
<td>I(x^3)</td>
<td>1</td>
<td>0.34</td>
<td>0.34</td>
<td>0.2573 0.6132</td>
</tr>
<tr>
<td>cos(x)</td>
<td>1</td>
<td>31.62</td>
<td>31.62</td>
<td>24.2370 3.492e-06 ***</td>
</tr>
<tr>
<td>Residuals</td>
<td>97</td>
<td>126.56</td>
<td>1.30</td>
<td></td>
</tr>
</tbody>
</table>

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Example in R (cont.)

```r
> f5 <- lm(y ~ x + cos(x))
> anova(f5)
```

Analysis of Variance Table

Response: y

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>1</td>
<td>692.94</td>
<td>692.94</td>
<td>533.564</td>
<td>&lt; 2.2e-16 ***</td>
</tr>
<tr>
<td>cos(x)</td>
<td>1</td>
<td>31.24</td>
<td>31.24</td>
<td>24.057</td>
<td>3.717e-06 ***</td>
</tr>
<tr>
<td>Residuals</td>
<td>98</td>
<td>127.27</td>
<td>1.30</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

```r
> lines(x,predict(f5), lwd=3, col="magenta")
```
Akaike’s Information Criterion (AIC)

General model selection idea:

Criterion = \underbrace{\text{Model Error}}_{\text{think of RSS—make as small as possible}} + \underbrace{\text{Penalty for Model Complexity}}_{\text{penalty increases with the number of parameters}}

Definition (AIC)

The AIC for \( k \) parameters is defined as

\[
AIC = \ln \hat{\sigma}_k^2 + \frac{n + 2k}{n}
\]

Definition (AICc — AIC for small sample sizes (small \( n \)))

The biased corrected AIC (AICc) for \( k \) parameters is defined as

\[
AIC = \ln \hat{\sigma}_k^2 + \frac{n + k}{n - k - 2}
\]
Bayesian Information Criterion (BIC) aka Schwarz’s Information Criterion (SIC)

### Definition (SIC/BIC)

The SIC for \( k \) parameters is defined as

\[
SIC = \ln \hat{\sigma}_k^2 + \frac{k \ln n}{n}
\]

The penalty term in the AIC is

\[
\frac{2k}{n}
\]

The penalty term in the SIC is

\[
\frac{k \ln n}{n}
\]

- The SIC penalizes larger models more severely than the AIC.
- The SIC is proven to be a **statistically consistent** model selection method whereas the AIC is not.
Mallows $C_p$

$C_p$ statistic is defined as

$$C_p = \frac{\text{RSS}_p}{\hat{\sigma}^2_q} - n + 2p$$

Pick model with smallest $C_p$ with the property that $C_p \approx p$. 

![Graph showing Mallows C_p vs number of variables]
Adjusted $R^2$

$\bar{R}^2$ statistic is defined as

$$\bar{R}^2 = 1 - \frac{RSS_p}{\sum(y_i - \bar{y})^2} \frac{n}{n - p}$$

Pick model with largest $\bar{R}^2$. 
Example

Cardiovascular Mortality

Temperature

Particulates
Scatterplot Matrix

> pairs(cbind(mort, temp, part))
Four Models

Data from 1970 to 1979 in LA.

- $M_t$ represents the cardiac mortality.
- $T_t$ represents the temperatures.
- $P_t$ represents the particulate pollution.

Four proposed models

1. $M_t = \beta_0 + \beta_1 t + w_t$
2. $M_t = \beta_0 + \beta_1 t + \beta_2 (T_t - T.) + w_t$
3. $M_t = \beta_0 + \beta_1 t + \beta_2 (T_t - T.) + \beta_3 (T_t - T.)^2 + w_t$
4. $M_t = \beta_0 + \beta_1 t + \beta_2 (T_t - T.) + \beta_3 (T_t - T.)^2 + \beta_4 P_t + w_t$
Example (cont.)

```r
> temp = temp - mean(temp)
> temp2 = temp^2
> trend = time(mort)
> fit = lm(mort ~ trend + temp + temp2 + part, na.action=NULL)
> summary(fit)

Call:
  lm(formula = mort ~ trend + temp + temp2 + part, na.action = NULL)

Residuals:
     Min       1Q   Median       3Q      Max
-19.0760  -4.2153  -0.4878   3.7435   29.2448

Coefficients:
                        Estimate Std. Error   t value     Pr(>|t|)
(Intercept)             2.831e+03   1.996e+02   14.19  < 2e-16 ***
trend                  -1.396e+00   1.010e-01   -13.82  < 2e-16 ***
temp                   -4.725e-01   3.162e-02   -14.94  < 2e-16 ***
temp2                  2.259e-02   2.827e-03    7.99    9.26e-15 ***
part                   2.554e-01   1.886e-02    13.54  < 2e-16 ***
---
Signif. codes:  *** 0.001 ** 0.01 * 0.05 . 0.1  1

Residual standard error: 6.385 on 503 degrees of freedom
Multiple R-Squared: 0.5954,   Adjusted R-squared: 0.5922
F-statistic: 185 on 4 and 503 DF,  p-value: < 2.2e-16
```
```r
> summary(aov(fit))

                  Df Sum Sq  Mean Sq F value Pr(>F)  
 trend             1 10666.9 10666.9 261.621 < 2.2e-16 ***
 temp              1  8606.6  8606.6 211.090 < 2.2e-16 ***
 temp2             1  3428.7  3428.7  84.094 < 2.2e-16 ***
 part              1  7476.1  7476.1 183.362 < 2.2e-16 ***
 Residuals         503 20508.4    40.8   

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```
### Example (cont.)

```r
> num = length(mort)
> AIC(fit)/num - log(2*pi) # AIC
[1] 4.721732

> AIC(fit, k=log(num))/num - log(2*pi) # BIC
[1] 4.771699

> log(sum(resid(fit)^2)/num)+(num+5)/(num-5-2) # AICc
[1] 4.722062
```

<table>
<thead>
<tr>
<th>Model</th>
<th>RSS</th>
<th>AICc</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>40,020</td>
<td>5.38</td>
</tr>
<tr>
<td>2</td>
<td>31,413</td>
<td>5.14</td>
</tr>
<tr>
<td>3</td>
<td>27,985</td>
<td>5.03</td>
</tr>
<tr>
<td>4</td>
<td>20,509</td>
<td>4.72</td>
</tr>
</tbody>
</table>