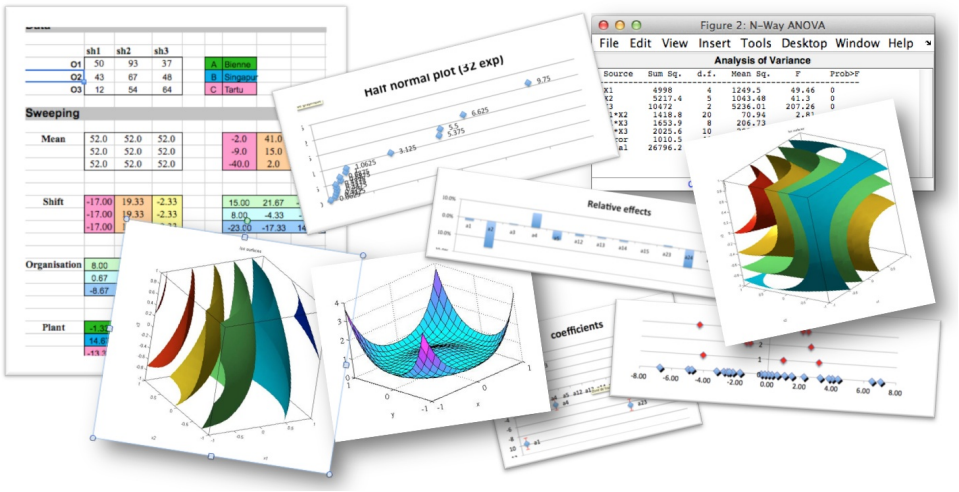
Design of Experiments



Jean-Marie Fürbringer
Doctoral Program in Robotics, Control and Intelligent Systems
Ecole Polytechnique Fédérale de Lausanne

Version fall 2020

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Design of experiments is a wonderful theoretical tool that acts as a backbone to many types of experimental research. Initiated by Sir Ronald Fisher around 1920, it gives statistical tools for the planning of an experimental campaign. It also provides a strategic perspective to research by giving handles for modeling phenomena and then to make specific objectives explicit. For experienced researchers this type of competence is often unconscious. For young researchers it can represent critical know-how.

The objective of the *Design of Experiments* course of the EPFL doctoral school is to help participants become familiar with the concepts of this theory, learn how to approach their experimental activities from a statistical point of view, and discover the properties of optimized designs in terms of efficiency. As a course companion, this text provides a base to facilitate first-timers in reading the specialized literature and help them communicate with DOE specialists.

With this document we review the concepts presented in the course and to bring some additional elements, mainly in the form of tables that will help you make your first steps in the experimental methodology. This will help you in your long-term learning. Secondly, this text is intended as a toolbox to which you can return when your memory and your understanding need refreshing. This text does not however replace the textbooks available for completing your knowledge.

The benefit an experimenter gets from a series of experiments depends on the care with which he chooses his experimental strategy. The strategy must take into account how the results will be analyzed: too often this is not done. The experimenter focuses his attention exclusively on the technical aspects of the measurements, without considering how he will be able to interpret the results. Such a procedure ends most of the time in very poor information. "*To the contrary, the end of everything must be kept in mind all the time*": this principle from the *7 Habits of Highly Effective People* by S. Covey [2] applies perfectly to experimental activity. The methodology of experimental design aims precisely to solve the

strategic problem and then gives an important place to the steps of planning and analysis.
Examples of the computing of the cases with Matlab are also provided. Those cases can also treated equally with an algorithmic software such as Python.
In completing the theoretical basis of this introductory course, take the time to consult some of the books and articles listed in the bibliography. The most classical references are *Statistics for Experimenters* by Box, Hunter and Hunter [3] and *Design and analysis of experiments* by Douglas Montgomery [4]. A recent book by Th. Ryan, *Modern Experimental Design*, is also of interest [5]. *Empirical model building and response surfaces* by Box and Draper [6] is more complex and is especially recommended for the inference of second-degree models and the elucidation of ridge systems.

1.1 An elementary example

The following elementary example shows how an optimised design allows one to get results of better quality than the classical one-factor-at-a-time approach.
Three objects, whose masses are of the same order of magnitude, have to be weighed using a scale with two plates as shown in figure 1.1.

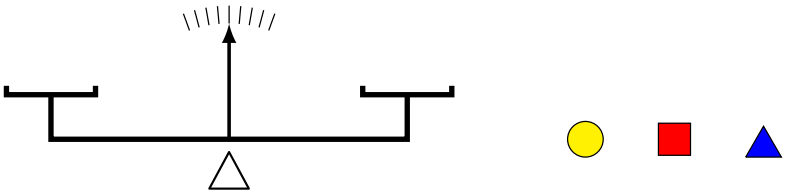


Figure 1.1: *Three objects to be weighed using a scale with two plates.*

The classical one-factor-at-a-time (OFAT) procedure is the following:

1. One measures the offset of the scale
2. One weighs the first object (meaning: recording the weight on the scale)
3. One proceeds in the same way for the second and the third objects

The weight of each object is computed by subtracting the offset from the second, third and fourth measurements.

If R_i is the result of the measurement i, with $i = 0, 1, ..., 3$, the weight of the object i, P_i is:

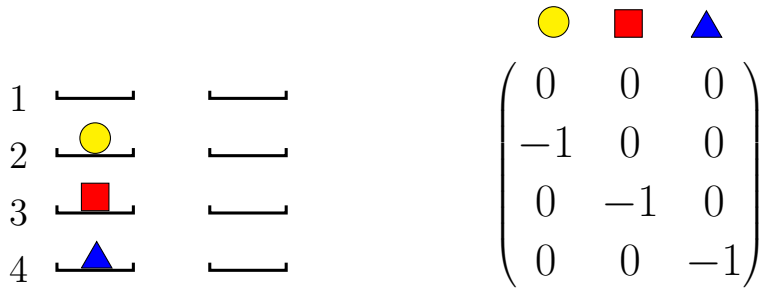
$$P_i = R_i - R_0 \tag{1.1}$$

The accuracy of the results can be estimated by the variance :

$$var(P_i) = var(R_i - R_0) = var(R_i) + var(R_0) \tag{1.2}$$

With the hypothesis that the experimental variance is the same in each experiment and that its value is σ^2, the equation 1.2 becomes:

$$var(P_i) = 2\sigma^2 \tag{1.3}$$

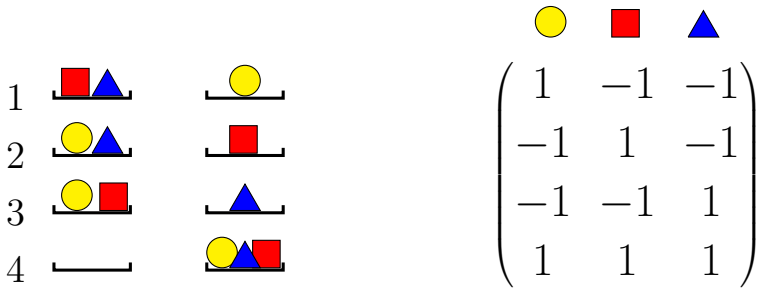


$$\begin{pmatrix} 0 & 0 & 0 \\ -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix}$$

Figure 1.2: *Graphical (left) and matrix (right) representation of the classical method for weighing three objects with a scale with two plates. The graphical representation shows on which plate the object is placed; the matrix representation has one column per object and one line per measurement to indicate the value that each factor takes in each experiment.*

Figure 1.2 provides two representations of the experiments that will be useful for understanding what happens. On the left hand side is the graphical representation and on the right hand side is the matricial one. The graphical representation shows two plates for each measurement and allows us to see on which plate an object is placed for the experiment. The matrix representation has one column per object and one line per experiment indicating the value that each factor takes for each experiment. In this case, the value '-1' indicates the left plate, the value '1' indicates the right plate, and the value '0' indicates that the corresponding object is not on the scale.
Plackett and Burman [8] propose an optimised design for this type of situation. Their design uses all the objects in each experiment, as well as the two plates.

The results of the four measurements allow us to develop a linear system of well-conditioned equations (figure 1.3). The weight of each object is computed by solving a system of four equations. The accuracy of the weight measured this way, with the same hypothesis on the measurement accuracy, is $\sigma^2/4$, which is eight times better than in the previous case.



$$\begin{pmatrix} 1 & -1 & -1 \\ -1 & 1 & -1 \\ -1 & -1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

Figure 1.3: *The best design uses the objects a maximum number of times. The graphical representation shows on which plate the objects are placed and the matrix representation counts one column per object and one line per experiment.*

Here is the proof:

$$\begin{aligned} \vec{P} &= \frac{1}{4} X^T \vec{R} \\ P_i &= \frac{1}{4} \sum_{i=1}^{4} x_{ij} \; R_j \\ var(P_i) &= var(\frac{1}{4} \sum_{i=1}^{4} x_{ij} \; R_j) = \frac{1}{16} \sum_{i=1}^{4} (x_{ij})^2 \; var(R_j) \\ var(P_i) &= \frac{1}{16} \sum_{i=1}^{4} (\pm 1)^2 \; var(R_j) = \frac{1}{4} \; \sigma^2 \end{aligned}$$

Why then does the Plackett-Burman design produce a more accurate result? The reason is that the new system transfers less experimental error (the variance). This comes from the fact that each object is measured four times instead of only once. The matrix of experiments presented in figure 1.3 is *balanced* and it is possible to see that each column counts the same number of '1' and '-1'. Figure 1.4 shows in graphical terms how the classical design is restricted to a smaller part of the experimental space than the Plackett-Burman design. The latter takes into account

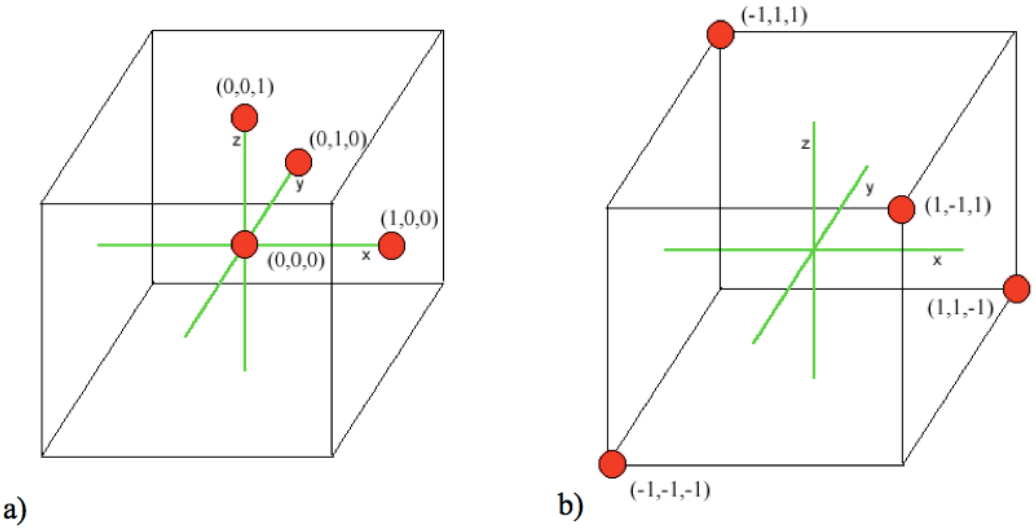


Figure 1.4: *a) Location of the measurement points in the experimental space when using a classical one-factor-at-a-time design. The central point corresponds to the measurement of the offset. The point at the centre of each face corresponds to the measure of each object. b) Location of the measurement points in a Plackett-Burman design.*

a space that is eight times bigger. The matrix used by Plackett and Burman is a Hadamard matrix, which has special properties. Hadamard matrices exist not only for three factors as in the present case, but also for more factors as presented in section 4.3. The theory of the design of experiments proposes several other types of optimal designs and a series of tools for the optimisation of experiment matrices [3, 4].

Chapter 2

Helicopter View and Mind Maps

2.1 Situation analysis

Even if the design of experiment is presented alongside a lot of mathematical and statistical concepts, the target is a strategical insight: what must be the positions of the experimental points in the experimental domain to ensure the best possible inference with a minimum of experimental effort? A good start is then to analyse the experimental domain and the model that is of interest. A helpful and efficient way to do that is through a mind map.
A generic map related to the improvement of a bicycle is presented in figure 2.1. With keywords and metadata, it aims to provide information on the objective of the study, the number and a few key information about the factors, the type of responses that will be analysed, the different models to be targeted and the different strategies foreseen. The anticipated pros and cons of each strategy can be also indicated. The map will be kept and adapted throughout the analysis and will inform the strategic decision taken at the start of the project and during its execution.
In this map there is no technological information. It focuses exclusively on the inference point of view. It is possible with one glance to understand the targeted objective and the different considered strategies.

- *The objective* is defined as the improvement of the bicycle.
- *The factors* are classified as design and environmental factors. Design factors correspond to characteristics of the bicycle that the experimenter wants to *design*. Environmental factors are those which can eventually be controlled in the experiment, but that will not be controlled outside of the laboratory. Factors can be classified in several ways: *discrete vs continuous, easy to change vs complicated* , etc. It is then also practical to indicate the range or the different values that the factors can take.

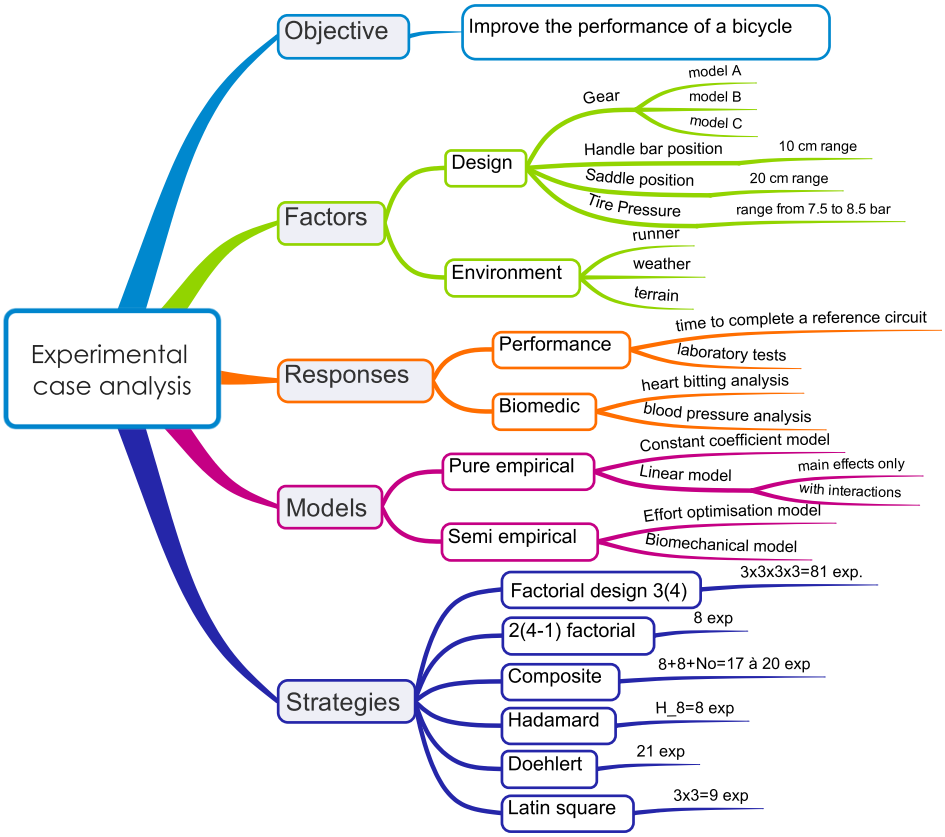


Figure 2.1: *Mind map of the experimental campaign aiming at the optimisation of a bicycle.*

- *The response* branch presents a classification of the results of the experiment that could be modeled. Sometimes one is chosen to represent the phenomenon, sometimes several are necessary to represent several *dimensions* of the problem.

- *The model* branch lists the alternative for modeling the phenomenon. It occurs that the previous experiments or the literature proposes one or more models. When no model has previously been proposed, or for the sake of simplicity, it is often interesting to consider an empirical model such as a first degree model or a first degree model with interactions (see subsection 4.1.1).

- *The strategy* branch lists the different designs that are considered in the initial analysis. A key piece of information is the number of experiments which are related to the cost of the campaign. Drawing relations between the designs

would eventually allow us to indicate a step by step strategy. For example starting with a small fractional design and completing it if necessary to a full factorial design if the interactions appear to be more important than initially expected.

This type of mind map can of course be done by hand in the laboratory notebook. Several software tools are available for the task such as *Freemind, iMindmap, Mindjet,* , etc.

2.2 Mind map of DOE

The conceptual map presented in figure 2.2 gives the organisation of the concepts for the three mains steps of the methodology: design, execution and analysis. The map can be used as scaffolding for you to put the concepts in place as you discover the methodology. This mind map has been done with *Cmap tool* .

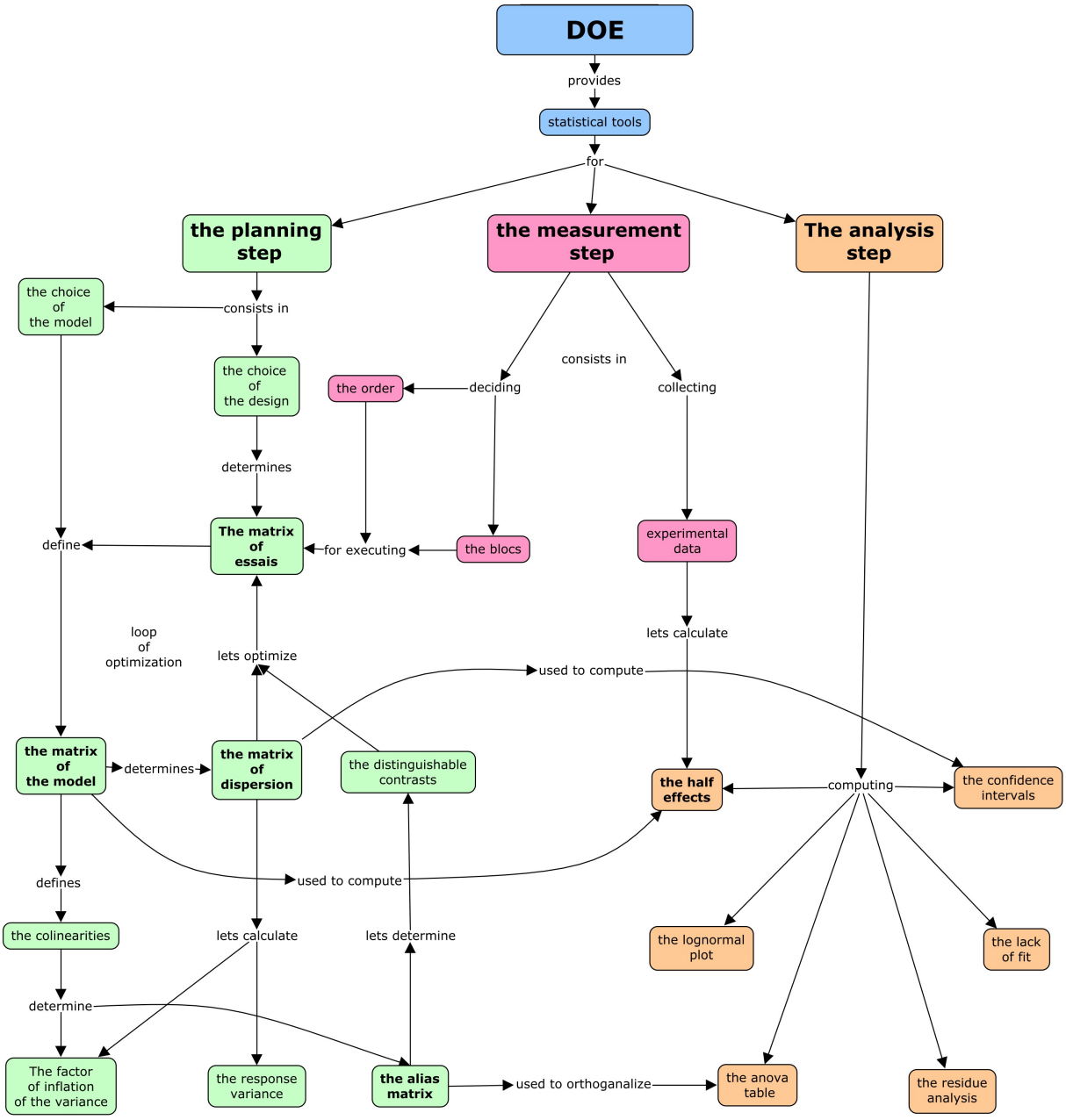


Figure 2.2: *DOE perspective showing how the different concepts are related and how they intervene in the experimental analysis.*

Chapter 3

Qualitative Factors

3.1 Constant coefficient model

A type of model that allows us to synthesise experiments made on qualitative factors are called *constant coefficient models*. Those models consider the direct effect of the factors that are called *main effects* and eventually interaction effects. The main effect is the direct influence of a factor on the modeled response. An interaction effect is the combined influence of two or more factors.

We can illustrate this with a simple example. Imagine that you like chocolate and strawberry ice-cream more than other flavors. But you dislike the chocolate ice-cream of the trademark A. For the other flavors of this trademark you do not have particular preference. Then, modeling your appreciation of several flavors and ice-cream trademarks, we would have a high main effect on the chocolate and on strawberry flavors, no effect on the trademark A and a significative interaction effect (in this case negative) for the chocolate ice-cream of trademark A.

3.1.1 The case of a workshop

We can now work with a more comprehensive case. An analysis is done in a workshop to determine the factors that influence the quality of the machining. An investment must be done implying a choice of machines and tools. The objective is then to determine the influence of the machines, of the tools and of the operators on the quality of the work and observe if some information can be obtained from the field. The response that will be analyzed is the surface quality of a given piece. This quality is defined here as the biggest defect h_{max} on the surface. The study consists in relating the quality to the factors such as the type of machines, the type of tools and the different operators.

A constant coefficient model without interactions of h_{max} can be defined as:

$$h_{max}(m,t,o) = \mu + \alpha_m + \beta_t + \gamma_o + \epsilon_{mtoi} \tag{3.1}$$

$$h_{max}(m,t,o) = \mu + \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} + \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{pmatrix} + \epsilon_{mtoi} \tag{3.2}$$

The indices m, t, o represent the machines, tools and operators. The index i represents eventual replicates. The real number μ is a constant representing the averaged surface defect. The real numbers α_m are a series of constant effects related to the different machines. The real numbers β_t are a series of constant effects related to the different tools. The real number γ_o are a series of constant effects related to the different operators. The real numbers ϵ_{mto} are the residues (the differences between the model and the measurements). In equation 3.2 the model has been detailed for a case with three machines, three tools and three operators. In a first step we applied a factorial strategy. The factorial strategy consists in performing measurement for each possible triplet (m, t, o). The results are given in table 3.1. We will see in the next section how to infer the model coefficients $\mu, \alpha_m, \beta_t, \gamma_o$ and the residues.

Table 3.1: *Results of a 27-run factorial experiment with three machines, three tools and three operators.*

Operators	Tool	Deckel	Schaublin	Maho
Charlie	1 mm	22.20 μm	30.30 μm	26.96 μm
	5 mm	15.73 μm	23.03 μm	23.98 μm
	20 mm	12.42 μm	17.46 μm	16.24 μm
Peter	1 mm	21.55 μm	29.05 μm	28.00 μm
	5 mm	16.83 μm	25.25 μm	24.78 μm
	20 mm	13.36 μm	19.76 μm	19.42 μm
Louis	1 mm	23.52 μm	30.46 μm	28.41 μm
	5 mm	17.57 μm	23.37μm	24.57 μm
	20 mm	10.55 μm	17.81 μm	17.57 μm

3.1.2 Sweeping

The determination of the coefficients of the constant coefficient model is called sweeping in reference to the iterative operation that is performed. The steps are:

1. calculation of the grand average
2. calculation of the main effects
3. calculation of the interaction effects

The order for determining the main effects of the different factors does not matter. But the *main effects* must be estimated before the interaction effects.
We can now consider the result vector $\vec{h}_{max}$ whose components are the results of the experiments. The dimension of this vector is 27 because the experiment counts 27 data points. The modeling procedure corresponds to the decomposition of this results vector in several orthogonal vectors. This decomposition is done by projecting the result vector on a specific base. This procedure is presented graphically in figure 3.1.

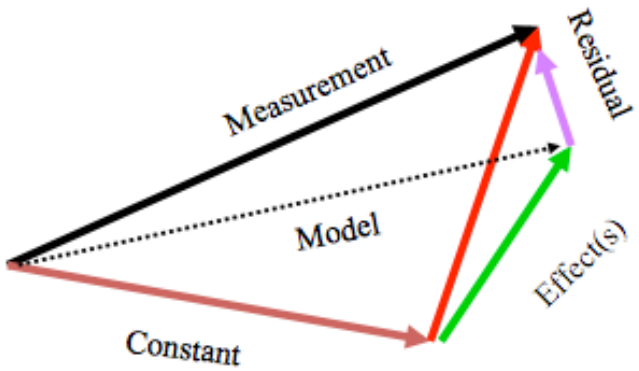


Figure 3.1: *Decomposition of the measurement result vector in constant, effect(s) and residual.*

The first element of this base is a vector whose components all have the same value $(1, 1, ..., 1)^T$. Then the first vector of the model called the *constant vector* has equal components and its value is the *grand mean*, the average of the whole data set.
The results vector $\vec{h}_{max}$ can now be decomposed in a constant vector $\vec{\mu}$ and a residual vector $\vec{\epsilon_\mu}$ such as:

$$\vec{h}_{max} = \vec{\mu} + \vec{\epsilon}_\mu \quad \text{with} \quad \vec{\mu} \perp \vec{\epsilon}_\mu \tag{3.3}$$

The next element of the decomposition are the vectors corresponding to the main effects. There is no importance of the order. Successively for each factor the former residue vector $\vec{\epsilon}$ is decomposed in two orthogonal vectors, one for the effect and one for the new residue. The components of the main effect vector of a given factor are obtained by the averages of the former residue according to the levels of the analysed factor.
For our example, we start with the factor 'machine'. Since this factor has three levels, although the vector corresponding to the main effect of the machines has 27 components, those components have only three distinct values: one value for the machine 'Deckel', one value for the machine 'Schaublin' and one value for the machine 'Daho'. These values are obtained by averaging the residual in three different groups, one for each machine. In the vector corresponding to the main effect of the machines, the suitable average value is placed following the order of the experiments. A new residue is calculated by subtracting the main effect vector from the former residue.

$$\vec{h}_{max} = \vec{\mu} + \vec{\alpha} + \vec{\epsilon}_{\mu\alpha} \tag{3.4}$$

The operation is repeated to treat all the factors. For our workshop example it gives:

$$\vec{h}_{max} = \vec{\mu} + \vec{\alpha} + \vec{\beta} + \vec{\gamma} + \vec{\epsilon}_{\mu\alpha\beta\gamma} \tag{3.5}$$

Notice that equation 3.2 must not be confused with equation 3.5. Both represent the same model. Nevertheless the former is a concise representation in which only the different values of the constant effects are represented. The latter is a vectorial equation. The application of this procedure to the workshop data gives as concise form:

$$h_{max}(m,t,o) = 21.5 + \left\{ \begin{array}{r} -4.41 \\ 2.57 \\ 1.84 \end{array} \right\} + \left\{ \begin{array}{r} 5.23 \\ 0.19 \\ -5.42 \end{array} \right\} + \left\{ \begin{array}{r} -0.56 \\ 0.51 \\ 0.05 \end{array} \right\} + \epsilon_{\mu\alpha\beta\gamma} \tag{3.6}$$

On a spreadsheet, the procedure can be applied easily. This has the advantage of showing precisely how the different coefficients are calculated. Figure 3.2 shows the decomposition of the data vector in several steps. Figure 3.3 shows the application of this procedure to the workshop example within Excel.The contrasts of colors show the different groups used to calculate the coefficients for each level of the factors.
An interesting way to present the results of the sweeping is a *dot plot*, as in figure 3.4. The base line is reserved for the residuals and the next lines show the values of the effects. At a glance it is then possible to see which effects are obviously bigger or not than the noise. In this example, it is evident that the operator effect is questionable as its span is significantly smaller than the residual span. A more precise analysis will be proposed with the ANOVA, but this type of graphic is a straightforward means to illustrate the results of the inference procedure.

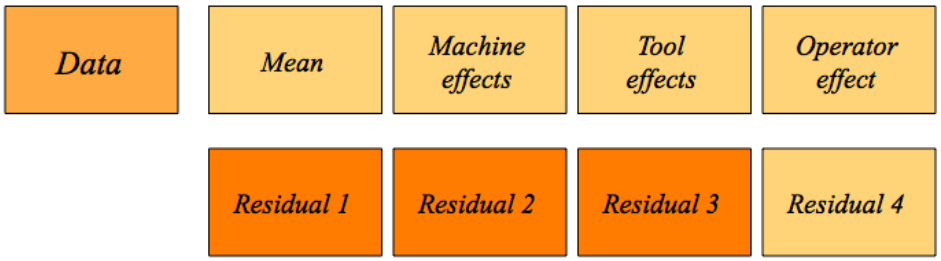


Figure 3.2: *Decomposition of the data vector in several steps.*

3.1.3 Interaction

The sweeping process can be continued to infer interaction coefficients. The targeted model is now :

$$h_{max}(m,t,o) = \mu + \alpha_m + \beta_t + \gamma_o + \alpha\beta_{mt} + \alpha\gamma_{mo} + \beta\gamma_{to} + \epsilon_{mtoi} \tag{3.7}$$

Terms like $\alpha\beta_{mt}$ have $M \times T$ numbers, if M is the number of machines and T the number of tools. There is one number per possible couple (m,t). The sweeping procedure is the same as previously: the effect is computed by averaging data that are equivalent between themselves. The main effects have been calculated by taking the average of all the data corresponding to a given level of a factor. Now interaction effects are calculated by taking the average of data corresponding to given levels of two variables. For example, the interaction coefficient $\alpha\beta_{13}$ is calculated averaging the adequate residue for the data corresponding to $m = 1$ and $t = 3$. This is possible only if there is more than one such data point. This is the case in our workshop example in which there are three data points for each couples.

3.2 Latin and Graeco-Latin squares

In the model given at equation 3.6 there are ten coefficients and seven degrees of freedom. This has been obtained with 27 experiments. The residue has then 20 degrees of freedom. This is good for diminishing the confidence interval. In practice however, we are interested to find a way to determine the effects with the least possible effort. This can be obtained by using magic squares. Magic squares are arrangements of levels that provide all the possible pairings, but not all triplets or higher order of combination. When the number of levels is the same for each variable, those squares are called *Latin squares* and *Graeco-Latin squares*. Their names come from their invention by the Swiss mathematician Leonhard Euler who

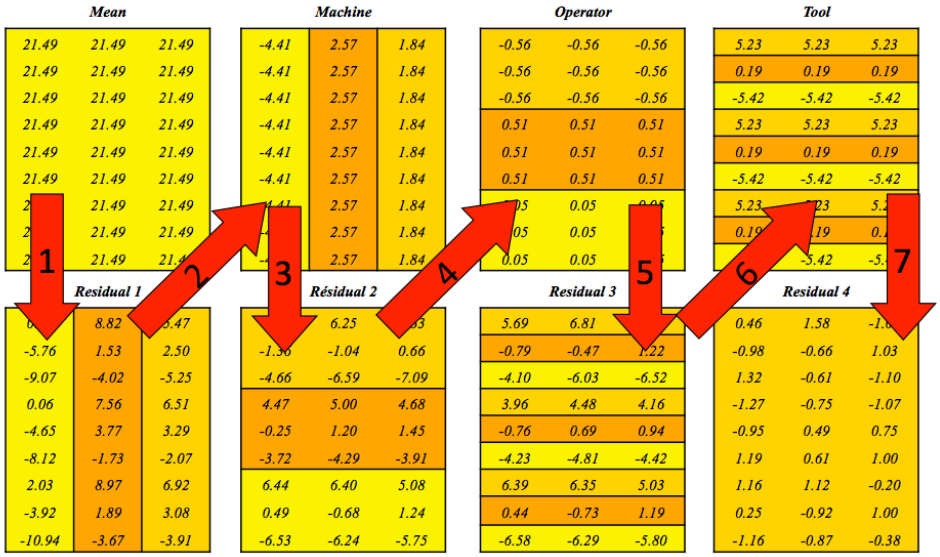


Figure 3.3: *Application of the sweeping procedure to the workshop example within Excel with 27 experiments.The contrasts of colors reveal the different groups used to compute the coefficients for each level of factors.*

used Greek and Latin letters to represent them as presented by Mac Neish in a paper of the Aannals of Mathematics [9]:

> *Euler Squares were first considered in a paper, 'Recherches sur une espèce de carrés magiques', Commentationes Arithmetica Collectae, 1849, vol. II, pp. 302-361. In this paper Euler proposed the following problem now well known as 'The problem of the 36 officers'. Six officers of six different ranks are chosen from each of six different regiments. It is required to arrange them in a solid square so that no officer of the same rank or of the same regiment shall be in the same row or in the same column. The problem is equivalent to that of arranging 36 pairs of integers, each less than or equal to six, in a square array so that the first(or second) numbers of the pairs in any row or column are all distinct,and no two pairs are identical.*

Those magic squares are orthogonal and can be combined. For our example, we take the 3×3 latin square (table 3.2). This would correspond to the nine experiments given in table 3.3. Each operator tests each machine and each tool. But not all the triplets are tested. This will allow us to estimate the main effects with only nine experiments, as shown in figure 3.6.

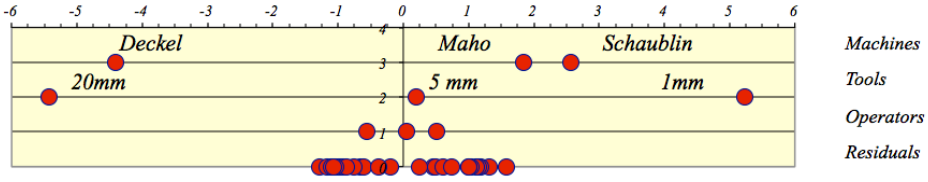


Figure 3.4: *Dotplot of the model obtained with 27 experiments in the workshop example.*

Table 3.2: 3×3 *Latin square*

a	b	c
b	c	a
c	a	b

A concise representation of the model is now given in equation 3.8 that must be compared with equation 3.6. The corresponding dot plots must also be compared (figures 3.4 and 3.5):

$$h_{max}(m,t,o) = 21.3 + \left\{ \begin{array}{r} -4.73 \\ 2.04 \\ 2.68 \end{array} \right\} + \left\{ \begin{array}{r} 5.30 \\ 0.14 \\ -5.44 \end{array} \right\} + \left\{ \begin{array}{r} -0.04 \\ 0.51 \\ -0.48 \end{array} \right\} + \epsilon_{\mu\alpha\beta\gamma} \quad (3.8)$$

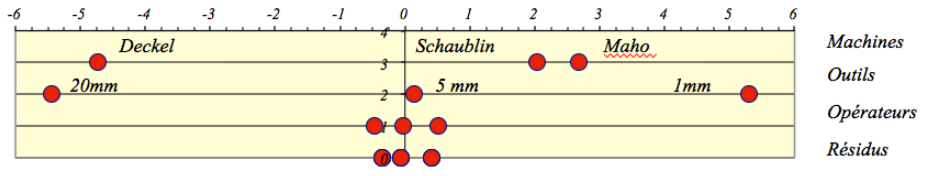


Figure 3.5: *Dot plot of the model obtained with nine experiments in the workshop example.*

Table 3.3: *Use of a* 3×3 *Latin square for testing three factors at three levels.*

	Deckel	Schaublin	Maho
1 mm	Charlie	Peter	Luis
5 mm	Peter	Luis	Charlie
10 mm	Luis	Charlie	Peter

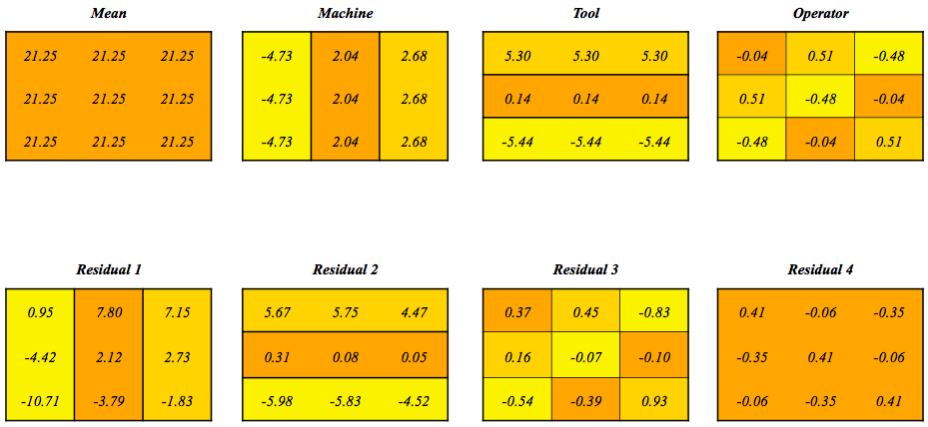


Figure 3.6: *Application of the sweeping procedure to the workshop example within Excel with nine experiments. The contrast of colors shows the different groups used to compute the coefficients for each level of factors.*

3.3 Results analysis and ANOVA

The question that we want to answer in this section is about the validity of the models obtained by the sweeping procedure. We have seen that we can obtain a model of nine coefficients with 27 or nine experiments. How do these models differ in term of accuracy? More generally what is the risk taken by using one or the other of these models?
We can begin by making some observations on figures 3.4 and 3.5:

- The residue of the data set of 27 experiments show several distinct values when the residues of the 9-experiment data show only three different values. This is due to the fact that the former residues have 20 degrees of freedom when the latter have only three degrees of freedom (we will return to this

point).

- The span of the effects of the machines and of the tools are significantly bigger than the span of the residue. This is not the case for the operator effects which do not differ significantly from the noise.

- Between the two models, the machines Maho and Schaublin have interchanged their positions. The difference between these two machines is very probably not significant, especially in the second model which has been determined with less data.

To determine the validity of a model in relation to experimental data, Ronald Fischer invented the ANOVA, the analysis of variance. Consider figure 3.1 where the vector of results has been decomposed in several vectors for the constant and the different effects and finally the residual. We are in a process of determining the possible causal relations between an answer and some factors. This possible cause-consequence link is determined by the correlations that exist between them and that are represented by the model. The constant is then related to unidentified factors that are not changed during the experiment. The coefficients of the model are related to the factors targeted by the set of experiments and are then varied during the experiments. The residue is due to unidentified factors that are not controlled and thus have changed during the experience. So the rationale on which the validity of the model is based is equivalent to saying: let's consider as probable causes of the change of the response the factors whose coefficients are significantly bigger than the residue. So the first step consists in comparing the size of the vectors representing the factors to the size of the vector of the residue. The size of the vectors can be represented by the euclidian norm

$$\|\vec{x}\| = \sqrt{\sum_{i=1}^{N} x_i^2} \tag{3.9}$$

if x_i are the components of a vector $\vec{x}$.
The square root conserving the relation of order between positive numbers, the equivalent comparison can be made between the sum of squares (SS). Hence, the sum of squares given in tables 3.4 are computed from the data in figures 3.3 and 3.6. From both cases it can be observed that the effects of the machines and of the tools are more than 100 times bigger than the residue. This gives consistency to the hypothesis that the machines and the tools have an effects on the quality of the pieces. To the contrary, in the 27-experiment case, the sum of squares for the effect of the operators is 5 times smaller that the residue. It is 1.6 times bigger for the 9-experiment case.

Table 3.4: *Sum of squares for the 27- and 9-experiment sets.*

(a) *27-experiment set*

Source	SS
Constant	12'465.7
Machine	264.5
Tool	511.0
Operator	5.2
Residue	24.2
Total	13'270.6

(b) *9-experiment set*

Source	SS
Constant	4'064.9
Machine	101.1
Tool	173.2
Operator	1.5
Residue	0.9
Total	4'341.6

But the different vectors that are compared do not have the same degree of freedom. A *fair* comparison between the vectors must take this into account and the comparison must be done between mean squares that are the sum of squares divided by the degree of freedom. As the effects of one factor sum to zero, constituting one constraint, the degree of freedom (DF) of a given factor is given by

$$DF = \text{nbr of levels} - 1 \tag{3.10}$$

The mean square (MS) is

$$MS = \frac{SS}{DF} \tag{3.11}$$

Table 3.5: *Sum of squares, degree of freedom and mean squares for the 27- and 9-experiment sets.*

(a) *27-experiment set*

Source	SS	DF	MS
Constant	12'465.7	1	12'465.70
Machine	264.5	2	132.23
Tool	511.0	2	255.50
Operator	5.2	2	2.62
Residue	24.2	20	1.21
Total	13'270.6	27	

(b) *9-experiment set*

Source	SS	DF	MS
Constant	4'064.9	1	4'064.91
Machine	101.1	2	50.55
Tool	173.2	2	86.61
Operator	1.5	2	0.74
Residue	0.9	2	0.44
Total	4'341.6	9	

The effects *machines* and *tools* are now clear as their mean squares are significantly bigger than the mean square of the residue. Comparing the two cases, it is possible to observe that the 9-experiment set seems to bring us to the same conclusion as

the 27- one, meaning that *operator* is not a significative factor. Let's go one step further to obtain a definitive confirmation of this hypothesis.
The Fisher ratio (F) is the ratio between the mean square of a given coefficient and the mean square of the residue. The last step consists in calculating the probability to get a given Fisher ratio. In the common hypothesis that the data is normally distributed following the distribution $N(\mu_i, \sigma^2)$ (hypothesis of homoscedasticity), a sum of squares of ν terms will follow the distribution $\chi^2(\nu)$ and the ratio of mean squares the F-distribution $F(\nu_1, \nu_2)$ where ν_1 and ν_2 are the degrees of freedom of the numerator and denominator[1]. The ANOVA table can now be completed as shown in table 3.6. The Fisher ratio of the constant is usually not presented; the whole line for the constant is even sometimes forgotten.

Table 3.6: *ANOVA table for the 27- and 9 experiment sets.*

(a) *27-experiment set*

Source	SS	DF	MS	F	p
Constant	12'465.7	1	12'465.70		
Machine	264.5	2	132.23	109.3	0.000%
Tool	511.0	2	255.50	211.19	0.000%
Operator	5.2	2	2.62	2.2	14.1%
Residue	24.2	20	1.21	1	
Total	13'270.6	27			

(b) *9-experiment set*

Source	SS	DF	MS	F	p
Constant	4'064.9	1	4'064.91		
Machine	101.1	2	50.55	113.7	0.000%
Tool	173.2	2	86.61	194.7	0.000%
Operator	1.5	2	0.74	1.7	21.5%
Residue	0.9	2	0.44	1	
Total	4'341.6	9			

The probability, called p-value, represents the probability of getting the ratio by chance. Hence the smaller the p-value, the higher is the significance of the considered effect. The standard criteria for accepting an effect is that $p < 5\%$. In specific industries such as pharmaceutics or aeronautics, the level of confidence

[1] On Excel 15 `p = loi.F.droite(x,` ν_1, ν_2`)`

can be higher. At the opposite end, in some research situations, especially when screening for factors, the criteria for rejecting of factors is sometimes higher.

MATLAB - anovan()

The built-in function [p, tbl, stats] = anovan(y, group) computes the n-way ANOVA table for testing the effects of multiple factors on the mean of the vector *y*. The vector *p* correspond to the p-values. The table *tbl* is the ANOVA table and the computed data is recorded in the structure *stats*. Here is a small example of the analysis of a latin square experiment:

The Latin square design 3×3:

	I	II	III
a	A	B	C
b	B	C	A
c	C	A	B

The data :

	I	II	III
a	-10.09	7.14	4.11
b	-7.33	2.75	8.19
c	-2.67	17.01	16.75

The code

```
Group_1 = reshape(repmat({'I','II','III'},3,1),9,1);
Group_2 = reshape(repmat({'a';'b';'c'},1,3),9,1);
Cat_3 = {'A','B','C'};
latin3 = [1 2 3;2 3 1;3 1 2];
Group_3 = reshape(Cat_3(latin3),9,1);
stats = anovan(Y,{Group_1,Group_2,Group_3},...
          "model","linear");
```

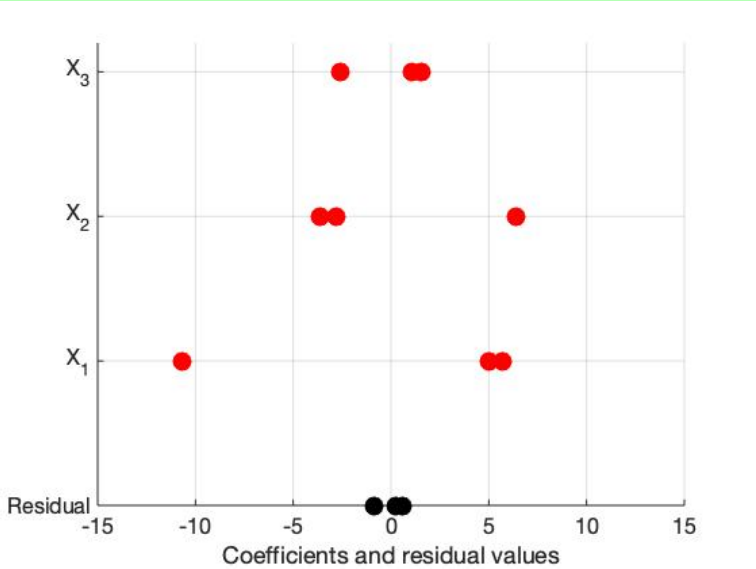
Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
X1	514.15	2	257.075	149.34	0.0067
X2	184.042	2	92.021	53.46	0.0184
X3	30.419	2	15.209	8.84	0.1017
Error	3.443	2	1.721		
Total	732.053	8			

Constrained (Type III) sums of squares.

Dotplot

```
var=[1 1 1 2 2 2 3 3 3]; % coefficients vs ...
   variables
sz=200;% size of the bullet
scatter(stats.coeffs(2:end),var,sz,'red',"filled")
grid on
axis([-15,15,0,3.2])
yticks(0:1:3)
yticklabels({'Residual','X_1','X_2','X_3'})
xlabel('Coefficients and residual values')
hold on
    scatter(stats.resid,zeros(1,9),sz,'k',"filled")
hold off
```



3.4 Some Latin and Graeco-Latin squares

The following is a list of common Latin squares of 3- 4- and 5 levels that can be combined in Graeco-Latin and hyper-Graeco-Latin squares.

Table 3.7: *Two orthogonal* 3×3 *Latin squares (a and b) that can be combined in a Graeco-Latin square(c) to test up to four factors of three levels.*

(a)

A	B	C
B	C	A
C	A	B

(b)

A	B	C
C	A	B
B	C	A

(c)

Aα	Bβ	Cγ
Bγ	Cα	Aβ
Cβ	Aγ	Bα

Table 3.8: *Three orthogonal* 4×4 *Latin squares that can be combined in one Graeco-Latin square or one hyper-Graeco-Latin square to test respectively up to four and five factors of four levels.*

(a)

A	B	C	D
B	A	D	C
C	D	A	B
D	C	B	A

(b)

A	B	C	D
D	C	B	A
B	A	D	C
C	D	A	B

(c)

A	B	C	D
C	D	A	B
D	C	B	A
B	A	D	C

Table 3.9: *Three orthogonal* 5×5 *Latin squares that can be combined in one Graeco-Latin square or one hyper-Graeco-Latin square to test respectively up to four and five factors of five levels.*

(a)

A	B	C	D	E
C	D	E	A	B
E	A	B	C	D
B	C	D	E	A
D	E	A	B	C

(b)

A	B	C	D	E
D	E	A	B	C
B	C	D	E	A
E	A	B	C	D
C	D	E	A	B

(c)

A	B	C	D	E
E	A	B	C	D
D	E	A	B	C
C	D	E	A	B
B	C	D	E	A

Chapter 4

Quantitative Factors

This chapter introduces the main ideas that allow us to transform data into information. As the objective of this course is the experimental strategy, only the main concepts are presented here. If you are not familiar with the multilineal regression, you are strongly recommended to consult a basic textbook to extend and develop the points provided here.
When analysing a system by the experimental approach, it is interesting to made it through successive empirical models. The objective is then to model the behaviour of the system by fitting of the experimental data obtained through a series of tests made by varying the inputs of the system, x_i, on a series of functions, one per output of interest, y_j:

$$y_j = f_j(x_1, \ldots, x_i, \ldots, x_N) \tag{4.1}$$

Graphically this can be represented by a bloc diagram as represented in Figure 4.1. This approach is sometimes called the black box approach.

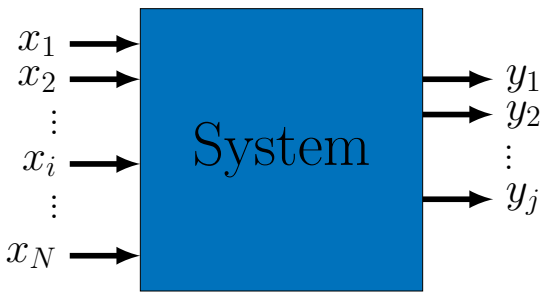


Figure 4.1: *Graphical representation of the bloc diagram of a black box*

The most simple model

4.1 Representing experiments with matrices

There are two main methods of data analysis and both will be used in this course:

1. The comparison of the data with a statistical distribution (such as the normal distribution)
2. The comparison of a subset of the data with another subset of the data (such as comparing the effects with the residual error)

Now let us introduce a few definitions necessary to develop the theory:

Response: A *response* is any manifestation or consequence of a phenomenon. It can be a qualitative or quantitative property. The *response* is the dependent variable, the consequence of the phenomenon under study.

Factor: A *factor* is any variable (or parameter) which has, in reality or all likelihood, an influence on the studied phenomenon. The factors are considered as the possible causes of the response. The *factors* are the independent variables.

Level: The *level* is the state, the value of a factor.

Standardisation of the factors: There is significant interest in working with coded factors centered on zero and varying in the interval $[-1, 1]$. It allows a direct comparison of the effects of each factor independently of their respective range and order of magnitude. As presented in figure 4.2, the coded variable x_i is achieved by the transformation of the natural variable u_i:

$$x_i = \frac{u_i - u_{i0}}{\Delta u_i} \quad with \quad \begin{cases} u_{io} = \frac{\max u_i + \min u_i}{2} \\ \\ \Delta u_i = \frac{\max u_i - \min u_i}{2} \end{cases} \tag{4.2}$$

The reverse function is:

$$u_i = u_{io} + x_i \, \Delta u_i \tag{4.3}$$

4.1.1 Matrix of experiments and matrix of the model

Matrices of experiments, and of the model, are the key concepts on which the theory is built. Let us continue with a few generic definitions :

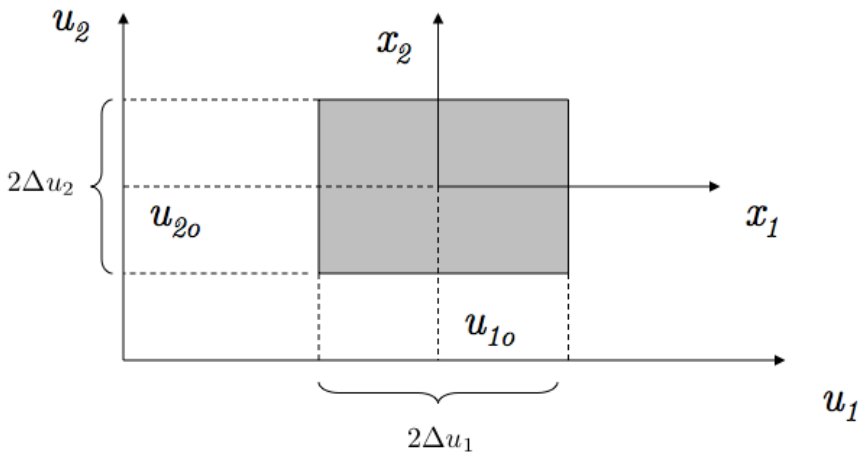


Figure 4.2: *Standardisation of the experimental domain.*

Matrix of experiments: The *matrix of experiments* E is the matrix of N rows and k columns whose element x_{ij} corresponds to the level of the factor j in the experiment i. There is then one row per experiment, and one column per factor. In the example below, a set of 9 experiments to test a solar panel in function of 3 factors, intensity of the radiation I, the angle of incidence θ and the outdoor temperature T, is given with the original values of the factors , W/m^2, deg, ^oC .

$$E = \begin{pmatrix} x_{11} & x_{12} & x_{13} \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ x_{N1} & x_{N2} & x_{1N3} \end{pmatrix} = \begin{matrix} I \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{matrix} \begin{matrix} \theta \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{matrix} \begin{matrix} T \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{matrix} \begin{pmatrix} 150 & 35 & 15 \\ 100 & 30 & -10 \\ 200 & 30 & -10 \\ 100 & 30 & 40 \\ 200 & 30 & 40 \\ 100 & 40 & -10 \\ 200 & 40 & -10 \\ 100 & 40 & 40 \\ 200 & 40 & 40 \end{pmatrix} = \begin{matrix} x_I \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{matrix} \begin{matrix} x_\theta \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{matrix} \begin{matrix} x_T \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{matrix} \begin{pmatrix} 0 & 0 & 0 \\ -1 & -1 & -1 \\ 1 & -1 & -1 \\ -1 & -1 & 1 \\ 1 & -1 & 1 \\ -1 & 1 & -1 \\ 1 & 1 & -1 \\ -1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \tag{4.4}$$

Empirical models: When dealing with quantitative factors it is possible (and interesting) to mathematically model the data. The parsimony principle, one of the

basic principles of science, asks us to give priority to the simplest possible model that adequately explains the data. In this perspective the following set of empirical models are of major importance (a graphical representation is given in figure 4.3):

- The constant model $y = \alpha_o + \epsilon_k$ that tends to model the data as a constant α_o with a residue ϵ_k for each of the k measurements.
- The first degree model $y = \alpha_o + \sum_i \alpha_i x_i + \epsilon_k$ that tends to model the data as a constant plus linear effects α_i for each of the x_i factors and a residue.
- The first degree model with interactions $y = a_o + \sum_i \alpha_i x_i + \sum_{i<j} \alpha_{ij} x_i x_j + \epsilon_k$ that integrates interaction effects α_{ij} to model the coupled effect of factors x_i and x_j.
- The second degree model $y = a_o + \sum_i \alpha_i x_i + \sum_{i \leq j} \alpha_i x_i x_j + \epsilon_k$ that integrates second degree effects α_{ii}.

Matrix of the model: Saying that an experimental result follows a given linear model is equivalent to saying that a linear system of equation relate the vector $\vec{\eta}$ of the results of N experiments and the vector $\vec{\alpha}$ with the coefficients of the model:

$$\vec{\eta} = X\vec{\alpha} \tag{4.5}$$

The model matrix X is built by the association of a linear model and a matrix of experiment E and has then a row per experiment and a column per coefficient. For a linear model with interactions for 2 factors x_1 and x_2, with N experiments the matrix of the model is a matrix $N \times 4$ as presented in equation 4.6.

$$\left.\begin{array}{l} y = a_o + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_{12} x_1 x_2 + \epsilon_k \\ E = \begin{pmatrix} x_{11} & x_{21} \\ \vdots & \vdots \\ x_{1N} & x_{2N} \end{pmatrix} \end{array}\right\} X = \begin{pmatrix} 1 & x_{11} & x_{21} & x_{11}x_{21} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1N} & x_{2N} & x_{1N}x_{2N} \end{pmatrix} \tag{4.6}$$

In Matlab: There is a built-in function `x2fx(X, mld_spec)` that builds the matrix of the model, where `X` is the essay matrix and `mld_spec` the model specification

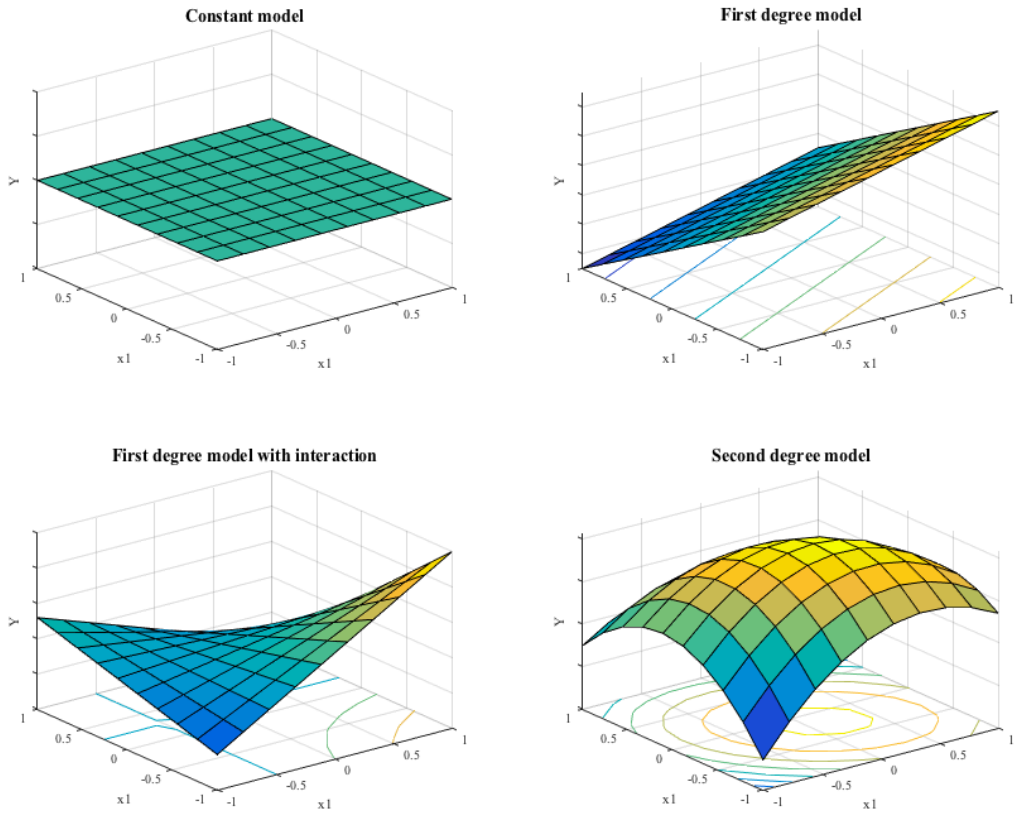


Figure 4.3: *Empirical models of increasing degree usable to model phenomena.*

with a column per factor and a line per coefficient. To build the model matrix of equation 4.6, the model specification would be the following:

$$\texttt{mld\_spec} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix} \tag{4.7}$$

4.1.2 Least squares fit, matrix of dispersion and matrix of correlation

Least squares resolution: In most of the cases, the matrix of the model is not square and thus not inversible. The coefficients α_i of the model can be obtained by way of a numerical process. The most current process is the least squares fit algorithm (LSF)[1]. For this algorithm, the linear system of equation 4.5 is replaced by

$$\vec{Y} = X\vec{\alpha} + \vec{\epsilon} \tag{4.8}$$

The hypotheses are that the measurements $\vec{Y}$ are random variables following a normal distribution $N(\eta(x), \sigma^2)$ and that the components of the vector of residue $\vec{\epsilon}$ follow a normal distribution $N(0, \sigma^2)$. The coefficients α_i of the model can then be estimated through the generalized least squares algorithm :

$$\hat{\vec{\alpha}} = \left(X^T X\right)^{-1} X^T Y \tag{4.9}$$

in Matlab: There is a built-in function `mdl = fitlm(X, y, modelspec)` that performs the multilinear regression of y on X. The output is a *linearModel class* object to which several *methods* can be applied to obtain standard plots, tests and diagnostics.

Matrix of dispersion: In this algorithm, the matrix of dispersion

$$(D)_{ij} = \left(X^T X\right)^{-1} \tag{4.10}$$

plays an important role. Its elements are the coefficients that describe the transfer of the experimental error to the model as seen in the calculation of the variance of the model coefficients here below:

$$var(\vec{\alpha}) = var\left(\left(X^T X\right)^{-1} X^T Y\right) = \left(X^T X\right)^{-1} var(Y) \tag{4.11}$$

Its analysis, that can be performed before the essays, provides a lot of information on the quality of the design of experiments.

[1] This presentation of the LSF is really simplified at the maximum. For a more rigorous presentation see [4] pp. 390.

Matrix of correlation: If the elements of the matrix of dispersion are divided by the square root of the corresponding diagonal elements, we obtain a matrix given the correlation coefficients of the factors taken two by two:

$$C_{ij} = \frac{D_{ij}}{\sqrt{D_{ii}D_{jj}}} \tag{4.12}$$

in Matlab: There is a built-in function `corrcov(D)` that transforms the dispersion matrix into a correlation matrix.

Variance inflation factors: The variance inflation factors (VIF) are indicators of colinearities between factors and predicts the quality of the estimation of the model coefficients. They are the diagonal elements of the inverse of the correlation matrix.

$$VIF(\vec{\alpha}) = diag(C^{-1}) \tag{4.13}$$

By definition they are bigger or equal to 1, 1 being the best possible VIF. The square root of the VIF tells you how much larger the standard error is, compared with what it would be if that variable were uncorrelated with the other predictor variables in the model. When the VIF are bigger than 10, it is recommended to avoid performing the measurement, because the experimental error will have a too important influence in the measurement.

Response variance function: The response variance function is a second degree fonction defining the variance of the model in the experimental domain.

$$var_Y(x) = f'(x) \left(X^T X\right)^{-1} f(x)\ \sigma_Y^2 \tag{4.14}$$

Before the experiment, when σ_Y^2 is not known, the function $var_Y(x)/\sigma_Y^2$ shows where, in the experimental domain, the model will be of better or worst quality.

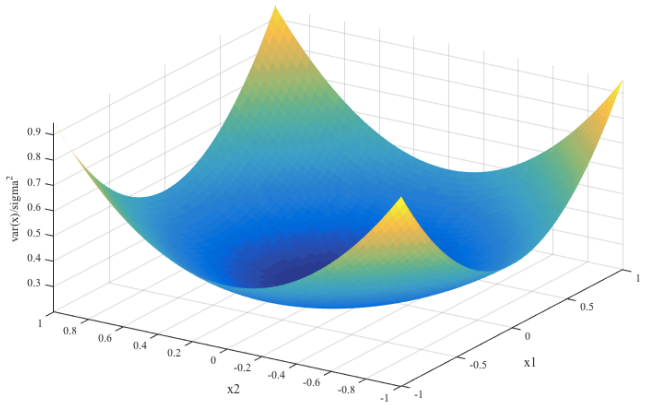


Figure 4.4: *Variance function of a linear model with interaction identified with a cubic-centered design.*

As example in figure 4.4 where the variance function of a design for two factors has been plotted, it is possible to observe that the model would be very accurate at the center of the domain and would lose accuracy when going away from it.

4.2 ANOVA and Orthogonality

In figure 3.1 it has been shown that the regression of a model corresponds to the projection of a response vector on a set of vectors. This becomes evident for a parametric model when the linear system is re-written in the vectorial form as in equation 4.15.

$$\vec{Y} = X\vec{\alpha} + \vec{\epsilon} = \sum_{j=1}^{N} \begin{bmatrix} x_{ij} \\ \vdots \\ x_{ij} \end{bmatrix} \alpha_j + \vec{\epsilon} \tag{4.15}$$

By definition the vector of residue $\vec{\epsilon}$ is always orthogonal to the vector of the model $\hat{\vec{Y}} = X\vec{\alpha}$. Applying the Pythagoras' theorem it is then possible to write

$$SS(Y_i) = SS(\hat{Y}_i) + SS(\epsilon_i) \tag{4.16}$$

Where the notation $SS(y)$ represents a sum of squares of y. Based on this breakdown into two orthogonal elements, an ANOVA table can be built as shown in table 4.1 Now we want to dig further and perform a breakdown of the model $\hat{Y}$ to be able to determine the level of confidence not only of the model as a whole, but also of its different coefficients.

Table 4.1: *ANOVA table of a parametric model.*

Source	SS	DF	MS	F	p
Model	$\hat{Y}^T\hat{Y}$	p	$\frac{1}{p}\hat{Y}^T\hat{Y}$	x	$F(x, p, N-p)$
Residue	$\epsilon^T\epsilon$	$N-p$	$\frac{1}{N-p}\epsilon^T\epsilon$		
Total	Y^TY	N			

We now separate the model into two parts:

$$\hat{Y} = \hat{Y}_1 + \hat{Y}_2 \quad \text{with} \quad \begin{cases} \hat{Y} = X\alpha \\ \hat{Y}_1 = X_1\alpha_1 \\ \hat{Y}_2 = X_2\alpha_2 \end{cases} \tag{4.17}$$

The sum of squares can be calculated:

$$\hat{Y}^2 = \left(\hat{Y}_1 + \hat{Y}_2\right)^2 = \hat{Y}_1^2 + \hat{Y}_1^T\hat{Y}_2 + \hat{Y}_2^T\hat{Y}_1 + \hat{Y}_2^2 \tag{4.18}$$

If the two parts are orthogonal their product is null and it is possible to write

$$\hat{Y}_1 \perp \hat{Y}_2 \quad \Rightarrow \; \hat{Y}_1'\hat{Y}_2 = \hat{Y}_2'\hat{Y}_1 = 0 \quad \Rightarrow \; \hat{Y}^2 = \hat{Y}_1^2 + \hat{Y}_2^2 \tag{4.19}$$

The breakdown of the sum of squares allows us to compute a new ANOVA as shown in table 4.2

Table 4.2: *ANOVA table of a parametric model separated into two orthogonal part with p_1 and p_2 the number of parameters of part 1, respectively part 2.*

Source	SS	DF	MS	F	p
Part 1	$\hat{Y}_1^T\hat{Y}_1$	p_1	$\frac{1}{p}\hat{Y}_1^T\hat{Y}_1$	x_1	$F(x_1, p_1, N-p)$
Part 2	$\hat{Y}_2^T\hat{Y}_2$	p_2	$\frac{1}{p}\hat{Y}_2^T\hat{Y}_2$	x_2	$F(x_2, p_2, N-p)$
Residue	$\epsilon^T\epsilon$	$N-p$	$\frac{1}{N-p}\epsilon^T\epsilon$		
Total	Y^TY	N			

This procedure can be generalized for more than two parts. What to do when the model has to be decomposed into non-orthogonal parts? Let's consider an experimental situation in which three measurements y_i are fitted on a linear model

$y = a_o + a_1 x + \epsilon$. The three measurements constitute the coordinates of the vector $\overrightarrow{OY}$ as shown in figure 4.5. The two regressors are the vector that support the grand mean $[1\ 1\ 1]^T$ and the vector $[x_1\ x_2\ x_3]^T$ defined by the values chosen for x in the three experiments and represented in the figure by I and X.The two regressors form a plane and the estimate of the model, $\hat{Y}$, is the projection of Y on this plane. The projection of $\hat{Y}$ on the two vectors of base determines the points A_1 and B_1 respectively and the distance OA_1 and OB_1 correspond to the coefficients of the model a_o and a_1.

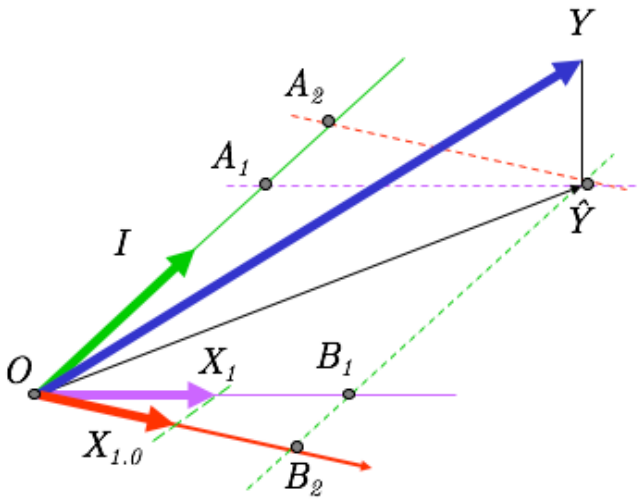


Figure 4.5: *3D representation of the least squares fit with two non-orthogonal predictors.*

The ANOVA requires an orthogonal breakdown. $\overrightarrow{O\hat{Y}}$ is orthogonal to $\overrightarrow{\hat{Y}Y}$, but $\overrightarrow{OA_1}$ is not orthogonal to $\overrightarrow{OB_1}$.The situation is presented with more clarity in 2D in figure 4.6. To have an orthogonal breakdown it is necessary to determine a new axis $\overrightarrow{OX_{1.0}}$ orthogonal to $\overrightarrow{OA_1}$ and in the plane OA_1B_1. With this change the projections of $\overrightarrow{O\hat{Y}}$ change also:

$$\overrightarrow{O\hat{Y}} = \overrightarrow{OA_1} + \overrightarrow{OB_1} = \overrightarrow{OA_2} + \overrightarrow{OB_2} \tag{4.20}$$

$$\overrightarrow{O\hat{Y}} = a_o\vec{I} + a_1\vec{X} = a_o^*\vec{I} + a_1^*\vec{X}_\perp \tag{4.21}$$

$$a_o^* = a_o + a_1 \cos\theta \tag{4.22}$$

$$a*_1 = a_1 \sin\theta \tag{4.23}$$

The new breakdown can be written:

$$\left(\overrightarrow{OY}\right)^2 = \left(\overrightarrow{OA_2} + \overrightarrow{OB_2}\right)^2 = \left(\overrightarrow{OA_2}\right)^2 + \left(\overrightarrow{OB_2}\right)^2 \tag{4.24}$$

and the sum of squares can be corrected. Table 4.3 presents the new ANOVA scheme. In practice, the corrected coefficient a_o^* for the constant is equivalent to the coefficient obtained when regressing the data with only one regressor (the constant in that case). The corrected sum of squares for the linear effect is computed indirectly by difference between the model sum of squares, $\hat{Y}^T\hat{Y}$, and the corrected sum of squares of the constant $a_o^{*2}I^TI$. This procedure can be continued step by step in the same way when there are more than two coefficients in the model.

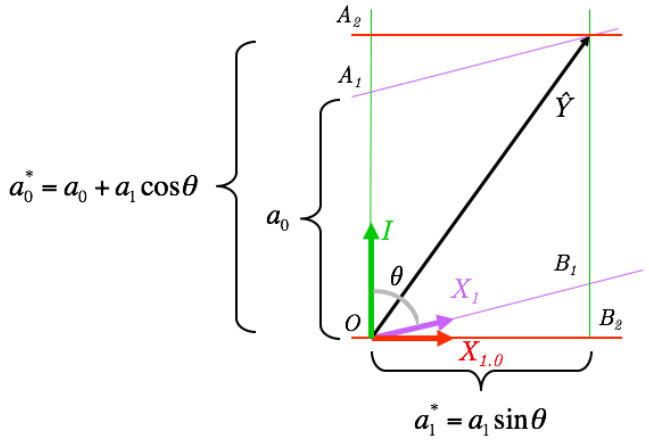


Figure 4.6: *3D representation of the least squares fit with two non-orthogonal predictors.*

Table 4.3: *ANOVA table of a linear model $y = a_o + a_1 x + \epsilon$ with corrected sum of squares to mitigate non-orthogonality of the regressors.*

Source	SS	SS*	DF	MS	F	p
Model	$\hat{Y}^T\hat{Y}$	-	2	$\frac{1}{2}\hat{Y}^T\hat{Y}$	x	$F(x,2,1)$
Residue	$\epsilon^T\epsilon$	-	1	$\epsilon^T\epsilon$		
Measurements	Y^TY	-	3			
Constant	$a_o^2 I^T I$	$a_o^{*2} I^T I$	1	...		
Linear effect	$a_1^2 X^T X$	$\hat{Y}^T\hat{Y} - a_o^{*2} I^T I$	1	...		
Residue	$\epsilon^T\epsilon$	$\epsilon^T\epsilon$	1	...		
Total	-	Y^TY	3			

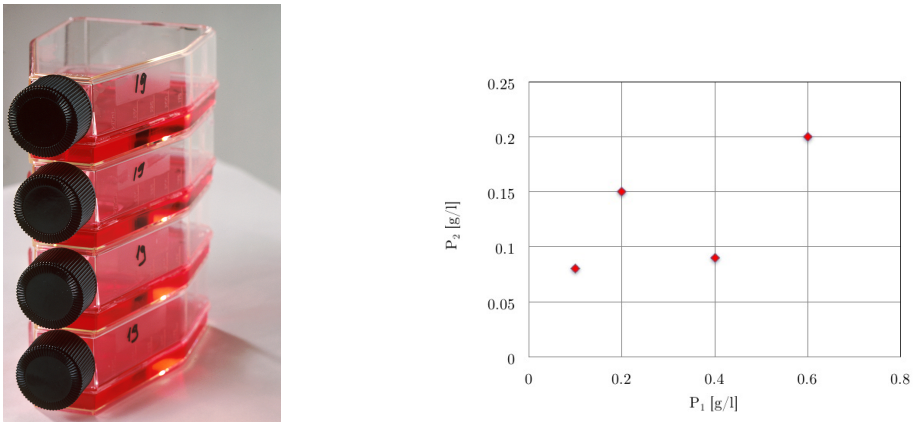
4.2.1 Example of a non-orthogonal situation

The preceding concepts are now illustrated with an experimental situation to test the interaction between two factors. We want to test if the conjugate use of two sugar substitutes P_1 and P_2 presents a combined negative effect for health. The experiments would consist in feeding different dishes of cultured cells with different sodas having different concentrations of the sugar substitute to be tested (figure 4.7a). After a period the dishes are analyzed to measure the level of an indicator related to the risk of diabetes. Table 4.4a presents the height experiments and the results. Figure 4.7b presents a scatterplot of the points of measurements in the experimental domain. Table 4.4b presents the matrix of the model:

$$y = a_o + a_1\ x_1\ +\ a_2\ x_2\ +\ a_{12}\ x_2 x_2\ + \epsilon \tag{4.25}$$

Applying the least squares fit algorithm given in equation 4.9, the coefficients of the model with interaction can be computed. This would give the values of the coefficients that appear at the bottom of table 4.4b. These coefficients are used to calculate the estimates $\hat{Y}$. The difference between measurement results and the estimates gives the residue ϵ, also presented in the same table. A representation of the model as a surface is given at figure 4.8. An ANOVA analysis of the model as a whole can be done based on the sum of squares (table 4.5). The very small probability ($p = 1.1 \times 10^{-8}$) of obtaining the Fisher ratio by chance concedes a good level of validity to the chosen model. But would it be possible to conclude

that the hypothesis on the combined effect of the sugar substitutes is validated? No, because, even if the model is coherent with the experimental data,there is still no evidence that the term of interaction is significant in itself. To answer that question, it is necessary to go one step further by computing the level of significance of each term. The difficulty to do so lies in the fact that the regressors (the vectors constituted by the columns of the model matrix) are not orthogonal, as can be observed in figure 4.9, and the sum of squares must thus be corrected.



(a) *(b)*

Figure 4.7: *(a) Dishes for the culture of cells. (b) Scatterplot of the standardized experimental points.*

Following the procedure presented earlier in this chapter and summarized in table 4.3, the corrected sum of squares of our example can be computed step by step, as illustrated in table 4.6, as follows:

1. Compute the total sum of squares of the measurements, $SS(Y) = \sum_i Y_i^2$
2. Compute the corrected sum of squares for the constant a_o, $SS(a_o) = \frac{1}{N}(\sum_i Y_i)^2$
3. Compute the first residue by subtraction, $SS(\epsilon_1) = SS(Y) - SS(a_o)$
4. Compute the second residue, $SS(\epsilon_2) = SS(Y) - SS(a_o, a_1)$
5. Compute the corrected sum of squares for the first linear coefficient a_o by subtraction, $SS(a_1|a_o) = SS(a_o, a_1) - SS(a_o)^2$
6. Proceed the same way to get the corrected sum of squares of a_2, $SS(a_2|a_o, a_1) = SS(a_o, a_1, a_2) - SS(a_o, a_1)$
7. And a_{12}, $SS(a_{12}|a_o, a_1, a_2) = SS(a_o, a_1, a_2) - SS(a_o, a_1, a_2)$

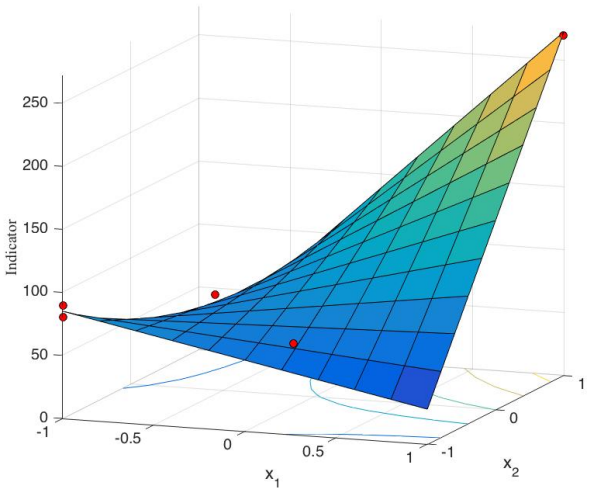


Figure 4.8: *Representation of the model as a surface in the standardized experimental domain with the experimental results (red dots).*

Table 4.4: *Sugar substitute experiments. (a) Experimental data constituted by the concentration of products P_1 and P_2 in $[g/l]$ and the corresponding level of the indicator of diabetes. (b) Matrix of the linear model with interaction and the computed coefficients.*

(a)

Run	P_1	P_2	Y
1	0.1	0.08	80.40
2	0.2	0.15	70.82
3	0.4	0.11	67.11
4	0.6	0.2	270.00
5	0.1	0.08	89.72
6	0.2	0.15	58.94
7	0.4	0.11	53.66
8	0.6	0.2	275.31
			$SS(Y)$
			179'082

(b)

I	X_1	X_2	X_1X_2	$\hat{Y}$	ϵ
1	-1.00	-1.00	1.00	85.6	-4.66
1	-0.60	0.17	-0.10	64.88	5.94
1	0.20	-0.83	-0.17	60.39	6.73
1	1.00	1.00	1.00	272.66	-2.65
1	-1.00	-1.00	1.00	85.06	4.66
1	-0.60	0.17	-0.10	64.88	-5.94
1	0.20	-0.83	-0.17	60.39	-6.72
1	1.00	1.00	1.00	272.66	2.66
a_o	a_1	a_2	a_{12}	$SS(\hat{Y})$	$SS(\epsilon)$
97.70	52.62	41.18	81.15	178'863	219

Table 4.5: *ANOVA table of a linear model with interaction for the experiments with the sugar substitute.*

Source	SS	DF	MS	F	p
Model	178'863	4	44'715	818.4	1.1×10^{-8}
Residue	219	4	54.6		
Total	179'082	8			

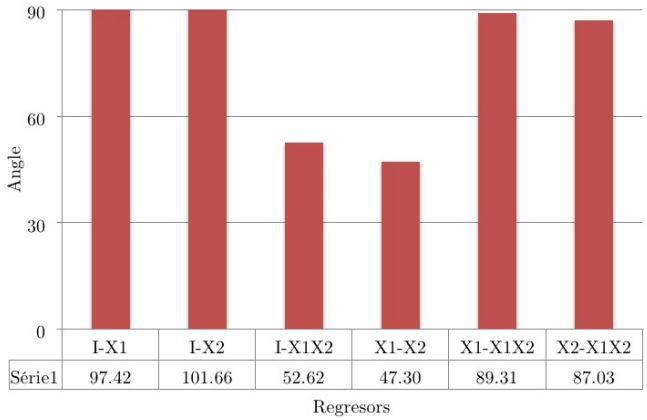


Figure 4.9: *Angles in degrees between the regressors.*

Now corrected sums of squares are available for each coefficient and the last residue and the ANOVA table can be completed as shown before. This shows that the coefficient of interaction has a p-value of 4.6×10^{-7}, which is fully acceptable. It would be then possible to conclude that the experiment has shown the effectiveness of a combined effect of the two sugar substitutes.

[2] $SS(A|B)$ is read sum of squares of A knowing B

Table 4.6: *ANOVA table of a linear model with interaction with corrected sum of squares to mitigate non-orthogonality of the regressors.*

Source	SS	SS*	DF	MS	F	p
$\hat{Y}$	178'863		4	44'715	818.4	1.1×10^{-8}
ϵ	219		4	54.6		
Total	①179'082		8			
a_o		②116'635				
ϵ_1		③62'447				
Total		179'082				
a_o		116'635				
a_1		⑤37'034				
ϵ_2		④25'413				
Total		179'082				
a_o		116'635				
a_1		37'034				
a_2		⑥9'139				
ϵ_3		16'275				
Total		179'082				
a_o		116'635	1			
a_1		37'034	1	37'034	678	3.8×10^{-8}
a_2		9'139	1	9'139	167	2.4×10^{-6}
a_{12}		⑦16'056	1	16'056	294	4.6×10^{-7}
ϵ_3		219	4	54.6		
Total		179'082	8			

4.2.2 Alternative sum of squares

The procedure presented in the previous sections for correcting the sum of squares is called the *sequential* sum of squares. The sums of squares of the different parts of the model are computed each time, based on the sums of squares of the previous parts. This method is also called *Type I*. In this procedure, the order of the coefficients has an influence on the final p-values. To avoid the latter, a *Type II* procedure has been developed and is proposed as an option by most standard softwares. The type II sum of squares procedure calculates the p-values in the most unfavorable situation for each coefficient, with which we mean that each coefficient would be placed ofter all the other ones.
Table 4.7 presents the ANOVA of the experiments of the sugar substitutes realized with type-II sums of squares. It is possible to observe that the p-values are bigger than in the type-I case, this procedure being more conservative.

Table 4.7: *ANOVA with type II SS of a linear model with interaction for the experiments with the sugar substitute.*

Source	SS	DF	MS	F	p
a_o	15'911	1			
a_1	7'130	1	7'130	130	0.034%
a_2	4'700	1	4'700	86	0.075%
a_{12}	16'122	1	16'122	295	0.007%
Residue	219	4	54.6		
Total	179'082	8			

in Matlab: The method `tbl = anova(mdl, anovatype, sstype)` can be applied to a linearModel object *mdl*. The *sstype* parameter allows us to select the type of sum of squares.

4.2.3 The use of the alias matrix

The procedure of orthogonalization can be formalized using the matrix of alias. Let's consider an experimental situation with an output Y obtained by n experiments.

Now let's assume that we are interested in regressing this data on two complementary models whose model matrices are X_1 and $[X_1 X_2]$ so that:

$$\hat{Y}_a = X_1\alpha \tag{4.26}$$

$$\hat{Y}_b = [X_1\ X_2]\begin{bmatrix}\alpha_1\\ \alpha_2\end{bmatrix} = X_1\ \alpha_1 + X_2\ \alpha_2 \tag{4.27}$$

α and $[\alpha_1\alpha_2]$ being the coefficients of the two models. The second model corresponding to the first model plus a complement, there are the same number of elements in α than in α_1. If the alias matrix is defined as:

$$A = \left(X_1^T\ X_1\right)^{-1}\left(X_1^T\ X_2\right) \tag{4.28}$$

it is possible to write the following transformations:

$$\alpha = \alpha_1 + A\alpha_2 \tag{4.29}$$

$$X_{2.1} = X_2 - X_1 A \tag{4.30}$$

and the decomposition in orthogonal components is:

$$\hat{Y} = X_1\left(\alpha_1 + A\ \alpha_2\right) + \left(X_2 - X_1\ A\right)\alpha_2 = X_1\ \alpha + X_{2.1}\ \alpha_2 \tag{4.31}$$

The corresponding ANOVA table, now based on a sequential sum of squares, can be written as presented in table 4.8

Table 4.8: *ANOVA (type I SS) using the alias matrix.*

Source	SS	SS^*
α_1	$\alpha_1^T X_1^T X_1 \alpha_1$	$(\alpha_1 + A\alpha_2\ )^T X_1^T X_1\,(\alpha_1 + A\alpha_2\ )$
α_2	$\alpha_2^T X_2^T X_2 \alpha_2$	$\alpha_2^T\ (X_2 - X_1 A)^T\,(X_2 - X_1 A)\,\alpha_2$
Residue	$\epsilon^T\epsilon$	$\epsilon^T\epsilon$
Total	$Y^T Y$	$Y^T Y$

For more than two parts, the procedure can be reformulated as follows:

1. Compute the coefficient estimates α for the full number of regressor N_R with the LSF algorithm (equation 4.9)

2. Compute the full model estimates $\hat{Y} = X\alpha$ and the corresponding residue, and collect the sums of squares $SS(\hat{Y})$ and $SS(\epsilon)$.

3. Compute the coefficient estimates $\alpha_{(N_R-1)}$ for $(N_R - 1)$ regressors as defined in equation 4.29 which gives:

$$\alpha_{(N_R-1)} = [\alpha_{j\leq(N_R-1)}] + A_{(N_R)}\ \alpha_{N_R} \tag{4.32}$$

with the matrix of alias between the model of N_R coefficients and the one with $(N_R - 1)$ coefficients:

$$A_{(N_R)} = \Big(\ [X_{i,j\leq(N_R-1)}]^T\ [X_{i,j\leq(N_R-1)}]\ \Big)^{-1} [X_{i,j\leq(N_R-1)}]^T\ X_{N_R} \tag{4.33}$$

4. Compute the model estimates $\hat{Y}_{(N_R-1)} = [X_{i,j\leq(N_R-1)}]\ \alpha_{(N_R-1)}$ and the corresponding sum of squares $SS(\hat{Y}_{(N_R-1)})$.

5. Compute the corrected sum of squares of α_{N_R}

$$SS(\alpha_{N_R}|[\alpha_{j\leq(N_R-1)}]) = SS(\hat{Y}_{(N_R-1)}) - SS(\hat{Y}) \tag{4.34}$$

6. For the type I sum of squares, go to 3 for $N_R - 2$ regressors, get

$$SS(\alpha_{N_R-1}|[\alpha_{j\leq(N_R-2)}]) = SS(\hat{Y}_{(N_R-2)}) - SS(\hat{Y}_{(N_R-1)})$$

and so on until reaching all the corrected sums of squares. For type II, go to 3 for $N_R - 1$ regressors and so on till every regressor has been excluded once.

7. Settle the ANOVA table with the collected sums of squares.

4.2.4 Lack of fit, goodness of fit and parsimony principle

The ANOVA analysis provides a way to determine if the data is coherent with a given model. The result of this analysis depends heavily on the design. The more factors a model includes, the more explanatory the ANOVA table will appear, but with a risk of over-fitting the data. Hence an equilibrium must be found as Vandekerckhove et al [10] note:

> Throughout history, prominent philosophers and scientists have stressed the importance of parsimony. For instance, in the Almagest (...) Ptolemy writes: *"We consider it a good principle to explain the phenomena by the simplest hypotheses that can be established, provided this does not contradict the data in an important way."* Ptolemy's principle of parsimony is widely known as Occams razor(...); the principle

> is intuitive as it puts a premium on elegance. In addition, most people feel naturally attracted to models and explanations that are easy to understand and communicate. Moreover, the principle also gives ground to reject propositions that are without empirical support (...).
>
> However, the principle of parsimony finds its main motivation in the benefits that it bestows those who use models for prediction. To see this, note that empirical data are assumed to be composed of a structural, replicable part and an idiosyncratic, non-replicable part. The former is known as the *signal*, and the latter is known as the *noise*. Models that capture all of the signal and none of the noise provide the best possible predictions to unseen data from the same source. Overly simplistic models, however, fail to capture part of the signal; these models underfit the data and provide poor predictions. Overly complex models, on the other hand, mistake some of the noise for actual signal; these models overfit the data and again provide poor predictions. Thus, parsimony is essential because it helps discriminate the signal from the noise, allowing better prediction and generalization to new data.
>
> Occam's razor is named after the English philosopher and Franciscan friar Father William of Occam (c.1288-c.1348), who wrote *"Numquam ponenda est pluralitas sine necessitate"* (plurality must never be posited without necessity), and *"Frustra fit per plura quod potest fieri per pauciora"* (it is futile to do with more what can be done with less). Occam's metaphorical razor symbolizes the principle of parsimony: by cutting away needless complexity, the razor leaves only theories, models, and hypotheses that are as simple as possible without being false. Throughout the centuries, many other scholars have espoused the principle of parsimony; the list predating Occam includes Aristotle, Ptolemy, and Thomas Aquinas (...), and the list following Occam includes Isaac Newton (*"We are to admit no more causes of natural things than such as are both true and sufficient to explain their appearances. Therefore, to the same natural effects we must, so far as possible, assign the same causes."*), Bertrand Russell, Albert Einstein (*"Everything should be made as simple as possible, but no simpler"*), and many others.

One tool to find the equilibrium between under- and over-fitting the the concept of *goodness of fit* or its opposite *the lack of fit*.
Let's consider an experimental situation with one output y that we want to model with one factor x so that $y = f(x)$. A set of experiments have been realized to determine this relation as presented in the table 4.9. We can observe in figure 4.10 that the ten data points are constituted by experiments for five different values of x that have been duplicated.

Table 4.9: *Set of 10 data points. The column $\bar{y}$ corresponds to the value of y averaged for each value of x. The column $\hat{y}$ corresponds to the estimated value of y by a linear model $y = a_o + a_1\ x$. The last row SS corresponds to the sum of squares.*

Runs	x	y	$\bar{y}$	$y - \bar{y}$	$\hat{y}$	$y - \hat{y}$	$\bar{y} - \hat{y}$
1	-1	11.60	11.43	0.17	9.09	2.51	2.35
2	-1	11.27	11.43	-0.17	9.09	2.18	2.35
3	-0.5	5.05	6.63	-1.58	7.51	-2.46	-0.88
4	-0.5	8.21	6.63	1.58	7.51	0.70	-0.88
5	0	3.19	2.86	0.33	5.93	-2.74	-3.07
6	0	2.53	2.86	-0.33	5.93	-3.40	-3.07
7	0.5	3.63	3.75	-0.12	4.35	-0.72	-0.60
8	0.5	3.86	3.75	0.12	4.35	-0.49	-0.60
9	1	5.56	4.98	0.57	2.77	2.78	2.21
10	1	4.41	4.98	-0.57	2.77	1.64	2.21
SS		449.27	443.33	5.94	401.42	47.85	41.91

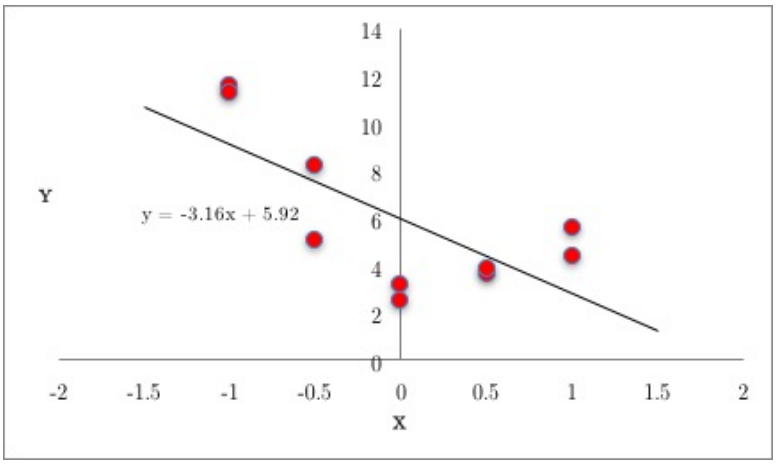


Figure 4.10: *Scatterplot of the ten data points.*

The question is now to determine if a linear model is acceptable for modeling such data. Table 4.10 presents an ANOVA analysis with some additional rows. The row *model* shows a low p-value that could be acceptable following the usual criteria (p-value $< 5\%$). The duplication of the experiments has allowed the determination of a pure error sum of squares. It can now be distinguished from the modeling error sum of squares called *lack of fit* obtained by subtracting the pure error to the sum of squares of the residue. When comparing these two sums of squares, a p-value of 1.1% is obtained. This allows us the reasonably confident assumption that there is a lack of fit and that it would be better to look for a better model.
To perform a lack-of-fit analysis it is necessary to have repetition of a few experimental points (to compute the pure error sum of squares) and to have more experimental points than the number of parameters of the model.

Table 4.10: *ANOVA table with lack of fit analysis.*

Source	SS	df	MS	F	p
Model	401.42	2	200.71	33.56	0.013%
Residue	47.85	8	5.98		
Lack of fit	41.91	3	13.97	11.76	1.1%
Pure error	5.94	5	1.19		

Now the theoretical base is established, we will present a series of optimized designs.

4.3 Plackett-Burman design

The properties of the Plackett-Burman (PB) group of designs are very similar to the factorial design [8]. These designs allow us one to estimate efficiently the main effects of the factors of a system supposed to be without any interaction (additive model). These matrices allow an efficient screening of a large number of factors with a minimum of runs.
A PB design requires N runs for estimating the main effects of a maximum of $N-1$ factors. Matlab has a *hadamard* routine that computes PB matrix for the cases where n, $n/12$, or $n/20$ is a power of 2.
The Plackett-Burman matrices are composed, like factorial matrices, of 1 and -1. The construction of such a design starts from a generator. The generator is a list of $+$ and $-$ signs that constitute the first line of the matrix. A case of seven factors is presented in 4.11.

Table 4.11: *Construction of a PB_8 from a generator.*

instruction	codification	matrix						
generator	+ + + − + − −	1	1	1	-1	1	-1	-1
N-2 next lines	− + + + − + −	-1	1	1	1	-1	1	-1
obtained by	− − + + + − +	-1	-1	1	1	1	-1	1
circular	+ − − + + + −	1	-1	-1	1	1	1	-1
permutations	− + − − + + +	-1	1	-1	-1	1	1	1
	+ − + − − + +	1	-1	1	-1	-1	1	1
	+ + − + − − +	1	1	-1	1	-1	-1	1
A line of −	− − − − − − −	-1	-1	-1	-1	-1	-1	-1

Plackett and Burman present a list of generators for cases inferior to 100 factors. This list is reproduced partially in table 4.12

Table 4.12: *A few generators of Plackett-Burman design.*

$N = 8$	+ + + − + − −
$N = 12$	+ + − + + + − − − + −
$N = 16$	+ + + + − + − + + − − + − − −
$N = 20$	+ + − − + + + + − + − + − − − − + + −
$N = 24$	+ + + + + − + − + + − − + + − − + − + − − − −
$N = 40$	Double of design $N = 20$

in Matlab: The build-in function `X = hadamard(n)` provides a model matrix X of a Plackett-Burman design of n runs. This function handles only the cases where n, $n/12$, or $n/20$ is a power of 2.

4.4 Factorial design

Unfortunately, the most frequent *design* used in laboratories consists in varying one factor at a time (OFAT): this has the disadvantage of neglecting the interactions between factors. In this regard, the factorial design is a major improvement to the experimental strategy. The advantage of the factorial design in comparison to the OFAT design resides precisely in the fact that all the factors are varied simultaneously, but in a structured way.
Factorial design is an ideal candidate for the identification of the polynomial model of first degree with interactions, as presented in equation 4.35. A distinction is usually made between designs of experiments after the number of levels of all the factors, usually two or three. There are however composite designs that mix factors with two levels and factors with three levels [11, 12, 13];
Let us consider a linear model with interactions (as, for example, the Taylor's series of a complicated function):

$$Y(x) = a_o + \sum_{i=1}^{N} a_i x_i + \sum_{i<j}^{N} a_{ij} x_i x_i + \sum_{i<j<k}^{N} a_{ijk} x_i x_i x_k + \cdots + a_{1\cdots N} x_1 \cdots x_N \quad (4.35)$$

This polynomial counts 2^N coefficients $a_o, a_1 \cdots a_N$ and each factor appears only at the first degree. The optimal design for determining the 2^N coefficients a_i of this model is a factorial design. The coefficients $a_o, a_1, \cdots, a_{12}, \cdots, a_{1\cdots N}$ are called the effects of the factors x_i. A distinction is usually made between:

- the constant effect a_o
- the main effects $a_1 \cdots a_N$
- the first order (two-by-two) interaction effects
- the higher order interaction effects

The a_i are the *half-effects* because they correspond to the variation between the center of the domain and the border. The a_i are commonly, if imprecisely, referred to as the effects. In order to explore an experimental space of N factors, the inferior and superior limits of each are considered. This is equivalent to considering a system of N factors at two levels that then counts 2^N possible states. There are several possibilities to represent such a system. For the sake of computation, the state of a factor is represented by an index indicating if the factor is in one or the other state. Usually the indices used are '-1' and +1', which has the advantage of creating a transitive group with the operation multiplication of column[3]. The state

[3]Tagushi designs are coded with 0, 1, 2, $\cdots$.

of the system is then fully determined by a state vector that contains the indices of each factor and, for each factor, the real physical values corresponding to the indices. This is shown in Table 4.13, where the line *experiment* is equivalent to the line *code*.

Table 4.13: *Example of coding for a factorial design.*

Factors	U[W]	T[° C]	P[pa]	D[m]
Minimum	10	250	100000	−0.02
Maximum	20	350	150000	0.02
Experiment	20	250	150000	0.02
Coded	1	−1	1	1

The nomenclature of factorial designs is s^N where s is the number of states (levels) for each factor and N the number of factors. The value s^N corresponds to the number of experiments of the design. A factorial matrix E can be systematically constructed as follows:

- The first column is filled with -1 in the first half and with 1 for the second half,
- The second column is filled in its first and third quarters with -1 and with 1 in its second and fourth quarters,
- We proceed in such a way for the other columns, alternating -1 and 1 following the successive fractions corresponding to $(\frac{1}{2})^n$ till $(\frac{1}{2})^N$.

The matrix of experiments E of a design 2^3 is then the following:

$$E = \begin{bmatrix} -1 & -1 & -1 \\ -1 & -1 & 1 \\ -1 & 1 & 0 \\ -1 & 1 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \\ 1 & 1 & 1 \end{bmatrix} \tag{4.36}$$

We can see that each column has the same number of 1 and of -1. The matrix of the model X is constructed from the matrix of experiments E. The matrix of the model X has the same number of columns as the number of coefficients of the model. The column corresponding to the interaction $x_i x_j$, which is used to calculate the coefficient a_{ij}, is the product of the columns i and j of the matrix of experiments. The product of two columns is a column with the same number of elements as the multiplied columns and whose elements are the products, two by two, of the original elements as shown in equation 4.37.

$$\begin{bmatrix} a_1 \\ \vdots \\ a_N \end{bmatrix} \begin{bmatrix} b_1 \\ \vdots \\ b_N \end{bmatrix} = \begin{bmatrix} a_1 b_1 \\ \vdots \\ a_N b_N \end{bmatrix} \tag{4.37}$$

The next step is to set up the matrix of the model for a model of three factors and a factorial design 2^3. The model has one constant effect, three main effects and four interaction terms:

$$Y(x) = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 + a_{12} x_1 x_2 + a_{13} x_1 x_3 + a_{23} x_2 x_3 + a_{123} x_1 x_2 x_3 \tag{4.38}$$

The first column of the matrix of the model, which corresponds to the coefficient a_0, is a column of 1. The next three columns, which correspond to the main effects a_1, a_2, a_3, are the three columns of the matrix of experiments (equation 4.36). The four last columns are for the interaction effects $a_{12}, a_{13}, a_{23}, a_{123}$ and are produced by multiplying the columns of the matrix of experiments as they correspond to the interacting factors.

$$X = \begin{bmatrix} 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 \\ 1 & 1 & 1 & -1 & 1 & -1 & -1 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \tag{4.39}$$

If the response of the system is organized in a vector $\vec{R}$ and the coefficients of the model in a vector $\vec{\alpha}$, then the system of equations is written and solved as follows:

$$\vec{R} = X\vec{\alpha} \tag{4.40}$$

$$\vec{\alpha} = (X^T X)^{-1} X^T \vec{R} \tag{4.41}$$

The model matrix (as well as the matrix of experiments) of a factorial design is

a *Hadamard matrix* and has the property to be proportional to the inverse of its transpose. This significantly simplifies equation 4.41, avoiding a matrix inversion:

$$\{(X^T X) = N I_N\} \Rightarrow \{\vec{\alpha} = \frac{1}{N} X^T \vec{R}\} \tag{4.42}$$

For the factorial designs, the number of coefficients (and of experiments) grows exponentially with the number of factors. For that reason, these plans rapidly become too costly and then unusable. If the coefficients of highest orders of interaction can be neglected, it is then possible to realize only a fraction of the factorial design, as we will see in the next chapter. But let us first practice the concept presented in this chapter in an example.

in Matlab: Several functions can be used to generate a full factorial design 2^N:

- `E = fact_mat(n)` is a function of LISA, the library of the course and generates an essay matrix E coded with -1 and 1.
- `E = ff2n(n)` is a built-in function that generates an essay matrix coded with 0 and 1. A standard full factorial matrix can be obtained with the linear transformation $E = (E - 0.5) * 2$;
- `E = fullfact(levels)` gives factor settings E for a full factorial design with n factors, where the number of levels for each factor is given by the vector levels of length n.

4.4.1 Example

A student wants to optimize her bicycle. She is aware of three factors about her preparation and her material:

1. The height of the saddle relative to the pedal can vary between 75 cm and 80 cm.
2. Her diet can be composed of dry meat and bananas (animal proteins and magnesium) or a mix of cereals (vegetable proteins and starch).
3. The derailleur gears: she has two models available, A and B.

Our student wants to determine if there is a main factor and the optimal set for improving her performance. In this perspective she plans a design of experiments (Table 4.14). A run consists in measuring the time necessary to run the usual 50 km she cycles daily as training. She plans to fit the data on the following model:

$$\begin{aligned} time = \quad & meantime \\ + \quad & \text{saddle effect} + \text{diet effect} + \text{gear effect} \\ + \quad & \text{interaction saddle}\times\text{diet} + \text{interaction saddle}\times\text{gear} + \text{interaction diet}\times\text{gear} \\ + \quad & \text{interaction saddle}\times\text{diet}\times\text{gear} \end{aligned}$$

Table 4.14: *Factorial runs for determining the effects of the three factors supposed to influence the student's performance.*

Run	Saddle	Diet	Gear	X_1	X_2	X_3
1	80 cm	dry meat & bananas	A	1	1	1
2	80 cm	dry meat & bananas	B	1	1	−1
3	80 cm	cereals	A	1	1	1
4	80 cm	cereals	B	1	1	−1
5	75 cm	dry meat & bananas	A	1	1	1
6	75 cm	dry meat & bananas	B	1	1	−1
7	75 cm	cereals	A	1	1	1
8	75 cm	cereals	B	1	1	−1

The matrix of the model is given in equation 4.39. The student makes a random permutation to determine the order of the runs. She conducts the experiments at a frequency of one per day, taking care to keep all external conditions as constant as possible (climate, time schedule, daily activities, etc.). She gets the following results (sorted in the original order and not the order of execution):

$$\vec{Y} = \begin{bmatrix} 92.1 & 138.3 & 117.9 & 155.7 & 104.7 & 96.9 & 129.3 & 125.1 \end{bmatrix}^T \tag{4.43}$$

Applying equation 4.42, she can make the following inference :

$$\vec{\alpha} = \begin{bmatrix} 120 & 6 & -12 & -9 & 1.2 & -12 & -0.6 & -1.5 \end{bmatrix}^T \tag{4.44}$$

In figure 4.11 the height results have been placed in the experimental space. This type of representation is especially indicated for presenting results to people with a

minimum of knowledge about the design of experiments. We can observe that the specific effects (the differences from one edge to another) for the saddle and the gear are not the same, indicating the presence of a significant interaction. In table 4.15 the effects are presented in a relative value a_i/a_0 which shows the importance of each coefficient in relation to the constant effect a_0. The relative half-effects give a rapid insight into which factors and interaction are important and which are less. So remember that the levels have been coded -1 and 1. So all the coefficients are without dimension and can be compared. The same data is presented graphically in figure 4.12.

Table 4.15: *Half-effects and relative half-effects for the bicycle experiments.*

Factor	-	saddle	diet	gear	saddle ×diet	saddle ×gear	diet ×gear	saddle ×diet ×gear
Coef.	a_o	a_1	a_2	a_3	a_{12}	a_{13}	a_{23}	a_{123}
half-effect	120	6	−12	−9	1.2	−12	−0.6	−1.5
Rel. effect	-	5%	−10%	−7.5%	1%	−10%	−.5%	−1.25%

4.5 Fractional factorial design of 2 levels

Fractional factorial designs are an important category of factorial designs because they allow us to avoid the exponential expansion of the number of experiments with the number of factors. This chapter presents the basic concepts of fractional factorial designs. For a deeper understanding consult the book *Statistics for experimenters* by Box [3]. It provides a very comprehensive and didactical explanation, fully illustrated by practical examples. In the book by Montgomery [4], chapter 8 covers to the same topics with a broader view.
The number of experiments N_e of a fractional factorial design is a power of 2 and the number of experiments is still used as a base for the nomenclature for those designs:

$$N_e = 2^{N-r} \tag{4.45}$$

N being the number of factors and r being an integer smaller than N.
To understand how this works, we can start from a property of the full factorial design: each column of a full factorial design is *independent* of the other columns.

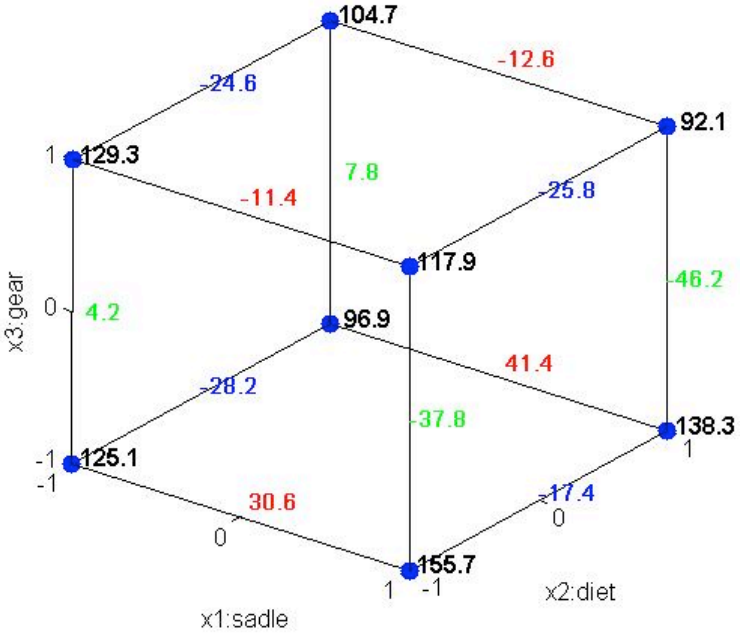


Figure 4.11: *Experimental data and specific effects. The change for the different specific effects (4.2, 7.8, -46.2, -37.8) indicates strong interaction.*

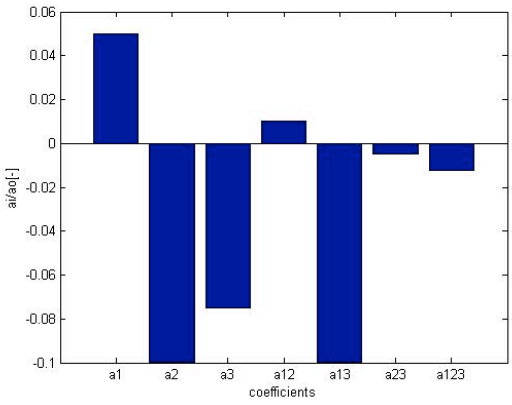


Figure 4.12: *Bar chart graphic of the half-effects for the bicycle experiments.*

This means that none of the columns of the experiment matrix can be obtained as the product of a set of the other columns. This independence is related to the lack of co-variance between the determined coefficients of the model that characterizes an orthogonal design. Fractional factorial designs are based on the fact that the coefficients of the highest levels of interaction can be neglected and that their corresponding columns in the model matrix can be used for introducing additional factors.

4.5.1 Construction of a fractional factorial design

The construction of a fractional factorial design can be done as follows:

1. Set-up of a full factorial matrix $2^{(N-r)}$ for $N - r$ factors
2. Build r generators based on the r highest interactions that will be associated with r additional factors. These generators are usually written as $j = k...n$ where each letter corresponds to a factor. Later on, we will see how to work with the generators and how to determine their properties.
3. Determine the list of the contrasts and alias groups that define the list of linear combinations of the coefficients that can be estimated by the design.

The consequence of associating additional factors with particular interactions is that the system of equations is now under-determined. It is then no longer possible to estimate all the coefficients of the model independently, but only some linear combinations of them.

in Matlab: The built-in function [`E`, `conf`] = `fracfact`(`gen`) creates the two-level fractional factorial design defined by the generator string *gen*. the generator is a string indicating the construction of the columns (example: 'a b c abc' for the generator $I = abc$). *conf* is a cell array of strings containing the confounding pattern for the design.

4.5.2 The alias concept

The group $(A, *)$ made with the set A of the columns of a model matrix of a full factorial design and the operation *column multiplication* represented by the symbol $*$, constitutes a commutative group. An example will help us to understand

the mechanism of the generators and alias. Consider a system of five factors $\{X1, X2, X3, X4, X5\}$ and a generator of size 5:

$$5 = 1\ 2\ 3\ 4 \tag{4.46}$$

The first four columns of the matrix of experiments correspond to the columns of a full factorial matrix of sixteen experiments. The fifth column is built by multiplying the first four columns. The product of a column by itself gives a column of 1 represented by the identity symbol I, associated with the constant coefficient α_o, which is the identity element of the group $(A, *)$. A group structure also requires that each element has an inverse. In this type of group each element (each column) is its own inverse:

$$5\ 5 = 5^2 = I \tag{4.47}$$

then

$$1\ 2\ 3\ 4\ 5\ = I \tag{4.48}$$

This relation indicates that the coefficient α_{12345} is an alias of α_o. By multiplying the two sides of equation 4.46 by the column 1, the following relation is obtained:

$$1\ 5\ = 2\ 3\ 4 \tag{4.49}$$

This means that α_{15} is aliased with α_{234}. Continuing the same way, equation 4.46 lets us define the sixteen alias set which also determines the sixteen contrasts that can be obtained with this fractional design - they are listed in table 4.16.

Table 4.16: *Set of contrasts of a factorial design* 2^{5-1} *defined by the generator* $5 = 1234$.

$\{\ \alpha_o\ ,\ \alpha_{12345}\ \}$				
$\{\ \alpha_1\ ,\ \alpha_{2345}\ \}$	$\{\ \alpha_2\ ,\ \alpha_{1345}\ \}$	$\{\ \alpha_3\ ,\ \alpha_{1245}\ \}$	$\{\ \alpha_4\ ,\ \alpha_{1235}\ \}$	$\{\ \alpha_5,\ \alpha_{1234}\ \}$
$\{\ \alpha_{12}\ ,\ \alpha_{345}\ \}$	$\{\ \alpha_{13}\ ,\ \alpha_{245}\ \}$	$\{\ \alpha_{14}\ ,\ \alpha_{235}\ \}$	$\{\ \alpha_{15}\ ,\ \alpha_{234}\ \}$	$\{\ \alpha_{23},\ \alpha_{145}\ \}$
$\{\ \alpha_{24}\ ,\ \alpha_{135}\ \}$	$\{\ \alpha_{25}\ ,\ \alpha_{134}\ \}$	$\{\ \alpha_{34}\ ,\ \alpha_{125}\ \}$	$\{\ \alpha_{35}\ ,\ \alpha_{124}\ \}$	$\{\ \alpha_{45},\ \alpha_{123}\ \}$

Another generator would give another distribution of the coefficients within the sets, but the same number of sets. In a first step, five factors have been considered that involve 32 coefficients (of the linear model with interactions). Now, with a generator, sixteen sets of aliases have been identified corresponding to sixteen independent contrasts that can be represented by the symbol l_i to differentiate them

Table 4.17: *Relation between the contrasts and the coefficients of the linear model with interaction in a factorial design 2^{5-1} defined by the generator $5 = 1234$.*

$l_0 = \alpha_0 + \alpha_{12345}$	$l_8 = \alpha_{14} + \alpha_{235}$
$l_1 = \alpha_1 + \alpha_{2345}$	$l_9 = \alpha_{15} + \alpha_{234}$
$l_2 = \alpha_2 + \alpha_{1345}$	$l_{10} = \alpha_{23} + \alpha_{145}$
$l_3 = \alpha_3 + \alpha_{1245}$	$l_{11} = \alpha_{24} + \alpha_{135}$
$l_4 = \alpha_4 + \alpha_{1235}$	$l_{12} = \alpha_{25} + \alpha_{134}$
$l_5 = \alpha_5 + \alpha_{1234}$	$l_{13} = \alpha_{34} + \alpha_{125}$
$l_6 = \alpha_{12} + \alpha_{345}$	$l_{14} = \alpha_{35} + \alpha_{124}$
$l_7 = \alpha_{13} + \alpha_{245}$	$l_{15} = \alpha_{45} + \alpha_{123}$

from the non-aliased coefficients α_i. If the design is executed without introducing any additional generator, it is possible to estimate the sixteen coefficients l_i that will have the relationship with the model coefficients α_i, as presented in Table 4.17.

Observe that if the 3×3 and 4×4 interactions (α_{ijk} and α_{ijkl}) are negligible, then the design allows us to estimate the main effects α_i and the 2×2 interactions α_{ij} without bias. Moreover, and if necessary, complementary experiments can be made (the second half of the full factorial design) so that the alias would be removed and all the coefficients be estimated.

An additional generator would let us further divide the number of experiments by two, but would also complicate the alias structure as a consequence.

The number of independent contrasts that can be estimated corresponds to the rank of the model matrix. Usually, interactions of highest levels are neglected and are not even mentioned in the alias table. In our example this would mean that only the coefficients α_i and α_{ij} would be mentioned in the alias table.

When writing down the alias table we must be cautious of the fact that several generators induce aliases that do not appear in the individual analysis of each generator. As an example, the two generators $I = 1235$ and $I = 1246$ of the 2^{6-2} design induce the alias $34 = 56$ that does not appear straight ($3 = 125$ and $4 = 126$, then $34 = 125126 = 56$). To guarantee that no alias has been missed, we need to check that all the main and interaction coefficients are present once, and only once, in the alias set. A quick way to do that is by writing down all the interactions as in the table 4.18 and tracing them successively.

Table 4.18: *Triangle of interactions* 2×2.

12	13	14	15
	23	24	25
		34	35
			45

4.5.3 The resolution of a fractional design

The resolution R is an important concept for selecting a fractional design. It describes the type of alias induced by the reduction of the full factorial design. The resolution indicates which levels of interaction are aliased. The most common resolution levels are:

III: A design of resolution $R = III$ confounds no main effects α_i between them, but confounds main effects α_i with interaction coefficients of first level α_{ij}. (and second level coefficients α_{ijk} with the constant effect α_i).

IV: A design of resolution $R = IV$ confounds neither main effects α_i between them, nor with first level interaction coefficients α_{ij} ; but it confounds first level interaction coefficients α_{ij} between them (and main effects α_i with second level coefficients α_{ijk}).

V: A design of resolution $R = V$ confounds neither main effects α_i between them, nor with first level interaction coefficients α_{ij} ; nor first level interaction coefficients α_{ij} between them; but it confounds first level interaction coefficients α_{ij} with second level coefficients α_{ijk}.

The resolution of a design corresponds to the size of the smallest generator and can be indicated by a Roman number as an index of the design: e.g. 2^{3-1}_{III} is a 2-level factorial design for three factors, with one generator and its resolution is $R = III$.

4.5.4 Fractional factorial table

Fractional factorial designs up to 11 factors are listed in figure 4.13 . This is a copy of a page of *Statistics for experimenters* by Box [3], which is famous among DOE users. I have heard a statistician saying that “ with this page in your pocket you can find a job in fifteen minutes " , and I tend to agree!

	3	4	5	6	7	8	9	10	11
4	2^{3-1}_{III} ±3=12								
8	2^3	2^{4-1}_{IV} ±4=123	2^{5-2}_{III} ±4=12 ±5=13	2^{6-3}_{III} ±4=12 ±5=13 ±6=23	2^{7-4}_{III} ±4=12 ±5=13 ±6=23 ±7=123				
16	2^3 2 times	2^4	2^{5-1}_{V} ±5=1234	2^{6-2}_{IV} ±5=123 ±6=234	2^{7-3}_{IV} ±5=123 ±6=234 ±7=134	2^{8-4}_{IV} ±5=234 ±6=134 ±7=123 ±8=124	2^{9-5}_{III} ±5=123 ±6=234 ±7=134 ±8=124 ±9=1234	2^{10-6}_{III} ±5=123 ±6=234 ±7=134 ±8=124 ±9=1234 ±10=12	2^{11-7}_{III} ±5=123 ±6=234 ±7=134 ±8=124 ±9=1234 ±10=12 ±11=13
32	2^3 4 times	2^4 2 times	2^5	2^{6-1}_{VI} ±6=12345	2^{7-2}_{IV} ±6=1234 ±7=1245	2^{8-3}_{IV} ±6=123 ±7=124 ±8=2345	2^{9-4}_{IV} ±6=2345 ±7=1345 ±8=1245 ±9=1235	2^{10-5}_{IV} ±6=1234 ±7=1235 ±8=1245 ±9=1345 ±10=2345	2^{11-6}_{IV} ±6=123 ±7=234 ±8=345 ±9=134 ±10=145 ±11=245
64	2^3 8 times	2^4 4 times	2^5 2 times	2^6	2^{7-1}_{VII} ±7=123456	2^{8-2}_{V} ±7=1234 ±8=1256	2^{9-3}_{IV} ±7=1234 ±8=1356 ±9=3456	2^{10-4}_{IV} ±7=2346 ±8=1346 ±9=1245 ±10=1235	2^{11-5}_{IV} ±7=345 ±8=1234 ±9=126 ±10=2456 ±11=1456
128	2^3 16 times	2^4 8 times	2^5 4 times	2^6 2 times	2^7	2^{8-1}_{VIII} ±8=1234567	2^{9-2}_{VI} ±8=13467 ±9=23567	2^{10-3}_{V} ±8=1237 ±9=2345 ±10=1346	2^{11-4}_{V} ±8=1237 ±9=2345 ±10=1346 ±11=1234567

Figure 4.13: *Two-level fractional factorial design for k variables and N runs.*

4.6 Composite design

The central composite design (CCD) is obtained by the aggregation of a *star design* to a factorial design as shown in figure 4.14. A matrix of experiments of a composite design has three types of rows as shown in table 4.19.

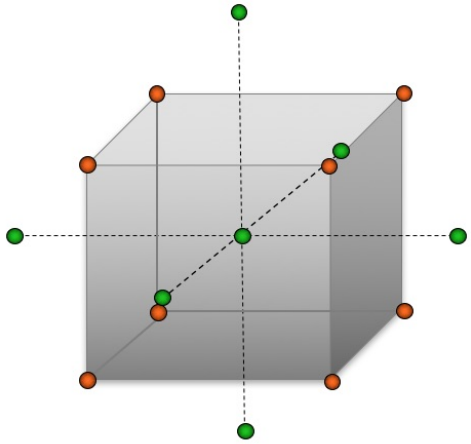


Figure 4.14: *Extension of a 2^3 factorial design (red dots) to a composite design of 15 runs with 6 star points and one or more central points.*

For response surface methodology, two important properties of the experimental design have to be considered: the *isovariance per rotation* and the *orthogonality*. The isovariance per rotation, also called *rotatability*, means that the variance of the estimated model at a given point will depend only on the distance of this point to the center of the domain. In other words, a design is rotatable when the variances of the model does not depend on the design orientation. To ensure this property, the axial distance α of the star design must follow the following equation

$$\alpha = (N_{fact})^{\frac{1}{4}} \tag{4.50}$$

where N_{fact} is the number of factorial experiments (experiments placed at the corner of the parallelepipedic domain).
Orthogonality has been presented extensively in section 4.2. In second degree fonction regression, there is always a covariance between the constant a_o and the second degree coefficients a_{ii}. But it is possible to get orthogonality between the other coefficients. In such a perspective, a central composite design is orthogonal if the following relation is respected:

$$\alpha = \left[\frac{1}{4} N_{fact} \left(\sqrt{N_{fact} + N_\alpha + N_o} - \sqrt{N_{fact}}\right)^2\right]^{1/4} \tag{4.51}$$

Table 4.19: *Structure of a central composite design.*

	Number of runs	x_1	x_2	...	...	x_N
Factorial Matrix	N_{fact}	-1	-1	...	...	-1
(full or fractional)		-1	-1	...	...	1
		⋮	⋮			⋮
		1	1	...	...	1
Center points	N_o	0	0	...	...	0
		⋮	⋮			⋮
Star runs	N_α	$-\alpha$	0	...	...	0
		α	0	...	...	0
		0	$-\alpha$	0	...	0
		0	α	0	...	0
		⋮				⋮
		0	...	...	0	α
		0	...	...	0	$-\alpha$

with N_α the number of points of the star design and N_o the number of points placed at the center of the domain.

To make a design approximately rotatable and orthogonal at the same time, we would first fix the axial distance for rotatability, and then add center points, so that [14]:

$$N_o > 4\sqrt{N_{fact}} + 4 - N_\alpha \tag{4.52}$$

In classical texts, the orthogonal blocking is also covered for this design(See Box [6] section 14.3).

To illustrate this type of design we can take an example proposed by Montgomery [4]. It consists in the modeling of a chemical process. The variables are the time and the temperature. Within the neighborhood of a probable maximum, a CCD is performed. The data is reproduced in table 4.21. The model matrix for a quadratic model is given at equation 4.53 and the dispersion matrix at equation 4.54. The design is almost orthogonal, the only extra-diagonal terms being between the constant a_o and the quadratic terms a_{ii}. Figure 4.15 shows the variance function in which we can observe the quite constant and low variance in the center of the experimental space.

Table 4.20: *Relation between the axial distance α and the number of points at the center N_o.*

Number of factors	2	3	4	5	5	6	6
Factorial design	2^2	2^3	2^4	2^{5-1}	2^5	2^{6-1}	2^6
Number of factorial runs	4	8	16	16	32	32	64
Number of star runs	4	6	8	10	10	12	12
Value of α if $N_o = 1$	1.0	1.22	1.41	1.55	1.60	1.72	1.76
if $N_o = 2$	1.08	1.29	1.48	1.61	1.66	1.78	1.82
if $N_o = 3$	1.15	1.35	1.55	1.66	1.72	1.84	1.89
if $N_o = 4$	1.21	1.41	1.61	1.72	1.78	1.90	1.94
$\alpha = (N_{fact})^{\frac{1}{4}}$	1.41	1.68	2	2	2.38	2.38	2.82
if $\alpha =$	1.41	1.68	2	2	2	2	2.37
then for rotatability, $N_o =$	5	6	7	7	6	6	9
then for orthogonality, $N_o =$	8	10	12	12	10	10	15

$$X = \begin{pmatrix} 1 & -1 & -1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1.41 & 0 & 0 & 2 & 0 \\ 1 & -1.41 & 0 & 0 & 2 & 0 \\ 1 & 0 & 1.41 & 0 & 0 & 2 \\ 1 & 0 & -1.41 & 0 & 0 & 2 \end{pmatrix} \tag{4.53}$$

$$D = \begin{pmatrix} 0.2 & 0 & 0 & 0 & -0.1 & -0.1 \\ 0 & 0.125 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.125 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.25 & 0 & 0 \\ -0.1 & 0 & 0 & 0 & 0.14 & 0.019 \\ -0.1 & 0 & 0 & 0 & 0.019 & 0.14 \end{pmatrix} \tag{4.54}$$

Table 4.21: *Data from a chemical experiment.*

run	time [min]	temperature $[^oC]$	yield [%]	viscosity $[Pa \cdot s]$	molecular weight	x_1	x_2
1	80	170	76.5	62	2940	-1	-1
2	80	180	77	60	3470	-1	1
3	90	170	78	66	3680	1	-1
4	90	180	79.5	59	3890	1	1
5	85	175	79.9	72	3480	0	0
6	85	175	80.3	69	3200	0	0
7	85	175	80	68	3410	0	0
8	85	175	79.7	70	3290	0	0
9	85	175	79.8	71	3500	0	0
10	92.07	175	78.4	68	3360	1.414	0
11	77.93	175	75.6	71	3020	-1.414	0
12	85	182.07	78.5	58	3630	0	1.414
13	85	167.93	77	57	3150	0	-1.414

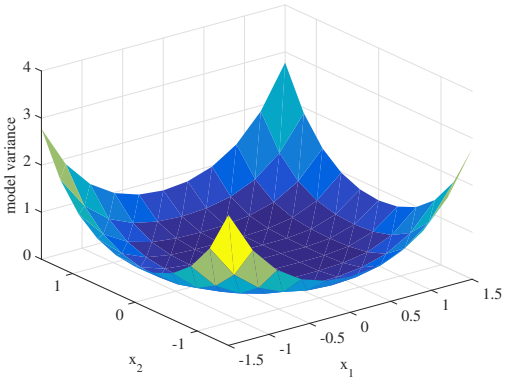


Figure 4.15: *Variance function of a composite design for a quadratic model of two factors with one single run at the center of the domain (9 runs) and* $\alpha = 1.41$.

In Matlab: There is a built-in function `E = ccdesign(n)` that generates a central composite design of n factors. The algorithm offers the possibility to favor the orthogonal or the isovariance per rotation characteristics of the design.

4.7 Doehlert design

Equiradial designs are constituted by points distributed on a sphere. To achieve this, regular geometric figures such as the pentagon, the hexagon, the octagon and the icosahedra are used. In this case also, the number of points at the center has an influence on the properties of the design. The most interesting case is the hexagonal design, also called a Doehlert network, which allows us to move the center of interest easily one step further along a gradient. This design is a good candidate for an optimization process.
At two dimensions, the matrix of experiments is constituted by the seven points presented in figure 4.16 and represented by the matrix of experiments given in equation 4.55. The number of levels is not the same for all the factors. The first factor is tested at five levels $\{-1, -0.5, 0, 0.5, 1\}$, and the second factor is tested at three levels only $\{-0.866, 0, 0.866\}$. Figure 4.17 shows the variance function over the whole experimental domain (radius 1.41). Comparing with the composite case presented in figure 4.15, observe the difference of geometry and of values. The Doehlert design has higher variance values at the border of the domain, mainly because of a smaller number of runs.

$$E = \begin{pmatrix} 1 & 0 \\ \frac{1}{2} & \frac{\sqrt{3}}{2} \\ -\frac{1}{2} & \frac{\sqrt{3}}{2} \\ -1 & 0 \\ -\frac{1}{2} & -\frac{\sqrt{3}}{2} \\ \frac{1}{2} & -\frac{\sqrt{3}}{2} \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0.5 & 0.866 \\ -0.5 & 0.866 \\ -1 & 0 \\ -0.5 & -0.866 \\ 0.5 & -0.866 \\ 0 & 0 \end{pmatrix} \tag{4.55}$$

The Doehlert design has two interesting properties:

Ease in shifting the design, giving the possibility to explore the neighborhood of the original experimental domain with a minimum of additional points. In figure 4.18 the seven original points (in red) are completed by three points (in green) which could allow the shift of the domain in six possible directions (three

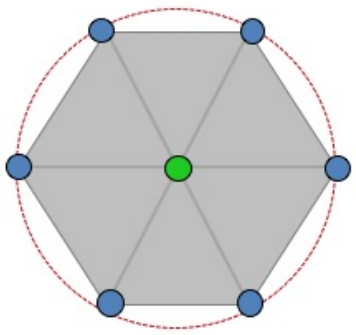


Figure 4.16: *Extension of a 2^2 factorial design (blue dots) to a Doehlert design.*

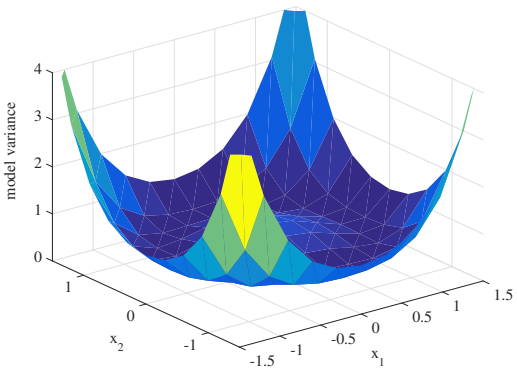


Figure 4.17: *Variance function of a Doehlert design for a quadratic model of two factors with one single run at the center of the domain (7 runs) and a radius $\rho = 1.41$.*

are represented in the figure). This property, illustrated here for a case in two dimensions, is also available in a domain with more dimensions. It is useful for optimisation, letting the experimenter follows an ascending or descending slope to reach a local extremum.

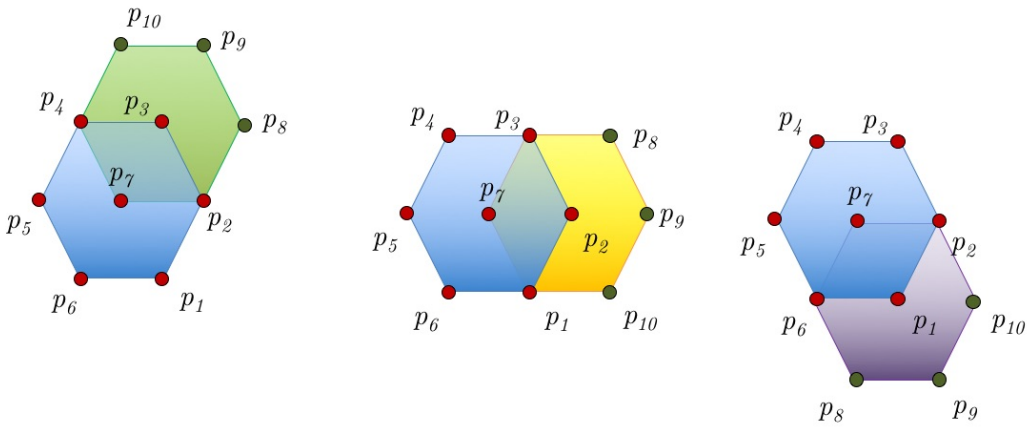


Figure 4.18: *Examples of the extension of a Doehlert design of two factors in 7 runs (red dots) with 3 additional runs (green dots).*

Ease in including new factors, making it possible to introduce additional factors after the start of the experiments. In figure 4.19 we can observe that a 2D-7 runs design can be converted to a 3D-13 runs design by adding six runs distributed in two triangles up and down of the original plane. This property is identifiable in table 4.22 giving the values of the experimental points up to five factors, and where we observe that the values of the factors positioned at the end of the list stay constant at a mid range value in the first experiments.

In Matlab: To date, there is no built-in function to generate a Doehlert design. The library of the course, LISA, offers a function `E = doehlert(n)` that generates a Doehlert design of n factors.

Table 4.22: *Coordinates of the Doehlert network up to 5 factors.*

runs	x_1	x_2	x_3	x_4	x_5
1	0	0	0	0	0
2	-1	0	0	0	0
3	1	0	0	0	0
4	-0.5	-0.866	0	0	0
5	0.5	0.866	0	0	0
6	-0.5	0.866	0	0	0
7	0.5	-0.866	0	0	0
8	-0.5	-0.2887	-0.8165	0	0
9	0.5	0.2887	0.8165	0	0
10	-0.5	0.2887	0.8165	0	0
11	0	-0.5774	0.8165	0	0
12	0.5	-0.2887	-0.8165	0	0
13	0	0.5774	-0.8165	0	0
14	-0.5	-0.2887	-0.2041	-0.7906	0
15	0.5	0.2887	0.2041	0.7906	0
16	-0.5	0.2887	0.2041	0.7906	0
17	0	-0.5774	0.2041	0.7906	0
18	0	0	-0.6124	0.7906	0
19	0.5	-0.2887	-0.2041	-0.7906	0
20	0	0.5774	-0.2041	-0.7906	0
21	0	0	0.6124	-0.7906	0
22	-0.5	-0.2887	-0.2041	-0.1581	-0.7746
23	0.5	0.2887	0.2041	0.1581	0.7746
24	-0.5	0.2887	0.2041	0.1581	0.7746
25	0	-0.5774	0.2041	0.1581	0.7746
26	0	0	-0.6124	0.1581	0.7746
27	0	0	0	-0.6325	0.7746
28	0.5	-0.2887	-0.2041	-0.1581	-0.7746
29	0	0.5774	-0.2041	-0.1581	-0.7746
30	0	0	0.6124	-0.1581	-0.7746
31	0	0	0	0.6325	-0.7746

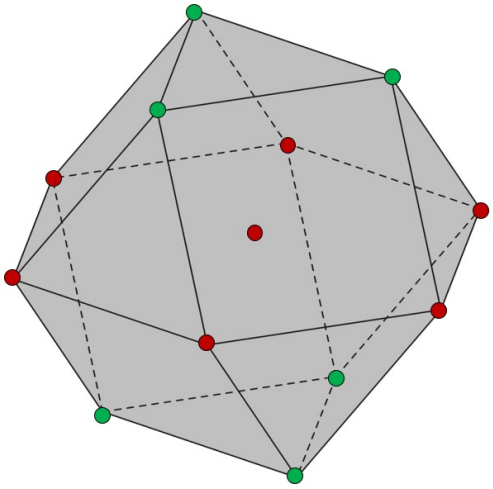


Figure 4.19: *Extension of a Doehlert design of two factors in 7 runs (red dots) to three factors with 6 additional runs(green dots).*

4.8 Box-Behnken design

Box-Behnken have proposed a 3-level design to collect data to fit a second degree model. It is constituted by the points at the center of the ridges with some additional center points. Figure 4.20 illustrates the situation for three factors. This is a spherical design as each external point lays at a distance of $\sqrt{2}$ of the central point. These characteristics confer a rotatable property to this design.

This type of design exists only for three or more factors. It can be decomposed in a series of factorial designs of two factors, the other ones remaining in their middle values as illustrated in table 4.23 for the case of a 3-factor design.

The number of runs is then obtained by

$$N_{run} = 2^2 \cdot \binom{N}{2} + N_o \tag{4.56}$$

In conclusion, this design has the advantage of a lower cost (reduced number of runs) in comparison to a 3^N or a composite design. Another advantage is the possibility to separate blocs. Table 4.24 provides a comparison between response surface designs and figure 4.21 compares them in term of number of runs for different number of factors.

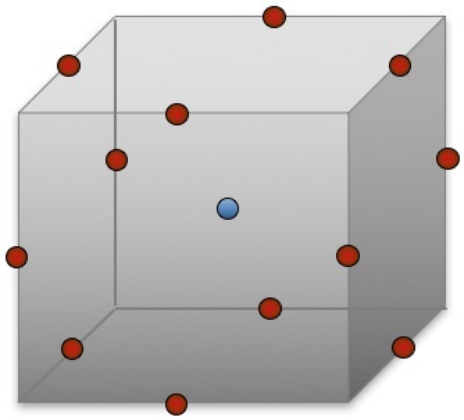


Figure 4.20: *Axonometry of a 13 run Box-Behnken design for 3 factors.*

Table 4.23: *A three-variable Box-Behnken design parted in three blocs of 5 runs. Each bloc corresponding to a 2^2 factorial design, plus a central point.*

run	x_1	x_2	x_2
1	-1	-1	0
2	-1	1	0
3	1	-1	0
4	1	1	0
5	0	0	0
6	-1	0	-1
7	-1	0	1
8	1	0	-1
9	1	0	1
9	0	0	0
11	0	-1	-1
12	0	-1	1
13	0	1	-1
14	0	1	1
15	0	0	0

Table 4.24: *Comparison of three classical 2nd-degree designs.*

Design	runs	Advantages	Disadvantages
3-level factorial	3^N	medium variance	high cost
Composite	$2^N + 2N + N_o$	low variance	medium cost
Doehlert	$N^2 + N + N_o$	low cost, extensions	high variance at the vertex
Box-Behnken	$4\binom{N}{2} + N_o$	low cost, blocking	high variance at the vertex

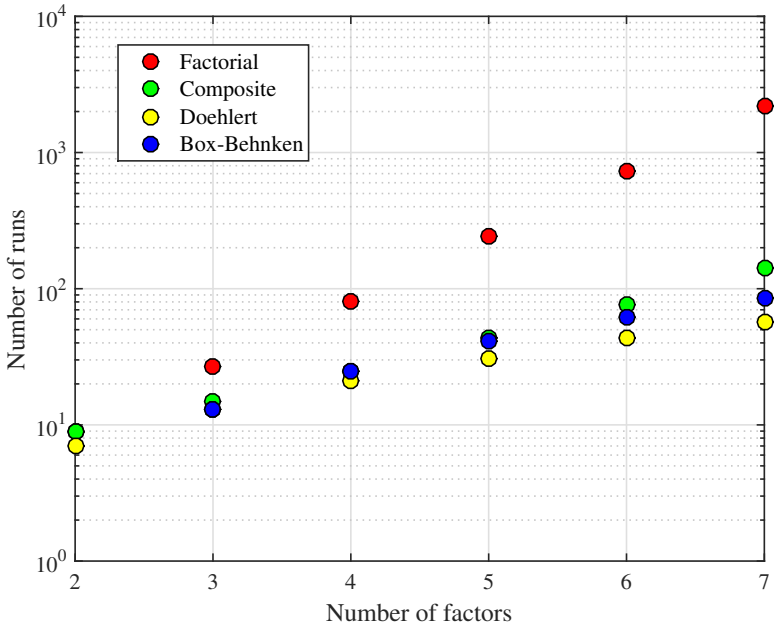


Figure 4.21: *Comparison of the number of runs for classical response surface designs.*

In Matlab: There is a built-in function `E = bbdesign(n)` that generates a Box-Behnken design of n factors. The algorithm offers the possibility to fix the number of central points and also the size of the blocs .

4.9 Extension

When in a hurry to get a first estimation of the second degree coefficients, a design called *extension* can be an interesting alternative. This is nevertheless a quick-fix to rapidly, but not very accurately, estimate the second degree curvature. Two options are presented in figure 4.23 with their respective variance function. The extension can be made by the frame (a), extending in orthogonal directions points that are positioned on a diagonal, or centrally (b) extending one point in two orthogonal directions. In two dimensions, there are four possible extended zones which will be chosen in function of the first results and of the objective of the study (figure 4.22).

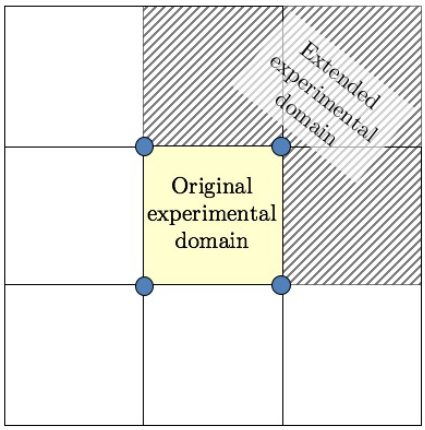
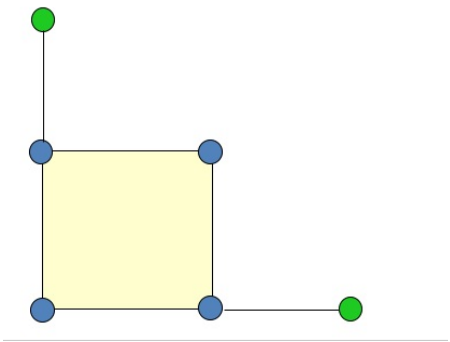


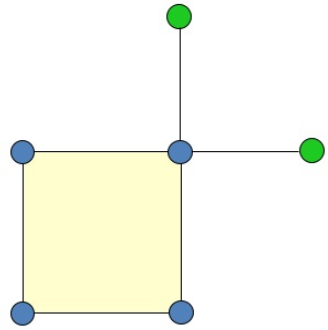
Figure 4.22: *Experimental domain (in yellow) with the extended zones. One is hatched in the upper-right corner. The other possible zones of extension are placed symmetrically at each corner.*

The variance functions show the characteristics of each design. Obviously the quality of the model is better in the neighborhood of the experimental points. The framed extension then provides a quite low variance in a diagonal band that is positioned at the lower-left corner (for the setup used here as illustration). But the variance increases rapidly when approaching the upper-right corner. The central extension would give better confidence for the model at the center of the

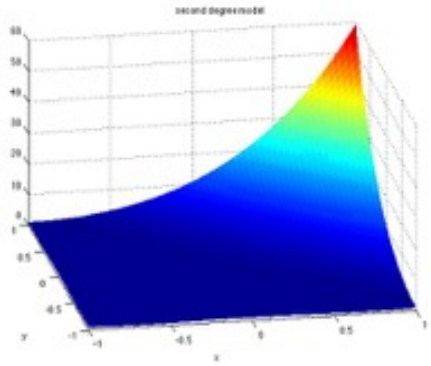
Figure 4.23: *2D extension designs with their respective variance function.*



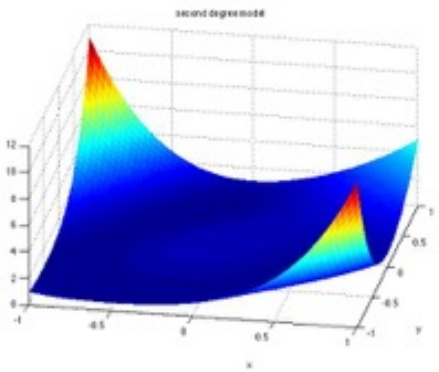
(a) *2D frame extension.*



(b) *2D central extension.*



(c) *Variance function of a 2D frame extension.*



(d) *Variance function of a 2D central extension.*

new domain, with a deterioration when approaching the lower-right corner or the upper-left corner where there is no experimental point.
The comparison of the two designs shown in table 4.25 indicates that the central extension is more interesting, from a statistical point of view.
To illustrate the usage of this design, we can consider the experiments with the sugar substitutes already presented in 4.2.1. This time, a factorial design 2^2 has been conducted (table 4.26). The regression of this data on a model with interaction $y = a_0 + a_1x_1 + a_1x_2 + a_{12}x_1x_2$ gives the standardized coefficients. Applying equation 4.2, coefficients for the function over the original domain can be computed and a corresponding response surface can be plotted, as shown in figure 4.24a:

$$y = \left(a_o - \frac{a_1\bar{u}_1}{\Delta u_1} - \frac{a_2\bar{u}_2}{\Delta u_2}\right) + \left(\frac{a_1}{\Delta u_1} - \frac{a_{12}\ \bar{u}_2}{\Delta u_1 \Delta u_2}\right) u_1 + \left(\frac{a_2}{\Delta u_2} - \frac{a_{12}\ \bar{u}_1}{\Delta u_1 \Delta u_2}\right) u_2 + \frac{a_{12}}{\Delta u_1 \Delta u_2} \tag{4.57}$$

Table 4.25: *Comparison of the two designs at the level of the trace and the determinant of the dispersion matrix, the variance inflation factors of the quadratic terms and of the maximum of the variance function for the extended experimental domain.*

Design	Trace	Determinant	$VIF(a_{ii})$	$\max\{var(\hat{Y})\}$
Framed extension	15	0.0625	6	60 σ
Central extension	9	0.0625	4.5	12 σ

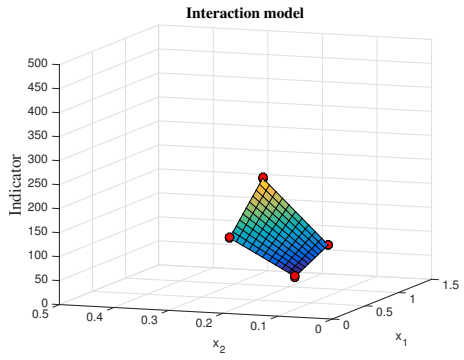
In a second step, it is interesting to determine if the surface response has a significative curvature as well as to determine if the maximum observed at the point $[0.6, 0.2]$ is a real maximum or only a point on an increasing slope. As we are dealing with toxicity, a model with a local maximum is not likely, but a curvature is a rational hypothesis, indicating a non-linear response of the cells. A rapid way to determine possible curvature would be performing a measurement at the center of the domain, point $[0.35, 0.14]$ in the original domain, $[0, 0]$ in the standard one. But this single measurement would not be sufficient to determine the two second degree coefficients a_{11} and a_{22}. Two additional measurements at least are necessary. The question is then to decide between a long term strategy or a short term.
For a long term strategy, the next step would consist in completing the factorial design to a composite design (section 4.6) or a Doehlert design (section 4.7) as shown in figure 4.25.
For a short term strategy, an *extension design* would be a good solution. In figure 4.24, the bar charts presented in (c) and (d) show clearly the advantage of the central extension in comparison with the framed extension. Table 4.27 gives the additional points and the coefficients resulting from a regression on a second degree model $y = a_0 + a_1x_1 + a_1x_2 + a_{12}x_1x_2 + a_{11}x_1^2 + a_{22}x_2^2$. The curvature of the response surface is confirmed, which would indicate a stabilisation of the toxicity for product 2.

Table 4.26: *Data of two replicates of a factorial design 2^2 for experiments with sugar substitutes with x_1 and x_2 corresponding to the standardized coordinates and a_o, a_1, a_2, a_{12} to the coefficients of a linear model with interactions.*

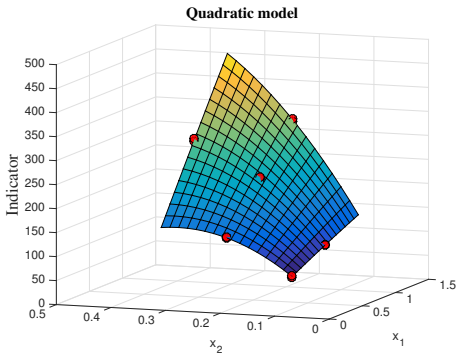
	Product 1 [g/l]	Product 2 [g/l]	Indicator #	I -	x_1 -	x_2	x_1x_2
1	0.1	0.08	80.40	1	-1	-1	1
2	0.1	0.2	150.82	1	-1	1	-1
3	0.6	0.08	117.11	1	1	-1	-1
4	0.6	0.2	250.02	1	1	1	1
5	0.1	0.08	76.05	1	-1	-1	1
6	0.1	0.2	152.35	1	-1	1	-1
7	0.6	0.08	115.81	1	1	-1	-1
8	0.6	0.2	246.62	1	1	1	1
			Coefficients	a_o	a_1	a_2	a_{12}
			standard	148.6	33.7	51.3	14.6
			SE	0.7	0.7	0.7	0.7
			original	29.5	-1.5	514	975

Table 4.27: *Data of two replicates of a central extension of a factorial design 2^2 for the experiments with sugar substitutes with x_1 and x_2 corresponding to the standardized coordinates and $a_o, a_1, a_2, a_{12}, a_{11}, a_{22}$, to the coefficients of a quadratic model.*

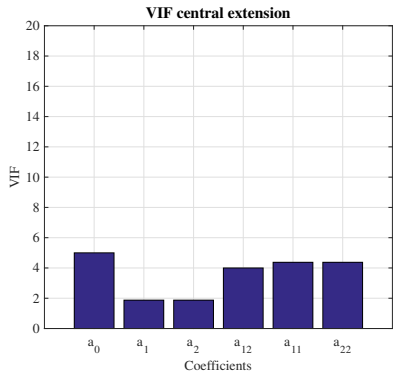
Run	Product 1 [g/l]	Product 2 [g/l]	Indicator #	I	x_1	x_2	x_1x_2	x_1^2	x_1^2
1	0.1	0.08	80.40	1	-1	-1	1	1	1
2	0.1	0.2	150.82	1	-1	0	0	1	0
3	0.6	0.08	117.11	1	0	-1	0	0	1
4	0.6	0.2	250.02	1	0	0	0	0	0
5	0.1	0.08	76.05	1	-1	-1	1	1	1
6	0.1	0.2	152.35	1	-1	0	0	1	0
7	0.6	0.08	115.81	1	0	-1	0	0	1
8	0.6	0.2	246.62	1	0	0	0	0	0
9	0.6	0.2	246.62	1	0	1	0	0	1
10	0.6	0.2	246.62	1	0	1	0	0	1
11	0.6	0.2	246.62	1	1	0	0	1	0
12	0.6	0.2	246.62	1	1	0	0	1	0
		Coefficients		a_o	a_1	a_2	a_{12}	a_{11}	a_{22}
		standardized		248.3	95.2	102.4	58.5	-1.5	-29.5
		standard error		2.0	1.4	1.4	4.0	2.4	2.4
		original		-3.7	2.7	1087.5	975	-6	-2049



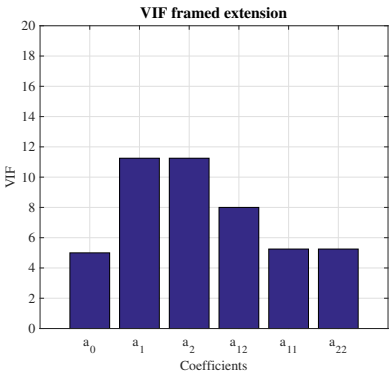
(a) Interaction response surface.



(b) Quadratic response surface.

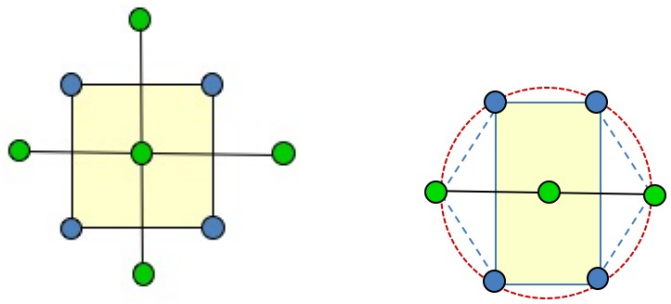


(c) VIF for the central extension.



(d) VIF for the framed extension.

Figure 4.24: *Response surface and VIF for the extension of a 2^2 factorial design.*

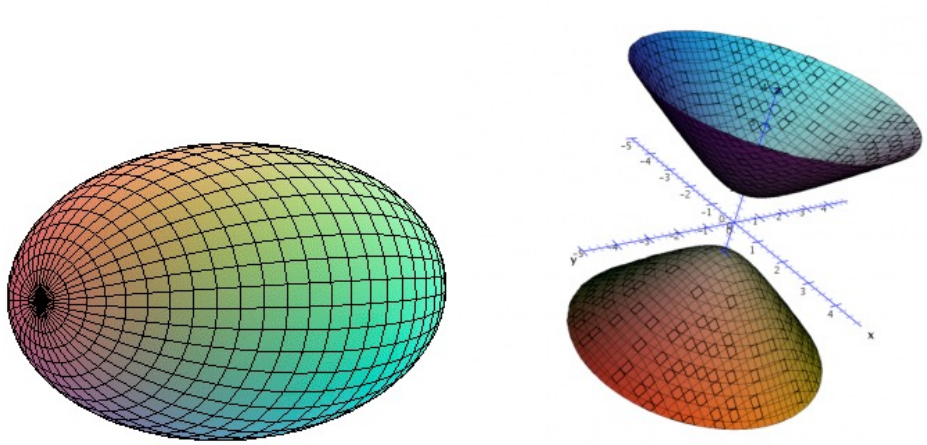


(a) ***(b)***

Figure 4.25: *Extension of a 2^2 factorial design (blue dots) to a composite design(a) and a Doehlert design (b).*

4.10 Canonical analysis

When fitting a second degree model, the resulting geometry will belong to the conic family. This means, for example in three dimensions (for three factors), that the isolines of the model are ellipsoid or a hyperboloid as presented in figure 4.26.



(a) ***(b)***

Figure 4.26: *Ellipsoid (a) and hyperboloid(b).*

The canonical analysis consist in placing this geometric figure in the experimental space and orienting its axes. The model of second degree can be written classically as:

$$y = a_o + \sum_{i=1}^{N} a_i x_i + \sum_{i \leq j}^{N} a_{ij} x_i x_j \tag{4.58}$$

This equation can be re-written in a vectorial form as

$$y = a_o + \begin{pmatrix} a_1 & \dots & a_N \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix} + \begin{pmatrix} x_1 & \dots & x_N \end{pmatrix} \begin{pmatrix} a_{11} & & \frac{1}{2}a_{1N} \\ & \ddots & \\ \frac{1}{2}a_{1N} & & a_{NN} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix} \tag{4.59}$$

$$y = a_o + \vec{a} \cdot \vec{x} + \vec{x}^{\,T} A \, \vec{x} \tag{4.60}$$

To place the figure within the experimental space, let‘s start by determining the *fix-point*, which is the point corresponding to the extremum of y, the center of the ellipsoid or the point at the center of the segment joining the extremes of the

hyperbolic surfaces for the hyperboloid. This fix-point x_s is situated where $\frac{\partial y}{\partial x_i} = 0$:

$$\begin{aligned} x_s =& -\frac{1}{2} a\ A^{-1} \\ y_s =& a_o + \frac{1}{2}\vec{a}\cdot\vec{x} \end{aligned} \tag{4.61}$$

The orientation of the figure is done by the way of the eigenvectors of the matrix A. As A is a $N \times N$ symmetric matrix , it has n real eigenvalues λ_i and the N eigenvectors are orthogonal. Once these elements have been determined, it is possible to re-write the model in a canonical form:

$$\hat{y} = y_s + \sum_{i=1}^{N} \lambda_i\ \tilde{X}_i^2 \tag{4.62}$$

where the $\tilde{X}_i$ are the coordinates of a point of the experimental space in the base of the eigenvectors.

In Matlab: The built-in function $[\mathtt{V},\mathtt{D}] = \mathtt{eig}(\mathtt{A})$ returns diagonal matrix D of eigenvalues and matrix V whose columns are the corresponding right eigenvectors, so that $\mathtt{A} * \mathtt{V} = \mathtt{V} * \mathtt{D}$.
The library of the course, LISA, offers a function `viz_quad`() that analyses a quadratic function and draws a slice plot and a isosurface plot of the model

Here is a small example: Imagine that we have an experimental situation with three factors that are investigated with a composite design as presented in table 4.28.
Fitting this result with the least squares algorithm on a second degree model provides the following function:

$$\begin{aligned} y(x_1,x_2,x_3) \ =& \ 9.26 - 1.61\ x_1 - 1.14\ x_2 + 0.88\ x_3 \\ &+1.67\ x_1\ x_2 - 2.03\ x_1\ x_3 + 1.37\ x_2\ x_3 \\ &+1.12\ x_1^2 - 3.27\ x_2^2 - 2.06\ x_3^2 \end{aligned} \tag{4.63}$$

which is not straightforward to interpret. The function can be written as follows:

$$y = a_o + \begin{pmatrix} -1.61 \\ -1.14 \\ 0.88 \end{pmatrix} \vec{x} + \vec{x} \begin{pmatrix} 1.12 & 0.83 & -1.01 \\ 0.83 & -3.27 & 0.69 \\ -1.01 & 0.69 & -2.06 \end{pmatrix} \vec{x} \tag{4.64}$$

Table 4.28: *A three-variable standardized composite design and the corresponding experimental results.*

Run	x_1	x_2	x_3	Result
1	-1	-1	-1	7.70
2	-1	-1	1	10.41
3	-1	1	-1	1.02
4	-1	1	1	8.68
5	1	-1	-1	4.67
6	1	-1	1	-1.27
7	1	1	-1	4.13
8	1	1	1	4.22
9	1.215	0	0	10.55
10	-1.215	0	0	11.84
11	0	1.215	0	1.01
12	0	-1.215	0	8.41
13	0	0	1.215	8.62
14	0	0	-1.215	4.39
15	0	0	0	8.40

Then, applying 4.61, we get the coordinates of the fix-point, which lies within the standardized experimental domain

$$x_s = \begin{pmatrix} 0.64 \\ -0.03 \\ -0.11 \end{pmatrix} \tag{4.65}$$

and the value of the function at the fix-point:

$$y_s = 8.71 \tag{4.66}$$

The calculation of the eigenvectors and eigenvalues provides:

$$\lambda_1 = -3.87 \qquad \lambda_2 = -1.84 \qquad \lambda_3 = 1.50$$

$$\vec{v}_1 = \begin{pmatrix} 0.24 \\ -0.86 \\ 0.46 \end{pmatrix} \quad \vec{v}_2 = \begin{pmatrix} 0.15 \\ 0.50 \\ 0.85 \end{pmatrix} \quad \vec{v}_3 = \begin{pmatrix} 0.96 \\ 0.13 \\ -0.25 \end{pmatrix} \tag{4.67}$$

Figure 4.27 shows that the model corresponds to a hyperboloid structure with a fix-point within the domain.

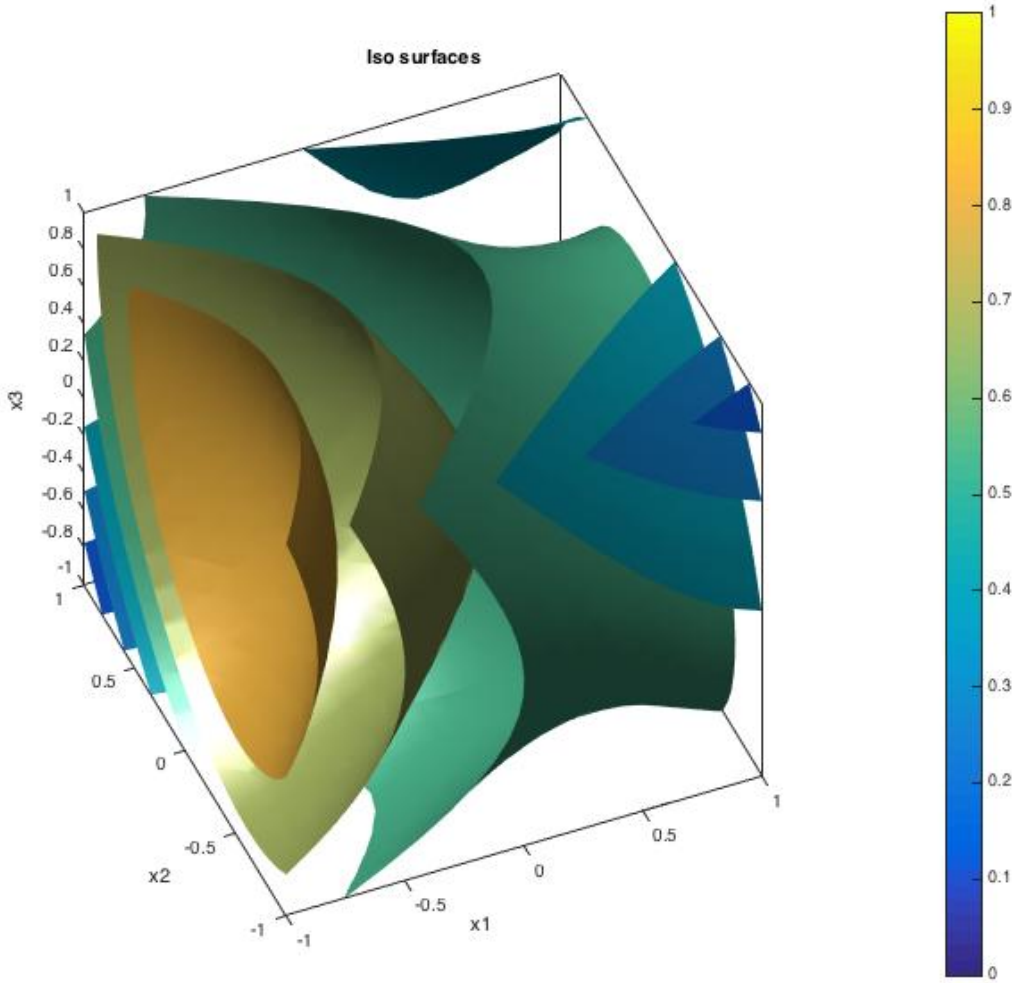


Figure 4.27: *Isosurfaces of the model produced with the LISA function* `viz_quad()` *and showing an hyperboloid structure.*

Chapter 5

Notes

5.1 The order of the factors in the sweeping does not matter

Consider a table of data in thee dimensions x_{ijk} with $i = 1 : n$, $j = 1 : m$ and $k = 1 : r$ corresponding to r replicates of a set of experiments of two variables with respectively n and m levels. We want to summarize it by the model $x_{ij} = \mu + \alpha_j + \beta_i + \alpha\beta_{ij} + \epsilon_{ijk}$. The grand mean μ corresponds to

$$\mu = \frac{1}{nmr} \sum_i \sum_j \sum_k x_{ijk} \tag{5.1}$$

The first residue $R(1)_{ijk}$ is obtained by subtracting the grand mean to each element:

$$R(1)_{ijk} = x_{ijk} - \mu = x_{ijk} - \frac{1}{nmr} \sum_i \sum_j \sum_k x_{ijk} \tag{5.2}$$

The effects of the *column* variable α_j are computed by averaging each column for all the replicates:

$$\alpha_j = \frac{1}{nr} \sum_i \sum_k R(1)_{ijk} = \frac{1}{nr} \sum_i \sum_k (x_{ijk} - \mu) = \frac{1}{nr} \sum_i \sum_k x_{ijk} \; - \; \mu = \mu_j - \mu \tag{5.3}$$

With μ_j the average of the column j for all the replicates. The second residue $R(2)_{ijk}$ is computed by subtracting the effect α_j to the first residue $R(1)_{ijk}$ then

$$R(2)_{ijk} = R(1)_{ijk} - \alpha_j = (x_{ijk} - \mu) - (\mu_j - \mu) = x_{ijk} - \mu_j \tag{5.4}$$

The effects of the *row* variable β_i are now computed by averaging the residue $R(1)_{ijk}$ for each row and for all replicates, so:

$$\beta_i = \frac{1}{mr}\sum_j\sum_k R(2)_{ijk} = \frac{1}{mr}\sum_j\sum_k (x_{ijk}-\mu_j) = \frac{1}{mr}\sum_j\sum_k x_{ijk}-\mu = \mu_i-\mu \quad (5.5)$$

The residue $R(3)_{ijk}$ is calculated the same way as the other residues:

$$R(3)_{ijk} = R(2)_{ijk} - \beta_i = R(1)_{ijk} - \alpha_j - \beta_i = x_{ijk} - \mu - \alpha_j - \beta_i \quad (5.6)$$

The interaction effects are computed by averaging $R(3)_{ijk}$ for each row, for each column for all the replicates:

$$\alpha\beta_{ij} = \frac{1}{r}\sum_k R(3)_{ijk} = \frac{1}{r}\sum_k x_{ijk} - \mu - \alpha_j - \beta_i \quad (5.7)$$

And finally the last residue is

$$\epsilon_{ijk} = x_{ij} - \mu - \alpha_j - \beta_i - \alpha\beta_{ij} \quad (5.8)$$

The sweeping can be simplified by subtracting the grand mean to the average of the rows and the columns.

5.2 The confidence region around the solution of the LSF is an ellipse

Let consider that the true model to be identified is given in equation 4.5 and the solution obtained through LSF in equation 4.9. If the measurements y_i are normally distributed and under the hypothesis of homoscedasticity is is possible to write that

$$\hat{\alpha} \sim N\left(\alpha, (X^TX)^{-1}\sigma^2\right) \quad (5.9)$$

Let now compare the true model η with the estimated model $\hat{Y}$

$$\eta - \hat{Y} = X\alpha - X\hat{\alpha} = X(\alpha - \hat{\alpha}) \quad (5.10)$$

It can be demonstrated (see [6] pp. 74-76) that the distance between them divided by the square of the MSE s^2 follows a Fisher distribution :

$$(\alpha - \hat{\alpha})^T(X^TX)(\alpha - \hat{\alpha})/ps^2 \sim F(p, \nu) \quad (5.11)$$

with p the number of parameters of the model and ν the degree of freedom of s^2.

Now let define $F_\beta(p,\nu)$ the β-significance level of the F distribution. The equation

$$(\alpha-\hat{\alpha})^T(X^TX)(\alpha-\hat{\alpha})/ps^2 = F_\beta(p,\nu) \tag{5.12}$$

defines an ellipsoidal $1-\beta$ confidence region for α. That means that there is a probability $1-\beta$ to find the true α within this region. If the matrix of information X^TX is diagonal, then this ellipsoid will be aligned with the axis of the parameter space.

Chapter 6

Bibliography

[1] Gunter, B.H. How statistical design concepts can improve experimentation in the physical sciences. *Computers in Physics* **1993**, *7*, 262–272.

[2] Covey, S.R. *The seven habits of highly effective people: restoring the character ethic*, 1st fireside ed ed.; Fireside Book: New York, 1990.

[3] Box, G.E.P.; Hunter, J.S.; Hunter, W.G. *Statistics for Experimenters, An introduction to design, data analysis and model building*, first ed.; Wiley Series in Probability and Mathematical Statistics, John Wyley and Son, 1978.

[4] Montgomery, D.C. *Design and analysis of experiments*, 7th edition ed.; John Wyley and Son, 2009.

[5] Ryan, T.P. *Modern Experimental Design*; Vol. 1, *Wiley Series in Probability and Statistics*, Wyley, 2007.

[6] Box, G.E.P.; Draper, N.R. *Empirical model-building and response surfaces*; Wiley Series in Probability and Mathematical Statistics, John Wyley and Son, 1987.

[7] Box, J.F. *R. A. Fisher, the Life of a Scientist*; John Wiley & Sons, 1978.

[8] Plackett, R.L.; Burman, J.P. The design of optimum multifactorial experiments. *Biometrika* **1946**, pp. 305–325.

[9] MacNeish, H.F. Euler Squares. *Annals of Mathematics* **1922**, *23*, pp. 221–227.

[10] Vandekerckhove, J.; Matzke, D.; Wagenmakers, E. Model Comparison and the Principle of Parsimony. *The Oxford Handbook of Computational and Mathematical Psychology* **2015**.

[11] Margolin, B.H. Orthogonal main-effect 2 n 3 m designs and two-factor interaction aliasing. *Technometrics* **1968**, *10*, 559–573.

[12] Chen, J.; Sun, D.; Wu, C. A catalogue of two-level and three-level fractional factorial designs with small runs. *International Statistical Review* **1993**, *61*, 131–131.

[13] Box, G.E.; Behnken, D.W. Some new three level designs for the study of quantitative variables. *Technometrics* **1960**, *2*, 455–475.

[14] Khuri, A.I.; Cornell, J.A. *Response Surfaces: Designs and Analyses*; Marcel Dekker, Inc.: New York, NY, USA, 1987.