Analyse de données fonctionnelles robuste dans des trames d'ondelettes ajustées

SUJET

Le problème scientifique est de **traiter, dans un formalisme homogène, des données de nature hétérogène, complémentaires, correspondant à un même phénomène physique, pour les simplifier (et spécialement de les classifier en groupes cohérents), les combiner, en extraire des modèles pertinents**. En effet, les méthodes statistiques exploitent traditionnellement des tableaux de données, organisées en lignes et colonnes. Or, beaucoup de problèmes présentent des structures de données plus complexes, comme les données fonctionnelles.

Les données fonctionnelles correspondent à ce que le langage courant désigne par l'appellation de « courbes ». Notre objectif global est de maîtriser et de contribuer au développement de nouveaux outils permettant d'étudier le comportement de ces données. Pourquoi cela ? D'abord, les données considérées figurent dans des espaces dont les dimensions sont souvent très élevées. Les techniques statistiques classiques sont essentiellement linéaires et ne rendent compte que d'un aspect limité des mesures. D'autres représentations sont à explorer. Ensuite, ces techniques perdent le caractère fonctionnel des mesures et des nouvelles techniques statistiques (sous le terme ombrelle d'analyse de données fonctionnelles) émergent pour en tenir compte. L'objectif appliqué est ici à la fois de gagner en performance et de mieux interpréter les données. Ce type de données abonde à IFPEN.

Pour illustrer notre propos, nous développons ci-dessous une problématique IFPEN où les données fonctionnelles jouent un rôle majeur.

RMN et distillation Les données RMN et de distillation sont récupérées de plus en plus souvent sur des échantillons de produits pétroliers lourds et se présentent sous forme de courbes. La RMN s'intéresse aux liaisons chimiques, la distillation est la mesure d'une qualité physique de l'échantillon. Les experts du domaine pensent que les informations recueillies par la distillation et par la spectrométrie RMN sont complémentaires. Les analyser conjointement serait un plus, à la fois pour l'extraction de descripteurs pertinents, et pour la conception de modèles prédictifs.

Quels sont les problèmes ? Avant toutes choses, les mesures chimiques sont sujettes à erreurs, de manipulation ou d'instrumentation. On fabrique ainsi des '*outliers*', qu'il est toujours très difficile de détecter vu la dimension du problème, mais qui faussent notablement les résultats ou les modèles obtenus.

Ensuite, les informations portées par la RMN et la distillation sont de natures extrêmement différentes. Les faire travailler mathématiquement ensemble demande une mise en forme non négligeable. Ainsi les données RMN, vu leur caractère 'piqué', peuvent être représentées par des ondelettes tandis que les distillations plus lisses feraient plus naturellement l'objet de modèles '*splines*'.

Les inégalités d'échantillonnage, la robustesse aux bruits, déformations et décalages sont aussi des enjeux importants pour ce type de données.

Positionnement de la proposition par rapport à l'état de l'art, et apport original

L'intérêt pour les techniques d'analyse de données fonctionnelles a réellement démarré avec [Ramsay-2006]. Deux articles récents traitent de l'état de l'art en: [Jacques-2014] et [Wang-2016]. Aujourd'hui, plusieurs packages en langage R y sont dédiés. Le site https://cran.r-project.org/web/views/FunctionalData.html donne une idée de la richesse du sujet. La maitrise de cette gamme récente d'outils représente en elle-même une quantité de travail considérable. En revanche, elle est susceptible de bénéficier à un large panel d'applications IFPEN, permettant d'accélérer, et de rendre plus automatique et robuste le traitement de données fonctionnelles, en croissance forte.

La combinaison de méthodes de réduction de dimension avec des décompositions sur des bases bien adaptées aux signaux analysés (Fourier, polynômes, *splines*, ondelettes [Giacofci-2013]) fait partie des tendances émergentes. Ces bases peuvent mettre en relief des caractéristiques localisées

(données RMN de pics), amplifier la diversité entre signaux présentant des différences ténues (comme entre les courbes de distillation), accroître la parcimonie [Müller-2008] ou la résistance aux perturbations aléatoires. Cependant, d'autres types de représentations plus redondantes (unions de bases, trames, dictionnaires) sont réputées (dans les applications de l'analyse harmonique) offrir de meilleures performances, notamment si l'on s'intéresse à des mesures « moralement » analogues, mais présentant, pour des raisons expérimentales ou phénoménologiques des décalages le long de la variable ordinale. La propriété recherchée est ici l'invariance (ou l'équivariance) par décalage, et éventuellement par échelle (pour des signaux acquis à des rythmes différents). Ce type d'invariance revêt un intérêt spécifique en classification [Wang-2014] des signaux de forme similaire, mais décalés ou dilatés.

Par ailleurs, d'un point de vue issu des statistiques robustes ou des méthodes de réduction de dimension parcimonieuses, les méthodes basées sur la variance ou l'énergie [Jolliffe-2016] manquent souvent de robustesse aux *outliers*, du fait de l'utilisation de moments d'ordre 2 (matrices de covariance). Il semble intéressant de considérer des techniques de régression robustes, soit par un choix de la fonction de coût adaptée, soit par pénalisation robuste (par exemple de type lasso) ou parcimonieuse. Les techniques d'ACP robustes, pourvues d'invariance par rotation et proposées dans [Ding-2006], semblent intéressantes, en regards des outils d'analyse fonctionnelle mentionnés plus haut.

L'apport original attendu se situe donc dans le développement de méthodes plus invariantes et plus robustes, combinant *in fine* l'analyse de données fonctionnelles avec des trames idoines et des techniques de régression robuste en grande dimension.

Ce sujet ne semble pas aujourd'hui être abordé frontalement par la littérature, peut-être du fait son positionnement au carrefour de disciplines. Il est à noter cependant que l'article [Han-2017] fait un pas dans ces directions, et cite justement les articles publiés antérieurement par IFPEN sur la conception de trames d'ondelettes (C. Chaux *et al.*, cf. bibliographie).

L'apport de la thèse se situera également dans la confrontation de telles méthodes récentes et innovantes à des jeux de données hétérogènes, notamment issus de la chimiométrie, pour lesquelles le besoin de modélisation est clairement identifié.

Bibliographie

[Chaux-2007] Caroline Chaux, Jean-Christophe Pesquet, Laurent Duval, 2007, Noise covariance properties in Dual-Tree Wavelet Decompositions,IEEE Transactions on Information Theory

[Chaux-2008] Caroline Chaux, Jean-Christophe Pesquet, Laurent Duval, 2008, A Nonlinear Stein Based Estimator for Multichannel Image Denoising, IEEE Transactions on Signal Processing

[Ding-2006] Chris Ding, Ding Zhou, Xiaofeng He, Hongyuan Zha, 2006, R1-PCA: Rotational Invariant L1-norm Principal Component Analysis for Robust Subspace Factorization, Proc. International Conference on Machine Learning

[Febrero-Bande 2012] Febrero-Bande, M., Oviedo de la Fuente, M. (2012). Statistical Computing in Functional Data Analysis: The R Package fda.usc. Journal of Statistical Software, 51(4), 1-28.

[Giacofci-2013] Joyce Madison Giacofci, 2013, Curve clustering and functional mixed models, PhD thesis

[Han-2017] Han X., Huang Z., Wang S., Chen X., Xu K;, Chen D., 2017, New insights to improve resolution and reliability of Raman spectral analysis using higher-density multiscale regression Chemometrics and Intelligent Laboratory Systems

[Jolliffe-2016] Jolliffe IT, Cadima J. 2016 Principal component analysis: a review and recent developments. *Phil. Trans. R. Soc.*

[Müller-2008] Hans-Georg Müller, 2008, Functional Modeling and Classification of Longitudinal Data, Chapman and Hall

[Ramsay-2006] J. O. Ramsay, B. W. Silverman, 2006, Functional Data Analysis

[Wang-2014] X. Wang, A. Qu, 2014, Efficient classification for longitudinal Data, Computational Statistics & Data Analysis

[Wang-2016] Jane-Ling Wang, Jeng-Min Chiou, Hans-Georg Müller, 2016, Review of functional data analysis, Annu. Rev. Stat. Appl.

Description du poste
PhD position at IFP Energies nouvelles (IFPEN)
Functional data analysis represents a recent corpus of methods aiming at the study of datasets of curves or more generally functions. Such methods are being increasingly used for the empirical analysis of experimental data. Examples at IFPEN are physico-chemical, transportation or geosciences data. Such methods notably allow to combine informations of distinct morphological nature, to classify them or to detect abnormal behaviors. Their use of projection kernels and dimension reduction methods takes advantage of curve regularity information. Among standard projection kernels are splines or orthogonal wavelets, that better emphasize subtle variations in the data.

This subject aims at extending the range of functional data analysis tools through the combination of two approaches: one from harmonic analysis, the other from robust statistics. For the first one, experimental data analysis being often noise-prone, exhibiting time shifts between phenomena of interest, it is judicious to resort to frame operators. Their redundancy unveils better information concentration (sparsity), noise robustness, shift invariance. For the second one, data being subject to abnormal behavior (outliers), robust regression methods (robust PCA, Lasso) for increased resistance.

At stake in this subject are both the theoretical aspects of contronting these two approaches (especially taming correlation induced by the usage of redundancy) and their implementation on ever increasing data volume, to develop a complete methodology for efficient high-throughput data analysis.

Poste de thèse à IFP Energies nouvelles (IFPEN)

L'analyse de données fonctionnelles forme un corpus récent de méthodes visant à étudier des ensembles de données se présentant sous la forme de courbes ou plus généralement de fonctions. Ces méthodes connaissent un intérêt croissant pour l'analyse empirique de données expérimentales, dont des exemples à IFPEN sont des données d'analyse physico-chimique, dans le domaine des transports et en géoscoences. Elles permettent notamment de combiner des informations de nature morphologiques distinctes, de les classifier, de détecter des comportements anormaux, par l'usage de noyaux de projections et de méthodes de réduction de dimension, permettant de mieux tirer profit des informations sur la régularité des courbes. Parmi les noyaux de projections courants, les splines ou des ondelettes orthogonales peuvent être employées, pour mieux mettre en évidence des variations subtiles dans les données.

Ce sujet vise à étendre la gamme des outils d'analyse de données fonctionnelles par la combinaison de deux approches, l'une issue de l'analyse harmonique, l'autre des statistiques robustes. Pour la première, l'analyse de données expérimentales étant souvent entachées de bruits, présentant des décalages temporels entre les phénomènes d'intérêt, il est judicieux de recourir à des opérateurs de trame, dont la redondance permet de meilleures propriétés de concentration de l'information (parcimonie), de robustesse aux bruits et d'invariance aux décalages. Pour la seconde, les données étant sujettes à des comportements aberrants (outliers), il est utile de mettre en oeuvre des méthodes de régression robustes (ACP robuste, Lasso) pour y résister.

L'enjeu de ce sujet réside à la fois dans l'aspect théorique de confrontation de ces deux approches (notamment la prise en compte des aspects de corrélation induits par l'usage de la redondance) et dans la mise en application sur des volumes de données toujours croissants, pour développer une méthodologie d'analyse haut-débit efficiente.

Spécialités / spécialisations et/ou connaissances ou compétences particulières recherchées
Statistiques, Traitement du signal, Optimisation, Mathématiques Appliquées
Mots clés
Analyse de données fonctionnelles, réduction de dimension, analyse harmonique, ondelettes, trames

Contacts
Francois.Wahl@univ-lyon1.fr
Marteau@math.univ-lyon1.fr
Laurent.Duval@ifpen.fr