Méthodes statistiques et de deep learning pour la discrimination de spectres de masse MALDI-TOF

Marie Bourlioux[1], Paul-Marie Grollemund[1], Caroline Arous[2], Adrien Geneste[2] et Anna Grizon[2]

[1] *LMBP, Université Clermont Auvergne, France,*
marie.bourlioux@doctorant.uca.fr *paul_marie.grollemund@uca.fr*
[2] *UMRF, Université Clermont Auvergne, INRAE, France,*
caroline.arous@inrae.fr *adrien.geneste@inrae.fr* *anna.grizon@inrae.fr*

Résumé. Une approche utilisant les méthodes de *Functional Data Analysis* (FDA) est proposée pour la discrimination de spectres de masse, issus de l'appareil MALDI-TOF (*Biotyper Matrix Assisted Laser Desorption Ionization - Time of Flight*). L'analyse des spectres produits permet l'identification de micro-organismes, enjeu microbiologique majeur avec des applications médicales, pharmaceutiques ou agroalimentaires. Cependant, une limite importante subsiste : si l'identification est fiable au niveau de l'espèce, elle reste limitée au niveau de la souche. Les principales techniques d'analyse des spectres reposent sur leur discrétisation, ce qui induit une perte d'information. C'est pourquoi une analyse plus fine des spectres est menée, en modélisant chacun d'eux comme une réalisation d'un processus aléatoire fonctionnel, afin d'exploiter la structure continue intrinsèque des spectres. On apporte ainsi une évaluation de différentes méthodes, incluant modèles statistiques d'analyse fonctionnelle et réseaux de neurones conçus pour traiter des entrées fonctionnelles, en particulier concernant leur robustesse dans un cadre de faible effectif. Une exactitude moyenne de 90% est atteinte pour une classification de six souches de l'espèce *Streptococcus thermophilus* avec sept réplicats biologiques par souche. Ces performances suggèrent que la modélisation fonctionnelle des spectres permet d'exploiter efficacement l'information contenue dans les données et d'atteindre un compromis entre une production de données en faible quantité et de bonnes performances de discrimination des souches.

Mots-clés. données fonctionnelles, biostatistiques, spectrométrie de masse, analyse discriminante, réseaux de neurones fonctionnels

Abstract. An approach based on Functional Data Analysis (FDA) is proposed for discrimination of mass spectra acquired using Biotyper Matrix Assisted Laser Desorption Ionization–Time of Flight (MALDI-TOF). The analysis of these spectra enables the identification of microorganisms, which represents a major challenge in microbiology, with medical, pharmaceutical or agri-food applications. However, an important limitation remains : while current methods allow reliable identification at the species level, they fail to achieve satisfactory discrimination at the strain level. Most existing mass spectrometry analysis techniques rely on discretization of the spectra, leading to a loss of information. To address this issue, a finer analysis of the spectra is performed by modeling each spectrum as a realization of a functional random process, thereby exploiting the intrinsic continuous structure of the data. Several methods, as statistical functional models and neural networks designed to handle functional inputs, are evaluated, especially in terms of model robustness in a small sample setting. An average classification accuracy of 90% can be achieved for the discrimination of six strains of *Streptococcus thermophilus*, using seven biological replicates per strain. These performances suggest that functional modeling of mass spectra allows efficient exploitation of the continuous information

contained in the data and provides a favorable trade-off between limited data acquisition and high strain-level discrimination performance.

Keywords. functional data, biostatistics, mass spectrometry, discriminant analysis, functional neural networks

1 Introduction

D'après l'agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail, la consommation de fromage au lait cru serait impliquée dans respectivement 34%, 37% et 60% des épidémies de salmonellose, de listériose et d'infections à *Escherichia coli* entérohémorragiques entre 2012 et 2022 (Anses, 2022). Or, ces bactéries représentent un risque sanitaire majeur pour les consommateurs et en particulier les personnes à risque telles que les jeunes enfants ou les personnes âgées. En revanche, le fromage au lait cru contient des communautés microbiennes indigènes qui seraient à l'origine d'un développement de saveurs plus variées que pour le lait pasteurisé (Yoon et al., 2016). La connaissance de la composition microbienne de ces produits laitiers représente ainsi un enjeu majeur pour les filières fromagères. C'est l'une des raisons pour lesquelles sont menées des recherches sur les fromages traditionnels à microbiote complexe et notamment sur la maîtrise de la qualité nutritive, sensorielle et sanitaire de ces produits. Les études conduites cherchent à comprendre et valoriser le cheminement de la fabrication du fromage dans son ensemble : de la production du lait cru jusqu'au consommateur, en passant par la transformation fromagère. Cette dernière constitue une étape capitale durant laquelle l'ajout de ferments contribue à la fois au développement des arômes et à la sécurité sanitaire. Or, la production de ferments requiert notamment une identification précise des micro-organismes prélevés et manipulés au cours des expérimentations.

Au-delà de l'étude des fromages au lait cru, ces préoccupations s'inscrivent dans une problématique plus large d'analyse et de gestion de la diversité microbienne. En effet, cela met en évidence l'enjeu de déréplication des prélèvements microbiens, visant à s'assurer que deux micro-organismes isolés d'un même échantillon et identifiés comme appartenant à la même espèce correspondent bien à des entités génétiquement distinctes. Une telle démarche contribue à limiter les redondances analytiques, optimiser les collections d'isolats et affiner la caractérisation de la diversité au sein d'une espèce. Par ailleurs, elle présente également un intérêt en diagnostic moléculaire, notamment pour déterminer si une bactérie pathogène isolée chez deux patients correspond à une même souche, ce qui constitue une question centrale en épidémiologie et en investigation de foyers infectieux.

Dans ce contexte, des technologies, à l'image de la spectrométrie de masse, se développent dans les laboratoires de microbiologie en réponse à cette problématique. C'est notamment le cas de l'appareil *Matrix Assisted Laser Desorption Ionization - Time of Flight* (MALDI-TOF) qui permet d'obtenir des spectres de masse avec en abscisses le rapport masse/charge (m/z) des protéines et en ordonnées l'intensité relative des molécules ionisées. Chaque micro-organisme est associé à une empreinte spectrale spécifique, permettant l'identification. Toutefois, les performances obtenues avec cet instrument de spectrométrie restent limitées puisque l'appareil se restreint à l'identification de micro-organismes au niveau de l'espèce et ne permet pas une identification fiable au niveau de la souche.

La discrimination effectuée avec le MALDI-TOF utilise des profils spectraux caractérisés par des variations importantes. Effectivement, l'analyse des spectres repose généralement sur leur discrétisation, entraînant de fait une perte d'information. Ce type d'approche ne permet pas d'exploiter la continuité intrinsèque des spectres, leur forme globale ou encore la dépendance entre les points voisins sur le domaine. Afin de dépasser cette limite, on propose d'utiliser l'ensemble de l'information portée par les spectres de masses, en faisant appel aux outils de *Functional Data Analysis* (FDA) dont le développement a été grandement influencé par Ramsay et Silverman (2005). En effet, les spectres de masse peuvent être considérés comme des réalisations d'une variable aléatoire fonctionnelle X, à valeurs dans un espace de fonctions (espace de Hilbert séparable $\mathcal{H}$, muni du produit scalaire $\langle\,.\,,\,.\,\rangle$ et de la norme associée $\|\,.\,\|$).

2 Méthodes de discrimination de données fonctionnelles

2.1 Functional Partial Least Squares - Discriminant Analysis (FPLS-DA)

Afin de classifier les spectres et d'exploiter la structure fonctionnelle des données, on utilise la méthode *Functional Partial Least Squares - Discriminant Analysis* (FPLS-DA), notamment introduite par les travaux de Preda et Saporta (2005).

Chaque fonction est approximée par une décomposition dans une base orthonormée de fonctions $\{\phi_m\}_{m=1}^{M}$:

$$X_i^{[M]}(t) = \sum_{m=1}^{M} c_{im}\,\phi_m(t) \qquad \mathbf{c}_i = (c_{i1}, \ldots, c_{iM})^\top.$$

Le produit scalaire de l'espace de Hilbert $\mathcal{H}$ s'écrit alors $\langle X_i^{[M]}, X_j^{[M]} \rangle = \mathbf{c}_i^\top \mathbf{c}_j$. Dans ce cadre, une direction fonctionnelle $w \in \mathcal{H}$, décomposée dans la base des $\{\phi_m\}_{m=1}^{M}$, telle que

$$w(t) = \sum_{m=1}^{M} a_m\,\phi_m(t) \text{ avec le vecteur de coefficients } \mathbf{a} = (a_1, \ldots, a_M)^\top$$

induit un score fonctionnel $s_i = \langle X_i, w \rangle = \mathbf{c}_i^\top \mathbf{a}$. Par conséquent, la FPLS-DA se ramène à une PLS-DA multivariée sur la matrice de coefficients $C = (\mathbf{c}_1, \ldots, \mathbf{c}_n)^\top \in \mathbb{R}^{n \times M}$. En effet, le problème revient à maximiser la covariance entre les scores $\mathbf{s}$ et la variable réponse Y, c'est-à-dire trouver $\max_{\|\mathbf{a}\|=1} \mathrm{Cov}^2(C\mathbf{a}, Y)$. Cette approche met ainsi en évidence des composantes latentes qui concentrent l'information discriminante en tenant compte des corrélations entre variables explicatives, ce qui améliore la stabilité et l'interprétabilité par rapport à des méthodes purement basées sur la variance.

La décomposition dans une base orthonormée de fonctions fournit donc une approximation finie-dimensionnelle de la FPLS-DA. Ainsi, la FPLS-DA peut être interprétée comme une PLS de C vers la matrice indicatrice Y et constitue un compromis entre discrimination et réduction de dimension, particulièrement adapté au cas où $M \gg n$.

2.2 Functional Neural Networks (FNN)

En parallèle de l'approche statistique offerte par la FPLS-DA, les performances de réseaux de neurones fonctionnels, méthode proposée par Heinrichs et al. (2023), sont également évaluées. Ces réseaux traitent les données d'entrées comme une fonction afin de détecter des motifs caractéristiques et discriminants dans les spectres pour classifier les différentes souches de l'étude.

Convolution sur données fonctionnelles. Dans les couches convolutionnelles du réseau, le produit de convolution est utilisé :

$$H^{(l+1)}(t) = \sigma\Big((H^{(l)} * \kappa^{(l)})(t) + b^{(l)}\Big)$$

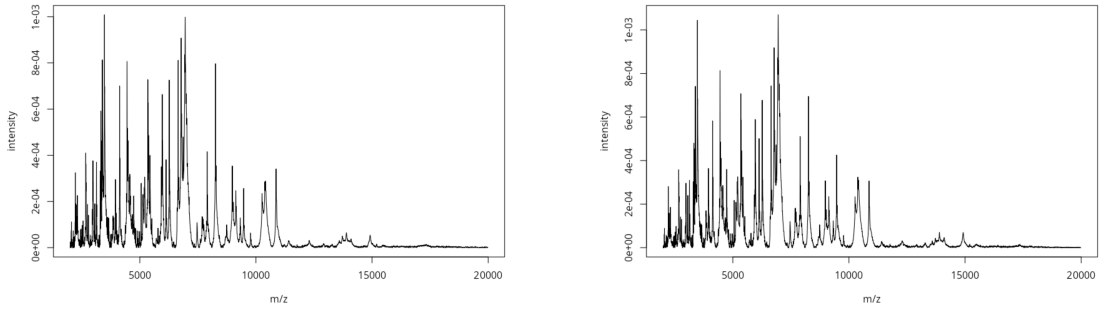
dans lequel $H^{(l)} : [0,T] \to \mathbb{R}^{d_l}$ est l'entrée de la couche l (avec $H^{(0)} = x$), $\kappa^{(l)} : [0,T] \to \mathbb{R}^{d_l \times d_{l+1}}$ est le noyau fonctionnel appris, $b^{(l)} \in \mathbb{R}^{d_{l+1}}$ est le vecteur de biais et σ la fonction d'activation. En pratique, la fonction κ est décomposée dans une base de fonctions pour avoir une analyse locale des spectres.

Architecture du réseau. Une des architectures utilisées pour classifier les spectres est constituée d'une première couche qui prend en entrée les spectres de masse $x(t) \in \mathbb{R}^p$, puis de deux couches de convolution fonctionnelle avec des dimensions respectives d_1, d_2 afin d'extraire les motifs des spectres d'entrée et enfin, d'une couche dense qui consiste en une projection linéaire $\mathbb{R}^{d_2} \to \mathbb{R}^K$, activée par softmax pour faire la classification en K classes, c'est-à-dire $\hat{y} = \text{softmax}(Wz + b)$ avec $W \in \mathbb{R}^{K \times d_2}$ et $b \in \mathbb{R}^K$.

Méthodes d'interprétation des résultats. Pour avoir une meilleure interprétation des résultats fournis par les réseaux de neurones, différentes méthodes liées au domaine de l'*eXplainable Artificial Intelligence* (XAI) sont ici utilisées, dont les trois suivantes : *occlusion sensitivity*, *Gradient Class Activation Map* et *activation maximisation* respectivement présentées dans les articles de Ancona et al. (2017), Selvaraju et al. (2016) et Schlegel et al. (2024).

3 Résultats de l'analyse des données microbiologiques

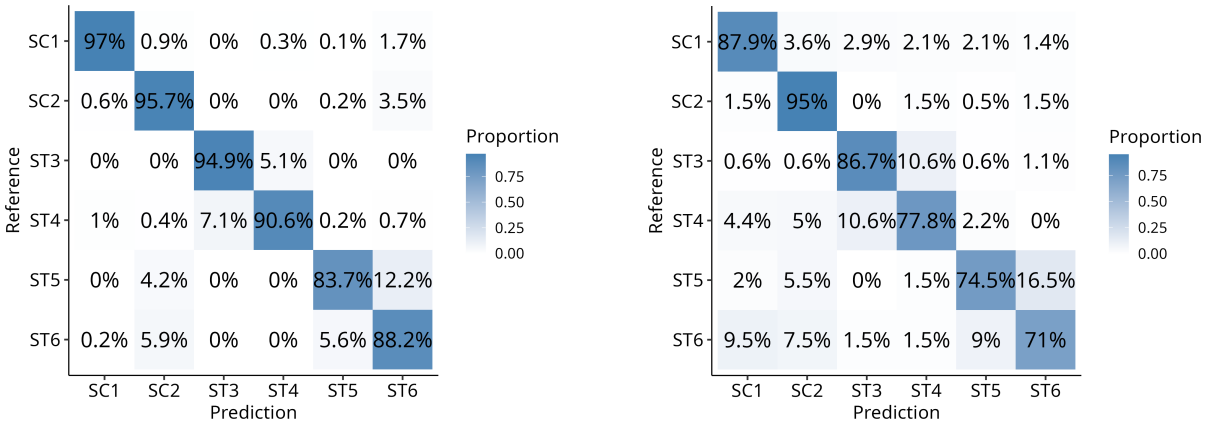
Pour cette étude, 20 réplicats biologiques de 6 souches de *Streptococcus thermophilus* (isolats provenant de fermes de la zone Saint-Nectaire, ferments commerciaux...) sont utilisés pour mesurer les performances des méthodes statistiques et de machine learning dans ce cas complexe, dans lequel la variabilité ne repose pas que sur des différences instrumentales ou une hétérogénéité de préparation d'échantillon. Ces réplicats biologiques sont chacun obtenus par moyennisation de triplicats techniques.



(a) Souche ST3 (b) Souche ST4

Figure 1 – Spectre de masse des souches de *Streptococcus thermophilus* ST3 et ST4

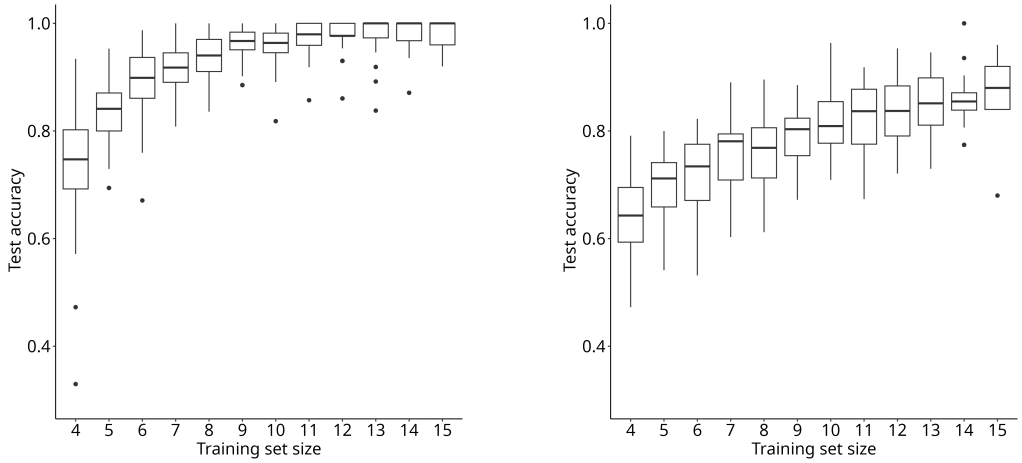
La figure 1 représente par exemple des réplicats biologiques de deux souches différentes, ST3 et ST4. Bien qu'il s'agisse de deux souches distinctes, leurs spectres partagent des variations similaires, rendant leur discrimination difficile.



(a) FPLS-DA (b) FNN

FIGURE 2 – Matrices de confusion normalisée par ligne, obtenue en moyennisant les résultats de classification sur les ensembles de test issus de différentes partitions train-test

Les performances de classification des deux méthodes mentionnées ci-dessus sont comparées, aboutissant à une meilleure classification (en termes d'exactitude, spécificité, rappel...) avec la FPLS-DA qu'avec les réseaux de neurones. Ce résultat peut s'expliquer par le fait que la FPLS-DA exploite la décomposition fonctionnelle permettant une réduction de dimension particulièrement bien adaptée aux petits échantillons, tandis que le FNN nécessite par nature une plus importante quantité de données. La figure 2 représente graphiquement cela avec des rappels par classe plus élevés pour la FPLS-DA, mais hétérogènes selon les souches à identifier.



(a) FPLS-DA (b) FNN

FIGURE 3 – Exactitude sur l'ensemble de test

Au-delà de l'obtention d'une méthode d'identification fiable, la détermination du nombre minimal de réplicats biologiques nécessaires pour atteindre un certain niveau de performance constitue un enjeu méthodologique majeur pour les biologistes utilisant le MALDI-TOF, notamment au regard du coût de production des données. Dans cette perspective, on étudie l'évolution de l'exactitude sur l'ensemble de test en fonction du nombre de réplicats biologiques par souche présents dans l'ensemble d'entraînement, afin d'identifier un compromis optimal entre

coût expérimental et pouvoir prédictif. La figure 3 montre les résultats du processus de bootstrap avec 100 répétitions et permet de voir que pour atteindre une exactitude de test de 90% en moyenne, cela nécessite au minimum 7 réplicats biologiques par souche avec la FPLS-DA.

Le réseau de neurones fonctionnels appliqué sur les réplicats biologiques donne une moins bonne exactitude moyenne que la FPLS-DA, à taille d'ensemble d'entraînement égale. Toutefois, les méthodes d'interprétation des réseaux de neurones permettent notamment de distinguer les parties du spectre importantes dans la prise de décision du réseau.

4 Conclusion

Des méthodes fonctionnelles de décomposition en une base de fonctions permettent d'exploiter la structure fonctionnelle des spectres de masse MALDI-TOF et d'aboutir après classification à de bonnes performances d'identification des réplicats biologiques des souches étudiées. Les méthodes de machine learning employées donnent de moins bons résultats de classification, ces méthodes nécessitant une plus importante quantité de données d'entraînement, ce qui est plus difficile à concilier avec l'objectif d'économie de données sous-jacent. Ce travail contribue ainsi à trouver une solution optimale et raisonnée entre la garantie d'obtenir de bonnes performances d'identification et la production de spectres en quantité soutenable et viable pour les utilisateurs de spectromètres de masse.

Notre présentation visera à exposer et discuter de résultats issus de nouvelles expérimentations microbiologiques, menées sur des données plus hétérogènes, pour évaluer la capacité de généralisation des méthodes utilisées. Nous aborderons également plus en détail les approches d'explicabilité appliquées aux réseaux de neurones, notamment en identifiant les régions du spectre qui contribuent à la classification et qui apportent une interprétation biologique cohérente. Enfin, au-delà de la mise en exergue des analyses statistiques et modélisations effectuées, l'interprétation des résultats obtenus sera replacée dans leur contexte biologique réel, permettant d'en discuter la pertinence et d'examiner leur impact sur la compréhension et l'identification des variations entre souches.

Bibliographie

Ancona, M., Ceolini, E., Öztireli, C., & Gross, M. (2017). Towards better understanding of gradient-based attribution methods for deep neural networks. *arXiv preprint arXiv :1711.06104*.

Anses. (2022, avril). Fromages au lait cru : quels risques pour la santé et comment mieux les prévenir ?

Heinrichs, F., Heim, M., & Weber, C. (2023). Functional Neural Networks : Shift invariant models for functional data with applications to EEG classification. *International Conference on Machine Learning*, 12866-12881.

Preda, C., & Saporta, G. (2005). PLS discriminant analysis for functional data. *ASMDA'05 XIth Int. Symp. on Applied Stochastic Models and Data Analysis*, 653-661.

Ramsay, J. O., & Silverman, B. W. (2005). *Functional data analysis* (2. ed.). Springer.

Schlegel, U., Keim, D. A., & Sutter, T. (2024). Finding the deepdream for time series : Activation maximization for univariate time series. *arXiv preprint arXiv :2408.10628*.

Selvaraju, R. R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., & Batra, D. (2016). Grad-CAM : Why did you say that ? *arXiv preprint arXiv :1611.07450*.

Yoon, Y., Lee, S., & Choi, K.-H. (2016). Microbial benefits and risks of raw milk cheese. *Food Control*, *63*, 201-215. https://doi.org/10.1016/j.foodcont.2015.11.013