Thesis for the Degree of Doctor of Philosophy

Statistical methods for Designed experiments and Spectroscopic data

Tobias Adolfsson

CHALMERS | GÖTEBORG UNIVERSITY

Department of Mathematical Statistics
Chalmers University of Technology and Göteborg University
SE-412 96 Göteborg, Sweden

Göteborg, March 2002

Statistical methods for Designed experiments and
Spectroscopic data
Tobias Adolfsson
ISBN 91-7291-136-0

Doktorsavhandlingar vid Chalmers tekniska högskola
Ny serie nr 1818
ISSN 0346-718X

Department of Mathematical Statistics
Chalmers University of Technology
SE-412 96 Göteborg
Sweden
Telephone +46 (0)31-772 1000

Abstract

This thesis consists of six papers related to saturated orthogonal designs, spectroscopic and high dimension data analysis. The first two papers deals with testing procedures for saturated orthogonal designs. Both the presented methods controls the multiple level of significance. In paper C a regression method is constructed for both multivariate and univariate situations. The correctness of the subspace of regression can be tested using the multiple testing technique constructed in paper A. Paper D gives a method for testing the chemical rank of spectroscopic data. In paper E the effectiveness of high-dimension data gathered in a random way is examined and compared to an optimality criteria. Paper F deals with partial least squares and cross validation. The paper explores, and partly explains, some of the strange behaviour of partial least squares.
Keywords: Saturated orthogonal designs, Multiple inference, Step-down testing, Partial least squares, Principal components regression
2000 Mathematical subject classification:62F03, 62F25, 62H25, 62J05, 62K15

Acknowledgements

I would like to thank the national project CPDC (Chemical Process Design and Control) and ASN (A Stochastic Network) for partly funding this thesis. I would like to thank my supervisor, Prof. Sture Holm, who is a man with great intuition and competence. Supervising is more than scientific cooperation, and I am glad that Sture knows this as well as mathematical statistics.

There is a bunch of people that has helped me making my time as a Phd-student a good time. Some of these people are Dan Mattsson, Ola Engvall, Robert Bohlin, "the ECMI-bunch/lunch", Fredrik Lundin, Kerstin Wiklander, Leif Cardell, Lennart Falk, Sven Järner etc. One person with a special place on this list, and in my heart, is Anna, who stood by me through this work's ups and downs. **Thanks!!!**

Göteborg, February 2002

Tobias Adolfsson

List of papers

The thesis consists of the following papers.

A: Adolfsson T, Variance estimation and multiple inference testing in saturated orthogonal designs.

B: Holm s, Adolfsson T, Mark S, A step-down test for effects in unreplicated factorial trials.

C: Adolfsson T, Using regression function information in determining the same.

D: Adolfsson T, A testing procedure for determining the chemical rank of spectroscopic absorbtion matrices - a heuristic approach.

E: Holm S, Adolfsson T, Efficiency of non-designed investigations.

F: Adolfsson T, Partial least squares and Cross Validation in a statistical perspective.

Chapter 1

Foreword

Mathematical statistics is a broad science. It spans over areas as spatial processes, queue systems , quality control, experimental design and many many more. In these thesis I will try to make contributions in areas such that the methods presented could be put into direct applications. When constructing a method for applications one should make sure that the method is easy to use and can be put into a more or less automated process of application. This is important since the average person applying the method in general is not a statistical expert. Most of the methods presented in the different papers can be more or less automated. The scope of the included papers are not really to describe automated processes of inference in certain situations. The intent is to, as far as possible, describe the theoretical features and underlying assumptions and there consequences. A application interested person reading this should not be alarmed by asymptotic calculations, strange distributions and mystic notations. The general ideas used in these papers are (at least to me) rather simple and straightforward. The assumptions needed for the methods to work are commonly used in various applications and should not state a problem.

Chapter 2

Spectroscopy

I will give a short introduction to the field of spectroscopy since many applications arises from this field. The dimension test method of paper D can be useful in analysis of spectroscopy data. The method of spectroscopy has been used for quite some time in chemistry. I will briefly give a the chemical motivation for using this method. In dilutions of chemical substances one can under certain conditions assume that the absorption of light is directly proportional to the concentration of a substance given the wavelength of the light. This is called **Beer's law** and can be state as follows: let λ denote the wavelength of the light used, let A_λ denote the absorption at the wavelength λ, let ϵ_λ^i be the molar absorptivity of species i at wavelength λ, let b be the optical pathlength and let c_i denote the concentration of the ith species. Let n_{sp} be the number of species present in the sample. Then under certain conditions the relationship

$$A_\lambda = \sum_{i=1}^{n_{sp}} \epsilon_\lambda^i b c_i = \sum_{i=1}^{n_{sp}} a_\lambda^i c_i$$

hold. In the last equality the molar absorptivity and the optical pathlength has been group together. Measurements of n_{sp} species in n_w wavelength and for n_{sa} samples can be collected in a matrix

$$X = \sum_{i=1}^{n_{sp}} c_i a_i' + E = CA + E.$$

We have that $c_i \in \mathbb{R}^{n_{sa} \times 1}$, $a \in \mathbb{R}^{n_w \times 1}$ and $E \sim N_{n_{sa}}(0, \sigma^2 I_{n_w})$. The samples might be derived in different ways. In the more traditional analysis sample usually means different solutions. Another way of deriving the sample is by *in situ* measurements. This means that we start a chemical reaction and measure the spectral absorptivity at different time points and hence the time points constitutes the sample. One such sample consists of measurements of absorption from many wavelengths. The test apparatus scans the wavelengths so fast that for most reactions one sample can be assumed to come from the same time point. One of the greatest advantages with *in situ* spectroscopy is that one basically do not interfere with the reaction which might be the case in other forms of analysis where quenching is needed. This is of great importance for unstable process mixtures, e.g. super-saturated solutions or when reactive intermidietes are present. In chemistry one splits spectroscopy into different groups depending on the wavelengths that are used. This is important to know since the different wavelength area exhibits rather different properties.

2.1 InfraRed (IR) spectroscopy

Usually infrared light is defined as light with wavelength in the range 0.78-1000μm. This corresponds to wavenumbers within 12800-10cm^{-1} which is a more commonly used unit in chemistry. The infrared light is then divided into three sub-regions: Near-IR (NIR) 12800-4000cm^{-1}, mid-IR 4000-200cm^{-1} and far-IR 200-10cm^{-1}. This grouping is made since there are different absorptions in the different wavenumber intervals. The theory of absorption is that of quantum mechanics. Quantum mechanics tells us that energies of vibrational motions are quantised, that is, can only take certain discrete values. When a molecule is hit by electro magnetic radiation energy matching the energy separating two energy states it may absorb the energy and change state. One usually talks of three different types of absorption: **fundamental absorption**, **overtones** and **combination bands**. Fundamental absorption occurs when a molecule absorbs energy to move from one vibrational mode to the state with the closest higher energy. The most common fundamental absorption is from the ground state to the first excited state. Overtones occur when the molecule moves from the ground state to a state with higher energy than the first excited state. When two vibrational modes are change due to absorption of one quantum a combinational band occurs. Since the three intervals has different levels of energy they will be dominated by different absorptions. NIR, which has the highest energy, most absorption is due to overtones and combination bands associated with

C-H, O-H and N-H bonds. In mid-IR, which has lower energy, fundamental transitions dominate even though not all fundamental transitions lay in the mid-IR band. There are also overtones and combination bands in mid-IR but these have lower intensity.

Mid-IR is a very useful analysis tool since it contains large amount of structural information and since nearly all compounds have absorption in this region. This makes it possible to use the technique for identification and determination of molecular structure in organic synthesis and production of pharmaceuticals. When monitoring a process there is a potential for identifying reaction intermediates, products and by-products from the spectra.

One major drawback with mid-IR is that the solvent used for the reaction is frequently not transparent in mid-IR. This means that one needs techniques that uses very short transmission path-lengths or other method like the attenuated total reflection technique (ATR).

2.2 Raman spectroscopy

Just as infrared spectroscopy, Raman spectroscopy is a vibrational technique. The physical principals used is however different. In Raman spectroscopy, the sample is irradiated by a monochromatic laser beam. When the laser beam hits the sample it will be scattered and a part of this scattering will be measured with a spectrometer. The scatterings are split into three types: Rayleight scattering, Stokes scattering and anti-Stoke scattering. The Rayleight scattering is of exactly the same energy as the incident radiation. The Stokes scattering is of lower energy energy than the incident radiation and the anti-Stokes scattering is of higher energy than the incident radiation. The decrease or increase of wavenumber is due to absorbation or emittance of energy quanta from changes in vibrational energy. The Rayleight scattering is the most intense of the the scatterings but contains no relevant information. Hence one measures Stokes or anti-Stokes, or both, scattering. It is more common to measure the Stokes scattering since this has a higher intensity.

Raman scatterings are also proportional to the concentration of the molecule. It is hence possible to analyse the data in the same way as absorbances in transmission spectroscopy. Without going into chemical detail one could say that Raman spectroscopy and IR spectroscopy complementary to each other. This is the case since they measure rather different properties of the molecules (Raman spectroscopy measures change in polarisability while IR spectroscopy has to do with the dipole moment of the molecule). Substances with low absorbance in IR might well have strong effect in Raman spectroscopy. There might however be problems with near-infrared lasers if floroscent substances are present.

Raman spectroscopy has some big practical advantages in process monitoring since laser light easily travels in optical fibers and the laser beam can be focused through a glass wall of a reactor.

2.3 Attenuated total reflection (ATR)

When monitoring chemical manufacturing processes using IR methods one might run into problems. In many chemical processes of interest the concentration of the species of interest is large. This will have the effect that one has to use short optical path-lengths. When using NIR the path-length is often long enough for practical use (0.5-5 mm). The absorptivity in mid-IR however is larger which will give very short path-lengths (10-30 μm). It is practically impossible to get enough flow through the cell in a *in situ* measurement since particles might be present. The same problem arises with other spectroscopic methods.

Attenuated total reflection is a way of getting round these problems. The main idea is to let a beam of some kind be reflected against the sample. This is done by insertion of a probe into the reaction vessel. Needless to say the probe must be made by very corrosion resistant materials.

The theory of ATR is rather difficult and it differs from the other spectroscopy methods. The reflection is never 100% even if the angel of incidence is larger than the critical angel. The data from the ATR experiment will not be linear. One usually uses a log transformation which makes the data reasonably linear.

Chapter 3

Multiple Inference

Multiple inference aims at providing a theory within which detail questions of sub-hypothesis can be answered within a given predescribed risk. Four of the included papers makes reference to multiple inference and in particular step-down methods. We will here state the most basic definitions of multiple inference and pay particular attention closed test with the special case step-down tests.

3.1 Definitions

Let $X \sim F_\theta$ be a stochastic variable where $\theta \in \Omega$. We define a statistical hypothesis as a statement about a parameter in a distribution. Hence, H:$\theta \in \Omega_0 \subset \Omega$ is a hypothesis. We can now define a family of hypothesis by $H_\lambda : \theta \in \Omega$ for $\lambda \in \Lambda$. The multiple level of significance will now be a way of defending a statement concerning a set of hypothesis.

Definition 1 *Let $H_\lambda : \theta \in \Omega$ for $\lambda \in \Lambda$ be a set of hypothesis about F_θ. Let $I_\theta = \{\lambda \in \Lambda : \theta \in \Omega_\lambda\}$ be the set of indexes of true hypothesis in the parameter point θ. Let C_λ be the set where H_λ is rejected and let T be an outcome. The multiple level of significance α is given by*

$$\alpha = \sup_{\theta \in \Omega} \mathbb{P}_\theta\{T \in C_\lambda \text{ for any } \lambda \in I_\theta\}.$$

The multiple level of significance given above is also called *family wise error rate (FWE)*. In words the definition above says that the multiple level of significance is the supremum over the parameter space of the probability of making **any** false rejections of the tested hypotheses.

One great difference between ordinary level of significance and multiple level of significance is the concept of a family of hypothesis. The family of hypotheses should be chosen such that the hypotheses natrually can be discussed in a common framework. Not only the basic hypotheses should be of interest but also their interrelations. Since all rejected hypotheses can be defended within the same risk we can also defend all logical consequences of the rejected hypotheses *within the same risk*.

The definition of multiple level of significance given here is not the only criteria used for error protection. Another criteria that has been proposed is *per family error rate*. The per family error rate is the supremum of the expected number of falsely rejected hypothesis. There is other methods to control multiple error rates but they are of little practical use.

3.2 Closed tests

Let the possible parameters θ in a problem be Ω. As mentioned before a hypothesis H_1 may be written $H_1 : \theta \in \Omega_1 \subset \Omega$. A finite family of hypothesis $\mathcal{H}$ is then given by $\mathcal{H} = \{H_i, i = 1, \ldots, m\}$ where $H_i : \theta \in \Omega_i \subset \Omega$. Let the conjunction $\wedge$ between two hypothesis be the intersection of the hypothesis sets i.e. $H_i \wedge H_j \Leftrightarrow \Omega_i \cap \Omega_j$.

Definition 2 (Closed hypothesis family) *A family of hypothesis $\mathcal{H}$ is said to be* closed under conjunction *if*

$$H_i \in \mathcal{H}, H_j \in \mathcal{H} \Rightarrow H_i \wedge H_j \in \mathcal{H}.$$

Given a closed family of hypothesis we can now construct a closed test.

Definition 3 (Closed tests) *Let $\mathcal{H}$ be a finite family of hypothesis closed under conjunction. Assume that for each $H_i \in \mathcal{H}$ we can define a preliminary individual test of level α. Then a closed test of level α is constructed by rejecting a hypothesis $H_i \in \mathcal{H}$ if and only if the preliminary test rejects H_i itself as well as all hypothesis in $\mathcal{H}$, which imply H_i, that is all hypothesises $H_j : \Omega_j \subset \Omega_i$.*

One can prove that closed tests gives tests with multiple level of significance α. It might seem strange that we get the correct multiple level of significance since we do all tests at level α. It is however natural since we put a lot of restrictions on the testing procedure since all sub-hypothesis of a hypothesis has to be rejected before the hypothesis itself can be rejected.

Definition 4 *If the truth of H_j implies the truth of H_i, then H_i is said to be a* component *of H_j.*

A Hypothesis, which has no component hypothesis is said to be minimal.

A multiple test is said to be coherent *if it has the property that if a hypothesis H_i is a component in the hypothesis H_j and H_j is not rejected then H_i is not rejected.*

A multiple test is said to be consonant *if it has the property whenever some hypothesis is rejected, then some of the minimal hypothesis that it implies are rejected as well.*

By the construction of closed test we see that all such tests must be coherent since a hypothesis in a closed test procedure is finally rejected only if all hypothesis implying it is also rejected. Closed test are however not necessarily consonant.

3.3 Step-down tests

Step-down tests are a special case of closed tests. In step-down tests the minimal hypothesis play a very central role. The construction of the test makes sure that everything is formulated in terms of minimal hypothesises so that step-down tests are automatically consonant. If we have a family of basic hypothesis of size m we, in general, have to make 2^{m-1} preliminary tests of all conjunctions. In a step-down procedure we only have to calculate m test statistics.

The general step-down procedure is constructed as follows. Let $H_1, H_2, \ldots, H_m$ be m hypothesis, which may individually be tested by use of test statistics $T_1, T_2, \ldots, T_m$. Assume that we reject for large values of the T_i:s. Let $I \subseteq \{1, 2, \ldots, m\}$ be the set of indices of the true hypothesis. Assume that the distribution of

$$\max_{i \in I} T_i$$

only depends on I through $|I|$ i.e. the number of indices in I. Let v_r be the $1 - \alpha$ quantile of the distribution of

$$\max_{i \in I} T_i$$

when $|I| = r$. We now order the statistics T_i into $T^{(1)} \leq T^{(2)} \leq \ldots \leq T^{(m)}$ and let $H^{(1)}, H^{(2)}, \ldots, H^{(m)}$ be the correspong hypothesis. A general step-down procedure is then defined by

- If $T^{(m)} < v_m$ make no rejections and stop,
  if $T^{(m)} \geq v_m$ reject $H^{(m)}$ and continue.
- If $T^{(m-1)} < v_{m-1}$ make no further rejections and stop,
  if $T^{(m-1)} \geq v_{m-1}$ reject $H^{(m-1)}$ and continue.

$$\vdots$$

- If $T^{(1)} < v_1$ make no further rejections and stop,
  if $T^{(1)} \geq v_1$ reject $H^{(1)}$ and stop.

Since step-down tests play an important role in the thesis I will, for completion, prove that the general step-down procedure has the correct multiple level of significance.

Theorem 1 *The general step-down test with level α is a multiple test with the multiple level of significance α.*

Proof: Let I be the set of indeces of the true hypotheses and let $|I| = m_0$. If we could perform an ordinary test of the conjunction of all hypotheses with index in I, we would reject this conjunction if $\max_{i \in I} T_i \geq v_{m_0}$. The probability that this happens is at most α. The opposite case, that is when all the T_i:s with $i \in I$ are smaller than v_{m_0}, has probability at least $1 - \alpha$. If this occurs, none of the true hypotheses can be rejected in the step down test. None of the true hypothesis can be rejected in the first $m - m_0$ steps since $i < j$ implies that $v_i \leq v_j$. If the procedure is not stopped already earlier, it will stop in step $m - m_0 + 1$ where then just the hypotheses with index in I remains. **Q.E.D.**

Chapter 4

A short summary of papers

Both papers A and B handles step-down methods for effect testing in saturated factorial experiments. The term effect saturated is here used for experiments where all degrees of freedom are consumed by effect estimation and hence we have no information left for variance estimation. The standard way of dealing with this type of problem is to make a robust estimation of the standard deviation with which one normalises the observations so that a test can be performed. An overwiev of such testing procedures can be found in Haaland and O'Connel (1995) or Berk and Picard (1991). The methods of papers A and B also falls into this framework.

Paper A

In paper A two different variance estimation methods are examined and shoosen between. Both are based on maximum likelihood methods. The asymptotic variance of the estimated variance is calculated. A step-down testing procedure is constructed that has the correct multiple level of significance using the most practical of the two estimates. Individual confidence intervals for the effects are constructed. The method, Censored Variance estimation Modulus Ratio (CVMR), is compared with the GMR method from Voss (1988) which is one of the most important modern methods in the field. By simulations the power is estimated in various situations and the power of the CVMR method is seen to be higher than that of GMR. Tables for critical values and confidence intervals are included.

Paper B

Paper B uses a method based on the graphical methods that where the first to be used for identifying effects in saturated factorial designs. The method follows in the lines of Daniel (1959), Zahn (1975) and others but is constructed so that the multiple level of significance is correct. The testing procedure is also here step-down. The method uses a 50% left coverage band for the absolute values of the effect. The coverage band is calculated by using ordered samples of $U(0,1)$ variables which makes the calculations portable to any other continuous distribution. The left coverage band is then used to estimate the standard variation by fitting the data to the coverage band. A test for presence of true effects is then constructed for the normalised data. Critical values for the test statistica are calculated by simulation and tabulated.

Paper C

A new multivariate and univariate regression method for use in information sparse, badly conditioned, situations is constructed in paper C. The method is constructed to work well in spectrometrical situations in chemistry but will apply well to other situations as well. The method is based on estimating the regression function size on an orthogonalisation of the design matrix. The goal of the method is to find a subspace of the design matrix in which the information about the dependent variable is maximal and solving the problem on this space. The testing CVMR procedure of paper A is show to be applicable in the univariate case and a generalisation of CVMR is shown to be applicable in the multivariate case. The regression method uses information in both dependent and independent variables for solving the regression problem but is still easy to handle theoretically. Prediction properties are examined and the theoretical criteria for including a basevectors are given.

Paper D

In this paper a test for the singular values in a normally distributed matrix is constructed. The construction is proven to have have the correct multiple level of significance given some assumptions. These assumptions are not proven to be correct but rather made likely to be correct. The lack of rigour is due to the fact that it is very difficult to make formal statements about singular values of matrices without very detailed information about the matrices involved. The assumptions however seems to be fulfilled for spektroscopical data.

Paper E

The paper is based on the observation that, if we choose points at random inside a n-dimensional sphere, most observations will be close to the boundary of the sphere if n is large. To spread the observations as much as possible is actually one possible criteria for optimality and the relationship between these are examined by asymptotical calculations. The calculations are made for the uniform distribution and results for a cut-off normal distribution is simulated.

Paper F

This paper consists of a summary of results previously published in Adolfsson (1999).

The thesis put into perspective

The use of factorial designs are well know in the field of statistics. Factorial designs is an important tool to use when constructing efficient experiments since they are information dence. Since experiments are expensive to perform in reality the experiment plans used are often reduced and without replicates. The most commonly used method for deciding on which factors to proceed with in a follow-up study is graphical. The methods given in papers A and B are both easy to use an could well help factorial designs to increase in popularity.

Paper C gives a new method for using in regression situations when ordinary least squares fails. The problem in other regression method to us for these situations is that these methods relies on selection processes that are not reliable. As shown in paper F for instance cross validation might destroy the structure of the problem in such a way that the selection process might fail in a dramatical way. I strongly believe that the user should examine the problem he or she wishes to solve carefully before choosing method. The method presented here solves some types of problems, but not all problems. I belive that there are no universally best method that applies to all problems and that one should be caucus against methods that are marketed as such. The selection process used in paper C is proven to be safe if the assumptions made in the model are correct.

Since there are very few satisfactory methods for choosing the number of components in PCR, I present a heuristic method that could be used for this problem. The correctness of the method is difficult to prove since it is very difficult to make statments of the eigenvalues of sums of matrices. All simulations performed has pointed towards that the theorem holds but the proof of that the assumptions hold is an open problem.

In paper E a result about undesigned experiments is given. The result states that for cut-off distributions of high dimension an undesigned experiment will not be far from optimal. One optimality consept is to place observations with wide spread. We then use the fact that high dimensional cut-off, rotation invariant, distributions will place a lot of mass in the boundary of the distribution. The result shows that in high dimensions randomly spread observation points has suprisingly near-optimal "designs".

Paper F is a shorted down version of Adolfsson (1999). It restates some important features of PLS and cross validation. It has been discovered that PLS, when run with cross validation as a stopping rules, might give poor result when data is sparse and structured. It is also known that estimated confidence intervalls gives very poor emperical coverage probability. These "features" are, at least partly, explained.

Bibliography

T. Adolfsson. Partial least squares and its implementations from a statistical point of view. Technical report, Chalmers University of Technology and Göteborg University, 1999.

K. N. Berk and R. R. Picard. Significans tests for saturated orthogonal arrays. *Journal of Quality Technology*, 1991.

C. Daniel. Use of half-normal plots in interpreting factorial two-level experiments. *Technometrics*, 1:311–341, 1959.

P. Haaland and M. O'Connel. Inference for effect-saturated fractional factorials. *Technometrics*, 1995.

D. Voss. Generalized modulus-ratio tests for analysis of factorial designs with zero degrees of freedom for error. *Commun. Statist.-Theory Meth*, 1988.

D.A. Zahn. An empirical study of the half-normal plot. *Technometrics*, 1975.

Paper A

Variance estimation and multiple inference testing in saturated orthogonal designs

Tobias Adolfsson
Göteborg University and Chalmers University of Technology

17th January 2002

Abstract

This article concerns variance estimation and a step down multiple inference testing procedure in the case of orthogonal designs. The testing procedure is compared with other testing procedures and simulations show a considerable increase of power compared with other methods.

1 Introduction

To do experiments within industrial or technical research is often very expensive. One way of maximising the information gained while minimising the cost is to use statistical Design Of Experiments (DOE). An often used tool is, for instance, two level orthogonal design. Since time and money often are limiting factors the designs used are often reduced and without replicates. This leads to that all the degrees of freedom in the experiment will be consumed by the effect estimation. If this is the case we have no degrees of freedom left for estimating the variance and hence any formal inference will have to be of different type than the normal.

Let us assume that we are to perform a 2^{8-4}_{IV} designed experiment. The factors involved are called A, B, C, D, E, F G, and H. The person performing the experiment thinks it to be plausible that three main effect might be present. Let's say that A, B, C are believed to be true effects. If this is true the only true effects that there can be are from the factors and the interactions A, B, C, AB, AC, BC and ABC. If we have access to replicates we could now estimate the variance and make inference like tests or confidence intervals for the effects. But since we believe that only seven of the 15 effects should be non-zero we actually have access to at least eight observations that could be used to estimate the noise. When using designed experiments as a screening process for optimising a process or product the situation mentioned above is often the case. In this case we could actually do inference. One should mention that when working with designed experiments, things like confounding patters are very important but in this article the focus will be on analysis of orthogonal designs not on DOE.

When analysing a saturated orthogonal experiment there are a number of different methods. Various graphical methods are available like normal probability plots and Pareto charts see, for instance, Box et al. (1978) for an overview. There have been attempts to formalise the graphical procedures, see Daniel (1959, 1976). In Daniel (1959) a half-normal plot was used when the modulus was plotted. This work was continued by Zahn (1975) who studied the empirical properties and simulated bounds for a testing procedure. For an overview of different methods for effect testing in saturated orthogonal designs see Haaland and O'Connel (1995) or Berk and Picard (1991).

All of these methods are based on the normation of the estimated effects by some kind of variance estimate. In Voss (1999) a more parametrical approach is used to construct the generalised modulus ratio (GMR) test. The GMR test is proven to maintain the correct multiple level of significance. In the proof of this property Voss uses a stochastic ordering lemma "borrowed" from the literature on ranking and selection. Voss also shows that the same method of proof will apply to the method in Berk and Picard (1991). Confidence intervals are also constructed and are shown to maintain, at least, the stated level of confidence. In the case of no true effect the confidence intervals are exact. The GMR test uses a simple but rather crude estimate for the variance. In this article we will use a slightly more complicated variance estimation process under the assumption that the measurement errors are normal. We will show that the test procedure that arises from this estimation will also maintain the correct multiple level of significance. We will also note that the proof of maintaining the correct level of significance might be translated to other distributions than the normal distribution. The asymptotic behaviour of the variance estimate will also be examined.

The setting is that we have $Y_i \sim N(\mu_i, \sigma)$, $i = 1, \ldots, n$ such that the Y_i are independent. We then use $|Y_i| = X_i$ as our sample. Assume that $\mu_i = 0$

say at least r of the X_i:s. The idea is to use the smallest r observations from the sample, which we hope contains no effect, and use these to estimate the variance. We will look at the r smallest observations of the sample and view these as a censured sample from a normal distribution and simply use the ml-estimates for σ as our estimate. The notation is not limited to saturated designs since the method might be applied to other problems.

2 Variance estimation

We will examine two different ways of estimating the variance in saturated designs. The methods used are actually more general than the orthogonal design setting, see Adolfsson (2002). In order to choose between the different estimates we will calculate the asymptotic variance of the estimates i.e. the Cramér-Rao limit. Both of the estimation methods have their positive and negative features. The first method views the observations as a censored sample which means that we take both the total number of effects and the number of effects used for estimating the variance into account. In the second method we view the sample as a sample from a cut off half-normal distribution. We then need to estimate both the cut off parameter and the variance. The support of the distribution is then dependent on the cut off parameter which makes asymptotic variance calculations difficult. The second method is however actually more correct in the setting when there exists effects since it really does not care what observations it misses. The true distribution to be used is more complicated since we in the first test have the situation used by the first method. If this test rejects we do have a sample from the r smallest out of n where we have one randomly selected observation has been excluded etc. This distribution changes in every step and hence so will the variance estimate. The use of such an estimate for use in a step down test procedure with correct multiple level of significance is, to my knowledge, an open problem.

2.1 Variance estimation method 1

As stated previously our sample consists of $X_j = |Y_j|$ where $Y_i \sim N(0, \sigma^2)$ $j = 1, \ldots, n$. We now order these observations into $0 < X_{(1)} < X_{(2)} < \ldots < X_{(n)}$ and for notational convenience set $X_i = X_{(i)}$. Now we take the r smallest observations and use these for estimating σ in the ml-sense. The density function for these censured data is

$$\binom{p}{r}\left(\prod_{i=1}^{r}\frac{2}{\sigma}\varphi\left(\frac{x_i}{\sigma}\right)\right)\left(2-2\Phi\left(\frac{x_r}{\sigma}\right)\right)^{n-r}$$

if x_r is the largest of the r smallest observations. This pdf is also the likelihood function for σ. So the log likelihood function is

$$l(\sigma,x)=k(n,r)-r\ln(\sigma)+(n-r)\ln(1-\Phi\left(\frac{x_r}{\sigma}\right))-\frac{1}{2\sigma^2}\sum_{i=1}^{r}x_i^2. \qquad (1)$$

Using that

$$\frac{\partial}{\partial\sigma}\varphi\left(\frac{x_i}{\sigma}\right)=\frac{x_i^2}{\sigma^3}\varphi\left(\frac{x_i}{\sigma}\right) \text{ and } \frac{\partial}{\partial\sigma}\Phi\left(\frac{x_r}{\sigma}\right)=-\frac{x_r}{\sigma^2}\varphi\left(\frac{x_r}{\sigma}\right)$$

we get

$$\frac{\partial l}{\partial\sigma}=0\Rightarrow -\frac{r}{\sigma}+(n-r)\frac{\frac{x_r}{\sigma^2}\varphi\left(\frac{x_r}{\sigma}\right)}{1-\Phi\left(\frac{x_r}{\sigma}\right)}+\frac{1}{\sigma^3}\sum_{i=1}^{r}x_i^2=0. \qquad (2)$$

A more convenient way of writing this equation is (both from a theoretical and practical point of view)

$$\sigma^2=\frac{r}{n-r}(\sigma^2-S_r^2)\frac{1-\Phi\left(\frac{x_r}{\sigma}\right)}{\frac{x_r}{\sigma}\varphi\left(\frac{x_r}{\sigma}\right)}. \qquad (3)$$

To get the ml-estimator we now have to solve this nonlinear equation. For comments on numerical methods for solving this see section 2.4.

One question that has to be answered is how large r has to be in relation to n for the estimate to be good enough. One way of getting a glimpse at this is to look at the asymptotic variance of the ml-estimate. It is known that under some regularity conditions a ml-estimator is asymptotically efficient (i.e asymptotically achieves the Cramér-Rao lower bound). The Cramér-Rao bound for the variance for the parameter $\tau(\theta)$ is given by

$$\frac{(\tau'(\theta))^2}{n\mathbb{E}_\theta\left[(\frac{\partial}{\partial\theta}\ln(f(X|\theta)))^2\right]}.$$

Since in our case $\tau(\sigma) = \sigma$ we only have to work with the denominator. We approximate the alternative expression for the denominator with

$$
\begin{aligned}
-\frac{\partial^2 l}{\partial \sigma^2} &= \frac{1}{\sigma^2}\left(3\sum_{i=1}^{r}\frac{x_i^2}{\sigma^2} + (n-r)\frac{\frac{x_r}{\sigma}\varphi\left(\frac{x_r}{\sigma}\right)}{1-\Phi\left(\frac{x_r}{\sigma}\right)}\right. \\
&\cdot \left.\left(2 - \frac{x_r^2}{\sigma^2} + \frac{\frac{x_r}{\sigma}\varphi\left(\frac{x_r}{\sigma}\right)}{1-\Phi\left(\frac{x_r}{\sigma}\right)}\right) - r\right). \qquad (4)
\end{aligned}
$$

If we let $r/n = p \in (0,1)$ and $r, n \to \infty$ we get

$$
u_r = \frac{x_r}{\sigma} \to u_\infty = \Phi^{-1}\left(\frac{p}{2} + \frac{1}{2}\right). \qquad (5)
$$

One can also see that

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{r}\frac{X_i^2}{\sigma^2} &= \frac{1}{n}\sum_{i=1}^{r}U_i^2 = \frac{1}{n}\sum_{i=1}^{n}U_i \mathbb{I}_{\{U_i \le U_r\}} \qquad (6) \\
&\to \mathbb{E}[U^2 | U < u_\infty] = \int_0^{u_\infty} u^2 2\varphi(u)\, du
\end{aligned}
$$

using the law of large numbers. Using (5) and (6) we get an approximation of (4) valid for large n and r, namely

$$
\begin{aligned}
-\frac{\partial^2 l}{\partial \sigma^2} &\approx \frac{n}{\sigma^2}\left(2u_\infty\varphi(u_\infty)\left(2 + 2\frac{u_\infty\varphi(u_\infty)}{1-p} - u_\infty^2\right)\right. \\
&+ \left. 3\int_0^{u_\infty} u^2 2\varphi(u)\, du - p\right) \\
&= \frac{n}{\sigma^2}f(p).
\end{aligned}
$$

The asymptotic factor function $f(p)$ is plotted in figure 1. This can be compared to the variance of the variance estimate that one would derive if one used the complete set of data. Calculating this is a simple task and one gets

$$
-\frac{\partial^2 l_1}{\partial \sigma^2} \approx \frac{n}{\sigma^2}\left(3\mathbb{E}[U^2] - 1\right) = \frac{2n}{\sigma^2}.
$$

Some elementary calculus gives that the

$$
2u_\infty\varphi(u_\infty)\left(2 + 2\frac{u_\infty\varphi(u_\infty)}{1-p} - u_\infty^2\right)
$$

part of $f(p)$ goes to zero as p approaches 1 from below, and hence the two asymptotic variances are equal at $p = 1$.

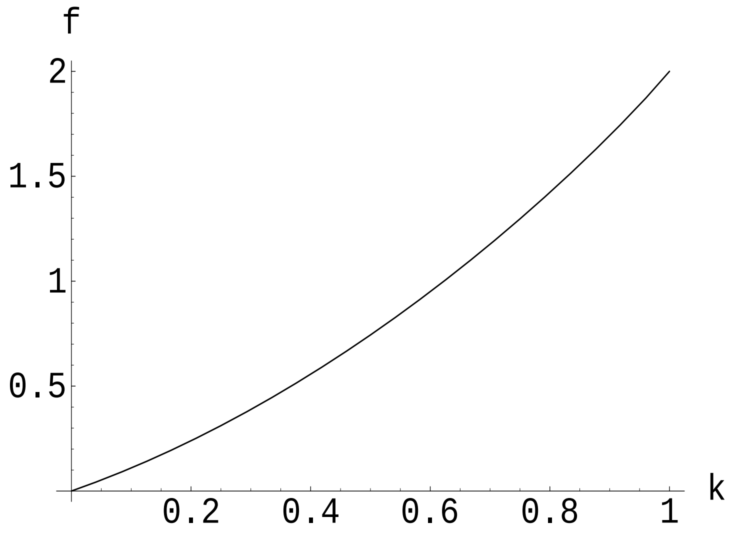

Figure 1: Asymptotic variance factor for the first method

2.2 Variance estimation method 2

The method presented here has the advantage of being more robust for the random censoring that will occur in the testing procedure that we are to construct. On the other hand it is more dependent on how large a proportion of the sample that we use for estimation since it uses less information than the previous method.

The idea now is to simply cut off the distribution at some level Δ. We then estimate σ (and Δ) in the maximum likelihood sense. So we now see the observations as observations from a distribution living only on $(0, \Delta)$ and hence the likelihood function is

$$L(\sigma, \Delta, x) = \prod_{i=1}^{r} \frac{\frac{2}{\sigma}\varphi\left(\frac{x_i}{\sigma}\right)}{2\left(\Phi\left(\frac{\Delta}{\sigma}\right) - 1/2\right)}, \tag{7}$$

and the log-likelihood function is then

$$l(\sigma, \Delta, x) = -n \ln \sigma + r - \frac{1}{2\sigma^2} \sum_{i=1}^{r} x_i^2 - n \ln \left(\Phi \left(\frac{\Delta}{\sigma} \right) \right) + k \tag{8}$$

where k is a known constant. From (7) we see that Δ should be estimated with the lowest possible value that makes sense, i.e $\Delta = x_r$. We put this into

(8), differentiate and get

$$\frac{\partial l}{\partial \sigma} = -\frac{n}{\sigma} - \frac{1}{2\sigma^2}\sum_{i=1}^{r} x_i^2 + n\frac{\frac{x_r}{\sigma^2}\varphi\left(\frac{x_r}{\sigma}\right)}{\Phi\left(\frac{x_r}{\sigma}\right) - 1/2}.$$

Setting this to zero and solving gives

$$\sigma^2 = \frac{(\sigma^2 - s_r^2)(\Phi\left(\frac{x_r}{\sigma}\right) - 1/2)}{\frac{x_k}{\sigma}\varphi\left(\frac{x_r}{\sigma}\right)}. \tag{9}$$

This is also a nonlinear equation that has to be solved numerically. In the case of this estimate the asymptotics are not as clear as in the previous case. The problem is that the set for which $f_X(x) \neq 0$ is dependent on the parameter Δ. This might result in that the Cramér-Rao bound might not be valid for σ and super efficient estimates might exist. However, if Δ and σ estimates are independent the Cramér-Rao bound will hold for σ. In the present case we can not at all be sure that the asymptotic results hold but never the less they will be presented, **keep this in mind!** Following the methodology in the previous section we calculate

$$\begin{aligned} -\frac{\partial^2 l}{\partial \sigma^2} &= \frac{n}{\sigma^2}\left(1 - \frac{3}{n}\sum_{i=1}^{r}\frac{x_i^2}{\sigma^2} + \frac{x_r}{\sigma}\left(\frac{-2\varphi\left(\frac{x_r}{\sigma}\right) + \frac{x_r^2}{\sigma^2}\varphi\left(\frac{x_r}{\sigma}\right)}{\Phi\left(\frac{x_r}{\sigma}\right) - 1/2}\right.\right. \\ &+ \left.\left.\frac{x_r}{\sigma}\left(\frac{\varphi\left(\frac{x_r}{\sigma}\right)}{\Phi\left(\frac{x_r}{\sigma}\right) - 1/2}\right)^2\right)\right). \end{aligned}$$

Using the same notation and arguments as in the previous section we get

$$\begin{aligned} -\frac{\partial^2 l}{\partial \sigma^2} &\to \frac{n}{\sigma^2}\left(3\int_0^{u_\infty} u^2\varphi(u)\,du - p - u_\infty\varphi(u_\infty)\left(-4 + u_\infty^2\right.\right. \\ &+ \left.\left.2\frac{u_\infty\varphi(u_\infty)}{p}\right)\right). \end{aligned}$$

Let us write this on the form

$$-\frac{\partial^2 l}{\partial \sigma^2} \to \frac{n}{\sigma^2}f(p).$$

This is valid since u_∞ is a function of p. We would like $f(p) \to 0$ as $p \to 0$ and $f(p) \to 2$ as $p \to 1$. The latter limit is trivial but for the first limit to hold we need to see that

$$\lim_{p\to 0}\frac{u_\infty(p)}{p}=\lim_{p\to 0}\frac{u'_\infty(p)}{1}=\lim_{p\to 0}\frac{1}{p'(u_\infty)}=\lim_{p\to 0}\frac{1}{2\varphi\left(u_\infty\right)}=\frac{\sqrt{2\pi}}{2}=\sqrt{\frac{\pi}{2}}$$

using l'Hopital's rule and implicit derivatives. We can now conclude that the function f has the same behaviour in the endpoints as the function in the previous section. This function does however grow more rapidly close to 2 and is hence less effective in some sense. Figure 2 shows the plot of $f(p)$.

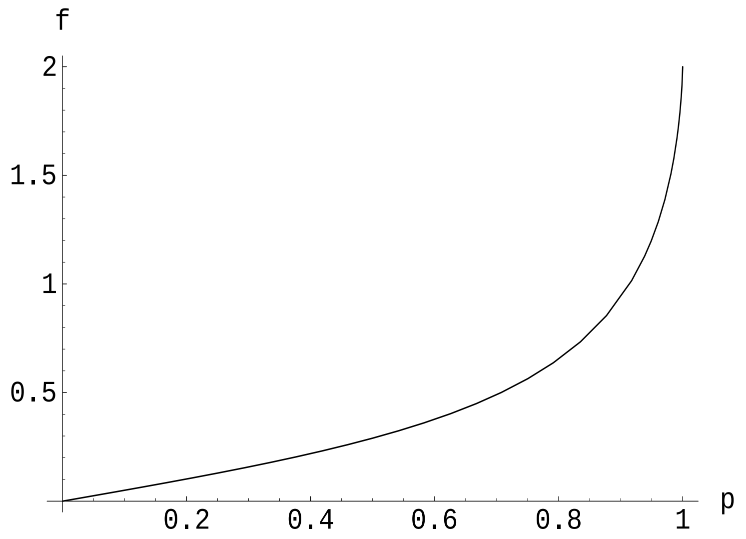

Figure 2: Asymptotic variance factor of the second method

2.3 Choosing variance estimation method

As we could see from the asymptotic variances, method one, viewing the problem as a censored sample, is the better of the two. This is natural since we actually have more information in this case. We have included in our model that we have n observations from the beginning. In the case of the second method we view the sample as from a cut off half-normal distribution. If the sample proportion used for estimation is small, which probably would be the case in saturated orthogonal designs, the cut off level will be low. This imply that the sample will come from a region of the normal distribution close to the origin. Since we are trying to estimate the curvature of the normal distribution this puts us in a rather difficult position since the normal distribution is rather flat in this region. The flatness of this region makes the estimates unreliable and hence we choose to work with the first of the variance estimations techniques.

2.4 Numerical comments on variance estimation

The function (1) and equation (2) might look a bit complicated but they are quite easy to handle. The simplest method for assessing the ml-estimator is probably to solve (2). When this function is plotted it gives an L-shaped curve which is known to give a complicated equation to solve using normal Newton methods based on estimation of the derivatives. I propose that a simple interval splitting algorithm is used. One could use Möbius approximations and solve these instead but this seems unnecessary. One should check whether or not the algorithm one is about to use uses derivatives or not since methods using derivatives might give very poor convergence (or even no convergence). There is an important point to this: many estimations using the ML-method give rather difficult numerical problems and one should be careful and make sure that the optimising tool that is used is actually suitable for the task.

3 Testing for true effects in saturated orthogonal designs

We are going to construct a step down test for testing for true effects that maintains the correct multiple level of significance and different types of confidence intervals that both keeps the correct multiple level of significance and some that only keep the individual level. In the first section we will work out the theory needed for showing these results. Hence the underlying model has the structure (since we are dealing with normal distributions folded round the y-axis we can assume that $\mu \geq 0$)

$$\mu_{(1)} \geq \mu_{(2)} \geq \ldots \geq \mu_{(i)} \geq \mu_{(i+1)} = \ldots = \mu_{(n)} = 0.$$

The problem is now to find i such that $\mu_{(i)} > 0$ and $\mu_{(i+1)} = 0$.

3.1 Theory

To show that the test that we are about to construct gives the correct multiple level of significants we need a lemma. The main idea in the proof of correct multiple level of significance is the use of stochastic orderings. To show this monotonicity of the test statistic we need to examine the variance estimates further. In the proof of this lemma we also need a fact known from reliability theory namely that the failure rate of a normal distribution is monotonically increasing. An observation to be made is that the same arguments will do for other distributions in the exponential family with monotony increasing failure rates (see Adolfsson (2002) for the gamma distribution).

Lemma 1 *The estimate σ^2 given by (3) is non-decreasing in S_r^2 for fix x_r.*

Proof: We will use a contra positive proof. We will show that if σ^2 decreases so will S_r^2. We rewrite (3) in the form

$$\sigma^2 = \frac{r}{n-r}(\sigma^2 - S_r^2) f\left(\frac{x_r}{\sigma}\right).$$

Rearranging the equation gives

$$\sigma^2 \left(1 - \frac{n-r}{rf\left(\frac{x_r}{\sigma}\right)}\right) = S_r^2.$$

We observe that $f(x) = 1/(xz(x))$ where $z(x)$ is the failure rate function for a normal distribution. We know now that $f(x)$ is a decreasing function. The proof is now easy: σ^2 decreases $\Rightarrow$ x_r/σ increases $\Rightarrow$ $f(x_r/\sigma)$ decreases $\Rightarrow$ $\frac{n-r}{rf\left(\frac{x_r}{\sigma}\right)}$ increases $\Rightarrow$ $1 - \frac{n-r}{rf\left(\frac{x_r}{\sigma}\right)}$ decreases $\Rightarrow$ $\sigma^2\left(1 - \frac{n-r}{rf\left(\frac{x_r}{\sigma}\right)}\right)$ decreases $\Rightarrow$ S_r^2 decreases. **Q.E.D.**

Lemma 2 *The estimate σ^2 given by (3) is non-decreasing in x_r.*

Proof: Once more we rearrange equation (3) to get

$$\frac{n-r}{rf\left(\frac{x_r}{\sigma}\right)} + \frac{1}{r} \cdot \frac{x_r^2}{\sigma^2} = 1 - \frac{k}{\sigma^2} \tag{10}$$

where k is a constant. We observe that σ is a non-constant function of x_r. Now take $x_r' \in \mathbb{R}^+$ such that $x_r < x_r'$. We observe that when x_r/σ

increases then (10) gives that σ increases. We are now going to force out a contradiction. Let $\sigma' = \sigma(x_r')$, assume that

$$\frac{x_r}{\sigma} > \frac{x_r'}{\sigma'}.$$

According to (10) we must have $\sigma' < \sigma$ but this gives that $x_r > x_r'$, a contradiction, and hence

$$\frac{x_r}{\sigma} \leq \frac{x_r'}{\sigma'}.$$

Once again we use (10) and conclude that $\sigma \leq \sigma'$. **Q.E.D.**

A basic part of the proof of correct multiple level of significants is stochastic ordering. The following definition and lemma is needed.

Definition 1 (Lehmann (1986)) *A family of distribution functions $F_\theta(x)$ on $\mathbb{R}$, for θ a real parameter, is said to be* stochastically increasing *if $\theta < \theta'$ implies $F_\theta(x) \geq F_{\theta'}(x)$ $\forall x$. Similarly, the family of distributions is said to be* stochastically decreasing *if $\theta < \theta'$ implies $F_\theta(x) \leq F_{\theta'}(x)$ $\forall x$.*

Lemma 3 (Stochastic ordering lemma (Voss (1999))) *Let $F_\theta(x)$, for real parameters θ, be a stochastically increasing family of distribution functions on the real line. Let $X_1, X_2, \ldots, X_k$ be independent random variables, where the distribution function of X_i is $F_{\theta_i}(x_i)$. For any fixed i $(1 \leq i \leq k)$, if the statistic $t = t(x_1, x_2, \ldots, x_k)$ is non-increasing (nondecreasing) function of x_i when all x_j for $j \neq i$ are held fixed, then the distribution of $T = t(X_1, X_2, \ldots, X_k)$ is stochastically decreasing (increasing) in θ_i.*

3.2 Test construction

We are going to construct a test based on the statistic

$$T_i = \frac{\max\{X_1, \ldots, X_i\}}{\hat{\sigma}}$$

which we are going to compare with critical values $c_\alpha(i, r, n)$ from the distribution of

$$C_i = \frac{\max\{X_1, \ldots, X_i\}}{\hat{\sigma}}, \quad i = 1, \ldots, n$$

such that

$$\mathbb{P}\left\{T_i > c_\alpha(i, r, n)\right\} = \alpha$$

under the null-hypothesis. The indexes r and n in $c_\alpha(i,r,n)$ is due to the fact that $\hat{\sigma} = \hat{\sigma}(r,n)$. With this we will be able to construct a testing method that maintains a correct multiple level of significant.

For this test to have any practical relevance we need to show that $\hat{\sigma}/\sigma$ is independent of σ. We note that the distribution of X_i/σ is independent of σ and using (2) we get that

$$\frac{\hat{\sigma}}{\sigma} = \frac{\frac{X_r}{\sigma}(p-r)\varphi\left(\frac{X_r}{\hat{\sigma}}\right)}{\left(1-\Phi\left(\frac{X_r}{\hat{\sigma}}\right)\right)\left(r-\frac{1}{\hat{\sigma}^2}\sum_{i=1}^r X_i^2\right)}.$$

Substituting $Z_i = X_i/\sigma$ and $q = \hat{\sigma}/\sigma$ we get

$$q = \frac{Z_r(p-r)\varphi\left(\frac{Z_r}{q}\right)}{\left(1-\Phi\left(\frac{Z_r}{q}\right)\right)\left(r-\frac{1}{q^2}\sum_{i=1}^r Z_i^2\right)}$$

and we see that this solution (in q) is independent of σ. This tells us that the test statistic T_i is independent of σ and hence that the test may be of practical use.

The test procedure is to reject the hypothesis $H_{0,n} : \mu_n = 0$ and conclude $H_{1,n} : \mu_n \neq 0$ if $t_n > c_\alpha(n,r,n)$ otherwise **no** rejections are made. If we reject in the first test we test $H_{0,n-1} : \mu_{n-1} = 0$ against $H_{1,n-1} : \mu_{n-1} \neq 0$ and reject if $t_{n-1} > c_\alpha(n-1,r,n)$. If no rejection is made in the second step we stop and only conclude that $\mu_n \neq 0$, otherwise we conclude that $\mu_n \neq 0$ and $\mu_{n-1} \neq 0$ and we continue. The procedure is now repeated until the first time we can not reject. Hence, no hypothesis can be rejected if not all of its predecessors have been rejected. One observation to be made is that the deeper we go down the rejection chain the more difficult it will be to reject. This is due to our variance estimate since if we have rejected a false hypothesis the variance estimation is incorrect. The correct variance estimate should be one where we have the r smallest out of n where one observation has been randomly excluded. We will over estimate the variance and hence our method will be conservative when true effects are present.

We now show that this test procedure maintains the correct level of significants.

Theorem 1 *If X_i $(i = 1,\ldots,n)$ are independent and $X_i = |Y_i|$, $Y_i \sim N(\mu_i, \sigma^2)$, then the test procedure above maintains the multiple level of significants α. The multiple significants level is exactly α in the null case.*

Proof: Let h denote the number of effects μ_i that are zero. The error rate is exactly α in the null case, for which $h = n$, from the definition of $c_\alpha(n, r, n)$. If $h = 0$ we can not do any false rejections. Hence we assume that $1 \leq h < n$. Without loss of generality we relabel the effects so that $\mu_i = 0$ for $i \leq h$ and $\mu_i \neq 0$ otherwise. A false assertion can only be made if

$$t_h = \frac{\max_{i \leq h}\{x_i\}}{\hat{\sigma}} > c_\alpha(h, r, n).$$

Now according to lemma 1 and lemma 2 t_h is a non-increasing function of x_j for each $j > h$, so the distribution of T_h is stochastically decreasing in each μ_j for each $j > h$ by the stochastic ordering lemma. We note that the critical values are chosen such that

$$\mathbb{P}\{T_h > c_\alpha(h, r, n)\} = \alpha$$

when $\mu_i = 0$ for all i. Hence,

$$\mathbb{P}\{T_h > c_\alpha(h, r, n)\} \leq \alpha$$

for **any** parameter configuration $(\mu_1, \mu_2, \ldots, \mu_n)$. **Q.E.D.**

n	k										
8	4	4.584	3.732	2.846	1.967						
9	4	4.789	4.160	3.307	2.549	1.747					
10	5	4.228	3.747	3.058	2.434	1.758					
11	5	4.456	3.956	3.359	2.755	2.168	1.567				
12	6	4.102	3.627	3.148	2.698	2.163	1.589				
13	6	4.238	3.817	3.334	2.956	2.432	1.979	1.477			
14	7	4.003	3.634	3.291	2.824	2.376	1.969	1.491			
15	7	4.184	3.772	3.407	3.080	2.583	2.172	1.848	1.384		
16	8	3.998	3.659	3.313	2.960	2.573	2.214	1.854	1.438		
17	8	4.040	3.742	3.406	3.048	2.709	2.442	2.061	1.737	1.355	
18	9	3.912	3.666	3.352	3.009	2.677	2.382	2.094	1.745	1.386	
19	9	3.953	3.701	3.426	3.106	2.834	2.557	2.247	1.975	1.649	1.316
20	10	3.906	3.629	3.358	3.045	2.814	2.554	2.236	1.988	1.687	1.351

Table 1: Table of critical values for 5% significance given n and k.

3.3 Simulating critical values for the testing procedure

The speed of our computers is still growing by the hour. This gives that simulations will be easier and easier to perform. But on the other hand we will also want to solve larger and larger problems. It is my experience that the implementation of the simulation method makes quite a difference. To generate the table 1, which contains critical values for the testing procedure in the previous section with an level of significance equal to 5% (based on 10000 simulations each), I used a specially designed C-program. Comparing the executional times between MATLAB and my code is comparing days to minutes. One should be aware that even though our computers are fast, and getting faster, we can still gain a lot from good algorithms.

4 Construction of confidence intervals

The construction of multiple confidence intervals are closely related to the method used for making intervals using the studentizied maximum modulus distribution. The intervals will only be calculated for effect not used in the variance estimate $\hat{\sigma}^2$. One could actually make intervals for all effects but then one needs to calculate a new distribution. We would have to construct a new distribution where we make sure that the X_i in the numerator is not used in the denomerator. By this we would avoid dependence between numerator and denomerator in the pivot distribution needed. Of course intervals constructed this way will be conservative.

The construction of a confidence interval that maintains the multiple level of significance is straight forward. We will use the same notation as in the previous section. The pivotal distribution will be that of

$$\max_i \frac{X_i}{\hat{\sigma}} \tag{11}$$

assuming that $\mu_i = 0$. The interval is derived from

$$1 - \alpha \quad = \quad \mathbb{P}\left\{ \max_i \frac{|Y_i - \mu_i|}{\hat{\sigma}} < z_\alpha \right\}. \tag{12}$$

Since this is valid for all μ_i:s the intervals will maintain the correct level of significance.

For constructing individual confidence intervals one needs to calculate a new pivotal distribution. To make this we need a new notation σ_i. The distribution needed is that of

$$\frac{|Y_i - \mu_i|}{\hat{\sigma}_i} \tag{13}$$

where $\hat{\sigma}_i$ is the estimate of σ using the r smallest observations *excluding* X_i. We need to do this to be able to use the stochastic ordering lemma to show that this distribution is stochastically decreasing in all $\mu_j,\ j \neq i$ which garanties that we maintains the individual level of significance.

Lemma 4 *The distribution of Eq. 13 is stochastically decreasing in* $|\mu_i|$ *for all* $i \neq j$.

Proof: The proof follows from the stochastic ordering lemma and lemmas 1 and 2. **Q.E.D.**

The lemma tells us that we can construct the confidence intervals from the distribution of Eq. 13 by setting $\mu_i = 0$ for all i.

Theorem 2 *The confidence intervals for* μ_i *below keeps at least the confidence level* $100(1-\alpha)\%$*:*

$$x_i \pm z_\alpha(r,n)\hat{\sigma}_i$$

where α *is chosen such that* $\mathbb{P}\left\{X_i/\hat{\sigma}_i > z_\alpha(r,n)\right\} = \alpha$.

In table 2 the 0.05, 0.01 and 0.001 percentiles of the confidence interval-distribution is shown. The table is based on 10,000 simulations each. The confidence intervals will by necessity be conservative if large true effects are present. Assume that we have a 2^4 full factorial design and use the 8 smallest absolute values for estimating the variance and construct the confidence intervals. We further assume that there is one very large effect. This effect is so large that it in no way influences the variance estimation. Then the confidence intervals should rather be based on a table for $n = 14$ and $r = 8$ and hence the confidence intervals will be conservative.

5 Power simulations

The numerical study will limit itself to comparing with Voss (1999). In Voss (1999) one can find comparisons between GMR and other methods. If one

n	k	5%	1%	0.1%	n	k	5%	1%	0.1%
3	1	2.880	15.007	142.234	27	13	1.858	2.530	3.500
4	2	1.551	3.616	10.399	28	14	1.847	2.506	3.462
5	2	2.123	5.082	15.345	29	14	1.857	2.518	3.501
6	3	1.736	3.256	7.430	30	15	1.854	2.493	3.392
7	3	1.979	3.716	8.339	31	15	1.858	2.506	3.450
8	4	1.779	2.970	6.190	32	16	1.856	2.487	3.359
9	4	1.905	3.176	6.516	33	16	1.867	2.498	3.366
10	5	1.783	2.769	4.889	34	17	1.857	2.487	3.340
11	5	1.878	2.954	5.306	35	17	1.867	2.486	3.351
12	6	1.811	2.707	4.455	36	18	1.865	2.469	3.281
13	6	1.862	2.802	4.517	37	18	1.873	2.489	3.304
14	7	1.812	2.637	4.164	38	19	1.867	2.482	3.309
15	7	1.858	2.742	4.329	39	19	1.877	2.487	3.314
16	8	1.823	2.608	3.925	40	20	1.870	2.466	3.231
17	8	1.853	2.670	4.121	41	20	1.874	2.472	3.271
18	9	1.829	2.572	3.853	42	21	1.874	2.460	3.244
19	9	1.852	2.632	3.900	43	21	1.878	2.469	3.229
20	10	1.840	2.570	3.730	44	22	1.875	2.470	3.221
21	10	1.856	2.586	3.761	45	22	1.883	2.484	3.235
22	11	1.834	2.532	3.619	46	23	1.876	2.463	3.219
23	11	1.856	2.563	3.643	47	23	1.883	2.467	3.255
24	12	1.841	2.537	3.505	48	24	1.879	2.465	3.216
25	12	1.854	2.539	3.476	49	24	1.886	2.471	3.215
26	13	1.844	2.518	3.505	50	25	1.881	2.460	3.196

Table 2: Confidence intervals percentiles given n and k.

is interested in the performance of our method compared with others than GMR Voss (1999) is a good place to start.The GMR method is performed as follows: Decide on the number of estimates believed to be merely random noise. Let us say that there should be at least k such estimates X_i. We form the Quasi Mean Square Error (QMSE) according to

$$\text{QMSE} = \frac{1}{k}\sum_{i=1}^{k} X_{(i)}^2.$$

We now base the step down test on

$$T_j = \frac{X_{(j)}}{\text{QMSE}}.$$

The largest of these T_j:s namely T_1 is compared with a suitable critical value $c(1, n, k, \alpha)$. If $T_1 < c(1, n, k, \alpha)$ we make no rejections and stop. If $T_1 > c(1, n, k, \alpha)$ we reject the hypothesis that the largest effect is a null effect and continue to test if $T_2 > c(2, n, k, \alpha)$ etc.

The simulations are made by choosing a specific set of non-zero β_i and running repeated tests on these. Since we have a step down procedure the power for each individual hypothesis will be decreasing since for instance the second hypothesis can only be rejected if the first is rejected. We will count the number of correct rejections for the true effects and the number of cases where *a*ny false rejections are made. In all of the simulations we use $n = 15$ and $k = 8$ i.e. a total of 15 estimated effects from which the eight smallest are used for the variance estimation.

5.1 Four equal effects

We now use four equal effects all the same size as the noise level of experiment. If the standard deviation of the measurement error is σ then $\beta_i = \sigma$ for four randomly selected i:s. In this setting both methods have very little power. This is however natural since the true effect is small compared with the noise level. The estimated Type I error probability is also given. This will be smaller then the designated $\alpha = 0.05$ since our variance estimate is too big when true effects are present. This conservativeness is however smaller than that of the GMR testing procedure. This is probably due to the more exact variance estimation method. The method presented in this paper will be abbreviated CVMR (Censored Variance estimation Modulus Ratio). In

Method	H_{01}	H_{02}	H_{03}	H_{04}	$\mathbb{P}\{\text{Type I error}\}$
CVMR	0.0874	0.0378	0.0185	0.0099	0.01275
GMR	0.0932	0.0284	0.00865	0.00245	0.00735

Table 3: Estimated power and type I error probability

table 3 the estimated power for each of the possible true hypothesis is given together with the estimated probability of a type I error. The table is based on 20,000 simulation where four β_i were set to equal σ.

In the next simulation four β_i:s were set to 2σ. The result of this simulations is shown in table 4.

Method	H_{01}	H_{02}	H_{03}	H_{04}	$\mathbb{P}\{\text{Type I error}\}$
CVMR	0.411	0.312	0.260	0.188	0.03775
GMR	0.419	0.256	0.151	0.0723	0.0042

Table 4: Estimated power and type I error probability

5.2 Four unequal effects

For X_i we now put $\mu_i = i\sigma$ and do as in the previous section. The result of 40,000 simulations is shown in table 5. The table clearly shows that the power is greater in the CVMR method than in the GMR method. This is probably due to the fact that the variance estimation is less conservative than the one used in GMR.

Method	H_{01}	H_{02}	H_{03}	H_{04}	$\mathbb{P}\{\text{Type I error}\}$
CVMR	0.933	0.764	0.410	0.0875	0.0314
GMR	0.931	0.661	0.237	0.0239	0.0051

Table 5: Estimated power and type I error probability

5.3 Conclusions concerning power

The CVMR method shows to have greater power when it comes to rejecting more than the first hypothesis. The variance estimation method used in CVMR uses information from the tail of the distribution in terms of the failure rate function. This stabilises the estimation and makes it more robust against the random censoring that the presence of true effects gives raise to. To gain more robustness against random censoring we need to use more advanced estimation methods that are adaptive. The use of adaptive methods for testing is an open problem when trying to maintain the correct multiple level of significance.

References

T. Adolfsson. *Using regression function information in determining the same.* PhD thesis, Chalmers University of Technology and Göteborg University, 2002.

K. N. Berk and R. R. Picard. Significans tests for saturated orthogonal arrays. *Journal of Quality Technology*, 1991.

G. E. P. Box, W. Hunter, and S. Hunter. *Statistics for experimenters.* Wiley, 1978.

C. Daniel. Use of half-normal plots in interpreting factorial two-level experiments. *Technometrics*, 1959.

C. Daniel. *Applications of Statistics to Industrial Exprimentation.* Wiley, New York, 1976.

P. Haaland and M. O'Connel. Inference for effect-saturated fractional factorials. *Technometrics*, 1995.

E. L. Lehmann. *Testing Statistical Hypotheses.* Springer-Verlag, 1986.

D. Voss. Analysis of orthogonal saturated designs. *J. Stat. Plan. and Inf.*, 1999.

D.A. Zahn. An empirical study of the half-normal plot. *Technometrics*, 1975.

Paper B

A step-down test for effects in unreplicated factorial trials

Sture Holm
Tobias Adolfsson
Chalmers University of Technology
Sigyn Mark
Trollhättan-Uddevalla University College

February 27, 2002

1 Introduction

Two-level fractional factorial designs are important statistical tools in particular for screening experiments and in situations where the cost of experimentation is high. It is common that no replicates are taken. Also full factorial designs may often be performed without replicates. This means that there does not exist any within series estimate of experimental error variance.

The problem has been overcome in different ways. If the fraction is not too small, the estimates of interactions of some high orders, not aliased with main effects or low order interactions may be used for variance estimation.

A common subjective method is to make a normal or half normal plot of all effect estimates and then select as important (significant) the effects with estimates deviating much from a straight line. This method by Daniel (1959) is commonly adopted in textbooks, e.g. Box et al. (1978) and Montgomery (1984).

There exist also some methods which have a warranted level of significance for declaring effects to be nonzero when they are in fact zero. The most important of those methods is Voss (1999, 1988). He constructs a variance estimate based on the half of the estimates with the smallest moduli.

The aim of the present paper is to present an alternative method with warranted level. It is based on another type of estimate, which makes it flexible with respect to the true number of zero and nonzero effects. Further this type of estimate makes it possible to construct a step-down test with a multiple level of significance. Principles and definitions for step-down test may be found in Hochberg and Tamhane (1987).

2 Coverage bounds

The test method to be presented later is based on coverage bounds for order statistics. In this section we will make a general discussion of coverage bounds in order to prepare for the method presentation in the next section.

Let $X^{(1)}, X^{(2)}, \ldots, X^{(n)}$ be order statistics in a sample of size n from a distribution with a cumulative distribution function $F(x)$. A left one-sided coverage band with coverage probability q is a sequence of constants $a_1, a_2, \ldots, a_n$ such that $\mathbb{P}\left\{X^{(k)} \geq a_k \text{ for } k = 1, 2, \ldots, n\right\} = q$. The band is said to be balanced if the individual missing probabilities $\mathbb{P}\left\{X^{(k)} < a_k\right\}$ are the same for all $k = 1, 2, \ldots, n$. In some application it is reasonable to have a coverage bound, which is balanced, but in others it is more reasonable to have a varying individual coverage probability for the order statistics. A right one-sided balanced coverage band may be defined analogously.

One may also define coverage bounds on both sides simultaneously, which then gives a coverage band. A two-sided coverage band with coverage probability q consists of two sequences of constants $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$ such that $\mathbb{P}\left\{a_k \leq X^{(k)} \leq b_k \text{ for } k = 1, 2, \ldots, n\right\} = q$. The band is said to be balanced if the individual missing probabilities $\mathbb{P}\left\{X^{(k)} < a_k\right\}$ and $\mathbb{P}\left\{X^{(k)} > b_k\right\}$ are the same on both sides and for all $k = 1, 2, \ldots, n$.

An attempt to use balanced coverage bands for out-lier detection is made by Atkinson (1981). He uses a half normal plot to detect outliers and influential observations. On the normal probability paper is constructed a band and the position of order statistics are compared to that band. The statistics are of a type given by Cook (1977). The band is constructed as an envelop of the order statistics for 20 simulated series of statistics. The reason for having 20 series seems to be related to some idea of standard coverage probability, since for each order the band reaches from the smallest to the largest of corresponding order statistics in the 20 simulated series. The method is further developed by Flack and Flores (1989). They made more accurate bands by basing it on a larger simulation and they also discussed other similar constructions.

We now consider the calculation of the coverage probability of a left one-sided coverage bound. The calculation can be made for the $U(0,1)$ distribution by first making the transformation $U_i = F(X_i)$ for $i = 1, 2, \ldots, n$ giving the ordered statistics $U^{(1)}, U^{(2)}, \ldots, U^{(n)}$. Then $U^{(k)}$ has a Beta distribution with parameters k and $n+1-k$. Now let $a_1, a_2, \ldots, a_n$ be any constants with the ordering $a_1 < a_2 < \ldots < a_n$ for which we want to calculate the probability $\mathbb{P}\left\{U^{(k)} \geq a_k \text{ for } k = 1, 2, \ldots, n\right\}$.

For each individual order statistic $U^{(k)}$ the probability $\mathbb{P}\left\{U^{(k)} < a_k\right\}$ is obtained by the Beta distribution. We introduce events $A_n = \{U^{(n)} < a_n\}$ and $A_k = \left\{U^{(k)} < a_k \text{ and } U^{(k+1)} \geq a_{k+1}\right\}$ for $k = 1, 2, \ldots, n-1$, which can be interpreted as a passage over the bound. Further we introduce the events $B_n = A_n$ and $B_k = A_k \cap \left(\cup_{j=k+1}^{n} A_j^c\right)$ for $k = 1, 2, \ldots, n-1$, which can be interpreted as the first passage over the bound when going from above in index. Thus now

$$\mathbb{P}\left\{U^{(k)} \geq a_k \text{ for } k = 1, 2, \ldots, n\right\} = 1 - \sum_{k=1}^{n} \mathbb{P}\left\{B_k\right\}.$$

The probabilities $\mathbb{P}\{B_k\}$ may be calculated successively with a simple equation system by using conditioning.

Since $A_k = \cup_{j=k}^{n}\{A_k \cap B_j\}$ we get

$$\mathbb{P}\{B_n\} = \mathbb{P}\{A_n\}, \quad \mathbb{P}\{B_{n-1}\} = \mathbb{P}\{A_{n-1}\}$$

and

$$\begin{aligned}
\mathbb{P}\{A_k\} &= \mathbb{P}\{A_k \cap B_k\} + \sum_{j=k+1}^{n} \mathbb{P}\{A_k \cap B_j\} \\
&= \mathbb{P}\{B_k\} + \sum_{j=k+1}^{n} \mathbb{P}\{B_j\}\,\mathbb{P}\{A_k | B_j\} \text{ for } k = 1, 2, \ldots, n-2.
\end{aligned}$$

Here

$$\begin{aligned}
\mathbb{P}\{A_k\} &= \binom{n}{k} a_k^k (1 - a_{k+1})^{n-k} \text{ and} \\
\mathbb{P}\{A_k | B_j\} &= \binom{j}{k} \frac{a_k^k (a_j - a_{k+1})^{j-k}}{a_j^j} \text{ for } j > k
\end{aligned}$$

which is proved in Lemma 1 in the appendix. Thus the equation system for determining $\mathbb{P}\{B_k\}$ is

$$\begin{aligned}
\mathbb{P}\{B_n\} &= a_n^n \\
\mathbb{P}\{B_{n-1}\} &= \binom{n}{n-1} a_{n-1}^{n-1}(1-a_n)^1 \\
\mathbb{P}\{B_k\} &= \binom{n}{k} a_k^k (1-a_{k+1})^{n-k} \\
&- \sum_{j=k+1}^{n} \mathbb{P}\{B_j\} \binom{j}{k} \frac{a_k^k (a_j - a_{k+1})^{j-k}}{a_j^j} \text{ for } k = n-2, n-3, \ldots, 1
\end{aligned}$$

By doing the calculation successively in the indicated order we do not even need to solve an equation system. From this result we can easily get the same type of result for random variables $X_1, X_2, \ldots, X_n$ with any continuous cumulative distribution function $F(x)$ by considering the transformation $U_k = F(X_k)$.

If $X^{(1)}, X^{(2)}, \ldots, X^{(n)}$ are the order statistics the coverage probability $\mathbb{P}\left\{X^{(k)} \geq a_k^* \text{ for } k = 1, 2, \ldots, n\right\}$ of a left one-sided coverage band with constants $a_1^*, a_2^*, \ldots, a_n^*$ is obtained by using $a_k = F(a_k^*)$ in the above calculation for the $U(0,1)$ distribution.

For the statistical problems to be treated later we want a reversed calculation. That is, for a given internal relation between the individual coverage probabilities för the different order statistics, we want to determine the absolute size of these probabilities in order to get a given total coverage probability. For instance if we want a balanced coverage bound we use the same individual coverage probability, which should be determined so that the total coverage probability is equal to some given value. In our statistical problems there are reasons for having individual coverage, which increases with the order of the statistic. Using a linear increase the individual probability $\mathbb{P}\left\{U^{(k)} \leq a_k\right\}$ is proportional to k. In order to get a given total coverage probability we need to determine this proportionality constant.

The problem of getting a given total coverage probability for a given structure of individual coverage probabilities is easily solved by iteration. From some starting values of the individual coverage probabilities, with the given internal structure, we calculate the constants a_k by the Beta distribution and then calculate the total coverage probability be the equation system for $\mathbb{P}\{B_k\}$. If this total coverage probability is smaller than intended the individual coverage probabilities are increased with the relative sizes preserved and if the total coverage probability is to big they are decreased. When there are results on both sides of the required total coverage the calculation is iterated in new points estimated linearly from the previous cases most close to the required result. The procedure converges to acceptable approximation in a few steps.

3 Estimation by use of coverage bound

A left coverage bound with a coverage probability of 0.5 is in a sense an estimate of a "natural left tangent" of the empirical cumulative distribution function. This coverage bound is situated in a median position to the left. In particular in one-parameter problems this may be used to get good parameter estimates.

When using the common method by Daniel (1959) one selects intuitively a lower part of the ordered values of the effect estimates to fit a straight line on the half normal plot and then declare as significant those factors whose absolute values of effect estimates deviate much from this line. We will now here use an objective and adaptive method with the same intention. In our method the estimate of the standard deviation is obtained by a scale factor required for aligning the ordered sample to the right of a left coverage bound with coverage probability 50%. In order to get good estimates we will use a coverage bound which has a linearly increasing individual coverage probability. This means that there is greater probability that a great ordered absolute value will be used for the estimate.

We describe the details of the construction in order to be precise. Let $a_1, a_2, \ldots, a_n$ be the boundary constants for a left bound with linearly increasing individual coverage probability and total coverage probability 50% for the case of $U(0,1)$ distribution. Then calculate the boundary constants $a_k^* = \Phi^{-1}((a_k+1)/2)$ for the distribution of ordered absolute values $X^{(1)}, X^{(2)}, \ldots, X^{(n)}$ of observations which are normally distributed with parameters $(0,\ \sigma)$. Then σ may be estimated by $\hat{\sigma} = \min_{1 \le k \le n}(X^{(k)}/a_k^*)$. When using the estimate in a situation with non-replicated factorial design, we have sample sizes of the type $n = 2^m - 1$ and the random variables are the absolute values of the estimates. If there are no true effects the estimate is median unbiased. If there are true effects, there is a tendency that the biggest absolute values deviate more or less to the right and σ is estimated by absolute values of lower order. Then the estimate is no more median un-

k	1	2	3	4	5	6	7
Bound	0.00062	0.0142	0.0452	0.0881	0.1390	0.1950	0.2575
8	9	10	11	12	13	14	15
0.3228	0.3915	0.4632	0.5379	0.6155	0.6966	0.7825	0.8771

Table 1: Bounds for a 50% left coverage bound for order statistics in a series of 15 ordered statistics from a uniform [0;1] distribution.

biased. We will use this type of estimate in the effect testing method in the next section.

In practice one needs to make a minor correction to the method. Since we always work with some rounding off of the observed quantities we might by coincidence get some estimates, which are 0. This numerical problem would completely destroy the estimate and give . It can in practice be avoided by adjusting the fitting with the size of the smallest observable unit. If this is d it means that we use the estimate $\hat{\sigma} = \min_{1 \leq k \leq n}((X^{(k)} + d)/a_k^*)$. This adjustment gives a minor median bias of the estimate.

4 Testing for effects in nonreplicated factorial designs

A full or fractional two-level experiment has $n = 2^k$ trials for some natural number k. We exemplify the discussion here with $k = 4$ which means 16 trials. In such a trial there is one estimate of the general mean and 15 effect estimates. Our aim is to get a reasonable test for effects which has a required small multiple level of significance $\alpha = 0.05$, i.e. a probability of at most 0.05 for declaring as significantly non-zero any effect which is in fact zero. This should be true for all possible combinations of zero and non-zero effects.

We need the 50% coverage bound constants for 15 ordered observations of absolute values from a $N(0,1)$ distribution and we can get those for the corresponding constants for the case of $U(0,1)$ distribution. In the calculation we use linearly increasing individual coverage probabilities. The individual probabilities, which give total coverage 50% are $0.1398 * i/15$. The result of the points a_i are given in Table 1.

By the simple transformation we get the corresponding constants for the

k	1	2	3	4	5	6	7
Bound	0.00078	0.0178	0.0567	0.1106	0.1751	0.2481	0.3285
8	9	10	11	12	13	14	15
0.4163	0.5123	0.6177	0.7353	0.8696	1.029	1.233	1.543

Table 2: Constants for a 50% left coverage bound for ordered absolute values of 15 normal (0;1) random variables.

case of absolute values of normal (0,1) random variables. The result is given in Table 2.

Using the constants in Table 2 and the estimate $\hat{\sigma}$ obtained by aligning the sample of absolute values of estimates to them we may now also make a test for effects. In the subjective method by Daniel (1959) one looks for much deviation among the biggest absolute values of the estimates. We follow the same line of thought in an objective method with prescribed risk of making any false statements on nonzero true effects. This is done by calculating a bound which will act as a rejection boundary, and for all effects with a greater absolute value of the estimate we will declare that there is a true effect.

The test we propose is a step down multiple test with a prescribed multiple level of significance a, which we will now describe in more detail.

Consider first a test of the hypothesis that all random variables are normally distributed with expectation 0 and the same (unknown) variance σ^2. First the absolute values of the estimates are ordered and then the scale is changed to align the cumulative distribution function with the 50 coverage bound. Next we consider the maximal value of this scale changed series, i.e. the scale changed maximal order statistic, which is used as a test s-tatistic. The distribution of this test statistic under the null hypothesis is rather complicated but a simulation easily gives e.g. the 95% quantile in the distribution, which then serves as a rejection limit.

In order to get a step down test we also study the distribution of the maximal ordered statistic in smaller samples of different sizes m after scale changes to alignment to the m smallest boundary values of the 50% left coverage bound defined above for the full sample. The 95% quantiles in these distributions are used as rejection limits.

In the closed test we may go from above successively and declare effects significant as long as the ordered statistics (absolute values) are above the

k	1	2	3	4	5	6	7
Bound	-	0.021	0.133	0.30	0.49	0.69	0.94
8	9	10	11	12	13	14	15
1.18	1.43	1.71	2.00	2.27	2.60	2.93	3.32

Table 3: Upper bounds for step down test with alignment to a 50% left coverage bound for ordered absolute values of 15 normal random variables. Multiple level of significance is 5%.

corresponding limits. Observe that the procedure has to stop the first time an order statistic is below the corresponding limit. In Theorem 1 in the appendix is proved that this test procedure has a multiple level of significance α.

This type of procedure can easily be constructed for other sample sizes. In two-level factorial designs the cases of 31 and 63 effects are particularily interesting. The very construction with a left coverage band, scale alignment and step down test limits can be used also for other scale families than the normal distribution.

When performing the test in practice we need to consider the rounding off results. Due to rounding there may be a considerable relativ variation in the small estimates in practical applications. This extra variation may destroy the estimate of the standard deviation if some estimate just happens to be extremely small. Occationally an estimate may even get the rounded value 0, which completely destroys the esimate, since it makes it formally equal to 0. A simple way to get rid of the problem in practice is to add the maximal rounding error to the absolute values of the estimates. It is easily seen that this makes the test procedure just slightly more conservative, and it means that the rounding can never decrease the estimate of σ.

5 An application example

As an illustration we use an example from Box et al. (1978). The estimates are given in their table 10.9 on page 331. In this case the measurement values are rounded to 0.25 units and the definition of the effects means that the maximal error in an estimate is 0.125. Thus after taking the absolute values of estimates we add 0.125 before we estimate the standard deviation and normalize the estimates. The estimated standard deviation in this case

1.80, which occurs in the alignment of the ordered absolute value 0.75 number 8 with the alignment value 0.4163.

Comparing the normalized absolute values of estimates with the rejection bound shows that in a multiple test with multiple level of significance 5% we may reject three hypothesis. In notations of Box et al. (1978) we can demonstrate existance of the main effects 2 (with estimate 24.0),1 (with estimate -8.0) and 4 (with estimate -5.5). The next effect is not significant for the multiple level of significans 0.05.

It may be illustrative to use a half normal plot for the test bounds as well as the normalised estimates. In figure 1 a plot of the data and the rejection bounds are presented.

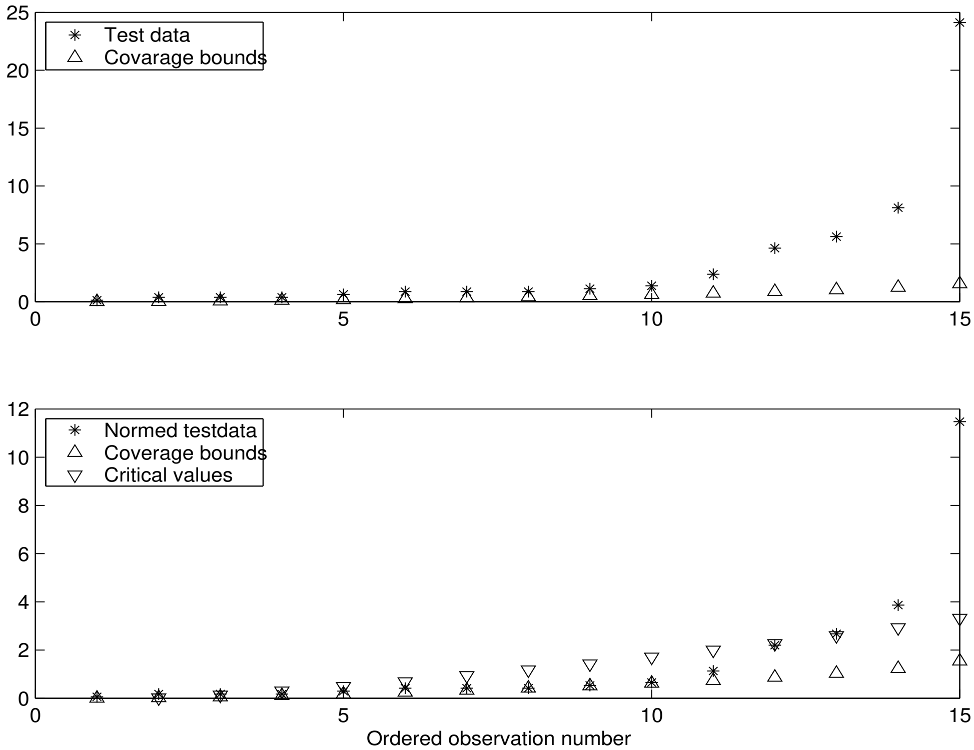

Figure 1: The top plot contains the raw data from Box et al. (1978) page 331and the 50% coverage bounds. The bottom plot contains the normed data, the coverage bounds, and the rejection bounds

6 Discussion

The statistical problem of determining significant effects in a reduced factorial design without replicates needs some method to obtain an error estimate from the very set of effect estimates we are investigating. The method we used here is based on a kind of tangent estimates, which is rather much related to a subjective technique of looking for a bend in a cumulative absolute normal probability plot.

Of course the technique means using order statistics to get the estimates. Usually order statistics are not the most efficient ones. On the other hand we have here the "sample size adaption" problem which can be suitably handled for order statistics. Further our method is well suited for illustration. The effect estimates can be seen together with the lower fitting boundary and the upper rejection boundary.

As already mentioned the method can be used for other sample sizes after preparation of a table. It can not only be used for reduced factorial designs but also for other problems with independent test statisics and these need not be normally distributed. An example may be some type of life time distributions with scale parameter, where it could be used after preparation of a suitable table. It is to be noted that for the table preparation there is the possibility of simulation beside the analytical calculation.

A Proofs

Lemma 1

$$\begin{aligned} \mathbb{P}\{A_k\} &= \binom{n}{k} a_k^k (1-a_{k+1})^{n-k} \\ \mathbb{P}\{A_k|B_j\} &= \binom{j}{k} \frac{a_k^k (a_j - a_{k+1})^{j-k}}{a_j^j} \quad \textit{for } j > k \end{aligned}$$

using the notation in section 2.

Proof: In the first equation we have that

$$\begin{aligned} \mathbb{P}\{A_k\} &= \mathbb{P}\left\{U^{(k)} < a_k \text{ and } U^{(k+1)} \geq a_{k+1}\right\} \\ &= \mathbb{P}\{k \text{ of the } U_i\text{:s } < a_k \text{ and } (n-k) \text{ of the } U_i\text{:s } \geq a_{k+1}\} \\ &= \binom{n}{k} a_k^k (1-a_{k+1})^{n-k} \end{aligned}$$

since the U_i:s are independent $U(0,1)$. For the second part we notice that

$$\mathbb{P}\left\{A_k|B_j\right\} = \mathbb{P}\left\{A_k|U^{(j)} < a_j\right\}$$

so that the $U^{(k)}$, $k = 1, \ldots, j$ conditioned on B_j are an ordered sample from $U(0, a_j)$. Hence we have

$$\begin{aligned}
\mathbb{P}\left\{A_k|B_j\right\} &= \binom{j}{k}\left(\frac{a_k}{a_j}\right)^k\left(\frac{a_j - a_k}{a_j}\right)^{j-k} \\
&= \binom{j}{k}\frac{a_k^k(a_j - a_{k+1})^{j-k}}{a_j^j}
\end{aligned}$$

Q.E.D.

Theorem 1 *The multiple test procedure described in section 4 has a multiple level of significance α.*

Proof: Denote the set of indices for the true hypotheses by I and the number of true hypotheses by m. Consider the ordered absolute values of the estimates x_i to these hypoteses and denote their maximal absolute value by z.

The proof is to a great extent based on conditioning on this z. Next consider the estimates x_i belonging to the false hypoteses, thus characterized by $\mu_i \neq 0$. Given Z, there is a random number N of the x_i-variables getting outcomes of their absolute values below z. These variables can in principle correspond to any of the $\mu_i \neq 0$.

The test procedure is based on ratios between ordered absolute variables. For the test actually made to reject the hypotheses corresponding to z, the ratio of absolute values of smaller index in the ordered series and z are used. A rejective error can only occur if there is rejection in this step.

Now condition on N as well as z. Decreasing values of the ordered absolute values of x_i with index less then $m + N$ increase the risk of making an error.

There is a stochastic order between the conditional distributions given an upper boundary z for the with $\mu = 0$ and any case with $|\mu| \neq 0$. Cases with $|\mu| \neq 0$ have stochastically greater conditional distributions. See lemma 2. Thus in a conditional calculation the risk is increased if, in the cases of false hypotheses, the distributions are substituted by the distribution of the true cases. Neither m nor N are known, but the sum $m+N$ appears as the number of remaining variables in the test really done.

Thus we now get an upper estimate of the risk of error by integrating over z in the conditional error obtained for $m+N$ variables with the same distribution. Observe however that the distribution of z is that of a maximal variable among m variables, where m is not known. The final trick is now to make the integration of the distribution of the maximum of the known number of $m+N$ variables (of the type in true hypotheses), which is stochastically larger than that of the unknown number of m variables. This gives again an overestimation of the risk since the conditional distributions are monotonically decreasing in the conditioning value z. See lemma 3.

The table used is based on calculations for this distribution. In order for a rejective error to occur, the procedure must reject in earlier steps and reject in the step corresponding to z. Since the risk in that step is bounded by α, the multiple risk is bounded by α. **Q.E.D.**

Note. There exists cases for which the values of the table are not conservative. If all false hypothesis have expectations μ_i with $|\mu_i|$ approaching ∞, the risk approaches α.

Lemma 2 *Suppose that U_1 is $N(0,1)$ distributed and U_2 is $N(\mu,1)$ distributed where $\mu \neq 0$. Then the distribution of $|U_1|$ given $|U_1| < z$ is stochastically smaller than the distribution of $|U_2|$ given $|U_2| < z$.*

Proof: We make the proof for $\mu > 0$, since the result depends only on the absolute value of μ. The cumulative distribution function of $|U_2|$ is equal to

$$\Phi(x-\mu) - \Phi(-x-\mu)$$

for any μ and all $x > 0$. The cumulative distribution function of $|U_1|$ is obtained as the special case $\mu = 0$. The probability density function in the general case is then

$$\varphi(x-\mu) + \varphi(-x-\mu) = 2\varphi(x)\, e^{-\mu^2/2} \cosh(\mu x)$$

and for the special case $\mu = 0$, the distribution of $|U_1|$, it is

$$2\varphi(x).$$

In order to prove the lemma we now consider the probabilities of the two intervals $(0,x)$ and (x,z) for the general case. We have for any z and any x,

$0 < x < z$, and $\mu > 0$

$$\begin{aligned}
\mathbb{P}_\mu \{0 < |U_2| < x\} &= \int_0^x 2\varphi(v)\, e^{-\mu^2/2} \cosh(\mu v) dv \\
&< e^{-\mu^2/2} \cosh(\mu x) \int_0^x 2\varphi(v)\, dv
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{P}_\mu \{x < |U_2| < z\} &= \int_x^z 2\varphi(v)\, e^{-\mu^2/2} \cosh(\mu v) dv \\
&> e^{-\mu^2/2} \cosh(\mu x) \int_0^x 2\varphi(v)\, dv.
\end{aligned}$$

Thus we have

$$\frac{\mathbb{P}_\mu \{0 < |U_2| < x\}}{\mathbb{P}_\mu \{x < |U_2| < z\}} < \frac{\mathbb{P}_0 \{0 < |U_1| < x\}}{\mathbb{P}_0 \{x < |U_1| < z\}}$$

which gives the stochastic ordering requirement

$$\frac{\mathbb{P}_\mu \{0 < |U_2| < x\}}{\mathbb{P}_\mu \{0 < |U_2| < z\}} < \frac{\mathbb{P}_0 \{0 < |U_1| < x\}}{\mathbb{P}_0 \{0 < |U_1| < z\}}.$$

Q.E.D.

Lemma 3 *Conditional distributions of the modulus $V = |U|/v$ of random $N(0,1)$ variables U, given $|U| \leq v$ is stochastically decreasing in v.*

Proof: Let Φ denote the cumulative distribution function for the $N(0,1)$ distribution. We have to show that if $v_1 < v_2$ then

$$\frac{\Phi(v_1 x) - 0.5}{\Phi(v_1) - 0.5} < \frac{\Phi(v_2 x) - 0.5}{\Phi(v_2) - 0.5}$$

for all $x, 0 < x < 1$. The left and right member coincide at $x = 0$ and $x = 1$. For any $v > 0$ and any $x, 0 < x < 1$, we have, however,

$$\begin{aligned}
\frac{\partial}{\partial v} \frac{\Phi(vx) - 0.5}{\Phi(v) - 0.5} &= \frac{x\varphi(vx)(\Phi(v) - 0.5) - \varphi(v)(\Phi(vx) - 0.5)}{(\Phi(v) - 0.5)^2} \\
&> \frac{x\varphi(vx)(\Phi(v) - 0.5) - \varphi(v)\, x(\Phi(v) - 0.5)}{(\Phi(v) - 0.5)^2} > 0
\end{aligned}$$

since $\Phi(v)$ is concave for $v > 0$ and $\varphi(v)$ is decreasing for $v > 0$. This proves the statement. **Q.E.D.**

References

A. C. Atkinson. Two graphical displays for outlying and influential observations in regression. *Biometrika*, 1981.

G. E. P. Box, W. Hunter, and S. Hunter. *Statistics for experimenters*. Wiley, 1978.

R. D. Cook. Detection of influential observations in linear regression. *Technometrics*, 1977.

C. Daniel. Use of half-normal plots in interpreting factorial two-level experiments. *Technometrics*, 1:311–341, 1959.

V. F. Flack and R. A. Flores. Using simulated envelopes in the evaluation of normal probability plots of regression residuals. *Technometrics*, 1989.

Y. Hochberg and A. Tamhane. *Multiple comparison procedures*. John Wiley, New York, 1987.

D.D Montgomery. *Design and analysis of experiments*. John Wiley, New York, 1984.

D. Voss. Generalized modulus-ratio tests for analysis of factorial designs with zero degrees of freedom for error. *Commun. Statist.-Theory Meth*, 1988.

D. Voss. Analysis of orthogonal saturated designs. *J. Stat. Plan. and Inf.*, 1999.

Paper C

Using regression function information in determining the same

Tobias Adolfsson

February 28, 2002

Abstract

There exists a number of methods for regression in the case when we have a poorly conditioned design matrix. Most of these methods use various regularisations on the design matrix. We will construct a regression method that uses information in the response as well as the design matrix. The method allows the design matrix to be ill conditioned. The method presented is shown to be theoretically simple compared with other methods using response information. Procedures for dimensionality testing are derived and predictive properties are thoroughly examined. Variance estimation in the case of effect sparsity is asymptotically examined and the asymptotics are studied by simulations.

1 Introduction

Many new and interesting regression methods have been born from the field of chemometrics. These methods are often subspace regression methods, i.e instead of solving the complete problem they solve an easier problem with lower dimension that approximates the original problem. These methods are used since

a the dimension is often very large and

b there is a wish to reduce the noise.

The underlying models and theory are quite easy to understand but to check its correctness is a rather difficult problem. Since one is working with subspace regression it is of crucial importance to get the correct dimension of the subspace. If a wrong decision is made one can make two basic mistakes: too small dimension or too large dimension. In the former case the regression function will be inadequate in the sense that information is lost. If the dimension is too large the variance of the estimates will be too large. As always there is a trade off between variance and unbiasedness. By increasing the bias one could reduce the variance and hence increase the precision of the estimates. If, on the other hand, the bias is too big the answer will be useless. The methods mainly used for assessing the dimension is cross validation, Aikake's Cp, various bootstrap methods etc, see Denham (2000) for a view of the methods used in PLS. One of the problems of estimating the dimension is that the methods are often constructed in such a way that there is no natural method to choose dimension. Our regression method is constructed to use the information gained in the experiment in an efficient way as well as being able to draw conclusions about the dimensionality of the subspaces. The idea is to let the size of the regression function on the normed explanatory variables decide the subspaces. When this space is drained of information, using a method based on the Singular Value Decomposition (SVD), we can construct a test for that the derived subspace really includes the effects we want. For more information on SVD and related topics see Horn and Johnson (1985, 1991). The rational of this method can be viewed from a perspective that might not be the usual in statistics. When solving a regression problem in statistics we usually write it as finding β that minimises $||Y - X\beta||_2$, this is however equivalent to minimising $||\varepsilon||_2$. The drawback with this method is that we force a minimisation on the hole of ε, this might not be necessary or even correct. It might well be the case that we should rather minimise ε projected onto another space, i.e. $||P\varepsilon||_2$ would be the correct quantity to minimise. If this is the case we would actually allow the error to become large in the directions where we cannot control it, then simply ignore it and make it small in the directions that we can control, and finally include these into our model. The principal idea in this article is to let the regression function tell us where it does make a difference and in this way help us to minimise the error where we actually can do this. This keeps us faithful to the old ideas of least squares but on a space where the regression function helps us to understand that it is meaningful. We will show that the criteria for which directions to include will be very simple.

There is a method based on a testing procedure for Principal Components Regression (PCR) called Significance Regression (SR) which will yeild similar estimations, see Faber (2001). SR does not control the multiple level of significance as our method will. The method of derivation is also different. I am not aware of any multivariate eqvivalent of SR whereas our method can be directly transfered to the multivariate case.

2 The model and basic methods

Let us begin with the usual univariate regression model

$$Y = X\beta + \varepsilon \tag{1}$$

where $\varepsilon \sim N(0, \sigma^2 I_n)$, $X \in \mathbb{R}^{n\times p}$, $\beta \in \mathbb{R}^{p\times 1}$ and obviously $Y \in \mathbb{R}^{n\times 1}$. We consider the design matrix X as fixed but not controlled i.e. we have not really designed X but rather it contains far more columns than needed. The last note about not controlled is not necessary but if X contains highly structured data we can use better methods for analysis than the one we are going to construct.

When an experiment has been performed we know X and Y. Any analysis that we perform can only be based on this matrix and vector. As X is thought of as containing the means of which we control the outcomes in Y (disregarding random errors) we see that every meaningful estimate of β must lie in the column space of X. This of course extends to the usual definition of estimability. We hence believe that we from X can get Y apart from a random error. The Ordinary Least Squares (OLS) solution is obtained by finding $\beta \in \mathbb{R}^{p\times 1}$ such that $||Y - X\beta||_2$ is minimised. In some applications X might not have full rank or at least X has a very poor condition number. If this is the case it is well known that OLS will, at least on average, give poor estimates. There are different methods for getting around this problem. One way is to regularize X as in, for instance ridge regression, and hence create a new problem which is easier to solve. We could also use a smaller part of X, i.e. a subspace of the column space of X, as an approximation of X and solve this smaller problem. In both these approaches we make regularizations or put restrictions on X while the problem is about Y or β depending on what we want as end result. In Principal Component Regression (PCR) we use the singular value decomposition (SVD) to give us the dominate space in which we solve our problem. Using a rank k truncated SVD to estimate a matrix gives the best approximation in Euclidean norm but once again I stress that we are not trying to solve any problem for X! It might well be that not all of X:s column space is needed to solve the problem but should we not let Y help decide what parts of X to use? There are methods that incorporate Y into the subspace selection process. One such example is Partial Least Squares (PLS) which stepwise chooses the maximum covariance between X and Y when looking for the correct subspace. But PLS has proven to be rather complicated when it comes to assessing properties as bias or variance of the estimates. See Helland (1988, 1990); Helland and Almøy (1994), von Rosen (1994), Frank and Friedman (1993) and Adolfsson (1999) for more information on PLS. Some effort has lately been put into testing procedure to determine the number of factors in PLS see Aziz and Cléroux (2001) and Denham (2000).

3 A new way of choosing the subspace for regression

The idea is once again choose to a subspace on which we solve the problem. This time however we use the regression function to help us pick out the right space. Using the model (1) we make a SVD of X, that is

$$X = U\Lambda V', \quad U'U = I_n, \quad V'V = I_p.$$

where

$$\Lambda = \begin{pmatrix} D & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \in \mathbb{R}^{n\times p}, \quad D = \mathrm{diag}(\lambda_1, \ldots, \lambda_r),$$

$r \leq \min(n,p)$ and $\lambda_1 \geq \ldots \geq \lambda_r > 0$. For notational convenience we adopt the the notation

$$\Lambda^{-1} = \begin{pmatrix} D^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \in \mathbb{R}^{p\times n}.$$

The matrix U_rU_r', where U_r is the matrix formed by the r first columns, is the orthogonal projection onto the column space of X which is the space where the estimate of β must lie. The OLS estimate of β is $\hat{\beta} = (X'X)^{-1}X'Y$. We now want to study the behaviour of this estimate on the level curves of $X'X$ or rather the normalised level curves. From the SVD of X we have that $X'X = V\Lambda^2V'$. We now introduce an artificial variable $Z = XV\Lambda^{-1}$ and solve the equation

$$Y = Z\gamma + \varepsilon$$

instead. This equation has the solution

$$\hat{\gamma} = Z'Y = \Lambda^{-1}V'X'Y = \Lambda^{-1}\Lambda U'Y = U'Y.$$

This vector describes the direction in the column space of X on which the estimated regression function is maximal. This however only gives us one direction which in general is far to small a space for solving the problem. We will however use this general idea to get the space on which we will solve the problem. Instead of choosing only the maximal direction in $S = col(X)$ we will choose the K-dimensional orthogonal base of a subspace R of S that best approximates the maximum normed regression function in a $||\cdot||_2$ sense.

We set $U_r = [u_1, \ldots, u_r]$ and project Y onto these vectors one at a time, hence $r_i = u_i u_i' Y, \quad i = 1, \ldots, r$. Breaking this down into parts we examine the scalar $u_i'Y$ more carefully. According to (1) we have

$$\begin{aligned} u_i'Y &= u_i'(X\beta + \varepsilon) = u_i'(U\Lambda V'\beta + \varepsilon) = e_i'\Lambda V'\beta + u_i'\varepsilon = \lambda_i e_i' V'\beta + u_i'\varepsilon \\ &= \lambda_i v_i'\beta + u_i'\varepsilon = f_i(\beta) + u_i'\varepsilon. \end{aligned} \tag{2}$$

We see that $\lambda_i v_i'\beta$ is a scalar describing how great effect the regression function has along the vector u_i and that $u_i'\varepsilon \sim N(0, \sigma^2 \sum_j u_{ij}^2) = N(0, \sigma^2)$ since u_i has length 1. Further more, letting $\tilde{\varepsilon}_i = u_i'\varepsilon$, we have that $\tilde{\varepsilon}_i$ and $\tilde{\varepsilon}_j$ are uncorrelated (and hence independent) for $i \neq j$ since $u_i \perp u_j$ for $i \neq j$. The vectors $r_1, \ldots, r_r$ now span S but they also carry information about the validness of the space in the sense that there length is proportional to the size of the estimated regression function in the space. We form a new matrix by $R = [r_1, \ldots, r_r]$ and do a SVD on this matrix. In matrix analysis one normally says that a square matrix T is orthogonal if $T'T = I$, I personally think this is a orthonormal matrix rather than an orthogonal. In this nonstandard notation we see that R is an orthogonal matrix, this makes the SVD simple since we can do it by inspection:

$$R = U_R D P_n, \quad D = \text{diag}(d_{(i)}).$$

The matrix U_R has the same columns as U but not necessarily in the same order and the d_i:s are the absolute values of $f_i(\beta) + \tilde{\varepsilon}_i$ ordered according to their size and the matrix P_n is a permutation matrix. This representation is useful since this tells us that if we use the TSVD with dimension k of R we can not do better in $||\cdot||_2$-sense when estimating the space containing the biggest estimated regression function. Another big advantage of this approach is that the distributional properties are easy to handle since we have independence. In a setting where n is moderately large and we from prior knowledge know that large portions of X does not effect Y we can easily derive a testing procedure, for choosing the correct k, which maintains the correct multiple level of significance (see Adolfsson (2002b)). Assuming that we have now fixed a subspace R on which to solve the problem we now just project the equation onto this subspace and solve the problem there.

4 Multivariate regression

With some minor modification the same method and arguments goes through for multivariate regression. We now assume that we have a model

$$Y = XB + E, \quad E \sim N_{r,n}(0, \sigma^2 I_n, I_r)$$

where $Y = (Y_1, \ldots, Y_r) \in \mathbb{R}^{n \times r}$, $B = (\beta_1, \ldots, \beta_r) \in \mathbb{R}^{p \times r}$ and $E = (\varepsilon_1, \ldots, \varepsilon_r)$. Using the same notation as in the previous section we have

$$\sum_{j=i}^{r} (u_i' Y_j)^2 = \sum_{j=1}^{r} (\lambda_i v_i' \beta_j + u_i' \varepsilon_j)^2 = \sum_{j=1}^{r} (\lambda_i v_i' \beta_j + \tilde{\varepsilon}_i)^2 \tag{3}$$

if we once more set $X = U \Lambda V'$. Let us now assume that there are no true regression effects. We will then have that each of the $\sum_{j=i}^{r} (u_i' Y_j)^2 t$ is χ_r^2 and independent. If we now use the k (say) smallest of these variables to estimate the standard deviation σ we will have the following likelihood function

$$L(\sigma) = \binom{n}{k} \prod_{i=1}^{k} \left(\frac{1}{\sigma^2} f_{\chi_r^2} \left(\frac{x_i}{\sigma^2} \right) \right) \left(1 - F_{\chi_r^2} \left(\frac{x_k}{\sigma^2} \right) \right)^{n-k}.$$

The log-likelihood function, disregarding constants, will then be

$$\begin{aligned} l(\sigma) &= -2k \ln(\sigma) + \sum_{i=1}^{k} \left(\frac{k}{2} - 1 \right) \ln \left(\frac{x_i}{\sigma^2} \right) - \sum_{i=1}^{k} \frac{x_i}{2\sigma^2} \\ &+ (n-k) \ln \left(1 - F_{\chi_r^2} \left(\frac{x_k}{\sigma^2} \right) \right). \end{aligned}$$

Setting $\frac{\partial l}{\partial \sigma} = 0$ we get

$$\sigma^2 = \frac{k}{2(n-k)} (r\sigma^2 - S_k^2) \frac{1 - F_{\chi_r^2} \left(\frac{x_k}{\sigma^2} \right)}{\frac{x_r}{\sigma^2} f_{\chi_r^2} \left(\frac{x_r}{\sigma^2} \right)}$$

where

$$S_k^2 = \frac{1}{k}\sum_{i=1}^{k} x_i.$$

Using lemma 1 and lemma 2 in Adolfsson (2002b) and noting that the failure rate of the gamma distribution is non decreasing for the parameter values of interest we can construct a test for true regression effects that maintains the correct multiple level of significance.

4.1 Efficiency of the variance estimate for multiple testing

If we have a regression experiment where we know that a large portion of X is non informative the estimate of σ above will probably be rather good, but how large portion do we need? One way of answering this question would be to examine the asymptotic variance of the estimate. It is well known that, under certain regularity conditions, we asymptotically achieve the Cramér-Rao limit for ml-estimates. We will now calculate this for the estimation method given above. We have that

$$-\frac{\partial^2 l}{\partial \sigma^2} = \frac{3}{\sigma^2}\sum_{i=1}^{k}\frac{x_i}{\sigma^2} - \frac{kr}{\sigma^2} - \frac{2}{\sigma^2}\frac{\frac{x_k}{\sigma^2} f_{\chi_r^2}\left(\frac{x_k}{\sigma^2}\right)}{1-F_{\chi_r^2}\left(\frac{x_k}{\sigma^2}\right)}\left(\frac{x_k}{\sigma^2} - r - 1 - \frac{\frac{x_k}{\sigma^2} f_{\chi_r^2}\left(\frac{x_k}{\sigma^2}\right)}{1-F_{\chi_r^2}\left(\frac{x_k}{\sigma^2}\right)}\right).$$

We now assume that $k/n = p \in (0,1)$ and let $k, n \to \infty$. By the law of large numbers we will then have that $x_k/\sigma^2 \to z_\infty = F_{\chi_r^2}^{-1}(p)$, $\frac{1}{n}\sum_{i=1}^{k} X_i/\sigma^2 \to \mathbb{E}[Z|Z < z_\infty]$, $(Z \sim \chi_r^2)$ and hence, $F_{\chi_r^2}(x_k/\sigma^2) \to p$. We now have the following asymptotic expression

$$\begin{aligned}
-\frac{1}{n}\frac{\partial^2 l}{\partial \sigma^2} &= \frac{1}{\sigma^2}\Bigg(3\mathbb{E}[Z|Z < z_\infty] - pr \\
&+ z_\infty f_{\chi_r^2}(z_\infty)\left(2\frac{z_\infty f_{\chi_r^2}(z_\infty)}{1-F_{\chi_r^2}(z_\infty)} + r + 1 - z_\infty\right)\Bigg) \\
&= \frac{1}{\sigma^2} f(p).
\end{aligned}$$

It is easy to see that $-\frac{1}{n}\frac{\partial^2 l}{\partial \sigma^2} \to 0$ as $p \to 0$ but the limit $p \to 1$ might not be as obvious. We have however that

$$\frac{f_{\chi_r^2}(z_\infty)}{1-F_{\chi_r^2}(z_\infty)} = \frac{c(z_\infty)^{\alpha-1}e^{-\frac{z_\infty}{2}}}{\int_{z_\infty}^{\infty} cu^{\alpha-1}e^{-\frac{u}{2}}du}$$

$$
\begin{aligned}
&= \frac{1}{\int_0^\infty \left(1+\frac{v}{z_\infty}\right)^{\alpha-1} e^{-\frac{v}{2}} dv} \\
&\leq \frac{1}{\int_0^\infty e^{-\frac{v}{2}} dv} = 2
\end{aligned}
$$

since $\alpha \geq 1$ and $z_\infty \geq 0$ and hence it is clear that $-\frac{1}{n}\frac{\partial^2 l}{\partial \sigma^2} \to \frac{2r}{\sigma^2}$ as $p \to 1$. Since the ordinary estimate of the variance using all observations has the same variance as our estimate when $p = 1$ we do not seem too aim to much off target. A plot of the variance factor $f(p)$ for different degrees of freedom can be found in figure 1.

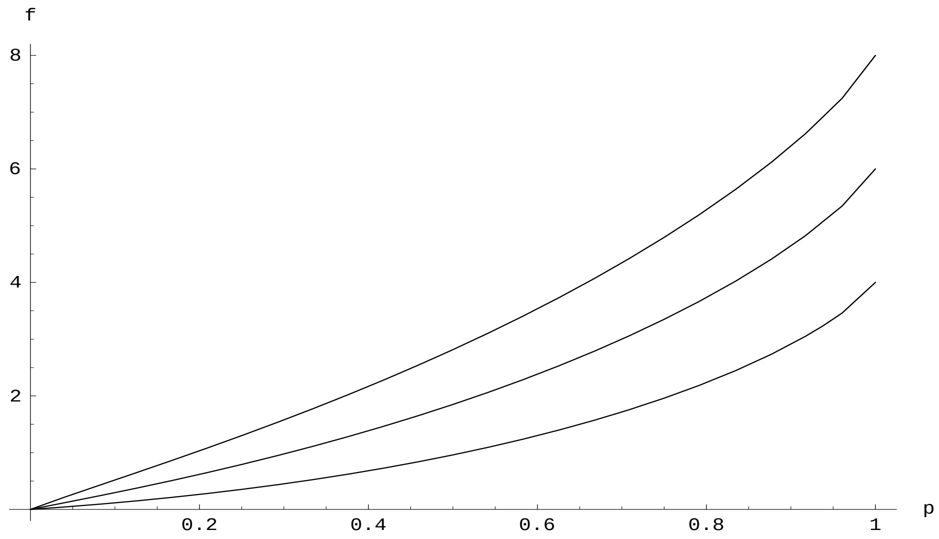

Figure 1: Asymptotic variance factor for 2, 3 and 4 degrees of freedom.

5 Prediction properties

In applications such as, for instance calibration, the predictive ability and properties are of great importance. We will now examine the Mean Square Error of Prediction (MSEP). The MSEP is defined as

$$
\text{MSEP} = \mathbb{E}\left[||Y_0 - \hat{Y}_0||_2^2 \right].
$$

Let us assume that the new observations that we want to predict are new observations in the same sample points as used for the estimation. In the case of calibration this would make sense since we would calibrate where we want to measure. Let $\beta_{\mathcal{I}} = \mathbb{E}\left[\hat{\beta}_{\mathcal{I}}\right]$ where $\hat{\beta}_{\mathcal{I}}$ is the estimate using $u_i : i \in \mathcal{I}$, $|\mathcal{I}| = k$ where $\mathcal{I}$ is a (fixed) set of indexes $1 \leq |\mathcal{I}| \leq n$. We then get

$$\text{MSEP} \propto \mathbb{E}\left[(\hat{\beta}_{\mathcal{I}} - \beta_{\mathcal{I}})' X' X (\hat{\beta}_{\mathcal{I}} - \beta_{I})\right] + \mathbb{E}\left[(\beta - \beta_{\mathcal{I}})' X' X (\beta - \beta_{\mathcal{I}})\right]. \quad (4)$$

We now rewrite the estimate $\beta_{\mathcal{I}}$ using the following observation: If $X = U \Lambda V'$ we get the OLS solution as $\beta = X^{+} Y$ where X^{+} is the Moore-Penrose generalised inverse. This inverse is given by $X^{+} = V \Lambda^{-1} U'$ (see Horn and Johnson (1985) p.421) and hence we have that

$$\hat{\beta} = \sum_{i=1}^{n} \frac{u_i' Y}{\lambda_i} v_i.$$

We then have that

$$\hat{\beta}_{\mathcal{I}} = \sum_{i \in \mathcal{I}} \frac{u_i' Y}{\lambda_i} v_i = \sum_{i \in \mathcal{I}} \frac{\lambda_i v_i' \beta + \tilde{\varepsilon}_i}{\lambda_i} v_i = \sum_{i \in \mathcal{I}} v_i' \beta v_i + \sum_{i \in \mathcal{I}} \frac{\tilde{\varepsilon}_i}{\lambda_i} v_i.$$

One can now immediately see that

$$\mathbb{E}\left[\hat{\beta}_{I}\right] = \sum_{i \in \mathcal{I}} v_i' \beta v_i \text{ and } \mathbb{E}\left[\beta - \hat{\beta}_{\mathcal{I}}\right] = \sum_{i \notin \mathcal{I}} v_i' \beta v_i.$$

Noting that $X'X = V \Lambda^2 V'$ we have that

$$\begin{aligned} \mathbb{E}\left[(\hat{\beta}_{\mathcal{I}} - \beta_{\mathcal{I}})' X' X (\hat{\beta}_{\mathcal{I}} - \beta_{\mathcal{I}})\right] &= \mathbb{E}\left[\left(\sum_{i \in \mathcal{I}} \frac{\tilde{\varepsilon}_i}{\lambda_i} v_i\right)' V \Lambda^2 V' \left(\sum_{i \in \mathcal{I}} \frac{\tilde{\varepsilon}_i}{\lambda_i} v_i\right)\right] \\ &= \sum_{i \in \mathcal{I}} \mathbb{E}\left[\tilde{\varepsilon}_i^2\right] = |\mathcal{I}| \sigma^2, \end{aligned} \quad (5)$$

and

$$\begin{aligned} (\beta - \beta_{\mathcal{I}})' X' X (\beta - \beta_{\mathcal{I}}) &= \left(\sum_{i \notin \mathcal{I}} v_i' \beta v_i\right)' V \Lambda^2 V' \left(\sum_{i \notin \mathcal{I}} v_i' \beta v_i\right) \\ &= \sum_{i \notin \mathcal{I}} (\lambda_i v_i' \beta)^2. \end{aligned} \quad (6)$$

Recalling that $u_i Y = \lambda_i v_i' \beta + \tilde{\varepsilon}_i$ the relationship between effect and variance, with respect to prediction, is now clear. If we include a new direction u_l we will decrease the squared bias with a factor $(\lambda_l v_l' \beta)^2$, according to (6), but at the same time we will increase the variance with a factor σ^2 due to (5). These observations together with (4) gives our decision rule for prediction: if the absolute value of an effect is less then the standard deviation it should be excluded, and if the effect exceeds the standard deviation it should be included! This is obvious since if we include a direction that includes an effect less then σ the increase in variance is σ^2 but the decrease in bias is less then σ^2 and vice versa. This also gives the implication that when the testing procedure is used the power should be great for alternatives where the effect exceeds the standard deviation for other alternatives we should not include the effect.

6 With a stochastic X

Let X be a stochastic matrix with a distribution that is independent of ε. There exists a SVD of $X = U\Lambda V'$ where U, Λ and V are independent of ε. In the case when $X\beta = 0$ eq. (2) gives

$$\frac{1}{\sigma^2}(u_i' Y)^2 = \frac{1}{\sigma^2}(u_i'\varepsilon)(\varepsilon' u_i) = \frac{1}{\sigma^2}\frac{u_i'\varepsilon\varepsilon' u_i}{u_i' u_i} \sim \chi_1^2$$

since $\varepsilon\varepsilon'$ has a Pseudo-Wishart distribution see, for instance, Kshirsagar (1972) for a proof. Note that this result is valid for any stochastic orthogonal set of vectors $u_i, \ldots, u_n$ which is independent of ε.

Thus the null case is clear. When there are true regression effects present it is not clear what happens. Since the u_i:s are stochastic the resulting distribution will be of the type $\chi_r^2(\delta)$ but δ is stochastic. In simulations this has show to have small effect on the method. The result in section 5 are however no longer valid.

7 A spectroscopic example

The data set used in this example consists of 440 measurements of organic solvent residue in penicillium samples. The concentration of four different solvents where measured in each sample. The different samples where then analysed in a NIR-spectrometer and the absorbation at 1050 wavelengths where recorded for each sample. Since the the absorbation follows Beer's law we have the setup

$$Y = X\beta, Y \in \mathbb{R}^{440\times 4}, X \in \mathbb{R}^{440\times 1050}.$$

The goal is to estimate β, i.e. the to calibrate the spectrometer. The condition number of a matrix is the quotient of the largest and the smallest singular values of the matrix. The larger the condition number the more ill-condition is the matrix. The condition number of X is $6.0 \cdot 10^5$ which means that X is rather ill-conditioned. The singular values $\lambda_3, \ldots, \lambda_{440}$ of X is displayed in figure 2. Using eq. (3) in section 4 to form the information

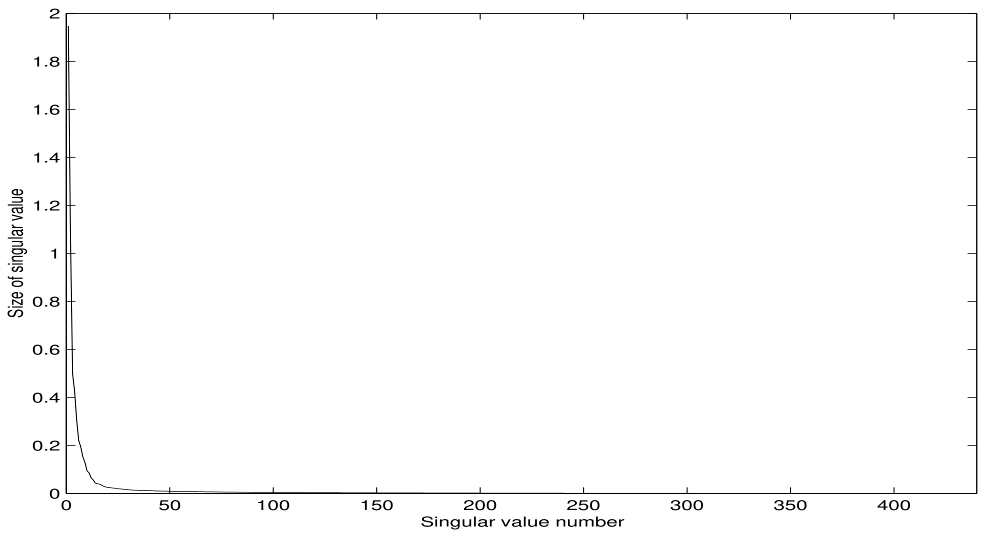

Figure 2: The singular values of X.

in each of the singular vectors u_i we can study the size of the singular values when they are reordered by the information. In figure 3 we see the singular values ordered by the information criteria and vice versa. The figure should be viewed column vise. The top right plot displays $\lambda_3, \ldots, \lambda_{40}$ and the bottom right plot shows the u_iY:s ordered according to the λ:s. The top left plot shows the λ:s ordered according to the size of the u_iY plotted in the bottom leftmost plot. The two largest effects are excluded in all plots since the coincide and are so large that they make smaller observations difficult to plot. In these plots one for instance see that one of the largest singular values occurs, when sorted by the u_iY:s, on 23:th place. Large singular values is no guarantee that there is information present.

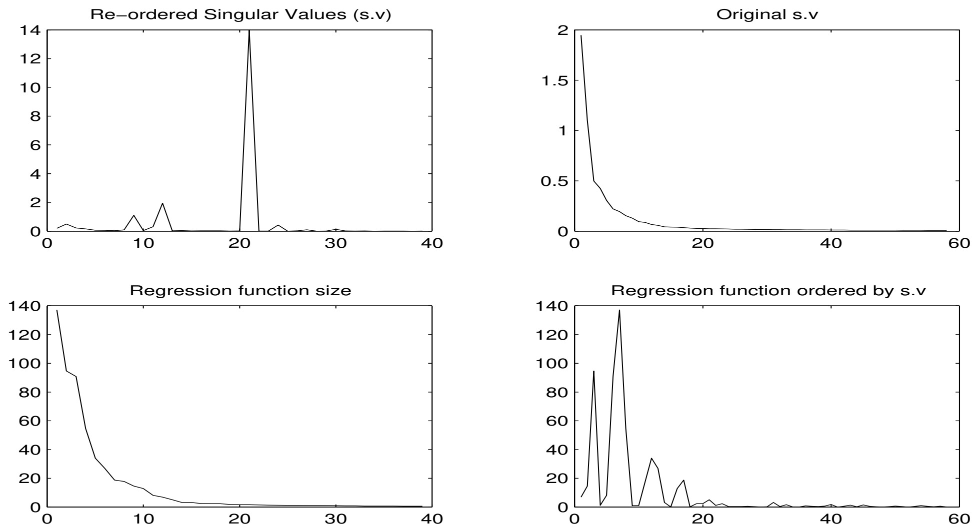

Figure 3: Reordering of the singular values.

Comparing this regression method to that of PCR we see that the residual sum of squares is smaller for the same number of base vectors (components). Using the step-down test procedure with the smallest 250 observations of the statistics for variance estimation we get that we can declare the first 71 base vectors as significant. Using the test procedure in Adolfsson (2002a) which is used for determining the number of components to use in PCR we get 27.

8 Conclusions

The regression method purposed in this paper shows great potential both theoretically and practically. In the case of a non-stochastic X matrix the predictive properties is easy to asses. The method uses information from both the dependent and independent variables but in a way that is easy to understand. It should be noted that any orthogonalisation of X can be used as long as it is independent of Y. If we are fitting a linear model where we have particular interest in quadratic terms of the x_i:s we could choose an orthogonalisation where we the quadratic terms are one of the orthogonal base vectors. This might be fruitful when we have prior knowledge of the problem. From a computational point of view it is $\max(n, p)$, $X \in \mathbb{R}^{n \times p}$ that determines the complexity of the problem. In, for instance, spectroscopical problems the limiting factor is normally the number of samples in the calibration set. if the sample size is moderate (< 500) the method is fast to calculate. If we have prior knowledge of the number of the rank of similar problems we could use the step-down method described in the paper. The step-down method does not disturb the structure of either X or Y which might be the case in cross validation, see Adolfsson (1999). On a reasonable fast computer the simulations for the step-down method is done rather fast.

References

T. Adolfsson. Partial least squares and its implementations from a statistical point of view. Technical report, Chalmers University of Technology and Göteborg University, 1999.

T. Adolfsson. *A testing procedure for dermening the chemical rank of spectrometrical absorption matrices - a heuristic approach.* PhD thesis, Chalmers University of Technology and Göteborg University, 2002a.

T. Adolfsson. *Variance estimation and multiple inference testing in saturated orthogonal designs.* PhD thesis, Chalmers University of Technology and Göteborg University, 2002b.

L. Aziz and R. Cléroux. The pls multivariate regression model: testing the significans of successive pls components. *J. Chemometrics*, 2001.

M. C. Denham. Choosing the number of factors in partial least squares regression: estimating and minimizing the mean squared error of prediction. *J. Chemometrics*, 14:351–361, 2000.

N. M. Faber. Critical evaluation of a significance test for partial least squares regression. *Anal. Chim. Acta*, pages 235–240, 2001.

I. E. Frank and J. H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 1993.

I. S. Helland. On the structure of partial least squares regression. *Comm. Statist. B-simulation Comput.*, 1988.

I. S. Helland. Partial least squares regression and statistical models. *Scand. J. Statist.*, 1990.

I. S. Helland and T. Almøy. Comparison of prediction methods when only a few components are relevant. *J. Amer. Statist. Assoc.*, 1994.

R. A. Horn and J. A Johnson. *Matrix analysis.* Cambridge university press, 1985.

R. A. Horn and J. A Johnson. *Topics in matrix analysis.* Cambridge university press, 1991.

A. M. Kshirsagar. *Multivariate Analysis*. Marcel Dekker, 1972.

D. von Rosen. Partial least squares and a linear model. In *Proceedings at the International Conference on Linear Statistical Inference LINSTAT 93*, 1994.

Paper D

A testing procedure for determining the chemical rank of spectroscopic absorption matrices - a heuristic approach

Tobias Adolfsson
Göteborg University and Chalmers University of Technology

28th February 2002

Abstract

We will discus a method for determining location shifts in normally distributed matrices for use with apsorption matrices. A testing methodology will be produced and argued for.

1 Determining chemical rank

In spectrometrical investigations the absorption of light of different wavelengths are measured. The laws of chemistry gives that the rank of the matrix of absorption is equal to the number of photocromatic species in the sample. If we where analysing a pure sample of the chemical of interest and the measurement apparatus would be without measurment errors this would be a true. In reallity no solutions are pure and no apparatus is without errors and hence the rank of the matrix will differ from the chemical rank. This problem is closely relatad to the problem of determining the number of components to use in Principal Component Regression (PCR). The law that controls the absorption i.e. Beer's law, makes it natrual to use factorisation methods.

When using sub-space methods, i.e. methods solving the problem on a smaller space than the original space, one needs a method to estimate the size of the sub-space. In many applications the predictive properties are the propperties that is wished to be optimal. In many methods to asses a stop criteria different estimates of the Mean Squared Error of Prediction (MSEP) is used. The most common way of of estimating MSEP is through cross validation, see Denham (2000) for an overview. It is known that cross validation might give incorrect estimates of the minimum of the MSEP, see Adolfsson (1999) or Faber (2001).

In the case of PCR there is a stepdown test that can be used. This test uses the fact that the eigenvalues and eigenvectors used is given by a deterministic matrix which used in a correct way will produce a t-test. The same method has been used for PLS but here the basevectors for the regressionspace is not deterministicly choosen so the theory does not hold, see Faber (2001). It must be recognised that when testing in such situations as this, the concept of multiple inference must play a central role. In many investigations a large number of hypothesis are tested and it is plausible that many rejections might occur only by chance if the inference is based on individual tests.

One should also be aware (as a statistician) that the research communities that use these method are more result than theory oriented. Statisticians rather want to test in a fashion that can be understod while reasearchers tend to want a good answer. In deciding for a method to asses the regression space one must remember that in solving a collinear regression problem the main idea is not to include all theoretical effects but rather to include all the effects that are big enough to stabilise the problem. Including to small effects will give such increase in variance that the solution becomes useless in a practical sense. A test should hence be constructed to be powerfull against alternatives where the effect is as large as the level of variance. The lack of such method is due to the complexness of the distributions arising in these problems. An approach that becomes more and more tractable with the growing speed of computers is to use a statistic with the correct theoretical properties and that can be simulated. It is my belief that this is the necessary way of moving this field forward.

One should be aware of the problem of estimating the chemical rank of the matrix and that of estimating the rules that governs the concentrations in a spectrometrical trial. The chemical rank gives the number of photocromatically active species in the trial. This is not necessarily the same as the number of photocromatilly active *interesting* species.

2 Intuitive background of the test procedure

We start with some well known facts from elemetary linear algebra. Let A be a $n \times n$ real matrix s.t. $A' = A > 0$. We consider the function $f(\mathbf{x}) : \mathbb{R}^{\mathbf{n}\times\mathbf{n}} \to \mathbb{R}$ where $f(\mathbf{x}) = \mathbf{x}'A\mathbf{x}$. If wee set $n = 2$ and plot the function for $|\mathbf{x}| = 1$ we get an ellipsoid of the type shown in figure 1. Since $A > 0$

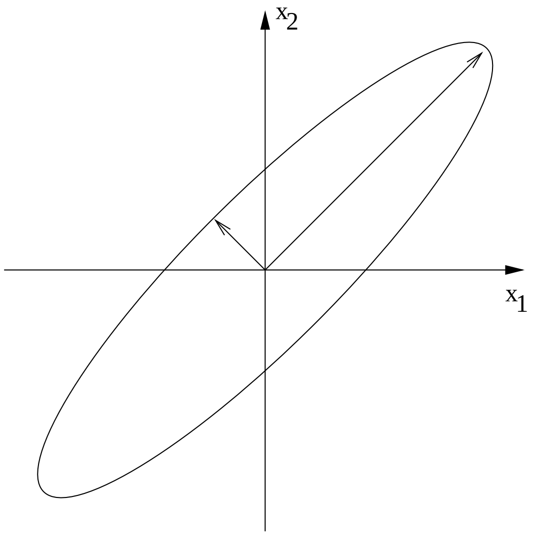

Figure 1: A plot of $f(\mathbf{x})$

we know that the eigenvalues $\lambda_1 > \lambda_2$ of A are positive. The length of the largest of the principal axis of the ellipsoid is given by λ_1 and the shortest principal axis by λ_2. The direction of the axis correspond to the eigenvectors of the matrix. A measure of "the stretch" of the ellipsoid is given by λ_1/λ_2 which, in numerical literature, is reffered to as the condition number of the matrix A. The same thing obviously hold for $n > 2$ but plotting becomes impossble (at least for $n > 3$). In the case $n > 2$ we let $\lambda_1 > \ldots > \lambda_n > 0$ be the eigenvalues of A and the condition number is now given by λ_1/λ_n. A good thing with the condition number and similar measures is that of the matrix is stretched out by a scalar, i.e. $A = \alpha B$, the condition number of A and B are the same.

The time has now come to start with stochastic matrixes. Let us assume that $W \sim W_p(\mathbf{0}, I_p\sigma^2, n)$, $n > p$. A matrix with this Wishart distribution will a.s. be positively definite and its eigenvalues are a.s. distinct. The distribution of the eigenvalues is known but rather complicated to use for direct calculations. We will try to avoid going into detail of the distribution but rather se what we can do in general. One well known fact of the Wishart distribution is that if $W_1 \sim W_p(\mathbf{0}, I_p, n)$ then $\sigma^2 W_1 \overset{\mathcal{D}}{=} W$. This gives that if λ_i is a (stochastic) eigenvalue of W_1 then $\sigma^2\lambda_i$ is a (stochastic) eigenvalue of W. We now know that any statistic on the form $\frac{\lambda_1}{\lambda_i}$ does not depend on σ. To make the qoutient more stable we study

$$R = \frac{\lambda_1}{\sum_{i=n-r}^{n} \lambda_i} \tag{1}$$

which also is independent of σ. This is also a measure of the skewness of the ellipsoid but more stabilised than using just λ_n in the denominator. The only thing that we know about this statistic is that it is independent of σ but we also now how to generate a sample of stocastic variable with this distribution in a computer.

3 Theoretical properties of the test procedure

We will construct our test procedur using the statistic defined in (1). As mentioned before this statistica gives a measure of how skew the level curves of the quadratic form given by the matrix is. The statistic gives a bound for the maximal natrual variation and the idea is to see what directions exceeds this natrual variation bound. This problem is in chemometrics known as the problem of determining the chemical rank of a matrix. The test procedure will not be proven to work with the correct multipel level of significans but will be made likely to work. The problem with stating a formal proof of the level of significans is due to the problem of making statements about the singular values of the sum of two matrices. I will however give a intuitive explanation of why this should work. The first assumptions that is made is that the matrix of measurements is actually a sum of two matrices. Let Y be a $n \times p$ matrix where each row comes from independent $N_n(\mathbf{0}, \sigma^2 I_p)$ distributions. Let M be a deterministic matrix of size $n \times p$. Let the rank of M be r. Let the matrix of measurements $X = M + Y$. The problem is to find r given X. For computional convenience I choose to work with the singular values of X. There is no *theoretical* difference since the singular values of X

is mearly the square root of the eigenvalues of $X'X$ but practically we save a matrix multiplication in computing.

As mentioned earlier the correctness of the testing procedure is hard to make rigorous. However, there are some reasons why the method should work. There is a version of the Courant-Fisher theorem for singular values that will come in handy to understand how the method works.

Theorem 1 *Let A be a $n \times m$ matrix, let $\sigma_1(A) \geq \sigma_2(A) \geq \ldots$ be the ordered singular values of A, and let k be a given integer with $1 \leq k \leq \min\{m,n\}$. Then*

$$\sigma_k(A) = \min_{w_1,\ldots,w_{k-1}\in\mathbb{C}^n} \max_{\substack{x\in\mathbb{C}^n \\ ||x||_2=1 \\ x\perp w_1,\ldots,w_{k-1}}} ||Ax||_2 \tag{2}$$

$$= \max_{w_1,\ldots,w_{n-k}\in\mathbb{C}^n} \min_{\substack{x\in\mathbb{C}^n \\ ||x||_2=1 \\ x\perp w_1,\ldots,w_{n-k}}} ||Ax||_2 \tag{3}$$

$$= \min_{\substack{S\subset\mathbb{C}^n \\ \dim\{S\}=k}} \max_{\substack{x\in S \\ ||x||_2=1}} ||Ax||_2 \tag{4}$$

$$= \max_{\substack{S\subset\mathbb{C}^n \\ \dim\{S\}=k}} \min_{\substack{x\in S \\ ||x||_2=1}} ||Ax||_2 \tag{5}$$

From this we see that, if $\sigma_1(A)$ is the biggest singular value, (2) above reduces to $\sigma_1(A) = \max_{||x||_2=1} ||Ax||_2$. If we to the matrix A add a "dominant direction", which corresponds to a true (large) effect then $\sigma_1(A)$ will increase. If this dominant direction is unequal to the direction x used for calculating $\sigma_1(A)$ any succeeding $\tilde{\sigma}_k(A), \quad k \neq 1$ will be larger than the original $\sigma_k(A)$! This is a key element in the reasoning for multipel inference. We will now study this phenomena in greater detail. Let us assume that Y is defined as earlier. The product YY'/n is a estimation of $\sigma^2 I_p$. The rows of Y will not show any systematic variations since they estimate zero i.e. $\frac{1}{n}\sum_{i=1}^{n} Y_{ij} \to 0$ a.s.. If we now add a matrix $M = (m_1', m_2', \ldots, m_n')'$ to Y where the rows in M have the structure $m_i = \alpha m_j$, $\alpha > 0$ for some $i \neq j$ the column sums no longer estimate zero (or at least need not). This assumption might seem strange but is a higly relevant assumption when dealing with spectroscopic data. The assumption is meerly stating that Beer's law is walid which is a assumption that is made in most (not to say all) spectroscopic investigations.

If the rank of M is $r < \min(n,p)$ this will tend to make a r-dimensional space have a larger spread than the other directions. This will happen since, if we once again look at $X'X$, the other dimensions meerly estimates a ball with radius σ^2. This means that the matrix M only affects a r dimensional space and leaves the rest of the dimensions unchanged. This will change the directions of the largest singular values if the size of the m_i:s is large enough compared with σ^2. Since the directions for the large singular values are changed the optimisation in theorem 1 for the smaller singular values that does not contain any true effects will be in other directions which now might include the larger directions in Y and hence the singular values will be larger.

Theorem 2 (Multipel test procedure) *Let $Y \sim N_n(\mathbf{0}, \sigma^2 I_p)$ with $n < p$ and $M \in \mathbb{R}^{n\times p}$ be such that the assumptions above are fulfilled. Let $\sigma_1(X) \geq \ldots \geq \sigma_p(X) > 0$ be the singular values of $X = Y + M$. Then testing the hypothesis $H_{0i} : \sigma_i(X) = \sigma_i(Y)$ against $H_{1i} : \sigma_i(X) = \sigma_i(\tilde{Y}) + \mu(M)$ by the procedure reject H_{0i} if $\sigma_i(X)/Q_X > e_\alpha$, where $Q_X = \sum_{i=n}^{n-k} \sigma_i(X)$ and e_α is chosen such that*

$$\mathbb{P}\left\{\frac{\sigma_1(Y)}{Q_Y} > e_\alpha\right\} = \alpha,$$

maintains the multipel level of significans α.

Proof: By the definition of e_α the overall conjunction trivially has the correct level of significans. So let us now assume that some H_{1i} is true. From the assumptions above we have that $Q_Y \leq Q_X$. Let $\tilde{Y}$ be a s.v. such that $\sigma_i(X) = \sigma_i(\tilde{Y}) + \sigma_i(M)$. Let us now study the artificial test based on $\tilde{Y}$. This test satisfy

$$\begin{aligned}
\mathbb{P}\left\{\frac{\sigma_i(\tilde{Y})}{Q_{\tilde{Y}}} \leq e_\alpha \forall i\right\} &= 1 - \mathbb{P}\left\{\exists i : \frac{\sigma_i(\tilde{Y})}{Q_{\tilde{Y}}} > e_\alpha\right\} \\
&= 1 - \mathbb{P}\left\{\frac{\max_i \sigma_i(\tilde{Y})}{Q_{\tilde{Y}}} > e_\alpha\right\} \\
&\geq 1 - \mathbb{P}\left\{\frac{\sigma_1(Y)}{Q_Y} > e_\alpha\right\} = 1 - \alpha.
\end{aligned}$$

By construction we have that $\sigma_i(X)/Q_X \leq \sigma_i(\tilde{Y})/Q_{\tilde{Y}}$ if H_{0i} is true and hence the real testing procedur will give no more rejections than the artificial one.

Q.E.D.

4 Simulations results

Simulations where used to examine the performance of the testing procedure. The following algorithm was used:

1. Generate $B_1', \ldots B_r' \sim N_p(0,1)$ (base vectors) and $U_1, \ldots U_r \sim U_n(0,1)$. Set $M = U'B$.

2. FOR I=1 TO number of samples DO

   (a) SET $E = N_n(0, \sigma^2 I_p)$ AND $X = M + E$.

   (b) CALCULATE data=SVD(X).

   (c) SET $S = \sum_{i=n-k+1}^{n} \lambda_i$ AND STAT= $\frac{\lambda_1}{S}$.

   (d) DO TEST.

   (e) SAVE number of rejections.

3. calculate statistics.

This means that we will have a different deterministic matrix each time we run the program. In the simulations $n = 47$, $p = 28$ and $k = 14$ where used. When r was set to zero the rejection rate was α (in the simulations set to 10%). With $r = 1$ the false rejection rate was meerly 1% and when $r = 2$ the false rejection rate was 0.1%. For $r \geq 3$ no false rejections where made. The power is good when the true rank of the matrix M is small compared to the total number of singular values. This can be explained with use of theorem 1. The true effects move the space for the minimisation of the smallest singular values and hence the smallest singular values will be to large, thus making the testing procedure conservative.

We now set $n = 50$, $p = 40$, $k = 25$, $r = 10$, and $\sigma = 0.5$. The mean number of the estimated rank is now 8.65. A plot of the 28 largest normed singular values and the rejection boundary for $\alpha = 0.10$ of one test realisation can be found in figur 2. Note that the plotted normed singular values have no distinct bend even thou it contains skifts. In this case it would difficult to estimate the dimension by inspection.

5 An example

The following example is taken from Hjorth (1994) and consern paper industri. A complete description of the dataset can be found on pages 74-76

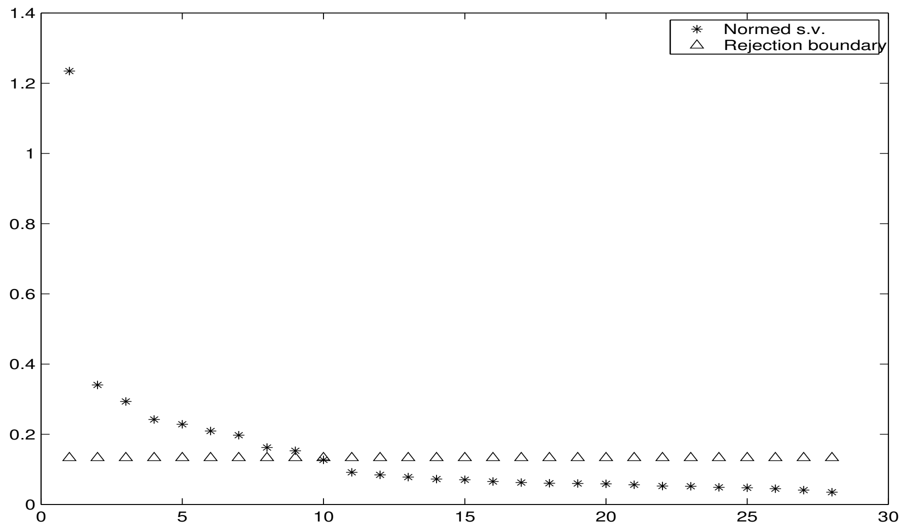

Figure 2: Plot of statistic and boundary for simulated data

in Hjorth (1994). The data is 48 observations on a 28-dimensional vector of variates. The testing procedure using $\alpha = 0.10$ and $r = 14$ gives that the estimated true rank of the data is 8. A plot of the statistics and the boundary can be found in figur 3

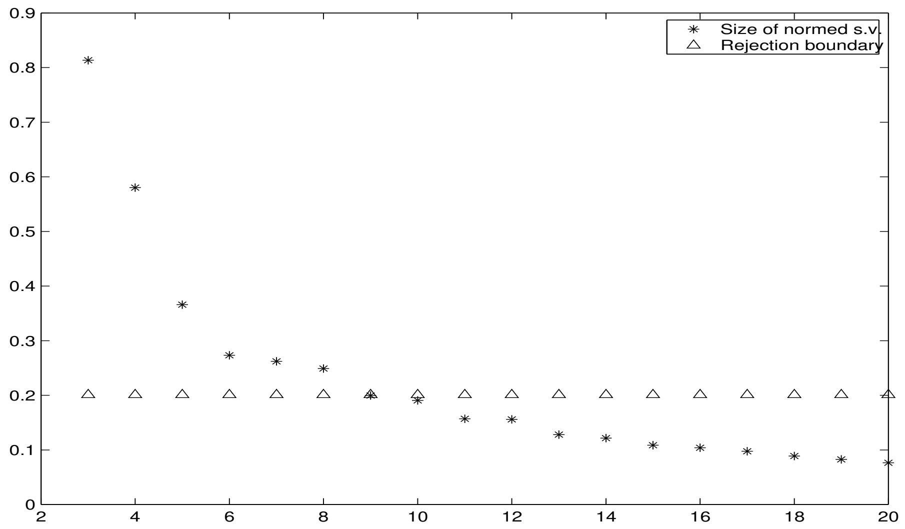

Figure 3: Plot of statistic and boundary for paper data

References

T. Adolfsson. Partial least squares and its implementations from a statistical point of view. Technical report, Chalmers University of Technology and Göteborg University, 1999.

M. C. Denham. Choosing the number of factors in partial least squares regression: estimating and minimizing the mean squared error of prediction. *J. Chemometrics*, 14:351–361, 2000.

N. M. Faber. Critical evaluation of a significance test for partial least squares regression. *Anal. Chim. Acta*, pages 235–240, 2001.

U. J. S. Hjorth. *Computer intense statistical methods.* Chapman & Hall, 1994.

Paper E

Efficiency of non-designed investigations

Sture Holm &Tobias Adolfsson
Göteborg University & Chalmers University of Technology

AMS Subject Classification: 62K05, 62P25.
Key words and phrases: Design, G-optimality.

Abstract

This paper shows how a completely undesigned experiment will be asymptotically G-optimal. It also gives variance results for finite sample sizes.

1 Introduction

In experimental design theory the problem of determining the points where experiments would be performed in order to get efficient estimates of some parameters is studied. In social sciences and in other fields it is often impossible to determine fixed experimental points. Points appearing at random have to be used. Are the estimates in such non-designed investigations much worse than estimates in well planned investigations of the same size (i.e. the same number of experimental points)? That is the question we will address in this paper.

2 Optimal designs

There are numerous optimality criteria in experimental designs. Box and Draper (1987) make a nice comment on this multitude by a section named "Alphabetic optimality" indicating that there are almost as many optimality criteria as there are letters in the alphabet. In our discussion we will only consider G-optimality and G-efficiency. The G-criterion is based on the maximum of the variance of the estimate of a regression function within a regression range. For the G-criterion there is the following very simple inequality.

Theorem 1 *Let A be a regression range for a (possibly multidimensional) variable x and suppose that a random variable Y has an m-parameter linear regression on x with error having a constant distribution with finite variance, denoted σ^2. If there are n independent observations of pairs (x_i, Y_i) with $x_i \in A$, $i = 1, \ldots, n$ and if $\hat{\mu}(x)$ denotes the estimate of the regression function in the point x, then*

$$\sup_{x \in A} Var[\hat{\mu}(x)] \geq \frac{m\sigma^2}{n}$$

See e.g. Atkinson and Donev (1992). This is a very important theorem for applications. If a design has a maximal variance equal to the bound in the theorem it is G-optimal. Further, if a design has a maximal variance close to the bound, it may be good enough in practice even if it is not optimal.

3 Random design points

In many applications, in particular in the social sciences, it is impossible to make observations in predetermined points. One has to accept the points obtained for individuals in a population.

Consider for example a population where the “design points” are normally distributed. Consider further a “cut off” of that distribution by disregarding the proportion of the distribution below some level of the normal distribution density and let the remaining set be the regression range A. When studying the variance of the estimate of the parameter linear regression function in such a regression range we may transform the problem to a sphere centred in the origin. This is due to scaling invariance properties of these problems.

In order to get information about the variance properties of this non-designed experiment, we can study properties of such a normalised problem. One natural distribution is then a cut-off normal distribution. Another natural distribution is the uniform distribution over a sphere centred in the origin. Both these are special cases of rotatable distributions.

4 Uniform distribution on a sphere

Assume that x_i has a uniform distribution in the p-dimensionell unit sphere and that x_i and x_j are independent for all $i \neq j$. We observe n such p-dimensionell random variables which now is our "design". For each such setting we observe y_i and try to estimate a linear relationship between these, that is the model

$$Y = X\beta + \varepsilon, \quad \varepsilon \sim N_n(0, I\sigma^2).$$

We may want to estimate a second order polynomial function of the type

$$\mu_{y_i} = \beta_0 + \sum_j x_{ij}\beta_{L,j} + \sum_{j<k} x_{ij}x_{ik}\beta_{M,jk} + \sum_j x_{ij}^2 \beta_{Q,j}.$$

We then have $Y = (y_1, \dots, y_n)'$, where v' denotes the transpose of the vector v, and X is built up from the parts

$$\mathbf{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^{n\times 1}, L = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \in \mathbb{R}^{n\times p}$$

$$M = \begin{pmatrix} x_{11}x_{12} & x_{11}x_{13} & \dots & x_{1(p-1)}x_{1p} \\ x_{21}x_{22} & x_{21}x_{23} & \dots & x_{2(p-1)}x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1}x_{n2} & x_{n1}x_{n3} & \dots & x_{n(p-1)}x_{np} \end{pmatrix} \in \mathbb{R}^{n\times(\frac{1}{2}p(1-p))}$$

$$Q = \begin{pmatrix} x_{11}^2 - \mu_2 & x_{12}^2 - \mu_2 & \dots & x_{1p}^2 - \mu_2 \\ x_{21}^2 - \mu_2 & x_{22}^2 - \mu_2 & \dots & x_{2p}^2 - \mu_2 \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1}^2 - \mu_2 & x_{n2}^2 - \mu_2 & \dots & x_{np}^2 - \mu_2 \end{pmatrix} \in \mathbb{R}^{n\times p},$$

where $\mu_2 = E[x_{ij}^2]$. We now construct the matrix $X = (\mathbf{1} \quad L \quad M \quad Q)$ and hence

$$X'X = \begin{pmatrix} \mathbf{1}'\mathbf{1} & \mathbf{1}'L & \mathbf{1}'M & \mathbf{1}'Q \\ L'\mathbf{1} & L'L & L'M & L'Q \\ M'\mathbf{1} & M'L & M'M & M'Q \\ Q'\mathbf{1} & Q'L & Q'M & Q'Q \end{pmatrix}.$$

We are interested in the asymptotic behaviour of $\frac{1}{n}X'X$ as $n \to \infty$. The limit equals the expected values of the entries in $X'X$ which have to be calculated. Some observations will make the calculations easier. First since the distribution of x_i is rotatable we can easily show that

$$E[x_{ij}x_{ik}] = 0. \tag{1}$$

Also due to the spheric symmetry we have that

$$\frac{x_{ij} + x_{ik}}{\sqrt{2}} \stackrel{\mathcal{D}}{=} x_{ij}.$$

Using this we see that

$$\begin{aligned} E\left[\left(\frac{x_{ij} + x_{ik}}{\sqrt{2}}\right)^4\right] &= \frac{1}{4}(2E[x_{ij}^4] + 6E[x_{ij}^2 x_{ik}^2]) \\ &= E[x_{ij}^4] \Leftrightarrow E[x_{ij}^2 x_{ik}^2] = 3E[x_{ij}^4]. \end{aligned}$$

If we denote $R = ||x_i||$, we must have $E[x_{ii}^2] = \frac{1}{p}E[R^2]$ and

$$E[x_{ii}^2] = \frac{1}{p}E[R^2] = \frac{1}{p}\int_0^1 r^2 dr^p = \int_0^1 r^2 r^{p-1} dr = \frac{1}{p+2}. \tag{2}$$

We also have that $R^4 = (x_{i1}^2 + \ldots + x_{ip}^2)^2 = \sum_{j+1}^p x_{ij}^4 + \sum_{j \neq k} x_{ij}^2 x_{ik}^2$ and hence

$$\begin{aligned} E[R^4] &= \int_0^1 r^4 dr^p = \frac{1}{p}\int_0^1 r^4 r^{p-1} dr = \frac{p}{p+4} = E[x_{ij}^4]\left(p + \frac{1}{3}p(p-1)\right) \\ &\Rightarrow \; E[x_{ij}^4] = \frac{3}{(p+4)(p+2)} \text{ and } E[x_{ij}^2 x_{ik}^2] = \frac{1}{(p+4)(p+2)}. \end{aligned} \tag{3}$$

The expected values of the i:th diagonal element in $L'L$ is $E[x_{ii}^2]$ and the off-diagonal elements have expectation 0 due to (1). Thus $L'L/n$ converges to a diagonal matrix with diagonal elements $\frac{1}{p+2}$.

Next consider the $M'M$ matrix. The element-wise expectation of the matrix will have diagonal elements of the type $E[x_{ij}^2 x_{ik}^2] = \frac{1}{(p+4)(p+2)}$ and off-diagonal elements 0. Thus $M'M/n$ tends to a diagonal matrix with elements $\frac{1}{(p+4)(p+2)}$.

The structure of the $Q'Q$ matrix is also simple but this matrix is not diagonal. The expectation of the diagonal elements are

$$E[(x_{ij}^2 - \mu_2)^2] = E[x_{ii}^4] - \mu_2^2 = \frac{2(p+1)}{(p+4)(p+2)^2}$$

which is easily calculated by using (2) and (3). The off-diagonal elements are $E[(x_{ij}^2-\mu_2)(x_{ik}^2-\mu_2)]$. A calculation similar to the one above gives

$$E[(x_{ij}^2-\mu_2)(x_{ik}^2-\mu_2)] = \frac{-2}{(p+4)(p+2)^2}.$$

We have now calculated the asymptotic behaviour of all block parts of $X'X/n$ and are now ready to calculate $(X'X/n)^{-1}$. This is fairly simple since all but one of the sub-matrices in $(X'X)$ are either diagonal or zero. Elementary linear algebra gives that the asymptotic inverse must also be a block matrix. The inverses of the matrices $\mathbf{1}'\mathbf{1}$, $L'L$ and $M'M$ are trivial since these matrices are diagonal. The $Q'Q$ matrix can be written on the form $aI+bJ$ where I is the identity matrix, J is a $p\times p$ matrix with all elements equal to one and a,b are real numbers. The inverse of such a matrix is also easy to calculate. We can now conclude that

$$\left(\frac{1}{n}X'X\right)^{-1} \to \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & (p+2)I & 0 & 0 \\ 0 & 0 & (p+4)(p+2)I & 0 \\ 0 & 0 & 0 & \frac{(p+4)(p+2)}{2}(I+\frac{1}{2}J) \end{pmatrix} = \Gamma.$$

This will be used for calculating the variance function of our non-designed experiment. We are interested of points in the sphere and the variance in a point at radius r equals $z\Gamma z$, where $z=(1,r_1,\ldots,r_n,r_1r_2,\ldots r_{p-1}r_p,r_1^2,\ldots,r_n^2)$ such that $r=(r_1^2+\ldots+r_n^2)^{1/2}\leq 1$.The variance equals

$$v(p,r) = \sigma^2\left(1+\frac{p(p+4)}{4}+\frac{3}{4}(p+4)(p+2)r^4-\frac{r^2}{2}(p+2)^2\right).$$

This function has it's maximum when $r=1$ and it equals

$$v(p,1) = \frac{\sigma^2}{2}(p^2+7p+10)$$

which has to be compared with the bound for the maximal variance

$$o(p) = \sigma^2\left(\frac{p(p-1)}{2}+2p+1\right) = \frac{1}{2}\left(p^2+3p+2\right).$$

For instance if $p=10$ the maximal variance is 90 and the bound is 66. The difference between the maximum of the variance function and the optimal variance is somewhat less than 10% when $p = 40$. These difference is far less then one would expect given the fact that no design is made.

For a first order polynomial model the results are simply obtained from the above calculations. For $p = 10$ the maximum of the variance function equals 13, while the bound is 11, and for $p = 40$ the maximum equals 43, while the bound is 41. In this case the "random design" is even more efficient compared to the optimal design.

5 Truncated normal distribution

The calculation of the variance function for any rotatable distribution of the "design points" is similar to the one for uniform distribution on a sphere. The most relevant example is perhaps a normal distribution truncated at some level of the normal density. For a 10 parameter case with truncation at 90 % in a normalised distribution it turns out that the variance functions are $\sigma\ 2(1 + 1.114r^2)$ and $\sigma^2(6.29 - 0.0647r^2 + 0.0601r^4)$ for the first and second order models, respectively. The maxima within the regression ranges are 18.8 and 159 to be compared to the bounds 11 and 66. For a 40 parameter case of the same type the variance functions are $\sigma^2(1 + 1.050r^2)$ and $\sigma^2(21.30 - 0.01561r^2 + 0.5461r^4)$ for the first order model and the second order model. The maxima within the regression ranges are 55.4 and 1485 to be compared to the bounds 41 and 861.

It is seen that the difference between the undesigned experiment and the optimal experiment is larger here than in the case with uniform distribution as expected. The difference is, however, not extremely large.

References

A. C. Atkinson and A. N. Donev. *Optimum Experimental Designs.* Oxford Science Puplications, 1992.

G. E. P. Box and N. R. Draper. *Emperical Model-Building and Response Surfaces.* John Wiley & Sons, 1987.

Paper F

Partial Least Squares regression and Cross Validation in a statistical perspective

Tobias Adolfsson
Göteborg University and Chalmers University of Technology

February 28, 2002

Abstract

Partial Least Squares (PLS) has gained great prominence in the field of chemometrics. As in all stepwise methods, the stopping criteria for PLS is crucial. We will examine the properties of cross-validation as a stopping rule and make comments on its usability.

1 Introduction

Partial Least Squares (PLS) first occurred in Wold et al. (1984) as a method to solve the problem of collinearity in linear regression. It was pointed out that the algorithm presented was a version of the Conjugate Gradient Method (CGM) but motivated in a different way. CGM is a well known numerical method for solving large systems of linear equations. The main idea of CGM is to use an iterative method that cuts down the error optimally in each step. On a computer with infinite floating number precision the algorithm will give the Ordinary Least Squares (OLS) estimate in a finite number of steps. The method cuts down the error by solving the problem on Krylow subspaces.

There has been a substantial effort put into making the statistical properties of the method clearer. This work was initialised in Helland (1988, 1990), Helland and Almøy (1994) and von Rosen (1994). These papers try to lay a fundament for the statistical aspects of PLS and the implicit modelling used. There has been attempts to calculate the bias introduced by the method, see Stoica and Söderström (1998), and to construct confidence intervals for the effects estimated by PLS, see Phatak et al. (1993). Since PLS uses a rather complex non-linear estimation process it is difficult to make precise and accurate models for confidence intervals. In Faber and Kowalski (1997) a method for confidence intervals estimation was given but this method was proven wrong in Morsing and Ekman (1998). The standard inference machinery used in statistics is not easily transformed into the situations occurring in PLS.

Since PLS is a iterative method one has to stop the algorithm at some point. A commonly used tool for stopping is cross validation where some observations are held back in the estimation process and then are predicted by the current solution, hence cross validation is a method for estimating the Mean Squared Error of Prediction (MSEP). Since PLS is often used in regression like situations the MSEP is a natural tool to examine. For an overview of different methods to estimate the MSEP in PLS, see Denham (2000).

2 PLS and optimal error reduction

Assume that we have the model

$$Y = X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I_n). \tag{1}$$

Our goal is to estimate β given $X \in \mathbb{R}^{n\times p}$ and $Y \in \mathbb{R}^{p\times 1}$. One way of deriving the CGM, and hence PLS, is to use the optimal error cut down. As shown in Helland (1988) the subspace on which the problem is solved is a Krylow space. If we let e_i denote the error in step i of the PLS algorithm w.r.t the OLS solution, and let d_i be the added search direction in the i:th step we get that e_i must be chosen from the space

$$\mathcal{E}_i = e_0 + span\{d_0, d_1, \ldots, d_{i-1}\}.$$

Let $||v||_A^2 = v'Av$ be an energy norm in the matrix A. The error is chosen such that

$$e_i = \min_{e\in\mathcal{E}_i} ||e||_{X'X}.$$

We now use the fact that PLS uses search directions that are $X'X$-orthogonal, i.e $d_i X'Xd_j = 0 \;\; \forall i \neq j$ (i.e. conjugate). Then the error may be written as

$$||e_i||_{X'X} = \sum_{j=i}^{p-1} \rho^2 d_j X'Xd_j \tag{2}$$

where $\rho \in \mathbb{R}$. It can easily be shown that every other vector e in $\mathcal{E}_i$ must contain these vectors and hence the optimality of the error term follows. Since the residuals r_j are chosen from the Krylow space

$$\mathcal{K}_i = r_0 + span\{(X'X)r_0, (X'X)^2 r_0, \ldots, (X'X)^i r_0\},$$

and the relationship between the residuals and the errors is $r_i = -(X'X)e_i$, it follows that the error e_i must be chosen from the Krylow space

$$\mathcal{K}_i = e_0 + span\{(X'X)e_0, (X'X)^2 e_0, \ldots, (X'X)^i e_0\}.$$

This means that, for a fixed i, the error must have the structure

$$e_i = \left(I + \sum_{j=1}^{i} \psi_j (X'X)^j \right) e_0$$

where the coefficients ψ_j are chosen so that the energy norm with respect to $X'X$ is minimised. To make the way of calculating the ψ_j more transparent we rewrite the expression in terms of eigenvectors for $X'X$. Let $P_i(\lambda)$ denote a polynomial function of order i that takes either scalars or matrices as arguments. If we let v_j, $j = 1, \ldots, p$ denote the orthogonal unit eigenvectors of $X'X$ we have that

$$P_i(X'X)v_j = P_i(\lambda_j)v_j$$

where λ_j denotes the corresponding eigenvalue. From this we see that when the order of this polynomial reaches the number of distinct eigenvalues of $X'X$ the error becomes zero. If we now rewrite the starting error e_0 as a sum of scalars times eigenvectors, i.e

$$e_0 = \sum_{j=1}^{p} \xi_j v_j,$$

the energy norm of the error will have the structure

$$||e_i||^2_{X'X} = \sum_{j=1}^{p} \xi_j^2 [P_i(\lambda_j)]^2 \lambda_j.$$

The polynomial $P_i(\cdot)$ is chosen such that $||e_0||^2_{X'X}$ is minimal with the restraint that $P_i(0) = 1$. Since $r_0 = X'y$ and $r_i = -(X'X)e_i$ we must have that $e_0 = -(X'X)^{-1}X'y$. Hence $e_0 = -\beta - (X'X)^{-1}X'\epsilon$, here ϵ denotes the random error. If we now assume that the standard unit base in $\mathbb{R}^p$ will do as an eigenbase we see that

$$\xi_j^2 \overset{\mathcal{D}}{=} \chi_1^2(\beta_j)\tilde{\sigma}_j^2,$$

where $\chi_1^2(\beta_j)$ is a non-central χ^2-distribution with one degree of freedom and non-central parameter β_j and $\tilde{\sigma}_j^2 = var(((X'X)^{-1}X'\epsilon)_j)$. In view of (2) we see that, if M_i denotes the model in step i, $M_i \perp_{X'X} e_i$ and hence Pythagoras theorem gives that

$$||\beta_{OLS}||_{X'X}^2 = ||M_i||_{X'X}^2 + ||e_i||_{X'X}^2. \tag{3}$$

We can now find a upper bound for the error in step i by

$$\begin{aligned} ||e_i||_{X'X}^2 &\leq \min_{P_i} \max_{\lambda} [P_i(\lambda)]^2 \sum_j \xi_j^2 \lambda_j \\ &= \min_{P_i} \max_{\lambda} [P_i(\lambda)]^2 ||e_0||_{X'X}^2. \end{aligned}$$

We can now see that the error is, in some sense, bounded by the eigenstructure of the $X'X$ matrix and the error. If the polynomials P_i were known we would be able to calculate the distribution of the length of the error but the polynomial is optimal with respect to the observed data and hence in a general model stochastic. The structure of the polynomial can however explain some of the unusual shrinkage properties of PLS, see Butler and Denham (2000).

3 Shrinkage

In Goutis (1996) a geometrical proof, showing that PLS gives shrinkage estimators in average, is given. In Frank and Friedman (1993) it is shown that there are some directions in which the PLS regression coefficients exceeds the coefficients of OLS. An analytic proof of the average shrinkage can easily be performed if a "good" norm is used. In view of section 2 we define an energy norm in $X'X$, i.e.

$$||a||_{X'X}^2 = a'X'Xa, \tag{4}$$

where X is defined as in (1). Observe that this is the complete X matrix. We can now prove that the PLS regression vector is shrinking in this norm.

Theorem 1 (Shrinkage in energy norm) *With the norm* $||\cdot||_{X'X}$ *defined as in (4) the following holds*

$$||\hat{\beta}_s||_{X'X} \leq ||\hat{\beta}_{s+1}||_{X'X} \leq ||\hat{\beta}||_{X'X}.$$

Proof:
First we observe that

$$\begin{aligned} ||\hat{\beta}_s||^2_{X'X} &= \hat{\beta}_s' X'X\hat{\beta}_s = (X\hat{\beta}_s)'(X\hat{\beta}_s) \\ &= (P_{P_{\mathcal{W}_s}X}Y)'(P_{P_{\mathcal{W}_s}X}Y) = ||P_{P_{\mathcal{W}_s}X}Y||^2_2. \end{aligned}$$

We also see that

$$\begin{aligned} ||\hat{\beta}_{s+1}||^2_{X'X} &= (P_{P_{\mathcal{W}_{s+1}}X}Y)'(P_{P_{\mathcal{W}_{s+1}}X}Y) \\ &= ([P_{P_{\mathcal{W}_s}X} + P_{P_{\mathcal{W}_{s+1}\setminus\mathcal{W}_s}X}]Y)'([P_{P_{\mathcal{W}_s}X} + P_{P_{\mathcal{W}_{s+1}\setminus\mathcal{W}_s}X}]Y) \\ &= ||P_{P_{\mathcal{W}_s}X}Y||^2_2 + ||P_{P_{\mathcal{W}_{s+1}\setminus\mathcal{W}_s}X}Y||^2_2 \geq ||P_{P_{\mathcal{W}_s}X}Y||^2_2 = ||\hat{\beta}_s||^2_{X'X}. \end{aligned}$$

The last inequality in the theorem is now trivial since, if k denotes the number of distinct eigenvalues of $X'X$, $P_{P_{\mathcal{W}_k}X}Y = \hat{Y}$ were $\hat{Y}$ is the OLS estimate.
Q.E.D.

4 Variance estimation problems in PLS

In the last years some efforts has been made to estimate the variance of the PLS regression coefficients. The papers in this field are mainly concerned with the use of Taylor-expansions with various order of the error term. To make the variances understandable the variance is often shown in the original parameters. Variance estimation is an important topic since statistics as confidence intervals and tests are often based on variance estimates. Using incorrect variance estimates can, or will, lead to catastrophic inference. In this section I will prove that these methods always under estimates the variance for the original parameters.

Theorem 2 (Under estimation of variances) *Assume that the PLS algorithm is stopped at step* s*, where* $s < k$*. Then* $\exists c \in \mathbb{R}^{p\times 1}$ *such that* $c'\hat{\beta}_{OLS}$ *is estimable,* $var(c'\hat{\beta}_{OLS}) \neq 0$ *and* $var(c'\hat{\beta}^s_{PLS}) = 0$.

Proof:
Since $\hat{\beta}^s_{PLS}$ is given by solving

$$P_{\mathcal{W}_s} Y = P_{\mathcal{W}_s} X\beta$$

in the OLS sense and $s < k$ $Q_s = col\{X\} \setminus \mathcal{W}_s \neq \emptyset$. We choose $c \in Q_s$ (observe that $c \in col\{X\}$). We now directly have that $c\hat{\beta}_{OLS}$ is estimable and has positive variance. But since the PLS regressor is based on a projection onto a s dimensional subspace of $col\{X\}$ the covariance matrix of $\hat{\beta}^s_{PLS}$ can only vary in this subspace or more mathematically put, $\Sigma^{latent}_{PLS} \in \mathcal{W}_s$ gives that $\Sigma^{original}_{PLS} \in \mathcal{W}_s$. Since $\mathcal{W}_s \perp Q_s$ we have

$$\mathrm{var}(c'\hat{\beta}^s_{PLS}) = c'\Sigma_{PLS} c = 0.$$

Q.E.D.

In some commercial implementations estimated confidence intervals has had very low empirical coverage probability. The theorem above might well be one of the reasons why this happens. When estimating in the PLS setting one should remember that there are two kinds of variations to take into account. One is the spread of the Y vector that is projected onto a subspace of the X matrix, the other is the variation of the projection space. The relative size of the variation of these two sources is not clear. It is common that the variation of the projection space is not taken into account and this might be a large contributer to the total variance of the method!

5 Cross validation

In (1) we now choose a subset of X that we are not to use when estimating β. Let

$$X = \begin{pmatrix} x'_1 \\ x'_2 \\ \vdots x'_n \end{pmatrix} \in \mathbb{R}^{n \times p},$$

and let

$$X_r = \begin{pmatrix} x'_{[1]} \\ x'_{[2]} \\ \vdots \\ x'_{[r]} \end{pmatrix} \in \mathbb{R}^{r \times p}$$

be the matrix containing the rows that are held back. We now use the Woodbury formula for a rank r correction of an inverse

$$(A - BD^{-1}C')^{-1} = A^{-1} + A^{-1}B(D - C'A^{-1}B)^{-1}C'A^{-1}.$$

Using this formula we can calculate the OLS estimate of β when r rows are held back and denote it $\hat{\beta}_{n-r}$. We have

$$\begin{aligned}
\hat{\beta}_{n-r} &= ((X'X)^{-1} + (X'X)^{-1}X_r'(I - X_r(X'X)^{-1}X_r')^{-1}X_r(X'X)^{-1}) \\
&\cdot \quad (X'Y - X_r'Y_r) \\
&= \hat{\beta} + (X'X)^{-1}X_r'(I - K_r)^{-1}X_r\hat{\beta} \\
&- \quad (X'X)^{-1}X_r'Y_r - (X'X)^{-1}X_r'(I - K_r)^{-1}K_rY_r
\end{aligned}$$

where $\hat{\beta}$ is the usual OLS estimate of β and $K_r = X_r(X'X)^{-1}X_r'$. We now calculate the cross validation estimates of the error when predicting the y_i:s in the cross validation subset

$$\begin{aligned}
X_r\hat{\beta}_{n-r} - Y_r &= X_r\hat{\beta} + K_r(I - K_r)^{-1}X_r\hat{\beta} - (I + K_r + K_r(I - K_r)^{-1}K_r)Y_r \\
&= (I + K_r(I - K_r)^{-1})X_r\hat{\beta} - (I + K_r + K_r(I - K_r)^{-1}K_r)Y_r.
\end{aligned}$$

Simplifying this expression we get

$$X_r\hat{\beta}_{n-r} - Y_r = (I - K_r)^{-1}(X_r\hat{\beta} - Y_r). \tag{5}$$

We now need the following definition:

Definition 1 (Estimatability) *A linear combination $c'\beta$ is estimable if $\exists a : E[a'Y] = c'\beta$.*

From this definition we get the following theorem:

Theorem 3 (Rao) *All linear parametric functions are estimable, if and only if $rank(X) = p$, the number of unknown parameters β_j.*

We can now draw conclusion about the estimability problems that might arise in cross validation.

Theorem 4 (Estimability and K_r) *If the full model is estimatible the following holds in the cross validation procedure*

$$X_r\beta \text{ is not estimable} \Leftrightarrow \det(I - X_r(X'X)^{-1}X_r') = 0.$$

Proof:

$$\begin{aligned}
& X_r\beta \text{ not estimable} \\
\Leftrightarrow\ & \exists v \neq \emptyset, v \in \mathbb{R}^{p\times 1}, v \in \mathcal{R}(X_r) \text{ s.t. } X_{n-r}'X_{n-r}v = \emptyset \text{ and } X_r'X_r'v \neq \emptyset \\
\Leftrightarrow\ & \exists v \neq \emptyset, v \in \mathbb{R}^{p\times 1} \text{ s.t. } X'Xv = (X_{n-r}'X_{n-r} + X_r'X_r)v = X_r'X_rv \\
\Leftrightarrow\ & \exists v \neq \emptyset, v \in \mathbb{R}^{p\times 1} \text{ s.t. } (X'X)^{-1}X_r'X_rv = v \\
\Leftrightarrow\ & \exists v \neq \emptyset, v \in \mathbb{R}^{p\times 1} \text{ s.t. } X_r(X'X)^{-1}X_r'X_rv = X_rv, \quad X_rv \neq \emptyset \\
\Leftrightarrow\ & \exists \lambda_i \text{ eigenvalue to } X_r(X'X)^{-1}X_r' \text{ s.t. } \lambda_i = 1.
\end{aligned}$$

The last three $\Leftarrow$ are based on

$$\begin{aligned}
\det(I - K_r) = 0 \quad & \Leftrightarrow \quad \exists w \text{ s.t. } K_rw = w, \quad w \neq \emptyset \\
& \Leftrightarrow \quad X_r(X'X)^{-1}X_r'w = w \neq \emptyset \\
& \Rightarrow \quad X_r'w \neq \emptyset \text{ i.e. } w \in \mathcal{N}(X_r')^{\perp} = \mathcal{R}(X_r) \\
& \Rightarrow \quad \exists v \neq \emptyset \text{ s.t } w = X_rv, v \in \mathcal{R}(X_r).
\end{aligned}$$

Q.E.D.

If $I - K_r$ is close to singular we have that

$$||X_r\hat{\beta}_{n-r} - Y_r|| = \frac{1}{||(I - K_r)||} \cdot ||X_r\hat{\beta} - Y_r||,$$

and we might get a blow up of the cross validation error. If the data in X is structured the problem mentioned above can easily occur and if the data is not structured this phenomena might still occur.

5.1 Cross validation on growing column spaces

We will now limit us to only holding back one row at the time and not a matrix as in the previous section. The equivalent of eq. 5 is then

$$x_r'\hat{\beta}_n - i - y_i = \frac{1}{1 - k_i}(x_r'\hat{beta} - y_i)$$

where $k_i = x_r'(X'X)^{-1}x_i$. If we assume that $X'X$ has rank p we have that

$$\text{trace}(X(X'X)^{-1}X') = \sum_{i=1}^{n} \lambda_i = p \tag{6}$$

where λ_i denotes the i:th eigenvalue of $X(X'X)^{-1}X'$. Since $X'X > 0$ we can derive an upper bound for the diagonal elements

$$\begin{aligned} (X(X'X)^{-1}X')_{ii} &= e_i'X(X'X)^{-1}X'e_i \\ &\leq \min_i \lambda_i = 1, \end{aligned} \tag{7}$$

where

$$e_i = (\underbrace{0 \ldots 0 \;\; 1}_{i \text{ elements}} \; 0 \ldots 0)'.$$

Assume that when we add a new column to X we also increase the rank of X by one i.e increasing p to $p+1$ makes rank$(X) = p+1$. Let X_p denote the matrix with p columns and set $P_p = X_p(X_p'X_p)^{-1}X_p'$. From eq. 6 we get that $\sum k_i$ increases as $p \mapsto p+1$. Since $P_{p+1} = P_p + P_1$, where P_1 is a rank one projection, we have that

$$P_{p+1} - P_p = P_1.$$

Since P_1 is a rank one projection matrix we know that its diagonal elements are non negative and that one of its eigenvalues is one while the rest of the eigenvalues are zero. We have now proven that

Theorem 5 (One step cross validation) *Let X, Y and β be defined according to (1) with $n > p$. Then the cross validation error in estimating y_i is*

$$\frac{1}{1-k_i}(e_i'\hat{\beta} - y_i).$$

Where k_i has the following properties

- $0 \leq k_i \leq 1$
- *$k_i(p)$, i.e. k_i is dependent on p and $k_i(p) \leq k_i(p+1)$ $\forall i$ and for at least one i we must have $k_i(p) < k_i(p+1)$.*

This theorem shows an important aspect of cross validation used on methods using increasing projective spaces. The cross validation error estimate increase with the dimension of the column space. The error might even become large when *the added dimension contains important information*. It is well known that when cross validation is used in PLS there has been a tendency for the algorithm to stop too early. Many implementations now includes additional stopping criterias to prevent the algorithm from stopping too early. Maybe one should look for a new stopping criteria.

References

N. A. Butler and M. C. Denham. The perculiar shrinkage properties of partial least squares regression. *J. R. Statist. Soc. B*, 62:585–593, 2000.

M. C. Denham. Choosing the number of factors in partial least squares regression: estimating and minimizing the mean squared error of prediction. *J. Chemometrics*, 14:351–361, 2000.

K. Faber and B. R. Kowalski. Propagandation of measurment errors for the validation of predictions obtained by principal component regression and partial least squares. *J. Chemometrics*, 1997.

I. E. Frank and J. H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 1993.

C. Goutis. Partial least squares algorithm yields shrinkage estimators. *Annals of Statistics*, 1996.

I. S. Helland. On the structure of partial least squares regression. *Comm. Statist. B-simulation Comput.*, 1988.

I. S. Helland. Partial least squares regression and statistical models. *Scand. J. Statist.*, 1990.

I. S. Helland and T. Almøy. Comparison of prediction methods when only a few components are relevant. *J. Amer. Statist. Assoc.*, 1994.

T. Morsing and C. Ekman. Comments on construction of confidence intervals in connection with partial least squares. *J. Chemometrics*, 1998.

A. Phatak, P. M. Reilly, and A. Penlidis. An approach to interval estimation in partial least squares regression. *Analytica Chimica Acta*, 1993.

P. Stoica and T. Söderström. Partial least squares: A first-order analysis. *Scand. J. Statist.*, 1998.

D. von Rosen. Partial least squares and a linear model. In *Proceedings at the International Conference on Linear Statistical Inference LINSTAT 93*, 1994.

S. Wold, A. Ruhe, H. Wold, and W. J. Dunn. The collinearity problem in linear regression. the partial least squares (pls) approach to generalized inverses. *SIAM J.Sci.Statist. Comp.*, 1984.