04/2023:52100

5.21. MÉTHODES CHIMIOMÉTRIQUES APPLIQUÉES AUX DONNÉES ANALYTIQUES

Le chapitre qui suit est publié à titre d'information. Il constitue une introduction à l'utilisation des techniques de chimiométrie et de science des données pour le traitement des données analytiques. L'objectif est d'apporter des indications sur les bonnes pratiques et exigences attachées à l'analyse des données.

1. GÉNÉRALITÉS

1-1. *INTRODUCTION*

1-1-1. **Objet du chapitre**

Le présent chapitre constitue une introduction à l'utilisation de techniques chimiométriques pour le traitement des données analytiques. Ces techniques sont applicables en recherche, contrôle qualité ou fabrication dans le domaine pharmaceutique. L'objectif est ici d'apporter des indications sur les bonnes pratiques et exigences attachées à la chimiométrie, et de présenter une sélection de techniques bien établies de chimiométrie et de science des données, mais non de traiter de façon exhaustive l'ensemble des techniques existantes, celles-ci ne cessant d'évoluer et de se perfectionner. Une brève description du principe de chacune des méthodes sélectionnées est fournie, avec leurs principales applications, aspects critiques et limitations, sans entrer toutefois dans le détail des algorithmes mathématiques. Un glossaire est proposé en fin de chapitre.

1-1-2. **Définition**

La chimiométrie est définie comme une discipline chimique qui utilise les mathématiques et les statistiques et la logique formelle pour (a) concevoir ou sélectionner des procédures expérimentales à performance optimale, (b) extraire des données analysées le maximum d'informations chimiques pertinentes et (c) obtenir des connaissances sur les systèmes chimiques.

Considérée d'un point de vue plus général, la chimiométrie n'est pas seulement un outil d'analyse des données chimiques, mais un instrument de compréhension des systèmes selon une approche empirique, lorsque l'on ne peut s'appuyer sur les connaissances ou théories disponibles pour expliquer des observations ou comportements. L'approche chimiométrique s'appuie uniquement sur les données recueillies, en leur appliquant des techniques de modélisation multivariée qui conduisent à l'établissement de modèles numériques ensuite utilisés pour la prédiction indirecte des propriétés visées. Les algorithmes mathématiques utilisés en chimiométrie sont communs à d'autres domaines de la science des données, comme l'apprentissage automatique (« *machine learning* ») ou la fouille de données (« *data mining* »).

1-1-3. **Contexte**

La chimiométrie possède de nombreuses applications, tant qualitatives que quantitatives. Elle peut aider l'analyste à structurer les données recueillies et à faire apparaître des relations cachées entre les variables au sein du système. Cependant, il est important de souligner que cette approche empirique, aussi puissante soit-elle, ne peut remplacer une théorie vérifiée ou établie, lorsqu'elle existe.

La chimiométrie a révolutionné la spectroscopie dans le proche infrarouge (SPIR) (*2.2.40*), avant de devenir un outil indispensable du contrôle analytique des procédés (PAT, pour « *process analytical technology* ») (*5.25*) et de la qualité par la conception (QbD, pour « *quality by design* ») pour l'amélioration du contrôle des procédés et du contrôle de la qualité des produits en général. Les méthodes chimiométriques sont aujourd'hui utilisées dans de très nombreux domaines scientifiques et techniques, notamment dans les sciences de la vie et de la santé – agroalimentaire, pharmacie, chimie, biochimie, génomique – mais aussi pour certaines applications industrielles dans les domaines de la pétrochimie, du textile, des mesures physiques, des cosmétiques, et leur champ d'application ne cesse de s'étendre.

Les principes mathématiques associés sont connus depuis le début du XXe siècle, mais la chimiométrie est réellement née avec le développement des technologies numériques et les progrès réalisés dans l'élaboration d'algorithmes mathématiques. Les techniques et méthodes qu'elle utilise reposent souvent sur la représentation géométrique, la transformation et la modélisation des données, même si elles ont ultérieurement bénéficié de développements mathématiques et théoriques.

L'expansion de la chimiométrie au-delà de ses premiers domaines d'application s'appuie sur l'évolution des algorithmes d'analyse des données, la multiplicité des sources de données et des capteurs, la facilité et la rapidité d'accès à un nombre croissant d'ensembles de données. La quantité des signaux et des informations pouvant être collectés atteint désormais des niveaux qui relèvent du modèle des mégadonnées (« *big data* »).

La fouille de données est une autre activité apparentée qui a pour but d'extraire des modèles et des connaissances à partir de grandes quantités de données existantes, souvent collectées à d'autres fins que la modélisation chimiométrique. Les algorithmes chimiométriques sont adaptés à l'exploration de données, car ils permettent de capter des variations provenant de toutes les sources pouvant être corrélées (variables colinéaires). Les considérations énoncées ci-après dans la section 1-2-2. Données s'appliquent aussi à la fouille de données pour assurer une bonne qualité d'information.

1-1-4. **Introduction à la chimiométrie**

En chimiométrie, contrairement à la chimie analytique classique, l'évaluation d'une propriété s'opère exclusivement sur la base des informations livrées par les échantillons examinés. Des algorithmes sont appliqués directement aux données, et l'utilisation de modèles permet d'extraire les informations recherchées (étape de modélisation ou étalonnage). La chimiométrie est associée à l'analyse multivariée, une méthode statistique qui, contrairement à beaucoup d'autres, dépend généralement peu d'hypothèses sur la distribution des données, puisqu'elle ne fait que rarement intervenir des tests sur de telles hypothèses. Dans le cadre de la modélisation, les variations les plus critiques des propriétés étudiées peuvent être amplifiées, tandis que les moins pertinentes, dues à des facteurs de perturbation (qu'elles soient d'origine physique, chimique, expérimentale ou instrumentale), sont ramenées à un minimum.

En chimiométrie, un modèle n'est pas une représentation formelle ou simplifiée d'un phénomène de nature physique, chimique, etc. Pour caractériser la performance d'un modèle, son aptitude à prédire les propriétés visées doit être évaluée. Le modèle (ou étalonnage) optimal est celui qui fournira la meilleure estimation de ces propriétés ou qui en tirera les meilleures conclusions. Pour être utile, un modèle doit être fiable et pouvoir être utilisé pour la prise de décision, par exemple. L'adoption d'un modèle pour un processus de prise de décision doit reposer sur des procédures d'évaluation acceptables, fiables et bien comprises.

L'analyse univariée considère individuellement les variables mesurées dans un système. Dans la réalité, cependant, on a souvent affaire à des systèmes complexes, où interviennent entre les variables des interactions et effets de combinaison qu'il est impossible de séparer. L'analyse multivariée traite simultanément de nombreuses variables, et recompose les relations à l'intérieur des ensembles de données, ou entre ces ensembles (typiquement présentés sous forme de matrices),

pour en extraire l'information. Des combinaisons linéaires sont souvent appliquées aux données initiales pour en capturer autant que possible la partie explicable (la partie dite « déterministe ») et, dans l'idéal, le bruit n'est pas inclus au modèle. Dès lors qu'il est convenablement validé, le modèle peut être utilisé pour prédire de nouvelles valeurs et se substituer aux mesures matérielles chronophages.

Des techniques de projection comme l'analyse en composantes principales (ACP), la régression sur composantes principales (RCP) et la régression des moindres carrés partiels (PLS, pour « *partial least squares* ») sont recommandées. L'approche suivie diffère cependant selon que les données ont été générées à l'aide de plans d'expériences (DoE, pour « *design of experiments* »), ou collectées de façon aléatoire à partir d'une population donnée (données non organisées). Dans le premier cas, les variables sont orthogonales par construction, et les méthodes statistiques telles que la régression linéaire multiple (RLM) sont donc bien adaptées à la description des données. En revanche, dans le cas de matrices de données non organisées, les variables sont rarement orthogonales mais plutôt colinéaires, et l'analyse multivariée constitue alors l'outil de choix, notamment avec des données spectroscopiques.

1-1-5. **Analyse qualitative et quantitative des données**

L'analyse quantitative des données consiste principalement à effectuer un étalonnage puis à l'appliquer à des échantillons nouveaux et inconnus ; l'étalonnage est l'opération par laquelle peut être estimée la relation mathématique qui lie les propriétés d'intérêt (concentration par exemple) avec les variables mesurées.

L'analyse qualitative des données, d'autre part, peut être divisée en deux types : l'analyse dite « exploratoire » et la classification.

En analyse exploratoire, la mise en œuvre des outils d'analyse multivariée a pour objectif de permettre une visualisation des données en vue de construire des hypothèses, choisir une procédure analytique ou un plan d'échantillonnage, établir la meilleure façon de réaliser l'analyse multivariée des données existantes ainsi que des futures données du même type. Une fois cette phase exploratoire réalisée, l'analyse peut se poursuivre par une seconde phase, de classification, où les échantillons sont organisés en groupes ou classes spécifiques.

La classification est le processus par lequel on détermine si des échantillons appartiennent à la même classe que ceux qui ont été utilisés pour construire le modèle. S'il existe un bon ajustement entre un échantillon inconnu et une classe particulière, on considère qu'il appartient à cette classe. De nombreuses tâches relèvent de la classification, par exemple le classement de substances selon leur qualité ou leur grade physique. Les tests d'identification constituent un cas particulier de classification, où l'on compare des échantillons inconnus à des échantillons de référence, soit directement, soit indirectement via un modèle chimiométrique par exemple.

Analyse de données supervisée ou non supervisée

Les techniques chimiométriques peuvent être classées en deux catégories : l'analyse de données supervisée et l'analyse de données non supervisée. Les applications les plus courantes de la chimiométrie sont des approches supervisées dans lesquelles chaque échantillon des ensembles de données d'apprentissage est associé à une classe ou à une valeur cible attendue (connue). Parmi les approches supervisées, on peut notamment citer la méthode des *k* plus proches voisins, la régression linéaire, les séparatrices à vaste marge, les arbres de décision et forêts aléatoires. Dans le cadre des approches non supervisées, comme le partitionnement (« *clustering* »), les *K*-moyennes et l'ACP, les échantillons ne sont pas associés à des classes, aucun résultat n'est à prévoir et l'objectif est d'identifier les relations cachées au sein des données.

1-1-6. **Notation utilisée dans le chapitre**

X, Y	ensembles de données
X	variable indépendante
Y	variable dépendante
$\boldsymbol{X}$, $\boldsymbol{Y}$	matrices
$\boldsymbol{x}$, $\boldsymbol{y}$	vecteurs
x, y	valeurs scalaires
i, j	indices
x_i	$i^{ème}$ valeur du vecteur $\boldsymbol{x}$
$x_{i,j}$	$i^{ème}$, $j^{ème}$, valeur de la matrice $\boldsymbol{X}$
$\boldsymbol{X}^T$	matrice transposée de la matrice $\boldsymbol{X}$
$\boldsymbol{X}^{-1}$	matrice inverse *(si elle existe)* de la matrice $\boldsymbol{X}$
$\hat{\boldsymbol{X}}$	estimation de la matrice $\boldsymbol{X}$
$\lvert\boldsymbol{X}\rvert$	déterminant de la matrice (carrée) $\boldsymbol{X}$
$\lVert\boldsymbol{x}\rVert$	norme du vecteur $\boldsymbol{x}$
b	coefficient de l'équation de régression

1-2. *BONNES PRATIQUES CHIMIOMETRIQUES*

1-2-1. **Étapes de l'analyse chimiométrique**

La mise en œuvre des techniques chimiométriques varie selon les exigences spécifiques attachées au système à analyser. Cependant, la procédure suivante est généralement valable pour l'analyse de données non organisées :

- définir l'objectif précis de la collecte de données et les résultats attendus de l'analyse, pour la formulation du problème ;
- vérifier l'origine et la disponibilité des données ; s'assurer que les données collectées couvrent le domaine de variation des variables ou attributs explorés ;
- si les données disponibles ne couvrent pas le domaine de variation attendu, procéder à la préparation et à l'analyse d'échantillons permettant d'obtenir les données expérimentales requises ;
- choisir soigneusement les variables (l'utilisation de variables pertinentes peut réduire l'erreur de prédiction) ; une connaissance du domaine et une vérification minutieuse sont toutefois requises ;
- si nécessaire, appliquer aux données brutes des transformations et prétraitements mathématiques ;
- construire le modèle et l'optimiser ;
- mettre le modèle à l'épreuve et en vérifier les performances sur de nouveaux échantillons ou données ;
- valider la méthode conformément aux usages et exigences pharmaceutiques en vigueur.

1-2-2. **Données**

1-2-2-1. *Qualité de l'échantillon*

La sélection pertinente de l'échantillon accroît la probabilité d'extraire des informations utiles des données analytiques. La qualité des résultats sera accrue s'il est possible de procéder à des ajustements de certains paramètres ou variables sélectionnés, selon un plan d'expériences. Il est possible de recourir à la méthode des plans d'expériences pour introduire des variations inter-échantillons systématiques et contrôlées, tant pour les analytes que pour les facteurs d'interférence. Les préalables habituels, lors d'une modélisation, sont de déterminer quelles sont les variables nécessaires pour décrire adéquatement les échantillons, quels échantillons sont similaires entre eux, ou si l'ensemble de données contient des sous-groupes d'échantillons.

1-2-2-2. *Tableaux de données, représentation géométrique*

Les réponses ou résultats de mesure obtenus à partir des échantillons, constituent une collection de valeurs numériques représentant l'intensité du signal, qui correspondent aux variables indépendantes (il faut toutefois noter que ces variables ne sont pas forcément linéairement indépendantes selon les définitions mathématiques). Le mode de représentation optimal de ces valeurs est un tableau dans lequel, par convention, chaque ligne contient les données associées à un échantillon. Une collection de lignes-échantillons constitue une matrice, dont les colonnes constituent les variables. Les échantillons peuvent être associés à des descripteurs, généralement appelés « données *Y* » (variable dépendante), qui reflètent leurs caractéristiques, à savoir la valeur d'une propriété ou d'un attribut, physique ou chimique. Il est possible d'ajouter en colonne les valeurs de cette propriété à la matrice des réponses des échantillons, et ainsi de combiner les réponses et les attributs de chaque échantillon.

Lorsque n objets sont décrits par m variables, le tableau des données correspond à une matrice $n \times m$. Chacune des m variables représente un vecteur contenant n valeurs correspondant aux n objets. Chaque objet apparaît comme un point décrit par ses m coordonnées (une pour chacun des axes correspondant aux m variables) dans un espace à m dimensions.

1-2-2-3. *Évaluation préliminaire des données*

Avant de réaliser une analyse multivariée, il est possible de procéder à une évaluation de la qualité de la réponse des échantillons au moyen d'outils statistiques. L'emploi d'outils graphiques est recommandé pour une première évaluation visuelle des données, en utilisant, par exemple une représentation sous forme d'histogrammes et/ou de diagrammes en boîte pour l'étude de la distribution des données, ou de nuages de points pour la détection de corrélations. Les statistiques descriptives sont utiles pour obtenir une évaluation rapide de chaque variable considérée séparément, avant de procéder à une analyse multivariée. On peut par exemple utiliser la moyenne, l'écart type, la variance, la médiane, le minimum, le maximum, les quartiles inférieur et supérieur pour évaluer les données et détecter des valeurs hors-limites ou aberrantes, ou encore une dispersion ou une asymétrie anormales. Ces statistiques révèlent tout élément suspect dans un tableau de données, et indiquent si une transformation peut être utile. Les statistiques bidimensionnelles, telles que la corrélation, montrent comment les variations de deux variables peuvent être corrélées dans un tableau de données. La vérification de ces statistiques est également utile pour réduire la taille du tableau de données en éliminant les redondances.

1-2-2-4. *Points aberrants*

Un point aberrant est un échantillon qui diffère grandement du reste des données et qui peut donc ne pas être correctement décrit par le modèle choisi. Les points aberrants peuvent être liés aux données *X* ou *Y*. Ils sont le reflet d'une interférence non attendue dans les données initiales ou d'une erreur de mesure. Dans le cadre de la prédiction, un écart important entre les valeurs prédites et les valeurs observées met en question la pertinence de la procédure de modélisation, et l'intervalle couvert par les données initiales. Les points aberrants peuvent être le résultat de modifications de l'interaction instrument/échantillons, ou de l'existence d'échantillons sortant du champ du modèle. Si cette nouvelle source de variabilité est confirmée et pertinente, les données correspondantes constituent une source d'informations utile. Il est recommandé de les étudier pour déterminer s'il y a lieu d'améliorer la robustesse de l'étalonnage, ou alors d'ignorer les résultats aberrants parce qu'ils apparaissent non critiques ou sans relation avec le processus (erreur de manipulation, par exemple).

Dans le cas de la classification, il convient de vérifier la présence de points aberrants sur chaque classe séparément.

1-2-2-5. *Erreurs de données*

Les erreurs de données peuvent être de différents types : erreur aléatoire dans les valeurs de référence des attributs, erreur aléatoire dans les réponses collectées, erreur systématique dans la relation entre les 2 types de données. Les sources possibles d'erreur d'étalonnage sont les erreurs spécifiques liées par exemple à la procédure de référence, puis la non-homogénéité des échantillons, et enfin la non-représentativité du groupe des échantillons utilisés pour l'étalonnage. La sélection du modèle lors de l'étalonnage ne représente généralement qu'une fraction de la variance ou erreur attribuable à la technique analytique modélisée. Cependant, il est difficile d'évaluer si cette erreur est supérieure à l'erreur associée à la procédure de référence, ou inversement.

1-2-2-6. *Prétraitement des données*

Le prétraitement des données fait actuellement partie intégrante de la modélisation chimiométrique. Les données brutes ne se prêtant pas toujours de manière optimale à l'analyse chimiométrique, il peut être nécessaire de les prétraiter à l'aide de transformations mathématiques pour améliorer la qualité de l'analyse. Des interférences peuvent provenir par exemple du bruit de fond, d'une dérive de la ligne de base, de mesures effectuées dans des conditions différentes, des échantillons, du matériel ou de l'opérateur. Les interférences entraînent une variation des données qui n'est pas liée à la propriété considérée.

Le prétraitement des données répond à deux objectifs. En premier lieu, il consiste à rendre comparables les différentes données, pour faciliter leur interprétation. Le prétraitement des données vise également à corriger et à supprimer les variations systématiques, les signaux aléatoires et les variables non informatives qui interfèrent avec les informations physiques et chimiques pertinentes et en rendent l'extraction plus complexe.

Il existe par ailleurs une large gamme de transformations applicables aux données *X* et aux données *Y* avant l'analyse multivariée afin d'améliorer la modélisation. Des méthodes par colonnes ou par lignes peuvent être utilisées, selon l'objectif et selon le type de données.

Les transformations telles que le centrage et le changement d'échelle des variables sont appliquées indépendamment à chaque variable de la même colonne dans l'ensemble des données. Par exemple, le centrage par rapport à la moyenne supprime une différence systématique en soustrayant la valeur moyenne d'une variable à chaque valeur de cette variable tout en préservant les similitudes et dissemblances entre les échantillons. Après centrage par rapport à la moyenne, les nouvelles valeurs de la variable sont centrées sur zéro, c'est-à-dire que la moyenne de chaque variable est égale à zéro. Les changements d'échelle permettent de modifier la variance des variables. En particulier, le changement d'échelle automatique, après le centrage initial par rapport à la moyenne, consiste à diviser chaque variable d'une colonne par l'écart-type de cette colonne, ce qui rend la variance de chaque variable égale à un (créant ainsi une équivalence pour l'analyse). Le but du changement d'échelle automatique est d'éviter que le modèle ne se concentre à tort sur les variables ayant les variances les plus élevées. Ces méthodes mathématiques sont souvent appliquées dans les techniques de projection telles que l'ACP ou la PLS afin de pouvoir comparer la contribution de chaque variable au résultat obtenu.

D'autres méthodes visent à corriger le signal, notamment en normalisant les échantillons, en améliorant le rapport signal/bruit (débruitage et élimination du bruit de fond) et en compensant les déplacements des pics dans les signaux (alignement du signal). Elles consistent en une transformation par ligne des variables.

L'objectif de la normalisation des données analytiques est de permettre la comparaison des échantillons en éliminant les biais. Pour l'essentiel, la normalisation met les données à l'échelle en leur appliquant un facteur constant (surface du signal, vecteur unitaire, moyenne, maximum, etc.). La méthode de normalisation par la variance (SNV, pour « *standard normal variate* ») est largement utilisée pour l'analyse des données spectroscopiques. La transformation SNV s'applique séparément à chaque spectre et, pour chacun des spectres, la moyenne et l'écart-type interne de toutes les variables sont calculés. La valeur moyenne est soustraite de l'absorbance pour chaque point et le résultat est divisé par l'écart-type. Une partie du bruit ou de la dispersion est ainsi éliminée des spectres. Cependant, la normalisation des données doit être appliquée avec précaution car elle entraîne une altération de la structure de corrélation des données. Cette méthode de normalisation est destinée à éliminer les variations entre les échantillons dues aux différences de trajet de la lumière, en corrigeant à la fois la pente et la dérive de la ligne de base. De plus, les méthodes de correction de diffusion (comme par exemple la correction multiplicative de diffusion (MSC, pour « *multiplicative scatter correction* »), l'extension de la correction multiplicative de diffusion (EMSC, pour « *extended multiplicative scatter correction* »)) sont fréquemment utilisées pour éliminer les effets de diffusion de la lumière des spectres dans le proche infrarouge.

Le bruit du signal est généralement supprimé en lissant le signal à l'aide de différents filtres (par exemple, filtres moyens, médians ou gaussiens, algorithme de Savitzky-Golay). L'influence de la composante ligne de base peut également être éliminée en combinant lissage et dérivations. La dérivée première compense l'effet causé par une ligne de base additive tandis que la dérivée seconde supprime également l'effet d'une ligne de base linéaire. Il est ainsi possible non seulement de compenser une partie du bruit, mais aussi de rendre plus visibles les différences entre les spectres dans l'ensemble de données.

L'alignement des signaux instrumentaux est particulièrement utile lors de l'utilisation de signaux chromatographiques ou de résonance magnétique nucléaire (RMN). Il a pour but de compenser les déplacements de pics éventuellement associés à des conditions expérimentales ou instrumentales instables. Par exemple, la déformation optimisée suivant les corrélations, désignée par l'acronyme COW (pour « *correlation optimised warping* »), est une technique classique où les déplacements de pics sont corrigés par étirement et par compression de sections sélectionnées (le « *warping* ») dans le signal cible et le signal qui est aligné.

1-2-3. **Maintenance des modèles chimiométriques**

Les méthodes chimiométriques doivent faire l'objet de réévaluations périodiques visant à vérifier que leurs performances restent satisfaisantes. Une réévaluation portant sur les paramètres critiques doit en outre être effectuée lors de tout changement concernant les conditions d'application du modèle chimiométrique (processus, source des échantillons, conditions de mesure, instrumentation, logiciels, etc.).

L'objectif de la maintenance/mise à jour des modèles chimiométriques est d'assurer la pérennité des applications. L'étendue de la validation (sélection des paramètres pertinents) est à définir sur la base d'une analyse de risque, pour la procédure analytique et le modèle chimiométrique considérés.

1-3. *ÉVALUATION ET VALIDATION DES MODÈLES CHIMIOMÉTRIQUES*

1-3-1. **Introduction**

Le terme « validation », outre le sens qui est communément le sien dans le contexte réglementaire (« validation d'une procédure analytique »), désigne également en chimiométrie et en science des données des étapes spécifiques du calcul nécessaires pour sélectionner le meilleur modèle possible pour un ensemble particulier de données et conditions préalables.

Pour développer un modèle candidat et estimer ses performances, et si l'on dispose d'un volume de données suffisant, l'approche recommandée est de diviser ces données en 3 sous-ensembles :

1) un groupe d'apprentissage, aussi appelé groupe d'étalonnage, pour la construction des modèles,
2) un groupe test, aussi appelé groupe test interne ou groupe de validation interne, pour optimiser et sélectionner un modèle candidat avec une faible erreur de prédiction,
3) un groupe test généré indépendamment, aussi appelé groupe test externe ou groupe de validation externe, pour évaluer les performances du modèle final sélectionné.

L'organigramme de la figure 5.21.-1 montre comment les ensembles de données sont généralement divisés pour construire, évaluer et valider un modèle dans le but ultime d'une validation réglementaire.

1-3-2. **Groupes tests, principes généraux**

Une fois qu'un modèle est construit avec le groupe d'apprentissage, il est conseillé de tester le modèle de régression avec les données des groupes tests internes et externes.

Le groupe test interne est généré en mettant de côté une fraction des échantillons d'étalonnage qui ne seront pas utilisés pour construire le modèle, mais conservés pour le valider et pour simuler la prédiction des données futures. Les deux partitions sont ensuite utilisées lors de la modélisation : le modèle est construit selon certains réglages et en utilisant les données d'apprentissage, tandis que les données du groupe test interne servent à vérifier la pertinence des différents réglages. Le réglage le plus performant sur les deux partitions est

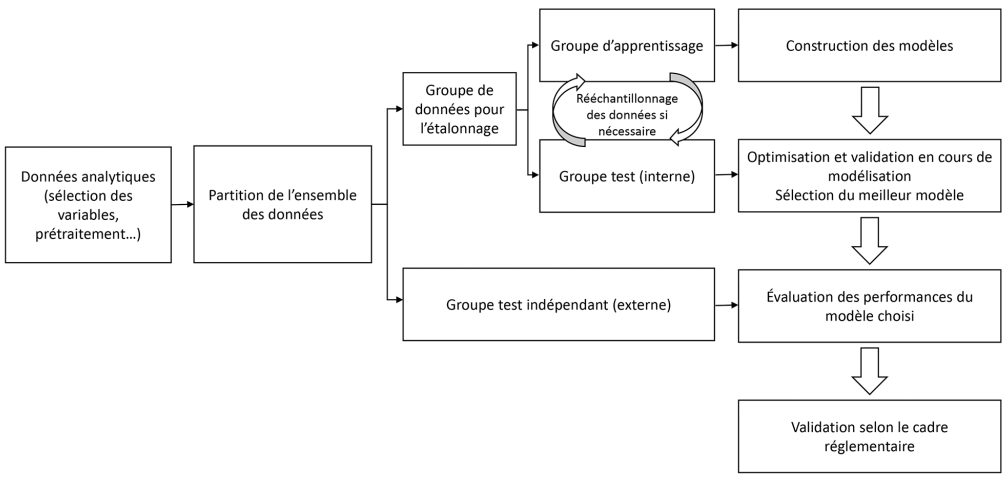


Figure 5.21.-1. – *Organigramme du procédé de validation d'un modèle chimiométrique supervisé*

ensuite sélectionné pour construire le modèle avec l'ensemble des données d'étalonnage. Ainsi, le groupe test interne sert uniquement à optimiser les réglages du modèle dans le but de sélectionner un réglage final qui fonctionnera bien avec les données futures.

Le groupe test externe est un ensemble d'échantillons n'intervenant qu'après que le modèle final et ses réglages ont été choisis. Il ne doit être utilisé ni pour la construction ni pour la sélection du modèle. Son seul but est d'évaluer le modèle final retenu et d'en évaluer la capacité à prédire les données futures. Le groupe test externe est également appelé groupe test indépendant car ni la construction du modèle, ni sa sélection n'en dépendent.

1-3-3. **Dimension des ensembles de données**

La taille de l'ensemble de données requis pour construire le modèle d'étalonnage dépend du nombre d'analytes et de propriétés interférentes que devra gérer le modèle. En règle générale, la taille du groupe d'apprentissage utilisé pour l'établissement du modèle doit être plus grande dans le cas où les variations interférentes sont aléatoires que dans celui où toutes les interférences majeures sont connues (ou qu'il serait possible de les faire varier conformément à un plan d'expériences). Le nombre minimum d'échantillons nécessaire pour couvrir l'intervalle d'étalonnage peut être estimé à partir du plan d'expériences correspondant. Quant à la taille du groupe test indépendant, elle est normalement de l'ordre de 20-40 pour cent de celle du groupe utilisé lors de la construction du modèle d'étalonnage, et doit couvrir les mêmes variations que les données d'étalonnage. Cependant, si l'on dispose d'un nombre important d'échantillons représentatifs, l'estimation de la performance de prédiction sera d'autant plus fiable que le groupe test indépendant est vaste (plus de 40 pour cent).

1-3-4. **Partition des ensembles de données**

Il existe de nombreuses techniques courantes permettant de diviser les échantillons d'étalonnage en groupes d'apprentissage et groupes test internes. La technique la plus évidente est la partition simple, c'est-à-dire une division directe en deux ensembles fixes : un groupe d'apprentissage et un groupe test. Lorsque le nombre d'échantillons est limité, il est préférable de partitionner de manière répétée les échantillons d'étalonnage, c'est-à-dire de procéder à un rééchantillonnage. Les échantillons d'étalonnage peuvent être partitionnés de manière itérative en un groupe test interne et le groupe d'apprentissage restant. Le rééchantillonnage permet d'utiliser plus efficacement les données disponibles, car les résultats de la validation sont moyennés sur de nombreuses partitions en groupe test interne et groupe d'apprentissage. Un schéma de rééchantillonnage très répandu est la validation croisée à K blocs (également appelée validation croisée *leave-subset-out* ou *leave-block-out*, ou encore validation croisée segmentée).

La plupart du temps, le partitionnement de l'ensemble des données en un groupe d'étalonnage et un groupe test indépendant peut être effectué au moyen d'algorithmes de sélection d'échantillons qui permettent d'éviter les biais et d'assurer la représentativité du groupe test par rapport à l'ensemble des échantillons. Il existe plusieurs algorithmes de la sorte, des plus simples aux plus complexes. Parmi eux, les algorithmes Kennard-Stone et Duplex sont régulièrement utilisés et efficaces pour sélectionner des sous-ensembles d'échantillons.

L'algorithme Kennard-Stone procède à une projection des échantillons dans l'espace de données multidimensionnel défini par toutes les variables de l'ensemble de données. L'algorithme commence par sélectionner l'échantillon le plus proche du point moyen dans l'espace de données pour le groupe test, il sélectionne ensuite l'échantillon le plus éloigné du premier échantillon puis, à chaque itération, sélectionne l'échantillon le plus éloigné des points déjà retenus, jusqu'à ce que le nombre prédéfini d'échantillons pour le groupe test externe soit atteint. Il est également possible d'appliquer cet algorithme en commençant par l'échantillon le plus éloigné du point moyen de l'espace de données.

Tout comme l'algorithme Kennard-Stone, l'algorithme Duplex projette les échantillons dans l'espace de données, mais il sélectionne les échantillons deux par deux, en commençant par les deux échantillons les plus éloignés l'un de l'autre qu'il affecte à un premier groupe. La deuxième paire la plus éloignée est sélectionnée pour un second groupe, et les sélections se poursuivent ainsi en respectant une attribution alternée des paires d'échantillons entre le premier et le second groupe, jusqu'à ce que le nombre prédéfini d'échantillons soit atteint dans le second groupe, qui sera le groupe test. Le premier groupe et les échantillons restants seront utilisés pour la modélisation.

1-3-5. **Indicateurs de performance**

En analyse de classification, l'erreur de prédiction du modèle est définie par rapport au taux de classification correcte, c'est-à-dire le pourcentage d'échantillons correctement classés en termes de « vrais-positifs » et « vrais-négatifs ».

En analyse quantitative, la construction d'un modèle de régression consiste à établir une relation mathématique entre les valeurs de la variable dépendante Y et les valeurs correspondantes de la variable indépendante X. Les données X peuvent représenter une collection de signaux – c'est à dire des réponses obtenues pour un certain nombre d'échantillons d'étalonnage – et les données Y les valeurs d'un attribut – c'est à dire de la propriété considérée dans les échantillons d'étalonnage.

1-3-5-1. *Erreur quadratique moyenne de prédiction*

L'exploration de la relation entre X et Y s'effectue au moyen des échantillons d'étalonnage dont les valeurs x et y ont été relevées et sont connues. Le modèle construit avec les données d'apprentissage est utilisé pour prédire les valeurs y du groupe de validation, qui sont ensuite comparées aux valeurs y observées (mesurées indépendamment avec la procédure de référence). La différence entre la valeur prédite et la valeur observée est appelée le résidu et peut être utilisée pour calculer une mesure de l'incertitude sur les prédictions futures. La détermination la plus courante du résidu est la racine de l'erreur quadratique moyenne de prédiction (ou RMSEP, pour « *root mean square error of prediction* »). Cette valeur est une estimation de l'incertitude moyenne qui peut être attendue sur la prédiction des valeurs y de nouveaux échantillons (mêmes unités que la variable dépendante). Comme il n'est pas posé d'hypothèse sur la distribution statistique de l'erreur dans le cadre de la modélisation, l'erreur de prédiction ne peut pas être utilisée directement pour déterminer un intervalle statistique valide pour les valeurs prédites. Les estimations d'erreurs fournies par la RMSEP sont toutefois fiables si les échantillons d'étalonnage sont représentatifs des échantillons futurs, et un intervalle d'incertitude approximatif pour les valeurs y prédites serait $\pm L \times RMSEP$, L déterminant la couverture de l'intervalle. La valeur L doit être fixée par l'opérateur. Un choix courant est $L = 2$. Ce choix, quel qu'il soit, doit refléter les exigences spécifiques attachées à la procédure analytique considérée.

La fidélité des modèles chimiométriques peut s'avérer supérieure à celle des procédures de référence utilisées pour l'acquisition des données d'étalonnage et de test externe. Ceci est typiquement le cas pour les déterminations de teneur en eau par SPIR et PLS, le semi-microdosage (*2.5.12*) étant la procédure de référence.

1-3-5-2. *Erreur type d'étalonnage et coefficient de détermination*

Des indicateurs statistiques peuvent également être calculés pour aider à évaluer la qualité de l'ajustement du modèle d'étalonnage aux données d'étalonnage. Deux de ces indicateurs sont l'erreur type d'étalonnage (ETE) et le coefficient de détermination R^2.

L'ETE s'exprime dans la même unité que les variables dépendantes. Elle reflète l'erreur associée à la modélisation et ne peut pas servir à estimer les futures erreurs de prédiction. Elle indique si l'exactitude du calcul effectué à l'aide de l'équation d'étalonnage sera suffisante pour l'objectif pour lequel elle a été générée. En pratique, l'ETE doit être comparée à l'erreur associée à la procédure de référence (erreur type du laboratoire ou ETL, voir glossaire). L'ETE est en général plus élevée que l'ETL, en particulier si la modélisation n'intègre pas toutes les interférences potentielles dans les échantillons, ou si d'autres phénomènes physiques entrent en jeu.

Le coefficient de détermination R^2 est une grandeur sans dimension qui représente la qualité de l'ajustement du modèle d'étalonnage aux données d'étalonnage. Il peut prendre des valeurs comprises entre 0 et 1. Une valeur proche de 0 indique que l'étalonnage a échoué à établir la relation entre les données et les valeurs de référence. Lorsque le coefficient de détermination augmente, les données X deviennent de meilleurs prédicteurs des valeurs de référence.

1-3-6. **Validation en cours de modélisation**

L'objectif de la validation en cours de modélisation est d'évaluer plusieurs modèles construits avec des réglages différents, et de faciliter le choix d'un modèle performant. Comme mentionné précédemment, en chimiométrie et en science des données, ce procédé consistant à tester des modèles candidats est appelé « validation interne du modèle ». Les algorithmes sont généralement itératifs et reposent sur des hyperparamètres contrôlant la complexité du modèle. Le procédé par lequel on détermine quand mettre fin aux itérations ou quand arrêter d'augmenter la complexité du modèle requiert le calcul des indicateurs de performance des modèles candidats sur le groupe test interne. Ces indicateurs de performance sont spécifiques de la technique chimiométrique utilisée et de la nature des données analytiques traitées, ainsi que de l'objectif de la méthode dans sa globalité, sous ses aspects analytique et chimiométrique.

Les échantillons sélectionnés sont soit spécifiquement affectés à cette fin (groupe test interne), soit sélectionnés itérativement à partir de l'ensemble des échantillons d'étalonnage, en appliquant une validation croisée par exclusion des observations selon une méthode spécifique, comme par exemple par randomisation avec sous-échantillonnage itératif.

La validation croisée est une méthode qui simule la validation interne en divisant les données disponibles en K sous-ensembles de taille identique (validation croisée à K blocs, également appelée validation *leave-subset-out*). Le modèle est entraîné avec K-1 partitions, et évalué avec la partition laissée de côté. Cette procédure est répétée K fois, chaque fois avec une autre des K partitions exclues. Pour obtenir une exactitude plus élevée, ou si les données sont très limitées, la validation croisée peut en outre être combinée à une randomisation. Une série de validations croisées à K blocs est exécutée et, avant chaque partition, les données sont mélangées de manière aléatoire. Une erreur de prédiction est finalement obtenue pour chaque échantillon de l'ensemble d'étalonnage et la racine de l'erreur quadratique moyenne de la validation croisée (RMSECV, pour « *root mean square error of cross-validation* ») est calculée.

L'algorithme de validation croisée à K blocs le plus simple est celui de la validation croisée « *leave-one-out* » (LOOCV), dans lequel K est égal à N (nombre d'échantillons dans l'ensemble d'étalonnage). Dans la LOOCV, un échantillon est prélevé dans le groupe d'étalonnage et un modèle est construit avec les échantillons restants. Le modèle obtenu est ensuite utilisé pour prédire l'échantillon sélectionné. Cette procédure est répétée en utilisant à chaque fois un autre échantillon. A la fin, une erreur de prédiction est obtenue pour chaque échantillon du groupe d'étalonnage. L'erreur de prédiction pour l'ensemble du groupe d'étalonnage est alors calculée comme la RMSECV. Lorsque la validation croisée est utilisée pour sélectionner le modèle optimal, le choix se portera sur le modèle ayant la plus faible RMSECV. La LOOCV présente l'inconvénient de nécessiter beaucoup de temps en raison du nombre élevé de calculs, et de pouvoir conduire à un modèle trop spécifique, à savoir un modèle qui soit trop lié au groupe d'étalonnage au point de ne plus pouvoir prédire/classer correctement les nouveaux échantillons. Une solution plus appropriée pourrait être la validation croisée de type « *leave-subset-out* », comme par exemple une validation croisée à K blocs dans laquelle le groupe d'étalonnage est divisé aléatoirement en K parties, chaque partie contenant $\frac{N}{K}$ observations (figure 5.21.-2).

1-3-7. **Évaluation de la performance des modèles**

A l'issue de l'optimisation, une fois un modèle adéquat obtenu, son aptitude à l'emploi est évaluée. A ce stade intervient un groupe test d'échantillons indépendants, non utilisés précédemment pour la construction, l'optimisation ou la sélection du modèle, afin d'évaluer la performance prédictive de ce dernier. Ce procédé d'évaluation d'un modèle est appelé validation externe (sans rapport, là encore, avec la validation en vertu du cadre réglementaire en vigueur). Le critère de performance du modèle optimisé est donné par l'erreur de prédiction en calculant la RMSEP.

1-3-8. **Validation réglementaire**

Les principes et considérations attachés à cette validation font l'objet de lignes directrices établies au niveau international sur la validation des procédures analytiques. Cependant, en raison de la nature particulière du traitement et de l'évaluation des données dans le cadre des méthodes chimiométriques en général, des aspects supplémentaires sont à prendre en compte pour la validation des procédures analytiques. Dans ce contexte, la validation porte à la fois sur la vérification des performances de la procédure analytique et sur l'évaluation du

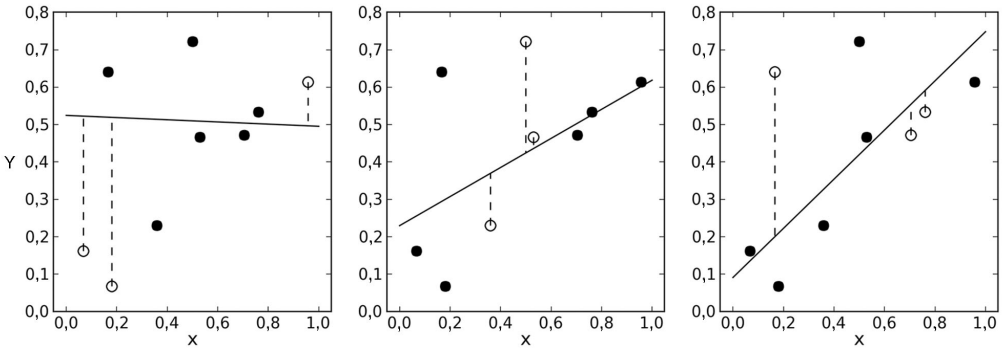


Figure 5.21.-2. – *Un exemple de validation croisée 3-blocs est ici donné comme illustration. Sous-ensemble de données du modèle candidat (données d'apprentissage) = ●. Sous-ensemble utilisé pour la validation = ○. On laisse de côté un tiers des données disponibles lors de la construction du modèle candidat. Dans cet exemple, les tiers laissés de côté sont choisis au hasard et chaque objet n'est laissé de côté qu'une seule fois. Les erreurs (résidus) des données de validation (tracés en pointillés) sont collectées pour constituer l'erreur cumulée de validation croisée*

modèle. Dans certains cas particuliers, il peut être nécessaire de ne valider que le modèle chimiométrique.

1-3-8-1. *Modèles qualitatifs*

La spécificité et la robustesse sont les principaux paramètres de validation des modèles qualitatifs et doivent à ce titre être examinés. Toute exception doit être scientifiquement justifiée.

Spécificité

Lors de la validation, il doit être démontré que le modèle possède un pouvoir discriminant suffisant entre la substance à examiner et un ensemble de substances appropriées, par exemple des substances manipulées sur le même site de fabrication avec lesquelles existe un risque de confusion. Si, au-delà de l'identité chimique des substances à identifier, certaines caractéristiques (telles que polymorphisme, taille des particules ou teneur en eau) sont susceptibles d'être pertinentes pour l'identification, leur impact sur le modèle devrait alors être évalué. La sélection des substances à inclure dans la validation de la spécificité doit reposer sur des critères logistiques (proximité et accessibilité par rapport au processus, avec une attention particulière portée aux substances d'aspect similaire), des critères chimiques (similarité de structure par exemple) ainsi que des critères physiques, s'ils sont critiques. Une fois défini cet ensemble de substances, l'objectif est d'établir le pouvoir discriminant de la méthode chimiométrique et sa capacité à rejeter ces substances. Il faut donc, pour chacune d'elles, analyser et évaluer un jeu d'échantillons représentatif de la variabilité type de la substance. Si la spécificité du modèle chimiométrique s'avère insuffisante, il convient alors de définir de nouveaux hyperparamètres pour adapter le modèle. Il peut par exemple être nécessaire de modifier la technique de modélisation et de revalider la nouvelle méthode.

Il convient d'effectuer une revalidation de la spécificité chaque fois qu'apparaissent de nouvelles circonstances susceptibles d'affecter l'identification (par exemple nouvelles substances traitées sur le même site). Cette revalidation peut toutefois se limiter aux nouvelles circonstances en cause et ne doit pas nécessairement considérer l'ensemble des aspects non affectés par le changement intervenu.

Si certaines propriétés des substances évoluent dans le temps (cas par exemple de lots de substances dont la granulométrie ou la teneur en eau augmente ou diminue) et si ces changements deviennent significatifs, il convient de les prendre également en compte dans la validation. On peut par exemple, à cet effet, amender le protocole et le rapport de validation, sans nécessairement procéder à une revalidation exhaustive du modèle chimiométrique.

On peut évaluer la spécificité en estimant le nombre de faux positifs et de faux négatifs obtenus lors de la classification du groupe test.

Robustesse

Pour valider la robustesse du modèle, il convient d'établir une liste aussi complète que possible des paramètres critiques (par exemple certains paramètres du processus tels que température, humidité ou performances de l'instrumentation analytique). On met alors à l'épreuve la procédure analytique en faisant varier ces paramètres. Il peut être intéressant de recourir aux plans d'expériences pour effectuer cette évaluation.

On peut évaluer la robustesse en estimant le nombre de classifications correctes, de rejets corrects, de faux positifs et de faux négatifs obtenus lors de la classification d'échantillons dans diverses conditions.

1-3-8-2. *Modèles quantitatifs*

Sauf exception justifiée, il convient d'examiner les paramètres de validation suivants : spécificité, linéarité, étendue de mesure, exactitude, fidélité, robustesse.

Spécificité

Il est important de détecter les points aberrants au regard de l'espace d'étalonnage. Pour ce faire, on peut utiliser le coefficient de corrélation entre l'échantillon et le barycentre (noyau) du groupe d'étalonnage (données X), ainsi que le test T^2 de Hotelling.

Linéarité

La validation de la linéarité est à effectuer par corrélation des résultats du modèle chimiométrique avec ceux d'une procédure analytique de référence. Elle doit couvrir toute l'étendue de mesure de la procédure. La validation de la linéarité est à effectuer sur un groupe d'échantillons sélectionnés à cette fin, ne faisant pas partie du groupe d'étalonnage. En première approche, une validation croisée de type *leave-subset-out* peut être effectuée sur le groupe d'étalonnage, mais ne doit pas remplacer l'évaluation réalisée avec un groupe test indépendant. La linéarité peut être évaluée au moyen du défaut d'ajustement, du coefficient de corrélation, de la pente et de l'ordonnée à l'origine de la représentation des valeurs prédites en fonction des valeurs de référence.

Étendue de mesure

L'intervalle des valeurs de référence pour l'analyte définit l'étendue du modèle chimiométrique, tandis que sa borne inférieure détermine les limites de détection et de quantification de la procédure analytique. Des témoins doivent être utilisés pour s'assurer que les résultats situés en-dehors de l'intervalle sont reconnus comme tels et identifiés. Sur l'intervalle couvert par le modèle, les critères d'acceptation pour l'exactitude et la fidélité doivent être satisfaits.

Exactitude

On peut évaluer l'exactitude du modèle chimiométrique en comparant les résultats analytiques respectivement obtenus avec ce modèle et avec une procédure de référence validée. Cette évaluation est à effectuer sur l'étendue définie du modèle chimiométrique, au moyen d'un groupe test indépendant. Il peut être utile ici aussi de procéder à une évaluation par validation croisée de type *leave-subset-out*, mais ici encore celle-ci ne doit pas remplacer l'évaluation réalisée au moyen d'un groupe test indépendant.

Fidélité

La détermination de la fidélité de la procédure analytique s'effectue par estimation de l'écart type des résultats de mesure à l'aide du modèle chimiométrique. La fidélité peut être considérée sous l'angle de la répétabilité (plusieurs mesures effectuées sur le même échantillon par la même personne et le même jour) et de la fidélité intermédiaire (plusieurs mesures effectuées sur le même échantillon par des personnes différentes et à des jours différents). Son évaluation est à effectuer pour différentes valeurs de l'analyte (la concentration, par exemple) couvrant l'étendue du modèle chimiométrique, ou tout au moins pour une valeur cible.

Robustesse

Les principes s'appliquant à la validation de la robustesse sont les mêmes que pour les méthodes qualitatives. Il convient d'accorder une attention particulière aux effets que peuvent avoir les paramètres critiques sur l'exactitude et sur la fidélité du modèle chimiométrique. Il peut être intéressant d'utiliser des plans d'expériences pour effectuer cette étude.

Il est également possible de mettre le modèle chimiométrique à l'épreuve en utilisant des échantillons contenant l'analyte à une concentration non comprise dans l'étendue de mesure de la procédure, ou des échantillons d'identité différente. La validation doit établir que ces échantillons d'épreuve sont clairement reconnus comme aberrants.

2. TECHNIQUES CHIMIOMÉTRIQUES

Une sélection non exhaustive de méthodes chimiométriques, dont une classification est proposée figure 5.21.-3, est présentée et discutée ici.

2-1. *ANALYSE EN COMPOSANTES PRINCIPALES*

2-1-1. **Introduction**

La complexité des informations contenues dans les grands ensembles ou tableaux de données rend l'interprétation humaine de ces informations extrêmement difficile sans l'assistance de méthodes spécifiques. L'ACP permet de révéler et de visualiser par projection les principales variations au sein des données. Elle peut révéler si un échantillon diffère d'un autre et dans quelle mesure, quelles sont les variables qui pèsent le plus dans cette différence, si leur contribution s'exerce dans le même sens, si elles sont corrélées ou indépendantes. Elle peut faire émerger des schémas ou regroupements dans les ensembles de données. Elle peut également servir à estimer la quantité d'information utile contenue dans les tableaux de données, en l'isolant du bruit ou des variations non pertinentes.

2-1-2. **Principe**

L'ACP est une méthode linéaire de projection des données qui opère une compression des données en les décomposant en « variables latentes ». Cette procédure de décomposition permet d'obtenir une matrice avec des colonnes de vecteurs orthogonaux (les « *scores* ») et une autre matrice avec des lignes de vecteurs orthonormés (les « *loadings* »). Les composantes principales (CP), ou variables latentes, sont une combinaison linéaire des axes de variables initiaux. Chaque variable latente peut être interprétée via sa connexion aux variables originelles. Les données présentées sont les mêmes mais dans un nouveau système de coordonnées. Les interrelations entre échantillons sont révélées par leurs projections dans le nouveau système de coordonnées (c'est-à-dire les scores). Les échantillons similaires dans le nouveau système de coordonnées se regroupent. La distance entre échantillons est une mesure de leur similarité (ou de leur dissemblance).

Le tableau de données originel est transformé en une nouvelle matrice, organisée différemment, dont la structure révèle des relations entre lignes et colonnes qui étaient cachées dans la matrice originelle (figure 5.21.-4). Cette nouvelle structure représente la partie « expliquée » des données originelles. La procédure opère une modélisation des données originelles jusqu'à parvenir à une erreur résiduelle qui constitue la partie « non expliquée » des données que l'étape de décomposition a permis de réduire au minimum.

L'idée de base est de remplacer un tableau complexe de données par une version simplifiée comportant un moins grand nombre de dimensions, mais s'ajustant aux données originelles de façon suffisamment proche pour pouvoir être considérée comme une bonne approximation, comme le montre mathématiquement la figure 5.21.-5. L'extraction d'informations à partir d'un tableau de données consiste à explorer la variation inter-échantillons, c'est à dire à découvrir ce qui rend des échantillons dissemblables, ou similaires. Deux échantillons peuvent être décrits comme similaires s'ils possèdent des valeurs proches pour la plupart des variables. D'une perspective géométrique, l'ensemble des mesures effectuées pour un échantillon définit un point dans un espace multidimensionnel, avec autant de dimensions qu'il y a de variables. Si 2 points possèdent des coordonnées proches, ils sont situés dans une même région ou dans un même volume. L'ACP permet de réduire le nombre des dimensions en respectant la distance entre échantillons (les échantillons similaires restent proches et les échantillons dissemblables éloignés, comme ils l'étaient dans l'espace multidimensionnel originel) mais en opérant une compression dans un système de coordonnées différents et de dimensionnalité réduite.

Le principe de l'ACP consiste à identifier dans l'espace des données les directions qui décrivent les variations les plus importantes au sein de l'ensemble des données, la variation étant définie comme une variance [statistique]. Chaque direction des *loadings* est une combinaison linéaire des variables originelles qui contribuent le plus à capturer les variations inter-échantillons. Les CP sont par construction orthogonales entre elles. Elles sont par ailleurs rangées selon un ordre où chacune est porteuse de davantage d'information que celle qui la suit (les composantes sont classées par ordre décroissant de variance expliquée). Leur interprétation est donc priorisée et commence par les premières CP ; celles-ci, concentrant les variations majeures, constituent de ce fait un système alternatif de moindre complexité qui permet mieux d'interpréter la structure des données. En règle générale, l'information pertinente se trouve concentrée dans les premières CP, ici appelées CP primaires, et les suivantes sont plutôt susceptibles d'exprimer le bruit. En pratique, un critère de sélection est utilisé pour s'assurer que le bruit n'est pas confondu avec de l'information, ce critère doit également être couplé à une méthode telle que la validation croisée ou l'évaluation des *loadings*, pour déterminer le nombre de CP à utiliser pour l'analyse. Les relations entre échantillons peuvent alors être visualisées sur un graphe ou une série de graphes des scores (ou cartes factorielles). Les résidus $\widehat{\boldsymbol{E}}$ contiennent les variations non incluses dans le modèle, ces variations constituent une mesure de la qualité de l'ajustement du modèle aux échantillons ou variables. La décision de retenir un certain nombre de composantes dans le modèle est un compromis entre simplicité et qualité de l'ajustement ou performance du modèle.

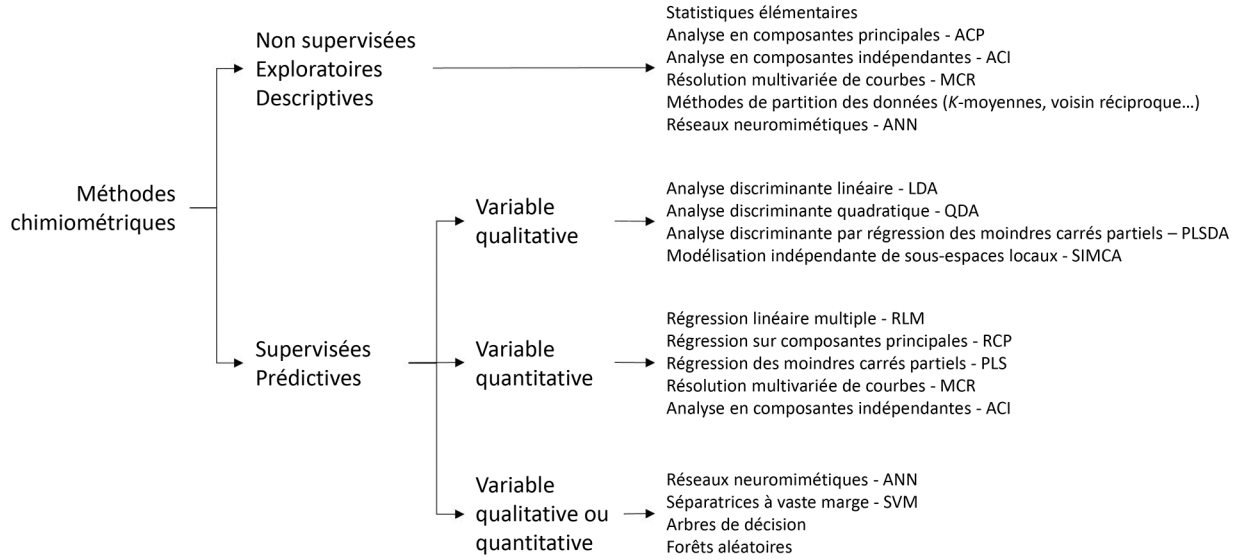


Figure 5.21.-3. – *Classification des méthodes chimiométriques présentées et discutées dans ce chapitre*

2-1-3. **Évaluation du modèle**

Le pourcentage de variance expliquée R^2 indique quelle proportion des variations initiales au sein des données est décrite par le modèle. Il exprime la proportion de structure que celui-ci a identifiée dans les données. Les pourcentages des variances expliquée et résiduelle expriment la qualité de l'ajustement réalisé par le modèle. Les modèles présentant un pourcentage de variance résiduelle faible (proche de 0 pour cent) ou un pourcentage de variance expliquée totale élevé (proche de 100 pour cent) expliquent la majeure partie de la variabilité présente dans les données. Avec les modèles simples, la variance résiduelle est ramenée à zéro avec un petit nombre de composantes. Si tel n'est pas le cas, on peut soupçonner l'existence d'un bruit important dans les données, mais il se peut également que la structure des données soit trop complexe pour pouvoir être appréhendée par un petit nombre de composantes. Les variables à faible variance résiduelle et forte variance expliquée pour une composante particulière sont bien appréhendées par le modèle. Les variables à forte variance résiduelle – pour toutes les composantes ou pour les composantes primaires – ne présentent avec les autres variables qu'une relation faible ou modérée. Si certaines ont une variance résiduelle très supérieure aux autres – pour toutes les composantes ou pour les composantes primaires – on peut les exclure du calcul, et ainsi obtenir un modèle mieux adapté au but recherché. Le pourcentage de variance expliquée est déterminé sur un groupe test indépendant afin de tester le modèle sur des données qui n'ont pas été utilisées pour sa construction.

2-1-4. **Aspects critiques**

En règle générale, l'ACP met en évidence les principales variations au sein de l'ensemble des données et elle peut ne pas singulariser les variations relativement mineures. Toutefois, si des variations plus faibles sont pertinentes, il peut être utile de procéder à des prétraitements spécifiques des données ou à une sélection de données, ou encore d'envisager une autre procédure chimiométrique.

2-1-5. **Utilisations possibles**

L'ACP est une méthode non supervisée et, de ce fait, un bon outil d'analyse exploratoire des données. Elle peut être utilisée, entre autres, pour la visualisation ou la compression des données, la mise en évidence de regroupements ou tendances, la détection de points aberrants.

En analyse exploratoire, la modélisation ACP peut être appliquée en une seule fois à l'ensemble du tableau des données ou, si l'on souhaite visualiser plus finement le point d'apparition d'une nouvelle variation, elle peut être appliquée

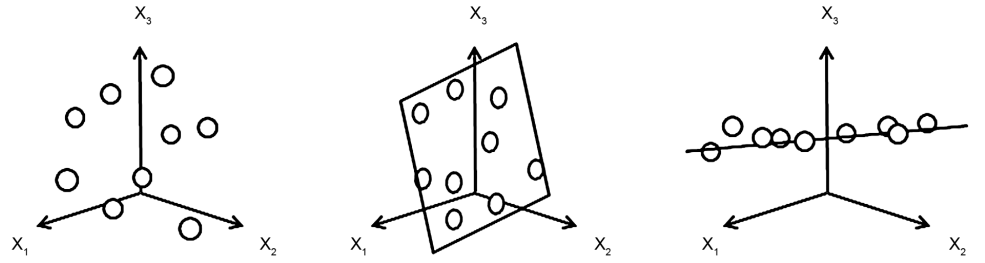
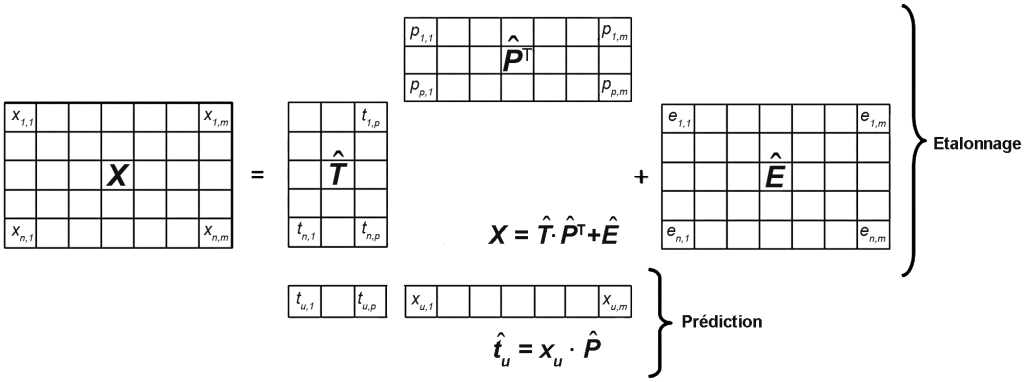


Figure 5.21.-4. – *Les données dans X peuvent être vues comme un tableau avec les coordonnées de l'échantillon sur chaque ligne. Si elle comporte 3 coordonnées ou moins, chaque observation peut être représentée géométriquement par un point dans le nuage de points montrant l'espace originel des données. Les représentations géométriques de 3 ensembles de données X sont ici présentées. A gauche, les observations sont représentées dans l'espace multivarié à trois dimensions. Les deuxième et troisième graphiques révèlent des structures (respectivement un plan et une ligne). Si les observations comportent plus de 3 coordonnées, des structures semblables peuvent toujours être présentes, mais elles sont difficilement visibles sans des outils comme l'ACP*



$\boldsymbol{X}$	=	matrice des données originelles, à n lignes et m colonnes
$\hat{\boldsymbol{T}}$	=	matrice des scores, à n lignes et p colonnes
$\hat{\boldsymbol{P}}^T$	=	matrice des *loadings*, à p lignes et m colonnes
$\hat{\boldsymbol{E}}$	=	matrice résiduelle (de même dimension que la matrice $\boldsymbol{X}$)
m	=	nombre de points de données (variables)
n	=	nombre de mesures (échantillons)
p	=	nombre de composantes (échantillons)
$\boldsymbol{x}_u$	=	données relatives à un échantillon inconnu
$\hat{\boldsymbol{t}}_u$	=	valeur des scores pour un échantillon inconnu

Figure 5.21.-5. – *Décomposition de la matrice* $\boldsymbol{X}$ *pour l'analyse en composantes principales (ACP)*

à des données évoluant dans le temps avec une fenêtre de largeur fixe ou mobile dans le cadre d'une analyse factorielle évolutive (EFA, pour « *evolving factor analysis* »), qui permet d'identifier par exemple le point où se manifeste une nouvelle composante principale (c'est-à-dire une nouvelle source de variation) dans une série d'échantillons consécutifs.

L'ACP constitue également la base de diverses méthodes de classification (SIMCA, pour « *soft independent modelling of class analogies* », par exemple) et de régression (RCP, par exemple). Grâce à la propriété qu'elle possède de capter les variations majeures dans les composantes principales primaires, l'ACP peut constituer le préliminaire d'un calcul de régression alors effectué sur la base d'un nombre réduit de variables dépendantes, les variables initiales étant remplacées par un petit nombre de variables latentes. Différentes techniques permettent d'utiliser les composantes principales comme données indépendantes dans le cadre d'une régression : RCP, MCR (pour « *multivariate curve resolution* »), réseaux neuromimétiques.

L'ACP constitue également une technique de base pour la maîtrise statistique des procédés multivariée (MSPM) (voir chapitre général *5.28. Maîtrise statistique des procédés multivariée*).

2-2. *ANALYSE EN COMPOSANTES INDEPENDANTES*

2-2-1. **Introduction**

L'analyse en composantes indépendantes (ACI) est une méthode de séparation aveugle des sources qui vise à extraire des signaux sources et leurs proportions dans les données. En chimiométrie, le terme « source » se rapporte en première approximation aux contributions des composants chimiques individuels *m* au signal global de chaque échantillon *n*, analysé comme un mélange. L'ACI peut être considérée comme une extension de l'ACP. Les résultats de l'ACP sont très utiles pour l'exploration des données, mais l'interprétation des « *loadings* » peut être complexe car ils correspondent rarement à des signaux sources purs. Alors que l'ACP optimise la matrice de variance des données, qui représente des statistiques de second ordre, l'ACI propose d'optimiser des statistiques d'ordre supérieur, comme le kurtosis. Ainsi, l'ACP fait apparaître des composantes non corrélées tandis que l'ACI identifie, comme composantes, des sources distinctes.

2-2-2. **Principe**

L'ACI a été conçue pour extraire les signaux sources d'un ensemble de signaux mêlés dans des proportions inconnues. Les mesures analytiques peuvent être définies comme une combinaison de plusieurs signaux sources indépendants. La combinaison linéaire des signaux mêlés peut s'écrire au moyen des équations suivantes (sans contribution du bruit) :

$$\boldsymbol{X} = \boldsymbol{AS}$$

$$\mathbf{x}_{\boldsymbol{n}} = a_{n1}\mathbf{s}_{\boldsymbol{1}} + a_{n2}\mathbf{s}_{\boldsymbol{2}} + \ \ldots\ + \mathrm{a}_{nm}\mathbf{s}_{\boldsymbol{m}}$$

où $\boldsymbol{X}$ est la matrice constituée de n signaux mêlés ($\mathbf{x}_{\boldsymbol{1}}$ à $\mathbf{x}_{\boldsymbol{n}}$), $\boldsymbol{S}$ est la matrice résultant du mélange de m signaux sources ($\mathbf{s}_{\boldsymbol{1}}$ à $\mathbf{s}_{\boldsymbol{m}}$) appelés composantes indépendantes (avec $n > m$) et $\boldsymbol{A}$ représente la matrice de mixage contenant des proportions (a_{11} à a_{nm}) des signaux purs dans chacun des signaux mêlés de $\boldsymbol{X}$.

L'ACI estime une transformation linéaire des données qui maximise l'indépendance entre les signaux extraits au moyen d'un critère reflétant l'indépendance statistique entre les sources (maximisation de la caractéristique non gaussienne, minimisation de l'information mutuelle ou utilisation de la méthode d'estimation de vraisemblance maximale). Plusieurs algorithmes (JADE, FastICA, par exemple) permettent d'estimer une matrice de démixage $\boldsymbol{W}$, donnant ainsi accès aux composantes indépendantes (matrice $\boldsymbol{S}$) et à leurs proportions (matrice $\boldsymbol{A}$).

2-2-3. **Aspects critiques**

Pour la décomposition ACI, la détermination du nombre de composantes indépendantes constitue l'étape critique de l'analyse des données. Les composantes indépendantes extraites différeront en fonction de leur nombre : un nombre trop faible se traduira par l'obtention de signaux non purs, alors qu'un nombre trop élevé de composantes indépendantes est susceptible de décomposer des signaux purs en plusieurs contributions et de produire des signaux bruités. En outre, il importe de noter que l'ACI ne classe pas les composantes indépendantes dans un ordre particulier et que, par conséquent, la même importance est accordée à la première et à la dernière composante.

L'ACI par blocs est un algorithme utilisé pour déterminer le nombre optimal de signaux à extraire. Il consiste, dans un premier temps, à diviser la matrice initiale en plusieurs blocs de taille similaire contenant des échantillons représentatifs de l'ensemble des données. Plusieurs modèles ACI sont ensuite calculés avec un nombre croissant de composantes indépendantes pour chaque bloc. Les composantes indépendantes correspondant aux signaux sources devant être présentes dans tous les blocs, elles doivent aussi être fortement corrélées. L'étude des corrélations permet de déterminer le nombre significatif de composantes indépendantes à sélectionner dans le modèle ACI.

2-2-4. **Utilisations possibles**

L'ACI peut être utilisée pour extraire des informations sur des composants purs à partir de divers signaux analytiques comme des spectres de masse, des spectres infrarouges ou des chromatogrammes. L'ACI est tout particulièrement prisée dans le domaine de l'imagerie chimique pour la décomposition des données spectroscopiques.

2-3. *MESURES DE SIMILARITE*

Le principal usage des algorithmes décrits ci-après est la mesure du degré de similarité entre un objet et un groupe ou entre un objet et le centre des données.

2-3-1. **Coefficient de corrélation et similarité cosinus**

Le coefficient de corrélation et la similarité cosinus, souvent aussi appelée corrélation, sont des méthodes statistiques utilisées pour comparer des données et en déterminer le degré de similarité. Le coefficient de corrélation est décrit par la formule ci-après. La similarité cosinus mesure l'angle entre une paire d'observations, ici considérées comme des vecteurs. Un coefficient de corrélation ou un score de similarité cosinus compris entre − 1 et + 1 est calculé pour une paire d'observations, une correspondance parfaite se traduisant par un score de + 1 et des caractéristiques spectrales totalement opposées par un score de − 1 (figure 5.21.-6). Quelle que soit la méthode utilisée (coefficient de corrélation ou similarité cosinus), il est important de noter que les observations comparées peuvent ne pas être à la même échelle en raison de la normalisation appliquée par les deux méthodes (voir le graphique situé à gauche de la figure 5.21.-6, où la valeur de corrélation est égale à 1,0 malgré une légère différence d'échelle).

Le coefficient de corrélation $r_{i,j}$ entre deux vecteurs $\boldsymbol{x}_i$ et $\boldsymbol{x}_j$ de même dimension m est donné par l'équation :

$$r_{i,j} = \frac{\sum_{k=1}^{m}\left(x_{i,k} - \bar{x_i}\right)\left(x_{j,k} - \bar{x_j}\right)}{\sqrt{\sum_{k=1}^{m}\left(x_{i,k} - \bar{x_i}\right)^2 \sum_{k=1}^{m}\left(x_{j,k} - \bar{x_j}\right)^2}}$$

La moyenne x est ici soustraite, alors que le calcul de la similarité cosinus ne comprend pas la soustraction d'une moyenne. Par conséquent, l'échelle entre les points extrêmes + 1 et − 1 sera numériquement différente selon les méthodes.

Les deux méthodes (similarité cosinus et coefficient de corrélation) permettent deux types de comparaison entre les ensembles de données :

- comparaison entre eux de deux échantillons sélectionnés,
- comparaison d'un ou plusieurs échantillons sélectionnés, représentés par des vecteurs, avec une bibliothèque de données de référence (d'un groupe ou d'une classe).

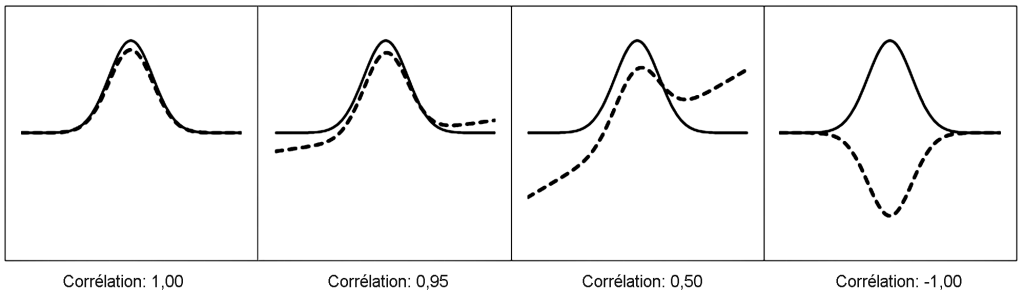


Figure 5.21.-6. – *Exemples de coefficients de corrélation illustrés par des correspondances de formes diverses*

Les données de référence peuvent être la moyenne d'un groupe présentant des caractéristiques types.

2-3-2. **Mesures de distance**

Dans l'espace d'objets, l'ensemble des objets à m dimensions (représentés par des vecteurs lignes) est constitué de points dans un espace à m dimensions plus ou moins voisins les uns des autres et pouvant former des groupes ou des amas. La mesure de la distance entre ces points exprime le degré de dissimilarité entre les objets. De même, la mesure de la distance d'un point au centre d'un groupe apporte des informations sur l'appartenance de cet objet au groupe. Les mesures de distance décrites ci-après illustrent la façon dont les objets peuvent être comparés.

2-3-2-1. *Distance euclidienne*

La distance euclidienne $ed_{i,j}$ entre 2 vecteurs $\boldsymbol{x}_i$ et $\boldsymbol{x}_j$ est égale à :

$$ed_{i,j} = \sqrt{\sum_{k=1}^{m} (x_{i,k} - x_{j,k})^2}$$

De même, il est possible de calculer la distance euclidienne $ed_{i,c}$ entre le vecteur $\boldsymbol{x}_i$ et le centre c d'un ensemble de données disposées en une matrice $\boldsymbol{X}$. Le vecteur $\boldsymbol{x}_j$ ci-dessus est alors simplement remplacé par le vecteur $\overline{\boldsymbol{x}}$ des moyennes par colonne de la matrice $\boldsymbol{X}$ (c'est-à-dire le centre géométrique). La distance euclidienne $ed_{i,c}$ en notation matricielle est alors donnée par :

$$ed_{i,c} = \sqrt{(x_i - \overline{x})(x_i - \overline{x})^{\mathrm{T}}}$$

2-3-2-2. *Distance de Mahalanobis*

La distance de Mahalanobis *md* prend en compte la corrélation entre les variables en utilisant l'inverse de la matrice de variance-covariance.

La matrice de variance-covariance $\boldsymbol{C}_x$ est donnée par l'équation :

$$C_x = (1/(n-1))\boldsymbol{X}_C^T \boldsymbol{X}_C$$

où $\boldsymbol{X}_c$ est la matrice $n \times m$ centrée sur la moyenne de chaque colonne. $\boldsymbol{C}_x$ est donc une matrice carrée dans laquelle la variance de chaque variable figure sur la diagonale, et la covariance entre variables de chaque côté de la diagonale. La distance de Mahalanobis du vecteur $\boldsymbol{x}_i$ au centre c de l'ensemble de données $\overline{\boldsymbol{x}}$, également appelée T^2 de Hotelling, est exprimée par l'équation suivante :

$$md_{i,c} = \sqrt{(x_i - \overline{x})\,\boldsymbol{C}_x^{-1}(x_i - \overline{x})^{\mathrm{T}}}$$

La matrice $\boldsymbol{C}_x^{-1}$ est l'inverse de la matrice de variance-covariance. Si la dimension m devient trop grande et que les variables mesurées sont corrélées, on obtient une matrice $\boldsymbol{C}_x$ mal conditionnée, voire singulière, qu'il devient impossible d'inverser. La solution consiste alors généralement à remplacer les variables brutes par des variables latentes (scores ACP ou PLS, par exemple) pour le calcul de la distance. La distance de Mahalanobis peut être utilisée pour détecter les valeurs aberrantes dans les données. La procédure est décrite dans le chapitre *5.28* pour les nouveaux objets (prédiction), dans le contexte des cartes de contrôle T^2 de Hotelling.

La distance de Mahalanobis peut également servir à calculer l'effet de levier h d'un modèle de régression comme suit :

$$h = \frac{1}{n} + \frac{(md_{i,c})^2}{(n-1)}$$

Les points présentant un effet de levier élevé exercent une influence importante sur le modèle (s'ils font partie du groupe d'apprentissage) et sont mal intégrés dans l'espace d'étalonnage, car situés à la limite, voire en dehors de celui-ci. Une valeur est considérée comme élevée si elle excède $2\,m/n$. Si des variables latentes sont utilisées, on considère généralement que m est le nombre de variables latentes. Ce calcul ne constitue qu'une approximation grossière de la valeur critique, en particulier pour la PLS.

2-3-2-3. *Aspects critiques*

Les distances euclidiennes n'expriment la proximité ou la distance réelle entre les points (données) que lorsque les variables sont strictement non corrélées. L'existence de corrélations entre les variables implique que celles-ci contiennent, au moins partiellement, les mêmes informations, et que la dimensionnalité de l'espace des données est en fait inférieure au nombre des variables. Les distances de Mahalanobis autorisent des corrections de corrélation, mais leur calcul suppose que la matrice de variance-covariance soit inversible. Or, dans certaines situations de forte colinéarité au sein de l'ensemble des données, la matrice est singulière et ne peut pas être inversée. Ceci est notamment vrai dans le cas des données spectroscopiques, où la résolution élevée des spectromètres introduit une redondance car ils délivrent des signaux sensiblement identiques pour des mesures réalisées à des longueurs d'onde consécutives. Une autre contrainte liée à l'inversion de la matrice de variance-covariance est la nécessité d'avoir un nombre de variables inférieur au nombre d'objets ($n > m$).

Les distances peuvent aussi être calculées dans l'espace des composantes principales, approche qui comporte plusieurs avantages : réduction de la dimensionnalité, orthogonalité des CP, hiérarchisation des CP. Comme ce sont les CP primaires qui portent la quantité maximale d'information, on peut attendre de l'élimination des CP secondaires, non porteuses de sens, une réduction des données avec une faible perte d'information.

Si le nombre de CP conservées est suffisant pour que le modèle reflète fidèlement les données, la distance euclidienne des points au centre des données sera identique qu'elle soit calculée à partir des scores des CP, ou des coordonnées initiales. Cela paraît logique si l'on considère que le calcul par ACP ne transforme pas les données mais en extrait simplement les variables latentes pour décrire l'espace des données, sans introduire de distorsion. Il en va de même lorsque l'on utilise les distances de Mahalanobis, les valeurs étant identiques que l'on travaille dans l'espace des données originelles ou dans celui des CP. La seule différence réside dans la simplification du calcul des distances de Mahalanobis. Grâce à l'orthogonalité des CP, ces distances peuvent être calculées comme les distances euclidiennes sur l'ensemble des scores réduits, multipliés par un facteur $\sqrt{n-1}$.

2-3-3. **Analyse discriminante**

2-3-3-1. *Principe*

On distingue deux types d'analyse discriminante : l'analyse discriminante linéaire (LDA, pour « *linear discriminant analysis* ») et l'analyse discriminante quadratique (QDA, pour « *quadratic discriminant analysis* »). Dans les deux cas, l'appartenance d'un objet x_i à l'un des K groupes ou classes prédéfinis dans l'ensemble des données est déterminée par la fonction de classement, exprimée par l'équation suivante dans le cas de la LDA :

$$cf(x_{i,K}) = (x_i - \overline{x}_K)^T \boldsymbol{C}^{-1} (x_i - \overline{x}_K) + \ln |\boldsymbol{C}| - 2 \ln (\pi_K)$$

où π_K est la probabilité préexistante du groupe K, égale au nombre d'objets contenus dans ce groupe divisé par le nombre total d'objets composant le groupe d'apprentissage, $\boldsymbol{C}$ la matrice de variance-covariance et $|\boldsymbol{C}|$ son déterminant.

2-3-3-2. *Aspects critiques*

La LDA présuppose que la matrice de variance-covariance est identique pour toutes les classes, tandis que la QDA estime une matrice de variance-covariance pour chaque classe. Le nombre de paramètres à estimer est de ce fait beaucoup plus élevé, il est donc recommandé de ne recourir à la QDA que si l'on dispose de données suffisantes.

2-3-3-3. *Utilisations possibles*

La LDA peut être utilisé pour des tâches de classification plus simples.

2-4. *SIMCA*

2-4-1. **Introduction**

La méthode SIMCA (pour « *soft independent modelling of class analogies* »), ou modélisation indépendante de sous-espaces locaux) est une méthode de classification supervisée des données. Elle nécessite l'utilisation d'un groupe d'apprentissage composé d'échantillons présentant des attributs connus, de même que leur assignation à différentes classes. Les classes peuvent se recouvrir partiellement et comporter des éléments communs. SIMCA assigne les échantillons en utilisant deux approches différentes. Chaque échantillon est affecté à l'une des classes les plus proches, avec des limites de confiance généralement fixées à 95 pour cent, ou il est affecté à aucune ou à une seule des classes, si le seuil repose sur la maximisation de la spécificité et de la sensibilité de la classe.

2-4-2. **Principe**

On procède en premier lieu à l'établissement de modèles ACP pour chaque classe. Les échantillons du groupe d'apprentissage font l'objet d'une analyse ACP (voir section correspondante) et un modèle en composantes principales est généré pour chacune des classes par validation croisée. Le nombre de composantes principales pertinentes peut être ajusté séparément pour chacune des classes. On peut ainsi réduire l'ensemble de données correspondant à chaque classe au modèle ACP pertinent.

Cette analyse préliminaire permet d'opérer ensuite une classification de nouveaux objets sur la base des modèles ACP individuels établis. Chaque nouvel objet est confronté à chacun de ces modèles, et il est assigné à une classe lorsque sa distance résiduelle au modèle correspondant est inférieure à la limite associée à cette classe (voir figure 5.21.-7). Cette distance peut être calculée par des métriques telles que la distance euclidienne ou la distance de Mahalanobis. Par conséquent, un objet peut appartenir à une ou à plusieurs classes si la distance correspondante est inférieure au seuil fixé. Si un objet présente une distance supérieure au seuil pour toutes les classes SIMCA, il est classé comme aberrant.

2-4-3. **Aspects critiques**

La SIMCA reposant essentiellement sur les principes de l'ACP, le processus de validation de la méthode est le même que pour l'ACP. Il faut en outre prendre en considération et analyser les éventuels recouvrements entre classes. Par exemple, à une molécule peuvent correspondre plusieurs groupes chimiques qui apparaissent dans son profil spectroscopique.

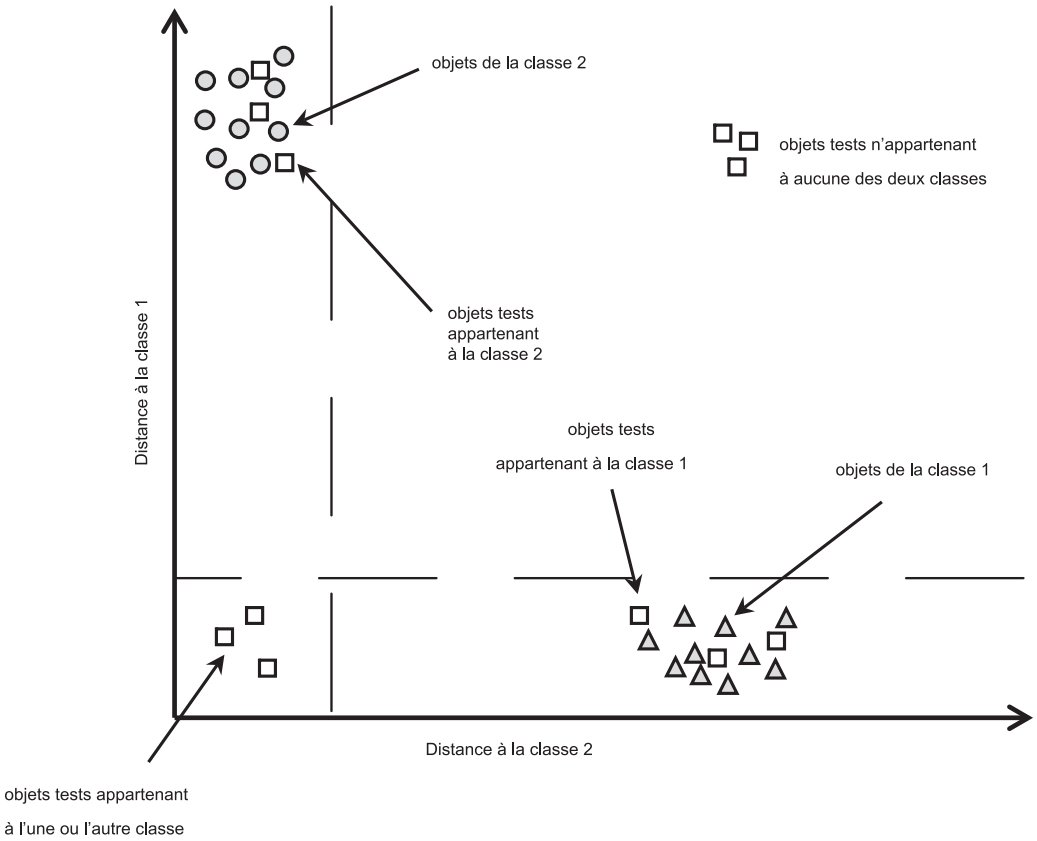


Figure 5.21.-7. – *Graphe représentant les 4 classifications possibles d'objets tests dans le cadre d'une analyse SIMCA à 2 classes (□ = échantillons inconnus à classer, ∆ = échantillons de la classe 1, ○ = échantillons de la classe 2)*

Le fait de regrouper ces données en sous-groupes chimiques entraîne naturellement des superpositions, leur séparation totale n'étant pas possible. En outre, il est nécessaire de disposer de paramètres statistiques pour évaluer le modèle, généralement la fidélité, la sensibilité, la spécificité, l'exactitude, le taux de classification correcte et la proportion d'échantillons non attribués.

2-4-4. **Utilisations possibles**

La SIMCA est un outil très utilisé pour la classification des données spectroscopiques issues par exemple de la spectroscopie dans le proche infrarouge, de la spectrométrie de masse ou d'autres techniques analytiques comme la chromatographie ou l'imagerie chimique. La SIMCA se prête davantage que l'ACP à classe unique à la discrimination de classes difficiles à séparer, au prix du maintien de plusieurs modèles ACP. Toutefois, les modèles ACP à classe unique sont plus faciles à maintenir.

2-5. *PARTITIONNEMENT (CLUSTERING)*

2-5-1. **Introduction**

Les outils de partitionnement des données (*clustering*), permettent de révéler des modes d'auto-organisation des données en groupes distincts et/ou d'apprécier le degré de similarité entre des objets. L'objectif est d'identifier, dans un ensemble de données, des groupes ou classes (*clusters*), c'est à dire des sous-ensembles d'objets ou points proches les uns des autres. Les points appartenant à un groupe partagent certaines caractéristiques qui les différencient des points d'un autre groupe. La caractérisation d'un groupe peut s'effectuer sur la base de trois grandes propriétés : taille, forme et distance au groupe le plus proche. Le partitionnement est utilisé à la fois à des fins d'exploration et de confirmation. Il s'agit d'une méthode non supervisée, à la différence de l'analyse discriminante qui procède à une classification supervisée en assignant un objet sans étiquette à un groupe d'objets prédéfini.

2-5-2. **Principe**

Il existe de nombreuses approches de l'analyse de partitionnement, toutes nécessitent de choisir une mesure de la similarité/dissemblance ou de la distance entre les objets, la métrique euclidienne étant typiquement retenue. Une première distinction peut être faite entre partitionnement hiérarchique et non hiérarchique. Un partitionnement hiérarchique conduit à une représentation graphique arborescente des données exprimant la hiérarchie des regroupements, tandis qu'un partitionnement non hiérarchique révèle des regroupements sans imposer de structure hiérarchique. De nombreux algorithmes sont décrits dans la littérature ; chacun effectue une partition des données selon une voie spécifique ou par optimisation d'un critère particulier. Cette distinction simple et exclusive entre deux modes de partitionnement est cependant incomplète, car il existe des algorithmes mixtes proches à la fois de la hiérarchisation et du simple regroupement. Le partitionnement hiérarchique peut procéder par identification des regroupements selon un processus récursif d'agrégation (classification ascendante) ou de division (classification descendante), pour construire une structure en arbre (dendrogramme). Si l'on opère par agrégation, on commence en considérant chaque point comme un groupe d'effectif 1, puis on fusionne progressivement par paires les groupes les plus similaires, jusqu'à regroupement de tous les points en une classe unique (figure 5.21.-8). Si l'on procède au contraire par division, on considère au départ l'ensemble des points comme un groupe unique que l'on divise progressivement par paires jusqu'à obtention de groupes d'effectif 1. Ces deux modes aboutissent à une structure hiérarchique des groupes, chaque groupe étant composé de sous-groupes. Les algorithmes utilisés peuvent aussi différer par le mode de calcul de la similarité inter-groupes (méthode des liens, ou *linkage* en anglais). Les algorithmes d'agglomération selon la distance minimale (« lien simple » ou *single linkage*) et les algorithmes d'agglomération selon la distance maximale (« lien complet » ou *complete linkage*) évaluent la similarité inter-groupes en calculant la distance entre toutes les paires d'objets appartenant à des classes différentes, mais dans le premier cas la distance calculée est la plus petite distance séparant 2 objets de 2 groupes différents tandis que dans le second cas la distance calculée est la plus grande distance séparant 2 objets de 2 groupes différents. L'algorithme de Ward, également appelé algorithme de variance minimale, évalue la similarité inter-groupes sur la base de la diminution de la variance intra-groupe lorsque l'on fusionne les 2 groupes les plus similaires. Cet algorithme aboutit généralement à des partitions plus équilibrées que les deux autres critères mentionnés précédemment.

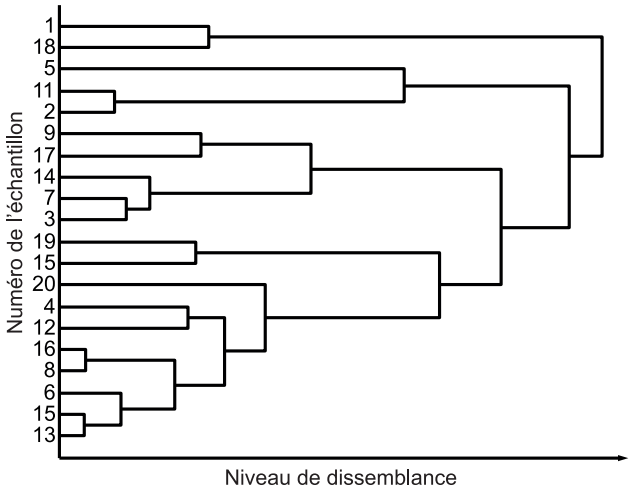


Figure 5.21.-8. – *Dendrogramme obtenu par classification hiérarchique ascendante à partir de points individuels*

Le partitionnement non hiérarchisé n'est pas aussi facile à catégoriser et décrire que le partitionnement hiérarchique. Divers algorithmes conduisent à des schémas de classification différents. Un aperçu des différentes catégories d'algorithmes est fourni ci-après. A côté de méthodes simples reposant sur les distances (recherche d'arbre recouvrant ou voisin réciproque, par exemple), il existe des méthodes plus sophistiquées comme l'algorithme des *K*-moyennes (méthode de partition classique), les algorithmes fondés sur des modèles tels que l'algorithme espérance-maximisation, les algorithmes utilisant la notion de densité tels que le DBSCAN (pour « *density-based spatial clustering of applications with noise* »), les algorithmes de grille tels que le STING (pour « *statistical information grid* »).

L'algorithme de recherche d'arbre recouvrant de poids minimal (algorithme de Kruskal par exemple) se rapproche des algorithmes de la théorie des graphes ; il procède par connexion de tous les points, par des lignes tracées entre les points les plus proches, puis, lorsque tous les points sont connectés, par rupture des lignes les plus longues de façon à conserver les groupes de points étroitement connectés. L'algorithme du voisin réciproque suit une procédure itérative pour assigner un point à un groupe lorsque la distance entre ce point et son voisin immédiat (appartenant à un groupe) est inférieure à une valeur seuil prédéfinie. Les méthodes de recherche des voisins réciproques, comme celle des *k* plus proches voisins, sont généralement utilisées dans le cadre d'un apprentissage supervisé à des fins de prédiction et de classification. La classification et la prédiction sont couvertes dans les sections suivantes (séparatrices à vaste marge et autres méthodes de régression).

Les approches non hiérarchiques de l'analyse de partitionnement telles que le classique algorithme des *K*-moyennes nécessitent de choisir *a priori* le nombre des groupes ainsi que la position initiale de leurs centres. Un critère d'erreur quadratique mesure la somme des distances quadratiques entre chaque objet et le barycentre (noyau) du groupe correspondant. La procédure commence par une partition aléatoire initiale et progresse par réassignation progressive des objets à des groupes, jusqu'à minimisation du critère. Certaines variantes de l'algorithme des *K*-moyennes permettent de diviser ou fusionner les groupes, et donc de trouver le nombre optimal de groupes même lorsque l'on part d'une partition initiale arbitraire. Les algorithmes fondés sur des modèles partent de modèles supposés pour parvenir au meilleur ajustement aux données. L'algorithme espérance-maximisation (EM), par exemple, assigne chaque objet à un groupe particulier en fonction de la probabilité d'appartenance associée à cet objet. Cette probabilité suit une loi gaussienne multivariée, et l'algorithme procède par ajustement itératif aux données en utilisant l'estimation de vraisemblance maximale. Il est considéré comme une extension de l'algorithme des *K*-moyennes, car la somme quadratique résiduelle utilisée pour la convergence des *K*-moyennes est semblable au critère de vraisemblance maximale.

Les algorithmes utilisant la notion de densité, comme le DBSCAN, assimilent les groupes à des régions de haute densité séparées par d'autres régions de densité faible ou nulle. L'algorithme examine le voisinage de chaque objet pour évaluer le nombre d'objets situés dans un rayon spécifié, et un groupe est détecté lorsqu'un nombre défini d'objets sont situés à proximité.

Les algorithmes de grille tels que le STING divisent l'espace des données en un nombre fini de cellules. La distribution des objets dans les différentes cellules est calculée en termes de moyenne, variance, minimum, maximum et type de distribution. Il existe plusieurs niveaux de cellules, apportant plusieurs niveaux de résolution. Chaque cellule d'un niveau particulier correspond à l'union de 4 cellules filles du niveau inférieur.

2-5-3. **Aspects critiques**

Les algorithmes sont tributaires des paramètres que l'utilisateur doit spécifier pour initialiser le regroupement des données. Par exemple, l'algorithme des *K*-moyennes nécessite de choisir un nombre prédéfini de groupes. La partition opérée variera selon le nombre de groupes choisi et la position des centres choisis par l'algorithme. Il est donc important d'exécuter l'algorithme plusieurs fois, en utilisant différents centres initiaux choisis au hasard. La métrique utilisée pour le calcul des distances exerce elle aussi une influence sur le regroupement des données. Avec les distances euclidiennes, les *K*-moyennes définiront des groupes sphériques, là où les distances de Mahalanobis donneraient des groupes ellipsoïdaux. La forme du groupe peut également être modifiée par les prétraitements appliqués aux données avant l'analyse de partitionnement. Une méthode classique de prétraitement des données consiste à les normaliser à l'intervalle unitaire [0, 1] (en soustrayant le minimum et en divisant par l'intervalle de chaque variable). La normalisation permet de garantir une même pondération pour toutes les variables. Les algorithmes utilisant la notion de densité peuvent traiter des groupes de forme arbitraire, mais leur point faible réside dans le traitement de données de haute dimensionnalité où les objets sont distribués de façon dispersée entre les différentes dimensions. Dans le cas des algorithmes de partitionnement hiérarchique, il n'est pas nécessaire de spécifier à l'avance le nombre de groupes. Le nombre de groupes est déterminé en coupant le dendrogramme à un niveau prédéfini. Les groupes résultants peuvent être fortement dépendants du type de lien choisi dans l'algorithme.

Lorsqu'un objet est considéré comme appartenant à un groupe avec une certaine probabilité, certains algorithmes tels que ceux utilisant la densité permettent un partitionnement souple, ou diffus. Dans la région frontière entre deux groupes adjacents, certains objets peuvent appartenir aux deux groupes.

2-5-4. **Utilisations possibles**

L'analyse de partitionnement aide à la compréhension de la structure des données, en regroupant par classes des objets qui partagent les mêmes caractéristiques. Le partitionnement hiérarchique permet en outre d'établir une classification dans l'espace des données. Ce type d'analyse est utilisé dans de très nombreux domaines, en particulier, pour faire émerger des informations de grands ensembles de données. L'objectif est d'extraire des informations cachées et inexploitées de grands volumes de données brutes (fouille de données), en recherchant des associations, tendances et relations entre les variables.

2-6. *ARBRES DE DÉCISION ET FORÊTS ALÉATOIRES*

2-6-1. **Arbres de décision**

Un arbre de décision (CART, pour « *classification and regression trees* ») est une technique d'analyse de données produisant une structure comparable à un organigramme. Utilisés au cours de l'apprentissage et pour la prédiction, ces arbres peuvent être utilisés pour une régression ou une discrimination et sont développés selon une stratégie appelée « diviser pour régner » qui suit des règles logiques de type « si-alors ». Un arbre est composé de nœuds de décision et se termine par des feuilles, également appelées nœuds terminaux. Chaque chemin allant de la racine à une feuille donnée correspond à une série de nœuds de décision successifs et peut donc se traduire par une séquence d'étapes « si-alors », ce qui le rend facile à interpréter.

L'apprentissage d'un arbre de décision s'effectue pas à pas en sélectionnant à chaque étape la variable la plus explicative sur un jeu d'étalonnage. Si cette variable est quantitative, un seuil optimal est déterminé de façon à séparer la population en deux groupes : les échantillons inférieurs au seuil sont affectés dans la partie gauche du nœud et ceux supérieurs au seuil qui sont affectés à la partie droite du nœud. Si la variable est une catégorie, alors la séparation se fait en fonction des catégories. Le choix de la variable et du seuil optimal sont déterminés en fonction d'un critère à optimiser, comme par exemple l'erreur quadratique dans le cadre d'un modèle de régression. Les divisions se succèdent jusqu'à ce que chaque nœud ne contienne plus que très peu de points. En règle générale, pour terminer le processus, les algorithmes d'arbres de décision font des choix à chaque division en vue d'obtenir la perte la plus faible (généralement l'erreur quadratique moyenne dans le cadre de la régression). L'étalonnage des arbres de décision est non paramétrique. Il existe différents modèles et algorithmes d'arbres de décision, selon le test de division choisi, l'interpolation effectuée aux feuilles, la fonction de perte, etc.

Pour effectuer des prédictions à partir de nouvelles données, le test de division est appliqué aux données d'entrée en commençant par la racine et, à chaque nœud de décision, l'une des branches est suivie en fonction du résultat. La recherche s'arrête lorsque l'on arrive à une feuille, c'est-à-dire lorsque l'instance d'apprentissage la plus proche des données d'entrée est trouvée ou interpolée. La prédiction finale est alors généralement la moyenne des échantillons d'étalonnage de la feuille dans le cas d'une régression, ou bien la classe majoritairement représentée dans le cas d'une discrimination.

2-6-2. **Forêts aléatoires**

Pour remédier aux problèmes d'instabilité et de déséquilibre rencontrés lors de la construction des arbres de décision, une approche courante consiste à utiliser la technique des forêts aléatoires (RF, pour « *random forest* »), dans laquelle on entraîne un ensemble d'arbres de décision à partir de

sous-ensembles des données d'apprentissage initiales préparés de manière aléatoire (« *bootstrap* »), puis on détermine la moyenne. La prédiction finale d'un échantillon est calculée sur l'ensemble des arbres de la forêt, par exemple en moyennant les résultats pour le cas de la régression. Les RF font partie des méthodes d'apprentissage dites « ensemblistes ».

Lors de la croissance des arbres, l'algorithme des forêts aléatoires introduit une dimension aléatoire supplémentaire. Au lieu de rechercher la meilleure similarité entre les variables lors de la division d'un nœud, la RF recherche la meilleure variable parmi un sous-ensemble aléatoire de variables. Il en résulte une plus grande diversité d'arbres, qui se traduit par un biais plus élevé pour une variance plus faible. Les arbres de décision individuels sont ensuite combinés. Les RF exploitent une technique appelée « *bagging* » (contraction de *bootstrap aggregating*) pour générer plusieurs arbres de décision. Le *bagging* est une technique d'échantillonnage avec remplacement d'un sous-ensemble des données d'apprentissage. Un échantillon aléatoire des données est prélevé, avec remplacement, et utilisé pour entraîner un arbre de décision. Pour chaque sous-ensemble prélevé, environ 2/3 de l'ensemble d'apprentissage est conservé. Une validation croisée avec le reste de l'ensemble d'apprentissage est effectuée à chaque cycle afin de déterminer les caractéristiques les plus utiles pour l'arbre spécifique en cours de construction. Ce processus est répété pour entraîner un autre arbre, et ainsi de suite.

La prédiction pour un point correspondant à une donnée unique est obtenue en calculant le résultat pour chaque arbre de décision, puis en faisant la moyenne de sortie (agrégation) pour obtenir le résultat final pour le point en question.

En établissant la moyenne de plusieurs arbres de décision construits à partir de différentes parties de l'ensemble d'apprentissage, les RF produisent de meilleurs modèles de classification avec une exactitude globale accrue et un moindre risque de surajustement. Le calcul de la moyenne des résultats de nombreux arbres construits ensemble réduit également l'instabilité inhérente à chacun des arbres pris individuellement.

2-6-3. **Aspects critiques**

L'instabilité est l'une des faiblesses des arbres de décision. Les points de division peuvent considérablement changer lorsque des points (données) sont ajoutés ou supprimés. Les décisions prises à un nœud dépendent des points de division précédents. Autre inconvénient, les arbres de décision gèrent mal les problèmes de balance des effectifs. Par exemple, des performances médiocres en matière d'exactitude de la prédiction sur un petit sous-ensemble des données d'apprentissage n'affecteront pas sensiblement le modèle final. Pour remédier à ce problème, il faudrait équilibrer les classes et les plages des variables réponses avant de construire un arbre de décision.

Des améliorations partielles de la robustesse globale des arbres de décision peuvent être obtenues en arrêtant le processus de construction avant d'atteindre les feuilles terminales, ou en développant l'arbre dans son intégralité pour le réduire par la suite. Cependant, ces améliorations se font aux prix d'une moindre exactitude de l'ajustement aux données d'apprentissage pendant le processus de modélisation, et d'une perte d'interprétabilité.

Les RF présentent un problème de dépendance entre les arbres, lié à la sélection aléatoire des données avec remplacement. L'échantillonnage des données se faisant avec remplacement, certaines données peuvent se retrouver dans plusieurs arbres. La même donnée risque d'être présente plusieurs fois dans le même arbre, le rendant inutile et il convient de veiller à développer des arbres présentant des résultats aussi peu corrélés que possible.

Les arbres de décision ont un inconvénient majeur en régression qui est la surestimation des valeurs faibles extrêmes et la sous-estimation des valeurs élevées extrêmes, qui vient du fait que la prédiction résulte de la moyenne des échantillons de la feuille. Par contre, la distribution des variables prédictives n'a pas d'impact majeur sur le modèle, par exemple une variable ayant beaucoup de valeurs à 0 va être problématique en PLS mais sera très bien gérée par des arbres de décision.

2-6-4. **Utilisations possibles**

Les arbres de décision sont des techniques largement appliquées en classification et en régression. Le surajustement peut devenir critique lors de l'application de ces techniques, mais peut être évité en modélisant au moyen de forêts aléatoires, au détriment de l'interprétabilité.

2-7. *RÉSOLUTION MULTIVARIÉE DE COURBES*

2-7-1. **Introduction**

La résolution multivariée de courbes (MCR) est apparentée à l'ACP, mais là où l'ACP recherche des directions qui représentent la variance maximale et sont orthogonales entre elles, la MCR est utilisée pour estimer des profils de contribution (scores MCR) et des profils de composantes pures (*loadings* MCR). La MCR est également connue sous l'appellation SMCR (pour « *self-modelling curve resolution* »). L'optimisation des paramètres MCR met souvent en œuvre l'algorithme des moindres carrés alternés (ALS, pour « *alternating least square* »).

2-7-2. **Principe**

La MCR-ALS estime les profils de contribution $\boldsymbol{C}$ et les profils de composantes pures $\boldsymbol{S}$ à partir de la matrice des données $\boldsymbol{X}$, selon l'équation $\boldsymbol{X} = \boldsymbol{C}\boldsymbol{S}^{\mathrm{T}} + \boldsymbol{E}$, en utilisant la régression classique des moindres carrés (potentiellement sous contrainte) à chaque itération. Dans la première étape, la matrice $\boldsymbol{C}$ est estimée en utilisant une estimation initiale de la matrice $\boldsymbol{S}$, puis l'estimation de $\boldsymbol{C}$ sert à obtenir une estimation actualisée de $\boldsymbol{S}$ qui est ensuite à nouveau utilisée pour affiner $\boldsymbol{C}$ (d'où le terme d'alterné). On procède ainsi jusqu'à convergence. La procédure ALS peut incorporer des informations connues sur le système physicochimique étudié et les utiliser pour contraindre les composantes/facteurs. Par exemple, ni la contribution ni l'absorbance ne peuvent par définition être négatives. Cette contrainte peut être utilisée pour extraire des profils de composantes pures et des contributions à partir d'ensembles de données se comportant correctement. D'autres types de contraintes peuvent être appliqués : égalité, unimodalité, système clos, bilan de masse, etc.

Il est souvent possible d'obtenir une estimation exacte des spectres des composantes pures ou des profils de contribution. Ces estimations peuvent être utilisées comme valeurs initiales pour l'optimisation par la méthode des moindres carrés alternés, sous contrainte. Comme indiqué précédemment, chaque itération produit alors une nouvelle estimation de la matrice des profils $\boldsymbol{S}$ et des profils de contribution $\boldsymbol{C}$. La connaissance physicochimique du processus peut servir à vérifier le résultat, les profils purs et les profils de contribution résolus devant pouvoir être expliqués par les connaissances disponibles. Si les résultats obtenus avec la MCR ne s'accordent pas avec les informations disponibles sur le système, l'application d'autres contraintes est nécessaire.

2-7-3. **Aspects critiques**

La sélection du nombre de composantes est un élément important du calcul ALS, pour l'obtention d'une solution robuste, et l'EFA, éventuellement avec fenêtre mobile de largeur fixe, peut par exemple en apporter une bonne estimation. Les contraintes imposées peuvent par ailleurs être « rigides » ou « souples », les contraintes rigides étant appliquées de façon stricte tandis que les contraintes souples laissent place à des écarts par rapport à la valeur restreinte. En général, en raison des ambiguïtés inhérentes à la solution obtenue, il est nécessaire de traduire par simple régression linéaire les scores de la MCR en, par exemple, une teneur en substance active. Il faut donc connaître la teneur d'au moins un échantillon. L'existence d'une corrélation entre les

variations de deux entités chimiques (ou davantage) induit une déficience de rang : tel est le cas par exemple lorsqu'une espèce se forme alors que l'autre est consommée, ou lorsque deux entités sont consommées à la même vitesse pour en produire une troisième. Le résultat est que la variation des espèces individuelles se trouve masqué. L'analyse simultanée de données obtenues à partir d'expériences indépendantes, en faisant varier les conditions expérimentales, ou à partir de mesures combinées effectuées par deux techniques différentes est en général plus satisfaisante du point de vue stratégique que des analyses individuelles successives des expériences pour lever une déficience de rang.

2-7-4. **Utilisations possibles**

La MCR peut être appliquée lorsque la procédure analytique utilisée livre des données multivariées avec une réponse linéaire ou linéarisable. L'avantage est alors qu'un seul étalon suffit par analyte. Cet avantage est d'autant plus manifeste que les mesures sont au moins partiellement sélectives. Avec des données ne satisfaisant pas aux conditions de linéarité et de sélectivité, l'emploi de plusieurs étalons par analyte est parfois nécessaire pour réaliser un étalonnage. Lorsque l'on ne dispose pas d'une réponse analytique pure pour un analyte, il est également possible d'estimer les vecteurs de départ en appliquant l'ACP, par exemple, à des mélanges d'analytes avec rotation varimax du système de coordonnées ACP. L'application de la MCR en association avec les moindres carrés alternés permet également à l'algorithme de faire varier librement les profils d'analytes. Cette possibilité peut être mise à profit pour la modélisation de profils difficiles à estimer séparément (ligne de base par exemple).

2-8. *RÉGRESSION LINÉAIRE MULTIPLE*

2-8-1. **Introduction**

La régression linéaire multiple (RLM) étend les modèles de régression linéaire inverse univariée à plusieurs variables dépendantes. Elle cherche à expliquer la relation entre une seule variable de réponse (données *Y*) et plusieurs variables indépendantes (données *X*) par une fonction de régression linéaire. Elle pourrait également être étendue à plusieurs variables de réponse (une matrice $\boldsymbol{Y}$ et non plus un vecteur $\boldsymbol{Y}$). Cependant, l'ajustement de réponses multivariées est semblable à l'ajustement séquentiel de chaque réponse individuelle, car la RLM ne tient pas compte de la structure potentielle de la matrice de réponse, contrairement à la régression PLS qui peut exploiter la structure des matrices de réponse.

2-8-2. **Principe**

Dans le cas de *m* variables indépendantes et *n* objets, le modèle est le suivant :

$$y_i = b_0 + b_1 x_{i,1} + \ldots + b_j x_{i,j} + \ldots + b_m x_{i,m} + f_i, \qquad i = 1, \ldots, n,$$

ce que l'on peut exprimer en notation matricielle par la relation :

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{f}$$

L'objectif est de trouver une estimation du vecteur des coefficients de régression $\boldsymbol{b}$ – désigné par $\hat{\boldsymbol{b}}$ – qui permet la minimisation de la somme des erreurs quadratiques entre la fonction de régression estimée et les réponses observées y_i :

$$\hat{\boldsymbol{b}} = \arg \min_b \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i \boldsymbol{b} \right)^2$$

Les estimations respectives des paramètres du modèle $\hat{\boldsymbol{b}}$ sont obtenues en résolvant un ensemble d'équations simultanées (dites équations normales) dont la solution est :

$$\hat{\boldsymbol{b}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$$

L'estimation des paramètres du modèle implique l'inversion de la matrice $(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})$. Si la dimension *m* devient trop grande et que les variables mesurées sont corrélées, on obtient une matrice $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ mal conditionnée, voire singulière. Une matrice $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ mal conditionnée donnerait un estimateur $\hat{\boldsymbol{b}}$ très variable, ce qui n'est pas souhaitable, et une matrice $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ singulière rendrait même impossible l'obtention d'une solution unique pour $\hat{\boldsymbol{b}}$.

2-8-3. **Aspects critiques**

La RLM exige des variables *X* non corrélées pour pouvoir estimer correctement le paramètre $\boldsymbol{b}$ du modèle. Dans le cas des échantillons pharmaceutiques, l'existence d'une matrice complexe dans laquelle les espèces interagissent à différents degrés rend difficile la sélection des variables non corrélées. Par exemple, en spectroscopie ultraviolette, les valeurs de l'absorbance sélective observées pour un analyte donné sont fortement corrélées car elles dépendent des variations de la concentration de l'analyte. On peut également observer un comportement similaire avec les mélanges si les concentrations de l'analyte varient dans le même sens. Dans les deux cas, des corrélations multiples entre les longueurs d'onde, appelées multicolinéarités, apparaissent dans les données *X*, ce qui rend difficile, voire impossible, l'estimation fiable de $\hat{\boldsymbol{b}}$. Les prédictions d'un tel modèle, très variables, ne sont donc pas fiables.

De ce fait, la RLM n'est employée en chimiométrie que pour des problèmes de régression simples et de faible dimension, ou pour évaluer des matrices $\boldsymbol{X}$ conçues spécifiquement à l'aide de la technique du DoE. Dans ce dernier cas, les matrices $\boldsymbol{X}$ sont conçues de manière à ce que $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ puisse facilement être inversé afin d'obtenir des estimateurs $\hat{\boldsymbol{b}}$ fiables. Pour cela, il faut que l'analyste puisse fixer les valeurs *X* selon les indications du DoE.

La RLM présente un certain nombre de contraintes et tendances :

- le nombre des variables *X* doit être inférieur au nombre des échantillons ($m < n$), condition nécessaire pour que la matrice $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ puisse être inversée,
- en cas de multicolinéarité entre les variables *X*, les coefficients *b* ne sont pas fiables et le modèle peut être instable,
- la RLM tend à produire un surajustement.

Pour éviter un surajustement, on utilise souvent la RLM en combinaison avec une sélection de variables. La sélection des variables *X* optimales repose généralement sur des estimateurs de l'erreur de prédiction obtenus, par exemple, par validation croisée. Étant donné que la sélection des variables elle-même tend à produire un surajustement, le modèle sélectionné doit faire l'objet d'une évaluation rigoureuse (recours nécessaire à un groupe test important ou à de nombreux groupes tests indépendants). Le surajustement dans la sélection des variables est en grande partie évité en utilisant les récentes techniques de régression régularisée ayant la capacité intrinsèque de sélectionner des variables, comme le « LASSO » (pour « *least absolute shrinkage and selection operator* »).

2-8-4. **Utilisations possibles**

La RLM est généralement adaptée aux matrices/groupes de données *X* simples et de faible dimension, pour lesquels le degré de corrélation entre les variables *X* est faible. Lorsque les matrices deviennent plus complexes, il est préférable de recourir à des méthodes telles que la RCP et la PLS pour construire des modèles d'étalonnage fiables.

2-9. *RÉGRESSION SUR COMPOSANTES PRINCIPALES*

2-9-1. **Introduction**

La RCP est une extension de l'ACP à des applications quantitatives. Elle peut également être considérée comme une extension de la RLM au cas des matrices $\boldsymbol{X}$ mal conditionnées et singulières. Elle procède en deux étapes. La première est la transformation par ACP de la matrice d'étalonnage $\boldsymbol{X}$ pour obtenir une matrice des scores $\hat{\boldsymbol{T}}$ et une matrice des *loadings* $\hat{\boldsymbol{P}}$. La seconde consiste à appliquer une régression linéaire multiple à la matrice des scores, pour les composantes

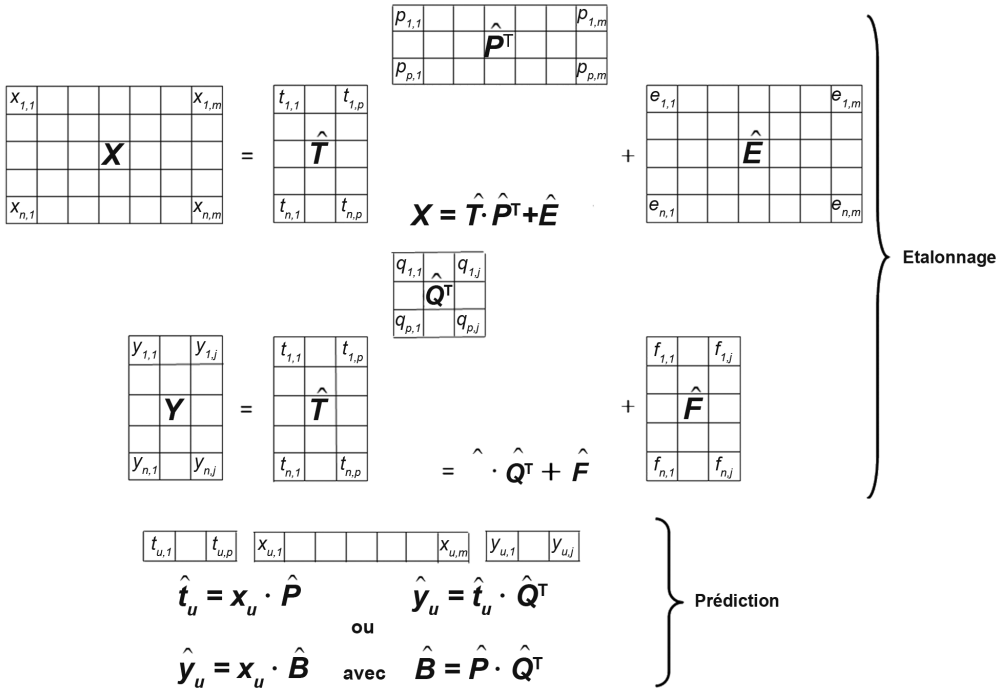
principales, afin d'établir la relation entre les données *Y* et les données *X*.

2-9-2. **Principe**

Comme pour l'ACP, la matrice d'étalonnage est décomposée en une matrice des scores et une matrice des *loadings* de façon à minimiser la somme des carrés des résidus stockés dans la matrice $\hat{\boldsymbol{E}}$, qui ne comporte dans l'idéal que les erreurs aléatoires (bruit). Pour l'étalonnage quantitatif, il faut recourir à une matrice supplémentaire $\boldsymbol{Y}$ des données analytiques de référence des échantillons d'étalonnage. Étant donné que la matrice de données originale $\boldsymbol{X}$ est remplacée par la matrice des scores orthogonale $\hat{\boldsymbol{T}}$, la RLM se prête bien à l'estimation des coefficients de régression, désignés ici par la matrice $\hat{\boldsymbol{Q}}^{\mathrm{T}}$. L'objectif est à nouveau de trouver des coefficients de régression permettant de minimiser la somme des carrés des résidus dans la matrice $\hat{\boldsymbol{F}}$. Dans la figure 5.21.-9, la RPC est présentée pour des réponses multivariables (à savoir *j* valeurs de propriété par objet). Comme la RLM est utilisée pour estimer $\hat{\boldsymbol{Q}}^{\mathrm{T}}$, la structure potentielle dans la matrice $\boldsymbol{Y}$ n'est là encore pas exploitée et la $j^{ème}$ colonne de la matrice $\hat{\boldsymbol{Q}}^{\mathrm{T}}$ ne changera pas si la régression repose uniquement sur la $j^{ème}$ variable de réponse. $\hat{\boldsymbol{Q}}^{\mathrm{T}}$ relie la matrice des scores $\hat{\boldsymbol{T}}$ à $\boldsymbol{Y}$. Dans de nombreux cas, il est plus pratique de travailler dans l'espace initial des données d'étalonnage. Pour cela, $\hat{\boldsymbol{Q}}^{\mathrm{T}}$ peut être rétrotransformé en coefficients de régression $\hat{\boldsymbol{B}}$ dans l'espace de données initial en utilisant les *loadings* $\hat{\boldsymbol{P}}$.

2-9-3. **Aspects critiques**

Un des points critiques du développement de modèles RCP est la sélection du nombre optimal de CP. A cet égard, le tracé, du nombre de CP par rapport à la variance $\boldsymbol{Y}$ résiduelle pour les objets mis de côté, comme dans certains systèmes de validation croisée, est un outil de diagnostic extrêmement utile pour définir le nombre optimal de CP. Le nombre de CP pour lesquelles on observe la variance $\boldsymbol{Y}$ résiduelle minimale dans la validation du modèle est ensuite utilisé dans l'évaluation du modèle pour prédire les objets du groupe test



$\boldsymbol{X}$	=	matrice des données originelles, à *n* lignes et *m* colonnes
$\hat{\boldsymbol{T}}$	=	matrice des scores, à *n* lignes et *p* colonnes
$\hat{\boldsymbol{P}}^{\mathrm{T}}$	=	matrice des loadings, à *p* lignes et *m* colonnes
$\hat{\boldsymbol{E}}$	=	matrice résiduelle (de même rang que la matrice $\boldsymbol{X}$)
$\boldsymbol{Y}$	=	matrice de propriété, à *n* lignes et *j* colonnes
$\hat{\boldsymbol{Q}}^{\mathrm{T}}$	=	matrice de corrélation, à *p* lignes et *j* colonnes
$\hat{\boldsymbol{F}}$	=	matrice résiduelle (de même taille que la matrice Y)
$\hat{\boldsymbol{B}}$	=	matrice des coefficients de régression
m	=	nombre des points de données (variables)
n	=	nombre des mesures (échantillons)
j	=	nombre des valeurs de la propriété par échantillon
p	=	nombre de composantes principales (facteurs)
$\boldsymbol{x}_u$	=	données relatives à un échantillon inconnu
$\hat{\boldsymbol{y}}_u$	=	valeurs prédites pour un échantillon inconnu
$\hat{\boldsymbol{t}}_u$	=	valeurs des *scores* pour un échantillon inconnu

Figure 5.21.-9. – *Décomposition des matrices pour la régression sur composantes principales (RCP)*

indépendant. La prise en compte de CP supplémentaires, au-delà de ce point, n'apporte pas, en général, d'amélioration de la prédiction, elle entraîne au contraire un surajustement du modèle d'étalonnage aux données, donnant lieu à un modèle inutilement complexe dont les performances prédictives sont amoindries.

Malgré l'intérêt de la RCP comme outil gérant la colinéarité des données *X*, cette technique a un point faible qui tient à la décomposition indépendante de la matrice ***X*** par rapport aux réponses ***Y***. Cette approche risque de prendre en compte une variation des données *X* qui n'est pas nécessairement pertinente pour un modèle de régression optimal puisqu'elle n'est pas liée aux données *Y*. De plus, les informations des données *X* liées à *Y*, mais contenues dans les composantes principales d'ordre supérieur, peuvent être écartées lors du processus de sélection du nombre optimal de CP.

Afin d'améliorer les performances du modèle d'étalonnage, il peut être judicieux de procéder pas à pas pour sélectionner les composantes principales, en modifiant l'ordre des CP (par exemple, sélection de CP2 au lieu de CP1).

2-9-4. **Utilisations possibles**

La RCP est une technique d'analyse multivariée qui offre de nombreuses possibilités pour la détection de mesures erronées et l'optimisation des modèles d'étalonnage quantitatifs. En spectroscopie, par exemple, elle apporte des solutions stables pour le spectre total ou pour des régions spectrales étendues des données d'étalonnage. Cependant, elle nécessite en général un plus grand nombre de composantes principales que la PLS, et les limites et inconvénients discutés plus haut expliquent que cette dernière soit aujourd'hui la méthode dominante pour la modélisation quantitative des données spectroscopiques.

2-10. *RÉGRESSION DES MOINDRES CARRÉS PARTIELS*

2-10-1. **Introduction**

La régression des moindres carrés partiels (PLS, pour « *partial least squares* » ou « *projection on latent structures* », projection sur structures latentes) est aujourd'hui l'algorithme le plus utilisé en régression multivariée.

La PLS met en relation deux ensembles de données (*X* et *Y*) indépendamment de leur colinéarité. Elle extrait simultanément les variables latentes des blocs de données *X* et *Y*, tout en maximisant la covariance entre ces blocs. Pour simplifier, on pourrait décrire la PLS comme une double analyse ACP appliquée simultanément aux données *X* et *Y*, où la structure des données *Y* est utilisée pour la recherche des composantes principales des données *X*. La quantité de variance modélisée, c'est à dire la partie expliquée des données, est maximisée pour chaque composante. La partie non expliquée des données est formée par les résidus, dont l'importance constitue une mesure de la qualité de la modélisation.

2-10-2. **Principe**

La différence majeure entre la PLS et la RCP est que la PLS opère une décomposition simultanée des matrices ***X*** et ***Y*** pour inférer les composantes (notées facteurs PLS, facteurs, ou variables latentes). Les facteurs les plus importants contiennent donc les informations qui à la fois décrivent les variations majeures dans *X* et les corrèlent autant que possible avec *Y*. Ce sont précisément ces informations qui sont pertinentes pour la prédiction des valeurs *Y* d'échantillons inconnus. En pratique, la PLS peut être appliquée soit à une seule variable ***Y*** (PLS1) soit à l'étalonnage simultané de plusieurs variables ***Y*** (PLS2).

La description détaillée des algorithmes PLS n'entre pas dans le champ du présent chapitre, mais le schéma de la figure 5.21.-10 en donne une représentation simplifiée. Les flèches tracées entre les matrices des scores $\widehat{\boldsymbol{T}}$ et $\widehat{\boldsymbol{U}}$ symbolisent l'interaction entre leurs éléments lors de l'itération. Alors que la matrice ***Y*** est décomposée en matrice des *loadings* $\widehat{\boldsymbol{Q}}$ et matrice des scores $\widehat{\boldsymbol{U}}$, la décomposition de la matrice ***X*** fournit, en plus de la matrice des *loadings* $\widehat{\boldsymbol{P}}$ et la matrice des scores $\widehat{\boldsymbol{T}}$, une matrice des poids $\widehat{\boldsymbol{W}}$ qui représente l'interrelation entre les données *X* et *Y*.

Pour coupler la décomposition de la matrice ***X*** avec la matrice ***Y*** en première estimation des valeurs de la matrice de scores $\widehat{\boldsymbol{T}}$, on utilise les données *Y* pour « guider » la décomposition de la matrice ***X***. En interchangeant les valeurs des scores des matrices $\widehat{\boldsymbol{U}}$ et $\widehat{\boldsymbol{T}}$, on peut opérer une modélisation interdépendante des données *X* et *Y*, et ainsi réduire l'influence des importantes variations de *X* non corrélées à *Y*. Des modèles d'étalonnage plus simples comportant un nombre réduit de facteurs peuvent en outre être développés. Comme la RCP, la PLS utilise les variances résiduelles pour déterminer le nombre optimal de composantes principales permettant de modéliser l'information et d'éviter un surajustement.

2-10-3. **Aspects critiques**

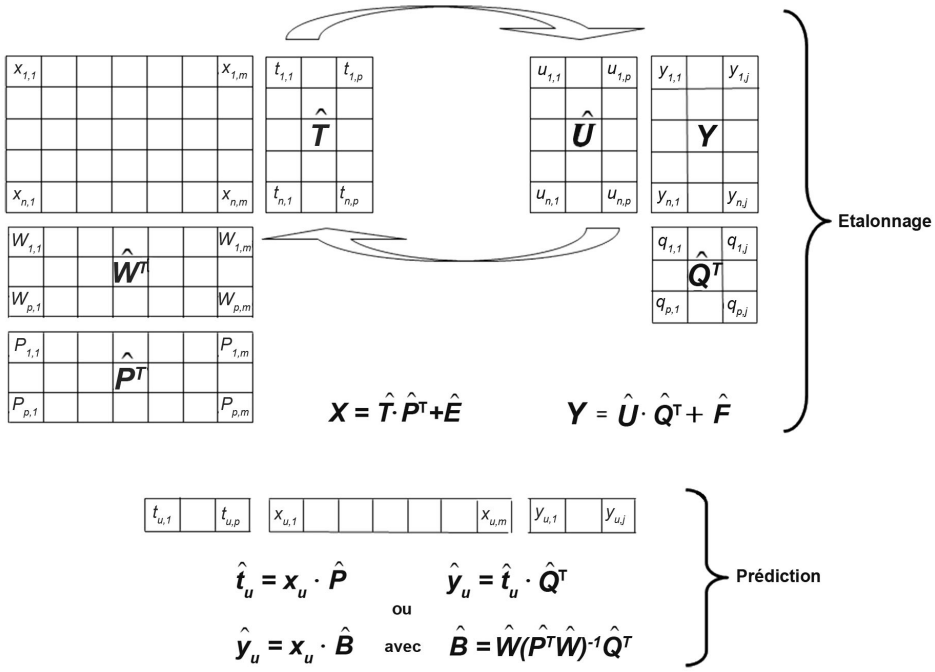
Une étape critique de la PLS est la sélection du nombre de facteurs. Un nombre trop faible de facteurs donnera un modèle incapable d'expliquer de façon satisfaisante la variabilité de l'ensemble des données d'apprentissage, tandis qu'à l'inverse un nombre trop élevé de facteurs conduira à une surmodélisation et à l'instabilité de l'étalonnage résultant (figure 5.21.-11). L'estimation du nombre optimal de facteurs intervient pendant la validation de l'étalonnage. La figure 5.21.-11 montre l'évolution de l'erreur d'étalonnage A d'un modèle et deux cas d'erreur de prédiction (B et C) selon le nombre de facteurs utilisés dans le modèle. L'erreur d'étalonnage décroît de façon continue lorsque le nombre de facteurs augmente. Dans le cas B, l'erreur de prédiction se caractérise par l'absence de minimum, alors que dans le cas C un minimum est observé. En l'absence de minimum, on peut choisir le nombre de composantes au moment où l'erreur cesse de décroître significativement.

Pour ce qui est du choix entre un modèle PLS1 ou PLS2, la modélisation PLS1 s'impose lorsqu'une seule variable ***Y*** est à considérer ; Si l'étude porte sur plusieurs variables ***Y***, on peut soit faire appel à un modèle PLS2 unique soit calculer des modèles PLS1 individuels pour les différentes variables ***Y***. En règle générale, un modèle PLS2 est l'approche recommandée pour un criblage ou dans le cas de variables ***Y*** fortement corrélées. Dans les autres cas, l'emploi de modèles PLS1 séparés pour les différentes variables ***Y*** permettra des prédictions plus satisfaisantes.

2-10-4. **Utilisations possibles**

La PLS s'est imposée comme première alternative à la RCP pour l'étalonnage quantitatif parce qu'elle intègre la structure des données *Y* dans la décomposition de la matrice d'étalonnage ***X***. Les facteurs les plus importants sont donc porteurs d'informations relatives à la fois aux sources de variation dans les données *X* et à leur corrélation avec les données *Y*. Les modèles ainsi établis sont en général plus simples et comportent moins de facteurs que ceux issus de la RCP. La PLS offre également des possibilités d'interprétation et des capacités de diagnostic supérieures pour l'optimisation des performances de l'étalonnage. Elle permet en outre de gérer la présence de bruit, tant dans les données *X* que *Y*.

L'analyse discriminante par PLS est un cas spécial de PLS où une régression est appliquée à la matrice ***X*** pour la transformer en matrice ***Y*** composée de 0 et de 1. Les 0 et 1 indiquent à quelle classe appartiennent ou non les échantillons. Cette technique est utilisée pour des applications semi-quantitatives, par exemple l'estimation de la composition des pixels en imagerie chimique.



$\boldsymbol{X}$	=	matrice des données originelles, à *n* lignes et *m* colonnes
$\boldsymbol{Y}$	=	matrice propriété, à *n* lignes et *j* colonnes
$\hat{\boldsymbol{T}}$	=	matrice des scores, à *n* lignes et *p* colonnes
$\hat{\boldsymbol{P}}^T$	=	matrice des loadings, à *p* lignes et *m* colonnes
$\hat{W}_T$	=	matrice des poids, à *p* lignes et *m* colonnes
$\hat{\boldsymbol{E}}$	=	matrice résiduelle (de même taille que la matrice $\boldsymbol{X}$)
$\hat{\boldsymbol{Q}}^T$	=	matrice des loadings de données $\boldsymbol{Y}$
$\hat{\boldsymbol{U}}$	=	matrice des scores de données $\boldsymbol{Y}$
$\hat{\boldsymbol{F}}$	=	matrice résiduelle (de même taille que la matrice $\boldsymbol{Y}$)
$\hat{\boldsymbol{B}}$	=	matrice des coefficients de régression
m	=	nombre des points de données (variables)
n	=	nombre des mesures (échantillons)
j	=	nombre des valeurs de la propriété par échantillon
p	=	nombre de facteurs
$\boldsymbol{x}_u$	=	données relatives à un échantillon inconnu
$\hat{\boldsymbol{y}}_u$	=	valeurs prédites pour un échantillon inconnu
$\hat{\boldsymbol{t}}_u$	=	valeurs des scores pour un échantillon inconnu

Figure 5.21.-10. – *Décomposition des matrices des données dans la régression PLS*

2-11. *SÉPARATRICES À VASTE MARGE POUR LA CLASSIFICATION SUPERVISÉE*

2-11-1. **Introduction**

Dans cette section, une classification binaire supervisée dans un espace de données (appelé espace objet) sera considérée. Dans ce cas, la frontière de décision peut être assimilée à une ligne, une surface ou un hyperplan, séparant les données en espaces correspondant à deux catégories ou classes, par exemple.

Pour réaliser une classification, les techniques multivariées procèdent par réduction de la dimensionnalité et de la complexité des caractéristiques de l'ensemble des données. En admettant que des modèles linéaires simples puissent être utilisés sans réduire la dimensionnalité, une frontière de séparation pourrait être calculée comme une fonction linéaire des données. Cependant, un tel résultat n'est pas unique et une infinité de frontières de séparation pourraient être calculées. En conséquence, la séparation linéaire n'est ni robuste ni flexible.

Il est également possible d'utiliser des séparatrices à vaste marge (SVM) pour calculer les hyperplans dans l'espace des données et effectuer une classification linéaire. Avec les SVM, la séparation en classes repose sur le calcul de la plus grande marge séparant les deux classes de données, et la meilleure ligne de décision, appelée frontière, est située au milieu de la marge. L'estimation de la marge la plus grande possible entre deux classes de données est une procédure plus exigeante que le simple calcul d'un hyperplan linéaire unique de séparation.

La marge est définie par un ensemble de vecteurs ou de points particuliers dans l'espace des données, appelés « vecteurs de support », à l'origine du nom (anglais) de cette technique particulière (« *support vector machines* »).

2-11-2. **Principe**

2-11-2-1. *SVM linéaires*

Une frontière de décision optimale est calculée en maximisant la distance entre l'hyperplan et les points les plus proches d'un groupe d'appartenance connu de l'ensemble des données d'apprentissage. Cette étape est appelée maximisation de la marge (figure 5.21.-12) et rend possible la classification correcte de nouveaux échantillons en dehors de l'ensemble des données d'apprentissage.

La marge est définie par deux hyperplans parallèles situés à égale distance de la frontière de décision. La séparation/discrimination optimale est obtenue par maximisation de la marge entre les deux groupes. Pour chaque hyperplan définissant la marge optimale, il faut au moins un point, c'est-à-dire un vecteur de support, dans l'espace de redescription. La zone située à l'intérieur de la marge est appelée zone d'indécision. Le déplacement d'un vecteur de support affecte la zone d'indécision et la frontière de décision, mais le déplacement de tout autre point n'affecte pas la frontière.

La distance de chaque point d'apprentissage à l'hyperplan de décision est calculée. Dans le cas d'une séparation en deux classes, le signe affecté à la distance indique l'appartenance au groupe et sa valeur correspond au niveau de certitude de la classification. Lors de la modélisation, la distance des points d'apprentissage à l'hyperplan séparateur contribue au poids qui leur est attribué. Ainsi, les points très distants auront un moindre poids ; les points situés à distance inférieure à un certain seuil ne seront pas pris en considération afin d'éviter un surajustement.

La séparation complète en deux groupes est obtenue avec une marge dure. En général, une séparation parfaite est irréalisable. Il est également possible que des points correctement classés soient situés dans la zone d'indécision.

Ainsi, pour les groupes non séparables, un certain degré de chevauchement est autorisé, et une marge souple est calculée. Des « variables ressort » (*slack variables*) sont affectées aux données ; elles prennent la valeur 0 si la classification est correcte, une valeur positive dans le cas contraire. L'hyperplan optimal s'obtient en maximisant la marge tout en maintenant à un minimum le nombre des points d'apprentissage mal classés (figure 5.21.-12). Un paramètre de régularisation doit être optimisé pour contrôler ce nombre afin d'obtenir un modèle précis et robuste, qui évite le surajustement.

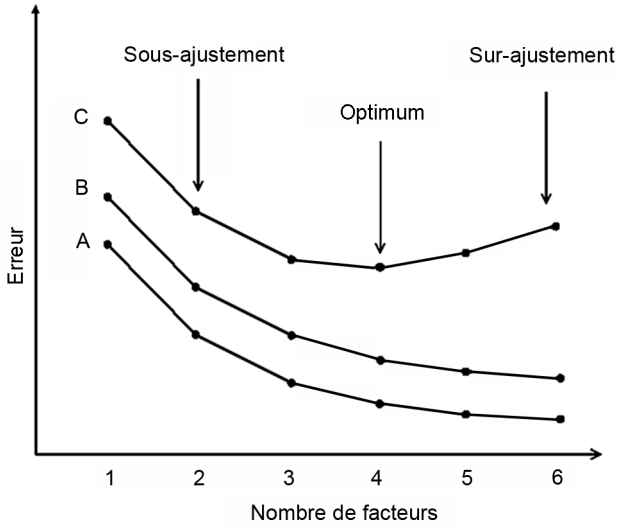


Figure 5.21.-11. – *Effet de l'addition de facteurs à un modèle : l'exactitude de l'étalonnage augmente (l'erreur résiduelle diminue) mais la performance prédictive du modèle peut se dégrader. Le nombre optimal de facteurs est un compromis. (A : étalonnage, B : validation croisée, C : prédiction)*

2-11-2-2. *SVM non linéaires*

Alors que les modèles linéaires sont limités à des espaces de faible dimensionnalité, d'autres modèles gagnent en flexibilité si des caractéristiques sont ajoutés à l'espace initial des données, par exemple en introduisant des interactions entre les points ou des variables polynomiales. Lorsque les données d'apprentissage sont transformées, au moyen de fonctions noyaux, en un espace de redescription de plus grande dimensionnalité, la SVM peut efficacement être étendue aux cas non linéaires.

En pratique, la SVM à noyaux procède au traitement des données en deux étapes. La méthode projette d'abord les données du groupe d'apprentissage dans un espace de redescription de dimensionnalité très supérieure à celle de l'espace des données initial. La projection des données d'un espace à un autre est appelée mappage. Dans un espace de redescription de dimensionnalité élevée possédant certaines propriétés (espace de Hilbert, par exemple), l'utilisation d'une méthode de séparation linéaire comme la SVM avec des marges souples devient possible. La SVM à noyaux procède alors comme une SVM linéaire (section 2-11-2-1), car le problème de classification, une fois reformulé, devient linéaire.

Cependant, il est difficile de trouver un mappage explicite pour un problème de séparation donné et le calcul dans un espace de dimensionnalité supérieure serait irréalisable sans

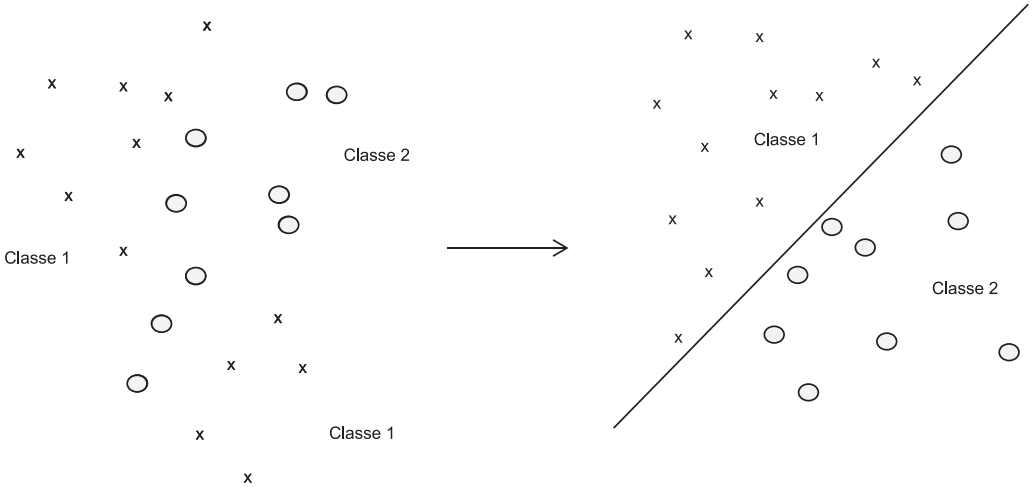


Figure 5.21.-12. – *Transformation de l'espace des objets, avec 2 classes non séparables, en un espace de redescription où une séparation est réalisable*

simplifier l'étape d'optimisation. De plus, il serait nécessaire de définir et de calculer la transformation inverse de l'espace de redescription vers l'espace des données initial.

Il est possible d'éviter ces complications, tout d'abord en utilisant la formulation alternative (double) des SVM, puis en recourant à l'astuce dite du noyau (ou substitution du noyau). Pour transformer l'espace des données en espace de redescription, les données initiales sont étendues par un ensemble de fonctions de base. La sélection de fonctions de base particulières permet de reformuler toute la procédure d'optimisation, de sorte qu'il suffit de calculer les distances entre les paires de points dans l'espace de redescription. Il s'ensuit que seuls les produits des variables étendues (appelés produits scalaires ou produits intérieurs) font encore partie de la procédure d'optimisation. Cette opération écrite en termes de produits intérieurs peut être avantageusement remplacée par une fonction noyau sans faire de simplifications.

Ainsi, une fonction noyau est une opération facile à calculer qui fait directement correspondre deux points quelconques de l'espace des données initial au produit scalaire des points correspondants dans l'espace de redescription, en évitant complètement le calcul explicite de la nouvelle représentation.

Les fonctions noyaux ont des caractéristiques qui peuvent être développées dans la théorie des espaces de Hilbert. Dans le cas des SVM, les fonctions noyaux populaires sont le noyau quadratique (de degré 2), le noyau polynomial (de degré 3 et plus) et la fonction de base radiale, également connue sous le nom de noyau gaussien. Il est aussi possible d'utiliser un noyau linéaire et sigmoïdal. Le noyau nécessite généralement l'optimisation d'un paramètre, comme le degré polynomial pour les noyaux polynomiaux, ou la largeur de la courbe de Gauss pour les fonctions de base radiale. Ce paramètre est critique, tout comme le paramètre de régulation, pour trouver un équilibre entre surajustement et sousajustement, et ainsi obtenir un modèle précis et robuste.

2-11-3. **Aspects critiques**

Il existe pour le calcul des SVM de nombreux algorithmes qui peuvent conduire à des résultats différents. Ainsi, l'optimisation réalisée dépendra des algorithmes utilisés. Les critères de contrôle appliqués peuvent varier, de sorte qu'ils opéreront différemment lors des itérations ou conduiront à des calculs instables, sensibles à l'influence de données redondantes ou non porteuses d'information.

Lors du calcul des SVM, les points d'apprentissage qui se situent nettement à l'intérieur de la frontière délimitant la classe à laquelle ils appartiennent n'ont qu'un effet faible ou nul sur la position du plan de décision. Celui-ci vise principalement les points difficiles à séparer, et non les objets clairement distincts. De ce fait, la SVM est sensible aux valeurs redondantes et aux points atypiques tels que les valeurs aberrantes, par exemple. Il peut donc être pertinent de procéder à une sélection ou un tri de variables spécifiques avant de réaliser la SVM. En outre, avec les SVM, les caractéristiques des données doivent varier selon une échelle de valeurs similaire. Il convient donc de normaliser les données, pour éviter le traitement de données d'entrée dans des échelles totalement différentes, qui crée des conditions médiocres pour l'optimisation des frontières.

Le modèle de SVM identifié comme optimal doit être mis à l'épreuve via une validation. Il faut, à cet effet, disposer d'un ensemble de données indépendant non utilisé lors des itérations et veiller à ce que ces données soient convenablement équilibrées (éviter la concentration des échantillons difficiles dans l'ensemble d'apprentissage et des échantillons faciles dans l'ensemble test, ou vice versa). Dans la pratique, si le jeu de données présente un effectif trop faible, l'optimisation des paramètres pourra s'effectuer par validation croisée. Il sera cependant toujours nécessaire de tester le modèle sur un jeu de données indépendant pour s'assurer de sa robustesse.

Avec les SVM, des lignes de décision peuvent être calculées, même si les données présentent peu de caractéristiques. Les SVM donnent de bons résultats, mais restent toutefois très sensibles au paramétrage et aux données. En pratique, les SVM nécessitent un prétraitement minutieux des données et un réglage fin et simultané de tous les paramètres pendant le calcul. Ceci peut être considéré comme un inconvénient. Bien que les SVM soient utiles dans certaines situations, les modèles basés sur les arbres (forêts aléatoires), ou le renforcement du gradient (*gradient boosting*), pour lesquels le traitement préalable est moindre, voire inexistant, peuvent constituer une alternative intéressante.

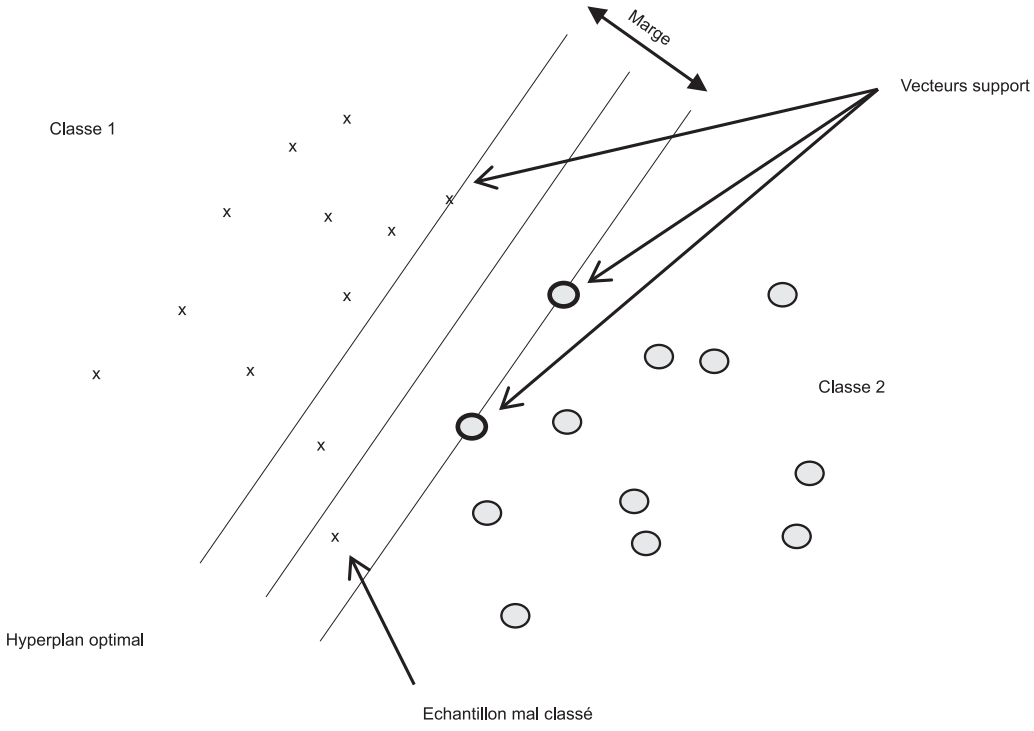


Figure 5.21.-13. – *Réalisation de la séparation des classes 1 et 2 dans l'espace de redescription, avec une tolérance sur les erreurs de classement*

2-11-4. **Utilisations possibles**

Les SVM présentent des performances de pointe sur des problèmes de classification simples. Elles constituent un outil de classification puissant et peuvent être appliquées à une grande variété de types de données. Elles fonctionnent avec des valeurs numériques et nominales, mais sont principalement utilisées dans le cadre de la classification supervisée binaire. Elles peuvent être généralisées à des problèmes de classification multiclasses ou étendues à la régression, bien que ces applications ne soient pas traitées ici. Le processus d'optimisation en SVM a pour moteur les objets difficiles à classer, non les objets clairement distincts (voir figure 5.21.-13). Les paramètres importants dans la modélisation SVM sont le paramètre de régulation, la fonction noyau et ses paramètres associés. Tous ces paramètres doivent être ajustés simultanément. Les SVM peuvent être utilisées pour la séparation de classes d'objets clairement identifiés. Elles donnent de bons résultats sur les grands ensembles de données tels que peuvent en fournir, par exemple, la spectroscopie dans le proche infrarouge, la résonance magnétique nucléaire, l'imagerie chimique ou la fouille de données en cours de processus, dans les situations où l'ACP et les méthodes apparentées échouent en raison du comportement non linéaire dans les données.

Les SVM peuvent être utilisées avec des ensembles de données de faible dimensionnalité, aux caractéristiques réduites, et sont alors particulièrement utiles, car travailler sur l'espace de redescription permet d'augmenter considérablement la dimensionnalité. Cependant, les SVM ont montré qu'elles étaient difficiles à étendre à des ensembles de données contenant de nombreux objets : si les SVM donnent de bons résultats avec de grands ensembles de données (10 000 points, par exemple), elles ne sont pas nécessairement adaptées à 100 000 points et plus. Par conséquent, les SVM peuvent ne pas fournir de bons résultats pour des problèmes tels que la classification d'images.

Le principal atout des SVM est la séparation de données (d'échantillons) associées à des signaux hautement corrélés, tels que des polymorphes, excipients, adultérants et produits falsifiés, ou en particulier si les mesures sont effectuées avec des échelles et des unités similaires.

2-12. *RÉSEAUX NEUROMIMÉTIQUES*

2-12-1. **Introduction**

Les réseaux neuromimétiques ou réseaux de neurones artificiels (ANN, pour « *artificial neural networks* »), encore appelés réseaux connexionnistes, sont des outils de calcul généralistes dont le développement a été initialement inspiré par l'étude des réseaux de neurones biologiques. Ils sont aujourd'hui largement utilisés dans des domaines très divers faisant appel au traitement de données par des systèmes informatisés. Selon l'architecture de ces réseaux, les méthodes utilisées pour construire les modèles ANN ainsi que leurs applications peuvent être extrêmement diverses. Ces réseaux sont généralement utilisés pour la modélisation supervisée et non supervisée, à des fins qualitatives et quantitatives. Historiquement, le premier modèle ANN fut le perceptron, une machine physique destinée à la reconnaissance d'images, utilisant un ensemble de poids combinés à des caractéristiques d'images pour obtenir des résultats de classification. Le perceptron était initialement conçu comme un ANN linéaire à une seule couche, et ne pouvait donc fonctionner que sur des données de classification linéairement séparables. Des versions plus élaborées ont ensuite été développées pour donner naissance au perceptron multicouche, doté d'une plus grande puissance de traitement. En tant qu'outils d'étalonnage et de classification multivariés, les réseaux neuromimétiques sont plutôt utilisés pour cartographier des relations non linéaires. La non-linéarité est obtenue par l'utilisation de fonctions d'activation telles que la fonction sigmoïde à tangente hyperbolique combinée aux poids dans des couches multiples. Il a été prouvé qu'en utilisant cette méthode, il est possible de modéliser n'importe quelle fonction non linéaire.

Le perceptron à une couche est le réseau neuronal non rétroactif le moins compliqué, car il ne comporte pas de boucles. Une autre catégorie de réseaux neuronaux est le réseau de neurones récurrents qui utilise des boucles de rétroaction pouvant être décrites à l'aide de graphes orientés cycliques ou acycliques, selon le type de réseau.

2-12-2. **Principe**

2-12-2-1. *Généralités*

Un réseau neuromimétique est un modèle de traitement des données dont l'élément de base est le neurone artificiel, que l'on peut comprendre comme une fonction mathématique utilisant en entrée la somme d'un vecteur pondéré et d'un biais. Le vecteur, appelé « entrée » du neurone, est soit obtenu directement à partir d'un échantillon faisant partie de l'ensemble des données, soit calculé à partir des sorties des neurones précédents. L'utilisateur choisit la forme de la fonction appliquée, dite d'activation. Les pondérations et les biais sont déterminés par un processus d'apprentissage à partir d'exemples connus, d'exemples générés automatiquement à partir d'un ensemble de règles ou, par exemple, d'un système chimique en cours d'observation. Les réseaux comportent souvent un grand nombre de neurones disposés en couches, composées chacune de neurones opérant en parallèle. Ils sont connectés aux neurones de la couche précédente dont ils reçoivent les données d'entrée, et à ceux de la couche suivante, auxquels ils envoient les données de sortie (figure 5.21.-14). La sortie d'un neurone constitue donc l'entrée des neurones de la couche suivante. La première couche est particulière puisqu'elle reçoit directement ses entrées de l'utilisateur, et transfère directement les données à la couche suivante sans appliquer de fonction d'activation. La dernière couche est elle aussi particulière car sa sortie est directement utilisée, sans autre traitement, comme sortie du modèle. Il existe un nombre illimité de façons de connecter des neurones entre eux, en faisant varier leur nombre ainsi que celui des couches de neurones. Ces connexions définissent l'architecture du réseau et peuvent répondre à toute exigence de modélisation des données, aussi complexe soit-elle.

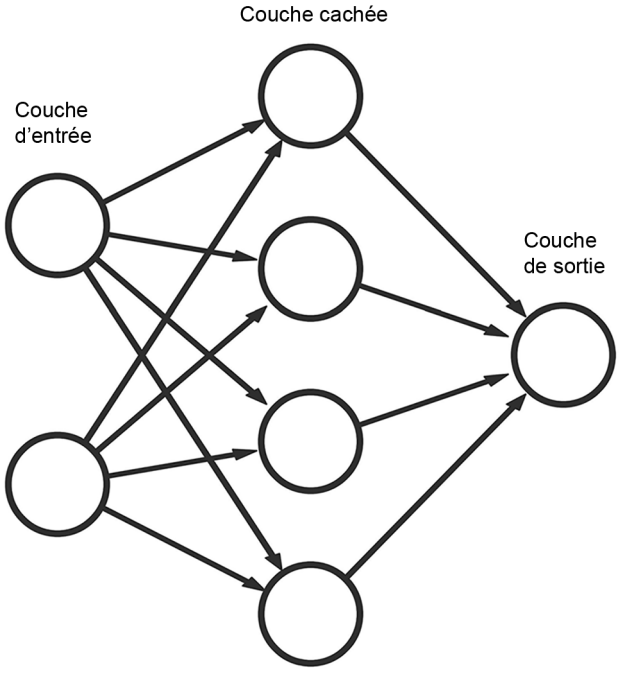


Figure 5.21.-14. - *Disposition type des couches de neurones et interconnexions*

2-12-2-2. *Réseaux neuronaux multicouches non rétroactifs*

Les réseaux neuromimétiques multicouches non rétroactifs (dits MLFF, pour « *multi-layer feed-forward* ») comportent une couche d'entrée et une couche de sortie, séparées par une ou plusieurs couches de neurones cachées. Il n'existe pas de limite au nombre de couches cachées que l'on peut intégrer au modèle, mais un modèle possédant une seule couche cachée est capable de gérer la plupart des tâches d'étalonnage multivarié réalisées en chimiométrie. Dans le modèle MLFF, les neurones sont connectés à la totalité des neurones des couches voisines. La fonction d'activation la plus appliquée initialement était une fonction sigmoïde ou une fonction tangente hyperbolique, mais d'autres fonctions (linéaire par exemple) peuvent aussi être utilisées. L'unité linéaire rectifiée (ReLU) est une fonction d'activation linéaire par morceaux de plus en plus répandue.

Les poids et biais initiaux peuvent être des nombres aléatoires, mais peuvent aussi être initialisés par d'autres algorithmes. Un algorithme d'apprentissage courant pour déterminer les poids et biais finals est la descente de gradient utilisant un algorithme de rétropropagation ou ses variantes. Dans cet algorithme, l'erreur de prédiction, calculée par différence entre la valeur de sortie et la valeur effective, est rétro-propagée et entraîne le calcul des modifications requises pour ajuster les poids et biais afin de minimiser l'erreur de prédiction. La descente de gradient stochastique est une évolution de la descente de gradient, dans laquelle le gradient est estimé à partir de sous-ensembles de données sélectionnés de manière aléatoire. Ainsi, il n'est pas nécessaire de disposer de grandes quantités de données d'apprentissage au même moment dans la mémoire de l'ordinateur, mais la convergence sera légèrement ralentie. Au cours de l'apprentissage du réseau, des unités cachées ou visibles peuvent être temporairement désactivées (« *dropout* ») pour rendre les modèles plus robustes et ne pas dépendre d'un seul chemin à travers le réseau pour parvenir à un résultat donné. Une autre amélioration importante consiste à appliquer une normalisation par lots (« *batch normalisation* ») qui augmente la stabilité numérique et les performances des réseaux neuronaux en mettant à l'échelle les activations.

Pour obtenir d'un réseau neuromimétique les performances voulues, il est nécessaire de l'optimiser. Cette opération porte souvent sur la détermination du nombre de couches, du nombre de neurones par couche, des fonctions d'activation associées à chaque couche et neurone, l'initialisation des poids, la vitesse d'apprentissage, etc.

2-12-2-3. *Réseau neuronal convolutif*

Les réseaux neuronaux convolutifs peuvent être considérés comme des empilements de filtres traitant les données d'entrée par convolutions (figure 5.21.-15). La structure du réseau s'inspire du cortex visuel animal, qui favorise la recherche des formes fondamentales générales associées aux données d'entrée dans les premières couches, alors que les couches suivantes s'attachent à la reconnaissance des caractéristiques modélisées. Les convolutions sont souvent organisées en tuiles qui se chevauchent partiellement, les tuiles réduisant les degrés de liberté par rapport à une couche entièrement connectée. Avant de passer à une autre couche, une opération de réduction de la dimensionnalité – appelée mise en commun ou « *pooling* » – intervient. Par exemple, dans ce qu'on appelle un « *maxpooling* », la sortie maximale d'une sélection de tuiles adjacentes de la couche précédente est introduite dans un neurone de la couche suivante, donnant une couche suivante avec moins de neurones. Il peut y avoir plusieurs couches de convolution avant une mise en commun. Les convolutions et mises en commun combinées agissent comme des extracteurs de caractéristiques. Si la convolution a été réalisée avec plusieurs filtres différents, les sorties de ces derniers sont concaténées (on appelle cela le « *flattening* »). Les dernières couches sont souvent des couches entièrement connectées (formant un réseau neuromimétique traditionnel) dont le but est de classifier ou de quantifier les résultats.

2-12-2-4. *Réseau de neurones récurrents*

Le réseau neuronal récurrent est une classe de réseaux neuromimétiques adaptée à la modélisation de séquences telles que le texte écrit ou parlé, les structures chimiques, les mesures de procédés de fabrication ou d'autres systèmes présentant des comportements dynamiques temporels. Ces réseaux disposent de plus de degrés de liberté et de structures plus denses que les réseaux neuronaux convolutifs. Ils consistent en des connexions en boucle où l'entrée courante est combinée avec les éléments d'entrées précédentes dans une séquence afin de pouvoir modéliser les relations entre les entrées précédentes et courantes.

2-12-2-5. *Réseaux antagonistes génératifs*

Le réseau adversatif génératif est une catégorie de systèmes d'apprentissage automatique dans laquelle deux réseaux neuronaux sont mis en concurrence pour se perfectionner dans la réalisation d'une tâche donnée. Ils peuvent ainsi « s'auto-former » s'ils disposent d'un ensemble de règles suffisant pour travailler.

2-12-2-6. *Apprentissage par transfert*

L'apprentissage par transfert (*transfer learning*) avec les réseaux neuromimétiques consiste à stocker les connaissances acquises lors de la résolution d'un problème et à les appliquer à un problème différent mais connexe. Cette méthode a été appliquée avec succès avec les réseaux neuronaux convolutifs, par exemple, dans lesquels les données d'entrée sont des images. Il est possible d'utiliser les parties fondamentales du réseau formé et d'adapter les couches ultérieures à une nouvelle tâche, à condition que les données d'entrée soient du même type. Ainsi, si le travail réalisé sur un modèle entraîné a permis de couvrir suffisamment de cas pour rendre ce modèle généralisable et de bonne qualité, il est possible d'utiliser le même modèle en réentraînant les couches ultérieures, voire le modèle complet, pour l'adapter à une tâche très différente, sous réserve que les données d'entrée soient du même type. Dans cette situation, la quantité de données nécessaires pour le réapprentissage est bien inférieure à la quantité de données requises pour former le modèle de réseau primaire. Cette méthode peut également être utilisée pour la mise à jour de modèles.

2-12-2-7. *Cartes auto-organisatrices*

L'objectif des cartes auto-organisatrices (algorithme SOM, pour « *self-organising map* ») est la création d'une carte topologique où les observations proches les unes des autres présentent entre elles des similarités de propriétés plus étroites qu'avec les observations plus distantes. Les neurones de la couche de sortie sont généralement disposés sous la forme

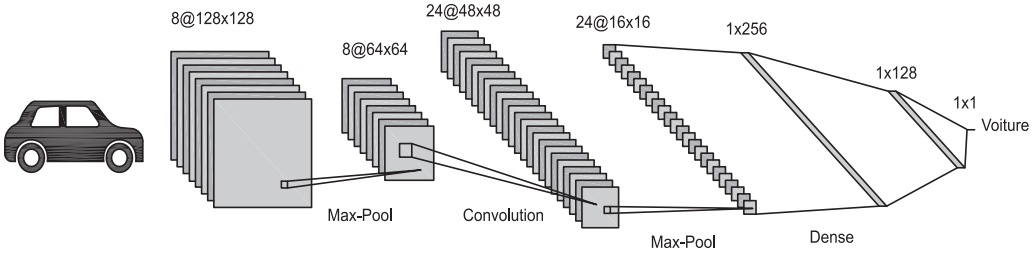


Figure 5.21.-15. – *Exemple de réseau neuronal convolutif configuré en modèle de classification avec des images en entrée et une sortie simple, indiquant l'attribut de chacune des images*

d'une carte bidimensionnelle où chaque neurone peut être figuré par un carré ou un hexagone. L'entraînement s'effectue par un apprentissage compétitif opérant différemment de l'algorithme de rétro-propagation précédemment décrit. Le résultat final est une carte bidimensionnelle des observations.

2-12-3. **Aspects critiques**

Les écueils les plus fréquents de l'utilisation de réseaux neuromimétiques sont le sous-apprentissage et le sur-apprentissage. En situation de sur-apprentissage, le modèle est capable de faire une très bonne prédiction sur l'ensemble d'apprentissage, mais est trop lié à cet ensemble pour constituer un bon modèle de prédiction en dehors de celui-ci. En situation de sous-apprentissage, à l'inverse, l'apprentissage est trop bref pour que le modèle acquière des performances prédictives satisfaisantes. Ces deux écueils peuvent cependant être évités si l'on utilise les réseaux neuronaux. Il faut, pour pouvoir procéder à l'entraînement du modèle ANN, disposer d'un ensemble d'apprentissage de taille appropriée, c'est-à-dire comportant suffisamment d'observations ou d'échantillons pour représenter les variations attendues des données. Un autre ensemble de données distinct, tout aussi représentatif, destiné à la validation interne, est alors généralement nécessaire pour obtenir des informations sur l'état de sur- ou sous-apprentissage lors du développement du modèle. En général, comme les modèles ne sont pas linéaires, le nombre d'observations requis est plus élevé que pour des données comparables soumises à une modélisation linéaire, car les poids à calculer sont plus nombreux. Bien que cela ne soit pas une règle absolue, le nombre d'échantillons minimal conseillé est de 10 par poids à calculer. Le nombre de poids devient vite important selon le nombre de couches et de neurones. Il faut donc faire attention à la complexité et la taille du réseau qui peut vite conduire au sur-apprentissage du modèle. D'autres astuces pour éviter le sur-apprentissage sont le *dropout* qui consiste à retirer quelques poids aléatoirement lors de l'apprentissage, et la régularisation qui permet de limiter l'amplitude des poids ce qui rend le modèle plus stable. Dans le cadre de l'apprentissage par transfert, lorsque des modèles comportant des données d'entrée similaires sont réaffectés, la quantité supplémentaire de données nécessaires à la formation pour le nouvel objectif peut être très faible par rapport à la quantité initiale de données utilisées pour le développement du modèle de base.

Des pré- et post-traitements peuvent être utilisés pour atteindre les objectifs du modèle de réseau neuromimétique. Dans ce cas, des méthodes multivariées linéaires (ACP, PLS ou autres méthodes d'apprentissage automatique) peuvent être utilisées en entrée du modèle ANN pour réduire les degrés de liberté et interpréter de façon plus détaillée la façon dont le modèle obtient ses résultats. Le nombre de poids en est réduit car le nombre de neurones d'entrée est bien inférieur à celui des variables d'origine. Il est aussi possible d'adapter les degrés de liberté dans le réseau aux données disponibles afin d'utiliser une série de couches cachées comportant un nombre variable de neurones.

Il est important de disposer d'une plus grande visibilité quant au fonctionnement d'un modèle de réseau neuromimétique. En d'autres termes, ces modèles doivent être explicables. Cette tâche peut être abordée sous différents angles : l'interprétabilité du modèle, qui consiste à rendre le modèle lui-même interprétable, et l'interprétabilité des prédictions, où des prédictions particulières sont utilisées pour explorer le modèle.

L'interprétabilité du modèle peut être obtenue, par exemple, dans les réseaux neuronaux convolutifs d'analyse d'image en exerçant des couches de coefficients pour révéler sous forme d'images les caractéristiques modélisées pour chaque couche du modèle. L'interprétabilité des prédictions peut être obtenue, par exemple, en alimentant le modèle avec des ensembles de données d'entrée complexes et altérés dont les résultats sont connus. Pour obtenir une bonne image dans l'analyse de l'interprétabilité des prédictions, il peut être nécessaire d'automatiser la création de perturbations pour couvrir suffisamment le comportement du modèle dans les applications en situation réelle. Cette technique, appelée « *data augmentation* », est très couramment utilisée en analyse d'image.

2-12-4. **Utilisations possibles**

L'avantage des réseaux neuromimétiques en étalonnage multivarié réside dans leur capacité à modéliser des relations non linéaires. L'approche de base consiste à avoir des neurones intégralement connectés, c'est-à-dire que toutes les interactions entre variables sont automatiquement répercutées. Il a été démontré que les réseaux comportant un nombre de neurones suffisant pouvaient représenter toute relation entre entrées et sorties, aussi complexe soit-elle.

Les réseaux neuronaux convolutifs sont couramment utilisés pour des tâches d'analyse d'images. Les couches de convolution peuvent être interprétées en termes d'analyse d'image traditionnelle non neuromimétique et de convolutions de structure similaire, mais avec une moindre marge d'adaptation. Les spectres pouvant être considérés comme des images unidimensionnelles, les réseaux neuronaux convolutifs sont également directement applicables aux données de spectroscopie et les poids des couches peuvent être interprétés d'une manière semblable que les *loadings* avec l'ACP et la PLS. En particulier, la convolution peut être comparée au prétraitement des données spectroscopiques.

Les réseaux de neurones récurrents sont utiles dès lors qu'une séquence d'évènements peut bénéficier d'une modélisation, comme par exemple les procédés chimiques, biotechnologiques ou pharmaceutiques.

L'apprentissage par transfert est très utile dans le domaine de l'imagerie, par exemple lorsqu'il est appliqué à de nouvelles situations avec le même type de données d'entrée, comme des images pour le contrôle qualité ou encore, dans le contexte du traitement du cancer, des images de taches cutanées à classer en lésions bénignes ou malignes. Dans le cas de la chimiométrie, cela peut être très utile pour la mise à jour de modèles après, par exemple, une maintenance d'instrument ou pour un transfert entre instruments.

Les cartes auto-organisatrices sont utilisées pour visualiser des données de dimensionnalité élevée tout en préservant la topologie des données originelles. Elles procèdent par apprentissage non supervisé, et sont surtout utiles comme outils d'exploration d'ensembles de données lorsque l'on ne dispose d'aucune connaissance préalable des schémas et relations à l'œuvre dans les échantillons.

Les réseaux neuromimétiques de base intégralement connectés comportent souvent un nombre élevé de coefficients (poids et biais), ce qui leur confère la capacité de modéliser toute relation complexe au sein de l'ensemble des données, mais dans le même temps rend plus difficile l'interprétation de ces coefficients. Cependant, avec une modélisation plus développée, il est maintenant possible d'obtenir des interprétations du contenu des couches cachées grâce à la prolifération des sous-tâches pour les différentes couches cachées. Ainsi, lorsque les méthodes de modélisation linéaire s'avèrent trop peu flexibles et ne parviennent pas à fournir l'exactitude de prédiction ou de classification attendue, les réseaux neuromimétiques peuvent être une bonne alternative.

3. DOMAINES D'APPLICATION CONNEXES

3-1. *LA CHIMIOMÉTRIE DANS L'IMAGERIE CHIMIQUE*

Le chapitre général *5.24* décrit en détail l'imagerie chimique. La combinaison de l'imagerie chimique avec la chimiométrie est précieuse pour l'identification et la quantification des composés présents dans un produit et l'étude de leur distribution.

Les images hyperspectrales sont généralement constituées de centaines de spectres rassemblés dans un hypercube, à savoir une matrice tridimensionnelle comportant deux dimensions

spatiales et une dimension spectrale. Tant que l'hypercube peut être déplié, tous les outils chimiométriques peuvent être appliqués. Le processus de dépliage consiste à transformer le cube de données 3D en une matrice de données 2D : les deux dimensions spatiales sont concaténées en une seule colonne, la deuxième dimension correspondant aux données du signal. La chimiométrie est nécessaire à différentes étapes de l'analyse des données d'image : compression, prétraitement et traitement (exploration, résolution, classification ou régression).

Etant donné la grande taille des données expérimentales obtenues avec les instruments modernes (par exemple, l'imagerie par spectrométrie de masse), des méthodes ont été proposées pour réduire le nombre de valeurs du signal et comprimer les données : filtres en ondelettes, regroupement des données par classe (« *binning* ») ou approches basées sur la sélection de régions d'intérêt.

Les étapes de prétraitement sont les mêmes que pour les mesures analytiques classiques, et consistent principalement à corriger le bruit et la ligne de base. La connaissance des principes physiques et chimiques qui sous-tendent les mesures analytiques est importante pour sélectionner correctement le prétraitement approprié et éviter la perte d'informations ou l'introduction d'artefacts. Par exemple, les spectres Raman peuvent porter un signal de fluorescence qui masque souvent le signal pertinent. Une correction de la ligne de base doit alors être appliquée.

Par ailleurs, il arrive que les détecteurs d'imagerie mesurent des artefacts. Ils proviennent généralement de pixels défectueux ou du rayonnement cosmique (spectroscopie Raman). Ces particules d'énergie élevée saturent un détecteur sous la forme d'un pic intense et étroit. La présence d'un tel artefact peut fausser les résultats et doit absolument être éliminée en utilisant des algorithmes spécifiques basés sur un critère de largeur de pic ou en comparant le pixel contenant le pic avec les pixels « normaux » environnants. L'algorithme des voisins réciproques permet de supprimer ces pics en supposant que deux spectres consécutifs ne peuvent contenir un tel pic à la même position du signal. Le pic de rayonnement cosmique est alors remplacé par des valeurs interpolées obtenues à partir de spectres voisins normaux.

Le bruit de mesure contenu dans l'image hyperspectrale peut être réduit pour renforcer le signal chimique. Il est possible à cet effet d'utiliser une procédure de lissage. Les méthodes chimiométriques de filtrage du bruit peuvent également être appliquées en utilisant l'ACP qui décompose l'ensemble des données initiales en variables latentes. Les premières composantes principales sont liées aux variations chimiques tandis que les dernières contiennent du bruit. L'hypercube est recalculé en ne sélectionnant que les composantes principales pertinentes qui ne modélisent pas le bruit de mesure.

L'extraction des informations contenues dans l'hypercube peut se faire au moyen de toutes les méthodes chimiométriques décrites dans le présent chapitre. Certaines d'entre elles sont particulièrement appropriées : MCR-ALS, ACP, PLS, moindres carrés classiques, ACI, etc.

Comme évoqué précédemment, la MCR-ALS est une méthode très populaire de résolution d'un mélange de signaux sans connaissance préalable du système chimique. Le postulat de cette méthode de décomposition est que le signal de l'échantillon peut être considéré comme la somme pondérée du signal de chaque espèce chimique pure. En général, dans une image chimique, chaque pixel de l'image contient les mêmes signaux purs mais leurs contributions diffèrent d'un pixel à l'autre. Le processus itératif de cet algorithme utilise des contraintes (non négativité, unimodalité, système clos, sélectivité) pour finalement fournir des signaux purs (avec une réelle signification chimique) et leur contribution dans l'image hyperspectrale.

A la fin du processus chimiométrique, la matrice de données sera méthodiquement repliée pour reconstituer l'hypercube et ainsi donner accès, par exemple, à la carte de distribution des espèces chimiques.

3-2. *FUSION DE DONNÉES*

La fusion de donnée est un processus consistant à intégrer des blocs de données provenant de différentes sources analytiques dans un même modèle global. Elle permet généralement d'améliorer l'exactitude des prédictions et conduit à une meilleure interprétation des résultats. Pour être utiles, les données à fusionner doivent fournir des informations complémentaires.

Selon la stratégie adoptée pour fusionner les données, on considère généralement trois niveaux de fusion, correspondant au niveau du cycle analytique des données : bas niveau, niveau intermédiaire et haut niveau.

La fusion de bas niveau utilise les données de tous les blocs simplement concaténées par échantillon en une seule matrice comportant autant de lignes que d'échantillons analysés et autant de colonnes que de signaux mesurés par les différents instruments. La fusion de bas niveau génère un grand volume de données et de nombreuses variables, avec une possible prédominance d'une source de données sur les autres. Il sera alors toujours nécessaire de prétraiter les blocs de données individuels (par exemple, en procédant à une mise à l'échelle des blocs et à une sélection des variables) afin d'éviter que le bloc présentant la variance la plus élevée ne régisse le modèle.

La fusion de niveau intermédiaire utilise les caractéristiques extraites des différents blocs de données pour créer une nouvelle matrice plus petite qui sera traitée par les techniques chimiométriques souhaitées. Ces caractéristiques sont censées capturer uniquement les variations pertinentes dans les différents blocs de données. Il peut s'agir de variables originales significatives ou, dans la plupart des cas, de variables latentes (par exemple, les scores PCA, PLS ou PLS-DA). La fusion de niveau intermédiaire nécessite une combinaison optimale entre les caractéristiques extraites et le prétraitement appliqué, rendant parfois l'ensemble du processus long et fastidieux.

La fusion de haut niveau intervient au niveau décisionnel, c'est-à-dire qu'elle utilise les résultats de modèles chimiométriques distincts qui sont combinés pour fournir la décision finale. Une approche heuristique, basée sur des systèmes de vote ou de notation, ou une approche bayésienne, basée sur l'estimation des probabilités, peut être utilisée pour parvenir à la conclusion finale.

Pour chaque niveau, le défi consiste à trouver la combinaison optimale de prétraitement, de sélection de variables et d'extraction de caractéristiques permettant de décrire la variation significative des réponses instrumentales et de fournir le meilleur modèle final. Des méthodes chimiométriques spécifiques ont été développées pour tenir compte de la variance commune et spécifique provenant de chacun des blocs analysés suite à la réduction de la taille de l'ensemble de données par sélection des caractéristiques.

4. GLOSSAIRE

Aberrant : dans un ensemble de données numériques, qualifie un point (une valeur) statistiquement différent des autres. Se rapporte également à l'échantillon associé à cette valeur. Il existe des tests statistiques spécifiques pour la recherche de valeurs aberrantes.

Analyse exploratoire des données : processus visant à révéler des schémas non attendus ou latents, en vue de la construction d'hypothèses.

Attribut d'un échantillon : propriété (qualitative ou quantitative) d'un échantillon.

Bootstrap : technique consistant à générer plusieurs ensembles d'échantillons de taille *n* à partir d'un ensemble originel de même taille *n*, par sélection aléatoire d'échantillons avec remise.

Centrage : opération destinée à faciliter la comparaison et l'interprétation des données. Elle consiste à calculer la moyenne de chaque variable (si le centrage est réalisé par rapport à la moyenne) que l'on soustrait ensuite de toutes les valeurs prises par la variable ; dans un tableau comportant des colonnes de variables, cette opération est effectuée colonne par colonne.

Colinéaire/non colinéaire : une famille de vecteurs est colinéaire si l'un des vecteurs au moins peut être représenté comme une combinaison linéaire des autres vecteurs. A l'inverse, une famille de vecteurs est non colinéaire si aucun des vecteurs ne peut être comme une combinaison linéaire des autres.

Composante (syn. facteur, variable latente) : dans le contexte de la chimiométrie, variable hypothétique sous-jacente, non observée et non mesurée, qui reflète la variance des variables mesurées. Les variables sont des combinaisons linéaires des facteurs, et ces facteurs sont supposés non corrélés entre eux.

Échantillon : objet, observation ou individu sur lequel sont collectées des données.

Échantillonnage : processus par lequel on constitue un sous-ensemble à partir d'une population, pour estimer les propriétés de cette population.

Erreur quadratique moyenne de prédiction : fonction de la somme résiduelle des carrés prédictive qui permet d'estimer l'exactitude :

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2}{n}}$$

où $\hat{y}_i$ est la réponse prédite pour le $i^{ème}$ échantillon, y_i la réponse observée pour le $i^{ème}$ échantillon et n le nombre d'échantillons.

Erreur type d'étalonnage : fonction de la somme résiduelle des carrés prédictive qui permet d'estimer l'exactitude en prenant en compte le nombre de paramètres :

$$SEC = \sqrt{\frac{\sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2}{n-p}}$$

où n est le nombre d'échantillons de l'ensemble d'apprentissage, p le nombre de paramètres du modèle dont on cherche à établir l'estimation, $\hat{y}_i$ la $i^{ème}$ valeur ajustée lors de l'étalonnage et y_i la $i^{ème}$ valeur de référence. Dans une régression multiple à m variables, $p = m + 1$ (1 coefficient pour chacune des m variables, plus terme constant).

Erreur type du laboratoire : erreur en rapport avec la fidélité intermédiaire ou la reproductibilité, selon le cas.

Facteur : voir Composante.

Fouille de données (*data mining*) : processus d'exploration, d'extraction et de modélisation de grands volumes de données, pour en faire émerger des relations ou schémas non connus *a priori*.

Hotelling (test T^2) de : version multivariée de la statistique t de Student. Elle est en général utilisée pour tester si le vecteur moyen d'un ensemble de données multivariées prend une certaine valeur, ou pour comparer les moyennes des variables. Elle peut également servir à la détection de valeurs aberrantes. Un test statistique multivarié peut être réalisé au moyen de la statistique T^2 de Hotelling. Une ellipse de confiance peut alors être tracée dans le graphe des scores pour révéler les points aberrants, qui se situent en-dehors de l'ellipse.

Hyperparamètre : tout paramètre, défini par l'analyste, dont la valeur est utilisée pour configurer ou contrôler le processus de modélisation (par exemple, le nombre d'unités cachées dans un réseau neuromimétique, ou la sélection d'une régression PLS ou RCP).

Interférence : intervention de substances, phénomènes chimiques ou artefacts instrumentaux dans la mesure portant sur l'analyte visé. Il existe alors un risque de confusion entre analyte et facteur d'interférence si celui-ci ne fait pas l'objet de variations indépendantes ou du moins aléatoires par rapport à l'analyte.

Leave-one-out : procédure qui consiste à retirer un échantillon unique de l'ensemble des données de façon à générer un nouvel ensemble de données.

Leave-subset-out : procédure qui consiste à retirer un groupe d'échantillons de l'ensemble des données de façon à générer un nouvel ensemble de données.

Levier (effet de) : mesure du caractère extrême d'un point ou d'une variable par rapport à la majorité des autres points ou variables. Les points ou variables à effet de levier élevé sont susceptibles d'exercer une influence importante sur le modèle.

Loadings : estimation obtenue en concentrant l'information portée par plusieurs variables sur un petit nombre de composantes, représentées chacune par un axe dans un espace multidimensionnel. Chaque variable possède le long de chacun de ces axes un *loading* qui exprime dans quelle mesure elle est prise en compte par les composantes du modèle.

Modèle empirique : modèle qui s'appuie sur l'expérience et les données, sans présupposé d'une relation mathématique explicite ni recours à la théorie pour décrire le comportement d'un système.

Non supervisé : se dit d'un processus d'exploration des données ne reposant pas sur des hypothèses préalables.

Orthogonal : deux vecteurs sont orthogonaux si leur produit scalaire est nul.

Orthonormé : se dit de vecteurs orthogonaux et de norme 1 (unitaires).

Paramètre : tout résultat obtenu après le processus d'étalonnage du modèle (par exemple, coefficients et *loadings*).

Prédiction indirecte : processus d'estimation de la valeur d'une réponse sur la base d'un modèle multivarié et des données observées.

Propriété : voir Variable.

Rééchantillonnage : processus de réarrangement impartial et de sous-échantillonnage de l'ensemble initial des données. Il intervient dans le cadre des procédures de validation / optimisation opérant par calcul itératif d'une propriété et de l'erreur associée. La validation croisée et le bootstrap, qui génèrent des ensembles successifs de données d'évaluation par sous-échantillonnage répété, sont des exemples de méthodes de rééchantillonnage.

Resélection : réutilisation d'échantillons (voir Rééchantillonnage).

Résidu : mesure des variations non prises en compte par le modèle, ou écart entre les valeurs prédites et les valeurs de référence.

Scores : coordonnées des échantillons dans le nouveau système de coordonnées défini par les composantes principales. Les scores représentent la façon dont les échantillons sont en relation entre eux pour chaque variable de mesure.

***Score* réduit** : $j^{ème}$ valeur $t_{i,j}$ du score du $i^{ème}$ échantillon, divisée par la norme de la matrice des scores :

$$\left|t_{i,j}\right| = \frac{t_{i,j}}{\sqrt{\sum_{j=1}^{p} t_{i,j}^2}}$$

où *p* est le nombre de paramètres que comporte le modèle.

Sousajustement : problème inverse du surajustement.

Supervisé : se dit d'un processus de modélisation à partir de données étiquetées par classes ou par valeurs.

Surajustement : tendance, pour un modèle, à sur-décrire les variations au sein des données, parce qu'il prend en considération non seulement la structure pertinente, mais également des variations bruitées ou non porteuses d'information, et ainsi à conduire à des prédictions non fiables.

Variable (syn. : attribut, descripteur, propriété, caractéristique) : propriété d'un échantillon qui peut être évaluée.

Variable dépendante (syn : réponse, variable à expliquer) : variable (typiquement, données *Y*) qui est liée à une ou plusieurs autres variables par une relation mathématique formelle (explicite) ou empirique.

Variable indépendante (syn. : variable explicative, prédicteur, variable prédictive) : variable d'entrée (typiquement, données *X*) dont on peut déduire par une fonction mathématique la valeur d'autres variables dites dépendantes.

Variable latente : voir Composante.

Varimax (rotation) : rotation analytique orthogonale des facteurs ayant pour effet de maximiser la variance du carré des loadings qui leur sont associés, et donc d'amplifier les loadings et valeurs propres les plus élevés et d'atténuer les moins élevés pour chaque facteur.

5. ABRÉVIATIONS

ACI	Analyse en composantes indépendantes
ACP	Analyse en composantes principales
ALS	*Alternating least squares*, moindres carrés alternés
ANN	*Artificial neural network*, réseaux neuromimétiques
CP	Composante principale
DBSCAN	*Density-based spatial clustering of applications with noise*
DoE	*Design of experiments*, plan d'expériences
EFA	*Evolving factor analysis*, analyse factorielle évolutive
EM	Espérance-maximisation
EMSC	*Extended multiplicative scatter correction*, extension de la correction multiplicative de diffusion
ETE	Erreur type d'étalonnage
ETL	Erreur type du laboratoire
JADE	*Joint approximate diagonalisation of eigenmatrices*
LASSO	*Least absolute shrinkage and selection operator*
LDA	*Linear discriminant analysis*, analyse discriminante linéaire
LOOCV	Validation croisée *"leave-one-out" cross-validation*
MCR	*Multivariate curve resolution*, résolution multivariée de courbes
md	Distance de Mahalanobis
MLFF	*Multi-layer feed-forward*, réseaux neuronaux multicouches non rétroactifs
MSC	*Multiplicative scatter correction*, correction multiplicative de diffusion
MSPM	Maîtrise statistique des procédés multivariée
PAT	*Process analytical technology*, contrôle analytique des procédés
PLS	*Partial least squares régression*, régression des moindres carrés partiels
PLS-DA	*Partial least squares discriminant analysis*, analyse discriminante des moindres carrés partiels
QbD	*Quality by design*, qualité par la conception
QDA	*Quadratic discriminant analysis*, analyse discriminante quadratique
RCP	Régression sur composantes principales
ReLU	Unité linéaire rectifiée
RF	*Random forest*, forêts aléatoires
RLM	Régression linéaire multiple
RMN	Résonance magnétique nucléaire
RMSECV	*Root mean square error of cross-validation*, racine de l'erreur quadratique moyenne de validation croisée
RMSEP	*Root mean square error of prediction*, racine de l'erreur quadratique moyenne de prédiction
SIMCA	*Soft independent modelling of class analogy*
SMCR	*Self-modelling curve resolution*
SOM	*Self-organising map*, carte auto-organisatrice
SPIR	Spectroscopie dans le proche infrarouge
STING	*Statistical information grid*
SVM	*Support vector machines*, séparatrices à vaste marge, machines à vecteurs support