Chapter 8
Response Surface Designs

An important objective of the design of experiment is the comparison of treatments either whose nature can be qualitative or quantitative. The objective in both the cases is to detect structure of some form among the treatment effects. The methods of regression analysis can be used in case of the treatments are quantitative in nature.

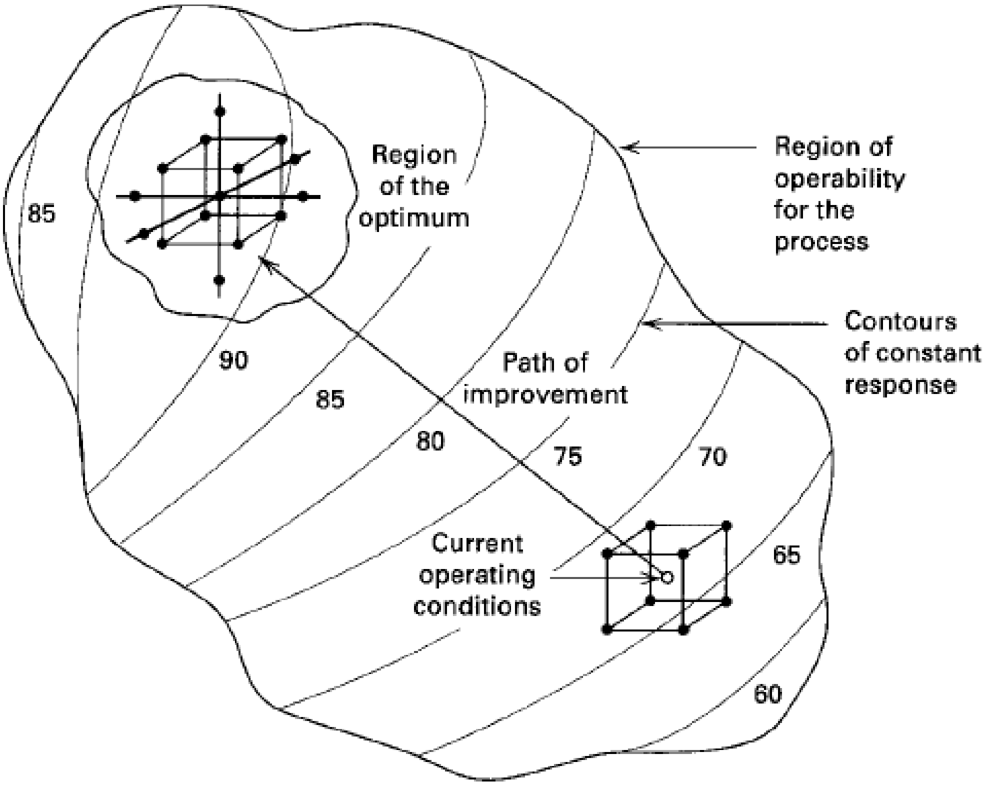
If the treatments are represented by the level of one treatment factor then the dependence of treatment effects on treatments can be represented by a response curve. If the treatments are level combinations of two or more treatment factors, then a response surface can be used. Such curves can be used to make judgments about treatment structure and to know the relationship between treatments and responses, or between input and output variables. Such knowledge of relationship is important if one wants to know the treatment combination which gives the optimal. The optimum can be defined in terms of highest or lowest response depending upon the situation. The exact relationship is never known to the experimenter but an attempt is made to approximate it. This can be achieved by using the methods of experimental design and regression analysis. Such methods are referred to as a **response surface methodology (RSM).**

We consider a simple example to illustrate the application of RSM. The relationship $y = x_1^2 + x_2^2$ can be represented as a two-dimensional surface in a three-dimensional space and this indicates the dependence of y on x_1 and x_2. If the units of the input variables are changed to as $x_1^* = 3x_1$ and $x_2^* = 5x_2$, then the relationship becomes

$$y = \frac{1}{9}x_1^{*2} + \frac{1}{25}x_2^{*2}.$$

Note that y is constant on the curves $x_1^2 + x_2^2 =$ constant, which is a circle. So on circles in the $(x_1, x_2)-$plane, y is constant on the curves $\frac{1}{9}x_1^{*2} + \frac{1}{25}x_2^{*2} =$ constant, which is now an ellipse in the $(x_1^*, x_2^*)-$plane. Obviously, these two surfaces are quite different from each other and this illustrate that the choice of surface depends on the choice of units of plotting also. This type of consideration are always kept in mind while doing with RSM.

RSM is a sequential procedure. Often when the experimenter is at a point on the response surface which is far away from the point of optimum, then there is a little curvature present in the system. In such a situation, the first order RSM will be appropriate. This is presented in the following figure:



Region of the optimum
Region of operability for the process
Contours of constant response
Path of improvement
Current operating conditions
85
90
85
80
75
70
65
60

The sequental nature of response surface methodology

The objective of experimenter is to lead along a path of improvement toward the general vicinity of the **optimum** in an efficient way. Once the experimenter is close to that region, a more elaborate model, e.g., such as second order model, may be employed. Then the analysis can be performed to locate the optimum.

Formulation of the problem

Suppose we have k quantitative factors $F_2, F_2, ..., F_k$ which affect the particular response. Each factor has continuous levels within a certain interval; e.g., F_i has levels X_i with $X_{iL} \le X_i \le X_{iU} (i = 1, 2, ..., k)$. A hypercube $\{X_{iL} \le X_i \le X_{iU}; (i = 1, 2, ..., k)\}$ can be defined as the **operational region** (OR). In the OR every level combination $(X_1, X_2, .., X_k)$ represents a feasible operating condition. Assuming that each such setting can be controlled (essentially without error) by the experimenter, a response is considered as a function of $(X_1 X_2 ... X_K)$ and is associated as

$$\eta = \phi\left(X_1, X_2, ..., X_k : \theta_1, \theta_2, ..., \theta_q) = \phi(X; \theta\right)$$

where $\theta_1, \theta_2, ..., \theta_q$ are parameters, $X = (X_1, X_2, ..., X_k)'$ and $\theta = (\theta_1, \theta_2, ..., \theta_q)'$. Now the true response $\eta = \eta(X_1, X_2, ..., X_k)$, and the form of the functional relationship ϕ at any given point in OR are unknown. Only observed responses $y = y(X)$ are available and attempt is made to approximate $\phi(X, \theta)$ by a polynomial function $f(X, \beta)$ in X. Then consider a model of the form

$$y(X) = f(X, \beta) + \varepsilon(X),$$

in place of $\eta = \phi(X, \theta)$ where $\beta = (\beta_1, \beta_2, ..., \beta_m)'$ are unknown parameters and $\varepsilon(X)$ represents random error.

Ideally, the experimenter wants to have $y(X)$ available for a sufficiently fine grid in OR so that approximate ϕ or a realization of ϕ can be adequately approximated. It is difficult in real experiments to do so. Instead of that, the experimenter has only a relatively small number of points (these are sometimes referred to as **runs** or **experiments**) and they are usually confined to a region called as **experimental region (ER)** or region of interest. Obviously, such an ER is contained in OR. The basic idea behind this is then the following.

- Based on the limited available knowledge about the process under study, the experimenter chooses an ER.
- Assuming that the response surface for ER is sufficiently smooth so that it can be approximated by a lower polynomial, say first or second degree polynomial.
- Then an appropriate treatment and error control design can be chosen to estimate the coefficients of the polynomial. From this, the response can be predicted for any point in ER.
- If one of these points attains the optimal response then (one may have reached an optimum which may only be an optimum either locally or globally.
- If the fitted response surface indicates that the optimum may only be outside ER then the experimenter can choose a new ER and repeat the whole process until the (predicted) optimum can be located more precisely.

This procedure leads to two sources of error:

(i) There can be experimental and sampling error in estimating the function $f(X; \beta)$ and

(ii) some bias may be introduced is approximating $\phi(X, \theta)$ due to the inadequacy of $f(X; \beta)$.

The objective of response surface designs is to minimize these errors. The basic requirements for such designs are as follows:

1. The design should allow $f(X,\beta)$ to be estimated with reasonable precision in ER under the assumption that a polynomial $f(X;\beta)$ of degree approximates $\phi(X;\theta)$ sufficiently well.
2. A provision in design should be there to check whether the chosen $f(X;\beta)$ provides a satisfactory fit to the response surface or whether a different polynomial is more appropriate.
3. The design should not contain large number of experimental points.
4. The design should be available for blocking of the experimental points.
5. One should be able to modify the design in case the polynomial of degree to which the polynomial is fitted is not found to be adequate and a polynomial of next higher degree needs to be fitted.

Now we discuss the basic tools and designs of RSM and point out the connection to treatment and error-control designs.

First-order models and designs

First-order regression model

The response surface function ϕ is approximated by a first-order polynomial within a small region based on the k input variables $X_1, X_2, ..., X_k$ as follows:

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i X_i + \varepsilon.$$

Here β_i is the regression coefficient associated with X_i and measures the change in the response y due to a change in the input variable X_i. This kind of information is available from the main effect from a factorial experiment where each factor has two levels. A good choice of a response surface design for such a situation is a 2^k factorial, or a fraction of it.

Suppose 2^k factorial is considered as a choice of response surface design then there are 2^k experimental points $(X_1, X_2, ..., X_k)_j$ say, with $j = 1,2,..,2^k$. With each level combination being replicated r times in a CRD, there are $N = r2^k$ experimental runs. The low and high level of the i^{th} factor are denoted by X_{i0} and X_{i1}, respectively. If the experimenter decides to use the coded levels X_i in place of X_{io} and X_{i1}, then the coded levels are expressed as

$$X_i = \frac{X_i - \overline{X}}{\frac{1}{2}(X_{i1} - X_{i0})}$$

With such a transformation, the low level becomes $x_{i0} = -1$ and high level becomes $x_{i1} = 1$. The model $y = \beta_0 + \sum_{i=1}^{k} \beta_i X_i + \varepsilon$ is rewritten as

$$y(x_1, x_2, ..., x_k)_\ell = \beta_0^* + \sum_{i=1}^{k} \beta_i^* x_i + \varepsilon(x_1, x_2, ..., x_k)_\ell,$$

where $x_i = \pm 1$, or in matrix notation as

$$y = (L,\ \ D)\beta^* + \varepsilon,$$

where y is the $N \times 1$ vector of observation, L is an $N \times 1$ vector of unity elements, D is the $N \times k$ design-model matrix consisting of elements of elements -1's and 1's, $\beta^* = (\beta_0^*, \beta_1^*, ..., \beta_k^*)'$ and ε is the $N \times 1$ vectors of errors. More specifically, let

$$D = (d_1, d_2, ..., d_k)$$

where each $d_i, i = 1, 2, ..., k$ is a $N \times 1$ vector and each d_i has $r2^{k-1}$ elements equal to -1 and $r2^{k-1}$ elements equal to 1. Thus $L'd_i = 0$ for every i. The d_i's are orthogonal to each other.

Least squares analysis

Using the least squares principle, the normal equations for the β_i^* of the model

$$y(x_1, x_2, ..., x_k)_\ell = \beta_0^* + \sum_{i=1}^{k} \beta_i^* x_i + \varepsilon(x_1, x_2, ..., x_k)_\ell,$$

are obtained as

$$(L,\ \ D)'(L,\ \ D)\hat{\beta}* = (L,\ \ D)'y$$

or $$N\hat{\beta}^* = \begin{pmatrix} L'y \\ D'y \end{pmatrix}.$$

Thus

$$\hat{\beta}_0^* = \frac{1}{N} \sum_{x_1, x_2, .., x_k} \sum_{\ell} y(x_1, x_2, ..., x_k)_\ell = \overline{y}$$

and

$$\hat{\beta}_i^* = \frac{1}{N} d_i' y$$
$$= \frac{1}{N}[(\text{sum of all observations with } x_i = 1) \text{ - } (\text{sum of all observations with } x_i = 1)].$$

Then

$$Var(\hat{\beta}_i^*) = \frac{1}{N}\sigma_\varepsilon^2$$

and

$$Cov(\hat{\beta}_i^*, \hat{\beta}_{i'}^*) = 0.$$

So, for any given point $z = (z_1, z_2, ..., z_k)'$ in the ER given by $\{-1 \le z_i \le 1; i = 1, 2, ..., k\}$, the predicted response are obtained as

$$\hat{y}(z) = \hat{\beta}_0^* + \sum_{i=1}^{k} \hat{\beta}_i^* z_i$$

with variance

$$Var[\hat{y}(z)] = \frac{1}{N}\left(1 + \sum_{i=1}^{k} z_i^2\right)\sigma_\varepsilon^2.$$

In order to know that which of the factors are influential and also to know the response surface given by $\hat{y}(z)$, we need to obtain an estimate of σ_ε^2. This is obtained through the analysis of variance as given in the following table with notations

$$y(x)_\ell = y(x_1, x_2, ..., x_k)_\ell$$
$$\sum_x = \sum_{x_1, x_2, ..., x_k}$$
$$D_1 = \text{SS(Total)} - N\sum_{i=1}^{k}\left(\hat{\beta}_i^*\right)^2$$
$$D_2 = D_1 - \text{SS}(PE)$$

ANOVA for first-order response surface design

Source	Degrees of freedom	Sum of squares
Regression	k	
β_1^*	1	$N(\hat{\beta}_1^*)^2$
β_2^*	1	$N(\hat{\beta}_2^*)^2$
$\vdots$	$\vdots$	$\vdots$
β_k^*	1	$N(\hat{\beta}_k^*)^2$
Error	$r2^k - k - 1$	$D_1 = SSE$
- Lack-of-fit error	$2^k - k - 1$	$D_2 = \text{SS}(LOF)$
- Pure error	$2^k(r-1)$	$\sum_x \sum_\ell \left[y(x)_\ell - \bar{y}(x)_o\right]^2 = \text{SS}(PE)$
Total	$N - 1$	$\sum_{xo\ell} \left[y(x)_\ell - \bar{y}(o)_o\right]^2$

Note that the SSE consists of following two parts:

(i) The usual pure error sum of squares for a CRD, denoted by SS(PE) and

(ii) The lack of fit (LOF) sum of the sums of squares for all the interactions for the 2^k factorial denoted here by SS(LOF). This sum of squares can be used to test whether the postulated model $y(x_1, x_2, ..., x_k)_\ell = \beta_0^* + \sum_{i=1}^{k} \beta_i^* x_i + \varepsilon(x_1, x_2, ..., x_k)_\ell$ provides a sufficiently good enough fit to the data.

To test whether the i^{th} factor contributes in explaining the response, we use the following F-statistic.

$$F_i = \frac{SS(\beta_i^*)}{MSE} \quad (i = 1, 2, ..., k)$$

which follows the F-distribution with 1 and $(N - k - 1)$ degrees of freedom . Suppose without loss of generality, only the first k_1 factors are important. Then instead of using the model $y(x_1, x_2, ..., x_k)_\ell = \beta_0^* + \sum_{i=1}^{k} \beta_i^* x_i + \varepsilon(x_1, x_2, ..., x_k)_\ell$, following model based on k_1 factors is used:

$$y(x_1, x_2, ..., x_k)_\ell = \beta_0^* + \sum_{i=1}^{k_1} \beta_i^* x_i + \varepsilon(x_1, x_2, ..., x_k)_\ell$$

and the predicted response then becomes

$$\hat{y}(z) = \hat{\beta}_0^* + \sum_{i=1}^{k_1} \hat{\beta}_i^* z_i$$

with the estimated variance as

$$\widehat{Var}\left[\hat{y}(z)\right] = \frac{1}{N}\left(1 + \sum_{i=1}^{k_1} z_i^2\right) MSE .$$

Then the responses for two different sets of input variables, $z = \left(z_1, z_2, ..., z_{k_1}\right)'$ and $w = \left(w_1, w_2, .., w_{k_1}\right)'$ are compared by considering the difference in the predicted values based on these input variable as

$$\hat{y}(z) - \hat{y}(w) = \sum_{i=1}^{k_1} \hat{\beta}_i^* (z_i - w_i)$$

and its estimated variance is given by

$$\widehat{Var}\left[\hat{y}(z) - \hat{y}(w)\right] = \frac{1}{N} \sum_{i=1}^{k_1} (z_i - w_i)^2 MSE .$$

Similarly, the experimenter can also consider the differences in responses if some of the input variables are kept constant at a desired level and the remaining input variables are varied to achieve optimum response in ER. Since the true response surface is being approximated and due to experimental error, there may not exist a single level combination which achieves the optimum response. Instead of this, there may exist a neighbourhood in which the optimum may lie and this optimum may not be significantly different from each other.

Alternative Design

It may not be a good idea to use the full 2^k factorial to estimate the parameters of a first-order response surface as this may involve large number of observations to handle. There are basically two ways to reduce the number of experimental points. One way is to replicate each design point $(x_1, x_2, ..., x_k)$ only once and in such case $\text{SS}(PE) = 0$ and $SSE = SS(LOF)$.

Another alternative is to use only a fraction of a 2^k factorial either as a single replicate or as a CRD with more than one replications. In either case, the experimenter has to choose a fraction such that all the k main effects are estimable with sufficient degrees of freedom for error so that comparisons like $\hat{y}(z) - \hat{y}(w)$ can be made with satisfactory statistical power as measured by its variance. This means that if we need to choose a very small fraction, then this can be achieved by fractional factorials, with several replications for each design point.

An important aspect in a 2^k factorial is that the blocking can be introduced easily without sacrificing the estimation of the main effects. This will help in reducing the experimental material as well as the cost and provide simplicity in the experimentation. We discussed this aspect in fractional factorial module.

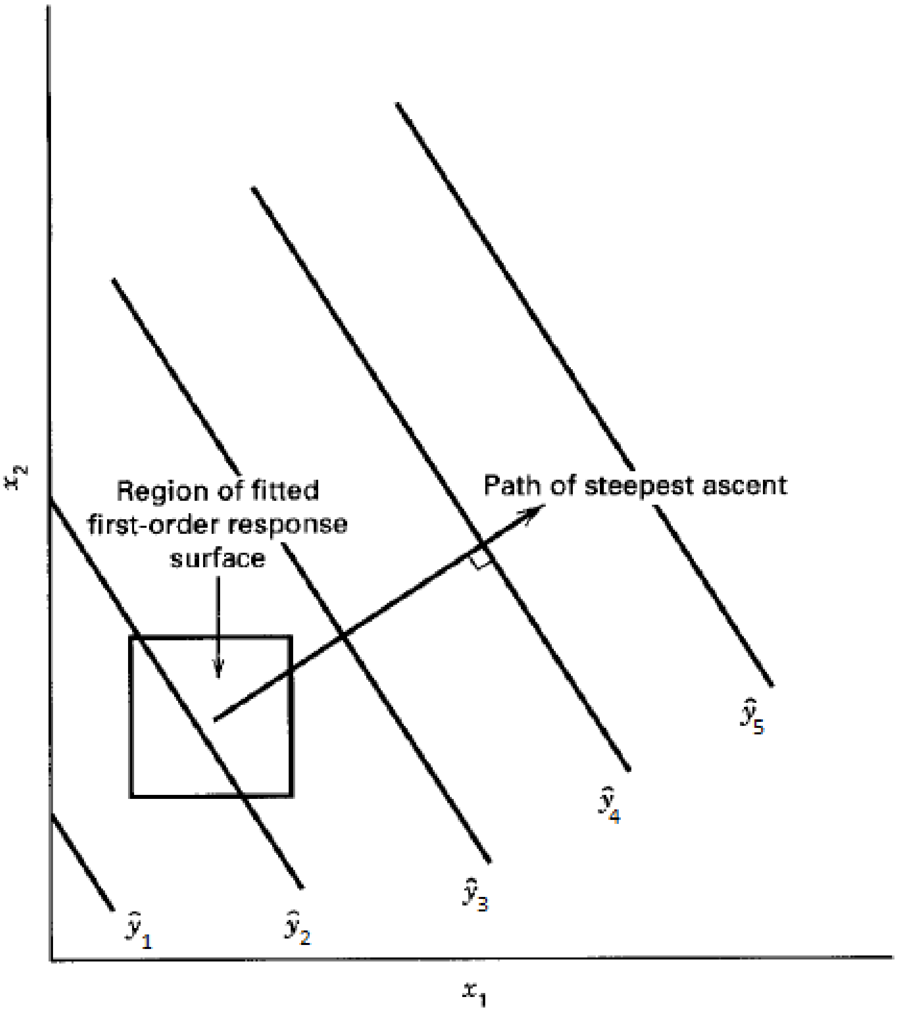
The method of steepest ascent

In many experimental conditions, the initial estimate of the optimum operating conditions for the system may be away from the actual optimum. In such conditions, one would like to move rapidly to the general vicinity of the optimum. It is expected to have a procedure which is simple to use and economically efficient. When the experimenter is remote away from the optimum, then usually it is assumed that a first-order model is an adequate approximation to the true surface in a small region of the $x's$.

The **method of steepest ascent** is a procedure for moving sequentially along the path of steepest ascent, i.e., in the direction of the maximum increase in the response. If minimization is desired, then this technique is called as the **method of steepest descent**. The fitted first-order model is

$$\hat{y} = \hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i x_i$$

and the first-order response surface can be represented as the contours of $\hat{y}$. The contours are a series of parallel lines such as shown in following figure:



First order response surface and path of steepest ascent

The direction of steepest ascent is the direction in which $\hat{y}$ increases most rapidly. Such direction is parallel to the normal to the fitted response surface. The experimenter usually take as the **path of steepest ascent** the line through the center of the region of interest and normal to the fitted surface. Thus, the steps along the path are proportional to the regression coefficients $\hat{\beta}_i$'s. The actual step size is determined by the experimenter based on process knowledge or other practical considerations.

The experiments are continued to be conducted along the path of steepest ascent until no further increase in response is observed. Then a new first-order model which may be a fit, a new path of steepest ascent determined, and the procedure continued. Finally, the experimenter will arrive in the vicinity of the optimum. This is judged by the lack of fit test of a first-order model. Some additional experiments are conducted to obtain a more precise estimate of the optimum at this point.

Analysis of a second-order response surface

When the experimenter is away from optimum, a lower order model is chosen to start with.

When the experimenter is relatively close to the optimum, then a model that incorporates curvature is usually required to approximate the response. In most cases, the second-order model is found to be suitable as

$$y = \beta_0 + \sum_{i=1}^{k} \beta_{ii} x_i + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \sum_{i<j}\sum \beta_{ij} x_i x_j + \varepsilon.$$

Now we discuss how to use this fitted model to find the optimum set of operating conditions for the $x's$ and to characterize the nature of the response surface.

Location of the stationary point

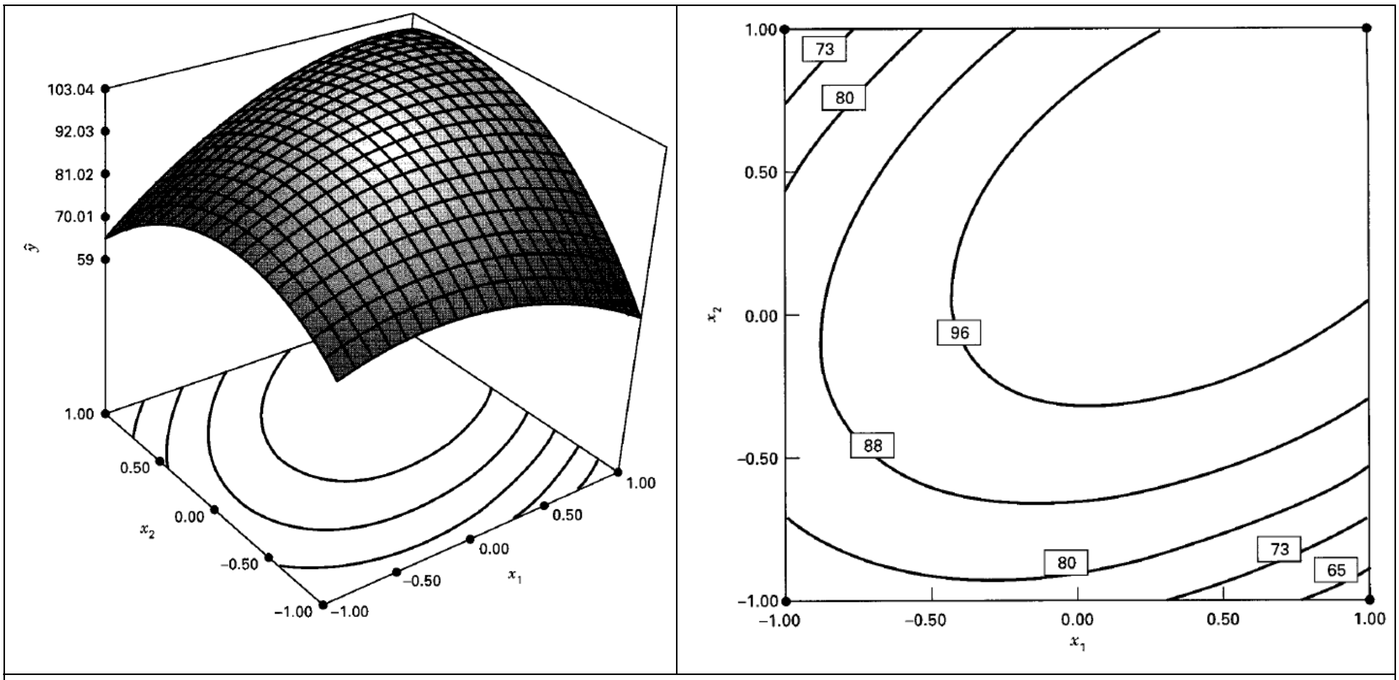
Suppose we wish to find the levels of $x_1, x_2, ..., x_k$ that optimize the predicted response. This point, if it exists, will be the set of $x_1, x_2, ..., x_k$ for which the partial derivatives $\frac{\partial \hat{y}}{dx_1} = \frac{\partial \hat{y}}{\partial x_2} = ... = \frac{\partial \hat{y}}{\partial x_k} = 0.$ This point, say $x_{1,s}, x_{2,s}, ..., x_{k,s}$, is called the **stationary point**. The stationary point could represent

(1) a point of maximum response,

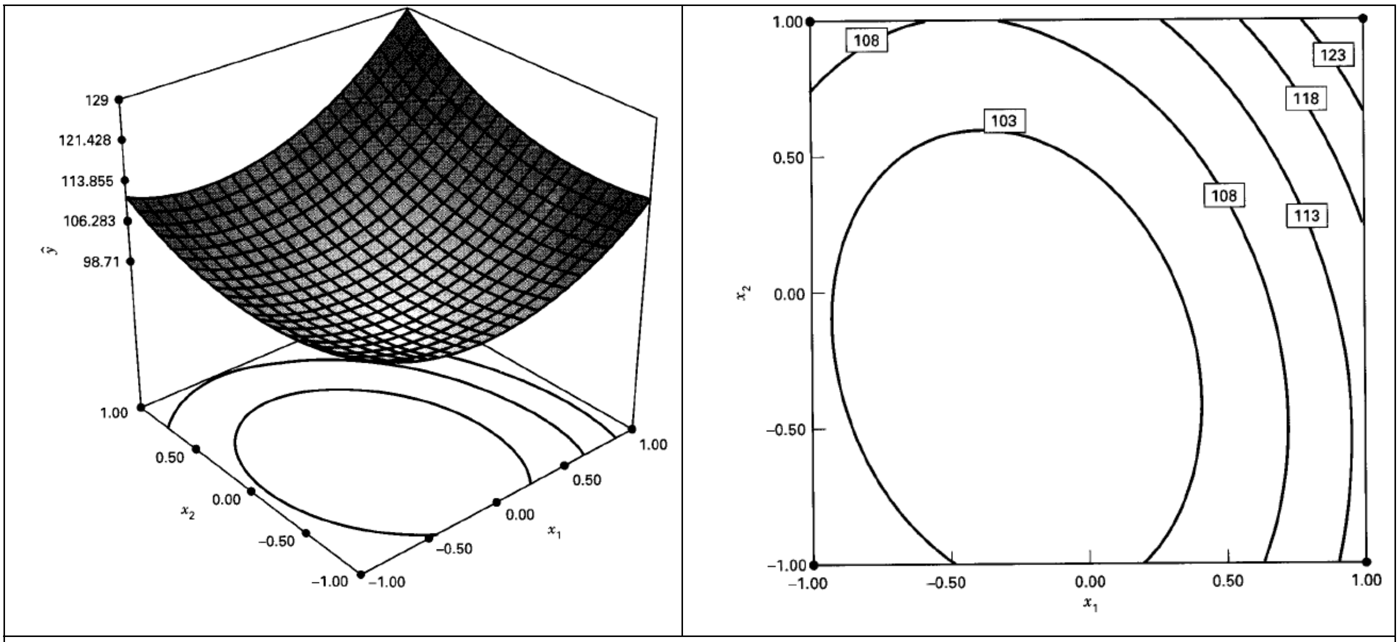
(2) a point of minimum response, or

(3) a saddle point.

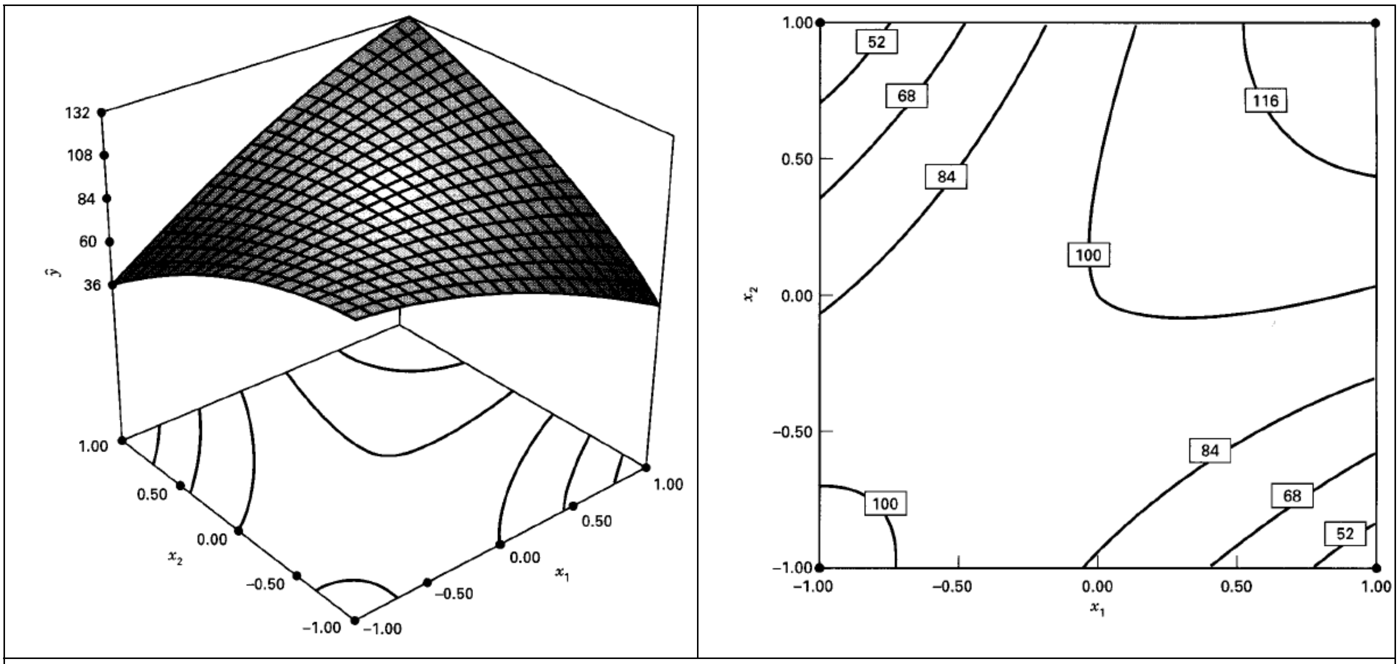
These three possibilities are shown in the following figures:



Response surface and contour plot illustrating with a maximum



Response surface and contour plot illustrating with a minimum



Response surface and contour plot illustrating a saddle point (or minimax)

Contour plots are very important in the study of the response surface. Contours are generated with the help of computer software. Such contours help the experimenter in characterizing the shape of surface. This also helps is locating the optimum with reasonably lower variability.

We may obtain a general mathematical solution for the location of the stationary point. The second-order model can be expressed in matrix notations as

$$\hat{y} = \hat{\beta}_0 + x'b + x'Bx$$

where

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix}, \quad b = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} \hat{\beta}_{11}, & \hat{\beta}_{12}/2, & ..., & \hat{\beta}_{1k}/2 \\ & \hat{\beta}_{22}, & ..., & \hat{\beta}_{2k}/2 \\ & & \ddots & \vdots \\ & & & \hat{\beta}_{kk} \end{bmatrix}$$

where b is a $(k \times 1)$ vector of the first-order regression coefficients and B is a $(k \times k)$ symmetric matrix whose main diagonal elements are the **pure** quadratic coefficients $(\hat{\beta}_{ii})$ and whose off-diagonal elements are one-half the **mixed** quadratic coefficients $(\hat{\beta}_{ij}, i \neq j)$. The stationary points obtained by solving $d\hat{y}/dx = 0$ as

$$\frac{\partial \hat{y}}{\partial x} = b + 2Bx = 0.$$

which gives the stationary point as

$$x_s = -\frac{1}{2}B^{-1}b$$

The predicted response at the stationary point is found by substituting x_s into $\hat{y} = \hat{\beta}_0 + x'b + x'Bx$ as

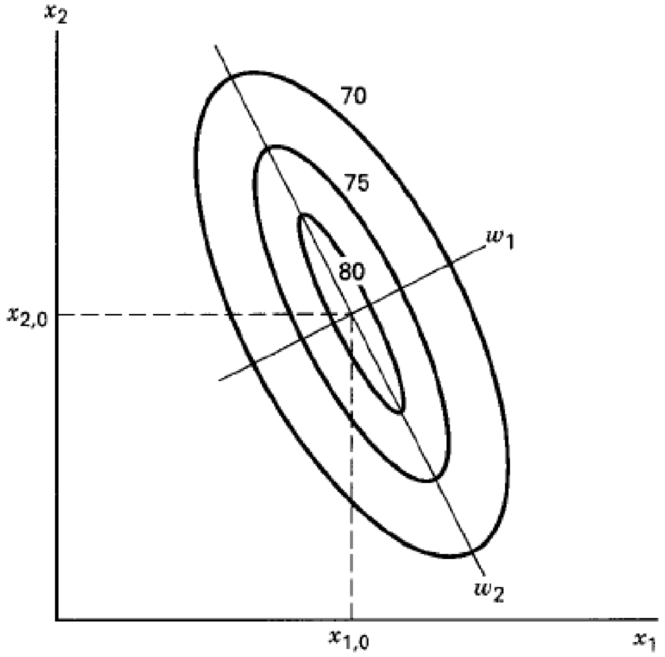
$$\hat{y}_s = \hat{\beta}_0 + \frac{1}{2}x_s^{'}b.$$

Characterizing the response surface

Once the stationary point is found then the response surface is characterized in the immediate vicinity of this point. The meaning of **characterize** is to determine whether the stationary point is a point of maximum or minimum response or a saddle point. The relative sensitivity of the response to the variables $x_1, x_2, .., x_k$ is also studied.

The most straightforward way to do this is to examine a contour plot of the fitted model. It is easier to study the contour plot if there are only two or three process variables (the $x's$), When there are relatively few variables, then the **canonical analysis** can be useful.

It is helpful first to transform the model into a new coordinate system with the origin at the stationary point x_s and then to rotate the axes of this system until they are parallel to the principal axes of the fitted response surface. This transformation is illustrated in the following figure:



Canonical form of the second order model

This results in the following fitted model

$$\hat{y} = \hat{y}_s + \lambda_1 w_1^2 + \lambda_2 w_2^2 + ... + \lambda_k w_k^2$$

where the w_i's are the transformed independent variables and the λ_i's are constants. This equation is called the **canonical form of the model** and λ_i's are the **eigenvalues or characteristic roots of the matrix** *B*.

The nature of the response surface can be determined from the stationary point and the **signs and magnitudes** of the λ_i's. First suppose that the stationary point is within the region of exploration for fitting the second-order model.

If all $\lambda_i > 0 \Rightarrow x_s$ is a point of minimum response;

If all $\lambda_i < 0 \Rightarrow x_s$ is a point of maximum response; and

It λ_i's have different signs, then x_s is a saddle point.

Furthermore, the surface is steepest in the w_i direction for which λ_i is the greatest.

Experimental designs for fitting response surfaces

Fitting and analyzing response surfaces is greatly facilitated by the proper choice of an experimental design.

When selecting a response surface design, some of the features of a desirable design are as follows:

1. It provides a reasonable distribution of data points throughout the region of interest.
2. The model adequacy and the lack of fit can be checked.
3. It allows the experiments to be performed in blocks.
4. It allows to built the higher order designs sequentially.
5. It provides an internal estimate of error.
6. It provides the precise estimates of the model coefficients.
7. It provides a good profile of the prediction variance throughout the experimental region.
8. It provides reasonable robustness against outliers or missing values.
9. It does not require a large number of runs.
10. It does not require too many levels of the independent variables.
11. It ensures simplicity of calculation of the model parameters.

All these features may not always be meeting in a design, so judgment based on experience must often be applied in design selection.

Design for fitting the first-order model

Consider the following first-order model in k variables for fitting

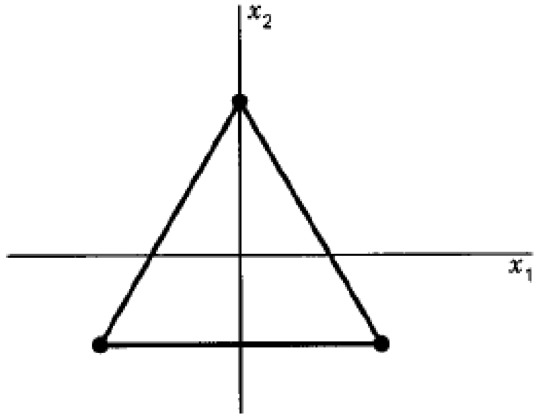
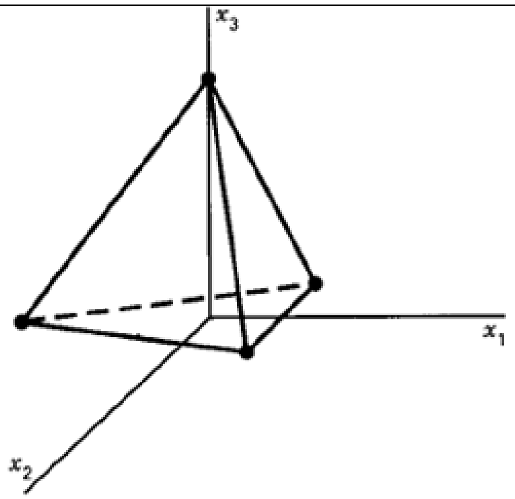
$$y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \varepsilon .$$

There is a unique class of designs that minimize the variance of the regression coefficients $\hat{\beta}_i$'s. These are the **orthogonal first–order designs**. A first-order design is orthogonal if the off-diagonal elements of the $(X'X)$ matrix are all zero. This implies that the cross-products of the columns of the X matrix sum to zero.

The 2^k factorial and fractions of the 2^k series in which main effects are not aliased with each other belongs to the class of orthogonal first-order designs. Assume that the low and high level of the k factors are coded as ± 1 levels to use is such designs.

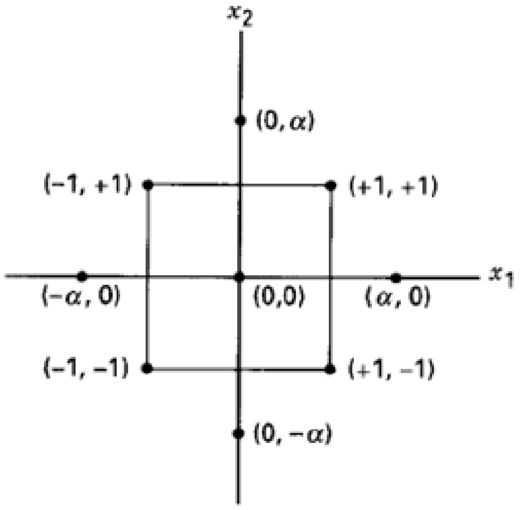
The 2^k design can not provide an estimate of the experimental error unless some runs are replicated. A method of including replication in the 2^k design is to augment the design with several observations at the center which is the point $x_i = 0, i = 1, 2, ..., k.$ The estimates of $\hat{\beta}_i$'s, $i \geq 1$ are not affected by adding the center points to the 2^k design. Only estimate of β_0 changes as it becomes the average of all the observations. The addition of center points does not alter the orthogonally property of the design.

Another orthogonal first-order design is the **simplex.** The simplex is a regularly sides figure with $k+1$ vertices in k dimensions. Thus, for $k = 2$ the simplex design is an equilateral triangle and for $k = 3$ it is a regular tetrahedron. Simplex designs in two and three dimensions are shown in the following figure:

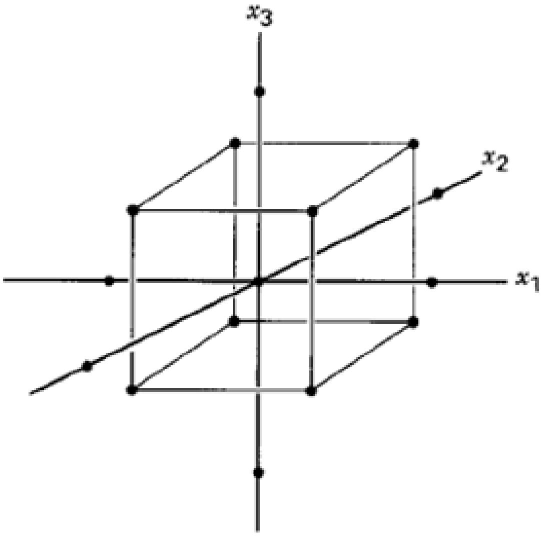
	
The simplex design for $k = 2$ variables	The simplex design for $k = 3$ variables

Designs for fitting the second-order model

The **central composite design** or **CCD** are used for fitting a second-order model. The CCD consists of a 2^k factorial with n_F runs, $2k$ axial or star runs , and n_c center runs. Following figure shows the CCD for $k=2$ and $k=3$ factors.



The central composite design for *k* = 2 variables



The central composite design for *k* = 3 variables

The CCD is developed through **sequential experimentation**. Suppose a 2^k is used to fit a first-order model and suppose this model exhibits lack of fit. Then axial runs is added to allow the quadratic terms to be incorporated into the model. The CCD is a very efficient design for fitting the second-order model. There are two parameters in the design that must be specified:

- the distance α of the axial runs from the design center and
- the number of center points n_c.

We now discuss the choice of these two parameters.

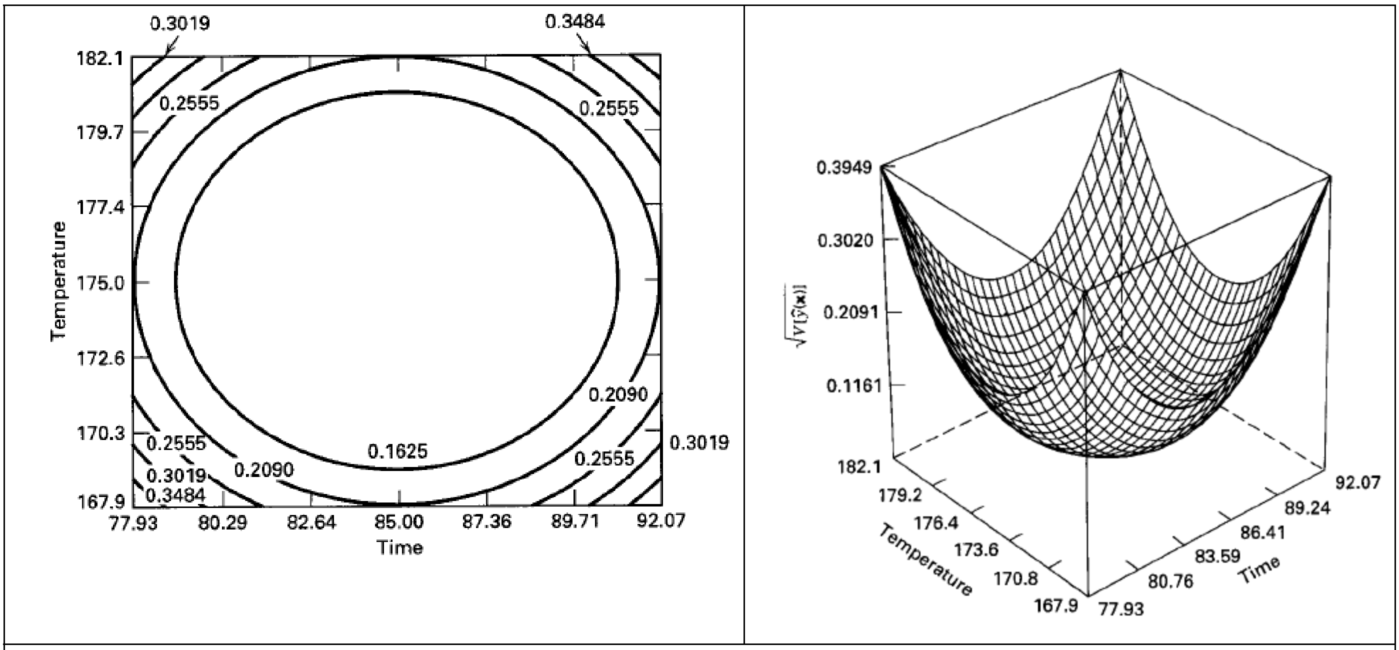
Readability

It is important for the second-order model to provide good predictions throughout the region of interest. One way to define “good” is to have the model which is a reasonably consistent and has stable variance of the predicted response at points of interest. The variance of the predicted response at some point x is

$$Var\left[\hat{y}(x)\right]=\sigma^2 x'(X'X)^{-1}x.$$

It is suggested that a second-order response surface design should be **rotatable**. This means that $Var\left[\hat{y}(x)\right]$ is the same at all points x that are at the same distance from the design center. This means that the variance of predicted response is constant on spheres.

Following figure shows contours of constant $\sqrt{Var\left[\hat{y}(x)\right]}$ for the second-order model fit using the CCD.



Contours of constant standard deviation $\sqrt{V\left[\hat{y}(x)\right]}$ of predicted response for the rotatble CCD

Notice that the contours of constant standard deviation of predicted response are concentric circles. A design with this property will leave the variance of $\hat{y}$ unchanged when the design is rotated about the center (0, 0,…,0). Hence it is termed as **rotatable design.**

Rotatability is an important criterion for the selection of a response surface design. The aim of RSM is optimization and the location of the optimum is unknown prior to running the experiment, so it makes sense to use a design that provides equal precision of estimation in all the directions. In fact, any first–order orthogonal design is rotatable.

A central composite design is made rotatable by the choice of α. The value of α for rotatability depends on the number of points in the factorial portion of the design. The choice $\alpha = \left(n_F\right)^{1/4}$ yields a rotatable central composite design where n_F is the number of points used in the factorial portion of the design.

The spherical CCD

Rotatability is a spherical property. It is an important design criterion when the region of interest is a sphere. It is not important to have the exact rotatability to have a good design. The best choice of α for a spherical region of interest from a prediction variance view point for the CCD is to set $\alpha = \sqrt{k}.$ This design called a **spherical CCD.** This puts all the factorial and axial design points on the surface of a sphere of radius $\sqrt{k}.$

Center runs in the CCD

The choice of α in the CCD is dictated primarily by the region of interest. When this region is a sphere, the design must include center runs to provide reasonably stable variance of predicted response. Generally, three to five center runs are recommended.

Blocking in response surface designs

When using the response surface designs, it is often necessary to consider blocking to eliminate nuisance variables. Such problem may occur when a higher order, say second-order design is assembled sequentially from lower order, say. Such necessity arises due to various reasons. For example, considerable time may elapse between the running of the first-order design and the running of the supplemental experiments which are required to build up a second-order design, and during this time, the test conditions may change which makes necessary to use blocking.

A response surface design is said to be **block orthogonally** if it is divided into blocks such that block effects do not affect the parameter estimates of the response surface model. If a 2^k or 2^{k-p} design is used as a first-order response surface design, the center points in these designs should be allocated among the blocks.

For a second-order design to block orthogonally, two conditions must be satisfied. If there are n_b observations in the b^{th} block, then these conditions are

1. Each block must be a first-order orthogonal design; that is,

   $$\sum_{u=1}^{n_b} x_{iu} x_{ju} = 0 \qquad i \neq j = 0,1,...,k \quad \text{for all } b$$

   where x_{iu} and x_{ju} are the levels of i^{th} and j^{th} variables in the u^{th} run of the experiment with $x_{0u} = 1$ for all u .

2. The fraction of the total sum of squares for each variable contributed by every block must be equal to the fraction of the total observations that occur in the block; that is,

$$\frac{\sum_{u=1}^{n_b} x_{iu}^2}{\sum_{u=1}^{N} x_{iu}^2} = \frac{n_b}{N} \quad i = 1, 2, \ldots, k \quad \text{for all } b$$

where N is the number of runs in the design.