CHALMERS | UNIVERSITY OF GOTHENBURG

PREPRINT 2008:24

On Super Saturated Experimental Design

SVEN AHLINDER
IVAR GUSTAFSSON

Department of Mathematical Sciences
Division of Mathematics
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Göteborg Sweden 2008

Preprint 2008:24

On Super Saturated Experimental Design

Sven Ahlinder, Ivar Gustafsson

Department of Mathematical Sciences
Division of Mathematics
Chalmers University of Technology and University of Gothenburg
SE-412 96 Göteborg, Sweden
Göteborg, June 2008

Preprint 2008:24
ISSN 1652-9715

Matematiska vetenskaper
Göteborg 2008

On Super Saturated Experimental Design

Sven Ahlinder Volvo (sven.ahlinder@volvo.com)

Ivar Gustafsson Chalmers (ivar@math.chalmers.se)

1. Abstract

Design of Experiments (DoE) is the scientific topic of how to perform series of investigations of an arbitrary object. Normally in DoE, the number of observations is larger than the number of variables. This gives a possibility for estimate statistical properties about the coefficients that describe the influence of the variables.

A modern vehicle is specified by a million parameters. The problem is that the companies can not afford to make a million experiments. This phenomenon has been identified already in the fifties and the answer was supersaturated designs.

Suppose a system of equations Ax=b. Here x is the underlying coefficients that describe the studied object, A is a series of experiments on the object and b is the result of the experiments.

A supersaturated experimental design is a plan to create an experimental series with fewer observations (corresponding to the rows of the matrix A) than unknown parameters (the elements of x). This leads to an underdetermined system of equations Ax=b. Solving an underdetermined system of linear equations could be done by a generalized inverse based on Singular Value Decomposition (SVD).

In normal, non saturated designs, the bottleneck of information is to replicate the measured results, given by a vector b. This results in methods for estimating predictions, $\hat{b}$, which have given experimental series A that have orthogonal columns since this gives the possibility to estimate the elements in x independently of each other.

In this paper a new method is introduced for creating supersaturated designs. It is simply assumed that the first order Taylor series describing the studied system is known. The first order Taylor series is the sum of all parameters, each multiplied by a coefficient. The coefficients are the elements of x. The coefficients give a steepest ascent direction to follow when optimizing the system. The method is focused on how to estimate the coefficients, the elements of a vector $\hat{x}$, using the experiments A on the "correct" vector x. For a non-saturated A, this is not a problem but in supersaturated A's this is vital.

If x lies in the row space of A, then the experimental series will be a success, but if x lies in the null space of A, it will be a complete failure. In all practical cases, x has components in both row space and null space. The best we can do is to take an A with a dimension of the row space as large as possible compared with the null space. This means that the rank, or the number of independent rows of A, should be as large as possible. This leads to the fact that in non-saturated designs, A should have linearly independent columns and in supersaturated designs, A should have linearly independent rows.

The total gain in lean optimization is expected to be $\sqrt{Number\ of\ rows * Number\ of\ columns}$. This means that investigating 4 times more parameters with the same amount of tests gives double the information.

2. Symbols:

A: The experiments conducted on a system, that is a matrix called the experimental design.
x: The vector of coefficients of the linear approximation of the system
b: The vector of results when conducting the experiments A on the system, i.e. b=Ax
A^+: (Generalized) inverse of A

$\hat{x}$: The estimated (calculated) coefficients of the linear approximation of the system, i.e. $\hat{x}$ = A+b

$\hat{b}$: The vector of results when conducting A on $\hat{x}$, i.e. $\hat{b}$ =A $\hat{x}$

s: The saturation (number of columns)/(rank)
||x||: A norm of a vector. $||x||_2$ is the 2-norm i.e. the Euclidian length.
||A||:A norm of a matrix. $||A||_2$ is the 2-norm i.e. the norm corresponding to the vector norm $||x||_2$.

3. Introduction

Design of Experiments, DoE is a discipline in statistics. In year 1935 the famous statistician R. A. Fischer published a book called “The design of experiments” [Ref 1] about how to design experimental series to be able to estimate the coefficients in linear polynomials with several variables. Those polynomials then served as models to be able to maximize the yield of crops. His basic ideas were like this:

Assume that the notation for a linear system of equations is Ax=b, where A is the left hand side, b is the right hand side and x are the coefficients to estimate. Then A is a list of which experiments we should perform to investigate our unknown system. If possible,A should have full rank and have orthogonal columns. This leads to independent estimation of the parameters in x, which is ideal. According to Fischer it is useful to add some extra rows to A, without destroying the orthogonality, to make the investigation less sensitive to measurement errors. Then the system of equations can be solved with the method of least squares, which will solve the problem min $||Ax-b||_2$.

In vehicle industry, DoE is used in several disciplines such as fuel consumption, emissions, noise, durability, handling, welding, painting, vibrations, air resistance, cooling, heating etc. The main problem is that a vehicle is defined by about a million variables, and the industry can not afford to do a million experiments. Due to this there is a growing interest in super saturated designs, where one does fewer experiments than the number of parameters to estimate and one gets an underdetermined system of equations to solve. In this case the parameters in x can not be independently estimated.

3.1 DoE and optimization

A gradient optimizer calculates the gradient of a system, as if it was a function, in the starting point p_0 and goes in this direction, which is the steepest ascent when maximizing. The gradient could be approximated by taking a small step δ from the starting point in each variable one at a time and calculate the influence on the function for each variable. In DoE terms all experiments of one gradient evaluation form a matrix A that is diagonal with diagonal elements δ. Since A= δI is square, $x=A^{-1}b= \delta^{-1}b$ which is the same as dividing all elements in b with δ.

Then a step, normally much larger than δ, is taken in the direction of the gradient. The larger step is normally repeated several times until an optimum is reached. Then the system (function) is evaluated again and a new direction is set.

Design of Experiments mainly works the same way as a gradient optimizer. You calculate the gradient $\hat{x}$ =A^{+}b and goes in that direction. The three major differences are:

DoE does not take a small step δ but gives the system a solid kick [2]

In DoE, A is not I, but has balanced columns.

A is mostly overdetermined in DoE, i.e. A has more rows than columns.

Normally there is only one step for each system evaluation, but a reasonably large one.

“Balanced columns” means that in each column of A there are equally many steps in the positive direction, “ones”, as in the negative direction, “minus ones”. This leads to the fact that only certain combinations of rows and columns are orthogonal and those are called experimental designs. One purpose of balanced designs is that the influence of the experimental error is divided by the number of rows. [2].

If A is made nonsaturated or over determined by increasing the number of rows the influence of errors is further decreased. Overdetermined systems of equations are often solved by the method of least squares.

The idea of a solid kick also relates to the errors in measurement. The longer step taken when investigating the gradient, the more exact the influence of the variable is estimated. On the other hand, if the step is too long, there may be an influential curvature and even a passed optimum. The gradient solver investigates only a small surrounding of the starting point and then takes a larger step. When taking the larger step you get into an area that is not investigated. One compromise is to let the step length in gradient investigation be as long as the planned step length in the improvement step.

Note that here; A is the experimental design which should be chosen with regard to properties as cost of experiments and orthogonality. This means that in DoE you decide in advance which system of equations you should solve by deciding which experiments you should perform on your system.

3.2 The Pareto principle

In 1906, Vilfredo Pareto observed that 80% of the property in Italy was owned by 20% of the population. In 1941, Dr. Joseph Moses Juran expanded the observation to "80% of your sales come from 20% of your clients". This is also known as the 80-20 rule or the vital few and the trivial many [7].

3.3 Optimization with super saturated designs.

A super saturated design, SSD, is an experimental series with fewer observations than variables. This leads to an underdetermined system of equations to solve. This can be done by a generalized inverse, $\hat{x}=A^{+}b$. The most common generalized inverse for underdetermined system of equations is the Moore-Penrose inverse, $A^{+}=A^{T}(AA^{T})^{-1}$ for full rank, and A^{+} is based on SVD in the general case, see below. Here $\hat{x}$ is the minimum norm solution of Ax=b, i.e. the solution with smallest norm $||\hat{x}||_2$

Except for the generalized inverse, optimization with SSD does not differ from ordinary DoE. $\hat{x}$ is estimated and a step along the steepest ascent is taken [3]. A reasonable step length should be chosen and the columns should be balanced as in normal DoE. One can hope that the direction of the step is good enough, at least compared to the limited effort of SSD.

3.4 Present methods

The traditional methods [4] studied for generating super saturated designs, mainly focus on low correlation between the columns. This is called $E(s^2)$ which means that the maximal correlation of the columns of A should be as small as possible. [4]. This is the typical case in Design of Experiments. However SSD is atypical in DoE as we will show later and the important concern in SSD is that A has linear independent rows.

3.5 An example of SSD optimization

Volvo has tested SSD optimizations on mathematical functions [3]. Some theoretical functions and a genuine engine model were tested. In all cases about 100 variables was used and 50 evaluations were performed. The 50 evaluations were divided into 3 steps with about 15 evaluations in each step. For

each step, centre of the investigation was moved in the direction of steepest ascent. In all cases the function value was improved for each step. Generally, 2000 random evaluations did not give such a good value as the 50 SSD evaluations which indicate that the method has a potential.

4. Theory

Assume a system with many variables to be optimized. This system could be a subsystem of a vehicle. Assume that in a reasonable small interval, the function could be approximated by a first order Taylor series. This means that a vector of linear coefficients, here called x, describes the interval reasonably good. To estimate x, **we** perform a list of experiments which are the rows of a matrix called A. The vector b is the result of the experiments and b=Ax. The estimation of x is denoted $\hat{x}$ and is defined by $\hat{x}=A^{+}b$ where A^{+} is a generalized inverse to be defined below.

4.1 Solving a system of equations

A linear system of equations must in general be square and have full rank to be solved without assumptions. For over-determined (long and thin) systems of equations, the least square solution is mostly used. For underdetermined (short and fat) systems of equations, the minimum norm solution can be used. Using singular value decomposition is an expensive way of solving systems of equations but it gives both least square and minimum norm solution [5]. Since Super Saturated Designs are aimed for expensive experiments, the cost of solving the systems of equations is not a problem

4.2 Singular Value Decomposition (SVD)

Any matrix A can be expressed as the product of three matrices USV^{T}, where U and V are orthogonal and S is quasi-diagonal. The diagonal elements of S are called the singular values. Since the inverse of an orthogonal matrix is its transpose, we may define a so called pseudoinverse by $A^{+}=VS^{+}U^{T}$ where S^{+} is the transpose of S with all non zero elements replaced by their inverses. Particularly there is a compact form for Singular Value Decomposition, SVD, where S is square which then may not be the cases for U and V.

Suppose A has m rows and n columns. Then A^{+} is the Moore-Penrose pseudo inverse and $\hat{x}=A^{+}b$ is the exact solution to Ax=b if m=n (with full rank), least square solution if m>n and minimum norm solution if m<n.

4.3 Estimations

Suppose x is a first order Taylor expansion of the behaviour of an object and A is the experiments conducted on the object. Let b be the results of the experiments. Then b=A*x.

If we want to estimate the first order Taylor expansion of the behaviour of the object, then $\hat{x}=A^{+}b$

If we then want to check if the object behaves linearly we may calculate $\hat{b}=A\hat{x}$ and see if $\hat{b}=b$.

The total calculation may be summarized as $\hat{b}=AA^{+}Ax$.

For m>n, A^{+} is a left inverse so $A^{+}A=I$. This means that $\hat{x}=A^{+}Ax=Ix=x$ and $\hat{b}=A\hat{x}=Ax=b$ if a first order Taylor expansion is a good description of the object that **we** let A operate on.

For m<n, A^{+} is a right inverse so $AA^{+}=I$. This means that $\hat{b}=AA^{+}Ax=IAx=Ib=b$, which means that the estimations made of the model $\hat{x}$ probably are correct. But note that $A^{+}A$ probably not is I. This means that $\hat{x}=A^{+}Ax$ probably not is x and the model is hard to interpret. Still $A\hat{x}=Ax$. If **we** have a new

observation O which is a linear combination of the rows of A, in the row-space of A, then $O\hat{x}$=Ox. Hence the quality of A is dependant of the (row) rank of A not the correlations of the columns that is indicated in [Ref 4].

4.3.1 Some properties

The norm of a matrix A is defined as ||A||=sup ||Ax|| / ||x|| for a vector norm ||x||. Let the norms be 2-norm. It follows that 0≤ ||Ax||≤ ||A||*||x||. ||Ax||=0 if x is in the null space of A and ||Ax||=||A||*||x|| if x is parallel to the column in V^T that corresponds to the largest singular value of A. [Ref 5]

Let $C=A^+A=VS^+U^TUSV^T=VS^+SV^T$. S^+S is diagonal so C is symmetric. Furthermore the diagonal of S^+S consists of zeroes or ones. Thus, by the spectral theorem, we conclude that the eigenvalues of C are zeroes or ones. It follows that $||C||=\sqrt{\lambda_{\max}(C^2)}=1$, where $\lambda_{\max}$ is the largest eigenvalue.

Since $x \in Nul(A) \Rightarrow x \in Nul(C)$ we also get ||Cx|| = 0 if $x \in Nul(A)$

Consider now the compact SVD of $A=U_1S_rV_1^T$, where S_r is positive, diagonal of size rxr, where r=rank(A). Then $A^T=V_1S_rU_1^T$ and the range spaces $R(A^T)$ of A^T and $R(V_1)$of V_1 are the same.

Further $C=A^+A=V_1S_r^{-1}U_1^TU_1S_rV_1^T=V_1V_1^T$ is then the orthogonal projection on $R(A^T)$=Row (A), the row space of A. In particular, if x belongs to Row(A), then $\hat{x}$=Cx=x.

From $C=V_1V_1^T$, with V_1 having r linearly independent columns, it follows that C has r eigenvalues equal to 1 and n-r eigenvalues equal to 0. Thus r=rank(C) is the sum of the eigenvalues of C, i.e. r=trace(C), the sum of the diagonal values of C.

5. Results

In this paper we conclude:

- Volvos data does not follow the Pareto principle.
- In SSD the rows should be linearly independent, not the columns.
- SSD is a bit of a chance, but a reasonably good one.
- The expected relative gain per parameter with SSD is 1/sqrt(columns/rank), generally $\sqrt{\frac{Number\,of\ rows}{Number\ of\ columns}}$ for full rank.
- The total gain in lean optimization is expected to be $\sqrt{Number\,of\ rows * Number\,of\ columns}$.
- SSD is good only for reasonably many variables.

5.1 Volvo data and Pareto principle

In Figure 1, the scaled accumulated coefficients of some massive investigations have been plotted in size order. The dot is the point that the Pareto principle indicates. The thick line is the normal probability distribution, i.e. x_i is N(0,1). The analytic function behind this graph is given in Appendix 1. We can see that the data gathered here is more close to a normal probability distribution than the Pareto principle. Therefore we are going to use normal probability distributed x in the further reasoning.

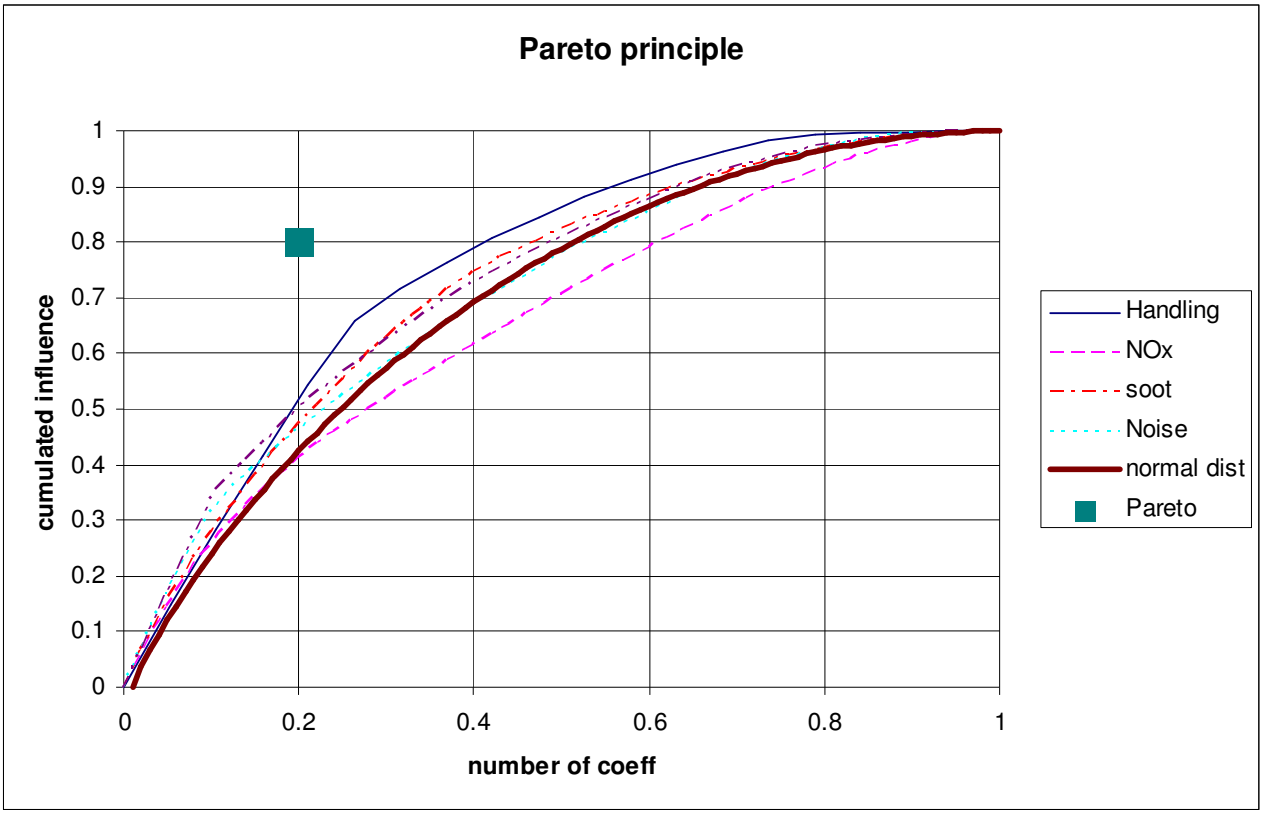

Figure 1 Pareto principle

5.2 Linearly independent rows

Since $\hat{x}$=Cx and $\hat{x}$ is the orthogonal projection of x on the row space of A, the important thing for a good design matrix A is to have full rank, m<n. The most common is to have as small correlation as possible between the columns of A, $E(s^2)$ [4], but $\hat{x}$ has maximal norm if x is in the row space of A since then $\hat{x}$=x so for estimating $\hat{x}$ in a supersaturated design, it is the independency of rows that is important, not the independency of the columns.

Orthogonal rows can easily be created by transposing classical designs. Another way is to concatenate a suitable number of unit matrices I so A=[I,I,I,…,I]. This is equivalent to lumping variables together so the first row is ones for the probably most influential variables and zeros for the other, then the next row is ones for some of the less probable variables and zeros for the others and so on.

5.3 Gain with Super Saturated Designs, SSD.

Recall that C is a projection matrix that projects x on the row space of A. A meter of the efficiency of a design A is the correlation coefficient between x and $\hat{x}$=Cx. If the correlation coefficient is 0.5, the gain is half the distance we are going. Let s be the saturation, s=(number of columns)/(rank). Then the correlation coefficient goes to 1/sqrt(s) when the number of columns goes high. This statement is proved in Appendix 2. For instance, if we have twice as many columns than (independent) rows, then s=2 and the correlation coefficient is $1/\sqrt{2} \approx 0.7$. This means that we expect to gain 70% with 50% of the effort if we take a reasonably large step in the direction of $\hat{x}$. This is proven only for normally distributed x.

The relative gain per parameter is $\sqrt{\frac{Number\ of\ rows}{Number\ of\ columns}}$. The gain for more columns is proportional to the number of columns so the total gain for a full rank investigation with super saturated design on a linear system is $\sqrt{Number\ of\ rows * Number\ of\ columns}$. This says that investigating some extra parameters is always a gain when doing an investigation.

Figure 2 shows that the less experiment are done, the more gain per experiment. We can see for instance that with a saturation of 4 you get 50% of the gain instead of expected 0.25.

Figure 3 shows that it takes a rank of at least 5 to get full effect of the method.

Figure 4 shows the spread in lean optimization compared to expected value, only simulations.21 rows is used for 184756 parameters or columns. We can see that in 10000 simulations the lowest value is 0.0055 when the easiest idea, number of rows/number of columns, is about 0.0001.

Since the null space of C has dimension at least n-m>0, there always is a possibility that ||Cx|| = 0. This happens if x is in the null space of A, and in this case you get no information of your experiments. The expected value of the information though is, relative to the number of experiments, higher than a full factorial experimental series.

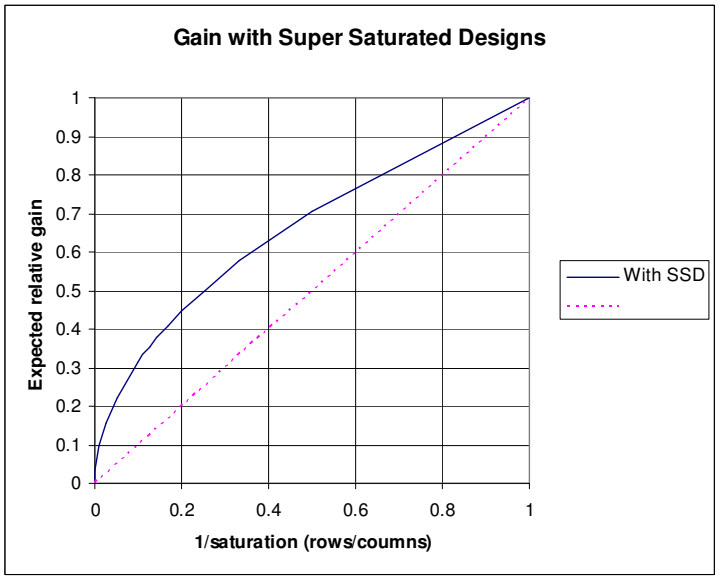

Figure 2 Gain with SSD

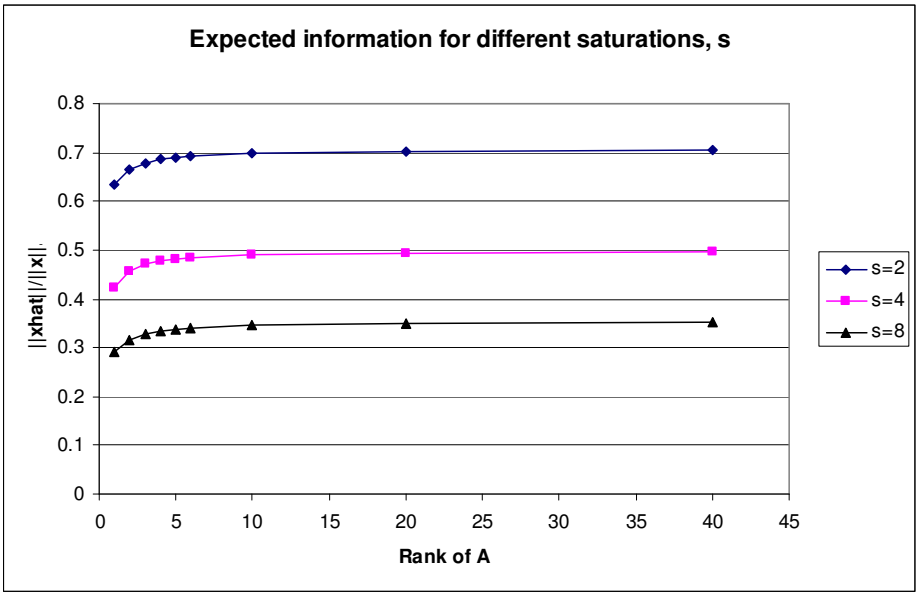

Figure 3 Rank for getting full gain

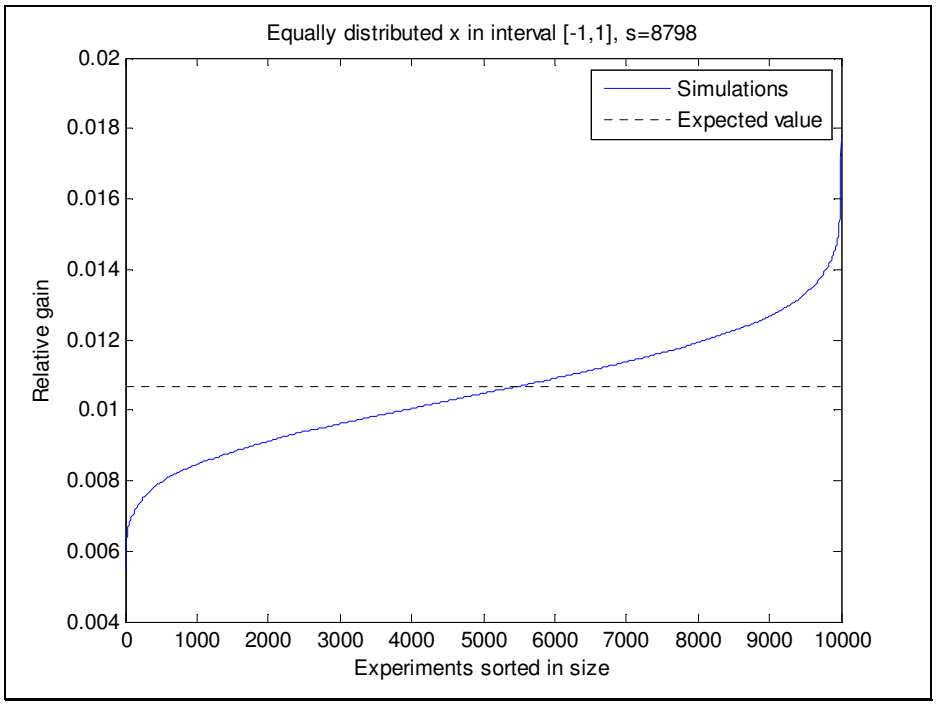

Figure 4 Equally distributed x

6. Discussion

It is some kind of symmetry that non saturated A:s should have independent columns and super saturated A:s should have independent rows. This follows from the idea of the information bottle neck which is the coefficients for non saturated designs and the observations for super saturated designs.

One can understand the result as that the first observation gives the most information and then the information per observation decreases.

The total gain in lean optimization is expected to be $\sqrt{Number\ of\ rows * Number\ of\ columns}$.

In this study, perturbations have not been studied. Probably such a study would result in that the rows in a super saturated A should not only be independent but orthogonal to minimize the influence of the perturbations on b.

7. References

Ref 1 R.A.Fisher The design of experiments ISBN 0028446909
Ref 2 Box Hunter Hunter Statistics for Experimenters ISBN 0-471-09315-7
Ref 3 Siomina Ahlinder Lean optimization using supersaturated experimental design, Applied Numerical Mathematics 58 (2008) 1-15
Ref 4 D.J.K.Lin, Generating systematic supersaturated designs, Technometrics 37(2)(1995) 213-225
Ref 5 J.W.Demmel Applied Numerical Linear Algebra ISBN 0-89871-389-7
Ref 6 Råde Westergren Beta Mathematics handbook for science and engineering ISBN 0-86238-406-0
Ref 7 Pareto principle http://en.wikipedia.org/wiki/Pareto_principle

8. Appendix 1

The thick line in Figure 1 is generated by reordering the elements of x in decreasing order of magnitude, defining a vector y by

$y_i = \sum_{k=1}^{i} |x_k|$ and then plot y.

In order to find an analytical function behind this graph we consider

$$F(x) = \int_z^\infty t e^{-t^2/2} dt$$

where the factor t in the integrand represents the size of an element and the factor $e^{-t^2/2}$ represent the corresponding probability density. The values x and z should be related by

$$x = \frac{2}{\sqrt{2\pi}} \int_z^\infty e^{-t^2/2} dt$$

where the integrand now just represents the probability density and $\frac{2}{\sqrt{2\pi}}$ is the required normalizing factor.

By the normal distribution function

$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt = 0.5 + \frac{1}{\sqrt{2\pi}} \int_0^x e^{-t^2/2} dt$ and since

$x = \frac{2}{\sqrt{2\pi}} \int_z^\infty e^{-t^2/2} dt = 1 - \frac{2}{\sqrt{2\pi}} \int_0^z e^{-t^2/2} dt$ we get

$2\Phi(z) = 1 + \frac{2}{\sqrt{2\pi}} \int_0^z e^{-t^2/2} dt = 2 - x$ or

$$z = \Phi^{-1}\left(\frac{2-x}{2}\right)$$

We also note that

$$\int_z^\infty te^{-t^2/2}dt = e^{-z^2/2}$$

so we may write the desired function

$$F(x) = e^{-\left\{\Phi^{-1}\left(\frac{2-x}{x}\right)\right\}^2/2}$$

For the inverse of this function we obtain a simpler, explicit formula.

Let $y = \int_z^\infty te^{-t^2/2}dt, \quad 0 \le y \le 1$

Then $y = e^{-z^2/2}, \quad \ln y = -z^2/2 \text{ and } \quad z = \sqrt{-2\ln y}.$

Further, $x = \frac{2}{\sqrt{2\pi}}\int_z^\infty e^{-t^2/2}dt = \frac{2}{\sqrt{2\pi}}\int_{\sqrt{-2\ln y}}^\infty e^{-t^2/2}dt$

so the inverse function of F is defined by

$$F^{-1}(y) = \frac{2}{\sqrt{2\pi}}\int_{\sqrt{-2\ln y}}^\infty e^{-t^2/2}dt$$

9. Appendix 2

Here we prove that the expected value of $\|\hat{x}\|/\|x\| = 1/\sqrt{s} = \sqrt{n/m}$,n>m. We assume that the matrix A has full rank; otherwise m should be replaced by the rank of A.

At first we recall some basic notations:

x is the estimated vector

$\hat{x}$ is the estimation of x

the norms are 2-norms

$A=USV^T$ is the singular value decomposition

$A^+=VS^+U^T$ is the Moore-Penrose inverse

$Ax=b$ and $\hat{x}=A^+Ax=Cx$.

Note that $\hat{x}=Cx$ is the orthogonal projection of x on the rows of A, so $\hat{x}^T(x-\hat{x})=0$ i.e. $\hat{x}^Tx=\hat{x}^T\hat{x}$.

Let α be the angle between x and $\hat{x}$. Then $\cos(\alpha) = (\hat{x}^T x)/(\|\hat{x}\|\|x\|) = \hat{x}^T\hat{x}/(\|\hat{x}\|\|x\|) = \|\hat{x}\|/\|x\|.$

Suppose that the components x_j of $x \in R^n$ are independent normally distributed random variables with expectation 0 and variance 1, i.e. x_j=N(0,1), j=1,…,n.

In the proof to follow we determine the expected norms E(||x||) and E(|| $\hat{x}$ ||) and then E(cos(v))=E(|| $\hat{x}$ ||)/E(||x||).

First we derive E(||x||) in the following lemma:

9.1 Lemma

The expected norm for $x \in R^n$ is

$$E(\|x\|) = \begin{cases} \dfrac{k!2^{k+1}}{\sqrt{2\pi}(2k-1)!!}, n = 2k+1 \\ \dfrac{\sqrt{2\pi}(2k-1)!!}{2(2k-2)!!}, n = 2k \end{cases}$$ Here n!!=n(n-2)(n-4)…

9.1.1 Proof

Since the probability density for each component x_j is $f(x_j) = \dfrac{1}{\sqrt{2\pi}} e^{-x_j^2/2}$ and the components are independent we get by standard probability analysis: E(||x||)= $\dfrac{1}{(2\pi)^{\frac{n}{2}}} \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \| x \| e^{-\|x\|^2/2} dx_1 \ldots dx_n$

By changing to spherical coordinates we readily get:

E(||x||)= $\dfrac{1}{(2\pi)^{\frac{n}{2}}} \int_0^{\infty} \int_0^{2\pi} \int_0^{\pi} \ldots \int_0^{\pi} r^n e^{-r^2/2} \sin\varphi_2 \sin^2\varphi_3 \ldots \sin^{n-2}\varphi_{n-1} d\varphi_{n-1} \ldots d\varphi_1 dr =$

$$= \frac{2\pi}{(2\pi)^{n/2}} \int_0^{\infty} \int_0^{\pi} \ldots \int_0^{\pi} r^n e^{-r^2/2} \sin\varphi_2 \sin^2\varphi_3 \ldots \sin^{n-2}\varphi_{n-1} d\varphi_{n-1} \ldots d\varphi_1 dr \qquad \text{(A2.1)}$$

For evaluating the integrals we use some standard definite integrals, see eg [Ref 6]:

$$\int_0^{\pi} \sin^n x dx = \begin{cases} \dfrac{2(n-1)!!}{n!!} & n = 1,3,5... \\ \dfrac{\pi(n-1)!!}{n!!} & n = 2,4,6... \end{cases} \qquad \text{(A2.2)}$$

and

$$\int_0^{\infty} x^n e^{-x^2/2} dx = \begin{cases} \dfrac{(2k-1)!!\sqrt{2\pi}}{2} & n = 2k \\ k!2^k & n = 2k+1 \end{cases} \qquad \text{(A2.3)}$$

By (A2.2 and A2,3) for the separate integrals in (A2.1) the formulas for the expected norm arise.

Q.E.D.

9.2 Theorem

Let $x \in R^n$ with x_i=N(0,1), i=1,2,…,n.

Then for $\hat{x}$=A^+Ax we get the expected norm:

(i) $E(||\hat{x}||)=E(||z||)$ where $z \in R^m$ and z_i=N(0,1), i=1,2,…,m
It follows that for the angle α between x and $\hat{x}$ holds, for n and m even or n and m odd,

(ii) $$E(\cos\alpha)=\frac{(m-1)!!(n-2)!!}{(m-2)!!(n-1)!!}$$

and for s=n/m we get

(iii) $$\lim_{m\to\infty} E(\cos\alpha)=s^{-1/2}$$

A similar formula to (ii) is valid also for the mixed odd- even case. Before giving the general proof we consider two special cases.

9.2.1 Example 1

The case of n=2, m=1. $\hat{x}$ is now the orthogonal projection of x on a line trough the origin. The norm is then $||\hat{x}||=||x||\cos\alpha$ and the value of the expected norm then becomes with r=||x||:

$$E(||\hat{x}||)=\frac{1}{\left(\sqrt{2\pi}\right)^2}\int_0^{\infty}\int_0^{2\pi} r^2\,|\cos\alpha|\,e^{-r^2/2}\,d\alpha dr =$$

$$=\frac{1}{2\pi}\int_0^{\infty} r^2 e^{-r^2/2}dr\int_0^{2\pi}|\cos\alpha|\,d\alpha=\frac{2}{\pi}\int_0^{\infty} r^2 e^{-r^2/2}dr$$

and by (A2,3) with k=1, n=2 we get

$$E(||\hat{x}||)=\frac{2}{\pi}\frac{\sqrt{2\pi}}{2}=\sqrt{2/\pi}$$

This value is the same as E(||z||) for $z \in R^m = R$ for the lemma with k=0, n=1. Thus the part (i) of the theorem is valid for this example.

9.2.2 Example 2

The case n=3, m=2. Here we use spherical coordinates for x=(x_1,x_2,x_3):

$$\hat{x}_1 = r\sin\theta\cos\varphi$$
$$\hat{x}_2 = r\sin\theta\sin\varphi$$
$$\hat{x}_3 = r\cos\theta$$

Assume $\hat{x}$ is the orthogonal projection of x on the (x_1,x_2)-plane. Then

$\hat{x}_1 = \sin\theta\cos\varphi$

$\hat{x}_2 = \sin\theta\sin\varphi$, with $||\hat{x}||=|r\sin\theta|$

$\hat{x}_3 = 0$

and then

$$E(||\hat{x}||)=\frac{1}{\sqrt{2\pi}^3}\int_0^\infty\int_0^{2\pi}\int_0^{\pi} |r\sin\theta|\, e^{-r^2/2} r^2\sin\theta\, d\theta d\varphi dr$$

where $r^2\sin\theta$ is the functional determinant. Further,

$$E(||\hat{x}||)=\frac{2\pi}{\left(\sqrt{2\pi}\right)^3}\int_0^\infty\int_0^{\pi} r^3 e^{-r^2/2}\sin^2\theta\, dr d\theta = \frac{\sqrt{2\pi}}{2}$$

since $\int_0^\infty \sin^2\theta\, d\theta = \frac{\pi}{2}$ and, by (A2.3), $\int_0^\infty r^3 e^{-r^2/2} dr = 2$

Finally, $E(\|\hat{x}\|) = \frac{\sqrt{2\pi}}{2} = E(\|z\|), z\in R^2, z_i = N(0,1), i = 1,2$ by the lemma with k=1, n=2. Thus for this example, the part (i) of the theorem is true.

9.2.3 Proof of the theorem

For simplicity, we restrict the proof to the case n even, the case n odd is treated similarly.

At first let m=n-1. Then in a spherical coordinate system based on the subspace Row(A) with dimension n-1, we have $||\hat{x}|| = |r\sin\varphi_{n-1}|$, compare with Example 2, and

$$E(||\hat{x}||)=\frac{1}{(2\pi)^{n/2}}\int_0^\infty\int_0^{2\pi}\int\cdots\int_0^{\pi} r^n e^{-r^2/2}\sin\varphi_2\sin^2\varphi_3\ldots\sin^{n-1}\varphi_{n-1}\, d\varphi_{n-1}\ldots d\varphi_2 d\varphi_1 dr =$$

$$\frac{2\pi}{(2\pi)^{n/2}}\int_0^\infty\int_0^{\pi}\cdots\int_0^{\pi} r^n e^{-r^2/2}\sin\varphi_2\sin^2\varphi_3\ldots\sin^{n-1}\varphi_{n-1}\, d\varphi_{n-1}\ldots d\varphi_3 d\varphi_2 dr$$

By using the formulas (A2.2) and (A2.3) for the separate integrals we obtain

$E(\|\hat{x}\|) = \frac{\sqrt{2\pi}(n-2)!!}{\pi(n-3)!!}$ and this is equal to

$E(\|z\|), z\in R^{n-1}, z_i = N(0,1), i = 1,2,\ldots,n-1$ by the lemma and the part (i) of the theorem is verified in this case.

Secondly, we consider n even and a general m<n. By generalizing the idea above of building up a spherical coordinate system successively from Row(A) of dimension m we get:

$$E(||\hat{x}||)=\frac{2\pi}{(2\pi)^{n/2}}\int_0^\infty\int_0^{\pi}\cdots\int_0^{\pi} r^n e^{-r^2/2}\sin\varphi_2\ldots\sin^{m-1}\varphi_m\sin^{m+1}\varphi_{m+1}\ldots\sin^{n-1}\varphi_{n-1}\, d\varphi_{n-1}\ldots d\varphi_2 dr$$

and by using once more the formulas (A2.2) and (A2.3) and the lemma we can readily prove that

$$E(\|\hat{x}\|) = E(\|z\|), z \in R^m, z_i = N(0,1), i = 1,2,...,m$$

i.e. part (i) of the theorem is proved.

For part (ii) of the theorem we just conclude that

E(cos(α))=E(|| $\hat{x}$ ||)/E(||x||)=E(||z||)/E(||x||)
where $x \in R^n$ x_j=N(0,1), j=1,…,n and $z \in R^m$ z_i=N(0,1), i=1,…,m
and from the lemma we then get for n and m even or n and m odd

$$E(\cos(\alpha)) = \frac{(m-1)!!(n-2)!!}{(m-2)!!(n-1)!!}$$

For part (iii), let n/m=s. Then

$$E(\cos(\alpha)) = \frac{(m-1)!!(sm-2)!!}{(m-2)!!(sm-1)!!} = \prod_{i=k}^{sk-1}\left(\frac{2i}{2i+1}\right)$$ where k=m/2

Denote this product by P_k. Then $\{P_k\}_{k=1}^{\infty}$ is an increasing series with upper limit $\lim_{k\to\infty}\left(\frac{2sk-2}{2sk-2+1}\right)^{sk-k} = \frac{1}{e^z}$

where $z = \frac{sk-k}{2sk-2}$

So, $P_k \to A$, $m \to \infty$, where the limit A<∞ exists.

Let now $P_k' = \prod_{i=k}^{sk-1}\frac{2i-1}{2i}$ Then

$$P_k P_k' = \frac{2k-1}{2sk-1} \to \frac{1}{s}, k \to \infty \qquad \text{A(2,7).}$$

One easily finds that $\frac{2k-1}{2k}P_k \le P_k' \le P_k \quad so\ P_k' \to A, k \to \infty$ as well. By (A2.3) we then conclude that

$A = \sqrt{\frac{1}{s}}$ and since k=m/2 the part (iii) is proven.

Q.E.D.