# Chapitre 23

# Théorie de la corrélation. Corrélations multiple et partielle

Ce chapitre généralise la corrélation simple entre deux variables quantitatives à plusieurs variables quantitatives.

## 23.1 Généralisation de l'espérance et de la variance pour $n$ variables aléatoires

Pour $n$ variables aléatoires, l'espérance d'une somme algébrique de variables aléatoires vaut :

$$\mathbb{E}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathbb{E}\left(X_i\right) \tag{23.1}$$

Pour $n$ variables aléatoires, la variance d'une somme algébrique de variables aléatoires vaut :

$$\mathbb{V}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} {a_i}^2 \mathbb{V}\left(X_i\right) + 2 \sum_{\substack{i=1 \\ i \neq j}}^{n} a_i a_j \text{cov}\left(X_i, X_j\right) \tag{23.2}$$

Dans ce cadre, deux variables aléatoires sont indépendantes si $\mathbb{V}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} {a_i}^2 \mathbb{V}\left(X_i\right)$.

## 23.2 Corrélation multiple

On appelle **corrélation multiple** la corrélation qu'il existe entre trois ou plusieurs variables.

Les principes fondamentaux qui caractérisent la corrélation multiple sont les mêmes que ceux de la corrélation simple.

Par exemple, soit $Y = a_0 + a_1 X + a_2 X^2 + \dots$. Le coefficient de détermination $r^2$ correspond toujours au rapport entre la variance $Y$ expliquée par l'ensemble des régresseurs et la variance totale de $Y$. Toutefois, $r^2$ doit être corrigée s'il y a $k$ régresseurs :

$$r^2{}_{\text{corrigé}} = \frac{(n-1)\, r^2 - k}{n-k-1} = \langle r^2 \rangle \tag{23.3}$$

avec $\langle r^2 \rangle$ la moyenne des coefficients de détermination, d'où

$$\sigma_e{}^2 = \left(1 - r^2{}_{\text{corrigé}}\right) s_Y{}^2 \tag{23.4}$$

## 23.3 Notation indicée

Soient $X_1, X_2, X_3, \dots$, les variables considérées. On désignera par $X_{11}, X_{12}, X_{13}, \dots$, les valeurs de la variable $X_1$, par $X_{21}, X_{22}, X_{23}, \dots$, les valeurs de la variable $X_2$, *etc*. Avec cette notation, la somme $X_{21} + X_{22} + X_{23} + \dots + X_{2N}$ s'écrira sous la forme $\sum_{j=1}^{N} X_{2j}$, $\sum X_{2j}$ ou simplement $\sum X_2$. Lorsque il n'y a aucune ambiguïté, c'est la dernière notation que l'on considère. C'est ainsi que la moyenne de $X_2$ peut s'écrire $\langle X_2 \rangle = \frac{1}{N} \sum X_2$.

## 23.4 Équation de régression. Plan de régression

Une équation de régression est une équation qui permet d'estimer une variable, $X_1$ par exemple, en fonction de variables indépendantes $X_2$, $X_3$, *etc*. Cette équation est l'équation de régression de $X_1$ en fonction de $X_2$, $X_3$, … On écrira parfois $X_1 = F(X_2, X_3, \dots)$ que l'on lit « $X_1$ est une fonction de $X_2$, $X_3$, *etc*. ».

L'équation de régression la plus simple entre trois variables, $X_1$ en fonction de $X_2$ et $X_3$, par exemple, a la forme :

$$X_1 = b_{1.23} + b_{12.3} X_2 + b_{13.2} X_3 \tag{23.5}$$

dans laquelle $b_{1.23}$, $b_{12.3}$ et $b_{13.2}$ sont des constantes.

Si $X_3$ reste constante dans l'équation n° 23.5, le graphe de $X_1$ en fonction de $X_2$ est une droite de pente $b_{13.2}$.

Si $X_2$ reste constante dans l'équation n° 23.5, le graphe de $X_1$ en fonction de $X_3$ est une droite de pente $b_{12.3}$.

**Remarque.** Les indices suivant chaque point indiquent les variables qui sont constantes dans chaque cas.

$X_1$ variant partiellement du fait des variations de $X_2$ ou $X_3$, on dit que $b_{12.3}$ et $b_{13.2}$ sont les **coefficients de régression partielle** de $X_1$ en $X_2$, avec $X_3$ constant et de $X_1$ en $X_3$, avec $X_2$ constant.

On dit que l'équation n° 23.5 est l'**équation de la régression linéaire** de $X_1$ en $X_2$ et $X_3$. Dans un système de coordonnées rectangulaires à trois dimensions, elle représente un plan que l'on appelle **plan de régression** et qui est la généralisation de la droite de régression entre deux variables.

## 23.5 Équations normales du plan de régression des moindres carrés

De même qu'il existe des droites de régression des moindres carrés représentant un ensemble de $N$ points $(X, Y)$ du diagramme de dispersion à deux dimensions, il existe des **plans de régression des moindres carrés** ajustant un ensemble de $N$ points $(X_1, X_2, X_3)$ d'un diagramme de dispersion à trois dimensions.

Le plan de régression des moindres carrés de $X_1$ en $X_2$ et $X_3$ a la forme de l'équation n° 23.5 dans laquelle $b_{1.23}$, $b_{12.3}$ et $b_{13.2}$ sont déterminés par la résolution simultanée des **équations normales**.

$$\left\{\begin{array}{l} \sum X_1 = b_{1.23}N + b_{12.3}\sum X_2 + b_{13.2}\sum X_3 \\ \sum X_1X_2 = b_{1.23}\sum X_2 + b_{12.3}\sum {X_2}^2 + b_{13.2}\sum X_2X_3 \\ \sum X_1X_3 = b_{1.23}\sum X_3 + b_{12.3}\sum X_2X_3 + b_{13.2}\sum {X_3}^2 \end{array}\right. \tag{23.6}$$

Celles-ci s'obtiennent en multipliant successivement l'équation n° 23.5 par 1, $X_2$ et $X_3$, et en sommant membre à membre.

Sauf avis contraire, chaque fois que l'on se réfère à une équation de régression, on sous-entend qu'il s'agit de l'équation de régression des moindres carrés.

Si $x_1 = X_1 - \langle X_1\rangle$, $x_2 = X_2 - \langle X_2\rangle$ et $x_3 = X_3 - \langle X_3\rangle$, on écrit plus simplement l'équation de $X_1$ en $X_2$ et $X_3$ sous la forme :

$$x_1 = b_{1.23}x_2 + b_{13.2}x_3 \tag{23.7}$$

dans laquelle $b_{1.23}$ et $b_{13.2}$ s'obtiennent par résolution simultanée des équations :

$$\left\{\begin{array}{l} \sum x_1x_2 = b_{1.23}\sum {x_2}^2 + b_{13.2}\sum x_2x_3 \\ \sum x_1x_3 = b_{1.23}\sum x_2x_3 + b_{13.2}\sum {x_3}^2 \end{array}\right. \tag{23.8}$$

Ces équations, qui sont équivalentes aux équations normales, s'obtiennent en multipliant successivement l'équation n° 23.7 par $x_2$ et $x_3$ et en sommet membre à membre.

## 23.6 Plan de régression et coefficient de corrélation

Soient $r_{12}$, $r_{13}$, $r_{23}$ les coefficients de corrélation linéaire respectifs des variables $X_1$ et $X_2$, $X_1$ et $X_3$ et $X_2$ et $X_3$ que l'on appelle parfois **coefficients de corrélation d'ordre zéro**. L'équation du plan de régression des moindres carrés est alors

$$\frac{x_1}{s_1} = \left( \frac{r_{12} - r_{13}r_{23}}{1 - {r_{23}}^2} \right) \frac{x_2}{s_2} + \left( \frac{r_{13} - r_{12}r_{23}}{1 - {r_{23}}^2} \right) \frac{x_3}{s_3} \tag{23.9}$$

dans laquelle les variables ont été centrées réduites $x_1 = X_1 - \langle X_1 \rangle$, $x_2 = X_2 - \langle X_2 \rangle$ et $x_3 = X_3 - \langle X_3 \rangle$, et $s_1$, $s_2$ et $s_3$ sont respectivement les écarts types $X_1$, $X_2$ et $X_3$.

**Remarque.** Si la variable $X_3$ est absente, si $X_1 = Y$ et $X_2 = X$, l'équation n° 23.9 se réduit à l'équation à deux variables.

## 23.7 Erreur quadratique moyenne d'un estimateur (écart type lié)

Par la généralisation évidente de l'équation à deux variables, on peut définir l'écart type de $X_1$ lié par $X_2$ et $X_3$ par :

$$s_{1.23} = \sqrt{\frac{\sum \left( X_1 - \hat{X}_1 \right)}{N}} \tag{23.10}$$

pour lequel $\hat{X}_1$ représente les valeurs $X_1$ estimées à partir des équations de régression 1 ou 5.

L'écart type lié d'un estimateur peut aussi s'exprimer en fonction des coefficients de corrélation $r_{12}$, $r_{13}$ et $r_{23}$, à partir de

$$s_{1.23} = s_1 \sqrt{\frac{1 - {r_{12}}^2 - {r_{13}}^2 + 2r_{12}r_{13}r_{23}}{1 - {r_{23}}^2}} \tag{23.11}$$

Il est possible de généraliser à trois dimensions l'erreur quadratique de deux variables. Pour cela, il suffit de remplacer les droites parallèles à la droite de régression par des plans parallèles au plan de régression. On obtient une meilleure estimation de l'erreur quadratique moyenne d'une population par $\hat{s}_{1.23} = s_{1.23}\sqrt{\frac{N}{N-3}}$.

## 23.8 Le coefficient de corrélation multiple

On définit le **coefficient de corrélation multiple** de la même manière que le coefficient de corrélation simple. Le coefficient de corrélation multiple de deux variables est :

$$R_{1.23} = \sqrt{1 - \frac{{s_{1.23}}^2}{{s_1}^2}} \tag{23.12}$$

dans lequel $s_1$ est l'écart type de la variable $X_1$ et dans lequel $s_{1.23}$ s'obtient par les équations précédentes. La quantité ${R_{1.23}}^2$ est appelée **coefficient de détermination multiple**.

**Remarque.** Lorsqu'on utilise une équation de régression linéaire, le coefficient de corrélation multiple prend le nom de **coefficient de corrélation linéaire multiple**. Sauf avis contraire, chaque fois que l'on se réfère à la corrélation linéaire multiple, cela sous-entend qu'il s'agit de corrélation multiple linéaire.

En fonction de $r_{12}$, $r_{13}$ et $r_{23}$, l'équation précédente s'écrit :

$$R_{1.23} = \sqrt{\frac{{r_{12}}^2 + {r_{13}}^2 - 2r_{12}r_{13}r_{23}}{1 - {r_{23}}^2}} \tag{23.13}$$

Le coefficient de corrélation multiple varie entre 0 et 1. Plus il est proche de 1, plus la relation linéaire entre les variables est exacte. Plus il est proche de zéro, moins cette relation est linéaire. Lorsque le coefficient de corrélation multiple est égal à 1, on dit que la **corrélation** est **totale**.

**Remarque importante.** Bien qu'il n'y ait pas de relation linéaire entre les variables dont le coefficient de corrélation est nul, il est possible qu'il y ait une **relation non linéaire** entre elles.

## 23.9 Substitution d'une variable expliquée

Les résultats précédents sont valables lorsque $X_1$ est la variable expliquée. Si c'est $X_3$ que l'on prend comme variable expliquée, il faut simplement permuter

les indices 1 et 3 dans les formules déjà obtenues.

Par exemple, l'équation de régression de $X_3$ en $X_1$ et $X_2$ est :

$$\frac{x_3}{s_3} = \left(\frac{r_{23} - r_{13}r_{12}}{1 - {r_{12}}^2}\right)\frac{x_2}{s_2} + \left(\frac{r_{13} - r_{23}r_{12}}{1{r_{12}}^2}\right)\frac{x_1}{s_1} \tag{23.14}$$

avec $r_{32} = r_{23}$, $r_{31} = r_{13}$, $r_{12} = r_{21}$.

## 23.10 Généralisation à plus de trois variables

On généralise la notion à plus de trois variables par analogie avec les résultats précédents. Ainsi, l'équation de régression linéaire de $X_1$ en $X_2$, $X_3$ et $X_4$ s'écrit :

$$X_1 = b_{1.234} + b_{12.34}X_2 + b_{14.23}X_4 + b_{13.24}X_3 \tag{23.15}$$

et représente un **hyperplan dans l'espace à quatre dimensions**. En multipliant successivement cette équation par 1, $X_2$, $X_3$ et $X_4$ et en sommant membre à membre, on obtient les équations normales à partir desquelles on calcule $b_{1.234}$, $b_{12.34}$, $b_{14.23}$ et $b_{13.24}$ qui, par substitution dans l'équation précédente donnent l'**équation de régression des moindres carrés** $X_1$ **en** $X_2$**,** $X_3$ **et** $X_4$.

## 23.11 Corrélation partielle

On a souvent besoin de mesurer une variable indépendante particulière, toutes les autres variables mises en jeu restant constantes. Pour cela, on définit un **coefficient de corrélation partielle** comme dans le cas de la corrélation simple. La seule différence est qu'il faut considérer les variations expliquées et non expliquées qui apparaissent avec et sans la variable indépendante particulière.

Si $r_{12.3}$ est le coefficient de corrélation partielle entre $X_1$ et $X_2$, avec $X_3$ constant, on trouve :

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{\left(1 - {r_{13}}^2\right)\left(1 - {r_{23}}^2\right)}} \tag{23.16}$$

Si $r_{12.34}$ est le coefficient de corrélation partielle entre $X_1$ et $X_2$, avec $X_3$ et $X_4$ constants, on trouve :

$$r_{12.34} = \frac{r_{12.4} - r_{13.4}r_{23.4}}{\sqrt{\left(1 - {r_{13.4}}^2\right)\left(1 - {r_{23.4}}^2\right)}} = \frac{r_{12.3} - r_{14.3}r_{24.3}}{\sqrt{\left(1 - {r_{14.3}}^2\right)\left(1 - {r_{24.3}}^2\right)}} \tag{23.17}$$

**Remarque.** Ces résultats sont commodes, car n'importe quel coefficient de corrélation partielle peut ultérieurement dépendre des coefficients de corrélation $r_{12}$, $r_{23}$, *etc*. c'est-à-dire des **coefficients de corrélation d'ordre zéro**.

Dans le cas de deux variables $X$ et $Y$, si les deux droites de régression ont pour équation $Y = a_0 + a_1 X$ et $X = b_0 + b_1 Y$, on a vu que $r^2 = a_1 b_1$. Cela se généralise aisément :

$$X_1 = b_{1.234} + b_{12.34} X_2 + b_{13.24} X_3 + b_{14.23} X_4 \tag{23.18}$$

et

$$X_4 = b_{4.123} + b_{41.23} X_1 + b_{42.13} X_2 + b_{43.12} X_3 \tag{23.19}$$

sont les équations de régression respectives de $X_1$ en $X_2$, $X_3$, $X_4$ et de $X_4$ en $X_1$, $X_2$, $X_3$, on a :

$$r_{14.23}{}^2 = b_{14.23} b_{41.23} \tag{23.20}$$

Ceci peut être un point de départ pour définir les coefficients de corrélation partielle.

## 23.12 Relation entre les coefficients de corrélations multiple et partielle

La correspondance entre les coefficients de corrélation multiple et les différents coefficients de corrélation partielle permet d'établir des résultats intéressants que l'on généralise facilement.

$$1 - R_{1.23}{}^2 = \left(1 - r_{12}{}^2\right)\left(1 - r_{13.2}{}^2\right) \tag{23.21}$$

$$1 - R_{1.234}{}^2 = \left(1 - r_{12}{}^2\right)\left(1 - r_{13.2}{}^2\right)\left(1 - r_{14.23}{}^2\right) \tag{23.22}$$

## 23.13 Régression multiple non linéaire

Les résultats de la régression multiple linéaire se généralisent également à la régression multiple non linéaire. On définit les coefficients de corrélation correspondants par les mêmes méthodes que ci-dessus.

## 23.14 Remarque : les tenseurs statistiques

L'écriture tensorielle a été introduite dans ce chapitre par la notation indicée de la sommation. Ceci est l'occasion d'introduire la notion de **tenseur**.

Un vecteur est un tenseur à un indice. On dit que c'est un tenseur d'ordre 1, car il ne possède qu'un ordre pour ces composantes.

Un tenseur est un produit de vecteurs.

Un tenseur est dit **contravariant** lorsque son indice est en exposant. Le vecteur est alors décrit par ses **composantes**.

Un tenseur est dit **covariant** lorsque son indice est en indice. Le vecteur est alors décrit par son **produit scalaire**.

La covariance est l'explication du processus contravariant – covariant. La covariance s'exprime par la contravariance.

Tout tenseur est une matrice, mais toute matrice n'est pas forcément un tenseur.

Si une matrice représente un objet, alors elle est un tenseur.

Un tenseur commande les lois de transformation de coordonnées.

La matrice de covariance $\sigma$ est un tenseur symétrique vérifiant l'égalité suivante :

$$\sigma_{ij} = \sigma_{ji} \tag{23.23}$$

Concrètement, un tenseur $\sigma_{ij}$ représente un « carré de données ». On dit que c'est un tenseur d'ordre 2. Si on considère les deux vecteurs associés $V$ et $P$, si l'on considère les valeurs contravariantes de $V$ et les valeurs covariantes de $P$, on peut construire la matrice suivante :

$$\begin{pmatrix} V^1P_1 & V^1P_2 & V^1P_3 \\ V^2P_1 & V^2P_2 & V^2P_3 \\ V^3P_1 & V^3P_2 & V^3P_3 \end{pmatrix} \tag{23.24}$$

qui est un tenseur d'ordre 2, noté avec la première représentation :

$$\begin{pmatrix} T^{11} & T^{12} & T^{13} \\ T^{21} & T^{22} & T^{23} \\ T^{31} & T^{32} & T^{33} \end{pmatrix} \tag{23.25}$$

Il est possible de proposer la combinaison inverse, à savoir les valeurs covariantes de $V$ et les valeurs contravariantes de $P$ :

$$\begin{pmatrix} V_1P^1 & V_1P^2 & V_1P^3 \\ V_2P^1 & V_2P^2 & V_2P^3 \\ V_3P^1 & V_3P^2 & V_3P^3 \end{pmatrix} \tag{23.26}$$

qui est un tenseur d'ordre 2 donnant une seconde représentation :

$$\begin{pmatrix} T_1^1 & T_1^2 & T_1^3 \\ T_2^1 & T_2^2 & T_2^3 \\ T_3^1 & T_3^2 & T_3^3 \end{pmatrix} \tag{23.27}$$

Un tenseur d'ordre 2 permet l'association d'un nombre à toute combinaison possible de deux vecteurs de base en termes de directions (Fig. 23.1).

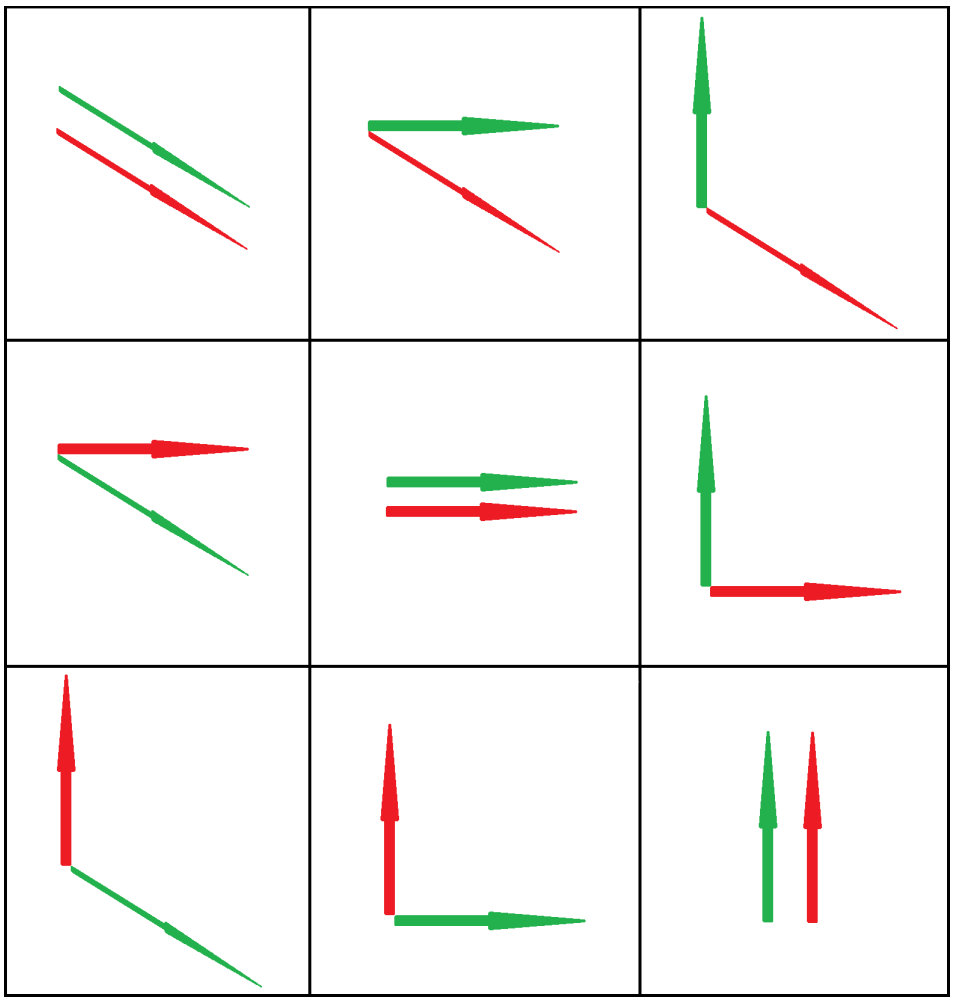

FIGURE 23.1 – Représentation des directions du tenseur d'ordre 2 de la représentation n° 1

Concrètement, un tenseur $\sigma_{ijk}$ représente un « cube de données ». On dit que c'est un tenseur d'ordre 3. Dans un tel tenseur, on associe un nombre à toute combinaison possible de trois vecteurs de base. On obtient le « cube de directions ».

# Bibliographie


[Carroll et Chang, 1970] CARROLL, J. D. et CHANG, J. J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of "eckert-young" decomposition. Psychometrika, 35:283–319.

[Christaller, 1950] CHRISTALLER, W. (1950). Das grundgerüst der raümlichen ordnung in europa. die systeme der europäischen zentralen orte. Frankfurter Geographische Hefte, 24(1):10–97.

[Escofier et Pagès, 2016] ESCOFIER, B. et PAGÈS, J. (2016). Analyses factorielles simples et multiples. Cours et études de cas. Sciences sup. Dunod, Paris.

[Gower, 1975] GOWER, J. C. (1975). Generalized procrustes analysis. Psychometrika, 40(1):39–51.

[Hurley et Cattell, 1962] HURLEY, J. et CATTELL, R. B. (1962). The procustes program : Producing direct rotation to test a hypothesized factor structure. Behavioral Science, 7(2):258–262.

[Husson et Pagès, 2006] HUSSON, F. et PAGÈS, J. (2006). Aspects méthodologiques du modèle indscal. Revue de statistique appliquée, 54(2):83–100.

[Lavit, 1988] LAVIT, C. (1988). Analyse conjointe de tableaux quantitatifs. Méthode + Programmes. Masson, Paris.

[Morand et Pagès, 2007] MORAND, E. et PAGÈS, J. (2007). Analyse factorielle multiple procustéenne. Journal de la société française de statistique, 148(2):65–97.

[Mosier, 1939] MOSIER, C. I. (1939). Determining a simple structure when loadings for certain tests are known. Psychometrika, 4:149–162.

[Pagès et Tenenhaus, 2002] PAGÈS, J. et TENENHAUS, M. (2002). Analyse factorielle multple et approche p.l.s. Revue de statistique appliquée, 50(1):5–33.

[Ten Berge, 1977] TEN BERGE, J. M. F. (1977). Orthogonal procustes rotation of two or more matrices. Psychometrika, 42(2):267–276.

[Torgerson, 1958] TORGERSON, W. S. (1958). Theory and Methods of Scaling. John Wiley & Sons, New York.

[Wold, 1975] WOLD, H. (1975). Modeling in complex situations with soft information. In Third World Congress of Econometric Society, August 21-26, Toronto, Toronto.

[Wold, 1982] WOLD, H. (1982). Soft Mdeling : the Basic Design and Some Extensions, pages 1–54. North-Holland Publishing Company, Amsterdam.

[Wold, 1985] WOLD, H. (1985). Partial Least Squares, pages 581–591. John Wiley & Sons, New York.