
5 Studying Influences and Optimizing Analytical Procedures

According to Sect. 3.5 the influencing of an analytical signal by the environment of the experiment (influence factors), i.e. interferences and other stimuli can be estimated either in a way basing on chemical facts (Eq. 3.16a) or in a statistical way (Eq. 3.16b,c). Therefore, two ways are feasible to study the significance of influences as a requirement of a subsequent optimization.

5.1

Testing the Significance of Influencing Factors

In analytical chemistry, the optimality criterion is frequently the relative increase of that share of the analytical gross signal that is caused by the analyte itself, namely $S_{AA}x_A/y_A$. According to Eq. (3.16a):

$$y_A = y_{A0} + S_{AA}x_A + \sum_{i=1}^N S_{Ai}x_i + \sum_{j=1}^m I_{Aj}x_j + e_{A'} \quad (5.1)$$

which is tantamount to the minimization of the interferences, $\sum S_{Ai}x_i$, and influencing factors, $\sum I_{Aj}x_j$. The quantities S_{Ai} are the cross sensitivities of the interferents i and I_{Aj} , the specific strength of the influence factors j . On the basis of this relationship the significance of interferents and factors can be studied using so-called *multifactorial designs*.

On the other hand, Eqs. (3.16a) and (5.1), respectively, can also be interpreted as being composed of the analyte signal and diverse (more or less) anonymous deviations e_i, e_j or e_{Aij} , respectively, see Eq. (3.16b,c):

$$y_A = y_{A0} + S_{AA}x_A + e_i + e_j + e_{A'} \quad (5.2)$$

On this basis, an *analysis of variance* (ANOVA) can be carried out to test the significance of the variations $e_i = \sum S_{Ai}x_i$, $e_j = \sum I_{Aj}x_j$, or more in detail, $e_B = S_{AB}x_B$, $e_C = S_{AC}x_C$ etc.

5.1.1

Analysis of Variance (ANOVA)

ANOVA was developed by FISHER [1925, 1935] as a statistical procedure that investigates influences (effects) of factors on a target quantity y according to a linear model which holds in the simplest case

Table 5.1. Measurement and evaluation scheme of one-way ANOVA

		Levels i of the factor a			
		1	2	...	m
Number j of single measurement	1	y_{11}	y_{21}	...	y_{m1}
	2	y_{12}	y_{22}	...	y_{m2}
	\vdots	\vdots	\vdots		\vdots
	n	y_{1n}	y_{2n}	...	y_{mn}
Sum		$S_1 = \sum y_{1j}$	$S_2 = \sum y_{2j}$...	$S_m = \sum y_{mj}$
Mean		\bar{y}_1	\bar{y}_2	...	\bar{y}_m
Overall mean		$\bar{\bar{y}}$			

Table 5.2. Variance components in one-way ANOVA

Source of variation	Sum of squares	Degrees of freedom	Variance	F -test
Between the factor levels	$SS_a = n \sum_{i=1}^m (\bar{y}_i - \bar{\bar{y}})^2$	$\nu_a = m - 1$	$s_a^2 = \frac{SS_a}{m - 1}$	$\hat{F} = \frac{s_a^2}{s_{res}^2}$
Residual (analytical random error)	$SS_{res} = \sum_{i=1}^m \sum_{j=1}^n (y_{ij} - \bar{y}_i)^2$	$\nu_{res} = m(n - 1)$	$s_{res}^2 = \frac{SS_{res}}{m(n - 1)}$	
Total	$SS_{total} = SS_a + SS_{res}$	$\nu_{total} = m \cdot n - 1$		

$$y_{ij} = \bar{\bar{y}} + \alpha_i + e_{ij} \quad (5.3)$$

with y_{ij} being the actual value, $\bar{\bar{y}}$ the overall mean, α_i an additive influence of the factor a at level i , and e_{ij} the residual deviation (*one-way analysis of variance*, see Sect. 4.3.4). By means of ANOVA it is possible to compare both variances and means where two different models exist:

- *Model I* (model “fixed”) that compares means on the basis of the null hypothesis $H_0: \mu_1 = \mu_2 = \dots = \mu_m$ and
- *Model II* (model “random”) by which variances are compared on the basis of the null hypothesis $H_0: \sigma_1^2 = \sigma_2^2 = \dots = \sigma_m^2$ (corresponding to $H_0: \sigma_a^2 = 0$)

and additionally a mixed one (see Sect. 4.3.4 and EISENHART [1947]).

Variance analysis should advantageously be carried out on the basis of balanced experiments where the number of observations per factor level is equal ($n_1 = n_2 = \dots = n_m = n$).

The measurement scheme of *One-way analysis of variance* is given in Table 5.1 for $i = 1 \dots m$ levels of the factor a (in analytical practice frequently a factor is studied only on two levels to compare, e.g., two laboratories, two operators, two different techniques, etc).

The variance components are calculated according to Table 5.2.

The estimated \hat{F} value has to be compared with the quantile of the F -distribution, $F_{1-\alpha, \nu}$, the tables of which can be found in textbooks of statistics (e.g., HALD [1960]; NEAVE [1981]; DIXON and MASSEY [1983]; GRAF et al. [1987]; SACHS [1992]). The influence of the factor a is significant when \hat{F} exceeds $F_{1-\alpha, \nu}$. In case of unbalanced experiments the different size of measurement series and, therefore, degrees of freedom have to be considered as a result of which both the evaluation scheme and the variance decomposition become more complicated (see DIXON and MASSEY [1983]; GRAF et al. [1987]).

By means of *Two-way ANOVA* two factors can be studied simultaneously. The model

$$y_{ij} = \bar{\bar{y}} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ij} \quad (5.4)$$

considers the influences α_i of the factor a at levels i , β_j of the factor b at levels j , and additionally the interactions (correlations), $(\alpha\beta)_{ij}$, of both the factors (y_{ij} actual value, $\bar{\bar{y}}$ overall mean, e_{ij} residual deviation, i.e., experimental error). The scheme of measurement and evaluation of two-way ANOVA is given in Table 5.3 and the corresponding variance decomposition in Table 5.4.

On the basis of two-way ANOVA two null hypotheses can be tested, namely

- $H_a: \alpha_1 = \alpha_2 = \dots = \alpha_m = 0$ by means of \hat{F}_a compared with $F_{1-\alpha, \nu_1=m-1, \nu_2=(m-1)(n-1)}$,
- $H_b: \beta_1 = \beta_2 = \dots = \beta_m = 0$ by means of \hat{F}_b compared with $F_{1-\alpha, \nu_1=n-1, \nu_2=(m-1)(n-1)}$.

Table 5.3. Evaluation scheme of two-way ANOVA (ab model: single measurements in each point; the index point marks that levels over which is actually added up or averaged, respectively)

	Levels i of the factor a				Sum	Mean
	1	2	...	m		
Levels j 1	y_{11}	y_{21}	...	y_{m1}	$S_{\bullet 1} = \sum y_{i1}$	$\bar{y}_{\bullet 1}$
of the 2	y_{12}	y_{22}	...	y_{m2}	$S_{\bullet 2} = \sum y_{i2}$	$\bar{y}_{\bullet 2}$
factor b \vdots	\vdots	\vdots		\vdots	\vdots	\vdots
n	y_{1n}	y_{2n}	...	y_{mn}	$S_{\bullet n} = \sum y_{in}$	$\bar{y}_{\bullet n}$
Sum	$S_{1\bullet} = \sum y_{1j}$	$S_{2\bullet} = \sum y_{2j}$...	$S_{m\bullet} = \sum y_{mj}$	$S_{\bullet\bullet}$	–
Mean	$\bar{y}_{1\bullet}$	$\bar{y}_{2\bullet}$...	$\bar{y}_{m\bullet}$	–	$\bar{y}_{\bullet\bullet}$

Table 5.4. Variance components in two-way ANOVA (*ab* model)

Source of variation	Sum of squares	Degrees of freedom	Variance	F-test
Between the levels of factor <i>a</i>	$SS_a = \sum_{i=1}^m \frac{S_{i\cdot}^2}{n} - \frac{S_{\cdot\cdot}^2}{m \cdot n}$	$\nu_a = m - 1$	$s_a^2 = \frac{SS_a}{m - 1}$	$\hat{F}_a = \frac{s_a^2}{s_{res}^2}$
Between the levels of factor <i>b</i>	$SS_b = \sum_{i=1}^n \frac{S_{\cdot i}^2}{m} - \frac{S_{\cdot\cdot}^2}{m \cdot n}$	$\nu_b = n - 1$	$s_b^2 = \frac{SS_b}{n - 1}$	$\hat{F}_b = \frac{s_b^2}{s_{res}^2}$
Total	$SS_{total} = \sum_{i=1}^m \sum_{j=1}^n y_{ij}^2 - \frac{S_{\cdot\cdot}^2}{m \cdot n}$	$\nu_{total} = m \cdot n - 1$	$s_{total}^2 = \frac{SS_{total}}{m \cdot n - 1}$	
Residual (analytical random error)	$SS_{res} = SS_{total} - SS_a - SS_b$	$\nu_{res} = (m - 1)(n - 1)$	$s_{res}^2 = \frac{SS_{res}}{(m - 1)(n - 1)}$	

As a rule, interactions $(\alpha\beta)_{ij}$ cannot be estimated if only single observations at each point of the measurement matrix are carried out. In case of single measurements, the residual error contains both experimental error and interactions, the separation of which is possible only in special cases, e.g., testing of homogeneity of solids when certain assumptions can be made. DANZER and MARX [1979] have investigated the homogeneity of steel samples by means of a destructive OES procedure. Consequently, no repeated measurements could be carried out and the residual error must be corrected by interaction terms estimated according to MANDEL [1961].

If possible, two-way ANOVA should be applied doing repetitions at each level. In case of double measurements the *2ab* model represented in Tables 5.5 and 5.6 is taken as the basis of evaluation and variance decomposition.

On the basis of this *2ab* ANOVA it is possible to test three null hypotheses, namely

- $H_a: \alpha_1 = \alpha_2 = \dots = \alpha_m = 0$ by means of \hat{F}_a compared with $F_{1-\alpha, \nu_1=m-1, \nu_2=(m-1)(n-1)}$,
- $H_b: \beta_1 = \beta_2 = \dots = \beta_n = 0$ by means of \hat{F}_b compared with $F_{1-\alpha, \nu_1=n-1, \nu_2=(m-1)(n-1)}$, and
- $H_{ab}: (\alpha\beta)_{11} = (\alpha\beta)_{12} = \dots = (\alpha\beta)_{mm} = 0$ by means of \hat{F}_{ab} compared with $F_{1-\alpha, \nu_1=n-1, \nu_2=mn}$.

In the case that interactions prove to be insignificant, it should be gone over to the *ab* model the estimations of which for the various variance components is more reliable than that of the *2ab* model. A similar scheme can be used for three-way ANOVA when the factor *c* is varied at two levels. In the general, three-way analysis bases on block-designed experiments as shown in Fig. 5.1.

Table 5.5. Evaluation scheme of two-way ANOVA ($2ab$ model: double measurements in each point)

		Levels i of the factor a				Sum	Mean
		1	2	...	m		
Levels j of the factor b	1	y_{11}, y_{12}	y_{21}, y_{22}	...	y_{m1}, y_{m2}	$S_{\bullet 1\bullet}$	$y_{\bullet 1\bullet}$
	2	y_{121}, y_{122}	y_{221}, y_{222}	...	y_{m21}, y_{m22}	$S_{\bullet 2\bullet}$	$y_{\bullet 2\bullet}$
	\vdots	\vdots	\vdots		\vdots	\vdots	\vdots
	n	y_{1n1}, y_{1n2}	y_{2n1}, y_{2n2}	...	y_{mn1}, y_{mn2}	$S_{\bullet n\bullet}$	$y_{\bullet n\bullet}$
Sum		$S_{1\bullet\bullet}$	$S_{2\bullet\bullet}$...	$S_{m\bullet\bullet}$	$S_{\bullet\bullet\bullet}$	–
Mean		$y_{1\bullet\bullet}$	$y_{2\bullet\bullet}$...	$y_{m\bullet\bullet}$	–	$y_{\bullet\bullet\bullet}$

Table 5.6. Variance components in two-way ANOVA ($2ab$ model)

Source of variation	Sum of squares	Degrees of freedom	Variance	F -test
Between the levels of factor a	$SS_a = \sum_{i=1}^m \frac{S_{i\bullet\bullet}^2}{2n} - \frac{S_{\bullet\bullet\bullet}^2}{2mn}$	$\nu_a = m - 1$	$s_a^2 = \frac{SS_a}{m - 1}$	$\hat{F}_a = \frac{s_a^2}{s_{res}^2}$
Between the levels of factor b	$SS_b = \sum_{i=1}^n \frac{S_{\bullet i\bullet}^2}{2m} - \frac{S_{\bullet\bullet\bullet}^2}{2mn}$	$\nu_b = n - 1$	$s_b^2 = \frac{SS_b}{n - 1}$	$\hat{F}_b = \frac{s_b^2}{s_{res}^2}$
Interaction between a and b	$SS_{ab} = SS_{total} - SS_a - SS_b - SS_{res}$	$\nu_{ab} = (m - 1)(n - 1)$	$s_{ab}^2 = \frac{SS_{ab}}{(m - 1)(n - 1)}$	$\hat{F}_{ab} = \frac{s_{ab}^2}{s_{res}^2}$
Residual (analytical random error)	$SS_{res} = \sum_{i=1}^m \sum_{j=1}^n \sum_{k=1}^2 y_{ijk}^2 - \sum_{i=1}^m \sum_{j=1}^n S_{ij\bullet}^2$	$m \cdot n$	$s_{res}^2 = \frac{SS_{res}}{m \cdot n}$	
Total	$SS_{total} = \sum_{i=1}^m \sum_{j=1}^n y_{ij}^2 - \frac{S_{\bullet\bullet\bullet}^2}{2mn}$	$\nu_{total} = 2m \cdot n - 1$	$s_{total}^2 = \frac{SS_{total}}{2mn - 1}$	

Following the scheme given in Fig. 5.1, the influence of three factors a , b and c can be studied on the basis of the linear model

$$y_{ijk} = \bar{\bar{y}} + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + e_{ijk} \quad (5.5)$$

The estimation of all the terms of Eq. (5.5) is possible for the balanced case and q repeated measurements in each cell of the data block represented in Fig. 5.1. Schemes for this and some reduced variants of three-way ANOVA are given in SCHEFFÉ [1961]; AHRENS [1967]; DUNN and CLARK [1974]; GRAF et al. [1987]; SACHS [1992].

By means of the three-way variance analysis according to the model at Eq. (5.5) and Table 5.7 the influence of the factors a , b and c can be tested as well as that of the interactions, i.e. the null hypotheses:

Table 5.7. Variance components in three-way ANOVA (*abc* model with repetitions)

Source of variation	Sum of squares	Degrees of freedom	Variance	F-test
Between the levels of factor <i>a</i>	$SS_a = npq \sum_{i=1}^m (y_{i\bullet\bullet} - y_{\bullet\bullet\bullet})^2$	$m - 1$	$s_a^2 = \frac{SS_a}{m - 1}$	$\hat{F}_a = \frac{s_a^2}{s_{res}^2}$
Between the levels of factor <i>b</i>	$SS_b = mpq \sum_{j=1}^n (y_{j\bullet\bullet} - y_{\bullet\bullet\bullet})^2$	$n - 1$	$s_b^2 = \frac{SS_b}{n - 1}$	$\hat{F}_b = \frac{s_b^2}{s_{res}^2}$
Between the levels of factor <i>c</i>	$SS_c = mnq \sum_{k=1}^p (y_{\bullet k\bullet} - y_{\bullet\bullet\bullet})^2$	$p - 1$	$s_c^2 = \frac{SS_c}{p - 1}$	$\hat{F}_c = \frac{s_c^2}{s_{res}^2}$
Interaction between <i>a</i> and <i>b</i>	$SS_{ab} = pq \sum_{i=1}^m \sum_{j=1}^n (y_{ij\bullet} - y_{i\bullet\bullet} - y_{j\bullet\bullet} + y_{\bullet\bullet\bullet})^2$	$(m - 1)(n - 1)$	$s_{ab}^2 = \frac{SS_{ab}}{(m - 1)(n - 1)}$	$\hat{F}_{ab} = \frac{s_{ab}^2}{s_{res}^2}$
Interaction between <i>a</i> and <i>c</i>	$SS_{ac} = nq \sum_{i=1}^m \sum_{k=1}^p (y_{ik\bullet} - y_{i\bullet\bullet} - y_{\bullet k\bullet} + y_{\bullet\bullet\bullet})^2$	$(m - 1)(p - 1)$	$s_{ac}^2 = \frac{SS_{ac}}{(m - 1)(p - 1)}$	$\hat{F}_{ac} = \frac{s_{ac}^2}{s_{res}^2}$
Interaction between <i>b</i> and <i>c</i>	$SS_{bc} = mq \sum_{j=1}^n \sum_{k=1}^p (y_{jk\bullet} - y_{j\bullet\bullet} - y_{\bullet k\bullet} + y_{\bullet\bullet\bullet})^2$	$(n - 1)(p - 1)$	$s_{bc}^2 = \frac{SS_{bc}}{(n - 1)(p - 1)}$	$\hat{F}_{bc} = \frac{s_{bc}^2}{s_{res}^2}$
Interaction between <i>a</i> , <i>b</i> and <i>c</i>	$SS_{abc} = q \sum_{i=1}^m \sum_{j=1}^n \sum_{k=1}^p (y_{ijk\bullet} - y_{ij\bullet\bullet} - y_{j\bullet k\bullet} - y_{i\bullet k\bullet} + y_{\bullet\bullet k\bullet} + y_{\bullet\bullet\bullet\bullet})^2$	$v_{abc} = (m - 1)(n - 1)(p - 1)$	$s_{abc}^2 = \frac{SS_{abc}}{v_{abc}}$	$\hat{F}_{abc} = \frac{s_{abc}^2}{s_{res}^2}$
Residual error	$SS_{res} = \sum_{i=1}^m \sum_{j=1}^n \sum_{k=1}^p \sum_{l=1}^q (y_{ijkl} - y_{ijk\bullet})^2$	$mnp(q - 1)$	$s_{res}^2 = \frac{SS_{res}}{mnp(q - 1)}$	
Total	$SS_{total} = \sum_{i=1}^m \sum_{j=1}^n \sum_{k=1}^p \sum_{l=1}^q (y_{ijkl} - y_{\bullet\bullet\bullet\bullet})^2$	$mnpq - 1$	$s_{total}^2 = \frac{SS_{total}}{mnpq - 1}$	

$y_{\bullet\bullet\bullet\bullet}$ mean over the *i* levels of factor *a*, $y_{\bullet\bullet j\bullet}$ mean over the *j* levels of factor *b*, etc.

$y_{ij\bullet\bullet}$ mean over the *i* levels of factor *a* and the *j* levels of factor *b*; $y_{\bullet\bullet k\bullet}$ mean over the *i* levels of factor *a* and the *k* levels of factor *c*, etc.

$y_{ijk\bullet}$ mean over the *i* levels of factor *a*, the *j* levels of factor *b*, and the *k* levels of factor *c*, etc.

$y_{\bullet\bullet\bullet\bullet}$ total mean

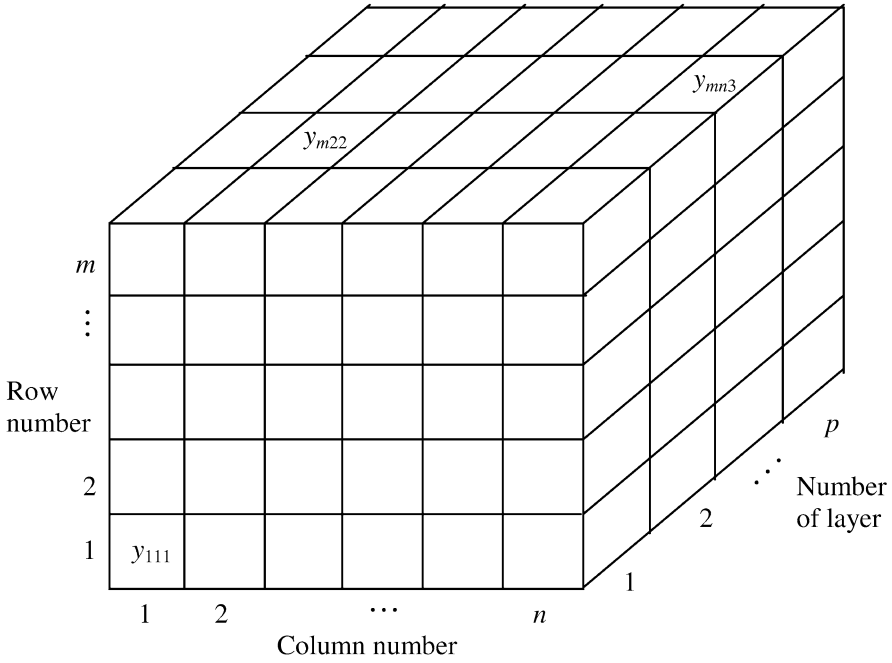


Fig. 5.1. Principle of three-way ANOVA: data are arranged in rows, columns, and layers (some examples of data pixels are given)

- $H_a: \alpha_1 = \alpha_2 = \dots = \alpha_m = 0$ by means of $\hat{F}_a \xleftrightarrow{\text{comp}} F_{1-\alpha, v_1=m-1, v_2=mnp-1}$
- $H_b: \beta_1 = \beta_2 = \dots = \beta_m = 0$ by means of $\hat{F}_b \xleftrightarrow{\text{comp}} F_{1-\alpha, v_1=n-1, v_2=mnp-1}$
- $H_c: \gamma_1 = \gamma_2 = \dots = \gamma_m = 0$ by means of $\hat{F}_c \xleftrightarrow{\text{comp}} F_{1-\alpha, v_1=p-1, v_2=mnp-1}$
- $H_{ab}: (\alpha\beta)_{11} = (\alpha\beta)_{12} = \dots = (\alpha\beta)_{mn} = 0$ by means of $\hat{F}_{ab} \xleftrightarrow{\text{comp}} F_{1-\alpha, v_1=(m-1)(n-1), v_2=mnp-1}$
- $H_{ac}: (\alpha\gamma)_{11} = (\alpha\gamma)_{12} = \dots = (\alpha\gamma)_{mp} = 0$ by means of $\hat{F}_{ac} \xleftrightarrow{\text{comp}} F_{1-\alpha, v_1=(m-1)(p-1), v_2=mnp-1}$
- $H_{bc}: (\beta\gamma)_{11} = (\beta\gamma)_{12} = \dots = (\beta\gamma)_{np} = 0$ by means of $\hat{F}_{ab} \xleftrightarrow{\text{comp}} F_{1-\alpha, v_1=(n-1)(p-1), v_2=mnp-1}$, and
- $H_{abc}: (\alpha\beta\gamma)_{111} = (\alpha\beta\gamma)_{121} = \dots = (\alpha\beta\gamma)_{mnp} = 0$ by $\hat{F}_{ab} \xleftrightarrow{\text{comp}} F_{1-\alpha, v_1=(m-1)(n-1)(p-1), v_2=mnp-1}$.

¹ $\xleftrightarrow{\text{comp}}$ symbolizes “compared with”

In some cases interactions are improbable and information on them is not needed. Then reduced variants of three-way ANOVA can be applied by which the effects of the main factors can be estimated more reliable (see DUNN and CLARK [1974]; GRAF et al. [1987]; SACHS [1992]). Concentrating on the main effects, the design of the experiments can be aimed at a minimum number of observations.

5.1.2

Experimental Design

Methods of variance analysis are helpful tools to evaluate effects of factors on the results of experiments afterwards. On the other hand, it may be advantageous to plan experiments in a comparative way (comparative experiments).

Statistical experimental design is characterized by the three basic principles: *Replication*, *Randomization* and *Blocking* (block division, planned grouping). *Latin square design* is especially useful to separate nonrandom variations from random effects which interfere with the former. An example may be the identification of (slightly) different samples, e.g. sorts of wine, by various testers and at several days. To separate the day-to-day and/or tester-to-tester (laboratory-to-laboratory) variations from that of the wine sorts, an $m \times m$ Latin square design may be used. In case of $m = 3$ all three wine samples (a , b , c) are tested by three testers at three days, e.g. in the way represented in Table 5.8:

Table 5.8. Latin square design for $m = 3$

	Tester 1	Tester 2	Tester 3
1st day	a	b	c
2nd day	b	c	a
3rd day	c	a	b

The results of the experiments are evaluated by means of three-way ANOVA in its simplest form, $m = n = p$ and $q = 1$. The significance of the sample effect can principally be guaranteed also in the case that both testers and days have significant influence (SHARAF et al. [1986]).

In contrast to common statistical techniques, by modern experimental design influencing factors are studied *simultaneously* (*multifactorial design*, MFD). The aim of MFD consists in an arrangement of factors in such a way that their influences can be quantified, compared and separated from random variations.

Frequently the signal intensity of the analyte A is the target quantity, the influences on which are described by Eq. (3.16a). Handling all the influences (interferences and other factors) in the same way and holding x_A at any constant value so that $\alpha_0 = y_{A0} + S_{AA}x_A$, Eq. (3.16a) can be written

Table 5.9. Design matrix for three factors at two levels (+ and – stand for +1 and –1)

Run	(z_0)	z_1	z_2	z_3	z_1z_2	z_1z_3	z_2z_3	$z_1z_2z_3$	Target value
1	+	+	+	+	+	+	+	+	y_1
2	+	+	+	–	+	–	–	–	y_2
3	+	+	–	+	–	+	–	–	y_3
4	+	+	–	–	–	–	+	+	y_4
5	+	–	+	+	–	–	+	–	y_5
6	+	–	+	–	–	+	–	+	y_6
7	+	–	–	+	+	–	–	+	y_7
8	+	–	–	–	+	+	+	–	y_8

$$y_A = \alpha_0 + \alpha_1x_1 + \alpha_2x_2 + \alpha_3x_3 + \alpha_{12}x_1x_2 + \alpha_{13}x_1x_3 + \alpha_{23}x_2x_3 + \alpha_{123}x_1x_2x_3 \quad (5.6)$$

in case of three influence factors. From the various types of treatment and design, *two-level factorial design* is mostly applied. That means that the influence factors are varied between a higher and a lower level, x_{\max} and x_{\min} . Using *complete factorial design* (CFD) the number of experiments is $N = 2^m$ for m factors. In the case of $m = 3$ as given in Eq. (5.6), $N = 8$ experiments have to be carried out.

Expediently, factorial design is done on the basis of transformed factors z_i , calculated from the x_i by

$$z_i = \frac{x_i - \bar{x}}{\frac{1}{2}(x_{\max} - x_{\min})} \quad (5.7)$$

where $\bar{x} = \frac{1}{2}(x_{\max} + x_{\min})$ so that $z_{\max} = \frac{x_{\max} - x_{\min}}{x_{\max} - x_{\min}} = +1$, $z_{\min} = \frac{x_{\min} - x_{\max}}{x_{\max} - x_{\min}} = -1$, and $\bar{z} = 0$.

With the transformation at Eq. (5.7), Eq. (5.6) becomes

$$y = a_0(z_0) + a_1z_1 + a_2z_2 + a_3z_3 + a_{12}z_1z_2 + a_{13}z_1z_3 + a_{23}z_2z_3 + a_{123}z_1z_2z_3 \quad (5.8)^2$$

The coefficients a_i are estimated from the results of experiments carried out according to a design matrix such as Table 5.9 which shows a 2^3 plan matrix. The significance of the several factors are tested by comparing the coefficients with the experimental error, to be exact, by testing whether the confidence intervals Δa_i include 0 or not. The experimental error can be estimated by repeated measurements of each experiment or – as it is done frequently in a more effective way – by replications at the centre of the plan (so-called “zero replications”), see Fig. 5.2.

The coefficients are estimated according to

$$a_i = \frac{1}{N} z_i^T y \quad (5.9)$$

² the target value y_A is symbolized here y

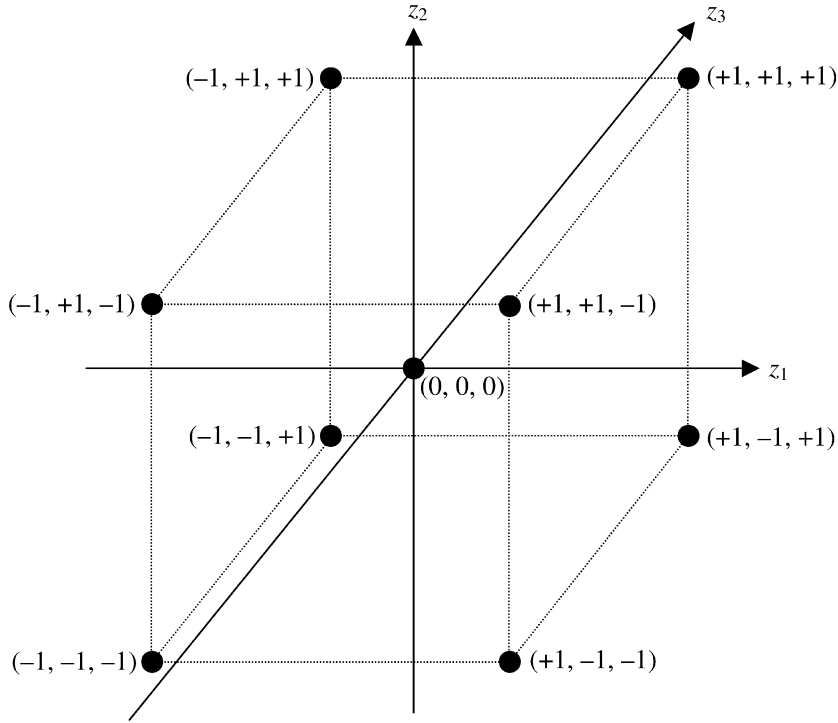


Fig. 5.2. Geometrical representation of a complete two level factorial matrix (three influence factors) with experiments in the centre point $(0,0,0)$

with \mathbf{z}_i^T the corresponding transposed \mathbf{z} vector and \mathbf{y} the vector of target values obtained as the result of the runs. N is the number of experiments ($N = 2^m$, here $N = 8$). As an example, the coefficients a_i in Eq. (5.8) corresponding to the design matrix at Eq. (5.9) are estimated by

$$\begin{aligned}
 a_0 &= \frac{1}{8} (+y_1 + y_2 + y_3 + y_4 + y_5 + y_6 + y_7 + y_8) \\
 a_1 &= \frac{1}{8} (+y_1 + y_2 + y_3 + y_4 - y_5 - y_6 - y_7 - y_8) \\
 a_2 &= \frac{1}{8} (+y_1 + y_2 - y_3 - y_4 + y_5 + y_6 - y_7 - y_8) \\
 a_3 &= \frac{1}{8} (+y_1 - y_2 + y_3 - y_4 + y_5 - y_6 + y_7 - y_8) \\
 a_{12} &= \frac{1}{8} (+y_1 + y_2 - y_3 - y_4 - y_5 - y_6 + y_7 + y_8) \\
 &\vdots \\
 a_{123} &= \frac{1}{8} (+y_1 - y_2 - y_3 + y_4 - y_5 + y_6 + y_7 - y_8)
 \end{aligned}$$

according to the vector expression

$$\mathbf{a} = \mathbf{Z}^T \mathbf{y} \quad (5.10)$$

with \mathbf{Z}^T being the transposed \mathbf{z} matrix.

Because the experimental expenditure increases strongly with the increasing number of influence factors, *fractional factorial design FFD* (*partial factorial design*) is applied in such cases. It is not possible to evaluate all the interactions by FFDs but only the main effects.

PLACKETT and BURMAN [1946] have developed a special fractional design which is widely applied in analytical optimization. By means of N runs up to $m = N - 1$ variables (where some of them may be dummy variables which can help to estimate the experimental error) can be studied under the following prerequisites and rules:

- The number of experiments (runs) must be a multiple of l^2 (l is the number of levels), that is $N = 8, 12, 16, 24, \dots$ in case of two-level experiments ($l = 2$).
- The first rows of the design matrixes are

$$N = 8: \quad + + + - + - -$$

$$N = 12: \quad + + - + + + - - + -$$

$$N = 16: \quad + + + + - + - + + - - + - - \text{ etc.}$$

in case of two-level design.

- The following rows of the design matrix is generated by shifting the first row cyclically one place ($N - 2$ times).
- The last row has minus in all factors.
- The procedure can be controlled as follows: each row contains $m/2$ times the higher level (+) and $(m/2 - 1)$ times the lower (-), the columns contain each $m/2$ times + and -.

Fractional factorial design is especially useful in case of a high number of influence variables from which the insignificant one have to be screened.

An example of a PLACKETT BURMAN plan for $l = 2$ levels, $m = 7$ influence factors (including dummy variables) and, therefore, $N = 8$ runs is given in Table 5.10.

The coefficients a_i of the main effects of the model

$$y = a_0 z_0 + a_1 z_1 + a_2 z_2 + a_3 z_3 + a_4 z_4 + a_5 z_5 + a_6 z_6 + a_7 z_7 \quad (5.11)$$

are obtained by the vector equation

$$\mathbf{a} = (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{y} \quad (5.12)$$

The coefficients characterize the effect of the belonging factor. The influence is significant in the case that

$$|a_i| \geq a_{crit} = \Delta a = s_a \cdot t_{1-\alpha, \nu} \quad (5.13)$$

Table 5.10. PLACKETT-BURMAN design matrix for $N = 8$ experiments and consequently $m = 7$ factors (including dummy variables) at two levels

Run	z_1	z_2	z_3	z_4	z_5	z_6	z_7	Target value
1	+	+	+	−	+	−	−	y_1
2	+	+	−	+	−	−	+	y_2
3	+	−	+	−	−	+	+	y_3
4	−	+	−	−	+	+	+	y_4
5	+	−	−	+	+	+	−	y_5
6	−	−	+	+	+	−	+	y_6
7	−	+	+	+	−	+	−	y_7
8	−	−	−	−	−	−	−	y_8

i.e., the influence of z_i (and therefore x_i) is insignificant if the confidence interval $a_i \pm \Delta a$ includes zero. Multifactorial experiments with a low number of factors can also be evaluated by ANOVA (see DANZER et al. [2001], Sect. 5.1.1).

If nonlinear effects are expected the variables must be varied at more than two levels. A screening plan comparable to the PLACKETT BURMAN design but on three levels is that of BOX and BEHNKEN [1960].

In case of special conditions, viz. internal correlations, interactions can be estimated in addition to the main effects by means of a 2^{m-1} design.

Multifactorial experiments are used in analytical chemistry for diverse applications, e.g., checking up *significant influences* before optimization procedures, recognizing *matrix effects*, and testing the *robustness* of analytical procedures (WEGSCHEIDER [1996]).

5.2

Optimization of Analytical Procedures

Analytical procedures should always run under optimum conditions. That means that for Eq. (5.6), which is here used only with two factors, the coefficients have to be chosen in such a way that y becomes an optimum

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 \stackrel{!}{=} \text{opt} \quad (5.14)$$

In analytical chemistry the target quantity y which has to be optimized is frequently the signal intensity, absolute or relative (signal-to-noise ratio), but occasionally other parameters like yields of extractions or chemical reactions, too. The classical way to optimize influences, e.g., in an optimization space as shown in Fig. 5.3a is to study the factors independently one after the other. In Fig. 5.3b,c it can be seen that an individual optimum will be found in this way.

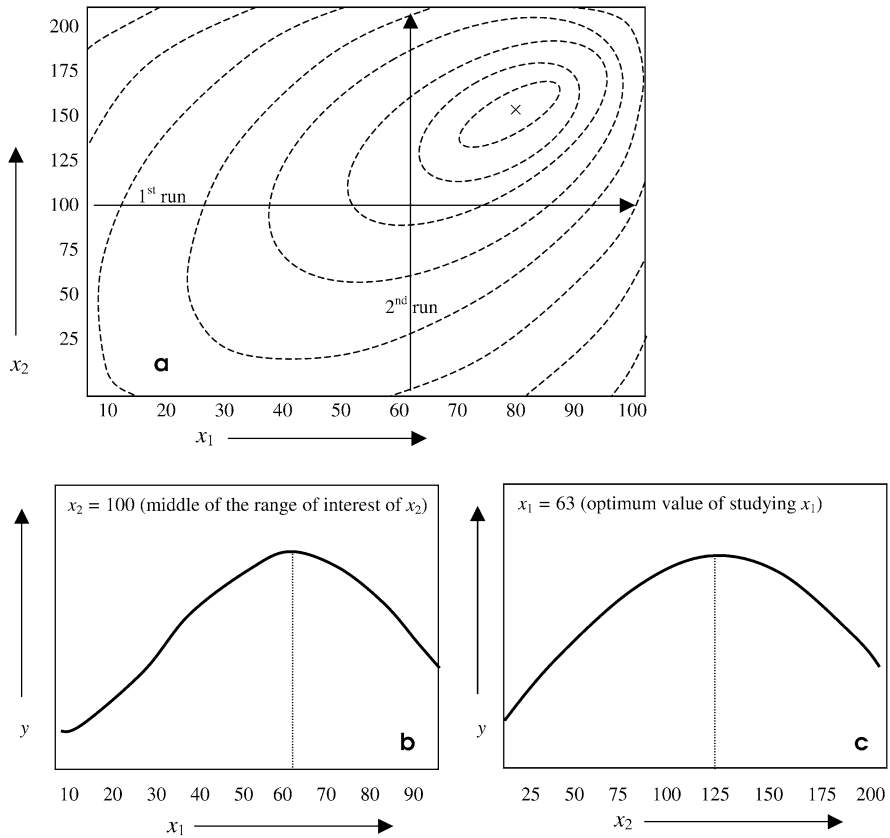


Fig. 5.3. Contour plot of the response surface $y = f(x_1, x_2)$; the optimum is situated at point \times (a); Response curves of y in dependence of x_1 as result of a first run (b) and x_2 as result of a second run (c)

However, the optima of x_1 and x_2 found in this way do not meet the global optimum of the response surface which is situated at $x_1 = 80$ and $x_2 = 150$. Because the global optimum is rarely found by such an obsolete proceeding, multivariate techniques of optimization should be applied.

The most reliable technique to find the global optimum by means of common methods is the transition from the quasi-two-dimensional approach (Fig. 5.3b,c) to a complete two-dimensional one. It consists of a certain number of experiments as shown in Fig. 5.4.

On the basis of the grid experiments a mathematical function $y = f(x_1, x_2, x_3, \dots)$, called the *response surface*, is estimated that characterizes the response as a function of the factors. In case of only two factors the response surface can be visualized by plots like that in Fig. 5.5.

Response surfaces are mostly described mathematically by polynomial approximations of 1st and 2nd degree. Grid search corresponds to a com-

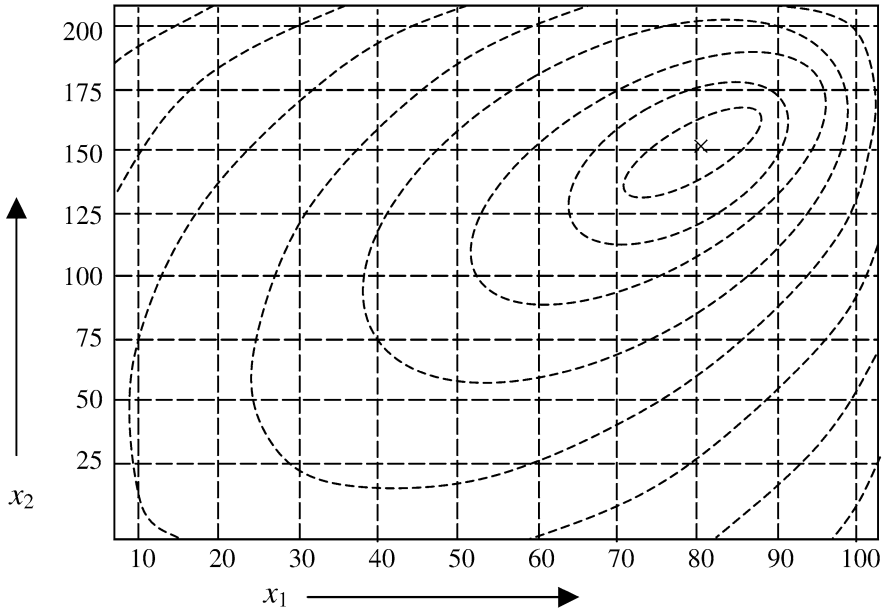


Fig. 5.4. Grid experiments for estimating the response surface $y = f(x_1, x_2)$

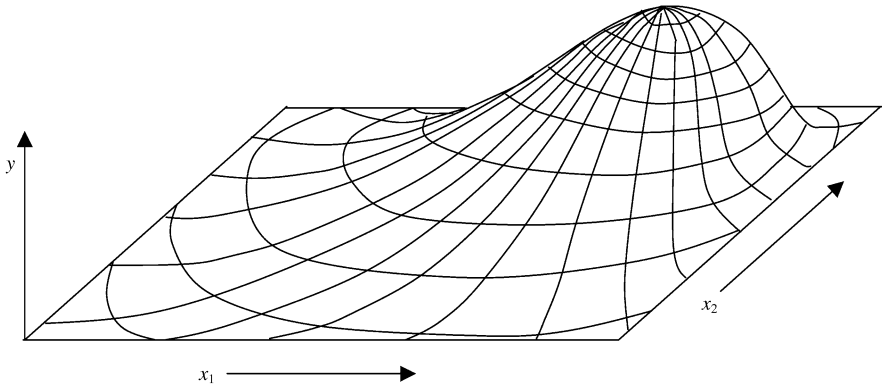


Fig. 5.5. Response surface as a result of grid experiments according to Fig. 5.4

plete factorial design. When the resolution of the variables is high enough, each optimum – the global one and all the local maxima and minima – can be found. But the high number of experiments imposes limits in this regard how it is generally in response surface technique.

The number of experiments can considerably be decreased by iterative optimization methods which starts at an area that can be selected by experience, supposition or randomly. This start area is moved step by

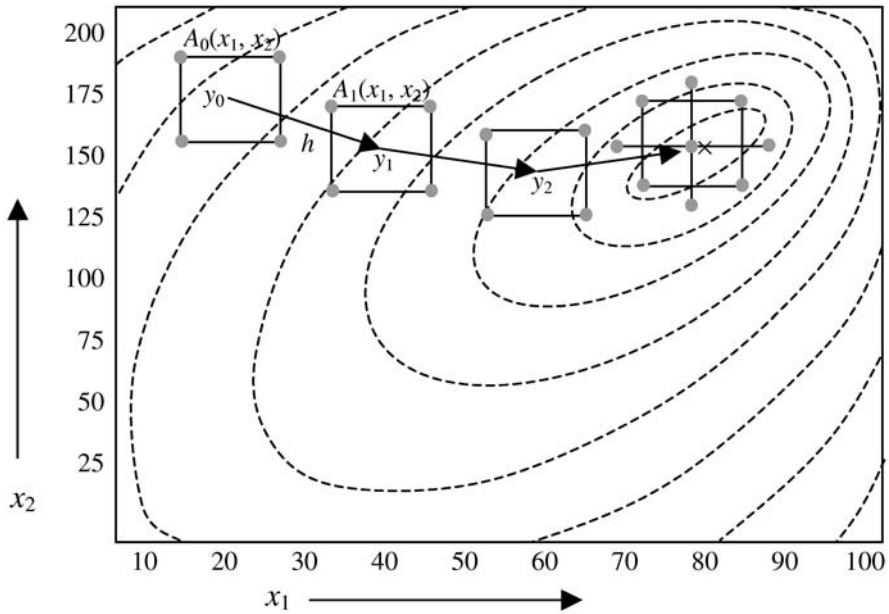


Fig. 5.6. Schematic representation of the Box-WILSON optimization with step width h

step in direction to the optimum following the gradient of the function $y = f(x_1, x_2, x_3, \dots)$, i.e. the steepest ascent³.

The well-known Box-WILSON optimization method (BOX and WILSON [1951]; BOX [1954, 1957]; BOX and DRAPER [1969]) is based on a linear model (Fig. 5.6). For a selected start hyperplane, in the given case an area $A_0(x_1, x_2)$, described by a polynomial of first order, with the starting point y_0 , the gradient $\text{grad}[y_0]$ is estimated. Then one moves to the next area in direction of the steepest ascent (the gradient) by a step width of h , in general

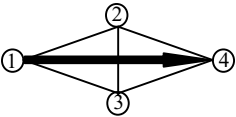
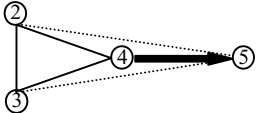
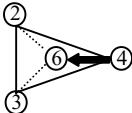
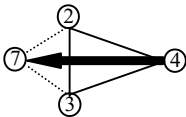
$$y_{i+1} = y_i + h \text{grad}[y_i] \quad (5.15)$$

Near the optimum both the step width and the model of the hyperplane are changed, the latter mostly from a first order model to a second order model. The vicinity of the optimum can be recognized by the coefficients a_1, a_2, \dots of Eq. (5.14) which approximate to zero or change their sign, respectively. For the second order model mostly a BOX-BEHNKEN design is used.

Because this proceeding is relatively expensive, an effective semi-quantitative method is widely used in optimization, the *sequential simplex optimization*. Simplex optimization is done without estimation of gradients and setting step widths. Instead of this, the progress of the optimization

³ steepest descent in case of minima

Table 5.11. Basic simplex operations

Operation	Movement	Condition
Reflexion		$y_1 < y_2 \approx y_3$
Expansion		$y_4 > y_2 \approx y_3$
Contraction		$y_4 \lesssim y_2 \approx y_3$
Strong contraction		$y_4 < y_2 \approx y_3$

procedure results directly from the quality of the preceded experimental values.

A simplex is a geometric figure formed by $p+1$ points in a p -dimensional space. In the two-dimensional case $y = f(x_1, x_2)$ the simplex is a triangle, in the three-dimensional case a tetrahedron etc. With regard to its form, the simplex may be regular, rectangular, or irregular. The simplex optimization starts with a set of $p + 1$ parameters (here $p + 1 = 3$). The movement of the simplex takes place according to the rules given in Table 5.11.

As an example, in Fig. 5.7 a simplex optimization is shown in a simplified way, i.e., only by reflexions and with simplexes of invariable size. The approach to the optimum is indicated by rotation or oscillation of the simplex. Then contractions should be included into the operations.

The optimum found by sequential proceeding, both by BOX-WILSON and simplex technique, is that local optimum situated nearest the starting point. It must not inevitably be identical with the global optimum. Therefore, it may be useful to repeat the optimization procedure one or several times.

5.3
Global Optimization by Natural Design

Some natural processes and principles have stimulated researchers to develop algorithms that imitate concepts of nature and are, therefore, summarized under the name *natural computation* (KATEMAN [1990]; LUCASIUS [1994]). The most prominent methods are:

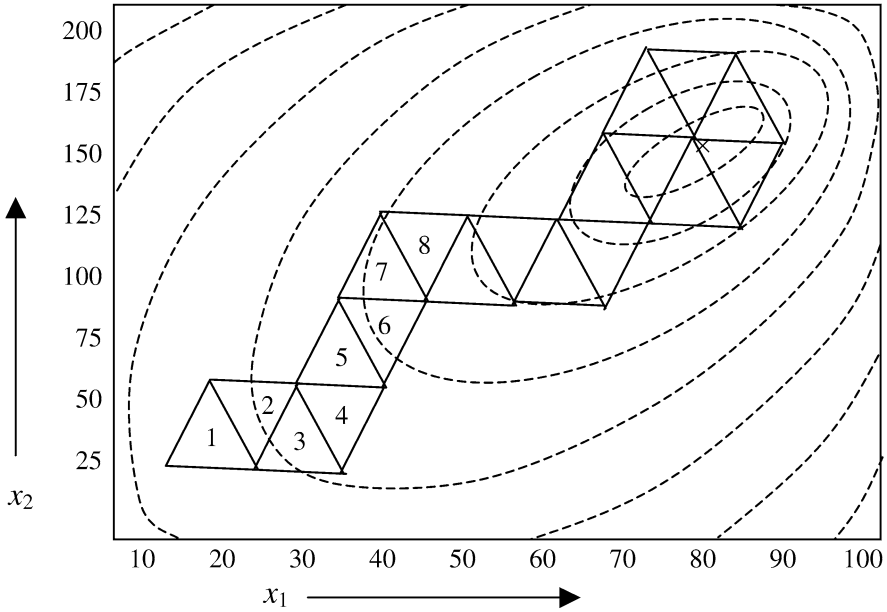


Fig. 5.7. Simplex optimization within a response surface of two factors x_1 and x_2 with simplexes of invariable size

- *Thermal computation* which comprises algorithms inspired by multi-particle systems like BROWNIAN motion (BROWNIAN search, Monte Carlo search) on the one hand and *simulated annealing* which is inspired by BOLTZMANN'S statistics on the other hand.
- *Evolutionary computation* which is learned by watching population dynamics; the most important programming are *genetic algorithms* which are inspired by the evolutionary processes of mutation, recombination, and natural selection in biology.
- *Connectionist computation* which is inspired by multi-cellular systems like artificial neural networks that mimic the way of working of the human brain.

Simulated annealing is a category of global optimization procedures the name of which is derived from the statistical mechanics simulating the atomic equilibrium according to the BOLTZMANN statistics. By means of simulated annealing it is searched for the most probable configuration of parameters on the basis of simulating the evolution of a substance to thermal equilibrium (KIRKPATRICK et al. [1983]; FRANK and TODESCHINI [1994]). The distribution of configurations \mathbf{x} can be described by the BOLTZMANN distribution

$$P(\mathbf{x}) = \frac{\exp(-C(\mathbf{x})/c)}{\sum_w \exp(-C(\mathbf{x}_w)/c)} \quad (5.16)$$

where $C(x)$ is the function to be optimized, x_w are other configurations and c is a control parameter. From the initial configuration x another configuration x_r in the neighborhood of x is generated by modifying one randomly selected variable. The new configuration is accepted when the difference $\Delta C(x_r, x) \leq 0$, otherwise the probability

$$P = \exp(-\Delta C(x_r, x)/c) \quad (5.17)$$

is compared with a random number generated from a uniform distribution $[0, 1]$. If P is larger than that random number, the new configuration is also accepted, otherwise it is declined. The iteration is continued until convergence is reached. Afterwards the optimization runs are continued with lowered control parameter c . More detailed information on simulated annealing can be found in VAN LAARHOVEN and AARTS [1987]; KALIVAS [1992].

Genetic Algorithms (GA) are the most important global optimization techniques. GA base on mimicking the evolution process by variation of populations according to the DARWIN rules [DARWIN 1859] such as *selection*, *reproduction*, *crossover (recombination)*, and *mutation*. Genetic algorithms have been pioneered by HOLLAND [1975], detailed representations can be found in GOLDBERG [1989]; DAVIS [1991], RECHENBERG [1973].

The initial data are binary coded in form of a bit sequence (*bit string*⁴). Start values of the variables x_1, x_2, \dots, x_m could be, e.g., 011001011, 100101100, ..., 010010101. This initial population is undertaken an evolution process such as schematically represented in Fig. 5.8.

In the course of each run corresponding to Fig. 5.8 the fitness of the members of the population is tested by means of an objective criterion (e.g., maximum correlation of a regression model or minimum random deviation of a response surface) that is compared with a break-off criterion fixed in advance. According to their fitness, the members from the present population (generation) are selected and reproduced by doubling. On the other hand, less fit members are omitted from the population. In the recombination step, parts of the bit-string are exchanged, namely by single-, two- or three-point-, uniform-, or circular crossover. In this way (by “mating”), from two parent bit strings two offspring are generated. Finally, by mutation only a small number of genes from the whole population is changed by flipping to the opposite value ($0 \rightarrow 1$ and $1 \rightarrow 0$, respectively).

For the selection and reproduction step the idea of “élitism” plays a role in so far as individuals of high quality should not become extinct. On the other hand, a larger number of élitists produces untimely a homogenisation of the population.

The advantage of Genetic Algorithms, in contrast to the traditional optimization methods, is the fact that a large number of variables can be included into the process. Also in the presence of local optima, GA can find rapidly the global optimum.

⁴ Also called “*chromosome*”; a bit in this chromosome is then called “*gen*”.

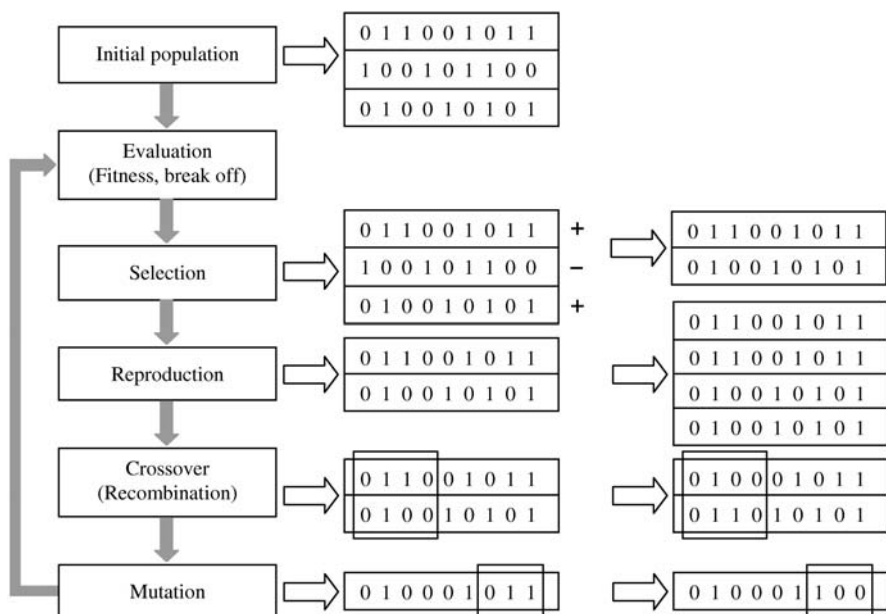


Fig. 5.8. Flow chart of a Genetic Algorithm

In many practical problems, interactions between the variables appear so that the absolute global optimum can be found heavily. As an example, wavelength selection in NIR determination of blood glucose (see Sect. 6.2.6) is considered. The aim of the selection is to find such combinations of wavelengths with which calibration models are obtained their prediction quality is as near at the global optimum as possible (DANZER et al. [2001], p 174). The number of combinations C for the selection of k wavelengths from n channels of the spectrometer is given by

$$C = \binom{n}{k} = \frac{n!}{k!(n-k)!} \quad (5.18)$$

So, for the selection of 15 wavelengths out of 86 the number of 2.1784×10^{16} combinations can be generated. Because of the existence of background, noise and strong multicollinearities in the NIR spectra, a large number of quasi-optimum solutions are available between which cannot be differentiated significantly. In Fig. 5.9 the wavelength selection 15 out of 86 is shown, carried out to improve the quality of the calibration model. Whereas the cross validated $PRESS$ value $s_{(cv)res}^2$ (see Eq. 6.105) is $295.6 \text{ mmol}^2/\text{L}^2$ when 86 equidistant wavelengths are used, $PRESS_{CV}$ improves to $147.0 \text{ mmol}^2/\text{L}^2$ for 15 GA-selected wavelengths.

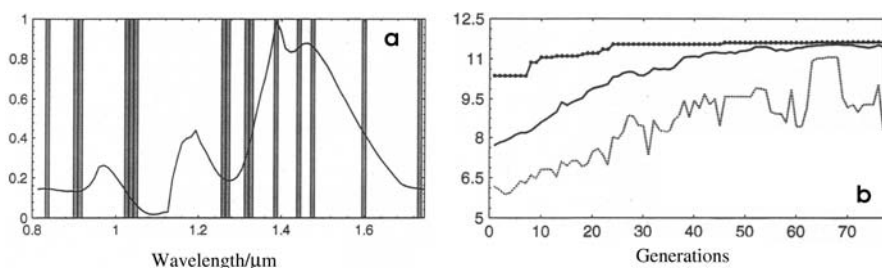


Fig. 5.9. Wavelength selection by Genetic Algorithm: **a** 15 optimum wavelength selected from 86 of the full spectrum; **b** relative fitness ($rf/10^4$) of the run in dependence of the number of generations, above: fitness of the best solution, middle: mean, below: fitness of the worst solution (FISCHBACHER et al. [1994/96;1995])

In connection with the NIR determination of blood glucose, GA has been used also to select spectra according to the quality of recordings (FISCHBACHER et al. [1994/96;1995]).

Optimization problems in general and, in particular in analytical chemistry, have been reported by GOLDBERG [1989]; DAVIS [1991]; LUCASIUS et al. [1991]; LUCASIUS and KATEMAN [1993]; WIENKE et al. [1992, 1993], and others.

In analytical chemistry, *Artificial Neural Networks* (ANN) are mostly used for calibration, see Sect. 6.5, and classification problems. On the other hand, feedback networks are usefully to apply for optimization problems, especially nets of HOPFIELD type (HOPFIELD [1982]; LEE and SHEU [1990]).

References

- Ahrens H (1967) *Varianzanalyse*. Akademie-Verlag, Berlin
- Box GPE (1954) *The exploration and exploitation of response surfaces: some general considerations and examples*. Biometrics 10:16
- Box GEP (1957) *Evolutionary operation: a method for increasing industrial productivity*. Appl Statist 6:81
- Box GEP, Behnken DW (1960) *Some new three level designs for the study of quantitative variables*. Technometrics 2:445
- Box GEP, Draper NR (1969) *Evolutionary operation*. Wiley, New York
- Box GPE, Wilson KB (1951) *On the experimental attainment of optimum conditions*. J R Statist Soc B 13:1
- Danzer K, Marx G (1979) *Application of the two-dimensional variance analysis for the investigation of homogeneity of solids*. Anal Chim Acta 110:145
- Danzer K, Hobert H, Fischbacher C, Jagemann K-U (2001) *Chemometrik – Grundlagen und Anwendungen*. Springer, Berlin Heidelberg New York
- Darwin C (1859) *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. John Murray, London

- Davis L (ed) (1991) *Handbook of genetic algorithms. Part I. A genetic algorithms tutorial*. Van Nostrand Reinhold, New York
- Dixon W, Massey FJ (1983) *Introduction to statistical analysis*. McGraw-Hill, New York
- Dunn OJ, Clark VA (1974) *Applied statistics - analysis of variance and regression*. Wiley, New York
- Eisenhart C (1947) *The assumptions underlying the analysis of variance*. Biometrics 3:1
- Fischbacher C, Jagemann K-U, Danzer K, Müller UA, Mertes B, Papenkorth L, Schüller J (1994/96) *Unpublished results*
- Fischbacher C, Jagemann K-U, Danzer K, Müller UA, Mertes B (1995) *Non-invasive blood glucose monitoring by chemometrical evaluation of NIR-spectra*. 24th EAS (Eastern Analytical Symposium), Somerset, NJ, November 12-16
- Fisher RA (1925) *Theory of statistical estimation*. Proc Cambridge Phil Soc 22:700
- Fisher RA (1935) *The design of experiments*. Oliver & Boyd, Edinburgh (7th edn 1960; 9th edn 1971, Hafner Press, New York)
- Frank IE, Todeschini R (1994) *The data analysis handbook*. Elsevier, Amsterdam
- Goldberg DE (1989) *Genetic algorithms in search, optimization and machine learning*. Addison-Wesley, New York
- Graf U, Henning H-U, Stange K, Wilrich P-Th (1987) *Formeln und Tabellen der angewandten mathematischen Statistik*. Springer, Berlin Heidelberg New York (3. Aufl)
- Hald A (1960) *Statistical tables and formulas*. Wiley, New York
- Holland JH (1975) *Adaption in natural and artificial systems*. University of Michigan Press, Ann Arbor, MI. Revised ed (1992) MIT Press, Cambridge, MA
- Hopfield JJ (1982) *Neural networks and physical systems with emergent collective computational abilities*. Proc Nat Acad Sci USA 79:2554
- Kalivas JH (1992) *Optimization using variations of simulated annealing*. Chemom Intell Lab Syst 15:1
- Kateman G (1990) *Evolutions in chemometrics*. Analyst 115:487
- Kirkpatrick S, Gelatt CD, Vecchi MP (1983) *Optimization by simulated annealing*. Science 220:671
- Lee BW, Sheu BJ (1990) *Combinatorial optimization using competitive Hopfield neural network*. Proc Internat Joint Conf Neural Networks II: 627, Washington, DC
- Lucasius, CB (1994) *Evolutionary computation: a distinctive form of natural computation with chemometric potential*. Chapter 9 in: Buydens LM, Melssen WJ: *Chemometrics. Exploring and exploiting chemical information*. Katholieke Universiteit Nijmegen
- Lucasius CB, Kateman G (1993) *Understanding and using genetic algorithms. Part 1. Concepts, properties and context*. Chemom Intell Lab Syst 19:1
- Lucasius CB, Blommers MJJ, Buydens LMC, Kateman G (1991) *A genetic algorithm for conformational analysis of DNA*. In: Davis L (ed) *Handbook of genetic algorithms*. Van Nostrand Reinhold, New York
- Mandel J (1961) *Non-additivity in two-way analysis of variance*. J Am Statist Assoc 56:878
- Neave HR (1981) *Elementary statistical tables*. Allen Unwin, London

- Plackett RL, Burman JP (1946) *The design of optimum multifactorial experiments*. Biometrika 33:305
- Rechenberg I (1973) *Evolutionsstrategie. Optimierung technischer Systeme nach Prinzipien der biologischen Evolution*. Frommann-Holzboog, Stuttgart
- Sachs L (1992) *Angewandte Statistik, 7th edn*. Springer, Berlin Heidelberg New York
- Scheffé H (1961) *The analysis of variance*. Wiley, New York
- Sharaf MA, Illman DL, Kowalski BR (1986) *Chemometrics*. Wiley, New York
- van Laarhoven PJM, Aarts EHL (1987) *Simulated annealing: theory and applications*. Reidel, Dordrecht
- Wegscheider W (1996) *Validation of analytical methods*. In: H. Günzler (ed) *Accreditation and quality assurance in analytical chemistry*. Springer, Berlin Heidelberg New York
- Wienke D, Lucasius C, Kateman G (1992) *Multicriteria target vector optimization of analytical procedures using a genetic algorithm. Part I. Theory, numerical simulations and application to atomic emission spectroscopy*. Anal Chim Acta 265:211
- Wienke D, Lucasius C, Ehrlich M, Kateman G (1993) *Multicriteria target vector optimization of analytical procedures using a genetic algorithm. Part II. Polyoptimization of the photometric calibration graph of dry glucose sensors for quantitative clinical analysis*. Anal Chim Acta 271:253

Analytical Chemistry

Theoretical and Metrological Fundamentals

Danzer, K.

2007, XXXII, 316 p., Hardcover

ISBN: 978-3-540-35988-3