Deuxième partie

Analyse exploratoire

Chapitre 5 Réduction de dimension

Notions : sélection de variables ; extraction de variables ; analyse en composantes principales ; analyse en composantes principales probabiliste.

Objectifs pédagogiques :

- Expliquer l'intérêt de réduire la dimension d'un jeu de données ;
- Faire la différence entre la sélection de variables et l'extraction de variables ;
- Projeter des données sur un espace de plus petite dimension ;
- Mettre en œuvre des méthodes d'extraction de variables.

5.1 Des séries statistiques aux jeux de données

Nous avons jusqu'à présent travaillé sur des séries statistiques contenant une seule variable. Cependant, dans la majorité des problèmes de sciences des données, nous disposons de plusieurs variables pour décrire chaque individu.

L'objet de nos études, à savoir le jeu de données, n'est donc plus un échantillon $(x_1, x_2, \ldots, x_n)$ d'une variable aléatoire rélle X, mais un échantillon d'un vecteur aléatoire à valeurs dans un espace $\mathcal{X}$. Nous considèrerons en général que $\mathcal{X} = \mathbb{R}^p$ et que notre jeu de données peut être décrit par une matrice $X \in \mathcal{X}^{n \times p}$. C'est par exemple la matrice de taille 31×8 des entrées du tableau 1.1.

Cela suppose que nous disposions d'une représentation p-dimensionnelle pertinente de nos données. Si celle-ci est assez évidente pour des données comme celles du tableau 1.1, ce n'est pas toujours le cas. En particulier, les variables qualitatives (comme la colonne « âge » du tableau 1.2) doivent être représentées par un (ou plusieurs) nombres réels. Nous verrons comment faire en pratique dans la PC3.

Enfin, nous supposons dans ce cours que nos données sont **structurées**, c'est-à-dire présentées sous forme vectorielle. Ce n'est pas le cas de nombreux types de données telles que du texte, des images, des séquences d'ADN, ou des molécules chimiques. La question de la représentation de ces données dites non-structurées dépasse le cadre de ce cours mais est très importante.

5.2 Notations

Nous essaierons à partir de maintenant de nous en tenir aux notations suivantes :

— Les lettres minuscules (x) représentent un scalaire ;

— les lettres minuscules surmontées d'une flèche ($\vec{x}$) représentent un vecteur ;

— les lettres majuscules (X) représentent une matrice, un événement ou une variable aléatoire ;

— les lettres calligraphiées ($\mathcal{X}$) représentent un ensemble ou un espace ;

— les *indices* correspondent à une variable tandis que les *exposants* correspondent à une observation : x_j^i est la j-ème variable de la i-ème observation, et correspond à l'entrée X_{ij} de la matrice X ;

— n est un nombre d'observations et p un nombre de variables.

5.3 Motivation ★

Le but de la réduction de dimension est de transformer une représentation $X \in \mathbb{R}^{n\times p}$ des données en une représentation $X^* \in \mathbb{R}^{n\times m}$ où $m \ll p$. Les raisons de cette démarche sont multiples.

Visualiser les données. Ce n'est pas tâche aisée avec un nombre très grand de variables. Comment visualiser n points en plus de 2 ou 3 dimensions ? Limiter les variables à un faible nombre de dimensions permet de visualiser les données plus facilement, quitte à perdre un peu d'information lors de la transformation.

Réduire les coûts algorithmiques du traitement des données. D'un point de vue purement computationnel, réduire la dimension des données permet de réduire d'une part l'espace qu'elles prennent en mémoire et d'autre part les temps de calcul. De plus, si certaines variables sont inutiles, ou redondantes, il n'est pas nécessaire de les obtenir pour de nouvelles observations : cela permet de réduire le coût d'acquisition des données.

Améliorer la qualité du traitement des données. Les algorithmes d'apprentissage supervisé ou de clustering sont généralement plus performants sur un faible nombre de variables. En effet, si certaines des variables ne sont pas pertinentes, elles risquent de biaiser les modèles appris.
De plus, les raisonnements développés en faible dimension pour construire un algorithme d'apprentissage supervisé ne s'appliquent pas nécessairement en haute dimension. C'est un phénomène connu sous le nom de **fléau de la dimension**, ou *curse of dimensionality* en anglais.
En effet, en haute dimension, les individus ont tendance à tous être éloignés les uns des autres. Pour comprendre cette assertion, plaçons-nous en dimension p et considérons l'hypersphère $\mathcal{S}(\vec{x},R)$ de rayon $R \in \mathbb{R}_+^*$ centrée sur une observation $\vec{x}$, ainsi que l'hypercube $\mathcal{C}(\vec{x},R)$ circonscrit à cette hypersphère. Le volume de $\mathcal{S}(\vec{x})$ vaut $\frac{2R^p\pi^{p/2}}{p\Gamma(p/2)}$, tandis que celui de $\mathcal{C}(\vec{x},R)$, dont le côté a pour longueur $2R$, vaut 2^pR^p. Ainsi

$$\lim_{p\to\infty} \frac{\text{Vol}(\mathcal{S}(\vec{x},R))}{\text{Vol}(\mathcal{C}(\vec{x},R))} = 0.$$

Cela signifie que la probabilité qu'un exemple situé dans $\mathcal{C}(\vec{x},R)$ appartienne à $\mathcal{S}(\vec{x},R)$, qui vaut $\frac{\pi}{4} \approx 0.79$ lorsque $p = 2$ et $\frac{\pi}{6} \approx 0.52$ lorsque $p = 3$, devient très faible quand p est grand : les données ont tendance à être éloignées les unes des autres.

Deux possibilités s'offrent à nous pour réduire la dimension de nos données :

— la **sélection de variables**, qui consiste à *éliminer* un nombre $(p-m)$ de variables de nos données ;

— l'**extraction de variables**, qui consiste à *créer* m nouvelles variables à partir des p variables dont nous disposons initialement.

5.4 Sélection de variables ★

La sélection de variables consiste à éliminer des variables peu informatives.

Dans le cas non-supervisé, il s'agit par exemple d'éliminer des variables

— dont la variance est très faible : leur valeur étant à peu près la même pour chaque individu, elle n'apporte aucune information permettant de distinguer deux individus ;

— qui sont corrélées à une autre variable : elles apportent alors la même information et sont redondantes.

Dans le cas supervisé, il s'agit aussi d'éliminer des variables qui ne sont pas pertinentes par rapport à la tâche de prédiction. On peut par exemple

— éliminer, par exemple à l'aide d'un test du chi2 comme vu dans la PC1, les variables indépendantes de l'étiquette à prédire. Remarquez néanmoins que deux variables chacune indépendante de l'étiquette peuvent être très informatives quand on les considère simultanément. Considérez par exemple, pour $\mathcal{X} = \{0,1\}^2$, un problème de classification binaire dans lequel l'étiquette y est donnée par $y = x_1 \oplus x_2$: les deux variables ensemble déterminent parfaitement y, mais chacune d'entre elle est uninformative ;

— chercher à éliminer des variables qui n'améliorent pas la performance d'un algorithme précis.

Nous reviendrons sur la sélection de variables supervisée quand nous parlerons du lasso (section 8.6).

5.5 Analyse en composantes principales ★

La méthode la plus classique pour réduire la dimension d'un jeu de données par extraction de variables est l'**analyse en composantes principales**, ou *ACP*. On parle aussi souvent de *PCA*, de son nom anglais *Principal Component Analysis*.

5.5.1 Maximisation de la variance

L'idée est de représenter les données de sorte à maximiser leur variance selon les nouvelles dimensions. Cela permet de pouvoir continuer à distinguer les individus les uns des autres dans leur nouvelle représentation (cf. figure 5.1). Ainsi, une ACP de la matrice $X \in \mathbb{R}^{n \times p}$ est une transformation linéaire orthogonale qui permet d'exprimer X dans une nouvelle base orthonormée, de sorte que la plus grande variance de X par projection s'aligne sur le premier axe de cette nouvelle base, la seconde plus grande variance sur le deuxième axe, et ainsi de suite. Les axes de cette nouvelle base sont appelés les **composantes principales**, abrégées en PC pour *Principal Components*.

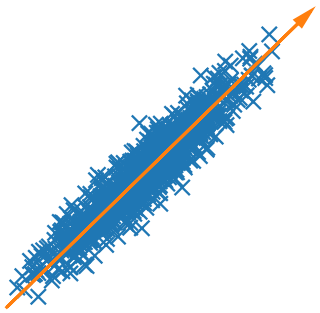


FIGURE 5.1 – La variance des données en deux dimensions est maximale selon l'axe indiqué par la flèche.

5.5.2 Standardisation

Dans la suite de cette section, nous supposons que les variables ont été **standardisées** de sorte à toutes avoir une moyenne de 0 et une variance de 1, pour éviter que les variables qui prennent de grandes valeurs aient plus d'importance que celles qui prennent de faibles valeurs. C'est un pré-requis de l'application de l'ACP. Cette standardisation s'effectue par :

$$x_j^i \leftarrow \frac{x_j^i - \overline{x_j}}{\sqrt{\frac{1}{n}\sum_{l=1}^{n}(x_j^l - \overline{x_j})^2}}, \tag{5.1}$$

où $\overline{x_j} = \frac{1}{n}\sum_{l=1}^{n} x_j^l$. On dira alors que X est **centrée** : chacune de ses colonnes a pour moyenne 0 et **réduite** : chacune de ses colonnes a pour variance 1.

Exemple

Considérons la matrice de données

$$X = \begin{bmatrix} 1.0 & 20.0 \\ 2.0 & 10.0 \\ 3.0 & 50.0 \\ 4.0 & 30.0 \\ 5.0 & 40.0 \end{bmatrix}.$$

La variance de la première colonne vaut 2.0 tandis que celle de la deuxième colonne vaut 200.0. Peut-on pour autant en conclure que la deuxième variable « varie » plus que la première, alors que les valeurs qu'elle prend sont simplement proportionnelles à celles prises par la première ?

La version standardisée de X est

$$\begin{bmatrix} -1.414 & -0.707 \\ -0.707 & -1.414 \\ 0.0 & 1.414 \\ 0.707 & 0.0 \\ 1.414 & 0.707 \end{bmatrix}.$$

5.5.3 Décomposition spectrale de la covariance

Proposition Soit $X \in \mathbb{R}^{n\times p}$ une matrice centrée de covariance empirique $\Sigma = \frac{1}{n}X^\top X$. Les composantes principales de X sont les vecteurs propres de Σ, ordonnés par valeurs propres décroissantes.

Preuve Considérons un vecteur $\vec{w} \in \mathbb{R}^p$. La projection de X sur $\vec{w}$ est le vecteur $X\vec{w} \in \mathbb{R}^n$. La moyenne de $X\vec{w}$ vaut 0 car les variables $(\vec{x}_1, \ldots, \vec{x}_p)$ sont elles-mêmes de moyenne nulle (X étant centrée). La variance de $X\vec{w}$ vaut alors

$$\mathrm{Var}(X\vec{w}) = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p} x_j^i w_j\right)^2 = \vec{w}^{\,\top}\Sigma\vec{w}.$$

Appelons maintenant $\vec{w}_1 \in \mathbb{R}^p$ la première composante principale de X. $\vec{w}_1$ est de norme 1 et tel que la variance de $X\vec{w}_1$ est maximale :

$$\vec{w}_1 \in \underset{\vec{w}\in\mathbb{R}^p}{\arg\max}\left(\vec{w}^{\,\top}\Sigma\vec{w}\right) \qquad \text{avec } \|\vec{w}_1\|_1{}^2 = 1. \tag{5.2}$$

Il s'agit d'un problème d'optimisation quadratique sous contrainte d'égalité, que l'on peut résoudre

(cf section 2.2.1 du poly d'Optimisation) en introduisant le multiplicateur de Lagrange $\alpha_1 > 0$ et en écrivant le lagrangien

$$L(\alpha_1,\vec{w}) = \vec{w}^{\top}\Sigma\vec{w} - \alpha_1\left(\|\vec{w}\|_1{}^2 - 1\right).$$

Le maximum de $\vec{w}^{\top}\Sigma\vec{w}$ sous la contrainte $\|\vec{w}_1\|_1 = 1$ est égal à $\min_{\alpha_1}\sup_{\vec{w}\in\mathbb{R}^p} L(\alpha_1,\vec{w})$. Le supremum du lagrangien est atteint en un point où son gradient s'annule, c'est-à-dire qui vérifie

$$2\Sigma\vec{w} - 2\alpha_1\vec{w} = 0.$$

Ainsi, $\Sigma\vec{w}_1 = \alpha_1\vec{w}_1$ et $(\alpha_1,\vec{w}_1)$ forment un couple (valeur propre, vecteur propre) de Σ.

Parmi tous les vecteurs propres de Σ, $\vec{w}_1$ est celui qui maximise la variance $\vec{w}_1^{\top}\Sigma\vec{w}_1 = \alpha_1\|\vec{w}_1\|_1 = \alpha_1$. Ainsi, α_1 est la plus grande valeur propre de Σ (rappelons que Σ étant définie par $X^{\top}X$ est semi-définie positive et que toutes ses valeurs propres sont positives.)

La deuxième composante principale de X vérifie

$$\vec{w}_2 = \arg\max_{\vec{w}\in\mathbb{R}^p}\left(\vec{w}^{\top}\Sigma\vec{w}\right) \qquad \text{avec } \|\vec{w}_2\|_1{}^2 = 1 \text{ et } \vec{w}^{\top}\vec{w}_1 = 0. \tag{5.3}$$

Cette dernière contrainte nous permet de garantir que la base des composantes principales est orthonormée.

Nous introduisons donc maintenant deux multiplicateurs de Lagrange $\alpha_2 > 0$ et $\beta_2 > 0$ et obtenons le lagrangien

$$L(\alpha_2,\beta_2,\vec{w}) = \vec{w}^{\top}\Sigma\vec{w} - \alpha_2\left(\|\vec{w}\|_1{}^2 - 1\right) - \beta_2\vec{w}^{\top}\vec{w}_1.$$

Comme précédemment, son supremum en $\vec{w}$ est atteint en un point où son gradient s'annule :

$$2\Sigma\vec{w}_2 - 2\alpha_2\vec{w}_2 - \beta_2\vec{w}_1 = 0.$$

En multipliant à gauche par $\vec{w}_1^{\top}$, on obtient

$$2\vec{w}_1^{\top}\Sigma\vec{w}_2 - 2\alpha_2\vec{w}_1^{\top}\vec{w}_2 - \beta_2\vec{w}_1^{\top}\vec{w}_1 = 0$$

d'où l'on conclut que $\beta_2 = 0$ et, en remplaçant dans l'équation précédente, que, comme pour $\vec{w}_1$, $2\Sigma\vec{w}_2 - 2\alpha_1\vec{w}_2 = 0$. Ainsi $(\alpha_2,\vec{w}_2)$ forment un couple (valeur propre, vecteur propre) de Σ et α_2 est maximale : il s'agit donc nécessairement de la deuxième valeur propre de Σ.

Le raisonnement se poursuit de la même manière pour les composantes principales suivantes. □

Preuve alternative Alternativement, on peut prouver ce théorème en observant que Σ, étant définie positive, est diagonalisable par un changement de base orthonormée : $\Sigma = Q^{\top}\Lambda Q$, où $\Lambda \in \mathbb{R}^{p\times p}$ est une matrice diagonale dont les valeurs diagonales sont les valeurs propres de Σ. Ainsi,

$$\vec{w}_1^{\top}\Sigma\vec{w}_1 = \vec{w}_1^{\top}Q^{\top}\Lambda Q\vec{w}_1 = (Q\vec{w}_1)^{\top}\Lambda\,(Q\vec{w}_1).$$

Si l'on pose $\vec{v} = Q\vec{w}_1$, il s'agit donc pour maximiser $\vec{w}_1^{\top}\Sigma\vec{w}_1$ de trouver $\vec{v}$ de norme 1 (Q étant orthonormée et $\vec{w}_1$ de norme 1) qui maximise

$$\sum_{j=1}^{p} v_j^2\lambda_j.$$

Pour tout $j = 1,\ldots,p$, on a $\lambda_j \geq 0$ (car Σ est définie positive) et $0 \leq v_j^2 \leq 1$ car $\|\vec{v}\|_1 = 1$. Ainsi,

$$\sum_{j=1}^{p} v_j^2\lambda_j \leq \left(\max_{j=1,\ldots,p}\lambda_j\right)\sum_{j=1}^{p} v_j^2 \leq \max_{j=1,\ldots,p}\lambda_j,$$

et ce maximum est atteint quand $v_j = 1$ et $v_k = 0\ \forall k \neq j$. On retrouve ainsi que $\vec{w}_1$ est le vecteur propre correspondant à la plus grande valeur propre de Σ, et ainsi de suite. □

5.5.4 Décomposition en valeurs singulières

Proposition Soit $X \in \mathbb{R}^{n \times p}$ une matrice centrée. Les composantes principales de X sont ses vecteurs singuliers à droite ordonnés par valeur singulière décroissante.

Preuve Factorisons X sous la forme $UDV^\top$ avec $U \in \mathbb{R}^{n \times n}$ et $V \in \mathbb{R}^{p \times p}$ orthogonales, et $D \in \mathbb{R}^{n \times p}$ diagonale. Alors

$$n\Sigma = X^\top X = VDU^\top UDV^\top = VD^2V^\top$$

et les valeurs singulières de X (les entrées de D) sont les racines carrées des valeurs propres de $n\Sigma$, tandis que les vecteurs singuliers à droite de X (les colonnes de V) sont les vecteurs propres de $n\Sigma$. □

En pratique, les implémentations de la décomposition en valeurs singulières (ou SVD) sont numériquement plus stables que celles de décomposition spectrale, et c'est ainsi que l'ACP est implémentée.

5.5.5 Choix du nombre de composantes principales

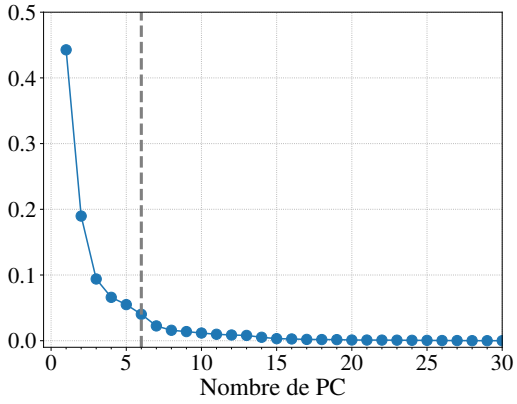
Réduire la dimension des données par une ACP implique de *choisir* un nombre de composantes principales à conserver. Pour ce faire, on utilise la **proportion de variance expliquée** par ces composantes : la variance de X s'exprime comme la trace de Σ, qui est elle-même la somme de ses valeurs propres.

Ainsi, si l'on décide de conserver les m premières composantes principales de X, la proportion de variance qu'elles expliquent est

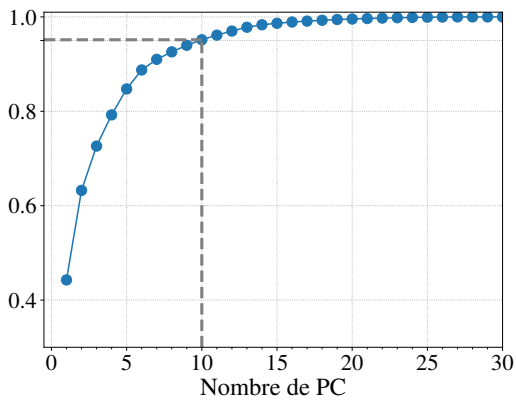
$$\frac{\alpha_1 + \alpha_2 + \cdots + \alpha_m}{\mathrm{Tr}(\Sigma)}$$

où $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_p$ sont les valeurs propres de Σ par ordre décroissant.

Il est classique de s'intéresser à l'évolution, avec le nombre de composantes, soit de la proportion de variance expliquée par chacune d'entre elles, soit à cette proportion cumulée. On peut représenter visuellement ces proportions sur un *scree plot* (figure 5.2), utilisé pour déterminer le nombre de composantes qui expliquent ensemble un pourcentage de la variance fixé a priori (95% sur la figure 5.2b), ou le nombre de composantes à partir duquel ajouter une nouvelle composante n'est plus informatif (« coude » sur la figure 5.2a).



(A) Pourcentage de variance expliqué par chacune des composantes principales. À partir de 6 composantes principales, ajouter de nouvelles composantes n'est plus vraiment informatif.



(B) Pourcentage cumulé de variance expliquée par chacune des composantes principales. Si on se fixe une proportion de variance expliquée de 95%, on peut se contenter de 10 composantes principales.

FIGURE 5.2 – Choix du nombre de PC à l'aide du pourcentage de variance expliquée.

5.6 Factorisation de la matrice des données ★

Soit $W \in \mathbb{R}^{p\times p}$ la matrice de toutes les composantes principales de $X \in \mathbb{R}^{n\times p}$. Posons $m < p$ le nombre de composantes principales choisies, et $\widetilde{W} \in \mathbb{R}^{p\times m}$ la matrice des m premières composantes principales de X. La nouvelle représentation dans $\mathbb{R}^m$ d'un individu $\vec{x} \in \mathbb{R}^p$ est donnée par sa projection sur $(\vec{w}_1,\vec{w}_2,\ldots,\vec{w}_m)$:

$$\vec{h} = \vec{x}^{\top}\widetilde{W}. \tag{5.4}$$

On obtient la représentation m-dimensionnelle des n individus de X par

$$\widetilde{H} = X\widetilde{W}. \tag{5.5}$$

La matrice $\widetilde{H} \in \mathbb{R}^{n\times m}$ peut être interprétée comme une **représentation latente** (ou cachée, *hidden* en anglais d'où la notation H) des données. C'est cette représentation que l'on a cherché à découvrir grâce à l'ACP.

5.6.1 Erreur de reconstruction

Si on utilise toutes les composantes, la représentation latente de X est donnée par

$$H = XW; H \in \mathbb{R}^{n\times p}. \tag{5.6}$$

Les colonnes de W étant des vecteurs orthonormés (il s'agit de vecteurs propres de $X^{\top}X$), on peut multiplier l'équation (5.6) à droite par W pour obtenir une factorisation de X :

$$X = HW^{\top}. \tag{5.7}$$

En se restreignant à $m < p$ composantes, la multiplication à droite par $\widetilde{W}^{\top}$ de la représentation latente H est une approximation de X :

$$Z = \widetilde{H}\widetilde{W}^{\top}. \tag{5.8}$$

$Z \in \mathbb{R}^{n\times p}$ peut être interprétée comme une **reconstruction** des données dans $\mathbb{R}^p$ à partir de leur représentation latente dans $\mathbb{R}^m$.

On peut alors calculer l'**erreur de reconstruction** comme la somme des carrés des distances entre les individus $\vec{x}^{\,i}$ et leur reconstruction $\vec{z}^{\,i}$:

$$\text{Err}_m = \sum_{i=1}^{n} \|\vec{x}^{\,i} - \vec{z}^{\,i}\|^2. \tag{5.9}$$

L'erreur de reconstruction vaut

$$\text{Err}_m = \sum_{i=1}^{n} \left\| \sum_{j=1}^{p} H_{ij}\vec{w}_j - \sum_{j=1}^{m} H_{ij}\vec{w}_j \right\|^2 = \sum_{i=1}^{n} \left\| \sum_{j=m+1}^{p} H_{ij}\vec{w}_j \right\|^2 = \sum_{i=1}^{n}\sum_{j=m+1}^{p} H_{ij}^2,$$

cette dernière égalité venant de ce que les vecteurs $\vec{w}_j$ sont orthogonaux et de norme 1. Ainsi, l'erreur de reconstruction est la somme des carrés des coefficients des dimensions qui n'ont pas été prises en compte.

Comme $H = XW$, on peut réécrire l'erreur de reconstruction comme

$$\text{Err}_m = \sum_{i=1}^{n}\sum_{j=m+1}^{p} \vec{w}_j^{\top}\vec{x}^{\,i}\vec{x}^{\,i\top}\vec{w}_j = \sum_{j=m+1}^{p} \vec{w}_j\Sigma\vec{w}_j^{\top}.$$

Ainsi, maximiser la variance $\sum_{j=1}^{m} \vec{w}_j\Sigma\vec{w}_j^{\top}$ est équivalent à minimiser l'erreur de reconstruction car $\sum_{j=1}^{p} \vec{w}_j\Sigma\vec{w}_j^{\top} = \text{trace}(\Sigma)$. C'est une autre justification de l'ACP.

5.6.2 Analyse factorielle

L'équation (5.7) s'inscrit dans le cadre plus général de **l'analyse factorielle**. Il correspond à considérer que les données sont les réalisations d'un vecteur aléatoire $(X_1,X_2,\dots,X_p)$ obtenues par

$$(X_1,X_2,\dots,X_p) = W(H_1,H_2,\dots,H_m) + \epsilon, \tag{5.10}$$

où $(H_1,H_2,\dots,H_m)$ est le vecteur aléatoire latent qui génère les données et ϵ un bruit gaussien : $\epsilon \sim \mathcal{N}(0,\Psi)$, avec $\Psi \in \mathbb{R}^{p\times p}$.

Supposons maintenant que $(H_1,H_2,\dots,H_m)$ est un vecteur aléatoire gaussien m-dimensionnel, d'espérance 0 (les variables latentes sont elles aussi centrées) et de covariance I_m où I_m est la matrice identité de dimensions $m \times m$. Alors $(X_1,X_2,\dots,X_p)$ est lui-même un vecteur aléatoire gaussien, d'espérance nulle et de covariance $WW^\top + \Psi$.

Si l'on suppose de plus que ϵ est un bruit isotropique, autrement dit que $\Psi = \sigma^2 I_p$, alors

$$(X_1,X_2,\dots,X_p) \sim \mathcal{N}(0,WW^\top + \sigma^2 I_p).$$

On peut alors estimer les paramètres W et σ^2 par maximum de vraisemblance ; c'est ce qu'on appelle l'**ACP probabiliste**.

L'ACP que nous venons de voir est un cas limite de l'ACP probabiliste, obtenu quand la covariance du bruit devient infiniment petite ($\sigma^2 \to 0$) [1].

On peut plutôt faire la supposition plus générale que Ψ est une matrice diagonale. Les valeurs de W et Ψ peuvent une fois de plus être obtenues par maximum de vraisemblance. C'est ce que l'on appelle **l'analyse factorielle**. Dans l'analyse factorielle, les composantes principales (les colonnes de W) ne sont pas nécessairement orthogonales. En particulier, il est donc possible d'obtenir des composantes dégénérées, autrement dit des colonnes de W dont toutes les coordonnées sont 0.

Pour aller plus loin

- Une variante populaire de l'analyse factorielle est la **factorisation positive de matrice** (ou NMF pour *non-negative matrix factorisation*), qui permet lorsque toutes les entrées de X sont positives, de chercher à la décomposer sous la forme HW où H et W ont elles aussi toutes leurs entrées positives. Cela facilite leur interprétation.
- Il existe de nombreuses approches de réduction de dimension non-linéaires, c'est-à-dire qui créent des composantes qui ne sont pas des composantes linéaires des variables initiales. Parmi elles :
  - le **positionnement multidimensionnel**, ou MDS pour *multidimensional scaling*, qui cherche à préserver la distance entre les individus. Dans le cas de la distance euclidienne, on se ramène à l'ACP ; mais il est possible d'utiliser d'autres distances, y compris des distances non-métriques.
  - le **t-SNE** (prononcé « ti-sni »), pour *t-Student Neighborhood Embedding*, qui cherche à approcher la loi des distances entre individus par une loi de Student.
  - le **UMAP**, pour *Uniform Manifold Approximation and Projection* qui suppose les individus uniformément distribués sur une variété riemanienne qu'il s'agit d'approcher.
- Enfin, nous verrons au chapitre 9 que la dernière couche cachée d'un réseau de neurones profond peut être considérée comme une nouvelle représentation des données prises en entrée par ce réseau de neurones. On parle ainsi parfois d'apprentissage de représentation (*representation learning*) plutôt que d'apprentissage profond.

1. Vous en trouverez la preuve dans l'article *Probabilistic principal components analysis*, M. E. Tipping & C. M. Bishop, Journal of the Royal Statistical Society Series B, 61 :611–622 (1999).