The Linear Regression Model: Model Structure Selection and Biased Estimators

by

J.G.M. Delissen

EUT Report 88-E-203
ISBN 90-6144-203-6
August 1988
THE LINEAR REGRESSION MODEL:
Model structure selection and biased estimators

by

J.G.M. Delissen

EUT Report 88-E-203
ISBN 90-6144-203-6

Eindhoven
August 1988
This report is submitted in partial fulfillment of the
requirements for the degree of electrical engineer (M.Sc.)
at the Eindhoven University of Technology. The work was
carried out from September 1987 until August 1988 in charge
of professor dr.ir. P. Eykhoff under supervision of
dr.ir. A.A.H. Damen.

CIP-GEGEVENS KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Delissen, J.G.M.

The linear regression model: model structure selection and biased estimators /
by J.G.M. Delissen. - Eindhoven: University of Technology, Faculty of Electrical
Engineering. - Fig. - (EUT report ISSN 0167-9708; 88-E-203)
Met lit. opg., reg.
ISBN 90-6144-203-6
SISD 656 UDC 519.71.001.3 NUGI 832
Trefw.: systeemidentificatie.
The linear regression model is one of the most frequently used models in statistics and there are applications of it in many different areas. In the field of system identification this model is often used for impulse response estimation. If the length of the actual impulse response is known the corresponding least squares estimator is known to be a minimum variance unbiased estimator. However, this is usually not the case and therefore it is more common than rare that the wrong model structure is used. In this report the consequences of (in)correct modelling will be studied. It will be shown that the estimates can be improved if a smaller model structure is chosen. Furthermore some criteria will be given which can be used for the selection of a model structure.

The asymptotic behaviour and sampling properties of these criteria will be studied in the case that the best out of two model structures (the true model and a predefined smaller model structure) has to be chosen. It will be shown that some criteria are asymptotically equivalent or will asymptotically choose the right model structure. Furthermore it will be shown that the risk involved with these criteria yield a range in the parameter space where it is smaller than the risk of the least squares estimator in the true model and a range where it is greater. As an exception to this rule a selection procedure will be given which is based on the James-Stein estimator.

In this report some other selection procedure will be given which is designed to improve the mean square error of estimating the impulse response. However, simulations will indicate that it is not superior to a common used similar model selection criterion.

Besides model structure selection some other ways are briefly investigated to improve the least squares estimator in the true model structure. Among them are principal component selection and ridge regression. Unfortunately they improve the least squares estimator only for a limited range in the parameter space. However, a class of biased estimators will be given, which will, under certain conditions, be uniformly better than the least squares estimator.

Delissen, J.G.M.
THE LINEAR REGRESSION MODEL: Model structure selection and biased estimators.
EUT Report 88-E-203

Correspondence about this report should be addressed to:

Dr. Ir. A.A.H. Damen,
Group Measurement and Control,
Faculty of Electrical Engineering,
Eindhoven University of Technology,
P.O. Box 513,
5600 MB Eindhoven,
The Netherlands
CONTENTS

NOTATIONAL CONVENTIONS.............................................. 4

1. INTRODUCTION.................................................................. 5

2. CONSEQUENCES OF (IN)CORRECT MODELLING......................... 7
   2.1 THE LEAST SQUARE ESTIMATOR.................................... 7
   2.2 PERFORMANCE MEASURES........................................... 7
   2.3 THE PROCESS IS IN THE MODEL SET.................................. 9
   2.4 THE PROCESS IS NOT IN THE MODEL SET.......................... 12
   2.5 EVALUATION.......................................................... 13

3. CRITERIA FOR SELECTING AN 'OPTIMAL' MODEL STRUCTURE........... 16
   3.1 NOTATION.................................................................. 16
   3.2 MODEL STRUCTURE SELECTION..................................... 17
   3.3 HYPOTHESIS ARGUMENTS............................................ 18
      3.3.1 CONFIDENCE INTERVALS...................................... 18
      3.3.2 HYPOTHESIS TESTING......................................... 18
   3.4 PREDICTION ARGUMENTS............................................ 20
      3.4.1 MALLOWS'S Cp................................................ 20
      3.4.2 AMEMIYA'S PC................................................. 21
      3.4.3 THE FINAL PREDICTION ERROR CRITERION.................. 22
   3.5 INFORMATION THEORETIC ARGUMENTS............................... 22
      3.5.1 AKAIKE'S INFORMATION THEORETIC CRITERION.............. 22
   3.6 BAYESIAN ARGUMENTS............................................. 24
      3.6.1 AKAIKE'S BAYESIAN INFORMATION CRITERION............... 25
      3.6.2 SCHWARZ CRITERION......................................... 26
   3.7 CROSS VALIDATION ARGUMENTS.................................... 27
      3.7.1 ALLAN'S PREDICTIVE SUM OF SQUARES....................... 27
      3.7.2 STOICA'S C1(m) AND C2(m).................................. 28
   3.8 STEIN RULE ARGUMENTS.......................................... 38

4. EVALUATION OF SELECTION CRITERIA................................ 43
   4.1 ASYMPTOTIC PROPERTIES......................................... 43
   4.2 F-TEST INTERPRETATION.......................................... 46
   4.3 THE PARAMETER SPACE............................................ 51
      4.3.1 THE MODIFIED Cp............................................ 51
   4.4 ORDERED-UNORDERED PARAMETER REDUCTION....................... 62
      4.4.1 COMPUTATIONAL PROBLEMS.................................. 62
      4.4.2 PRIOR KNOWLEDGE........................................... 64
   4.5 ALL FORMULAS CAN BE WRONG..................................... 66

5. BIASED ESTIMATORS.................................................. 68
   5.1 CANONICAL FORMS................................................ 68
   5.2 JAMES STEIN ESTIMATORS........................................ 69
   5.3 PRINCIPAL COMPONENT SELECTION................................ 69
   5.4 RIDGE REGRESSION................................................ 70

6. CONCLUSIONS.................................................................. 72

7. REFERENCES.................................................................... 73

APPENDIX A........................................................................ 76
APPENDIX B........................................................................ 79
APPENDIX C........................................................................ 80
APPENDIX D........................................................................ 81
NOTATIONAL CONVENTIONS

OPERATORS

arg min \( f(z) \): value of \( z \) that minimizes \( f(z) \)
\( z \)
cov(\( z \)): covariance matrix of random vector \( z \)
plim: probability limit
\( E(\cdot) \): expectation
\( \text{tr}(A) \): trace (sum of diagonal elements) of the matrix \( A \)
A\(^T\): transpose of matrix \( A \)
A\(^{-1}\): inverse of matrix \( A \)
\( (A'A)^{-1}A' \)
\( |z| \): norm of a vector = \( \sqrt{z'z} \)

SYMBOLS USED IN TEXT

d: dimension of the largest to be considered model structure
d_e: dimension of the true model structure
e: vector of process disturbances (except in section 3.7.2)
I: identity matrix (except in section 3.7.2)
L(\cdot): loss function
\( \text{LH}(\cdot) \): likelihood function
M: model structure
\( \text{MSE}(\cdot) \): mean square error function
n: number of output samples
p: dimension of the current model structure
\( p(\cdot) \): (joint) probability density function
\( p(\cdot|\cdot) \): conditional (joint) probability density function
\( \text{RM}(\cdot) \): risk matrix function
S_f: shrinkage factor
U: known design matrix
w: vector of model residuals
x: vector of undisturbed output samples
y: vector of disturbed output samples
Z: transformed U, used in canonical forms
\( \beta \): transformed \( \theta \), used in canonical forms
\( \theta \): vector of process parameters
\( \hat{\theta} \): vector used to parametrize models
\( \sigma^2 \): variance of process disturbances
\( \cdot \): used to indicate that this is an estimator
1. INTRODUCTION

The linear regression model is one of the most frequently used models in statistics and there are applications of it in many different areas. In matrix notation the model can be expressed as:

\[ y = U\theta + w \]  \hspace{1cm} (1.1)

where,

- \( y \): \( n \)-vector of observable random variables
- \( U \): known matrix of dimension \( n \times d \)
- \( \theta \): \( p \)-vector of unknown regression coefficients
- \( w \): \( n \)-vector of unknown random variables (disturbances)

In the field of system identification the linear regression model is often used for impulse response estimation. The model is especially useful when there is no prior information about a relationship between the Markov parameters. Let the input-output behaviour of a process be given by:

\[ y(k) = \sum_{i=0}^{d_c} h(i)u(k-i) + e(k) \] \hspace{1cm} (1.2)

where,

- \( y(k) \): disturbed output
- \( u(k) \): undisturbed input
- \( e(k) \): disturbance
- \( h \): impulse response
- \( d_c \): length of (correct) impulse response

Then, given the data set \( (y(1), \ldots, y(n), u(2-p), \ldots, u(n)) \) of measured input-output samples, we can write the input-output relation of the process in the following form:

\[
\begin{bmatrix}
  y(1) \\
  \vdots \\
  y(n)
\end{bmatrix}
= \begin{bmatrix}
  u(1) & u(2-p) \\
  \vdots & \ddots \\
  u(n) & u(n-p+1)
\end{bmatrix}
\begin{bmatrix}
  h(0) \\
  \vdots \\
  h(p-1)
\end{bmatrix}
+ \begin{bmatrix}
  e(1) \\
  \vdots \\
  e(n)
\end{bmatrix}
\] \hspace{1cm} (1.3)

\[ y = U\theta + e \]

\[ y = x + e \]

So we see that \( U \) is filled with the input samples and becomes a Toeplitz matrix. It is assumed that there are enough output samples \( (n \geq d_c) \). The input signal is supposed to be sufficiently rich, so that \( U \) has full rank. The vector \( y \) consists of the output samples and \( \theta \) becomes the unknown impulse response. The \( n \)-vector \( x \) denotes the undisturbed output.

Throughout this report it is assumed that the disturbances \( e_i \) are
independent, have zero mean, have variance $\sigma^2$ and are normally distributed (for some definitions see appendix A). Thus,

$$e_i \sim \mathcal{N}(0, \sigma^2)$$  \hspace{1cm} (1.4)

so,

$$e \sim \mathcal{N}(0, \sigma^2 I_n)$$  \hspace{1cm} (1.5)

$$y \sim \mathcal{N}(U\theta, \sigma^2 I_n)$$  \hspace{1cm} (1.6)

Here $I_n$ denotes an identity matrix of dimension $n \times n$.

The objective is to use this linear regression model (1.1) to estimate the impulse response $\theta$. One usually estimates this impulse response by minimizing the difference between the measured output $y$ and the estimated output $x$. This results in an estimator of the following form:

$$\hat{\theta} = (U'U)^{-1}U'y = U^+y$$  \hspace{1cm} (1.7)

When the length of the actual impulse response is known this estimator is a minimum variance unbiased estimator of $\theta$. But in practice the length of the impulse response is unknown, in fact it may be infinite. In such situations we are forced to estimate the dimension of the model. Then we are usually dealing with an incorrect model.

The aim of this study is to investigate what the consequences of incorrect modelling are and to evaluate criteria for selecting a model. Furthermore I will briefly mention some ways to improve the least square estimator. Then it is assumed that the process is in the model set.
2. CONSEQUENCES OF (IN)CORRECT MODELING

In this chapter I will introduce the least squares estimator, define some functions that are going to be used to measure the performance of an estimator and discuss the properties of the least squares estimator when the process is (in)correctly modelled. The properties that will be derived are dependent on the assumption that the model is selected without reference to the actual data. However, since this is normally not the case the results should be used with caution. I will return to this point later.

2.1 THE LEAST SQUARES ESTIMATOR

As already mentioned, the objective is to obtain an estimation of the unknown Markov parameters. Throughout this chapter we restrict ourselves to the estimator of \( \theta \) that minimizes the residual sum of squares (RSS):

\[
\text{RSS} = (y - \hat{x})' (y - \hat{x})
\]

Here \( \hat{x} \) denotes the estimated undisturbed output \( x \) (\( \hat{x} = U^* \)). Thus given the model \( y = U^* + w \), where \( \theta \) is left as a variable and \( w \) is the corresponding vector of residuals (so RSS=\( w'w \)), minimizing of RSS results in the optimizing condition:

\[
\frac{\delta \text{RSS}}{\delta \theta} = 2(U'U)^{-1}U'y = 0
\]

When \( U \) has full rank this leads to the optimizing vector,

\[
\hat{\theta} = (U'U)^{-1}U'y
\]

We call the estimator given in (2.3) the least squares estimator of \( \theta \) (under the loss function RSS).

2.2 PERFORMANCE MEASURES

If one has some estimate of \( \theta \), say \( \hat{\theta} \), it is desirable to know how good that estimator is. I will now discuss the form of functions that are to be used in evaluating the performance of an estimator. Although there are many alternatives, I will only discuss the mean square error (MSE) function and the risk matrix (RM).

First of all I will discuss the mean square error function for measuring the distance between the actual impulse response and the estimated impulse response.

Let the squared difference between \( \theta \) and \( \hat{\theta} \) be denoted as \( L(\theta) \). Thus:

\[
L(\theta) = (\hat{\theta} - \theta)'(\hat{\theta} - \theta)
\]

Then the corresponding mean square error function is:

\[
\text{MSE} = \mathbb{E}(L(\theta))
\]
Rewrite $\text{MSE}(\theta)$ as,

$$\text{MSE}(\theta) = (E(\hat{\theta}) - \theta)^T (E(\hat{\theta}) - \theta) + E( (\hat{\theta} - E(\hat{\theta}))^T (\hat{\theta} - E(\hat{\theta})) ) \quad (2.6)$$

Thus we see that this loss function consists of a bias part (first term) and a variance part (second term).

Sometimes, for example when prediction is the objective, it is desirable to measure the performance of an estimator by looking how well the undisturbed output $X$ has been estimated. In that case the $\text{MSE}$ function will be defined as follows.

Let the squared difference between $X$ and $\hat{X}$ be denoted as $L(X)$. Thus:

$$L(X) = (X - \hat{X})^T (X - \hat{X}) \quad (2.7)$$

Then the corresponding mean square error function is,

$$\text{MSE}(X) = E(L(X)) \quad (2.8)$$

Similar to (2.6) we can split this function into a bias part and a variance part.

To establish a relation between (2.5) and (2.8) let us consider the risk matrix of the estimator $\theta$.

$$\text{RM}(\theta) = E((\hat{\theta} - \theta)(\hat{\theta} - \theta)^T) \quad (2.9)$$

Then it can be seen that,

$$\text{MSE}(\theta) = \text{tr}(\text{RM}(\theta)) \quad (2.10)$$

$$\text{MSE}(x) = \text{tr}(U^* \text{RM}(\theta) * U') \quad (2.11)$$

Suppose that there are two estimators $\hat{\theta}_a$ and $\hat{\theta}_b$ of $\theta$. Then we say that for a given performance measure $\hat{\theta}_a$ is a better estimator than $\hat{\theta}_b$ when the following condition holds,

a: for $\text{MSE}(\theta)$ if $\text{MSE}(\theta, \hat{\theta}_a) < \text{MSE}(\theta, \hat{\theta}_b)$ \quad (2.12)

b: for $\text{MSE}(x)$ if $\text{MSE}(x, \hat{x}_a) < \text{MSE}(x, \hat{x}_b)$ \quad (2.13)

c: for $\text{RM}(\theta)$ if $\text{RM}(\theta, \hat{\theta}_a) - \text{RM}(\theta, \hat{\theta}_b)$ is positive semidefinite (p.s.d) and nonzero. \quad (2.14)

It is easy to see that if condition (2.14) is true, then conditions (2.12) and (2.13) are also true (because of (2.10) and (2.11)). But the reverse doesn't hold. If condition (2.12) and/or (2.13) are true, the condition (2.14) doesn't have to be true. Thus the condition (2.14) is much more severe than condition (2.12) and (2.13).
2.3 THE PROCESS IS IN THE MODEL SET

I will now discuss some properties of the least squares estimator when the process is in the model set. Suppose that the length of the actual impulse response is $d_c$ (the subscript $c$ stands for correct). Let the length of the impulse response in the model be equal to $p$ where $p \geq d_c$. Then we can write,

$$\text{process } y = U_f \theta_f + e = U_f \theta_c + e \quad \text{with } \theta_c = 0 \quad (2.15)$$

$$\text{model } y = U_f \phi_f + w = U_f \phi_c + w$$

where,

- $U_f = (U U_r)$ is an $n \times p$ matrix
- $U = n \times d_c$ matrix
- $U_r = n \times (p-d_c)$ matrix
- $\theta_f = (\theta' \ \theta_r')'$ and $\phi_f = (\phi' \ \phi_r')'$ are $p$-vectors
- $\theta$ and $\phi$ are $d_c$-vectors
- $\theta_c = 0$ and $\phi_c$ are $(p-d_c)$-vectors

Since $\theta_f$ is not known we leave it in the model as a variable ($\phi_f$). Then $w$ is the corresponding $n$-vector of residuals.

CASE ($p=d_c$):

First, let us assume that $p=d_c$ so $\theta_f = \theta$. Because the process is in the model set and we assumed to know the exact probability density function of $e$, we can derive the maximum likelihood estimator (MLE) of $\theta$.

Because,

$$y \in N(\theta_0, \sigma^2 I_n)$$

we can write the probability of obtaining the observed vector $y$ as,

$$p(y; \theta, \sigma^t) = (2\pi\sigma^t)^{-n/2} \exp \left\{ -(y-U\theta)^t(y-U\theta)/2\sigma^t \right\} \quad (2.16)$$

Because $\theta$ is unknown we leave it as a variable ($\theta$). Then we obtain the likelihood function $L_H(\theta | y)$ for $\theta$,

$$L_H(\theta | y) = p(y; \theta) \quad (2.17)$$

To obtain the maximum likelihood estimator of $\theta$ we have to optimize $L(\theta | y)$. This results in:

$$\theta_{ml} = \arg \max_\theta L_H(\theta | y) = (U'U)^{-1}U'y = U'y \quad (2.18)$$

Thus we see that the MLE of $\theta$ is equal to the least squares estimator of $\theta$ (under the loss function $L_y$). Because the process is in the model set the following properties of the LSE $\hat{\theta} = U'y$ can be derived:
PROPERTIES

a: unbiased, \( \text{E}(\hat{\Theta}) = \Theta \)  

\[ (2.19) \]

b: \( \text{cov}(\hat{\Theta}) = \text{E}( (\hat{\Theta} - \text{E}(\hat{\Theta}))(\hat{\Theta} - \text{E}(\hat{\Theta}))') = \sigma^2(U'U)^{-1} \)  

\[ (2.20) \]

c: minimum variance unbiased estimator  

(see Arnold '81)  

\[ (2.21) \]

d: \( \hat{\Theta} \in \mathcal{N}(\Theta, \sigma^2(U'U)^{-1}) \) because \( \hat{\Theta} \) is a linear combination of \( \gamma \) and (a) and (b) hold.  

\[ (2.22) \]

Thus our performance measures become:

\[ e: \text{MSE}(\Theta) = \sigma^2 \text{tr}((U'U)^{-1}) \]  

\[ (2.23) \]

\[ f: \text{MSE}(\hat{\Theta}) = \sigma^2 \text{tr}((U(U'U)^{-1}U') = \sigma^2 \sigma_i \]  

\[ (2.24) \]

g: \( \text{RM}(\hat{\Theta}) = \sigma^2(U'U)^{-1} \)  

\[ (2.25) \]

In the derivation of the maximum likelihood estimator of \( \Theta \) we assumed that the variance of the noise \( \sigma^2 \) was known. In practice, this is usually not the case. Then we have to estimate \( \sigma^2 \). The likelihood function (LM) for \( \Theta \) and \( \sigma^2 \) can be written as:

\[ LH(\Theta, \sigma^2 | y) = p(y; \Theta, \sigma^2) \]  

\[ (2.26) \]

Here \( \Theta \) and \( \sigma^2 \) are left as variables \( \Theta \) and \( \sigma^2 \). To find the maximum likelihood estimator of \( \Theta \) and \( \sigma^2 \) we have to optimize (2.26). Thus,

\[ \hat{\Theta}, \hat{\sigma}^2 = \arg \max_{\Theta, \sigma^2} LH(\Theta, \sigma^2 | y) \]  

\[ (2.27) \]

This leads to the optimizing estimators:

\[ \hat{\Theta} = U^+ y \quad \text{and} \quad \hat{\sigma}^2 = (y' - \hat{x})' (y' - \hat{x}) / n = \hat{e}' \hat{e} / n \]  

\[ (2.28) \]

The with the estimator \( \hat{\Theta} \) corresponding residuals will be denoted as \( \hat{e} \). Thus the estimate of \( \Theta \) remains the same. The estimate of \( \sigma^2 \) can be written as,

\[ \hat{\sigma}^2 = \frac{||x||^2}{n} \]  

\[ (2.29) \]

Here \( ||x||^2 \) is defined as \( x'x \). Then we see that \( \phi^2 \) is a biased estimate of \( \sigma^2 \) because,

\[ \text{E}(\phi^2) = \text{E}(||\gamma||^2 - ||uu^+y||^2) / n \]

\[ = \text{E}(||\gamma||^2 / n - \text{E}(||uu^+y||^2) / n \]

\[ = \sigma^2 (n-d_c) / n \]  

\[ (2.30) \]

Consider now the following estimate of \( \sigma^2 \),

\[ \hat{\sigma}^2 = \frac{\hat{e}'n / (n-d_c) - \frac{1}{n} (1_n'uu^+) y^2 / (n-d_c)}{n} \]

\[ (2.31) \]

This estimator has the following properties,
a: unbiased (see (2.30) and (2.31)) \hfill (2.32)
b: minimum variance unbiased estimator (see Arnold '81) \hfill (2.33)
c: \hat{\sigma}^2 and \hat{\theta} are independent, because \hfill (2.34)
\hat{e}=(I_n-UU')y and \hat{\theta}=U'y are normally distributed with covariance \hfill E((u-E(u))(\hat{\theta}-E(\theta)')) = \Omega_n,dc

Let P=(I_n-UU'). Because P is an orthogonal projector on the null space of U', it has \((n-d_c)\) eigenvalues which are one and \(d_c\) eigenvalues which are zero. Therefore \(\hat{e}'P\hat{e}\) can be written as \(e'V'Ve\), where \(V\) is an \(n\times n\) orthonormal matrix and \(T\) is an \(n\times n\) diagonal matrix whose diagonal elements are the eigenvalues of \(P\). If we rewrite \(e'V'Ve\) as \(r'Tr\), where \(r \sim N(0,\sigma^2I_n)\), it can be seen that \((n-d_c)\sigma^2/\sigma^2\) is distributed as a Chi^2-distribution (for some distributions see appendix A) with \((n-d_c)\) degrees of freedom, thus
\(n-d_c)\sigma^2/\sigma^2 \sim \text{Chi}^2(n-d_c) \hfill (2.35)

CASE \(p > d_c\):

So far we discussed the case when the length of the actual impulse response was known. Let us now discuss what happens when the model is chosen too large, thus \(n > p > d_c\). In that case our estimator of \(\theta_f = (\theta_0, \theta_r)'\) becomes,
\[ \hat{\theta}_f = (U_f'U_f)^{-1}U_f'y \hfill (2.36) \]

This estimator has the following properties,

PROPERTIES

a: unbiased, \(E(\hat{\theta}_f) = \theta_f \hfill (2.37)\)
b: \(\text{cov}(\hat{\theta}_f) = \sigma^2(U_f'U_f)^{-1} \hfill (2.38)\)
c: \(\hat{\theta}_f \sim N(\theta_f, \sigma^2(U_f'U_f)^{-1})\), because of (a) \& (b) \hfill (2.39)
d: \(\text{MSE}(\theta_f) = \sigma^2 \text{tr}((U_f'U_f)^{-1}) \hfill (2.40)\)
e: \(\text{MSE}(x) = \sigma^2p \hfill (2.41)\)
f: \(\text{RM}(\theta_f) = \sigma^2(U_f'U_f)^{-1} \hfill (2.42)\)

However, this estimator is no longer a minimum variance estimator, because if we compare the covariance matrix of the estimator
\[ \hat{\theta}_f = (U_f'y, 0')', \hfill (2.43) \]
given by,
\[
\begin{bmatrix}
\sigma^2 \left( U'U \right)^{-1} & 0_{d_c, p-d_c} \\
0_{p-d_c, d_c} & 0_{p-d_c, p-d_c}
\end{bmatrix}
\]  \hspace{1cm} (2.44)

(see previous section), with the covariance matrix of \( \hat{\Theta}_f \) as given in (2.39), we see that the difference between these two matrices,

\[
\text{cov}(\hat{\Theta}_f) - \text{cov}(\hat{\Theta}_f) = \sigma^2 \begin{bmatrix}
(U^*U_f)(U^*U_f)' - (U^*U_f)B & -B(U^*U_f)'
-B(U^*U_f) & B
\end{bmatrix}
\]  \hspace{1cm} (2.45)

with \( B = (U_f' I_{n-p} U_f)^{-1} \) (see appendix B)

is at least positive semidefinite, since \( B \) is assumed to be positive definite \( (U_f \) has full rank).

The estimator of \( \sigma^2 \) (see (2.31)) where \( U \) and \( d_c \) are replaced by \( U_f \) and \( p \) is in this larger model still unbiased and independent of \( \hat{\Theta}_f \), but also in this case it is no longer a minimum variance estimator because of reasons similar to (2.45).

### 2.4 THE PROCESS IS NOT IN THE MODEL SET

In the previous section we discussed the situation that the process is in the model set. Let us now assume that the process is not in the model set. Thus we assume that the dimension of the model is chosen too small. Then we get,

\[
\text{process } \gamma = U_1 \hat{\Theta}_1 + U_2 \hat{\Theta}_2 + e = U \hat{\Theta} + e
\]

\[
\text{model } \gamma = U_1 \hat{\Theta}_1 + w = U_1 \hat{\Theta}_1 + U_2 \hat{\Theta}_2 + w = U \hat{\Theta} + w \text{ so } \hat{\Theta}_2 = 0
\]

Here we assume that:

- \( U = (U_1, U_2) \) is an \( n \times d_c \) matrix
- \( U_1 \) is an \( n \times p \) matrix
- \( U_2 \) is an \( n \times (d_c - p) \) matrix
- \( \hat{\Theta} = \left( \hat{\Theta}_1, \hat{\Theta}_2 \right)' \) and \( \hat{\Theta} = \left( \hat{\Theta}_1, \hat{\Theta}_2 \right)' \) are \( d \)-vectors
- \( \hat{\Theta}_1 \) and \( \hat{\Theta}_1 \) are \( p \)-vectors
- \( \hat{\Theta}_2 \) and \( \hat{\Theta}_2 = 0 \) are \( (d_c - p) \)-vectors

Although one may argue about the type of estimator to use in this situation, we take the least squares estimator as an estimate of \( \hat{\Theta}_1 \), since we are investigating the consequences of incorrect modelling when using this estimator, thus,

\[
\hat{\Theta}_1 = (U_1' U_1)^{-1} U_1' y = U_1^+ y ; \hat{\Theta}_2 = 0
\]  \hspace{1cm} (2.47)

so,
\[
\hat{\Theta} = (\hat{\Theta}_1' \ 0')'
\]  
(2.48)

Then the following properties can be derived:

**Properties**

a: biased, \( E(\hat{\Theta}_1) = \Theta_1 + U_1^T \Theta_2 \) \( E(\hat{\Theta}_2) = 0 \)  
(2.49)

b: \( \text{cov}(\hat{\Theta}_1) = \sigma^2(U_1^T U_1)^{-1} \)  
(2.50)

c: \( \hat{\Theta}_1 \sim N(\Theta_1 + U_1^T \Theta_2, \sigma^2(U_1^T U_1)^{-1}) \)  
(2.51)

Thus our performance measures become now,

d: \( \text{MSE}(\Theta) = |\Theta_2|^2 + |U_1^T \Theta_2|^2 + \sigma^2 \text{tr}(U_1^T U_1)^{-1} \)  
(2.52)

e: \( \text{MSE}(x) = |(1_n - U_1^T U_1)^T U_2 \Theta_2|^2 + \sigma^2 p \)  
(2.53)

\[
\text{RM}(\Theta) =
\begin{bmatrix}
\sigma^2(U_1^T U_1)^{-1} + (U_1^T U_2 \Theta_2)(U_1^T U_2 \Theta_2)' & -U_1^T U_2 \Theta_2 \\
-U_2^T (U_1^T U_2 \Theta_2)' & \Theta_2^T \Theta_2
\end{bmatrix}
\]  
(2.54)

(see appendix B)

We will now investigate the effect of the too small model on the estimate of \( \sigma^2 \). Suppose \( \sigma^2 \) is estimated with the estimator given in (2.31), where \( U \) and \( d_e \) are replaced by \( U_1 \) and \( p \). Then, in our incorrect model \( \sigma^2 \) is estimated by,

\[
\hat{\sigma}^2 = |(1_n - U_1^T U_1)^T y|^2 / (n-p)
\]  
(2.55)

But because,

\[
E(\hat{\sigma}^2) = |(1_n - U_1^T U_1)^T U_2 \Theta_2|^2 / (n-p) + \sigma^2
\]  
(2.56)

we see that this estimator is normally upwards biased.

### 2.5 Evaluation

In the previous sections we considered the consequences of (in)correct modelling. The aim of this section is to evaluate these consequences.

Suppose our estimated model is too large. As we saw in (2.46), the difference between the covariance matrices of the too large model and the correct model is positive semidefinite (and is non zero). Thus, although
too big a model set leads to unbiased estimates, one pays with greater variance. Consequently, since in this case the risk matrices are equivalent with the covariance matrices, our three performance measures \((2.5) \ (2.8)\) and \((2.9)\) will prefer the estimators from the correct model.

Let us now suppose that the chosen model is too small. If we compare the estimate of \(\Theta\) in the too small model \((\Theta)\) with the estimate of \(\Theta\) in the correct model \((\Theta = U^T y)\) we can see that the difference between the risk matrices of both models is given by,

\[
RM(\Theta, \Theta) - RM(\Theta, \Theta) = \begin{bmatrix}
U_1^TU_2(\sigma^2B-\Theta_2^T)(U_1U_2)^T & -U_1^TU_2(\sigma^2B-\Theta_2^T) \\
-(\sigma^2B-\Theta_2^T)(U_1U_2)^T & (\sigma^2B-\Theta_2^T)
\end{bmatrix}
\] \(2.57\)

with \(B = (U_2^T(I_n-U_1U_1^T)U_2)^{-1}\) (see appendix B)

Here we see that when,

\[
(\sigma^2B-\Theta_2^T) \text{ is p.d.,}
\] \(2.58\)

thus when the covariance matrix of estimating \(\Theta_2\) in the correct model minus the bias matrix of \(\Theta_2\) is p.d., the difference between the risk matrices is at least p.s.d. (see appendix B). The condition \((2.58)\) is equivalent to,

\[
(\sigma^2B-\Theta_2^T) \text{ is p.d.} \iff \Theta_2B^{-1}\Theta_2 / \sigma^2 < 1
\] \(2.59\)

(see appendix B)

If condition \((2.59)\) is true, it means that our three performance measures indicate that we can get a better estimate in the incorrect model structure. But even when \((2.59)\) doesn't hold our estimates can become better in the incorrect model structure according to the mean square error functions if,

\[
MSE(\Theta, \Theta) - MSE(\Theta, \Theta) = \]

\[
|\Theta_2| + |U_1^TU_2^T| + \sigma^2(\text{tr}(U_1^TU_1)^{-1}) \cdot \text{tr}(U_1^TU_1)^{-1} < 0
\] \(2.60\)

or,

\[
MSE(x, \tilde{x}) - MSE(x, \tilde{x}) =
\]

\[
|(I_n-U_1U_1^T)U_2^T| + \sigma^2(p-d_c) < 0
\] \(2.61\)

From these evaluations we can conclude that too large a model always gives worse estimates because of the increasing variance, but too small a model can give better estimates when the introduced bias is smaller than the amount of decreasing variance.

Our primary goal was to get a good estimate of the impulse response...
according to a particular performance measure. Thus, although the length of
the actual impulse response may be known, it may be better to choose a
smaller model. In that case we define the best model as the model that
minimizes:

\[
MSE(\theta) : \| \theta_2 \|^2 + \| U_1^+ U_2 \theta_2 \|^2 + \sigma^2 \text{tr}( (U_1^+ U_1)^{-1} )
\]  
\[
MSE(x) : \| (I_n - U_1 U_1^+ ) U_2 \theta_2 \|^2 + \sigma^2 p
\]

If the performance measure is the risk matrix, it is more difficult to
define a best model. Therefore we will define a set of 'good' models rather
than a best model. This set is defined as follows:

Let \( d_{\text{max}} \) be the dimension of the model with,

\[
RM(\theta, \hat{\theta}_p) - RM(\theta, \hat{\theta}_{d_{\text{max}}}) \text{ is p.s.d. and non zero for } p > d_{\text{max}}
\]

Let \( d_{\text{min}} \) be the dimension of the model with,

\[
RM(\theta, \hat{\theta}_p) - RM(\theta, \hat{\theta}_{d_{\text{min}}}) \text{ is p.s.d. and non zero for } p < d_{\text{min}}
\]

Where \( \hat{\theta}_p, \hat{\theta}_{d_{\text{max}}}, \) and \( \hat{\theta}_{d_{\text{min}}} \) are the estimates of \( \theta \) as given in (2.48) in the
model with dimension \( p, d_{\text{max}} \) and \( d_{\text{min}} \). Then the set of 'good' models is
defined as the set of models with dimension \( d_{\text{min}} \leq p \leq d_{\text{max}} \).

Since we used the least square estimator to estimate \( \hat{\theta}_1 \), all above
"optimal models" are confined to the class of least square estimates.

In order to find the best model we have to cope with the fact that the
actual impulse response is unknown. Logically, because that is what we
wanted to estimate. Thus we are forced to consider estimates of the
dimension of the model. In the next chapters I will discuss some criteria
that have been proposed for finding an 'optimal model'.
3. CRITERIA FOR SELECTING AN 'OPTIMAL' MODEL STRUCTURE

In this chapter I will present a review of criteria that have been proposed for selecting the dimension of a linear regression model.

3.1 NOTATION

Throughout this chapter we consider a set of possible model structures. This set consists of models with increasing dimension. Let the highest dimension to be considered be \( d \). Then there are \( d+1 \) possible models which vary in dimension from 0 to \( d \). Furthermore it is assumed that the length of the actual impulse response (\( d_c \)) is smaller than \( d \). Let the dimension of the varying model be denoted as \( p \), then we write,

\[
\text{process: } y = U\theta + e = U_1\theta_1 + U_2\theta_2 + e \quad \text{where elements of } \theta_2 \text{ may be zero}
\]

\[
\text{model: } y = U_1\theta_1^* + w \quad \text{with } \theta_1^* = U_1^* y \text{ and } e = (I_n - U_1 U_1^*) y
\]

where

- \( U = (U_1 \quad U_2) \) is an \( n \times d \) matrix
- \( U_1 \) is an \( n \times p \) matrix
- \( U_2 \) is an \( n \times (d-p) \) matrix
- \( \theta = (\theta_1 \quad \theta_2)^t \) is a \( d \times 1 \) vector
- \( \theta_1 \) is a \( p \times 1 \) vector
- \( \theta_2 \) is a \( (d-p) \times 1 \) vector

Let the residual sum of squares (RSS\( p \)) in the model with dimension \( p \) be defined as,

\[
\text{RSS}_p = e'e = (y^*x)'(y^*x) = \left\| (I_n - U_1 U_1^*) y \right\|^2
\]

Since in practice \( \sigma^2 \) is usually not known, I will define 4 different estimates of \( \sigma^2 \) when working with a model of dimension \( p \).

a: \( \hat{\sigma}^2(p) = \text{RSS}_p/(n-p) \)

b: \( \hat{\sigma}^2(p) = \text{RSS}_p/n \)

c: \( \hat{\sigma}^2(d) = \text{RSS}_d/(n-d) \)

d: \( \hat{\sigma}^2(d) = \text{RSS}_d/n \)

All these estimators and \( \hat{\sigma}^2 \) are independent (see 2.34). The estimators (b) and (d) are the MLE estimates of \( \sigma^2 \) when the dimension of the model is respectively \( p \) and \( d \), provided that \( d \geq p > d_c \). The estimators (a) and (c) are the corrected estimates of \( \sigma^2 \) (see 2.31) in the model of dimension \( p \) and \( d \). As we already saw the estimator (a) is usually upwards biased for \( p < d_c \) (see 2.47).
3.2 Model Structure Selection

In the field of system identification it is very popular to perform a whiteness test of the residuals or to look at the behaviour of the residuals loss function in order to select a model structure. These methods can be described as subjective ones because the outcome of these methods highly depend on the interpretation of a particular plot. For example, in case of linear regression models a frequently used plot is the plot of $R_p^2$ versus $p$. Here $R_p^2$ is the squared multiple correlation coefficient and it is defined as,

$$R_p^2 = 1 - \frac{RSS_p}{y'y}$$ (3.7)

This plot may yield a locus of maximum $R_p^2$ which remains quite flat as $p$ is decreased and then turns sharply downward. The value of $p$ at which this 'knee' is, is used to indicate the dimension of the selected model. Unfortunately this knee isn't so clear when the signal to noise ratio is bad. Then it becomes very difficult to determine an 'optimal' model.

Another 'subjective' method is the plot of the residual mean square ($RMS_p$) versus $p$, where

$$RMS_p = \frac{RSS_p}{n-p} = \sigma^2(p)$$ (3.8)

As we saw in the previous chapter the expectation of $RMS_p$ is equal to $\sigma^2$ when the process is in the model set. If not, it is usually biased upwards. When interpreting this plot the choice of $p$ is based on (see Hocking '76):

- $a$: minimum $RMS_p$
- $b$: the value of $p$ such that $RMS_p = RSS_d$
- $c$: the value of $p$ where the $RMS_p$ increases sharply for further reduction of $p$

An unfortunate aspect of these two measures is that they do not consider the gain of improved estimation when using an incorrect model.

In the remainder of this chapter I will only discuss procedures for selecting the dimension of a linear regression model which yield a solution that is free from interpretation. This means that the model structure is chosen that optimizes a particular criterion.

In the last two decades, many criteria have been proposed for model structure selection (for good reviews see Hocking '76, Thompson '78, Amemiya '80, Judge et al. '80). The aim of this chapter is to summarize some of these criteria to have some guidelines for deciding where to cut off the tail of an impulse response.

Most of the selection procedures to be discussed have a very different background. There are criteria which are designed on hypothesis arguments, on prediction arguments, information theoretic arguments, bayesian arguments, cross-validation arguments or on Stein arguments. In the following I will discuss each of these arguments.
3.3 HYPOTHESIS ARGUMENTS

3.3.1 CONFIDENCE INTERVALS

As we already saw in chapter 2, the following holds for the full model.

\[ y \sim N(\mathbf{0}, \sigma^2 I_n) \]
\[ \hat{\theta} \sim N(\theta, \sigma^2 (U'U)^{-1}) \]
\[ (n-d)\hat{\sigma}^2(d)/\sigma^2 \sim \chi^2(n-d) \] (3.9a)

Thus we see that,
\[ (\hat{\theta} - \theta)'(U'U)(\hat{\theta} - \theta)/\sigma^2 \sim \chi^2(d) \] (3.9b)

because \( \sigma^2 \) is unknown and we estimate it with \( RSS_d/(n-d) \). Since \( RSS_d/(n-d) \) and \( \theta \) are independent, we get from the ratio of 3.9a and 3.9b,
\[ (\hat{\theta} - \theta)'(U'U)(\hat{\theta} - \theta)/\hat{\sigma}^2(d) \sim F(d, n-d) \] (3.10)

where \( F(d, n-d) \) denotes a F-distribution with \( d \) and \( n-d \) degrees of freedom (see appendix A).

Suppose we estimate the markov parameters from the full model. If we are only interested in the last \( (d-p) \) parameters, we can see that,
\[ (\hat{\theta}_2 - \theta_2)'(U_2'U_2)(\hat{\theta}_2 - \theta_2)/((d-p)\hat{\sigma}^2(d)) \sim F(d-p, n-d) \] (3.11)

It is now possible to derive a 100(1-\( \alpha \))% joint confidence region for \( \theta_2 \). This region is defined as the region in which \( \theta_2 \) lies with probability 100(1-\( \alpha \)). Because of (3.11) we see that a 100(1-\( \alpha \))% confidence region of \( \theta_2 \) (\( R(\theta_2) \)) can be defined as,
\[ R(\theta_2) = \left\{ \theta_2 \mid (\hat{\theta}_2 - \theta_2)'(U_2'U_2)(\hat{\theta}_2 - \theta_2)/((d-p)\hat{\sigma}^2(d)) < F^a(d-p,n-d) \right\} \] (3.12)

where \( F^a(d-p,n-d) \) is the solution of
\[ P \left( x < F^a(d-p,n-d) \right) = 1-\alpha \]
\[ x \in F(d-p,n-d) \] (3.13)

It is not difficult to construct a joint confidence region for 2 parameters, but if the number of parameters exceeds 2 the problem is somewhat harder to solve.

3.3.2 HYPOTHESIS TESTING

Let us assume that we estimate the parameters in the full model and we wish to test if \( \theta_2 = 0 \), so that a smaller model structure can be used to
estimate θ. Thus we have the following hypothesis,

\[
\begin{align*}
H_0 & \quad \theta_2 = 0 \\
H_1 & \quad \theta_2 \neq 0
\end{align*}
\]

In hypothesis testing we assume that the null hypothesis (H0) is true unless there is a convincing evidence that the hypothesis H1 is true. We should like to make a decision between H0 and H1 such that the risk of rejecting H0 when it is true is less than a certain number α.

A commonly used test in system identification (see Söderström '77) is the F-test. Suppose that H0 is true, thus \( \theta_2 = 0 \). Then we see with,

\[
\begin{align*}
\text{RSS}_p - \text{RSS}_d &= \| (I_n - u_1 u_1^*) y \|_1^2 - \| (I_n - u u^*) y \|_1^2 \\
&= \| u u^* y \|_1^2 - \| u_1 u_1^* y \|_1^2 \\
&= \| A A^* y \|_1^2
\end{align*}
\]

(3.14)

that \( (\text{RSS}_p - \text{RSS}_d) / \sigma^2 \) is distributed as,

\[
(\text{RSS}_p - \text{RSS}_d) / \sigma^2 \sim \chi^2(d-p)
\]

(3.15)

And because \( \text{RSS}_p - \text{RSS}_d \) and \( \text{RSS}_d \) are independent we see that,

\[
(\text{RSS}_p - \text{RSS}_d) / \| (d-p) \delta^2(d) \| \sim F(d-p,n-d)
\]

(3.16)

When H0 is not true this expression is distributed as a noncentral F distribution. In this F-test we accept H0 and thus take the smaller model structure to estimate \( \theta \) if,

\[
(\text{RSS}_p - \text{RSS}_d) / \| (d-p) \delta^2(d) \| < F^0(d-p,n-d)
\]

or equivalently,

\[
\| A A^* y \|_1^2 / \| (d-p) \delta^2(d) \| < F^0(d-p,n-d)
\]

(3.17)

This test is called a partial F-test, because it measures the contribution of regressor \( \theta_2 \) given that the other regressor \( \theta_1 \) is in the model set.

In the previous section we defined the joint confidence region of \( \theta_2 \) when we estimated \( \theta \) in the full model. Testing if \( \theta_2 \) could be set to zero by looking if the null-vector belongs to the joint confidence region of \( \theta_2 \) can reduce the number of parameters, however, \( \theta_1 \) would still be estimated in the full model. To establish a relation between the partial F-test and the joint confidence region, let us consider the following. A necessary condition for \( \theta_2 = 0 \) to belong to this region is (see 3.12),

\[
\theta_2 \in R(\theta_2) \quad \text{if} \quad \hat{\theta}_2 u_2^* u_2 \hat{\theta}_2 / \| (d-p) \delta^2(d) \| < F^0(d-p,n-d)
\]

(3.18)

and thus if,

\[
(\| A A^* y \|_1^2 + \| u_1 u_1^* A A^* y \|_1^2) / \| (d-p) \delta^2(d) \| < F^0(d-p,n-d)
\]

(3.19)
because,
\[
\left( \left[ U_2 A^\dagger y \right]' \right) = \left( \left[ (I_n - U_1 U_1^\dagger) U_2 A^\dagger y \right]' + \left[ U_1 U_1^\dagger U_2 A^\dagger y \right]' \right)
\]  \hspace{1cm} (3.20)

This means that when $\Theta_2 = 0$ belongs to the joint confidence region of $\Theta_2$ the condition (3.17) holds and thus $H_0$ is accepted. But the reverse is not true. If $H_0$ is accepted according to condition (3.17) it does not have to mean that $\Theta_2$ belongs to this joint confidence region. This is only true if $U$ is orthogonal.

How can we use this partial F-test in the selection of a model. A possibility is to take the smallest value of $p$ as the final dimension of the model for which the test (3.17) holds. Thus we have to determine the largest dimension of the vector $\Theta_2$ for which the partial F-test gives satisfactory results.

An important problem with hypothesis testing is that a value for $\alpha$ has to be chosen. A small $\alpha$ causes many parameters to be deleted from the full model. A large $\alpha$ causes many parameters to be included. The most widely used value of $\alpha$ is 5-10%.

3.4 PREDICTION ARGUMENTS

3.4.1 MALLows' $C_p$

Let us suppose that our main concern is to get a good prediction of future responses. Considers the prediction of future responses from the same design matrix $U$ as is used in the estimation. As a measure of the goodness of this prediction consider the mean square error of prediction (MSEP).

\[
\text{MSEP}(y_p) = E( (y_p - \hat{y}_p)'(y_p - \hat{y}_p) )
\]  \hspace{1cm} (3.21)

Here $y_p$ denotes the vector of real future responses, $\hat{y}_p$ denotes the vector of estimated future responses. Because we consider prediction from our current matrix $U$ we can write,

\[
\hat{y}_p = \hat{x}
\]

and thus MSEP can be written as,

\[
\text{MSEP}(y_p) = E( (\hat{y}_p - \hat{x})'(\hat{y}_p - \hat{x}) ) + n\sigma^2
\]  \hspace{1cm} (3.22)

Thus minimization of MSEP means in this case that we would like to minimize $\text{MSE}(x)$. As we already saw in (2.53) we can write $\text{MSE}(x)$ as,

\[
\text{MSE}(x) = \left( (I_n - U_1 U_1^\dagger) \Theta_2 \right)' + n\sigma^2
\]  \hspace{1cm} (3.23)

Consider now the expectation of the residuals sum of squares.

\[
E( \text{RSS}_p ) = \left( (I_n - U_1 U_1^\dagger) \Theta_2 \right)' + (n-p)\sigma^2
\]  \hspace{1cm} (3.24)

As an estimate of $\text{MSE}(x)/\sigma^2$ Mallows (173) considered the following
function,

\[ \text{Cp} = \frac{\text{RSSp}}{\hat{\sigma}^2} + \frac{2p}{n} \]  

(3.25)

Here \( \hat{\sigma}^2 \) is some estimate of \( \sigma^2 \). The most practicable estimates of \( \sigma^2 \) are \( \hat{\sigma}^2(d) \) and \( \hat{\sigma}^2(d) \). When \( \hat{\sigma}^2 = \hat{\sigma}^2(d) \) \( \text{Cp} \) is an unbiased estimate of (3.23). So the model structure that minimizes \( \text{Cp} \) should be chosen. Mallows pointed out that it is better to inspect a plot of \( \text{Cp} \) versus \( p \) than blindly choose the model that minimizes \( \text{Cp} \). Models leading to smaller \( \text{Cp} \) are preferred, but points close to the line \( \text{Cp} = p \) are likely to be for models with a small bias. In view of our performance measures we see that this model selection procedure is clearly an procedure that is designed to improve the performance measure \( \text{MSE}(x) \).

3.4.2 Amemiya's PC

Amemiya (180) also considered the problem of improving prediction. Suppose we have a vector of future input samples \( u_f \) and we want to predict the response to this input sequence. With the predictor \( x_p = u_f^t \beta_1 \), the mean square error function of prediction can be written as,

\[ \text{E}( \gamma_o^t x_p^t ) = \sigma^2 \left[ 1 + u_f^t (U_1^t U_1)^{-1} u_f \right] + \left( u_f^t \theta - u_f^t U_1^t \theta \right)^2 \]  

(3.26)

Where \( \gamma_o \) is the actual future response. Amemiya suggested that one should take further expectations of (3.26) regarding \( u_f \) as the random variable that satisfies,

\[ \text{E}( u_f u_f^t ) = U_f U_f^t / n \]  

(3.27)

Then we obtain,

\[ \text{E}( \gamma_o^t x_p^t ) = \left[ (I_n - U_1 U_1^t) \gamma_2 \right]^t / n + \sigma^2 (1 + p/n) \]  

(3.28)

Given this risk function he considered two criteria based on different decision strategies. The first one he called the Prediction Criterion (PC). This criterion is obtained by replacing \( \sigma^2 \) with its estimator \( \hat{\sigma}^2(p) \) and then minimizing the first term of (3.28) by putting it to zero. Thus,

\[ \text{PC} = \hat{\sigma}^2(p)(1 + p/n) = \left( \frac{1}{n} \right) \text{RSSp}(1 + 2p/(n-p)) \]  

(3.29)

The second criterion is obtained by estimating the first term rather than eliminating it. Because of relation (3.24) the first term can be estimated with,

\[ \left[ (I_n - U_1 U_1^t) \gamma_1 \right]^2 / n - \sigma^2(n-p)/n \]  

(3.30)

When we substitute this in (3.28) the second criterion becomes,

\[ \text{RSSp}/n + 2p \hat{\sigma}^2/n \]  

(3.31)
Because we have to estimate \( \sigma^2 \) we see that then \( \sigma^2 \) is estimated with \( \hat{\sigma}^2(d) \) or \( \hat{\sigma}^2(d) \) minimizing this criterion is equivalent to minimizing Mallows \( \text{Cp} \) when the same estimator of \( \sigma^2 \) is used. However when \( \sigma^2 \) is estimated with \( \hat{\sigma}^2(p) \) we see that this second criterion becomes exactly \( \text{PC} \).

Amemiya pointed out that his \( \text{PC} \) can be used as a selection criterion either in linear regression models with the error term having a general variance-covariance matrix or in the nonlinear regression model.

### 3.4.3 THE FINAL PREDICTION ERROR (FPE) CRITERION

In '69 Akaike introduced a procedure for fitting autoregressive models for prediction. Akaike suggested to take the model which minimizes the FPE, which is defined as follows,

\[
\text{FPE} = V(\hat{\theta})(\frac{n+\#\text{par}}{n-\#\text{par}}) \tag{3.32}
\]

Here \( \#\text{par} \) is the number of free parameters and \( V(\hat{\theta}) \) is the residuals sum of squares. This criterion reflects the prediction-error variance that one will obtain, on the average, when the model is applied as a predictor to other datasets than those used for the identification. Although it was originally designed for autoregressive model it has become a popular method even in non autoregressive models. When we use this method in a linear regression model we have,

\[
\text{FPE}_p = \text{RSS}_p(\frac{n+p}{n-p}) = \text{RSS}_p(1+2p/(n-p)) \tag{3.33}
\]

Thus we see that this method is equivalent with Amemiya's \( \text{PC} \), besides a factor \( 1/n \).

### 3.5 INFORMATION THEORETIC ARGUMENTS

#### 3.5.1 AKAIKE'S INFORMATION THEORETIC CRITERION

Let us consider the general process where the probability of obtaining the vector of observed output samples is given by, \( p(y;\Theta) \). Here \( \Theta \) denotes the unknown parameters of the process. Let us consider a set of estimates \( \hat{\Theta} \)'s of the vector of parameters \( \Theta \). The objective is to estimate \( \Theta \) in such a way that the a posteriori probability density function (PDF) \( p(y;\hat{\Theta}) \) is as close as possible to the real PDF. As a measure of agreement between the real PDF and the estimated PDF consider the following function,

\[
I(\Theta,\hat{\Theta}) = E\{-\ln(p(y;\hat{\Theta})/p(y;\Theta))\} = \\
-\int p(y;\Theta) \ln \left(\frac{p(y;\hat{\Theta})}{p(y;\Theta)}\right) dy \tag{3.34}
\]
Here the expectation is taken with respect to $y$. We consider $\theta$ as given and thus independent of $y$. So (3.34) is in fact the a priori expectation. $I(\theta, \bar{\theta})$ is called the Kullback-Leibler information distance and it has the following properties,

$$a: I(\theta, \bar{\theta}) \geq 0$$

$$b: I(\theta, \bar{\theta}) = 0 \text{ if and only if } \theta = \bar{\theta}$$

For an enhanced list of properties see Ponomarenko '81. Since there are only $n$ independent realizations $y_i$ ($i=1..n$) available, the sample mean of the log likelihood ratio is used to estimate $I(\theta, \bar{\theta})$, thus

$$\hat{I}(\theta, \bar{\theta}) = -1/n \sum_{i=1}^{n} \ln( \frac{p(y_i; \bar{\theta})}{p(y_i; \theta)} )$$  \hspace{1cm} (3.37)

This is a consistent estimate of $I(\theta, \bar{\theta})$. If the objective is to minimize (3.37) we see that this can be realized without knowing the true value of $\theta$, giving the well known maximum likelihood estimator of $\theta$.

Akaike ('73) proposed to minimize,

$$E(2^*I(\theta, \bar{\theta}))$$  \hspace{1cm} (3.38)

Here the expectation is taken with respect to the distribution of $\theta$. Thus, here we assume that $\bar{\theta}$ is no longer a constant but actually depends on $y$, and therefore has a distribution. In order to calculate this minimum he derived the following criterion (for some easier derivation than the original one see Amemiya '80),

$$\text{AIC} = -2 \ln( p(y; \bar{\theta}) ) + 2 \#\text{par}$$  \hspace{1cm} (3.39)

Where $\#\text{par}$ is the number of unknown parameters. This criterion is called Akaike's Information theoretic Criterion (AIC).

If we project these derivations on our problem it can be seen (see appendix C),

$$2I(\theta, \bar{\theta}) = (\bar{x} - x)'(\bar{x} - x) / \sigma^2$$  \hspace{1cm} (3.40)

and thus,

$$E(2^*I(\theta, \bar{\theta})) = \text{MSE}(\bar{x}) / \sigma^2$$  \hspace{1cm} (3.41)

Since,

$$p(y; \bar{\theta}) = (2\pi \sigma^2)^{-n/2} \exp \left( -\frac{(y - \bar{\theta})'(y - \bar{\theta})}{2\sigma^2} \right)$$  \hspace{1cm} (3.42)

we can see that,

$$\text{AIC} = n^* \ln(2\pi) + n^* \ln(\sigma^2) + \frac{1}{n^*} \|\bar{y} - y\|^2 / \sigma^2 + 2p$$  \hspace{1cm} (3.43)

Minimizing the third term of the AIC (within our model structure) leads to the least square estimator $\hat{\theta}_1$ of $\theta_1$. So when $\sigma^2$ is known this AIC leads to
\[ \sigma^2 \text{ known,} \]
\[ \min AIC = \min \left( \frac{RSSp}{\sigma^2} + 2p \right) \quad (3.44) \]

so it is equal to minimizing \( C_p \) (with known \( \sigma^2 \)). When \( \sigma^2 \) is unknown we must use the maximum likelihood estimation of \( \sigma^2 \) in that model. This means that we should take \( \tilde{\sigma}(p) \) as an estimate of \( \sigma^2 \). In that case the AIC criterion becomes,

\[ \sigma^2 \text{ unknown,} \]
\[ \min AIC = \min \left( n \ln(\tilde{\sigma}(p)) + 2p \right) \quad (3.45) \]

If we don't take the maximum likelihood estimation of \( \sigma^2 \), thus we don't follow the AIC exactly, but some other estimator (see 3.3-3.6) we get the following criteria,

\[ \tilde{\sigma}(p) \min \left( n \ln(\tilde{\sigma}(p)) + p \right) \quad (3.46) \]
\[ \tilde{\sigma}(d) \min \left( \frac{RSSp}{\tilde{\sigma}(d)} + 2p \right) \quad (3.47) \]
\[ \tilde{\sigma}(d) \min \left( \frac{RSSp}{\tilde{\sigma}(d)} + 2p \right) \quad (3.48) \]

Here we see that in the last two cases the criteria become equivalent with the \( C_p \) criterion if the same estimate of the noise would be used.

### 3.6 Bayesian Arguments

In the preceding chapters we assumed that there was no prior information about the actual impulse response. In this chapter it is assumed that there exist some prior density function about \( \theta \). Under this assumption, it is possible to derive a model selection procedure using a bayesian framework. Of course, since these methods are based on heuristic arguments the only justification for them is through performance in practice. I will discuss now a general model selection procedure based on bayesian arguments.

In our linear regression model we are searching for the best model structure out of \( d \) possible model structures. Assume that there exists a probability function for the model structure. Let the model structure with the dimension \( p \) be denoted as \( M_p \). The probability of the model structure \( M_p \) is denoted as,

\[ p(M_p) \text{ with } \sum_{p=1}^{d} p(M_p) = 1 \quad (3.49) \]

Furthermore we assume that the prior probability density function of the parameters \( \theta_1 \) in the model structure \( M_p \) is given by,

\[ p(\theta_1 | M_p) \quad (3.50) \]
Let,
\[ p(y|\Theta_1, M_p) \]  
(3.51)
de note the usual probability density function of \( y \) conditional on the 
model structure \( M_p \) with its parameters \( \Theta_1 \).

Then with the relations,
\[ p(y|M_p) = \int p(y|\Theta_1, M_p)^*p(\Theta|\Theta_1|M_p) \, d\Theta_1 \]  
(3.52)
\[ p(y) = \sum_{p=1}^d p(y|M_p)^*p(M_p) \]  
(3.53)
it is possible to derive the probability of the model structure \( M_p \) 
conditional on \( y \). Because,
\[ p(y|M_p) = p(y,M_p) / p(M_p) \]  
(3.54)
and
\[ p(M_p|y) = p(y,M_p) / p(y) \]  
(3.55)
we can see that,
\[ p(M_p|y) = p(y|M_p)^*p(M_p) / \sum_{p=1}^d p(y|M_p)^*p(M_p) \]  
(3.56)
When we substitute the measured \( y \), then (3.56) is the a posteriori 
probability for \( M_p \). Finding the \( M_p \) for which this a posteriori probability 
is maximal yields the Bayes solution to this problem. For this maximization 
problem \( p(y) \) is a constant and can therefore be skipped from (3.56).

When we assume that each model structure has equal probability we can see 
that maximizing of \( p(M_p|y) \) is equal to maximizing \( p(y|M_p) \).

There are several model selection criteria developed on this general 
selection procedure or on a variant of it. In the following I will discuss 
only a few.

3.6.1 AKAIKE'S BAYESIAN INFORMATION CRITERION (BIC)

Akaike (\'77) derived a criterion based on a mixture of classical and 
bayesian arguments. His suggestion may be interpreted as follows (see 
Amemiya \'80). Assume that the prior density function of \( \Theta_1 \) is given by,
\[ \Theta_1 \sim N(0, \Sigma(U_1^T U_1)^{-1}) \]  
(3.57)
Thus the covariance matrix of the prior is except for a factor equal to 
the covariance matrix of the least-squares estimate of \( \Theta_1 \). It is assumed 
that all model structures have equal probability.

Thus because,
\[ p(y|\Theta_1, M_p) = N(U_1 \Theta_1, \sigma^2 I_n) \]  
(3.58)
we have that,
\[ p(y|M_p) = N(0, \sigma^2 \ln + \rho U_1 (U_1^T U_1)^{-1} U_1^T) \] (3.59)

Akaike suggested to estimate \( \sigma^2 \) and \( \rho \) by maximizing (3.59). Substituting these estimates in (3.59) and taking minus two times the logarithm of it, he obtained,

\[
\text{BIC} = \frac{(n-p) \ln \left( (I_n - U_1 U_1^T) y \right)^T (I_n - U_1 U_1^T) y \right)}{(n-p)} + p \ln \left( U_1 U_1^T \right) \] (3.60)

The model that minimizes this BIC should be chosen.

3.6.2 SCHWARZ CRITERION (SC)

Schwarz (1978) studies the asymptotic behaviour of Bayes estimators under a special class of prior distributions. Because of the asymptotic nature of his derivation this prior distribution needs not be known exactly. In the large-sample limit, the leading term of the Bayes solution turns out to be the maximum likelihood estimator. In order to select a model Schwarz proposed to minimize the following criterion,

\[
SC = -2 \ln L(\theta | y) + \# \text{par} \ln(n) \] (3.61)

Here \( k \) is the number of free parameters and \( L(\theta | y) \) is the likelihood function for \( \theta \). Just as in the AIC we can get for our problem two different criteria dependent on the fact if \( \sigma^2 \) is known or not.

So,

\[
\begin{align*}
\sigma^2 \text{ known} \quad & \\
\min SC &= \min \left( \frac{RSS}{\sigma^2} + p \ln(n) \right) \quad (3.62)
\end{align*}
\]

\[
\begin{align*}
\sigma^2 \text{ unknown} \quad & \\
\min SC &= \min \left( n \ln( \sigma^2(p) ) + p \ln(n) \right) \quad (3.63)
\end{align*}
\]

When we compare this criterion with the AIC we see that this criterion is more parsimonious i.e. it deletes more parameters from the full model than the AIC does because \( \ln(n) \) is usually greater than 2 (for \( n > 7 \)).
A very common way to validate the selection of a subset of regressors is to collect additional data and to look how well this subset performs. Because it is not always possible to collect additional data, the current data set could be split into two groups. One for analysis (subset selection and estimation) and one for validation. Based on this principle some criteria have been proposed for subset selection.

### 3.7.1 Allan's Predictive Sum of Squares (PRESS)

Allan ('74) proposed a criterion that simulates cross-validation. Suppose we are working with a model of dimension \( p \). Let,

- \( U_i(\cdot) = U_1 \) matrix without the \( i \)-th row.
- \( u_1(\cdot) = \) the \( i \)-th row of \( U_1 \)
- \( y(\cdot) = \) output vector without the \( i \)-th output
- \( y(i) = \) \( i \)-th output

Here we use all but the \( i \)-th output samples to predict the \( i \)-th response. Let the predicted response be denoted as \( y_p(i) \). Then we can write for a least squares prediction,

\[
y_p(i) = u_1(i)^*U_1(i)^* y(i)
\]

(3.64)

Allan proposed to "predict" each output sample using the other \( n-1 \) output samples. The resulting "errors of prediction" are squared and summed to form \( \text{PRESS}(p) \). Here \( p \) stands for the dimension of the model.

\[
\text{PRESS}(p) = \sum_{i=1}^{n} [y(i) - y_p(i)]^2
\]

(3.65)

The proposed procedure is now to choose the model that minimizes \( \text{PRESS} \).

At first sight the criterion seems very complicated. Fortunately it is possible to rewrite it in a more easier form. It can be proved (see appendix D) that \( \text{PRESS}(p) \) is equivalent with,

\[
\text{PRESS}(p) = w_p D_p^{-2} w_p
\]

(3.66)

Where,

- \( w_p = (I_n - U_1 U_1^T) y \)
- \( D_p = \) the diagonal matrix whose diagonal elements are those of \( (I_n - U_1 U_1^T) \)
3.7.2. STOICA’S CI(m) and C2(m)

In the previous PRESS criterion the validation set consisted each time of one sample. Stoica et al. ('85) discussed some more general cross-validation criteria. For the sake of convenience I will adopt some notations used in that article.

Let the interval \( I = \{1, \ldots, n\} \) be divided into \( k-1 \) intervals of length \( m \) and 1 interval which is smaller than \( 2m \), where \( k \) is the largest integer not greater than \( n/m \). Then the intervals \( I_v \) can be defined as follows,

\[
I_v = \{(v-1)m+1, \ldots, vm\} \quad v = 1, \ldots, k-1
\]
\[
I_k = \{(k-1)m+1, \ldots, n\} \quad v = k
\]

Furthermore let the residuals as a function of the model parameters \( \theta \) be denoted as \( e(t, \theta) \) where \( t \) stands for the \( t \)-th residual. Let,

\[
\hat{\theta} = \arg \min_{\theta} V(\theta) \quad V(\theta) = \frac{1}{n} \sum_{t \in I} \varepsilon(t, \theta)
\]

Stoica et al. assumed the following conditions to hold,

A1. \( e(t, \theta) \) is sufficiently smooth (so its derivatives with respect to \( \theta \) exist and are finite)

A2. \( V_{\theta\theta}(\hat{\theta}) = \delta^2 V(\theta) / \delta \theta^2 \mid_{\theta=\hat{\theta}} \) is positive definite

A3. The residuals \( e(t, \theta) \) and

\[
\varepsilon_\theta(t, \theta) = \delta e(t, \theta) / \delta \theta
\]
\[
\varepsilon_{\theta\theta}(t, \theta) = \delta^2 e(t, \theta) / \delta \theta^2
\]

are stationary and ergodic processes for any \( \theta \). Moreover the sample moments are assumed to converge to the theoretical moments (as \( n \) tends to infinity) at a rate of order \( O(1/n) \).

**FIRST CROSS-VALIDATION CRITERION**

For cross-validatory assessment of the model structure, the following function is used,

\[
C_I = \sum_{v=1}^{k} \sum_{t \in I_v} e^2(t, \hat{\theta}_v)
\]

with,

\[
\hat{\theta}_v = \arg \min_{\theta} \sum_{t \in I_v} e^2(t, \theta) \quad v = 1, \ldots, k
\]

Thus each time \( m \) residuals are used to validate the estimate from the other \( n-m \) residuals. Stoica et al. proved that when the assumptions A1-A3
hold, for \( k \) large enough the following relation (for the SISO case) holds,

\[
\frac{1}{n} C_1 = C_1 + O(1/k^2m)
\]  

(3.72)

where,

\[
C_1(m) = \sum_{v=1}^{k} w_v(\hat{\Theta}) + 4/n^2 \sum_{v=1}^{k} w_v(\hat{\Theta}) v \Theta_v(\hat{\Theta})^{-1} w_v(\Theta)
\]

(3.73)

with,

\[
w_v(\Theta) = \sum_{t \in I_v} e(t, \dot{\Theta}) e(t, \Theta) \quad v = 1..k
\]

(3.74)

Therefore their first model structure selection rule was stated as follows,

Choose the model structure that leads to the smallest value of \( C_1(m) \).

SECOND CROSS-VALIDATION CRITERION

For the second cross-validatory assessment the following function is used,

\[
C_{\|} = \sum_{v=1}^{k} \sum_{t \in I_v} e^2(t, \hat{\Theta}_v)
\]

(3.75)

where,

\[
\hat{\Theta}_v = \arg \min_{\Theta} \sum_{t \in I_v} e^2(t, \Theta) \quad v = 1..k
\]

(3.76)

Thus in this function only \( m \) residuals are used to estimate \( \Theta \) and all \( n-m \) other residuals are used to validate this estimate. Stoica et al. proved that when \( A1-A3 \) are valid, for \( m \) and \( k \) large enough the following relation (for the SISO case) holds,

\[
\frac{1}{((k-1)n)} C_{\|} = C_2 + O(1/(\min(n,m)^{3/2})
\]

(3.77)

where,

\[
C_2(m) = \sum_{v=1}^{k} w_v(\hat{\Theta}) + 2k/n \sum_{v=1}^{k} w_v(\hat{\Theta}) v \Theta_v(\hat{\Theta})^{-1} w_v(\Theta)
\]

(3.78)

with,

\[
w_v(\Theta) = \sum_{t \in I_v} e(t, \dot{\Theta}) e(t, \Theta) \quad v = 1..k
\]

(3.79)

The second cross-validation criterion is stated as follows,

Choose the model structure which leads to the smallest value of \( C_2(m) \).
The advantage of \( C_1(m) \) and \( C_2(m) \) is that they are much easier to compute than \( \text{C}_1 \) and \( \text{C}_2 \), since in the last two criteria for each subset the 'new' estimate has to be computed. In \( C_1(m) \) and \( C_2(m) \) we use only one estimate \((\Theta)\).

In the following I will give the derivation of the two criteria when projected on the linear regression model.

**STOICA'S \( C_1(m) \) IN THE LINEAR REGRESSION MODEL**

In the linear regression model \((y=U_{1}\Theta+e)\) we have,

\[
e(t,\Theta) = y(t) - u_1(t)\Theta \tag{3.80}
\]

Here \( u_1(t) \) is the \( t \)-th row of \( U_1 \) and \( y(t) \) is the \( t \)-th element of \( y \). So we have,

\[
e_\Theta(t,\Theta) = -u_1(t)' \tag{3.81}
\]

\[
e_\Theta(t,\Theta) = 0 \tag{3.82}
\]

and,

\[
V(\Theta) = (y-U_{1}\Theta)'(y-U_{1}\Theta)/n \tag{3.83}
\]

\[
V_{\Theta\Theta} = 2U_{1}'U_{1}/n \tag{3.84}
\]

From these expressions we see that the assumptions A1-A2 are valid. The assumption A3 does not have to hold, since, for example, the statistical properties of \( e_\Theta(t,\Theta) \) depend entirely on the actual input. Therefore, when giving the derivation of the cross-validation criteria for the linear regression model and to establish the same asymptotic behaviour for those criteria as given in (3.72) and (3.77), the assumptions made in A-3 will be given in terms of assumptions concerning the actual input.

Because of (3.71), (3.81), (3.82) and the Taylor series expansion it can written that,

\[
0 = \delta/\delta\Theta 1/n \sum_{t\in I_1} e(t,\Theta) |_{\Theta=\hat{\Theta}_v}
\]

\[
= V_{\hat{\Theta}_v}(\hat{\Theta}_v) - 2/n \sum_{t\in I_1} e(t,\hat{\Theta}_v)e_\Theta(t,\hat{\Theta}_v) \tag{3.85}
\]

\[
= V_{\hat{\Theta}_v}(\hat{\Theta}_v) + V_{\hat{\Theta}_v}(\hat{\Theta}_v)(\hat{\Theta}_v - \Theta) +
\]

\[
-2/n \sum_{t\in I_1} e(t,\hat{\Theta}_v)e_\Theta(t,\hat{\Theta}_v) - 2/n \sum_{t\in I_1} e_\Theta(t,\hat{\Theta}_v)e_\Theta(t,\hat{\Theta}_v)(\hat{\Theta}_v - \Theta)
\]

\[
= 2/n u_1'u_1 (\hat{\Theta}_v - \Theta) + 2/n u\v e\v - 2/n u\v u\v + (\hat{\Theta}_v - \Theta)
\]

\[
= 2/n u_1'u_1 (\hat{\Theta}_v - \Theta) + 2/n u\v e\v
\]
where $Uv$ is a matrix whose rows are the vectors $u(t)$ for $t \in V$ and $ev$ is the corresponding vector of residuals. $Us$ is the matrix $U1$ without the rows $u(t)$ for $t \in V$. Because of relation (3.85) we can write,

$$\begin{align*}
(\hat{\Theta}_v - \hat{\Theta}) &= -(Us'Us)^{-1} Uv'ev \quad (= O(1/k/m)) \quad (3.86)
\end{align*}$$

This relation is exact. (For the order determination, Stoica et al. made use of the following,

$$1/m \sum_{t \in I} e(t, \hat{\Theta})e_g(t, \hat{\Theta}) = E( e(t, \hat{\Theta})e_g(t, \hat{\Theta}) ) + O(1/m) =$$

$$1/n \sum_{t \in I} e(t, \hat{\Theta})e_g(t, \hat{\Theta}) + O(1/\sqrt{n}) + O(1/m) = O(1/\sqrt{m})$$

because of assumption A-3. In our linear regression model we have,

$$1/m \sum_{t \in I} e(t, \hat{\Theta})e_g(t, \hat{\Theta}) = -1/m Uv'ev$$

and

$$1/n \sum_{t \in I} e(t, \hat{\Theta})e_g(t, \hat{\Theta}) = -1/n U_1'(I-U_1U_1')y = 0$$

Although the last expression is equal to zero this does not have to mean that $-1/mUv'ev$ is of order $O(1/\sqrt{m})$. However, if we assume that $ev$ is a vector with white noise (and thus assume that the process is in the model set, which is usually the case when $n$ tends to infinity), averaging the vector $ev$ by $-1/mUv'ev$ makes it to decrease with order $O(1/\sqrt{m})$. Thus when $n$ tends to infinity it is likely that the assumption that $-1/mUv'ev$ is of order $O(1/\sqrt{m})$ is true.)

In (3.86) the inverse of $Us'Us$ has to be computed for each subset. To decrease the computational task Stoica et al. used, in terms of our linear regression model, the following approximation,

$$\begin{align*}
\text{Approximation-1 :}
(\hat{\Theta}_v - \hat{\Theta}) &= -(U_1'U_1)^{-1} Uv'ev + O(1/k^2/m) \\
\end{align*}$$

Translating the assumptions (A-3) made in this approximation in terms of conditions on the actual input, it is assumed that,

$$(Us'Us)^{-1} = (U_1'U_1)^{-1} + O(m/n^4) \quad (3.88)$$

Let us now evaluate $C_I$. It can be written that,

$$e'(t, \hat{\Theta}_v) = e'(t, \hat{\Theta}) + 2 e(t, \hat{\Theta})e_g(t, \hat{\Theta})' (\hat{\Theta}_v - \hat{\Theta}) +$$

$$(\hat{\Theta}_v - \hat{\Theta})'e_g(t, \hat{\Theta})e_g(t, \hat{\Theta})'(\hat{\Theta}_v - \hat{\Theta})$$

$$= e'(t, \hat{\Theta}) - 2 e(t, \hat{\Theta})u(t)(\hat{\Theta}_v - \hat{\Theta}) +$$
\( (\hat{\theta}_v \cdot \hat{\theta})' u_1(t)' u_1(t)(\hat{\theta}_v \cdot \hat{\theta}) \)

So,
\[ \Sigma e^v(t, \hat{\theta}_v) = \Sigma (e(t, \hat{\theta}) - u_1(t)(\hat{\theta}_v \cdot \hat{\theta}))' \]

and,
\[ 1/n C_1 = 1/n \Sigma e^v(t, \hat{\theta}_v) = \]
\[ 1/n \Sigma (e(t, \hat{\theta}) - u_1(t)(\hat{\theta}_v \cdot \hat{\theta})') \]

This relation is exact. However, in the derivation of the first cross validation criterion Stoica et al. approximated (3.90) with,

**Approximation 2:**

\[ 1/n C_1 = 1/n \Sigma (e^v(t, \hat{\theta}) - 2e(t, \hat{\theta})u_1(t)(\hat{\theta}_v \cdot \hat{\theta}))' + O(1/k^m) \]  

Since \((\hat{\theta}_v \cdot \hat{\theta})\) is of order \(O(1/k^m)\) (see 21). When the approximations 1 and 2 are combined, we obtain the criterion,

\[ 1/n C_1 = 1/n \Sigma (e^v(t, \hat{\theta}) - 2e(t, \hat{\theta})u_1(t)'\]
\[ = 1/n \Sigma (ev'ev + 2ev'uv(U_1'U_1)^{-1}uv'ev + O(1/k^m)) + 0(1/k^m) \]

Thus in the linear regression model the first cross validation criterion is stated as follows,

\[ C_1(m) = 1/n \Sigma (ev'ev + 2ev'uv(U_1'U_1)^{-1}uv'ev) \]
The first approximation gives clearly a computational improvement compared to the real cross validation criterion, but the second approximation does not seem to give any computational improvement. Therefore we can consider to delete this approximation. This results in the following relation,

\[
\frac{k}{1/n} C_1 = \frac{1/n}{\Sigma} \sum_{v=1}^{k} \left| e_v + U_v(U_v^T U_v)^{-1} U_v e_v \right|^2 + O(1/k^m)
\]  

(3.93)

The difference between the real cross validation criterion and the approximated one remains of the order \(O(1/k^m)\), so in order to approximate \(1/n C_1\) it does not make any difference if we use approximation 2 or not (Although minimizing the first right term of (3.93) and \(C_1\) can lead to different model structures).

Stoica et al. showed that under certain conditions the first cross validation criterion is asymptotically equal to AIC. For the linear regression model, this can be seen as follows,

\[
\frac{k}{1/n} \sum_{v=1}^{k} e_v e_v + 2 \sum_{v=1}^{k} e_v U_v(U_v^T U_v)^{-1} U_v e_v = V(\hat{\theta}) + 2/n \sum_{v=1}^{k} \text{tr}( (U_v^T U_v)^{-1} \Sigma U_v e_v e_v U_v )
\]  

(3.94)

If we assume that, when \(n\) tends to infinity, \(e_v\) becomes a vector with white noise samples (with approximated variance \(V(\hat{\theta})\)), the last expression can approximated with,

\[
V(\hat{\theta}) + 2/n \sum_{v=1}^{k} \text{tr}( (U_v^T U_v)^{-1} U_v e_v e_v U_v ) = V(\hat{\theta}) + 2p/n V(\hat{\theta}) \left[1 + 2p/n \right]
\]  

(3.95)

So for large \(n\) we have,

\[
\ln(V(\hat{\theta}) \left[1 + 2p/n \right]) = \ln(V(\hat{\theta})) + \ln(1 + 2p/n) = \ln(V(\hat{\theta})) + 2p/n = AIC/n
\]  

(3.96)

(see 3.45)

When \(n\) tends to infinity the model is usually overestimated and thus the condition that \(e_v\) is a white noise vector is fulfilled.
STOICA'S C2(m) IN THE LINEAR REGRESSION MODEL

For the second cross-validation criterion with (3.76), (3.81) and (3.82) the following holds,

\[ Q = \sum_{t \in I_v} e(t, \hat{\theta}_v) e(t, \hat{\theta}_v) \]
\[ = \sum_{t \in I_v} e(t, \hat{\theta}) e(t, \hat{\theta}) + 2/m \sum_{t \in I_v} e(t, \hat{\theta}) e(t, \hat{\theta}')(\hat{\theta}_v - \hat{\theta}) \]
\[ = -U_s' e_s + U_s' U_s (\hat{\theta}_v - \hat{\theta}) \]

So,

\[ (\hat{\theta}_v - \hat{\theta}) = (U_s' U_s)^{-1} U_s' e_s \quad ( = O(1/m) ) \] (3.97)

where \( U_s \) is the matrix whose rows are the vectors \( u(t) \) for \( t \in I_v \) and \( e_s \) is the vector of the corresponding residuals. In (3.98) it is assumed that \( U_s' e_s = O(1/m) \) (see (3.86)).

To decrease the computational task Stoica et al. used, in terms of our linear regression model, the following approximation of (3.98),

\[ \text{Approximation-1} : \]
\[ (\hat{\theta}_v - \hat{\theta}) = n/m (U_1^T U_1)^{-1} U_1' e_s + O(1/m) \] (3.99)

Translating the assumptions (A-3) made in this approximation in terms of conditions on the actual input, it is assumed that,

\[ (U_s' U_s)^{-1} = n/m ((U_1^T U_1)^{-1} + O(1/m/m) \]

(3.100)

Let us now evaluate \( C_{11} \).

\[ e(t, \hat{\theta}_v) = e(t, \hat{\theta}) + 2 e(t, \hat{\theta}) e(t, \hat{\theta})' (\hat{\theta}_v - \hat{\theta}) \]
\[ + (\hat{\theta}_v - \hat{\theta})' e(t, \hat{\theta}) e(t, \hat{\theta})' (\hat{\theta}_v - \hat{\theta}) \]
\[ = e(t, \hat{\theta}) + 2 e(t, \hat{\theta}) u(t)(\hat{\theta}_v - \hat{\theta}) \]
\[ + (\hat{\theta}_v - \hat{\theta})' u(t)' u(t)(\hat{\theta}_v - \hat{\theta}) \]
\[ = (e(t, \hat{\theta}) - u(t)(\hat{\theta}_v - \hat{\theta}))' \]

(3.101)

So,

\[ \frac{1}{k} \sum_{v=1}^{k} e(t, \hat{\theta}_v) = e(t, \hat{\theta}) - u(t)(\hat{\theta}_v - \hat{\theta}) \]

(3.102)

\[ \frac{1}{k} \sum_{v=1}^{k} (e(t, \hat{\theta}) - u(t)(\hat{\theta}_v - \hat{\theta}))' = 0 \]

(3.103)
First let us consider term $T_3$.

$$
T_3 = \frac{1}{((k-1)n)} \sum_{v=1}^{k} e(t, \hat{v}) + \frac{1}{((k-1)n)} \sum_{v=1}^{k} \sum_{v=1}^{k} \left( \hat{v} \cdot \hat{v} \right) u(t) u(t) \left( \hat{v} \cdot \hat{v} \right)
$$

(3.102)

Here $U_v$ denotes the matrix $U$, without the rows $u_1(t)$ for $t \in I_v$. Stoica et al. used the following approximation of $T_3$,

$$
T_3 \approx \frac{1}{((k-1)n)} \sum_{v=1}^{k} \left( \hat{v} \cdot \hat{v} \right) U_v U_v \left( \hat{v} \cdot \hat{v} \right)
$$

(3.103)

Where it is assumed that,

$$
\frac{1}{n} U_v U_v = \frac{1}{n} U_1 U_1 + O(1/n)
$$

(3.105)

Let us now consider term $T_2$,

$$
T_2 = -2 \frac{1}{((k-1)n)} \sum_{v=1}^{k} e(t, \hat{v}) u(t) \left( \hat{v} \cdot \hat{v} \right)
$$

(3.104)
Stoica et al. used the following approximation,

\[
T_2 = O(1/n)
\]  

(3.107)

Because \((\hat{\Theta}_v \cdot \hat{\Theta})\) is of order \(O(1/n)\). Since \((\hat{\Theta}_v \cdot \hat{\Theta})\) is of order \(O(1/m)\), \(T_3\) is of order \(O(1/m)\) and thus much larger than \(T_2\) when \(n\) tends to infinity.

When we use the approximation-1 of \((\hat{\Theta}_v \cdot \hat{\Theta})\), \(T_3\) becomes,

\[
T_3 = \frac{1}{((k-1)n)} \sum_{v=1}^{k} \left( \frac{n}{m} (U_1 U_1)^{-1} U_1 \hat{e} s + O(1/m) \right) U_1 \hat{e} s + O(1/n) \\
+ \frac{1}{((k-1)n)} \sum_{v=1}^{k} \left( n^{1/3} \cdot m^3 \right) \hat{e} s U_1 U_1 \hat{e} s + O(1/n) \\
+ O(1/m^3) + O(1/n)
\]

\[
= \frac{k}{n} \sum_{v=1}^{k} \left( \hat{e} s U_1 U_1 \hat{e} s + O(1/min(n,m/m)) \right)
\]

\[
= \frac{2k}{n^2} \sum_{v=1}^{k} w v \hat{e} s = \frac{1}{w v} + O(1/min(n,m/m))
\]

(3.108)

Because of this last approximation and with approximation-2 it can be written that,

\[
1/((k-1)n) \Delta_{11} = C_2 + O(1/min(n,m/m))
\]

(3.109)

with,

\[
C_2 = V(\hat{\Theta}) + \frac{k}{n^2} \sum_{v=1}^{k} w v \hat{e} s = \frac{1}{w v}
\]
Thus the second cross-validation criterion for the linear regression model can be stated as,

\[
C_2 = \frac{1}{n} \sum es' es + k es' Us (U'U)^{-1} Us' es
\]

We see that this criterion is essentially the same as \(C_1\). The only difference is that the second term in \(C_2\) has a different weight factor. Therefore and for reasons similar to (3.95-3.96), it can be seen, as Stoica et al. proved, that this criterion is asymptotically equivalent to a generalized AIC. Since for large \(n\) and \(m\),

\[
C_2 = \ln(V(\theta)) + kp/n \quad (3.110)
\]

**SOME REMARKS**

As already mentioned, the assumption A-3 does not have to hold for the linear regression model. However, in order to derive the cross-validation criteria for the linear regression model and to establish the same asymptotical behaviour of those criteria as given in (3.72) and (3.77), we projected the assumptions made in A-3 on conditions about the actual input (see 22, 32 and 37). However, this means that the assumptions are all fairly weak. If the order of approximation is somewhat different as those given in (3.88), (3.100) and (3.105) it will only affect the order of approximation given in the two cross-validation criteria and not the general ideas.

The procedures given above depend on \(m\), and the choice of this parameter should be discussed. Although no precise rules on how to choose \(m\) can be given, Stoica et al. gave some ideas about the value of \(m\). For the first cross-validation criterion \(m\) should be chosen so as to indicate on how many future sampling points we intend to use our model. Suppose we wish to use the estimated model at some, say \(f\), future time instants. Then we may choose \(m\) such that,

\[
m/(n-m) = f/n \quad (3.111)
\]

However, since \(m/(n-m)\) must be small for \(C_1(m)\) to be a good approximation of \(C_1\), \(C_1(m)\) can be used to select a good 'short term' model.

For the second cross validation criterion Stoica et al. remarked that a good choice for \(k\) may be \(k=ln(n)\), since for this value \(C_2(m)\) becomes asymptotically equivalent to the well-known Schwarz criterion (SC) and therefore it will choose, as we will see, asymptotically the right model structure.
3.8 STEIN RULE ARGUMENTS

So far I discussed only the least squares method for obtaining an estimator for \( \theta \). It was said that when the process is in the model set, this estimator is a minimum variance unbiased estimator. Thus if we are willing to improve our estimator we have to restrict ourselves to the class of unbiased estimators.

Suppose we observe \( n \) independently normally distributed variables, \( r_1, \ldots, r_n \) with unknown means \( \mu_1, \ldots, \mu_n \) and known variance \( \sigma^2 \). It is common to estimate \( \mu_1, \ldots, \mu_n \) by \( r_1, \ldots, r_n \). However, Stein (1956) proved that this estimator is admissible for \( n \leq 3 \) and inadmissible for \( n > 3 \) under the loss function,

\[
\text{MSE}(\hat{\mu}, \mu) = E\{(\hat{\mu} - \mu)^2|\mu\}
\]

(An estimator \( \delta \) of \( \mu \) is said to be inadmissible if there exists an estimator \( \delta^* \) for which \( \text{MSE}(\mu, \delta^*) \leq \text{MSE}(\mu, \delta) \) for all \( \mu \) with strict inequality for some \( \mu \). An estimator is admissible if it is not inadmissible.)

James and Stein (1961) exhibited an estimator that is uniformly better than the least squares one. For a review on this topic see Draper and Van Nostrand (1979) or Judge and Bock (1978).

Suppose that we have,

\[
r = \mu + e \quad \text{and} \quad e \sim N(0, \sigma^2 I_n)
\]

As an estimator of \( \mu \) we take \( \hat{\mu} = r \). Then we have,

\[
E\{|\hat{\mu}|^2\} = |\mu|^2 + n^2\sigma^2
\]

Here we see that when \( n^2\sigma^2/\mu^2 \mu \) is large we would be over-estimating \( \mu \) by a very large amount. Thus in order to improve our estimator it suggests that the least squares estimator could be improved by shrinking it (= multiplying it by a scalar). This shrinking factor (\( S_f \)) should be small when \( n^2\sigma^2/\mu^2 \mu \) is large and should be nearly 1 when \( n^2\sigma^2/\mu^2 \mu \) is small. Of course, \( \mu \) is unknown. When taking its estimate the previous discussion suggest that we should consider estimates of the following form,

\[
\tilde{\mu} = (1 - c \sigma^2 / \mu^2 \mu) \hat{\mu} = (1 - c \sigma^2 / r^2) \hat{r}
\]

Where \( c \) is some constant. James and Stein proved that when \( \sigma^2 \) is known the estimator \( \tilde{\mu} \),

\[
\tilde{\mu} = (1 - c \sigma^2 / \mu^2 \mu) \hat{\mu}
\]

\( 0 < c < 2(n-2) \) \( c \)-optimal = \( (n-2) \)

is uniformly better, in the sense that \( \text{MSE}(\mu, \tilde{\mu}) < \text{MSE}(\mu, \hat{\mu}) \) for all values of \( \mu \), than the least squares estimator for \( n \geq 3 \). The optimal value for \( c \) is \( (n-2) \). When \( \sigma^2 \) is not known they proved that with an independent estimator of \( \sigma^2 \) (\( \hat{\sigma}^2 \)), where \( \hat{\sigma}^2 \mu \) must be distributed as \( \sigma^2 \) times a Chi-squared distribution with \( v \) degrees of freedom (thus be unbiased), the estimator

\[
0 < c < 2(n-2)
\]

is uniformly better, in the sense that \( \text{MSE}(\mu, \tilde{\mu}) < \text{MSE}(\mu, \hat{\mu}) \) for all values of \( \mu \), than the least squares estimator for \( n \geq 3 \). The optimal value for \( c \) is \( (n-2) \). When \( \sigma^2 \) is not known they proved that with an independent estimator of \( \sigma^2 \) (\( \hat{\sigma}^2 \)), where \( \hat{\sigma}^2 \mu \) must be distributed as \( \sigma^2 \) times a Chi-squared distribution with \( v \) degrees of freedom (thus be unbiased), the estimator

\[
0 < c < 2(n-2)
\]
\[ \tilde{\mu} = [1 - (c/(n+2))v^*\hat{\sigma}^2/(\hat{\mu} - \mu)]^* \mu \]  
(3.117)

\[ 0 < c < 2(n-2) \quad \text{c-optimal} = (n-2) \]

is also uniformly better than the least squares estimator under the quadratic loss function (3.112) for \( n \geq 3 \).

In 3.117 we used the origin as the place to shrink to. This is not necessary. Although the notation conceals it, we may choose any other point to shrink to. The estimator of the mean given by the James-Stein estimator in the following general form,

\[ \tilde{\mu} = [1 - (c/(v+2))v^*\hat{\sigma}^2/(\hat{\mu} - \mu)]^* (\hat{\mu} - \mu) + o \]

with,

\[ c = (n-2) \]
\[ v = \text{degrees of freedom in } \hat{\sigma}^2 \]
\[ o = \text{new origin to shrink to} \]

is also uniformly better than the least squares estimator. Thus a researcher can choose a value \( o \) that he thinks that \( \mu \) should be near. He then computes \( \tilde{\mu} \) with (3.118). If he is correct, his estimator will be better than \( \hat{\mu} \), but even when he is wrong, his estimator will be better than \( \hat{\mu} \).

How can we use these properties in model selection? We first have to make an assumption. Let us assume that the U matrix is column orthogonal with,

\[ U'U = Id/l \]  
(3.119)

here \( l \) is some number (which only depends on \( n \)). This would approximately be the case when \( U \) is filled with white noise and \( n \gg d \). Then we can see that \( \Theta \) is independently normally distributed. Thus,

\[ \hat{\Theta} \in \mathcal{N}(\Theta, l\hat{\sigma}^2 Id) \]  
(3.120)

Let now take \( \Theta \) the place of \( \mu \) in the previous derivations. This means that according to (3.116) the estimator,

\[ \tilde{\Theta} = [1 - (c/(v+2))v^*\hat{\sigma}^2/(\hat{\Theta} - \Theta)]^* \hat{\Theta} \]  
(3.121)

\[ \hat{\Theta} = Sf^* \hat{\Theta} \]

with,

\[ 0 < c < 2(d-2) \quad \text{c-optimal} = (d-2) \]
\[ \hat{\sigma}^2 = (RSSd/v \quad \text{v} = (n-d) \]

(which is independent from \( \Theta \)) is a uniformly better estimator than \( \hat{\Theta} \) for \( d \geq 3 \), according to the loss function MSE(\( \Theta \)). I will call this estimator the James-Stein estimator of \( \Theta \). An important point is that the James-Stein estimator itself is also not admissible. In relation (3.74) we can see that \( Sf < 1 \). If \( Sf > 0 \) then \( \Theta \) will be shrunk towards zero. However when \( Sf < 0 \) then \( \Theta \) will be shrunk past zero. Sclove ('68) mentioned that the estimator
\( \hat{\theta} = \max(0,Sf) \hat{\theta} \) \hspace{1cm} (3.122)

with
\[
Sf = \left[1-(c/(v+2))v^s\hat{\vartheta}^s/\hat{\theta}^s\right] \hspace{1cm} (3.123)
\]

\[ 0 < c < 2(d-2) \hspace{1cm} v=n-d \]

is uniformly better the James-Stein estimator (with the same value of c). I will call this estimator the positive James-Stein estimator. In this positive James-Stein estimator there exists no optimal value for c, but if c lies between 0 and 2(d-2) it gives better results than the least squares estimator.

The property (3.122) is an interesting property for hypothesis testing. Suppose the parameter set is divided, without reference to the data, into two subsets. One subset which is assumed to be important and another subset which is assumed to be unimportant. Because all the parameters are independent we can shrink the subset, which is supposed to be unimportant, towards zero. Then,

\[
\hat{\theta} = \begin{bmatrix} (\hat{\theta})^1 \\ \max(0,Sf) (\hat{\theta})^2 \end{bmatrix} \text{ with } \hat{\theta} = \begin{bmatrix} \hat{\theta}^1 \\ \hat{\theta}^2 \end{bmatrix} \hspace{1cm} (3.124)
\]

where
\[
Sf = \left[1-(c/(v+2))v^s\hat{\vartheta}^s/\hat{\theta}^s\right] \hat{\theta}^s (\hat{\theta})^2
\]

(\( \hat{\theta}^1 \)) is a p-vector
(\( \hat{\theta}^2 \)) is a (d-p)-vector
\[ 0 < c < 2(d-p-2) \hspace{1cm} v=n-d \]

is uniformly better estimator than \( \hat{\theta} \), under \( \text{MSE}(\hat{\theta}) \), if \( (d-p) \geq 3 \). This last relation suggests that we can set the last \( (d-p) \) parameters to zero if \( Sf < 0 \).

In this case \((U'U+I_d)/l\) a normal F-test would conclude that the last \((d-p)\) parameters could be set to zero if \( (3.124) \).

\[
(\hat{\theta})^2/((d-p)^s/(d)) < f^* d-p, n-d \hspace{1cm} (3.125)
\]

If we look at relation (3.124) we see that in the James-Stein case we set the last \((d-p)\) parameters to zero if,

\[
(\hat{\theta})^2/((d-p)^s/(d)) < c^*[1^s/(v+2)(d-p)] \hspace{1cm} (3.126)
\]

Thus the positive James-Stein rule can be seen as a F-test with a particular value of c. However, the main difference is that when the hypothesis is rejected, the subset will be shrunk towards zero instead of maintaining their original values. This will assure us that, unlike in the F-test case, we will get a uniformly better estimate. I will return on this point later.
The previous derived results for the parameter space only hold when $U'U = I_d$. What can we do when this relation does not hold? Let us consider the general model $y = U\theta + e = x + e$. Rewrite $U$ as,

$$U = V\Sigma W'$$

(singular value decomposition) \hspace{1cm} (3.127)

Here $V$ and $W$ are column unitary and orthonormal, and $\Sigma$ is a $n \times d$ diagonal matrix with $\text{diag}(\Sigma) = (\sigma_1, \ldots, \sigma_d)$, and $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_d > 0$. (since we assumed that $U$ had full rank). Therefore we can write,

$$U^* = \Sigma W'$$ \hspace{1cm} (3.128)

$$U'U = \Sigma^2 W'W = \Sigma W' W \Sigma = \Sigma^2$$ \hspace{1cm} (3.129)

Where $\Sigma^*$ is a $d \times n$ diagonal matrix with $\text{diag}(\Sigma^*) = (1/\sigma_1, \ldots, 1/\sigma_d)$. $D$ is a $d \times d$ diagonal matrix with the same elements as $\Sigma$. From this last expression we can see that the columns of $W$ are the eigenvectors of $U'U$ and the diagonal values of $\Sigma$ are the corresponding eigenvalues. Let us define,

$$Z = U + \Sigma^{-1}W'$$ \hspace{1cm} (3.130)

$$\beta = W\Sigma \alpha$$ \hspace{1cm} (3.131)

Then we can rewrite the general model as,

$$y = Z\beta + e$$ \hspace{1cm} (3.132)

This model will be called the orthonormal canonical form of the general model. In this canonical form the least squares estimator of $\beta$ is given by,

$$\hat{\beta} = (Z'Z)^{-1}Z'y = Z'y = WD\Sigma^{-1}$$ \hspace{1cm} (3.133)

because of (3.27-3.29). Since $Z'Z = I_d$ we see from this expression that,

$$\hat{\beta} \in N(\beta, \sigma^2 I_d)$$ \hspace{1cm} (3.134)

Thus in the $\beta$ space the least squares estimate of $\beta$ is independently normally distributed. This means that we can apply the results from 3.122 directly on them. So,

$$\hat{\beta} = \max(0, Sf) \hat{\beta}$$ \hspace{1cm} (3.135)

with

$$Sf = \left(1 - \frac{c}{(v+2)}\right)v\sigma^2 / x'x$$

(since $\hat{\beta}'\hat{\beta} = x'x$)

$$v = (n - d) \sigma^2 = RSSd/v \quad 0 < c < 2(d-2)$$

is a uniformly better estimator than $\hat{\beta}$ according to the loss function,

$$E\{ (\beta - \hat{\beta})'(\beta - \hat{\beta}) \} = E\{ (x - x)'(x - x) \}$$ \hspace{1cm} (3.136)

for $d \geq 3$. This means that the shrunken estimator in the original parameter space,
\[ \hat{\theta} = \mathbf{wD}^{-1} \mathbf{w}' \mathbf{\hat{\beta}} = \max(0, \mathbf{Sf}) \ast \hat{\theta} \]  

(3.137)

gives uniformly better results than \( \hat{\theta} \) according to the performance function \( \text{MSE}(x) \). Since the estimator in the \( \beta \) space was independently normally distributed, we can test if some subset of \( \beta \) parameters can be set to zero according to the previous given method. However, although it can lead to a reduction in the \( \beta \) space, it does not lead to a reduction in the \( \theta \) space because of the transformation (3.131). Thus a real model reduction cannot be achieved.

The estimator in (3.137) is better than \( \hat{\theta} \) under the loss function \( \text{MSE}(x) \). But how does it perform under \( \text{MSE}(\theta) \)? This problem will be discussed in chapter 5.
4. EVALUATION OF SELECTION CRITERIA

4.1 ASYMPTOTIC PROPERTIES

In this section I will derive the asymptotic properties of a general criterion that is based on $RSS_p$. When taking this derivation as an example, it is possible to say something about the asymptotic properties of particular criteria. I do not claim that this derivation is a good one, but I think it gives some insight into what happens when $n$ tends to infinity.

Let the criterion be given by,

$$ C = (x'x)'(x'x) + S = RSS_p + S $$

(4.1)

Here $S$ is a monotonically increasing function of $p$, that is built-in to penalize large models. For example in the Cp case $S$ is $2p^2$. Suppose we have two models $A$ and $B$,

$$ \text{model } A: \ y = U_1\theta_1 + U_2\theta_2 + e \quad \dim U = n*d $$

$$ \text{model } B: \ y = U_1\theta_1 + e \quad \dim U = n*p $$

(4.2)

We take that model as our final model that minimizes $C$. Our goal is to investigate what happens in our selection as $n$ tends to infinity. This will we done in two cases. In one case I assume that model $A$ is true and in the other case I assume that model $B$ is true. First let us suppose that model $A$ is true. Let,

$$ \text{RSS}_p - \text{RSS}_d = \|AA'Y\|^2 \text{ with } A = (I_n - U_1U_1')U_2 $$

(4.3)

See (3.14). Rewrite $A$ as,

$$ A = VW' $$

( singular value decomposition )

(4.4)

Here $V$ and $W$ are column unitary and orthonormal, and $\Sigma$ is an $n*(d-p)$ diagonal matrix with $\text{diag}(\Sigma) = (\sigma_1, .., \sigma(d-p))$, where $\sigma_1 \geq .. \geq \sigma(d-p)$. Then it follows that,

$$ A^* = W\Sigma^*V' $$

(4.5)

$$ AA^* = V\\Sigma V' $$

(4.6)

Here $\Sigma^*$ is a $(d-p)*n$ diagonal matrix with $\text{diag}(\Sigma^*) = (1/\sigma_1, .., 1/\sigma(d-p))$ and $\Sigma$ is an $n*n$ diagonal matrix whose first $(d-1)$ diagonal elements are 1 and the rest 0. Let us rewrite $T$ as

$$ T = DD' \text{ with } D = [I_{(d-p)} \ 0]' = n*(d-p) \text{ matrix} $$

(4.7)

Then with the following definitions,

$$ r = D'Ve \quad (r = (d-p)-\text{vector} = \text{noise vector}) $$

(4.8)

$$ b = D'W\theta_2 \quad (b = (d-p)-\text{vector} = \text{'bias' vector}) $$

(4.9)

where

$$ r \in N(0, \sigma^2I_{(d-p)}) $$

(4.10)
because \( r \) is a linear combination of normally distributed variables and \( \text{E}(r) = 0 \) and \( \text{cov}(r) = I_{(d-p)} \), we can rewrite (4.3) as,

\[
|\text{AA}'y|^2 = |\text{VEW}'\theta_2|^2 + 2e'\text{VEW}'\theta_2 + |\text{VDD}'\nu'|^2
\]

\[
= |\text{VEW}'\theta_2|^2 + 2e'\text{VDD}'\text{VEW}'\theta_2 + |\text{D}'\nu'|^2
\]

\[
= |b|^2 + 2r^\prime b + |r|^2
\]

Here,

\[
|r|^2 / \sigma^2 \sim \chi^2 (d-p)
\]

\[
2r^\prime b / \sqrt{(4a'bb')} \sim N(0,1)
\]

Let us now investigate the asymptotic behaviour of the selection criterion. First we have to make an assumption. In the following it is assumed that,

\[
|b|^2 = |(I_n - U_1U_1')U_2\theta_2|^2
\]

increases proportionally with \( n \). We say that this factor is of order \( n (= O(n)) \). This means that future input samples must contain enough information and that the input power does not change in time. Let us now investigate what happens with the selection of a model when \( n \) tends to infinity. Let \( Ca \) and \( Cb \) denote the value of the criterion with model \( A \) and \( B \). Then we can write,

\[
Cb - Ca = |\text{AA}'y|^2 + Sb - Sa
\]

Here \( Sb - Sa < 0 \). When \( Cb - Ca \) is positive we choose model \( A \). Thus when,

\[
Cb - Ca > 0 \iff b^\prime b + 2r^\prime b + r^\prime r + Sb - Sa > 0 \iff b^\prime b / F + 2b^\prime r / F + r^\prime r / F + (Sb - Sa) / F > 0
\]

with \( F = (4a'bb') (= O(n)) \) we see that,

\[
b^\prime b / F = 1/4a^2 > 0
\]

\[
\text{plim} \ 2b^\prime r / F = \text{plim} N(0,1) / \sqrt{F} = 0
\]

\[
\text{plim} r^\prime r / F = \text{plim} \chi^2 (d-p) * a^2 / F = 0
\]

and finally,

\[
\text{plim} (Cb - Ca) > 0 \text{ if and only if } \text{plim} (Sb - Sa) / F > 1/4a^2
\]
Thus we see that when model A is true and \( n \) tends to infinity the right model is chosen if the order of \( S_b - S_a \) is smaller than \( O(n) \).

Let us now assume that model B is true. Because \( \theta_2 = 0 \) we have,

\[
Cb - Ca = r'r + Sb - Sa \tag{4.21}
\]

When this factor is negative, model B is chosen. Thus,

\[
\operatorname{plim} Cb - Ca < 0 \quad \text{if and only if} \quad n \to \infty
\]

\[
\operatorname{plim} r'r + Sb - Sa < 0 \tag{4.22}
\]

Here we see that when \( S_a - S_b \) is an increasing function of \( n \) such that relation (4.22) is true, we asymptotically choose the right model.

Let us take an example. Suppose,

\[
C = RSSp + 2p \sigma^2 \tag{4.23}
\]

This is the \( C_p \) or AIC when \( \sigma^2 \) is known. First assume model A is true. Then we see that condition (4.20) is satisfied because,

\[
\lim_{n \to \infty} 2(p-d) \sigma^2 / F = 0 \tag{4.24}
\]

However when model B is true the condition (4.22) cannot be satisfied because \( 2(d-p) \sigma^2 \) is not an increasing function of \( n \). Although \( n \) tends to infinity there always exists a risk,

\[
P(\quad r'r > 2\sigma^2(d-p)\quad ) = P(\quad \chi^2(d-p) > 2(d-p)\quad ) \tag{4.25}
\]

that the wrong model is chosen.

Although we obtained this result for the case in which we had to make a choice between two model structures, we can say that in the general case in which we have \( d \) possible model structures the \( C_p \) criterion is not consistent as an estimator of the true model. There is always a risk that we overestimate the true model dimension.

Let us now investigate the SC criterion and assume \( \sigma^2 \) is known. Then,

\[
Cb - Ca = b'b + 2b'r + r'r + \sigma^2 \ln(n)(p-d) \tag{4.26}
\]

Here we see that the last term has order \( O(\ln(n)) \). Thus suppose model A is the right one. Then with,

\[
\lim_{n \to \infty} \ln(n)/n = 0 \tag{4.27}
\]

we see (4.20) that,

\[
\operatorname{plim} Cb - Ca > 0 \tag{4.28}
\]
Thus the right model is chosen. When model B is true we see that,

\[ \text{plim } C_b - C_a = \text{plim } r^r + \sigma^2 \ln(n)(p-d) < 0 \quad (4.29) \]

So also in this case the SC chooses the asymptotically the right model. By similar reasons it can be shown that Akaike's SIC also chooses asymptotically the right model.

In literature, some authors have studied the asymptotic behaviour of several criteria. Nishii (184) showed that AIC, Cp, FPE and PRESS are all asymptotically equivalent. Furthermore he proved that the SC gives a consistent estimator of the true model structure. Shibata (181) proved that when the number of parameters is infinite or increases with the sample size, his criterion,

\[ S_n(p) = (n+2p)RSS_p/n = (n+2p) \hat{\sigma}^2(p) \quad (4.30) \]

is an asymptotically optimal criterion, in the sense that it attains a lower bound for MSE(x). He showed that the Cp, AIC and the FPE are all asymptotically equivalent to this criterion. Furthermore he showed that in this case the SC is not asymptotically optimal in his sense.

### 4.2 F-TEST INTERPRETATION

In chapter 3 I discussed some criteria that have been proposed for model selection. Although some of them are known to be asymptotically equivalent, their small sample properties are all different. Since most of these criteria are simple functions of the residuals sum of squares (RSS) it is interesting and informative to interpret them through the F-test statistic.

First, let us consider two models A and B with,

- process \( y = U_1 \theta_1 + U_2 \theta_2 + e \)
- model A: \( y = U_1 \theta_1 + U_2 \theta_2 + w \) \quad (4.31)
- model B: \( y = U_1 \theta_1 + w \)

From chapter 2 (2.58-2.61) we know that,

a: Under the mean square error function in the 'parameter space' the model B gives better results when,

\[ \| \theta_2 \|^2 + \| U_1^T U_2 \theta_2 \|^2 + \sigma^2 \left( \text{tr}(U_1^T U_1)^{-1}) - \text{tr}(U^T U)^{-1}) \right) < 0 \quad (4.32) \]

b: Under the mean square error function in the 'measurement space' the model B gives better results when,

\[ \| (I_n - U_1 U_1^T) U_2 \theta_2 \|^2 + \sigma^2(p-d) < 0 \quad (4.33) \]
c: Under the risk matrix the model B gives better results when,

\[ \begin{vmatrix} (I_n - U_1^T U_1) U_2 \end{vmatrix} \cdot \sigma^2 < 0 \] (4.34)

Let us look at condition (4.32). Define,

\[ H = U_2^T (I_n - U_1^T U_1) U_2 \] (4.35)

\[ H \] is positive definite (since \( U \) has full rank) and therefore we can rewrite it as,

\[ H = WW^T \text{ with } D = W \Sigma \] (4.36)

Where \( W \) is orthonormal and \( \Sigma \) is a positive definite diagonal matrix. So \( D \) has full rank. Then we can write,

\[ \Theta_2 \Theta_2 = \Theta_2 (U_1^T U_1)^T (U_1^T U_2) \Theta_2 = \]

\[ \Theta_2 (I (d-p) + (U_1^T U_2)^T (U_1^T U_2) ) \Theta_2 = \]

\[ \Theta_2 D D^{-1} (D^{-1})^T \Theta_2 = \]

\[ \Theta_2 D G \Theta_2 \] (4.37)

with,

\[ G = D^{-1} (I (d-p) + (U_1^T U_2)^T (U_1^T U_2) ) (D^{-1})^T \]

(4.38)

Because \( G \) is symmetric it holds that,

\[ \max_{x \neq 0} x^T G x / x^T x = \text{tmax} \]

(4.39)

Where \( \text{tmax} \) is the largest eigenvalue of \( G \). Therefore,

\[ \Theta_2 D G \Theta_2 \leq \text{tmax} \Theta_2 D D^{-1} \Theta_2 = \text{tmax} \Theta_2 H \Theta_2 \] (4.40)

This means that for all possible \( \Theta_2 \) the condition (4.32) will be true if,

\[ \begin{vmatrix} (I_n - U_1^T U_1) U_2 \end{vmatrix} \cdot \sigma^2 < \text{tr} ( (U_1^T U_1)^{-1} ) \cdot \text{tr} ( (U_1^T U_1)^{-1} ) \] / \( \text{tmax} \) (4.41)

When in the conditions (4.33), (4.34) and (4.41) \( \Theta_2 \) and \( \sigma^2 \) are replaced with their estimators \( \hat{\Theta}_2 \) and \( RSSp / (n-d) \), we can see that these conditions can be stated in terms of the F-test statistics associated with testing \( \Theta_2 = 0 \). So with,

\[ F = (n-d) * (RSSp - RSSd) / ((d-p) * RSSd) \]

(4.33) (4.34) and (4.41) become respectively,
\[ F < 1 \quad (4.42) \]

\[ F < 1/(d-p) \quad (4.43) \]

and

\[ F < \sigma^2 \left( \text{tr} \left( (U'u_1)^{-1} \right) - \text{tr} \left( (U'u)^{-1} \right) \right) / \text{tmp} \quad (4.44) \]

Let us now look at some of the criteria of the previous section. Then we see that model B is selected when the following conditions hold,

**F-test:**

\[ F < F_a (d-p, n-p) \quad (4.45) \]

**Mallows' Cp (with \( \sigma^2 \) known):**

\[ F < 2 \quad (4.46) \]

**Amemiya's PC (=FPE):**

\[ F < 2n/(n+p) \quad (4.47) \]

**Akaike's AIC (\( \sigma^2 \) unknown):**

\[ F < \left[ \exp \left( \frac{2(d-p)}{n} \right) - 1 \right] \frac{(n-d)/(d-p)}{1} \quad (4.48) \]

**Schwarz' SC**

\[ \sigma^2 \text{ known :} \]

\[ F < \ln(n) \quad (4.49) \]

\[ \sigma^2 \text{ unknown :} \]

\[ F < \left[ \exp \left( \ln(n)(d-p)/n \right) - 1 \right] \frac{(n-d)/(d-p)}{1} \quad (4.50) \]

For BIC and PRESS no such expressions can be derived. Since most of the discussed criteria can be seen as an F-test with a particular value of \( \alpha \), it is interesting to study the sampling properties of such an F-test. This will be done in the special case that \( U'u=I \) (which would approximately be the case when \( U \) is filled with white noise).

Suppose we have a general F-test where model B is chosen when \( F \) is smaller than a certain number \( c \) and where model A is chosen when \( F \) is greater than \( c \). The objective is now to investigate the mean squared error of estimating \( \theta \) of the F-test as a function of the length of \( \theta \). Since this MSE(\( \theta \)) can be written as,

\[ \text{MSE}(\theta) = \text{E} \left( (\theta_1 \cdot \hat{\theta}_1)'(\theta_1 \cdot \hat{\theta}_1) \right) + \text{E} \left( (\theta_2 \cdot \hat{\theta}_2)'(\theta_2 \cdot \hat{\theta}_2) \right) \]

\[ = p*\sigma^2 + \text{MSE}(\theta_2) \quad (4.51) \]

When \( c=0 \), so model A is always chosen, then \( \text{MSE}(\theta_2)=(d-p)\sigma^2 \) and \( \text{MSE}(\theta) = d^*\sigma^2 \). From Sclove (1972) and Judge and Sock (1978) we know that for \( c>0 \),

\[ \text{MSE}(\theta_2) > (d-p)^*\sigma^2 \text{ if } |\theta_2|^2 > (d-p)^*\sigma^2 \quad (4.52) \]

\[ \text{MSE}(\theta_2) < (d-p)^*\sigma^2 \text{ if } |\theta_2|^2 < (d-p)^*\sigma^2/2 \quad (4.53) \]

\[ \text{MSE}(\theta_2) = (d-p)^*\sigma^2*L \text{ if } |\theta_2|^2 = 0 \quad (4.54) \]

where \( L=P \left( \chi_i^2(d-p+2)/\chi_i^2(n-d) > c^*(d-p)/(n-d) \right) \)
Furthermore as $\Theta^T \Theta$ increases and approaches infinity, the $\text{MSE}(\Theta_2)$ approaches the risk of the least square estimator.

Knowing these properties a plot of $\text{MSE}(\Theta_2)$ versus the length of $\Theta_2$ will yield the following characteristic (see Judge and Bock (78)).

![Graph](image)

**Figure 1: MSE($\Theta_2$) versus $|\Theta_2|^2/\sigma^2$ for the F-test.**

From figure 1 we can conclude that the F-test performs better than the least squares estimator ($c=0$) if the fault made in the hypothesis is small, but over a large range in the $\Theta_2$ parameter space it is inferior to the least squares estimator.

In section 3.8 I discussed the positive James-Stein estimator and said that under certain conditions it can be used in hypothesis testing. It was said that in those cases the estimator is uniformly better than the least squares estimator.

Sclove (72) combined the normal F-test with the positive James-Stein rule and constructed a modified F-test of the following form,

\begin{align*}
\text{choose model B when } & F \leq c, \\
\text{choose model A when } & F > c \quad \text{and estimate } \Theta_2 \text{ by} \\
\Theta_2 &= \frac{1 - (a(d-p-2)/(n-d+2))^* \text{RSS}/\Theta_2^T \Theta_2}{a} \Theta_2
\end{align*}

where $0 < a \leq 2$

He proved that this estimator is uniformly better, in $\text{MSE}(\Theta)$ sense, than the normal F-test with decision level $c$ (in this case where $U^T U = I_n$).

If we combine the results from the F-test, positive James-Stein estimator and the modified F-test in one plot we get the following characteristics (see Judge and Bock '78).
Conclusion:

In this section we saw that most criteria can be written in terms of F-test statistics with a particular value of $\sigma$. We studied the sample performance of such a F-test in the case that we had to select the best out of two models and that the general design matrix $U$ was orthonormal. A plot of the risk involved with such a F-test yields a range of the parameter space where the risk is smaller than the risk of the least squares estimator of the full model and a range where it is greater.

But even when $U$ is not orthonormal there is a range in the parameter space where the risk ($\text{MSE}(\hat{\theta})$) is smaller than the risk of the least squares estimator of the full model and a range where the risk is greater. (See Judge and Bock 178).

An exception to this general rule is the positive James-Stein estimator. Here the risk will, under certain conditions, be uniformly better than that of the LSE.
4.3 THE PARAMETER SPACE

As already mentioned, most of the discussed model selection procedures are simple functions of the residuals sum of squares and can be seen as criteria that are designed for improving the prediction and thus MSE(x). However, when parameter estimation is the object, for example for controlling purposes, we have to improve the estimators by looking at MSE(θ). When U is orthonormal minimizing of MSE(x) is equal to minimizing MSE(θ). But when U is not orthogonal they are not the same. Noticing this difference between MSE(x) and MSE(θ) it is possible to derive another class of ad hoc procedures for model selection.

4.3.1 THE MODIFIED Cp

As we saw, Mallows derived his Cp by giving an estimate of the MSE(x). Following his derivation it is possible to derive an unbiased estimate of MSE(θ). We can write MSE(θ) as (see 2.52).

\[
\text{MSE}(θ) = |Θ_2|^2 + |U_1^*U_2^* Θ_2|^2 + σ^2 \text{tr} \left( (U_1^*U_1)^{-1} \right)
\]  (4.56)

Since Θ_2 is unkown we replace it with its unbiased estimate. Let \( \hat{Θ}_2 \) denote the estimate of Θ_2 from the full model. First, let us look at the estimate of Θ in the full model.

\[
\hat{Θ} = \begin{bmatrix}
Θ_1 \\
Θ_2
\end{bmatrix} = U^+ γ = \begin{bmatrix} T_1 \\
T_2
\end{bmatrix} γ
\]  (4.57)

Here

\[
T_1 = U_1^* - U_1^* U_2 T_2
\]

is a p*n matrix

\[
T_2 = A^+
\]

is a (d-p)*n matrix

\[
A = (I_n - U_1^*U_1)U_2
\]

is a n*(d-p) matrix

(see appendix A)

So if we estimate Θ_2 by \( \hat{Θ}_2 \) from the full model we see that,

\[
E( |\hat{Θ}_2|^2 ) = |Θ_2|^2 + σ^2 \text{tr} (T_2^*T_2)
\]  (4.58)

\[
E( |U_1^*U_2^* \hat{Θ}_2|^2 ) = |U_1^*U_2^* Θ_2|^2 + σ^2 \text{tr}( (U_1^*T_1)'(U_1^*T_1))
\]  (4.59)

With these relations and with the estimate \( \text{RSSd}/(n-d) = \hat{σ}^2(d) \) of \( σ^2 \) we get that,

\[
|\hat{Θ}_2|^2 + |U_1^*U_2^* \hat{Θ}_2|^2 + \hat{σ}^2(d) \text{tr}( (U_1^*U_1)^{-1} )
\]

\[
- \hat{σ}^2(d) \text{tr}(T_2^*T_2) + \text{tr}( (U_1^*T_1)'(U_1^*T_1) )
\]  (4.60)

is an unbiased estimate of MSE(θ). When we add the following constant term
(constant in the sense that it does not depend on the dimension of the model p),
\[
2\sigma^2\left|\mathbf{U}^*e\right|^2 = 2\sigma^2(d)\cdot\text{tr}\left(\mathbf{U}'\mathbf{U}^{-1}\right)
\]  
(4.61)

to (4.60) we get an easier to handle unbiased estimate (Np = New Cp criterion) of \(\text{MSE}(\Theta)\),
\[
\text{Np} = \left|\hat{\Theta}_p - \hat{\Theta}_f\right|^2 + \sigma^2(d)\cdot\text{tr}\left(2\sigma^2(\mathbf{U}'\mathbf{U}^{-1})^{-1}\cdot\text{tr}\left(\mathbf{U}'\mathbf{U}^{-1}\right)\right)
\]
\[
= \left|\mathbf{U}'(1-\mathbf{U}'\mathbf{U})\mathbf{y}\right|^2 + \sigma^2(d)\cdot\text{tr}\left(2\sigma^2(\mathbf{U}'\mathbf{U}^{-1})^{-1}\cdot\text{tr}\left(\mathbf{U}'\mathbf{U}^{-1}\right)\right)
\]  
(4.62)

with
\[
\hat{\Theta}_p = \begin{bmatrix} \hat{\Theta}_1 \\ 0 \end{bmatrix} = \hat{\Theta} \text{ of model with dimension } p
\]
\[
\hat{\Theta}_f = \mathbf{U}'\mathbf{y} = \hat{\Theta} \text{ of full model}
\]

So this Np criterion looks very like the original Cp criterion. When \(\mathbf{U}'\mathbf{U} = \mathbf{I}_p\) minimizing of Np is equal to minimizing Cp.

Since the Np criterion is especially designed to minimize \(\text{MSE}(\Theta)\) the question arises if it is indeed better than Cp or not. As a measure of performance we should calculate the risk of estimating \(\Theta\) involved with these two criteria. However, this can be a very complicated problem. For example, let us assume that we use Np and Cp to select the best out of 2 possible model structures where,

- process: \(y = \mathbf{U}\mathbf{\theta} + \mathbf{e}\)
- model A: \(y = \mathbf{U}\mathbf{\theta} + \mathbf{w}\)
- model B: \(y = \mathbf{w}\) thus all parameters are set to zero

and assume that \(\mathbf{\theta}\) is known. Then we choose model B when,

- Np criterion
  \[
  \ln = \left|\mathbf{U}'\mathbf{y}\right|^2 - 2\sigma^2\cdot\text{tr}\left(\mathbf{U}'\mathbf{U}^{-1}\right) \leq 0
  \]  
  (4.63)

- Cp criterion
  \[
  \mathbf{L} = \left|\mathbf{y}\right|^2 - \left|\mathbf{1}_p - \mathbf{U}'\mathbf{U}\right|\mathbf{y}\right|^2 - 2\sigma^2(d) \leq 0
  \]  
  (4.64)

otherwise we choose model A. To obtain the risk involved with each of these criteria we have to compute,

- Np criterion
  \[
  \text{MSE}(\text{Np},\Theta) = E\left(\max(0,-\text{sign}(\ln))\cdot|\hat{\Theta}|^2\right) + E\left(\max(0,\text{sign}(\ln))\cdot|\hat{\Theta} - \hat{\Theta}_f|^2\right)
  \]  
  (4.65)
Cp criterion

\[ \text{MSE}(C_p, 0) = \mathbb{E}\{ \max(0, -\text{sign}(L_c) ||\theta||^2) \} + \mathbb{E}\{ \min(0, \text{sign}(L_c) ||\theta - \hat{\theta}||^2) \} \]

(4.66)

where,

\[ \text{sign}(h) = 1 \text{ if } h > 0 \]
\[ \text{sign}(h) = -1 \text{ if } h \leq 0 \]

Let us try to calculate \( \text{MSE}(C_p, 0) \) when \( \theta = 0 \). Define,

\[ U = \mathbb{V} \mathbb{W}' (s.v.d.) \]

(4.67)

so,

\[ U^* = \mathbb{W} \mathbb{V}' \]

(4.68)

Then we see that \( \text{MSE}(C_p, 0) \) becomes,

\[ \mathbb{E}\{ e' \mathbb{V} e \} \]

under the condition that \( e' \mathbb{W} \mathbb{V}' e > 2\sigma^2 \)

Define \( \Sigma = DD' \), with \( D = [I_\sigma \ 0] \) is a \( n\times d \) matrix, and \( \Sigma = DD' \) where \( T \) is a \( d\times d \) positive definite diagonal matrix. Then with,

\[ n = D'Ve \sim \mathcal{N}(0, \sigma^2 I_d) \]

(4.70)

we get,

\[ \text{MSE}(C_p, 0) = \mathbb{E}\{ n' T n \} \text{ for } n' T n > 2\sigma^2 \]

(4.71)

Similar for \( \text{MSE}(N_p, 0) \) we get

\[ \text{MSE}(N_p, 0) = \mathbb{E}\{ n' T n \} \text{ for } n' T n > 2\sigma^2 \text{tr}(T) \]

(4.72)

From these equations we see that they are equivalent when \( d = 1 \). For \( d > 2 \) these functions are very difficult to compute. Let us study the case when \( d = 2 \). Although the area of the ellipsoid formed by \( n' T n = 2\sigma^2 \text{tr}(T) \) is greater than that of the circle \( n' n = 2\sigma^2 \), the calculation of \( \text{MSE}(N_p, 0) \) is performed over an area where the values of \( F(n) = (a n_1' + b n_2')p(n_1)p(n_2) \) (here \( \text{diag}(T) = (a, b) \) ) are greater than those of the area in calculation of \( \text{MSE}(C_p, 0) \). To illustrate this I have made a 3-dimensional plot of the areas that have to be computed for some \( a \) and \( b \) (see figure 3). Simulations for various combinations of \( a \) and \( b \) confirmed this. All the simulations in this report were done with the package PC-Matlab. Some other simulations for \( d > 2 \) also indicated that \( \text{MSE}(C_p, 0) \leq \text{MSE}(N_p, 0) \). So in this ad hoc derivation we saw that \( \text{MSE}(C_p, 0) \leq \text{MSE}(N_p, 0) \).
Since these results only hold for $\theta = 0$ we expect that the calculations of the risk for $\theta \neq 0$ will not become any easier. Because an evaluation of the risk involved by choosing the best out of $d$ possible model would complicate the calculation again we are forced to make some comparison between these criteria by means of more simulations.

This has been done. I have done two types of simulations. First I simulated the $\text{MSE}(\theta)$ for the $\text{Ip}$ and the $\text{Cp}$ criterion assuming that there had to be made a choice between the full model and the model with dimension 0. This MSE has been simulated as a function of the length of $\theta$. Secondly I simulated the $\text{MSE}(\theta)$ for the $\text{Ip}$ and the $\text{Cp}$ criterion assuming that there had to be made a choice between 10 possible model structures (dimension 0 to 9). This MSE has again been simulated as a function of $\theta$. 

Figure 3: Some 3-dimensional plots of the risk computation if $\theta=0$
SIMULATIONS-1:

First I simulated the MSE as defined in (4.65) and (4.66) as a function of the length of \( \theta \), assuming a fixed orientation for \( \theta \). In that simulation I took some predefined matrix \( U \), calculated the undisturbed output \( x \) and disturbed it with a noise vector. From the resulting vector \( y \), I let the \( N_p \) and the \( C_p \) criterion decide what model structure to take. From that chosen model structure I calculated the resulting error between the actual \( \theta \) and the estimated \( \theta \). To obtain reasonable results, the resulting MSE has been averaged over 200 simulations. This procedure has been done for 100 different values of the length of \( \theta \).

The preceding procedure has been done for two different orientations of \( \theta \). In both cases the \( U \) matrix was the same. The difference between the two cases was that in one case the norm of the undisturbed output vector \( x \) was greater than that in the other case (for equal length of \( \theta \)).

Values:

\[
U = \begin{bmatrix}
1 & 0 & 1 \\
0 & 1 & 0 \\
1 & 1 & 0 \\
0 & 1 & 1
\end{bmatrix}
\]

(\text{with } \text{tr}((U^TU)^{-1}) = 0.1454), \text{ } \sigma^2 = 1

<table>
<thead>
<tr>
<th>\text{case 1}</th>
<th>\text{case 2}</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \theta )</td>
<td>( a^* [1 \ 1 \ 1]' )</td>
</tr>
<tr>
<td>( x^'x )</td>
<td>( a^*189 )</td>
</tr>
</tbody>
</table>

RESULTS:

The results are shown in figure 4a and 4b for case 1 and in figure 5a and 5b for case 2. In these figures I plotted the simulated mean square error as a function of the squared length of \( \theta \). This is done for the \( N_p \) criterion, the \( C_p \) criterion and for the case that always the full model (FM) would be chosen.

In figure 4a-4b we see that when \( \theta^*\theta \) is small (\(<0.08\)) both the \( N_p \) and \( C_p \) criterion give better results than the FM. The difference between \( N_p \) and \( C_p \) is relatively small in this region. But we see that for a small value of \( \theta^*\theta \) the \( C_p \) criterion is slightly better than the \( N_p \) criterion and when \( \theta^*\theta \) is increased the \( N_p \) criterion becomes slightly better. However, if the squared length is further increased (\( >0.08 \)) we see that both the \( C_p \) as the \( N_p \) criterion give worse results than the FM (compare it with...
the characteristic in figure 1). But we see that in this region the $N_p$ criterion gives much worse results than the $C_p$ criterion.

To have an impression of the goodness of the 200 simulations per value of $	heta$, we can compare the theoretical value of the MSE for the full model (=0.14) with the simulated MSE. We see that the simulated MSE varies around this value.

When we look at case 2 (figure 5a and 5b) we see that there is a clear preference for the $C_p$ criterion for small squared lengths of $\theta$. However when $\theta^2$ is increased the $N_p$ criterion turns out to be better although they both give much worse results than the full model.
From the previous figures we may conclude that for small values of $\Theta^T\Theta$ the 
Cp criterion gives better results than the Np criterion. But if $\Theta^T\Theta$ is increased it depends on the orientation of $\Theta$ which criterion will be 
better. If $(x'x)/(\Theta^T\Theta)$ is 'small' the Np criterion turns out to be better, 
if not the Cp criterion turns out to be better.

Thus we can not say that one criterion is better than the other since 
there are always points in the parameter space where the reverse holds, 
but the Cp criterion has the advantage that for small $\Theta^T\Theta$ it is always 
better than the Np criterion.

SIMULATIONS-2

In the previous section the Np-Cp criterion had to make a choice between 2 
model structures. In this section the Np-Cp criterion has to make a choice 
between 10 different model structures.

I constructed several $16 \times 9$ matrices $U$ in such a way that I could control 
the condition number (= largest singular value/smallest singular value of $U$) and the sum of squared elements (= I define this as the 'signal power') of it. Furthermore I considered two different orientations of $\Theta$. Per 
orientation I considered 21 different lengths of the actual $\Theta$. With a 
particular $U$ and $\Theta$ I calculated the undisturbed output $x$ and disturbed it 
with white noise to obtain the output vector $y$. As the full model I took 
the true model (So the highest to be regarded dimension of the model is the 
dimension of the true model = 9). From the vector $y$ I calculated in each of 
the 10 possible model structures (dimension 0 to 9) the value of some 
criteria. The model structure with the smallest value was chosen according 
to that criterion.

In this chosen model structure the difference between the actual and the 
estimated $\Theta$ was calculated. The simulated MSE($\Theta$) was approximated by 100 of 
such simulations.

Besides the Cp, Np and the FM (=always choosing the full model) criterion 
as in the previous section I considered two other criteria. In each model 
structure I calculated the squared difference between the actual and 
estimated $\Theta$ (thus I assumed that $\Theta$ was known). The model structure with the 
smallest squared difference between the actual and estimated $\Theta$ ($=L(\Theta)$) 
was chosen. This has also been done for the squared difference between the 
actual and estimated $x$ ($=L(x)$). These last two criteria can not be 
calculated in practice since the actual $\Theta$ or $x$ are not known. They are only 
computed to have some idea about the lower bounds.
VALUES:

The two orientations of the impulse responses were defined as follows. (see table 1 and figure 6)

Table 1: Impulse responses

<table>
<thead>
<tr>
<th>impulse response A</th>
<th>impulse response B</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ h(i) = a \cdot e^{-0.25i} ]</td>
<td>[ h(i) = a \cdot (\cdot -1)^i e^{-0.25i} ]</td>
</tr>
</tbody>
</table>

Figure 6: The two orientations of the impulse response

The \( U \) matrices were constructed in the following way. First I took a matrix and filled it with white noise samples. To control the condition number I added the same constant term to each element of it and finally I multiplied the matrix in such a way that I would obtain the desirable signal to noise ratio \( (x'x)^{-1}/noise^2 \). The variance of the noise \( \sigma^2 \) was set to 0.01 and the signal to noise ratio was set to 100. I considered 4 different \( U \) matrices. For some characteristics of these matrices see table 2. In this table is also the ratio \( (x'x)_A/(x'x)_B \) given, which is the ratio of the norms of the undisturbed output \( x \) with the impulse response \( A \) and \( B \) (for equal length of \( \theta(A) \) and \( \theta(B) \)).

Table 2: \( U \) characteristics

<table>
<thead>
<tr>
<th></th>
<th>case 1</th>
<th>case 2</th>
<th>case 3</th>
<th>case 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>( tr(U'U)^{-1} )</td>
<td>7.565</td>
<td>11.40</td>
<td>153.0</td>
<td>331.6</td>
</tr>
<tr>
<td>condition number</td>
<td>2.868</td>
<td>4.658</td>
<td>38.80</td>
<td>50.55</td>
</tr>
<tr>
<td>( \frac{(x'x)_A}{(x'x)_B} )</td>
<td>1.776</td>
<td>0.8648</td>
<td>11.92</td>
<td>14.30</td>
</tr>
</tbody>
</table>
RESULTS:

For each matrix $U$, I made a plot of the $\text{MSE}(\theta)$ (as a function of $\theta^2\theta$) for the two orientations of $\theta$. As already mentioned, these plots were made of 21 different values of $\theta^2\theta$ and each value of $\text{MSE}(\theta)$ was approximated by 100 simulations. For the results of case 1-4 see figure 7a, 7b-10a, 10b.

In each figure, the $\text{MSE}(\theta)$ was simulated for the $Wp, C_p, FM, L(x)$, and $L(\theta)$ criterion. The $\text{MSE}(\theta)$ in the $L(\theta)$ criterion is the lower bound for the $\text{MSE}(\theta)$ which can be achieved with a criterion.

---

**Figure 7a:** Impulse response A

**Figure 7b:** Impulse response B

**Figure 7:** Case 1, $\text{MSE}(\theta)$ for $L(\theta)$ (--), $L(x)$ (----), $C_p$(...), $Wp$ (---), $FM$(----)

**Figure 8a:** Impulse response A

**Figure 8b:** Impulse response B

**Figure 8:** Case 2, $\text{MSE}(\theta)$ for $L(\theta)$ (--), $L(x)$ (----), $C_p$(...), $Wp$ (---), $FM$(----)
In figure 7a-7b we have a matrix with a low condition number (=2.868) and we see that for both the impulse response A and B the Cp criterion gives better results than the Np criterion (for the given range in the parameter space). In figure 8a-8b we have another matrix with a low condition number (=4.658), but the range in the parameter space is enlarged. Here we see that for small values of $\theta$ the Cp criterion gives better results for both the actual impulse response A and B, but if $\theta$ is large it depends on the impulse response which criterion is better. However, this behaviour could be expected. In the previous section we saw that the Np criterion gave better results (in choosing between two model structures) when the norm of $x$ was relatively small. In this case where when $\theta$ is large not many parameters can be deleted. So we are in fact choosing between the full model with dimension 9 and the model with dimension 8 (or yet one dimension lower). Since $(x'x)_A$ is lower than $(x'x)_B$ we can therefore expect that the
61

The Np criterion can be better than the Cp criterion with impulse response A. In both the figures 7a-7b and 8a-8b we see that there is not much difference between the L(Ω) criterion and the L(x) criterion.

In figure 9a-9b we have a matrix with a large condition number (=38.80) and in these figures we see that for low values of Ω we the Cp criterion gives better results and for large Ω it depends on the actual impulse response which criterion is better. In this case we also see a large variation between the L(Ω) criterion and the L(x) criterion.

In figure 10a-10b we have another matrix with a large condition number (=50.55). Here we see that, for both the impulse response A and B, the Cp criterion is better than the Np criterion (for this range of the parameter space) and that there is a small difference between the L(Ω) and L(x) criterion.

From these figures we see that for small values of Ω the Cp criterion gives better results than the Np criterion. If Ω is large it depends on the actual impulse response which criterion gives better results (and thus how long Cp remains better).

Conclusions

From the previous simulations it turns out that this Np criterion, which was based on the idea that it might be better to base a selection criterion on the parameter space rather than on the measurement space, does not give a clear improvement above the Cp criterion. In fact, the results may become much worse.

Simulations indicated that when the actual parameters are indeed near zero the Cp criterion seems to be superior to the Np criterion. But even when the parameters are large the Cp criterion seems to be a good choice in many cases.

Therefore, if in practice the full model is chosen in such a way that the model is chosen too large and thus the last parameters are indeed near zero, the Cp criterion seems to be a good choice.

So evaluating these results I think that we can conclude that it is not advisable to use this Np criterion because of the following reasons,

(1) it could not prove its superiority to Cp
(2) it requires more computation than Cp
4.4 ORDERED-NONORDERED PARAMETER REDUCTION

Sofar I only discussed the case in which we didn't have any idea about the length of the actual impulse response and thus were forced to choose a model according to a certain criterion. However, the criteria in chapter 3 can also be used in other situations. For example for estimating the delay of an impulse response. In both cases we are testing if we can set of ordered parameters to zero. I will call this manner the ordered parameter reduction.

A further step would be to test if we could set any single parameter to zero. In that case we should choose the best subset of parameters according to a criterion. I will call this manner the nonordered parameter reduction (or subset selection). The aim of this chapter is to mention some topics that are highly related to nonordered parameter reduction.

4.4.1 COMPUTATIONAL PROBLEMS

Suppose the aim is to select the best subset out of $d$ possible parameters. Then there are,  
$$2^d$$  (4.73)  
possible subsets to examine. The subset that minimizes a certain criterion (for example $C_p$ of AIC) should be chosen. There are some algorithms for efficiently handling the computation of all possible subsets (for a review on this topic see Hocking '76). The underlying idea of most of these algorithms is to perform the calculation of all possible subsets in such a way that sequential subset models differ by only one variable.

Another idea is to use branch and bound techniques. Suppose we are searching for the best subset (in RSS sense) of dimension $p$. Thus we have to delete $(d-p)$ parameters from the full model. For example let $(d-p)=4$ and let RSS$(a,b,c,d)$ denote the residuals sum of squares when parameters $a, b, c, d$ are deleted from the full model. The underlying idea of these branch and bound algorithms is that if RSS$(e) \geq$ RSS$(a,b,c,d)$ then, for every set of parameters where $e$ is involved is RSS$(e, ?, ?, ?) \geq$ RSS$(a,b,c,d)$ and thus we should not search in that direction. However when RSS$(e) <$ RSS$(a,b,c,d)$ additional subsets have to be evaluated. After finding the best subset for each dimension it is possible to select the overall best subset according to one of the previous discussed selection criteria.

When the number of parameters is large it is usually not feasible to examine all possible subsets. Therefore various methods have been developed for evaluating only a small number of subsets by either adding or deleting parameters one at a time. These methods are generally referred to as stepwise methods. The basic ideas behind these methods are called forward selection (FS) and backward elimination (BE).
FORWARD SELECTION.

First we start with a model with no parameters. Then we select the parameter which gives the greatest reduction in the residuals sum of squares sense. When this reduction is smaller than the predefined value Fin the procedure is stopped. When not, we include this parameter in our model and we check if we can add another parameter in the model. Of course the number of variables included highly depends on Fin. One way of choosing Fin is to assume that the i-th variable is the last to enter variable. Then is,

\[
F = \frac{RSS_p - RSS_{p+1}}{RSS_{p+1}/(n-p-1)} \in F(1,d-p-1)
\]

where RSS_{p+1} denotes the RSS of the model where the i-th variable is included. Then the i-th variable is included if

\[
F > Fin = F_{a}^{2}(1,d-p-1)
\]

Of course we should again choose a value for a. Another possibility is to take Fin = 2 which is in accordance with the Cp rule.

BACKWARD ELIMINATION

This procedure is the reverse of the FS. In this case we start with all possible parameters and we check if we can delete some parameters one at the time. At any step the variable with the smallest reduction of the residuals sum of squares is chosen. Then the i-th variable is deleted from the model if,

\[
F = \frac{RSS_{p} - RSS_{p-i}}{RSS_{p}/(n-p)} \leq Fin
\]

Where RSS_{p-i} is the RSS of the model without the i-th variable. The same values for Fin as in FS can be used as a stopping rule.

Based on these two procedures several other procedures have been proposed such as the forward stepwise selection procedure, which is essentially the same as the FS procedure but now at each stage the possibility of deleting a variable is considered, or the backward stepwise selection procedure, which is essentially the same as the BE procedure but now at each stage the including of a variable is considered.

It should be noted that none of the stepwise methods assure us that we get the best overall subset. However there is a feeling that these methods will reveal subsets that are near optimal. The obvious advantage of these methods is that in the worst case only t+(t-1)+(t-2)+...+1 = t(t+1)/2 subsets have to be calculated.
4.4.2 PRIOR KNOWLEDGE

In this section I want to discuss the prior conditions which have to be met to assure us that the method of searching for the best subset will indeed give better results than taking all the parameters in our final model. This will be done in the simple case in which the \( U \) matrix is orthogonal and the variance of the noise is known. It is assumed that,

\[
U'U = \lambda^* I_d
\]  
(4.77)

Thus we can see that,

\[
\hat{\theta} \in \mathcal{N}(\theta, \sigma^2 I_d/l)
\]  
(4.78)

Let us denote \( s^2 = \sigma^2 / l \) as the variance of the noise in the parameter space. As a selection criterion consider the generalized AIC criterion,

\[
GAIC = \min RSS_p + a^* \sigma^2
\]  
(4.79)

This means that a variable is included in the model if its contribution to the decreasing of the RSS exceeds \( a^* \sigma^2 \). For the parameter space this condition means that a variable is included if,

\[
\hat{\theta}_i^T \theta_i > a^* \sigma^2 / k = a^* s^2
\]  
(4.80)

If this condition doesn't hold, the variable is excluded from the model. Because of the simplicity of the distribution of the parameters it is possible to calculate the resulting mean square error. Let us write,

\[
MSE(\theta) = \sum_{i=1}^{d} E(\hat{\theta}_i - \theta_i)^2 = \sum_{i=1}^{d} MSE(\theta_i)
\]  
(4.81)

Let \( MSE(\theta_i) \) denote the mean square error of the \( i \)-th variable. Since we don't know what the exact value of \( \theta_i \) is we leave it as a variable. Because of the normal distribution of \( \theta_i \) we can calculate \( MSE(\theta_i) \) as follows,

\[
MSE(\theta_i) = \int \int \theta_i^2 p(\theta_i - \theta_i^*) d\theta_i^* + s^2 - \int \int (\theta_i - \theta_i^*)^2 p(\theta_i - \theta_i^*) d\theta_i^* = \int \int \theta_i^2 p(\theta_i - \theta_i^*) d\theta_i^* + s^2 + (as - \theta_i^*)s'p(\theta_i - \theta_i^*) + (as + \theta_i^*)s'p(\theta_i^*) \]  
(4.82)
where,
\[ p(t) = \frac{1}{\sqrt{2\pi s^2}} \exp(-t^2/(2s^2)) \tag{4.83} \]
\[ F(t) = \int_{-\infty}^{t} p(r) \, dr \tag{4.84} \]

When we plot this \( \text{MSE}(\theta_1) \) versus \( \theta_1 \) in a figure we get the following results (see figure 11).

![Figure 11: MSE(\(\theta_1\)) versus \(|\theta_1|\)' for \(\theta'=0, \quad a'=1, \quad a'=2, \quad a'=4, a'=\text{infinity}\)\n
In this plot it is assumed that \(s'\) is equal to one, but this is not a severe restriction since the characteristics remain the same for a general \(s'\). In figure 11, \(\text{MSE}(\theta_1)\) was plotted for five different values of \(a\), namely 0, 1, 2, 4 and infinity. In this application these values correspond...
with a hypothesis test with significance level $\alpha$ of 100%, 32%, 16%, 5% and 0%. Of course a significance level of 100% means that we always include the variable and a significance level of 0% means that we always exclude the variable from the model. We see that the common characteristic is that when $|\theta_i|$ is small (depending on $a$) the GAIC is better than just including all the variables. However if $|\theta_i|$ is large the results become much worse. We can see that the overall behaviour (over all $\theta_i$) is worse than just including the variable.

So from this figure we must conclude that the GAIC will be better than the least-squares estimator from the full model, if there are enough parameters that are close to zero. So if some prior information about the parameters exists one can decide to perform a nonordered parameter reduction or not. Of course, in case of ordered parameter reduction similar conditions hold, but in those cases it is more likely that they are satisfied since most impulse responses tend to zero or have a delay time. So in that region the parameters are close to zero.

4.5 All Formulas Can Be Wrong

In the beginning of chapter 2, I said that the formulas that were going to be derived are based on the assumption that the model was selected without reference to the actual data. Furthermore we saw that almost all the selection criteria had their base in one or more of the derived formulas in section two. For example the Cp criterion is derived under the assumption that it is a (unbiased) estimate of the $\text{MSE}(x)$. However, since these selection criteria are used to select a model structure on the actual data the previous assumption can not be met, and so the derived formulas are invalid.

To illustrate that the usual properties of the least squares estimator are invalid, consider the example in section 4.4.2 where we discussed the nonordered parameter reduction when $U$ was orthonormal. Let us suppose that the $i$-th variable is included into the final model structure according to the $\min AIC$ condition. According to the properties of the least squares estimator the estimate of the $i$-th variable in this model structure should be unbiased. However if we calculate the actual expectation of the difference between the estimate and the actual $\theta_i$ we get,

$$
\mathbb{E}(\hat{\theta}_i - \theta_i) = \int (\hat{\theta}_i - \theta_i) p(\hat{\theta}_i, \theta_i) d\theta_i + \int (\hat{\theta}_i - \theta_i) p(\hat{\theta}_i, \theta_i) d\hat{\theta}_i
$$

as

$$
= s^T (p(\text{as} + \theta_i) - p(-\text{as} + \theta_i))
$$

(4.85)

So we see that we get a biased estimator unless $a=0$, which of course corresponds to taking all the variables in the final model.

In a paper on this topic Miller (1984) considered three types of bias that occurs in the least squares estimator when the model structure is chosen upon the data.
(1) omission bias
(2) stopping rule bias
(3) competition bias

The omission bias is the 'normal' bias that we get in the least squares estimator when we delete a variable from the full model (see 2.49). Stopping rule bias is the bias that we obtain when we are choosing the number of parameters to use. Competition bias is the bias that we get when we are choosing between subsets of the same size. The bias in the previous example should be regarded as a combination of stopping rule bias and competition bias.

The question is, how to eliminate these last two types of bias. Clearly a good solution would be to perform the model selection on one data set and to perform the estimation of the parameters on another data set, provided that the division of the data set into two halves is taken randomly. But of course in many fields, sample sizes are not large enough to make this method practical. Miller (’84) has mentioned some other solutions but it still remains a very difficult problem to tackle.
5. BIASED ESTIMATORS

In this chapter the primary concern is to look at methods that have been developed for improving the least squares estimates of the impulse response. Since the least squares estimator is known to be a minimum variance unbiased estimator we are forced to consider the class of unbiased estimators. We already discussed the James Stein estimator in chapter 3.8. Besides the James-Stein estimator there is class of estimators such as principal component regression and ridge regression where the object is to improve the least squares estimator when the matrix U suffers from multicollinearity, i.e. if there exists some near linear relationship between the columns of U. Because of the vast quantity of papers that exists in literature on these topics, the object is to discuss briefly some general ideas rather than to give an exhaustive review. In the following all results will be given under the assumption that the process is in the model set.

5.1 CANONICAL FORMS

In chapter 3.8 I already discussed the orthonormal canonical form of the general linear regression model (3.130-3.132). In this chapter I will present another canonical form which will be used for explanatory purposes. Consider the general model as given by,

\[ y = U\theta + e \] (5.1)

Let us rewrite \( U'U \) as (see 3.83)

\[ U'U = WTW' \] (5.2)

Here \( W \) is orthonormal and \( T \) is a positive definite diagonal matrix with \( \text{diag}(T) = (t_1, \ldots, t_d) \) \( t_1 \geq \ldots \geq t_d > 0 \) where the diagonal elements are the eigenvalues of \( U'U \). Define now,

\[ Z = U'W \] (5.3)

\[ \beta = W'\theta \] (5.4)

Then we have the transformed model,

\[ y = Z\beta + e \] (5.5)

Where,

\[ \hat{\beta} = Z'y = T^{-1}Z'y = W'\hat{\theta} \sim N(\beta, \sigma^2 T^{-1}) \] (5.6)

So we see that in the \( \beta \) space the parameters are independent, but they don't have equal variance. Furthermore it holds that,

\[ E( (\hat{\beta} - \beta)'(\hat{\beta} - \beta) ) = E( (\hat{\theta} - \theta)'(\hat{\theta} - \theta) ) = \text{MSE}(\theta) \] (5.7)

Thus a comparison between two estimator according to the MSE(\( \theta \)) function can also be done in the \( \beta \) space.
5.2 James Stein Estimators

From chapter 3.8 we know that the positive James-Stein estimator,
\[ \hat{\theta} = \max(0, S_f^{*} \cdot \theta) \]  
(5.8)

where
\[ S_f^{*} = [1 - (c/\nu) \cdot \hat{\sigma}^2 / (\hat{\nu} \cdot \hat{X}^T \cdot \hat{X})] \]
\[ \nu = (n - d), \quad \hat{\sigma}^2 = \text{RSS}_d / (n - d), \quad d \geq 3 \]

will be uniformly better than \( \hat{\theta} \) according to the loss function MSE(\( x \)) if,
\[ 0 < c < 2(d - 2) \]  
(5.9)

But how does this estimator perform under MSE(\( \theta \))? Judge and Bock (178) proved that this estimator will be uniformly better than the least squares estimator if,
\[ 0 < c < 2 \left( \frac{\text{tr} \left( (U'U)^{-1} \right)}{t_{\text{max}} - 2} \right) \]  
(5.10)

\( t_{\text{max}} \) is the largest eigenvalue of \( (U'U)^{-1} = (1/\sigma_d)^2 \)

provided that \( \text{tr} \left( (U'U)^{-1} \right) > 2t_{\text{max}} \). This means that we can always improve our estimator in MSE(\( x \)) sense and in some cases in MSE(\( \theta \)) sense.

5.3 Principal Component Selection

Let us consider the canonical form as described in (5.5),
\[ y = Z\beta + \epsilon \]  
(5.11)

The columns of \( Z \), \( z_1 \ldots z_d \) are called the principal components and the length of the i-th principal component corresponds with the i-th largest eigenvalue of \( U'U \), thus \( z_i'z_i = t_i \). Since the mean square error of the least squares estimator can be given by,
\[ \text{MSE}(\theta) = \sigma^2 \text{tr}( (U'U)^{-1} ) = \sigma^2 \text{tr}( T^{-1} ) \]  
(5.12)

small eigenvalues in \( T \) have a large contribution in the variance of \( \theta \). Since small eigenvalues mean small lengths of the corresponding columns of \( Z \), the idea is to drop these columns out of the model (5.11). Let us now partition \( Z \) into two parts \( Z_1 \), to be retained, and \( Z_2 \), to be deleted. Deletion of components in \( Z_2 \) means that the parameters \( \beta_2 \) (\( \beta \) is also partitioned into two parts) have implicitly been set to zero. The remaining parameters in \( \beta_1 \) are estimated with the least squares estimator. So,
\[ \hat{\beta}_1 = Z_1^T y \quad \text{and} \quad \hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} z_1^T y \\ 0 \end{bmatrix} \]  
(5.13)
This means that back in the original parameter space,

$$\hat{\theta} = W \hat{\beta} = [ W_1 W_2 ] \hat{\beta} = W_1 Z_1^T y \quad (5.14)$$

This estimator has the properties that, when $W_2^T \hat{\theta}$ is indeed equal to $0$, it is unbiased and has a mean square error of $\sigma^2 \text{tr}(T_1^{-1})$ where $T_1 = Z_1^T Z_1$, which is smaller than the original one. However, when this condition is not true, then over a large range in the parameter space this estimator is inferior to the least squares estimator.

### 5.4 Ridge Regression

As we saw in (5.12), the least squares estimator has a large variance when $U^T U$ has small eigenvalues. As a solution to this problem Hoerl and Kennard (1970) proposed the concept of ridge regression. The ridge estimator is found by solving a slightly modified version of the normal equations. Specifically, the ridge estimator is defined as the solution to,

$$\begin{align*}
(U^T U + k^{*} I_d) \tilde{\theta} &= U^T y \\
\tilde{\theta} &= (U^T U + k^{*} I_d)^{-1} U^T y
\end{align*} \quad (5.15) \text{ or } (5.16)$$

Where $k \geq 0$. The ridge estimator can be seen as a linear transformation of the least squares estimator since,

$$\tilde{\theta} = (U^T U + k^{*} I_d)^{-1} U^T y$$

$$= (U^T U + k^{*} I_d)^{-1} (U^T U) \hat{\theta} = Q \hat{\theta} \quad (5.17)$$

Therefore since $E(\tilde{\theta}) = Q \hat{\theta}$, the ridge estimator is biased. Furthermore it has variance $\sigma^2 Q (U^T U)^{-1} Q = \sigma^2 Q (U^T U)^{-1} Q$. Thus the mean square error can be written as,

$$\text{MSE}(\tilde{\theta}) = \sigma^2 \text{tr}(Q(U^T U)^{-1} Q + k^{*} \hat{\theta}'(U^T U + k^{*} I_d)^{-2} \hat{\theta}) \quad (5.18)$$

$$d$$

$$= \sigma^2 \sum_{i=1}^{d} \frac{t_i}{(t_i + k^{*})^2} + k^{*} \hat{\theta}'(U^T U + k^{*} I_d)^{-2} \hat{\theta} \quad (5.19)$$

Where $t_i$, $i=1...d$, are the eigenvalues of $U^T U$. When using this ridge estimator, the choice of $k$ should be such that the reduction in the variance term is greater than the increase in the squared bias. Hoerl and Kennard proved the existence of a $k > 0$ for which the MSE of $\tilde{\theta}$ is smaller than that of $\hat{\theta}$. A sufficient condition is that,

$$k < 2\sigma^2 / \beta \cdot \text{max} \quad (5.20)$$

Where $\beta \cdot \text{max}$ is the largest element of $\beta = U^T \theta$. (see 5.4). Theobald (1974) showed that for a general weighted squared error, $E((\tilde{\theta} - \theta)' W (\tilde{\theta} - \theta))$ where $W$ is some positive definite matrix, a necessary condition is that,

$$k < 2\sigma^2 / \theta^T \theta \quad (5.21)$$

The existence of a $k$ for which the ridge estimator has smaller variance than the least squares estimator has encouraged many authors to derive some
estimate for $k$, since $\beta_{\max}$ or $\theta$ are of course not known. For example, Hoerl and Kennard ('70) proposed to inspect the ridge trace, i.e., a plot of the elements of $\theta$ versus $k$. However, this is a rather subjective method. Mallows ('73) derived an biased estimate of the risk involved with the ridge estimator. As an estimate for $k$ he suggested to minimize

$$C_k = \text{RSS}_k / 2\sigma^2 - n + 2 + 2\text{tr}(U(U'U + k^2 1_d)^{-1}U')$$

where $\text{RSS}_k$ is the residuals sum of squares as a function of $k$. Hoerl and Baldwin ('75) suggested to take,

$$k = d^* \sigma^2 / \hat{\theta}^T \hat{\theta}$$

Here $k$ is the sample analog of $d^* \sigma^2 / 1_{d}^T \hat{\theta}$, which is the harmonic mean of $k_i = \sigma^2 / \beta_i$ ($i=1..d$). Many other estimators have been proposed. For a good review see Hocking ('76) or Judge et al. ('80).

Since all suggestions for $k$ are based on the actual data, Judge et al. ('80) pointed out that the resulting ridge estimator improves upon $\hat{\theta}$ only for a limited range of the parameterspace, and the region of improvement depends on the unknown parameters $\theta$ and $\sigma^2$. However if we leave the ridge estimators of the form,

$$\hat{\theta} = (I_d + k^2 (U^T U)^{-1})^{-1} \hat{\theta}$$

and consider the more general form,

$$\hat{\theta} = (I_d + k^2 C)^{-1} \hat{\theta}$$

where $C$ is a symmetric, positive definite matrix, Strawderman ('78) derives a class of estimators that are better than the least squares estimators. One of his results is that the estimator,

$$\hat{\theta} = (I_d + a (U^T U)^{-1} (U^T U \hat{\theta} + g U^T \hat{\theta} + h))^{-1} \hat{\theta}$$

with,

$$s = (y - U \hat{\theta})^T (y - U \hat{\theta}) = \text{RSS}_d$$

$$0 \leq a \leq 2(d-2)/(n-d+2) * \text{tmax}$$

$tmax$ is the largest eigenvalue of $\hat{\theta}^T U^T U \hat{\theta}$

$$h \geq 0, g \geq 2d/(n-p+2)$$

is a minimax estimator (this means that it minimizes the maximum risk) under the loss function,

$$L(\theta, a) = \sigma^2 / (1/\sigma^2) (\theta - a)^T \hat{\theta}^{-1}(\theta - a)$$

Thus when we take $B = I_d$, it holds that,

$$\text{MSE}(\hat{\theta}, \theta) \leq \text{MSE}(\hat{\theta}, \theta)$$

when the conditions given in 5.27 are fulfilled.
The linear regression model is often used for impulse response estimation. When the actual length of the impulse response is known, the corresponding least squares estimator is known to be a minimum variance unbiased estimator. However, the actual length of the impulse response is usually not known. Therefore it is more common than rare that the wrong model is used. Although too large a model leads to unbiased estimates one pays with greater variance. Too small a model leads to biased estimates, however, the estimates in this model can be better, according to some performance measure, than those in the correct model if the introduced bias is smaller than the amount of decreasing variance.

Many criteria have been proposed for the selection of an 'optimal' model. Most of the selection procedures have a very different background. There are criteria which are designed on hypothesis arguments (F-test), prediction arguments (Cp, PC, FPE), information theoretic arguments (AIC), Bayesian arguments (BIC, SC) or on cross-validation arguments (PRESS, C1(m) C2(m)). Although some criteria are asymptotically equivalent (Cp, AIC, FPE, PRESS, C1(m)) or choose asymptotically the right model structure (BIC, SC, C2(m)) their small sample properties are all different.

If these criteria are used to select the best out of two model structures (where the two model structures are the full model and some predefined smaller model structure) most of the criteria can be written in terms of an F-test with a particular significance level. A plot of the risk (MSE(0) or MSE(x)) involved with such an F-test yields a range in the parameter space where the risk is smaller than the risk of the least squares estimator of the full model and a range in the parameter space where it is larger. An exception to this rule is the positive Stein rule estimator. This estimator can also be seen as an F-test but in this case the estimates in the full model are shrunken if the hypothesis is rejected. This will assure us that, under certain conditions, the risk involved with this estimator will always be smaller than that of the least squares estimator in the full model.

Since most criteria are designed to improve prediction and thus MSE(x) another criterion has been proposed (Np). However, although it was designed as an unbiased estimate of the MSE(0), it could not prove its superiority, in simulated test cases, to the Cp criterion (in MSE(0) sense), which was designed as an (unbiased) estimate of MSE(x).

Improving the least squares estimator if the true model structure is known means that we have to consider biased estimators. Some biased estimator such as principal component selection and ridge regression only improve the least squares estimator for a limited range in the parameter space. There exists, however, a class of biased estimators, such as the James-Stein estimator and the Strawderman estimator, which are, under certain conditions, uniformly better (in MSE(0) sense) than the least squares estimator.
REFERENCES

Akaike, H.
Fitting autoregressive models for prediction.

Akaike, H.
Information theory and the extension of the maximum likelihood principle.
Ed. by B.W. Petrov and F. Csaki.

Akaike, H.
On entropy maximization principles.
In: Applications of statistics
Ed. by P.R. Krishnaiah
Amsterdam: North-Holland, p.27-41, (1977)

Allen, D.M.
The relationship between variable selection and data augmentation and a method for prediction.

Amemiya, T.
Selection of regressors.

Arnold, S.F.
The theory of linear models and multivariate analysis.

Draper, N.R & Van Nostrand, R.C.
Ridge regression and James-Stein Estimation: Review and Comments.

Hocking, R.R.
The analysis and selection of variables in linear regression.

Hoerl, A.E. & Kennard, R.W.
Ridge regression: Biased estimation of nonorthogonal problems.

Hoerl, A.E. & Baldwin, K.F
Bounds on minimum mean square error in ridge regression.
James, W. & Stein, C.
Estimation with quadratic loss.

Judge, G.G. & Bock, M.E.
The statistical implications of pre-test and Stein-rule estimators in econometrics.
Amsterdam: North-Holland (1978)

Judge, G.G. & Griffiths, WE. & Hill, R.C. & Lee, T.C.
The theory and practice of econometrics.

Mallows, C.L.
Some comments on Cp.
Technometrics, vol. 15 (1973), no. 4, p. 661-675

Miller, A.J.
Selection of subsets of regression variables

Nishii, R.
Asymptotic properties of criteria for selection of variables in multiple regression.

Ponomarenko, M.F.
Information theory and identification.
Dept. of Electrical Engineering, Eindhoven University of Technology, 1981.
EUT Report R1-E-122

Schwarz, G.
Estimating the dimension of a model.

Sclove, S.L.
Improved estimation for coefficients in linear regressions.

Sclove, S.L. & Morris, C. & Radhakrishnan, R.
Non-optimality of preliminary-test Estimators for the mean of a multivariate normal distribution.

Shibata, R.
An optimal selection of regression variable
Söderström, T.
On model structure testing in system identification.
Int. J. Control, vol. 26 (1977), no. 1, o. 1-18

Stein, C.
A necessary and sufficient for admissability.

Model-structure selection by cross-validation

Strawderman, W.E.
Minimax adaptive generalized ridge regression estimators.

Theobold, C.M.
Generalizations of mean square error applied to ridge regression.

Thompson, M.L.
Selection of variables in multiple regression:
Part 1. A review and evaluation.
Part 2. Chosen procedures, computations and examples.
APPENDIX A

This appendix contains some elementary mathematics that have been used in this report.

ALGEBRA

Suppose,

\[ U = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \]

is an \( n \times d \) matrix and has full rank.

\( U_1 = p \times n \) matrix

\( U_2 = (d-p) \times n \) matrix

Then is,

\[ U^T U = \begin{bmatrix} U_1^T U_1 & U_1^T U_2 \\ U_2^T U_1 & U_2^T U_2 \end{bmatrix} \quad (A-1) \]

and

\[ (U^T U)^{-1} = \begin{bmatrix} (U_1^T U_1)^{-1} + (U_1^T U_2) B (U_2^T U_2)^T & -(U_1^T U_2) B \\ -(U_2^T U_2)^T B & B \end{bmatrix} \quad (A-2) \]

with,

\[ B = (U_2^T (I_n - U_1^T U_1) U_2)^{-1} \]

and \( U_1^T = (U_1^T U_1)^{-1} U_1 \)

So with \( U^T = (U^T U)^{-1} U^T \) we get,

\[ U^T = \begin{bmatrix} T_1 \\ T_2 \end{bmatrix} = \begin{bmatrix} U_1^T - U_1^T U_2 T_2 \\ T_2 \end{bmatrix} \quad (A-3) \]

where,

\( T_1 = p \times n \) matrix

\( T_2 = A^+ \) is a \((d-p) \times n\) matrix with \( A = (I_n - U_1 U_1^T) U_2 \)

We can see that,

\[ U U^T = U_1 U_1^T + A A^+ \quad (A-4) \]

Let \( P = U U^T \) then we see that \( P \) is an orthogonal projector since the following conditions hold:

\[ P^2 = P \quad \text{and} \quad P^T = P \]

Let \( P \) be the orthogonal projector on a space spanned by the columns of \( U \), then \( 1-P \) is the orthogonal projector on the null space of the matrix \( U^T \).
POSITIVE (SEMI)DEFINITE

A matrix $H$ is said to be positive definite if

$$\text{for all } x \neq 0, \ x'Hx > 0 \text{ and } H' = H$$

(A-5)

A matrix $H$ is said to be positive semidefinite if,

$$\text{for all } x \neq 0, \ x'Hx \geq 0 \text{ and } H' = H$$

(A-6)

If $H$ is p.s.d then $T'H T$ is also p.s.d

(A-7)

If $H$ is p.d then $T'H T$ is also p.d. provided that for

$T$ holds that $Tx=0 \implies x=0$. Otherwise it is p.s.d.

DERIVATIVES

Let $L$ be a scalar and $\theta$ be a $d$-vector. Then $\delta L/\delta \theta$ and $\delta^2 L/\delta \theta \delta \theta'$ are defined as follows,

$$\delta L / \delta \theta = \begin{bmatrix} \delta L / \delta \theta_1 \\ \vdots \\ \delta L / \delta \theta_d \end{bmatrix}$$

(A-9)

$$\delta^2 L / \delta \theta \delta \theta' = \begin{bmatrix} \delta^2 L / \delta \theta_1 \delta \theta_1 & \ldots & \delta^2 L / \delta \theta_d \delta \theta_1 \\ \vdots & \ddots & \vdots \\ \delta^2 L / \delta \theta_1 \delta \theta_d & \ldots & \delta^2 L / \delta \theta_d \delta \theta_d \end{bmatrix}$$

(A-10)

DISTRIBUTIONS

We say that $x_i$ is normally distributed with mean $\mu_i$ and variance $\sigma_i^2$ if $x_i$ is a continuous random variable with density,

$$p(x_i) = \frac{1}{\sqrt{2\pi\sigma^i}} \exp\left(-\frac{(x_i-\mu_i)^2}{2\sigma^i}\right)$$

(A-11)

and we denote this as,

$$x_i \sim N(\mu_i, \sigma^i)$$

(A-12)

Let $x$ be a random vector with,

$$E(x) = \mu$$

(A-13)
and, 
\[ E( (x-\mu)(x-\mu)' ) = C \]  
(A-14)

Then we say that the vector \( x \) is normally distributed if,
\[ p(x) = (2\pi)^{-n/2} \sigma^n / (\text{det}(C))^{1/2} \exp\left( -\frac{1}{2} (x-\mu)' C^{-1} (x-\mu) \right) \]  
(A-15)

where \( \text{det}(C) \) is the determinant of \( C \). We denote this as,
\[ x \in \mathcal{N}(\mu, C) \]  
(A-16)

Let \( r \) be a linear combination \( x \), thus
\[ r = Qx \]  
(A-17)

Where the matrix \( Q \) denotes the linear transformation. Then \( r \) is normally distributed with mean \( Q\mu \) and covariance matrix \( QCQ' \). Thus,
\[ r \in \mathcal{N}(Q\mu, QCQ') \]  
(A-18)

Let \( x_1, \ldots, x_n \) be independent random variables such that \( x_i \in \mathcal{N}(0, 1) \). Define,
\[ \sum_{i=1}^{n} x_i^2 \]  
(A-19)

then we say that \( y \) is distributed as a Chi\(^2\) distribution with \( n \) degrees of freedom. Thus,
\[ y \in \text{Chi}^2(n) \]  
(A-20)

Let \( z \) be distributed as a Chi\(^2\) distribution with \( m \) degrees of freedom. Then we say that,
\[ \frac{y/n}{z/m} \]  
(A-21)

is distributed as an \( F \) distribution with \( n \) and \( m \) degrees of freedom and we denote this as,
\[ v \in \mathcal{F}(n, m) \]  
(A-22)
This appendix contains the calculation of the risk matrix and some related topics.

Let us now assume that the following relation holds,

\[ y = U\theta + e = U_1\theta_1 + U_2\theta_2 + e \quad e \sim \mathcal{N}(0, \sigma^2 I_n) \]

where \( U \) has full rank. Then with,

\[
\hat{\theta}_p = \begin{bmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{bmatrix} = \begin{bmatrix} u_1^\top y \\ 0 \end{bmatrix} = \begin{bmatrix} \theta_1 + U_1^\top U_2 \theta_2 + U_1^\top e \\ 0 \end{bmatrix}
\]

we get that \( E\{ \hat{\theta}_p - \theta \} (\hat{\theta}_p - \theta)^\top \) is equal to

\[
\begin{bmatrix}
\sigma^2 (U_1^\top U_1)^{-1} + (U_1^\top U_2 \theta_2) (U_1^\top U_2 \theta_2)^\top \\
(U_1^\top U_2 \theta_2) (U_1^\top U_2 \theta_2)^\top \\
(U_2 \theta_2)^\top (U_2 \theta_2)
\end{bmatrix}
\]

Then with \( \hat{\theta} = U^\top y \) we get that \( E\{ \hat{\theta} - \theta \} (\hat{\theta} - \theta)^\top = \sigma^2 (U^\top U)^{-1} \). So with relation (A-2) we see that,

\[
E\{ (\hat{\theta} - \theta)^\top (\hat{\theta} - \theta) \} - E\{ (\hat{\theta}_p - \theta) (\hat{\theta}_p - \theta)^\top \} =
\begin{bmatrix}
-U_1^\top U_2 (B \sigma^2 \theta_2 \theta_2^\top) (U_1^\top U_2)^\top \\
(U_1^\top U_2 (B \sigma^2 \theta_2 \theta_2^\top) (U_1^\top U_2)^\top \\
-(B \sigma^2 \theta_2 \theta_2^\top) (U_1^\top U_2)^\top \\
(B \sigma^2 \theta_2 \theta_2^\top)
\end{bmatrix}
\]

If we rewrite (B-2) as,

\[
\begin{bmatrix}
-U_1^\top U_2 \\
[U_1^\top U_2 (B \sigma^2 \theta_2 \theta_2^\top) (U_1^\top U_2)^\top] \\
1(d-p)
\end{bmatrix}
\]

then we see that if \( (B \sigma^2 \theta_2 \theta_2^\top) \) is p.d., the matrix given in (B-2) is at least p.s.d. (see A-8). A necessary condition for \( (B \sigma^2 \theta_2 \theta_2^\top) \) to be p.d is that \( B \) must be p.d. Let us rewrite \( B \) as \( B=WD\theta_2 \theta_2^\top W^\top \) then must,

\[
1 - \frac{1}{\sigma^2} \theta_2 \theta_2^\top W^\top W \theta_2 \theta_2^\top W^\top \leq 1
\]

be p.d.. The second term has rank 1, and the only nonzero eigenvalue has value \( \theta_2 \sigma^2 \). Thus because of condition (A-6) the following relation holds,

\[
(B \sigma^2 \theta_2 \theta_2^\top) \text{ is p.d. } \iff \theta_2 \sigma^2 \theta_2^\top \sigma^2 < 1
\]
APPENDIX C

PROVE THAT,

\[ 2*1(\Theta, \bar{\Theta}) = (\bar{x}-x)'(\bar{x}-x) / \sigma^2 \]

Proof:

Let,

\[ I(\Theta, \bar{\Theta}) = - \int p(y; \Theta) \ln \left( \frac{p(y; \Theta)}{p(y; \bar{\Theta})} \right) dy \]

with,

\[ p(y; \Theta) = (2\pi \sigma^2)^{-n/2} \exp \left( -(y-U\Theta)'(y-U\Theta)/2\sigma^2 \right) \]

Then we have,

\[ \delta \ln(p(y; \Theta)) / \delta \Theta = -(U'U)/2\sigma^2 \]

\[ \delta' \ln(p(y; \Theta)) / \delta \Theta = -(U'U)/\sigma^2 \]

and all higher derivatives are zero. Let \( \Theta_m \) be the value of \( \Theta \) that minimizes \( I(\Theta, \bar{\Theta}) \) (thus \( \Theta_m=\bar{\Theta} \)). Then, because of the Taylor expansions series and \( (C-1)-(C-4) \), we can write,

\[ I(\Theta, \bar{\Theta}) = I(\Theta, \Theta_m) + (\bar{\Theta} - \Theta_m)' \delta I(\Theta, \Theta_m) / \delta \Theta_m + 1/2*(\bar{\Theta} - \Theta_m)' \delta^2 I(\Theta, \Theta_m) / \delta \Theta_m \delta \Theta_m' * (\bar{\Theta} - \Theta_m) \]

But since,

\[ \delta^2 I(\Theta, \Theta_m) / \delta \Theta_m \delta \Theta_m' = \]

\[ \delta' - \int p(y; \Theta) \ln \left( \frac{p(y; \Theta)}{p(y; \bar{\Theta})} \right) dy / \delta \Theta_m \delta \Theta_m' = \]

\[ \int p(y; \Theta) U'U/\sigma^2 dy = U'U/\sigma^2 \]

we have that,

\[ 2*1(\Theta, \bar{\Theta}) = (\bar{\Theta} - \Theta)'U(\bar{\Theta} - \Theta)/\sigma^2 = (\bar{x}-x)'(\bar{x}-x)/\sigma^2 \]

Q.E.D.
APPENDIX D

PROVE THAT,

\[ \text{PRESS}(p) = \sum_{i=1}^{n} (y(i) - u1(i)^{\prime}U_1(i)^{\prime}y(i))^{\prime} = w_p D_p^{-2} w_p \]

with,

- \( U_1(i) \) = the \( i \)-th row of \( U_1 \)
- \( u1(i) \) = the \( i \)-th row of \( u1 \)
- \( y(i) \) = output vector without the \( i \)-th output
- \( y(i) \) = \( i \)-th output

\[ w_p = (I_n - U_1 U_1^{\prime}) y \]
\[ D_p = \text{the diagonal matrix whose diagonal elements are those of } (I_n - U_1 U_1^{\prime}) \]

Proof:

Let us consider the \( i \)-th term of \( \text{PRESS} \) and let us write,

\[ A = [U_1(i)^{\prime} u1(i)^{\prime}]^{\prime} = [X^{\prime} b^{\prime}]^{\prime} \]

where \( A \) is just another ordering of the rows of \( U \). Then we can write this \( i \)-th term as,

\[ \text{PRESS}(p,i) = [y(i) - bx'y(i)] \]

Let us now consider the \( i \)-th term of \( w_p \). We can write,

\[ w_p(i) = y(i) - bA^{\prime} \tilde{y} \]

where \( \tilde{y} = (y(i)^{\prime} y(i))^{\prime} \) is just another ordering of \( y \). Since,

\[ (A'^{A})^{-1} = (X'X)^{-1} - (X'X)^{-1}b'(I+b(X'X)^{-1}b')^{-1}b(X'X)^{-1} \]

\[ = M \]

we can write,

\[ A^{\prime} = M [X^{\prime} b^{\prime}] \]

so,

\[ bA^{\prime} \tilde{y} = bMX'y(i) + bMb'y(i) \]

and,

\[ AA^{\prime} = [X^{\prime} b^{\prime}]M [X^{\prime} b^{\prime}] \]
\[ (UU^{\prime})_{jj} = (AA^{\prime})_{nn} = bMb' \]

Thus the \( i \)-th component of \( w_p D_p^{-2} w_p \) can be written as,
\[
\frac{y(i) - bA^*y}{(1-bMb')} = \frac{y(i) - bMb'y(i) - bMX'y(i)}{(1-bMb')}
\]

But since (see D-1),

\[
bMX'y(i) = [b^V - b^Vb'(1+b'^Vb)^{-1}b^V]X'y(i)
\]

With \(V=(X'X)^{-1}\). Because \((1+b'^Vb)\) is a scalar we can see that if we multiply the last two expressions with it,

\[
bMX'y(i)*(1+b'^Vb) = b^VX'y(i)
\]

\[
(1-bMb')*(1+b'^Vb) = 1
\]

So,

\[
\frac{bMX'y(i)}{(1-bMb')} = bX^*y(i)
\]

But this means the i-th component of \(w_pD_p^{-2}w_p\) is equal to the PRESS(p,i). Thus PRESS(p) and \(w_pD_p^{-2}w_p\) are equal.

Q.E.D.
(171) Monnee, P., and M.H.A.J. Herben
MULTIPLE-BEAM GROUNDSTATION REFLECTOR ANTENNA SYSTEM: A preliminary study.

(172) Baartmans, M.J., and A.H.M. Akkermans
ERROR REDUCTION IN TWO-DIMENSIONAL PULSE-AREA MODULATION, WITH APPLICATION TO COMPUTER-GENERATED TRANSPARENTS.

(173) Zhu Yu-Cai
ON A BOUND OF THE MODELLING ERRORS OF BLACK-BOX TRANSFER FUNCTION ESTIMATES.

(174) Berkelaar, M.R.C.M., and J.F.M. Thenuwen
TECHNOLOGY MAPPING FROM BOOLEAN EXPRESSIONS TO STANDARD CELLS.

(175) Janssen, P.H.M.
FURTHER RESULTS ON THE McMillan Degree and the Kronecker Indices of ARMA Models.

(176) Janssen, P.H.M., and P. Stoica, T. Söderström, P. Eykhoff
MODEL STRUCTURE SELECTION FOR MULTIVARIABLE SYSTEMS BY CROSS-VALIDATION METHODS.

(177) Zarkova
ARCS IN CESIUM SEEDED NOBLE GASES RESULTING FROM A MAGNETICALLY INDUCED ELECTRIC FIELD.

(178) Zhu Yu-Cai
THE BOUNDS OF THE MODELLING ERRORS OF BLACK-BOX MIMO TRANSFER FUNCTION ESTIMATES.

(179) Lieshout, C.J.P. van, and L.P.P.P. van Ginneken
GM: A gate matrix layout generator.

(180) Zarkova
GRIDLESS ROUTING FOR GENERALIZED CELL ASSEMBLIES: Report and user manual.

(181) Bolten, M.H.J., and P.T.M. Vaessen
FREQUENCY SPECTRA FOR ADMITTANCE AND VOLTAGE TRANSFERS MEASURED ON A THREE-PHASE POWER TRANSFORMER.

(182) Zhu Yu-Cai
BLACK-BOX IDENTIFICATION OF MIMO TRANSFER FUNCTIONS: Asymptotic properties of prediction error models.

(183) Zhu Yu-Cai
THE INFLUENCE OF CONTACT SURFACE MICROSTRUCTURE ON VACUUM ARC STABILITY AND ARC VOLTAGE.

(184) Kadete, H.
ENHANCEMENT OF HEAT TRANSFER BY CORONA WIND.

(185) Hermans, P.A.M., and A.M.J. Kwaks, I.V. Bruza, J. Dijk
THE IMPACT OF TELECOMMUNICATION ON RURAL AREAS IN DEVELOPING COUNTRIES.

(186) Xu Yanhong
THE INFLUENCE OF CONTACT SURFACE MICROSTRUCTURE ON VACUUM ARC STABILITY AND ARC VOLTAGE.

(187) Kalse, F., and L. Stok, R. van den Born
DESIGN AND IMPLEMENTATION OF A MODULE LIBRARY TO SUPPORT THE STRUCTURAL SYNTHESIS.
(188) Jóźwiak, J.
THE FULL DECOMPOSITION OF SEQUENTIAL MACHINES WITH THE STATE AND OUTPUT
BEHAVIOUR REALIZATION.

(189) Pineda de Cuyz, J.
ALWAYS: A system for wafer yield analysis.

(190) Siuzdak, J.
OPTICAL COUPLERS FOR COHERENT OPTICAL PHASE DIVERSITY SYSTEMS.

(191) Bastiaansen, M.J.
LOCAL-FREQUENCY DESCRIPTION OF OPTICAL SIGNALS AND SYSTEMS.

(192) Worm, S.C.J.
A MULTI-FREQUENCY ANTENNA SYSTEM FOR PROPAGATION EXPERIMENTS WITH THE
OLYMPUS SATELLITE.

(193) Kersten, W.F.J. and G.A.P. Jacobs
ANALOG AND DIGITAL SIMULATION OF LINE-ENERGIZING OVERVOLTAGES AND COMPARISON
WITH MEASUREMENTS IN A 400 KV NETWORK.

(194) Hosselet, L.M.L.F.
MARTINUS VAN MARUM: A Dutch scientist in a revolutionary time.

(195) Bondarev, V.N.
ON SYSTEM IDENTIFICATION USING PULSE-FREQUENCY MODULATED SIGNALS.

(196) Liu Wen-Jiang, Zhu Yu-Cai and Cai Da-Wei
MODEL BUILDING FOR AN INGOT HEATING PROCESS: Physical modelling approach and
identification approach.