The linear regression model : model structure selection and biased estimators

tu/e
Research Report
ISSN 0167-9708
Coden: TEUEDE

Eindhoven
University of Technology
Netherlands

Faculty of Electrical Engineering

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

Eindhoven University of Technology Research Reports
EINDHOVEN UNIVERSITY OF TECHNOLOGY

Faculty of Electrical Engineering
Eindhoven The Netherlands

ISSN 0167- 9708 Coden: TEUEDE

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
SISO 656 UDC 519.71.001.3 NUGI 832
Trefw.: systeemidentificatie.

SUMMARY

The linear regression model is one of the most frequently used model 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 repsonse. However, similations 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 uniformely better than the least squares estimator.

Delissen, J.G.M.
THE LINEAR REGRESSION MODEL: Model structure selection and biased estimators.
Faculty of Electrical Engineering, Eindhoven University of Technology, The Netherlands, 1988.
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

OPERATORS

$\arg\min_{z} f(z)$:value of z that minimizes f(z)

cov(z)	:covariance matrix of random vector z
plim	:probability limit
E( )	:expectation
tr(A)	:trace (sum of diagonal elements) of the matrix A
A'	:transpose of matrix A
A^{-1}	:inverse of matrix A
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_c	: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( )	:Loss function
LH( )	:Likelihood function
M	:model structure
MSE( )	:Mean Square Error function
n	:number of output samples
p	:dimension of the current model structure
p( )	:(joint) probability density function
p(|)	:conditional (joint) probability density function
RM( )	:Risk Matrix function
Sf	: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
β	:transformed θ, used in canonical forms
θ	:vector of process parameters
ϕ	:vector used to parametrize models
σ^2	:variance of process disturbances
$\hat{}$, $\tilde{}$	: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:

$$y = U\theta + w \tag{1.1}$$

where,

- y: n-vector of observable random variables
- U: known matrix of dimension n*d
- θ: 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 usefull when there is no prior information about a relationship between the markov parameters. Let the input-output behaviour of a process be given by :

$$y(k) = \sum_{i=0}^{d_c-1} h(i)u(k-i) + e(k) \tag{1.2}$$

here,

- 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),..,y(n),u(2-p),..,u(n)) of measured input-output samples, we can write the input-output relation of the process in the following form :

$$\begin{bmatrix} y(1) \\ \vdots \\ y(n) \end{bmatrix} = \begin{bmatrix} u(1) & \cdots & u(2-p) \\ \vdots & & \vdots \\ u(n-p+1) & \cdots & u(n-p+1) \end{bmatrix} \begin{bmatrix} h(0) \\ \vdots \\ h(p-1) \end{bmatrix} + \begin{bmatrix} e(1) \\ \vdots \\ e(n) \end{bmatrix} \tag{1.3}$$

$$y = U\Theta + e$$

$$y = x + e$$

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 Θ 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 σ^2 and are normally distributed (for some definitions see appendix A). Thus,

$$e_i \in N(0,\sigma^2) \tag{1.4}$$

so,

$$e \in N(\underline{0},\sigma^2 I_n) \tag{1.5}$$

$$y \in N(U\theta,\sigma^2 I_n) \tag{1.6}$$

Here I_n denotes an identity matrix of dimension n*n.

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

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

When the length of the actual impulse response is known this estimator is a minimum variance unbiased estimator of θ. 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 MODELLING

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 θ that minimizes the residual sum of squares (RSS):

$$RSS = (y-\hat{x})'(y-\hat{x}) \tag{2.1}$$

Here $\hat{x}$ denotes the estimated undisturbed output x ($\hat{x}=U\hat{\theta}$). Thus given the model $y=U\phi+w$, where ϕ 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:

$$\delta RSS / \delta \phi = 2(U'U)\phi - 2U'y = \underline{0} \tag{2.2}$$

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

$$\hat{\theta} = (U'U)^{-1}U'y = U^{+}y \tag{2.3}$$

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

2.2 PERFORMANCE MEASURES

If one has some estimate of θ, 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 $\hat{\theta}$ and θ be denoted as $L(\theta)$.
Thus:

$$L(\theta) = (\hat{\theta}-\theta)'(\hat{\theta}-\theta) \tag{2.4}$$

Then the corresponding mean square error function is,

$$MSE(\theta)=E(L(\theta)) \tag{2.5}$$

Rewrite MSE(θ) as,

$$MSE(\theta) = (E\{\hat{\theta}\}-\theta)'(E\{\hat{\theta}\}-\theta) + E\{ (\hat{\theta}-E\{\hat{\theta}\})'(\hat{\theta}-E\{\hat{\theta}\}) \} \quad (2.6)$$

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 MSE function will be defined as follows.

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

$$L(x) = (\hat{x}-x)'(\hat{x}-x) \quad (2.7)$$

Then the corresponding mean square error function is,

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

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 θ.

$$RM(\theta) = E\{ (\hat{\theta}-\theta)(\hat{\theta}-\theta)' \} \quad (2.9)$$

Then it can be seen that,

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

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

Suppose that there are two estimators $\hat{\theta}_a$ and $\hat{\theta}_b$ of θ. 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,

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

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

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

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 supcript 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,

$$\begin{aligned} &\text{process} \quad y = U_f\theta_f + e = U\theta + U_r\theta_r + e \text{ with } \theta_r = \underline{0} \\ &\text{model} \quad\;\; y = U_f\hat{\theta}_f + w = U\hat{\theta} + U_r\hat{\theta}_r + w \end{aligned} \tag{2.15}$$

here,

$U_f = (U\ U_r)$ is an $n*p$ matrix
U = $n*d_c$ matrix
U_r = $n*(p-d_c)$ matrix

$\theta_f = (\theta'\ \theta_r')'$ and $\hat{\theta}_f = (\hat{\theta}'\ \hat{\theta}_r')'$ are p-vectors
θ and $\hat{\theta}$ are d_c-vectors
$\theta_r = \underline{0}$ and $\hat{\theta}_r$ are $(p-d_c)$-vectors

Since θ_f is not known we leave it in the model as a variable ($\hat{\theta}_f$). 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 θ.

Because,

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

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

$$p(y;\theta,\sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\{ -(y-U\theta)'(y-U\theta)/2\sigma^2 \} \tag{2.16}$$

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

$$LH(\hat{\theta}|y) = p(y;\hat{\theta}) \tag{2.17}$$

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

$$\hat{\theta}_{ml} = \arg\max_{\hat{\theta}} LH(\hat{\theta}|y) = (U'U)^{-1}U'y = U^{+}y \tag{2.18}$$

Thus we see that the MLE of θ is equal to the least squares estimator of θ (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, $E\{ \hat{\theta} \} = \theta$ (2.19)

b: $cov(\hat{\theta}) = E\{ (\hat{\theta}-E\{ \hat{\theta} \})(\hat{\theta}-E\{ \hat{\theta} \})' \} = \sigma^2(U'U)^{-1}$ (2.20)

c: minimum variance unbiased estimator (see Arnold '81) (2.21)

d: $\hat{\theta} \in N(\theta,\sigma^2(U'U)^{-1})$ because $\hat{\theta}$ is a linear combination of y and (a) and (b) hold. (2.22)

Thus our performance measures become :

e: $MSE(\theta) = \sigma^2 tr\{ (U'U)^{-1} \}$ (2.23)

f: $MSE(x) = \sigma^2 tr\{ U(U'U)^{-1}U' \} = \sigma^2 d_c$ (2.24)

g: $RM(\theta) = \sigma^2(U'U)^{-1}$ (2.25)

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

$$LH(\Phi,\Omega^2|y)=p(y;\Phi,\Omega^2) \tag{2.26}$$

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

$$\hat{\theta},\tilde{\sigma}^2 = \arg \max_{\Phi,\Omega} LH(\Phi,\Omega^2|y) \tag{2.27}$$

This leads to the optimizing estimators:

$$\hat{\theta} = U^+y \quad \text{and} \quad \tilde{\sigma}^2 = (y-\hat{x})'(y-\hat{x})/n = \hat{e}'\hat{e}/n \tag{2.28}$$

The with the estimator $\hat{\theta}$ corresponding residuals will be denoted as $\hat{e}$. Thus the estimate of θ remains the same. The estimate of σ^2 can be written as,

$$\tilde{\sigma}^2 = \|(I_n-UU^+)y\|^2/n \tag{2.29}$$

Here $\|x\|^2$ is defined as x'x. Then we see that $\tilde{\sigma}^2$ is a biased estimate of σ^2 because,

$$\begin{aligned} E\{ \tilde{\sigma}^2 \} &= E\{ \|y\|^2 - \|UU^+y\|^2 \}/n \\ &= E\{ \|y\|^2 \}/n - E\{ \|UU^+y\|^2 \}/n \\ &= \sigma^2(n-d_c)/n \end{aligned} \tag{2.30}$$

Consider now the following estimate of σ^2,

$$\hat{\sigma}^2 = \tilde{\sigma}^2 n/(n-d_c) = \|(I_n-UU^+)y\|^2/(n-d_c) \tag{2.31}$$

This estimator has the following properties,

a: unbiased (see (2.30) and (2.31)) (2.32)

b: minimum variance unbiased estimator (see Arnold '81) (2.33)

c: $\hat{\sigma}^2$ and $\hat{\theta}$ are independent, because (2.34)

$\hat{e}=(I_n-UU^+)y$ and $\hat{\theta}=U^+y$ are normally distributed

with covariance $E\{ (w-E(w))(\hat{\theta}-E(\hat{\theta}))' \} = 0_{n,dc}$

Let $P=(I_n-UU^+)$. Because P is a 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 e'Pe can be written as e'V'TVe, where V is an n*n orthonormal matrix and T is an n*n diagonal matrix whose diagonal elements are the eigenvalues of P. If we rewrite e'V'TVe as r'Tr, where $r \in N(0,\sigma^2 I_n)$, it can be seen that $(n-d_c)\hat{\sigma}^2/\sigma^2$ is distributed as a Chi^2-distribution (for some distributions see appendix A) with $(n-d_c)$ degrees of freedom, thus

$$(n-d_c)\hat{\sigma}^2/\sigma^2 \in Chi^2(n-d_c) \tag{2.35}$$

CASE { $p > d_c$ }:

Sofar 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 $\theta_f = (\theta' \; \theta_r')'$ becomes,

$$\hat{\theta}_f = (U_f'U_f)^{-1}U_f'y \tag{2.36}$$

This estimator has the following properties,

PROPERTIES

a: unbiased, $E\{ \hat{\theta}_f \} = \theta_f$ (2.37)

b: $cov(\hat{\theta}_f) = \sigma^2(U_f'U_f)^{-1}$ (2.38)

c: $\hat{\theta}_f \in N(\theta_f,\sigma^2(U_f'U_f)^{-1})$,because of (a) & (b) (2.39)

d: $MSE(\theta_f) = \sigma^2 \; tr\{ (U_f'U_f)^{-1} \}$ (2.40)

e: $MSE(x) = \sigma^2 p$ (2.41)

f: $RM(\theta_f) = \sigma^2(U_f'U_f)^{-1}$ (2.42)

However, this estimator is no longer a minimum variance estimator, because if we compare the covariance matrix of the estimator

$$\tilde{\theta}_f = ((U^+y)' \; 0')' \tag{2.43}$$

given by,

$$\begin{bmatrix} \sigma^2(U'U)^{-1} & 0_{d_c,p-d_c} \\ 0_{p-d_c,d_c} & 0_{p-d_c,p-d_c} \end{bmatrix} \qquad (2.44)$$

(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,

$$\mathrm{cov}(\hat{\theta}_f)-\mathrm{cov}(\tilde{\theta}_f) = \sigma^2 \begin{bmatrix} (U^+U_r)B(U^+U_r)' & -(U^+U_r)B \\ -B(U^+U_r)' & B \end{bmatrix} \qquad (2.45)$$

with $B=(U_r'(I_n-UU^+)U_r)^{-1}$ (see appendix B)

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

The estimator of σ^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,

$$\begin{aligned} &\text{process } y= U_1\theta_1+U_2\theta_2+e = U\theta+e \qquad (2.46)\\ &\text{model } \quad y= U_1\phi_1+w = U_1\phi_1+U_2\phi_2+w = U\phi+w \text{ so } \phi_2=\underline{0} \end{aligned}$$

Here we assume that:

$U=(U_1\ U_2)$ is an $n*d_c$ matrix
U_1 is an $n*p$ matrix
U_2 is an $n*(d_c-p)$ matrix

$\theta=(\theta_1'\ \theta_2')'$ and $\phi=(\phi_1'\ \phi_2')'$ are d-vectors
θ_1 and ϕ_1 are p-vectors
θ_2 and $\phi_2=\underline{0}$ 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 θ_1, since we are investigating the consequences of incorrect modelling when using this estimator, thus,

$$\hat{\theta}_1 = (U_1'U_1)^{-1}U_1'y = U_1^+y \ ; \ \hat{\theta}_2 = \underline{0} \qquad (2.47)$$

so,

$$\hat{\theta} = (\hat{\theta}_1' \; 0')' \tag{2.48}$$

Then the following properties can be derived:

PROPERTIES

a: biased, $E(\hat{\theta}_1) = \theta_1 + U_1^+ U_2 \theta_2 \quad E(\hat{\theta}_2) = \underline{0}$ (2.49)

b: $\text{cov}(\hat{\theta}_1) = \sigma^2 (U_1' U_1)^{-1}$ (2.50)

c: $\hat{\theta}_1 \in N(\theta_1 + U_1^+ U_2 \theta_2, \; \sigma^2 (U_1' U_1)^{-1})$ (2.51)

Thus our performance measures become now,

d: $MSE(\theta) = |\theta_2|^2 + |U_1^+ U_2 \theta_2|^2 + \sigma^2 tr((U_1' U_1)^{-1})$ (2.52)

e: $MSE(x) = |(I_n - U_1 U_1^+) U_2 \theta_2|^2 + \sigma^2 p$ (2.53)

f:

$$RM(\theta) = \begin{bmatrix} \sigma^2 (U_1' U_1)^{-1} + (U_1^+ U_2 \theta_2)(U_1^+ U_2 \theta_2)' & -U_1^+ U_2 \theta_2 \theta_2' \\ -\theta_2 (U_1^+ U_2 \theta_2)' & \theta_2 \theta_2' \end{bmatrix} \tag{2.54}$$

(see appendix B)

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

$$\hat{\sigma}^2 = |(I_n - U_1 U_1^+) y|^2 / (n-p) \tag{2.55}$$

But because,

$$E(\hat{\sigma}^2) = |(I_n - U_1 U_1^+) U_2 \theta_2|^2 / (n-p) + \sigma^2 \tag{2.56}$$

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 θ in the too small model ($\hat{\theta}$) with the estimate of θ in the correct model ($\bar{\theta} = U^{+}y$) we can see that the difference between the risk matrices of both models is given by,

$$RM(\theta,\bar{\theta}) - RM(\theta,\hat{\theta}) = \begin{bmatrix} U_1^{+}U_2\{\sigma^2 B-\theta_2\theta_2'\}(U_1^{+}U_2)' & -U_1^{+}U_2\{\sigma^2 B-\theta_2\theta_2'\} \\ -\{\sigma^2 B-\theta_2\theta_2'\}(U_1^{+}U_2)' & \{\sigma^2 B-\theta_2\theta_2'\} \end{bmatrix} \qquad (2.57)$$

$$\text{with } B=\{U_2'(I_n-U_1U_1^{+})U_2\}^{-1} \quad \text{(see appendix B)}$$

Here we see that when,

$$\{\sigma^2 B-\theta_2\theta_2'\} \text{ is p.d.,} \qquad (2.58)$$

thus when the covariance matrix of estimating θ_2 in the correct model minus the bias matrix of θ_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,

$$\{\sigma^2 B-\theta_2\theta_2'\} \text{ is p.d. } \Longleftrightarrow \quad \theta_2' B^{-1}\theta_2 / \sigma^2 < 1 \qquad (2.59)$$

(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,

$$MSE(\theta,\hat{\theta}) - MSE(\theta,\bar{\theta}) =$$

$$|\theta_2|^2 + |U_1^{+}U_2\theta_2|^2 + \sigma^2(\text{tr}((U_1'U_1)^{-1}) - \text{tr}((U'U)^{-1})) < 0 \qquad (2.60)$$

or,

$$MSE(x,\hat{x}) - MSE(x,\bar{x}) =$$

$$|(I_n-U_1U_1^{+})U_2\theta_2|^2 + \sigma^2(p-d_c) \quad < \quad 0 \qquad (2.61)$$

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 ammount 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:

$$MSE(\theta) : \quad |\theta_2|^2 + |U_1^{+}U_2\theta_2|^2 + \sigma^2 tr\{ (U_1'U_1)^{-1} \} \tag{2.62}$$

$$MSE(x) : \quad |(I_n - U_1U_1^{+})U_2\theta_2|^2 + \sigma^2 p \tag{2.63}$$

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 dmax be the dimension of the model with,

$$RM(\theta,\hat{\theta}_p) - RM(\theta,\hat{\theta}_{dmax}) \quad \text{is p.s.d. and non zero for } p > dmax$$

Let dmin be the dimension of the model with,

$$RM(\theta,\hat{\theta}_p) - RM(\theta,\hat{\theta}_{dmin}) \quad \text{is p.s.d. and non zero for } p < dmin$$

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

Since we used the least square estimator to estimate θ_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,

$$\text{process:} \quad y=U\theta+e=U_1\theta_1+U_2\theta_2+e \quad \text{where elements of } \theta_2 \text{ may be zero} \tag{3.1}$$

$$\text{model:} \quad y=U_1\theta_1+w \quad \text{with} \quad \hat{\theta}_1 = U_1^+y \text{ and } \hat{e} =(I_n-U_1U_1^+)y$$

where

$U=(U_1\ U_2)$ is an n*d matrix
U_1 is an n*p matrix
U_2 is an n*(d-p) matix

$\theta=(\theta_1'\ \theta_2')'$ is a d-vector
θ_1 is a p-vector
θ_2 is a (d-p) vector

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

$$RSSp = \hat{e}'\hat{e} = (y-\hat{x})'(y-\hat{x}) = \|(I_n-U_1U_1^+)y\|^2 \tag{3.2}$$

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

$$\text{a:} \quad \hat{\sigma}^2(p) = RSSp/(n-p) \tag{3.3}$$

$$\text{b:} \quad \tilde{\sigma}^2(p) = RSSp/n \tag{3.4}$$

$$\text{c:} \quad \hat{\sigma}^2(d) = RSSd/(n-d) \tag{3.5}$$

$$\text{d:} \quad \tilde{\sigma}^2(d) = RSSd/n \tag{3.6}$$

All these estimators and $\hat{\theta}_1$ are independent (see 2.34). The estimators (b) and (d) are the MLE estimates of σ^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 σ^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 decribed 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,

$$R_p^2 = 1 - RSSp/(y'y) \tag{3.7}$$

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. Unfortunally 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 (RMSp) versus p, where

$$RMSp = RSSp/(n-p) = \hat{\sigma}^2(p) \tag{3.8}$$

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

a: minimum RMSp
b: the value of p such that RMSp ≈ RSSd
c: the value of p where the RMSp 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.

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

$$\hat{\theta} \in N(\theta,\sigma^2(U'U)^{-1})$$

$$(n-d)\hat{\sigma}^2(d)/\sigma^2 \in Chi^2(n-d) \qquad (3.9a)$$

Thus we see that,

$$(\hat{\theta}-\theta)'(U'U)(\hat{\theta}-\theta)/\sigma^2 \in Chi^2(d) \qquad (3.9b)$$

because σ^2 is unknown and we estimate it with RSSd/(n-d). Since RSSd/(n-d) and $\hat{\theta}$ are independent, we get from the ratio of 3.9a and 3.9b,

$$(\hat{\theta}-\theta)'(U'U)(\hat{\theta}-\theta)/d\hat{\sigma}^2(d) \in F(d,n-d) \qquad (3.10)$$

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,

$$(\hat{\theta}_2-\theta_2)'(U_2'U_2)(\hat{\theta}_2-\theta_2)/((d-p)\hat{\sigma}^2(d)) \in F(d-p,n-d) \qquad (3.11)$$

It is now possible to derive a 100(1-α)% joint confidence region for θ_2. This region is defined as the region in which θ_2 lies with probability 100(1-α)%. Because of (3.11) we see that a 100(1-α)% confidence region of θ_2 ($R(\theta_2)$) can be defined as,

$$R(\theta_2) = \{\, \theta_2 \mid (\hat{\theta}_2-\theta_2)'(U_2'U_2)(\hat{\theta}_2-\theta_2)/((d-p)\hat{\sigma}^2(d)) < F^{\alpha}(d-p,n-d) \,\} \qquad (3.12)$$

here $F^{\alpha}(d-p,n-d)$ is the solution of

$$\left[\begin{array}{l} P(\, x < F^{\alpha}(d-p,n-d)\,) = 1-\alpha \\ x \in F(d-p,n-d) \end{array}\right. \qquad (3.13)$$

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 θ_2=0, so that a smaller model structure can be used to

estimate θ. Thus we have the following hypothesis,

$$H0 \qquad \theta_2=0$$

$$H1 \qquad \theta_2<>0$$

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,

$$\begin{aligned} RSSp-RSSd &= |(I_n-U_1U_1^+)y|^2 - |(I_n-UU^+)y|^2 \\ &= |UU^+y|^2 - |U_1U_1^+y|^2 \\ &= |AA^+y|^2 \end{aligned} \tag{3.14}$$

with $A=(I_n-U_1U_1^+)U_2$ (see appendix A)

that $(RSSp-RSSd)/\sigma^2$ is distributed as,

$$(RSSp-RSSd)/\sigma^2 \in Chi^2(d-p) \tag{3.15}$$

And because RSSp-RSSd and RSSd are independent we see that,

$$(RSSp-RSSd)/((d-p)\hat{\sigma}^2(d)) \in F\ (d-p,n-d) \tag{3.16}$$

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 θ if,

$$(RSSp-RSSd)/((d-p)\hat{\sigma}^2(d)) < F^{\alpha}\ (d-p,n-d)$$

or equivalently,

$$|AA^+y|^2 \ /\ ((d-p)\hat{\sigma}^2(d) \qquad < F^{\alpha}\ (d-p,n-d) \tag{3.17}$$

This test is called a partial F-test, because it measures the contribution of regressor θ_2 given that the other regressor θ_1 is in the model set.

In the previous section we defined the joint confidence region of θ_2 when we estimated θ in the full model. Testing if θ_2 could be set to zero by looking if the null-vecor belongs to the joint confidence region of θ_2 can reduce the number of parameters, however, θ_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),

$$\theta_2 \in R(\theta_2) \text{ if } \quad \hat{\theta}_2'U_2'U_2\hat{\theta}_2 \ /\ ((d-p)\hat{\sigma}^2(d)) \qquad < F^{\alpha}\ (d-p,n-d) \tag{3.18}$$

and thus if,

$$(\ |AA^+y|^2 + |U_1U_1^+U_2A^+y|^2\ )\ /\ ((d-p)\hat{\sigma}^2(d)) \quad < F^{\alpha}\ d-p,n-d \tag{3.19}$$

because,

$$\{ \|U_2A^+y\|^2 \} = \{ \|(I_n-U_1U_1^+)U_2A^+y\|^2 + \|U_1U_1^+U_2A^+y\|^2 \} \qquad (3.20)$$

This means that when $\theta_2=0$ belongs to the joint confidence region of θ_2 the condition (3.17) holds and thus H0 is accepted. But the reverse is not true. If H0 is accepted according to condition (3.17) it does not have to mean that θ_2 belongs to this joint confidence region. This is only true is when 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 θ_2 for which the partial F-test gives satisfactory results.

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

3.4 PREDICTION ARGUMENTS

3.4.1 MALLOWS' Cp

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).

$$MSEP(y_p)= E\{ (y_p-\hat{x}_p)'(y_p-\hat{x}_p) \} \qquad (3.21)$$

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

$$\hat{x}_p = \hat{x}$$

and thus MSEP can be written as,

$$MSEP(y_p) = E\{ (\hat{x}-x)'(\hat{x}-x) \} + n\sigma^2$$

$$= MSE(x) + n\sigma^2 \qquad (3.22)$$

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

$$MSE(x) = \|(I_n-U_1U_1^+)U_2\theta_2\|^2 + p\sigma^2 \qquad (3.23)$$

Consider now the expectation of the residuals sum of squares.

$$E\{ RSSp \} = \|(I_n-U_1U_1^+)U_2\theta_2\|^2 + (n-p)\sigma^2 \qquad (3.24)$$

As an estimate of $MSE(x)/\sigma^2$ Mallows ('73) considered the following

function,

$$\boxed{Cp = RSSp/\hat{\sigma}^2 + 2p-n} \qquad (3.25)$$

Here $\hat{\sigma}^2$ is some estimate of σ^2. The most practible estimates of σ^2 are $\hat{\sigma}^2(d)$ and $\bar{\sigma}^2(d)$. When $\hat{\sigma}^2=\bar{\sigma}^2(d)$ Cp is an unbiased estimate of (3.23). So the model structure that minimizes Cp should be chosen. Mallows pointed out that it is better to inspect a plot of Cp versus p than blindly choose the model that minimizes Cp. Models leading to smaller Cp are prefered, but points close to the line 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 MSE(x).

3.4.2 AMEMIYA'S PC

Amemiya ('80) 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 $x_p=u_f'\hat{\theta}_1$, the mean square error function of prediction can be written as,

$$E\{ (y_a-x_p)^2 \} = \sigma^2[1+u_f'(U_1'U_1)^{-1}u_f] + [u_f'\theta-u_f'U_1^+U\theta]^2 \qquad (3.26)$$

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

$$E\{ u_f u_f' \} = U'U/n \qquad (3.27)$$

Then we obtain,

$$E\{ (y_a-x_p)^2 \} = \|(I_n-U_1U_1^+)U_2\theta_2\|^2/n + \sigma^2(1+p/n) \qquad (3.28)$$

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 σ^2 with its estimator $\hat{\sigma}^2(p)$ and then minimizing the first term of (3.28) by putting it to zero. Thus,

$$\boxed{PC= \hat{\sigma}^2(p)(1+p/n) = (1/n)*RSSp(1+2p/(n-p))} \qquad (3.29)$$

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,

$$\|(I_n-U_1U_1^+)y\|^2/n - \sigma^2(n-p)/n \qquad (3.30)$$

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

$$RSSp/n + 2p\sigma^2/n \qquad (3.31)$$

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

Amemiya pointed out that his 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,

$$FPE = V(\bar{\theta})((n+\#par)/(n-\#par)) \tag{3.32}$$

here #par is the number of free parameters and $V(\bar{\theta})$ 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,

$$\boxed{FPEp = RSSp((n+p)/(n-p)) = RSSp(1+2p/(n-p))} \tag{3.33}$$

Thus we see that this method is equivalent with Amemiya's 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 θ denotes the unknown parameters of the process. Let us consider a set of estimates $\bar{\theta}$'s of the vector of parameters θ. The objective is to estimate θ in such a way that the a posteriori probability density function (PDF) $p(y;\bar{\theta})$ 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,

$$I(\theta,\bar{\theta}) = E\{ -\ln(p(y;\bar{\theta})/p(y;\theta)) \} =$$

$$-\int p(y;\theta) \ln (p(y;\bar{\theta})/p(y;\theta)) \, dy \tag{3.34}$$

Here the expectation is taken with respect to y. We consider $\bar{\theta}$ 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,

$$a: I(\theta,\bar{\theta}) \geq 0 \tag{3.35}$$

$$b: I(\theta,\bar{\theta}) = 0 \text{ if and only if } \theta=\bar{\theta} \tag{3.36}$$

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

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

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 θ, giving the well known maximum likelihood estimator of θ.

Akaike ('73) proposed to minimize,

$$E(2*I(\theta,\hat{\theta})) \tag{3.38}$$

Here the expectation is taken with respect to the distribution of $\bar{\theta}$. 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),

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

Where #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),

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

and thus,

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

Since,

$$p(y;\theta) = (2\pi\sigma^2)^{-n/2} \exp (-(y-U\theta)'(y-U\theta)/2\sigma^2) \tag{3.42}$$

we can see that,

$$AIC = n*\ln(2\pi) + n*\ln(\sigma^2) + |\bar{y}-y|^2 / \sigma^2 + 2p \tag{3.43}$$

Minimizing the third term of the AIC (within our model structure) leads to the least square estimator $\hat{\theta}_1$ of θ_1. So when σ^2 is known this AIC leads to

σ^2 known,

$$\min AIC = \min \{ RSSp/\sigma^2 + 2p \} \tag{3.44}$$

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

σ^2 unknown,

$$\min AIC = \min \{ n*\ln(\tilde{\sigma}^2(p)) + 2p \} \tag{3.45}$$

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

$$\hat{\sigma}^2(p) \quad \min \{ n*\ln(\hat{\sigma}^2(p)) + p \} \tag{3.46}$$

$$\hat{\sigma}^2(d) \quad \min \{ RSSp/\hat{\sigma}^2(d) + 2p \} \tag{3.47}$$

$$\tilde{\sigma}^2(d) \quad \min \{ RSSp/\tilde{\sigma}^2(d) + 2p \} \tag{3.48}$$

Here we see that in the last two cases the criteria become equivalent with the Cp 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 reponse. In this chapter it is assumed that there exist some prior density function about θ. 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,

$$p(M_p) \qquad \text{with} \sum_{p=1}^{d} p(M_p) = 1 \tag{3.49}$$

Furthermore we assume that the prior probability density function of the parameters (θ_1) in the model structure M_p is given by,

$$p(\theta_1 | M_p) \tag{3.50}$$

Let,

$$p(y|\theta_1,M_p) \tag{3.51}$$

denote the usual probability density function of y conditional on the model structure M_p with its parameters θ_1.

Then with the relations,

$$p(y|M_p) = \int p(y|\theta_1,M_p)*p(\theta_1|M_p)\, d\theta_1 \tag{3.52}$$

$$p(y) = \sum_{p=1}^{d} p(y|M_p)*p(M_p) \tag{3.53}$$

it is possible to derive the probability of the model structure M_p conditional on y. Because,

$$p(y|M_p) = p(y,M_p) \;/\; p(M_p) \tag{3.54}$$

and

$$p(M_p|y) = p(y,M_p) \;/\; p(y) \tag{3.55}$$

we can see that,

$$p(M_p|y) = p(y|M_p) * p(M_p) \;/\; \sum_{p=1}^{d} p(y|M_p)*p(M_p) \tag{3.56}$$

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 θ_1 is given by,

$$\theta_1 \in N(0,\tau^2(U_1'U_1)^{-1}) \tag{3.57}$$

Thus the covariance matrix of the prior is except for a factor equal to the covariance matrix of the least-squares estimate of θ. It is assumed that all model structures have equal probability.

Thus because,

$$p(y|\theta_1,M_p) = N(U_1\theta_1,\sigma^2 In) \tag{3.58}$$

we have that,

$$p(y|M_p) = N(0, \sigma^2 I_n + \tau^2 U_1 (U_1' U_1)^{-1} U_1') \qquad (3.59)$$

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

$$\boxed{\begin{aligned} BIC = \quad & (n-p)*\ln(\; |(I_n - U_1 U_1^+)y|^2 \; / \; (n-p) \;) \\ & p*\ln(\; |U_1 U_1^+ y|^2 \; / \; p \;) \end{aligned}} \qquad (3.60)$$

The model that minimizes this BIC shoold be chosen.

3.6.2 SCHWARZ CRITERION (SC)

Schwarz ('78) 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,

$$SC = -2\ln LH(\theta|y) + \#par*\ln(n) \qquad (3.61)$$

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

$$\boxed{\begin{aligned} & \sigma^2 \text{ known} \\ & \min SC = \min(\; RSSp/\sigma^2 + p*\ln(n) \;) \end{aligned}} \qquad (3.62)$$

$$\boxed{\begin{aligned} & \sigma^2 \text{ unknown} \\ & \min SC = \min(\; n*\ln(\; \tilde{\sigma}^2(p) \;) + p*\ln(n) \;) \end{aligned}} \qquad (3.63)$$

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).

3.7 CROSS-VALIDATION ARGUMENTS

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,

$U_1(/i)$ = U_1 matrix without the i-th row.
$u1(i)$ = the i-th row of U_1
$y(/i)$ = 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,

$$y_p(i) = u1(i) * U_1(/i)^+ \, y(/i) \tag{3.64}$$

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 PRESS(p). Here p stands for the dimension of the model.

$$PRESS(p) = \sum_{i=1}^{n} [\, y(i) - y_p(i) \,]^2 \tag{3.65}$$

The proposed procedure is now to choose the model that minimizes PRESS.

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

$$\boxed{PRESS(p) = w_p' D_p^{-2} w_p} \tag{3.66}$$

Where,

$w_p = (I_n - UU_1^+)y$

D_p = the diagonal matrix whose diagonal elements are those of $(I_n - U_1 U_1^+)$

3.7.2. STOICA'S C1(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 adapt some notations used in that article.

Let the interval I={ 1..n } be divided in 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 Iv can be defined as follows,

$$Iv = \{ (v-1)m+1,..,vm \} \quad v = 1,..,k-1 \tag{3.67}$$

$$Ik = \{ (k-1)m+1,..,n \} \quad v = k \tag{3.68}$$

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

$$\hat{\theta} = \arg\min_{\phi} V(\phi) \qquad V(\phi) = 1/n \sum_{t\epsilon I} e^2(t,\phi) \tag{3.69}$$

Stoica et al. assumed the following conditions to hold,

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

A2. $V_{\phi\phi}(\hat{\theta}) = \delta^2 V(\phi) / \delta\phi^2 \,|_{\phi=\hat{\theta}}$ is positive definite

A3. The residuals $e(t,\phi)$ and

$$e_{\phi}(t,\phi) = \delta\, e(t,\phi) / \delta\phi$$

$$e_{\phi\phi}(t,\phi) = \delta^2\, e(t,\phi) / \delta\phi^2$$

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

FIRST CROSS-VALIDATION CRITERION

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

$$C_I = \sum_{v=1}^{k} \sum_{t\epsilon I_v} e^2(t,\hat{\theta}_v) \tag{3.70}$$

with,

$$\hat{\theta}_v = \arg\min_{\phi} \sum_{t\epsilon I - I_v} e^2(t,\phi) \quad v=1..k \tag{3.71}$$

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, <u>for k large enough</u> the following relation (for the SISO case) holds,

$$1/n\ C_I = C_1 + O(1/k^2 m) \tag{3.72}$$

where,

$$C_1(m) = V(\hat{\theta}) + 4/n^2 \sum_{v=1}^{k} wv(\hat{\theta})' V_{\phi\phi}(\hat{\theta})^{-1} wv(\hat{\theta}) \tag{3.73}$$

with,

$$wv(\hat{\theta}) = \sum_{t \epsilon I_v} e(t,\hat{\theta}) e_\phi(t,\hat{\theta}) \qquad v=1..k \tag{3.74}$$

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

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

<u>SECOND CROSS-VALIDATION CRITERION</u>

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

$$C_{II} = \sum_{v=1}^{k} \sum_{t \epsilon I - I_v} e^2(t,\hat{\theta}_v) \tag{3.75}$$

where,

$$\hat{\theta}_v = \arg\min_{\phi} \sum_{t \epsilon I_v} e^2(t,\phi) \qquad v=1..k \tag{3.76}$$

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

$$1/((k-1)n)\ C_{II} = C_2 + O(\ 1/\min(n, m^{3/2})\ ) \tag{3.77}$$

where,

$$C_2(m) = V(\hat{\theta}) + 2k/n^2 \sum_{v=1}^{k} wv(\hat{\theta})' V_{\phi\phi}(\theta)^{-1} wv(\hat{\theta}) \tag{3.78}$$

with,

$$wv(\hat{\theta}) = \sum_{t \epsilon I_v} e(t,\hat{\theta}) e_\phi(t,\hat{\theta}) \qquad v=1..k \tag{3.79}$$

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 C_I and C_{II}, 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 estimtate $(\hat{\theta})$.

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 ($y=U_1\theta_1+e$) we have,

$$e(t,\phi) = y(t)-u1(t)\phi \tag{3.80}$$

Here u1(t) is the t-th row of U_1 and y(t) is the t-th element of y. So we have,

$$e_{\phi}(t,\phi) = -u1(t)' \tag{3.81}$$

$$e_{\phi\phi}(t,\phi) = 0 \tag{3.82}$$

and,

$$V(\phi) = (y-U_1\phi)'(y-U_1\phi)\ /n \tag{3.83}$$

$$V_{\phi\phi} = 2U_1{}'U_1\ /n \tag{3.84}$$

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 $e_{\phi}(t,\phi)$ 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,

$$\begin{aligned}
\underline{0} &= \delta/\delta\phi\ \ 1/n \sum_{t\epsilon I-I_V} e^2(t,\phi)\ \Big|_{\phi=\hat{\theta}_V} \\
&= V_{\phi}(\hat{\theta}_V) - 2/n \sum_{t\epsilon I_V} e(t,\hat{\theta}_V)e_{\phi}(t,\hat{\theta}_V) \\
&= V_{\phi}(\hat{\theta}) + V_{\phi\phi}(\hat{\theta})(\hat{\theta}_V-\hat{\theta}) + \\
&\quad -2/n \sum_{t\epsilon I_V} e(t,\hat{\theta})e_{\phi}(t,\hat{\theta}) - 2/n \sum_{t\epsilon I_V} e_{\phi}(t,\hat{\theta})e_{\phi}(t,\hat{\theta})'(\hat{\theta}_V-\hat{\theta}) \\
&= 2/n\ U_1{}'U_1\ (\hat{\theta}_V-\hat{\theta}) + 2/n\ Uv'ev - 2/n\ Uv'Uv\ (\hat{\theta}_V-\hat{\theta}) \\
&= 2/n\ Us'Us\ (\hat{\theta}_V-\hat{\theta}) + 2/n\ Uv'ev
\end{aligned} \tag{3.85}$$

where Uv is a matrix whose rows are the vectors u1(t) for $t \epsilon I_v$ and ev is the corresponding vector of residuals. Us is the matrix U1 without the rows u1(t) for $t \epsilon I_v$. Because of relation (3.85) we can write,

$$(\hat{\theta}_v - \hat{\theta}) = -(Us'Us)^{-1}\, Uv'ev \quad \{ = O(1/k\sqrt{m}) \} \tag{3.86}$$

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

$$1/m \sum_{t \epsilon I_v} e(t,\hat{\theta}) e_\phi(t,\hat{\theta}) = E\{ e(t,\hat{\theta}) e_\phi(t,\hat{\theta}) \} + O(1/\sqrt{m}) =$$

$$1/n \sum_{t \epsilon I} e(t,\hat{\theta}) e_\phi(t,\hat{\theta}) + O(1/\sqrt{n}) + O(1/\sqrt{m}) = O(1/\sqrt{m})$$

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

$$1/m \sum_{t \epsilon I_v} -e(t,\hat{\theta}) e_\phi(t,\hat{\theta}) = -1/m\, Uv'ev$$

and

$$1/n \sum_{t \epsilon I} -e(t,\hat{\theta}) e_\phi(t,\hat{\theta}) = -1/n\, U_1'(I - U_1 U_1^{\dagger})y = 0$$

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,

Approximation-1 :

$$(\hat{\theta}_v - \hat{\theta}) = -(U_1'U_1)^{-1}\, Uv'ev + O(1/k^2\sqrt{m}) \tag{3.87}$$

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

$$(Us'Us)^{-1} = (U_1'U_1)^{-1} + O(m/n^2) \tag{3.88}$$

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

$$e^2(t,\hat{\theta}_v) = e^2(t,\hat{\theta}) + 2\, e(t,\hat{\theta}) e_\phi(t,\hat{\theta})' (\hat{\theta}_v - \hat{\theta}) +$$

$$(\hat{\theta}_v - \hat{\theta})' e_\phi(t,\hat{\theta}) e_\phi(t,\hat{\theta})' (\hat{\theta}_v - \hat{\theta})$$

$$= e^2(t,\hat{\theta}) - 2\, e(t,\hat{\theta}) u1(t) (\hat{\theta}_v - \hat{\theta}) +$$

$$(\hat{\theta}_v-\hat{\theta})'u1(t)'u1(t)(\hat{\theta}_v-\hat{\theta})$$

So,

$$\sum_{t\epsilon I_v} e^2(t,\hat{\theta}_v) = \sum_{t\epsilon I_v} \{ e(t,\hat{\theta}) - u1(t)(\hat{\theta}_v-\hat{\theta}) \}^2 \qquad (3.89)$$

and,

$$1/n\ C_I = 1/n \sum_{v=1}^{k} \sum_{t\epsilon I_v} e^2(t,\hat{\theta}_v) =$$

$$1/n \sum_{v=1}^{k} \sum_{t\epsilon I_v} \{ e(t,\hat{\theta}) - u1(t)(\hat{\theta}_v-\hat{\theta}) \}^2 \qquad (3.90)$$

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

Approximation-2 :

$$1/n\ C_I = 1/n \sum_{v=1}^{k} \sum_{t\epsilon I_v} \{ e^2(t,\hat{\theta}) - 2e(t,\hat{\theta})u1(t)(\hat{\theta}_v-\hat{\theta}) \} + O(1/k^2m) \qquad (3.91)$$

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

$$1/n\ C_I = 1/n \sum_{v=1}^{k} \sum_{t\epsilon I_v} \{e^2(t,\hat{\theta}) - 2e(t,\hat{\theta})u1(t)^* [-(U_1'U_1)^{-1} Uv'ev + O(1/k^2\sqrt{m})] \} + O(1/k^2m)$$

$$= 1/n \sum_{v=1}^{k} (ev'ev + 2ev'Uv(U_1'U_1)^{-1}Uv'ev) + O(1/k^2m)$$

$$= V(\hat{\theta}) + 4/n^2 \sum_{v=1}^{k} wv(\hat{\theta})'V_{\theta\theta}(\hat{\theta})^{-1}wv(\hat{\theta}) + O(1/k^2m)$$

$$= C_1 + O(1/k^2m) \qquad (3.92)$$

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

Choose the model structure that minimizes,

$$C_1(m) = 1/n \sum_{v=1}^{k} \{ ev'ev + 2ev'Uv(U_1'U_1)^{-1}Uv'ev \}$$

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,

$$1/n\ C_I = 1/n \sum_{v=1}^{k} \left| ev + Uv(U_1'U_1)^{-1}Uv'ev \right|^2 + O(1/k^2m) \qquad (3.93)$$

The difference between the real cross validation criterion and the approximated one remains of the order $O(1/k^2m)$, so in order to approximate $1/n\ C_I$ 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,

$$1/n \sum_{v=1}^{k} ev'ev + 2ev'Uv(U_1'U_1)^{-1}Uv'ev =$$

$$V(\hat{\theta}) + 2/n*tr\{ (U_1'U_1)^{-1} * \sum_{v=1}^{k} Uv'evev'Uv \} \qquad (3.94)$$

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

$$V(\hat{\theta}) + 2tr\{ (U_1'U_1)^{-1} * 1/n * \sum_{v=1}^{k} U_v'V(\hat{\theta})I_mUv =$$

$$V(\hat{\theta}) + 2/n\ V(\hat{\theta})\ tr\{ (U_1'U_1)^{-1}\ (U_1'U_1) \} =$$

$$V(\hat{\theta}) + 2p/n\ V(\hat{\theta}) = V(\hat{\theta})\ [1 + 2p/n\ ] \qquad (3.95)$$

So for large n we have,

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

$$\approx \ln\{ V(\hat{\theta}) \} + 2p/n = AIC/n \qquad (3.96)$$

(see 3.45)

When n tends to infinity the model is usually overestimated and thus the condition that ev is a white noise vector is fulfilled.

STOICA'S $C_2(m)$ IN THE LINEAR REGRESSION MODEL

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

$$\begin{aligned} \underline{0} &= \sum_{t\epsilon I_v} e(t,\hat{\theta}_v)e_\phi(t,\hat{\theta}_v) \\ &= \sum_{t\epsilon I_v} e(t,\hat{\theta})e_\phi(t,\hat{\theta}) + 2/m \sum_{t\epsilon I_v} e_\phi(t,\hat{\theta})e_\phi(t,\hat{\theta})'(\hat{\theta}_v-\hat{\theta}) \\ &= -Us'es + Us'Us\,(\hat{\theta}_v-\hat{\theta}) \end{aligned} \tag{3.97}$$

So,

$$(\hat{\theta}_v-\hat{\theta}) = (Us'Us)^{-1}\,Us'es \quad (= O(1/\sqrt{m})) \tag{3.98}$$

where Us is the matrix whose rows are the vectors $u1(t)$ for $t\epsilon I_v$ and es is the vector of the corresponding residuals. In (3.98) it is assumed that $Us'es = O(\sqrt{m})$ (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),

> Approximation-1 :
>
> $(\hat{\theta}_v-\hat{\theta}) = n/m\,(U_1'U_1)^{-1}\,Us'es + O(1/m)$ (3.99)

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

$$(Us'Us)^{-1} = n/m\,((U_1'U_1)^{-1} + O(1/(m\sqrt{m}) \tag{3.100}$$

Let us now evaluate C_{II},

$$\begin{aligned} e^2(t,\hat{\theta}_v) &= e^2(t,\hat{\theta}) + 2\,e(t,\hat{\theta})e_\phi(t,\hat{\theta})'\,(\hat{\theta}_v-\hat{\theta}) + \\ &\quad (\hat{\theta}_v-\hat{\theta})'e_\phi(t,\hat{\theta})e_\phi(t,\hat{\theta})'(\hat{\theta}_v-\hat{\theta}) \\ &= e^2(t,\hat{\theta}) - 2\,e(t,\hat{\theta})u1(t)(\hat{\theta}_v-\hat{\theta}) + \\ &\quad (\hat{\theta}_v-\hat{\theta})'u1(t)'u1(t)(\hat{\theta}_v-\hat{\theta}) \\ &= (\,e(t,\hat{\theta}) - u1(t)(\hat{\theta}_v-\hat{\theta})\,)^2 \end{aligned} \tag{3.101}$$

So,

$$1/((k-1)n)\,C_{II} = 1/((k-1)n) \sum_{v=1}^{k} \sum_{t\epsilon I-I_v} e^2(t,\hat{\theta}_v) =$$

$$1/((k-1)n) \sum_{v=1}^{k} \sum_{t\epsilon I-I_v} (\,e(t,\hat{\theta}) - u1(t)(\hat{\theta}_v-\hat{\theta})\,)^2 =$$

$$1/((k-1)n) \sum_{v=1}^{k} \sum_{t\epsilon I-I_v} e^2(t,\hat{\theta}) +$$

$$1/((k-1)n) \sum_{v=1}^{k} \sum_{t\epsilon I-I_v} -2e(t,\hat{\theta})u1(t)(\hat{\theta}_v-\hat{\theta}) +$$

$$1/((k-1)n) \sum_{v=1}^{k} \sum_{t\epsilon I-I_v} (\hat{\theta}_v-\hat{\theta})'u1(t)'u1(t)(\hat{\theta}_v-\hat{\theta}) =$$

$$V(\hat{\theta}) + T2 + T3 \qquad (3.102)$$

First let us consider term T3.

$$T3 = 1/((k-1)n) \sum_{v=1}^{k} \sum_{t\epsilon I-I_v} (\hat{\theta}_v-\hat{\theta})'u1(t)'u1(t)(\hat{\theta}_v-\hat{\theta})$$

$$= 1/((k-1)n) \sum_{v=1}^{k} (\hat{\theta}_v-\hat{\theta})\, Uv'Uv(\hat{\theta}_v-\hat{\theta}) \qquad (3.103)$$

Here Uv denotes the matrix U_1 without the rows u1(t) for $t\epsilon I_v$. Stoica et al. used the following approximation of T3,

Approximation-2

$$T3 = 1/((k-1)n) \sum_{v=1}^{k} (\hat{\theta}_v-\hat{\theta})'U_1'U_1(\hat{\theta}_v-\hat{\theta}) \qquad (3.104)$$

$$+ O(1/n)$$

Where it is assumed that,

$$1/n\ Uv'Uv = 1/n\ U_1'U1 + O(m/n) \qquad (3.105)$$

Let us now consider term T2,

$$T2 = -2/((k-1)n) \sum_{v=1}^{k} \sum_{t\epsilon I-I_v} e(t,\hat{\theta})u1(t)(\hat{\theta}_v-\hat{\theta}) =$$

$$-2/((k-1)n) \sum_{v=1}^{k} \sum_{t\epsilon I} e(t,\hat{\theta})u1(t)(\hat{\theta}_v-\hat{\theta})$$

$$+ 2/((k-1)n) \sum_{v=1}^{k} \sum_{t\epsilon I_v} e(t,\hat{\theta})u1(t)(\hat{\theta}_v-\hat{\theta}) =$$

$$2/((k-1)n) \sum_{v=1}^{k} \sum_{t\epsilon I_v} e(t,\hat{\theta})u1(t)(\hat{\theta}_v-\hat{\theta}) \qquad (3.106)$$

Stoica et al. used the following approximation,

Approximation-2
T2 = O(1/n)

(3.107)

Because $(\hat{\theta}_v-\hat{\theta})$ is of order $O(1/\sqrt{m})$. Since $(\hat{\theta}_v-\hat{\theta})$ is of order $O(1/\sqrt{m})$, T3 is of order O(1/m) and thus much larger than T2 when n tends to infinity.

When we use the approximation-1 of $(\hat{\theta}_v-\hat{\theta})$, T3 becomes,

$$T3 = 1/((k-1)n) \sum_{v=1}^{k} \{n/m\ (U_1'U_1)^{-1}Us'es + O(1/m)\}'U_1'U_1*$$
$$\{n/m\ (U_1'U_1)^{-1}Us'es + O(1/m)\}$$
$$+ O(1/n)$$

$$= 1/((k-1)n) \sum_{v=1}^{k} n^2/m^2\ es'Us\ (U_1'U_1)^{-1}\ Us'es$$
$$+ O(1/m\sqrt{m}) + O(1/n)$$

$$= k/n \sum_{v=1}^{k} es'Us\ (U'U)^{-1}\ Us'es + O(1/\min(n,m\sqrt{m}))$$

$$= 2k/n^2 \sum_{v=1}^{k} wv'V_{\phi\phi}^{-1}*wv + O(1/\min(n,m\sqrt{m})) \qquad (3.108)$$

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

$$1/((k-1)n)\ C_{II} = C_2 + O(1/\min(n,m\sqrt{m}) \qquad (3.109)$$

with,

$$C_2 = V(\hat{\theta}) + k/n^2 \sum_{v=1}^{k} wv'V_{\phi\phi}^{-1}wv$$

Thus the second cross-validation criterion for the linear regression model can be stated as,

> Choose the model structure that minimizes,
>
> $$C_2 = 1/n \sum_{v=1}^{k} es'es + k\, es'Us\, (U'U)^{-1}\, Us'es$$

We see that this criterion is essentially the same as C_1. The only difference is that that 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,

$$C2 \approx \ln(V(\hat{\theta})) + kp/n \qquad (3.110)$$

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,

$$m/(n-m) \approx f/n \qquad (3.111)$$

However, since m/(n-m) must be small for C1(m) to be a good approximation of CI, C1(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 C2(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 θ. 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, $r_1..r_n$ with unknown means $\mu_1..\mu_n$ and known variance σ^2. It is common to estimate $\mu_1..\mu_n$ by $r_1..r_n$. However, Stein ('56) proved that this estimator is admissable for n<3 and inadmissable for n≥3 under the loss function,

$$MSE(\mu,\hat{\mu}) = E\{ (\hat{\mu}-\mu)'(\hat{\mu}-\mu) \} \qquad (3.112)$$

(An estimator δ of μ is said to be inadmissable if there exists an estimator δ^* for which $MSE(\mu,\delta^*) \leq MSE(\mu,\delta)$ for all μ with strict inequality for some μ. An estimator is admissable if it is not inadmissable.)

James and Stein ('61) exhibited an estimator that is uniformly better than the least squares one. For a review on this topic see Draper and Van Nostrand ('79) or Judge and Bock ('78).

Suppose that we have,

$$r = \mu + e \qquad e \in N(0,\sigma^2 I_n) \qquad (3.113)$$

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

$$E\{ |\hat{\mu}|^2 \} = |\mu|^2 + n*\sigma^2 \qquad (3.114)$$

Here we see that when $n*\sigma^2/\mu'\mu$ is large we would be over-estimating μ 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 (Sf) should be small when $n*\sigma^2/\mu'\mu$ is large and should be nearly 1 when $n*\sigma^2/\mu'\mu$ is small. Of course, μ is unknown. When taking its estimate the previous discussion suggest that we should consider estimates of the followng form,

$$\tilde{\mu} = [1 - c\sigma^2/\hat{\mu}'\hat{\mu}]* \hat{\mu} = [1 - c\sigma^2/r'r]* r \qquad (3.115)$$

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

$$\tilde{\mu} = [1 - c\sigma^2/\hat{\mu}'\hat{\mu}]* \hat{\mu} \qquad (3.116)$$

$$0 < c < 2(n-2) \qquad c\text{-optimal} = (n-2)$$

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

$$\tilde{\mu} = [1 - ((n-2)/(v+2))v^*\hat{\sigma}^2/\hat{\mu}'\hat{\mu}]^* \hat{\mu} \qquad (3.117)$$

$$0 < c < 2(n-2) \qquad c\text{-optimal} = (n-2)$$

is also uniformly better than the least squares estimator under the quadratic loss function (3.112) for n≥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,

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

$$= Sf(o)^* (\hat{\mu}-o) + o \qquad (3.118)$$

with,

$c = (n-2)$
v = degrees of freedom in $\hat{\sigma}^2$
o = 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 μ 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,

$$U'U = I_d / l \qquad (3.119)$$

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 >> d. Then we can see that $\hat{\theta}$ is independently normally distributed. Thus,

$$\hat{\theta} \in N(\theta, l\sigma^2 I_d) \qquad (3.120)$$

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

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

$$\tilde{\theta} = Sf^* \hat{\theta}$$

with,

$$0 < c < 2(d-2) \qquad c\text{-optimal} = (d-2)$$

$$\hat{\sigma}^2 = l^*RSSd/v \ \{ = l^*\hat{\sigma}^2(d) \} \qquad v=(n-d)$$

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

$$\bar{\theta} = \max(0, Sf)^* \hat{\theta} \tag{3.122}$$

with

$$Sf=[1-(c/(v+2))v^*\hat{\sigma}^2/\hat{\theta}'\hat{\theta}] \tag{3.123}$$

$$0 < c < 2(d-2) \qquad v=n-d$$

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,

$$\bar{\theta} = \begin{bmatrix} (\hat{\theta})1 \\ \max(0, Sf2)\ (\hat{\theta})2 \end{bmatrix} \text{ with } \hat{\theta} = \begin{bmatrix} (\hat{\theta})1 \\ (\hat{\theta})2 \end{bmatrix} \tag{3.124}$$

where

$$Sf2=[1-(c/(v+2))v^*\hat{\sigma}^2/(\hat{\theta})2'(\hat{\theta})2]$$

$(\hat{\theta})1$ is a p-vector
$(\hat{\theta})2$ is a (d-p)-vector

$$0 < c < 2(d-p-2) \qquad v=n-d$$

is uniformly better estimator than $\hat{\theta}$, under MSE(θ), if (d-p) ≥ 3. This last relation suggests that we can set the last (d-p) parameters to zero if Sf2 < 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 (see (3.17)),

$$(\hat{\theta})2'(\hat{\theta})2 \ /\{(d-p)\hat{\sigma}^2(d)\} < F^{\alpha}\ d-p,n-d \tag{3.125}$$

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,

$$(\hat{\theta})2'(\hat{\theta})2 \ /\{(d-p)\hat{\sigma}^2(d)\} < c^*l^*v/\{(v+2)(d-p)\} \tag{3.126}$$

Thus the positive James-Stein rule can be seen as a F-test with a particular value of α. However, the main difference is that when the hypothesis is rejected, the subset will be shrunken 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 $U'U = I_d / l$. What can we do when U this relation does not hold?
Let us consider the general model $y=U\theta+e=x+e$. Rewrite U as,

$$U=V\Sigma W' \text{ (singular value decomposition)} \tag{3.127}$$

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

$$U^+ = W\Sigma^*V' \tag{3.128}$$

$$U'U = W\Sigma'\Sigma W' = WTW' = WDDW' \tag{3.129}$$

Where Σ^* is a d*n diagonal matrix with $diag(\Sigma^*)=(1/\sigma_1,..,1/\sigma_d)$. D is a d*d diagonal matrix with the same elements as Σ. From this last expression we can see that the columns of W are the eigenvectors of $U'U$ and the diagonal values of T are the corresponding eigenvalues. Let us define,

$$Z = U^*WD^{-1}W' \tag{3.130}$$

$$\beta = WDW'^* \theta \tag{3.131}$$

Then we can rewrite the general model as,

$$y = Z\beta + e \tag{3.132}$$

This model will be called the <u>orthonormal canonical form</u> of the general model. In this canonical form the least squares estimator of β is given by,

$$\hat{\beta} = (Z'Z)^{-1}Z'y = Z' y = WDW' \hat{\theta} \tag{3.133}$$

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

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

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

$$\tilde{\beta} = \max(0,Sf)\ \hat{\beta} \tag{3.135}$$

with

$$Sf = [1-(c/(v+2))v^*\hat{\sigma}^2/\hat{x}'\hat{x}]$$

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

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

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

$$E((\hat{\beta}-\beta)'(\hat{\beta}-\beta)) = E((\hat{x}-x)'(\hat{x}-x)) \tag{3.136}$$

for $d \geq 3$. This means that the shrunken estimator in the original parameter space,

$$\tilde{\theta} = WD^{-1}W'\tilde{\beta} = \max(0,Sf)^{*}\ \hat{\theta} \qquad (3.137)$$

gives uniformly better results than $\hat{\theta}$ according to the performance function MSE(x). Since the estimator in the β space was independently normally distributed, we can test if some subset of β parameters can be set to zero according to the previous given method. However, although it can lead to a reduction in the β space, it does not lead to a reduction in the θ 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 MSE(x). But how does it perform under MSE(θ)? 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 RSSp. 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,

$$C = (\hat{x}-x)'(\hat{x}-x) + S = RSSp + S \tag{4.1}$$

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 $2p\hat{\sigma}^2$. Suppose we have two models A and B,

$$\begin{aligned} &\text{model A}: \quad y=U_1\theta_1 + U_2\theta_2 + e \quad \dim U = n*d \\ &\text{model B}: \quad y=U_1\theta_1 + e \qquad\qquad \dim U_1 = n*p \end{aligned} \tag{4.2}$$

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,

$$RSSp-RSSd = \|AA^+y\|^2 \quad \text{with } A=(I_n-U_1U_1^+)U_2 \tag{4.3}$$

See (3.14). Rewrite A as,

$$A = V\Sigma W' \qquad (\text{ singular value decomposition }) \tag{4.4}$$

Here V and W are column unitary and orthonormal, and Σ 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,

$$A^+ = W\Sigma^*V' \tag{4.5}$$

$$AA^+ = VTV' \tag{4.6}$$

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

$$T = DD' \text{ with } D = [\, I_{(d-p)} \; 0 \,]' = n*(d-p) \text{ matrix} \tag{4.7}$$

Then with the following definitions,

$$r=D'V'e \qquad (\text{ r=(d-p)-vector = noise vector}) \tag{4.8}$$

$$b=D'\Sigma\, W'\theta_2 \quad (\text{ b=(d-p)-vector = 'bias' vector}) \tag{4.9}$$

where

$$r \in N(0,\ \sigma^2 I_{(d-p)}) \tag{4.10}$$

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

$$|AA^{+}y|^2 = |V\Sigma W'\theta_2|^2 + 2e'V\Sigma W'\theta_2 + |VDD'V'e|^2$$

$$= |\Sigma W'\theta_2|^2 + 2e'VDD'\Sigma W'\theta_2 + |D'V'e|^2$$

$$= |b|^2 + 2r'b + |r|^2 \qquad (4.11)$$

Here,

$$|r|^2 / \sigma^2 \quad \epsilon \text{ Chi}^2 (d-p) \qquad (4.12)$$

$$2r'b / \sqrt{(4\sigma^2 b'b)} \quad \epsilon N(0,1) \qquad (4.13)$$

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,

$$|b|^2 = |(I_n - U_1U_1^{+})U_2\theta_2|^2 \qquad (4.14)$$

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,

$$Cb-Ca = |AA^{+}y|^2 + Sb-Sa \qquad (4.15)$$

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

$$Cb-Ca > 0 \quad <==>$$

$$b'b + 2r'b + r'r + Sb-Sa > 0 \quad <==>$$

$$b'b /F + 2b'r /F + r'r /F + (Sb-Sa) /F > 0 \qquad (4.16)$$

with $F = (4\sigma^2 b'b)$ (= O(n)) we see that,

$$b'b /F = 1/4\sigma^2 > 0 \qquad (4.17)$$

$$\operatorname*{plim}_{n\to\infty} 2b'r /F = \operatorname*{plim}_{n\to\infty} N(0,1) / \sqrt{F} = 0 \qquad (4.18)$$

$$\operatorname*{plim}_{n\to\infty} r'r /F = \operatorname*{plim}_{n\to\infty} \text{Chi}^2 (d-p)*\sigma^2 / F = 0 \qquad (4.19)$$

and finally,

$$\operatorname*{plim}_{n\to\infty} Cb-Ca > 0 \text{ if and only if}$$

$$\operatorname*{plim}_{n\to\infty} (Sb-Sa) / F > -1/4\sigma^2 \qquad (4.20)$$

Thus we see that when model A is true and n tends to infinity the right model is chosen if the order of Sb-Sa 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,

$$\operatorname*{plim}_{n\to\infty} Cb-Ca < 0 \text{ if and only if}$$

$$\operatorname*{plim}_{n\to\infty} r'r + Sb-Sa < 0 \tag{4.22}$$

Here we see that when Sa-Sb 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 Cp or AIC when σ^2 is known. First assume model A is true. Then we see that conditon (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,

$$P(r'r > 2\sigma^2(d-p)) = P(Chi^2 (d-p) > 2(d-p)) \tag{4.25}$$

that the wrong model is chosen.

Although we obtained this result for the case in which we had to made a choice between two model structures, we can say that in the general case in which we have d possible model structures the Cp 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 σ^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,

$$\operatorname*{plim}_{n\to\infty} Cb-Ca > 0 \tag{4.28}$$

Thus the right model is chosen. When model B is true we see that,

$$\operatorname{plim}_{n\to\infty} Cb-Ca = \operatorname{plim}_{n\to\infty} r'r + \sigma^2 \ln(n)(p-d) < 0 \qquad (4.29)$$

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

In literature, some authors have studied the asymptotic behaviour of several criteria. Nishii ('84) 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 ('81) proved that when the number of parameters is infinite or increases with the sample size his criterion,

$$Sn(p) = (n+2p)RSSp/n = (n+2p)\,\bar{\sigma}^2(p) \qquad (4.30)$$

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,

$$\begin{aligned} &\text{process} && y=U_1\theta_1 + U_2\theta_2 + e \\ &\text{model A:} && y=U_1\phi_1 + U_2\phi_2 + w \\ &\text{model B:} && y=U_1\phi_1 + w \end{aligned} \qquad (4.31)$$

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

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

$$|\theta_2|^2 + |U_1^{+}U_2\theta_2|^2 + \sigma^2\left(tr\left((U_1'U_1)^{-1} \right) - tr\left((U'U)^{-1} \right) \right) < 0 \qquad (4.32)$$

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

$$|(I_n - U_1U_1^{+})U_2\theta_2|^2 + \sigma^2(p-d) < 0 \qquad (4.33)$$

c: Under the risk matrix the model B gives better results when,

$$\|(I_n - U_1U_1^+)U_2\theta_2\|^2 - \sigma^2 < 0 \qquad (4.34)$$

Let us look at condition (4.32). Define,

$$H = U_2'(I_n - U_1U_1^+)U_2 \qquad (4.35)$$

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

$$H = W\Sigma\Sigma'W' = DD' \quad \text{with } D = W\Sigma \qquad (4.36)$$

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

$$\theta_2'\theta_2 + \theta_2'(U_1^+U_2)'(U_1^+U_2)\theta_2 =$$

$$\theta_2'(I_{(d-p)} + (U_1^+U_2)'(U_1^+U_2))\theta_2 =$$

$$\theta_2'D (D^{-1}(I_{(d-p)} + (U_1^+U_2)'(U_1^+U_2))(D^{-1})') D'\theta_2 =$$

$$\theta_2'D(G)D\theta_2 \qquad (4.37)$$

with,

$$G = D^{-1}(I_{(d-p)} + (U_1^+U_2)'(U_1^+U_2))(D^{-1})' \qquad (4.38)$$

Because G is symmetric it holds that,

$$\max_{x\neq 0} \; x'Gx / x'x = tmax \qquad (4.39)$$

Where tmax is the largest eigenvalue of G. Therefore,

$$\theta_2'D(G)D\theta_2 \leq tmax\ \theta_2'DD'\theta_2 = tmax\ \theta_2'H\theta_2 \qquad (4.40)$$

This means that for all possible θ_2 the condition (4.32) will be true if,

$$\|(I_n - U_1U_1^+)U_2\theta_2\|^2 < \sigma^2(tr((U_1'U_1)^{-1}) - tr((U'U)^{-1})) / tmax \qquad (4.41)$$

When in the conditions (4.33), (4.34) and (4.41) θ_2 and σ^2 are replaced with their estimators ($\hat{\theta}_2$ and RSSd/(n-d)), we can see that these conditions the 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,

$$F < 1 \tag{4.42}$$

$$F < 1/(d-p) \tag{4.43}$$

and

$$F < \sigma^2(\text{tr}((U_1'U_1)^{-1}) - \text{tr}((U'U)^{-1})) / \text{tmax} \tag{4.44}$$

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 : $F < F^{\alpha}(d-p,n-p)$ (4.45)

Mallows' Cp (with $\hat{\sigma}^2(d)$) : $F < 2$ (4.46)

Amemiya's PC (=FPE) : $F < 2n/(n+p)$ (4.47)

Akaike's AIC (σ^2 unknown) :

$$F < [\exp(2(d-p)/n) - 1]*[(n-d)/(d-p)] \tag{4.48}$$

Schwarz' SC

σ^2 known :

$$F < \ln(n) \tag{4.49}$$

σ^2 unknown :

$$F < [\exp(\ln(n)(d-p)/n) - 1]*[(n-d)/(d-p)] \tag{4.50}$$

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 α it is interessing to study the sampling properties of such a 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 θ of the F-test as a function of the length of θ_2. Since this MSE(θ) can be written as,

$$\begin{aligned} MSE(\theta) &= E((\theta_1-\hat{\theta}_1)'(\theta_1-\hat{\theta}_1)) + E((\theta_2-\hat{\theta}_2)'(\theta_2-\hat{\theta}_2) \\ &= p*\sigma^2 + MSE(\theta_2) \end{aligned} \tag{4.51}$$

When c=0, so model A is always chosen, then $MSE(\theta_2)=(d-p)\sigma^2$ and $MSE(\theta) = d*\sigma^2$. From Sclove ('72) and Judge and Bock ('78) we know that for c<>0,

$MSE(\theta_2) > (d-p)*\sigma^2$ if $|\theta_2|^2 > (d-p)*\sigma^2$ (4.52)

$MSE(\theta_2) < (d-p)*\sigma^2$ if $|\theta_2|^2 < (d-p)*\sigma^2/2$ (4.53)

$MSE(\theta_2) = (d-p)\sigma^2*L$ if $|\theta_2|^2 = 0$ (4.54)

where $L=P(\text{Chi}^2(d-p+2)/\text{Chi}^2(n-d) > c*(d-p)/(n-d))$

Furthermore as $\theta_2'\theta_2$ increases and approaches infinity, the $MSE(\theta_2)$ approaches the risk of the least square estimator.

Knowing these properties a plot of $MSE'\theta_2)$ versus the length of θ_2 will yield the following characteristic (see Judge and Bock ('78)).

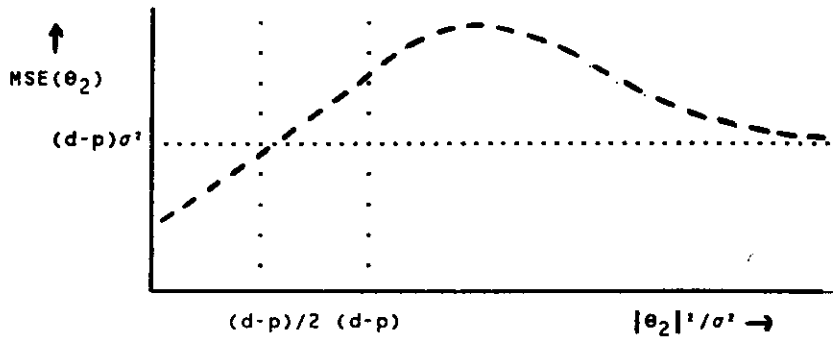

Figure 1: $MSE(\theta_2)$ versus $|\theta_2|^2/\sigma^2$ for the F-test.

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 θ_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,

choose model B when $F \leq c$ (4.55)
choose model A when $F > c$ and estimate θ_2 by

$$\tilde{\theta}_2 = [1- (a(d-p-2)/(n-d+2))*FRSS/\hat{\theta}_2'\hat{\theta}_2]* \hat{\theta}_2$$

where $0 < a \leq 2$

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

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).

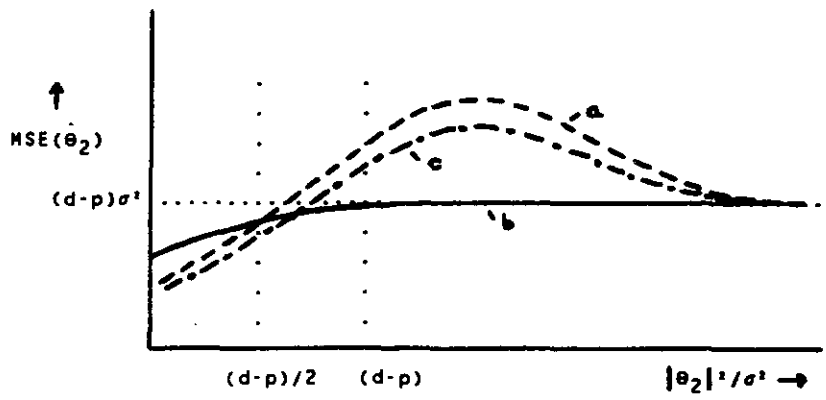

Figure 2: MSE($\hat{\theta}_2$) versus $|\theta_2|^2/\sigma^2$; a=F-test, b=positive James-Stein estimator, c=modified F-test.

Conclusion:

In this section we saw that most criteria can be written in terms of F-test statistics with a particular value of α. 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 (MSE(θ)) 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 '78).

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).

$$MSE(\theta) = |\theta_2|^2 + |U_1^+U_2\theta_2|^2 + \sigma^2 tr\{ (U_1'U_1)^{-1} \} \qquad (4.56)$$

Since θ_2 is unkown we replace it with its unbiased estimate. Let $\hat{\theta}_2$ denote the estimate of θ_2 from the full model. First, let us look at the estimate of θ in the full model.

$$\hat{\theta} = \begin{bmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{bmatrix} = U^+ y = \begin{bmatrix} T_1 \\ T_2 \end{bmatrix} y \qquad (4.57)$$

Here

$T_1 = U_1^+ - U_1^+U_2T_2$ is a p*n matrix
$T_2 = A^+$ is a (d-p)*n matrix
$A = (I_n - U_1U_1^+)U_2$ is a n*(d-p) matrix

(see appendix A)

So if we estimate θ_2 by $\hat{\theta}_2$ from the full model we see that,

$$E\{ |\hat{\theta}_2|^2 \} = |\theta_2|^2 + \sigma^2 tr\, T_2'T_2 \qquad (4.58)$$

$$E\{ |U_1^+U_2\hat{\theta}_2|^2 \} = |U_1^+U_2\theta_2|^2 + \sigma^2 tr\{(U_1^+ - T_1)'(U_1^+ - T_1)\} \qquad (4.59)$$

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

$$|\hat{\theta}_2|^2 + |U_1^+U_2\hat{\theta}_2|^2 + \hat{\sigma}^2(d)*tr\{ (U_1'U_1)^{-1} \}$$

$$-\hat{\sigma}^2(d)*(tr\{ T_2'T_2 \} + tr\{ (U_1^+ - T_1)'(U_1^+ - T_1) \}) \qquad (4.60)$$

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),

$$2*|U^{+}e|^{2} - 2*\hat{\sigma}^{2}(d)*tr\{ (U'U)^{-1} \} \qquad (4.61)$$

to (4.60) we get an easier to handle unbiased estimate (Np='New' Cp criterion) of MSE(θ),

$$Np = | \hat{\theta}_p - \hat{\theta}_f |^{2} + \hat{\sigma}^{2}(d)*(tr\{ 2*(U_1'U_1)^{-1} \}-tr\{ (U'U)^{-1} \})$$

$$= |U^{+}(I-U_1U_1^{+})y|^{2} + \hat{\sigma}^{2}(d)*(tr\{ 2*(U_1'U_1)^{-1} \}-tr\{ (U'U)^{-1} \}) \qquad (4.62)$$

with

$$\hat{\theta}_p = \begin{bmatrix} \hat{\theta}_1 \\ 0 \end{bmatrix} = \begin{bmatrix} U_1^{+}y \\ 0 \end{bmatrix} = \hat{\theta} \text{ of model with dimension p}$$

$$\hat{\theta}_f = U^{+}y = \hat{\theta} \text{ of full model}$$

So this Np criterion looks very like the original Cp criterion. When $U'U = I_n/l$ minimizing of Np is equal to minimizing Cp.

Since the Np criterion is especially designed to minimize MSE(θ) the question arises if it is indeed better than Cp or not. As a measure of performance we should calculate the risk of estimating θ 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: $y=U\theta + e$
model A: $y=U\theta + w$
model B: $y=w$ thus all parameters are set to zero

and assume that <u>σ^2 is known</u>. Then we choose model B when,

Np criterion

$$Ln=|U^{+}y|^{2} - 2*\sigma^{2} tr\{ (U'U)^{-1} \} \leq 0 \qquad (4.63)$$

Cp criterion

$$Lc=|y|^{2} - |(I_n-UU^{+})y|^{2} - 2\sigma^{2}(d) \leq 0 \qquad (4.64)$$

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

Np criterion

$$MSE(Np,\theta)= E\{ \max(0,-\text{sign}(Ln))*|\theta|^{2} \} + E\{ \max(0, \text{sign}(Ln))*|\theta-\hat{\theta}_f|^{2} \}$$

(4.65)

Cp criterion

$$MSE(Cp,\theta)= E\{ \max(0,-\text{sign}(Lc))*|\theta|^2 \} + E\{ \min(0, \text{sign}(Lc))*|\theta-\hat{\theta}_f|^2 \} \quad (4.66)$$

where,

sign(h) = 1 if h>0
sign(h) = -1 if h≤0

Let us try to calculate MSE(Cp,θ) when $\underline{\theta=0}$. Define,

$$U=V\Sigma W' \text{ (s.v.d.)} \quad (4.67)$$

so,

$$U^{+}=W\tilde{\Sigma}V' \quad (4.68)$$

Then we see that MSE(Cp,0) becomes.

$$E\{ e'V\tilde{\Sigma}'W'W\tilde{\Sigma}V'e \} \quad (4.69)$$

under the conditon that $e'V\Sigma W'W\tilde{\Sigma}V'e > 2d\sigma^2$

Define $\Sigma\tilde{\Sigma} = DD'$, with $D=[\ I_d\ 0]'$ is a n*d matrix, and $\tilde{\Sigma}'\tilde{\Sigma}=DTD'$ where T is a d*d positive definite diagonal matrix. Then with,

$$n=D'V'e \ \epsilon\ N(0,\sigma^2 I_d) \quad (4.70)$$

we get,

$$MSE(Cp,0) = E\{ n'Tn \} \quad \text{for } n'n > 2d\sigma^2 \quad (4.71)$$

Similar for MSE(Np,0) we get

$$MSE(Np,0) = E\{ n'Tn \} \quad \text{for } n'Tn > 2\sigma^2 tr\{ T \} \quad (4.72)$$

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 ellips formed by $n'Tn = 2\sigma^2\ tr\{ T \}$ is greater than that of the circle $n'n = 2d\sigma^2$, the calculation of MSE(Np,0) is performed over an area where the values of $F(n)=(a*n_1^2 + b*n_2^2)*p(n_1)*p(n_2)$ (here diag(T)=(a b)) are greater than those of the area in calculation of MSE(Cp,0). To illustate 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 MSE(Cp,0) ≤ MSE(Np,0). So in this ad hoc derivation we saw that MSE(Cp,0) ≤ MSE(Np,0)

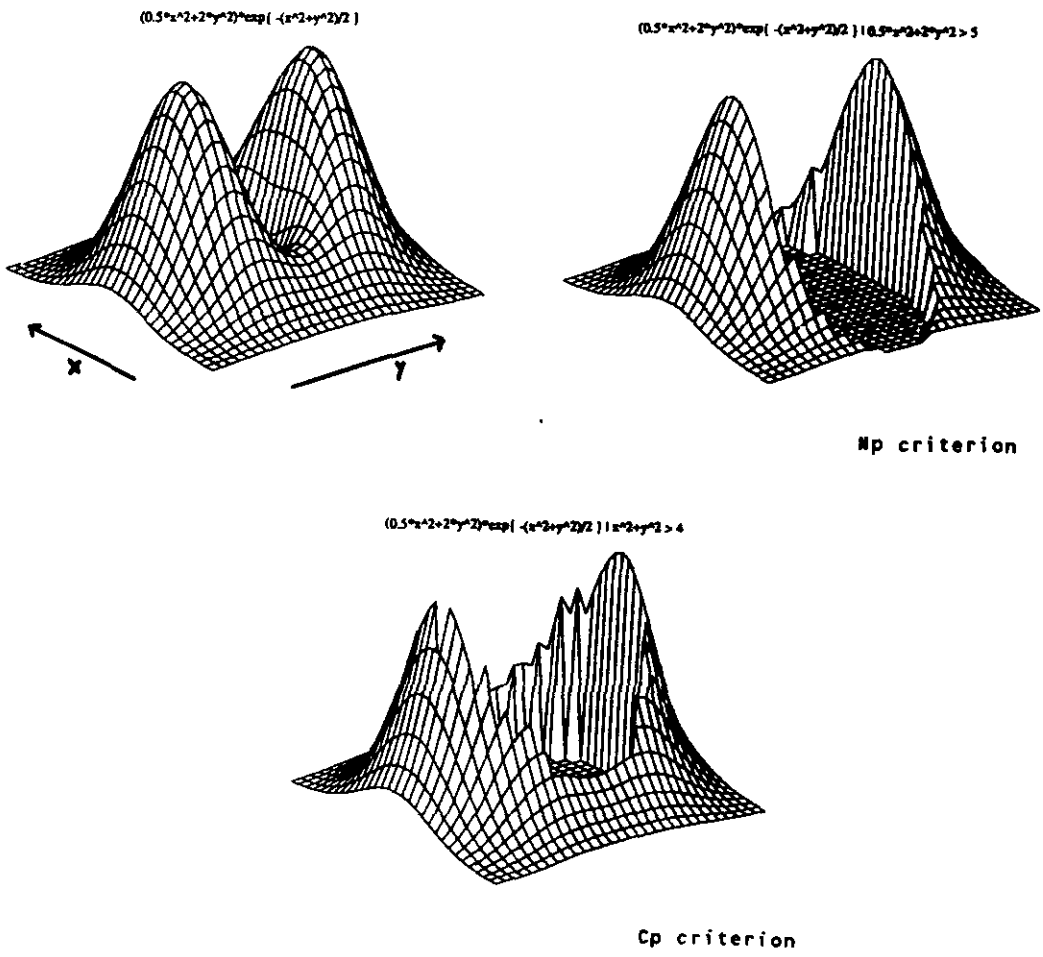

Figure 3: Some 3-dimensional plots of the risk computation if $\theta=0$

Since these results only hold for $\theta=0$ we expect that the calculations of the risk for $\theta<>0$ 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 MSE(θ) for the Mp and the 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 θ. Secondly I simulated the MSE(θ) for the Mp and the 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 θ.

SIMULATIONS-1:

First I simulated the MSE as defined in (4.65) and (4.66) as a function of the length of θ, assuming a fixed orientation for θ. 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 Np and the Cp criterion decide what model structure to take. From that chosen model structure I calculated the resulting error between the actual θ and the estimated θ. 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 θ.

The preceding procedure has been done for two different orientations of θ. 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 θ).

Values:

$$U = 10/3 * \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}$$

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

	case 1	case 2
θ	$a*[1 \quad 1 \quad 1]'$	$a\sqrt{1.5}*[1 \quad 0 \quad -1]'$
$x'x$	a^2*189	$a^2*33.4$

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 θ. This is done for the Np criterion, the Cp 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 Np and Cp criterion give better results than the FM. The difference between Np and Cp is relative small in this region. But we see that for a small value of $\theta'\theta$ the Cp criterion is slightly better than the Np criterion and when $\theta'\theta$ is increased the Np criterion becomes slightly better. However, if the squared length is further increased (>0.08) we see that both the Cp as the Np criterion give worse results than the FM (compare it with

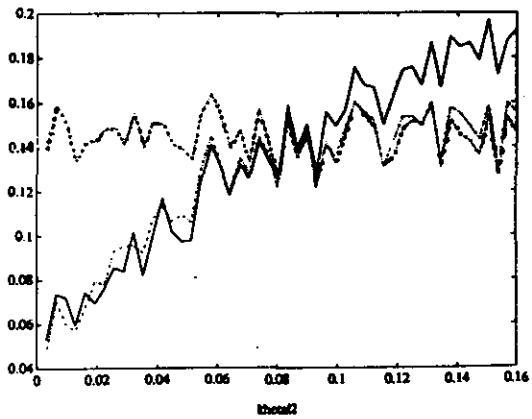

Figure 4a: $0 < \theta'\theta < 0.16$

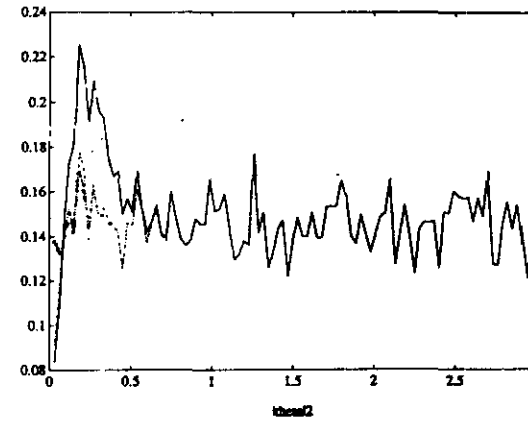

Figure 4b: $0 < \theta'\theta < 3$

Figure 4: Case 1, MSE(θ) for Np(-) Cp(-.) FM(..)

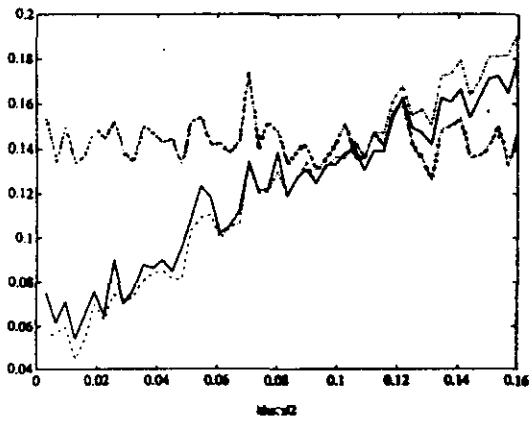

Figure 5a: $0 < \theta'\theta < 0.16$

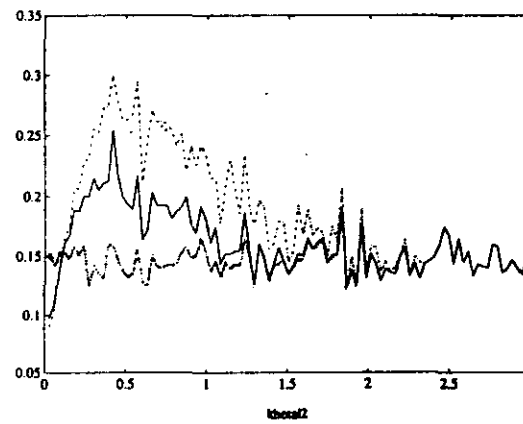

Figure 5b: $0 < \theta'\theta < 3$

Figure 5: Case 2, MSE(θ) for Np(-) Cp(-.) FM(..)

the characteristic in figure 1). But we see that in this region the Np criterion gives much worse results than the Cp criterion.
To have an impression of the goodness of the 200 simulations per value of θ, 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 Cp criterion for small squared lengths of θ. However when $\theta'\theta$ is increased the Np 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 $\theta'\theta$ the Cp criterion gives better results than the Np criterion. But if $\theta'\theta$ is increased it depends on the orientation of θ which criterion will be better. If $(x'x)/(\theta'\theta)$ is 'small' the Np criterion turns out to be better, if not the Cp citerion 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 $\theta'\theta$ 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 contructed several 16*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 θ. Per orientation I considered 21 different lengths of the actual θ. With a particular U and θ 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 θ was calculated. The simulated MSE(θ) was approximated by 100 of such simulations.

Besides the Cp, Np and the FM (=always chosing 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 θ (thus I assumed that θ was known). The model structure with the smallest squared difference between the actual and estimated θ (=L(θ)) 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 θ 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 where defined as follows. (see table 1 and figure 6)

Table 1: Impulse responses	
impulse response A	impulse response B
$h(i)=a*exp(-0.25*i)$	$h(i)=a*(-1)^i exp(-0.25*i)$

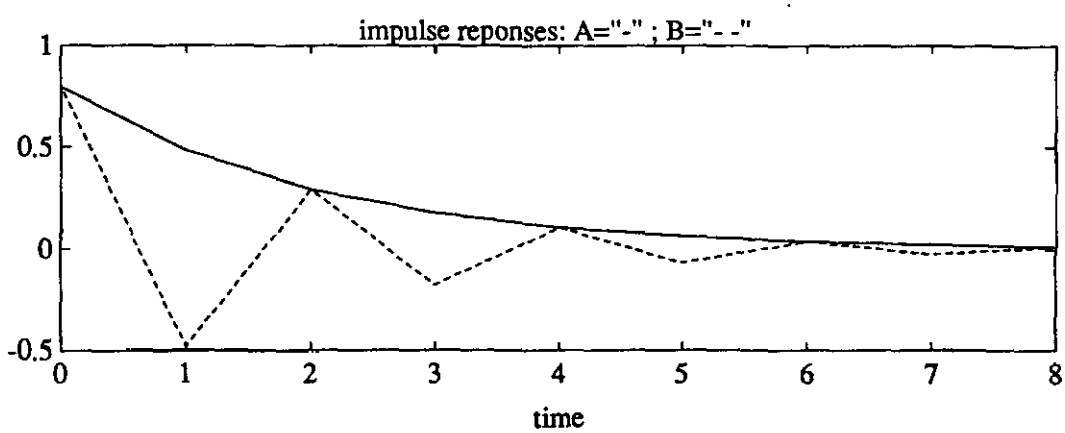

Figure 6: The two orientations of the impulse response

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 elememt of it and finally I multiplied the matrix in such a way that I would obtain the desirable signal to noise ratio (= $\text{sum}((U_{ij})^2)/n\sigma^2$). The variance of the noise (σ^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				
	case 1	case 2	case 3	case 4
$tr((U'U)^{-1})$	7.565	11.40	153.0	331.6
condition number	2.868	4.658	38.80	50.55
$(x'x)_A/(x'x)_B$	1.776	0.8648	11.92	14.30

RESULTS:

For each matrix U I made a plot of the MSE(θ) (as a function of θ'θ) for the two orientations of θ. As already mentioned these plots were made of 21 different values of θ'θ and each value of MSE(θ) was approximated by 100 simulations. For the results of case 1-4 see figure 7a,7b-10a,10b.
In each figure the MSE(θ) was simulated for the Np,Cp,FM,L(x), and L(θ) criterion. The MSE(θ) in the L(θ) criterion is the lowerbound for the MSE(θ) which can be achieved with a criterion.

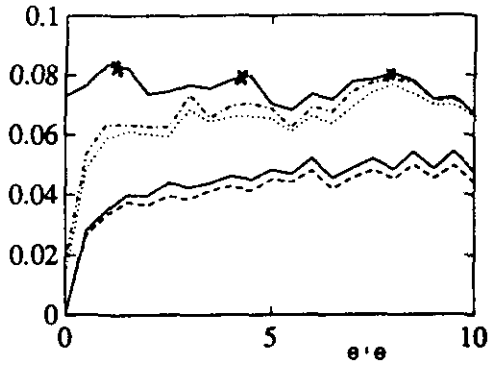

Figure 7a: impulse response A

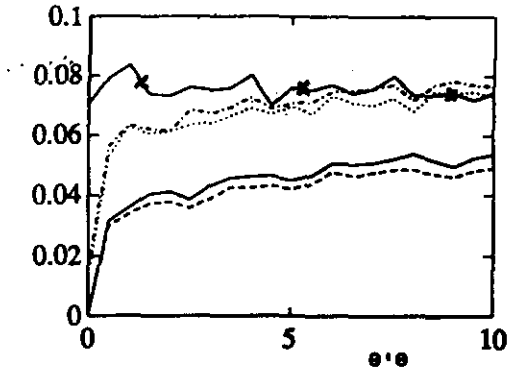

Figure 7b: impulse response B

Figure 7: Case 1, MSE(θ) for L(θ)(- - -) L(x)(——) Cp(...) Np(-.-.) FM(—✕—)

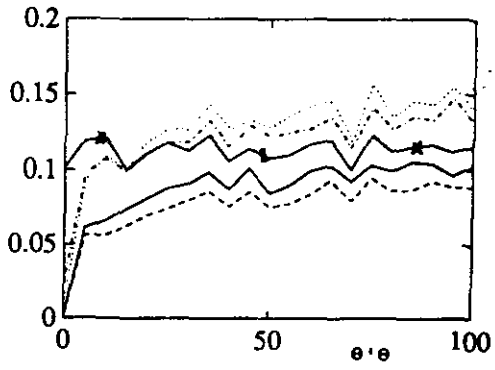

Figure 8a: impulse response A

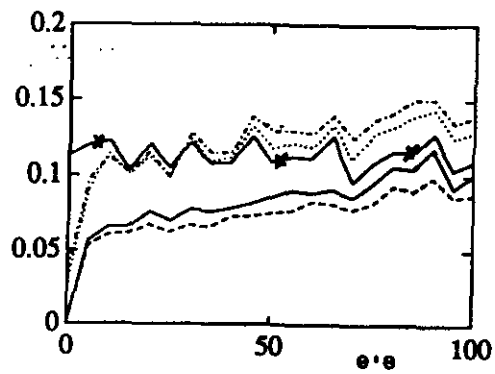

Figure 8b: impulse response B

Figure 8: Case 2, MSE(θ) for L(θ)(- - -) L(x)(——) Cp(...) Np(-.-.) FM(—✕—)

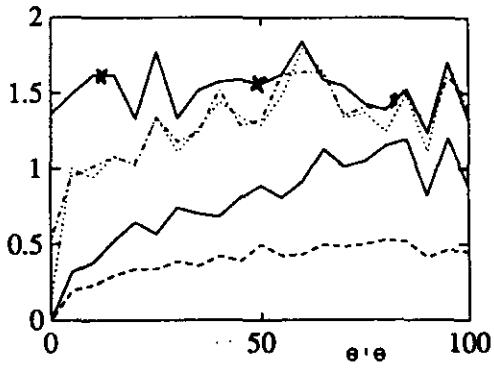

Figure 9a: impulse response A

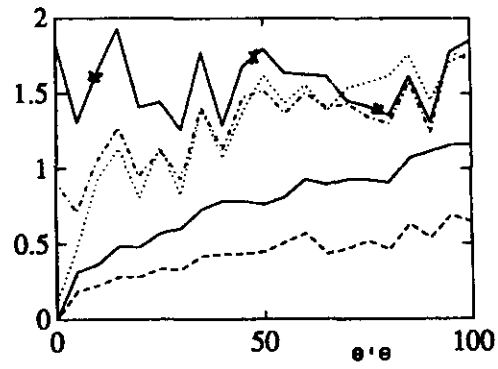

Figure 9b: impulse response B

Figure 9: Case 3, MSE(θ) for L(θ)(- - -) L(x)(——) Cp(...) Np(-.-.) FM(—✱—)

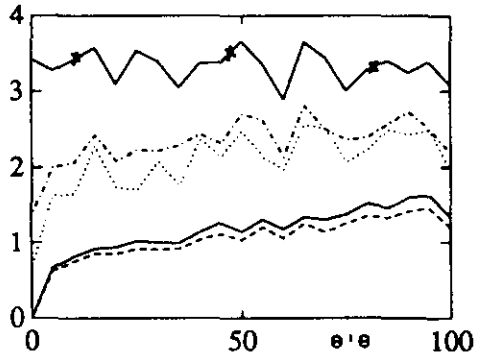

Figure 10a: impulse response A

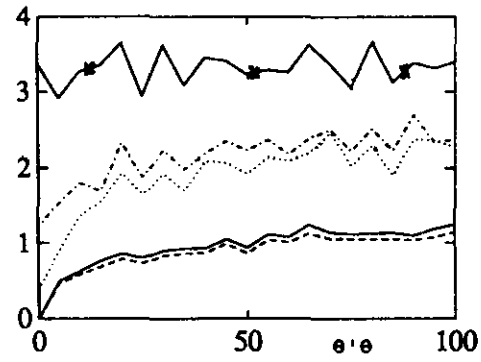

Figure 10b: impulse response B

Figure 10: Case 4, MSE(θ) for L(θ)(- - -) L(x)(——) Cp(...) Np(-.-.) FM(—✱—)

In figure 7a-7b we have a matrix with a low condition number (=2.868) and we see that for both the impulse reponse 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 θ'θ the Cp criterion gives better results for both the actual impulse response A and B, but if θ'θ 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 relative small. In this case where when θ'θ is large not many parameters can be deleted. So we are in fact chosing 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

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(θ) 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 θ'θ the Cp criterion gives better results and for large θ'θ it depends on the actual impulse response which criterion is better. In this case we also see a large variation between the L(θ) 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(θ) and L(x) criterion.

From these figures we see that for small values of θ'θ the Cp criterion gives better results than the Np criterion. If θ'θ 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 ot 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,

$$2^d \tag{4.73}$$

possible subsets to examine. The subset that minimizes a certain criterion (for example Cp 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 a 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,

$$F = \frac{RSSp - RSSp+i}{RSSp+i/(n-p-1)} \in F(1,d-p-1) \qquad (4.74)$$

where RSSp+i denotes the RSS of the model where the i-th variable is included. Then the i-th variable is included if

$$F > Fin = F^{\alpha}(1,d-p-1) \qquad (4.75)$$

Of course we should again choose a value for α. 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 residuls sum of squares is chosen. Then
the i-th variable is deleted from the model if,

$$F = \frac{RSSp-i - RSSp}{RSSp/(n-p)} < Fin \qquad (4.76)$$

Where RSSp-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 wich the U matrix is orthogonal and the variance of the noise is known. It is assumed that,

$$U^{t}U=l*I_{d} \tag{4.77}$$

Thus we can see that,

$$\hat{\theta} \in N(\theta,\sigma^{2}I_{d}/l) \tag{4.78}$$

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,

$$GAIC = \min RSSp + a^{2}*\sigma^{2} \tag{4.79}$$

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

$$\hat{\theta}_{i}^{t}\hat{\theta}_{i} \quad > a^{2}*\sigma^{2}/k = a^{2}*s^{2} \tag{4.80}$$

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,

$$MSE(\theta)= \sum_{i=1}^{d} E(\hat{\theta}_{i} - \theta_{i})^{2} = \sum_{i=1}^{d} MSE(\theta_{i}) \tag{4.81}$$

Let $MSE(\theta_{i})$ denote the mean square error of the i-th variable. Since we don't known what the exact value of θ_{i} is we leave it as a vaiable. Because of the normal distribution of $\hat{\theta}_{i}$ we can calculate $MSE(\theta_{i})$ as follows,

$$MSE(\theta_{i})=$$

$$\int_{-as}^{as} \theta_{i}^{2}\, p(\hat{\theta}_{i}-\theta_{i})\, d\hat{\theta}_{i} + s^{2} - \int_{-as}^{as} (\hat{\theta}_{i}-\theta_{i})^{2}\, p(\hat{\theta}_{i}-\theta_{i})d\hat{\theta}_{i} =$$

$$(\theta_{i}^{2}-s^{2})\{ F(a)-F(-a)\} + s^{2} +$$

$$(as-\theta_{i})s^{2}p(as-\theta_{i}) + (as+\theta_{i})s^{2}p(a\sigma+\theta_{i}) \tag{4.82}$$

where,

$$p(t) = 1/\sqrt{(2\pi s^2)}\ \exp(\ -t^2/(2s^2)\ ) \tag{4.83}$$

$$f(t) = \int_{-\infty}^{t} p(r)\ dr \tag{4.84}$$

When we plot this $MSE(\theta_i)$ versus θ_i in a figure we get the following results (see figure 11).

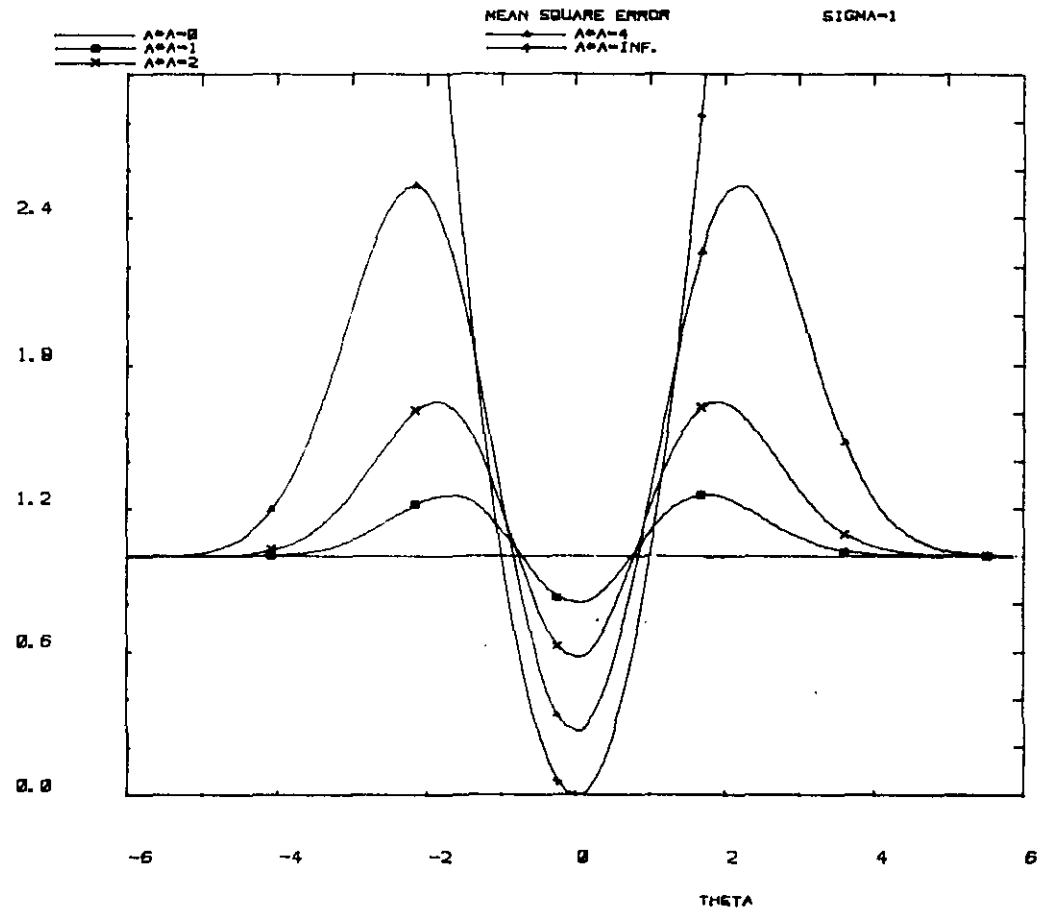

Figure 11: $MSE(\theta_i)$ versus $|\theta_i|^2$ for $a^2=0$, $a^2=1$, $a^2=2$, $a^2=4$, $a^2=$infinity

In this plot it is assumed that s^2 is equal to one, but this is not a severe restriction since the characteristics remain the same for a general s^2. In figure 11, $MSE(\theta_i)$ 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 α 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 θ_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 chaper 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 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 θ_i we get,

$$E\{ \hat{\theta}_i - \theta_i \} = \int_{as}^{\infty} (\hat{\theta}-\theta_i)\, p(\hat{\theta}_i-\theta_i)\, d\hat{\theta}_i + \int_{-\infty}^{-as} (\hat{\theta}_i-\theta i)\, p(\hat{\theta}_i-\theta_i) d\hat{\theta}_i$$

$$= s^2\{ p(as+\theta_i) - p(-as+\theta_i) \} \qquad (4.85)$$

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 ('84) considered three types of bias that occures 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 omision 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. Competion 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 chaper 3.8 I already discussed the orthonormal canonical form of the general linear regression model (3.130-3.132). In this chaper I will present another canonical form which will be used for explanatory purposes. Consider the general model as given by,

$$y = U\theta + e \tag{5.1}$$

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

$$U'U = WTW' \tag{5.2}$$

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

$$Z = U*W \tag{5.3}$$

$$\beta = W'\theta \tag{5.4}$$

Then we have the transformed model,

$$y = Z\beta + e \tag{5.5}$$

Where,

$$\hat{\beta} = Z^{+}y = T^{-1} Z'y = W'\hat{\theta} \quad \epsilon N(\beta,\sigma^2 T^{-1}) \tag{5.6}$$

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

$$E\{ (\hat{\beta}-\beta)'(\hat{\beta}-\beta) \} = E\{ (\hat{\theta}-\theta)'(\hat{\theta}-\theta) \} = MSE(\theta) \tag{5.7}$$

Thus a comparision between two estimator according to the MSE(θ) function can also be done in the β space.

5.2 JAMES STEIN ESTIMATORS

From chapter 3.8 we know that the positive James-Stein estimator,

$$\tilde{\theta} = \max(0, Sf) * \hat{\theta} \tag{5.8}$$

with

$$Sf = [1 - \{c/(v+2)\} v^* \hat{\sigma}^2 / (\hat{x}'\hat{x})]$$

$$v = (n-d) \ , \ \hat{\sigma}^2 = RSSd/(n-d), \quad d \geq 3$$

will be uniformly better than $\hat{\theta}$ according to the loss function MSE(x) if,

$$0 < c < 2(d-2) \tag{5.9}$$

But how does this estimator perform under MSE(θ). Judge and Bock ('78) proved that this estimator will be uniformly better than the least squares estimator if,

$$0 < c < 2\{ tr\{ (U'U)^{-1} \}/tmax - 2 \} \tag{5.10}$$

tmax is the largest eigenvalue of $(U'U)^{-1} = (1/\sigma_d)^2$

provided that $tr\{ (U'U)^{-1} \} > 2tmax$. This means that we can always improve our estimator in MSE(x) sense and in some cases in MSE(θ) sense.

5.3 PRINCIPAL COMPONENT SELECTION

Let us consider the canonical form as decribed in (5.5),

$$y = Z\beta + e \tag{5.11}$$

The columns of Z, $z_1 .. z_d$ 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,

$$MSE(\theta) = \sigma^2 tr\{ (U'U)^{-1} \} = \sigma^2 \ tr\{ T^{-1} \} \tag{5.12}$$

small eigenvalues in T have a large contribution in the variance of $\hat{\theta}$. 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 Z1, to be retained, and Z2, to be deleted. Deleting of components in Z_2 means that the parameters $\beta 2$ (β is also partitioned into two parts) have implicitly been set to zero. The remaining parameters in β_1 are estimated with the least squares estimator. So,

$$\hat{\beta}_1 = Z_1^+ y \qquad \text{and} \qquad \hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} Z_1^+ y \\ 0 \end{bmatrix} \tag{5.13}$$

This means that back in the original parameter space,

$$\hat{\theta} = W \hat{\beta} = [\, W_1 \; W_2 \,] \hat{\beta} = W_1 Z_1^{+} y \tag{5.14}$$

This estimator has the properties that, when $W_2'\theta$ is indeed equal to 0, it is unbiased and has a mean square error of $\sigma^2 tr\{ T1^{-1} \}$ where $T1 = Z_1'Z_1$, 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 estinator has a large variance when $U'U$ has small eigenvalues. As a solution to this problem Hoerl and Kennard ('70) 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,

$$(U'U+k^*I_d)\,\tilde{\theta} = U'y \tag{5.15}$$

or

$$\tilde{\theta} = (U'U+k^*I_d)^{-1}U'y \tag{5.16}$$

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

$$\begin{aligned}\tilde{\theta} &= (U'U+k^*I_d)^{-1}U'\,y \\ &= (U'U+k^*I_d)^{-1}(U'U)\hat{\theta} \;=\; Q\hat{\theta}\end{aligned} \tag{5.17}$$

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

$$MSE(\tilde{\theta}) = \sigma^2 tr\{\, Q(U'U)^{-1}Q \,\} + k^{*2}\theta'(U'U+k^*I_d)^{-2}\theta \tag{5.18}$$

$$= \sigma^2 \sum_{i=1}^{d} t_i/(t_i+k)^2 + k^{*2}\theta'(U'U+k^*I_d)^{-2}\theta \tag{5.19}$$

where t_i, $i=1..d$, are the eigenvalues of $U'U$. 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,

$$k < 2\sigma^2/\beta\text{-max} \tag{5.20}$$

where β-max is the largest element of $\beta = W'\theta$. (see 5.4). Theobald ('74) showed that for a general weighted squared error, $E\{ (\hat{\theta}-\theta)'W(\hat{\theta}-\theta) \}$ where W is some positive definite matrix, a necessary condition is that,

$$k < 2\sigma^2/\theta'\theta \tag{5.21}$$

The existance 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 βmax or θ 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 $\tilde{\theta}$ 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

$$C_k = RSSk/2\hat{\sigma}^2 \quad -n + 2 + 2tr(U(U'U+k*I_d)^{-1}U' \tag{5.22}$$

where RSSk is the residuals sum of squares as a function of k. Hoerl and Baldwin ('75) suggested to take,

$$k = d*\hat{\sigma}^2/\hat{\theta}'\hat{\theta} \tag{5.23}$$

Here k is the sample analog of $d*\sigma^2/\theta'\theta$, 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 θ and σ^2. However if we leave the ridge estimators of the form,

$$\tilde{\theta} = (U'U + k*I_d)^{-1}U'y = (I_d + k*(U'U)^{-1})^{-1}\hat{\theta} \tag{5.24}$$

and consider the more general form,

$$\tilde{\theta} = (I_d + k*C)^{-1}\hat{\theta} \tag{5.25}$$

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,

$$\tilde{\theta} = [I_d + asBU'U/(\hat{\theta}'U'U\hat{\theta} + gs + h)]^{-1}*\hat{\theta} \tag{5.26}$$

with,

$$s=(y-U\hat{\theta})'(y-U\hat{\theta})=RSSd$$

$$0 \le a \le 2(d-2)/((n-d+2)*tmax)$$

tmax is the largest eigenvalue of $[B^{-1}U'U]$

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

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

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

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

$$MSE(\tilde{\theta},\theta) \le MSE(\hat{\theta},\theta)$$

when the conditions given in 5.27 are fulfilled.

6. CONCLUSIONS

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(Θ) 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(Θ), it could not prove its superiority, in simulated test cases, to the Cp criterion (in MSE(Θ) 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, uniformely better (in MSE(Θ) sense) than the least squares estimator.

7. REFERENCES

Akaike, H.
Fitting autoregressive models for prediction.
Ann. Inst. Statist. Math, vol. 21 (1969), p. 243-247.

Akaike, H.
Information theory and the extension of the maximum likelihood principle.
In: Proc. 2nd Int. Symp. on Information Theory; Tsahkador, Armenia, 2-8 September 1971.
Ed. by B.N. Petrov and F. Csaki.
Budapest: Akademiai Kiado, p. 267-281, (1973)

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.
Technometrics, vol. 16 (1974), no. 1, p. 125-127.

Amemiya, T.
Selection of regressors.
Inter. Econ. Review, vol. 21 (1980), no. 2, p. 331-354.

Arnold, S.F.
The theory of linear models and multivariate analysis.
New York: John Wiley & Sons, Inc (1981)

Draper, N.R & Van Nostrand, R.C.
Ridge regression and James-Stein Estimation: Review and Comments.
Technometrics, vol. 21 (1979), no. 4, p. 451-465.

Hocking, R.R.
The analysis and selection of variables in linear regression.
Biometrics,, vol. 32 (1976), p. 1-49.

Hoerl, A.E. & Kennard, R.W.
Ridge regression: Biased estimation of nonorthogonal problems.
Technometrics, vol. 12 (1970), no. 1, p. 55-67.

Hoerl, A.E. & Baldwin, K.F
Bounds on minimum mean square error in ridge regression.
Commun. Statist. Theor. meth., vol A7 (1978), p. 1209-1218.

James, W. & Stein, C.
Estimation with quadratic loss.
In: Proc. 4th Berkeley Symp. ,vol I,
Ed. by J. Neyman.
Berkeley & Los Angeles: University of California Press, p.361-379 (1961)

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.
New York: John Wiley & Sons, Inc (1980)

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
J.R. Statist. Soc. A, vol.147 (1984), p. 389-425

Nishii, R.
Asymptotic properties of criteria for selection of variables in multiple regression.
The Annals of Statistics, vol. 12 (1984), no. 2, p. 758-765

Ponomarenko, M.F.
Information theory and identification.
Dept. of Electrical Engineering, Eindhoven University of Technology, 1981.
EUT Report 81-E-122

Schwarz, G.
Estimating the dimension of a model.
Annals of Statistics, vol. 6 (1978), p. 461-464.

Sclove, S.L.
Improved estimation for coefficients in linear regressions.
American Statistical Association Journal, vol. 63 (1968), p.596-606.

Sclove, S.L. & Morris, C. & Radhakrishanan, R.
Non-optimality of preliminary-test Estimators for the mean of a multivariate normal distribution.
Ann. Math. Statist., vol. 43 (1972), p.1481-1490

Shibata, R.
An optimal selection of regression variable
Biometrika, vol. 68 (1981), no. 1, p45-54.

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.
Ann. Math. Statist., vol. 26 (1955), p. 518-522.

Stoica, P. & Eykhoff, P. & Janssen, P. & Söderström, T.
Model-structure selection by cross-validation
Int. J. Control, vol. 43 (1986), p. 1841-1878.

Strawderman, W.E.
Minimax adaptive generalized ridge regression estimators.
Journal of the American Statistical association, vol. 73 (1978), p. 401-414.

Theobold, C.M.
Generalizations of mean square error applied to ridge regression.
J.R. Statist. Soc. B, vol. 36 (1974), p. 103-106

Thompson, M.L.
Selection of variables in multiple regression:
Part 1. A review and evaluation.
Inter. Statist. Review, vol. 46 (1978), p. 1-19.
Part 2. Chosen procedures, computations and examples.
Inter. statist. review, vol. 46 (1978), p. 129-146.

APPENDIX A

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

ALGEBRA

Suppose,

$U = [\, U_1 \; U_2 \,]$ = n*d matrix and has full rank
U_1 = p*n matrix
U_2 = (d-p)*n matrix

Then is,

$$U'U = \begin{bmatrix} U_1'U_1 & U_1'U_2 \\ U_2'U_1 & U_2'U_2 \end{bmatrix} \tag{A-1}$$

and

$$(U'U)^{-1} = \begin{bmatrix} (U_1'U_1)^{-1}+(U_1^+U_2)B(U_1^+U_2)' & -(U_1^+U_2B) \\ -(U_1^+U_2B)' & B \end{bmatrix} \tag{A-2}$$

with,

$$B=(U_2'(I_n-U_1U_1^+)U_2)^{-1} \quad \text{and} \quad U_1^+=(U_1'U_1)^{-1}U_1'$$

So with $U^+=(U'U)^{-1}U'$ we get,

$$U^+ = \begin{bmatrix} T_1 \\ T_2 \end{bmatrix} = \begin{bmatrix} U_1^+ - U_1^+U_2T_2 \\ T_2 \end{bmatrix} \tag{A-3}$$

where,

T_1 = p*n matrix
$T_2 = A^+$ is a (d-p)*n matrix with $A=(I_n-U_1U_1^+)U_2$

We can see that,

$$UU^+ = U_1U_1^+ + AA^+ \tag{A-4}$$

Let $P=UU^+$ than we see that P is an orthogonal projector since the following conditions hold:

$$P^2 = P \quad \text{and} \quad P' = P$$

Let P be the orthogonal projector on a space spanned by the columns of U, then I-P is the orthogonal projector on the null space of the matrix U'.

POSITIVE (SEMI)DEFINITE

A matrix H is said to be positive definite if

$$\text{for all } x \neq 0, \quad x'Hx > 0 \quad \text{and} \quad H' = H \tag{A-5}$$

A matrix H is said to be positive semidefinite if,

$$\text{for all } x \neq 0, \quad x'Hx \geq 0 \quad \text{and} \quad H' = H \tag{A-6}$$

If H is p.s.d then $T'HT$ is also p.s.d (A-7)

If H is p.d then $T'HT$ is also p.d. provided that for T holds that $Tx=0 \Longrightarrow x=0$. Otherwise it is p.s.d. (A-8)

DERIVATIVES

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

$$\delta L / \delta\theta = \begin{bmatrix} \delta L / \delta\theta_1 \\ \vdots \\ \delta L / \delta\theta_d \end{bmatrix} \tag{A-9}$$

$$\delta^2 L / \delta\theta\,\delta\theta' = \begin{bmatrix} \delta^2 L / \delta\theta_1\,\delta\theta_1 & \cdots & \delta^2 L / \delta\theta_d\,\delta\theta_1 \\ \vdots & & \vdots \\ \delta^2 L / \delta\theta 1\,\delta\theta_d & \cdots & \delta^2 L / \delta\theta_d\,\delta\theta_d \end{bmatrix} \tag{A-10}$$

DISTRIBUTIONS

We say that x_i is normally distributed with mean μ_i and variance σ^2 if x_i is a continuous random variable with density,

$$p(x_i) = 1/\sqrt{(2\pi\sigma^2)}\ \exp\{ -(x_i-\mu_i)^2/(2\sigma^2) \} \tag{A-11}$$

and we denote this as,

$$x_i \in N(\mu_i, \sigma^2) \tag{A-12}$$

Let x be a random vector with,

$$E\{ x \} = \mu \tag{A-13}$$

and,

$$E\{ (x-\mu)(x-\mu)' \} = C \qquad (A\text{-}14)$$

Then we say that the vector x is normally distributed if,

$$p(x) = (2\pi)^{-n/2} * 1/(\det(C))^{\frac{1}{2}} * \exp[-\tfrac{1}{2}(x-\mu)' C^{-1}(x-\mu)] \qquad (A\text{-}15)$$

where det(C) is the determinant of C. We denote this as,

$$x \in N(\mu, C) \qquad (A\text{-}16)$$

Let r be a linear combination x, thus

$$r = Qx \qquad (A\text{-}17)$$

Where the matrix Q denotes the linear transformation. Then r is normally distributed with mean $Q\mu$ and covariance matrix QCQ'. Thus,

$$r \in N(Q\mu, QCQ') \qquad (A\text{-}18)$$

Let $x_1, \ldots x_n$ be independent random variables such that $x_i \in N(0,1)$. Define,

$$y = \sum_{i=1}^{n} x_i^2 \qquad (A\text{-}19)$$

then we say that y is distributed as a Chi^2 distribution with n degrees of freedom. Thus,

$$y \in Chi^2(n) \qquad (A\text{-}20)$$

Let z be distributed as a Chi^2 distribution with m degrees of freedom. Then we say that,

$$v = \frac{y/n}{z/m} \qquad (A\text{-}21)$$

is distributed as an F distribution with n and m degrees of freedom and we denote this as,

$$v \in F(n,m) \qquad (A\text{-}22)$$

APPENDIX B

This appendix contains the calculation of the risk matrix and some related topics.

Let us now assume that the following relation holds,

$$y = U\theta + e = U_1\theta_1 + U_2\theta_2 + e \qquad e \in N(0,\sigma^2 I_n)$$

Where U has full rank. Then with,

$$\hat{\theta}_p = \begin{bmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{bmatrix} = \begin{bmatrix} U_1^+ y \\ 0 \end{bmatrix} = \begin{bmatrix} \theta_1 + U_1^+ U_2\theta_2 + U_1^+ e \\ 0 \end{bmatrix}$$

we get that $E\{ (\hat{\theta}_p-\theta)(\hat{\theta}_p-\theta)' \}$ is equal to

$$\begin{bmatrix} \sigma^2(U_1'U_1)^{-1}+(U_1^+U_2\theta_2)(U_1^+U_2\theta_2)' & -(U_1^+U_2\theta_2)\theta_2' \\ -\theta_2(U_1^+U_2\theta_2)' & \theta_2\theta_2' \end{bmatrix} \qquad (B\text{-}1)$$

Then with $\hat{\theta} = U^+y$ we get that $E\{ (\hat{\theta}-\theta)(\hat{\theta}-\theta)' \} = \sigma^2(U'U)^{-1}$. So with relation (A-2) we see that,

$$E\{ (\hat{\theta}-\theta)(\hat{\theta}-\theta)' \} - E\{ (\hat{\theta}_p-\theta)(\hat{\theta}_p-\theta)' \} =$$

$$\begin{bmatrix} U_1^+U_2(B\sigma^2-\theta_2\theta_2')(U_1^+U_2)' & -U_1^+U_2(B\sigma^2-\theta_2\theta_2') \\ -(B\sigma^2-\theta_2\theta_2')(U_1^+U_2)' & (B\sigma^2-\theta_2\theta_2') \end{bmatrix} \qquad (B\text{-}2)$$

If we rewrite (B-2) as,

$$\begin{bmatrix} -U_1^+U_2 \\ I_{(d-p)} \end{bmatrix} [\, B\sigma^2-\theta_2\theta_2' \,] \begin{bmatrix} -U_1^+U_2 \\ I_{(d-p)} \end{bmatrix}'$$

then we see that if $(B\sigma^2-\theta_2\theta_2')$ is p.d., the matrix given in (B-2) is at least p.s.d. (see A-8). A necessary condition for $(B\sigma^2-\theta_2\theta_2')$ to be p.d is that B must be p.d. Let us rewrite B as $B=WDD'W'$ then must,

$$I - 1/\sigma^2 * D^{-1}W'\theta_2\theta_2'WD^{-1}$$

be p.d.. The second term has rank 1, and the only nonzero eigenvalue has value $\theta_2'B^{-1}\theta_2/\sigma^2$ (eigenvector = $D^{-1}W'\theta_2$). Thus because of condition (A-6) the following relation holds,

$$(B\sigma^2-\theta_2\theta_2') \text{ is p.d} \quad <==> \quad \theta_2'B^{-1}\theta_2/\sigma^2 < 1$$

APPENDIX C

PROVE THAT,

$$2*I(\theta,\tilde{\theta}) = (\tilde{x}-x)'(\tilde{x}-x) \,/\, \sigma^2$$

Proof:

Let,

$$I(\theta,\Phi) = -\int p(y;\theta)\, \ln\{\, p(y;\Phi)/p(y;\theta)\,\}\, dy \qquad (C\text{-}1)$$

with,

$$p(y;\Phi) = (2\pi\sigma^2)^{-n/2} \exp\{\, -(y-U\Phi)'(y-U\Phi)/2\sigma^2\,\} \qquad (C\text{-}2)$$

Then we have,

$$\delta \ln\{\, p(y;\Phi)\,\}/\,\delta\Phi = -(U'U)\Phi/2\sigma^2 + U'y/2\sigma^2 \qquad (C\text{-}3)$$

$$\delta^2 \ln\{\, p(y;\Phi)\,\}/\,\delta\Phi\,\delta\Phi' = -(U'U)/\sigma^2 \qquad (C\text{-}4)$$

and all higher derivatives are zero. Let θm be the value of Φ that minimizes $I(\theta,\Phi)$ { thus $\theta m=\theta$ }. Then, because of the Taylor expansions series and (C-1)-(C-4), we can write,

$$I(\theta,\tilde{\theta}) = I(\theta,\theta m) + (\tilde{\theta}-\theta m)'\,\delta\, I(\theta,\theta m)\,/\delta\theta m$$

$$+ 1/2*(\tilde{\theta}-\theta m)'*\delta^2\, I(\theta,\theta m)\,/\,\delta\theta m\,\delta\theta m'\,*(\tilde{\theta}-\theta m)$$

$$= 0 + 0 + 1/2*(\tilde{\theta}-\theta m)*\delta^2\, I(\theta,\theta m)/\delta\theta m\,\delta\theta m'*(\tilde{\theta}-\theta m)$$

But since,

$$\delta^2\, I(\theta,\theta m)\,/\delta\theta m\,\delta\theta m' =$$

$$\delta^2 - \int p(y;\theta)\, \ln\{\, p(y;\Phi)/p(y;\theta)\,\}\, dy \quad /\delta\theta m\,\delta\theta m' =$$

$$\int p(y;\theta)\, U'U/\sigma^2\, dy \;= U'U/\sigma^2$$

we have that,

$$2*I(\theta,\tilde{\theta}) = (\tilde{\theta}-\theta)U'U(\tilde{\theta}-\theta)/\sigma^2 = (\tilde{x}-x)'(\tilde{x}-x)/\sigma^2$$

Q.E.D.

APPENDIX D

PROVE THAT,

$$PRESS(p) = \sum_{i=1}^{n} [\, y(i) - u1(i) * U_1(/i)^{+} y(/i) \,]^2 = w_p' D_p^{-2} w_p$$

with,

$U_1(/i)$ = U_1 matrix without the i-th row.
$u1(i)$ = the i-th row of U_1
$y(/i)$ = output vector without the i-th output
$y(i)$ = i-th output

$$w_p = (I_n - UU_1^{+})y$$

D_p = the diagonal matrix whose diagonal elements are those of $(I_n - U_1 U_1^{+})$

Proof:

Let us consider the i-th term of PRESS and let us write,

$$A = [U_1(/i)' \; u1(i)']' = [\, X' \; b' \,]'$$

Where A is just another ordering of the rows of U. Then we can write this i-th term as,

$$PRESS(p,i) = [y(i) - bX^{+}y(/i)]$$

Let us now consider the i-th term of w_p. We can write,

$$w_p(i) = y(i) - bA^{+} \tilde{y}$$

Where $\tilde{y} = (\, y(/i)' \; y(i)' \,)'$ is just another ordening of y. Since,

$$(A'A)^{-1} = (X'X)^{-1} - (X'X)^{-1} b'(1 + b(X'X)^{-1}b')^{-1} b(X'X)^{-1}$$

$$= M \qquad (D\text{-}1)$$

we can write,

$$A^{+} = M \, [\, X' \; b' \,]$$

so,

$$bA^{+} \tilde{y} = bMX'y(/i) + bMb'y(i)$$

and,

$$AA^{+} = [\, X' \; b' \,]' M \, [\, X' \; b' \,]$$

$$(UU^{+})_{ii} = (AA^{+})_{nn} = bMb'$$

Thus the i-th component of $w_p' D_p^{-1} w_p$ can be written as,

$$\frac{y(i) - bA^{+-}y}{(1-bMb')} = \frac{y(i)-bMb'y(i) - bMX'y(/i)}{(1-bMb')}$$

$$= y(i) - \frac{bMX'y(/i)}{(1-bMb')}$$

But since (see D-1),

$$bMX'y(/i) = [bV-bVb'(1+b'Vb)^{-1}bV]X'y(/i)$$

$$(1-bMb') = [1-bVb' + bVb'(1+b'Vb)^{-1} bVb']$$

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,

$$bMX'y(/i)*(1+b'Vb) = bVX'y(/i)$$

$$(1-bMb') *(1+b'Vb) = 1$$

So,

$$\frac{bMX'y(/i)}{(1-bMb')} = bX^{+} y(/i)$$

But this means the i-th component of $w_p'D_p^{-2}w_p$ is equal to the PRESS(p,i). Thus PRESS(p) and $w_p'D^{-2}w_p$ are equal.

Q.E.D.

Eindhoven University of Technology Research Reports
Faculty of Electrical Engineering

ISSN 0167-9708
Coden: TEUEDE

(171) Monnee, P. and M.H.A.J. Herben
MULTIPLE-BEAM GROUNDSTATION REFLECTOR ANTENNA SYSTEM: A preliminary study.
EUT Report 87-E-171. 1987. ISBN 90-6144-171-4

(172) Bastiaans, M.J. and A.H.M. Akkermans
ERROR REDUCTION IN TWO-DIMENSIONAL PULSE-AREA MODULATION, WITH APPLICATION TO COMPUTER-GENERATED TRANSPARENCIES.
EUT Report 87-E-172. 1987. ISBN 90-6144-172-2

(173) Zhu Yu-Cai
ON A BOUND OF THE MODELLING ERRORS OF BLACK-BOX TRANSFER FUNCTION ESTIMATES.
EUT Report 87-E-173. 1987. ISBN 90-6144-173-0

(174) Berkelaar, M.R.C.M. and J.F.M. Theeuwen
TECHNOLOGY MAPPING FROM BOOLEAN EXPRESSIONS TO STANDARD CELLS.
EUT Report 87-E-174. 1987. ISBN 90-6144-174-9

(175) Janssen, P.H.M.
FURTHER RESULTS ON THE McMILLAN DEGREE AND THE KRONECKER INDICES OF ARMA MODELS.
EUT Report 87-E-175. 1987. ISBN 90-6144-175-7

(176) Janssen, P.H.M. and P. Stoica, T. Söderström, P. Eykhoff
MODEL STRUCTURE SELECTION FOR MULTIVARIABLE SYSTEMS BY CROSS-VALIDATION METHODS.
EUT Report 87-E-176. 1987. ISBN 90-6144-176-5

(177) Stefanov, B. and A. Veefkind, L. Zarkova
ARCS IN CESIUM SEEDED NOBLE GASES RESULTING FROM A MAGNETICALLY INDUCED ELECTRIC FIELD.
EUT Report 87-E-177. 1987. ISBN 90-6144-177-3

(178) Janssen, P.H.M. and P. Stoica
ON THE EXPECTATION OF THE PRODUCT OF FOUR MATRIX-VALUED GAUSSIAN RANDOM VARIABLES.
EUT Report 87-E-178. 1987. ISBN 90-6144-178-1

(179) Lieshout, G.J.P. van and L.P.P.P. van Ginneken
GM: A gate matrix layout generator.
EUT Report 87-E-179. 1987. ISBN 90-6144-179-X

(180) Ginneken, L.P.P.P. van
GRIDLESS ROUTING FOR GENERALIZED CELL ASSEMBLIES: Report and user manual.
EUT Report 87-E-180. 1987. ISBN 90-6144-180-3

(181) Bollen, M.H.J. and P.T.M. Vaessen
FREQUENCY SPECTRA FOR ADMITTANCE AND VOLTAGE TRANSFERS MEASURED ON A THREE-PHASE POWER TRANSFORMER.
EUT Report 87-E-181. 1987. ISBN 90-6144-181-1

(182) Zhu Yu-Cai
BLACK-BOX IDENTIFICATION OF MIMO TRANSFER FUNCTIONS: Asymptotic properties of prediction error models.
EUT Report 87-E-182. 1987. ISBN 90-6144-182-X

(183) Zhu Yu-Cai
ON THE BOUNDS OF THE MODELLING ERRORS OF BLACK-BOX MIMO TRANSFER FUNCTION ESTIMATES.
EUT Report 87-E-183. 1987. ISBN 90-6144-183-8

(184) Kadete, H.
ENHANCEMENT OF HEAT TRANSFER BY CORONA WIND.
EUT Report 87-E-184. 1987. ISBN 90-6144-6

(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.
EUT Report 87-E-185. 1987. ISBN 90-6144-185-4

(186) Fu Yanhong
THE INFLUENECE OF CONTACT SURFACE MICROSTRUCTURE ON VACUUM ARC STABILITY AND ARC VOLTAGE.
EUT Report 87-E-186. 1987. ISBN 90-6144-186-2

(187) Kaiser, F. and L. Stok, R. van den Born
DESIGN AND IMPLEMENTATION OF A MODULE LIBRARY TO SUPPORT THE STRUCTURAL SYNTHESIS.
EUT Report 87-E-187. 1987. ISBN 90-6144-187-0

Eindhoven University of Technology Research Reports
Faculty of Electrical Engineering

ISSN 0167-9708
Coden: TEUEDE

(188) Jóźwiak, J.
THE FULL DECOMPOSITION OF SEQUENTIAL MACHINES WITH THE STATE AND OUTPUT BEHAVIOUR REALIZATION.
EUT Report 88-E-188. 1988. ISBN 90-6144-188-9

(189) Pineda de Gyvez, J.
ALWAYS: A system for wafer yield analysis.
EUT Report 88-E-189. 1988. ISBN 90-6144-189-7

(190) Siuzdak, J.
OPTICAL COUPLERS FOR COHERENT OPTICAL PHASE DIVERSITY SYSTEMS.
EUT Report 88-E-190. 1988. ISBN 90-6144-190-0

(191) Bastiaans, M.J.
LOCAL-FREQUENCY DESCRIPTION OF OPTICAL SIGNALS AND SYSTEMS.
EUT Report 88-E-191. 1988. ISBN 90-6144-191-9

(192) Worm, S.C.J.
A MULTI-FREQUENCY ANTENNA SYSTEM FOR PROPAGATION EXPERIMENTS WITH THE OLYMPUS SATELLITE.
EUT Report 88-E-192. 1988. ISBN 90-6144-192-7

(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.
EUT Report 88-E-193. 1988. ISBN 90-6144-193-5

(194) Hosselet, L.M.L.F.
MARTINUS VAN MARUM: A Dutch scientist in a revolutionary time.
EUT Report 88-E-194. 1988. ISBN 90-6144-194-3

(195) Bondarev, V.N.
ON SYSTEM IDENTIFICATION USING PULSE-FREQUENCY MODULATED SIGNALS.
EUT Report 88-E-195. 1988. ISBN 90-6144-195-1

(196) Liu Wen-Jiang, Zhu Yu-Cai and Cai Da-Wei
MODEL BUILDING FOR AN INGOT HEATING PROCESS: Physical modelling approach and identification approach.
EUT Report 88-E-196. 1988. ISBN 90-6144-196-X