Contributions à l'analyse de données non vectorielles

Nathalie Villa-Vialaneix

Université Toulouse 1 Capitole

École Doctorale Mathématiques, Informatique et Télécommunications de Toulouse

Contributions à l'Analyse de Données Non Vectorielles

Nathalie Vialaneix

Manuscrit en vue de l'obtention de
l'Habilitation à Diriger des Recherches

présenté et soutenu publiquement le 13 novembre 2014 devant :

Philippe Besse	Professeur, INSA, Toulouse
Dianne Cook	Professeur, Iowa State University
Marie-Laure Martin-Magniette	Directrice de Recherche, AgroParisTech
Jean-Michel Poggi	Professeur, Université Paris Descartes
Anne Ruiz-Gazen	Professeur, Toulouse School of Economics
Jean-Philippe Vert	Chercheur, Mines ParisTech & Institut Curie

sur la base des rapports rédigés par Mme Dianne Cook, M. Jean-Michel Poggi et M. Jean-Philippe Vert.

Remerciements

En premier lieu, je tiens à exprimer ma gratitude à Anne Ruiz-Gazen de m'avoir épaulée dans les diverses étapes cette habilitation, depuis la rédaction de ce manuscrit jusqu'à la soutenance. Son soutien, sa bienveillance, son dynamisme et son optimisme ont été une aide précieuse.

Je veux aussi remercier chaleureusement Dianne Cook, Jean-Michel Poggi et Jean-Philippe Vert d'avoir pris le temps d'évaluer ce document de synthèse. Je leur suis reconnaissante de m'avoir appuyée dans cette étape professionnelle qui me tenait à cœur. Je veux aussi remercier Philippe Besse, Marie-Laure Martin-Magniette et Josiane Mothe pour avoir accepté de faire partie du jury de soutenance car c'est toujours un plaisir d'échanger avec eux.

La recherche est évidemment un travail collectif et je n'aurais pu avancer dans cette voie sans l'aide des nombreuses personnes avec lesquelles j'ai collaboré depuis le début de ma thèse. Il est toujours délicat de citer nommément certains collaborateurs - et que tous les autres sachent que j'ai conscience de leur devoir beaucoup - mais je tenais à remercier, en particulier, Fabrice Rossi tant notre collaboration a été durable et fructueuse et s'est prolongée bien au-delà de relations professionnelles. Il a indéniablement beaucoup compté dans ma formation scientifique ; son amitié et son humour m'ont accompagnée dans les moments les plus difficiles.

Je dois également beaucoup aux équipes de recherche qui m'ont accueillie ces dernières années : l'équipe SAMM de l'université Paris 1 m'a offert un cadre scientifique épanouissant et a tout fait pour faciliter mon intégration malgré la distance. Merci donc à l'intégralité de l'équipe, et en particulier à son ancienne directrice, Marie Cottrell, son directeur actuel, Jean-Marc Bardet, et aux membres de l'axe dit « du mal » qui ont organisé pour moi des séances de travail en visio-conférences dans des conditions techniques osées : les efforts consentis m'ont beaucoup aidée à ne pas me sentir isolée. Je suis également très reconnaissante aux membres de l'unité MIA-T de l'INRA de Toulouse pour la qualité de leur accueil lors de ma délégation au sein du laboratoire durant l'année 2012/2013 et, tout particulièrement, à Christine Cierco-Ayrolles qui m'a encouragée à faire cette demande : cette année a été une année d'épanouissement scientifique pour moi. Le soutien du laboratoire ainsi que celui de mes collaboratrices de l'équipe GenPhySE, Magali San Cristobal et Laurence Liaubet, m'ont permis de finalement intégrer l'unité en février 2014 et de participer à l'encadrement des thèses de Jérôme Mariette et Valérie Sautron avec lesquels travailler est un plaisir. À l'INRA, j'ai trouvé un environnement scientifique stimulant et un environnement professionnel chaleureux, en particulier en occupant le bureau de Céline, avec laquelle je partage le goût des activités féminines délicates.

J'ai eu également la chance de pouvoir apporter ma petite pierre à l'animation de la SFdS, grâce à Jean-Michel Poggi, son président de l'époque, et cette expérience a été une source de rencontres stimulantes.

Plusieurs pages seraient probablement nécessaires pour remercier les personnes qui m'ont encouragée ces dernières années mais il faut savoir conclure et la conclusion de cette page de remerciements est naturellement tournée vers Jean, mon compagnon depuis plus de vingt ans, dont le soutien inconditionnel ne s'est jamais démenti. Il est indéniable que je lui dois bien plus que la conclusion de ce modeste travail.

Table des matières

Note technique *: Ce manuscrit a été écrit à l'aide du logiciel libre LaTeX à partir du modèle « The Legrand Orange Book » mis à la disposition de tous par Mathias Legrand. La bibliographie a été réalisée avec le programmes libres biblatex et biber. La plupart des graphiques ont été réalisés avec le logiciel libre R et en particulier avec les packages* **igraph**, **ggplot2** *et* **SOMbrero**. *La mise à disposition, de manière libre, de ces outils facilite quotidiennement notre vie scientifique et cette note technique a pour but de remercier collectivement les personnes qui participent à leur développement.*

Introduction

Dans de nombreux problèmes réels d'analyse de données, les observations collectées ne sont pas des données numériques et vectorielles classiques. Une première stratégie pour aborder ce type de questions est de simplifier celles-ci en les résumant par une représentation vectorielle puis d'utiliser des méthodes d'analyse statistique classiques (apprentissage supervisé, fouille non supervisée de données). Une alternative souvent préférée à cette approche simplificatrice est d'adapter les méthodes d'analyse à la structure particulière des données, que celles-ci soient des données représentées par des courbes (données fonctionnelles), des arbres ou des graphes (données relationnelles ou données hiérarchiques) ou bien d'autres types de données non vectorielles.

Ce mémoire résume mes activités de recherche dans cette dernière direction. De manière plus précise, je me suis intéressée, au cours de ma thèse, à l'analyse de données fonctionnelles, c'est-à-dire à l'analyse de données qui peuvent être décrites par des courbes et qui sont fréquemment modélisées sous la forme d'observations d'une variable aléatoire à valeur dans un espace de Hilbert. J'ai étudié l'adaptation de méthodes neuronales et à noyau à ce type de données. Tout en maintenant une activité dans ce domaine, je me suis peu à peu intéressée à d'autres types de données non vectorielles, à savoir des données relationnelles, modélisées sous la forme de graphes. J'ai investi mes compétences et connaissances sur les méthodes neuronales et les méthodes à noyau pour étudier ce type de données.

Mes activités de recherche actuelles se situent à l'interface entre statistique et informatique, sur les thématiques de la fouille de données et de l'apprentissage pour des données complexes et non vectorielles. Au delà des aspects de développement méthodologique, une part non négligeable de mes activités est consacrée à l'application de ces méthodes sur des problématiques concrètes issues de divers domaines d'application : sciences humaines, sociales et environnementales, suite aux collaborations nouées avec des chercheurs de l'Université Toulouse 2 (Jean Jaurès) durant ma thèse et, plus récemment, génomique et biologie des systèmes. Mon intégration récente à l'INRA (comme chargée de recherche, depuis février 2014) promet une intensification de ce dernier type d'applications. Les thématiques abordées dans ce manuscrit ainsi que les liens qui existent entre elles, sont schématisées dans la figure 2 que je reprendrai à plusieurs reprises dans ce mémoire.

L'ensemble des travaux présentés ici a donné lieu à 25 publications dans des revues nationales ou internationales à comité de lecture ainsi qu'à des publications dans des actes de conférences. La liste de mes publications est donnée dans le chapitre B en annexe. Les publications dans des revues sont découpées en 4 grands ensembles thématiques selon qu'elles sont des publications méthodologiques sur l'analyse de graphes, des publications méthodologiques sur l'analyse de données fonctionnelles, des applications en sciences humaines, sociales et sciences de l'environnement ou bien des applications en biologie des systèmes et génomique. La répartition des publications selon ces 4 grandes thématiques est donnée dans la figure 1 [1].

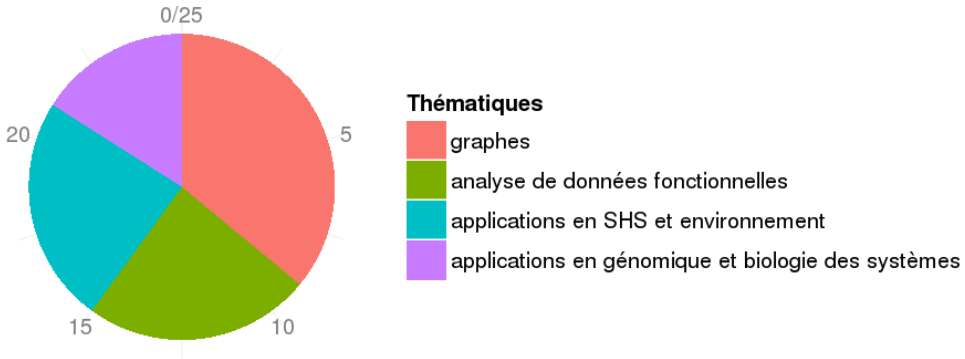

FIGURE 1 – Répartition des thématiques des publications dans des revues à comité de lecture.

De manière similaire, ce mémoire est organisé de manière thématique : dans le chapitre 1, je présente mes contributions à l'analyse de données relationnelles qui constitue ma thématique de recherche la plus active actuellement. Ce chapitre est découpé en deux grandes parties qui correspondent, respectivement, à des contributions pour la fouille de données relationnelles et pour l'inférence de réseau. Dans le chapitre 2, je présente mes contributions à l'analyse de données fonctionnelles ; pour simplifier le propos, j'ai résumé les résultats théoriques obtenus dans cette partie et n'ai inclus aucune démonstration. Les développements complets sont inclus dans les articles cités. À la fin de chacune des trois grandes parties de ce manuscrit (classification non supervisée & visualisation de graphes, inférence de réseau, analyse de données fonctionnelles), j'ai inclus une présentation des perspectives de mes travaux de recherche dans le domaine. La conclusion de ce manuscrit (page 77) fait la synthèse de mon projet de recherche. Une annexe contient

1. Bien sûr, cette répartition est relativement subjective car il est parfois assez difficile de différencier ce qui est de l'ordre du « méthodologique » de ce qui est de l'ordre de l'« application », l'un et l'autre étant étroitement mêlés dans plusieurs travaux.

un court CV qui synthétise l'évolution de ma carrière, mes activités d'encadrement et d'animation ainsi que mes participations à des contrats de recherche.

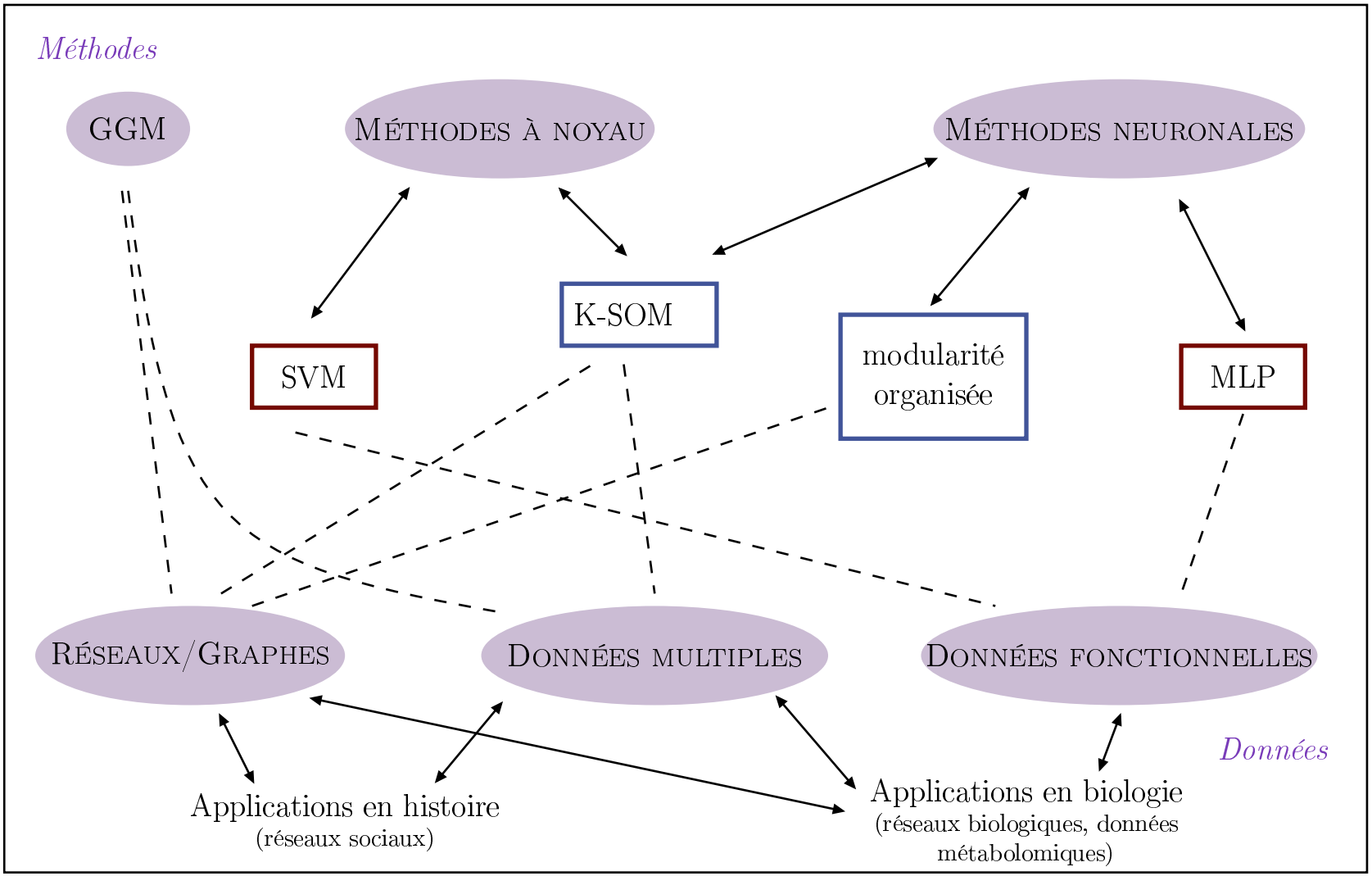

FIGURE 2 – Thématiques de recherche de ce manuscrit, organisées en méthodes (en haut) et données & applications (en bas), et leurs relations. Les approches supervisées sont entourées d'un rectangle rouge, les approches non supervisées d'un rectangle bleu. SVM : Support Vector Machine. K-SOM : Kernel Self-Organizing Maps. MLP : Multi-Layer perceptrons. GGM : Modèle Graphique Gaussien. Par « Données multiples », je fais référence à des données comportant plusieurs groupes de variables, éventuellement de types différents, ou à des données provenant de plusieurs groupes d'individus.

1 — Analyse et inférence de graphes

1.1 Introduction

Dans de nombreuses applications, les données ne sont pas décrites par des variables numériques ou qualitatives mais par leurs relations les unes aux autres. Ce type de données, dites *relationnelles* et communément appelées *réseaux*, est fréquemment modélisé par des graphes, c'est-à-dire par la donnée d'un ensemble de n sommets $V = \{x_1, \ldots, x_n\}$, modélisant des entités, et d'un ensemble d'arêtes E qui modélisent les relations entre ces entités. Cet exposé se restreint au cas de graphes non orientés, simples (sans boucle et arête multiple) et éventuellement pondérés. Dans ce dernier cas, les poids peuvent être représentés par une matrice W, de taille $n \times n$, symétrique, à diagonale nulle et à coefficients positifs. Ces données se retrouvent naturellement dans de nombreuses applications (M.E.J. Newman 2003; Dorogovtsev and Mendes 2003), les plus connues étant probablement les réseaux de l'internet (WWW : Wasserman and Faust 1994; Albert et al. 1999; Huberman and Adamic 1999; Scott 2000), les réseaux sociaux (Freeman 2004; Borgatti et al. 2009), comme les réseaux de collaborations (M.E.J. Newman 2001) ou les réseaux sociaux de l'internet (Wellman et al. 1996; Adamic and Glance 2005; Traud et al. 2011) et les réseaux biologiques (réseau d'interactions protéine-protéine, réseaux métaboliques, réseaux de régulation génique... voir Barabási et al. 2011).

Mes travaux de recherche ont trait à l'analyse statistique sur les réseaux et abordent les deux principaux aspects de celle-ci : en premier lieu, la fouille de données, destinée à extraire de l'information pertinente d'un réseau donné et plus récemment, l'inférence, qui consiste à reconstruire, à partir de données observées, le graphe de dépendance entre les variables. La section 1.2 se situe au cœur de la première thématique, en présentant des travaux relatifs à la classification non supervisée de sommets dans les graphes et l'utilisation de la classification pour la visualisation de graphes. Ces méthodes s'avèrent utiles pour guider l'utilisateur dans son exploration d'un grand réseau (certaines applications pouvant conduire à la manipulation de graphes de plusieurs centaines, plusieurs milliers, voire plusieurs dizaines de milliers de sommets) : la classification permet de découper le graphe en grands ensembles et d'aider à mettre en lumière sa structure globale, par l'analyse des relations existant entre ces grands ensembles. Les revues de références (Fortunato 2010; Schaeffer 2007) donnent un panorama complet des méthodes de classification de sommets d'un graphe. La section 1.3 présente la problématique de l'inférence

de réseau, spécifique au cadre biologique, où un graphe de dépendances entre variables est reconstruit à partir d'observations de ces variables. Mes travaux se restreignent au cas du modèle graphique gaussien (D. Edwards 1995) et abordent la question de l'intégration de données de natures ou d'échantillons multiples dans l'inférence. Dans les deux sections de ce chapitre, les problématiques et méthodes sont illustrées sur des cas d'études réels, en histoire (la section 1.2.4 présente l'études de données historiques, issues d'un grand corpus de documents du Moyen-Âge) ou en biologie.

1.2 Classification non supervisée & visualisation

1.2.1 Motivation et contribution personnelle

La notion de *communauté* dans les réseaux est une notion qui a été étudiée en premier dans le domaine des sciences sociales : en effet, il est généralement admis de manière assez naturelle (Freeman 2004) que les groupes humains sont structurés en sous-groupes sociaux cohésifs. Du point de vue de l'objet mathématique « graphe », la définition de ces communautés n'est pas complètement uniforme et peut varier selon le domaine d'application. Cependant, de manière assez consensuelle, la notion de communautés fait référence à des groupes de sommets denses (*ie* avec un grand nombre d'arêtes à l'intérieur du groupe) et connectés entre eux par un nombre faible (comparativement) d'arêtes. De nombreuses études sur des réseaux sociaux (M.E.J. Newman 2003; Porter, Onnela, et al. 2009; Traud et al. 2011) ont montré les relations entre ces groupes et des caractéristiques décrivant les individus, validant la pertinence des méthodes de recherche de communautés dans des cas réels. Certaines études font aussi état d'une structure modulaire hiérarchique complexe (Porter, Mucha, et al. 2007). L'exemple célèbre du club de karaté de Zachary (Zachary 1977) montra que la recherche de communautés dans un réseau social simple pouvait effectivement mettre en valeur des phénomènes sociaux important au sein du groupe de personnes étudié (et dans le cas de cette étude, anticiper ou expliquer la scission du club de karaté en deux groupes). Ces questions ont progressivement gagné de l'attention dans d'autres domaines d'application que celui des sciences sociales, notamment en biologie où certains travaux ont mis en valeur une relation entre communautés dans les graphes (plutôt appelés *modules* dans ce contexte d'application) et groupes fonctionnels (voir (Guimerà and Amaral 2005) pour un exemple d'application à un réseau métabolique).

Aussi, les propositions de méthodes de classification non supervisée des sommets d'un graphe, destinées à retrouver une partition de sommets en groupes densément connectés, ont connu un développement très important dans la littérature récente où elles sont souvent appelées « *méthodes de détection de communautés* ». Les revues (Schaeffer 2007; Porter, Onnela, et al. 2009; Fortunato 2010) proposent trois états de l'art des méthodes de classification non supervisées dans les graphes ainsi que des applications de ces méthodes sur des données issues de domaines d'application variés. Également, (Danon et al. 2005; Lancichinetti and Fortunato 2009) comparent les performances de différentes méthodes de classification non supervisée en terme de qualité de la classification obtenue et de complexité de l'algorithme. Parmi les méthodes les plus utilisées, on trouve l'optimisation d'un critère de qualité spécifique aux graphes, appelé *modularité* et introduit dans (M.E.J. Newman and Girvan 2004). Cette optimisation est un problème NP complet et de nombreuses méthodes d'approximation de la résolution de ce problème ont été proposées (M.E.J. Newman 2006; Reichardt and Bornholdt 2006; Blondel et al. 2008; Noack and Rotta 2009), pour n'en citer que quelques-unes. Parmi les approches couramment utilisées pour la classification non supervisée de sommets dans un graphe, on rencontre

aussi le « *spectral clustering* » (classification spectrale, (Ng et al. 2002; Luxburg 2007)), qui est basée sur la décomposition spectrale du *laplacien* du graphe, une matrice dont les propriétés algébriques sont fortement reliées à la structure du graphe.

Mes travaux en classification non supervisée se positionnent sur le développement de méthodologies combinant classification avec visualisation : l'objectif de la visualisation de graphes (Di Battista et al. 1999) est de fournir à l'utilisateur une représentation d'ensemble du graphe qui soit à la fois esthétique et une aide à l'interprétation. La plupart des algorithmes de représentation de graphes sont basés sur des modèles de forces (Fruchterman and Reingold 1991) et se concentrent sur un rendu esthétique qui favorise des arêtes courtes et de tailles uniformes. (Noack 2007) fait remarquer que ce type d'approches a pour conséquence de concentrer les sommets de forts degrés au centre de la figure et, de ce fait, ne correspond pas à la manière intuitive qu'un utilisateur a de comprendre les relations existant dans un grand réseau. En effet, l'analyste recherchera au contraire à extraire les grands ensembles et à avoir une vue macroscopique des relations existant entre eux, puis se focalisera sur les détails de tel ou tel ensemble d'intérêt. Cette démarche est proche de ce qui est fait en classification non supervisée de sommets et il est donc naturel de combiner les deux approches (classification et visualisation) comme outil d'exploration d'un graphe. Pour ce faire, plusieurs approches sont possibles :

1. effectuer une classification non supervisée des sommets dans un premier temps et représenter le *graphe des classes* dans un deuxième temps. Le graphe des classes est un graphe simplifié dans lequel chaque sommet représente une classe (Herman et al. 2000). Ces méthodes peuvent être utilisées en combinaison avec une classification hiérarchique des sommets pour permettre une exploration de plus en plus fine du graphe (Auber et al. 2003; Archambault et al. 2010; Seifi et al. 2010), qui est implémentée de manière interactive dans certains logiciels de visualisation de graphes (voir par exemple, Tulip [1] (Auber 2003) ou Gephi [2] (Bastian et al. 2009)) ;
2. effectuer une classification non supervisée des sommets dans un premier temps et représenter le graphe dans son ensemble, en utilisant la donnée des graphes comme contrainte sur la représentation, dans un second temps. Cette approche a particulièrement été étudiée dans le milieu des années 1990 sous le nom de *clustered graph visualization* (Bourqui et al. 2007; Eades and Feng 1996; Eades and Huang 2000) ;
3. effectuer classification et visualisation en même temps en introduisant dans la recherche de communautés des contraintes liées à la représentation du graphe des classes qui en résultera. (Noack 2007) propose également une approche alternative qui est proche de celle-ci en optimisant un modèle d'énergie conçu pour représenter à proximité les sommets de zones denses du graphe.

Mes contributions dans ce champ se situent principalement sur la troisième approche avec le développement d'une extension des cartes auto-organisatrices pour des données décrites par des noyaux (voir la section 1.2.2) . En particulier, cette approche est utile pour analyser des graphes mais elle peut être aussi utilisée pour l'analyse de données non vectorielles (ou vectorielles) dans un cadre assez général. Une approche similaire, mais spécifique aux graphes, est décrite dans la section 1.2.3 où une extension de la modularité est proposée pour représenter un graphe simplifié sur une grille. Cette section présente également une application de la classification basée sur le critère de modularité à la visualisation hiérarchique d'un grand graphe.

1. http://tulip.labri.fr
2. http://gephi.org

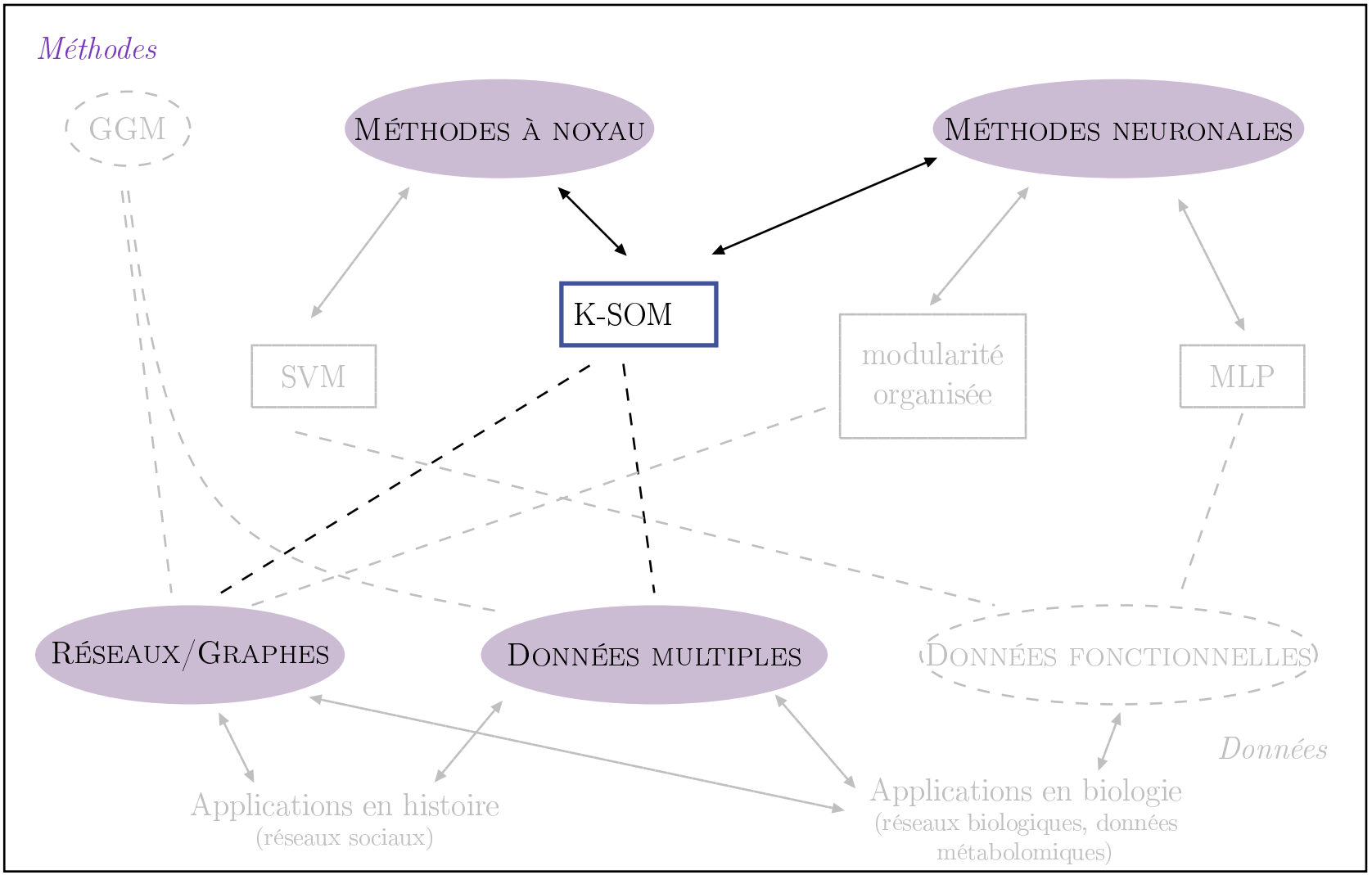

Figure 1.1 – Contributions présentées dans la section 1.2.2 « Approches à noyau ».

1.2.2 Approches à noyau

Cette première partie présente principalement les travaux des articles (Boulet, Jouve, et al. 2008; Massoni et al. 2013; Olteanu, Villa-Vialaneix, and Cierco-Ayrolles 2013; Olteanu and Villa-Vialaneix 2015; Mariette et al. 2014; Boelaert et al. 2014). Les thématiques abordées dans cette partie sont résumées dans la figure 1.1 qui est une simplification de la figure 2 dans laquelle les thématiques non abordées ont été grisées. Mes principaux collaborateurs sur ces sujets ont été, depuis 2007, Fabrice Rossi (professeur dans l'équipe SAMM, Université Paris 1) et, depuis 2012, Madalina Olteanu (maîtresse de conférences dans l'équipe SAMM, Université Paris 1). Actuellement, la thèse de Jérôme Mariette (Unité MIA-T, INRA de Toulouse), que je co-encadre, s'inscrit dans la poursuite du développement de cette thématique.

Définir une dissimilarité ou un noyau pour les graphes

Lorsque les objets d'étude ne sont pas des données numériques standard, comme dans le cas des graphes où les objets d'étude sont des entités (les sommets) décrites par leurs relations, il est commun de les décrire par une mesure de similarité ou de dissimilarité. Dans le cas des graphes, une dissimilarité classique est la longueur du plus court chemin dans le graphe, reliant deux sommets du graphe. Ces mesures de dissimilarité sont généralement symétriques et à valeurs positives mais peuvent ne pas être euclidiennes. Une autre approche consiste à utiliser un *noyau* qui est une mesure de similarité possédant quelques propriétés additionnelles qui en font son intérêt. Le noyau est une application $K : V \times V \to \mathbb{R}$ (V désigne l'ensemble des sommets du graphe ou, par extension, n'importe quel espace abstrait) tel que

$$\forall x,x' \in V, \qquad K(x,x') = K(x',x),$$

et

$$\forall N \in \mathbb{N} \text{ et } \forall (\alpha_i)_{i=1,\dots,N} \subset \mathbb{R} \text{ et } \forall (x_i)_{i=1,\dots,N} \subset V, \sum_{i,j=1}^{N} \alpha_i \alpha_j K(x_i,x_j) \geq 0.$$

L'intérêt des noyaux est qu'ils définissent, de manière implicite, un cadre euclidien pour l'espace V sur lequel il sont définis. En effet, (Aronszajn 1950) montre que pour tout noyau K, il existe un espace de Hilbert $(\mathcal{H},\langle .,. \rangle_{\mathcal{H}})$ et une application $\phi : V \to \mathcal{H}$ tels que le noyau correspond exactement au produit scalaire de $\mathcal{H}$ pour les données transformées par ϕ :

$$\forall x,x' \in V, \qquad K(x,x') = \langle \phi(x),\phi(x') \rangle_{\mathcal{H}}. \tag{1.1}$$

Cette propriété de *reproduction de l'espace de Hilbert* $\mathcal{H}$, a servi de justification théorique pour adapter beaucoup de méthodes d'analyse de données classiques au cadre des données décrites par des noyaux. En effet, utilisant la propriété (1.1), toute méthode d'analyse de données (classification supervisée ou non supervisée, régression) peut être adaptée au cadre non vectoriel de manière naturelle, à partir du moment où elle n'est basée que sur des calculs de normes et de produits scalaires : il suffit, en effet, de remplacer ceux-ci par leur équivalent dans l'*espace image* $\mathcal{H}$ en faisant référence à celui-ci de manière implicite, simplement au travers du noyau K. C'est notamment le principe sur lequel sont basées les machines à vecteurs de support (SVM (Vapnik 1995), voir section 2.4 pour mes travaux sur le sujet dans le cadre de l'analyse de données fonctionnelles). Ces approches, dites *méthodes à noyau* ont été utilisées avec succès dans de nombreux domaines d'application dont la biologie computationnelle (Schölkopf et al. 2004).

Pour les graphes, plusieurs noyaux ont été proposés dans la littérature, la plupart basés sur le *laplacien* du graphe qui est la matrice L, de dimension $n \times n$, telle que :

$$\forall i,j = 1,\dots,n, \qquad L_{ij} = \begin{cases} -W_{ij} & \text{si } i \neq j \\ d_i & \text{sinon} \end{cases},$$

où d_i est le degré du sommet x_i (*ie*, le nombre d'arêtes afférentes au sommet x_i ou $d_i = \sum_{j \neq i} W_{ij}$ dans le cadre d'un graphe pondéré). Cette matrice est fortement connectée à la structure du graphe : par exemple, (Luxburg 2007) montre que les vecteurs propres associés à la valeur propre 0 de la matrice permettent de retrouver les composantes connexes du graphe. (Heuvel and Pejic 2001; Boulet, Jouve, et al. 2008) montrent d'autres propriétés structurelles du graphe liés à la décomposition spectrale du laplacien. Dans un cadre très général, ces propriétés structurelles ont été utilisées pour justifier une approche de classification non supervisée basée sur le laplacien et appelée *classification spectrale* (« spectral clustering »).

Plusieurs noyaux ont été définis à partir de versions régularisées du laplacien d'un graphe, parmi lesquels :

- le *noyau de la chaleur* (R.I. Kondor and Lafferty 2002) : $K_\beta(x_i,x_j) = [K_\beta]_{ij}$ avec $K_\beta = e^{-\beta L}$ dont on peut démontrer qu'il correspond à un processus de diffusion de la chaleur le long des arêtes du graphe (le paramètre β définissant l'intensité de la diffusion). Ce noyau a été utilisé de nombreuses fois avec succès en biologie computationnelle (voir, par exemple, (Yamanishi, J.P. Vert, Nakaya, et al. 2003) pour une application à la classification non supervisée dans un réseau génomique, (Yamanishi, J.P. Vert, and Kanehisa 2005) pour une application à l'inférence de réseaux enzymatiques) ;
- le *noyau du temps moyen de parcours* (Fouss et al. 2006) : $K = L^+$ où L^+ est l'inverse généralisée du laplacien. Là aussi, ce noyau a une interprétation concrète simple : il permet de calculer le temps moyen nécessaire avec une marche aléatoire le long des arêtes pour relier deux sommets du graphe. (Pons and Latapy 2006) utilisent une idée similaire pour calculer une mesure de dissimilarité entre sommets d'un graphe de manière rapide.

Un cadre général pour ce type de noyaux, dérivés du laplacien, est décrit dans (Smola and R. Kondor 2003).

Parfois, les données ne sont pas décrites pas un noyau mais par une mesure de dissimilarité. De manière similaire au cadre du noyau, cette dissimilarité peut être plongée dans un espace euclidien si elle réalise la condition suivante (Schoenberg 1935; Young and Householder 1938; Krislock and Wolkowicz 2012) : la matrice d'éléments

$$s_{ij} = \left(\delta(x_i,x_n)^2 + \delta(x_j,x_n)^2 - \delta(x_i,x_j)^2\right)/2$$

est positive. Dans ce cas, s peut-être utilisé directement comme noyau, ainsi que la matrice d'éléments

$$\tilde{s}(i,j) = -\frac{1}{2}\left(\delta^2(x_i,x_j) - \frac{1}{n}\sum_{k=1}^{n}\delta^2(x_i,x_k) - \frac{1}{n}\sum_{k=1}^{n}\delta^2(x_k,x_j) + \frac{1}{n^2}\sum_{k,k'=1}^{n}\delta^2(x_k,x_{k'})\right)$$

comme suggéré dans (Lee and Verleysen 2007). Lorsque cela n'est pas le cas, (Y. Chen et al. 2009) propose de faire subir à la matrice de similarités utilisée, un pré-traitement consistant à supprimer du spectre les vecteurs propres associés aux valeurs propres négatives du spectre de la matrice ou bien à utiliser une reconstruction basée sur l'opposée des valeurs propres négatives. Les similarités obtenues ne sont alors plus identiques aux similarités de départ et une approche alternative s'appuie sur le concept

d'espace pseudo-euclidien décrit dans (Pękalska and Duin 2005) qui montrent que si $\Delta = (\delta(x_i,x_j))_{i,j=1,\dots,n}$ est une matrice de dissimilarité symétrique entre éléments x_i et x_j de V alors il existe deux espaces euclidiens $(\mathcal{E},\langle .,.\rangle_{\mathcal{E}})$ et $(\mathcal{F},\langle .,.\rangle_{\mathcal{F}})$ et une application de plongement $\psi : x \in \mathcal{G} \to (\psi|_{\mathcal{E}}(x),\psi|_{\mathcal{F}}(x)) \in \mathcal{E} \otimes \mathcal{F}$ tels que

$$\delta(x_i,x_j) = \|\psi|_{\mathcal{E}}(x_i) - \psi|_{\mathcal{E}}(x_j)\|^2_{\mathcal{E}} - \|\psi|_{\mathcal{F}}(x_i) - \psi|_{\mathcal{F}}(x_j)\|^2_{\mathcal{F}}. \tag{1.2}$$

De manière similaire à l'équation (1.1), l'équation précédente donne un cadre général pour étendre les méthodes d'analyse de données basées sur des calculs de normes et de produits scalaires aux données décrites par des mesures de dissimilarité.

Carte auto-organisatrice pour données décrites par un noyau ou une mesure de dissimilarité

L'algorithme de cartes auto-organisatrices (parfois appelées *cartes de Kohonen* ou *SOM*) a été proposé par T. Kohonen (Kohonen 1995). C'est une méthode d'analyse de données non supervisée qui allie classification non supervisée et projection des données sur un espace de faible dimension. De manière plus précise, les données sont projetées sur une *carte* qui est une grille, souvent régulière et rectangulaire, généralement de dimension 2 ou 1, composée de *neurones* ou *unités*. La grille est munie d'une topologie qui définie une « distance » entre unités. Les données sont alors classées dans les unités (qui constituent donc chacune une classe) de manière à ce que la topologie de celles-ci dans l'espace initial soit préservée : deux observations voisines dans l'espace des données sont classées dans la même unité (comme pour tout algorithme de classification non supervisée) ou dans des unités voisines sur la carte. Chaque unité est représentée dans l'espace d'origine par un *prototype* qui est un centre de gravité généralisé des observations de cette unité et des unités voisines (les observations sont prises en compte avec une pondération dépendant de la distance, sur la grille, avec l'unité dans laquelle elles sont classées). Dans le cadre numérique, l'algorithme alterne de manière itérative :

- **une étape d'affectation** qui consiste à affecter une ou des observations à l'unité dont le prototype est le plus proche ;
- **une étape de représentation** qui consiste à remettre à jour les prototypes à partir des modifications effectuées dans l'étape précédente.

L'apprentissage est généralement effectué de deux manières possibles (qui sont déclinées en de très nombreuses variantes) : en version « déterministe » (appelé aussi *batch* : dans ce cas, l'étape d'affectation concerne toutes les observations du jeu de données) ou en version « stochastique » (appelé aussi *on-line* : dans ce cas, à chaque itération, une seule observation, tirée au hasard, est traitée et l'étape de représentation correspond à une pseudo-descente de gradient stochastique ; des résultats théoriques de convergence, sur des cartes de dimension 1, sont données dans (Cottrell, Fort, and Pagès 1998) pour cette version de l'algorithme). (Fort et al. 2002) discutent les avantages et inconvénients des deux approches : la version déterministe de l'apprentissage est généralement plus rapide mais au détriment de la qualité de l'organisation des données sur la carte.

Lorsque les données ne sont pas vectorielles, la question de la définition des prototypes dans l'espace initial ne peut être réalisée de manière classique. Plusieurs extensions de l'algorithme de carte auto-organisatrice ont été proposées dans ce cadre. Une première approche utilise une méthode proche de l'analyse des correspondances multiples (AFCM) pour étendre les cartes auto-organisatrices à des données catégorielles (Cottrell and Letrémy 2005). D'autres approches, utilisables dans le cadre de l'analyse de graphe, nécessitent uniquement la connaissance d'une mesure de dissimilarité entre les données. Elles sont basées sur le *principe de la médiane* (Kohohen and Somervuo 1998) qui remplace le calcul traditionnel des prototypes par une optimisation effectuée

sur le jeu de données initial (un prototype correspond alors à une observation du jeu de données et la « distance » entre prototypes et observations découle alors directement de la connaissance de la mesure de dissimilarité entre paires d'observations). Un des principaux désavantages de cette approche est qu'elle est particulièrement restrictive et dépend fortement de la qualité de représentation des données traitées avec des effets de sous-optimisation importants sur l'étape de représentation. Pour augmenter la flexibilité de cette méthode, (Conan-Guez et al. 2006) proposent de représenter chaque unité par plusieurs prototypes, tous choisis parmi les données initiales mais cette approche peut considérablement augmenter les temps de calcul alors que les prototypes sont toujours contraints à être choisis parmi les données initiales.

Une alternative aux algorithmes basés sur le principe de la médiane se rapproche du cadre euclidien standard. Deux approches assez similaires ont été développées :

- lorsque les données sont décrites par un noyau K, l'algorithme de cartes auto-organisatrices à noyau a été proposé, pour sa version stochastique, dans (Mac Donald and Fyfe 2000; Andras 2002) et pour sa version déterministe dans (Villa and Rossi 2007; Boulet, Jouve, et al. 2008) ;
- lorsque les données sont décrites par une mesure de dissimilarité Δ, non nécessairement euclidienne, l'algorithme de cartes auto-organisatrices dit « relationnel » a été proposé, pour sa version stochastique dans (Olteanu, Villa-Vialaneix, and Cottrell 2012; Olteanu and Villa-Vialaneix 2015), et pour sa version déterministe dans (Hammer, Hasenfuss, et al. 2007; Rossi, Hasenfuss, et al. 2007; Hammer and Hasenfuss 2010).

Le récent article (Rossi 2014) fait une revue des différentes versions de l'algorithme de cartes auto-organisatrices pour données non vectorielles, établit les liens entre ces différentes versions et en discute les limites et les perspectives. Ici, nous nous restreindrons à la présentation des algorithmes proposés dans (Villa and Rossi 2007; Boulet, Jouve, et al. 2008; Olteanu and Villa-Vialaneix 2015) et montrerons ensuite comment ces approches peuvent être utilisées pour représenter de manière simplifiée des graphes et être ainsi une aide pour la compréhension de leur structure. Pour ce faire, nous définissons préalablement quelques notations relatives aux cartes auto-organisatrices. Dans la suite, la grille sera supposée être composée de U unités dont les prototypes seront notés $(p_u)_{u=1,\dots,U}$. La grille est également munie d'une relation topologique entre unités, classiquement appelée « distance » que nous noterons d (d est donc une application de $\{1,\dots,U\}\times\{1,\dots,U\} \to \mathbb{R}^+$). Une distance naturelle $d(u,u')$ sur une grille peut être la longueur du plus court chemin entre les unités u et u' sur la grille ou bien la distance euclidienne entre leurs positions sur la grille. Enfin, pour une observation x_i, $f(x_i)$ désignera l'unité (*ie*, la classe, pour reprendre le vocabulaire utilisée en classification non supervisée) dans laquelle x_i est affecté.

La **version déterministe de l'algorithme de carte auto-organisatrice à noyau** consiste à proposer une représentation des prototypes dans l'espace image $(\mathcal{H},\langle .,. \rangle_{\mathcal{H}})$. En effet, contrairement à l'espace initial dans lequel évoluent les données (les sommets du graphe, par exemple), l'espace image est un espace vectoriel standard muni des opérations usuelles. Les prototypes s'expriment alors comme des combinaisons convexes des images par ϕ des données initiales :

$$p_u = \sum_{i=1}^{n} \gamma_{ui} \phi(x_i) \qquad \text{où} \qquad \gamma_{ui} \geq 0 \text{ et } \sum_i \gamma_{ui} = 1.$$

La *phase d'affectation* d'une donnée x_i consiste donc à rechercher le prototype le plus proche, au sens de la distance dans l'espace image $\mathcal{H}$, en utilisant un calcul des distances

basé sur la seule connaissance du noyau K :

$$\begin{aligned}\|\phi(x_i) - p_u\|^2_{\mathcal{H}} &= \|\phi(x_i) - \sum_j \gamma_{uj}\phi(x_j)\|^2_{\mathcal{H}} \\ &= K(x_i,x_i) - 2\sum_j \gamma_{uj}K(x_i,x_j) + \sum_{jj'} \gamma_{uj}\gamma_{uj'}K(x_j,x_{j'}).\end{aligned}$$

La *phase de représentation des prototypes* consiste ensuite à remettre à jour tous les prototypes en calculant le centre de gravité généralisé des données :

$$\forall\, u = 1,\ldots,U, \qquad p_u = \arg\min_{p=\sum_i \gamma_i\phi(x_i)} \sum_{i=1}^{n} H(d(f(x_i),u))\|\phi(x_i) - p\|^2_{\mathcal{H}} \tag{1.3}$$

où H est une fonction de voisinage telle que $H : \mathbb{R}^+ \to \mathbb{R}^+$, $H(0) = 1$ et $\lim_{x\to+\infty} H(x) = 0$, qui généralement, décroît au cours de l'apprentissage. L'équation (1.3) a une solution très simple qui ne nécessite pas non plus par la connaissance de l'espace image ni de l'application de plongement ϕ :

$$\forall\, u = 1,\ldots,U \text{ et } \forall\, i = 1,\ldots,n, \qquad \gamma_{ui} = \frac{H(d(f(x_i),u))}{\sum_{j=1}^{n} H(d(f(x_j),u))}.$$

La méthode complète est décrite dans l'algorithme 1. (Villa and Rossi 2007) discutent

Algorithme 1 SOM à noyau, version déterministe

1: $\forall\, u = 1,\ldots,U$ et $\forall\, i = 1,\ldots,n$, initialiser γ^0_{ui} aléatoirement dans [0,1] tel que $\sum_{i=1}^{n} \gamma^0_{ui} = 1$ **Résultat :** $p^0_u = \sum_i \gamma^0_{ui}\phi(x_i)$
2: **Pour** $l = 1 \to L$ **Faire**
3: *affectation* $\forall\, i = 1,\ldots,n$, affecter x_i :

$$f^l(x_i) = \arg\min_u \|\phi(x_i) - p^{l-1}_u\|^2_{\mathcal{H}}$$

4: *représentation* $\forall\, u = 1,\ldots,U$, mettre à jour p_u :

$$p^l_u = \sum_i \gamma^l_{ui}\phi(x_i) \qquad \text{où} \qquad \gamma^l_{ui} = \frac{H^l(d(f^l(x_i),u))}{\sum_{j=1}^{n} H^l(d(f^l(x_j),u))}$$

5: **Fin Pour**
6: **Résultat :** $(p^L_u)_u$ et $(f^L(x_i))_i$

les relations entre cet algorithme et l'algorithme standard dans le cadre euclidien ainsi que ces relations avec l'algorithme basé sur le principe de la médiane.

Lorsque les données ne sont pas décrites par un noyau mais par une mesure de dissimilarité, non nécessairement euclidienne, (Hammer and Hasenfuss 2010) suggèrent d'utiliser un principe similaire et d'exprimer également les prototypes par une combinaison convexe de leurs images dans l'espace pseudo-euclidien sous-jacent :

$$p_u = \sum_i \gamma_{ui}\psi(x_i) \qquad \text{où} \qquad \gamma_{ui} \geq 0 \text{ et } \sum_i \gamma_{ui} = 1.$$

La *phase d'affectation* d'une donnée x_i, choisie au hasard, qui consiste à rechercher le prototype le plus proche au sens de la dissimilarité δ, se réduit donc à

$$f(x_i) = \arg\min_{u=1,\ldots,U} \Delta_i\gamma_u - \frac{1}{2}\gamma_u^T\Delta\gamma_u$$

où Δ_i est la $i$ème ligne de la matrice $\Delta = (\delta_{ij})_{i,j=1,\dots,n}$. En version stochastique, la *phase de représentation* consiste ensuite à mettre à jour les prototypes par une pseudo-descente de gradient :

$$p_u^{\text{new}} = p_u^{\text{old}} + \mu H\left(d\left(f(x_i),u\right)\right)\left(\psi(x_i) - p_u^{\text{old}}\right), \tag{1.4}$$

où μ est un paramètre qui en général décroît au cours du temps t (classiquement à la vitesse $1/t$). La calcul de l'équation (1.4) ne nécessite pas la connaissance de l'espace image et de la fonction de plongement ψ mais se réduit à une remise à jour des coefficients γ_u :

$$\gamma_u^{\text{new}} = \gamma_u^{\text{old}} + \mu H\left(d\left(f(x_i),u\right)\right)\left(\mathbf{1}_i - \gamma_u^{\text{old}}\right),$$

où $\gamma_u = (\gamma_{u1},\dots,\gamma_{un})^T$ et $\mathbf{1}_i$ est le vecteur de dimension n dont le seul coefficient non nul est le $i$ème. La méthode complète est décrite dans l'algorithme 2. De manière

Algorithme 2 SOM relationel, version stochastique

1: $\forall\, u = 1,\dots,U$ et $\forall\, i = 1,\dots,n$, initialiser aléatoirement γ_{ui}^0 dans [0,1] tel que $\sum_{i=1}^n \gamma_{ui}^0 = 1$ **Résultat :** $p_u^0 = \sum_i \gamma_{ui}^0 \psi(x_i)$
2: **Pour** $l = 1 \to L$ **Faire**
3: Choisir au hasard une observation x_i parmi $(x_j)_j$
4: *affectation* affecter x_i :

$$f^l(x_i) = \arg\min_{u=1,\dots,U}\left(\Delta_i \gamma_u^{l-1} - \frac{1}{2}(\gamma_u^{l-1})^T \Delta \gamma_u^{l-1}\right)$$

5: *représentation* $\forall\, u = 1,\dots,U$,

$$p_u^l = \sum_i \gamma_{ui}^l \psi(x_i) \qquad \text{où} \qquad \gamma_u^l = \gamma_u^{l-1} + \mu(l) H^l(d(f^l(x_i),u))\left(\mathbf{1}_i - \gamma_u^{l-1}\right)$$

6: **Fin Pour**
7: **Résultat :** $(p_u^L)_u$ et $(f^L(x_i))_i$

rigoureuse, la phase de représentation n'est pas une vraie phase de descente de gradient, car l'algorithme de carte auto-organisatrice ne possède pas de véritable fonction de coût. Toutefois, (Heskes 1999) prouve que, dans le cadre d'une taille de voisinage fixe, et avec une étape d'affectation modifiée, l'algorithme de carte auto-organisatrice minimise une énergie obtenue à partir de la formule de la médiane généralisée.

(Olteanu and Villa-Vialaneix 2015) soulignent que les complexités des deux versions (déterministe et stochastique) des algorithmes relationels et à noyau, sont comparables, de l'ordre de $\mathcal{O}(Un^2)$) mais que le nombre d'itérations nécessaires pour stabiliser l'algorithme déterministe est généralement inférieur à celui nécessaire pour stabiliser son équivalent stochastique. Toutefois, la meilleure organisation des données sur la carte compense ce désavantage. Formellement parlant, la convergence de l'algorithme de cartes auto-organisatrices n'a été prouvée que dans des cas très restreints (Cottrell and Fort 1987; Cottrell, Fort, and Pagès 1998) et qui ne sont pas généralisables au cadre pseudo-euclidien (lorsque la dissimilarité n'est pas euclidienne) comme souligné dans (Hammer, Gisbrecht, et al. 2011) pour l'algorithme Neural Gaz. Des preuves de la convergence de la version modifiée proposée par (Heskes 1999) existent toutefois mais là encore, ne sont pas extensibles au cadre pseudo-euclidien.

Mise en œuvre et exemple d'application en visualisation de graphes

Une partie des méthodes décrites dans la section précédente ont été implémentées et rendues publiques dans un package R[3] appelé **SOMbrero**[4]. L'implémentation du package a débuté dans le cadre du stage de Laura Bendhaïba, (Bendhaïba et al. 2013; Boelaert et al. 2014) ; **SOMbrero** propose une implémentation de la version stochastique de l'algorithme de carte auto-organisatrice, qui est prévue pour traiter trois types de données :

- des données numériques standard, multi-dimensionnelles ;
- des données décrites par une table de contingence qui sont traitées à l'aide de l'algorithme Korresp (Cottrell, Letrémy, and Roy 1993) ;
- l'algorithme relationel comme décrit dans (Olteanu and Villa-Vialaneix 2015).

Le package incorpore de nombreuses fonctionnalités, notamment :

- de nombreux graphiques pour analyser la carte obtenue (effectifs des classes, résumés des individus et des prototypes par classe, ajout de variables extérieures, représentation des distances entre prototypes) ;
- une fonctionnalité pour obtenir une classification non supervisée a posteriori des prototypes, appelée « super-classes » et pour représenter cette classification ;
- des critères de qualité (erreur de quantification, qui est le calcul de la variance intra-classe généralisée des observations, erreur topographique (Polzlbauer 2004) qui détermine la qualité de l'organisation de la carte en calculant la fréquence d'observations pour laquelle la seconde meilleure unité n'est pas dans le voisinage direct de l'unité à laquelle l'observation a été affectée).

Les deux premiers algorithmes (pour données numériques et tables de contingence) ont été implémentées en s'inspirant d'une partie des heuristiques des programmes originaux de Patrick Letremy (SAS/IML, voir http://samm.univ-paris1.fr/Programmes-SAS-de-cartes-auto).

L'implémentation a été pensée de manière à ce que l'utilisation soit simplifiée pour l'utilisateur, avec la possibilité d'appeler chacune de ces fonctionnalités en seulement une ligne de commande (et des valeurs par défaut choisies de manière pertinente). Des exemples reprenant des jeux de données standard ou originaux ont également été incorporés au package, sous forme de vignettes décrivant les commandes et analysant les résultats. En particulier, l'exemple fourni pour illustrer l'algorithme relationnel est basé sur l'étude d'un graphe et montre comment l'algorithme de carte auto-organisatrice peut être utilisé pour fournir à l'utilisateur une vision simplifiée du graphe et l'aider à en embrasser d'un coup d'œil sa structure macroscopique avant une analyse plus détaillée.

En guise d'exemple, un graphe simple est étudié qui est décrit dans (Knuth 1993). Les sommets de ce graphe sont les 77 personnages du roman « Les misérables » de Victor Hugo. Les 254 arêtes du graphe modélisent la co-apparition de deux personnages donnés dans le même chapitre du roman[5]. Le graphe de co-apparitions est représenté dans la figure 1.2. En calculant une matrice de dissimilarités qui correspond à la longueur du plus court chemin entre paires de sommets du graphe (non pondéré), l'algorithme de carte auto-organisatrice relationnel permet de traiter les données : chaque sommet du graphe est alors affecté à une unité d'une grille que nous avons choisie rectangulaire et de dimension 5×5. Une fois cette classification obtenue, il est possible d'en tirer une représentation simplifiée du graphe en représentant le *graphe des classes* comme suit :

3. R est un logiciel libre de programmation statistique ; voir http://www.r-project.org.

4. disponible sur R-Forge : http://sombrero.r-forge.r-project.org. Dernière version : 0.1-2-beta, Février 2014.

5. Le graphe est téléchargeable à http://people.sc.fsu.edu/~jburkardt/datasets/sgb/jean.dat.

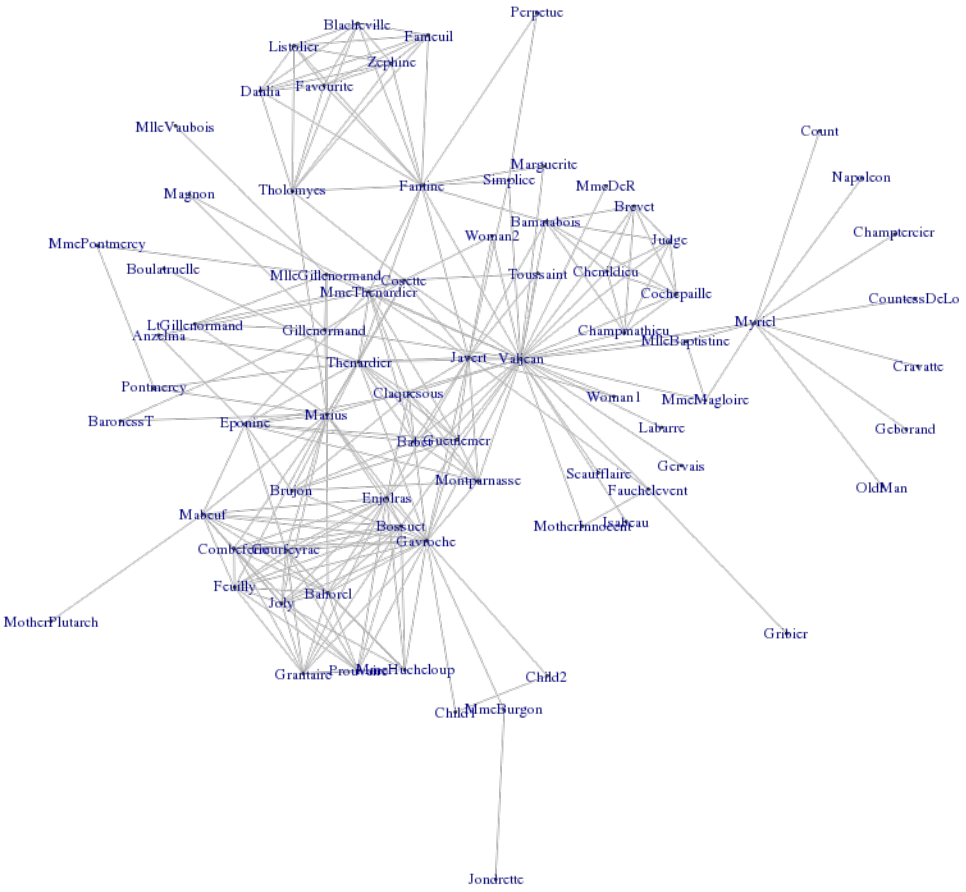

FIGURE 1.2 – Graphe de co-apparitions des personnages du roman « Les Misérables »

- chaque unité de la grille est représentée par un disque dont l'aire est proportionnelle au nombre de sommets classés dans cette unité ;
- les unités sont jointes par des arêtes dont l'épaisseur est proportionnelle au nombre total d'arêtes joignant deux sommets de chacune des deux classes.

Les résultats sont donnés dans les figures 1.3 (représentation simplifiée) et 1.4 (classification des 77 personnages sur la grille). Ils ont été obtenus à partir des commandes suivantes :

```
data(lesmis)
mis.som <- trainSOM(x.data = dissim.lesmis,
                    type = "relational",
                    proto.init="random")
plot(mis.som, what = "add", type = "graph", var = lesmis,
     print.title=TRUE)
plot(mis.som, what = "obs", type = "names", scale=c(1,0.5))
```

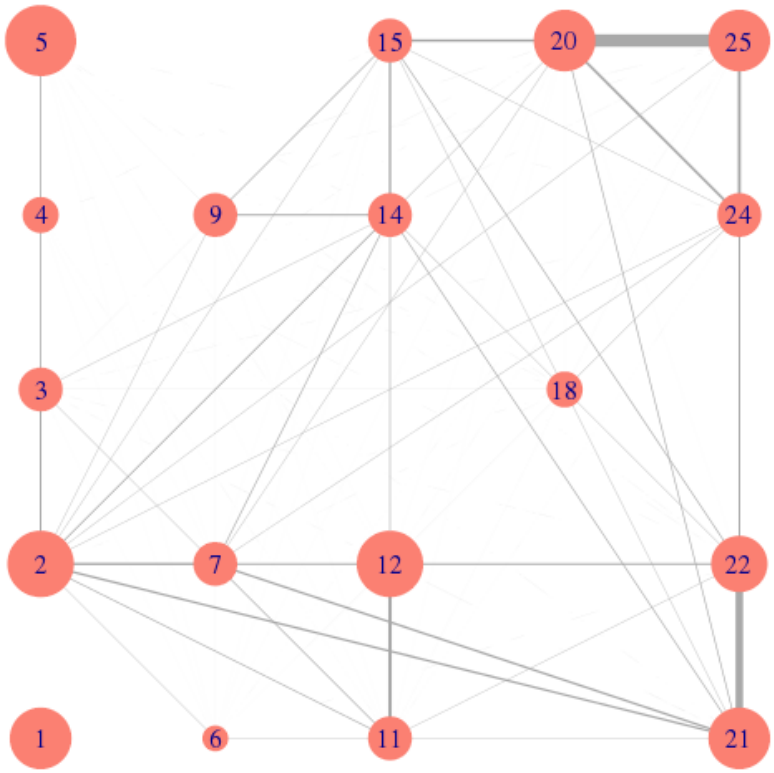

FIGURE 1.3 – Représentation simplifiée (graphe des classes) de la carte obtenue pour le graphe « Les Misérables » par l'algorithme de carte auto-organisatrice stochastique relationel tel qu'implémenté dans le package **SOMbrero**

On y retrouve les sous-histoires relatives au roman et plusieurs classes sont organisées autour d'un personnage principal. Les relations sur la carte permettent donc d'appréhender les liens entre les divers personnages. Si on numérote les classes de 1 à 25, de bas en haut puis de gauche à droite, en haut à gauche, la classe 5 est organisée autour de l'évèque monseigneur Myriel, qui constitue la première partie du roman et influencera le destin futur de Valjean. Valjean est situé dans la classe 2 (sur la gauche), avec des connexions vers toutes les autres parties de la carte. Parmi les personnages qui lui sont les plus proches se trouve Javert (classe 7, deuxième classe en bas et à gauche), le policier qui le poursuit, et Fantine (classe 11, en bas, au centre) à qui il vient en aide. Cosette, la pupille de Valjean, et Marius, son amoureux, sont dans les classes 14 et 15 (en haut au centre). L'approche de simplification de la représentation d'un graphe, illustrée ici sur un exemple jouet simple qui peut être compris directement par visualisation directe du graphe, prend tout son sens pour l'analyse de graphes plus complexes (car plus grand), comme discuté dans la section 1.2.4.

À noter que **SOMbrero** dispose aussi d'une interface graphique (interface web développée à l'aide du package **shiny**) accessible en ligne à `http://shiny.nathalievilla.org/sombrero` ou bien directement en local, en chargeant le package **SOMbrero** dans R et en exécutant la ligne de commande :

```
sombreroGUI()
```

Une copie d'écran de l'interface graphique est fournie dans la figure 1.5

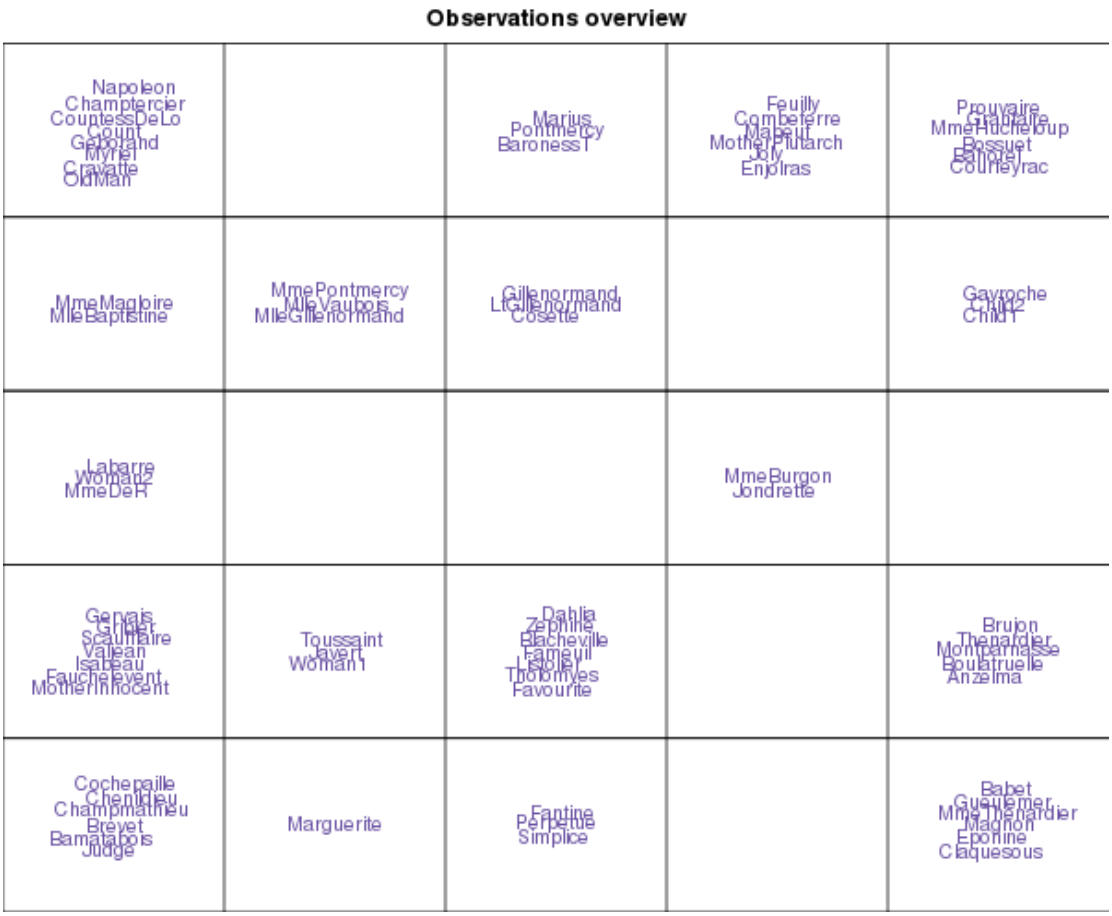

FIGURE 1.4 – Classification des divers personnages du graphe « Les Misérables » par l'algorithme de carte auto-organisatrice stochastique relationel tel qu'implémenté dans le package **SOMbrero**

Intégrer des informations extérieures

Les applications réelles fournissent des données de plus en plus complexes et notamment, pour le cas de l'analyse de réseaux, il n'est pas rare de disposer d'informations supplémentaires, sur les sommets ou les arêtes du graphe. Dans cette partie, nous supposerons connues un certain nombre de variables, appelées *étiquettes*, qui décrivent les sommets du graphe. Ces variables peuvent être éventuellement regroupées en groupes « thématiques ». De manière plus précise, on notera $(x_i^{(1)})_{i=1,\dots,n}$, …, $(x_i^{(D)})_{i=1,\dots,n}$, D groupes de variables décrivant les sommets $x_1,\dots,x_n$ du graphe, ces variables pouvant être de nature quelconque (ou bien elles mêmes des sommets d'un autre graphe).

En sciences sociales, croiser les informations additionnelles sur les sommets du graphe avec la classification est une pratique courante : cette opération est habituellement menée sous l'angle de l'*assortativité* : il s'agit de comprendre si des sommets d'un groupe donné partagent des caractéristiques communes après avoir effectué une classification non supervisée des sommets (voir (Traud et al. 2011) pour le calcul de la significativité d'un coefficient d'assortativité qui met en relation classes du réseau facebook© de plusieurs universités américaines) et divers types de caractéristiques décrivant les étudiants impliqués dans ce réseau ou bien (Laurent and Villa-Vialaneix 2011) pour l'utilisation d'indices issus de la statistique spatiale pour étudier la significativité du lien entre struc-

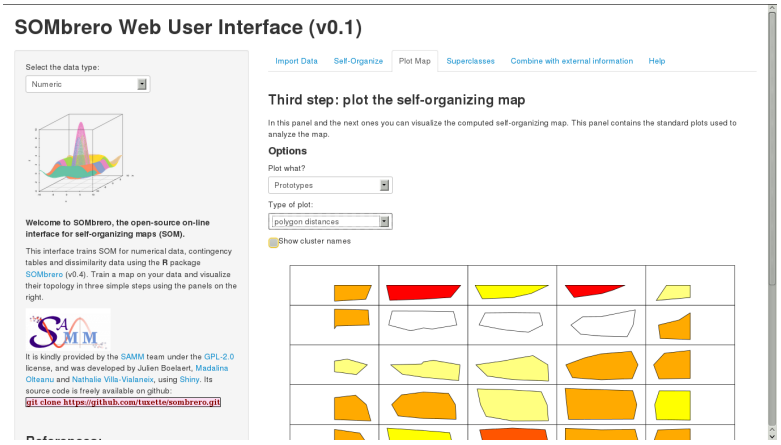

FIGURE 1.5 – Interface web du package **SOMbrero**.

ture d'un réseau et valeur des variables décrivant les sommets.

Dans (Olteanu, Villa-Vialaneix, and Cierco-Ayrolles 2013; Olteanu and Villa-Vialaneix 2015), nous abordons cette question sous l'angle de l'*intégration des informations supplémentaires* pour construire une carte auto-organisatrice. En classification non supervisée, cette question a déjà été abordée par d'autres auteurs de diverses manières : (Steinhaeuser and Chawla 2008) effectue une classification principalement basée sur les étiquettes des sommets qui est ensuite corrigée par un principe de seuillage basé sur les poids des arêtes entre sommets. (Ester et al. 2006; Moser et al. 2007; Ge et al. 2008) formalisent cette question sous la forme d'un problème d'optimisation basé sur des distances entre étiquettes proches de l'algorithme des k-moyennes. À l'inverse, d'autres auteurs favorisent la structure du graphe dans leur classification, comme (Cruz et al. 2011; H. Li et al. 2008). Enfin, d'autres auteurs cherchent, comme nous, à équilibrer les contributions des différents types de données : (Combe et al. 2012; Combe et al. 2013) combinent deux critères (un critère de modularité et un critère d'entropie) pour obtenir un critère global à optimiser tenant compte des différents objectifs. (Hanisch et al. 2002; Zhou et al. 2009) combinent diverses dissimilarités en une dissimilarité globale qui est utilisée pour la classification. Dans le cadre des cartes auto-organisatrices, diverses méthodologies ont également été proposées pour combiner des informations : (Lebbah et al. 2005) combinent informations numériques et binaires en se basant sur deux énergies de quantification qui sont optimisées en parallèle. (Ghassany et al. 2012) introduisent un critère de collaboration, après la phase d'apprentissage des différentes cartes qui correspondent chacune à un groupe de variables.

Nous abordons cette question de manière différente en supposant connu un noyau pour chaque groupe d'étiquettes, $K^{(d)}$ $(d = 1,\ldots,D)$, qui décrit la similarité $K^{(d)}\left(x_i^{(d)},x_{i'}^{(d)}\right)$ entre les étiquettes du groupe d des sommets x_i et $x_{i'}$ du graphe ou bien une dissimilarité qui décrit la dissimilarité entre ces mêmes étiquettes. Pour des questions de clarté du propos, nous nous restreignons dans cet exposé au cas où un noyau est connu mais l'approche est généralisable au cadre de dissimilarités comme décrit dans (Olteanu and Villa-Vialaneix 2015). L'idée principale consiste à combiner les diverses informations par le biais de la définition d'un noyau unique qui est la combinaison convexe des divers noyaux :

$$\forall\, i = 1,\ldots,n,\ \widetilde{K}\left(\tilde{x}_i,\tilde{x}_{i'}\right) = \sum_{d=0}^{D} \alpha_d K^{(d)}\left(x_i^{(d)},x_{i'}^{(d)}\right), \qquad \alpha_d \geq 0 \text{ et } \sum_d \alpha_d = 1, \tag{1.5}$$

où $x_i^{(0)} := x_i$, $K^{(0)} := K$ est un noyau sur les sommets du graphe initial $\mathcal{G}$, comme décrit dans les sections précédentes et $\tilde{x}_i = (x_i, x_i^{(1)}, \ldots, x_i^{(D)})$. (Yamanishi, J. Vert, et al. 2004; Yamanishi, J.P. Vert, and Kanehisa 2005) ont utilisé une approche similaire pour de l'inférence de réseaux (classification supervisée) qui intègre de l'information provenant de plusieurs sources de données recueillies à divers niveaux de l'échelle du vivant. Le choix des poids relatifs à chacun des noyaux y est basé sur une mesure de performance de la classification supervisée. De manière similaire, (Lanckriet et al. 2004; Rakotomamonjy et al. 2008) proposent de résoudre directement un problème d'optimisation dans lequel les poids $(\alpha_d)_d$ sont optimisés simultanément avec la résolution du problème d'optimisation classique de SVM supervisé. Dans le cadre non supervisé, une approche similaire est proposée par (Zhao et al. 2009) qui optimisent la combinaison linéaire sur un critère de qualité de la classification (voir aussi (Gönen and Alpaydin 2011) pour une revue des diverses approches permettant de combiner plusieurs noyaux).

Algorithme 3 Carte auto-organisatrice multi-noyaux

1: $\forall u = 1, \ldots, U$ et $\forall i = 1, \ldots, n$, initialiser aléatoirement γ_{ui}^0 dans [0,1] tel que $\sum_{i=1}^n \gamma_{ui}^0 = 1$
2: $\forall d = 0, \ldots, D$, initialialiser $\alpha_d^0 = \frac{1}{D+1}$ **Résultat :** $p_u^{\alpha,0} = \sum_{i=1}^n \gamma_{ui}^0 \sum_{d=0}^D \alpha_d^0 \phi^{(d)}(x_i^{(d)})$
3: **Pour** $l = 1 \to L$ **Faire**
4: Choisir au hasard une observation $\tilde{x}_i$ parmi $(\tilde{x}_j)_j$
5: *affectation* affecter $\tilde{x}_i$

$$f^l(\tilde{x}_i) \leftarrow \arg\min_{u=1,\ldots,U} \left\| \phi^{\alpha^{l-1}}(\tilde{x}_i) - p_u^{\alpha,l-1} \right\|_{\mathcal{H}^{\alpha^{l-1}}}$$

6: *représentation* $\forall u = 1, \ldots, U$,

$$\gamma_u^l \leftarrow \gamma_u^{l-1} + \mu(l) H^l \left(d(f^t(\tilde{x}_i), u) \right) \left(\mathbf{1}_i - \gamma_u^{l-1} \right)$$

7: *optimisation des poids*

$$\forall d = 0, \ldots, D,\ \alpha_d^l \leftarrow \alpha_d^{l-1} + \nu(t) \mathcal{D}_d^l$$

Résultat : $p_u^{\alpha,l} = \sum_{i=1}^n \gamma_{ui}^l \sum_{d=0}^D \alpha_d^l \phi^{(d)}(x_i^{(d)})$
8: **Fin Pour**
9: **Résultat :** α^L, $(p_u^{\alpha,L})_u$ et $(f^L(\tilde{x}_i))_i$

De manière similaire à (Rakotomamonjy et al. 2008), nous proposons d'optimiser la combinaison convexe des noyaux en intégrant une étape de pseudo-descente de gradient stochastique à l'algorithme. Cette idée est aussi similaire à celle de (Villmann et al. 2012) pour optimiser le paramètre d'un noyau dans les algorithmes LVQ. De manière plus précise, on détermine la dérivée, par rapport aux $(\alpha_d)_d$ de la fonction de coût

$$\mathcal{E}((\gamma_{ui})_{ui}, (\alpha_d)_d) = \sum_{u=1}^{U} \sum_{i=1}^{n} H(d(f(\tilde{x}_i), u)) \|\tilde{\phi}^{\alpha}(\tilde{x}_i) - p_u^{\alpha}\|^2_{\tilde{\mathcal{H}}^{\alpha}}$$

où $\left(\tilde{\mathcal{H}}^{\alpha}, \langle .,. \rangle_{\tilde{\mathcal{H}}^{\alpha}} \right)$ désigne l'espace de Hilbert associé au noyau défini dans l'équation (1.5),

$\tilde{\phi}^{\alpha}$ désigne la fonction de plongement sous jacente et

$$p_u^{\alpha} = \sum_{i=1}^{n} \gamma_{ui} \sum_{d=0}^{D} \alpha_d \phi^{(d)}(x_i^{(d)}) = \sum_{i=1}^{n} \gamma_{ui} \phi^{\alpha}(\tilde{x}_i)$$

avec $\phi^{(d)}$ la fonction de plongement associé au noyau $K^{(d)}$. Dans la version stochastique de l'algorithme de carte auto-organisatrice à noyau, à classification $(f(\tilde{x}_i))_i$ fixée, la contribution de l'observation choisie x_i à cette dérivée est : $\forall\, d = 0, \ldots, D$,

$$\mathcal{D}_d := \frac{\partial \mathcal{E}|_{x_i}}{\partial \alpha_d} = \sum_{u=1}^{U} H\left(d(f(\tilde{x}_i), u)\right) \left(K^{(d)}(x_i^{(d)}, x_i^{(d)}) - 2 \sum_{j=1}^{n} \gamma_{uj} K^{(d)}(x_i^{(d)}, x_j^{(d)}) + \sum_{j,j'=1}^{n} \gamma_{uj} \gamma_{uj'} K^{(d)}(x_j^{(d)}, x_{j'}^{(d)}) \right).$$

Utilisant cette dérivée, une étape de pseudo-descente de gradient est intégrée dans l'algorithme pour l'optimisation en ligne des poids $(\alpha_d)_d$ comme décrit dans l'algorihme 3.

Pour assurer que l'étape d'optimisation des poids respecte la contrainte de convexité des $(\alpha_d)_d$, une stratégie similaire à celle décrite dans (Luenberger 1984; Bonnans 2006; Rakotomamonjy et al. 2008) est utilisée : le gradient $(\mathcal{D}_d^l)_d$ est réduit et projeté de cette manière :

$$\widetilde{\mathcal{D}}_d = \begin{cases} 0 & \text{if } \alpha_d = 0 \text{ et } \mathcal{D}_d - \mathcal{D}_{d_0} > 0 \\ -\mathcal{D}_d + \mathcal{D}_{d_0} & \text{if } \alpha_d > 0 \text{ et } d \neq d_0 \\ \sum_{d \neq d_0,\ \alpha_d > 0} (\mathcal{D}_d - \mathcal{D}_{d_0}) & \text{sinon} \end{cases}$$

D'un point de vue pratique, on fait décroître le pas $\nu(t)$ à la vitesse habituelle ν_0/t avec une valeur initiale ν_0 suffisamment petite pour assurer la positivité des $(\alpha_d)_d$.

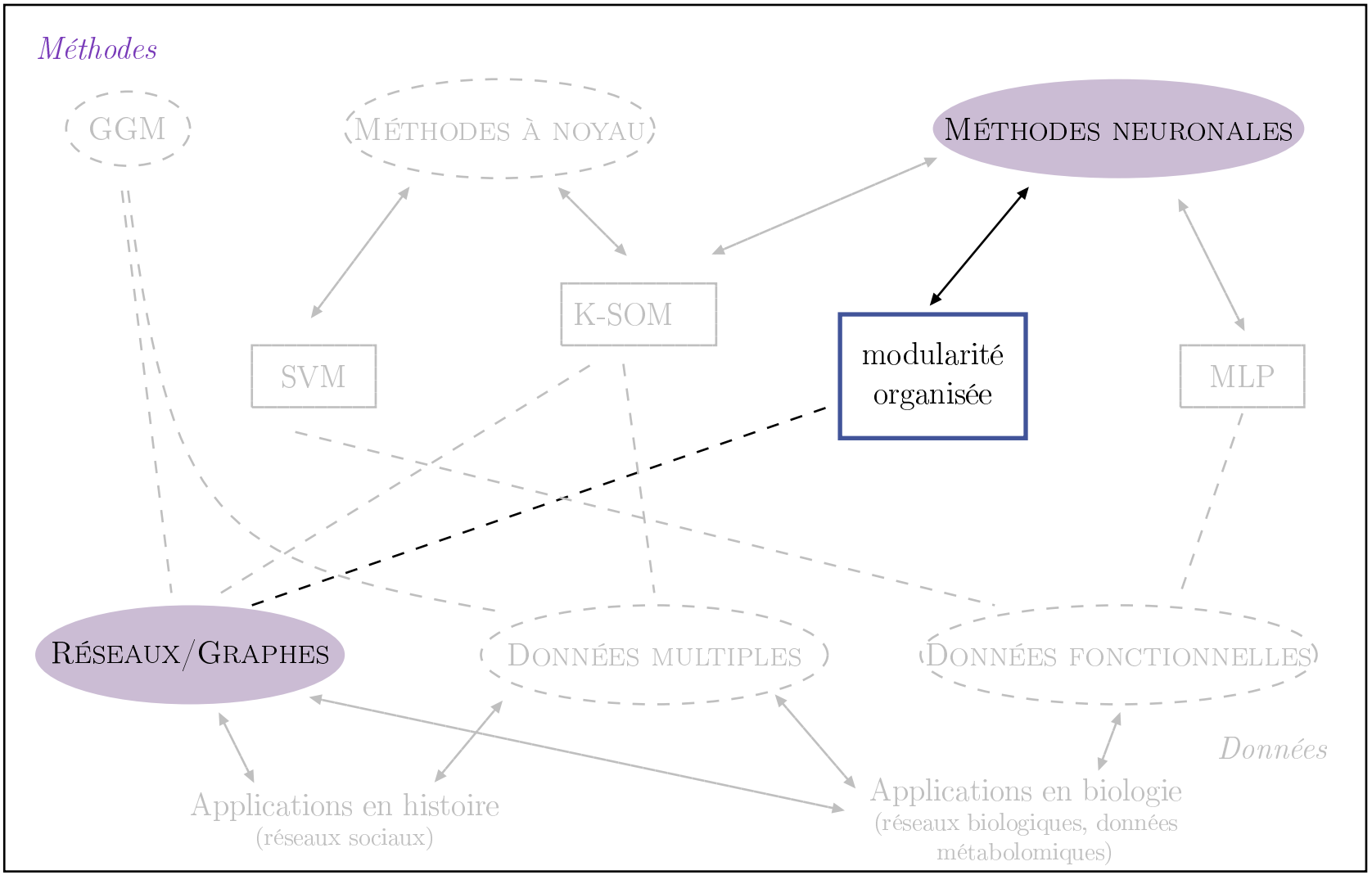

Figure 1.6 – Contributions présentées dans la section 1.2.3 « Approches basées sur la modularité »

1.2.3 Approches basées sur la modularité

Cette seconde partie présente principalement les travaux des articles (Rossi and Villa 2009; Rossi and Villa-Vialaneix 2010; Rossi and Villa-Vialaneix 2011b). Les thématiques abordées dans cette partie sont résumées dans la figure 1.6 qui est une simplification de la figure 2 dans laquelle les thématiques non abordées ont été grisées. Mon principal collaborateur sur ce sujet est Fabrice Rossi (actuellement professeur dans l'équipe SAMM, Université Paris 1).

La modularité comme critère de classification de sommets d'un graphe

Les travaux présentés dans la section précédente sont basés sur l'utilisation d'une approche générique pour des données non vectorielles décrites par un noyau ou une mesure de dissimilarité. Elles construisent une classification, organisée sur une carte, qui est basée sur le plongement du graphe dans un espace euclidien ou pseudo-euclidien. Dans la section actuelle, nous utilisons un autre type d'approches, basées sur un critère de qualité propre aux graphes, la *modularité* (M.E.J. Newman and Girvan 2004). Nous développons des méthodes qui permettent la visualisation du graphe en nous appuyant sur une classification obtenue par optimisation de la modularité ou d'un critère dérivé de celle-ci. Rappelons que, pour une partition donnée des sommets du graphe, $\mathcal{C}_1$, ..., $\mathcal{C}_C$, la modularité a pour expression

$$\mathcal{Q}(\mathcal{C}_1,\ldots,\mathcal{C}_C) = \frac{1}{2m}\sum_{k=1,\ldots,C}\sum_{x_i,\,x_j\in\mathcal{C}_k}\left(W_{ij} - \frac{d_i d_j}{2m}\right) \tag{1.6}$$

où m est le nombre d'arêtes (ou la somme des poids des arêtes $1/2\sum_{i,j=1}^{n} W_{ij}$ dans le graphe et les autres notations sont celles introduites précédemment (W_{ij} est le poids de l'arête entre les sommets x_i et x_j et d_i est le degré du sommet x_i, $d_i = \sum_{j\neq i} W_{ij}$). L'idée de ce critère de qualité d'une classification est qu'il mesure la pertinence de classer ensemble deux sommets du graphe en comparant le poids de l'arête qui les joint (ce poids étant égal à 0 si aucune arête ne relie les sommets considérés) à un modèle nul dans lequel les poids des arêtes ne dépendent que du degré des sommets considérés et non de la partition des sommets. Dans le modèle nul, les poids théoriques des arêtes, $P_{ij} = \frac{d_i d_j}{2m}$ sont proportionnels aux degrés des sommets afférents à l'arête et sont normalisés de telle manière que la somme des poids $(W_{ij})_{ij}$ est égale à la somme des poids théoriques $(P_{ij})_{ij}$. Ainsi, si le poids de l'arête (x_i,x_j), W_{ij}, est beaucoup plus grand que le poids théorique du modèle nul, P_{ij}, cette arête est considérée comme particulièrement « importante » et la partition $\mathcal{C}_1$, ..., $\mathcal{C}_C$ maximisant le critère $\mathcal{Q}$ aura tendance à classer x_i et x_j dans la même classe. Le fait de ne pas minimiser directement le nombre d'arêtes entre les sommets de classes différentes mais de tenir compte des degrés des sommets des graphes permet de mieux séparer les sommets de fort degrés (une arête afférente à un tel sommet ayant une importance moindre dans le critère de qualité) que pour des approches similaires au critère de coupe optimale comme la classification spectrale (Luxburg 2007). L'idée est de dire que les arêtes des sommets « les plus populaires » n'ont pas une signification aussi forte que les arêtes de sommets de plus faible degré.

Dans (Fortunato and Barthélémy 2007), les auteurs montrent que l'optimisation de la modularité peut induire des problèmes de résolution (certaines petites communautés significatives peuvent ne pas être détectées par optimisation de la modularité). Toutefois, malgré ce problème, la modularité reste une des mesures les plus utilisées pour l'obtention de communautés et elle a montré sa pertinence pour mettre en valeur la structure d'un réseau. Dans (Villa-Vialaneix, Liaubet, Laurent, Cherel, et al. 2013), lors d'un travail débuté dans le cadre du stage de Adrien Gamot, nous montrons notamment que les

groupes de gènes obtenus par optimisation de la modularité ont une cohérence forte en terme de groupe fonctionnel (c'est-à-dire de groupes de gènes partageant une fonction biologique commune). La maximisation de $\mathcal{Q}$ est un problème NP-complet et nécessite donc un algorithme de résolution heuristique. Pour ce faire, de nombreuses approches ont été proposées : l'approche initiale, décrite dans (M. Newman 2004), s'appuie sur une démarche de classification hiérarchique simple, elle est rapide mais en pratique conduit à des solutions sous efficaces (en terme de modularité de la classification trouvée). Une approche plus performante, mais aussi plus coûteuse en temps de calcul, est d'utiliser une optimisation par recuit simulé (Guimerà, Sales-Pardo, et al. 2004; Villa-Vialaneix, Liaubet, Laurent, Cherel, et al. 2013) ou par recuit déterministe (Lehmann and Hansen 2007). Utilisant une matrice de modularité, (M.E.J. Newman 2006) a proposé une méthode approchée basée sur une approche spectrale. Toutefois, le meilleur compromis entre temps de calcul (qui permet de traiter de très gros réseaux) et qualité de l'optimisation semble avoir été atteint par les algorithmes gloutons à raffinement hiérarchique décrits dans (Noack and Rotta 2009).

Dans la suite, je présenterai tout d'abord une approche basée sur une carte auto-organisatrice qui s'appuie sur une adaptation du critère de modularité présenté plus haut. L'optimisation du nouveau critère est effectuée par une approche par recuit déterministe. Dans un second travail, je présenterai comment, par une approche en deux temps, il est possible d'utiliser la modularité pour obtenir des représentations synthétiques du graphe. Dans ce travail, un algorithme similaire à celui de (Noack and Rotta 2009) est utilisé de manière hiérarchique pour explorer le graphe et un test de significativité d'une partition de sommets est proposé.

Un critère de modularité organisée

Dans cette partie, nous adaptons l'idée de carte topographique à un contexte qui est spécifique au graphe. Ce travail est décrit dans (Rossi and Villa 2009; Rossi and Villa-Vialaneix 2010). De la même manière que dans la section 1.2.2, nous supposerons donc que nous disposons d'une carte composée de U unités, $\{1,\dots,U\}$ munie d'une structure de voisinage. Cette structure de voisinage est ici modélisée par une mesure de similarité *a priori*, fournie sous la forme d'une matrice S, de dimensions $U \times U$ et telle que $S_{uu} = 1$ et $S_{uu'} = S_{u'u}$. Pour faire le lien avec les notations introduites dans la section 1.2.2, cette matrice peut être $S(u,u') = H(d(u,u'))$, soit par exemple, $S_{uu'} = \exp\left(-\eta d(u,u')\right)$ [6] (pour un $\eta > 0$), la différence étant que cette similarité est fixée et n'évolue pas au cours de l'algorithme contrairement à l'approche classique de cartes auto-organisatrices où H est généralement décroissante au cours de l'apprentissage. Nous introduisons alors le critère de *modularité organisée* (sur la carte) de la partition de sommets $\mathcal{C}_1$, ..., $\mathcal{C}_C$ comme

$$\mathcal{O}(f) = \frac{1}{2m} \sum_{i,j=1}^{n} S_{f(x_i),f(x_j)} \left(W_{ij} - P_{ij}\right) \tag{1.7}$$

où $f(x_i)$ est l'unité (ou classe) dans laquelle le sommet x_i est affecté sur la carte. Le principe de ce critère devient clair lorsque l'on ré-écrit l'expression de la modularité donnée dans l'équation (1.6) sous la forme

$$\mathcal{Q}(\mathcal{C}_1,\dots,\mathcal{C}_C) = \frac{1}{2m} \sum_{i,j=1}^{n} \mathbf{1}_{\{f(x_i)=f(x_j)\}} \left(W_{ij} - P_{ij}\right)$$

6. Dans (Rossi and Villa-Vialaneix 2010), nous utilisons une carte dont les unités sont localisées par un point dans $\mathbb{R}^2$ et pour distance entre ces unités, $d(u,u')$, la distance euclidienne.

où $\mathbf{1}_{\{f(x_i)=f(x_j)\}} = 1$ si et seulement si $f(x_i) = f(x_j)$ (x_i et x_j sont classés dans la même classe) et 0 sinon. La version organisée du critère de modularité de l'équation (1.7) favorise donc, de manière similaire à la modularité mais de façon plus souple, la classification des sommets connectés[7] du graphe dans des unités voisines sur la carte. De manière similaire à ce qui est proposé dans la section précédente (et illustré sur le graphe des Misérables), la classification des sommets sur la carte peut être utilisée pour proposer une représentation statique et simplifiée du graphe, la position des unités sur la grille fournissant une position naturelle pour la représentation des classes de sommets correspondantes.

Tout comme l'optimisation de la modularité, l'optimisation de $\mathcal{O}$ est un problème NP-complet. Dans (Rossi and Villa-Vialaneix 2010), nous proposons une approximation de cette optimisation par un algorithme de recuit déterministe. Pour cela, $\mathcal{O}$ est réécrite sous la forme :

$$\mathcal{O}(f) = \mathcal{F}(M) = \sum_{i,j=1}^{n} \sum_{u,u'=1}^{U} M_{iu} S_{uu'} M_{ju'} B_{ij}$$

où $M_{iu} = \begin{cases} 1 & \text{si } f(x_i) = u \\ 0 & \text{sinon} \end{cases}$ et $B_{ij} = \begin{cases} 0 & \text{si } i = j \\ \frac{1}{2m}(W_{ij} - P_{ij}) & \text{sinon} \end{cases}$. La distribution de Gibbs de notre problème s'écrit alors

$$\mathbb{P}(M) = \frac{1}{Z_P} \exp(\mathcal{F}(M)/T),$$

où Z_P est la constante de normalisation $\sum_M \exp(\beta \mathcal{F}(M))$ et $T > 0$ est la température du système. Cette distribution est approchée par l'introduction d'un *champ moyen*, $(E_{iu})_{i=1,\dots,n,\, u=1,\dots,U}$ qui pondère la matrice d'affectations M de telle sorte que la fonction de coût

$$\mathcal{G}(M,E) = \sum_{i=1}^{n} \sum_{u=1}^{U} M_{iu} E_{iu}$$

approche au mieux $\mathcal{F}(M)$. De manière plus précise, la matrice E est choisie de telle sorte à minimiser la divergence de Kullback-Leibler entre $\widetilde{\mathbb{P}}(M,E) = \frac{1}{Z_{\widetilde{P}}} \exp(\beta \mathcal{G}(M,E))$ ($Z_{\widetilde{P}} = \sum_M \exp(\mathcal{G}(M,E)/T)$) et $\mathbb{P}(M)$. La conséquence de l'utilisation de la distribution $\widetilde{\mathbb{P}}(M,E)$ au lieu de $\mathbb{P}(M)$ est que, sous cette distribution, M_{iu} et $M_{ju'}$ sont indépendants dès lors que $i \neq j$. Le calcul de $Z_{\widetilde{P}}$ devient donc numériquement facilement réalisable, contrairement à celui de Z_P dont la complexité combinatoire est trop élevé. Une approche de type EM est utilisée : celle-ci alterne une phase d'optimization (pour la recherche de E) et une phase de calcul d'espérance (pour le calcul de l'espérance de M sous la distribution $\widetilde{\mathbb{P}}$). La méthode est décrite dans l'algorithme 4.

Une analyse détaillée des performances de l'algorithme sur un exemple jouet (le réseau social du club de karaté de Zachary (Zachary 1977)) ainsi que des comparaisons avec d'autres méthodes sont décrites dans (Rossi and Villa-Vialaneix 2010). En particulier, une des classifications obtenues pour le graphe « Les Misérables » précédemment décrit dans la section 1.2.2 est donnée dans la figure 1.7 (à gauche). Les comparaisons montrent que la méthode d'optimisation de recuit organisée donne généralement de meilleurs résultats en terme de qualité de la classification (du point de vue de la valeur de la modularité) et en terme de qualité du rendu graphique (par rapport à la minimisation du nombre de paires de sommets qui se croisent sur le rendu graphique), que les approches de cartes auto-organisatrices à noyau. Par ailleurs, dans (Rossi and Villa-

7. ou plutôt « significativement » connectés comparativement au modèle nul.

Algorithme 4 Optimisation de la modularité organisée par recuit déterministe

1: Initialiser $\forall\, i = 1,\ldots,n$ et $\forall\, u = 1,\ldots,U$

$$E_{iu} = \frac{2}{U} \sum_{j \neq i} B_{ij} \sum_{u'=1}^{U} S_{uu'}$$

Résultat : E.

2: Initialiser $T^0 \leftarrow \alpha \frac{2\lambda_B \lambda_S}{U}$ où λ_B et λ_S sont les rayons spectraux des matrices B et S et $\alpha > 1$ **Résultat :** T^0

3: **Pour** $l = 1 \rightarrow L$ **Faire** *boucle de recuit*

4: *injection de bruit* $E \leftarrow E + \epsilon$ avec $\epsilon_{iu} \sim \mathcal{U}[0,1]$

5: **Répéter** *étape de type EM*

6: *étape E* : calculer $\mathbb{E}_{\widetilde{\mathbb{P}}}(M_{iu}) = \frac{\exp(E_{iu}/T^l)}{\sum_{u'} \exp(E_{iu'}/T^l)}$

7: *étape M* : calculer E par optimisation de la divergence de Kullback-Leibler :

$$E_{iu} = 2 \sum_{j \neq i} \sum_{u'} \mathbb{E}_{\mathbb{R}}(M_{ju'}) S_{uu'} B_{ij}$$

8: **Jusqu'à** Convergence de E

9: $T^l \leftarrow \nu T^{l-1}$ avec $\nu \simeq \frac{0{,}1T^0}{\alpha}$

10: **Fin Pour**

11: **Résultat :** $\forall\, i = 1,\ldots,n,\ f(x_i) = \max_{u=1,\ldots,U} E_{iu}$

Vialaneix 2010), nous proposons l'utilisation directe des sorties $(E_{iu})_{i=1,\ldots,n,\, u=1,\ldots,U}$ de l'algorithme de recuit déterministe pour produire une représentation dite « floue » du graphe sur la carte : supposons que les coordonnées de l'unité u dans le plan $\mathbb{R}^2$ soient données par $z^u = (z_1^u, z_2^u)$. Pour chaque sommet x_i du graphe, l'espérance de sa position z^{x_i} dans $\mathbb{R}^2$ est alors déterminée par :

$$\mathbb{E}_{\widetilde{\mathbb{P}}}(z^{x_i}) = \sum_u \mathbb{E}_{\widetilde{\mathbb{P}}}(M_{iu}) z^u.$$

Une classification ascendante hiérarchique est alors appliquée à l'ensemble des positions $(z^{x_i})_i$ qui est coupée à une hauteur donnée, ce qui fournit à la fois une classification plus fine que celle qui est obtenue directement sur la grille et des positions pour les classes dans le plan $\mathbb{R}^2$. Une application limitée de quelques itérations d'un algorithme de forces (de type Fruchterman & Reingold (Fruchterman and Reingold 1991)) est enfin effectuée pour ajuster les positions ainsi obtenues et éviter la superposition des classes et des arêtes. La visualisation finale, sur l'exemple « Les Misérables », est donnée dans la figure 1.7 (à droite).

Utiliser la classification pour représenter

Les approches décrites précédemment, basées sur des cartes topologiques, sont pratiquées en une seule étape qui combine classification et visualisation. Cependant, elles peuvent s'avérer trop lourdes d'un point de vue numérique pour des graphes de grandes tailles. Également, il est fréquent que pour des graphes de plusieurs milliers de sommets, l'utilisateur souhaite procéder à l'exploration de la structure de manière hiérarchique : par zooms successifs à l'intérieur des classes, il accède à des détails de plus en plus fins sur des zones d'intérêt. Comme dans ce qui précède, à chaque niveau de la hiérarchie, la

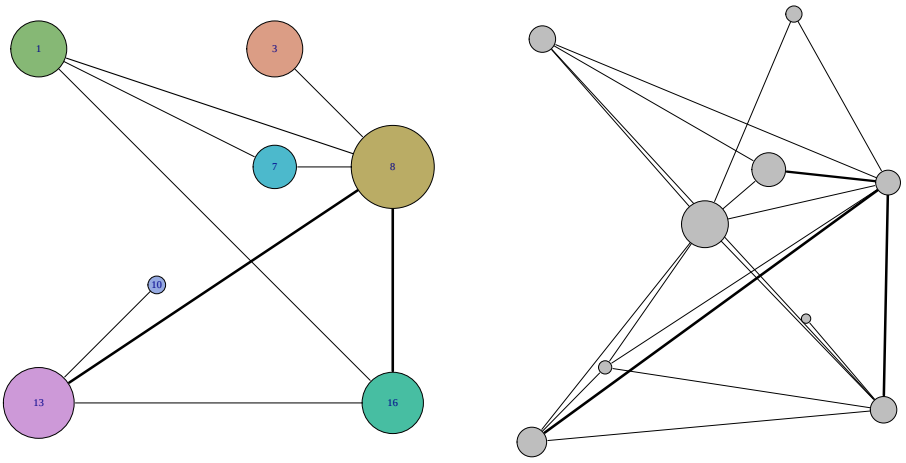

FIGURE 1.7 – Représentation simplifiée du graphe « Les Misérables » obtenue par optimisation de la modularité (à gauche) et représentation floue correspondante obtenue à partir des résultats de l'algorithme de recuit déterministe (à droite). La représentation de droite est plus précise (plus de classes, la granularité de la représentation est plus fine), mais au détriment d'une petite perte de lisibilité (plus d'arêtes qui se croisent, par exemple).

représentation du graphe est souvent simplifiée : les classes seules sont représentées ainsi que les liens qui existent entre elles, et non l'intégralité des sommets (Auber et al. 2003; Seifi et al. 2010; Archambault et al. 2010). L'approche que nous proposons dans (Rossi and Villa-Vialaneix 2011b) est proche de ces approches-ci. Comme les articles (Auber et al. 2003; Seifi et al. 2010; Archambault et al. 2010), notre contribution se base en effet sur une classification hiérarchique des sommets qui, dans notre cas, est effectuée par une méthode rapide d'optimisation de la modularité. Nos apports, dans ces travaux, touchent à plusieurs points méthodologiques :

- pour un graphe (ou un sous-graphe donné), **nous optimisons la modularité** grâce à un algorithme glouton à raffinement hiérarchique comme décrit dans (Noack and Rotta 2009). Par rapport à l'algorithme initial, nous proposons une simple modification qui est une étape de vérification de la connexité des classes obtenues. Comme souligné dans (Archambault et al. 2010), la connexité des classes est cruciale pour une représentation du graphe (simplifié) des classes qui n'induise pas l'utilisateur en erreur lors de l'interprétation de son organisation macroscopique ;
- partant du graphe initial, la modularité est tout d'abord optimisée pour obtenir une partition initiale du graphe puis **le processus est itéré** pour chacune des classes : pour une partition donnée du graphe ou d'un sous-graphe, la modularité est maximisée pour obtenir une partition plus fine de chacune des classes du graphe ou du sous-graphe. Ceci permet, notamment, de limiter le défaut de résolution de la modularité en forçant l'obtention de classes plus fines. Cette méthodologie est schématisée dans la figure 1.8. Le problème d'une telle approche est que chacune des étapes d'optimisation de la modularité fournit une partition des sommets du sous-graphe considéré, celle-ci pouvant être éventuellement dépourvue de sens véritable si le graphe n'a pas une structure modulaire claire. Pour aborder cette question, nous proposons une approche basée sur un test de permutations : **la significativité d'une partition d'un sous-graphe est estimée** en comparant

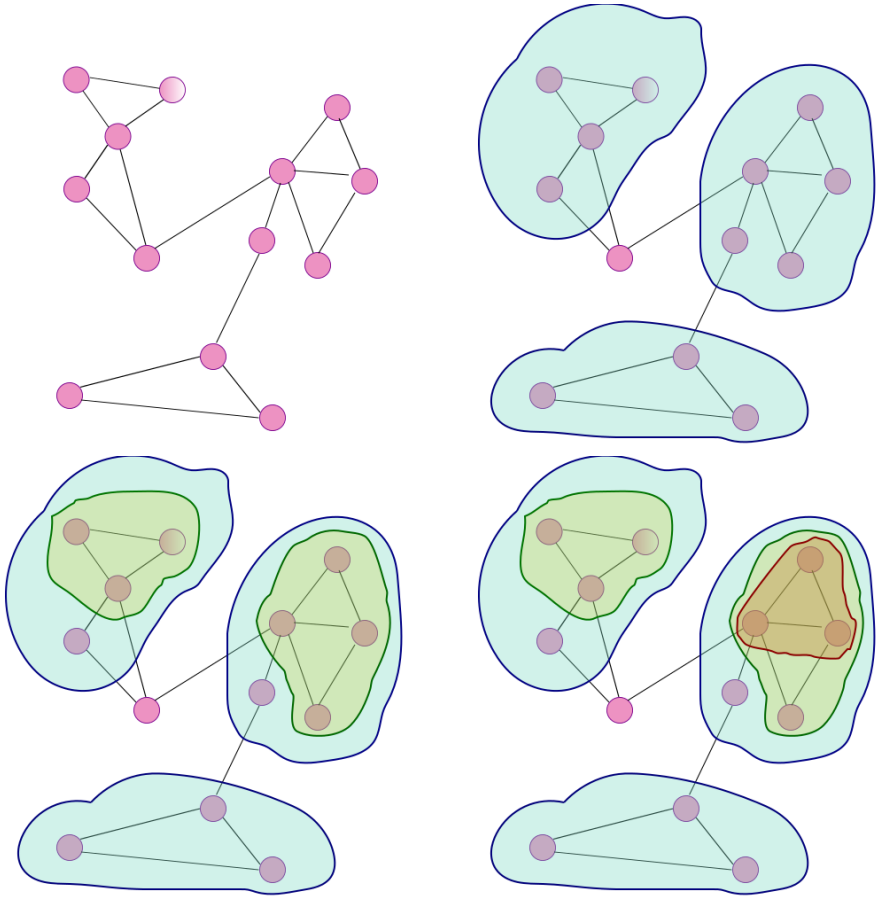

FIGURE 1.8 – Schématisation du processus de classification hiérarchique : partant d'un graphe (en haut à gauche), une première partition des sommets est obtenue par optimisation de la modularité (en haut à droite ; les sommets non entourés correspondent à une classe à part entière) puis chacune des classes de cette partition est à nouveau partitionnée (en bas à gauche) et le processus est itéré sur les classes de la partition ainsi obtenue (en bas à droite).

la modularité de cette partition avec la modularité maximale obtenue pour 100 graphes aléatoires de structures similaires et en ne conservant que les partitions dont la modularité est supérieure à toutes les modularités obtenues sur les 100 graphes aléatoires (modularité dite alors « significativement élevée »). Pour générer les graphes aléatoires de comparaison, nous nous appuyons sur un modèle dit *de configuration* (M.E.J. Newman 2003) qui est une distribution uniforme sur l'ensemble des graphes simples de même distribution des sommets que le graphe (ou le sous-graphe) partitionné. Pour ce faire, nous utilisons l'approche MCMC décrite dans (Roberts Jr. 2000) qui permet d'obtenir un graphe aléatoire de même distribution de degrés qu'un graphe cible, par permutations aléatoires de ses arêtes : les résultats de (Rao et al. 1996) montrent, en effet, que cette approche est une approximation asymptotique du tirage uniforme dans l'ensemble des graphes ayant une distribution de degrés fixée ;

- des **représentations successives des différents niveaux de la hiérarchie de partitions** sont alors construites, en partant de la classification la plus grossière pour aller vers la classification la plus fine. Pour respecter un principe général de

cohérence, l'éclatement d'une classe en sous-classes ne modifie pas le rendu du reste du graphe. Cette contrainte requiert donc d'**estimer pour la partition la plus grossière, l'espace nécessaire pour la représentation de toutes les sous-classes au niveau le plus fin**. Ceci est effectué en procédant de manière récursive : une visualisation de toutes les sous-classes est calculée de manière indépendante par un algorithme de forces adapté et l'espace nécessaire pour une « super-classe » regroupant plusieurs classes est approché par un cercle englobant toutes les sous-classes comme dans la figure 1.9.

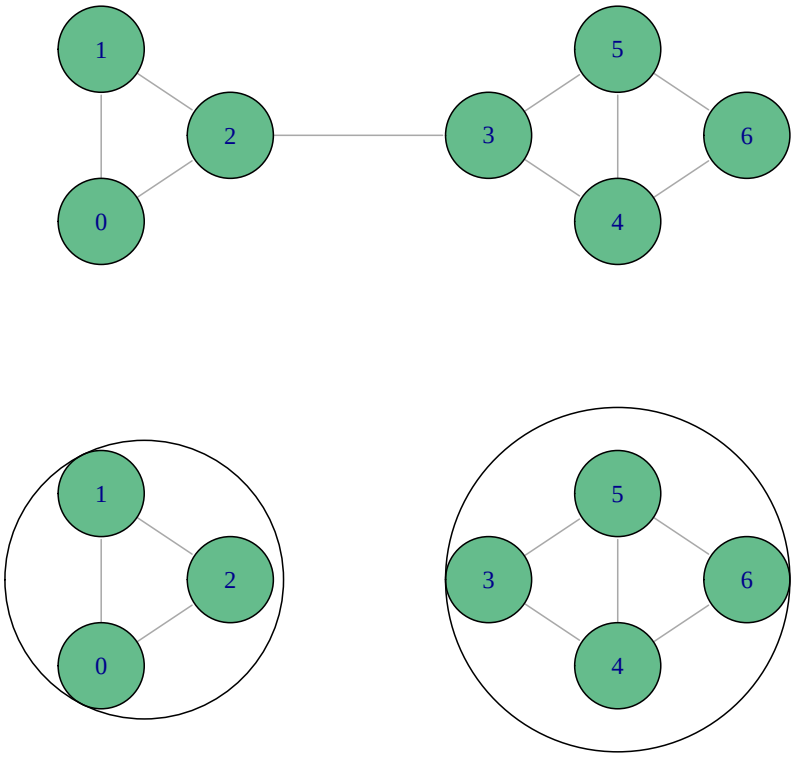

FIGURE 1.9 – Exemple d'estimation de l'occupation des classes : les sommets du graphe d'origine (en haut) sont partitionnés en deux classes dont les visualisations sont calculées indépendamment pour fournir une estimation d'occupation par des cercles englobants (en bas).

Les différentes visualisations sont effectuées en utilisant des algorithmes de forces du type de (Fruchterman and Reingold 1991) mais dans lesquels les forces ont été modifiées pour prendre en compte des tailles de sommets différentes (qui correspondent aux surfaces des classes, proportionnelles à leurs effectifs ou aux disques englobants). De manière plus précise, nous utilisons l'approche proposée dans (Tunkelang 1999) dans laquelle les forces attirant les sommets (analogie aux ressorts) ont une longueur au repos qui est non nulle mais assure le non chevauchement de cercles de rayons donnés qui peuvent être de longueurs différentes.
Enfin, les visualisations sont effectuées de manière récursive : la visualisation la plus grossière est tout d'abord calculée en tenant compte de l'estimation de l'espace

nécessaire au développement des sous-classes. Puis, **les sous-classes sont peu à peu développées** et leur visualisation est calculée en ajoutant une force attractive centrée, pour contraindre les sous-classes d'une même classe à rester autour de l'emplacement prévu pour la classe mère de la visualisation de niveau supérieur, et en ajoutant également des sommets virtuels, comme dans (Eades and Huang 2000), représentant les classes extérieures connectées aux sous-classes de la classe qui est à développer (ces sommets virtuels sont immobiles lors du calcul de la visualisation de la classe).

La méthode proposée est ainsi complètement automatisée et ne nécessite aucun ajustement de paramètre. L'utilisateur doit uniquement choisir le niveau maximal de raffinement envisagé dans la visualisation, mais ce paramètre n'a pas d'influence sur le calcul de la hiérarchie et plusieurs visualisations peuvent être comparées en faisant varier ce paramètre, sans devoir recalculer la classification hiérarchique.

Appliquée au graphe « Les misérables » décrit dans la section 1.2.2, la méthode fournit une classification à deux niveaux :

- au premier niveau, le plus grossier, la classification comprend 6 classes ;
- au second niveau, le plus fin, deux classes de la classification initiale sont partitionnées, respectivement en 3 et 2 sous-classes, soit un total de 9 classes.

La hiérarchie de visualisation peut alors être explorée en trois temps comme présenté dans la figure 1.10. La classification organise ici encore l'histoire du roman « Les Misérables » en sous-histoires avec des classes organisées respectivement autour de Valjean (partagée en trois sous-classes au niveau 2), de Gavroche (partagée en deux sous-classes au niveau 1), des Thénardiers, de Cosette et Marius, de Fantine et de Myriel. Une des limites de l'approche apparaît dans cette représentation : la classe de Valjean, personnage central du roman, en marron sur la figure 1.10, a une position légèrement excentrée due à une sur-estimation de la place nécessaire pour représenter son développement au niveau 2 (cette limite a été soulevée sur un exemple de plus grande taille dans (Rossi and Villa-Vialaneix 2011b)). Toutefois, l'approche prend tout son sens pour l'exploration de

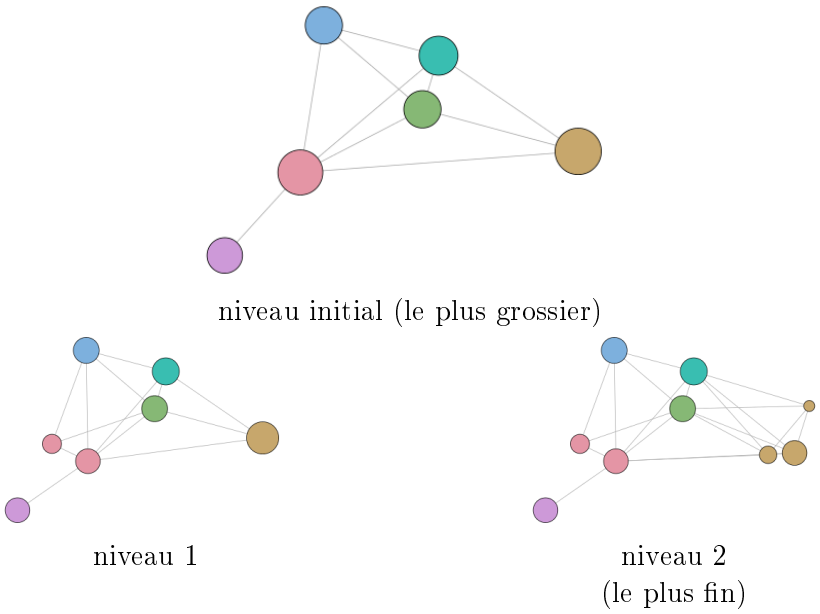

FIGURE 1.10 – Mise en œuvre de la représentation par classification hiérarchique pour le graphe « Les Misérables ».

graphes de grande taille pour lesquels la génération de représentations de plus en plus fines est très rapide et permet une bonne exploration du graphe comme présenté dans la section 1.2.4.

1.2.4 Application pour la fouille de données d'un graphe réel

Les méthodes décrites dans les sections précédentes ont été appliquées à des données réelles et, en particulier, elles sont été utilisées pour un projet mené en collaboration avec des historiens, en partie réalisé dans le cadre du projet « Graphes-Comp » financé par l'ANR[8]. Dans ce programme, un corpus de documents médiévaux, provenant des archives départementales du Lot (France)[9] a été étudié. Ce corpus est donc constitué d'un nombre important de documents dont les actes originaux ont été perdus mais qui ont pu nous parvenir grâce au travail de retranscription d'un feudiste[10]. Les documents du corpus sont tous des actes notariés, chacun décrivant une ou plusieurs transactions et présentant un certain nombre de caractéristiques communes : tout d'abord, les transactions concernent des lieux situés sur la seigneurie de Castelnau Montratier, localisée près de l'actuel village du même nom (Lot, France). Par ailleurs, toutes les transactions relevées par le feudiste décrivent des accords qui, bien que de natures différentes (vente, location, donation, bail à fief...), portent pour la plupart sur des terres et impliquent des rentes. Ces transactions ont été réalisées entre 1238 et 1768, avec une densité de transactions assez variable tout au long de la période. Les transactions ont été modélisées dans une base de données consultable en ligne sur le site web du projet : `http://graphcomp.univ-tlse2.fr` (la manière dont les sources ont été modélisées dans la base de données est brièvement décrite dans (Rossi et al. 2013)). De ces données, deux graphes peuvent être déduits :

- un graphe *biparti* modélisant les relations entre transactions et individus activement impliqués dans celles-ci (voir (Rossi et al. 2013)) ;
- un graphe des individus qui est la projection du graphe biparti précédent (pondéré ou non) : deux individus sont reliés par une arête si ils ont été simultanément impliqués dans la même transaction (voir (Boulet, Jouve, et al. 2008; Rossi and Villa-Vialaneix 2011b; Villa-Vialaneix, Jouve, et al. 2012)).

Dans (Boulet, Jouve, et al. 2008), une approche par carte auto-organisatrice à noyaux a été comparée à des approches algébriques permettant d'extraire de l'information du graphe des individus à partir du spectre de son Laplacien. La carte ainsi produite a fourni une représentation simplifiée du graphe, montrant sa division en trois grandes périodes temporelles (ce qui est consistant avec la connaissance historique puisque les sources et les familles impliquées dans les transactions connaissent un changement abrupt durant la guerre de Cent ans). Le travail a aussi mis en valeur l'imparfaite retranscription des sources dans la base de données. Dans (Rossi and Villa-Vialaneix 2011b), nous reprenons le graphe des individus pour affiner sa représentation avec l'approche hiérarchique décrite dans la section 1.2.3 qui est également mise en relation avec la date des transactions dans lesquelles les individus sont impliqués. Enfin, dans (Villa-Vialaneix, Jouve, et al. 2012), nous combinons l'information relationnelle fournie par le graphe des individus avec l'information spatiale connue sur les transactions pour montrer que ces deux types de données sont significativement dépendantes.

8. Programme Non Thématique, 2005/2009, Graphes-Comp, ANR-05-BLAN-0229.

9. Archives départementales du Lot, ed. by Gérard Miquel and Willy Luis `http://www.lot.fr/cg_archives.php`.

10. Les feudistes sont, au Moyen-Âge, des juristes spécialisés dans le droit féodal et les droits seigneuriaux.

Enfin, dans (Rossi et al. 2013), dans une perspective plus historique, nous montrons comment des études structurelles du graphe biparti peuvent aider à automatiser la recherche des erreurs de transcription et notamment à aider la désambiguisation des homonymes. Également, nous proposons la visualisation du graphe de la figure 1.11 dans laquelle visualisation (par l'algorithme décrit dans (Fruchterman and Reingold 1991)) et classification (par optimisation de la modularité) sont combinés. Chaque classe repré-

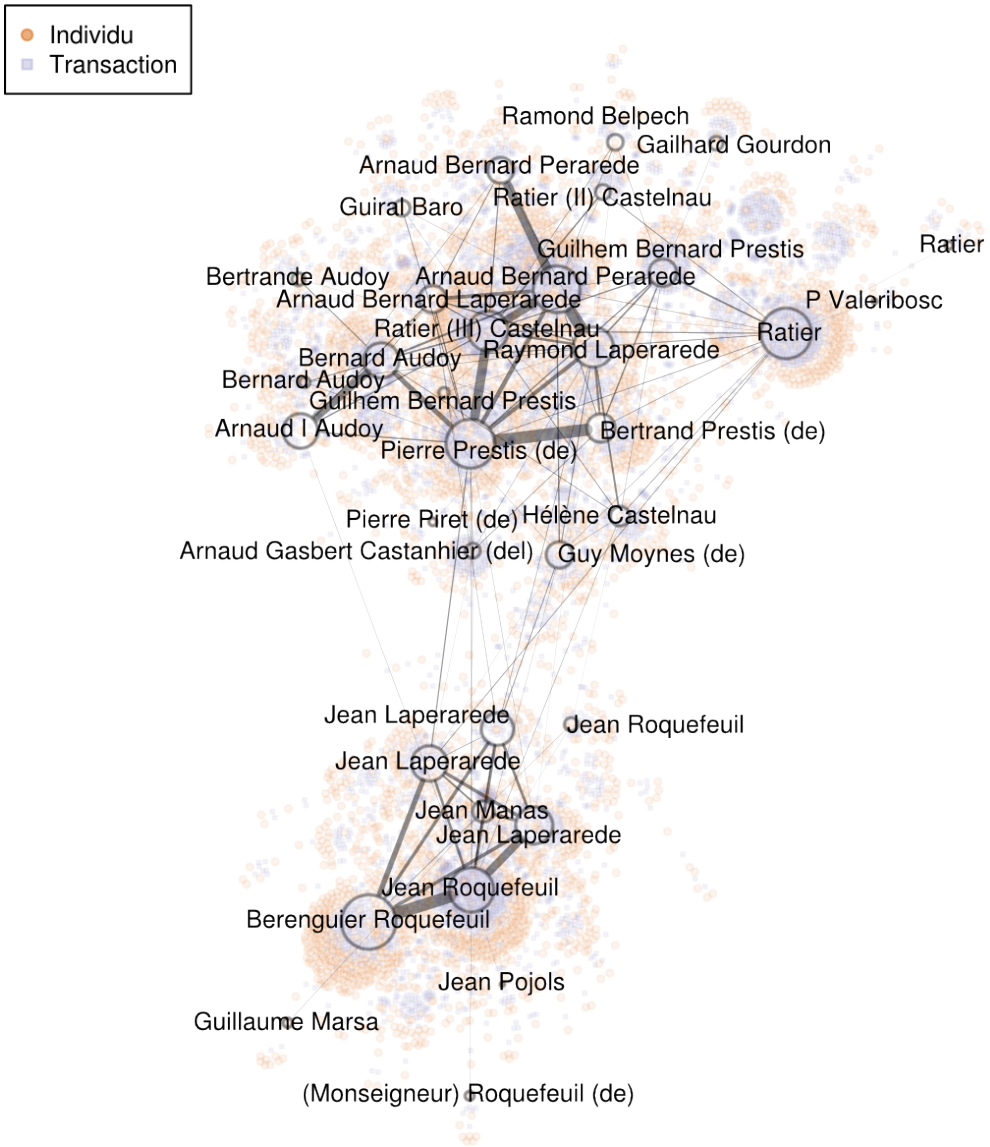

FIGURE 1.11 – Représentation du graphe biparti transactions/individus issu du corpus de documents médiévaux étudié dans le projet « Graphes-Comp ».

sentée sur la figure est étiquetée avec l'individu de plus fort degré qu'elle contient, ce qui permet de visualiser de manière très simple les relations entre les plus gros seigneurs de la région.

Ces travaux ont donné lieu à des articles dans quelques journaux destinés au grand public : sur le blog de Nature, *Nature News*[11], dans *Le Figaro*[12], dans le *Journal du*

11. http://www.nature.com/news/2008/080519/full/news.2008.839.html
12. par Yves Miserey, publié le 24/05/2008, http://bit.ly/1lb63sK

CNRS et sur le Blog de l'Opération *2013, Mathématiques pour la planète terre, 2013*[13].

1.2.5 Conclusions et perspectives

Cette section a présenté plusieurs approches permettant de fouiller la structure d'un graphe. Ces approches sont basées sur des combinaisons de classification non supervisée des sommets de méthodes de visualisation d'un graphe simplifié, dit graphe des classes. Elles utilisent des structures de cartes topologiques, qui définissent des positions *a priori* des classes sur une grille, ou bien des méthodes de représentations hiérarchiques. Les approches en une étape présentent l'avantage de construire une classification et une visualisation en même temps, fournissant une classification construite spécifiquement pour permettre une meilleure visualisation. Toutefois, elles peuvent s'avérer trop lourdes en temps de calcul. L'approche hiérarchique, au contraire, découple classification et représentation et fournit une solution automatisée très rapide mais elle est exclusivement limitée au cadre de graphes simples, contrairement aux approches basées sur des dissimilarités ou des noyaux qui peuvent être utilisées pour analyser des graphes étiquetés mais aussi des données très générales, non vectorielles.

Les perspectives de ce pan de mon travail de recherche sont l'extension de ces approches pour aborder un certain nombre de verrous d'importance pour la fouille de graphe :

- la première thématique d'importance est la **prise en compte d'informations additionnelles à la structure du graphe** : ces informations peuvent être des descripteurs des sommets (des *étiquettes*) ou bien des descripteurs des arêtes (au-delà du poids, des descripteurs qualitatifs qui permettent de construire des *multi-graphes*, c'est-à-dire des graphes contenant plusieurs ensembles d'arêtes). J'ai commencé à aborder cette thématique dans quelques travaux : comme décrit dans la section 1.2.2, (Massoni et al. 2013; Olteanu, Villa-Vialaneix, and Cierco-Ayrolles 2013; Olteanu and Villa-Vialaneix 2015) proposent l'utilisation de multi-noyaux et de multi-dissimilarités pour définir des classes et des cartes auto-organisatrices. Ces approches permettent de traiter de manière naturelle des graphes étiquetés mais pourraient aussi être utilisées pour analyser des multi-graphes de la même manière. Le choix de dissimilarités ou de noyaux appropriés à des types de données divers (numériques, qualitatives, graphes ou données structurées en général) reste encore largement un problème ouvert que je souhaite aborder dans les prochaines années.

  Dans (Laurent and Villa-Vialaneix 2011; Villa-Vialaneix, Liaubet, Laurent, Cherel, et al. 2013), nous avons également proposé l'utilisation de tests pour déterminer si des étiquettes décrivant les sommets avaient une distribution significativement corrélées à la structure du graphe. Dans (Villa-Vialaneix, Liaubet, Laurent, Cherel, et al. 2013) ce type de méthodes est notamment utilisé pour déterminer si un phénotype d'intérêt est significativement corrélé à la structure de co-expression d'un ensemble de gènes régulés par des eQTL. Dans (Laurent and Villa-Vialaneix 2012), nous avons également proposé une méthode de représentation globale des graphes qui utilise des étiquettes à valeurs dans un ensemble fini et qui est basée sur une approche factorielle (type AFC) : les sommets de même étiquette sont représentés plus proches sur le graphe. Cette dernière méthode, quoique prometteuse, ne permet néanmoins de représenter correctement que des graphes de tailles réduites car la représentation produite pour des grands graphes a des défauts de chevauchement des sommets. Ceux-ci pourraient probablement être corrigés en

13. http://mpt2013.fr/commerce-en-reseau-au-moyen-age/

combinant l'approche avec des algorithmes de force.
De manière plus générale, l'**intégration de données multiples**, non nécessairement limitées aux graphes, est une problématique qui sera probablement centrale dans mes activités de recherche future : en effet, celle-ci se pose dans plusieurs projets ANR auxquels je participe déjà et, de manière plus générale, elle revient fréquemment dans de nombreux problèmes en biologie des systèmes. (Lawrence et al. 2008) ont développé une interface graphique sous R (dans le cadre d'un package bioconductor) pour combiner des données numériques (données d'expression par exemple) avec des informations sur le plan d'expérience et des données d'annotation (voir aussi (Cook et al. 2007) pour une discussion sur l'importance de l'utilisation des graphiques pour comprendre les données d'expression). Je souhaiterais étendre ce type d'approches en développant des outils d'exploration permettant l'exploration de données non nécessairement vectorielles. Dans le cadre de la thèse de Jérôme Mariette, nous comptons aborder cette question avec des méthodes proches de celles proposées dans (Olteanu and Villa-Vialaneix 2015). Dans le cadre des projets financés par l'ANR « SusOStress » et « PigHeat » (sur lequel je collabore, entre autres, avec Laurence Liaubet et Magali San Cristobal, GenPhySE, INRA), ainsi que dans le cadre de la thèse de Valérie Sautron (qui s'inscrit dans le projet « SusOStress », co-encadrée avec Elena Terenina et Pierre Mormède, GenPhySE, INRA), l'objectif est d'aborder l'intégration de données obtenues à divers niveaux de l'échelle du vivant (transcriptome, métabolome, formulation sanguine...) dans le cadre de l'étude d'un phénomène biologique d'intérêt (la résistance au stress ou à la chaleur) pour la production agricole (élevage porcin). Ces questions méthodologiques sont cruciales pour répondre à des questions biologiques d'un intérêt majeur pour la production agricole et l'alimentation. Sur ces projets, nous étudions actuellement des approches par des méthodes d'analyses factorielles permettant d'analyser simultanément plusieurs tableaux de données en **prenant en compte des aspects temporels**. Des premiers travaux sur la prise en compte de la temporalité dans l'étude de graphes sont également en cours en collaboration avec Nathalie Viguerie (INSERM), dans le cadre d'un projet européen sur l'obésité, DiOGenes[14]. Dans ces travaux, nous tenons compte de la temporalité et d'informations obtenues à divers niveaux de l'échelle du vivant, pour obtenir des groupes d'éléments au fonctionnement similaire (par exemple, des groupes de gènes co-régulés) qui évolue au cours d'une diète basse calorie.
Enfin, à plus long terme, la problématique de l'intégration de données, vue sous l'angle de l'approche multi-graphes, sera étudiée lors du projet financé par l'ANR « memRNAse » lors duquel nous souhaitons confronter un réseau de co-expression avec un réseau bibliographique *a priori*.

- un deuxième aspect méthodologique d'importance est le **passage à l'échelle des méthodes proposées**. En effet, les données à traiter sont des plus en plus volumineuses en terme de nombre de variables (avec un faible nombre d'observations de chacune de ces variables) ou bien, au contraire, en terme de nombre d'individus à analyser (dans le cadre de grands graphes). Pour aborder ces questions, en collaboration avec Madalina Olteanu (SAMM, Université Paris 1) et dans le cadre de la thèse de Jérôme Mariette (co-encadrée avec Christine Gaspin, MIA-T, INRA), je m'intéresse actuellement à deux types d'approches : d'un côté des approches basées sur des techniques de ré-échantillonnage. Dans un travail préliminaire, (Brunet et al. 2013; Mariette et al. 2014), nous montrons comment des techniques de « bag-

14. http://www.diogenes-eu.org

ging » peuvent être utilisées pour stabiliser une classification des sommets d'un graphe, éventuellement étiqueté. Nous prévoyons un prolongement de ce travail, à la fois en affinant la proposition méthodologique pour développer une méthode rapide, robuste et parallélisable mais aussi en appliquant ce type d'approches pour proposer une typologie des ARN non codants.

Une approche alternative est l'utilisation d'approches *parcimonieuses* : en effet, dans les approches relationnelles et à noyaux décrites dans la section 1.2.2, les prototypes sont exprimés comme combinaison convexe de **toutes** les observations traitées, ce qui peut considérablement augmenter la complexité de la méthode si le nombre d'observations est grand et ce qui réduit également l'intérêt des prototypes, notamment en terme d'interprétation. Utiliser des prototypes exprimés seulement comme combinaison convexe d'un faible nombre d'exemples permet donc de retrouver une meilleure interprétation des classes et également de réduire la complexité de l'approche (voir (Hofmann et al. 2014) pour une étude approfondie de diverses stratégie permettant de réaliser cette tâche dans une méthode d'apprentissage supervisée utilisant des prototypes). Dans (Mariette et al. 2014), nous proposons une première étude basée sur des approches empiriques consistant à sélectionner, par une méthode de bagging, les observations les plus représentatives du jeu de données, la limite actuelle de l'approche étant que celles-ci sont sélectionnées globalement et non par prototype. L'extension de ce travail est en cours d'étude et sera appliqué au problème de typologie des ARN non codants. Une autre méthode pourrait être l'inclusion dans la méthode d'une pénalité L^1 qui permet d'effectuer une sélection de variables au cours de l'apprentissage. Cette approche sera étudiée dans un deuxième temps et est également celle que l'on souhaite intégrer dans les méthodes factorielles d'intégration de données temporelles à l'étude dans le cadre du projet « SusOStress ».

Une vision synthétique des projets de recherche et collaborations à venir sur la thématique de la fouille de graphes est donnée dans la figure 1.12.

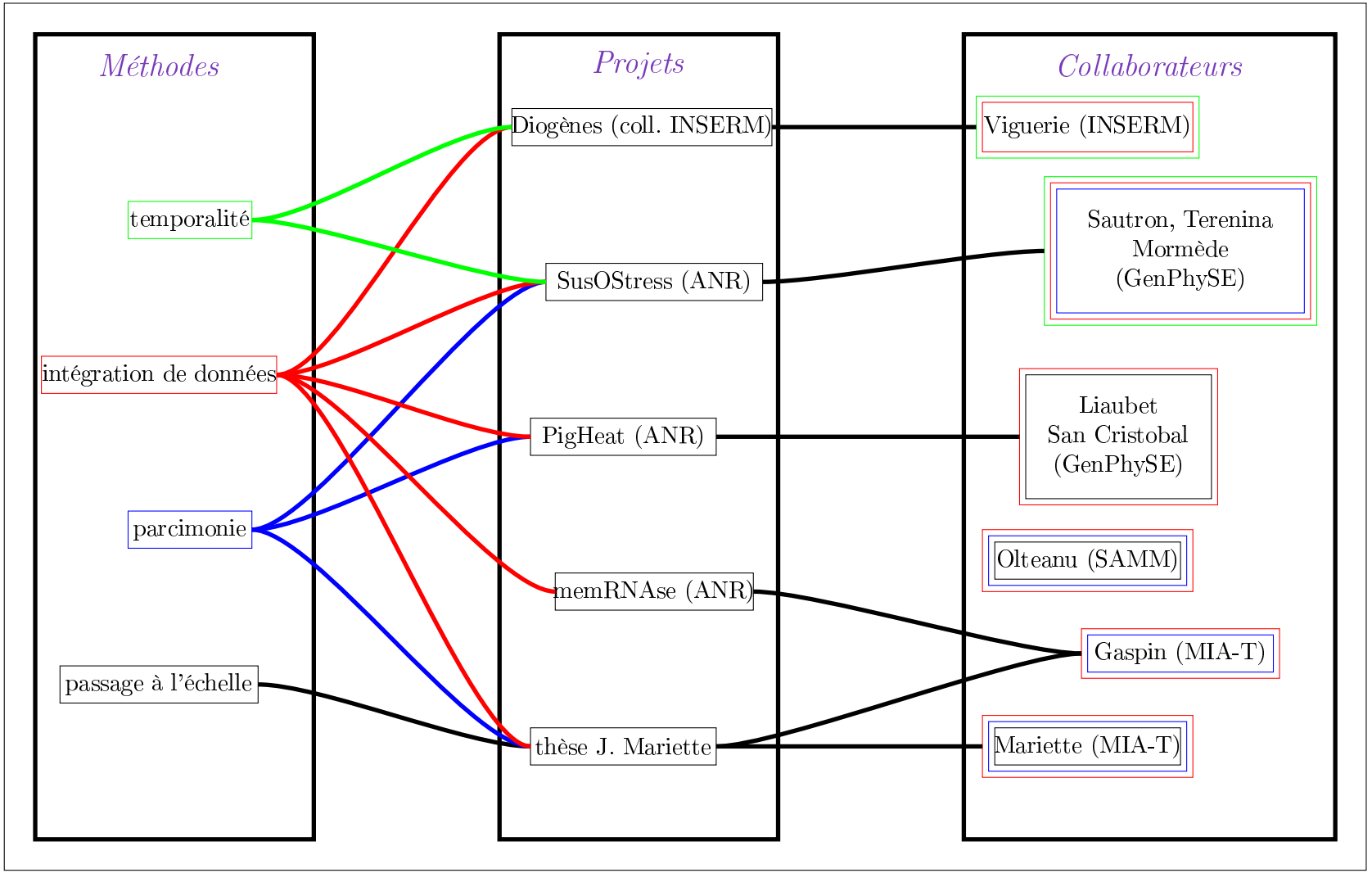

FIGURE 1.12 – Relations entre thématiques méthodologiques de recherche (à gauche, 4 couleurs), projets auxquels je participe (au centre) et collaborateurs (à droite). Les couleurs entourant les collaborateurs correspondent aux thématiques sur lesquelles des collaborations sont en cours.

1.2.6 Références

Adamic, L.A. and N. Glance (2005). « The political blogosphere and the 2004 US election: divided they blog ». In: *Proceedings of the 3rd LINKDD Workshop*. New York, NY, USA: ACM Press, pp. 36–43.

Albert, R., H. Jeong, and A.L. Barabási (1999). « Diameter of the world-wide web ». In: *Nature* 401.130, p. 9907038. DOI: 10.1038/43601.

Andras, P. (2002). « Kernel-Kohonen networks ». In: *International Journal of Neural Systems* 12, pp. 117–135.

Archambault, D., T. Munzner, and D. Auber (2010). « Tugging graphs faster: efficiently modifying path-preserving hierarchies for browsing paths ». In: *IEEE Transactions on Visualization and Computer Graphics* 17.3, pp. 276–289. DOI: 10.1109/TVCG.2010.60. URL: http://hal.archives-ouvertes.fr/index.php?action_todo=search&view_this_doc=inria-00413861&version=1.

Aronszajn, N. (1950). « Theory of reproducing kernels ». In: *Transactions of the American Mathematical Society* 68.3, pp. 337–404.

Auber, D. (2003). « Tulip: a huge graph visualisation framework ». In: *Graph Drawing Softwares*. Ed. by P. Mutzel and M. Jünger. Mathematics and Visualization. Springer-Verlag, pp. 105–126.

Auber, D., Y. Chiricota, F. Jourdan, and G. Melançon (2003). « Multiscale visualization of small world networks ». In: *INFOVIS'03*.

Barabási, A., N. Gulbahcel, and J. Loscalzo (2011). « Network medecine: a network-based approach to human disease ». In: *Nature Reviews Genetics* 12, pp. 56–68.

Bastian, M., S. Heymann, and M. Jacomy (2009). « Gephi: an open source software for exploring and manipulating networks ». In: *Proceedings of the Third International AAAI Conference on Weblogs and Social Media*. Ed. by E. et al. Adar. Menlo Park: AAAI Press, 2009, pp. 361–362. URL: http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154.

Bendhaïba, L., M. Olteanu, and N. Villa-Vialaneix (2013). « SOMbrero : cartes auto-organisatrices stochastiques pour l'intégration de données décrites par des tableaux de dissimilarités ». In: *2èmes Rencontres R BoRdeaux*. (June 27–26, 2013). Lyon, France.

Blondel, V., J.L. Guillaume, R. Lambiotte, and E. Lefebvre (2008). « Fast unfolding of communites in large networks ». In: *Journal of Statistical Mechanics: Theory and Experiment* P10008, pp. 1742–5468.

Boelaert, J., L. Bendhaïba, M. Olteanu, and N. Villa-Vialaneix (2014). « SOMbrero: an R package for numeric and non-numeric self-organizing maps ». In: *Advances in Self-Organizing Maps and Learning Vector Quantization (Proceedings of WSOM 2014)*. (July 2–4, 2014). Ed. by T. Villmann, F.M. Schleif, M. Kaden, and M. Lange. Vol. 295. Advances in Intelligent Systems and Computing. Mittweida, Germany: Springer Verlag, Berlin, Heidelberg, pp. 219–228.

Bonnans, F. (2006). *Optimisation Continue*. Paris, France: Dunod.

Borgatti, S.P., A. Mehra, D.J. Brass, and G. Labianca (2009). « Network analysis in the social sciences ». In: *Science* 323.5916, pp. 892–895. DOI: 10.1126/science.1165821.

Boulet, R., B. Jouve, F. Rossi, and N. Villa (2008). « Batch kernel SOM and related Laplacian methods for social network analysis ». In: *Neurocomputing* 71.7-9, pp. 1257–1273. DOI: doi:10.1016/j.neucom.2007.12.026.

Bourqui, R., D. Auber, and P. Mary (2007). « How to draw clustered weighted graphs using a multilevel force-directed graph drawing algorithm ». In: *Proceedings of the*

11th International Conference Information Visualization, 2007. IV'07. Pp. 757–764. DOI: 10.1109/IV.2007.65.

Brunet, F., J. Mariette, C. Cierco-Ayrolles, C. Gaspin, P. Bardou, and N. Villa-Vialaneix (2013). « Classification d'un graphe de co-expression avec des méta-données pour la détection de micro-RNAs ». In: *Modèles et l'Analyse des Réseaux : Approches Mathématiques et Informatiques (MARAMI 2013).* (Oct. 16–18, 2013). Saint-Étienne, France.

Chen, Y., E.K. Garcia, M.R. Gupta, A. Rahimi, and L. Cazzanti (2009). « Similarity-based classification: concepts and algorithm ». In: *Journal of Machine Learning Research* 10, pp. 747–776.

Combe, D., C. Largeron, E. Egyed-Zsigmond, and M. Géry (2012). « Getting clusters from structure data and attribute data ». In: *Proceedings of International Conference on Advances in Social Networks Analysis and Mining (IEEE/ACM)*, pp. 731–733.

— (2013). « ToTeM: une méthode de détection de communautés adaptées aux réseaux d'information ». In: *Proceedings of 13e Conférence Francophone sur l'Extraction et la Gestion des Connaissances (EGC)*, pp. 305–310.

Conan-Guez, B., F. Rossi, and A. El Golli (2006). « Fast algorithm and implementation of dissimilarity self-organizing maps ». In: *Neural Networks* 19.6-7, pp. 855–863.

Cook, D., H. Hofmann, E.K. Lee, H. Yang, B. Nikolau, and E. Wurtele (2007). « Exploring gene expression data, using plots ». In: *Journal of Data Science* 5, pp. 151–182.

Cottrell, M. and J.C. Fort (1987). « Étude d'un processus d'auto-organisation ». In: *Annales de l'IHP, section B* 23.1, pp. 1–20.

Cottrell, M., J.C. Fort, and G. Pagès (1998). « Theoretical aspects of the SOM algorithm ». In: *Neurocomputing* 21, pp. 119–138.

Cottrell, M. and P. Letrémy (2005). « How to use the Kohonen algorithm to simultaneously analyse individuals in a survey ». In: *Neurocomputing* 63, pp. 193–207.

Cottrell, M., P. Letrémy, and E. Roy (1993). « Analyzing a contingency table with Kohonen maps: a factorial correspondence analysis ». In: *Proceedings of International Workshop on Artificial Neural Networks (IWANN 93)*. Ed. by J. Cabestany, J. Mary, and A. (Eds.) Prieto. Lecture Notes in Computer Science. Springer Verlag, pp. 305–311.

Cruz, J.D., C. Bothorel, and F. Poulet (2011). « Entropy based community detection in augmented social networks," , 2011 International Conference, pp.163-168 doi: 10.1109/CASON.2011.6085937 ». In: *Proceedings of Computational Aspects of Social Networks (CASoN)*, pp. 163–168. DOI: 10.1109/CASON.2011.6085937.

Danon, L., A. Diaz-Guilera, J. Duch, and A. Arenas (2005). « Comparing community structure identification ». In: *Journal of Statistical Mechanics*, P09008.

Di Battista, G., P. Eades, R. Tamassia, and I.G. Tollis (1999). *Graph Drawing: Algorithms for the Visualization of Graphs*. Prentice Hall.

Dorogovtsev, S.N. and J.F.F. Mendes (2003). *Evolution of Networks. From biological Nets to the Internet and WWW*. Oxford University Press.

Eades, P. and Q.W. Feng (1996). « Multilevel visualization of clustered graphs ». In: *Proceedings of International Conference on Graph Drawing, Symposium on Graph Drawing*. Ed. by Stephen C. North. Vol. 1190. Lecture Notes in Computer Science. Berkeley, California, USA: Springer, pp. 101–112.

Eades, P. and M.L. Huang (2000). « Navigating clustered graphs using force-directed methods ». In: *Journal of Graph Algorithms and Applications* 4.3, pp. 157–181.

Edwards, D. (1995). *Introduction to Graphical Modelling*. New York: Springer.

Ester, M., R. Ge, B.J. Gao, Z. Hu, and B. Ben-Moshe (2006). « Joint cluster analysis of attribute data and relationship data: the connected k-center problem ». In: *SIAM International Conference on Data Mining*. ACM Press, pp. 25–46.

Fort, J.C., P. Letremy, and M. Cottrell (2002). « Advantages and drawbacks of the Batch Kohonen algorithm ». In: *Proceedings of 10th European Symposium on Artificial Neural Networks (ESANN 2002)*. Ed. by M. Verleysen. Bruges, Belgium, pp. 223–230.

Fortunato, S. (2010). « Community detection in graphs ». In: *Physics Reports* 486, pp. 75–174. URL: http://arxiv.org/pdf/0906.0612v2.

Fortunato, S. and M. Barthélémy (2007). « Resolution limit in community detection ». In: *Proceedings of the National Academy of Sciences*. Vol. 104. 1. doi:10.1073/pnas.0605965104; URL: http://www.pnas.org/content/104/1/36.abstract, pp. 36–41.

Fouss, F., L. Yen, A. Pirotte, and M. Saerens (2006). « An experimental investigation of graph kernels on a collaborative recommendation task ». In: *IEEE International Conference on Data Mining (ICDM)*, pp. 863–868. URL: http://www.isys.ucl.ac.be/staff/francois/Articles/Fouss2005c.pdf.

Freeman, L.C. (2004). *The Development Of Social Network Analysis: A Study In The Sociology Of Science*. Booksurge.

Fruchterman, T. and B. Reingold (1991). « Graph drawing by force-directed placement ». In: *Software, Practice and Experience* 21, pp. 1129–1164.

Ge, R., M. Ester, J.G. Byron, Z. Hu, B.K. Bhattacharya, and B. Ben-Moshe (2008). « Joint cluster analysis of attribute data and relationship data: the connected k-center problem, algorithms and applications ». In: *ACM Transactions on Knowledge Discovery from Data* 2.2, p. 7.

Ghassany, M., N. Grozavu, and Y. Bennani (2012). « Collaborative clustering using prototype-based techniques ». In: *International Journal of Computational Intelligence and Applications* 11.3, p. 1250017. DOI: 10.1142/S1469026812500174.

Gönen, M. and E. Alpaydin (2011). « Multiple kernel learning algorithms ». In: *Journal of Machine Learning Research* 12, pp. 2211–2268. URL: http://jmlr.org/papers/v12/gonen11a.html.

Guimerà, R. and L.A.N. Amaral (2005). « Functional cartography of complex metabolic networks ». In: *Nature* 433, pp. 895–900. DOI: 10.1038/nature03288.

Guimerà, R., M. Sales-Pardo, and L.A.N. Amaral (2004). « Modularity from fluctuations in random graphs and complex networks ». In: *Physical Review E* 70, 025101(R). DOI: 10.1103/PhysRevE.70.025101.

Hammer, B., A. Gisbrecht, A. Hasenfuss, B. Mokbel, F.M. Schleif, and X. Zhu (2011). « Topographic Mapping of Dissimilarity Data ». In: *Advances in Self-Organizing Maps (Proceedings of the 8th Workshop on Self-Organizing Maps, WSOM 2011)*. Ed. by J. Laaksonen and T. Honkela. Vol. 6731. Lecture Notes in Computer Science. Espoo, Finland: Springer, pp. 1–15.

Hammer, B. and A. Hasenfuss (2010). « Topographic mapping of large dissimilarity data sets ». In: *Neural Computation* 22.9, pp. 2229–2284.

Hammer, B., A. Hasenfuss, F. Rossi, and M. Strickert (2007). « Topographic processing of relational data ». In: *Proceedings of the 6th Workshop on Self-Organizing Maps (WSOM 07)*. Ed. by Bielefeld University Neuroinformatics Group. Bielefeld, Germany.

Hanisch, D., A. Zien, R. Zimmer, and T. Lengauer (2002). « Co-clustering of biological networks and gene expression data ». In: *Bioinformatics* 18.Suppl. 1, S145–S154.

Herman, I., G. Melançon, and M. Scott Marshall (2000). « Graph visualization and navigation in information visualisation ». In: *IEEE Transactions on Visualization and Computer Graphics* 6.1, pp. 24–43.

Heskes, T. (1999). « Energy functions for self-organizing maps ». In: *Kohonen Maps*. Ed. by E. Oja and S. Kaski. Amsterdam: Elsevier, pp. 303–315. URL: http://www.snn.ru.nl/reports/Heskes.wsom.ps.gz.

Heuvel, J. van den and S. Pejic (2001). « Using Laplacian eigenvalues and eigenvectors in the analysis of frequency assignment problems ». In: *Annals of Operations Research* 107.1-4, pp. 349–368.

Hofmann, D., F.M. Schleif, B. Paaß en, and B. Hammer (2014). « Learning interpretable kernelized prototype-based models ». In: *Neurocomputing* 141, pp. 84–96. ISSN: 0925-2312. DOI: 10.1016/j.neucom.2014.03.003. URL: http://www.sciencedirect.com/science/article/pii/S0925231214003968.

Huberman, B.A. and L.A. Adamic (1999). « Growth dynamics of the world-wide web ». In: *Nature* 401.131. DOI: 10.1038/43604.

Knuth, D.E. (1993). *The Stanford GraphBase: A Platform for Combinatorial Computing*. Reading, MA: Addison-Wesley.

Kohohen, T. and P.J. Somervuo (1998). « Self-organizing maps of symbol strings ». In: *Neurocomputing* 21, pp. 19–30.

Kohonen, T. (1995). *Self-Organizing Maps*. Ed. by Springer. Vol. 30. Springer Series in Information Science.

Kondor, R.I. and J. Lafferty (2002). « Diffusion kernels on graphs and other discrete structures ». In: *Proceedings of the 19th International Conference on Machine Learning*, pp. 315–322.

Krislock, N. and H. Wolkowicz (2012). « Handbook on Semidefinite, Conic and Polynomial Optimization ». In: ed. by M.F. Anjos and J.B. Lasserre. Vol. 166. International Series in Operations Research & Management Science. New York, Dordrecht, Heidelberg, London: Springer. Chap. Euclidean distance matrices and applications, pp. 879–914.

Lancichinetti, A. and S. Fortunato (2009). « Community detection algorithms: a comparative analysis ». In: *Physical Review E* 80, p. 056117. DOI: 10.1103/PhysRevE.80.056117.

Lanckriet, G.R.G., N. Cristianini, P. Bartlett, L. El Ghaoui, and M.I. Jordan (2004). « Learning the kernel matrix with semidefinite programming ». In: *Journal of Machine Learning Research* 5, pp. 27–72.

Laurent, T. and N. Villa-Vialaneix (2011). « Using spatial indexes for labeled network analysis ». In: *Information, Interaction, Intelligence (I3)* 11.1. URL: http://www.irit.fr/journal-i3/volume11/numero01/.

— (2012). « Analyse de données pour des graphes étiquetés ». In: *44èmes Journées de Statistique de la SFdS (JdS 2012)*. (May 21–25, 2012). Bruxelles, Belgique.

Lawrence, M., D. Cook, E.K. Lee, H. Babka, and E.S. Wurtele (2008). « `explorase`: multivariate exploratory analysis and visualization for systems biology ». In: *Journal of Statistical Software* 25.9. URL: http://www.jstatsoft.org/v25/i09.

Lebbah, M., A. Chazottes, F. Badran, and S. Thiria (2005). « Mixed Topological Map ». In: *Proceedings of the 13th European Symposium on Artificial Neural Networks (ESANN)*. Ed. by M. Verleysen. Bruges, Belgium, pp. 357–362.

Lee, J.A. and M. Verleysen (2007). *Nonlinear Dimensionality Reduction*. Information Science and Statistics. New York; London: Springer. URL: http://www.worldcat.org/search?qt=worldcat_org_all&q=9780387393506.

Lehmann, S. and L.K. Hansen (2007). « Deterministic modularity optimization ». In: *The European Physical Journal B* 60.1, pp. 83–88.

Li, H., Z. Nie, W.C.W. Lee, C.L. Giles, and J.R. Wen (2008). « Scalable community discovery on textual data with relations ». In: *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, pp. 1203–1212.

Luenberger, D. (1984). *Linear and Nonlinear Programming*. Dordrecht, The Netherlands: Addison-Wesley.

Luxburg, U. von (2007). « A tutorial on spectral clustering ». In: *Statistics and Computing* 17.4, pp. 395–416. URL: http://www.kyb.mpg.de/publications/attachments/luxburg06_TR_v2_4139[1].pdf.

Mac Donald, D. and C. Fyfe (2000). « The kernel self organising map. » In: *Proceedings of 4th International Conference on knowledge-based Intelligence Engineering Systems and Applied Technologies*, pp. 317–320.

Mariette, J., M. Olteanu, J. Boelaert, and N. Villa-Vialaneix (2014). « Bagged kernel SOM ». In: *Advances in Self-Organizing Maps and Learning Vector Quantization (Proceedings of WSOM 2014)*. (July 2–4, 2014). Ed. by T. Villmann, F.M. Schleif, M. Kaden, and M. Lange. Vol. 295. Advances in Intelligent Systems and Computing. Mittweida, Germany: Springer Verlag, Berlin, Heidelberg, pp. 45–54.

Massoni, S., M. Olteanu, and N. Villa-Vialaneix (2013). « Which distance use when extracting typologies in sequence analysis? An application to school to work transitions ». In: *International Work Conference on Artificial Neural Networks (IWANN 2013)*. (June 12–14, 2013). Puerto de la Cruz, Tenerife.

Moser, F., R. Ge, and M. Ester (2007). « Joint cluster analysis of attribute and relationship data without a-priori specification of the number of clusters ». In: *Proceedings of the 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. San Jose, CA, USA, pp. 510–519.

Newman, M. (2004). « Fast algorithm for detecting community structure in networks ». In: *Physical Review E* 69, p. 066133. DOI: 10.1103/PhysRevE.69.066133.

Newman, M.E.J. (2001). « The structure of scientific collaboration networks ». In: *Proceedings of the National Academy of Sciences of the United States of America* 98, p. 0007214. DOI: 10.1073pnas.021544898.

— (2003). « The structure and function of complex networks ». In: *SIAM Review* 45, pp. 167–256.

— (2006). « Finding community structure in networks using the eigenvectors of matrices ». In: *Physical Review, E* 74.036104. URL: http://arxiv.org/abs/physics/0605087.

Newman, M.E.J. and M. Girvan (2004). « Finding and evaluating community structure in networks ». In: *Physical Review, E* 69, p. 026113. DOI: 10.1103/PhysRevE.69.026113. URL: http://www.citebase.org/abstract?id=oai:arXiv.org:cond-mat/0308217.

Ng, A.Y., M.I. Jordan, and Y. Weiss (2002). « On spectral clustering: analysis and an algorithm ». In: *Advances in Neural Information Processing Systems*. Vol. 14, pp. 849–856.

Noack, A. (2007). « Energy models for graph clustering ». In: *Journal of Graph Algorithms and Applications* 11.2, pp. 453–480.

Noack, A. and R. Rotta (2009). « Multi-level algorithms for modularity clustering ». In: *SEA 2009: Proceedings of the 8th International Symposium on Experimental Algorithms*. Berlin, Heidelberg: Springer-Verlag, pp. 257–268. ISBN: 978-3-642-02010-0. DOI: http://dx.doi.org/10.1007/978-3-642-02011-7_24.

Olteanu, M. and N. Villa-Vialaneix (2015). « On-line relational and multiple relational SOM ». In: *Neurocomputing* 147. Forthcoming, pp. 15–30. DOI: 10.1016/j.neucom.2013.11.047.

Olteanu, M., N. Villa-Vialaneix, and C. Cierco-Ayrolles (2013). « Multiple kernel self-organizing maps ». In: *XXIst European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2013)*. (Apr. 24, 2023–Apr. 26, 2008). Ed. by M. Verleysen. Bruges, Belgium: i6doc.com, pp. 83–88.

Olteanu, M., N. Villa-Vialaneix, and M. Cottrell (2012). « On-line relational SOM for dissimilarity data ». In: *Advances in Self-Organizing Maps (Proceedings of WSOM 2012)*. (Dec. 12–14, 2012). Ed. by P.A. Estevez, J. Principe, P. Zegers, and G. Barreto. Vol. 198. AISC (Advances in Intelligent Systems and Computing). Santiago, Chile: Springer Verlag, Berlin, Heidelberg, pp. 13–22. ISBN: 978-3-642-35229-4. DOI: 10.1007/978-3-642-35230-0_2.

Pękalska, E. and R.P.W. Duin (2005). *The Dissimilarity Representation for Pattern Recognition. Foundations and Applications*. Singapore: World Scientific.

Polzlbauer, G. (2004). « Survey and comparison of quality measures for self-organizing maps ». In: *Proceedings of the Fifth Workshop on Data Analysis (WDA'04)*. Ed. by J. Paralic, G. Polzlbauer, and A. Rauber. Sliezsky dom, Vysoke Tatry, Slovakia: Elfa Academic Press, pp. 67–82.

Pons, P. and M. Latapy (2006). « Computing communities in large networks using random walks ». In: *Journal of Graph Algorithms and Applications* 10.2, pp. 191–218. URL: http://www.liafa.jussieu.fr/~pons/publi/communities.pdf.

Porter, M.A., P.J. Mucha, M. Newman, and A.J. Friend (2007). « Community structure in the United States house of representatives ». In: *Physica A* 386, pp. 414–438. DOI: 10.1063/1.2390556.

Porter, M.A., J.P. Onnela, and P.J. Mucha (2009). « Communities in networks ». In: *Notices of the American Mathematical Society* 56.9, pp. 1082–1097.

Rakotomamonjy, A., F.R. Bach, S. Canu, and Y. Grandvalet (2008). « SimpleMKL ». In: *Journal of Machine Learning Research* 9, pp. 2491–2521.

Rao, A.R., R. Jana, and S. Bandyopadhyay (1996). « A Markov Chain Monte Carlo method for generating random (0, 1)-matrices with given marginals ». In: *Sankhyã: The Indian Journal of Statistics, Series A (1961-2002)* 58.2, pp. 225–242.

Reichardt, J. and S. Bornholdt (2006). « Statistical mechanics of community detection ». In: *Physical Review, E* 74.016110.

Roberts Jr., J. M. (2000). « Simple methods for simulating sociomatrices with given marginal totals ». In: *Social Networks* 22.3, pp. 273–283. ISSN: 0378-8733. DOI: 10.1016/S0378-8733(00)00026-5.

Rossi, F. (2014). « How many dissimilarity/kernel self organizing map variants do we need? » In: *Advances in Self-Organizing Maps and Learning Vector Quantization (Proceedings of WSOM 2014)*. (July 2–4, 2014). Ed. by T. Villmann, F.M. Schleif, M. Kaden, and M. Lange. Vol. 295. Advances in Intelligent Systems and Computing. Mittweida, Germany: Springer Verlag, Berlin, Heidelberg, pp. 3–23. DOI: 10.1007/978-3-319-07695-9_1.

Rossi, F., A. Hasenfuss, and B. Hammer (2007). « Accelerating relational clustering algorithms with sparse prototype representation ». In: *Proceedings of the 6th Workshop on Self-Organizing Maps (WSOM 07)*. Bielefeld, Germany: Neuroinformatics Group, Bielefeld University.

Rossi, F. and N. Villa (2009). « Topologically ordered graph clustering via deterministic annealing ». In: *XVth European Symposium on Artificial Neural Networks, Compu-*

tational Intelligence and Machine Learning (ESANN 2009). (Apr. 22–24, 2009). Ed. by M. Verleysen. Bruges, Belgium: d-side publications, pp. 529–534. ISBN: 2-930307-09-9.

Rossi, F. and N. Villa-Vialaneix (2010). « Optimizing an organized modularity measure for topographic graph clustering: a deterministic annealing approach ». In: *Neurocomputing* 73.7-9, pp. 1142–1163. DOI: 10.1016/j.neucom.2009.11.023.

— (2011b). « Représentation d'un grand réseau à partir d'une classification hiérarchique de ses sommets ». In: *Journal de la Société Française de Statistique* 152.3, pp. 34–65. URL: http://publications-sfds.math.cnrs.fr/index.php/J-SFdS/article/view/82/73.

Rossi, F., N. Villa-Vialaneix, and F. Hautefeuille (2013). « Exploration of a large database of French notarial acts with social network methods ». In: *Digital Medievalist* 9. URL: http://www.digitalmedievalist.org/journal/9/villavialaneix/.

Schaeffer, S.E. (2007). « Graph Clustering ». In: *Computer Science Review* 1.1, pp. 27–64.

Schoenberg, I. (1935). « Remarks to Maurice Fréchet's article "Sur la définition axiomatique d'une classe d'espace distanciés vectoriellement applicable sur l'espace de Hilbert" ». In: *Annals of Mathematics* 36, pp. 724–732.

Schölkopf, B., K. Tsuda, and J.P. Vert (2004). *Kernel methods in computational biology*. London: MIT Press.

Scott, J.P. (2000). *Social Network Analysis: A Handbook*. Sage Publications Ltd. ISBN: 978-0761963394.

Seifi, M., M. Guillaume, M. Latapy, and B. Le Grand (2010). « Interactive multiscale visualization of huge graphs: application to a network of weblogs ». In: *8th Workshop on Visualization and Knowledge Extraction (EGC 2010)*.

Smola, A.J. and R. Kondor (2003). « Kernels and regularization on graphs ». In: *Proceedings of the Conference on Learning Theory (COLT) and Kernel Workshop*. Ed. by M. Warmuth and B. Schölkopf. Lecture Notes in Computer Science, pp. 144–158.

Steinhaeuser, K. and N.V. Chawla (2008). « Community detection in a large real-world social network ». In: *Social Computing, Behavioral Modeling, and Prediction*. Ed. by H. Liu, J.J. Salerno, and M.J. Young. Springer US, pp. 168–175.

Traud, A.L., E.D. Kelsic, P.J. Mucha, and M.A. Porter (2011). « Comparing community structure to characteristics in online collegiate social networks ». In: *SIAM Review* 53.3, pp. 526–543. DOI: 10.1137/080734315.

Tunkelang, D. (1999). « A Numerical Optimization Approach to General Graph Drawing ». CMU-CS-98-189. PhD thesis. School of Computer Science, Carnegie Mellon University. URL: http://reports-archive.adm.cs.cmu.edu/anon/1998/abstracts/98-189.html.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. New York, USA: Springer Verlag.

Villa, N. and F. Rossi (2007). « A comparison between dissimilarity SOM and kernel SOM for clustering the vertices of a graph ». In: *6th International Workshop on Self-Organizing Maps (WSOM 2007)*. (Sept. 3–6, 2007). Bielefeld, Germany: Neuroinformatics Group, Bielefeld University. ISBN: 978-3-00-022473-7. DOI: 10.2390/biecoll-wsom2007-139.

Villa-Vialaneix, N., B. Jouve, F. Rossi, and F. Hautefeuille (2012). « Spatial correlation in bipartite networks: the impact of the geographical distances on the relations in a corpus of medieval transactions ». In: *Revue des Nouvelles Technologies de l'Informa-*

tion SHS-1, pp. 97–110. URL: http://www.editions-hermann.fr/ficheproduit.php?prodid=1371.

Villa-Vialaneix, N., L. Liaubet, T. Laurent, P. Cherel, A. Gamot, and M. San Cristobal (2013). « The structure of a gene co-expression network reveals biological functions underlying eQTLs ». In: *PLoS ONE* 8.4, e60045. DOI: 10.1371/journal.pone.0060045.

Villmann, T., H. Sven, and M. Kästner (2012). « Gradient based learning in vector quantization using differentiable kernels ». In: *Advances in Self-Organizing Maps (Proceedings of WSOM 2012)*. (Dec. 12–14, 2012). Ed. by P.A. Estevez, J. Principe, P. Zegers, and G. Barreto. Vol. 198. AISC (Advances in Intelligent Systems and Computing). Santiago, Chile, pp. 193–204.

Wasserman, S. and K. Faust (1994). *Social Network Analysis: Methods and Applications*. Cambridge, UK: Cambridge University Press.

Wellman, B., J. Salaff, D. Dimitrova, L. Garton, M. Gulia, and C. Haythornthwaite (1996). « Computer networks as social networks: collaborative work, telework, and virtual community ». In: *Annual Review of Sociology* 22, pp. 213–238. DOI: 10.1146/annurev.soc.22.1.213.

Yamanishi, Y., J. Vert, and M. Kanehisa (2004). « Protein network inference from multiple genomic data: a supervised approach ». In: *Bioinformatics* 20.Suppl. 1, pp. i363–i370. DOI: 10.1093/bioinformatics/bth910.

Yamanishi, Y., J.P. Vert, and M. Kanehisa (2005). « Supervised enzyme network inference from the integration of genomic data and chemical information ». In: *Bioinformatics* 21.Supp. 1, pp. i468–i477. DOI: 10.1093/bioinformatics/bti1012.

Yamanishi, Y., J.P. Vert, A. Nakaya, and M. Kanehisa (2003). « Extraction of Correlated Gene Clusters from Multiple Genomic Data by Generalized Kernel Canonical Correlation Analysis ». In: *Bioinformatics* 19, pp. 323i–330i.

Young, G. and A. Householder (1938). « Discussion of a set of points in terms of their mutual distances ». In: *Psychometrika* 3, pp. 19–22.

Zachary, W.W. (1977). « An information flow model for conflict and fission in small groups ». In: *Journal of Anthropological Research* 33.4, pp. 452–473.

Zhao, B., J.T. Kwok, and C. Zhang (2009). « Multiple kernel clustering ». In: *Proceedings of the 9th SIAM International Conference on Data Mining (SDM)*. Sparks, Nevada, USA.

Zhou, Y., H. Cheng, and J.X. Yu (2009). « Graph clustering based on structural/attribute similarities ». In: *Proceedings of the VLDB Endowment*. Vol. 2. 1, pp. 718–729.

1.3 Inférence

1.3.1 Introduction

L'inférence de réseau est une thématique de recherche sur laquelle je travaille depuis peu de temps (2011 environ). L'évolution de mon travail vers cette thématique provient de mes collaborations de plus en plus nombreuses dans le domaine de la bio-statistique, et plus particulièrement de la génomique, avec des chercheurs de l'INRA (particulièrement Magali San Cristobal, Laurence Liaubet, Elena Terenina, Pierre Mormède de l'unité GenPhySE, INRA de Toulouse) et de l'INSERM (particulièrement Nathalie Viguerie de l'Institut des Maladies Métaboliques et Cardiovasculaires de l'INSERM de Toulouse). C'est une thématique qui devrait prendre une importance croissante dans mes activités futures suite à mon intégration à l'INRA et qui s'est déjà matérialisée

par ma participation au groupe d'animation du réseau méthodologique MIA de l'INRA « NETBIO » [15].

La question de l'inférence de réseau est devenue une approche importante et répandue en biologie des systèmes à cause du développement très important des techniques d'acquisition de données à haut débit. Ces techniques (type « biopuces » par exemple) permettent d'obtenir des données transcriptomiques dans lesquelles les expressions de plusieurs milliers de gènes sont mesurées simultanément sur un nombre restreint (classiquement quelques dizaines) d'individus. La méthodologie consiste à mesurer la quantité d'ARNs messagers (ARNm) qui correspondent à une liste donnée de gènes, qui sont trouvés dans une cellule donnée, dans des conditions expérimentales données (voir la figure 1.13 [16]pour une schématisation du processus de recueil des données) : l'ARNm est une copie de l'ADN utilisée comme intermédiaire par les cellules dans la synthèse des protéines (voir figure 1.13 à gauche [17] pour une schématisation du processus) et la mesure de la quantité d'un ARNm donné donne donc une mesure quantitative de l'activité du gène correspondant. Toutefois, le processus qui permet de passer de l'ADN à

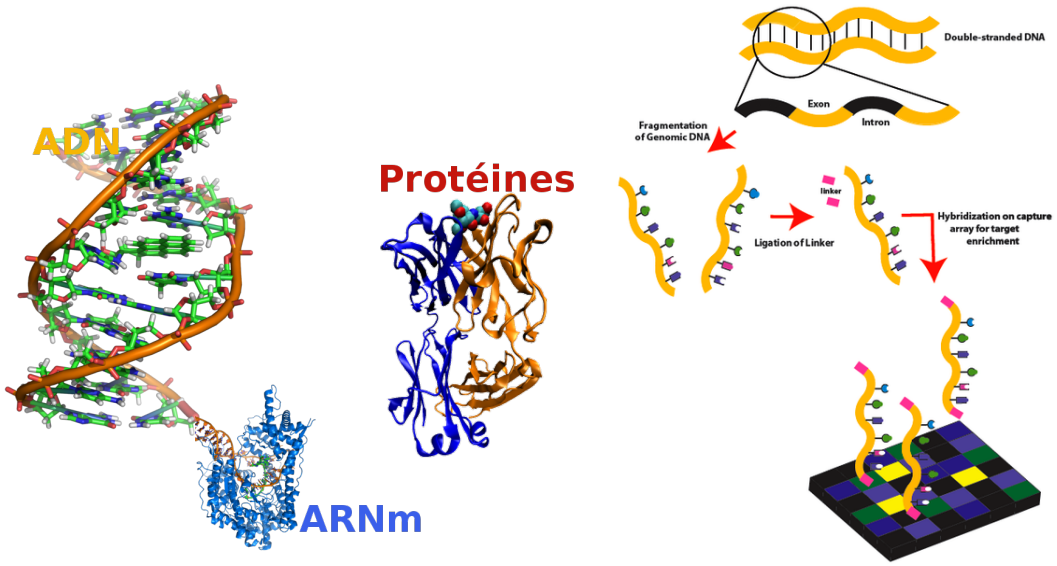

FIGURE 1.13 – Gauche : schématisation du processus biologique de processus de transcription de l'ADN en ARN messager qui permettra la production de protéines. Droite : schématisation des différentes étapes permettant le recueil de données transcriptomiques par technique de biopuces.

la création de protéines n'est pas direct et il existe des phénomènes d'activation et de répression dans les expressions des gènes : ainsi, tout phénomène biologique induit ou est le résultat d'une cascade complexe de régulations au niveau génomique (Abdollahi et al. 2007; Barabási et al. 2011). Dans ce contexte *inférer un réseau* signifie, à partir de données recueillie dans une expérience biologique (typiquement des données transcriptomiques) à tenter de reconstruire cette cascade d'interactions en définissant un graphe dont

- les sommets sont les gènes d'intérêt pour le phénomène biologique étudiés ;

15. « Inférence de réseaux biologiques », http://bit.ly/1pYUtcV.
16. L'image provient de Wikimedia Commons et est attribuable à SarahKusala.
17. L'image est un montage réalisé à partir d'images provenant de Wikimedia Commons et attribuables à Zephyris http://en.wikipedia.org/wiki/User:Zephyris et Thomas Splettstoesser http://commons.wikimedia.org/wiki/User:Splette.

- les arêtes modélisent les relations de régulation entre ces gènes.

Notons toutefois que l'inférence de réseaux biologiques n'est pas réduite à cette seule question des réseaux de co-expression génique mais concerne, d'une manière plus large, la modélisation de tous les systèmes d'intéractions qui peuvent se situer à divers niveaux du vivant. Nous élargirons cette discussion dans la partie conclusion 1.3.4 de cette section.

1.3.2 Motivation et contribution personnelle

Différentes approches ont été utilisées dans le contexte de l'inférence de réseaux : réseaux bayésiens (Pearl and Russel 2002) (voir aussi le package R **bnlearn** (Scutari 2010)), réseaux basés sur le calcul d'une information mutuelle (Meyer et al. 2008) (voir package R **minet**), réseaux issus de méthodes d'apprentissage à noyaux (par exemple (Yamanishi, J.P. Vert, and Kanehisa 2005) dans le cadre de l'inférence d'un réseau enzymatique à partir de données de différentes natures), réseaux inférés par des régressions basées sur des forêts aléatoires (Huynh-Thu et al. 2010)... Selon les cas, ces approches permettent de reconstruire des réseaux pondérés ou non, orientés ou non et contenant des cycles ou non (typiquement, les approches bayésiennes reconstruisent des réseaux orientés mais sans cycle et les approches basées sur des techniques d'apprentissage, des modèles de régression ou des calculs de mesures de ressemblance reconstruisent des réseaux non orientés mais avec éventuellement des cycles).

Parmi les approches les plus anciennes utilisées en biologie, les réseaux de corrélation, souvent appelés « *relevance network* » (Butte and Kohane 2000) procèdent par simple calcul des corrélations (de Pearson) deux à deux entre expression des gènes avant de procéder à un test statistique permettant de sélectionner les paires de gènes significativement corrélées et de déduire le réseau de celles-ci, comme illustré dans la figure 1.14.

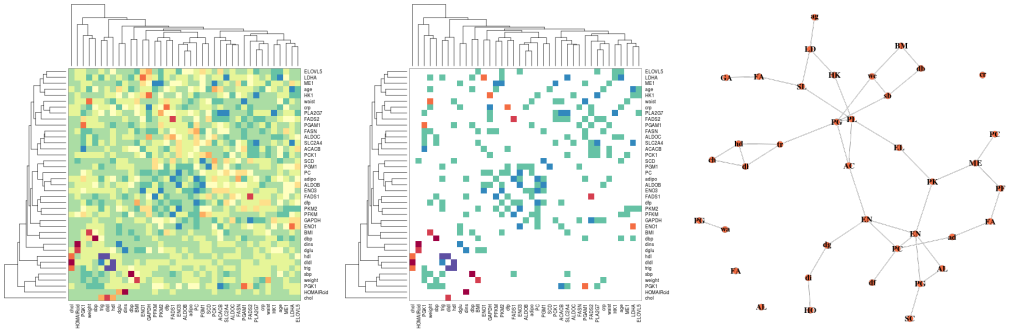

FIGURE 1.14 – Schématisation de la reconstruction d'un réseau de type « relevance network ».

Les limites de cette approche proviennent du fait que les arêtes ainsi inférées ne sont pas réellement adaptées à la modélisation de relations de régulation directes entre gènes. En effet, dans le cas simple où l'expression d'un gène X suit une loi normale $\mathcal{N}(\mu,\sigma^2)$ et où l'expression de deux gènes Y et Z serait définie par la simple relation linéaire suivante :

$$Y = aX + b + \epsilon_1 \qquad \text{et} \qquad Z = cX + d + \epsilon_2$$

(où ϵ_1 et ϵ_2 sont des variables aléatoires centrées, indépendantes de X et, par exemple, de distribution normale), les corrélations $\mathrm{Cor}(X,Y)$ et $\mathrm{Cor}(X,Z)$ sont effectivement élevées mais, de la même manière, la corrélation $\mathrm{Cor}(Y,Z)$ est également élevée alors qu'elle ne correspond pas à une information que le biologiste souhaite retenir. Pour contourner

ce problème, (Schäfer and Strimmer 2005) proposent de s'intéresser aux *corrélations partielles* : étant donné un ensemble de gènes dont les expressions $(X_j)_{j=1,\dots,p}$ suivent une loi de distribution $\mathcal{N}(0,\Sigma)$, on s'intéresse aux quantités

$$\text{Cor}\left(X_j,X_{j'}|(X_k)_{k\neq j,\,j'}\right)$$

et on définit un graphe $\mathcal{G} = (V,E)$ dont les sommets sont les gènes $V = \{1,\dots,p\}$ et dont les arêtes correspondent aux corrélations partielles non nulles :

$$(j,j') \in E \qquad \Leftrightarrow \qquad \text{Cor}\left(X_j,X_{j'}|(X_k)_{k\neq j,\,j'}\right) \neq 0.$$

On peut montrer que, dans le cadre de ce modèle, appelé *Modèle Graphique Gaussien* (D. Edwards 1995), ce problème est équivalent à déterminer la matrice de concentration $\mathbf{S} = \Sigma^{-1}$ puisque

$$\text{Cor}\left(X_j,X_{j'}|(X_k)_{k\neq j,\,j'}\right) = -\frac{\mathbf{S}_{jj'}}{\sqrt{\mathbf{S}_{jj}\mathbf{S}_{j'j'}}}.$$

Dans le contexte de l'inférence de réseaux de co-expression génique, une difficulté supplémentaire est induite par le fait que le nombre n d'observations des expressions des p gènes d'intérêt $\mathbf{X} = (x_{ij})_{i=1,\dots,n,\ j=1,\dots,p}$ est généralement faible devant le nombre de gènes p. La conséquence directe est qu'une estimation naïve de $\mathbf{S}$ à partir de l'inverse de la matrice de variance covariance empirique $\widehat{\Sigma} = \frac{1}{n}\mathbf{X}^T\mathbf{X}$ est de mauvaise qualité à cause du mauvais conditionnement de $\widehat{\Sigma}$ (ce problème est de même nature que celui que je décris dans le cadre de l'analyse des données fonctionnelles dans la section 2.3.1 de ce même manuscrit). Pour répondre à cette question,

1. (Schäfer and Strimmer 2005) proposent une approche en deux temps : l'inverse de Σ est tout d'abord estimée en utilisant une régularisation de la matrice de variance empirique $\widetilde{\Sigma} = \widehat{\Sigma} + \lambda\mathbb{I}_p$ (où $\mathbb{I}_p$ est la matrice identité). Ensuite, les coefficients les plus importants de cet estimé sont sélectionnés par un test bayésien de non nullité d'un ensemble de valeurs ;
2. une approche alternative permet d'effectuer estimation et sélection d'arêtes dans une même étape (Meinshausen and Bühlmann 2006; Friedman et al. 2008) : cette approche consiste à estimer j régressions linéaires

   $$X_j = \beta_j^T X_{-j} + \epsilon$$

   où X_{-j} est le vecteur aléatoire $(X_k)_{k\neq j}$ en utilisant une estimation par maximum de vraisemblance pénalisé par la norme L^1 (régression LASSO (Tibshirani 1996)) qui conduit à résoudre le problème d'optimisation suivant :

   $$\arg \min_{\beta_1,\dots,\beta_p:\ \beta_j\in\mathbb{R}^{p-1}} \sum_{i=1}^{n}\sum_{j=1}^{p}\left(x_{ij} - \beta_j^T\mathbf{X}_{i,-j}\right)^2 + \lambda\|\beta_j\|_{L^1}$$

   avec :
   - $\mathbf{X}_{i,-j}$ la i-ème ligne de la matrice $\mathbf{X}$ privée de sa j-ème colonne ;
   - $\|\beta_j\|_{L^1} = \sum_{k=1}^{p-1}|\beta_{jk}|$.

   La pénalisation par la norme L^1 induit la parcimonie des coefficients $(\beta_j)_j$, c'est-à-dire le fait que, sauf pour un nombre limité de coefficients k, $\beta_{jk} = 0$. Les coefficients non nuls correspondent exactement aux arêtes du graphe recherché car :

   $$\beta_{jj'} = -\frac{\mathbf{S}_{jj'}}{\mathbf{S}_{jj}}.$$

C'est dans le cadre de ce modèle que, jusqu'à présent, mes travaux en inférence de graphes se situent. De manière plus précise, ceux-ci abordent la question d'adapter ce type de modèles aux problématiques posées par les plans d'expérience plus complexes des projets en génomique qui, par exemple, intègrent des données issues de plusieurs conditions expérimentales, observées à plusieurs pas de temps ou faisant intervenir des éléments qui correspondent à divers niveaux de l'échelle du vivant (transcriptome, protéome, phénotypes...). Le paragraphe suivant décrit principalement le travail (Villa-Vialaneix et al. 2014a) qui aborde la question de l'inférence jointe de réseaux issus de données qui correspondent à plusieurs conditions expérimentales. Il correspond aux thématiques mises en valeur dans la figure 1.15 qui reprend la figure 2. Il faut noter que, du point de vue des méthodes statistiques utilisées, ces travaux, qui sont basés sur des modèles non linéaires, sont assez différents des méthodes sur lesquelles j'ai travaillé précédemment (méthodes non linéaires et/ou non paramétriques principalement), le point commun se situant au niveau de la prise en compte de la grande dimension ($p \ll n$), notamment par des approches pénalisées, comme je l'avais fait précédemment dans mes travaux de thèse sur l'analyse de données fonctionnelles (voir le chapitre 2 qui suit). Mes principaux collaborateurs sur cette thématique sont Matthieu Vignes (INRA de Toulouse, Unité MIA-T), Magali San Cristobal (INRA de Toulouse, Unité GenPhySE) et Nathalie Viguerie (INSERM de Toulouse, Laboratoire I2MC).

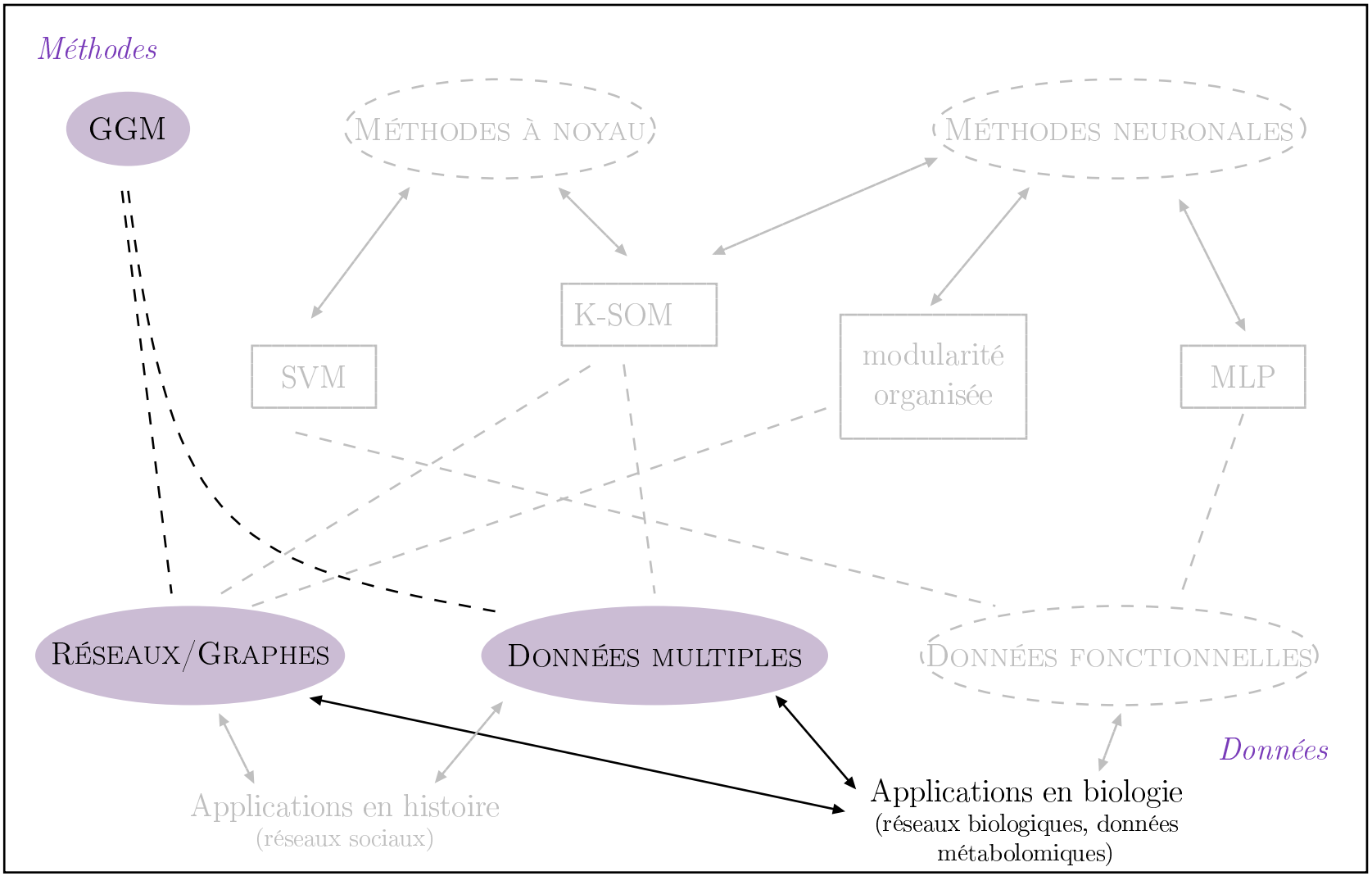

Figure 1.15 – Contributions présentées dans la section 1.3 « Inférence ».

1.3.3 Consensus LASSO

Le travail présenté dans cette section correspond au cas, fréquent en biologie des systèmes, où les données transcriptomiques sont collectées dans des conditions expérimentales variées afin d'essayer de comprendre l'impact de cette condition expérimentale d'intérêt sur le fonctionnement de l'organisme. J'ai été confrontée à ce type de données lors de deux collaborations distinctes :

- d'une part, lors d'une collaboration avec des chercheuses de l'unité GenPhySE, INRA (Laurence Liaubet, Magali San Cristobal), où des données d'expression avaient été récoltés dans le muscle de cochons de différentes races (un travail préliminaire a été réalisé dans le cadre du stage de Nicolas Edwards en juin 2013 (Villa-Vialaneix, N.A. Edwards, et al. 2012), qui a consisté à comparer diverses approches d'inférence jointes sur ces données) ;
- d'autre part, lors d'une collaboration avec Nathalie Viguerie (INSERM) dans le cadre du projet pan-européen d'études de l'obésité DiOGenes[18]. Dans ce projet, l'expression de 221 gènes a été mesurée dans le tissu lipidique de 204 femmes obèses *avant* et *après* un régime basse calorie (voir (Viguerie et al. 2012) pour une première analyse de ces données avec une approche d'inférence simple basée sur les ratio d'évolution des gènes entre les pas de temps qui a également été réalisée dans le cadre du stage de Nicolas Edwards).

Une approche courante, avec ce type de données, consiste à rechercher des *gènes différentiellement exprimés*, c'est-à-dire à rechercher des gènes dont l'expression est significativement différente selon la condition expérimentale (*ie*, entre deux pas de temps, entre différentes races). Une question plus difficile est alors de comprendre, non seulement comment la condition expérimentale d'étude influence l'expression individuelle de chacun des gènes, mais aussi comment elle impacte la manière dont les gènes interagissent entre eux, c'est-à-dire quelles paires de gènes sont liées (co-exprimés ou régulés l'un par l'autre), *indépendamment* de la condition et quels gènes sont liés dans une condition particulière.

Une approche naïve pour aborder cette question consiste à inférer un réseau différent dans chacune des conditions et ensuite à comparer les deux réseaux. Cependant, cette méthodologie n'exploite que partiellement l'information disponible faisant fi, notamment, de la similarité des processus biologiques sous-jacents dans les différentes conditions (l'hypothèse biologique selon laquelle il doit exister un fonctionnement commun indépendant de la condition expérimentale compte tenu de la similarité des expériences). Particulièrement dans le cas où le nombre d'échantillons disponibles est faible, ce qui est fréquent dans le cas de données transcriptomiques, une telle stratégie peut s'avérer assez inefficace. Dans le cadre du modèle graphique gaussien, plusieurs approches ont été proposées pour tenir compte de manière plus fine du protocole expérimental et proposer des procédures d'inférence *jointes* (Chiquet et al. 2011; Mohan et al. 2012; Danaher et al. 2013). Dans ce domaine, notre contribution a consisté à proposer une approche pénalisée dans laquelle une pénalité permet de contrôler explicitement les différences entre les divers réseaux inférés en calculant leur distance (en norme L^2) par rapport à un réseau dit « consensus ». De manière plus précise, si on note $(X_j^c)_{j=1,\dots,p} \sim \mathcal{N}(0,\Sigma^c)$ pour $c = 1,\dots,k$ les variables aléatoires modélisant l'expression des gènes $1, \dots, p$ dans les conditions $1, \dots, k$ et $(x_{ij}^c)_{i=1,\dots,n_c}$ les n_c observations indépendantes de ces variables pour chaque condition, le problème de l'estimation indépendante par modèle LASSO graphique est équivalent (voir Chiquet et al. 2011) à la résolution de p problèmes

18. http://www.diogenes-eu.org

d'optimisation quadratique indépendants :

$$\arg\min_{\beta_j} \frac{1}{2}\beta_j^T \widehat{\Sigma}_{\backslash j \backslash j}\beta_j + \beta_j^T \widehat{\Sigma}_{j\backslash j} + \lambda\|\beta_j\|_{L^1}, \tag{1.8}$$

où :

- $\beta_j = (\beta_j^1, \ldots, \beta_j^k) \in \mathbb{R}^{k(p-1)}$;
- $\widehat{\Sigma}_{\backslash j \backslash j}$ est la matrice diagonale par bloc $\mathbb{D}\text{iag}\left(\widehat{\Sigma}^1_{\backslash j \backslash j}, \ldots, \widehat{\Sigma}^k_{\backslash j \backslash j}\right)$ avec $\widehat{\Sigma}^c_{\backslash j \backslash j}$ la matrice de variance empirique des observations relatives à la condition c privée de la ligne et de la colonne j ;
- $\widehat{\Sigma}_{j\backslash j}$ est, de manière similaire, le vecteur de dimension $k(p-1)$, $\left(\widehat{\Sigma}^1_{j\backslash j}, \ldots, \widehat{\Sigma}^k_{j\backslash j}\right)$ où $\widehat{\Sigma}^c_{j\backslash j}$ correspond à la ligne j privée de la colonne j de la matrice de variance empirique des observations relatives à la condition c ;
- $\lambda > 0$ est un paramètre de régularisation.

Notre proposition (Villa-Vialaneix et al. 2014a) consiste en l'ajout aux problèmes de l'équation (1.8) d'une pénalité contrôlant les différences entre conditions :

$$\frac{1}{2}\beta_j^T \widehat{\Sigma}_{\backslash j \backslash j}\beta_j + \beta_j^T \widehat{\Sigma}_{j\backslash j} + \lambda\|\beta_j\|_1 + \mu \sum_{c=1}^{k} \left\|\beta_j^{\text{cons}} - \beta_j^c\right\|_2^2. \tag{1.9}$$

où μ est un deuxième paramètre de régularisation et où β_j^{cons} est un vecteur de $\mathbb{R}^{k(p-1)}$ qui permet de modéliser le fonctionnement commun inter-conditions. Nous proposons deux types de choix pour β_j^{cons} :

1. une valeur fixe qui permet par exemple, l'incorporation d'un a priori biologique qui peut être, par exemple, un réseau extrait d'une connaissance bibliographique ;
2. un consensus adaptatif, optimisé simultanément avec les $(\beta_j^c)_c$ à partir desquels il est défini de manière explicite. Un choix naturel pour β_j^{cons} est par exemple la moyenne des $(\beta_j^c)_c$:

$$\beta_j^{\text{cons}} = \sum_c \frac{n_c}{n}\beta_j^c.$$

Dans les deux cas, nous montrons que l'équation (1.9) peut être ré-écrite sous la forme d'un problème quadratique

$$\frac{1}{2}\beta_j^T \mathcal{Q}_j^1(\mu) + \beta_j^T \mathcal{Q}_j^2(\mu) + \lambda\|\beta_j\|_1$$

dans lesquelles les matrices $\mathcal{Q}_j^1(\mu)$ et $\mathcal{Q}_j^2(\mu)$ sont des matrices qui ont des valeurs explicites dépendantes (éventuellement) de μ et indépendantes de λ. Ceci permet d'obtenir des problèmes qui s'écrivent tous sous la forme

$$\mathcal{C}(\beta_j) + \lambda\mathcal{P}(\beta_j)$$

où $\mathcal{C}(\beta_j)$ est une fonction quadratique de β_j et $\mathcal{P}(\beta_j)$ est une pénalité L^1. La résolution de ces problèmes est mise en œuvre en utilisant des stratégies similaires à celles décrites dans (Osborne et al. 2000; Chiquet et al. 2011) qui utilisent des approches par « ensemble actif ». Elle est implémentée dans le package R **therese** disponible sur la forge R : `http://therese-pkg.r-forge.r-project.org`. Nous décrivons également, dans (Villa-Vialaneix et al. 2014a) une approche par bootstrap permettant d'améliorer dans certains cas les performances de cette approche, qui est similaire à la méthode BOLASSO (Bach 2008). La question d'améliorer la stabilité de l'inférence de réseau dans

le cadre du LASSO graphique se pose en effet à cause du faible nombre d'observations disponibles ; alternativement, (Haury et al. 2012) abordent ce problème en utilisant une approche par *sélection stable* (Meinshausen and Bühlmann 2010).

D'un point de vue pratique, nous avons démontré la pertinence de l'approche jointe comparée à des estimations indépendantes sur des données simulées. Les données du projet DiOGenes, décrites au début de cette section, y sont également analysées pour l'obtention de deux réseaux (voir figure 1.16). Certaines des interactions trouvées dans

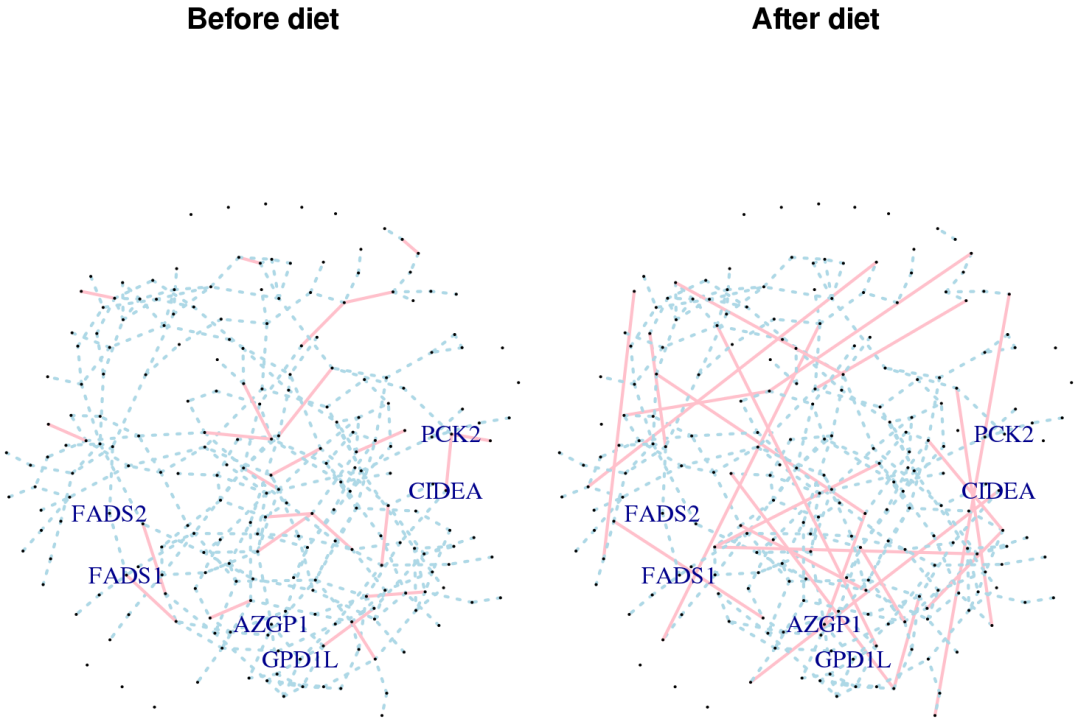

FIGURE 1.16 – Réseaux inférés par utilisation de l'approche jointe « Consensus LASSO » sur des données transcriptomiques relatives à l'obésité : réseau correspondant à l'état avant le régime (à gauche) et après le régime (à droite).

ce réseau (soit communes aux deux conditions, soit spécifiques à une condition) ont été validées par les biologistes comme cohérentes avec les connaissances actuelles sur le sujet.

1.3.4 Conclusions et perspectives

Dans cette section, j'ai présenté mes travaux de recherche les plus récents, qui se tournent vers l'inférence de graphes plutôt que leur analyse exploratoire. Cette évolution a été dictée par mon orientation vers les applications en biologie où, contrairement aux applications en sciences humaines et sociales, les graphes (ou réseaux) sont rarement donnés *a priori*. Dans certains cas, il est possible d'avoir à disposition un réseau *a priori* tiré d'études précédentes mais alors celui-ci est susceptible d'être imparfaitement représentatif de la réalité, soit qu'il est seulement connu en partie, soit qu'il correspond à une réalité "globale" qui ne rend pas compte des particularités d'un tissu donné dans une

condition expérimentale d'intérêt donnée. Ces problématiques sont d'un intérêt croissant en biologie et les problématiques que je souhaite aborder sur ce sujet dans les prochaines années font écho à celles que je décris dans la section 1.2.5 sur mes perspectives de recherche en fouille de données pour les graphes.

Comme pour l'analyse de graphes, l'intégration de données est un verrou méthodologique important pour s'adapter au mieux aux plans d'expérience de plus en plus complexes qui sont utilisés dans les projets en biologie intégrative. En particulier, on peut penser à :

- la **prise en compte d'informations de sources multiples** dans l'inférence : par exemple, (Yamanishi, J.P. Vert, and Kanehisa 2005) intègrent, en utilisant des noyaux, des informations génomiques et chimiques pour reconstruire un réseau enzymatique. (Charbonnier et al. 2010) proposent une approche par pénalisation pour introduire des contraintes sur le réseau inféré, contraintes qui permettent d'intégrer un *a priori* biologique. De manière proche, (Rapaport et al. 2007) introduisent une pénalisation basée sur la structure d'un réseau bibliographique *a priori* dans des méthodes de classification supervisée et non supervisée : ce type d'approches pourrait être étendu aux problèmes d'inférence. J'ai également commencé à aborder ce type de problèmes dans le travail présenté précédemment (Villa-Vialaneix et al. 2014a) qui permet d'intégrer un *a priori* biologique dans l'inférence sous la forme d'un réseau fixé auquel le réseau inféré devrait ressembler. Cette proposition a été développée dans le cadre de l'inférence jointe de réseaux à partir de données obtenues dans des conditions expérimentales distinctes mais liées ; elle pourrait être étendue à l'inférence simple et la problématique du projet, financé par l'ANR, « memRNAse » dans lequel l'inférence de réseau pourrait être dirigée par une connaissance biologique d'un réseau bibliographique *a priori* (qui existe pour l'organisme modèle sur lequel nous travaillons) ;
- des problèmes d'inférence faisant intervenir **plusieurs niveaux de l'échelle du vivant** : par exemple, dans (Montastier et al. 2014), nous avons commencé à aborder cette question en inférant un réseau permettant de modéliser les interactions entre gènes, acides lipidiques et phénotypes dans le cadre d'une étude sur le fonctionnement du tissu lipideux chez les obsèses. L'approche proposée est basée sur une méthodologie mélangeant Analyse Canonique des Corrélations dans sa version parcimonieuse (comme implémentée dans le package R **mixOmics** (Lê Cao et al. 2009)) avec un modèle graphique gaussien ; une approche similaire est présentée dans (Rengel et al. 2012) (utilisant une PLS parcimonieuse) pour inférer un graphe biparti gènes/phénotypes. Le travail effectué dans la pré-publication citée ci-dessus est encore assez préliminaire d'un point de vue méthodologique et l'approche développée pourrait être appronfondie pour permettre son utilisation à des problèmes plus généraux que celui traité, en étudiant de manière systématique diverses stratégies pour combiner les différents niveaux biologiques ;
- la **prise en compte de l'aspect temporel des données** : lorsque les données transcriptomiques sont collectées à divers pas de temps, une approche naturelle consiste à remplacer le modèle linéaire à observations indépendantes du modèle graphique gaussien par un modèle linéaire auto-régressif (Opgen-Rhein and Strimmer 2007; Shimamura et al. 2009; Charbonnier et al. 2010) : dans ce cas, le réseau est construit à partir des coefficients du modèle auto-regressif et modélise les relations de dépendances temporelles entre gènes. Ce modèle est adapté à la recherche de relations de régulation statiques lorsque les données sont collectées à des pas de temps proches et permettent donc d'observer des relations causales entre les

expressions à des pas de temps consécutifs. (Lebre et al. 2010; Jung et al. 2013) proposent des approches permettant d'inférer des réseaux qui varient au cours du temps, ces modèles restant basés sur des modèles de régression dans lesquels l'expression d'un gène au temps t est prédite par l'expression des autres gènes au temps $t-1$. Dans les données du projet « Diogènes », décrites dans la section précédente, les mesures transcriptomiques ont été effectuées à des pas de temps distants de plusieurs mois. L'hypothèse selon laquelle un modèle auto-régressif pourrait permettre de modéliser les relations de régulation ne paraît donc pas, dans ce cas-ci, très réaliste. Toutefois, l'approche Consensus LASSO est elle-aussi imparfaite : elle ne tient pas compte de la nature appariée des données, c'est-à-dire du fait que les mêmes individus sont observés entre les diverses expériences. Une extension de notre travail viserait à permettre d'intégrer cet aspect et offrirait une alternative aux approches existantes pour l'inférence de réseaux à partir de données temporelles lorsque les pas de temps sont suffisamment distants pour que l'hypothèse selon laquelle les relations de régulation sont directement observées entre deux pas de temps consécutifs ne soit pas réaliste.

1.3.5 Références

Abdollahi, A., C. Schwager, J. Kleeff, I. Esposito, S. Domhan, P. Peschke, K. Hauser, P. Hahnfeldt, L. Hlatky, J. Debus, J.M. Peters, H. Friess, J. Folkman, and P.E. Huber (2007). « Transcriptional network governing the angiogenic switch in human pancreatic cancer ». In: *Proceedings of the National Academy of Sciences* 104.31, pp. 12890–12895. DOI: 10.1073/pnas.0705505104. URL: http://www.pnas.org/content/104/31/12890.abstract.

Bach, F. (2008). « Bolasso: model consistent lasso estimation through the bootstrap ». In: *Proceedings of the Twenty-fifth International Conference on Machine Learning (ICML)*. URL: http://www.di.ens.fr/~fbach/fbach_bolasso_icml2008.pdf.

Barabási, A., N. Gulbahcel, and J. Loscalzo (2011). « Network medecine: a network-based approach to human disease ». In: *Nature Reviews Genetics* 12, pp. 56–68.

Butte, A. and I. Kohane (2000). « Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements ». In: *Proceedings of the Pacific Symposium on Biocomputing*, pp. 418–429.

Charbonnier, C., J. Chiquet, and C. Ambroise (2010). « Weighted-Lasso for structured network Inference from time course data ». In: *Statistical Applications in Genetics and Molecular Biology* 9.1. DOI: 10.2202/1544-6115.1519.

Chiquet, J., Y. Grandvalet, and C. Ambroise (2011). « Inferring multiple graphical structures ». In: *Statistics and Computing* 21.4, pp. 537–553.

Danaher, P., P. Wang, and D. Witten (2013). « The joint graphical lasso for inverse covariance estimation accross multiple classes ». In: *Journal of the Royal Statistical Society Series B*. Forthcoming. DOI: 10.1111/rssb.12033.

Edwards, D. (1995). *Introduction to Graphical Modelling*. New York: Springer.

Friedman, J., T. Hastie, and R. Tibshirani (2008). « Sparse inverse covariance estimation with the graphical lasso ». In: *Biostatistics* 9.3, pp. 432–441.

Haury, A.C., F. Mordelet, P. Vera-Licona, and J.P. Vert (2012). « TIGRESS: trustful inference of gene regulation using stability selection ». In: *BMC Systems Biology* 6.1, p. 145. ISSN: 1752-0509. DOI: 10.1186/1752-0509-6-145. URL: http://www.biomedcentral.com/1752-0509/6/145.

Huynh-Thu, V.A., A. Irrthum, L. Wehenkel, and P. Geurts (2010). « Inferring regulatory networks from expression data using tree-based methods ». In: *PLoS ONE* 5.9, e12776. DOI: doi:10.1371/journal.pone.0012776.

Jung, N., F. Bertrand, S. Bahram, L. Vallat, and M. Maumy-Bertrand (2013). « Cascade: a R package to study, predict and simulate the diffusion of a signal through a temporal gene network Bioinformatics first published online ». In: *Bioinformatics* 3, btt705. DOI: doi:10.1093/bioinformatics/btt705.

Lê Cao, K.A., I. González, and S. Déjean (2009). « *****Omics: an R package to unravel relationships between two omics data sets ». In: *Bioinformatics* 25.21, pp. 2855–2856. DOI: 10.1093/bioinformatics/btp515. URL: http://bioinformatics.oxfordjournals.org/cgi/content/abstract/25/21/2855.

Lebre, S., J. Becq, F. Devaux, M. Stumpf, and G. Lelandais (2010). « Statistical inference of the time-varying structure of gene-regulation networks ». In: *BMC Systems Biology* 4.1, p. 130. ISSN: 1752-0509. DOI: 10.1186/1752-0509-4-130. URL: http://www.biomedcentral.com/1752-0509/4/130.

Meinshausen, N. and P. Bühlmann (2006). « High dimensional graphs and variable selection with the Lasso ». In: *Annals of Statistic* 34.3, pp. 1436–1462.

— (2010). « Stability selection ». In: *Journal of the Royal Statistical Society Series B* 72.4, pp. 417–473. DOI: 10.1111/j.1467-9868.2010.00740.x.

Meyer, P.E., F. Lafitte, and G. Bontempi (2008). « minet: A R/Bioconductor package for inferring large transcriptional networks using mutual information ». In: *BMC Bioinformatics* 9.461.

Mohan, K., J.Y. Chung, S. Han, D. Witten, S.I. Lee, and M. Fazel (2012). « Structured learning of Gaussian graphical models ». In: *Proceedings of NIPS (Neural Information Processing Systems) 2012*. Lake Tahoe, Nevada, USA.

Montastier, E., S. Caspar-Bauguil, N. Villa-Vialaneix, P. Hlavaty, E. Tvrzicka, I. Gonzalez, W.H.M. Saris, D. Langin, M. Kunesova, and N. Viguerie (2014). « System model network for adipose tissue signatures related to weight changes in response to calorie restriction and subsequent weight maintenance ». Submitted for publication (first co-author among three first authors).

Opgen-Rhein, R. and K. Strimmer (2007). « Learning causal networks from systems biology time course data: an effective model selection procedure for the vector autoregressive process ». In: *BMC Bioinformatics* 8.Suppl 2, S3. ISSN: 1471-2105. DOI: 10.1186/1471-2105-8-S2-S3. URL: http://www.biomedcentral.com/1471-2105/8/S2/S3.

Osborne, M.R., B. Presnell, and B.A. Turlach (2000). « On the LASSO and its dual ». In: *Journal of Computational and Graphical Statistics* 9.2, pp. 319–337.

Pearl, J. and S. Russel (2002). *Bayesian Networks*. Cambridge, Massachussets, USA: Bradford Books (MIT Press).

Rapaport, F., A. Zinovyev, M. Dutreix, E. Barillot, and J.P. Vert (2007). « Classification of microarray data using gene networks ». In: *BMC Bioinformatics* 8, p. 35. DOI: 10.1186/1471-2105-8-35.

Rengel, D., S. Arribat, P. Maury, M.L. Martin-Magniette, T. Hourlier, M. Laporte, D. Varès, S. Carrère, P. Grieu, S. Balzergue, J. Gouzy, P. Vincourt, and N.B. Langlade (2012). « A gene-phenotype network based on genetic variability for drought responses reveals key physiological processes in controlled and natural environments ». In: *PloS One* 7, e45249. DOI: 0.1371/journal.pone.0045249.

Schäfer, J. and K. Strimmer (2005). « An empirical Bayes approach to inferring large-scale gene association networks ». In: *Bioinformatics* 21.6, pp. 754–764. DOI: 10.1093/bioinformatics/bti062.

Scutari, M. (2010). « Learning Bayesian networks with the bnlearn R package ». In: *Journal of Statistical Software* 35.3, pp. 1–22.

Shimamura, T., S. Imoto, R. Yamaguchi, A. Fujita, M. Nagasaki, and S. Miyano (2009). « Recursive regularization for inferring gene networks from time-course gene expression profiles ». In: *BMC Systems Biology* 3, p. 41. DOI: doi:10.1186/1752-0509-3-41.

Tibshirani, R. (1996). « Regression shrinkage and selection via the lasso ». In: *Journal of the Royal Statistical Society, series B* 58.1, pp. 267–288.

Viguerie, N., E. Montastier, J.J. Maoret, B. Roussel, M. Combes, C. Valle, N. Villa-Vialaneix, J.S. Iacovoni, J.A. Martinez, C. Holst, A. Astrup, H. Vidal, K. Clément, J. Hager, W.H.M. Saris, and D. Langin (2012). « Determinants of human adipose tissue gene expression: impact of diet, sex, metabolic status and cis genetic regulation ». In: *PLoS Genetics* 8.9, e1002959. DOI: 10.1371/journal.pgen.1002959.

Villa-Vialaneix, N., N.A. Edwards, L. Liaubet, and N. Viguerie (2012). « Comparison of network inference packages and methods for multiple network inference ». In: *1ères Rencontres R BoRdeaux*. (July 2–3, 2012). BoRdeaux, France.

Villa-Vialaneix, N., M. Vignes, N. Viguerie, and M. San Cristobal (2014a). « Inferring networks from multiple samples with concensus LASSO ». In: *Quality Technology and Quantitative Management* 11.1, pp. 39–60. URL: http://www.cc.nctu.edu.tw/~qtqm/qtqmpapers/2014V11N1/2014V11N1_F3.pdf.

Yamanishi, Y., J.P. Vert, and M. Kanehisa (2005). « Supervised enzyme network inference from the integration of genomic data and chemical information ». In: *Bioinformatics* 21.Supp. 1, pp. i468–i477. DOI: 10.1093/bioinformatics/bti1012.

2 — Analyse de données fonctionnelles

2.1 Introduction

Durant ma thèse de doctorat et durant les années qui ont suivi, je me suis intéressée à l'analyse de données dite *fonctionnelles*. Ces données se retrouvent dans nombre de problèmes réels, où des mesures d'une quantité physique sont effectuées de manière continue : mesures spectrométriques, courbes de croissance d'individus, courbes de température, enregistrements de voix... D'un point de vue mathématique, ces données sont souvent modélisées par une variable X vivant dans un *espace fonctionnel*, c'est à dire un espace de fonctions plus ou moins régulières, de dimension infinie : l'exemple typique consiste à considérer que les mesures proviennent de l'observation d'une variable aléatoire à valeur dans $L^2[0,1]$, l'ensemble des fonctions de carré intégrable définies sur $[0,1]$ (voir Ramsay and Silverman 1997; Ramsay and Silverman 2002; Ferraty and Vieu 2006 pour les principaux ouvrages de référence sur le sujet). En pratique, c'est une version discrétisée des fonctions, $X(t_1), \ldots, X(t_D)$ qui est observée, où les $(t_d)_{d=1,\ldots,D}$ peuvent être déterministes ou aléatoires, uniformément répartis ou irrégulièrement répartis, dépendre ou non de l'observation considérée mais beaucoup de travaux sur le sujet ne considèrent pas cette dimension de la question et travaillent directement à la prise en compte de variables observées vivant dans $L^2[0,1]$ (voir les articles (Cardot, Ferraty, and P. Sarda 2003; Rossi and Villa-Vialaneix 2011a) pour des exceptions).

2.2 Contribution personnelle

Dans ce domaine, je me suis essentiellement intéressée à développer des méthodes d'*apprentissage supervisé* pour *la discrimination et la régression fonctionnelles*, c'est à dire, des modèles permettant de prédire les valeurs d'une variable aléatoire cible Y, à valeur dans $\mathbb{R}$ (régression) ou dans $\{-1,1\}$ (discrimination binaire), à partir des valeurs de X. La solution de ce problème est construite à partir d'observations indépendantes du couple (X,Y) : (x_1,y_1), ... (x_n,y_n). Pour ce faire, j'ai surtout travaillé sur des méthodes issues de l'apprentissage statistique (Devroye et al. 1996; Vapnik 1998; Györfi et al. 2002) en utilisant, en particulier, des approches neuronales et des approches à noyau. Mes contributions dans ce domaine ont consisté à proposer des approches dites *inverses*, dans l'esprit de la régression inverse par tranches (K. Li 1991) pour des ré-

gressions fonctionnelles par perceptron multi-couches (section 2.3.1) ou basées sur un modèle inverse gaussien (section 2.3.2) : des résultats de consistance de ces deux approches ont été démontrés sous des hypothèses de travail classiques dans le contexte de la régression fonctionnelle mais ces hypothèses peuvent être contraignantes en pratique. Une alternative à ces approches utilise des méthodes à noyau, semblables aux Machines à Vaste Marge, SVM du cadre multi-dimensionnel (section 2.4) : les résultats de consistance obtenus pour ces approches sont des résultats de consistance universelle, qui ne requièrent pas ou peu d'hypothèses techniques sur la distribution du couple (X,Y).

Dans cette partie de mon habilitation, je vais présenter une version synthétique de ces travaux. Cette thématique de recherche est moins présente dans mes activités de recherche actuelle. Je présenterai donc, dans une courte partie de conclusion et perspectives (section 2.5), les relations qui peuvent exister entre cette thématique et mes préoccupations actuelles en recherche ainsi que la manière dont mes connaissances sur les données fonctionnelles et l'apprentissage statistique peuvent être utilisées dans le cadre de mon travail de recherche actuel.

Les travaux présentés dans cette partie sont principalement ceux des articles (Ferré and Villa 2005; Ferré and Villa 2006) (section 2.3.1), (Rossi and Villa 2006; Villa and Rossi 2006b; Rossi and Villa-Vialaneix 2011a) (section 2.4) et (Hernández, Biscay, Villa-Vialaneix, and Talavera 2014a) (section 2.3.2). Une partie de ces travaux est déjà présentée dans mon manuscrit de thèse (**villavialaneix_T2005**). Enfin, mes principaux collaborateurs sur le sujet sont Louis Ferré (professeur à l'Université Toulouse 2), qui a dirigé ma thèse puis, depuis 2004, Fabrice Rossi (professeur à l'Université Paris 1) et, depuis 2008, Noslen Hernandez (chercheuse au CENATAV, La Havane, Cuba) et Rolando Biscay (enseignant chercheur à l'Universidad de Valparaiso, Chili). Les activités de recherche que je présente dans cette partie sont résumées dans la figure 2.1 qui correspond à la figure 2 dans laquelle les thèmes non abordés ont été grisés.

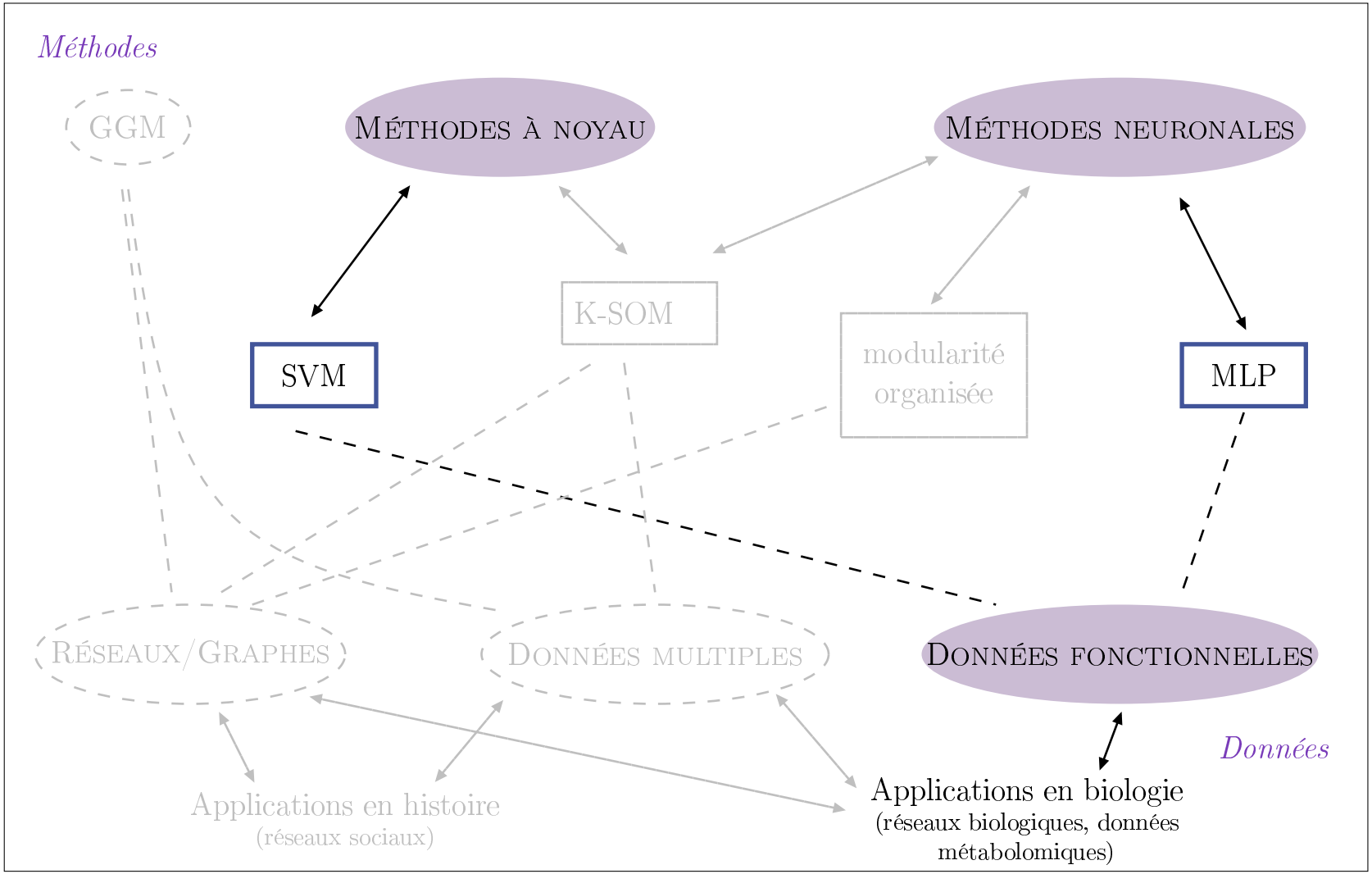

FIGURE 2.1 – Contributions présentées dans la partie 2 « Analyse de données fonctionnelles ».

2.3 Approches dites « inverses »

Une première partie de mes travaux de recherche en analyse des données fonctionnelles a porté sur des approches inverses, c'est-à-dire des approches où l'on s'appuie sur le modèle dans lequel Y explique X pour estimer la régression de Y en X. Dans ce contexte, durant ma thèse, j'ai travaillé sur un modèle semi-paramétrique combiné à un réseau de neurones (section 2.3.1) et, plus récemment, j'ai étudié un problème de calibration qui peut être utile en spectrométrie (section 2.3.2).

2.3.1 Régression inverse et perceptron multi-couches

Dans les articles (Ferré and Villa 2005; Ferré and Villa 2006), nous avons développé une méthode semi-paramétrique pour l'estimation de problèmes de discrimination $Y \in \{1,\ldots,C\}$ ou de régression $Y \in \mathbb{R}$ lorsque la variable explicative est à valeur dans un espace de Hilbert, $(\mathcal{H},\langle .,. \rangle)$, typiquement $\mathcal{H} = L^2[0,1]$ muni du produit scalaire ordinaire de cet espace. L'idée principale consiste à se placer dans le cadre de la régression inverse, introduite par (K. Li 1991) qui propose de se baser sur le modèle suivant :

$$Y = F\left(\langle X,\beta_1\rangle,\ldots,\langle X,\beta_d\rangle,\epsilon\right) \tag{2.1}$$

où F est une fonction inconnue (à estimer), ϵ est une variable aléatoire centrée et indépendante de X et les $(\beta_j)_{j=1,\ldots,d}$ sont (dans le cadre multi-dimensionnel) des vecteurs linéairement indépendants. Cette approche a été étendue au cadre fonctionnel dans les articles (Dauxois et al. 2001; Ferré and Yao 2003; Ferré and Yao 2005; Amato et al. 2006) qui établissent notamment des conditions suffisantes de validité du modèle (voir Théorème 2.1 de (Ferré and Yao 2003)). Ces conditions sont satisfaites, par exemple, pour des variables aléatoires elliptiques). Dans ce cas, les paramètres β_1, ..., β_d du modèle (2.1) sont à valeurs dans $\mathcal{H}$ (*ie*, des fonctions).

Le point clé de ce modèle est d'estimer l'espace EDR (*Effective Dimension Reduction*), qui est l'espace engendré par les $(\beta_j)_{j=1,\ldots,d}$, à partir d'observations i.i.d. du couple (X,Y). Si Γ_X est l'opérateur de variance de X (que l'on supposera, sans perte de généralité, centrée),

$$\Gamma_X = \mathbb{E}(X \otimes X), \qquad \text{où } X \otimes X : u \in \mathcal{H} \to \langle X,u\rangle X \in \mathcal{H},$$

alors, les travaux (Dauxois et al. 2001; Ferré and Yao 2003; Ferré and Yao 2005) montrent que l'espace EDR contient les vecteurs Γ_X-orthonormés de l'opérateur $\Gamma_X^{-1}\Gamma_{\mathbb{E}(X|Y)}$ associés à ses d premières valeurs propres. La difficulté, dans le cadre fonctionnel, est que Γ_X a un inverse non borné, c'est-à-dire non continu (les valeurs propres de cet opérateur, même si toutes positives, ont 0 pour point d'accumulation). Aussi, l'estimateur empirique de Γ_X,

$$\widehat{\Gamma}_X^n = \frac{1}{n}\sum_{i=1}^{n} x_n \otimes (x_n - \bar{x}), \qquad \text{avec } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

est mal conditionné et ne peut être utilisé pour estimer Γ_X^{-1}. (Ferré and Yao 2003; Ferré and Yao 2005) suggèrent de procéder par seuillage de l'opérateur Γ_X^{-1} comme proposé par (Bosq 1991) dans le cadre de séries temporelles ou par (Cardot, Ferraty, and P Sarda 1999) dans le cadre du modèle linéaire fonctionnel.

Nos apports dans ces travaux sont de deux natures :

- tout d'abord, dans (Ferré and Villa 2005; Ferré and Villa 2006), nous proposons d'**estimer l'espace EDR en utilisant une approche par régularisation** :

l'estimateur $\widehat{\Gamma}_X$ est pénalisé par un opérateur de régularisation. Dans le cadre fonctionnel, des opérateurs pertinents de régularisation peuvent être

$$\forall\, h,h' \in \mathcal{H}[h,h'] = \int_0^1 D^2h(t)D^2h(t)dt$$

où D^2 est l'opérateur retournant la dérivée d'ordre 2 du sous-ensemble $\mathcal{S}$ de $\mathcal{H}$ des fonctions deux fois différentiables. Γ_X^{-1} est donc estimé par l'inverse de l'opérateur $\widetilde{\Gamma}_X^{\alpha} : \ h \in \mathcal{H} \longrightarrow \Gamma_X h + \lambda[h,h]$, où λ est un paramètre de régularisation. Dans (Ferré and Villa 2006), nous démontrons la consistance de cette approche pour l'estimation de l'espace EDR ;

- ensuite, dans (Ferré and Villa 2006), nous proposons **l'estimation de la fonction F de l'équation (2.1) par une méthode de réseau de neurones, le perceptron multi-couches**. Ce travail étend les articles (T. Chen and H. Chen 1995; Sandberg and Xu 1996; Rossi and Conan-Guez 2005) qui proposent d'utiliser les perceptrons multi-couches pour l'estimation non paramétrique de problèmes de régression et de discrimination fonctionnelles. Les stratégies qui étendent les réseaux de neurones au cadre fonctionnel comprennent l'utilisation directe des courbes en entrée du réseau ou la projection préalable de celles-ci sur une base fonctionnelle fixe ou une base fonctionnelle dérivée d'une ACP fonctionnelle de X. Notre proposition est similaire à cette dernière approche mais nous suggérons d'utiliser en entrée du perceptron la projection des données sur l'espace EDR $\langle X,\beta_1\rangle$, ..., $\langle X,\beta_d\rangle$, comme illustré dans la figure 2.2. L'expression de la fonction de prédiction

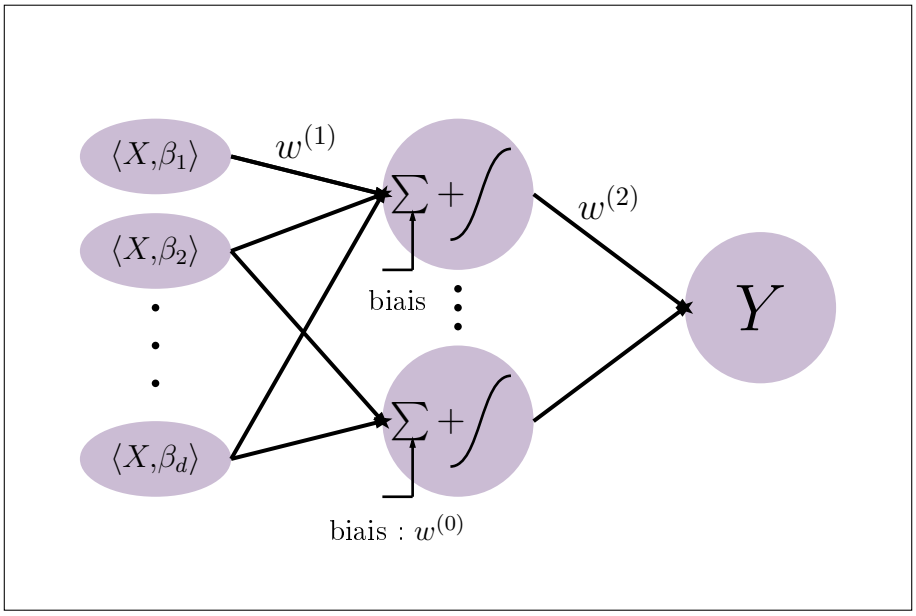

FIGURE 2.2 – Proposition de modèle de perceptron multi-couches fonctionnel dans le cas d'une approche par régression inverse.

est alors

$$\hat{F}_w(X) = \sum_{k=1}^{q} w_k^{(2)} G\left(\sum_{j=1}^{d} w_{kj}^{(1)} \langle X,\beta_j\rangle + w_k^{(0)}\right),$$

où les poids $(w_k^{(p)})_{k,p=0,1,2}$ sont des nombres réels et G est une fonction de lien fixée (la fonction sigmoïde, $G(x) = \frac{1}{1+e^{-x}}$ par exemple). Ces poids sont estimés (pour

q fixé) par minimisation empirique d'une fonction de coût $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$, qui peut, par exemple, être simplement $L(F_w(x),y) = (F_w(x) - y)^2$:

$$w^* := \arg \min_{w \in \mathbb{R}^{q(d+2)}} \sum_{i=1}^{n} L\left(F_w(x_i),y_i\right).$$

En complément, dans (Ferré and Villa 2006), nous donnons des hypothèses suffisantes (portant sur la régularité de L comme fonction de x, w et y) sous lesquelles les poids ainsi estimés convergent vers l'ensemble des poids optimaux minimisant $\mathbb{E}(L(F_w(X),Y))$. Les résultats de (Hornik 1991), qui montrent que les perceptrons multi-couches sont des approximateurs universels, permettent ainsi de conclure à la pertinence de l'estimation du modèle (2.1) par le biais de cette approche.

2.3.2 Régression inverse par estimation de densité (DBIR)

Dans cette partie, nous présentons un travail qui est basé également sur une approche inverse mais qui ne nécessite pas d'hypothèse forte sur la distribution de (X,Y) contrairement à l'hypothèse sous-jacente au modèle (2.1). Cette partie reprend les travaux (Hernández, Biscay, Villa-Vialaneix, and Talavera 2011; Hernández, Biscay, Villa-Vialaneix, and Talavera 2014a). L'approche proposée est une approche non paramétrique qui se place dans le cadre où le modèle inverse

$$X = r(Y) + \epsilon \tag{2.2}$$

(avec $r : \mathbb{R} \longrightarrow \mathcal{H}$ et ϵ est un processus aléatoire centré et indépendant de X) est justifié par le contexte applicatif. C'est notamment le cas en spectrométrie où ce type de modèle, reliant la courbe produite par le spectromètre, X, est expliquée par la quantité d'un des constituants, Y, de l'échantillon analysé. La problématique est bien de déterminer une estimation de la valeur de Y lorsque la valeur x de la variable fonctionnelle X est observée, en utilisant un ensemble d'apprentissage (x_1,y_1), ..., (x_n,y_n) d'observations i.i.d. du couple (X,Y). Cependant, contrairement aux approches classiques, l'estimation sera basée sur des hypothèses et une procédure d'estimation qui concernent le modèle inverse de l'équation (2.2).

Dans ce contexte applicatif, il est, en effet, assez naturel de supposer que ϵ, qui modélise un bruit ou une perturbation du signal enregistré par le spectromètre, suit une distribution gaussienne de moyenne zéro et d'opérateur de variance Ξ. Ainsi, la distribution conditionnelle $\mathcal{L}(X|Y = y)$ est également gaussienne, d'espérance $\mathbb{E}(X|Y)$ et d'opérateur de variance Ξ. En notant $(\lambda_j,\phi_j)_{j\geq 1}$ l'ensemble des valeurs propres et fonctions propres de Ξ (rangés par ordre décroissant de leurs valeurs propres) et en supposant en outre que l'hypothèse de régularité supplémentaire

$$\sum_{j\geq 1} \frac{r_j^2}{\lambda_j} < +\infty,$$

avec $r_j : y \in \mathbb{R} \longrightarrow \langle \phi_j,r(y) \rangle \in \mathbb{R}$, est vérifiée, la densité de la loi $\mathcal{L}(X|Y = y)$ a la formulation explicite suivante :

$$\forall\, x \in \mathcal{H},\, y \in \mathbb{R}, \qquad f(x|y) = \exp\left[\sum_{j\geq 1} \frac{r_j(y)}{\lambda_j}\left(x_j - \frac{r_j(y)}{2}\right)\right] \tag{2.3}$$

où $x_j = \langle x,\phi_j \rangle$. Une procédure d'estimation, basée sur l'estimation préalable de r puis de $f(x|y)$ permet alors de proposer un estimateur de $\mathbb{E}(Y|X=x)$ en utilisant la relation

$$\mathbb{E}(Y|X=x) = \frac{\int f(x|y) f_Y(y) y dy}{f_X(x)}. \tag{2.4}$$

Nous détaillons les étapes de cette estimation ci-dessous dans le cas particulier où $\mathcal{H} = L^2[0,1]$ pour simplifier :

1. **estimation de la fonction de régression $y \in \mathbb{R} \longrightarrow r(y)(t)$, en tout $t \in [0,1]$** : ce problème est un problème classique d'estimation non paramétrique unidimensionnelle. Une approche simple consiste à utiliser pour cela l'estimateur à noyau de Nadaraya-Watson :

$$\hat{r}(y)(t) = \frac{\sum_{i=1}^n K\left(\frac{y_i - y}{h}\right) x_i}{\sum_{i=1}^n K\left(\frac{y_i - y}{h}\right)}$$

   où K est un noyau de lissage (par exemple, $K(x) = e^{-x^2}$) et $h > 0$ est la fenêtre de lissage. Dans le cadre de notre travail et des simulations que nous avons effectuées, nous avons fixé la valeur de h pour toutes les valeurs de t et y ;
2. **décomposition spectrale de l'estimateur de Ξ** : Ξ peut être estimé à partir de l'étape précédente en déterminant des résidus estimés :

$$\forall\, i = 1,\ldots,n, \qquad \hat{\epsilon}_i = y_i - \hat{r}(y_i)$$

   puis $\widehat{\Xi} = \frac{1}{n}\sum_{i=1}^n \hat{\epsilon}_i \otimes \hat{\epsilon}_i$. La décomposition spectrale de cet opérateur est un problème qui ne présente aucune difficulté calculatoire ni théorique et conduit à obtenir les estimateurs $(\hat{\lambda}_j,\hat{\phi}_j)_{j=1,\ldots,p}$ des $p = p(n)$ premiers valeurs propres et vecteurs propres (par ordre décroissant des valeurs propres) ;
3. **estimation de $f(x|y)$** : en utilisant l'expression explicite de l'équation (2.3), on peut alors proposer l'estimateur suivant :

$$\hat{f}(x|y) = \exp\left[\sum_{j=1}^p \frac{\hat{r}_j(y)}{\hat{\lambda}_j}\left(\hat{x}_j - \frac{\hat{r}_j(y)}{2}\right)\right]$$

   où $\hat{r}_j(y) = \langle r(y),\hat{\phi}_j \rangle$ et $\hat{x}_j = \langle x,\hat{\phi}_j \rangle$ et, plus précisément, calculer $\hat{f}(x|y_i)$ pour tout $i = 1,\ldots,n$ au point d'intérêt observé x ;
4. en déduire **l'estimateur « plug-in » de** $\mathbb{E}(Y|X=x)$ en utilisant la relation de l'équation (2.4) :

$$\hat{y}(x) = \frac{\frac{1}{n}\sum_{i=1}^n \hat{f}(x|y_i) y_i}{\hat{f}_X(x)}.$$

Dans (Hernández, Biscay, Villa-Vialaneix, and Talavera 2014a), nous démontrons la consistance de cette approche, c'est-à-dire, la convergence en probabilité de $\hat{y}(x)$ vers $\mathbb{E}(Y|X=x)$ pour tout x tel que $f_X(x) \neq 0$. Ce résultat est basé sur des hypothèses qui concernent la régularité des densités et fonctions de régression impliquées, la vitesse de convergence vers 0 des valeurs propres λ_j et la vitesse de convergence vers $+\infty$ de $p = p(n)$.

2.4 Méthodes à noyau pour la discrimination

La section précédente présente des travaux qui s'appuient sur des modèles inverses. Toutefois, ce type d'approches nécessitent des hypothèses assez fortes portant, notamment, sur la dispersion des valeurs propres de l'opérateur de variance Γ_X ou de l'opérateur Ξ de la section 2.3.2, pour assurer leur consistance. Ces hypothèses peuvent être assez difficiles à vérifier en pratique.

Dans une approche plus proche de celle de l'apprentissage statistique (Vapnik 1998; Devroye et al. 1996), nous nous sommes ensuite intéressés à développer des approches permettant d'obtenir des résultats de *consistance universelle*, c'est-à-dire, qui ne requièrent pas ou peu d'hypothèses sur la distribution du couple (X,Y). De manière plus précise, dans (Rossi and Villa 2006), nous étendons le cadre des SVM pour la discrimination binaire au cas où les variables explicatives sont fonctionnelles. Plus tard, dans (Rossi and Villa-Vialaneix 2011a), nous proposons une approche consistante permettant de prendre en compte la discrétisation des variables fonctionnelles explicatives et de leur appliquer une transformation fonctionnelle, basée sur la dérivée. Cette méthode est proche de celle présentée dans (Rossi and Villa 2006), dans le sens où elle s'appuie sur le cadre commun des espaces de Hilbert à Noyau Reproduisant (RKHS, voir (Berlinet and Thomas-Agnan 2004)). Les travaux connexes à ce travail sont, par exemple, les articles (Preda 2007; Hernández, Biscay, and Talavera 2007) qui abordent le problème de la régression fonctionnelle par SVR (Support Vector Regression, qui sont une extension des SVM au cadre de la régression). Plus récemment, (Kadri, Rabaoui, et al. 2011; Kadri, Rakotomamonjy, et al. 2012) ont abordé le problème de la régression et de la classification fonctionnelles lorsque la variable *à expliquer* est fonctionnelle et ont utilisé une approche par noyau à valeur opérateur qui permet, dans sa version multiple, de combiner des variables explicatives multiples, éventuellement de natures différentes (fonctionnelles, discrètes, continues).

2.4.1 SVM pour la discrimination fonctionnelle

Dans ce domaine, notre article (Rossi and Villa 2006) a été précurseur : dans ce travail, connaissant un ensemble d'apprentissage $(x_i,y_i)_{i=1,\dots,n}$ issu d'un couple de variables aléatoires (X,Y) à valeur dans $\mathcal{H} \times \{-1,1\}$ où $\mathcal{H}$ est un espace de Hilbert, nous proposons la résolution du problème quadratique

$$\begin{aligned} \max_{\beta\in\mathbb{R}^n} \quad & \sum_{i=1}^n \beta_i - \sum_{i,j=1}^n \beta_i\beta_j y_i y_j K(x_i,x_j) \\ \text{tels que} \quad & \sum_{i=1}^n \beta_i y_i = 0, \\ & 0 \leq \beta_i \leq C, \forall\, i = 1,\dots,n \end{aligned} \tag{2.5}$$

où $K : \mathcal{H} \times \mathcal{H} \longrightarrow \mathbb{R}$ est un noyau sur $\mathcal{H}$ (voir section 1.2.2 pour une description plus précise des noyaux). Ce problème est le *problème dual* d'un problème d'optimisation dans l'espace image, $\mathcal{X}$, associé au noyau K dans lequel une discrimination linéaire à marge optimale est construite. Cette approche correspond donc au cadre classique des SVMs multi-dimensionnels et nos apports dans ces travaux concernent les points suivants :

- nous décrivons **quels types de noyaux peuvent être utilisés lorsque X est une fonction**. Certains de ces noyaux sont spécifiques au cadre fonctionnel et utilisent des pré-traitements fonctionnels (dérivée, projection basée sur une ACP fonctionnelle...). L'universalité de ces noyaux (c'est-à-dire, leur capacité à « couvrir » l'espace image), n'est toutefois pas démontrée ;

- nous proposons une **approche par validation permettant de sélectionner un noyau**, un ou des paramètres associés en procédant par projection des entrées fonctionnelles dans des espaces de dimension de plus en plus grande. Le choix des divers paramètres et de la dimension optimale de projection se fait par une procédure de validation qui pénalise plus fortement les projections sur des espaces de grande dimension. Cette approche est décrite dans l'algorithme 5 ;

Algorithme 5 SVM pour données fonctionnelles

Requis $(\Psi_j)_{j\geq 1}$, base Hilbertienne de $\mathcal{H}$
Requis $\forall\, d \in \mathbb{N}$, $\mathcal{I}_d$, ensemble fini de noyaux sur $\mathbb{R}^d$
Requis $\forall\, d \in \mathbb{N}$, $C_d > 1$
Requis $\forall\, d \in \mathbb{N}$, $\lambda_d > 0$ tel que $\sum_d |\mathcal{I}_d| e^{-2\lambda_d^2} < +\infty$
1: Partage des données en $\mathcal{D}_1 = \{(x_i,y_i)\}_{i=1,\dots,l}$ (apprentissage) et $\mathcal{D}_2 = \{(x_i,y_i)\}_{i=l+1,\dots,n}$ (validation)
2: **Pour** $a = (d,K_d,C) \in \mathbb{N} \times \mathcal{I}_d \times [0,C_d]$ **Faire**
3: $\quad \mathcal{P}_{V_d}$ est l'opérateur de projection de $\mathcal{H}$ sur l'espace engendré par $(\Psi_j)_{j=1\dots,d}$
4: $\quad$ Résolution du problème (2.5) avec le noyau $K = K_d \circ \mathcal{P}_{V_d}$, sur les données $\mathcal{D}_1$
Résultat : $F_a = \operatorname{sign}\left(\sum_{i=1}^{l} \beta_i K(x_i,.)\right)$
5: $\quad$ Calcul de l'erreur de validation sur $\mathcal{D}_2$:

$$\widehat{L}_{n-l} F_a = \frac{1}{n-l} \sum_{i=l+1}^{l} \mathbb{I}_{\{F_a(x_i) \neq y_i\}}$$

6: **Fin Pour**
7: Choix de la discrimination optimale :

$$a^* := \arg\min_a \widehat{L}_{n-l} F_a + \frac{\lambda_d}{\sqrt{n-l}}$$

Résultat : $\widehat{F}^n := F_{a^*}$

- nous démontrons la **consistance universelle** de cette approche, c'est-à-dire, le fait que sous des hypothèses très réduites (qui concernent essentiellement le fait que l'ensemble $\mathcal{I}_d$ contient au moins un noyau universel dans $\mathbb{R}^d$ et des conditions sur les vitesses de convergence respectives de n et $n-l$ vers $+\infty$), l'erreur de classification associée à la fonction de discrimination $\widehat{F}^n$,

$$L\widehat{F}^n = \mathbb{P}(\widehat{F}^n(X) \neq Y)$$

converge vers l'erreur de classification optimale pour le couple (X,Y)

$$\mathbb{P}(F^*(X) \neq Y) \qquad \text{avec } F^*(x) = \begin{cases} 1 & \text{si } \mathbb{P}(Y=1|X=x) > 1/2 \\ -1 & \text{sinon.} \end{cases}.$$

2.4.2 Utiliser les dérivées

En pratique, comme mentionné dans la section précédente, il est courant d'utiliser la dérivée en pré-traitement de données fonctionnelles, pour s'intéresser aux inflexions de la fonction plutôt que directement à ses valeurs. Toutefois, peu de travaux se sont intéressés aux performances asymptotiques d'un tel pré-traitement. Dans le cadre de notre

étude sur l'utilisation des méthodes à noyau pour les données fonctionnelles, nous avons proposé, dans (Villa and Rossi 2006b), l'utilisation d'un noyau basé sur une approche spline qui incorpore le pré-traitement par dérivées et avons démontré sa consistance universelle dans le cadre de la discrimination binaire. Ce travail est généralisé dans (Rossi and Villa-Vialaneix 2011a) où nous montrons qu'une approche similaire peut être combinée avec des méthodes de classification et de régression très générales : cette approche permet de contrôler à la fois la perte d'information due à l'utilisation, comme prédicteur, de la dérivée plutôt que de la fonction d'origine. Également, elle prend en compte le fait que les données ne sont pas observées directement mais au travers d'une discrétisation. De manière plus précise, comme illustré dans la figure 2.3 nous montrons qu'il est pos-

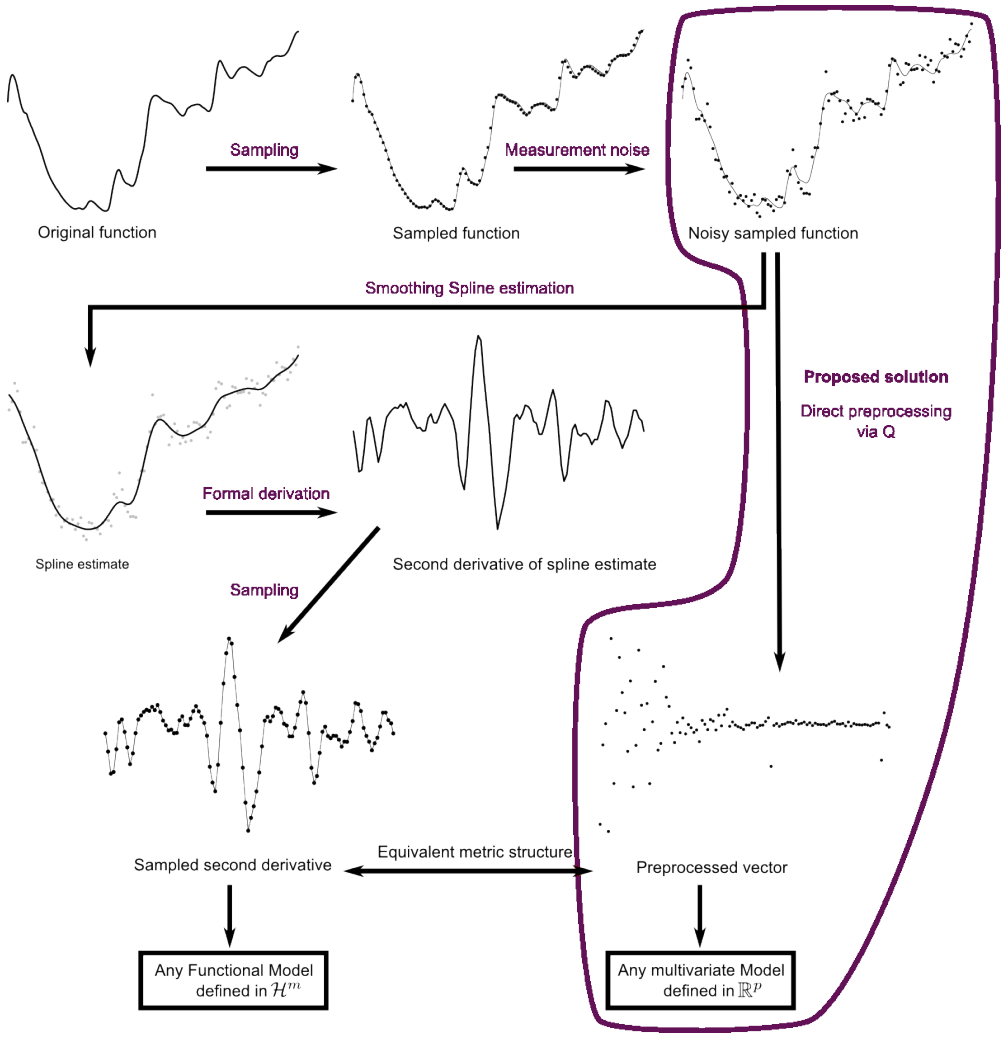

FIGURE 2.3 – Schéma de la méthodologie proposée dans l'article (Rossi and Villa-Vialaneix 2011a) pour l'utilisation des dérivées dans les modèles fonctionnels : l'approche habituelle est schématisée par le processus à gauche de la figure et l'approche proposée est entourée sur la partie droite de la figure.

sible de définir une transformation linéaire des observations fonctionnelles discrétisées qui, combinée avec une méthode de régression ou de discrimination multidimensionnelle classique, soit équivalente à une estimation spline des fonctions puis une régression ou une discrimination fonctionnelle sur les dérivées.

En effet, en utilisant les travaux de (Wahba 1990), on montre que, pour des fonctions à valeurs dans l'espace $\mathcal{H}^m$ ($m \geq 1$), ensemble des fonctions m fois dérivables sur [0,1], muni du produit scalaire

$$\forall\, x,x' \in \mathcal{H}^m, \qquad \langle x,x' \rangle_{\mathcal{H}^m} = \int_0^1 x^{(m)}(t)(x')^{(m)}(t)\, dt + \sum_{k=0}^{m-1} x^{(k)}(0)(x')^{(k)}(0),$$

et pour un ensemble de points $\tau_d = \{t_1, \ldots, t_d\}$ dans [0,1], il est possible de définir une matrice $\mathbf{Q}_{\lambda,\tau_d}$ (calculable explicitement), telle que

$$(\mathbf{x}_{\tau_d})^T \left(\mathbf{Q}_{\lambda,\tau_d}\right)^T \mathbf{Q}_{\lambda,\tau_d} \mathbf{x}'_{\tau_d} = \langle \hat{x}_{\lambda,\tau_d}, \hat{x}'_{\lambda,\tau_d} \rangle_{\mathcal{H}^m} \tag{2.6}$$

où

- $\mathbf{x}_{\tau_d}$ désigne les observations de la fonction x aux points de τ_d : $\mathbf{x}_{\tau_d}$ est donc le vecteur de $\mathbb{R}^d$ $(x(t))_{t\in\tau_d}$;
- $\hat{x}_{\lambda,\tau_d}$ désigne l'estimation spline de x aux points de τ_d et avec le paramètre de régularisation $\lambda > 0$: $\hat{x}_{\lambda,\tau_d}$ est donc la solution dans $\mathcal{H}^m$ du problème d'optimisation

$$\arg\min_{h\in\mathcal{H}^m} \frac{1}{|\tau_d|} \sum_{t\in\tau_d} (x(t_d) - h(t_d))^2 + \lambda \int_0^1 \left(h^{(m)}(t)\right)^2 dt.$$

L'équation (2.6) fait donc le lien entre observations d'une discrétisation des fonctions et la dérivée de leurs estimées par des splines de lissage. Ainsi, remplaçant le problème de régression ou de discrimination fonctionnel basé sur les observations $(x_i,y_i)_{i=1,\ldots,n}$ par un problème de régression ou de discrimination multi-dimensionnel (dans $\mathbb{R}^{|\tau_d|}$) basé sur $(\mathbf{Q}_{\lambda,\tau_d}(\mathbf{x}_i)_{\tau_d},y_i)_{i=1,\ldots,n}$, nous obtenons l'estimateur F_{n,τ_d} de la fonction de régression ou de la fonction de discrimination. Sous des hypothèses relatives aux points de τ_d, à la vitesse de convergence de $|\tau_d|$ et λ vers, respectivement $+\infty$ et 0, ainsi que sous des hypothèses peu restrictives sur les variables X et Y, nous montrons la consistance universelle de cette approche, c'est à dire

$$\lim_{d\to+\infty} \lim_{n\to+\infty} LF_{n,\tau_d} = L^*$$

où

- *dans le cadre de la discrimination* ($Y \in \{-1,1\}$), LF_{n,τ_d} est l'erreur associée à F_{n,τ_d}, $\mathbb{P}\left(F_{n,\tau_d}(\mathbf{X}^{\tau_d}) \neq Y\right)$, et L^* est l'erreur optimale $\mathbb{P}\left(F^*(X) \neq Y\right)$ associée à la fonction de discrimination $F^*(x) = \begin{cases} 1 & \text{si } \mathbb{P}(Y=1|X=x) > 1/2 \\ -1 & \text{sinon.} \end{cases}$;
- *dans le cadre de la régression* ($Y \in \mathbb{R}$), LF_{n,τ_d} est l'erreur quadratique associée à F_{n,τ_d} $\mathbb{E}\left([F_{n,\tau_d}(\mathbf{X}^{\tau_d}) - Y]^2\right)$ et L^* est l'erreur quadratique optimale $\min_{F^*:\mathcal{H}^m\to\mathbb{R}} \mathbb{E}\left([F^*(X) - Y]^2\right)$.

Utilisant les résultats de (Steinwart and Christmann 2008), nous avons montré comment cette approche peut être utilisée dans le cadre d'une régression fonctionnelle par SVR (Support Vector Regression). Dans des applications sur données réelles et données bruitées, nous avons montré que cette approche s'avérait particulièrement utile dans le cas de données bruitées.

2.5 Conclusion et perspectives

L'analyse des données fonctionnelles est une thématique de recherche sur laquelle je ne suis plus réellement active, particulièrement en ce qui concerne les aspects les

plus théoriques de mes travaux passés sur le sujet. Toutefois, cette thématique n'est pas totalement déconnectée de mes préoccupations actuelles et ce pour deux raisons principales :

1. la première est que, comme les graphes, les données fonctionnelles sont des données qui diffèrent des données multi-dimensionnelles standard. Dans mes travaux passés, j'ai opté pour des approches liées, notamment, à l'utilisation de noyaux et donc, d'un point de vue méthodologique, ces approches rejoignent les travaux que j'ai présentés dans la section 1.2.2 pour l'analyse et la représentation de graphes. La formation et l'expertise que j'ai acquises sur les méthodes d'apprentissage à noyau au travers de l'analyse des données fonctionnelles a déjà été mise à profit dans le cadre de données non vectorielles plus générales et ces compétences me sont utiles dans mon projet de recherche qui est centré sur l'analyse et l'intégration de données non vectorielles ;
2. la seconde est que dans le domaine d'application dans lequel mon intégration au sein de l'INRA m'oriente, ce type de données est très présent : le cadre fonctionnel est, en effet, assez naturel dans l'analyse de certaines données issues de la technologie haut débit, comme les données métabolomiques collectées par RMN (Résonance Magnétique Nucléaire, voir la figure 2.4 pour un exemple). Dans de précédents travaux, à visée applicative et réalisés en collaboration avec des collègues de l'INRA principalement, nous avons étudié comment la combinaison de pré-traitements basés sur des ondelettes avec des méthodes d'apprentissage linéaires ou non linéaires permettaient d'améliorer la qualité de prédiction d'un phénotype donné à partir d'un spectre RMN (Rohart, Paris, et al. 2012) ou permettait d'identifier des métabolites d'intérêt à partir d'un modèle prédictif en utilisant, de manière similaire à (Poggi and Tuleau 2006; Genuer et al. 2010), l'importance des variables telle que définie par (Breiman 2001) pour la sélection (Villa-Vialaneix, Hernandez, et al. 2014). De manière moins directe, les données collectées lors des gros projets de génomique sont conçus selon des plans d'expérience de plus en plus complexes : comme je l'ai déjà évoqué dans la section 1.2.5, certains projets intègrent notamment l'étude d'une dynamique d'évolution liée à un phénomène extérieur d'intérêt (injection d'une substance, perturbation externe, vieillissement...). Ce type de données peut être relié à l'analyse de données fonctionnelles, dans l'esprit de l'article (Déjean et al. 2007). Ainsi, mes connaissances des données fonctionnelles et de manière plus générale, des données de grande dimension ou des données non vec-

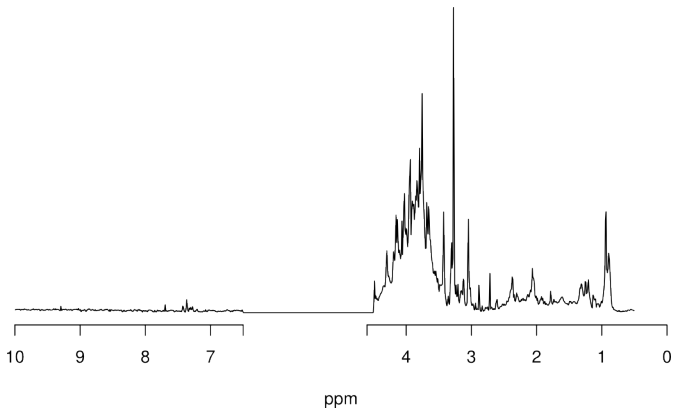

FIGURE 2.4 – Exemple d'un spectre données métabolomiques collectées par RMN, extrait de (Villa-Vialaneix, Hernandez, et al. 2014).

torielles ainsi que mon expertise sur les méthodes d'apprentissage qui peuvent être mise en œuvre pour les analyser devraient s'avérer utile pour la poursuite de mon projet de recherche. En particulier, ces sujets seront abordés dans la thèse de Valérie Sautron (que je co-encadre), qui porte sur l'intégration de données collectées à plusieurs pas de temps après un facteur externe de stress, certaines de ces données étant des données métabolomiques (thèse co-financée par l'ANR via le projet « SusOStress » et par la région Midi-Pyrénées).

2.6 Références

Amato, U., A. Antoniadis, and IK De Feis (2006). « Dimension reduction in functional regression with applications ». In: *Computational Statistics and Data Analysis* 50, pp. 2422–2446. DOI: doi:10.1016/j.csda.2004.12.007.

Berlinet, A. and C. Thomas-Agnan (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Boston, Norwell, MA, USA / Dordrecht, The Netherlands: Kluwer Academic Publisher.

Bosq, D. (1991). « Modelization, non-parametric estimation and prediction for continuous time processes ». In: *Nonparametric functional estimation and related topics, Nato ASI Series C*. Ed. by G. Roussas. Vol. 335. ASI Series. NATO, pp. 509–529.

Breiman, L. (2001). « Random Forests ». In: *Machine Learning* 45.1, pp. 5–32. URL: http://www.springerlink.com/content/u0p06167n6173512/fulltext.pdf.

Cardot, H., F. Ferraty, and P Sarda (1999). « Functional Linear Model ». In: *Statistics and Probability Letters* 45, pp. 11–22. URL: http://math.u-bourgogne.fr/IMB/cardot/CFS99.pdf.

Cardot, H., F. Ferraty, and P. Sarda (2003). « Spline estimators for the functional linear model ». In: *Statistica Sinica* 13, pp. 571–591. URL: http://www.inra.fr/Internet/Departements/MIA/T/cardot/Doc/CFSsinica.ps.gz.

Chen, T. and H. Chen (1995). « Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems ». In: *IEEE Transactions on Neural Networks* 6.4, pp. 911–917.

Dauxois, J., L. Ferré, and A.F. Yao (2001). « Un modèle semi-paramétrique pour variable aléatoire hilbertienne ». In: *Comptes Rendus Mathématique. Académie des Sciences. Paris* 327.I, pp. 947–952. DOI: doi:10.1016/S0764-4442(01)02163-2. URL: http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6VJ2-44GDVWJ-B&_user=722937&_rdoc=1&_fmt=&_orig=search&_sort=d&view=c&_acct=C000040378&_version=1&_urlVersion=0&_userid=722937&md5=4484eaa7a72c394bcbd6428ce3a5e1a5.

Déjean, S., P.G.P. Martin, A. Baccini, and P. Besse (2007). « Clustering time-series gene expression data using smoothing spline derivatives ». In: *EURASIP Journal on Bioinformatics and Systems Biology*, p. ID 70561. DOI: 10.1155/2007/70561.

Devroye, L., L. Györfi, and G. Lugosi (1996). *A Probabilistic Theory for Pattern Recognition*. New York: Springer-Verlag.

Ferraty, F. and P. Vieu (2006). *NonParametric Functional Data Analysis*. Springer.

Ferré, L. and N. Villa (2005). « Discrimination de courbes par régression inverse fonctionnelle ». In: *Revue de Statistique Appliquée* LIII.1, pp. 39–57. URL: http://www.numdam.org/numdam-bin/fitem?id=RSA_2005__53_1_39_0.

— (2006). « Multi-layer perceptron with functional inputs: an inverse regression approach ». In: *Scandinavian Journal of Statistics* 33.4, pp. 807–823. DOI: doi:10.1111/j.1467-9469.2006.00496.x.

Ferré, L. and A.F. Yao (2003). « Functional sliced inverse regression analysis ». In: *Statistics* 37.6, pp. 475–488.

— (2005). « Smoothed functional inverse regression ». In: *Statistica Sinica* 15.3, pp. 665–683. URL: http://www3.stat.sinica.edu.tw/statistica/J15N3/J15N35/J15N35.html.

Genuer, R., J.M. Poggi, and C. Tuleau-Malot (2010). « Variable selection using random forests ». In: *Pattern Recognition Letters* 31, pp. 2225–2236.

Györfi, L., M. Kohler, A. Krzyżak, and H. Walk (2002). *A Distribution-Free Theory of Nonparametric Regression*. New York: Springer.

Hernández, N., R.J. Biscay, and I. Talavera (2007). « Progress in Pattern Recognition, Image Analysis and Applications (Proceedings of 12th Iberoamericann Congress on Pattern Recognition, CIARP 2007, Valparaiso, Chile, November 13-16, 2007) ». In: ed. by L. Rueda, D. Mery, and J. Kittler. Vol. 4756. Lecture Notes in Computer Science. Berlin/Heidelberg: Springer. Chap. Support vector regression methods for functional data, pp. 564–573.

Hernández, N., R.J. Biscay, N. Villa-Vialaneix, and I. Talavera (2011). « A simulation study of functional density-based inverse regression ». In: *Revista Investigacion Operacional* 32.2, pp. 146–159. URL: http://rev-inv-ope.univ-paris1.fr/files/32211/32211-06.pdf.

— (2014a). « A non parametric approach for calibration with functional data ». In: *Statistica Sinica*. Forthcoming.

Hornik, K. (1991). « Approximation capabilities of multilayer feedfoward networks ». In: *Neural Networks* 4.2, pp. 251–257.

Kadri, H., A. Rabaoui, P. Preux, E. Duflos, and A. Rakotomamonjy (2011). « Functional regularized least squares classiffication with operator-valued kernels ». In: *Proceedings of International Conference on Machine Learning (ICML)*. Seattle, USA.

Kadri, H., A. Rakotomamonjy, F. Bach, and P. Preux (2012). « Multiple operator-valued kernel learning ». In: *Advances in Neural Information Processing Systems 25 (NIPS)*. Lake Tahoe, Nevada, USA.

Li, K.C. (1991). « Sliced inverse regression for dimension reduction ». In: *Journal of the American Statistical Association* 86.414, pp. 316–342.

Poggi, J.M. and C. Tuleau (2006). « Classification supervisée en grande dimension. Application à l'agrément de conduite automobile ». In: *Revue de Statistique Appliquée* LIV.4, pp. 41–60.

Preda, C. (2007). « Regression models for functional data by reproducing kernel Hilbert spaces methods ». In: *Journal of Statistical Planning and Inference* 137.3, pp. 829–840. DOI: doi:10.1016/j.jspi.2006.06.011.

Ramsay, J.O. and B.W. Silverman (1997). *Functional Data Analysis*. New York: Springer Verlag.

— (2002). *Applied Functional Data Analysis*. Springer Verlag.

Rohart, F., A. Paris, B. Laurent, C. Canlet, J. Molina, M.J. Mercat, T. Tribout, N. Muller, N. Iannuccelli, N. Villa-Vialaneix, L. Liaubet, D. Milan, and M. San Cristobal (2012). « Phenotypic prediction based on metabolomic data on the growing pig from three main European breeds ». In: *Journal of Animal Science* 90.12. DOI: 10.2527/jas.2012-5338.

Rossi, F. and B. Conan-Guez (2005). « Functional multi-layer perceptron: a nonlinear tool for functional data anlysis ». In: *Neural Networks* 18.1, pp. 45–60.

Rossi, F. and N. Villa (2006). « Support vector machine for functional data classification ». In: *Neurocomputing* 69.7-9, pp. 730–742. DOI: 10.1016/j.neucom.2005.12.010.

Rossi, F. and N. Villa-Vialaneix (2011a). « Consistency of functional learning methods based on derivatives ». In: *Pattern Recognition Letters* 32.8, pp. 1197–1209. DOI: 10.1016/j.patrec.2011.03.001.

Sandberg, I.W. and L. Xu (1996). « Network approximation of input-output maps and functionals ». In: *Circuits Systems Signal Processing* 15.6, pp. 711–725.

Steinwart, I. and A. Christmann (2008). *Support Vector Machines*. Information Science and Statistics. Springer.

Vapnik, V. (1998). *Statistical Learning Theory*. New York, USA: Wiley.

Villa, N. and F. Rossi (2006b). « Un résultat de consistance pour des SVM fonctionnels par interpolation spline ». In: *Comptes Rendus Mathématique. Académie des Sciences. Paris* 343.8, pp. 555–560. DOI: 10.1016/j.crma.2006.09.025.

Villa-Vialaneix, N., N. Hernandez, A. Paris, C. Domange, N. Priymenko, and P. Besse (2014). « On combining wavelets expansion and sparse linear models for regression on metabolomic data and biomarker selection ». In: *Communications in Statistics - Simulation and Computation*. Forthcoming. DOI: 10.1080/03610918.2013.862273.

Wahba, G. (1990). *Spline Models for Observational Data*. Philadelphia, Pennsylvania, USA: Society for Industrial and Applied Mathematics.

Conclusion générale

Les chapitres précédents résument l'évolution de mes thématiques de recherche : durant ma thèse et les premières années de ma carrière d'enseignante chercheuse, mes centres d'intérêt étaient plutôt tournés vers l'étude des propriétés théoriques de consistance de méthodes supervisées pour l'analyse de données fonctionnelles. Mon interaction avec mon environnement de recherche (dans une université de sciences humaines et sociales) et l'orientation initiale de mes thématiques de recherche vers l'apprentissage et les données non vectorielles ont fait petit à petit évoluer mon intérêt vers des préoccupations plus appliquées en fouille de données et en apprentissage pour des données non vectorielles plus générales, des graphes en particulier. Ces nouvelles préoccupations ont trouvé un cadre d'application riche en biologie, dans les études génétiques et de biologie des systèmes en particulier et j'ai présenté dans les chapitres 1.2.5, 1.3.4 et 2.5 mes perspectives et projets de recherche en relation avec mon intégration récente au sein du département MIA de l'INRA.

Pour résumer, les divers aspects de mes préoccupations futures s'articulent principalement autour de

l'intégration de données collectées à plusieurs niveaux de l'échelle du vivant (données métabolomiques, données transriptomiques, phénotypes...) et qui peuvent être de types multiples, non nécessairement vectoriels (données fonctionnelles, graphes, ...). Cette question pourra être abordée sous l'angle du développement de *méthodes d'exploration et de visualisation* mais aussi sous l'angle de l'*inférence de réseaux*. Dans tous les cas, il s'agit de trouver des stratégies adaptées, d'une part pour tenir compte de la nature propre (non vectorielle) de chacune des données et, d'autre part, pour combiner de manière équilibrée les informations apportées par les différents types de données et/ou éventuellement pour sélectionner certaines de ces données. D'un point de vue applicatif, cette thématique trouve des débouchées dans plusieurs domaines d'application en biologie et, actuellement, je cible principalement les *méthodes exploratoires pour l'analyse de données multi-omiques* (en lien avec plusieurs projets de recherche dans lesquels je suis impliquée), l'*annotation fonctionnelle de gènes* par intégration de plusieurs types de données transcriptomiques et de données d'ontologie, par exemple ou, d'un point de vue proche, l'étude de méthodes d'exploration de la *typologie des ARNs non codants*.

Pour aborder cette question, il sera en particulier nécessaire de travailler sur

le passage à l'échelle des méthodes car les données traitées sont souvent de grande dimension (« $n \ll p$ » ou bien n très grand dans le cadre des méthodes à noyau) et diverses stratégies peuvent alors être mise en œuvre : sélection de variables, pénalité parcimonieuse (L^1), techniques de ré-échantillonnage et de « bagging »...

l'interprétabilité des résultats car lorsque les données analysées sont non vectorielles, il convient de repenser les méthodes utilisées habituellement pour leur représentation. Si l'on prend le cas des méthodes basées sur des prototypes (comme les cartes auto-organisatrices), l'intérêt de ces approches dans un cadre exploratoire ou prédictif tient à l'existence de prototypes facilement utilisables pour l'interprétation et qui, dans un cadre multi-dimensionnel, peuvent être visualisés de manière simple. Dans le cadre de l'utilisation de noyaux, les prototypes sont décrits pas des combinaisons convexes des n individus, eux-mêmes caractérisés par une ou des données non vectorielles et la représentation et la compréhension de la signification de ces prototypes est alors en partie perdue.

la prise en compte de divers aspects du protocole expérimental dans l'exploration de ces données et dans l'inférence de réseaux. En particulier, les aspects temporels (en terme de classification de sommets dans un graphe, ou bien d'inférence de réseau) sont des informations à intégrer dans l'analyse, de même que l'appariemment des individus, observées dans plusieurs conditions par exemple.

Ces problématiques sont supportées par divers projets et collaborations (en grande partie résumés sur la figure 1.12), et en particulier (et de manière non exclusive) pour

- les projets financés par l'ANR BIOADAPT « PigHeat » et « SusOStress », qui visent à l'analyse génétique des réponses à la chaleur et au stress chez le porc et à la définition d'un modèle biologique pour ces réponses par intégration de données multi-omiques avec des aspects temporels. La thèse de Valérie Sautron se situe dans le cadre de ce dernier projet ;
- la thèse de Jérôme Mariette qui aborde les aspects d'intégration de données et en particulier les problématiques de passage à l'échelle et d'interprétabilité avec des applications visées dans l'analyse exploratoire de données multi-omiques et de caractérisation de la typologie d'ARN non codants ;
- ma collaboration active avec l'INSERM (Nathalie Viguerie) sur le projet DiOGenes qui vise également à l'intégration de données multi-omiques et temporelles pour l'analyse de la réponse à une diète basse calorie chez des obèses.

Les nombreuses questions soulevées par les projets actuels en biologie des systèmes, qui visent à comprendre des données de très grande dimension, complexes, de natures diverses et souvent bruitées, seront probablement au cœur de l'évolution de mes thématiques dans les prochaines années. Cet objectif s'accompagne d'un effort personnel pour comprendre le domaine d'application, la biologie, qui est un domaine de connaissances en constante évolution [1], au travers, notamment, d'un dialogue avec les biologistes avec lesquels je collabore.

1. Un exemple significatif est celui du « dogme fondamental de la biologie moléculaire », qui mettait l'accent sur la transcription et la traduction, qui a été récemment remis en cause et conduit à un intérêt croissant porté aux régions non codantes : celles-ci pourraient, en effet, avoir un rôle fondamental dans la régulation de la transcription notamment.

A — Bref Curriculum Vitae

de **Nathalie Villa-Vialaneix**, née le 28 juillet 1976 (38 ans), de nationalité française.

A.1 Formation et parcours professionnel

1999 Agrégation de mathématiques

1999/2000 Professeure agrégée stagiaire en lycée

2000/2006 PRAG à l'Université Toulouse II (anciennement Le Mirail)

en parallèle,

- Obtention du DEA de Mathématiques Appliquées de l'Université Toulouse III Paul Sabatier en 2002 (Mention Bien)
- Obtention du doctorat de Mathématiques de l'Université Toulouse Le Mirail le 21 octobre 2005 (Mention Très Honorable)

2006/2014 Maîtresse de Conférences à l'Université de Perpignan Via Domitia, IUT, Département STID (Carcassonne) et membre de l'équipe de recherche SAMM, Université Paris 1

- 2008/2009 : en disponibilité 1 an sur un poste de lecturer à Toulouse School of Economics
- 2012/2013 : en délégation 1 an à l'INRA, Unité MIA-T, Toulouse

2014-... Chargée de Recherche 1ère classe à l'INRA, Unité MIA-T, Toulouse

A.2 Encadrements

A.2.1 Encadrements de stages

Juin/Août 2014 Co-encadrement (avec Nathalie Viguerie, I2MC, INSERM de Toulouse) « Recherche de gènes différentiellement exprimés lors d'un régime basse calorie à partir de données biopuces ».

Juin/Août 2014 Co-encadrement (avec Christine Gaspin, MIA-T, INRA de Toulouse) du stage de Sayma Besbes (M1 Économétrie, Université Toulouse 1) « Alignement, normalisation et analyse de données RNASeq ». Stage dans le cadre du projet « memRNAse » financé par l'ANR.

Mai/Septembre 2013 Co-encadrement (avec Christine Cierco-Ayrolles, MIA-T, INRA de Toulouse) du stage de Florian Brunet (M2 Statistique & Économétrie,

Université Toulouse 1) « Intégration de données pour la classification non supervisée de données biologiques ».

Février/Juillet 2013 Co-encadrement (avec Madalina Olteanu, SAMM, Université Paris 1) du stage de Laura Bendhaïba (Diplôme d'ingénieur GIS, PolyTech'Lille) « Création d'un package R pour des cartes auto-organisatrices ».

Mai/Juin 2013 Co-encadrement (avec Elena Terenina, LGC, INRA de Toulouse) du stage de Arthur Gomez (DUT STID, Université de Perpignan Via Domitia) « Étude de la cinétique des réponses au stress chez le porc ».

Mai/Juin 2012 Co-encadrement (avec Magali San Cristobal et Laurence Liaubet, LGC, INRA de Toulouse) du stage de Nicolas Edwards (DUT STID, Université de Perpignan Via Domitia) « Analyse de données transcriptomiques par réseaux de co-expression génique ».

Avril/Juillet 2012 Co-encadrement (avec Christophe Sibertin-Blanc, IRIT, Université Toulouse 1 dans le cadre d'un projet financé par la Maison des Sciences de l'Homme de Toulouse) du stage de Soraya Popic (Licence SID, Université Toulouse 3 - Paul Sabatier) « Analyse Statistique de résultats de simulations issus de la plateforme SocLab », dans le cadre du projet « SocLab-Stat » financé par la MSH-T (Maison des Sciences de l'Homme de Toulouse).

Juin/Avril 2011 Co-encadrement (avec Taoufiq Dkaki, IRIT, Université Toulouse 2) du stage de Reda Semlal (M1 ENSEEIHT, Informatique & mathématiques appliquées) « Implémentation d'une plateforme d'analyse d'algorithmes de recherche d'information », dans le cadre du projet « Analyse de graphes » financé par FREMIT (Structure Fédérative de Recherche en Mathématiques et Informatique de Toulouse, FR 3424, UPS/CNRS).

Mai/Juin 2010 Co-encadrement (avec Taoufiq Dkaki, IRIT, Université Toulouse 2) du stage de Nicolas Paris (DUT STID, Université de Perpignan Via Domitia) « Création et alimentation d'une base de données en recherche d'information ».

Mai/Juin 2010 Co-encadrement (avec Martin Paegelow, GEODE, Université Toulouse 2) du stage de Thomas Palmer (DUT STID, Université de Perpignan Via Domitia) « Comparaison de méthodes de prédiction de l'occupation des sols ».

Avril/Août 2009 Co-encadrement (avec Magali San Cristobal, LGC, INRA de Toulouse) du stage de Adrien Gamot (M2 Statistique & Économétrie, Université Toulouse 1) « Analyse de grands réseaux biologiques : construction et structure ».

Avril/Juillet 2008 Stage de Xavier Noguera (Licence professionnelle bioinformatique, Université de Perpignan Via Domitia) « Développement de programmes R pour la représentation de graphes avec le logiciel libre Tulip ».

Avril/Juin 2007 Co-encadrement (avec Anne-Ruiz Gazen, Toulouse School of Economics) du stage de Houcine Zaghdoudi (M1 ISMAG, Université Toulouse 2) « Apprentissage pour la prédiction d'orages ».

Mai 2005 Co-encadrement (avec Martin Paegelow, GEODE, Université Toulouse 2) du stage de Jean Mendiboure (Maîtrise MASS, Université Toulouse 2) « Réseaux de neurones pour la prédiction de l'occupation des sols ».

A.2.2 Encadrements de thèses

Jérôme Mariette (2013-...) : thèse co-encadrée avec Christine Gaspin, Unité MIA-T, INRA de Toulouse.

Valérie Sautron (2013-...) : thèse co-encadrée avec Pierre Mormède & Elena Terenina, Unité GenPhySE, INRA de Toulouse dans le cadre du projet ANR « SusOStress ».

A.2.3 Participations à des comités et des jurys de thèse

2014-... Participation au comité de thèse de Valentin Voillet, « Approche intégrative du développement musculaire pour identifier des biomarqueurs précoces de survie néonatale » (encadrée par Laurence Liaubet & Magali San Cristobal), GenPhySE, INRA de Toulouse.

Juin 2013 Membre du jury de la thèse de Joseph El Gemayel, « Modèles de la rationalité des acteurs sociaux » (encadrée par Christophe Sibertin Blanc), IRIT, Université Toulouse 1.

2011/2013 Participation au comité de thèse d'Agnès Bonnet, « Étude de l'expression des gènes au cours des stades précoces de la folliculogenèse ovarienne chez les mammifères de rente (brebis) » (encadrée par Philippe Mulsant), Laboratoire de Génétique Cellulaire (LGC), INRA de Toulouse.

A.3 Contrats de recherche institutionnels et industriels

Projet *Internet 3D* financé par le département GA de l'INRA (2014) : *Réseau de gènes co-exprimés et Interactions géniques nucléaires. Responsable* : Yvette Lahbib-Mansais (GenPhySE, INRA de Toulouse)

Projet *memRNAse* financé par l'ANR Non Thématique 2013 (2014/2017) : *Localisation membranaire de la RNase E : rôle dans la mâturation, la surveillance et la dégradation des ARNm. Responsable* : Agamemnon Carpousis (LMGM, Université Toulouse III)

Projet financé par EDF (nov 2013/janv 2014) : *Création d'une interface* **shiny** *pour l'analyse de modèles méchanistes. Responsable de ce projet*

Projet *SUSoSTRESS* financé par l'ANR BIOADAPT 2012 (2013/2016) : *Génétique moléculaire de la réponse au stress et robustesse chez le Porc. Responsable* : Pierre Mormède (GenPhySE, INRA de Toulouse)

Projet *PigHeat* financé par l'ANR BIOADAPT 2012 (2013/2016) : *Adaptation des porcs à la chaleur : la voie génétique. Responsable* : David Renaudeau (UZR, INRA Antilles-Guyane)

Projet *SocLab-Stat* financé par la MSH de Toulouse (2012) : *Développement de méthodes d'analyse statistique de sortie d'un modèle de simulations agent. Responsable* : Christophe Sibertin Blanc (IRIT, Université Toulouse 1)

Projet *ModULand* financé par l'ANR Non Thématique 2011 (2012/2015) : *Usage des sols : modèles, dynamique et décisions. Responsable* : Christine Thomas-Agnan (Toulouse School of Economics, France)

Projet financé par le PEPII 2011 du CNRS (2012/2013) : *Couplage de Méthodes et Modèles pour la Biologie Intégrative. Responsable* : Marc Bailly-Bechet (Laboratoire de Biométrie et Biologie Évolutive, Université Lyon 1)

Projet IMRIT financé par FREMIT (Institut Fédératif de Recherche de Mathématiques et Informatique de Toulouse) (2007/2010) : *Comparaison de graphes et visualisation d'évolution. Responsable de ce projet*

Projet financé par le Central Institut for Meteorology and Geodynamics (Vienne, Autriche) pour le compte d'EUMETSAT (sept 2006/dec 2007) : *Refonte de l'algorithme de discrimination PGE11. Responsable* : Anne Ruiz-Gazen (Toulouse School of Economics, France)

Projet Graphes-Comp financé par l'ANR Non Thématique 2005 (2006/2008) : *Comparaison de grands graphes : Application à la recherche de réseaux de sociabilités paysannes au Moyen-Age. Responsable* : Bertrand Jouve

(Université Toulouse 2, Jean Jaurès, France)

Projet M05AH4 financé par ECOS-Nord Mexique (2005/2009) : *Modélisations prospectives de l'occupation du sol par approches géomatique et statistique. Responsables* : Martin Paegelow (Université Toulouse 2, Jean Jaurès, France), Jean-François Mas (Institut de Géographie, Université Nationale Autonome du Mexique, Mexique)

Projet BIA2003-01499 financé par la FEDER (2004/2006) : *Sistemas de Información Geográfica y modelización de la dinámica paisajística de la montaña mediterránea : Sierra Nevada y Pirineos Orientales franceses. Responsable* : Maria Teresa Camacho-Olmedo (Université de Grenade, Espagne)

Programme PEVS (2001/2003) : *Dynamique et modélisation d'anthroposystèmes montagnards méditerranéens : réchauffement climatique et scenarii environnementaux* (programme piloté par le comité scientifique MOdélisation, Transfert d'Informations, Valorisation pour l'Environnement du CNRS). *Responsable* : Martin Paegelow (Université Toulouse 2, Jean Jaurès, France)

A.4 Activités d'animation scientifique

Membre du comité de programme de manifestations scientifiques ASHS 2008, 2010, 2011 et 2012 ; MARAMI 2010, 2011, 2012 et 2013, FGG 2010, 2011, 2012, 2013 et 2014 ; ICANN 2013 et 2014 ; WSOM 2014 ; SFC 2014 ; JFRB 2014

Organisation de manifestations scientifiques Membre des comités d'organisation de : 3èmes Journées MASH (2006, Université Toulouse 2, Jean Jaures), MARAMI 2010 (Université Toulouse 3, Paul Sabatier), 11ème forum des jeunes mathématiciennes (2001, Université Toulouse 3, Paul Sabatier), MASHS 2009, 2011, 2013 & 2014 (Université Toulouse 2, Jean Jaures, Université Aix-Marseille & Université Paris 1), Ateliers du GDR MASCOT NUM 2013 & 2014 (IHP, Paris), 1/2 Journées satellites STID aux Journées de Statistique de la SFdS 2011, 2012 & 2013, Journées de la Société Française de Statistique (2013, Toulouse Business School, Toulouse), JFRB 2014 (IHP, Paris)

Rapports d'expertise scientifique Computational Statistics and Data Analysis, Information Sciences, Information Retrieval, Journal de la Société Française de Statistique, Neuro computing, Bernouilli, Electronic Journal of Applied Statistical Analysis, Neural Networks, Environmental Modelling and Software, IEEE Transactions on Neural Networks and Learning Systems, Pattern Recognition Letters, BMC Bioinformatics, Communication in Statistics, Computational Statistics... pour l'éditeur Taylor and Francis et pour l'ANR (appels à projets « Émergence » et « CONTINT »)

Membre du comité éditorial des revues « Statistique & Enseignement » (2012/2014, rédactrice en chef adjointe) et « CSBIGS » (2014/...)

Membre de groupes & associations scientifiques Membre du conseil (2013/...) et webmistress (2011/...) de la Société Française de Statistique ; Membre du bureau (2011/2014) du GDR MASCOT NUM ; Co-animatrice (2014/...) du réseau méthodologique MIA de l'INRA « NETBIO »

Membre titulaire élue du CNU 26ème section (2011/...)

A.5 Activités d'enseignement

Formation professionnelle et écoles chercheur

- Summer school on « Network Analysis and Applications » (21 juin/5 juillet 2014), School for advanced

sciences of Luchon (France). Intervention sur « Mining co-expression network » (26 juin)

- Formation professionnelle à la Statistique, niveau 3, organisée à l'INRA de Toulouse en 2012 (2 sessions) et 2013 (1 session). Cours pris en charge : « Machine Learning » (2012 uniquement) et « An introduction to network inference and mining » (2012 et 2013)
- ESSA Summer School 2012, Université Toulouse 1, Capitole, Toulouse : organisation d'un tutoriel « Using R to analyze simulation outputs »

2006/2014 MCF à l'Université de Perpignan Via Domitia, Département STID de l'IUT : cours de statistique et d'informatique en DUT STID et en licences professionnelles bioinformatique et biostatistique ; suivi de projets en statistique et informatique (bases de données, programmation web) $\simeq$ 192 *heures/an*

2008/2009 Disponibilité sur un poste de lecturer, Université Toulouse 1, Capitole : cours de statistique en licence sciences économiques, en master 2 Statistique & Économétrie et en master 2 Actuariat *192 heures*

2000/2006 PRAG à l'Université Toulouse 2, Jean Jaurès : cours de mathématiques et statistique en Licence MASS, en IUT d'informatique, à l'IUP NTIE, dans les parcours dits « non spécialistes » (histoire, psychologie...) $\simeq$ 396 *heures/an*

B — Liste des publications

B.1 Publications dans des revues internationales à comité de lecture

Analyse et inférence de graphes

Boulet, R., B. Jouve, F. Rossi, and N. Villa (2008). « Batch kernel SOM and related Laplacian methods for social network analysis ». In: *Neurocomputing* 71.7-9, pp. 1257–1273. DOI: doi:10.1016/j.neucom.2007.12.026.

Rossi, F. and N. Villa-Vialaneix (2010). « Optimizing an organized modularity measure for topographic graph clustering: a deterministic annealing approach ». In: *Neurocomputing* 73.7-9, pp. 1142–1163. DOI: 10.1016/j.neucom.2009.11.023.

Villa-Vialaneix, N., M. Vignes, N. Viguerie, and M. San Cristobal (2014a). « Inferring networks from multiple samples with concensus LASSO ». In: *Quality Technology and Quantitative Management* 11.1, pp. 39–60. URL: http://www.cc.nctu.edu.tw/~qtqm/qtqmpapers/2014V11N1/2014V11N1_F3.pdf.

Olteanu, M. and N. Villa-Vialaneix (2015). « On-line relational and multiple relational SOM ». In: *Neurocomputing* 147. Forthcoming, pp. 15–30. DOI: 10.1016/j.neucom.2013.11.047.

Analyse de données fonctionnelles

Ferré, L. and N. Villa (2006). « Multi-layer perceptron with functional inputs: an inverse regression approach ». In: *Scandinavian Journal of Statistics* 33.4, pp. 807–823. DOI: doi:10.1111/j.1467-9469.2006.00496.x.

Rossi, F. and N. Villa (2006). « Support vector machine for functional data classification ». In: *Neurocomputing* 69.7-9, pp. 730–742. DOI: 10.1016/j.neucom.2005.12.010.

Rossi, F. and N. Villa-Vialaneix (2011a). « Consistency of functional learning methods based on derivatives ». In: *Pattern Recognition Letters* 32.8, pp. 1197–1209. DOI: 10.1016/j.patrec.2011.03.001.

Hernández, N., R.J. Biscay, N. Villa-Vialaneix, and I. Talavera (2014a). « A non parametric approach for calibration with functional data ». In: *Statistica Sinica*. Forthcoming.

Application en SHS et sciences de l'environnement

Ruiz-Gazen, A. and N. Villa (2007). « Storms prediction: logistic regression vs random forest for unbalanced data ». In: *Case Studies in Business, Industry and Government Statistics* 1.2, pp. 91–101. URL: http://www.bentley.edu/centers/sites/www.bentley.edu.centers/files/csbigs/ruiz.pdf.

Villa, N., M. Paëgelow, M.T. Camacho Olmedo, L. Cornez, F. Ferraty, L. Ferré, and P. Sarda (2007). « Various approaches to predicting land cover in mountain areas ». In: *Communication in Statistics - Simulation and Computation* 36.1, pp. 73–86. DOI: 10.1080/03610910601096379.

Villa-Vialaneix, N., M. Follador, M. Ratto, and A. Leip (2012). « A comparison of eight metamodeling techniques for the simulation of N2O fluxes and N leaching from corn crops ». In: *Environmental Modelling and Software* 34, pp. 51–66. DOI: 10.1016/j.envsoft.2011.05.003.

Rossi, F., N. Villa-Vialaneix, and F. Hautefeuille (2013). « Exploration of a large database of French notarial acts with social network methods ». In: *Digital Medievalist* 9. URL: http://www.digitalmedievalist.org/journal/9/villavialaneix/.

Villa-Vialaneix, N., C. Sibertin-Blanc, and P. Roggero (2014). « Statistical exploratory analysis of agent-based simulations in a social context ». In: *Case Studies in Business, Industry and Government Statistics* 5.2, pp. 132–149. URL: http://publications-sfds.fr/index.php/csbigs/article/view/223.

Application en génomique et biologie des systèmes

Rohart, F., A. Paris, B. Laurent, C. Canlet, J. Molina, M.J. Mercat, T. Tribout, N. Muller, N. Iannuccelli, N. Villa-Vialaneix, L. Liaubet, D. Milan, and M. San Cristobal (2012). « Phenotypic prediction based on metabolomic data on the growing pig from three main European breeds ». In: *Journal of Animal Science* 90.12. DOI: 10.2527/jas.2012-5338.

Viguerie, N., E. Montastier, J.J. Maoret, B. Roussel, M. Combes, C. Valle, N. Villa-Vialaneix, J.S. Iacovoni, J.A. Martinez, C. Holst, A. Astrup, H. Vidal, K. Clément, J. Hager, W.H.M. Saris, and D. Langin (2012). « Determinants of human adipose tissue gene expression: impact of diet, sex, metabolic status and cis genetic regulation ». In: *PLoS Genetics* 8.9, e1002959. DOI: 10.1371/journal.pgen.1002959.

Villa-Vialaneix, N., L. Liaubet, T. Laurent, P. Cherel, A. Gamot, and M. San Cristobal (2013). « The structure of a gene co-expression network reveals biological functions underlying eQTLs ». In: *PLoS ONE* 8.4, e60045. DOI: 10.1371/journal.pone.0060045.

Villa-Vialaneix, N., N. Hernandez, A. Paris, C. Domange, N. Priymenko, and P. Besse (2014). « On combining wavelets expansion and sparse linear models for regression on metabolomic data and biomarker selection ». In: *Communications in Statistics - Simulation and Computation*. Forthcoming. DOI: 10.1080/03610918.2013.862273.

B.2 Publications dans des revues nationales à comité de lecture

Analyse et inférence de graphes

Laurent, T. and N. Villa-Vialaneix (2011). « Using spatial indexes for labeled network analysis ». In: *Information, Interaction, Intelligence (I3)* 11.1. URL: http://www.irit.fr/journal-i3/volume11/numero01/.

Rossi, F. and N. Villa-Vialaneix (2011b). « Représentation d'un grand réseau à partir d'une classification hiérarchique de ses sommets ». In: *Journal de la Société Française de Statistique* 152.3, pp. 34–65. URL: http://publications-sfds.math.cnrs.fr/index.php/J-SFdS/article/view/82/73.

Villa-Vialaneix, N., T. Dkaki, S. Gadat, J.M. Inglebert, and Q.D. Truong (2011). « Recherche et représentation de communautés dans un grand graphe : une approche combinée ». In: *Document Numérique* 14.1, pp. 59–80. DOI: 10.3166/dn.14.1.59-80.

Cottrell, M., M. Olteanu, F. Rossi, J. Rynkiewicz, and N. Villa-Vialaneix (2012). « Neural networks for complex data ». In: *Künstliche Intelligenz* 26.2, pp. 1–8. DOI: 10.1007/s13218-012-0207-2.

Villa-Vialaneix, N., B. Jouve, F. Rossi, and F. Hautefeuille (2012). « Spatial correlation in bipartite networks: the impact of the geographical distances on the relations in a corpus of medieval transactions ». In: *Revue des Nouvelles Technologies de l'Information* SHS-1, pp. 97–110. URL: http://www.editions-hermann.fr/ficheproduit.php?prodid=1371.

Analyse de données fonctionnelles

Ferré, L. and N. Villa (2005). « Discrimination de courbes par régression inverse fonctionnelle ». In: *Revue de Statistique Appliquée* LIII.1, pp. 39–57. URL: http://www.numdam.org/numdam-bin/fitem?id=RSA_2005__53_1_39_0.

Villa, N. and F. Rossi (2006b). « Un résultat de consistance pour des SVM fonctionnels par interpolation spline ». In: *Comptes Rendus Mathématique. Académie des Sciences. Paris* 343.8, pp. 555–560. DOI: 10.1016/j.crma.2006.09.025.

Hernández, N., R.J. Biscay, N. Villa-Vialaneix, and I. Talavera (2011). « A simulation study of functional density-based inverse regression ». In: *Revista Investigacion Operacional* 32.2, pp. 146–159. URL: http://rev-inv-ope.univ-paris1.fr/files/32211/32211-06.pdf.

Application en SHS et sciences de l'environnement

Paëgelow, M., N. Villa, L. Cornez, F. Ferraty, L. Ferré, and P. Sarda (2004). « Modélisations prospectives de l'occupation du sol. Le cas d'une montagne méditerranéenne ». In: *Cybergéo*, p. 295. DOI: 10.4000/cybergeo.2811.

Autre

Villa-Vialaneix, N. (2013). « J'ai testé pour vous... un MOOC ». In: *Statistique et Enseignement* 4.2, pp. 3–17. URL: http://publications-sfds.math.cnrs.fr/index.php/StatEns/article/view/241.

B.3 Éditoriaux

Villa-Vialaneix, N., L. Liaubet, and M. San Cristobal (2011). « What is a (good) gene network? » In: *Journal of Animal Breeding and Genetics* 128.1, pp. 1–2. DOI: 10.1111/j.1439-0388.2010.00916.x.

Cottrell, M., M. Olteanu, J. Rouchier, and N. Villa-Vialaneix (2012). « Éditorial du numéro spécial RNTI - MASHS 2011/2012 : Modèles et Apprentissage en Sciences Humaines et Sociales ». In: *Revue des Nouvelles Technologies de l'Information* SHS-1, pp. 97–110. URL: http://www.editions-hermann.fr/ficheproduit.php?prodid=1371.

B.4 Chapitres d'ouvrages collectifs

Follador, M., N. Villa, M. Paëgelow, F. Renno, and R. Bruno (2008). « Modelling Environmental Dynamics ». In: ed. by M. Paëgelow and M.T. Camacho-Olmedo. Environmental Science and Engineering. Berlin/Heidelberg: Springer. Chap. Tropical deforestation modelling: a comparative analysis of different predictive approaches. The case study of Peten, Guatemala, pp. 77–108. ISBN: 978-3-540-68489-3.

Paëgelow, M., M.T. Camacho-Olmedo, F. Ferraty, L. Ferré, P. Sarda, and N. Villa (2008). « Modelling Environmental Dynamics ». In: ed. by M. Paëgelow and M.T. Camacho-Olmedo. Environmental Science and Engineering. Berlin/Heidelberg: Springer. Chap. Prospective modelling of environmental dynamics. A methodological comparison applied to mountain land cover changes, pp. 141–168. ISBN: 978-3-540-68489-3.

B.5 Communications dans des conférences internationales avec comité de lecture et publication des actes

Rossi, F. and N. Villa (2005). « Classification in Hilbert spaces with support vector machines ». In: *XIth International Symposium on Applied Stochastic Models and Data Analysis (ASMDA 2005)*. (May 17–20, 2005). Ed. by J. Janssen and P. Lenca. Brest, France, pp. 635–642. ISBN: 2-908849-15-1.

Villa, N. and F. Rossi (2005). « Support vector machine for functional data classification ». In: *XIIIth European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2005)*. (Apr. 27–29, 2005). Ed. by M. Verleysen. Bruges, Belgium: d-side publications, pp. 467–472. ISBN: 2-930307-05-6.

Villa, N. and R. Boulet (2007). « Clustering a medieval social network by SOM using a kernel based distance measure. » In: *XVth European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2007)*. (Apr. 25–27, 2007). Ed. by M. Verleysen. Bruges, Belgium: d-side publications, pp. 31–36. ISBN: 2-930307-07-2.

Villa, N. and F. Rossi (2007). « A comparison between dissimilarity SOM and kernel SOM for clustering the vertices of a graph ». In: *6th International Workshop on Self-Organizing Maps (WSOM 2007)*. (Sept. 3–6, 2007). Bielefeld, Germany: Neuroinformatics Group, Bielefeld University. ISBN: 978-3-00-022473-7. DOI: 10.2390/biecoll-wsom2007-139.

Rossi, F. and N. Villa (2008). « Consistency of derivative based functional classifiers on sampled data ». In: *XVIth European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2008)*. (Apr. 23–25, 2008). Ed. by M. Verleysen. Bruges, Belgium: d-side publications, pp. 445–450. ISBN: 2-930307-08-0.

Villa, N. and F. Rossi (2008). « Recent advances in the use of SVM for functional data classification ». In: *Functional and Operatorial Statistics (Prooceedings of First International Workshop on Functional and Operatorial Statistics (IWFOS 2008)*. (June 5–7, 2008). Ed. by S. Dabo-Niang and F. Ferraty. Contributions to Statistics. Toulouse, France: Physica-Verlag HD, pp. 273–280. ISBN: 978-3-7908-2061-4. DOI: 10.1007/978-3-7908-2062-1_41.

Rossi, F. and N. Villa (2009). « Topologically ordered graph clustering via deterministic annealing ». In: *XVth European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2009)*. (Apr. 22–24, 2009). Ed.

by M. Verleysen. Bruges, Belgium: d-side publications, pp. 529–534. ISBN: 2-930307-09-9.

Hernández, N., R.J. Biscay, N. Villa-Vialaneix, and I. Talavera-Bustamante (2010). « A functional density-based nonparametric approach for statistical calibration ». In: *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications - 15th Iberoamerican Congress on Pattern Recognition (CIARP 2010)*. (Nov. 8–11, 2010). Ed. by I. Bloch and R.M. Cesar. Vol. 6419. Lecture Notes in Computer Science. Sao Paulo, Brazil: Springer, pp. 450–457. ISBN: 978-3-642-16686-0. DOI: doi:10.1007/978-3-642-16687-7.

Liaubet, L., N. Villa-Vialaneix, A. Gamot, F. Rossi, P. Chérel, and M. SanCristobal (2010). « The structure of a gene network reveals 7 biological functions underlying eQTLs in pig. » In: *World Congress on Genetics Applied to Livestock Production (WCGALP 2010)*. (Aug. 1–6, 2010). Ed. by Gesellschaft für Tierzuchtwissenschaften e. V. 0147. Leipzig, Germany. ISBN: 978-3-00-031608-1.

Rohart, F., N. Villa-Vialaneix, A. Paris, J. Molina, C. Canlet, D. Milan, B. Laurent, and M. SanCristobal (2010). « Phenotypic prediction based on metabolomic data: LASSO vs BOLASSO, primary data vs wavelet data ». In: *World Congress on Genetics Applied to Livestock Production (WCGALP 2010)*. (Aug. 1–6, 2010). Ed. by Gesellschaft für Tierzuchtwissenschaften e. V. 0157_PP3-55. Leipzig, Germany. ISBN: 978-3-00-031608-1.

Olteanu, M., N. Villa-Vialaneix, and M. Cottrell (2012). « On-line relational SOM for dissimilarity data ». In: *Advances in Self-Organizing Maps (Proceedings of WSOM 2012)*. (Dec. 12–14, 2012). Ed. by P.A. Estevez, J. Principe, P. Zegers, and G. Barreto. Vol. 198. AISC (Advances in Intelligent Systems and Computing). Santiago, Chile: Springer Verlag, Berlin, Heidelberg, pp. 13–22. ISBN: 978-3-642-35229-4. DOI: 10.1007/978-3-642-35230-0_2.

Massoni, S., M. Olteanu, and N. Villa-Vialaneix (2013). « Which distance use when extracting typologies in sequence analysis? An application to school to work transitions ». In: *International Work Conference on Artificial Neural Networks (IWANN 2013)*. (June 12–14, 2013). Puerto de la Cruz, Tenerife.

Olteanu, M., N. Villa-Vialaneix, and C. Cierco-Ayrolles (2013). « Multiple kernel self-organizing maps ». In: *XXIst European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2013)*. (Apr. 24, 2023–Apr. 26, 2008). Ed. by M. Verleysen. Bruges, Belgium: i6doc.com, pp. 83–88.

Boelaert, J., L. Bendhaïba, M. Olteanu, and N. Villa-Vialaneix (2014). « SOMbrero: an R package for numeric and non-numeric self-organizing maps ». In: *Advances in Self-Organizing Maps and Learning Vector Quantization (Proceedings of WSOM 2014)*. (July 2–4, 2014). Ed. by T. Villmann, F.M. Schleif, M. Kaden, and M. Lange. Vol. 295. Advances in Intelligent Systems and Computing. Mittweida, Germany: Springer Verlag, Berlin, Heidelberg, pp. 219–228.

Mariette, J., M. Olteanu, J. Boelaert, and N. Villa-Vialaneix (2014). « Bagged kernel SOM ». In: *Advances in Self-Organizing Maps and Learning Vector Quantization (Proceedings of WSOM 2014)*. (July 2–4, 2014). Ed. by T. Villmann, F.M. Schleif, M. Kaden, and M. Lange. Vol. 295. Advances in Intelligent Systems and Computing. Mittweida, Germany: Springer Verlag, Berlin, Heidelberg, pp. 45–54.

Sibertin-Blanc, C. and N. Villa-Vialaneix (2014). « Data analysis of social simulations outputs ». In: *Post-proceedings of 15th International Workshop on Multi-Agent-Based Simulation (MABS 2014)*. (May 5–6, 2014). Forthcoming. Paris, France.

B.6 Conférences invitées

Rossi, F. and N. Villa (2007). « Discrimination de fonctions par Machines à Vecteurs de Support ». In: *5èmes Journées de Statistique Fonctionnelle et Opérationnelle*. (June 21–22, 2007). Lille, France, pp. 22–23.

Villa, N., F. Rossi, and Q.D. Truong (2008). « Mining a medieval social network by kernel SOM and related methods ». In: *Modèles et Apprentissage en Sciences Humaines et Sociales (MASHS 2008)*. (June 5–6, 2008). Créteil, France.

Villa, N. and F. Rossi (2009). « Méthodes de classification organisée pour la recherche de communautés dans les réseaux sociaux ». In: *38ièmes Journées de Statistique de la SFdS (JdS 2009), 1/2 Journée Satellite STID*. (Mar. 27–28, 2009). Bordeaux, France. URL: http://hal.inria.fr/inria-00386797/fr/.

Villa-Vialaneix, N. and F. Rossi (2010a). « Classification and regression based on derivatives: a consistency result ». In: *II Simposio sobre Modelamiento Estadístico*. (Dec. 2–3, 2010). Valparaìso, Chile.

— (2010b). « Visualization of graphs by organized clustering: application to social and biological networks ». In: *Workshop on Challenging problems in Statistical Learning (STATLEARN)*. (Jan. 28–29, 2010). Paris, France.

Rossi, F., N. Villa-Vialaneix, and F. Hautefeuille (2011a). « Exploration of a large database of French charters with social network methods ». In: *International Medieval Congress (IMC 2011), Session 1607 "Problems and Possibilities of Early Medieval Diplomatic, II: Members and Margins"*. (July 11–14, 2011). Leeds, UK.

— (2011b). « Exploration of a large database of French notarial acts with social network methods ». In: *Digital Diplomatics 2011*. (Sept. 29–Oct. 1, 2011). Napoli, Italy.

Villa-Vialaneix, N. and T. Laurent (2013). « Permutation tests for labeled network analysis ». In: *7th International Conference of the ERCIM WG on Computational and Methodological Statistics (ERCIM 2014)*. London, UK.

Villa-Vialaneix, N., M. Vignes, N. Viguerie, and M. San Cristobal (2014b). « Inferring networks from multiple samples with consensus LASSO ». In: *ENBIS Spring Meeting*. (Apr. 10–11, 2014). Paris, France.

B.7 Autres conférences (nationales ou sans communication orale ou sans publication d'actes)

Bruno, R., M. Follador, M. Paëgelow, F. Renno, and N. Villa (2006). « Integrating remote sensing, GIS and prediction models to monitor the deforestation and erosion in Peten reserve, Guatemala ». In: *XIth International Congress for Mathematical Geology (IAMG 2006)*. (Sept. 3–8, 2006). Ed. by E. Pirard, A. Dassargues, and H.S. Havenith. Liège, Belgium.

Villa, N. and F. Rossi (2006a). « SVM fonctionnels par interpolation spline ». In: *38ièmes Journées de Statistique de la SFdS (JdS 2006)*. (May 29–June 2, 2006). Clamart, France.

Boulet, R., F. Hautefeuille, B. Jouve, P. Kuntz, B. Le Goffic, F. Picarougne, and N. Villa (2007). « Sur l'analyse de réseaux de sociabilité dans la société paysanne médiévale ». In: *Modèles et Apprentissage en Sciences Humaines et Sociales (MASHS 2007)*. (May 10–11, 2007). Brest, France.

Follador, M., F. Renno, R. Bruno, M. Paëgelow, N. Villa, and J.F. Mas (2007). « Remote sensing, GIS and predictive methods: a new approach to environmental and hazard problems ». In: *Sesto Forum Italiano di Scienze della Terra (GeoItalia 2007)*. (Sept. 12–14, 2007). Rimini, Italy.

Gamot, A., N. Villa, L. Liaubet, F. Rossi, G. Tosser-Klopp, P. Chérel, and M. San Cristobal (2009). « Are gene networks always meaningful? » In: *European Animal Disease Genomics Network of Excellence for Animal Health and Food Safety (EADGENE Days)*. (Oct. 13–15, 2009). Paris, France.

Villa, N., T. Dkaki, S. Gadat, J.M. Inglebert, and Q.D. Truong (2009). « Recherche et représentation de communautés dans des grands graphes ». In: *2ème Séminaire Veille Stratégique, Scientifique et Technologique (VSST 2009)*. (Mar. 30–31, 2009). Nancy, France.

Laurent, T. and N. Villa-Vialaneix (2010). « Analysis of the influence of a network on the values of its nodes: the use of spatial indexes ». In: *1ère Conférence Modèles et Analyse des Réseaux : Approches Mathématiques et Informatique (MARAMI 2010)*. (Oct. 10–11, 2010). Toulouse, France.

Villa-Vialaneix, N., M. Follador, and A. Leip (2010). « A comparison of three learning methods to predict N2O fluxes and N leaching ». In: *Modèles et Apprentissage en Sciences Humaines et Sociales (MASHS 2010)*. (June 10–11, 2010). Ed. by C. Bienacki, E. Masson, A. Lendasse, and E. Séverin. Lille, France: Multiprint Oy (Espoo, Finland), pp. 57–64. ISBN: 978-952-60-2177-4.

Leip, A., M. Follador, S. Tarantola, M. Busto, and N. Villa-Vialaneix (2011). « Sensitivity of the process-based model DNDC on microbiological parameters ». In: *Nitrogen and Global Change - Key findings & Future challenges*. (Apr. 11–14, 2011). Edinburgh, UK.

Villa-Vialaneix, N., L. Liaubet, T. Laurent, A. Gamot, P. Cherel, and M. San Cristobal (2011). « L'analyse d'un réseau de co-expression génique met en valeur des groupes fonctionnels homogènes et des gènes importants relatifs a un phénotype d'intérêt ». In: *Actes des 43èmes Journées de Statistique, Société Française de Statistique*. (May 23–27, 2011). Tunis, Tunisie.

Laurent, T. and N. Villa-Vialaneix (2012). « Analyse de données pour des graphes étiquetés ». In: *44èmes Journées de Statistique de la SFdS (JdS 2012)*. (May 21–25, 2012). Bruxelles, Belgique.

Villa-Vialaneix, N., N.A. Edwards, L. Liaubet, and N. Viguerie (2012). « Comparison of network inference packages and methods for multiple network inference ». In: *1ères Rencontres R BoRdeaux*. (July 2–3, 2012). BoRdeaux, France.

Villa-Vialaneix, N., F. Rossi, and F. Hautefeuille (2012a). « Exploration relationnelle d'un corpus d'actes notariés médiévaux ». In: *Colloque Configuration(s)*. (June 8–9, 2012). Paris, France.

— (2012b). « Spatial correlation in bipartite networks: the impact of the geographical distances on the relations in a corpus of medieval transactions ». In: *Modèles et Apprentissage en Sciences Humaines et Sociales (MASHS 2012)*. (June 4–5, 2012). Paris, France.

Bendhaïba, L., M. Olteanu, and N. Villa-Vialaneix (2013). « SOMbrero : cartes auto-organisatrices stochastiques pour l'intégration de données décrites par des tableaux de dissimilarités ». In: *2èmes Rencontres R BoRdeaux*. (June 27–26, 2013). Lyon, France.

Brunet, F., J. Mariette, C. Cierco-Ayrolles, C. Gaspin, P. Bardou, and N. Villa-Vialaneix (2013). « Classification d'un graphe de co-expression avec des méta-données pour la détection de micro-RNAs ». In: *Modèles et l'Analyse des Réseaux : Approches Mathématiques et Informatiques (MARAMI 2013)*. (Oct. 16–18, 2013). Saint-Étienne, France.

Leroux, D. and N. Villa-Vialaneix (2013). « sexy-rgtk: a package for programming RGtk2 GUI in a user-friendly manner ». In: *2èmes Rencontres R BoRdeaux*. (June 27–26, 2013). Lyon, France.

Villa-Vialaneix, N., M. Olteanu, and C. Cierco-Ayrolles (2013). « Carte auto-organisatrice pour graphes étiquetés ». In: *Colloque Extraction et Gestion de Connaissances (EGC 2013), ateliers Fouille de Grands Graphes (FGG)*. (Jan. 29–29, 2013). Toulouse, France.

Villa-Vialaneix, N. and M. San Cristobal (2013). « Consensus LASSO : inférence conjointe de réseaux de gènes dans des conditions expérimentales multiples ». In: *45e Journées de Statistique de la SFdS (JdS 2013)*. (May 27–31, 2013). Toulouse, France.

Hernández, N., R.J. Biscay, N. Villa-Vialaneix, and I. Talavera (2014b). « Density-based inverse calibration with functional predictors ». In: *11th International Conference on Operations Research (ICOR 2014)*. (Mar. 11, 2014–Mar. 14, 2010). Havana, Cuba. ISBN: 978-3-642-16686-0.

Olteanu, M. and N. Villa-Vialaneix (2014). « Self-organizing maps for clustering visualization of bipartite graphs ». In: *46e Journées de Statistique de la SFdS (JdS 2014)*. (June 2–6, 2014). Rennes, France.

Picheny, V., J. Vandel, M. Vignes, and N. Villa-Vialaneix (2014). « Reconstruction quality of a biological network when its constituting elements are partially observed ». In: *AI & Statistics*. (Apr. 22–25, 2014). L014. Reykjavik, Iceland.

Villa-Vialaneix, N. (2014). « J'ai testé pour vous... un MOOC ». In: *46e Journées de Statistique de la SFdS (JdS 2014)*. (June 2–6, 2014). Rennes, France.

B.8 Articles soumis ou en révision

Montastier, E., S. Caspar-Bauguil, N. Villa-Vialaneix, P. Hlavaty, E. Tvrzicka, I. Gonzalez, W.H.M. Saris, D. Langin, M. Kunesova, and N. Viguerie (2014). « System model network for adipose tissue signatures related to weight changes in response to calorie restriction and subsequent weight maintenance ». Submitted for publication (first co-author among three first authors).

B.9 Logiciels

Villa-Vialaneix, N., L. Bendhaïba, J. Boelaert, and M. Olteanu (2013). *SOMbrero: SOM Bound to Realize Euclidean and Relational Outputs*. R package. version 0.5 (11/20 2013). URL: http://sombrero.r-forge.r-project.org.

Villa-Vialaneix, N. and N. Edwards (2013). *therese: Trust the Holy Estimation of Regulatory nEtworks from Several Expression data*. R package. version 0.2 (10/17 2013). URL: http://therese-pkg.r-forge.r-project.org.