CheMOOCs

Glossaire

Jean-Claude Boulet
INRA, Montpellier, France

Table des matières

Acronymes utilisés dans CheMoocs **5**

Base d'un espace vectoriel **5**

Carte factorielle **5**

Cercle des corrélations **6**

Coefficient de corrélation **6**

Coefficient de détermination **7**

Coefficients de régression **7**

Combinaison linéaire **7**

Colinéarité de vecteurs **8**

Cosinus **8**

Covariance **8**

Décomposition d'une matrice **9**

Délimiteur décimal **9**

Discret / continu (pour une fonction ou variable) **9**

Distribution d'une variable **10**

Ecart-type **10**

Espace vectoriel **10**

Espace dual **11**

Histogramme **11**

Individu extrême, atypique, outlier **11**

Inertie **12**

Inverse, pseudo-inverse de Moore-Penrose d'une matrice **12**

Jeux d'étalonnage, de validation, de test **12**

Loadings **13**

Matrice **13**

Matrice conjonctive, disjonctive **13**

Matrice de corrélation **14**

Matrice des indicatrices **15**

Matrice identité **15**

Moyenne **15**

Normalisation d'un vecteur **15**

Norme d'un vecteur **15**

Orthogonalité entre vecteurs **16**

Produit scalaire **16**

Projection orthogonale **16**

Robustesse d'un modèle **18**

RMSEC, RMSECV, RMSEP **18**

Scores **18**

Scores plot **18**

Sous-espace vectoriel **19**

Standardisation d'un vecteur **19**

Transposée d'un vecteur, d'une matrice **19**

Variable latente **20**

Variance **20**

Vecteur **20**

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

Une *base* d'un espace vectoriel de dimension P est constituée de P vecteurs : $\{\mathbf{u}_1, \mathbf{u}_2, ...\mathbf{u}_P\}$ linéairement indépendant, c'est à dire qu'aucun ne peut être écrit comme une combinaison linéaire des autres.

De même, on définit une base d'un sous-espace vectoriel de $\mathbb{R}^P$, de dimension A, par 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.

Une base orthonormée ne contient que des vecteurs de norme 1 et tous orthogonaux entre eux. Si la matrice $\mathbf{P}$ de dimensions $(P \times A)$ contient en colonne les vecteurs d'une base orthonormée, alors $\mathbf{P}'\mathbf{P} = \mathbf{I}_A$. Les loadings de l'ACP forment une base orthonormée.

Carte factorielle

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.

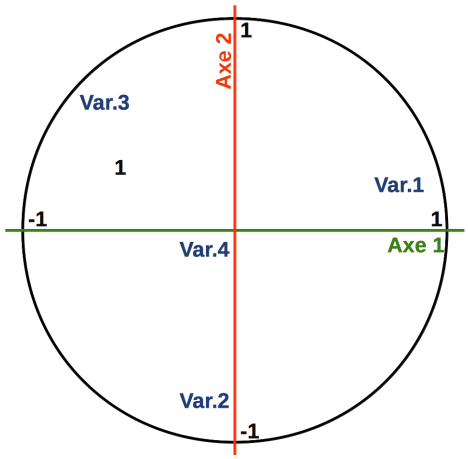

FIGURE 1 – Le cercle des corrélations pour 4 variables : Var1, Var2, Var3 et Var4 représentées sur le plan des composantes principales 1-2, ou axes 1-2.

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. Les deux vecteurs $\mathbf{u}$ et $\mathbf{v}$ sont utilisés pour donner deux directions, sur lesquelles appuient deux cotés d'un triangle rectangle ABC, rectangle en B. Notons $d(A, B)$ et $d(A, C)$ les distances respectives entre A et B et entre A et C. $\overrightarrow{AB}$ et $\overrightarrow{AC}$ sont aussi des vecteurs dont les normes $\|\overrightarrow{AB}\|$ et $\|\overrightarrow{AB}\|$ vérifient : $d(A, B) = \|\overrightarrow{AB}\|$ et $d(A, C) = \|\overrightarrow{AC}\|$. Le cosinus entre $\mathbf{u}$ et $\mathbf{v}$ est calculé ainsi :

$$cov(\mathbf{u}, \mathbf{v}) = \frac{d(A, B)}{d(A, C)} = \frac{\|\overrightarrow{AB}\|}{\|\overrightarrow{AC}\|}$$

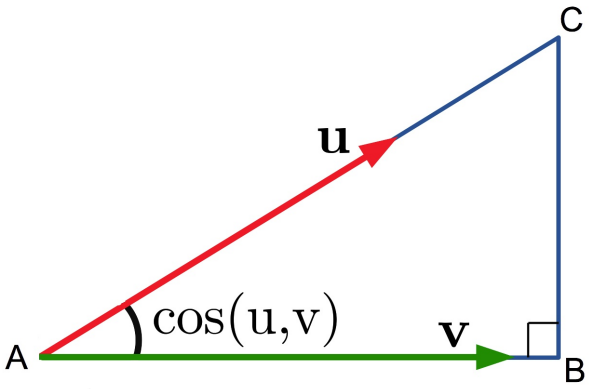

FIGURE 2 – Le cosinus entre deux vecteurs $\mathbf{u}$ et $\mathbf{v}$.

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 : $\pi = 3.14159 = 3,14159$

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.

Ex : 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-.

Distribution d'une variable

La *distribution dune 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

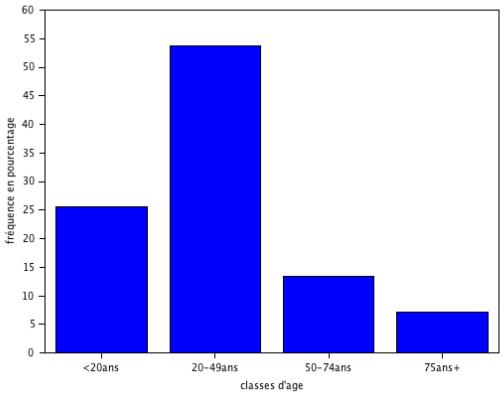

FIGURE 3 – Exemple d'une distribution : la population Française en 2000 (à gauche) et son histogramme (à droite). Source : INSEE

Ecart-type

Voir à *variance*.

Espace vectoriel

Un *espace vectoriel* est défini par des vecteurs munis d'une loi, comme l'addition de vecteurs ou la multiplication par un scalaire.

- 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 : $2\ \mathbf{x}_1 = \begin{pmatrix} 2 * 3.4 & 2 * 2.9 & 2 * 7.1 \end{pmatrix} = \begin{pmatrix} 6.8 & 5.8 & 14.2 \end{pmatrix}$.

Un vecteur pointe dans une direction de l'espace vectoriel.
Le nombre minimum de vecteurs nécessaires pour représenter tout l'espace vectoriel, c'est à dire toutes ses directions, est la dimension P de l'espace vectoriel : ses vecteurs sont définis dans $\mathbb{R}^P$.
Ex.1 : l'espace 3D que nous connaissons est un espace vectoriel de dimension 3, soit $\mathbb{R}^3$.
Ex.2 : une feuille de papier définit un espace vectoriel de dimension 2, soit $\mathbb{R}^2$.
Ex.3 : un spectre de 500 variables est défini dans un espace vectoriel de dimension 500, soit $\mathbb{R}^{500}$.

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.

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*.

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 son centre de gravité est la somme des distances de tous les points au centre de gravité.
Plus les points sont regroupés autour de leur centre de gravité, et plus le nuage de points est petit, compact.

Inverse, pseudo-inverse de Moore-Penrose d'une matrice

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* correspond à une disposition de nombres sous forme d'un tableau, avec des lignes et des colonnes, sans valeurs manquantes. Ex :
$\mathbf{X} = \begin{pmatrix} 3.4 & 2.9 & 7.1 \\ 5.0 & 1.2 & 5.8 \end{pmatrix}$ est une matrice de dimensions (2×3), 2 lignes et 3 colonnes.

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 4 – Rugby, vainqueurs 2005 – 2016 du tournoi des 6 nations

Prenons l'exemple du tableau de la figure 4.
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 5.

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

FIGURE 5 – Exemple de codage conjonctif (à gauche) et disjonctif (à droite) des données de la figure 4

Matrice de corrélation

Une *matrice de corrélation* représente les corrélations entre deux jeux de variables. Supposons qu'un premier jeu contienne les trois variables $\mathbf{x}_1$, $\mathbf{x}_1$ et $\mathbf{x}_1$ 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 particulière, notée $\mathbf{I}_P$ pour une dimension P. Elle est carrée (elle a P lignes et P colonnes), et ne contient que des valeurs 0 à l'exception de sa diagonale secondaire qui ne contient que des 1. Soit une matrice $\mathbf{X}$ de dimensions $(N \times P)$. Alors :
$\mathbf{X}\mathbf{I}_P = \mathbf{I}_N\mathbf{X} = \mathbf{X}$.
Exemple d'une matrice identité :

$$\mathbf{I}_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Moyenne

La *moyenne* de N nombres est la somme des N nombres divisée par N.

Normalisation d'un vecteur

Voir à *norme d'un vecteur*

Norme d'un vecteur

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

$$\|\mathbf{x}\| = \sqrt{\mathbf{x}'\mathbf{x}} = \sqrt{\sum_{i=1}^{N} x_i^2}$$

La *normalisation d'un vecteur* consiste à le diviser par sa norme, pour que le vecteur résultat ait pour norme 1. 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.

Produit scalaire de vecteurs, de matrices

Le *produit scalaire* de deux vecteurs 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 :

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

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

Ex : Le produit scalaire des lignes $\mathbf{X}_1 = \begin{pmatrix} 3.4 & 2.9 & 7.1 \\ 6.4 & 7.2 & 3.9 \end{pmatrix}$ avec les colonnes de $\mathbf{X}_2 = \begin{pmatrix} 4.3 & 1.4 \\ 9.6 & 3.7 \\ 5.5 & 0.9 \end{pmatrix}$ est :

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

Projection orthogonale

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 6. L'objet est constitué

des mains, en 3 dimensions. La projection est réalisée sur le rideau, en 2 dimensions. La direction orthogonale est celle de la lumière, de la bougie vers les mains.

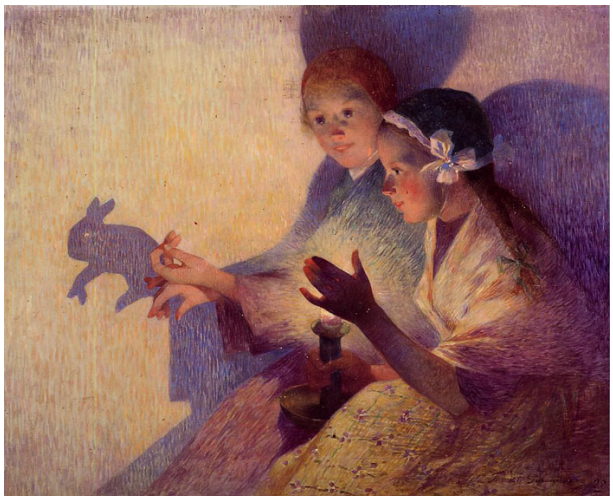

FIGURE 6 – Tableau de Ferdinand Loyen Du Puigaudeau (1864-1930) illustrant une projection orthogonale.

Le résultat est une perte d'information (on ne voit plus sur le drap que ce sont des mains qui sont représentées), 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 :

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

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

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

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

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.
Ces trois indices se calculent avec la même formule, similaire à un écart-type. Si $\mathbf{y}$ et $\widehat{\mathbf{y}}$ sont les vecteurs donnant les valeurs de référence et les valeurs prédites pour N échantillons, alors le RMSE se calcule ainsi :

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

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 :

$$RMSECV_{loo} = \sqrt{\frac{PRESS}{N}}$$

Scores

Voir *Décomposition d'une matrice*.

Scores plot

Voir *Carte factorielle*.

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'absorbance 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. :

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

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

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

La transposée s'applique à l'addition de matrices, au produit de matrices, au produit d'une matrice par un scalaire.

Soient a un scalaire, et $\mathbf{A}$, $\mathbf{B}$ et $\mathbf{C}$ trois matrices telles que le produit $\mathbf{AB}$ et la somme $\mathbf{A}+\mathbf{C}$ soient possibles. Alors :
$(a\mathbf{A})' = (a\mathbf{A})^T = a\mathbf{A}' = a\mathbf{A}^T$
$(\mathbf{AB})' = (\mathbf{AB})^T = \mathbf{B}'\mathbf{A}' = \mathbf{B}^T\mathbf{A}^T$
$(\mathbf{A}+\mathbf{C})' = (\mathbf{A}+\mathbf{C})^T = \mathbf{A}' + \mathbf{C}' = \mathbf{A}^T + \mathbf{C}^T$

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

Soit le vecteur $\mathbf{x}$ contenant N valeurs $\{x_i\}$, leur moyenne étant $\bar{x}$. La *variance* est la somme des carrés des écarts entre les x_i et $\bar{x}$, divisée par N.
En termes mathématiques :

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

En pratique, 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.

Vecteur

Un *vecteur* est un élément d'un espace vectoriel. Il est représenté par une disposition de nombres le long d'une ligne ou bien le long d'un colonne, au choix. Toutefois, la représentation en colonne est privilégiée pour harmoniser les notations. La longueur du vecteur est donnée par le nombre d'élé-

ments, elle donne aussi la dimension de l'espace vectoriel dans lequel le vecteur est défini.

Ex :

$\mathbf{x}_1 = \begin{pmatrix} 3.4 & 2.9 & 7.1 \end{pmatrix}$ est un vecteur-ligne de dimensions (1×3) ou de longueur 3.

$\mathbf{x}_2 = \begin{pmatrix} 3.4 \\ 2.9 \\ 7.1 \end{pmatrix}$ est un vecteur-colonne de dimensions (3×1) ou de longueur 3.

$\mathbf{x}_1$ et $\mathbf{x}_2$ sont deux représentations du même vecteur défini dans un espace vectoriel de dimension 3.

Un spectre acquis sur P longueurs d'ondes est un vecteur de longueur P, défini dans un espace vectoriel de dimension P.