CheMOOCs

Glossaire

Jean-Claude Boulet
INRA, Montpellier, France

Jean-Michel Roger
IRSTEA, Montpellier, France

V17.10

Table des matières

Acronymes utilisés dans CheMoocs

Acronyme FR	Acronyme EN	Signification
ACP	PCA	analyse en composantes principales
AFD	FDA, LDA	analyse factorielle discriminante
CAH		classification ascendente hiérarchique
	CART	arbre de classification et de régression
	CV	validation croisée
kppv	knn	k plus proches voisins
	MLR	régression linéaire multiple
PIR	NIR	proche infra-rouge
SPIR		spectroscopie proche infra-rouge
	PCR	régression sur composantes principales
	PLSR	régression moindres carrés partiels, ou projection sur structures latentes
	PRESS	somme des carrés des erreurs de prédiction en validation croisée leave-one-out
	RMSEC	racine-carrée de l'erreur d'étalonnage
	RMSECV	racine-carrée de l'erreur de validation croisée
	RMSEP	racine-carrée de l'erreur de prédiction
	SVD	décomposition en valeurs singulières

Base d'un espace vectoriel ou d'un sous-espace vectoriel

Une *base* $\mathcal{U}$ d'un espace vectoriel $\mathcal{E}$ de dimension P est constituée de P vecteurs : $\mathcal{U} = \{\mathbf{u}_1, \mathbf{u}_2, ...\mathbf{u}_P\}$ linéairement indépendants, c'est à dire qu'aucun ne peut être écrit comme une combinaison linéaire des autres. Tous les vecteurs de $\mathcal{E}$ peuvent s'écrire sous forme d'une combinaison linéaire des vecteurs de $\mathcal{U}$, et cette combinaison linéaire est unique.
De même, une base d'un sous-espace vectoriel de $\mathbb{R}^P$, de dimension A, est constituée de A vecteurs définis dans $\mathbb{R}^P$ et linéairement indépendants.
Une base n'est pas unique : de très nombreuses bases (une infinité) peuvent être utilisées pour définir le même espace vectoriel.

Carte factorielle - score plot

La *carte factorielle* ou *score plot* en Anglais représente les coordonnées des observations sur le plan formé par deux axes principaux, généralement les axes 1 et 2.

Dans cette représentation, chaque point représente une observation. Les points sont donc distincts les uns des autres, comme le sont les échantillons qu'ils représentent.

Cercle des corrélations

Le *cercle des corrélations* est utilisé en ACP. Il consiste à représenter les corrélations de chacune des variables initiales sur un plan formé de deux composantes principales, souvent les deux premières.

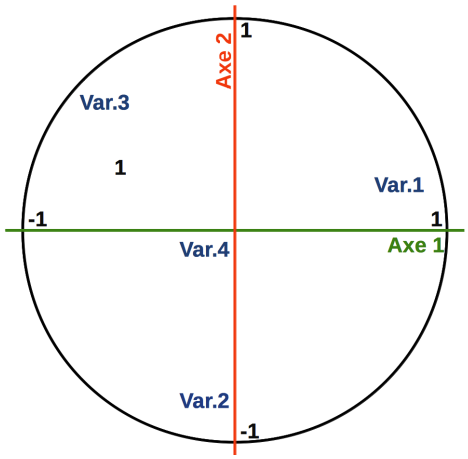

FIGURE 1 – Le cercle des corrélations pour 4 variables : Var1, Var2, Var3 et Var4 représentées sur le plan des composantes principales 1-2, ou axes 1-2.

Selon l'exemple de la figure 1, Var1 est bien expliquée par l'axe 1, avec une forte corrélation positive ; Var2 est bien expliquée par l'axe 2, avec une forte corrélation négative ; Var3 est bien expliquée par les axes 1 et 2, du fait de sa proximité avec le cercle ; enfin Var4 n'est pas du tout expliquée par les deux premières composantes, elle doit l'être par d'autres composantes.

Coefficient de corrélation

Le *coefficient de corrélation* selon Pearson permet, comme la covariance, de mesurer comment deux variables représentées ici par les vecteurs $\mathbf{x}$ et $\mathbf{y}$ varient dans le même sens, ou pas. Il est noté r et sa valeur est comprise entre 1 (forte corrélation positive) et -1 (forte corrélation négative). Une valeur de 0 indique que les variables varient indépendamment l'une de l'autre. La corrélation entre une variable et elle-même est 1.

Soient $\bar{x}$ et $\bar{y}$ les moyennes de $\mathbf{x}$ et $\mathbf{y}$, x_i et y_i leurs valeurs pour l'indice i. Le coefficient de corrélation

est :

$$r(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i-\bar{x})^2\sum_{i=1}^{N}(y_i-\bar{y})^2}}$$

Le *coefficient de détermination* R^2, compris entre 0 et 1, est le carré du coefficient de corrélation.

$$R^2_{\mathbf{x},\mathbf{y}} = r^2(\mathbf{x},\mathbf{y})$$

Coefficient de détermination

Voir à *corrélation*

Coefficients de régression

Soit une matrice de spectres $\mathbf{X}$ de dimensions $(N \times P)$ et une grandeur quantitative $\mathcal{Y}$ (ex : gluten) dont les valeurs prédites à partir de $\mathbf{X}$ donneront $\widehat{\mathbf{y}}$. Les coefficients de régression, ou b-coefficients, forment un vecteur de dimension $(P \times 1)$ noté $\mathbf{b}$ qui vérifie :

$$\widehat{\mathbf{y}} = \mathbf{X}\mathbf{b} + \mathbf{E}$$

$\mathbf{E}$ étant l'erreur. La formule s'écrit aussi avec β et ϵ au lieu de $\mathbf{b}$ et $\mathbf{E}$:

$$\widehat{\mathbf{y}} = \mathbf{X}\beta + \epsilon$$

Combinaison linéaire

Des vecteurs $\{\mathbf{u}_1, \mathbf{u}_2, ...\mathbf{u}_P\}$ sont reliés par une *combinaison linéaire* s'il existe les nombres $\{a_1, a_2, ...a_P\}$ tels que :

$$a_1\mathbf{u}_1 + a_2\mathbf{u}_2 + ... + a_P\mathbf{u}_P = \overrightarrow{0}$$

$\overrightarrow{0}$ étant le vecteur nul. Dans le cas contraire, les vecteurs sont dits indépendants.

Colinéarité de vecteurs

Deux vecteurs $\mathbf{x}_1$ et $\mathbf{x}_2$ sont *colinéaires* si on peut trouver un nombre k tel que : $\mathbf{x}_1 = k\mathbf{x}_2$. Deux vecteurs colinéaires pointent la même direction de l'espace, mais pas nécessairement le même sens.

Cosinus

Le *cosinus* est utilisé pour mesurer l'angle entre deux vecteurs. Il est compris entre -1 (deux vecteurs colinéaires dans des sens opposés) et 1 (deux vecteurs colinéaires dans le même sens). Il vaut 0 pour deux vecteurs orthogonaux.
La mesure du cosinus est illustrée avec la figure 2. Le cosinus entre $\mathbf{u}$ et $\mathbf{v}$ est égal au rapport de la distance entre A et B divisée par la distance entre A et C. Plus simplement, la formule de calcul du cosinus à partir de $\mathbf{u}$ et $\mathbf{v}$ est :

$$cos(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u}'\mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}$$

Cette formule est basée sur l'utilisation du produit scalaire.

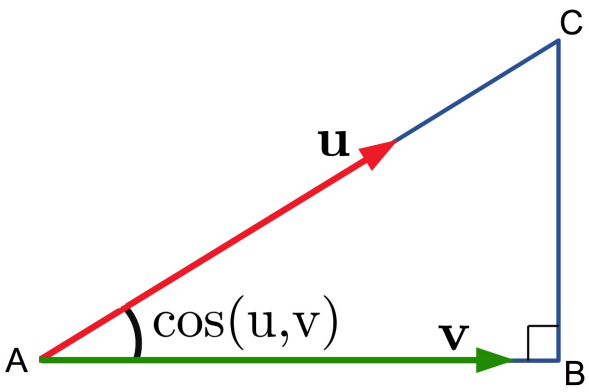

FIGURE 2 – Le cosinus entre deux vecteurs $\mathbf{u}$ et $\mathbf{v}$.

Covariance

La *covariance* permet, comme le coefficient de corrélation, de mesurer comment deux variables représentées ici par les vecteurs $\mathbf{x}$ et $\mathbf{y}$ varient dans le même sens, ou pas. La valeur de la covariance est dépendante des unités prises pour mesurer les N valeurs de $\mathbf{x}$ et de $\mathbf{y}$. Elle peut prendre tout

type de valeurs.

Soient $\bar{x}$ et $\bar{y}$ les moyennes de $\mathbf{x}$ et $\mathbf{y}$, x_i et y_i leurs valeurs pour l'indice i. La formule de la covariance est :

$$cov(\mathbf{x}, \mathbf{y}) = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})$$

Décomposition d'une matrice

Il est fréquent en chimiométrie de décomposer une matrice de spectres $\mathbf{X}$ selon l'équation :

$$\mathbf{X} = \mathbf{TP}' + \mathbf{E}$$

La matrice $\mathbf{P}$ représente les *loadings* (les vecteurs-propres des composantes principales en ACP). Les loadings forment une base de l'espace dans lequel évoluent les spectres de $\mathbf{X}$. S'ils sont correctement définis (avec assez d'échantillons par exemple) ils doivent être indépendants des échantillons, c'est à dire pouvoir être utilisés pour décrire un nouvel échantillon.

La matrice $\mathbf{T}$ représente les *scores* (les coordonnées des observations sur les premiers axes, en ACP). Ils sont très dépendants d'un échantillon à l'autre puisqu'ils représentent les différences entre échantillons.

La matrice $\mathbf{E}$ est parfois notée ϵ. Elle représente l'erreur, et doit être aussi faible que possible.

Les décompositions de matrices se retrouvent dans plusieurs méthodes, par exemple : ACP, ICA, MCR-ALS, PLSR.

Délimiteur décimal

Le point et la virgule seront utilisés indistinctement comme *délimiteurs décimaux*. Nous n'utilisons pas de délimiteur de milliers. Ex : $\pi = 3.14159 = 3,14159$

Dimension d'un espace vectoriel

La *dimension d'un espace vectoriel* est le nombre minimum de vecteurs nécessaires pour générer tout l'espace vectoriel, c'est à dire toutes ses directions.

Ex.1 : l'espace 3D que nous connaissons est un espace vectoriel de dimension 3, soit $\mathbb{R}^3$.

Ex.2 : une carte de Paris est un espace vectoriel de dimension 2 - si on reste dans le plan de la carte-.

Ex.3 : un spectre de 1101 variables est défini dans un espace vectoriel de dimension 1101, soit $\mathbb{R}^{1101}$.

Discret / continu (pour une fonction ou variable)

Une *fonction continue* peut prendre toutes les valeurs réelles possibles dans une certaine plage. Une *fonction discrète* n'est définie que pour un certain nombre de valeurs.
Une absorbance à une longueur d'onde donnée est toujours une variable continue puisque l'absorbance est une valeur réelle. Le même raisonnement est appliqué pour un spectre, soit un ensemble de longueurs d'onde ordonnnées. Un spectre est par nature continu : il est possible de connaitre l'absorbance de n'importe quelle longueur d'onde, par exemple pour $\lambda = 1785.4564nm$. En pratique, on ne gardera pas toutes les valeurs, mais juste un certain nombre. Avec le λ précédent, les plus proches seront l'absorbance à $\lambda = 1785nm$ et l'absorbance à $\lambda = 1786nm$ si on échantillonne tous les $1nm$ -d'où la discrétisation-. Dit autrement, le spectre a été discrétisé puisqu'on ne garde que les longueurs d'onde en nm correspondant à des nombres entiers, toutefois les absorbances à $\lambda = 1785nm$ et à $\lambda = 1786nm$ restent deux variables continues.

Distance entre deux points

Soient deux points $\mathbf{x}$ et $\mathbf{z}$ représentés dans l'espace $\mathbb{R}^N$ par leurs coordonnées $(x_1, x_2, ...x_n)$ et $(z_1, z_2, ...z_n)$ disposées sous forme de vecteurs-colonne. Soit $\mathbf{M}$ une métrique définie dans $\mathbb{R}^N$. La distance entre $\mathbf{x}$ et $\mathbf{z}$ au sens de $\mathbf{M}$, notée $d_M(\mathbf{x}, \mathbf{z})$, est définie ainsi :

$d_M(\mathbf{x}, \mathbf{z}) = \sqrt{(\mathbf{x} - \mathbf{z})'\mathbf{M}(\mathbf{x} - \mathbf{z})}$

Note : la distance peut aussi se calculer dans $\mathbb{R}^P$. Deux métriques sont utilisées en chimiométrie : la métrique Euclidienne usuelle, $\mathbf{M}$ est la matrice-identité $\mathbf{I}$, et la métrique de Mahalanobis, $\mathbf{M}$ est une matrice de variance-covariance.

Dans le cas de la métrique Euclidienne usuelle, la formule précédente se simplifie en :

$d_M(\mathbf{x}, \mathbf{z}) = \sqrt{(\mathbf{x} - \mathbf{z})'(\mathbf{x} - \mathbf{z})}$

Cela revient à : (1) faire la différence entre les N éléments des deux vecteurs, deux à deux ; (2) mettre au carré les N différences obtenues ; (3) faire la somme des N carrés ; (4) prendre la racine carrée de la somme.

La distance Euclidienne usuelle est celle que nous utilisons tous dans notre espace à 3 dimensions, la longueur du trajet le plus court entre ces deux points. En chimiométrie la même notion est étendue

à un espace de dimension N ou P.

Distribution d'une variable

La *distribution d'une variable* est la relation entre des classes de mesure, qui peuvent être des plages de valeurs prises par une variable, et le nombre d'observations ou bien la fréquence d'appartenance des observations à chaque classe. La distribution est souvent représentée par un histogramme, comme dans l'exemple présenté figure 3.

Classe d'age	*Pourcentage*
moins de 20 *ans*	25, 6
$20 - 59$ *ans*	53, 8
$60 - 74$ *ans*	13, 4
75 *ans et plus*	7, 2

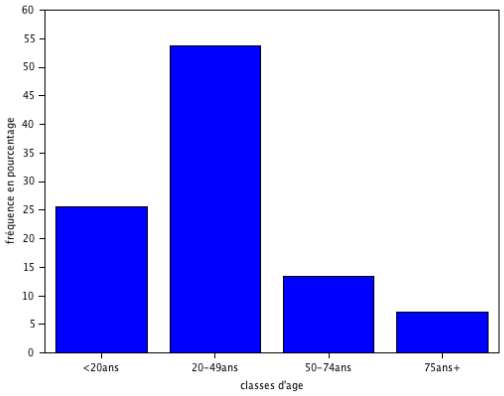

FIGURE 3 – Exemple d'une distribution : la population Française en 2000 (à gauche) et son histogramme (à droite). Source : INSEE

Une distribution très répandue en biologie est la distribution normale, ou distribution Gaussienne, avec un aspect caractéristique en forme de cloche comme observable sur la figure 4.

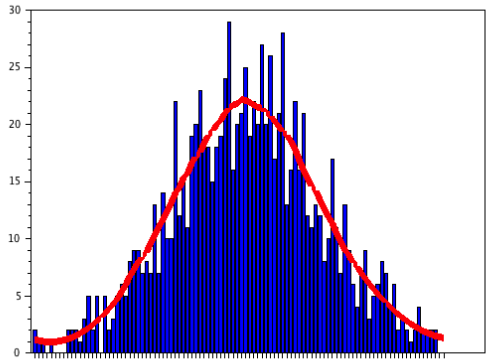

FIGURE 4 – Exemple d'une distribution normale ou Gaussienne : en rouge la forme continue attendue pour un très grand nombre d'observations, en bleu la forme discrète observée pour 1000 observations réparties en 100 classes.

Ecart-type

Voir à *variance*.

Espace dual

Soit une matrice $\mathbf{X}$ de dimensions $(N \times P)$. Les espaces vectoriels définis dans $\mathbb{R}^P$ avec les vecteurs-ligne de $\mathbf{X}$ et dans $\mathbb{R}^N$ avec les vecteurs-colonne de $\mathbf{X}$ sont des *espaces duaux* qui partagent certaines propriétés : même origine des données (la matrice $\mathbf{X}$), même dimension, et après ACP les scores d'un espace sont les loadings de l'autre espace.

Espace vectoriel

Les vecteurs de même dimension peuvent être additionnés ou multipliés par un scalaire. Par exemple :

- l'addition de $\mathbf{x}_1 = \begin{pmatrix} 3.4 & 2.9 & 7.1 \end{pmatrix}$ et $\mathbf{x}_2 = \begin{pmatrix} 4.3 & 9.6 & 5.5 \end{pmatrix}$ est : $\mathbf{x}_1 + \mathbf{x}_2 = \begin{pmatrix} 3.4+4.3 & 2.9+9.6 & 7.1+5. \end{pmatrix}$ $\begin{pmatrix} 7.7 & 12.5 & 12.6 \end{pmatrix}$.
- La multiplication de $\mathbf{x}_1 = \begin{pmatrix} 3.4 & 2.9 & 7.1 \end{pmatrix}$ par 2 donne : $2\ \mathbf{x}_1 = \begin{pmatrix} 2*3.4 & 2*2.9 & 2*7.1 \end{pmatrix} =$ $\begin{pmatrix} 6.8 & 5.8 & 14.2 \end{pmatrix}$.

Ces opérations d'addition et de multiplication par un scalaire (nombre) sont des combinaisons linéaires entre vecteurs et permettent de créer de nouveaux vecteurs : par exemple $\mathbf{x}_3 = 2 * \mathbf{x}_1 - \mathbf{x}_2$

.

Un espace vectoriel regroupe tous les vecteurs défini par un ensemble donné de vecteurs et toutes les combinaisons linéaires possibles entre ces vecteurs. On dit que l'espace vectoriel est généré par cet ensemble de vecteurs.

Fiabilité d'une mesure

Une mesure est *fiable* lorsque le résultat obtenu est précis, et que la répétition de la mesure acquise sur le même échantillon mis dans des conditions différentes de jour, d'opérateur, etc, donne le même résultat.

Hétéroscédasticité, homoscédasticité

L'*hétéroscédasticité* et l'*homoscédasticité* sont deux termes opposés, employés pour décrire l'évolution de la variance des erreurs en fonction des valeurs mesurées. Soit une variable $\mathcal{Y}$ dont les valeurs exactes pour N observations sont représentée par un vecteur $\mathbf{y}$. Les valeurs prédites pour ces mêmes N observations sont représentées par $\widehat{\mathbf{y}}$. Pour tout i, l'erreur de prédiction est : $e_i = y_i - \widehat{y}_i$. Il y a homoscédasticité si la variance des erreurs e_i est indépendante des valeurs de $\mathcal{Y}$. Par exemple, la mesure de la taille d'individus au moyen d'un mètre-ruban est homoscédastique puisque l'erreur dépend de la lecture de l'opérateur, pas de la taille de l'individu. Mais si différents objets de tailles très différents sont mesurés avec un pied à coulisse, un mètre-ruban et un décamètre selon les cas, l'incertitude sera d'autant plus petite que l'objet mesuré sera petit, la mesure sera donc hétéroscédastique.

Histogramme

Un *histogramme* est une figure représentant en abscisse des classes, et en ordonnée des nombres ou des fréquences d'appartenance à des classes de mesure. Voir un exemple à la définition de *distribution*.

Homoscédasticité

Voir à *hétéroscédasticité*.

Individu extrême, atypique, outlier

Un *individu extrême, atypique, outlier* est un individu qui se distingue des autres, par exemple par la forme de son spectre ou sa composition chimique. Cet individu peut être parfaitement valide : ainsi un chauffeur transportant une équipe de basket dans son bus aurait une taille atypique par rapport aux autres personnes présentes dans le bus. Si l'individu est valide, il doit être gardé parmi les données.
Parfois, on constate que l'individu est extrême parce qu'une erreur a été commise, soit sur les mesures spectrales, soit sur la caractérisation de son état. Dans ce cas seulement, l'individu peut être supprimé du jeu de données.

Inertie

L'*inertie* d'un nuage de points par rapport à un axe est la somme des distances de chacun des points à l'axe, portées au carré, et pondérées de la masse de chaque point (généralement 1) :

$$I = \Sigma m_i r_i^2$$

Plus l'inertie est faible, et plus les points sont regroupés autour de l'axe. *A contrario*, quelques points éloignés de l'axe peuvent augmenter considérablement l'inertie, car leur distance intervient au carré.

Inverse, pseudo-inverse de Moore-Penrose, matrice mal conditionnée

Soit $\mathbf{A}$ une matrice carrée, c'est à dire que son nombre de colonnes est égal à son nombre de lignes, nombre que nous noterons P. L'*inverse* de $\mathbf{A}$ est notée $\mathbf{A}^{-1}$ et vérifie : $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_P$, avec $\mathbf{I}_P$ la matrice-identité de dimension P. L'inverse de $\mathbf{A}$ n'est pas calculable si son rang n'est pas égal à sa dimension, c'est à dire P, et si $\mathbf{A}$ n'est pas carrée.
Prenons un exemple. L'inverse de $\mathbf{A} = \begin{pmatrix} 3 & 6 \\ 4 & 2 \end{pmatrix}$ est $\mathbf{A}^{-1} = \begin{pmatrix} -0.111111 & 0.333333 \\ 0.222222 & -0.166667 \end{pmatrix}$ puisque $\begin{pmatrix} 3 & 6 \\ 4 & 2 \end{pmatrix}$ x $\begin{pmatrix} -0.111111 & 0.333333 \\ 0.222222 & -0.166667 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

Pour toute matrice $\mathbf{B}$ de dimensions $(N \times P)$, il est possible de calculer sa *pseudo-inverse de Moore-Penrose*, notée $\mathbf{B}^+$, de dimensions $(P \times N)$, qui vérifie les 5 propriétés suivantes :
$\mathbf{B}\mathbf{B}^+\mathbf{B} = \mathbf{B}$;
$\mathbf{B}^+\mathbf{B}\mathbf{B}^+ = \mathbf{B}^+$;
$(\mathbf{B}\mathbf{B}^+)' = \mathbf{B}\mathbf{B}^+$;
$(\mathbf{B}^+\mathbf{B})' = \mathbf{B}^+\mathbf{B}$;
rang de $\mathbf{B}^+$ = rang de $\mathbf{B}$.
Si $\mathbf{B}$ est une matrice carrée et est inversible (rang = dimension), alors $\mathbf{B}^+ = \mathbf{B}^{-1}$.
Les inverses et pseudo-inverses sont utilisées dans les projections orthogonales et dans les projections obliques.

Jeux d'étalonnage, de validation, de test

Le *jeu d'étalonnage* est utilisé pour construire un modèle d'étalonnage.
Le *jeu de validation* est utilisé pour valider un ou plusieurs modèles issus de l'étalonnage. Si la prédiction est mauvaise, il faut refaire d'autres étalonnages.
Le *jeu de test* est utilisé pour valider un ou plusieurs modèles issus de l'étalonnage et ayant passé avec succès le jeu de validation. Il renseigne sur la robustesse du ou des modèles obtenus. Il ne doit pas être utilisé pour construire un nouveau modèle d'étalonnage.

Loadings

Voir *Décomposition d'une matrice*.

Matrice

Une *matrice* est un tableau de nombres de N lignes et P colonnes. Par exemple : $\mathbf{X} = \begin{pmatrix} 3.4 & 2.9 & 7.1 \\ 5.0 & 1.2 & 5.8 \end{pmatrix}$ est une matrice de dimensions (2×3), 2 lignes et 3 colonnes.

Les matrices sont souvent obtenues en regroupant (concaténant) plusieurs vecteurs. Ici par exemple,

$\mathbf{X}$ regroupe les vecteurs-colonne $\begin{pmatrix} 3.4 \\ 5.0 \end{pmatrix}$, $\begin{pmatrix} 2.9 \\ 1.2 \end{pmatrix}$ et $\begin{pmatrix} 7.1 \\ 5.8 \end{pmatrix}$ ou bien les vecteurs-ligne $\begin{pmatrix} 3.4 & 2.9 & 7.1 \end{pmatrix}$ et $\begin{pmatrix} 5.0 & 1.2 & 5.8 \end{pmatrix}$

Matrice conjonctive, disjonctive

Les *matrices conjonctives* et les *matrices disjonctives* sont utilisées pour coder l'appartenance d'individus à des classes au moyen de nombres qui seront repris dans des calculs.

2005	2006	2007	2008	2009	2010
Galles	*France*	*France*	*Galles*	*Irlande*	*France*
2011	2012	2013	2014	2015	2016
Angleterre	*Galles*	*Galles*	*Irlande*	*Irlande*	*Angleterre*

FIGURE 5 – Rugby, vainqueurs 2005 – 2016 du tournoi des 6 nations

Prenons l'exemple du tableau de la figure 5.

Le *codage conjonctif* s'obtient en attribuant les valeurs de 1, 2, 3 et 4 aux nations suivantes : Galles, France, Irlande et Angleterre. L'Italie et l'Ecosse, n'ayant pas gagné, ne sont pas représentés

Le *codage disjonctif* s'obtient en créant une matrice de 12 lignes et 4 colonnes. Les lignes correspondent aux années, les colonnes aux 4 nations, dans l'ordre : Galles, France, Irlande et Angleterre. La matrice est remplie de 0 sauf lorsque la ligne et la colonne correspondent à un vainqueur, auquel cas on met la valeur 1 ; il n'y a qu'une valeur 1 par ligne.

Le résultat donne la figure 6.

$$\begin{pmatrix} 1 \\ 2 \\ 2 \\ 1 \\ 3 \\ 2 \\ 4 \\ 1 \\ 1 \\ 3 \\ 3 \\ 4 \end{pmatrix} \quad \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

FIGURE 6 – Exemple de codage conjonctif (à gauche) et disjonctif (à droite) des données de la figure 5

Matrice de corrélation

Une *matrice de corrélation* représente les corrélations entre deux jeux de variables. Supposons qu'un premier jeu contienne les trois variables $\mathbf{x}_1$, $\mathbf{x}_2$ et $\mathbf{x}_3$ et un deuxième jeu les deux variables $\mathbf{y}_1$ et $\mathbf{y}_2$. Notons $r_{xi,yj}$ le coefficient de corrélation entre $\mathbf{x}_i$ et $\mathbf{y}_j$. Une matrice de corrélation est :

$$\begin{pmatrix} r_{x1,y1} & r_{x1,y2} \\ r_{x2,y1} & r_{x2,y2} \\ r_{x3,y1} & r_{x3,y2} \end{pmatrix}$$

Lorsque la matrice de corrélation est réalisée sur un seul jeu de variables (le deuxième jeu est aussi le premier), la matrice de corrélation est carrée, symétrique, et sa diagonale secondaire est composée de 1, comme dans l'exemple ci-dessous :

$$\begin{pmatrix} 1.00 & 0.67 & -0.39 & 0.05 \\ 0.67 & 1.00 & 0.92 & -0.71 \\ -0.39 & 0.92 & 1.00 & 0.23 \\ 0.05 & -0.71 & 0.23 & 1.00 \end{pmatrix}$$

Matrice des indicatrices

Une *matrice des indicatrices* est une matrice disjonctive, voir description à *matrices conjonctives, disjonctives* .

Matrice identité

La matrice identité est une matrice carrée (autant de lignes que de colonnes) remplie de 0 sauf des 1 sur la diagonale.

$$\text{Ex : } \mathbf{I} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

est la matrice identité de $\mathbb{R}^4$.

Matrice mal conditionnée

Une matrice *mal conditionnées* (*ill-conditionned* en Anglais) est une matrice rendue inversible grâce au bruit de l'information qu'elles contiennent.

Soit la matrice $\mathbf{X} = \begin{pmatrix} 1 & 0.99999 \\ 2 & 2.00001 \end{pmatrix}$ et son inverse $\mathbf{X}^{-1} = \begin{pmatrix} 66667 & -33333 \\ -66666.667 & 33333.333 \end{pmatrix}$.

Le calcul de $\mathbf{X}^{-1}$ a été possible, et pourtant on se rend compte que les deux colonnes de $\mathbf{X}$ sont identiques à une très faible valeur près.

Comme la matrice $\mathbf{X}$ est construite à partir de mesures, par exemple des spectres, ses valeurs ont une certaine incertitude ; ce ne sont pas des valeurs exactes. Ainsi dans cet exemple 1 et 2 ne sont pas significativement différent de 0.99999 et 2.00001, en conséquence $\mathbf{X}$ n'aurait jamais dû être inversible.

Le problème des matrices mal conditionnées est qu'un résultat est obtenu, et qu'il a de grandes chances d'être faux, tout comme les conclusions qui vont s'appuyer dessus. Il faut donc prendre des précautions pour ne pas inverser des matrices mal conditionnées.

Moyenne

Soient N nombres $\{x_1, x_2, ...x_n\}$ formant un vecteur $\mathbf{x}$ de longueur N dont on veut calculer la moyenne. En toute rigueur, le calcul dépend de ce que représentent ces N nombres.

S'ils représentent la *totalité* des valeurs d'une population, par exemple les notations obtenues par un élève tout au long de l'année, la moyenne est obtenue en faisant la somme de ces N nombres puis

en divisant par leur nombre (N) :

$$moyenne = \bar{x} = \frac{x_1 + x_2 + ... + x_N}{N} = \frac{\sum_{i=1}^{N} x_i}{N}$$

S'ils représentent un *échantillonnage* pris dans une population, par exemple la taille de 1000 personnes choisies au hasard dans un pays, la moyenne est obtenue en divisant la somme des N nombres par $N - 1$:

$$moyenne = \bar{x} = \frac{x_1 + x_2 + ... + x_N}{N - 1} = \frac{\sum_{i=1}^{N} x_i}{N - 1}$$

En chimiométrie, le nombre d'observations N est suffisamment important pour que la différence entre les résultats obtenus avec N et avec $N - 1$ puissent être considérée comme négligeable. Ainsi, en pratique, on simplifie souvent les formules en divisant systématiquement par N.

A noter que la même logique est appliquée pour le calcul de la variance ou des différentes erreurs RMSEC.

Norme d'un vecteur, normalisation

La *norme* d'un vecteur représente la longueur du vecteur dans l'espace. Elle est directement liée au produit scalaire. Soit un vecteur-colonne $\mathbf{x}$ contenant N valeurs x_i, $i = 1$ à N. Sa norme est notée $\|\mathbf{x}\|$. Elle est donnée par la formule :

$$\|\mathbf{x}\| = \sqrt{\mathbf{x}'\mathbf{x}} = \sqrt{\mathbf{x}^T\mathbf{x}} = \sqrt{\sum_{i=1}^{N} x_i^2}$$

Ainsi, la norme d'un vecteur est la racine carrée du produit scalaire du vecteur avec lui-même.

La normalisation est la division d'un vecteur par sa norme.

A ne pas confondre avec la standardisation, qui consiste à diviser un vecteur par l'écart-type de ses valeurs, pour que les valeurs du vecteur résultat aient un écart-type de 1.

Orthogonalité entre vecteurs

Deux *vecteurs orthogonaux* sont deux vecteurs dont le produit scalaire est nul.

La notion d'orthogonalité dans l'espace $\mathbb{R}^P$ est exactement la même que celle que nous avons dans

l'espace à 3 dimensions, où par exemple la verticale et l'horizontale forment un angle droit, donc sont deux directions orthogonales.

Précision d'une méthode analytique

La *précision d'une méthode analytique* est en relation avec l'*incertitude de mesure* : une mesure précise a nécessairement une incertitude de mesure faible, et *vice-versa*. L'incertitude de mesure désigne habituellement la plage autour de la valeur estimée ou mesurée autour de laquelle la valeur vraie (inconnue) se situe avec 95% de chances. En métrologie, l'incertitude de mesure est : $+/-2Sr$ avec Sr l'écart-type de répétabilité de la méthode.

Produit matriciel

Le *produit matriciel*, ou produit de deux matrices $\mathbf{X}_1$ et $\mathbf{X}_2$, est la matrice obtenue en calculant les produits scalaires de tous les vecteurs-colonne de la transposée de $\mathbf{X}_1$ par tous les vecteurs-colonne de $\mathbf{X}_2$.

Par exemple, le produit scalaire entre $\mathbf{X}_1 = \begin{pmatrix} 3.4 & 2.9 & 7.1 \\ 6.4 & 7.2 & 3.9 \end{pmatrix}$ et $\mathbf{X}_2 = \begin{pmatrix} 4.3 & 1.4 \\ 9.6 & 3.7 \\ 5.5 & 0.9 \end{pmatrix}$ donne :

$$\mathbf{X}_1.\mathbf{X}_2 = \begin{pmatrix} 3.4 & 2.9 & 7.1 \\ 6.4 & 7.2 & 3.9 \end{pmatrix} . \begin{pmatrix} 4.3 & 1.4 \\ 9.6 & 3.7 \\ 5.5 & 0.9 \end{pmatrix} = \begin{pmatrix} 81.55 & 21.88 \\ 118.09 & 39.11 \end{pmatrix}$$

Produit scalaire

Le *produit scalaire* concerne deux vecteurs de même dimensions. C'est un nombre obtenu par la somme des produits 2 à 2 des éléments des deux vecteurs.

Ex : Le produit scalaire de $\mathbf{x}_1 = \begin{pmatrix} 3.4 \\ 2.9 \\ 7.1 \end{pmatrix}$ et $\mathbf{x}_2 = \begin{pmatrix} 4.3 \\ 9.6 \\ 5.5 \end{pmatrix}$ est :

$$\mathbf{x}_1^T.\mathbf{x}_2 = \begin{pmatrix} 3.4 & 2.9 & 7.1 \end{pmatrix} . \begin{pmatrix} 4.3 \\ 9.6 \\ 5.5 \end{pmatrix} = (3.4 * 4.3) + (2.9 * 9.6) + (7.1 * 5.5) = 81.51$$

Projection orthogonale, projection oblique

Une *projection orthogonale* consiste à représenter un objet défini dans un espace de dimension P dans un autre espace de dimension plus petite A ($A < P$), selon une direction qui est orthogonale à ce dernier espace.
Visuellement, un exemple de projection orthogonale est donné par l'image 7. L'objet est constitué des mains, en 3 dimensions. La projection est réalisée sur le rideau, en 2 dimensions. La direction orthogonale au rideau est celle de la lumière, de la bougie vers les mains.

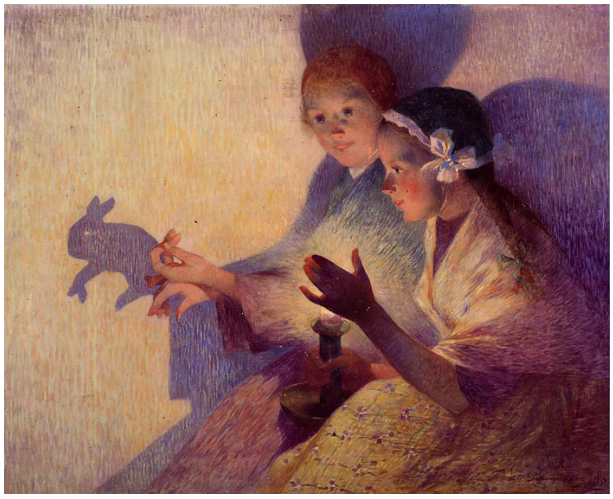

FIGURE 7 – Tableau de Ferdinand Loyen Du Puigaudeau (1864-1930) illustrant une projection orthogonale.

Le résultat est une perte d'information (on ne reconnait plus les mains sur le rideau), mais aussi la mise en avant d'une autre information (le contour des mains a la forme d'un lapin).

Aspects mathématiques des projections orthogonales

Soit une matrice $\mathbf{X}$ de spectres de dimensions $(N \times P)$ et soit également un sous-espace vectoriel défini par les vecteurs-colonne d'une matrice $\mathbf{P}$ de dimensions $(A \times P)$. La projection orthogonale de

$\mathbf{X}$ *sur* le sous-espace défini par les colonnes de $\mathbf{P}$,notée $\mathbf{X_P}$, est :

$$\mathbf{X_P} = \mathbf{XP}(\mathbf{P'P})^{+}\mathbf{P'}$$

$(\mathbf{P'P})^{+}$ est la pseudo-inverse de Moore-Penrose de $\mathbf{P'P}$ (voir à *inverse, pseudo-inverse d'une matrice*). De même, la projection de $\mathbf{X}$ *orthogonalement au* sous-espace défini par les colonnes de $\mathbf{P}$, notée $\mathbf{X_{P\perp}}$, est :

$$\mathbf{X_{P\perp}} = \mathbf{X} - \mathbf{X_P} = \mathbf{X}(\mathbf{I}_P - \mathbf{P}(\mathbf{P'P})^{+}\mathbf{P'})$$

Les *projections obliques* sont obtenues en introduisant une métrique, représentée par une matrice carrée $\mathbf{S}$ symétrique, semi-définie positive. Elles s'écrivent :

$$\mathbf{X_P} = \mathbf{XSP}(\mathbf{P'SP})^{+}\mathbf{P'}$$

$$\mathbf{X_{P\perp}} = \mathbf{X}(\mathbf{I}_P - \mathbf{SP}(\mathbf{P'SP})^{+}\mathbf{P'})$$

QQplot

Le *QQplot* ou *quantile-quantile plot* représente les variables (en abscisse) et leurs valeurs (en ordonnée) classées par ordre croissant. Des lignes horizontales positionnées de part et d'autre de la moyenne à 1 écart-type, 2 écarts-type,...permettent d'identifier les variables les plus importantes.

Rang d'une matrice

Le *rang d'une matrice* est la dimension du sous-espace vectoriel engendré par ses vecteurs-ligne ou ses vecteurs-colonne. C'est un nombre entier. La dimension de l'espace vectoriel généré par les lignes est de même dimension que l'espace vectoriel généré par les colonnes. Ainsi, le rang est inférieur ou égal à la plus petite dimension de la matrice.

Robustesse d'un modèle

La *robustesse d'un modèle* est sa capacité à garder une bonne capacité de prédiction quand il est soumis à des variations d'environnement (comme la température) ou d'échantillons.

RMSEC, RMSECV, RMSEP, PRESS

Le *RMSEC* ou *root mean square error of calibration* est l'erreur d'étalonnage.
Le *RMSECV* ou *root mean square error of cross-validation* est l'erreur de validation croisée.
Le *RMSEP* ou *root mean square error of prediction* est l'erreur de prédiction.
Soient $\mathbf{y}$ et $\widehat{\mathbf{y}}$ les vecteurs donnant les valeurs de référence et les valeurs prédites pour N échantillons. Soit A le nombre de dimensions de PLSR ou PCR utilisées pour construire le modèle, et C la valeur de centrage, 1 si modèle centré, 0 si modèle non centré. Le calcul du RMSEP est :

$$RMSEP = \sqrt{\frac{(\widehat{\mathbf{y}} - \mathbf{y})'(\widehat{\mathbf{y}} - \mathbf{y})}{N}} = \sqrt{\frac{\sum_{i=1}^{N}(\widehat{y}_i - y_i)^2}{N}}$$

La même formule est souvent utilisée pour calculer le RMSEC et RMSECV. Toutefois, un calcul plus juste est obtenu en calculant plus précisément les degrés de liberté, soit : $ddl = N - A - C$. Cela donne :

$$RMSEC \text{ ou } RMSECV = \sqrt{\frac{(\widehat{\mathbf{y}} - \mathbf{y})'(\widehat{\mathbf{y}} - \mathbf{y})}{N - A - C}} = \sqrt{\frac{\sum_{i=1}^{N}(\widehat{y}_i - y_i)^2}{N - A - C}}$$

Le calcul précis est surtout utile lorsque le nombre d'observations est faible.

Le *PRESS* ou *Predicticted Residual Error Sum of squares* est la somme des carrés des erreurs de prédiction obtenues en validation croisée en enlevant un seul individu (leave-one-out ou loo) :

$$PRESS = (\widehat{\mathbf{y}}_{\mathbf{loo}} - \mathbf{y})'(\widehat{\mathbf{y}}_{\mathbf{loo}} - \mathbf{y}) = \sum_{i=1}^{N}(\widehat{y}_{i,loo} - y_i)^2$$

Le $RMSECV$ obtenu en validation croisée s'écrit :

$$RMSECV_{loo} = \sqrt{\frac{PRESS}{N}} \qquad ou \qquad RMSECV_{loo} = \sqrt{\frac{PRESS}{N - A - C}}$$

Scores

Voir *Décomposition d'une matrice*.

Scores plot

Voir *Carte factorielle*.

Sous-espace vectoriel

Un vecteur de dimension P (ex : un spectre de P longueurs d'onde) est défini dans l'espace vectoriel $\mathbb{R}^P$, de dimension P.
Supposons que nous choisissions A vecteurs de $\mathbb{R}^P$, $A < P$, tous pointant dans des directions différentes. Par combinaison linéaire de ces A vecteurs, nous pouvons construire un espace vectoriel de dimension A qui sera un *sous-espace vectoriel* de $\mathbb{R}^P$, car compris dans $\mathbb{R}^P$.

Ceci est très utile en spectroscopie. L'ensemble des spectres mesurés sur un produit donné n'occupe pas tout $\mathbb{R}^P$, dans la mesure où les spectres ont des formes semblables. En pratique, $A << P$. Il devient possible de ne travailler que dans le sous-espace vectoriel, ce qui facilite les calculs et améliore l'interprétation.

Splot

Le *Splot* est une figure permettant d'identifier les variables d'une matrice $\mathbf{X}$ qui sont les plus importantes lors de la décomposition de $\mathbf{X}$ en scores et loadings ($\mathbf{X} = \mathbf{TP'} + \mathbf{E}$).
La matrice $\mathbf{X}$ contient N observations et P variables spectrales. La matrice de scores $\mathbf{T}$ contient A colonnes. Pour chaque colonne $\mathbf{t}_i$ extraite de $\mathbf{T}$, $i = 1$ à A, le Splot représente les corrélations entre $\mathbf{t_i}$ et les P colonnes de $\mathbf{X}$ versus les covariances entre $\mathbf{t_i}$ et les P colonnes de $\mathbf{X}$. Le terme Splot vient du fait que la forme obtenue rappelle souvent un S. Les variables importantes sont représentées aux deux extrémités du S.

Surajustement, sous-ajustement

Lors de la construction d'un modèle de régression ou de classification, les premières variables ou dimensions introduites dans le modèle apportent une information utile pour la quantification ou la classification. Au fur et à mesure de l'incorporation de variables ou dimensions, l'information utile va diminuer logiquement puisque si le choix est bien fait, la variable ou dimension i doit apporter plus d'information utile que la $i + 1$ mais moins que la $i - 1$. Moins d'information utile, mais plus

de bruit incorporé au modèle. Ainsi, il existe une plage de dimensions dans laquelle se trouvent les modèles optimaux. En dessous de cette plage, les modèles sont dits *sous-ajustés*, en dessus ils sont dits *surajustés* ou *overfitted* en Anglais, comme illustré Figure 8.

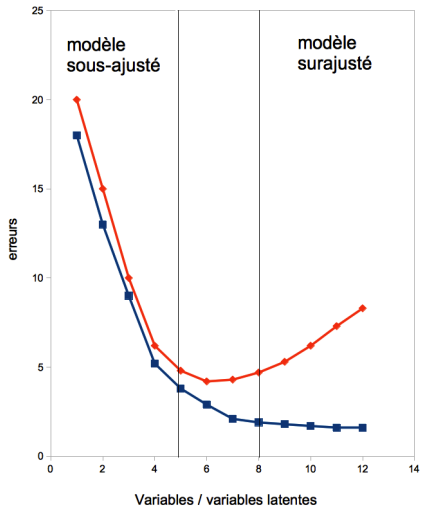

FIGURE 8 – Illustration du sous et du surajustement. La courbe en bleu représente l'erreur standard d'étalonnage ($RMSEC$) et la courbe rouge l'erreur standard de validation croisée ($RMSECV$).

Le problème de surajustement se pose pour tous les modèles de régression et de classification. Il cause des problèmes insurmontables pour appliquer la MLR (régression linéaire multiple) en spectroscopie, puisque les spectres ont forcément beaucoup de variables comparé aux dimensions réelles des modèles, et qu'en plus il s'ajoute au problème de l'inversion de matrices non inversibles ou mal conditionnées.

Sous-espace vectoriel

Un spectre de P longueurs d'onde, ou vecteur de longueur P, est défini dans l'espace vectoriel $\mathbb{R}^P$, de dimension P.
Toutefois, l'ensemble des spectres mesurés sur un produit donné n'occupent pas tout $\mathbb{R}^P$, dans la mesure où les spectres ont des formes semblables et leurs valeurs sont continues (deux valeurs d'ab-

sorbance pour des longueurs d'onde proches sont forcément peu différentes) ce qui interdit certaines combinaisons.

Ces spectres n'occupent qu'une partie de $\mathbb{R}^P$. On dit qu'ils occupent un *sous-espace vectoriel* de $\mathbb{R}^P$. Si ce sous-espace vectoriel est défini avec une base de A vecteurs, sa dimension est A. En général, $A << P$.

Standardisation d'un vecteur

La *standardisation* d'un vecteur consiste à le diviser par son écart-type, de manière à ce que les valeurs du vecteur résultant aient un écart-type de 1. A ne pas confondre avec la normalisation, qui consiste à diviser un vecteur par sa norme.

Transposée d'un vecteur, d'une matrice

La *transposée* transforme des lignes en colonne, et *vice-versa*. On la note avec *prime* ou avec ^{T}

Ex. :

$$\mathbf{x}_1 = \begin{pmatrix} 3.4 & 2.9 & 7.1 \end{pmatrix} \text{ et } \mathbf{x'}_1 = \mathbf{x}_1^T = \begin{pmatrix} 3.4 \\ 2.9 \\ 7.1 \end{pmatrix}$$

$$\mathbf{x}_2 = \begin{pmatrix} 5.7 \\ 0.9 \\ -4.5 \end{pmatrix} \text{ et } \mathbf{x'}_2 = \mathbf{x}_2^T = \begin{pmatrix} 5.7 & 0.9 & -4.5 \end{pmatrix}$$

$$\mathbf{X} = \begin{pmatrix} 5.7 & 4.2 & 6.8 & 5.2 \\ 0.9 & 9.9 & 5.0 & 2.3 \\ -4.5 & 1.1 & 6.3 & 3.7 \end{pmatrix} \text{ et } \mathbf{X'} = \mathbf{X}^T = \begin{pmatrix} 5.7 & 0.9 & -4.5 \\ 4.2 & 9.9 & 1.1 \\ 6.8 & 5.0 & 6.3 \\ 5.2 & 2.3 & 3.7 \end{pmatrix}$$

Notez que $\left(\mathbf{X}^T\right)^T = \mathbf{X}$.

Validation croisée

La *validation croisée* est une méthode de validation qui utilise le jeu d'étalonnage (N observations). Elle est donc utilisée pour construire un modèle, choisir les dimensions, pas pour valider un

modèle. Un bloc d'échantillons est retiré du jeu d'étalonnage. Un modèle est construit avec les données restantes, puis il est utilisé pour prédire les échantillons précédemment retirés. Au total le jeu initial est découpé en K blocs, la procédure est donc répétée K fois. Il existe plusieurs méthodes de sélection des échantillons : (a) un échantillon à la fois ou leave-one-out ($K = N$), (b) partition du jeu d'étalonnage en K blocs gardant l'ordre initial des observations (Jack knife), (c) partition en K blocs dont les observations sont tirées aléatoirement, (d) stores Vénitiens : la première observation est dans le bloc 1, la deuxième dans le bloc 2,...la k^{ieme} dans le bloc K, la $(k+1)^{ieme}$ dans le bloc 1, la $(k+2)^{ieme}$ dans le bloc 2...et ainsi de suite.

Variable latente

Plusieurs méthodes de chimiométrie sont basées sur la *décomposition de matrices*. L'hypothèse est que les spectres sont expliqués par un petit nombre de signaux, qu'on a appelés *loadings*. Ces signaux sont induits par des *variables latentes*. Un exemple simple : un spectre contenant de l'éthanol et du glucose en solution aqueuse sera obtenu à partir des signaux purs de trois variables latentes chimiques : l'eau, l'éthanol et le glucose. Le principe est étendu à des spectres plus complexes, avec obtention de signaux basés uniquement sur des variables latentes mathématiques (ACP, PLSR) ou cherchant à retrouver des variables latentes en relation avec la chimie (ICA, MCR-ALS).

Variance

Soient N valeurs $\{x_i\}$ formant un vecteur $\mathbf{x}$ dont la moyenne est $\bar{x}$. La *variance* est la somme des carrés des écarts entre les x_i et leur moyenne, divisée par N :
Selon l'origine des $\{x_i\}$, la variance pour des valeurs représentant toute une population est :

$$variance = var(\mathbf{x}) = \frac{\sum_{1}^{N}(x_i - \bar{x})^2}{N}$$

Et la variance pour des valeurs constituant un échantillonage tiré d'une population est :

$$variance = var(\mathbf{x}) = \frac{\sum_{1}^{N}(x_i - \bar{x})^2}{N-1}$$

La première formule est la plus utilisée (voir discussion à la définition de la moyenne).
La variance indique si les valeurs sont plutôt proches de leur moyenne (variance faible), ou si elles en sont plutôt éloignées (variance forte).
L'*écart-type* est la racine-carrée de la variance. Il est habituellement noté σ ou s.

Vecteur

Un *vecteur* est un tableau de nombres ne comportant qu'une colonne ou une ligne. Toutefois, la représentation en colonne est privilégiée pour harmoniser les notations.
Ex :
$\mathbf{x}_1 = \begin{pmatrix} 3.4 & 2.9 & 7.1 \end{pmatrix}$ est un vecteur-ligne de dimensions (1×3).
$\mathbf{x}_2 = \begin{pmatrix} 3.4 \\ 2.9 \\ 7.1 \end{pmatrix}$ est un vecteur-colonne de dimensions (3×1).