DIPLÔME NATIONAL DE DOCTORAT

(Arrêté du 25 mai 2016)

Date de la soutenance : **15 juin 2023**

Nom de famille et prénom de l'auteur : **Madame AL SOUKI Louna**

Titre de la thèse : *« Régression de données fonctionnelles avec prédiction et interprétabilité : inférence des propriétés en chimiométrie avec des moindres carrés partiels sparse (PLS) »*

Résumé

La chimie analytique joue un rôle crucial dans divers domaines car elle couvre l'identification, la quantification et la caractérisation des substances chimiques. Elle est essentielle pour comprendre la composition et le comportement de la matière et pour développer de nouveaux matériaux et technologies. Elle traite spécialement des mélanges complexes formés par un ensemble de molécules différentes, notamment les mélanges pétroliers, la source d'énergie la plus utilisée dans le monde depuis la révolution industrielle. Ils sont utilisés pour produire de l'essence, du diesel et d'autres combustibles pour les voitures, les camions, les avions et les bateaux. Leur caractérisation peut être établie à l'aide des propriétés physico-chimiques globales telles que la densité, la viscosité, le point d'éclair, le point d'écoulement, etc.. Ces méthodes sont normalisées et peuvent avoir un impact significatif sur les processus de raffinage, de transport et de stockage du pétrole. Elles peuvent également influencer la composition et la qualité des produits qui en dérivent. Cependant, les analyses de référence nécessitent des ressources importantes et sont coûteuses, limitant ainsi le nombre d'analyses pour le suivi des processus. De fait, il est donc nécessaire de disposer d'analyses rapides.

Les méthodes d'analyse rapide utilisent principalement la spectroscopie, qui exploite les propriétés physiques des produits. Cette approche présente plusieurs avantages, tels que la miniaturisation, les faibles coûts de fonctionnement et la rapidité. La spectroscopie comprend plusieurs types comme l'infrarouge, la résonance magnétique nucléaire, etc. Chacun diffère par la plage de longueurs d'onde utilisée, le type d'interaction impliquée ou le type de substance étudiée. Toutefois, ils ont tous en commun d'utiliser un profil de signal représenté par des données fonctionnelles. Cependant, leur précision est relativement inférieure et les résultats peuvent ne pas être aussi exhaustifs que ceux obtenus à partir de méthodes plus standardisées. Pour cela, les techniques de la chimiométrie peuvent résoudre ce problème en créant un modèle prédictif pour chaque propriété. De ce fait, cette thèse a deux objectifs principaux : le premier consiste à prédire les propriétés physico-chimique de nouveaux mélanges à partir de mélanges de référence, tandis

que le deuxième objectif est d'apporter des éclairages supplémentaires sur les parties du signal qui est relié le plus la propriété d'intérêt.

Les N spectres obtenus à partir de mesures physico-chimiques sont considérés comme des données fonctionnelles. Ces dernières décrivent des phénomènes ou des variables qui varient continuellement dans le temps ou dans l'espace. Les spectres peuvent être discrétisés en un nombre fini de point P et stockés dans une matrice $\mathbf{X} \in \mathbb{R}^{N \times P}$. Les propriétés macroscopiques à prédire sont des données traditionnelles qui sont des valeurs numériques regroupées dans un vecteur $\mathbf{y} \in \mathbb{R}^{N}$. Pour résoudre ce problème, on utilise généralement des techniques de calibration multivariée pour l'analyse prédictive, et la modélisation de régression pour établir une relation mathématique entre les données des spectres $\mathbf{X}$ et les propriétés macroscopiques y. Cette thèse se place dans un contexte de régression linéaire où la relation entre les deux parties sont représentées par un vecteur de coefficient de régression $\boldsymbol{\beta}$. Les spectres de mesures physico-chimiques ont généralement un nombre d'observations N dans $\mathbf{X}$ plus petit que le nombre de variables P représentant la longueur d'onde. Ainsi, nous traitons principalement des cas où $P \gg N$. La méthode OLS classique utilisée généralement en régression linéaire minimise l'erreur quadratique de prédiction. Cependant, avec les problèmes de grande dimension, l'estimation OLS n'est plus applicable. Il est important de réduire le nombre de variables afin de pouvoir les visualiser et les analyser plus facilement. Toutefois, les techniques de réduction de dimension peuvent entraîner une perte de précision. La solution consiste à accepter une certaine perte de précision en échange d'une simplification des données.

IFPEN a fourni des données réelles pour appuyer cette étude. Chaque donnée comprend des propriétés physico-chimiques standardisées pour une variété de coupes de pétrole ainsi qu'un ou plusieurs spectres d'analyse rapide associés. Dans le chapitre 2, on présente les données réelles utilisées dans ce manuscrit et rendues publiques dans un article pour encourager le développement d'autres études scientifiques. Elles sont composées de spectres infrarouges qu'on utilise pour prédire la densité des coupes considérées. En outre, nous avons proposé un générateur de données qui permet de reproduire des ensembles de données similaires aux données réelles. Les données simulées sont généralement importantes dans les projets de data science car elles permettent aux scientifiques de tester des hypothèses et de mener des expériences sans avoir à se limiter aux données réelles disponibles. Nous avons simulé des spectres fonctionnels représentés par une matrice $\mathbf{X}$ explicative en utilisant des mélanges de Gaussiennes et avons lié linéairement la réponse $\mathbf{y}$ à un nombre petit de variable de $\mathbf{X}$**.** De cette façon, nous avons pu évaluer la précision des prédictions en utilisant des modèles linéaires et évaluer l'interprétabilité lors de la création de régressions parcimonieuses.

Avant de commencer l'analyse et la modélisation des données, une procédure d'évaluation a été mise en place. L'un des objectifs principaux de la thèse est d'obtenir une précision de prédiction élevée. Dans un contexte de régression, la prédiction est souvent évaluée en séparant les données en ensembles de calibration et de validation. Le choix d'observation de calibration peut être fait aléatoirement sans grande connaissance préalable de la population. Dans cette procédure, chaque observation a une chance égale d'être retenue. Cependant, il n'est jamais garanti d'obtenir une calibration représentative de l'ensemble de l'échantillon. L'algorithme Kennard et Stone est moins aléatoire et largement utilisé en chimiométrie. Il sélectionne de manière séquentielle des observations de calibration uniformément espacées par rapport aux valeurs dans $\mathbf{X}$. Or en régression, la réponse $\mathbf{y}$ porte aussi des informations importantes. D'où la deuxième contribution de cette thèse : un algorithme de calibration-validation nommé CalValXy détaillé dans le chapitre 3. Il sélectionne les observations de calibration en utilisant à la fois les informations de $\mathbf{X}$ et de $\mathbf{y}$. Pour évaluer la performance de la calibration, nous avons utilisé des mesures telles que la distance euclidienne et la distance Φ_2 qui quantifie la similitude entre l'ensemble de calibration et les

données, ainsi que des métriques telles que l'erreur quadratique moyenne (RMSE), l'erreur absolue moyenne (MAE) et le coefficient de détermination (R^2) pour évaluer la précision de prédiction. Les résultats obtenus ont montré que la calibration construite avec cet algorithme couvre de manière uniforme l'espace expérimental et fournit des prédictions en régression précises par rapport aux méthodes de référence. Nous avons également cherché à fournir des informations sur le signal en fonction de la prédiction des propriétés, c'est-à-dire en considérant d'évaluer l'interprétabilité des résultats. Nous avons choisi de relever ce défi en détectant des informations à l'aide d'indicateurs de parcimonie. La parcimonie dans un modèle de régression fait référence à la présence d'un nombre relativement faible de coefficients non nuls dans le modèle. Les modèles de régression parcimonieux sont plus faciles à interpréter car ils ne comprennent que les variables les plus importantes dans le modèle, ce qui permet de localiser plus facilement les facteurs les plus pertinents qui permettent une bonne prédiction. Ils nécessitent également moins de calculs pour construire le modèle de régression. Cela peut être particulièrement bénéfique lorsqu'on traite des données de grande dimension. Nous proposons d'évaluer la parcimonie en calculant la mesure de comptage ℓ_0 des coefficients de régression $\boldsymbol{\beta}$ et en les comparant graphiquement aux spectres originaux pour évaluer la localisation.

Les techniques de réduction de dimension impliquent la transformation de données d'un espace de grande dimension vers un espace de faible dimension tout en préservant les informations clés des données d'origine. Deux catégories d'approches sont couramment utilisées : les méthodes projectives et les méthodes pénalisées. Les méthodes de projection sont basées sur la synthèse de la matrice de données $\mathbf{X}$ originale dans un espace de dimension inférieure, utilisant souvent des techniques telles que la régression PLS (moindres carrées partiels), couramment utilisées en chimiométrie. Le principe de la PLS consiste à résumer $\mathbf{X}$ en une matrice de score $\mathbf{T}$ en maximisant la covariance entre $\mathbf{T}$ et $\mathbf{y}$. La PLS a montré son efficacité grâce à la simplicité de son algorithme et la précision dans ses prédictions. Cependant, les résultats manquent d'interprétabilité. Les méthodes de pénalisation, quant à elles, consistent à régulariser les coefficients de régression pour une meilleur interprétabilité. Le lasso est une technique souvent utilisée. Sa régularisation de type ℓ_1 permet de produire des résultats parcimonieux visant une bonne interprétabilité. Néanmoins, le lasso possède des désavantages dans des situations de grande dimension comme sa saturation quand N variables sont sélectionnées. Pour bénéficier des avantages des méthodes de projection et de pénalisation, leur combinaison a été proposée dans la littérature nommée sparse PLS. Comme troisième contribution de cette thèse détaillée dans le chapitre 4, une nouvelle approche généralisée appelée Dual sparse PLS a été développée. Cette méthode est inspirée par les sPLSs et applique une réduction adaptative. Elle est basée sur une norme duale de la pénalité sélectionnée. Nous avons proposé quatre types de normes inspirés des techniques connues : lasso, groupe lasso, moindres carrés et ridge, qui présentent des performances de calibration et de validation quasiment équivalentes aux modèles de référence, avec moins de composantes formant $\mathbf{T}$. Un test de référence comparatif a été réalisé à l'aide des données simulés et de données réelles de spectroscopie proche infrarouge. Il a été constaté que les coefficients obtenus indiquent l'emplacement exact des zones de données influentes. Cela fournit une meilleure interprétation du modèle de prédiction formé.

Un package R appelé dual.spls a été développé pour mettre en œuvre la régression Dual-sPLS et la rendre accessible à la communauté scientifique pour résoudre des problèmes réels. En plus de la modélisation, ce package propose des fonctions supplémentaires pour faciliter son utilisation autonome, notamment des données réelles, un algorithme de simulation de données, une méthode de calibration et validation CalValXy et des outils d'évaluation. Le chapitre 5 propose un tutoriel d'utilisation de dual.spls avec des exemples graphiques et les lignes de codes.

Ce travail de thèse a donc permis de gérer un projet de data science dans le domaine de la

chimiométrie. Avant de proposer des solutions, nous nous sommes mis des objectifs selon la problématique de caractérisation du pétrole à l'aide de données fonctionnelles de grande dimension. Ensuite, dans un premier lieu, nous avons regrouper les données qui peuvent nous être utile et surtout, nous avons contribué à aider la communauté scientifique en publiant des spectres infrarouges réels avec leurs densités associées en tant que source ouverte pour d'autres travaux. Dans un deuxième lieu, nous avons proposé une méthode de calibration validation qui permet de crée des modèles à partir d'un sous ensemble représentatif des observations selon l'espace de $\mathbf{X}$ et $\mathbf{y}$. Enfin, nous avons conçu une méthode de régression qui présente de nombreux avantages :

- les prédictions correspondent ou dépassent les méthodes de référence ou comparables,
- dans les différentes propositions de normes que nous avons examinées, elles produisent en outre des représentations parcimonieuses des données simulées et réelles,
- elles offrent une localisation interprétable des caractéristiques du point de vue des données fonctionnelles,
- elles permettent enfin le regroupement des variables : la possibilité de rassembler les variables explicatives en sous-ensembles plus significatifs (échantillons contigus autour d'un pic, bandes spectrales disjointes associées à un composé) pour pouvoir combiner différentes modalités physico-chimiques.

Mots Clés : *Régression moindre carrées, lasso, ridge, régression, parcimonie, norme duale, chimiométrie, machine learning*