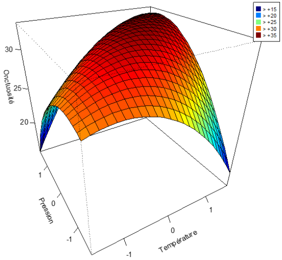
Plans pour surfaces de réponses

François Husson

UP de mathématiques appliquées
Agrocampus Ouest



Modèle de régression linéaire simple

Définition du modèle :

$$\begin{cases} \forall i = 1, ..., n \quad Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \\ \forall i = 1, ..., n \quad \varepsilon_i \ \text{ i.i.d. } \ , \ \mathbb{E}(\varepsilon_i) = 0, \ \ \mathbb{V}(\varepsilon_i) = \sigma^2 \\ \forall i \neq k \quad cov(\varepsilon_i, \varepsilon_k) = 0 \end{cases}$$

Estimation de β_0 et β_1 par moindres carrés :

$$\underset{(\hat{\beta}_0 \ , \ \hat{\beta}_1)}{\arg\min} \sum_{i=1}^{n} \Big(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \Big)^2$$

Dériver pour obtenir $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$ et $\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$

$$\mathbb{V}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2} = \frac{\sigma^2}{(n-1)\mathbb{V}(x)}$$

$\Rightarrow$ variance faible si n grand et si les x sont très dispersés

Modèle de régression linéaire multiple

Sous forme indicée :

$$\begin{cases} \forall i = 1, ..., n \quad & Y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \varepsilon_i \\ \forall i = 1, ..., n \quad & \varepsilon_i \ \text{ i.i.d.} \ , \quad \mathbb{E}(\varepsilon_i) = 0, \quad \mathbb{V}(\varepsilon_i) = \sigma^2 \\ \forall i \neq k \quad & cov(\varepsilon_i, \varepsilon_k) = 0 \end{cases}$$

Matriciellement :

$$Y = X\beta + E \quad \text{avec} \ \ \mathbb{E}(E) = 0, \ \ \mathbb{V}(E) = \sigma^2 Id$$

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ \vdots & \vdots & & \vdots & & \vdots \\ 1 & x_{i1} & & x_{ij} & & x_{ip} \\ \vdots & \vdots & & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_j \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Estimation des paramètres du modèle

Critère des moindres carrés

$$
\begin{aligned}
\hat{\beta} &= \arg\min_{\beta} \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip})\right)^2 \\
&= (X'X)^{-1}X'Y \quad \text{si } X'X \text{ est inversible}
\end{aligned}
$$

Propriétés

$$
\begin{aligned}
\mathbb{E}(\hat{\beta}) &= \beta \\
\mathbb{V}(\hat{\beta}) &= (X'X)^{-1}\sigma^2
\end{aligned}
$$

Prédiction

$$
\begin{aligned}
\hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_j x_{ij} + \ldots + \hat{\beta}_p x_{ip} \\
\mathbb{V}(Y_{x_o}) &= \sigma^2 \left(1 + x_0'(X'X)^{-1}x_0\right)
\end{aligned}
$$

Démarche en plan d'expériences

Facteurs :

- x_1 : température de cuisson (120^o à 140^o)
- x_2 : durée de cuisson (40 à 60 minutes)

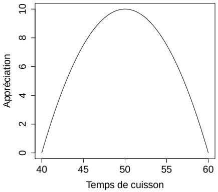
Variable d'intérêt Y : moelleux de pain de mie

- Quels sont les effets des facteurs x_1 et x_2 ? Quel est le rôle des variables dans la variation de la réponse ?
- Optimalité : y a-t-il des paramètres qui optimise la variable Y ? $\Rightarrow$ on veut une réponse avec le minimum d'incertitude

Modèle pour des surfaces de réponse

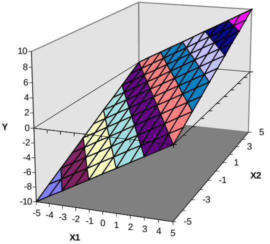
$$Y_i = \beta_0 + \underbrace{\beta_1 x_{i1} + \beta_2 x_{i2}+}_{\text{effets linéaires}} \underbrace{\beta_{11} x_{i1}^2 + \beta_{22} x_{i2}^2}_{\text{effets quadratiques}} \underbrace{+\beta_{12} x_{i1} x_{i2}+}_{\text{interaction}} \varepsilon_i$$

Effets quadratiques : très souvent présents en pratique

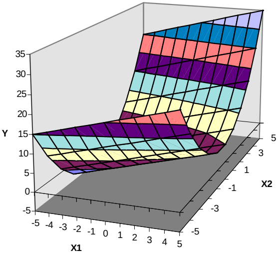


Interaction entre 2 variables quanti : l'effet d'une variable x_1 sur Y dépend d'une autre variable x_2

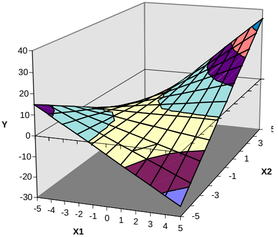
Surfaces de réponses pour deux facteurs x_1 et x_2



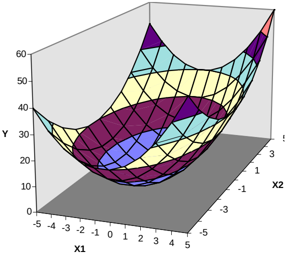
A) $y = x_1 + x_2$



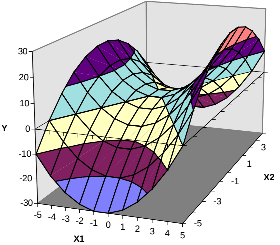
B) $y = x_1 + x_2 + x_2^2$



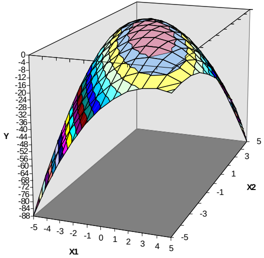
C) $y = x_1 + x_2 + x_1 x_2$



D) $y = x_1 + x_2 + x_1^2 + x_2^2$



E) $y = x_1 + x_2 + x_1^2 - x_2^2$



F) $y = 1.5 x_1 x_2 - x_1^2 - x_2^2$

Construction d'un plan continu

Problème : optimiser une recette de galette pour minimiser le nombre de galettes qui se déchirent (Y). 2 facteurs quantitatifs, la quantité de farine (entre 45 % et 55 %) et la température de cuisson (entre 180 et 220 degrés), étudiés selon un plan en 10 essais

Modifier les valeurs de F_1 et F_2 pour que la prévision de Y en tout point soit la plus précise possible

`https://husson.github.io/img/plan_CC.xlsx`

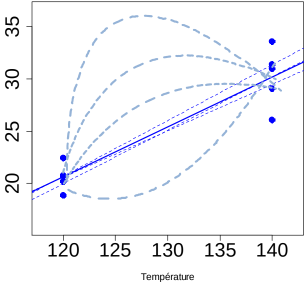
Qualité d'un plan

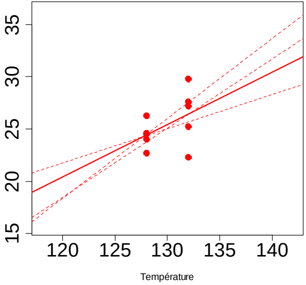
$$\mathbb{V}(\hat{\beta}) = (X'X)^{-1}\sigma^2$$

$\Rightarrow$ qualité du plan connue avant de faire les expériences

- essais au bord du domaine : maximiser la dispersion des x
- essais au centre : tester la linéarité
- orthogonalité entre facteurs : si 2 facteurs, $\mathbb{V}(\hat{\beta}_1) = \frac{\sigma^2}{n \times (1 - r_{12})\mathbb{V}(x_1)}$

Si $r_{12} = 0 \Rightarrow \mathbb{V}(\hat{\beta}_1) = \mathbb{V}(\hat{\beta}_1)^{(regsimple)}$ sinon $\mathbb{V}(\hat{\beta}_1) \nearrow$





Codage

$$x_{new} = \frac{x - (x_{max} + x_{min})/2}{(x_{max} - x_{min})/2} \implies x_{new} \in [-1, 1]$$

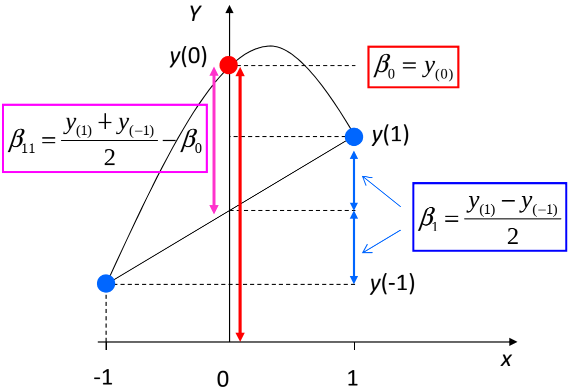
- permet de s'affranchir des unités
- plans faciles à construire (tables de plan)
- interprétation facile des coefficients du modèle

$$Y = \beta_0 + \beta_1 x + \beta_{11} x^2 \quad \begin{cases} Y_{(0)} = \beta_0 \\ Y_{(+1)} = \beta_0 + \beta_1 + \beta_{11} \\ Y_{(-1)} = \beta_0 - \beta_1 + \beta_{11} \end{cases}$$

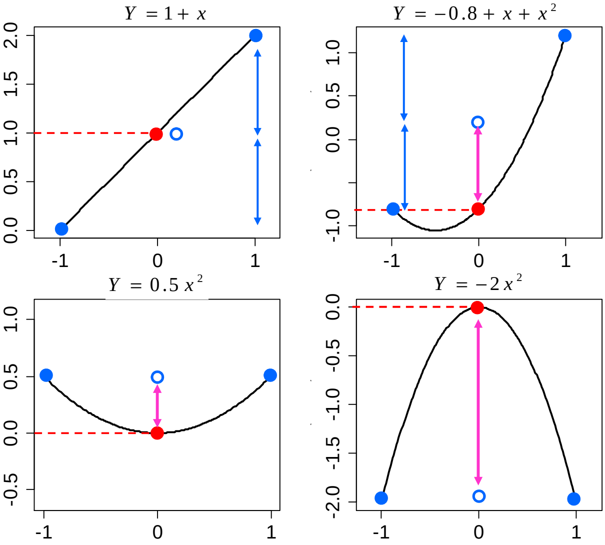
- β_0 : valeur de Y au centre du domaine
- β_1 : $Y_{(+1)} - Y_{(-1)} = 2\beta_1 \implies \beta_1 = \frac{Y_{(+1)} - Y_{(-1)}}{2}$
- β_{11} : $Y_{(+1)} + Y_{(-1)} = 2\beta_0 + 2\beta_{11} \implies \beta_{11} = \frac{Y_{(+1)} + Y_{(-1)}}{2} - \beta_0$

Interprétation des coefficients en régression quadratique

$$Y = \beta_0 + \sum_j \beta_j x_j + \sum_j \beta_{jj} x_j^2 + \sum_{j \neq k} \beta_{jk} x_j x_k + \varepsilon$$



Interpretation des coefficients en régression quadratique



$$\boxed{\beta_0 = y_{(0)}}$$

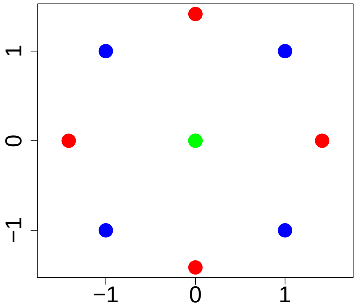
$$\boxed{\beta_{11} = \frac{y_{(1)} + y_{(-1)}}{2} - \beta_0}$$

$$\boxed{\beta_1 = \frac{y_{(1)} - y_{(-1)}}{2}}$$

Construction d'un plan composite centré à k facteurs

- Plan factoriel complet ou fractionnaire $n_f = 2^{k-p}$
- Points en étoile avec $\alpha = \sqrt[4]{n_f} = n_f^{1/4}$
- Points au centre

Nb d'expériences : $2^{k-p} + 2k + n_0$



Exemple avec 2 facteurs

$$\begin{bmatrix} 1 & 1 \\ 1 & -1 \\ -1 & 1 \\ -1 & -1 \\ \sqrt[4]{4} = \sqrt{2} & 0 \\ -\sqrt{2} & 0 \\ 0 & \sqrt{2} \\ 0 & -\sqrt{2} \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix}$$

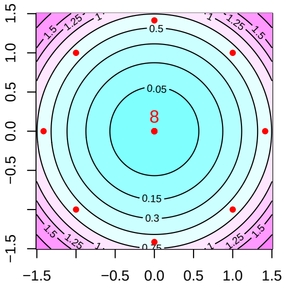
Plan composite centré avec le package rsm

```
> library(rsm)
> planccd <- ccd(2)   # donne le plan standard
> planccd<-ccd(2, coding=list (x1~(Temp-130)/10, x2~(Duree-50)/10))
> planccd
   run.order std.order     Temp      Tps Block
1          1         6 130.0000 50.00000     1
2          2         7 130.0000 50.00000     1
3          3         1 120.0000 40.00000     1
4          4         5 130.0000 50.00000     1
5          5         4 140.0000 60.00000     1
6          6         2 140.0000 40.00000     1
7          7         8 130.0000 50.00000     1
8          8         3 120.0000 60.00000     1
9          1         6 130.0000 50.00000     2
10         2         7 130.0000 50.00000     2
11         3         3 130.0000 35.85786     2
12         4         1 115.8579 50.00000     2
13         5         2 144.1421 50.00000     2
14         6         8 130.0000 50.00000     2
15         7         5 130.0000 50.00000     2
16         8         4 130.0000 64.14214     2
```

Ici, $n_0 = 8$ points au centre

Propriétés du plan composite centré

- Isovariance par rotation : (obtenue si $\alpha = n_f^{1/4}$) précision du plan dépend de la distance au centre, pas de la direction
- Précision uniforme : la précision est identique à la distance 1 dans tout le domaine (si bon nombre de points au centre)



Variance de la prévision

- Corrélation des effets : tous les effets sont orthogonaux mais il y a une corrélation entre effets quadratiques en fonction de n_0

En pratique :

- répartir les points au centre parmi toutes les expériences
- s'adapter à la réalité terrain : faire toutes les expériences à 140^o pour éviter de changer 15 fois la température du four

Nombre d'essais du PCC

Nombre de facteurs (k)	2	3	4	5	6
Plan factoriel complet ou fractionnaire	2^2	2^3	2^4	2^{5-1}	2^{6-1}
Nombre de points du plan factoriel : $n_f = 2^{k-p}$	4	8	16	16	32
Niveau codé des points axiaux : $\alpha = \sqrt[4]{n_f}$	1.414	1.682	2	2	2.378
Nombre de points axiaux : $n_\alpha = 2k$	4	6	8	10	12
Nombre de points au centre : n_0					
cas de l'orthogonalité	8	9	12	10	15
cas de la précision uniforme	5	6	7	6	9
Nombre total de points ($n_f + n_\alpha + n_0$)					
orthogonalité	16	23	36	36	59
précision uniforme	13	20	31	32	53

Vérification de la qualité du plan

La qualité d'un plan dépend des essais, du modèle et est mesurée par $(X'X)^{-1}$

```
> library(rsm)
> plan <- ccd(2)
> X <- model.matrix(~x1+x2+I(x1^2)+I(x2^2)+I(x1*x2),data=plan)
> t(X)%*%X
            (Intercept) x1 x2 I(x1^2) I(x2^2) I(x1 * x2)
(Intercept)          16  0  0       8       8          0
x1                    0  8  0       0       0          0
x2                    0  0  8       0       0          0
I(x1^2)               8  0  0      12       4          0
I(x2^2)               8  0  0       4      12          0
I(x1 * x2)            0  0  0       0       0          4
> solve(t(X)%*%X)
            (Intercept)    x1    x2 I(x1^2) I(x2^2) I(x1 * x2)
(Intercept)      0.1250 0.000 0.000 -0.0625 -0.0625       0.00
x1               0.0000 0.125 0.000  0.0000  0.0000       0.00
x2               0.0000 0.000 0.125  0.0000  0.0000       0.00
I(x1^2)         -0.0625 0.000 0.000  0.1250  0.0000       0.00
I(x2^2)         -0.0625 0.000 0.000  0.0000  0.1250       0.00
I(x1 * x2)       0.0000 0.000 0.000  0.0000  0.0000       0.25
```

Modèle de régression

$$Y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \sum_{j=1}^{k} \beta_{jj} x_{ij}^2 + \sum_{j=1}^{k} \sum_{l=j+1}^{k} \beta_{jl} x_{ij} x_{il} + \varepsilon_i$$

Décomposition de la variabilité :

- effets linéaires seuls
- effets quadratiques seuls
- interactions seules
- résiduelle
  qui se décompose en 2 termes (car n_0 vraies répétitions, pts au centre) :
  - erreur pure : variance des Y pour pts au centre ($n_0 - 1$ ddl) : estimation de la véritable répétabilité expérimentale
  - erreur d'ajustement : erreur résiduelle moins l'erreur pure ($ddl_{\text{ajustement}} = ddl_{\text{résiduelle}} - ddl_{\text{erreur pure}}$)

Modèle de régression : tests

- Tests des effets linéaires, quadratique ou des interactions

H_0 : pas d'effet d'une variable ou d'un groupe de variables
H_1 : effet de la variable ou du groupe de variables

$$F_{var} = \frac{CM_{var}}{CM_{residuelle}} \qquad \text{sous} H_0,\ \mathcal{L}(F_{var}) = F_{ddl_{var}}^{ddl_{residuelle}}$$

- Test d'ajustement du modèle :

H_0 : le modèle est bien ajusté
H_1 : les écarts au modèle ne peuvent pas s'expliquer uniquement par la variabilité résiduelle

$$F_{ajust} = \frac{CM_{ajust}}{CM_{pure}} \qquad \text{sous} H_0,\ \mathcal{L}(F_{ajust}) = F_{ddl_{pure}}^{ddl_{ajust}}$$

$\Rightarrow$ une erreur d'ajustement significative incite à changer de modèle (ajout d'effets quadratiques, etc.)

Plan composite centré avec le package rsm

Plan pour 2 facteurs :

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{11} x_{i1}^2 + \beta_{22} x_{i2}^2 + \beta_{12} x_{i1} x_{i2} + \varepsilon_i$$

```
> library(rsm)
> set.seed(1234)
> plan <- ccd(2, coding=list (x1~(Temp-130)/10, x2~(Duree-50)/10))
> Y <- c(1, 5, 4, 7, 8, 8, 4, 5, 2, 5, 4, 5, 5, 9, 7, 5)
> CR.rsm <- rsm(Y~SO(x1,x2),data=plan) ## SO pour 2nd order
> summary(CR.rsm)                      ##  FO(x1,x2)+TWI(x1,x2)+PQ(x1,x2)

Analysis of Variance Table
Response: Y
            Df Sum Sq Mean Sq F value    Pr(>F)
FO(x1, x2)   2 49.792 24.8958 67.1341   1.6e-06   ## effets linéaires
TWI(x1, x2)  1  9.000  9.0000 24.2694 0.0005991   ## interaction
PQ(x1, x2)   2  6.500  3.2500  8.7640 0.0063261   ## effets quadratiques
Residuals   10  3.708  0.3708
Lack of fit  3  1.833  0.6111  2.2815 0.1662512   ## erreur d'ajustement
Pure error   7  1.875  0.2679                     ## erreur pure

Multiple R-squared:  0.9463,    Adjusted R-squared:  0.9194
F-statistic: 35.21 on 5 and 10 DF,  p-value: 4.911e-06
```

Plan composite centré avec le package rsm

```
             Estimate Std. Error t value  Pr(>|t|)
(Intercept)  4.62500    0.21530 21.4815 1.066e-09 ***
x1           2.23744    0.21530 10.3921 1.116e-06 ***
x2           1.10355    0.21530  5.1256 0.0004470 ***
x1:x2       -1.50000    0.30448 -4.9264 0.0005991 ***
x1^2         0.50000    0.21530  2.3223 0.0426035 *
x2^2         0.75000    0.21530  3.4835 0.0058867 **
```

Recherche de l'optimum :

$$\begin{cases} \frac{\partial \hat{Y}}{\partial x_1} = 0 \\ \frac{\partial \hat{Y}}{\partial x_2} = 0 \end{cases} \qquad \begin{cases} 2.237 - 1.5 x_2 + 2 \times 0.5 \times x_1 = 0 \\ 1.104 - 1.5 x_1 + 2 \times 0.75 \times x_2 = 0 \end{cases}$$

$$x_2 = (2.237 + x_1)/1.5$$

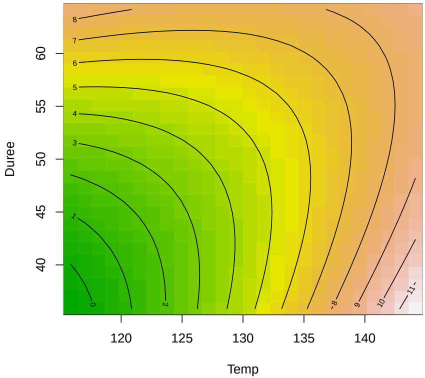
$$1.104 - 1.5 x_1 + 1.5 \times (2.237 + x_1)/1.5 = 0 \quad \Rightarrow \quad x_1 = 6.682 \Rightarrow x_2 = 5.946$$

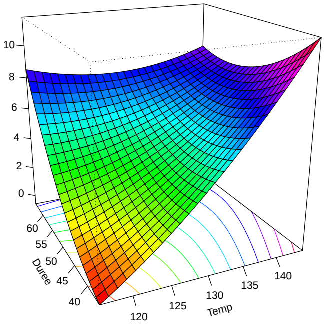
```
Stationary point of response surface:               ## optimum
      x1       x2
6.681981 5.946278

Eigenanalysis:                  ## vp ttes < 0 ==> point stationnaire = maximum
$values                         ## vp ttes > 0 ==> point stationnaire = minimum
[1]  1.3853453 -0.1353453   ## vp >0 et <0 ==> point stationnaire = point selle
```

Représentation des surfaces de réponse

```
> contour(CR.rsm,~x1+x2,image=TRUE)
> persp(CR.rsm,~x1+x2,col=rainbow(50), contours="colors")
```





Pb de visualisation avec 3 variables ou plus : tracer le graphe pour 2 variables les autres étant fixées à leur valeur centrale ou à l'optimum

Construction séquentielle du plan

1. construire le plan factoriel et les points au centre
2. à partir des points au centre, l'erreur pure permet de savoir si le travail réalisé est bon
3. les points au centre permettent de savoir si les effets sont linéaires ou non ; si non linéaires, ajouter les points en étoile
4. peut-on supposer que les effets quadratiques sont nuls ?

Plan de Box-Benhken

Mode de construction :

- construire un plan complet pour chaque couple de 2 facteurs, les autres facteurs étant à la moyenne
- ajouter des points au centre

Avantages :

- 3 niveaux par variable (vs 5 pour PCC)
- travail séquentiel possible : permet de rajouter des facteurs (fixés au niveau moyen avant)

```
> library(rsm)
> Benhken <- bbd(3)
```

Exemple avec 3 facteurs

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