



HAL
open science

Approches topologiques pour l'analyse exploratoire de données et l'aide à la décision

Michael Aupetit

► **To cite this version:**

Michael Aupetit. Approches topologiques pour l'analyse exploratoire de données et l'aide à la décision. Intelligence artificielle [cs.AI]. Ecole Doctorale de l'Université Paris Sud XI (ED 427), 2012. tel-01167056

HAL Id: tel-01167056

<https://hal.science/tel-01167056>

Submitted on 24 Jun 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Copyright



**THÈSE D'HABILITATION À DIRIGER DES
RECHERCHES DE L'UNIVERSITÉ PARIS-SUD XI**

Spécialité

Informatique (ED427)

Présentée par

Dr. Michaël AUPETIT¹

Sujet de la thèse :

**Approches topologiques pour l'analyse
exploratoire de données et l'aide à la décision**

soumise aux rapporteurs

Pr. Guy MELANCON	Université Bordeaux I, France
Pr. Michel VERLEYSEN	Université Catholique de Louvain, Belgique
Pr. Djamel ZIGHED	Université Lumière Lyon II, France

soutenue le 11 juillet 2012 à Orsay devant le jury composé de :

D.R. Michèle SEBAG	Univ. Paris-Sud XI	Président
Pr. Michel VERLEYSEN	Univ. Catholique de Louvain	Rapporteur
Pr. Djamel ZIGHED	Univ. Lyon II	Rapporteur
Pr. Younes BENNANI	Univ. Paris XIII	Examineur
D.R. Jean-Daniel FEKETE	INRIA Saclay	Examineur
Pr. Gérard GOVAERT	UT Compiègne	Examineur

1. <http://michael.aupetit.free.fr>

Avant-propos

Depuis au moins les premières pierres taillées de l'ère Paléolithique, les hommes n'ont cessé de créer des artefacts, moyens d'agir sur leur environnement et moyens de l'observer au-delà de leurs capacités propres. Ils ont développé ces outils pour les assister dans leur quête viscérale de compréhension (sciences) et de maîtrise (techniques) de ce monde dont ils font partie. Cette compréhension du monde est nécessaire pour en prédire les états, et la maîtrise qui en découle est le moyen de ne plus le subir mais de l'asservir pour réduire les souffrances qu'il nous assène par nature. En plus de transformer le Monde, les hommes ont aussi pu réparer, corriger et augmenter leur propre corps par des orthèses et des prothèses biologiques, chimiques, mécaniques ou numériques. La multiplication de ces moyens d'agir et d'observer entraîne un accroissement exponentiel des données désormais capturées dont la masse est supposée assurer les hommes de contenir toute l'information utile à leur quête. Cette massification des données impose de développer des méthodes d'analyse et de traitement toujours plus efficaces pour que les hommes qui les étudient ou appuient leurs décisions sur elles puissent continuer à le faire et à le faire mieux. J'ai proposé différentes approches dans les champs de l'analyse descriptive et de la modélisation prédictive afin de rendre plus intelligible la chaîne de traitement de l'information du capteur à l'écran. J'ai placé la Topologie au coeur de mes travaux. En effet, je considère qu'elle forme le substrat essentiel à l'interprétabilité de l'information, c'est-à-dire à la transmission du sens dans cette chaîne, et *in fine* à la compréhension et à la maîtrise du Monde par l'Homme.

Le développement permanent des artefacts techniques pour tenter de mieux comprendre et maîtriser le Monde, entretient la croissance de sa complexité, à la fois parce que grâce aux artefacts développés pour la science et la technologie, nous accédons à des mécanismes toujours plus précis et plus nombreux qui le gouvernent, mais aussi parce que les artefacts techniques engendrés par ces connaissances nouvelles, font partie intégrante du Monde lui-même et en modifient le fonctionnement. Aux causes naturelles s'entremêlent les causes artificielles. C'est pour certains auteurs [31] une nouvelle ère qui a commencée au 19e siècle, l'ère de l'Anthropocène, dans laquelle l'intelligence des hommes les a dotés de moyens techniques capable de modifier durablement leur écosystème et en particulier les mécanismes de la Sélection Naturelle qui ont engendrée cette intelligence. Il est vraisemblable que ce développement aboutira à l'émergence de machines intelligentes, conscientes et émotionnelles capables d'explorer ce monde par elles-mêmes et de communiquer aux hommes leurs conclusions sur les lois qui le gouvernent, comme le font déjà quelques machines encore rudimentaires mais efficaces [111]. Ces machines sont la forme ultime de ces orthèses dont les hommes cherchent à se doter depuis le Paléolithique pour tenter de dépasser leur condition. Avant que n'advienne ce moment singulier que Kurzweil appelle la Singularité [72], où les machines autonomes seront en mesure de dépasser l'homme et de développer pour elles-mêmes des connaissances et des techniques, des questions éthiques seront posées qui mèneront soit à l'abandon du développement de telles machines, soit à la nécessité impérieuse de les maîtriser. Il sera alors impératif que les processus internes et les produits de ces machines soient intelligibles aux hommes afin qu'ils en comprennent le sens et en conservent le contrôle. C'est ainsi tout l'enjeu de mes recherches actuelles et futures.

Table des matières

1	Motivations	1
1.1	Premiers pas	1
1.2	Un parcours entre recherche fondamentale et besoins applicatifs	1
1.3	Des machines et des hommes	2
2	Problématique	3
2.1	Analyser et prévoir	3
2.1.1	Les requêtes	3
2.1.2	Les observations	4
2.1.3	Les modèles	4
2.2	Le besoin d'interprétabilité	6
2.3	La Topologie au coeur des mes travaux	12
2.3.1	Un point de vue topologique	12
2.3.2	Autres exemples de l'utilité de la topologie	17
2.3.3	Le caractère essentiel et générique de la Topologie	20
2.4	Voir pour comprendre	24
2.4.1	Du signal à la décision	24
2.4.2	Deux modes de représentation graphique	26
2.4.3	L'inférence visuelle	28
2.4.4	Le principe de fiabilité pour permettre l'inférence	29
2.4.5	En résumé	30
2.5	Problématique scientifique et axes de recherches	32
2.5.1	La problématique scientifique principale	32
2.5.2	Le paradigme de la visualisation multidimensionnelle <i>in situ</i>	33
2.5.3	L'apprentissage automatique de la topologie	33
2.6	Aperçu de mes principales contributions	35
3	Visualiser la topologie d'un nuage de points	37
3.1	Les données	37
3.2	L'analyse descriptive par projection	37
3.2.1	L'inférence à partir des représentations graphiques	37
3.2.2	Le cas des projections dans le plan	38
3.2.3	Les distorsions	39
3.2.4	Le problème des distorsions	40
3.2.5	Montrer plus que les distorsions : la mesure de proximité	44

3.2.6	Le paradigme <i>WinSitu</i> de visualisation <i>in situ</i>	46
3.2.7	Le paradigme <i>WinSitu</i> face à l'état de l'art . .	47
3.3	Applications	51
4	Extraire la topologie d'un nuage de points	52
4.1	Quelques notions de topologie	52
4.2	Le cas des nuages de points	55
4.3	Analyser la connexité intra et inter classes	55
4.3.1	Le graphe des classes	55
4.3.2	Quel graphe de proximité choisir?	56
4.3.3	Une représentation graphique sans perte d'in- formation topologique	57
4.3.4	Aspects calculatoires	58
4.3.5	Liens avec l'état de l'art	59
4.4	Les limites de l'approche descriptive	60
4.5	Vers une approche générative	61
4.5.1	La reconstruction de formes	61
4.5.2	Le Topology Representing Network et le Wit- ness Complex	61
4.5.3	Les modèles génératifs	64
4.6	Le graphe génératif	66
4.6.1	Contexte	66
4.6.2	Le cadre génératif	67
4.6.3	Définition du modèle de graphe génératif . . .	68
4.6.4	La vraisemblance et sa maximisation avec l'al- gorithme EM	70
4.6.5	Emergence de la topologie et sélection de mo- dèle par le critère BIC	70
4.6.6	Exemples d'applications	72
5	Conclusion et perspectives	77
5.1	En résumé	77
5.2	Visualisation d'information	78
5.2.1	Topologie et distortions	78
5.2.2	Paradigme WinSitu et critères de fiabilité et d'authenticité	80
5.2.3	ProxiViz et au-delà	81
5.2.4	Formaliser l'interprétabilité	82
5.2.5	Visualiser <i>in situ</i> d'autre types de données . .	83
5.3	Modélisation <i>in situ</i>	83
5.3.1	Au-delà des graphes	83
5.3.2	Quel paradigme d'apprentissage?	84
5.3.3	Les structures multi-échelles	86

5.3.4	Gérer la complexité algorithmique	87
5.3.5	Complexité et topologie	88
5.3.6	Modèles interprétables	90
5.3.7	Autres types de données	91
5.3.8	Autres applications	92
5.4	Futurs possibles	93
5.4.1	Des données à la connaissance	93
5.4.2	Des machines conscientes	95

Références	107
-------------------	------------

1 Motivations

1.1 Premiers pas

La lecture des romans d'Isaac Asimov [7] et de ses trois Lois de la Robotique ont nourris ma passion pour la Robotique et l'Intelligence Artificielle. J'ai construit un premier robot en 1985 à partir de jouets récupérés, doté d'une fonction rudimentaire d'évitement d'obstacle. Puis un second en 1990 et un troisième en 1993, pour l'essentiel des machines télécommandées. Concevoir un système cognitif pour ces machines est rapidement devenu la motivation première de mes études...

1.2 Un parcours entre recherche fondamentale et besoins applicatifs

Après l'obtention en 1998 du diplôme d'ingénieur en informatique de l'École pour les Etude et la Recherche en Informatique et Electronique (EERIE) option "Intelligence Artificielle", située à Nîmes, et du Diplôme d'Etudes Approfondies "Systèmes Automatiques et Microélectroniques" (SyAM) de l'Université de Montpellier 2, j'ai soutenu mon Doctorat en Génie Industriel de l'Institut National Polytechnique de Grenoble en décembre 2001.

J'ai effectué mes recherches dans le cadre de mon post-doctorat dès janvier 2002 puis en tant qu'ingénieur-chercheur à partir de mars 2004, au laboratoire Détection et Sismologie Opérationnelle (LDSO), enfin à partir de décembre 2008 au laboratoire Information, Modèles et Apprentissage (LIMA). Le LDSO est situé à Bruyères-le-Châtel, Essonne, il fait partie du Département Analyse Surveillance Environnement, au sein de la Direction des Applications Militaires du Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA). Le LIMA est situé à Saclay, Essonne, il fait partie du département Capteurs Signaux et Informations (DCSI) de l'institut LIST au sein de la Direction de la Recherche Technologique du CEA.

Au LDSO, les analystes géophysiciens dépouillent quotidiennement des milliers de signaux sismiques, infrasonores et hydro-acoustiques reçus sur des réseaux de plusieurs centaines de capteurs répartis autour du globe. L'une de leurs missions est d'alerter les autorités en cas d'événement sismique de forte magnitude en France métropolitaine, pour l'organisation des secours. Pour cela, il doivent repérer les secousses à l'origine de ces signaux, et en déterminer la localisation temporelle et spatiale, ainsi que la magnitude. Ils déterminent aussi l'origine anthropique ou naturelle de ces événements, puis les

archivent en base de données. A partir de ces archives, ma mission était de faciliter le travail des analystes géophysiciens, d'une part en automatisant les tâches sans valeur ajoutée² pour leur permettre de focaliser leur expertise sur les tâches complexes ; et d'autre part de leur proposer des outils d'aide à la décision pour les assister dans la réalisation de ces tâches complexes. J'ai en particulier contribué au développement d'un outil d'aide à la décision pour la révision du bulletin sismique [90].

Au LIMA, chaque projet aborde un métier différent généralement en lien avec un client industriel extérieur. J'ai travaillé sur un projet de conception d'un outil d'aide à la décision basé sur la représentation graphique d'informations. Dans ce projet, des conteneurs portuaires sont sondés à l'aide d'un canon à neutron, la matière soumise au flux neutronique génère des particules dont le spectre en énergie est mesuré et localisé spatialement. De ces spectres sont estimés les proportions d'une quinzaine d'éléments chimiques présents dans la zone sondée. L'interface que nous avons réalisée doit à partir de ces seules proportions, permettre aux officiers des douanes de décider l'ouverture du conteneur pour une fouille approfondie en quête de substances illicites ou dangereuses. Nous avons proposé différentes méthodes de visualisation d'information pour rendre intelligible à l'officier l'information reconstruite à partir de ces mesures [14].

1.3 Des machines et des hommes

La confrontation avec des experts métiers de différents horizons (géophysique, biologie, mesures physiques...) non spécialistes des méthodes de fouille de données, montre que l'interprétabilité des informations présentées par les systèmes d'aide à la décision est primordiale pour qu'ils accordent toute leur confiance à ces systèmes et les utilisent *in fine*. Je me suis focalisé sur l'étude scientifique et le développement technique de méthodes originales de représentation des informations, interprétables directement par l'expert métier. Ces études et développements m'ont conduit à aborder les différents domaines scientifiques sur lesquels se fonde l'Intelligence Artificielle : les Mathématiques, l'Informatique et les Sciences Cognitives.

2. les tirs de carrière génèrent de nombreux tremblements de terre de faible magnitude, qu'il faut traiter afin de constituer des cartes de risque sismique fiables, mais qui n'ont pas le caractère d'urgence d'une alerte.

2 Problématique

2.1 Analyser et prévoir

Les systèmes de mesure génèrent des masses de données à la fois en termes de nombre de variables (capteurs) et de nombre d'individus (taille de l'échantillon). Ces données doivent être analysées interactivement par un analyste. Les outils d'aide à la décision ont un rôle crucial : ils doivent appuyer la décision de l'analyste et pour cela lui rendre interprétables les informations contenues dans les données brutes pour lui permettre de construire un modèle du phénomène physique mesuré.

Ce modèle et ses paramètres (l'hypothèse) génèrent des états que l'analyste confronte aux observations de ceux du système modélisé (les données brutes). L'écart entre les observations et la prédiction réalisée grâce au modèle, permet d'infirmer ou de confirmer l'hypothèse. L'hypothèse tant qu'elle n'est pas invalidée par de nouvelles observations, a valeur de vérité. Et si elle est formulée dans un langage interprétable par l'analyste, elle lui fournit une explication plausible du comportement du système. La démarche scientifique est basée sur ce processus interactif de modélisation où l'analyste agit sur les paramètres du modèle dont la confrontation aux observations modifie en retour ses croyances jusqu'à la convergence vers un modèle à la fois précis et interprétable.

2.1.1 Les requêtes

Face aux états pris par le système observé, l'analyste a plusieurs types de requêtes auxquelles le modèle de ce système peut répondre [40] :

- **Description** : Que s'est-il passé ? Le modèle doit fournir une description du système et de ses états interprétable directement par l'analyste (segmentation d'un marché ; corrélation de mesures...) ou exploitable automatiquement pour accéder efficacement à l'information (structures de données, graphes, arbres de partitionnement...);
- **Explication** : Pourquoi cela s'est-il passé ? Le modèle doit permettre de relier et de comparer objectivement les observations à des faits connus pour leur donner un sens et tester des hypothèses (variables discriminantes ; tests statistiques ; analyse en composantes principales ; analyse de sensibilité...);
- **Prédiction** : Que va-t-il se passer ? Le modèle doit permettre d'inférer les états futurs possibles du système à partir de son

état actuel et de ses états observés dans le passé (météo, finance...);

- **Décision** : Comment agir sur le système pour que tel événement ait ou n'ait pas lieu? Le modèle peut fournir une aide à la décision, *i.e.* proposer à l'analyste les actions à mener sur des variables de contrôle pour placer le système dans un état désiré à partir d'un état donné et d'un contexte (recommandation pour la révision de bulletin sismiques [90], contrôle de conteneurs [14], pilotage de processus, maintenance prédictive...). Ou bien il peut effectuer lui-même ces actions dans des cas où leur robustesse et leur efficacité sont éprouvées au sens de critères objectifs acceptés par les entités qui prennent la responsabilité de cette automatisation (contrôle de systèmes complexes comme le pilotage automatique d'un avion de ligne par exemple).

2.1.2 Les observations

Pour aborder les réponses techniques apportées à ces requêtes, il faut préciser quelques notions. Dans le cadre de la Statistique [101], on considère que les états du système sont perçus au travers d'un certain nombre de variables aléatoires. Un ensemble de valeurs prises simultanément pour ces variables forment un individu, une observation de l'état du système. L'ensemble des individus forme un échantillon d'une population plus large et inobservable censée décrire tous les états possibles du système mesurables dans cet espace de variables. Le domaine couvert par la population dans cet espace des variables est probabilisé par l'application en tout point du domaine d'une fonction densité de probabilité à valeurs dans l'espace des réels positifs, le domaine est le support de cette fonction. L'intégrale de Lebesgues de cette fonction sur le domaine est fixée par convention à l'unité : la probabilité qu'une observation appartienne au domaine vaut 1 et exprime la totale certitude que cela se produise. Un objectif majeur du statisticien est de caractériser la fonction densité de probabilité de la population à partir de l'échantillon, *i.e.* d'estimer la loi à l'origine des observations afin d'expliquer les observations passées et d'estimer la probabilité d'observations futures.

2.1.3 Les modèles

Suivant la définition de Tukey [116], on peut distinguer parmi les méthodes statistiques les méthodes d'analyse exploratoire de données basées sur des modèles descriptifs, et les méthodes d'inférence

statistique ou d'analyse confirmatoire, issues de modèles prédictifs.

Le modèle descriptif fournit une ou plusieurs mesures de l'échantillon sous une forme conventionnelle (symbole, valeur moyenne ou écart-type d'une variable...) ou perceptuelle (graphique, nuage de points, histogramme, diagramme noeud-lien...). Je considère que ces mesures servent de support à la formulation mentale et subjective par l'analyste d'une hypothèse prédictive ultérieure ou d'une décision (choix d'une action à mener). Parmi les requêtes exposées précédemment, le cas descriptif répond aux requêtes de description ou d'explication de l'analyste et *in fine* d'aide à la décision. Individus et variables n'ont pas de rôle a priori, l'analyste cherche des relations entre eux mesurées à l'aune de différents modèles de représentation (modèles non supervisés).

Au contraire, le modèle prédictif met en relation des variables explicatives et des variables à expliquer, définies comme telles a priori. Je considère qu'un modèle prédictif est une hypothèse objectivée car exprimée explicitement par l'analyste dans un formalisme mathématique et éventuellement implémentée sous forme d'un algorithme dans un programme informatique, qui sert à la prédiction ou la décision automatique et dont on peut mesurer objectivement la capacité à reproduire les observations. Le cas prédictif répond aux requêtes de prédiction et de décision. Individus et variables ont des rôles clairement identifiés (variables explicatives et à expliquer, individus prototypiques et atypiques...) et il s'agit de construire par optimisation d'une fonction d'énergie, minimisation d'un critère d'erreur entre prédictions et observations, utilisation de tests statistiques ou acquisition de connaissances d'experts, le modèle capable de rendre compte des relations entre variables explicatives et variables à expliquer à partir des individus prototypiques.

Les modèles descriptifs ont un rôle d'encodage de l'information, ils projettent les données dans des espaces topologiques spécifiques (vectoriels, arbres, graphes, hyper-graphes, complexes simpliciaux...) et peuvent aussi être employés seuls (compression de fichiers, structure de données accélérant l'accès à l'information). Ils m'intéressent surtout en tant que moyen de fournir à l'analyste "un point de vue" particulier sur les données. En effet, l'encodage présenté à l'analyste n'est pas une prédiction dont la confrontation au réel donnera valeur à l'hypothèse sous-jacente, c'est plutôt une transposition des données dans un référentiel dans lequel l'analyste peut former mentalement des hypothèses à leur confronter (Figure 1a). Dans ce cas, le modèle décisionnel *in fine* n'est pas implémenté sur une machine donc objectif, mais mental donc subjectif, il s'enrichit implicitement des connaissances externes aux données que possèdent l'analyste qui

est alors seul juge de l'écart de prédiction donc de la pertinence de l'hypothèse que le modèle implémente. La possibilité offerte à l'analyste d'intégrer à son modèle mental des connaissances externes aux données, se fait aux dépens de l'objectivité du modèle, donc de la possibilité de le transmettre *verbatim* à d'autres ou d'en automatiser l'exploitation.

Au contraire, dans le cas prédictif, ce qui est présenté objectivement à l'analyste est une instance des variables à prédire devant s'approcher au mieux des valeurs que prendraient ces variables mesurées sur le système réel. L'essentiel du sens est porté implicitement par le modèle, le modèle doit donc être interprétable par l'analyste, *i.e.* l'analyste doit s'en former une représentation mentale dont ce modèle est pour lui l'implémentation objective (Figure 1d).

Qu'elles soient descriptives ou prédictives, les méthodes de l'analyse de données utilisées en interaction avec l'analyste, ont finalement le même objectif de permettre à celui-ci de décider soit d'agir sur la représentation qui ne le satisfait pas (les paramètres du modèle prédictif, ceux du modèle descriptif ou de sa propre représentation mentale) soit d'en tirer parti pour agir sur le monde extérieur.

La figure 1a illustre le cas du modèle descriptif. Le modèle mental de la chaîne de traitement est supposé maîtrisé par l'analyste, sa certitude est forte. Le modèle mental du système étudié est en cours de construction, sa certitude augmente avec l'expérience en réduisant l'écart subjectif entre observations et attentes

La figure 1b illustre le cas de l'analyse avec un modèle prédictif non interprétable (boîte noire). La réponse attendue du modèle mental du modèle prédictif est incertaine car le modèle mental de ce modèle n'est pas connu. L'interprétation des écarts entre les observations issues du système réel et du modèle prédictif censé le représenter est difficile.

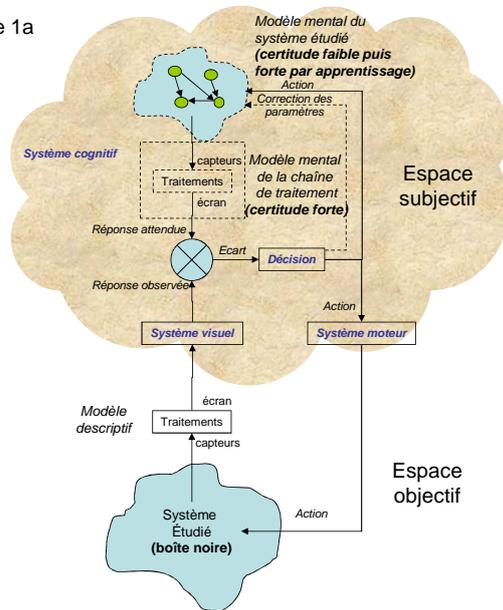
La figure 1c montre que pour rendre utilisable un modèle prédictif non interprétable, il faut d'abord s'en faire un modèle mental, donc appliquer l'analyse descriptive au modèle prédictif lui-même. Je considère qu'un modèle est interprétable s'il permet cette modélisation mentale de son fonctionnement.

Le figure 1d illustre comment le modèle prédictif rendu interprétable permet de modéliser et de comprendre le système réel.

2.2 Le besoin d'interprétabilité

Un modèle n'est exploitable par l'analyste que s'il lui est interprétable. On peut illustrer cette nécessité pour l'analyste de comprendre le modèle qu'il utilise par l'analogie avec une planisphère, représen-

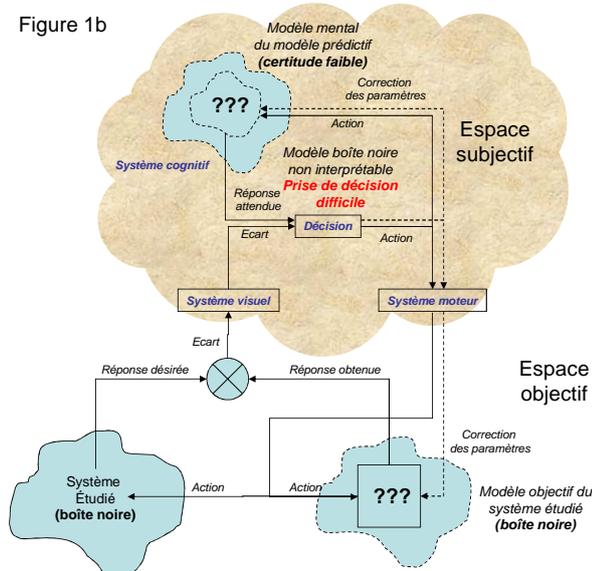
Figure 1a



Analyse descriptive

La sortie du système étudié est confrontée à celle du modèle mental de ce système. L'écart entre les deux permet de corriger le modèle et d'acquiescer une certitude forte de son fonctionnement.

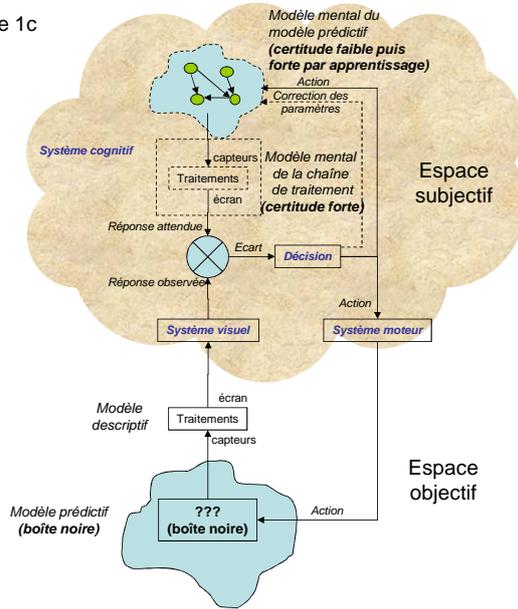
Figure 1b



Analyse prédictive avec un modèle non interprétable

Le modèle boîte noire n'est pas interprétable par l'analyste, la réponse attendue du modèle est très incertaine, sa confrontation à la sortie désirée du système étudié ne permet pas de corriger le modèle. Une étape préalable de compréhension du modèle doit être effectuée (Figure 1c).

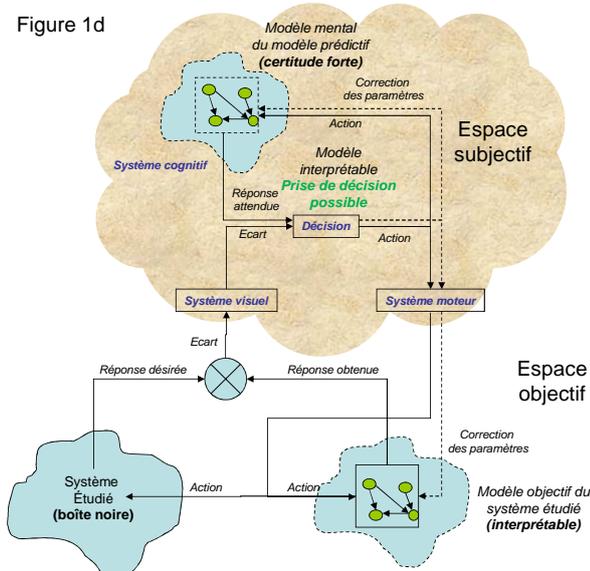
Figure 1c



Analyse descriptive du modèle prédictif

Le modèle boîte noire doit être compris par l'analyste par apprentissage. Ce n'est possible qu'avec certaines familles de modèles que l'on peut interpréter.

Figure 1d



Analyse prédictive avec un modèle interprétable

Le modèle rendu interprétable par apprentissage de l'analyste (Figure 1c) permet la compréhension du système étudié et la prise de décision.

tation synthétique d'une partie très restreinte (à tous points de vue) du globe terrestre. Utiliser cette carte nécessite de comprendre ce qu'elle modélise bien (la présence d'eau en un lieu donné avec un certain degré de précision), ce qu'elle ne modélise pas (la vitesse instantannée du vent, la continuité de la surface du globe sur les bords de la carte) ou ce qu'elle modélise mal (les distances entre points éloignés sur le globe). Il faut a priori pour cela comprendre la projection utilisée, la légende (signification des couleurs, repères et symboles) ainsi que les partis pris de la représentation (frontières des états, lieux jugés importants par le cartographe...). Dans un cas plus général, les données brutes issues de la mesure d'un système physique (le globe terrestre) observé par différents capteurs, sont transformées par de multiples traitements pour finalement livrer à l'analyste une représentation synthétique (la carte) de l'information qu'elles portent. Cette information telle qu'elle est portée et déformée par le modèle doit aider l'analyste à prendre ses décisions. Il faut donc qu'il comprenne le modèle et la nature de ces déformations afin de saisir la part de réalité reproduite par celui-ci et la distinguer de sa part artificielle pour prendre des décisions efficaces et pertinentes car fondées sur des informations vraies.

La performance du modèle peut se mesurer à l'écart entre son comportement et celui du système réel modélisé. Mais ce critère de qualité s'il permet de jauger les déformations que le modèle induit, n'est pas suffisant. En effet, comme proposé précédemment, l'analyste doit se forger une représentation mentale dont ce modèle est l'implémentation objective, et pour cela il doit agir, s'investir dans la conception du modèle afin de le comprendre (Figure 1c).

J'adhère ainsi au cadre conceptuel proposé initialement par James J. Gibson [51] dans les années 70. Gibson a proposé une approche écologique de la perception visuelle qui a marqué un tournant dans le domaine de la psychologie cognitive : la perception visuelle n'a lieu au niveau de la rétine que si d'une part un stimulus en provenance de l'environnement stimule les photorécepteurs de la rétine, et d'autre part que si ce stimulus est obtenu de manière active par l'observateur engagé vers un but. Cette perception visuelle dynamique permet la compréhension de l'environnement en supprimant les multiples ambiguïtés d'une perception passive grâce aux mouvements des yeux, de la tête et du corps. Cette thèse a été développée et étendue au domaine de la cognition en général [76] : il n'y a pas de perception et de compréhension du monde sans action sur lui, l'interaction est au coeur du processus de compréhension.

L'analyste doit donc interagir avec le modèle pour le comprendre, mais cela ne suffit pas. Il faut aussi que le modèle soit interpré-

table, c'est-à-dire qu'il soit tel que les interactions exploratoires de l'analyste lui permette de s'en forger efficacement une représentation mentale. De nombreux travaux ont été menés sur l'interprétabilité des modèles dans le domaine des systèmes d'inférence floue [4, 30]. Par ailleurs, cette question est au coeur des interfaces Hommes-Machines [24]. La thèse de [99] donne aussi différents critères d'interprétabilité. Je donne ici une synthèse basée sur ces différentes sources et sur ma propre expérience, des critères que je crois nécessaire pour rendre interprétable un modèle. Un modèle est d'autant plus interprétable qu'il est à la fois :

- **Simple** pour être appréhendable cognitivement : le nombre d'informations à appréhender simultanément doit être limité car l'empan mnésique (taille de la mémoire de travail) chez l'homme est compris entre 5 et 9 unités d'information [91]. On peut cependant par apprentissage agréger plusieurs informations élémentaires en une nouvelle unité d'information (on mémorise un court instant les 10 chiffres d'un numéro de téléphone en les rassemblant par paires ou triplets). Un modèle complexe devrait donc être décomposé en un nombre limité de sous-modèles eux-mêmes ainsi décomposés récursivement afin de permettre à l'analyste d'appréhender progressivement le modèle complet par apprentissage et agrégation successives de ses sous-modèles. La parcimonie est une propriété de certains modèles qui suivent le principe du rasoir d'Occam : *Pluralitas non est ponenda sine neccesitate*, qui dit que de deux modèles à la précision équivalente le plus simple doit être retenu. Cette propriété paraît nécessaire mais elle n'est pas suffisante pour rendre un modèle interprétable comme l'illustre la différence de principe entre les modèles parcimonieux Support Vector Machine (SVM) [102] et Relevance Vector Machine [113]. Ces deux modèles sont utilisés en classification supervisée, le premier exhibant des données voisines de la frontière des classes comme support de celle-ci, le second définissant les "centres" des classes comme support. Le SVM qui sélectionne les individus frontières donc "atypiques" paraît moins interprétable que le RVM qui synthétise les classes par leurs individus moyens ou "prototypiques". Dans un système d'inférence floue, le nombre de règles actives, ainsi que le nombre de fonctions d'appartenance à des termes linguistiques ainsi que le nombre de celles qui sont actives pour une observation, doivent être eux aussi limités [4] ou bien ces facteurs être hiérarchisées comme indiqué ci-dessus. Enfin, les modèles d'apprentissage automatique appelés "Deep Belief Network" [20]

sont des modèles dont l'architecture est constituée de multiples niveaux d'abstraction dont l'objectif est de rendre possible la modélisation de tâches de haut-niveau que l'on rencontre par exemple dans les traitements de la vision, du langage ou des mouvements chez les humains. A l'image des processus naturels d'apprentissage observés chez les animaux, l'apprentissage de ces modèles se base lui aussi sur la modélisation progressive de tâches de complexité croissante.

- **Transparent** pour tout montrer, ne rien cacher à l'analyste : les modèles composés doivent l'être par une combinaison explicite de modèles simples, explicite en ce qu'elle est montrée en tous ces termes à l'analyste et qu'aucune information sur le traitement réalisé ne lui est masquée au moins le temps nécessaire à leur modélisation mentale.
- **Prévisible** pour donner confiance : les mêmes causes ou des causes similaires doivent produire les mêmes effets ou des effets similaires. L'état du modèle ne doit pas changer sans raison explicite et les changements d'états du modèle doivent être prévisibles, attendus par l'analyste sinon il révisera ses croyances ou remettra en cause le modèle afin que cela soit le cas. Le modèle doit rester sous la maîtrise de l'analyste, il ne doit pas donner l'impression d'avoir son propre libre arbitre ni paraître chaotique (*i.e.* imprévisible au-delà d'un certain horizon temporel) si le processus générant les données ne l'est pas. Il n'est pas nécessairement déterministe, mais au moins suffisamment stable sur un horizon de temps raisonnable et face à des entrées et paramètres identiques, car c'est un signe de fiabilité et de précision qui donne confiance à l'analyste. La continuité est une propriété favorisant la prévisibilité.
- **Complet** pour signaler les événements atypiques : le modèle doit répondre, il ne doit pas rester inerte face à une nouvelle donnée. Si une donnée sort de son domaine de définition, un message doit le signaler.
- **Contextualisé** pour donner une référence : le modèle doit fournir une échelle permettant d'évaluer si la sortie est proche ou éloignée d'une limite, d'une ou plusieurs valeurs de références, *i.e.* permettant de situer cette valeur relativement à un contexte. Cette échelle peut être explicite et prendre la forme de bornes, d'un écart-type, d'un intervalle ou d'une ellipse de confiance, ou implicite si la sortie est fournie sous forme d'une proportion.
- **Sensé** pour relier à la référence : toutes les variables d'entrée tout comme les combinaisons entre variables d'entrée ou inter-

médiaires ainsi que leurs paramètres, doivent faire sens pour l'analyste. En effet, les variables de sortie ne peuvent acquérir un sens que par leur relation interprétable aux variables d'entrée déjà porteuses de sens. Le sens attribué aux variables d'entrées découle lui-même d'un enchaînement constitué empiriquement par l'analyste au fil du temps et de sa propre expérience fortuite ou contrôlée, et relié *in fine* à ses propres stimuli sensoriels internes (proprioception) et externes [33].

Les modèles descriptifs ont pour rôle l'encodage de l'information dans un référentiel devant être accessible aux représentations mentales de l'analyste afin que ses hypothèses y soient falsifiables par confrontation à ces représentations et fournissent des bases solides sur lesquelles il appuiera son raisonnement puis sa décision. De plus, afin de permettre à l'analyste d'enrichir de ses propres connaissances les modèles prédictifs et d'en interpréter plus facilement la sortie, il faut là aussi encoder les informations internes à ces modèles dans un espace accessible à ses représentations mentales. En d'autres termes, ces modèles, qu'ils soient descriptifs ou prédictifs, doivent être interprétables pour être utilisables et utiles à l'analyste en tant qu'outils d'aide à la décision (Figures 1a et 1d).

Mes travaux se placent dans le cadre d'applications d'aide à la décision qui demandent de concevoir des modèles interprétables pour que l'analyste les comprennent et précis pour que leur utilisation représente une valeur ajoutée à son expertise.

Les caractères *sensé* et *prévisible* de l'interprétabilité évoquent la notion de continuité de laquelle découle celle de connexité : sens d'une entité acquis par sa relation (connexité) à une entité faisant déjà sens ; enchaînement des causes et des effets (connexité) ; causes voisines aux effets voisins (continuité). Or cette notion est fondamentale en Topologie.

2.3 La Topologie au coeur des mes travaux

2.3.1 Un point de vue topologique

Notre capacité humaine à généraliser et prédire par nos seuls sens et notre propre expérience l'état de notre environnement aux instants suivants, se restreint aux changements apprenables de celui-ci induits par une certaine régularité ou continuité des phénomènes qui s'y déroulent et pour lesquels des causes voisines ont le plus souvent des effets voisins. Il n'y a pas de généralisation donc de prédiction possible sans des structures régulières sous-jacentes³. La

3. Les systèmes dynamiques chaotiques ne sont pas prédictibles au-delà d'un faible horizon temporel mais ils exhibent eux aussi une certaine structure et peuvent être caractérisés en

Topologie est une branche des mathématiques qui fournit les outils théoriques pour définir et caractériser ces structures.

Reprenons l’analogie avec le globe terrestre. Nous sommes contraints sans autre artifice de nous déplacer à pied à la surface de ce globe. L’ensemble de nos états physiques est soumis à cette contrainte structurelle dont la nature est topologique : qu’importe le chemin sinueux ou droit parcouru et sa longueur, nous ne pouvons aller de Paris à Sydney sans traverser un océan. Supposons que l’on mesure notre position sur le globe à l’aide de différents capteurs fixes mesurant *in fine* la distance les séparant de nous (le système de positionnement global (GPS) fonctionne sur ce principe). Ces capteurs en tant que points de mesure intégrant une grandeur physique durant un temps borné par la durée de la mesure, disloquent ces structures, ils découpent le continuum physique en un échantillon discret, un nuage de points dont chacun est caractérisé par un ensemble de mesures simultanées, ses coordonnées dans l’espace des capteurs, et dont aucun n’est plus directement relié aux autres si ce n’est dans la succession de leurs instants de mesure.

L’information topologique originelle est perdue puisque d’une structure continue nous observons un nuage de points par nature isolés dont la topologie est triviale sans autre propriété saillante que le nombre de ces points. Cependant cette information pourrait être reconstruite connaissant la nature de la fonction H de transfert des capteurs. En effet par la distance quadratique à quatre points fixes en positions indépendantes donc formant un système de coordonnées, on peut déterminer la position d’un cinquième point dans ce repère et positionner par une fonction H^{-1} de projection sur le globe les points du nuage échantillon. Supposons donc que nous disposions de quatre tels capteurs. On peut alors se convaincre que deux points voisins sur le globe sont à des distances similaires de chaque capteur donc ont des images voisines par H dans l’espace des capteurs et réciproquement. La fonction de transfert H de l’espace d’origine muni d’un repère cartésien vers l’espace à quatre dimensions des distances quadratiques mesurées par les capteurs est une application bijective continue de réciproque H^{-1} continue, c’est un homéomorphisme. On peut alors comprendre que si un point de la surface du globe à 4 voisins suivant les 4 points cardinaux, son image a pour voisins les points images de ces 4 points dans l’espace des capteurs, en d’autres termes comme un point de la surface du globe ne peut se mouvoir qu’en combinant deux directions indépendantes, par exemple Nord-Sud et Est-Ouest, il en est de même de son image dans l’espace des

particulier par la topologie de leurs attracteurs sous-jacents [52].

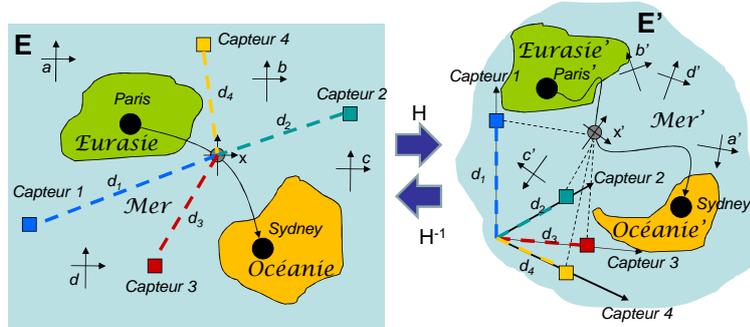
capteurs. Les points images voisins de l'image d'un point dans l'espace des capteurs sont donc situés au voisinage d'un sous-espace de dimension 2. Ce sous-espace est une d -variété, un espace topologique homéomorphe à une partie de \mathbb{R}^d . Sa dimension $d = 2$ est parfois appelée dimension intrinsèque pour la différencier de celle de l'espace ambiant des capteurs dans lequel elle est plongée. Cette 2-variété est l'image par H de la surface du globe, elle a la topologie de la sphère usuelle. S'il existe deux îles sur un lac à la surface du globe, alors leurs images par H sont aussi deux îles sur un lac sur cette 2-variété. Les distances géodésiques mesurées à la surface du globe ne sont généralement pas commensurables avec celles géodésiques mesurées sur cette 2-variété, mais on peut envisager d'analyser ce qu'il en est de la connexité : s'il n'est par exemple pas possible d'aller de Paris à Sydney sans traverser une étendue d'eau, alors ce n'est pas possible non plus en se déplaçant sur la variété image (Figure 2).

Ces caractéristiques sont des invariants topologiques, c'est-à-dire des propriétés invariantes par homotopie (préservation de la connexité) ou par homéomorphisme (préservation de la connexité, de l'orientation, de la dimension...) qui caractérisent tous les éléments d'une même famille d'espaces topologiques et qui permettent de les classer. Les groupes d'homologie sont des invariants dont on peut extraire les nombres de Betti. Les nombres de Betti sont un moyen de préciser la connexité en termes du nombre de cycles et de leur ordre (trous dans une surface ou tunnels dans une surface sans bord, cavité dans un volume...), propriété généralisable aux variétés de dimension supérieure. On se reportera aux ouvrages [57, 93, 124] pour les définitions formelles.

En résumé, l'information géométrique ne peut être préservée que si H est une isométrie (translation, rotation, symétrie) ou à un facteur d'échelle près, une similitude. Tandis que pour préserver la topologie, il suffit que H soit un homéomorphisme ou une homotopie, une famille bien plus vaste de transformations contenant les similitudes. Dans ce cas H préserve entre autres invariants topologiques l'orientation, la connexité au sens large et la dimension intrinsèque. Il est même possible de préserver l'essentiel de l'information topologique hormis la dimension intrinsèque et l'orientation si H est une homotopie, une famille encore plus large contenant les homéomorphismes. Le fondement de mes travaux s'appuie sur cette observation.

En effet, la fonction de transfert associée au système de mesure n'est pratiquement jamais une similitude où la mesure serait directement proportionnelle à la grandeur mesurée, mais elle est généralement une homotopie où pour simplifier deux grandeurs voisines

Figure 2 **Préservation de l'information topologique par homéomorphisme**



Un homéomorphisme est une application continue H de réciproque H^{-1} continue d'un espace topologique dans un autre. La fonction H transforme chaque point x de l'espace réel à gauche en un point image x' dans l'espace des capteurs à droite ayant pour coordonnées les distances quadratiques d_i entre x et un nombre fini de capteurs $Capteur i$ en position générale dans E . $E' = H(E)$ est un espace topologique homéomorphe à E . Les structures topologiques présentes dans E existent dans $H(E)$: ici les continents Eurasiens et Australiens séparés par la mer. Par contre les angles et les distances (géométrie) ne sont pas nécessairement préservés.

Figure 3a **Topologie discrète induite par une partition d'un espace topologique**

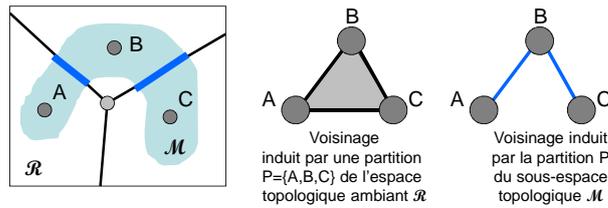


Figure 3b

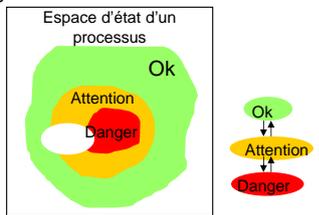
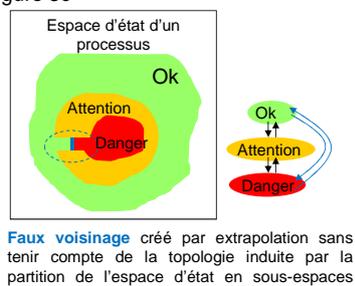


Figure 3c



produisent des mesures voisines. L'information topologique est donc généralement mieux préservée par la chaîne de mesure que l'information géométrique. C'est fondamentalement parce que le phénomène originel mesuré est au moins continu par morceaux, et que le système de mesure préserve cette continuité, qu'il est possible de construire un modèle capable de généraliser, de prédire, d'inférer par continuité (interpolation ou extrapolation) la valeur ou la classe à associer à une nouvelle observation.

Considérons maintenant un modèle capable de prédire qu'à telle position géographique se trouve de l'eau ou de la terre, et considérons que les seules mesures disponibles pour situer un point à la surface du globe sont les distances aux quatre capteurs. A chaque point de la 2-variété image de la surface du globe dans l'espace des capteurs est associé soit l'étiquette "terre", soit "mer". Dans un monde parfait, aucun point image d'un point de la surface du globe n'existe en dehors de cette 2-variété, dans la réalité, nous nous déplaçons au voisinage du sol avec une précision de quelques mètres suivant le véhicule terrestre que nous utilisons, et les distances mesurées le sont avec un certain degré de précision et sont perturbées par un bruit de mesure, les points images sont donc en pratique situés au voisinage de cette 2-variété. Les méthodes usuelles de modélisation cherchent à associer à tout point de l'espace des capteurs une étiquette de classe "terre" ou "mer" par des méthodes d'apprentissage dites "supervisées" qui utilisent les points déjà observés et leur étiquette de classe et interpolent ou extrapolent cette information au voisinage de ces points [110, 58]. Elles permettent de prédire la classe correcte en tout point du globe et font généralement quelques erreurs au voisinage de la frontière entre les classes. Ce sont des modèles prédictifs que l'on sait aujourd'hui rendre très performants [102]. Cependant elles ne transmettent pas une information importante pour l'analyste : peut-on aller de Paris à Sydney sans emprunter ni la mer ni les airs ?

La question est moins anecdotique qu'il n'y paraît. Prenons l'exemple de la surveillance d'un système critique (centrale nucléaire, avion de ligne...) dont les états de fonctionnement définissent des sous-espace de l'espace des capteurs dont il est couvert. Si Paris représente la zone de fonctionnement normale, l'océan une zone de fonctionnement critique servant d'alerte, et Sydney une zone à fort risque de panne ou d'accident, il devient primordial de savoir s'il est possible de passer directement de la zone normale à la zone dangereuse sans passer par la zone critique génératrice de l'alerte préventive, c'est-à-dire s'il existe des états normaux et dangereux voisins sur la variété dans l'espace des capteurs. De même on veut savoir s'il

existe des moyens de piloter le système pour ressortir de la zone critique sans passer par la zone dangereuse voire obtenir directement les trajectoires à suivre. La figure 3a explique la notion de topologie induite par un sous-espace décrite dans [39]. Les figures 3b et 3c illustrent comment la topologie induite par les sous-espaces d'état permet d'apporter une réponse à ces problèmes.

Ce principe est utilisé dans [122] pour la planification de trajectoire d'un bras robot (Figure 4a).

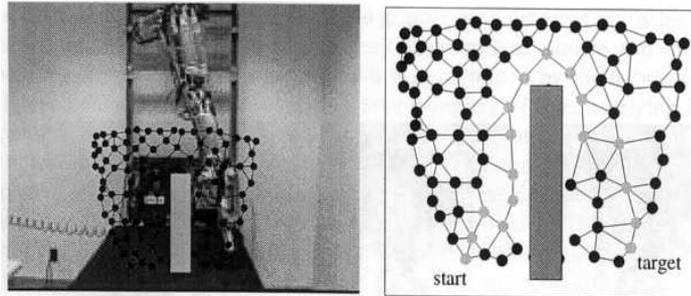
Ces requêtes font appel à la connaissance de propriétés topologiques par nature comme l'adjacence, le chevauchement, l'inclusion, l'entrelacement, la séparation ou la continuité que ces structures peuvent avoir, et requièrent un modèle capable d'extraire cette information à partir d'un échantillon de ces sous-espaces d'état. Nous avons proposé des modèles permettant cela dans [15, 10, 46] dont nous nous servons pour l'analyse exploratoire de données.

Par ailleurs, les similitudes sont nécessaires pour préserver la densité de probabilité ou la géométrie des structures intrinsèques au phénomène observé, tandis que les homotopies suffisent à préserver l'essentiel de la topologie de ces structures. Or les homotopies forment une classe bien plus large de transformations que les similitudes, il est donc *a priori* beaucoup plus probable que la transformation réalisée par la chaîne de mesure fasse partie de la première classe que de la seconde. Ainsi lorsque l'on extrait les informations topologiques, statistiques et géométriques à la sortie de la chaîne de mesure, il est aussi beaucoup plus probable que les informations topologiques aient été moins déformées par cette chaîne - qu'elles soient plus fidèles à la réalité - que les autres. Donc outre son utilité, l'information topologique est plus probablement fiable et moins dépendante de la chaîne de mesure, que celles statistiques ou géométriques. J'illustre maintenant d'autres cas typiques pour lesquels cette information topologique est essentielle.

2.3.2 Autres exemples de l'utilité de la topologie

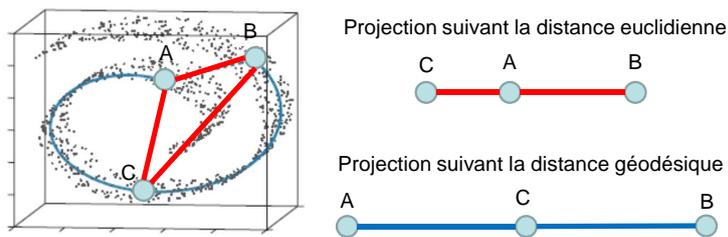
Reprenons l'analogie du globe terrestre et considérons le cas de points sans étiquette dont certains sont sur le continent australien et les autres sur le continent européen, clairement séparés par l'étendue océanique exempte de points. Peut-on détecter que l'image des ces points dans l'espace des capteurs forme deux classes distinctes, ou du point de vue topologique que l'image de la population génératrice de ce nuage est constituée de deux composantes connexes (Figure 2)? C'est un problème de classification automatique, de segmentation ou de partitionnement que l'on retrouve par exemple en marketing

Figure 4a **Topologie de l'espace d'état d'un bras robot**



Les positions prises par l'extrémité d'un bras robot sont observées par deux caméras. Un modèle de mémoire associative et topologique (Zeller, Sharma et Schulten, 1996) permet d'associer à chaque position du bras dans le champ des deux caméras, la valeur des angles des différentes articulations. L'information topologique est aussi encodée par le modèle sous forme d'un graphe (topologie induite par le sous-espace atteignable de l'espace ambiant), on sait ainsi quelle position est voisine de quelle autre à la fois dans le champ des caméras et dans l'espace des angles des articulations. On peut ainsi piloter le bras robot d'une position à une autre (*start* et *target* à droite) du domaine atteignable par une recherche de plus court chemin sur le graphe. Les obstacles sont évités car détectés implicitement comme l'absence d'observations dans cette région de l'espace d'état, ce qui induit une absence de liens dans le graphe.

Figure 4b **Projeter les données en limitant les distorsions topologiques**



Pour projeter un nuage de points, il est plus important de préserver les distances géodésiques, le long des structures sous-jacentes au nuage de point, que les distances de l'espace ambiant, « à vol d'oiseau ».

lorsqu'il s'agit de segmenter les clients afin de leur adresser une campagne de communication ciblée.

Si l'on souhaite maintenant dans une approche descriptive, visualiser ces points et leur structure, il est intéressant de les projeter dans un plan. Les méthodes de projection cartographiques sont fondées sur la connaissance de la nature sphérique du globe. Mais si nous ignorons a priori la topologie de cette structure (*e.g.* supposons-la torique) et ne disposons que d'un nuage de points image d'un échantillon de cette structure dans l'espace à 4 dimensions des capteurs, il devient important de reconstituer un modèle de même topologie que cette structure à partir du nuage de points, pour tenir compte de cette information (*e.g.* calculer des distances géodésiques) lors de la projection [73, 108] (Figure 4b). De même il peut être pertinent de débruiter les observations perturbées par le système de mesure en les projetant sur leur structure sous-jacente dans l'espace des capteurs. Ce qui correspond à la reconstruction préalable de cette structure ignorant le bruit de mesure puis à la projection des observations bruitées sur celle-ci. Enfin dans ce même cadre, il est souvent plus efficace de réaliser des traitements dans un espace de faible dimension que dans l'espace multidimensionnel des capteurs. On peut réduire par projection la dimension de l'espace ambiant, *i.e.* projeter le nuage de points dans un sous-espace de l'espace ambiant, simplement connexe et de dimension intrinsèque homogène. Cependant, une telle projection ne tient pas compte de la dimension intrinsèque potentiellement variable localement de la structure sous-jacente au nuage de points, ni de sa géométrie potentiellement non linéaire, ni d'autres particularités topologiques, l'espace de projection sera alors d'une dimension trop grande préservant plus d'information que nécessaire, ou trop petite perdant une partie de l'information, sans qu'aucune dimension ne soit optimale. Par exemple, une bouteille de Klein ne peut être plongée sans recoupement que dans un espace de dimension supérieure à 3, bien que sa dimension intrinsèque soit 2, si l'on peut modéliser cette variété à partir d'un échantillon et réaliser les traitements au sein d'un espace métrique s'appuyant sur cette structure modèle seule non plongée, indépendamment d'un espace ambiant (Figure 5), on peut alors réduire la dimension des entrées du système de traitement postérieur sans perte d'information topologique. Par exemple, si cette structure est un complexe simplicial, on peut utiliser les coordonnées barycentrique pour cela.

Considérons maintenant un autre exemple lié à nos déplacements dans un bâtiment. Tous les usagers de ce bâtiment construisent mentalement une carte de celui-ci à partir de leur propre expérience des lieux. Le nombre, la nature et la sensibilité des capteurs de chacun

sont différents ; leur expérience subjective ayant depuis leur naissance imprégné leur tissu neuronal traitant et mémorisant ces informations est différente ; leur fréquence de passage dans chaque pièce et couloir, leur façon de parcourir les lieux sont différentes. Pour toutes ces raisons, la représentation mentale que chacun s'est forgée du bâtiment est radicalement différente de celle des autres - elle est profondément subjective. Pourtant, toutes ces représentations mentales possèdent un invariant commun avec le bâtiment lui-même : un invariant topologique encodant la relation d'adjacence physique des pièces et des couloirs, la substance élémentaire et fondamentale, nécessaire et suffisante pour permettre à chacun de naviguer dans le bâtiment en décomposant un trajet complexe en une succession de trajets simples (Figure 6).

Par analogie, on peut ainsi comparer en termes de leurs caractéristiques topologiques, deux nuages de points issus d'un même système réel observé par deux systèmes de mesure distincts. Par leur nature topologique, ces caractéristiques rendent commensurables deux phénomènes *a priori* techniquement non comparables dans le cadre statistique classique, car observés dans deux espaces de description radicalement différents.

Prenons enfin un dernier exemple où il s'agit d'étiqueter tous les points observés ne connaissant l'étiquette que de quelques-uns. Dans ce cas les points sans étiquette sont connus au moment de la construction du modèle, seule leur étiquette ne l'est pas. C'est le cas pratique pour lequel l'étiquetage effectué par un expert est une opération longue et coûteuse, et où il est souhaitable de procéder à un étiquetage automatique des points non étiquetés à partir des quelques points dont l'étiquette est déjà fournie. Les méthodes d'apprentissage semi-supervisé exploitent les points étiquetés mais à la différence des méthodes supervisées qui n'exploitent que ceux-ci, elles modélisent aussi la structure sous-jacente aux points non étiquetés le long de laquelle elles propagent les étiquettes de proche en proche (Figure 7). Ces méthodes sont plus performantes que celles qui ne tiennent pas compte de cette structure lorsqu'elle existe [19]. Le modèle, généralement un graphe, doit permettre d'approcher au mieux les distances géodésiques [1], pour que ces approches soient efficaces.

2.3.3 Le caractère essentiel et générique de la Topologie

La géométrie précise les propriétés métriques d'un espace topologique prédéfini. En effet, avant de mesurer la longueur du chemin entre deux points, il faut déterminer si ce chemin existe dans l'es-

Figure 5

Un espace de projection de dimension hétérogène

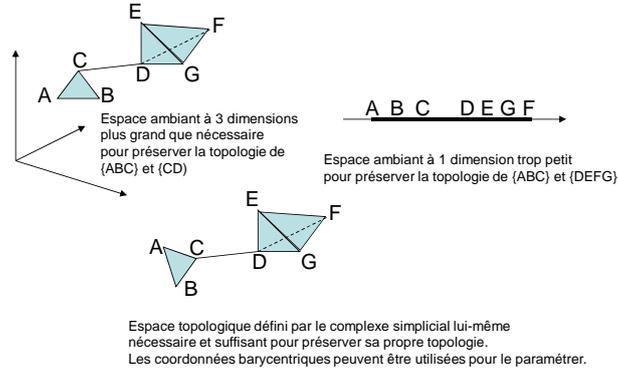
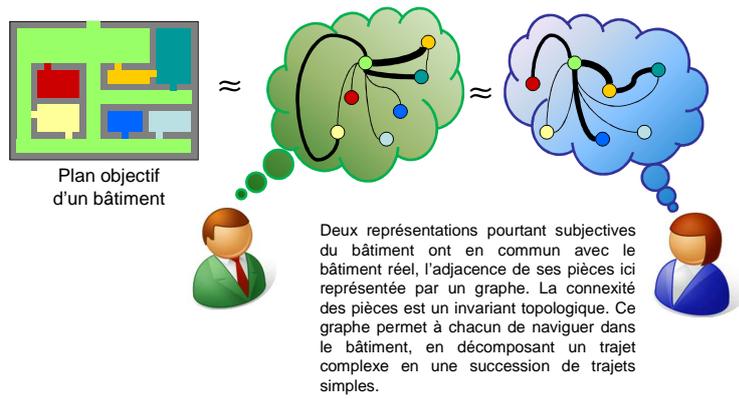


Figure 6

Un invariant topologique très pratique

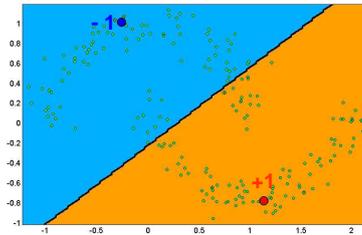


pace considéré. En ce sens, la topologie caractérise les objets que l'on perçoit avec plus de généralité : une table est plus génériquement décrite comme 4 composantes unidimensionnelle (les pieds) connectées sur le bord et sur la même face d'une composante bidimensionnelle (le plan) que par un ensemble de mesures géométriques précises d'angles, de longueurs, de surfaces, de volumes ou de densités qui viendront *in fine* compléter l'information topologique primordiale. Par ailleurs, les perceptions visuelles de nature topologique (séparation, connexité, continuité...) sont préattentives [119], traitées très rapidement et automatiquement par les premières couches neuronales du système visuel allégeant d'autant la charge cognitive. Transmettre l'information topologique sous une forme graphique fidèle est donc un moyen de focaliser la charge cognitive de l'analyste sur le traitement des informations non topologiques résiduelles ou celles externes aux données qu'il faut leur corrélérer pour leur donner un sens. De plus, les neurones du cortex visuel s'auto-organisent sous l'influence des stimuli sensoriels [98, 68], de telle sorte que des informations voisines sont encodées par des régions voisines du cortex, qui intègre donc nativement une part d'information topologique au sein de ses connections dendritiques. C'est une preuve que notre système visuel préserve une part de l'information topologique perçue de notre environnement. Je pense que cet encodage topologique n'a pas résisté par hasard à la pression de sélection Darwinienne des espèces par leur environnement, et qu'il joue un rôle important dans l'appréhension et le traitement analytique de ce que nous percevons de notre environnement, efficace pour notre survie et nos prises de décisions. D'autres auteurs [96, 36] ont aussi proposé que les structures grammaticales du langage sont elles-mêmes basées sur des structures topologiques émergeant de nos perceptions primitives du temps et de l'espace. Je suis donc persuadé que la part topologique de l'information contenue dans les données observées, transmise à notre cortex cérébral par la chaîne de mesure externe puis par le système visuel, joue un rôle essentiel dans la boucle sensori-motrice [122, 115] qui pilote l'interaction analytique [51] et finalement mène à l'interprétabilité de l'information globale.

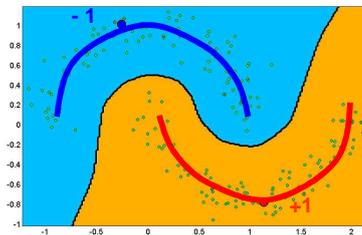
Ainsi l'information topologique propre au phénomène physique observé est seule capable de persister le long de la chaîne de mesure qui s'étend des capteurs au système visuel. Notre système visuel lui-même la préserve au moins en partie et la transmet au cortex cérébral signe que la part topologique de l'information est essentielle à son traitement cognitif. Si le langage du cerveau est avant tout topologique, alors je crois pertinent que les machines parlent aussi ce langage afin de faciliter nos interactions avec elles.

Figure 7

Variétés de classes d'un classifieur



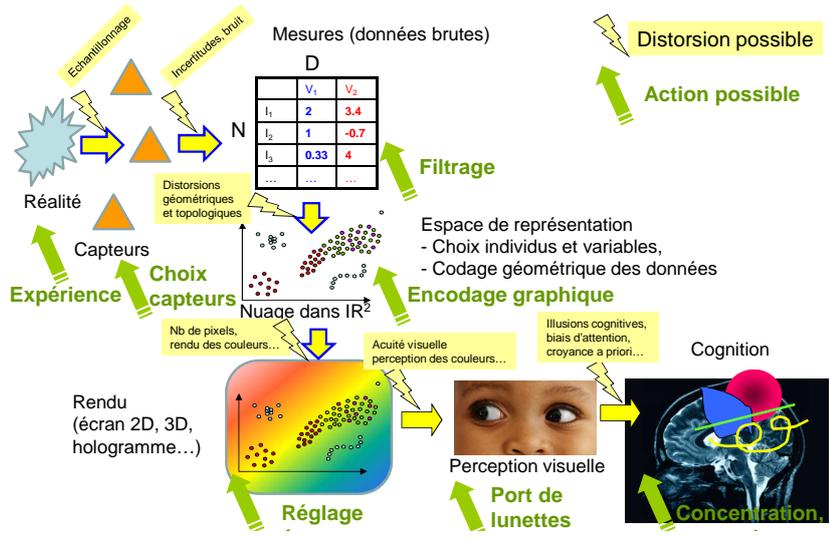
Approche supervisée classique
 Seules les données étiquetées (points bleu et rouge) sont considérées durant la phase d'apprentissage. L'espace topologique induit par les données non étiquetées n'est pas pris en compte, la frontière des classes obtenue produit une mauvaise capacité de généralisation.



Approche semi-supervisée
 L'espace topologique induit par les nombreuses données non étiquetées (points verts) est considéré. Il contraint l'apprentissage de la frontière des classes pour obtenir une meilleure capacité de généralisation.

Figure 8

Chaîne de traitement pour la visualisation



Aussi, la conviction qui guide mes travaux est que cette information topologique est fondamentale pour construire des outils d'aide à la décision interprétables, capables de rendre l'orthèse que constitue cette chaîne transparente à l'analyste, afin que ces outils deviennent des prolongements qui lui soient cognitivement intégrés, comme le piano fait corps avec le pianiste virtuose.

2.4 Voir pour comprendre

Les informations issues des mesures "brutes" ou de traitements sur elles, sont très généralement portées à la connaissance de l'analyste par des moyens graphiques, le sens visuel étant le plus développé chez l'homme pour percevoir et analyser son environnement.

2.4.1 Du signal à la décision

La chaîne de transmission entre les données brutes et le système cognitif de l'analyste comporte de nombreux maillons (Figure 8) : (A) les capteurs ; (B) les données brutes ; (C) la représentation ; (D) le rendu ; (E) le médium physique ; (F) le système visuel ; (G) le système cognitif.

Une partie de cette chaîne qui s'étend du maillon (B) au maillon (D), est appelée "visualization pipeline" et a été décrite dans [29] dans le cas de la modalité visuelle, mais elle pourrait se décliner pour d'autres modalités sensorielles. Dans cette chaîne, les données brutes (B) issues de mesures de grandeurs physiques de l'environnement (dont la chaîne elle-même fait partie) par différents capteurs (A), subissent une transformation (C) par les outils de fouille de données et par l'action de l'analyste qui choisit des méthodes de traitement descriptives ou prédictives, et les individus et les variables sur lesquels les appliquer. Les résultats de ces traitements sont encodés sous forme de primitives graphiques (points, lignes, couleur, position...), qui sont envoyées vers un transducteur (D) qui leur donne leur forme physique (illumination des pixels d'un écran...) que nos sens (F) perçoivent à travers le médium physique (air, eau...) et d'éventuelles corrections intermédiaires (lunettes) (E). Enfin notre système cognitif (G) interprète ces informations pour agir sur l'ensemble de la chaîne ou sur d'autres systèmes pour refermer la boucle sensori-motrice.

Chaque point de cette chaîne de mesure est soumise à des perturbations indépendantes les unes des autres, altérant l'information initiale :

- (A) Les capteurs échantillonnent l'environnement à la fois en termes de nombre d'individus (fréquence et durée d'acquisition) et de nombre de variables (le nombre de capteurs) rompant la continuité éventuelle du phénomène mesuré. Ces mesures peuvent être imprécises, incertaines, manquantes, erronées, perturbées. Elles subissent des filtrages et amplifications diverses au niveau du système d'acquisition de chaque capteur.
- (B) Les données brutes sont une projection de ces mesures dans l'espace numérique discret de l'ordinateur (quantification). Elles peuvent ne pas être stockables (flux continu) du fait d'une fréquence d'acquisition ou d'un nombre de capteurs trop important compte tenu du matériel disponible, ou être stockées de manière distribuée, dans ces cas, la gestion des données peut nécessiter leur compression, entraînant des pertes ou des altérations de l'information.
- (C) L'analyste emploie des outils d'analyse qui fournissent un point de vue particulier de ces données brutes qui souvent en masque une partie (sélection d'un modèle donc d'un espace de représentation, sélection des individus et des variables traitées et projetées dans cet espace), ou n'en conserve qu'une synthèse avec perte (moyenne d'un ensemble au lieu de l'ensemble complet). Les primitives géométriques et les variables graphiques utilisées pour encoder ces données projetées, participent elles aussi à la dégradation de l'information (choix de la taille ou de la couleur et d'une échelle associée, pour coder une valeur numérique) suivant l'efficacité avec laquelle elles seront perçues par l'analyste.
- (D) Le transducteur physique (écran) de ces informations numériques a aussi ses propres contraintes (rendu des couleurs, luminosité, nombres de pixels, fréquence de rafraichissement...).
- (E) Le médium physique traversé de l'écran à l'oeil peut lui-aussi déformer ou dégrader le signal (turbidité, changement d'indice de réfraction par les fluctuations thermiques...).
- (F) Notre système de perception visuelle est lui-même un transducteur imparfait dont les caractéristiques fines sont propres à chaque individu, en termes de sensibilité aux couleurs et aux mouvements, d'acuité visuelle, de sensibilité aux illusions perceptives (contraste, distance, perspective...).
- (G) Enfin, au niveau cognitif la qualité de notre modèle de l'environnement dépend de notre capacité à explorer, à détecter des signes, à maintenir notre attention, de notre propension à sur-interpréter les informations du fait de croyances a priori sur ce qui est attendu (illusions cognitives)....

Chacune de ces perturbations peut-être réduite en améliorant la technologie (capteurs plus précis, plus variés, plus nombreux, numériseurs, écrans...), la méthode (traitement du signal, modèles mathématiques) ou par l'entraînement de l'analyste (apprendre à observer, à analyser, à s'affranchir des a priori...).

Le domaine de l'analyse exploratoire se restreint généralement au domaine s'étendant de la collecte des données brutes numériques en entrée (B) à leur représentation qui doit être rendue sur l'écran (C). Son objectif est la recherche d'espaces de représentation pertinents (les paires de lunettes chaussées par l'analyste) qui mettent en avant telles ou telles caractéristiques des données. En amont se situent les domaines de la physique des capteurs, de l'électronique d'acquisition et de traitement du signal, en aval le domaine des interfaces homme-machine qui doivent traduire l'espace de représentation en manifestation physique, et des sciences cognitives qui étudient la perception et la cognition chez l'homme et permettent d'améliorer tous les éléments amonts en conséquence. Cette chaîne de visualisation est en fait incluse dans une boucle de rétroaction, l'analyste percevant ces informations modifie interactivement dans la mesure de ses possibilités, l'espace de représentation pour explorer les données, ajoute ou modifie les capteurs et les traitements associés, agit sur le phénomène physique observé, choisit d'autres interfaces, modifie sa propre représentation mentale de cet environnement, ses hypothèses et les moyens de les valider. Il s'aide de supports physiques ou numériques pour tracer son parcours analytique et en réalise une synthèse pour la soumettre à la critique de ses pairs ou assister le décideur dans sa prise de décision. Cette fouille visuelle interactive appelée Visual Analytics par les anglo-saxons est le nom actuel que porte ce domaine de recherche. Il est pluridisciplinaire puisqu'il s'intéresse à la chaîne complète [67].

2.4.2 Deux modes de représentation graphique

On peut distinguer schématiquement deux modes de représentation graphique de l'information issue des traitements [119] :

- **Le mode symbolique ou conventionnel** : on peut représenter une valeur numérique issue d'un modèle descriptif ou prédictif (une moyenne, une variance, une corrélation, la qualité d'un estimateur, son degré d'incertitude, le taux de vrais positifs, la classe d'appartenance...) sous forme de symboles représentant ce nombre, ou bien représenter par des lettres un texte ou encore par des noeuds et des liens, des structures topologiques plus complexes (arbres, graphes, hyper-graphes...).

L'information par sa nature symbolique est alors très synthétique et sa lisibilité est très robuste aux distorsions induites par leur rendu graphique, dans la limite cependant de la résolution de l'écran d'affichage, de la disposition sans chevauchement des objets sur l'écran et de la charge cognitive induite par la quantité d'objets affichés. Elle est interprétable objectivement à conditions que les analystes qui partagent cette information en partagent et comprennent également le référentiel (alphabet, sémantique...). Elle est donc un moyen privilégié de communication limitant les ambiguïtés d'interprétation pré-attentives et s'affranchissant des conditions perturbées de transmission physique de la machine à l'analyste et *in fine* des analystes entre eux. La complexité des structures modélisées est absorbée et intégrée par le modèle dont l'analyste ne perçoit sous forme symbolique que le résultat fruste de la mesure. Mais si le modèle est interprétable par l'analyste, cette information symbolique peut suffire à faire sens.

- **Le mode analogique ou perceptuel** : on peut représenter des informations beaucoup plus complexes (formes, signaux, images...) en utilisant des variables graphiques plus diverses (position, couleur, forme, taille, texture...). Le résultat transmis à l'analyste est alors plus riche, par exemple si l'on songe aux millions de pixels d'une image, leur représentation symbolique sous forme de trois tableaux de nombres est incompréhensible par l'analyste alors que leur encodage sous forme de pixels colorés disposés en grille est plus facilement interprétable et communique donc beaucoup plus d'informations. Il est aussi plus sensible aux distorsions induites par son système visuel. Mais cette représentation peut être bénéfique si elle exploite les facultés pré-attentives de ce système visuel à appréhender efficacement une telle masse d'informations non symboliques. Elle peut aussi être plus attractive pour l'analyste par nature habitué à interpréter des formes géométriques perçues visuellement, et lui permet plus facilement d'agréger mentalement d'autres informations, externes à celles représentées pourvu qu'elles s'expriment facilement dans le même référentiel visuel. Cependant, le prix à payer est la subjectivité de la représentation, puisque l'abstraction symbolique support du raisonnement analytique est effectuée mentalement par l'analyste avant d'être éventuellement communiquée à ses pairs.

En pratique, sauf à représenter un unique symbole, le mode symbolique invoque nécessairement le mode analogique au moins dans

sa dimension spatiale, par exemple, les lettres d'un texte sont habituellement disposées dans l'ordre imposé par les mots et phrases du texte et suivant des lignes droites parallèles pour que le texte soit lisible. Les approches sont donc hybrides mêlant les deux modes de représentation en associant un traitement cognitif de haut niveau à une perception pré-attentive de bas niveau, comme par exemple le simple tableau de nombres, la localisation de symboles sur une carte ou encore la légende traduisant une primitive graphique (longueur, couleur, texture...) en symbole.

2.4.3 L'inférence visuelle

Afin de transmettre l'information topologique, je cherche des modèles capables de – ou permettant mentalement à l'analyste de – construire un espace topologique probabilisé, population plausible de l'échantillon observé. Or les représentations graphiques ont immédiatement une portée géométrique, statistique et topologique que le système de perception visuelle perçoit naturellement. En effet, considérons le cas classique d'un nuage de points représentés graphiquement dans un repère cartésien à deux dimensions (Figure 9). Dans le cadre statistique, une population (l'espace topologique probabilisé) est observée au travers d'un échantillon (le nuage de points) à partir duquel on veut la caractériser. Les points de ce nuage sont les individus, définis par leurs valeurs suivant 2 variables aléatoires représentées ici par les axes du repère cartésien. Je porte trois regards subjectifs et complémentaires sur ce nuage :

- le premier est de nature *statistique*, on détecte la densité plus ou moins forte des points dans l'espace et les points atypiques ou prototypiques (modaux) de ce point de vue ;
- le second est de nature *topologique*, le nuage de points se découpe visuellement en formes plutôt ponctuelles, linéiques ou surfaciques, connectées ou non, percées de trous ou non, emboîtées ou non, superposées ou non... ;
- le troisième est un regard de nature *géométrique* sur les points, leurs distances relatives, et sur ces formes, leur courbure, leur taille relative, leur convexité...

De ces trois regards, seul le regard géométrique porte sur le nuage de points en tant qu'ensemble de points individuels distribués dans le plan. Par contre tous ont une portée immédiatement structurelle : de la densité du nuage de points émergent des formes caractérisées par leur topologie et subséquemment leur géométrie. Pourtant la seule information objectivement affichée est la position absolue des points dans le repère, de laquelle se déduisent leurs positions rela-

tives et leurs distances euclidiennes relatives. Ces trois caractères de la population - statistique, topologique et géométrique - sont donc inférés par notre système visuel à partir du nuage de points et définissent pour l'analyste les paramètres d'un modèle mental qu'il se forge de la structure sous-jacente aux données telles qu'il les perçoit, l'idée qu'il se fait de la population dont le nuage de points est un échantillon.

2.4.4 Le principe de fiabilité pour permettre l'inférence

On peut voir les méthodes de représentation graphique d'information comme consistant à corrélérer au mieux la structure topologique, statistique et géométrique que l'on souhaite transmettre, à celle intrinsèque au substrat de rendu graphique, généralement un écran à 2 dimensions simplement connexe (propriété topologique) constitué d'une matrice de pixels de taille finie et de distribution spatiale plane et régulière (propriété géométrique), et sans pixels superposés, donc uniforme (propriété statistique), muni de trois dimensions supplémentaires par la couleur des pixels. Et ceci afin que d'un point de vue topologique ou qualitatif, les objets qui apparaissent immédiatement (pré-attentif) comme voisins donc similaires à l'écran (pixels voisins dans la matrice et/ou couleur voisine) soient aussi le plus souvent effectivement voisins suivant la structure à transmettre (qu'il s'agisse de la topologie linéique d'un signal ou d'un texte, ou de celle plus complexe d'un arbre, d'un graphe ou d'une forme à deux dimensions ou plus...), et que d'un point de vue géométrique et statistique ou quantitatif, le degré de similarité ou la densité perçus reflètent ceux originels. La charge cognitive (non pré-attentive) de l'analyste est diminuée d'autant plus que ce principe est respecté, car il limite l'effort supplémentaire nécessaire pour reconstituer mentalement les propriétés topologiques des structures (et au second plan leurs propriétés géométriques et statistiques), qui ne seraient pas transmises par ce biais. Les modes de représentation graphiques qui maximisent cette corrélation sont donc les vecteurs privilégiés pour transmettre à l'analyste une information sur la structure en particulier topologique des données et permettre à celui-ci de l'explorer. Ces modes fournissent des représentations graphiques fiables, desquelles l'analyste peut inférer légitimement des propriétés des données d'origine. On remarque que ce principe de fiabilité est consistant puisqu'il s'applique au cas trivial de données directement définies par leur représentation graphique (représentation noeud-lien d'un graphe par exemple), et qu'il est optimal dans ce cas puisque de telles données sont *ipso facto* représentées sans distorsion. La figure 10 illustre le

principe de fiabilité.

J'énonce ainsi un principe élémentaire que devraient respecter les modèles de représentation graphique afin de préserver au mieux l'information à transmettre à l'analyste. Ce principe de fiabilité me sera utile par la suite, car les nuages de points objets de mes traitements sont observés dans des espaces de dimension supérieure à 2 et ne peuvent donc généralement pas se représenter en l'état dans un plan sans perte d'information. La sphère usuelle, 2-variété simplement connexe sans bord, fournit un exemple classique de ce problème de distorsion (invoquer pour solution un espace de représentation à 3 dimensions spatiales ne ferait que repousser légèrement les frontières de l'ensemble très réduit des formes représentables sans distorsion dans un tel espace).

Je m'intéresse par la suite au cas des représentations graphiques dans un plan, d'une part pour la raison pratique que les artefacts d'affichage sont encore pour quelques temps des écrans dont l'espace des variables graphiques se limite à deux dimensions d'espace, et trois de couleur auxquelles on peut encore adjoindre la dimension temporelle, d'autre part parce que l'ajout d'une troisième dimension spatiale à ces artefacts ne change pas fondamentalement les problèmes de fond auxquels je m'intéresse.

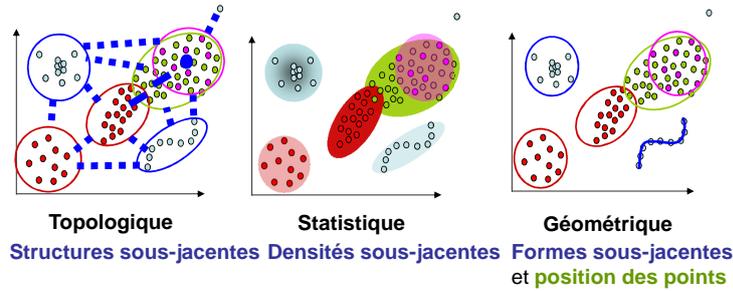
2.4.5 En résumé

Nous arrivons au terme du tronc contextuel commun expliquant les travaux de recherches que j'ai engagés.

En résumé, je cherche des modèles en mesure d'assister l'analyste dans sa prise de décisions. Pour cela, ces modèles doivent être interprétables. Certains critères d'interprétabilité évoquent des notions topologique de connexité et de continuité. L'information topologique est la plus capable de persister au travers de la chaîne de mesure, des capteurs aux neurones. Elle est aussi interprétable car perçue pré-attentivement par notre système visuel et utilisée par notre cortex cérébral pour organiser les informations qu'il perçoit et probablement contribuer à son activité analytique. Enfin elle est utile à la résolution de nombreux problèmes pratiques posés à l'analyste. Que le modèle qui la transmet soit descriptif ou prédictif, le passage par un mode de représentation graphique interface de la machine à l'homme est encore pour longtemps incontournable. Cette représentation graphique est en mesure de préserver en grande partie l'information topologique au même titre que les autres éléments de la chaîne de mesure, et permet une perception pré-attentive de l'information topologique allégeant d'autant la charge cognitive. La

Figure 9

Trois points de vue complémentaires

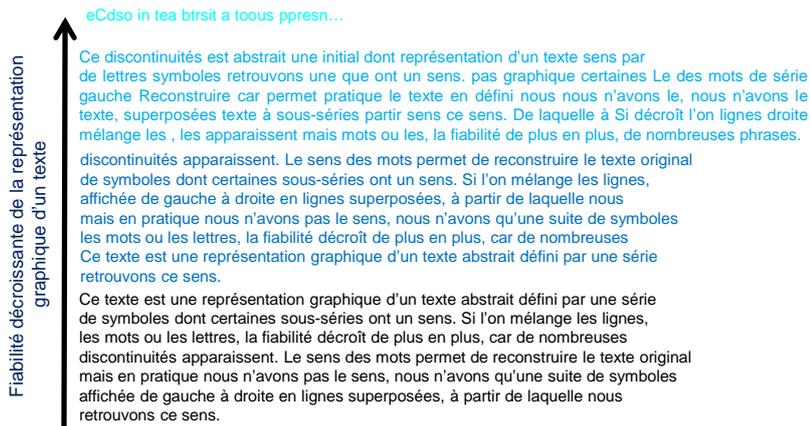


Inférence sur la population (statistique inférentielle)

Mesure sur l'échantillon (statistique descriptive)

Figure 10

Le principe de fiabilité



Le principe de fiabilité impose que les entités perçues comme similaires dans leur représentation graphique (par leur position, leur forme, leur couleur...) soient le plus souvent possible, effectivement similaires à l'origine. Ici, on prend l'exemple d'un texte dont on détériore la fiabilité (du bas vers le haut)

problématique subsistante est donc la suivante : comment par le biais de modèles prédictifs ou descriptifs, extraire l'information topologique et la transmettre sous une forme *in fine* graphique immédiatement perceptible par l'analyste. Des réponses à cette problématique d'extraction d'information topologique découleront les moyens d'explorer scientifiquement la problématique du lien entre topologie et interprétabilité et de développer des méthodes efficaces exploitant l'information topologique pour assister l'analyste dans sa prise de décisions lorsqu'il est confronté à des observations multivariées issues de la mesure de phénomènes réels.

On notera que lorsque je considère les modèles prédictifs, je ne considère pas l'information topologique comme une variable à prédire à partir d'une base d'exemples qui contiendrait des individus ou groupes d'individus déjà munis d'une telle caractéristique. Mais je cherche à m'appuyer sur ces modèles prédictifs soit pour transmettre à l'analyste l'information topologique qu'ils induisent sur les classes prédites, soit parce que prendre en compte cette information au sein du modèle en plus d'en expliquer la structure, peut en améliorer les performances.

J'ai donc exploré deux principaux axes de recherche, l'un descriptif est centré sur les méthodes de visualisation par projection, l'autre prédictif se base sur des méthodes de modélisation automatique dans l'espace multidimensionnel.

2.5 Problématique scientifique et axes de recherches

2.5.1 La problématique scientifique principale

Je considère que l'espace des capteurs préserve au moins en partie l'information topologique originelle, et que cette information est potentiellement au fondement de la résolution de nombreux problèmes décisionnels ainsi qu'un support probablement essentiel de l'interprétabilité, la problématique scientifique principale que j'explore est alors d'extraire et de transmettre cette information à l'analyste. Et puisque les observations forment un nuage de point et non un ensemble de variétés continues qui porteraient en elles directement cette information topologique, il faut donc soit définir des méthodes capables d'estimer automatiquement les invariants topologiques de la population à partir de cet échantillon et les transmettre à l'analyste (approche prédictive), soit permettre à l'analyste de réaliser mentalement cette estimation à partir d'une représentation adéquate de l'échantillon (approche descriptive). Dans les deux cas, le mode de représentation graphique est le vecteur privilégié de transfert de

l'information topologique à l'analyste. La figure 11 illustre les deux axes de recherches que j'ai suivis.

2.5.2 Le paradigme de la visualisation multidimensionnelle *in situ*

Mon premier axe de recherche concerne les représentations graphiques et se place naturellement dans le cadre des modèles descriptifs. Ces représentations doivent permettre une inférence sur la population réelle, dont on obtiendra la caractérisation la plus fidèle à partir des données brutes, *i.e.* les données contenant l'information qui est *a priori* la plus complète disponible sur cette population. La question fondamentale qui se pose est donc celle de l'authenticité de cette représentation : en quoi ce que l'on observe est lié aux données originelles ? Quelle information est préservée ? Dans quelle mesure l'est-elle ? Comment cette information sur la qualité de l'information représentée est-elle portée à notre connaissance ? Que peut-on inférer sur la population originelle, à partir de la vue de cette représentation ?

Je me suis intéressé à ces questions dans le cas particulier des méthodes de représentation graphique par projection non linéaire dont j'ai décrit les différentes distorsions qu'elles induisent presque toujours. J'ai montré en particulier l'inanité de toute inférence basée sur ces techniques lorsque ces distorsions sont ignorées, et proposé un nouveau paradigme de visualisation *in situ* et une implémentation de ce paradigme rendant possible l'inférence à partir de telles projections à condition d'en compléter la représentation.

2.5.3 L'apprentissage automatique de la topologie

Mon second axe adresse la problématique complémentaire du premier, explorant l'extraction de l'information topologique par des méthodes automatiques :

- La première approche se situe dans le cadre des modèles descriptifs. Elle modélise la connexité des classes de données étiquetées directement *in situ*, dans l'espace d'origine multidimensionnel, par un graphe de synthèse appelé "graphe des classes" qui peut alors être représenté graphiquement sans distorsions suivant le paradigme de la visualisation *in situ*.
- La seconde approche s'intègre à un modèle génératif soit descriptif (modèle de densité de probabilité ou de classification automatique) soit prédictif (modèles de discrimination), qui permet d'extraire une structure topologique vraisemblable de la population *in situ* dont le nuage de point est un échantillon.

Figure 11 Deux axes de recherches

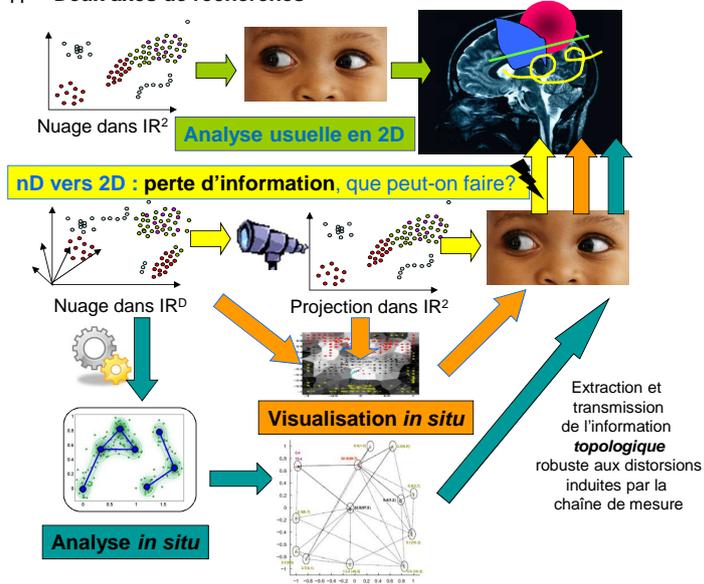
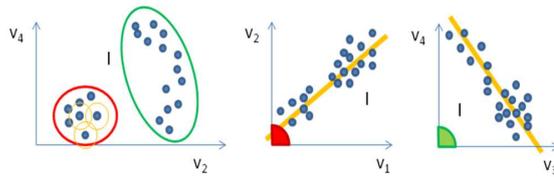


Figure 12 Exemple d'inférences basées individus et basées variables



Les individus I (points bleus) sont positionnés dans l'espace $V=(V_1, V_2, V_3, V_4)$. A gauche, la métrique euclidienne (cercles de rayon unité en orange) est utilisée pour mesurer la similarité entre individus dans l'espace (V_2, V_4) . A droite, les individus sont positionnés dans (V_1, V_2) et (V_3, V_4) . Un exemple d'inférence sur I sachant (V_2, V_4) peut être que deux **classes d'individus** apparaissent (ellipses verte et rouge) si l'on connecte les points distants de moins de 2 unités de longueur. Un exemple d'inférence sur V sachant I peut être que les variables V_1 et V_2 sont fortement corrélées, et que V_3 et V_4 le sont aussi (lignes oranges) tandis que V_2 et V_4 ne le sont pas, donc l'existence de deux **classes de variables** (V_1, V_2) (coin rouge) et (V_3, V_4) (coin vert) est plausible.

Ce type de modèle trouve des applications non seulement en analyse exploratoire de données où il fournit une extension probabiliste au graphe des classes, mais il est aussi une solution de principe à l'ensemble des applications décrites en introduction.

Afin de mener à bien ces travaux, j'ai été amené à l'étude et l'utilisation de méthodes issues de différents domaines des mathématiques et des sciences de l'information, en particulier : la Géométrie, l'Analyse, la Topologie, l'Apprentissage Statistique, la Fouille de Données, et la Visualisation d'Informations.

2.6 Aperçu de mes principales contributions

Cette synthèse s'appuie principalement sur les publications suivantes

- [12] **WinSitu, un nouveau paradigme pour l'analyse exploratoire de données basée sur des projections.** Michaël Aupetit. Apprentissage et Visualisation - Revue des Nouvelles Technologies de l'Information (RNTI A4) pp. 79-98, Editions Hermann (2010).

Résumé : Présentation des critères d'authenticité et de fiabilité comme fondements du paradigme *WinSitu* de visualisation *in situ*. Une version anglaise est en cours de révision pour le journal "Knowledge Discovery and Data Mining".

- [10] **Visualizing distortions and recovering topology in continuous projection techniques.** Michaël Aupetit. Neurocomputing 70(7-9) : 1304-1330 (2007).

Résumé : Comment détecter visuellement les défauts locaux d'une représentation graphique à deux dimensions d'un nuage de points multi-dimensionnels et reconstruire interactivement la topologie intra et inter classes de ce nuage. J'ai nommé ultérieurement ProxiViz cette méthode.

- [80] **CheckViz : sanity check and topological clues for linear and non linear mappings.** Sylvain Lespinats, Michaël Aupetit. Computer Graphics Forum journal 30(1) :113-125, Wiley (2011).

Résumé : Comment avoir une vue synthétique de l'ensemble des défauts d'une représentation graphique à deux dimensions d'un nuage de points multi-dimensionnels.

- [48] **Un graphe génératif pour la classification semi-supervisée.** Pierre Gaillard, Michaël Aupetit, Gérard Govaert. Revue Ingénierie des Systèmes d'Information, Vol. 15 (2010).

Résumé : Comment réaliser un modèle semi-supervisé ne nécessitant aucun réglage arbitraire de méta-paramètres.

- [46] **Learning topology of a labeled data set with the supervised generative Gaussian graph**. Pierre Gaillard, Michaël Aupetit, Gérard Govaert. *Neurocomputing* 71(7-9) : 1283-1299 (2008).

Résumé : Comment modéliser automatiquement la connexité d'un nuage de points multi-dimensionnels étiquetés à l'aide d'un modèle génératif.

- [15] **High-dimensional labeled data analysis with topology representing graphs**. Michaël Aupetit, Thibaud Catz. *Neurocomputing* 63 : 139-169 (2005).

Résumé : Comment modéliser et visualiser la connexité intra et inter classes d'un nuage de points multi-dimensionnels étiquetés à l'aide d'un modèle géométrique basé sur l'hypothèse d'un classifieur au sens du plus proche voisin.

Je présente maintenant ces travaux plus en détail.

3 Visualiser la topologie d'un nuage de points

3.1 Les données

Les données que je considère dans mes travaux et dans ce document sont des valeurs numériques issues de mesures physiques ou du traitement de telles mesures ou de données structurées (signaux, images, textes, arbres, graphes...). Ces valeurs sont les coordonnées d'individus vecteurs multi-dimensionnels situés dans un espace euclidien défini par un ensemble de variables aléatoires, au voisinage de sous-espaces localement euclidiens appelés variétés topologiques que je cherche à modéliser et caractériser. Je dispose donc d'un ensemble $S_I = \{I_1, \dots, I_N\} \subseteq I$ de N individus et d'un ensemble $S_V = \{V_1, \dots, V_D\} \subseteq V$ de D variables aléatoires réelles, des données $\{x_i^v\}_{i \in S_I, v \in S_V} \in \mathbb{R}^{N \times D}$. On peut voir ces données comme un nuage de N points-individus de \mathbb{R}^D (ou comme un nuage de D points-variables de \mathbb{R}^N mais je n'ai pas traité ce second cas). Ces données sont fournies sous forme d'un tableau individus-variables \mathcal{T} ($N \times D$) ou d'une matrice de similarités inter-individus \mathcal{I} ($N \times N$) (ou inter-variables \mathcal{V} ($D \times D$) dans le second cas).

3.2 L'analyse descriptive par projection

3.2.1 L'inférence à partir des représentations graphiques

Ces données peuvent être représentées en projetant un ensemble d'individus $S_I \subseteq I$ dans un espace dirigé par un ensemble de variables $S_V \subseteq V$. Dans cette représentation, on peut choisir l'ensemble S_V des variables ainsi que l'ensemble S_I des individus à représenter. Deux types d'inférences sont alors possibles (Figure 12) :

- Inférence basée individus (Inférence sur S_I sachant S_V) : Des classes d'individus similaires, ou des individus prototypiques ou atypiques, sont mis en évidence par l'ensemble de variables S_V .
- Inférence basée variables (Inférence sur S_V sachant S_I) : des classes de variables similaires, ou des variables prototypiques ou atypiques, sont mises en évidence par l'ensemble d'individus S_I .

Définir ces classes nécessite d'abord de mesurer les similarités entre les éléments (individus ou variables) pour déterminer les constituants de ces classes. Cette mesure peut être déjà fournie sous forme de la matrice \mathcal{I} ou \mathcal{V} ou doit être calculée à partir du tableau \mathcal{T} . Il existe dans la littérature de très nombreuses mesures de similarité entre individus ou entre variables, les plus classiques sont

par exemple la distance euclidienne pour évaluer la similarité entre individus : $\mathcal{I}_{i,j} = \sqrt{\sum_{v \in S_V} (x_i^v - x_j^v)^2}$, ou le coefficient de corrélation de Pearson pour évaluer la similarité entre variables : $\mathcal{V}_{v,w} = \frac{1}{|S_I| \sigma_v \sigma_w} \sum_{i \in S_I} (x_i^v - \bar{x}^v)(x_i^w - \bar{x}^w)$ avec $\sigma_u = \sqrt{\frac{1}{|S_I|} \sum_{i \in S_I} (x_i^u - \bar{x}^u)^2}$ et $\bar{x}^u = \frac{1}{|S_I|} \sum_{i \in S_I} x_i^u$.

3.2.2 Le cas des projections dans le plan

Si l'on se restreint à un ensemble S_V de 2 variables, la mesure de similarité peut être évaluée visuellement à partir de la représentation graphique des individus S_I comme des points dans un repère cartésien orthonormé à deux dimensions définies par les variables S_V et muni de la norme euclidienne. Dans cette représentation, la corrélation linéaire apparaît comme un alignement du nuage S_I selon une droite du plan dont la pente indique l'amplitude de cette corrélation. De même, la présence de classes d'individus de S_I similaires (ou d'individus de S_I atypiques), spécifiques à ce sous espace S_V , apparaissent aussi immédiatement comme des groupes séparés de points rapprochés (ou comme des points isolés). Au-delà de cette représentation cartésienne basée sur deux des variables de V , il existe de nombreuses méthodes de représentation, en particulier juxtaposant ou imbriquant de telles représentations cartésiennes, ou utilisant d'autres modes de représentation tels que les coordonnées parallèles [61] ou les courbes d'Andrews [92] par exemple.

Je me suis intéressé aux méthodes représentant spécifiquement les individus sous forme d'un nuage de points du plan comme la représentation cartésienne ou la projection sur les deux premiers axes principaux obtenus par l'Analyse en Composantes Principales [63].

Parmi ces méthodes dites de projection ⁴, celles continues ⁵ et non linéaires ⁶ qui m'intéressent ont fait l'objet de nombreux travaux

4. Toutes les représentations graphiques sont le résultat de la projection des variables à visualiser (le fond) dans l'espace des variables graphiques (la forme), mais les méthodes dites de projection ou de réduction de dimension utilisées pour une représentation graphique font généralement référence aux variables spatiales de l'espace de représentation, *i.e.* des variables homogènes correspondant aux axes horizontal et vertical de l'écran.

5. Est dite "continue" une projection qui à chaque individu assigne un point image distinct dans l'espace de projection. Les cartes auto-organisées de Kohonen [68] sont au contraire des méthodes de projection dites "discrètes" car elles incluent une phase de quantification vectorielle associant à un ensemble d'individus voisins un même point image prototypique de cet ensemble (un neurone de la carte).

6. Est dite "linéaire" une méthode de projection qui définit un espace de projection dont les axes sont des combinaisons linéaires des axes de l'espace d'origine. L'Analyse en Composante Principale [63], la projection sur deux des axes d'origines, ou le Grand Tour [6] sont des exemples de méthodes "continues" et "linéaires".

(voir par exemple [34][81] [71] [100] [114]). A chaque individu est associé un point image dans l'espace plan de projection. La méthode de projection tente alors par exemple par la minimisation d'une fonction d'énergie (ou de stress), de positionner ces points images de telle sorte que leurs distances relatives reproduisent au mieux celles contenues dans la matrice \mathcal{I} mesurées entre les individus d'origine. Les méthodes diffèrent essentiellement par l'information privilégiée qu'elles tentent de préserver : les petites distances dans l'espace de projection plutôt que les grandes par exemple pour l'Analyse en Composantes Curvilignes (ACC) [34] ; les taux de rappel et de précision pour NeRV [117] ; la densité de probabilité pour le Stochastic Neighbors Embedding [82]... Une analyse comparative de la plupart de ces méthodes de projection est proposée dans [74].

3.2.3 Les distorsions

Dans l'article [10], je décris deux types de distorsions possibles induites par ces méthodes de projection :

- Distorsions géométriques (Figures 13a et 13b) :
les distances peuvent être compressées ou étirées par la projection (une demi-sphère en caoutchouc pourra être mise à plat en étirant le voisinage de son bord et en comprimant le voisinage de son pôle). Elles sont mesurables directement à partir de l'échantillon original et de sa projection.
- Distorsions topologiques (Figures 13c et 13d) :
les compressions et étirements peuvent être tels qu'un recollement ou une déchirure surviennent modifiant localement la topologie de la structure sous-jacente (une demi-sphère en papier se déchirera le long de son bord lors de cette aplatissement, alors que les deux hémisphères d'une sphère en caoutchouc se recolleront l'une sur l'autre lors de l'aplatissement). Elles sont induites par l'application qui projette l'espace d'origine dans l'espace de projection. Il n'y aucune distorsion topologique si cette application est un homéomorphisme.

Les distorsions géométriques sont bien connues, mais je les ai distinguées des distorsions topologiques pour la première fois dans ce travail. Toutes ces distorsions sont des artefacts de la projection qui peuvent avoir une cause structurelle (induite par la structure sous-jacente aux données qui ne peut se projeter dans l'espace de projection sans déformation, comme une sphère sur un plan) ou technique (induite par la méthode de projection inadaptée ou dont le processus d'optimisation reste piégé dans un optimum local créant des distorsions qui n'existeraient pas à l'optimum global). On peut agir sur

la méthode de projection pour réduire ou supprimer les artefacts techniques mais on ne peut éviter les artefacts structurels puisque ces structures ne sont généralement pas homéomorphes à l'espace de projection et que l'on ne peut même pas choisir un espace plus adapté puisque l'on ignore *a priori* la nature des structures topologiques sous-jacentes aux données que l'on cherche à modéliser.

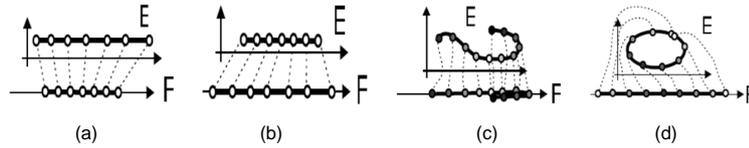
Dans la pratique, on sait quantifier les distorsions géométriques mais on ne peut quantifier les distorsions topologiques sans avoir d'abord modélisé à partir de l'échantillon l'espace topologique sous-jacent à la population d'origine et celui sous-jacent à la population projetée, puis extrait leurs invariants topologiques et mesuré leurs écarts.

3.2.4 Le problème des distorsions

Les méthodes de projections non linéaires ont une particularité : contrairement aux projections linéaires, les axes de l'espace de projection n'ont aucun lien explicite avec les axes ou variables S_V de l'espace d'origine. Le seul lien qui subsiste entre l'espace de projection et l'espace d'origine concerne les individus puisque leurs similarités originelles \mathcal{I} sont estimées par la matrice de distances $\hat{\mathcal{I}}$ objectives entre leurs images, distances elles-mêmes perçues subjectivement comme une matrice $\tilde{\mathcal{I}}$ par l'analyste (Figure 14). L'inférence sur S_I décrite ci-dessus (recherche d'individus typiques ou prototypiques, ou de classes d'individus), reste donc la seule envisageable. Cependant, qu'importe le critère précis utilisé pour mesurer la distorsion entre $\hat{\mathcal{I}}$ et \mathcal{I} que la méthode de projection tente de minimiser, si l'écart $\Xi = \mathcal{I} - \hat{\mathcal{I}}$ n'est pas nul, ces distorsions existent ($\Xi_{ij} > 0$ indique une compression, $\Xi_{ij} < 0$ est signe d'un étirement). On pourrait également considérer une mesure d'erreur de la forme $\Xi_{ij} = |\hat{\mathcal{I}}_{ij} - \mathcal{I}_{ij}|$ ou toute autre mesure d'erreur, sans invalider le raisonnement suivant.

Tant que ces distorsions ne sont pas quantifiées et que cette quantification n'est pas portée à la connaissance de l'analyste, inférer une quelconque caractéristique de \mathcal{I} à partir de $\hat{\mathcal{I}}$ seulement n'a pas de sens : on ne peut réaliser cette inférence tant que l'on ne connaît pas aussi la matrice Ξ , et donc, la réaliser visuellement à partir de $\tilde{\mathcal{I}}$ seulement est d'autant moins possible. Par conséquent, la chaîne d'inférence $\tilde{\mathcal{I}} \rightarrow \hat{\mathcal{I}} \rightarrow \mathcal{I} \rightarrow S_I$ est rompue, on ne peut effectuer l'inférence sur S_I à partir de $\tilde{\mathcal{I}}$. En particulier rien ne différencie de ce point de vue, le résultat d'une projection non linéaire de celui d'une projection aléatoire. Bien sûr, les projections d'individus échantillons d'une population structurée, obtenues par l'optimisa-

Figure 13 **Distorsions géométriques et topologiques**



Les individus (points blancs) sont affichés dans l'espace d'origine E et leur image par une projection dans l'espace de projection F. La ligne noire est la structure topologique sous-jacente au voisinage de laquelle résident les individus dans E. (a) montre les compressions, (b) les étirements, (c) les recouvrements ou faux voisinages et (d) les déchirures. (a) et (b) sont des **distorsions géométriques** tandis que (c) et (d) sont des **distorsions topologiques**. La projection en (a) et (b) est un homéomorphisme mais pas en (c) et (d).

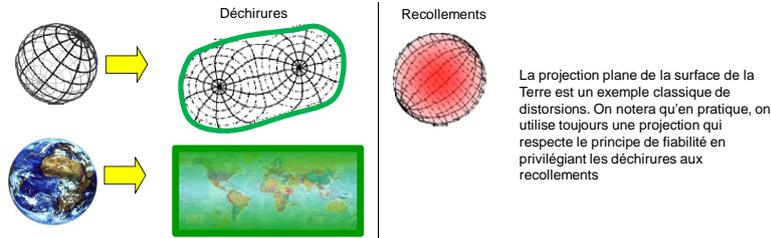


Figure 14 **Inférences et distorsions dans la chaîne de visualisation**



La matrice de similarités d'origine I entre individus, est projetée sous forme d'un nuage de points dans un espace à deux dimensions où une métrique définie par l'analyste permet de calculer \hat{I} estimation de I . La matrice \hat{I} est rendue graphiquement sur l'écran d'affichage et finalement perçue par l'analyste comme une matrice de similarité \tilde{I} estimation de \hat{I} . L'analyste effectue une inférence basée individus à partir de \tilde{I} seulement, unique information accessible à ses yeux, bien qu'il serait idéal (mais impossible) qu'il base son inférence sur I qui contient directement l'information authentique.

tion d'une fonction d'énergie, produisent plus souvent qu'une projection aléatoire, des ensembles de points dont notre système visuel s'empresse d'imaginer la forme. Mais rien ne permet d'inférer de cette seule projection que l'une de ces formes subjectives, existe intègre ou altérée dans l'espace d'origine. En particulier, savoir que la méthode de projection utilisée privilégie la préservation des petites distances, de la variance ou de la densité, n'est d'aucun secours, tant qu'aucune quantification de cette préservation n'est fournie de manière interprétable à l'analyste en sus du nuage de points projeté lui-même. La cause n'en est pas la subjectivité de l'analyste, mais bien l'absence de lien entre l'espace de projection et l'espace d'origine : on ne peut remonter de $\hat{\mathcal{I}}$ à \mathcal{I} sans connaître précisément Ξ .

Pourtant la matrice Ξ est parfaitement connue, la principale difficulté réside en pratique dans sa représentation graphique afin de la porter à la connaissance de l'analyste d'une manière interprétable. En effet, Ξ contient N^2 éléments qu'il faudrait représenter au voisinage des N points projetés. Les principales méthodes de l'état de l'art consistent soit à décrire le contenu de Ξ complètement mais dans une représentation graphique séparée (cas du diagramme de Shepard) et sans pouvoir identifier directement les couples de points concernés, soit à observer sous forme symbolique une statistique $\hat{\Xi}$ calculée sur Ξ comme par exemple la moyenne du carré des écarts $\hat{\Xi} = \sum_{ij} (\Xi_{i,j})^2$ (stress global). Une revue détaillée des critères de l'état de l'art est proposée dans [80].

Nous avons proposé récemment avec Sylvain Lespinats [80] de représenter simultanément au voisinage de chaque point i projeté une statistique plus détaillée $\hat{\Xi}_i = (\Xi_i^+, \Xi_i^-)$, montrant les compressions impliquant cet individu i par la somme des carrés des écarts pour de petites distances dans l'espace de projection $\Xi_i^+ = \sum_j \Xi_{i,j}^2 | \hat{\mathcal{I}}_{i,j} < \sigma$ (critère minimisé par l'Analyse en Composantes Curvilignes [34]) et les étirements impliquant cet individu i par la somme des carrés des écarts pour de petites distances dans l'espace d'origine $\Xi_i^- = \sum_j \Xi_{i,j}^2 | \mathcal{I}_{i,j} < \sigma$ (critère minimisé par la méthode de Sammon [100]), où σ est le paramètre d'échelle réglant la taille du voisinage considéré. Les cellules de Voronoï⁷ de chaque point sont colorées à partir d'une échelle de couleur bidimensionnelle perceptuellement uniforme⁸ permettant de distinguer les deux types de distorsions (blanc

7. La cellule de Voronoï V_i du point x_i de \mathbb{R}^D est l'ensemble $\{x \in \mathbb{R}^D | \forall j \in I, (x - x_i)^2 \leq (x - x_j)^2\}$. Ces cellules sont calculables en temps $O(N \log(N))$ dans le plan [18]. Les raisons du choix de cette région pour la représentation graphique sont discutées dans [10, 80].

8. Une échelle de couleurs est perceptuellement uniforme lorsque les rapports de distances spatiales entre deux couleurs quelconques et une couleur de référence dans la représentation graphique de l'échelle de couleur, sont identiques aux rapports des similarités perçus entre les

pour les points dont la distance à tous les autres dans l'espace de projection est identique à celle d'origine, vert pour les points subissant des étirements, pourpre pour ceux subissant des compressions et noir pour ceux subissant simultanément les deux types de distorsions). Cette méthode de qualification d'une représentation est appelée CheckViz, elle fournit une vue d'ensemble de $2N$ distorsions locales agrégées (Ξ_i^+ et Ξ_i^- pour chacun des N points i). La méthode CheckViz est illustrée sur la figure 15.

CheckViz permet d'inférer que les structures visualisées dans les zones blanches exemptes de distorsions existent aussi entre les individus dans l'espace d'origine duquel dérive \mathcal{I} . Deux autres inférences sont aussi possibles. Lorsque tous les points de deux groupes de points séparés sont colorés en pourpre (aucun étirement), alors cette séparation existe aussi entre ces points dans l'espace d'origine. Lorsqu'un groupe de points contient des points de deux classes différentes, et que tous ces points sont colorés en vert (aucune compression), alors ces deux classes de points se chevauchent effectivement dans l'espace d'origine.

J'avais par ailleurs dans l'article [10] déjà exploré différentes représentations globales d'une statistique agrégée $\hat{\Xi}$ de Ξ moins complètes que la précédente. J'avais aussi proposé de visualiser directement dans l'espace de projection les compressions et étirements locaux $\Xi_{\bullet} = \Xi_{i,j|V_i \cap V_j \neq \emptyset}$, représentation exacte mais partielle de Ξ impliquant uniquement les paires d'individus dont les points images par la projection ont des cellules de Voronoï adjacentes. La valeur de la distorsion était visualisée par la coloration des cellules de Voronoï des segments de droite reliant ces points projetés. Comme seules les paires de points projetés voisins sont considérées, cette méthode ne permettait de visualiser que les recollements mais pas les déchirures qui impliquent des paires d'individus originellement voisins mais projetés séparés l'un de l'autre.

Ces différentes approches ont pour principal intérêt de montrer une partie des distorsions Ξ au sein même de l'espace de projection au voisinage direct des points ou couples de points impliqués matérialisant $\hat{\mathcal{I}}$. Elles permettent de corrélérer visuellement les structures de points observées avec la distorsion locale, et donc d'évaluer si ces structures sont authentiques ou bien des artefacts de la projection.

Cependant, l'inférence sur \mathcal{I} n'est ni directe ni complète puisqu'elle passe par une reconstruction mentale de \mathcal{I} à partir de $\hat{\mathcal{I}}$ (ce qui est perçu de $\hat{\mathcal{I}}$ au travers du nuage de points projeté) et de $\hat{\Xi}$ (ce qui est perçu des distorsions Ξ représentées soit par une statistique

couleurs elles-mêmes.

$\hat{\Xi}$ soit par une partie Ξ_{\bullet} exacte mais incomplète).

J'ai donc proposé dans [10] puis exposé plus en détail dans [12, 13] un changement de paradigme.

3.2.5 Montrer plus que les distorsions : la mesure de proximité

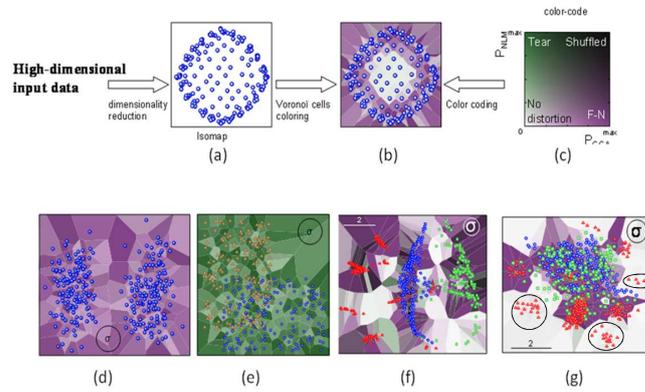
L'idée principale que j'ai explorée est qu'il faut réintroduire dans l'espace de projection une partie au moins de l'information originelle. Ici cette information est portée par \mathcal{I} . J'ai donc proposé de visualiser \mathcal{I} sous la forme exacte bien qu'incomplète d'une de ses colonnes $\mathcal{I}_{\bullet} = \mathcal{I}_{\bullet,s}$ associée à un individu s arbitrairement sélectionné par l'analyste. Il y a N éléments sur cette colonne, chacun associé à un individu. La valeur d'un élément de cette colonne est représentée par la coloration de la cellule de Voronoï de l'individu correspondant. Les individus à la cellule claire sont originellement proches de l'individu sélectionné, tandis que les individus à la cellule foncée sont originellement éloignés de cet individu. La figure 16 montre le résultat obtenu sur des données jouets.

La structure topologique originelle des données peut être extraite grâce à cette méthode. Plus précisément, l'ajout des valeurs $\mathcal{I}_{\bullet,s}$ sur la vue représentant exactement $\hat{\mathcal{I}}$, focalise l'attention sur les points blancs tous proches du point sélectionné donc tous proches les uns des autres (on ne peut par contre rien conclure quant à la similarité originelle entre points sombres). Apparaissent immédiatement les déchirures impliquant le point sélectionné et son voisinage originel, comme des groupes de points blancs (originellement proches de s) isolés les uns des autres par des régions contenant des points noirs (originellement éloignés de s). Apparaissent immédiatement les recolllements impliquant le point sélectionné et son voisinage originel, comme des groupes de points blancs (originellement proches de s) parsemés de points noirs (originellement éloignés de s). Les distorsions géométriques sont moins interprétables car la corrélation des distances $\mathcal{I}_{\bullet,s}$ perçues au travers de l'échelle de couleur, avec celle des distances $\tilde{\mathcal{I}}_{\bullet,s}$ perçues spatialement comme la distance euclidienne entre chaque point et le point sélectionné, n'est pas immédiate, ces deux informations étant représentées par des variables graphiques non homogènes (espace et couleur).

La valeur visualisée au voisinage de chaque point i , appelée mesure de proximité, étant la mesure de distance ou de similarité d'origine $\mathcal{I}_{i,s}$ entre ce point i et le point de référence s sélectionné par l'analyste, est donc la valeur authentique fournie en entrée, avant la projection et les artefacts qu'elle engendre. Ainsi, on réinjecte dans l'espace de projection une information intègre sur les données d'ori-

Figure 15

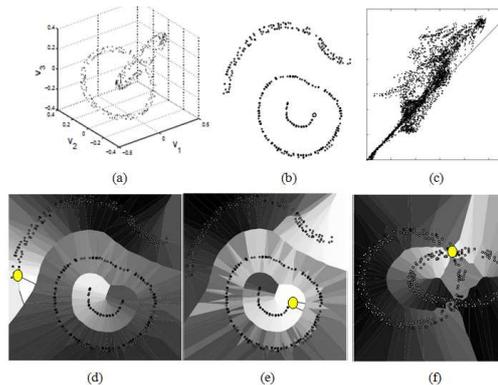
CheckViz



(a) La projection nue de données multivariées n'est pas utilisable en l'état pour inférer des propriétés sur ces données dans leur espace d'origine. (b) Cette même projection est colorée en fonction des distorsions locales. (c) La carte des couleurs perceptuellement uniforme (Etirements purs en vert, compressions pures en pourpre). (d) La règle de Séparation-Vraie. (e) La règle de Chevauchement-Vrai. (f) Les données Oil-flow projetées avec ISOMAP et (g) avec l'ACC. En (f) la règle de Séparation-Vraie s'applique entre les points verts et bleus : les deux classes sont effectivement séparées dans l'espace d'origine. En (g) la règle Aucune-Distorsion s'applique aux groupes de points rouges encadrés : ces groupes existent effectivement dans l'espace d'origine.

Figure 16

ProxiViz et le critère d'authenticité



(a) Les individus d'origine sont issus de deux anneaux entrelacés sans contact dans \mathbb{R}^3 , projetés par ACC dans (b,d,e) et par ACP dans (f). En (b) La projection nue obtenue n'est pas utilisable pour l'inférence. En (c) Le diagramme de Shepard de (b) où les points au-dessus de la diagonale montrent les étirements, et ceux au-dessous montrent les compressions. L'analyste sait par ailleurs que l'ACC favorise les étirements et déchirures, tandis que l'ACP ne peut générer que des compressions ou recollements. En (d,e,f) la similarité authentique de tous les individus à un individu de référence choisi arbitrairement par l'analyste (point jaune) est encodée graphiquement par la nuance de gris (luminance) des cellules de Voronoï de leur image par la projection (polygones). Plus la cellule est claire (resp. foncée) plus l'individu est proche (resp. loin) de l'individu de référence dans l'espace d'origine. En (d) et (e) l'un des anneaux a été déchiré en deux parties par la projection. En (f) les deux anneaux se recollent

gine que la projection *boîte noire* a altérée, ce qui contribue à rendre interprétable la vue obtenue.

Nous avons nommé ProxiViz cette méthode de représentation complémentaire de la méthode CheckViz. CheckViz donne une vue globale des compressions Ξ_i^+ et étirements Ξ_i^- en chaque point i , alors que ProxiViz fournit une vue locale des proximités originelles $\mathcal{I}_{\bullet,s}$ à un point de référence s . CheckViz guide l'analyste pour sélectionner les points de référence auxquels appliquer ProxiViz.

3.2.6 Le paradigme *WinSitu* de visualisation *in situ*

La méthode ProxiViz d'analyse exploratoire par projection s'inscrit dans un nouveau paradigme : la projection n'est plus une fin en soi, car elle n'est pas interprétable ainsi, mais elle est un *moyen*. En effet, elle permet de construire un écran formé d'une mosaïque de pixels (les cellules de Voronoï) colorés par une information *authentique*. Cet écran est comme une fenêtre ouverte sur les données telles qu'elles sont *in situ* dans l'espace d'origine, aussi ai-je nommé ce paradigme de l'analyse exploratoire : "WInSitu" (pour "Window In Situ" ou fenêtre sur l'espace *in situ*). La coloration des cellules de Voronoï par des valeurs exactes éléments de \mathcal{I} , apporte le complément d'information nécessaire à l'interprétabilité de cette projection qui seule ne montre que $\hat{\mathcal{I}}$, et la rend finalement utilisable pour l'analyse exploratoire.

Dans ce paradigme il y a donc trois ingrédients essentiels afin de permettre à l'analyste de faire une inférence sur les individus à partir de leur représentation graphique [13] :

1. **Principe de co-visualisation** : la co-visualisation sur une même représentation graphique de la matrice de similarité initiale \mathcal{I} ou de son estimation $\hat{\mathcal{I}}$ avec une information supplémentaire sous les deux conditions suivantes ;
2. **Principe de fiabilité** : une méthode de projection linéaire ou non, continue ou non, de l'ensemble des individus, tendant à minimiser les distorsions Ξ , prenant en compte la matrice \mathcal{I} et son image $\hat{\mathcal{I}}$ par la projection ;
3. **Principe d'authenticité** : une mesure authentique, sans perte d'information sur les données d'origine \mathcal{I}_{\bullet} , \mathcal{V}_{\bullet} ou \mathcal{T}_{\bullet} , co-visualisée avec les individus projetés, dans l'espace de projection, ou bien un terme d'erreur accompagné de règles d'inférences claires basées sur lui qui permettent de reconstruire tout ou partie de \mathcal{I} .

La figure 17 illustre la complémentarité des deux derniers de ces trois principes.

Dans l'ancien paradigme où la projection est une fin en soi, la course à la projection parfaite est sans fin, il faudra toujours faire un choix sur l'information qui doit être préservée et celle qui sera perdue, chaque méthode pouvant être meilleure que les autres suivant le critère choisi pour les comparer [79]. Dans le nouveau paradigme, il n'est plus primordial d'obtenir une projection parfaite, il suffit qu'elle tende à limiter les distorsions et en particulier les recollements comme décrit plus tôt dans le principe de fiabilité décrit à la section 2.4.4, et qu'elle soit raisonnablement efficace dans cette tâche, afin de former un écran constitué de groupes de pixels adjacents qui soient aussi localement ordonnés (*i.e.* voisins) dans l'espace d'origine. En effet, les conséquences des distorsions sur l'interprétabilité ne sont pas symétriques : maximiser la corrélation entre le voisinage perçu pré-attentivement dans l'espace de représentation, et le voisinage originel de la structure à représenter, correspond à minimiser prioritairement les distorsions de type recollement qui justement rompent cette corrélation, et donc privilégier les méthodes de projection telles l'Analyse en Composantes Curvilignes, qui tendent à dissiper sous forme de déchirures les distorsions inévitables. La figure 17b montre qu'un cas extrême, où déchirures et recollements sont à leur paroxysme du fait d'une projection aléatoire, ne permet aucune inférence sur I .

Ce principe de fiabilité s'applique encore dans une technique que nous avons récemment développée avec Sylvain Lespinats. Cette technique de projection supervisée, appelée ClassiMap [78], tend à concentrer les déchirures entre points de classes différentes, et les recollements entre points de même classe. Ainsi les classes sont mieux séparées dans l'espace de projection à condition qu'elles soient effectivement séparable dans l'espace d'origine. Là encore, ce que l'on voit est cohérent avec ce que l'on attend dans la mesure du possible.

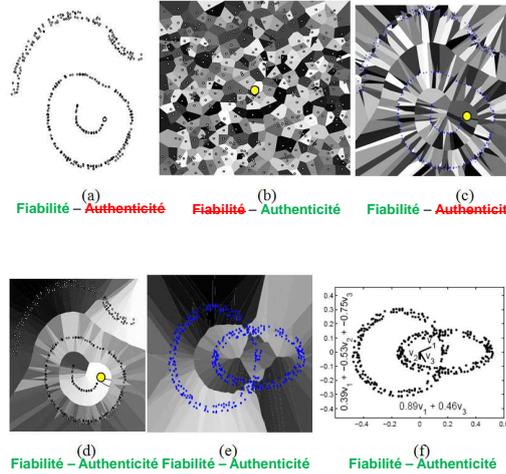
Enfin, nos récents travaux réalisés avec Nicolas Heulot et Jean-Daniel Fekete, confirment ce principe de fiabilité et montrent que les utilisateurs font significativement moins d'erreur dans une tâche de comptage de classes à partir d'une projection sujette aux déchirures qu'avec une projection générant des recollements [59].

3.2.7 Le paradigme *WinSitu* face à l'état de l'art

Le paradigme WinSitu n'est pas nouveau en soi car il est en fait appliqué dans la majorité des techniques de représentation graphique d'information. Par exemple, les méthodes de projections li-

Figure 17

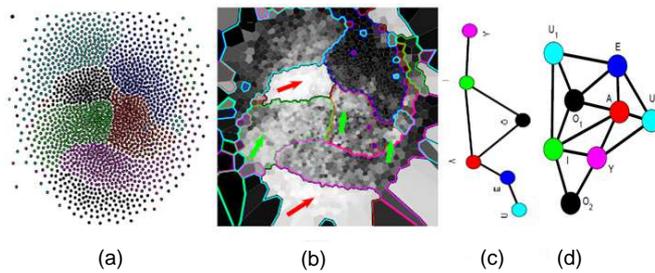
Le paradigme WinSitu



Le paradigme WinSitu est illustré avec ProxiViz et l'ACC, et avec l'ACP. En (a) la projection nue par ACC des anneaux entrelacés (Figure 16b). En (b) la similarité authentique (nuance de gris) à un point de référence (point jaune) co-visualisée avec une projection aléatoire des anneaux. En (c) une similarité aléatoire (nuances de gris) co-visualisée avec la projection par ACC (fiabilité) (a). En (d), la similarité authentique au même point de référence (point jaune) qu'en (b) et (c) co-visualisée avec la projection par ACC (fiabilité) (a). **Aucune des représentations graphiques (a), (b) ou (c) n'est interprétable et utilisable pour l'inférence.** Le paradigme WinSitu pose que la co-visualisation (d) de l'authenticité et de la fiabilité est nécessaire pour rendre possible et potentiellement fructueuse une inférence basée individus. En (e), la composante V_2 est co-visualisée en niveau de gris avec la projection par ACP des individus. Cette représentation graphique ainsi que celle de la figure 16f implémentent le paradigme WinSitu, mais ici l'information authentique est une colonne de T plutôt qu'une ligne de I. En (f), la représentation classique par ACP covisualisée avec les axes combinaisons linéaires explicites (information authentique) des variables d'origine, représentés au centre de la vue, est elle aussi interprétable.

Figure 18

Application de ProxiViz



(a) Les données ISOLET (1800 individus et 617 variables, 6 classes associées aux 6 voyelles) sont projetées par ACC dans le plan. (b) ProxiViz permet de reconstituer par exploration visuelle interactive une connectivité intra-classe et inter-classe originelle plausible encodée par le graphe des classes en (c). Ce graphe diffère fortement de celui que l'on reconstituerait naïvement à partir de la représentation graphique (a).

néaires en font partie par construction, l'information authentique visualisée étant dans ce cas l'exact positionnement des points projetés par rapport aux axes (principaux pour l'ACP) combinaisons linéaires explicites de variables d'origine AT , ou bien étant fournie par la projection directe dans le même espace des vecteurs unitaires associés à chaque variable d'origine ; les coordonnées parallèles ou en étoile en font aussi partie puisqu'elles montrent les valeurs exactes contenues dans le tableau \mathcal{T} . De même, les représentations graphiques de type noeud-liens de données structurées comme les graphes ou les arbres implémentent aussi ce paradigme puisque les liens et sommets montrent exactement la matrice d'adjacence \mathcal{I} . Nous discutons plus en détail dans [12] les liens avec les cartes auto-organisées de Kohonen. On notera que pour ces dernières, bien qu'il s'agisse de méthodes de projection discrètes où des prototypes représentent les individus, on peut considérer qu'elles réalisent *in fine* une projection continue des prototypes vers l'espace de projection de la carte, et en cela, qu'elles n'échappent pas aux différents types de distorsions évoquées ici ni à leurs conséquences sur l'interprétabilité de la projection.

L'originalité de ce paradigme réside en fait dans sa déclaration explicite en tant que principe nécessaire pour rendre interprétable une représentation graphique. Son utilisation implicite et quasi-générale dans les méthodes de l'état de l'art, est la meilleure preuve de sa nécessité. Les méthodes qui ne le mettent pas en oeuvre, comme les méthodes brutes de projections non linéaires, auxquelles on peut par exemple ajouter la méthode des courbes d'Andrew, sont toutes non interprétables en ce qu'elles rompent les liens sémantiques entre l'espace de représentation et l'espace d'origine, représentant graphiquement des valeurs estimées $\hat{\mathcal{I}}$, sans au minimum montrer l'erreur d'estimation Ξ pourtant connue, et dans le cadre WinSitu, sans montrer de valeur authentique \mathcal{I}_\bullet .

Dans le contexte de l'analyse visuelle d'information, l'*expressivité*, l'*effectivité* [85] et la *vérité* [95] sont habituellement des critères de sélection d'une bonne méthode de visualisation :

- l'**expressivité** d'une méthode de visualisation est sa capacité à montrer toute l'information voulue et seulement elle : on ne fait pas de la décoration, tout ce qui ne porte pas d'information est banni, et tout ce que l'on veut montrer doit être encodé graphiquement ;
- l'**efficacité** est la propension d'une méthode de visualisation à être interprétée efficacement par le système visuel humain : ce qui est montré n'est pas nécessairement ce qui est perçu, il est par exemple plus facile de comparer un écart entre deux

valeurs numériques codées sous forme de la longueur de deux segments de droites placés en vis-à-vis, que codées sous forme d'une couleur sur une échelle "arc-en-ciel".

- la **vérité** [95] caractérise les méthodes montrant dans la même représentation toute l'information disponible sur une donnée, *e.g.* à la fois la valeur et son incertitude. C'est la transposition au domaine graphique du principe qui veut que l'on accompagne toujours la donnée d'une valeur estimée par la valeur de son incertitude (écart-type, intervalle de confiance...). Les méthodes de l'état de l'art ainsi que CheckViz proposée avec Sylvain Lespinats, montrant en plus de $\hat{\mathcal{I}}$, les écarts Ξ soit sous une forme agrégée $\hat{\Xi}$ soit sous une forme exacte mais incomplète Ξ_{\bullet} , vérifient le critère de vérité.

Le paradigme WinSitu invoque les principe de *fiabilité* et d'*authenticité* comme nouveaux critères essentiels d'évaluation :

- La **fiabilité** caractérise la propension d'une méthode de représentation à représenter comme perceptuellement similaires les objets qui le sont par nature. Elle peut être vue comme une forme particulière de l'efficacité car la fiabilité est supposée induire une réduction de la charge cognitive pour mener une inférence sur les structures sous-jacentes aux objets d'origine à partir de leur représentation. Cependant elle n'est pas liée au choix des variables graphiques mais à celui de la méthode de projection. Elle vient en amont de l'efficacité dans la chaîne de visualisation.
- L'**authenticité** caractérise la propension d'une méthode de représentation à véhiculer sans distorsion de fond les informations originelles \mathcal{I} . Elle qualifie ce qui doit être montré pour permettre l'inférence sur les données d'origine au travers de la chaîne visuelle, là où l'expressivité qualifie comment ce qui doit être montré doit l'être sans artifice ni manquement. Les méthodes de projection non linéaires sont parfaitement expressives, elles montrent bien $\hat{\mathcal{I}}$ sous forme d'un nuage de points, seulement montrer uniquement $\hat{\mathcal{I}}$ ne permet pas d'analyser \mathcal{I} et donc d'en inférer des caractéristiques de S_I . L'authenticité se distingue de la vérité en ce que la vérité accompagne un estimateur $\hat{\mathcal{I}}$ de \mathcal{I} d'une incertitude exacte et complète Ξ , exacte et partielle Ξ_{\bullet} , ou estimée $\hat{\Xi}$, mais ne montre pas \mathcal{I} directement ni même une de ses parties \mathcal{I}_{\bullet} . L'authenticité peut aussi être vue comme une forme particulière de l'efficacité mais elle traite de la nature de l'information à montrer (le quoi) plutôt que de la forme graphique (le comment) que cette information devrait prendre pour être interprétée efficacement.

Grâce à ce nouveau paradigme et son implémentation dans *ProxiViz*, les projections non-linéaires deviennent effectivement utilisables pour l'inférence de structures topologiques, et nos expériences montrent qu'elles sont aussi efficaces pour cela [59].

3.3 Applications

Nous avons montré dans [10] que ProxiViz implémentant le paradigme de la visualisation *in situ* permet l'analyse des connexités intra-classes et inter-classes des données "Isolet"[94] de dimension 617. L'expérience est résumée sur la figure 18.

Les données "Oil flow" [22] de dimension 12, sont aussi analysées par ProxiViz dans [17] (Figure 25b) et en complément d'autres méthodes dans [47].

Une autre application de ProxiViz est proposée dans un projet d'interface graphique d'aide à la décision pour le contrôle douanier de conteneurs maritimes [14]. Dans ce cas le point de référence n'est pas affiché sur la projection mais fourni comme requête dans l'espace originel. On ne voit sur la projection que la similarité des données de référence à cette requête. Les données sont la signature chimique du contenu des conteneurs sondés par un appareil de mesure spécifique. La projection représente uniquement les conteneurs déjà analysés. Ils sont d'autant plus lumineux que leur contenu est proche en termes d'éléments chimiques, de celui actuellement en cours de vérification (la requête). Se dessine alors sur la carte une forme caractéristique qu'à l'usage le douanier mémorise et reconnaît, assistant sa décision future d'ouvrir ou non le conteneur pour un contrôle plus poussé.

Nous avons vu comment on peut extraire visuellement une information topologique à partir d'une représentation graphique. Je présente maintenant le deuxième axe de mes travaux dans lequel je cherche à extraire automatiquement *in situ* dans l'espace multidimensionnel des données, cette information topologique.

4 Extraire la topologie d'un nuage de points

4.1 Quelques notions de topologie

Je précise ici un peu plus les notions topologiques évoquées en introduction. Le terme de topologie en Apprentissage Automatique se réfère habituellement soit à la structure des réseaux de neurones en couches, ou des modèles graphiques (réseaux bayésiens par exemple), soit à des modèles de représentation porteurs d'une information topologique plus ou moins directement liée aux données, comme les cartes auto-organisées de Kohonen [69]. Ici, je parle de la topologie de variétés supposées sous-jacentes à un nuage de points.

La topologie est le domaine des sciences qui étudie les structures ou formes, définit des critères permettant de les classer, et des moyens pratiques de calculer ces critères [57, 93]. Ces formes sont des espaces topologiques. Un espace topologique est un couple (E, T) , où E est un ensemble et T sa topologie, *i.e.* un ensemble de parties de E contenant l'ensemble vide et E lui-même, et dont les éléments sont tels que l'union et l'intersection d'une partie quelconque d'entre eux fait aussi partie de T . Les éléments de T sont les "ouverts" de (E, T) . Nous nous intéressons aux variétés topologiques, espaces topologiques localement homéomorphes à l'espace euclidien \mathbb{R}^D .

Un homéomorphisme entre deux espaces topologiques est une application bijective continue et de réciproque continue entre ces deux espaces.

Les formes qui nous intéressent sont définies par une union de variétés. Une d -variété est un espace topologique connexe localement homéomorphe à \mathbb{R}^d , sauf pour les points de son bord (si elle en a un) dont le voisinage est homéomorphe à $\mathbb{R}^+ \times \mathbb{R}^{d-1}$. Le bord d'une d -variété est l'union de $(d - 1)$ -variétés sans bord.

Deux formes peuvent avoir la même topologie, *i.e.* être homéomorphes, sans être pour autant superposables (leur plongement dans un même espace métrique diffère). Un exemple type d'homéomorphisme est celui qui existe entre une tasse en céramique et un rond de serviette, on peut passer de l'un à l'autre par déformation continue sans modifier leur topologie. Les invariants topologique sont des descripteurs de formes qui permettent de les classer : les formes d'une même classe sont obtenues par certaines transformations (homéomorphisme, homotopie...) de l'une à l'autre qui préservent ces descripteurs, d'où leur nom d'invariants. Deux descripteurs élémentaires d'un espace topologique sont la dimension intrinsèque (par opposition à la dimension ambiante de l'espace dans lequel la forme

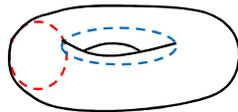
est plongée) et le nombre de composantes connexes. Les nombres de Betti sont des invariants qui caractérisent la connexité au sens large des variétés, le premier nombre de Betti B_0 est le nombre de composantes connexes, le second B_1 le nombre de tunnels, le troisième B_2 le nombre de cavités... Par exemple, une sphère usuelle (surface d'une boule usuelle) est connexe, sans bord, et comporte un vide ($(B_0, B_1, B_2) = (1, 0, 1)$), un cylindre ou un anneau sont connexes, possèdent deux cercles pour bords, et comportent un tunnel ($(B_0, B_1, B_2) = (1, 1, 0)$), un tore creux est connexe, sans bord, et comporte deux tunnels et une cavité ($(B_0, B_1, B_2) = (1, 2, 1)$)... La figure 19 illustre les nombres de Betti d'un tore et d'un cylindre.

Pour calculer effectivement ces invariants, il est nécessaire de trianguler la variété objet de l'étude, *i.e.* de construire un complexe simplicial homéomorphe à cette variété, donc un espace topologique discret, manipulable numériquement de même topologie que la forme continue que l'on souhaite étudier, ce qui est toujours possible. Un complexe simplicial est un ensemble particulier de simplexes. Un d -simplexe est une d -variété définie par un ensemble de $d + 1$ sommets, dont l'enveloppe convexe dans un espace vectoriel forme la réalisation géométrique : un 0-simplexe est un point de cet espace, un 1-simplexe est un segment de droite, un 2-simplexe un triangle plein, un 3-simplexe un tétraèdre plein... Les faces d'un simplexe sont aussi des simplexes, mais de dimension inférieure. Du point de vue abstrait ou combinatoire, un d -simplexe est l'ensemble de ses sommets. Un complexe simplicial est un ensemble de simplexes tel que l'intersection de deux quelconques d'entre eux est soit vide soit un simplexe contenu dans cet ensemble. En d'autres termes, les simplexes d'un complexe simplicial soit ne se touchent pas, soit sont en contact par l'une de leurs faces.

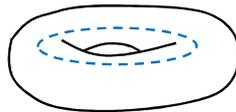
On peut définir des groupes d'homologie à partir d'un complexe simplicial. Les nombres de Betti sont les rangs de ces groupes d'homologie [26], ils caractérisent l'équivalence topologique en termes d'homologie qui est une caractérisation plus restreinte que l'homéomorphisme : deux formes peuvent ainsi avoir mêmes nombres de Betti alors qu'elles ne sont pas homéomorphes, par exemple un cercle, un ruban en anneau, un ruban de Moebius et un tore plein appartiennent à la même classe d'homologie, ils ont les mêmes nombres de Betti $B_0 = 1, B_1 = 1, B_2 = 0 \dots B_{\forall k > 2} = 0$, mais ne sont pas homéomorphes, en particulier parce qu'ils n'ont pas la même dimension intrinsèque : 1 pour le cercle, 2 pour les rubans et 3 pour le tore plein dans notre espace usuel à 3 dimensions, et que le ruban de Moebius a subi une torsion par rapport au ruban en anneau.

Figure 19

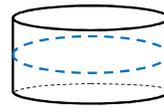
Exemples de nombres de Betti



Tore vide : une composante connexe ($B_0=1$), deux 1-cycles (boucles bleue et rouge) ($B_1=2$) et une cavité ($B_2=1$).



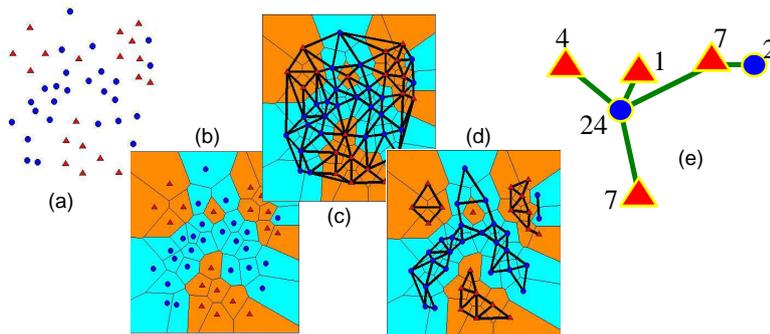
Tore plein : une composante connexe ($B_0=1$), un 1-cycle (boucle bleue) ($B_1=1$) et aucune cavité ($B_2=0$).



Cylindre : une composante connexe ($B_0=1$), un 1-cycle (boucle bleue) ($B_1=1$) et aucune cavité ($B_2=0$). Le cylindre et le tore plein ne sont pas homéomorphes mais font partie du même groupe d'homologie.

Figure 20

Graphe des classes automatique



(a) Un nuage de points étiquetés. (b) Les cellules de Voronoï associées. (c) Le graphe de Delaunay de ce nuage. (d) Elagage des arêtes mixtes. (e) Graphe des classes : chaque sommet représente une composante connexe du graphe (d). Deux sommets sont liés si les composantes étaient liées à l'étape (c). Ce graphe résume la topologie des classes au sens du Plus Proche Voisin.

4.2 Le cas des nuages de points

Nous avons à notre disposition des outils mathématiques comme les nombres de Betti, issus de la Topologie Algébrique définissant des invariants de formes, critères de classification de ces formes. Nous avons des outils issus de la Topologie Algorithmique, permettant de calculer ces invariants à partir de complexes simpliciaux [107], et ces invariants valent pour ceux des variétés dont ces complexes simpliciaux sont des triangulations.

Comme en pratique, nous n'avons accès qu'à un nuage de points, il nous reste deux fonctions de transfert à déterminer pour extraire ses invariants topologiques : du nuage de points à la forme sous-jacente, et de cette forme à sa triangulation. Nous considérons que la forme sous-jacente au nuage de points est un ensemble de variétés que nous appelons *variétés latentes* car elles ne sont pas directement observées. En fait il n'est pas nécessaire de modéliser le nuage de points par une forme puis de la trianguler, car la triangulation est un complexe simplicial, donc est en elle-même une forme particulière. Donc le principal problème posé est de choisir une triangulation construite à partir de ce nuage qui permette d'extraire une information topologique intéressant l'analyste.

Nous avons étudié deux cas, une approche descriptive dans [15], basée sur la triangulation du nuage de points étiquetés, afin d'étudier la topologie des classes associées, et une approche prédictive dans [9, 46, 48], basée sur un modèle génératif dont la structure source est un complexe simplicial.

4.3 Analyser la connexité intra et inter classes

Plutôt que de chercher à extraire visuellement l'information topologique après avoir projeté les données, donc introduit des artefacts, j'ai proposé dans [15] de l'extraire *in situ*, directement dans l'espace d'origine, puis de la rendre interprétable, accessible à l'analyse visuelle sans risque de perte d'information due à la projection. L'information topologique que j'ai proposé d'extraire est la connexité intra-classe et inter-classe d'individus étiquetés, dans un espace multi-dimensionnel euclidien.

4.3.1 Le graphe des classes

Pour mesurer cette connexité, il est naturel de chercher à représenter les individus par un graphe, structure dont la connexité est explicite et calculable. Les graphes de proximité [62] sont des graphes dont la connexité dépend de la position des sommets dans

un espace métrique. Deux sommets sont reliés s'ils respectent certaines conditions de voisinage dans cet espace. J'ai proposé d'utiliser un tel graphe pour analyser la topologie des classes associées aux étiquettes des individus. L'idée principale était, une fois construit le graphe reliant les individus voisins, d'élaguer les arêtes mixtes reliant des individus de classes différentes. On détermine alors le nombre de composantes connexes de chaque classe (connexité intra-classe) et les liens de voisinage entre elles (connexité inter-classes) matérialisés en négatif par les arêtes mixtes élaguées.

Afin de rendre visualisable ces informations, j'ai défini un graphe de synthèse, dont chaque sommet correspond à une composante connexe du graphe élagué initial. Dans ce graphe que j'appelle "graphe des classes", deux sommets sont liés si les deux composantes du graphe initial étaient liées par au moins une des arêtes mixtes finalement élaguées. Les sommets et arêtes du graphe des classes peuvent être valués par le nombre de sommets ou d'arêtes que chacun représente dans le graphe initial. La figure 20 illustre la construction du graphe des classes.

4.3.2 Quel graphe de proximité choisir ?

Un classifieur est une fonction de l'espace de représentation des individus vers l'espace des classes. Cette fonction de classification partitionne l'espace d'origine en régions associées chacune à une classe unique. J'appelle ces régions "variétés de classe", bien qu'il puisse s'agir d'une union de variétés, voir de formes plus complexes que des variétés suivant les propriétés de la fonction de classification. Pour définir une topologie des classes, et permettre l'étude des relations d'adjacence entre celles-ci, il faut permettre un recouvrement de ces variétés le long de leur frontière commune (on parle alors de décomposition cellulaire plutôt que de partition au sens strict). C'est cette intersection non vide entre ces ensembles qui définit leurs relations de voisinage, donc leur topologie, et en permet l'étude. La frontière entre deux variétés de classe est aussi le lieu où chacune des deux classes est également possible, ce qui donne sens à leur intersection.

Une même classe peut être distribuée dans différentes régions non connexes de l'espace d'origine. On dit que cette classe est morcelée, et ce morcellement caractérise la connexité intra-classe de cette variété de classe. L'ensemble des relations de voisinage entre variétés de classe différentes, caractérise la connexité inter-classe de ces variétés.

Dans notre cas, nous n'avons pas de variétés de classe *a priori*,

puisque seuls les individus sont étiquetés, il existe donc une infinité de classifieurs, et donc de variétés de classe, en mesure d'expliquer ces étiquettes, *i.e.* dont les individus étiquetés forment un échantillon bien classé. Et il existe $2^{N(N-1)/2}$ graphes non orientés possibles pouvant connecter les individus les uns aux autres. Mais l'un de ces graphes et l'un de ces classifieurs sont liés : le graphe de Delaunay des individus et le classifieur au sens du plus proche voisin euclidien de ces individus.

Le classifieur au sens du plus proche voisin, est caractérisé par des variétés de classe unions des cellules de Voronoï des individus de cette classe au sens de la métrique euclidienne⁹. La triangulation de ces variétés s'obtient naturellement par la construction du complexe de Delaunay des individus [18]. Le complexe de Delaunay des individus contient un k -simplexe K si et seulement si l'intersection des cellules de Voronoï des sommets de K est non vide. En élaguant ensuite toutes les faces mixtes (simplexes ayant au moins deux sommets de classe différente), on obtient le complexe de Delaunay restreint aux variétés de classe au sens du plus proche voisin [39], dont on sait ainsi qu'il appartient à la même classe d'homotopie que ces variétés de classe. Il préserve l'arc-connexité au sens où pour tout chemin permettant de passer d'un point de départ à un point d'arrivée d'une même variété de classe, sans jamais changer de classe, alors il existe un tel chemin sur le graphe de Delaunay élagué, partant de la donnée la plus proche du point de départ et arrivant à la donnée la plus proche du point d'arrivée. Dans la publication [8], je me contente de considérer le graphe de Delaunay, mais le complexe simplicial de Delaunay restreint aux variétés de classe, en ajoutant au graphe, les triangles, tétraèdres et simplexes de dimensions supérieures, fournirait en plus une estimation de la dimension intrinsèque locale des variétés de classes et la possibilité de calculer les nombres de Betti.

4.3.3 Une représentation graphique sans perte d'information topologique

Le graphe initial a une représentation géométrique dans l'espace d'origine multidimensionnel des individus, puisque chaque sommet-individu y est positionné. En revanche, les sommets du graphe des

9. L'usage de la métrique euclidienne suppose de normaliser les variables. Le problème du conditionnement des variables reste entier dans un cadre d'analyse exploratoire non supervisé. Rappelons que la métrique euclidienne est un cas particulier de la métrique de Mahalanobis[86] avec une matrice de variance identité. La matrice de variance représente un ensemble de paramètres qui avec le type de graphe, demeurent au choix de l'analyste et lui donneront un point de vue en conséquence sur les données.

classes sont attachés à des ensembles d'individus éléments des composantes connexes du graphe initial élagué, ils n'ont pas de position géométrique naturelle dans l'espace des individus. Le graphe des classes est donc un graphe abstrait. Afin de le visualiser il faut le plonger dans un espace métrique plan par exemple, ce qui peut se faire avec n'importe quelle méthode de projection linéaire ou non, ou de visualisation de graphe¹⁰.

L'intérêt majeur de cette modélisation, est que cette projection ne peut altérer l'information topologique sur la connexité des classes extraite dans l'espace multidimensionnel des individus, car elle est portée par la structure topologique (nombre de sommets et liens entre eux) du graphe des classes, qui est invariante à toute projection (sauf cas particuliers de projections superposant des sommets).

Cette méthode d'analyse est descriptive : elle nécessite le choix d'un point de vue a priori en termes de type de graphe de proximité et de métrique, et son résultat est interprété par l'intermédiaire d'une représentation graphique. J'ai montré que le choix de l'utilisation du graphe de Delaunay des individus permet l'analyse de la topologie des classes au sens du classifieur au plus proche voisin euclidien de ces individus, classifieur fondamental en apprentissage statistique. Plus précisément, je montre dans [15] grâce à un résultat de Edelsbrunner et Shah [39], que les composantes connexes du graphe de Delaunay élagué et les composantes connexes de la variété de classes qu'elles représentent appartiennent deux à deux à une même classe d'homotopie (Figures 3a et 20)

Cette méthode implémente le paradigme WinSitu dans la mesure où le choix de la métrique et du graphe étant effectués, le graphe des classes proposé permet de visualiser la topologie des classes *in situ*, sans distorsion due à la méthode de projection. Quelque soit la projection utilisée (la plus fiable possible pour limiter la charge cognitive), on sait que l'information de connexité montrée est authentique, exactement ce qu'elle est dans l'espace de représentation originel.

4.3.4 Aspects calculatoires

La complexité de calcul du graphe de Delaunay en grande dimension est un problème souvent présenté comme insurmontable dans la littérature. Ses détracteurs font en fait la confusion entre le graphe de Delaunay et la triangulation de Delaunay. La triangulation de

10. Ces méthodes cherchent par exemple une représentation plane minimisant le nombre de croisement d'arêtes, ou utilisent des modèles physique d'attraction-répulsion pour positionner les sommets du graphe dans le plan [65].

Delaunay est le nom commun du complexe simplicial de Delaunay, structure duale du complexe cellulaire de Voronoï, qui a chaque k -cellule de Voronoï (sommets, arêtes, faces, volumes...) associe un $(D - k)$ -simplexe de Delaunay (pour $D = 3$, respectivement les tétraèdres, triangles, arêtes, sommets). Le nombre d'éléments de cette structure est dans le cas général en $O(N^{\lceil D/2 \rceil})$, exponentiel avec la dimension de l'espace d'origine des individus. Le graphe de Delaunay n'est qu'un sous-ensemble de ce complexe simplicial, n'en contenant que les sommets (0-squelette) et arêtes (1-squelette), dont la complexité du calcul est en $O(N^3)$, et dont j'ai pu par hasard trouver une référence [2] dans la littérature¹¹. N'ayant pas connaissance à l'époque de cette référence (Voir aussi [89] pour un travail plus récent sur le sujet), j'ai étudié la perte d'information engendrée par l'utilisation du graphe de Gabriel¹² ($O(N^3)$) à la place du graphe de Delaunay. En particulier, je montre expérimentalement [15] que la probabilité d'être Gabriel pour le graphe de Delaunay de sommets uniformément répartis dans le d -cube unité, est strictement croissante et tend vers 1 quand la dimension d tend vers l'infini.

4.3.5 Liens avec l'état de l'art

Zighed *et al.* [123] ont travaillé sur les graphes de proximité dans le cadre de l'analyse et de la fouille de données. Ils ont proposé la statistique des arêtes coupées, comme un moyen d'évaluer le chevauchement des classes. Dans cette approche, on compte le nombre d'arêtes mixtes du graphe des voisins relatifs des individus, et l'on compare ce nombre à la distribution du nombre de ces arêtes mixte dans les graphes construits sur le même ensemble d'individus mais dont les étiquettes de classe sont mélangées aléatoirement en respectant leurs proportions initiales. Lorsque le nombre observé est proche de celui obtenu en moyenne par mélange aléatoire des étiquettes, on peut considérer que l'espace de représentation ne porte aucune information utile à la séparation des classes, les classes se chevauchent aléatoirement, elles ne présentent pas de structure apprenable. Plus que d'une méthode d'analyse géométrique du chevauchement des

11. Je m'explique l'absence de telles références par le fait que la triangulation de Delaunay est développée et utilisée par la communauté Géométrie Algorithmique, pour la modélisation d'objet en 2 ou 3 dimensions, dans ce cas les problèmes de complexité ne se posent pas, et les applications ne requièrent pas la connaissance du graphe seul. En pratique, le graphe est extrait du complexe simplicial, ce qui ne pose pas de problème en dimension faible mais devient inextricable en grande dimension ($D > 3$).

12. Le graphe de Gabriel est le sous-graphe de Delaunay le plus complet que l'on sache définir de manière générale, il connecte deux sommets par une arête dont le milieu coupe la frontière commune aux cellules de Voronoï de ces sommets, c'est aussi le graphe tel que les boules diamétrales à ses arêtes, ne contiennent aucun autre sommet que les extrémités de ces arêtes. Il a été étendu récemment aux complexes simpliciaux [37].

classes d'individus, il s'agit d'un test statistique de pertinence des variables de représentation. Cependant, les arêtes mixtes d'un sous-graphe de Delaunay des individus joue un rôle fondamental dans les deux approches, montrant l'intérêt de ce graphe pour l'analyse exploratoire de données étiquetées.

Une revue plus complète de l'état de l'art est faite dans l'article. Le graphe des classes en tant que résumé de la connexité des composantes connexes de classes, n'est pas nouveau. Cette méthode de synthèse d'un graphe a été développée par exemple dans [54]. Mon apport se situe dans le choix du graphe de Delaunay comme graphe à synthétiser, et dans la mise en évidence et l'exploitation du fait que ce graphe particulier construit sur les individus, et les variétés de classe du classifieur au plus proche voisin euclidien basés sur ces mêmes individus (classifieur fondamental en apprentissage statistique), appartiennent tous les deux à une même classe d'homotopie.

Dans l'article [8], j'ai aussi proposé des outils d'analyse supplémentaire, pour distinguer les individus frontières ou atypiques des autres individus, permettre la projection de points échantillons de la frontière des classes, ainsi que proposé d'extraire d'autres caractéristiques topologiques comme la profondeur.

4.4 Les limites de l'approche descriptive

Le classifieur au sens du plus proche voisin peut générer des variétés de classes très morcelées, lorsque les individus sont bruités. Le graphe des classes contient alors de nombreux sommets qui masquent la structure topologique essentielle des classes. De plus, les connexités intra-classe et inter-classe sont intimement liées puisque le morcellement d'une classe n'est dû qu'à la présence d'individus d'une ou plusieurs autres classes s'intercalant entre les individus de la première. La densité des individus n'est pas prise en compte dans le graphe des classes, deux classes peuvent être vues comme adjacentes, même si un vaste espace vide les sépare.

L'approche basée sur les modèles génératifs que je vais présenter maintenant, est un moyen de générer un graphe des classes permettant de s'affranchir de la présence de bruit et de distinguer les deux types d'adjacence, celle liée à la densité et celle liée aux classes. Les sommets du graphe ne sont plus des individus, mais des prototypes représentants ces individus. Un modèle génératif est adjoint à ce graphe afin de déterminer quels arcs doivent être élagués. Ce modèle est alors applicable aussi bien dans un cadre supervisé que non supervisé ainsi que comme une solution de principe à l'ensemble des applications évoquées à la section 2.3.

4.5 Vers une approche générative

4.5.1 La reconstruction de formes

Le problème de la reconstruction de surface est bien connu dans le domaine de la Géométrie Algorithmique. Un objet physique (maquette d'un produit, organe...) étant échantillonné (laser, scanner...), l'objectif est de reconstruire numériquement la surface continue de cet objet à partir de cet échantillon. Cette reconstruction passe par une phase de maillage triangulaire de tout ou partie de l'échantillon, la principale difficulté étant de définir une méthode et les conditions correspondantes d'échantillonnage de la surface de l'objet en fonction de ses caractéristiques géométriques (courbure...), afin d'assurer que la topologie du modèle soit identique à celle de la surface de l'objet, donc que le modèle soit bien une triangulation (au sens topologique) de celle-ci. On sait donc comment échantillonner l'objet pour que la méthode reconstruise un modèle de même topologie. C'est un processus opérationnalisable, dans la mesure où l'on peut contrôler l'échantillonnage, de manière similaire au cas des plans d'expérience en physique expérimentale (Où réaliser la prochaine mesure pour maximiser l'information acquise par le modèle ?) ou de l'apprentissage actif en Apprentissage Automatique (Quel est le prochain individu de l'échantillon à sélectionner pour maximiser l'information acquise par le modèle ?). En pratique dans le cadre de l'analyse exploratoire de données, l'échantillonnage n'est généralement pas contrôlé ni contrôlable, les conditions d'échantillonnage deviennent alors de simples hypothèses du modèle : on infère la fonction densité de probabilité de la population à partir de l'échantillon et d'hypothèses sur la famille paramétrique de fonctions de densité à laquelle est supposée appartenir la vraie fonction.

De nouvelles approches sont apparues en Géométrie Algorithmique pour traiter des formes autres que des surfaces, travailler en dimension supérieure à 3 et prendre en compte la présence de bruit [26, 28]. L'un de ces modèles appelé Witness Complex [106] a aussi été étudié initialement par la communauté de l'Apprentissage Automatique sous le nom de Topology Representing Network [88] et a constitué la base des mes travaux dans ce domaine.

4.5.2 Le Topology Representing Network et le Witness Complex

Les modèles descriptifs ou ceux utilisés en reconstruction de formes sont habituellement basés sur la construction d'un graphe ou d'un complexe simplicial dont les sommets sont les individus. Si l'on fait un parallèle avec les modèles statistiques, les fenêtres de Parzen

[109] pour l'estimation de densité basée sur la somme de fonctions noyaux de type gaussiennes centrées sur les individus, ou la version basée sur le volume minimal de la boule centrée sur ceux-ci et englobant leurs K plus proches voisins, sont de tels modèles non paramétriques basés sur les individus seuls et sur un méta-paramètre d'échelle ou de régularisation (la largeur des fenêtres ou le nombre K). De ces modèles dérivent leurs homologues paramétriques, les modèles de mélange de lois de probabilité comme les mélanges de gaussiennes [83], et les modèles de quantification vectorielle comme les K -moyennes [84].

Les K -moyennes fournissent un moyen de quantifier un échantillon en remplaçant ses N individus par K prototypes (individus virtuels vivant dans le même espace mais en moindre nombre). En Quantification Vectorielle [3, 87], l'objectif est de positionner les prototypes de manière à minimiser la distorsion qui subsiste lorsque l'on remplace chacun des N individus par son prototype le plus proche. La méthode des K -moyenne est la plus connue pour accéder à un minimum local de cette distorsion¹³. Un élément remarquable de ces méthodes de quantification vectorielle, est que par construction les prototypes ont tendance à se positionner au centre de gravité "local" donc "au milieu" plutôt que "sur les bords" du nuage de points.

De nombreux modèles dérivent de la quantification vectorielle, en remplaçant les variétés ponctuelles que sont les prototypes par des variétés linéaires comme l'analyse en composantes principales [63], ou non linéaires comme les courbes et surfaces principales [56], et leurs homologues discrètes que sont les cartes auto-organisées [68]. Aucun de ces modèles n'est cependant suffisamment flexible pour apprendre la connexité de la population puisqu'il l'impose *a priori*.

Le modèle Topology Representing Network [88] est le plus proche de la solution que je recherchais. Ce modèle construit un graphe à partir d'un ensemble de M prototypes positionnés sur le nuage de points par exemple par une méthode de quantification vectorielle. L'algorithme de construction appelé Competitive Hebbian Learning (**CHL**) connecte entre eux pour chaque individu son premier et son deuxième plus proches prototypes. Ces connexions forment un graphe appelé Triangulation Induite de Delaunay¹⁴ (**TID**) qui a la propriété remarquable d'être un sous-graphe du graphe de Delaunay des M prototypes. On constate sur des exemples jouets en

13. Nous avons étudié avec Lepetz et Nemoz-Gaillard dans [77] les propriétés de ces fonctions d'énergie et défini un pseudo-potentiel permettant de regrouper dans un cadre unique un grand nombre de méthodes de quantification vectorielle.

14. Ce n'est en fait pas une triangulation mais seulement un graphe, les facettes triangulaires ne sont pas construites par le CHL.

deux ou trois dimensions, que la TID tend à reproduire visuellement la structure topologique en terme d'arc-connexité de la population génératrice de l'échantillon. Le CHL est une méthode descriptive s'appuyant sur un ensemble de prototypes et les individus qu'ils représentent, elle n'optimise aucun critère spécifique. Elle ne fait que traduire en terme de liens entre prototypes (matrice d'adjacence), la position relative entre individus et prototypes.

Le CHL a été étendu au cas des complexes simpliciaux générant une structure appelée Witness Complex par De Silva et Carlsson [106]. Le Witness Complex d'un ensemble de prototypes et d'individus est un sous-complexe du complexe de Delaunay de ces prototypes. Cependant, ce modèle possède plusieurs défauts du point de vue de l'apprentissage statistique, que j'ai décrit dans [8, 9], en particulier : l'impossibilité de décrire une source ponctuelle d'individus par un sommet isolé du complexe (chaque composante connexe compte au moins deux prototypes) ; la sensibilité au bruit (un individu suffit à générer un simplexe) ; l'existence d'individus non exploités pour la construction du modèle (existence de zones mortes dans lesquelles la présence d'individus est ignorée pour les k -simplexes avec $k > 1$) ; l'absence de métrique permettant de mesurer la qualité du modèle ; et la non consistance mise en évidence par le fait que considérant un échantillon dense des simplexes d'un Witness Complex plongé, le Witness Complex construit à partir des mêmes sommets et de cet échantillon ne correspond pas nécessairement au Witness Complex générateur, en d'autres termes, un Witness Complex n'est pas nécessairement homéomorphe au sous-complexe de Delaunay de mêmes sommets population sous-jacente à l'échantillon aussi dense que l'on veut qui l'a généré. La non consistance est en fait une propriété de la Triangulation de Delaunay Restreinte [39]. Considérant un ensemble de prototypes, sa Triangulation de Delaunay Restreinte à une forme est équivalente au Witness Complex d'un échantillon dense de cette forme. Shah et Edelsbrunner montrent que la Triangulation de Delaunay Restreinte à une forme est homéomorphe à cette forme si celle-ci a certaines propriétés relatives à son intersection avec le complexe cellulaire de Voronoï des prototypes. A partir de ces propriétés il est trivial de montrer que la Triangulation de Delaunay Restreinte à son propre plongement ne lui est homéomorphe que si elle est un complexe de Gabriel, *i.e.* un complexe de Delaunay tel que l'intersection de toute k -face de ce complexe avec la $(D - k)$ -face du complexe cellulaire de Voronoï de ses sommets est non vide. Cette propriété dépend uniquement de la position relative des prototypes entre eux ce qui restreint la famille de complexes simpliciaux possibles pour modéliser par cette

approche la forme sous-jacente à un échantillon. La figure 21 illustre ces différentes limites.

Le Witness Complex apparaît donc comme un moyen simple et peu coûteux de générer un sous-complexe de Delaunay des prototypes, dont la topologie semble s'approcher sans que cela n'ait été quantifié, de celle de la population sous-jacente au nuage de points. Ses limites m'ont poussé à chercher une autre solution dans le contexte de l'apprentissage statistique de la topologie. Le besoin de consistance, le besoin d'une mesure de qualité de l'estimation et l'équivalence entre la méthode des K -moyennes dont la TID est un prolongement, et certains modèles de mélange de Gaussiennes¹⁵ m'ont conduit à imaginer une méthode de construction d'un complexe simplicial qui de même prolongerait ces modèles de mélange, *i.e.* une solution issue du domaine de l'Apprentissage Automatique basée sur les modèles génératifs.

4.5.3 Les modèles génératifs

Les modèles génératifs modélisent le processus de génération des données. Un modèle de mélange est un modèle génératif combinaison convexe de composants élémentaires chacun défini par une fonction densité de probabilité. Dans ce processus génératif, l'un des composants est sélectionné suivant une loi multinomiale de paramètre les proportions des composants du mélange, puis un individu de l'échantillon est généré suivant la loi associée au composant choisi (la loi Normale dans le cas gaussien). L'apprentissage consiste à ajuster les paramètres du modèle génératif (proportions du mélange, paramètres des composants, nombre de composants) afin de maximiser la vraisemblance du modèle étant donné l'échantillon, pénalisée par un terme fonction du nombre de paramètres et de la taille de l'échantillon afin de gérer le compromis biais-variance¹⁶.

Un modèle génératif a par construction deux propriétés qui font défaut au Witness Complex : un critère de qualité défini par la vraisemblance pénalisée, et la consistance parce qu'il est par construction asymptotiquement le meilleur modèle au sens de ce critère, des

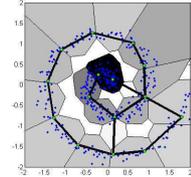
15. On peut démontrer que les K -moyennes sont strictement équivalent à un modèle de mélange de gaussiennes dont les variances de chaque composant sont diagonales et identiques, et la proportion des composants est identique, et pour lequel on optimise la vraisemblance par la méthode Classification Expectation-Maximisation [27].

16. Un modèle muni de trop nombreux degrés de liberté, permet une adéquation forte aux individus de l'échantillon ce qui génère un biais faible sur l'estimateur mais une grande variance suivant l'échantillon. On parle aussi de sur-apprentissage. Inversement un modèle trop peu complexe, insuffisamment flexible (constant ou linéaire par exemple), génère une variance faible face à différents échantillons d'une même population mais entraîne un biais trop élevé par rapport à l'échantillon fourni. Sélectionner le modèle de complexité optimale revient à gérer le compromis biais-variance [49].

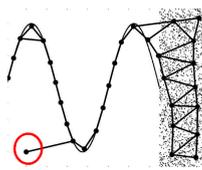
Figure 21

Limites du Witness Complex

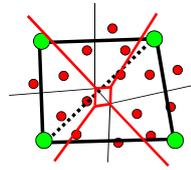
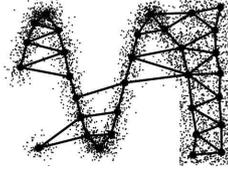
Principe : à chaque k -simplexe de Delaunay d'un ensemble de prototypes (points verts) correspond une cellule de Voronoï d'ordre $k+1$ (polygones gris). Un individu (point bleu) dans une cellule active le simplexe correspondant (trait noir)



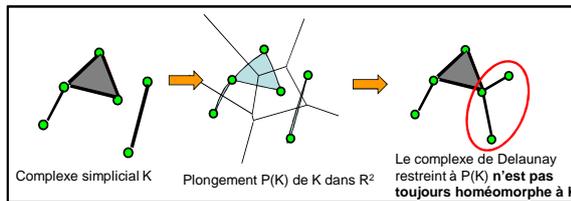
Pas de composante isolée



Sensibilité au bruit



Région d'influence trop petite pour permettre l'apparition de l'arête diagonale à partir de l'échantillon (points rouges)



Non consistance

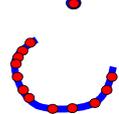
Figure 22

Un cadre génératif pour l'apprentissage automatique de la topologie

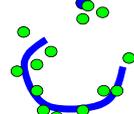
Hypothèses générales sur le processus statistique de génération des données



Des variétés principales inconnues...

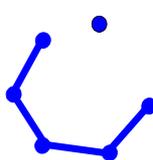


...desquelles sont tirés aléatoirement des individus avec une distribution de densité inconnue...

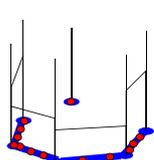


...corrompus par un bruit de nature inconnue menant aux observations

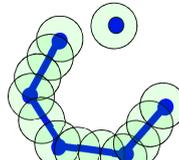
Hypothèses simplifiées – modèle du graphe génératif gaussien



Il existe un graphe G ayant la même connectivité...



...auquel on associe une distribution uniforme sur chaque arête, et multinomiale sur les sommets...



...convoluée avec un bruit gaussien centré et isovarié

données qu'il génère.

Dans le cas des modèles de mélanges de gaussiennes, ce processus peut se décomposer naturellement en deux parties auxquelles on peut attacher une sémantique différente : un terme source ou structurel, et un terme perturbateur. La partie structurelle est représentée par les moyennes ou centres des gaussiennes, générateurs ponctuels de données (les variétés latentes). La partie perturbatrice est représentée par une densité gaussienne convoluée à la partie structurelle, éparpillant chaque individu au voisinage de sa source suivant une loi normale. Ce terme perturbateur peut dans certains cas permettre d'extraire une information utile sur la corrélation locale des variables lorsque la matrice de variance de la loi normale est incluse dans les paramètres libres du modèle. Le modèle que j'ai proposé, comme le modèle d'analyse en Composante Principales Localisées [60], transfère cette information de corrélation vers la partie structurelle du modèle pour ne maintenir dans la partie perturbatrice que les résidus non informatifs.

Les modèles de mélange sont intéressants car ils sont explicatifs, ils fournissent un moyen de décrire à l'analyste la nature en termes de source et de perturbation de la population sous-jacente à l'échantillon, vu sous l'angle de ce modèle au maximum de vraisemblance. Ce sont aussi des modèles prédictifs permettant de mesurer la propension de chaque composant du modèle à avoir généré un nouvel individu et d'assigner à celui-ci les propriétés (classe d'appartenance par exemple) attachées au composant générateur le plus probable. En plus des méthodes classiques de validation croisée, il existe pour ces modèles probabilistes des critères de pénalisation comme le critère BIC [104] qui permettent de trouver un compromis objectif entre biais et variance sans nécessiter de rééchantillonnage coûteux en temps de calcul ni de partitionnement de la base de données n'exploitant pas toutes les données disponibles.

J'ai donc proposé un modèle de mélange dont le terme source n'est plus un ensemble de points mais un graphe voire un complexe simplicial.

4.6 Le graphe génératif

4.6.1 Contexte

L'Apprentissage de la Topologie est un domaine récent en Apprentissage Statistique. Les travaux de Martinetz et Schulten [88] sur le Topology Representing Network et ceux de Fritzke [42, 43] qui a proposé des versions du TRN à structures dynamiques dans les an-

nées 90 sont les premiers à abandonner le support trop contraint des cartes auto-organisées de Kohonen pour explorer cette voie dans le cadre de l'apprentissage automatique. J'ai proposé à Pierre Gaillard, Gilles Gasso, Frédéric Chazal et David Cohen-Steiner de se joindre à moi afin de relancer l'intérêt pour ce domaine en organisant l'atelier Topology Learning lors de la conférence NIPS 2007 [16]. Mon objectif est de développer des méthodes basées sur la théorie de l'Apprentissage Statistique pour retrouver les invariants topologiques des formes ou structures sous-jacentes à un nuage de points multi-dimensionnel.

J'ai proposé initialement dans [9] le principe d'un graphe génératif pour extraire la connexité d'un nuage de points. Avec Pierre Gaillard et Gérard Govaert, nous avons approfondi l'étude de ce modèle dans [47][48] et l'avons décliné dans différents cadres applicatifs. Je résume ici les fondements de notre approche du problème de l'Apprentissage de la Topologie.

Dans le cadre de la régression non linéaire, on estime une fonction complexe en combinant des fonctions élémentaires de base issues d'une famille de fonctions suffisamment riche pour que s'appliquent les théorèmes d'approximation universelle [32]. Pour modéliser des variétés inconnues, nous suivons la même approche en considérant un modèle obtenu par l'assemblage de variétés issues d'une famille suffisamment riche pour que l'on puisse obtenir la complexité nécessaire. Les complexes simpliciaux ont l'intérêt de fournir une telle famille puisqu'ils permettent de trianguler toute variété, et de permettre en plus le calcul des invariants topologiques que l'on souhaite extraire [107].

4.6.2 Le cadre génératif

Nous posons le problème de l'Apprentissage de la Topologie comme un problème génératif (Figure 22) : soit \mathcal{M} un espace topologique latent et \mathbb{R}^D l'espace Euclidien (espace des observations ou espace ambiant). Les variétés principales \mathcal{M}^{prin} sont définies comme l'image de \mathcal{M} par une fonction $f : f(\mathcal{M}) = \mathcal{M}^{prin}$. Cependant, en pratique, nous ne connaissons ni \mathcal{M} ni f , et nous devons nous contenter d'un ensemble fini de points \underline{x} représentant les M données observées, au lieu d'un ensemble de variétés \mathcal{M}^{prin} . Les points $x \in \underline{x}$ sont les images par l'application f de points "cachés" $\underline{z} \subset \mathcal{M}$, issus de \mathcal{M}^{prin} suivant une certaine fonction densité de probabilité p^{prin} , et potentiellement corrompus par un bruit de nature inconnue $\epsilon : \underline{x} = f(\underline{z}) + \epsilon$. Nous souhaitons extraire la connexité de \mathcal{M} à partir de l'observation du nuage de points \underline{x} . L'application f peut modi-

fier la géométrie de \mathcal{M} et éventuellement sa topologie. Cependant, nous supposons que f est un homeomorphisme, donc que la topologie de \mathcal{M} est la même que celle de \mathcal{M}^{prin} , et donc que nous n'avons pas besoin d'estimer f . Il suffit de nous focaliser sur l'extraction de la topologie de \mathcal{M}^{prin} . Pour cela, nous proposons de modéliser \mathcal{M}^{prin} avec un modèle statistique basé sur un graphe permettant l'extraction de cette topologie (en particulier la connexité) avec des méthodes de l'état de l'art [106]).

Nous introduisons le modèle génératif en simplifiant les hypothèses générales ci-dessus (Figure 22). Au lieu de considérer tout type de variété pour \mathcal{M}^{prin} , et du fait des propriétés désirables des complexes simpliciaux, nous supposons que \mathcal{M}^{prin} est un sous-graphe G du graphe de Delaunay¹⁷ de quelques prototypes positionnés dans l'espace ambiant. Au lieu de supposer toute densité de probabilité p^{prin} sur \mathcal{M}^{prin} , nous supposons que p^{prin} est uniforme le long de chaque arête du graphe G . Et au lieu de supposer n'importe quel type de bruit ϵ , nous supposons un bruit gaussien additif isovarié de moyenne nulle et de variance σ . Le modèle génératif obtenu est donc une somme pondérée de fonctions de densité basées sur les composantes élémentaires du graphe (ses arcs et ses sommets), convoluées avec un bruit gaussien. En tant que modèle génératif, et en considérant un réglage des paramètres et de la complexité du modèle par le critère BIC (parcimonie et mesure de la qualité), ce modèle est par construction le meilleur modèle de densité des points qu'il génère (consistance)¹⁸.

4.6.3 Définition du modèle de graphe génératif

Etant donné un ensemble de prototypes \underline{w} positionnés au voisinage d'un nuage de points (les individus) avec un modèle de mélange gaussien isovarié, le graphe de Delaunay (DG) des prototypes est construit¹⁹. Chaque arête et chaque sommet du graphe est la base d'un modèle génératif, de sorte que le graphe génère un mélange de densités gaussiennes. Le modèle de mélange résultant représente les données à partir de ces éléments génératifs que nous appelons "points Gaussiens" et "segments Gaussiens", constituant le "Graphe Génératif Gaussien" (GGG).

17. C'est donc le 1-squelette du complexe simplicial de Delaunay, pour lequel, comme pour tout complexe simplicial plongé dans l'espace des prototypes, aucun D -simplexe ne chevauche un autre hormis sur leur éventuelle face commune, en particulier G est un graphe planaire si $D = 2$.

18. Cela ne signifie cependant pas que le modèle est identifiable : il pourrait exister des modèles ayant même densité et même complexité donc même score BIC mais ayant des paramètres et éventuellement une topologie différents. Je discute de cet aspect dans les perspectives.

19. Les algorithmes pour construire le graphe de Delaunay sont fournis dans [18, 2]

La valeur de la densité d'un point Gaussien centré sur un prototype $w_j \in \underline{w}$ et de variance σ^2 , calculée en un point $x_i \in \underline{x}$ est définie par :

$$g_j^0(x_i; \sigma) = g^0(x_i|w_j; \sigma) = (2\pi\sigma^2)^{-D/2} \exp\left(-\frac{(x_i - w_j)^2}{2\sigma^2}\right) \quad (1)$$

Un segment gaussien normalisé est défini comme la somme d'un nombre infini de points gaussiens régulièrement répartis le long d'un segment de droite. Il s'agit donc de l'intégrale d'un point gaussien le long d'un segment de droite (Figure 23). La valeur en un point x_i du segment gaussien $[w_{a_j}, w_{b_j}]$ associé à la $j^{\text{ème}}$ arête $\{a_j, b_j\}$ de longueur L_j du graphe de Delaunay, de variance σ^2 est donnée par :

$$\begin{aligned} g_j^1(x_i; \sigma) &= g^1(x_i|\{w_{a_j}, w_{b_j}\}; \sigma) = \frac{1}{(2\pi\sigma^2)^{\frac{D}{2}} L_j} \int_{w_{a_j}}^{w_{b_j}} \exp\left(-\frac{(x_i - t)^2}{2\sigma^2}\right) dt \\ &= \frac{\exp\left(-\frac{(x_i - q_j^i)^2}{2\sigma^2}\right)}{(2\pi\sigma^2)^{\frac{D-1}{2}}} \cdot \frac{\operatorname{erf}\left(\frac{Q_j^i}{\sigma\sqrt{2}}\right) - \operatorname{erf}\left(\frac{Q_j^i - L_j}{\sigma\sqrt{2}}\right)}{2L_j} \end{aligned} \quad (2)$$

où $L_j = \|w_{b_j} - w_{a_j}\|$, $Q_j^i = \frac{(x_i - w_{a_j})(w_{b_j} - w_{a_j})}{L_j}$ et $q_j^i = w_{a_j} + (w_{b_j} - w_{a_j}) \frac{Q_j^i}{L_j}$ est la projection orthogonale de x_i sur la droite passant par w_{a_j} et w_{b_j} . Dans le cas où $w_{a_j} = w_{b_j}$, nous posons $g_j^1(x_i; \sigma) = g^0(x_i|w_{a_j}; \sigma)$.

Dans l'équation (2), la partie gauche du produit représente le bruit gaussien orthogonal au segment, et la partie droite le bruit gaussien intégré le long du segment (convolution). Les fonctions g^0 et g^1 sont positives et l'on démontre que leur intégrale sur \mathbb{R}^D vaut 1, donc que ce sont des fonctions densités de probabilité.

Un point gaussien est associé à chaque prototype de \underline{w} et un segment gaussien à chaque arête du graphe de Delaunay (DG). Le mélange de gaussiennes est obtenu par une somme pondérée des N_0 points gaussiens et N_1 segments gaussiens, de telle sorte que la somme des poids $\underline{\pi}$ vaut 1 et qu'ils soient positifs ou nuls :

$$p(x_i; \Theta) = \sum_{d=0}^1 \sum_{i=1}^{N_d} \pi_j^d g_j^d(x_i; \sigma) \quad (3)$$

avec $\sum_{d=0}^1 \sum_{j=1}^{N_d} \pi_j^d = 1$ et $\pi_j^d \geq 0 \forall j, d$ et où $\Theta = \{\underline{\pi}, \underline{w}, \sigma, DG\}$ représente l'ensemble des paramètres du modèle. Le poids π_j^0 (resp. π_j^1) est la probabilité a priori qu'une donnée x soit tirée du point gaussien associé à w_j (resp. du segment gaussien associé à la $j^{\text{ème}}$ arête de DG).

Ainsi l'espace latent est un ensemble de points et de segments, qui sont plongés dans l'espace ambiant par la réalisation géométrique du graphe de Delaunay. De plus, en toute généralité, les segments

latents sont définis de longueur 1, ce qui permet de définir la distribution a priori d'une variable cachée t pour chaque point et chaque segment. Dans notre cas, la distribution a priori sur t pour un point est la distribution de Dirac, et pour un segment, la distribution uniforme.

4.6.4 La vraisemblance et sa maximisation avec l'algorithme EM

La fonction $p(x_i; \underline{\pi}, \underline{w}, \sigma, DG)$ est la densité de probabilité au point x_i sachant les paramètres du modèle. Nous mesurons la vraisemblance P des données \underline{x} par rapport aux paramètres $\Theta = \{\underline{\pi}, \underline{w}, \sigma, DG\}$ du modèle GGG :

$$P(\Theta; \underline{x}) = \prod_{i=1}^M p(x_i; \underline{\pi}, \underline{w}, \sigma, DG) \quad (4)$$

Afin de maximiser la vraisemblance P par rapport aux paramètres $\underline{\pi}$, \underline{w} et σ , nous utilisons le cadre *EM* [35]. Dans mes travaux préliminaires [9] je n'ai dérivé les règles que pour l'optimisation de $\underline{\pi}$. Pierre Gaillard dans sa thèse [45] a résolu les difficultés techniques en se plaçant dans le cadre GEM pour déterminer les règles d'optimisation de \underline{w} et σ . D'autres méthodes d'optimisation sont envisageables. EM a l'avantage de ne pas nécessiter de réglage de métaparamètres, et pour inconvénient d'être lent à converger. Le développement *in extenso* des règles de mise à jour des paramètres est fourni dans [45]. Le principe du GGG est présenté sur la figure 24.

4.6.5 Emergence de la topologie et sélection de modèle par le critère BIC

Pour obtenir le graphe représentant la topologie (TRG) à partir du modèle génératif, l'idée clef est de supprimer du graphe DG des prototypes, les éléments gaussiens qui n'ont aucune chance d'avoir généré des données, *i.e* les éléments associés à un poids faible : $\pi_j^d < \gamma$.

Soit $G = \{\sigma, \underline{w}, E, \underline{\pi}\}$ le graphe génératif défini par sa variance du bruit σ^2 , ses N_0 sommets \underline{w} , son ensemble d'arcs E et ses proportions $\underline{\pi}$. Et soit $G_{N_0, \gamma} = \{\sigma, \underline{w}, E, \underline{\pi} | \pi_j^d \geq \gamma\}$, le graphe génératif qui contient seulement les éléments génératifs ayant une proportion plus élevée que le seuil $\gamma : \pi_j^d \geq \gamma, \forall d \in \{0, 1\}$.

En réglant le paramètre γ de 1 à 0, on obtient une séquence de graphes génératifs emboîtés allant de l'ensemble vide au graphe DG complet :

$$G_1 = \emptyset \subseteq \dots \subseteq G_\gamma \subseteq \dots \subseteq G_0 = DG \quad (5)$$

Figure 23

Les composants du modèle génératif

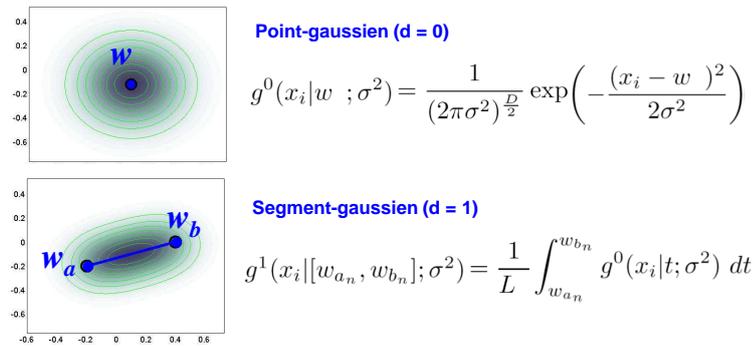
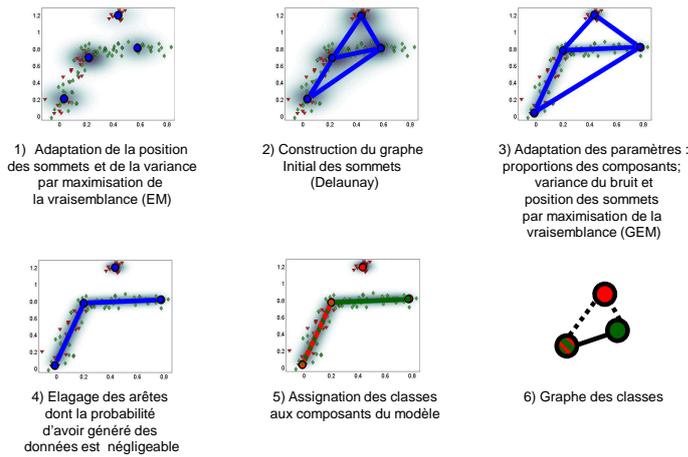


Figure 24 Principe de l'apprentissage de la connexité avec le GGG par maximisation de la vraisemblance



Supposons que l'un des graphes de la séquence ait la "bonne" topologie, *i.e.* celle inconnue des variétés principales, nous utilisons le Critère d'Information Bayésien (BIC) pour sélectionner ce graphe. Le critère BIC [104] est adapté et satisfaisant en pratique pour la sélection du nombre de composants d'un modèle de mélange classique [97, 41]. Ici nous l'utilisons pour régler le compromis entre la vraisemblance $P(\underline{x}; G_\gamma)$ et la complexité du graphe génératif G_γ . Donc nous supposons que la topologie du modèle génératif d'un ensemble de points est un estimateur de la topologie des variétés principales de cet ensemble, dont la qualité est mesurée par le critère BIC de vraisemblance pénalisée :

$$BIC(N_0, G_\gamma) = -\log(P(\underline{x}; G_\gamma)) + \frac{v_\gamma}{2} \log(M) \quad (6)$$

où v_γ est le nombre de paramètres libres du modèle G_γ et M le nombre de données.

La vraisemblance de chaque graphe génératif G_γ de la séquence est optimisée en fonction des proportions²⁰ $\underline{\pi}$ de telle sorte que tous les graphes génératifs sont à leur maximum de vraisemblance. A la fin, le graphe représentant la topologie pour un nombre de prototypes N_0 donné, est celui défini par le graphe génératif G_{γ^*} minimisant le critère BIC à N_0 fixé (eq. (6)). On recommence pour chaque valeur N_0 et l'on retient finalement le TRG optimal associé au couple de paramètres (N_0^*, γ^*) qui minimise BIC.

4.6.6 Exemples d'applications

Nous avons décrit plusieurs applications dans différentes publications [46, 45, 48] qui confirment l'intérêt du graphe génératif.

Une application en classification automatique se base sur les composants connexes du graphe élagué pour constituer les classes. Nous avons sélectionné les images de 5 objets de la base d'objets fournie par l'Amsterdam Library of Object Images (ALOI) [50]. Pour chaque objet, un camera a enregistré 72 images en faisant tourner l'objet autour d'un axe vertical avec un pas angulaire de 5 degrés. La taille des images est réduite à 12×16 pixels. Nous nous attendons à trouver dans l'espace des pixels à 192 dimensions, 5 classes de 72 points chacune correspondant aux images d'un objet particulier. Comparé au Witness Complex, le GGG retrouve plus souvent les bonnes classes. Les modèles de mélange classiques fournissent trop

²⁰. Pour des raisons de temps de calcul, seuls les proportions $\underline{\pi}$ sont optimisées durant cette étape. En commençant par les proportions normalisés obtenus avec le GGG complet, le problème est convexe et l'algorithme converge rapidement. La variance du bruit est aussi supposée peu variable.

de composants (16 pour chaque objet) du fait de la structure en filaments non linéaires des images d'un même objet dans l'espace des pixels.

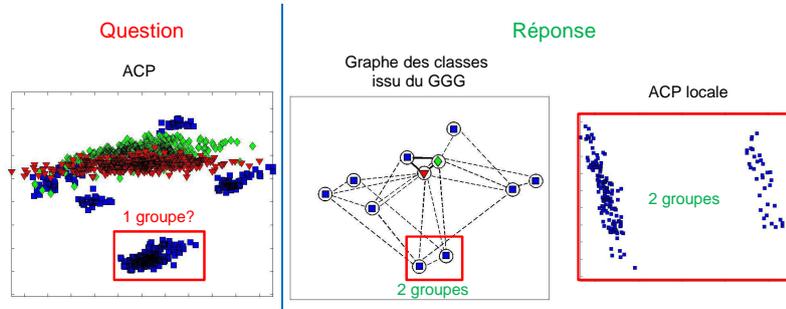
Pour l'analyse exploratoire de données étiquetées, nous construisons le GGG dans un cadre supervisé, chaque composant peut représenter simultanément plusieurs classes et leurs proportions ce qui permet de détecter un chevauchement des classe. Les composantes connexes du graphe sont extraites suivant la pureté des classes de chaque composant, puis le graphe des classes est construit en associant à chaque composante de classe un noeud et en reliant deux noeud par une lien de "contact" si les composantes de classes sont reliées après élagage, et par un lien de "voisinage" si elles sont reliées avant élagage mais plus après (deux classes en vis-à-vis de part et d'autre d'une région de faible densité). Ce graphe des classes fournit plus d'informations que le graphe des classes simple basé sur les données directement, puisqu'il indique le nombre de composantes de chaque classe (topologie intra-classes), le degré de chevauchement, le contact et le voisinage entre les classes (topologie inter-classes), alors que le graphe des classes simple n'indique ni le chevauchement ni le voisinage. Sur différents jeux de données, la structure topologique des classes révélée par le graphe des classes confirme ce que montre la méthode ProxiViz de visualisation des similarités in situ (Figures 25a, 25b et 25c).

En apprentissage semi-supervisé, certaines données sont étiquetées et la majorité ne le sont pas, il s'agit alors d'inférer la classe des données non étiquetées connaissant celle des données étiquetées, et disposant à l'avance de l'ensemble des données étiquetées ou non. Comparé au cas de l'apprentissage supervisé, on peut ici s'appuyer sur la structure des données pour propager de proche en proche les étiquettes de classe à partir des données étiquetées. Pierre Gaillard a proposé dans [48] de construire un GGG mais sans élagage. Les proportions du mélange associées aux arêtes du graphes indiquent comment propager les étiquettes de classes associées à chaque sommet (Figure 26). La méthode fournit des résultats au moins aussi bons que les meilleures méthodes de l'état de l'art, mais elle a pour avantage de ne pas nécessiter de fixer arbitrairement des méta-paramètres, en particulier pour déterminer le nombre de composants du modèle. En effet, le critère BIC est utilisé pour cela et il s'appuie sur l'ensemble des données indépendamment de leur absence ou présence d'étiquette, tandis que les méthodes de l'état de l'art nécessitent un réglage heuristique de la complexité du modèle, ou s'appuient sur une validation croisée peu fiable car elle se base sur le nombre de données étiquetées généralement très faible dans le

contexte semi-supervisé.

Pierre Gaillard a envisagé deux autres applications dans sa thèse. L'une consiste à débruiter les images de taille 28x28 pixels de la base de caractères manuscrits MNIST, en projetant chaque donnée sur le graphe du modèle GGG appris dans l'espace des pixels à 784 dimensions. Les premiers résultats montrent qu'un classifieur au sens des K-plus-proches-voisins est plus précis avec ce débruitage que sans, et qu'il est plus robuste au choix du paramètre K. Une autre application consiste à compléter les attributs géométriques des données avec des attributs topologiques. Pour cela on considère les chiffres manuscrit de la base MNIST, et l'on construit un GGG dans l'espace image à deux dimensions où une donnée est placée au milieu de chaque pixel de luminosité supérieure à un seuil. On détermine manuellement l'existence de 7 classes d'homotopie et le nombre d'images appartenant à chaque classe pour chaque chiffre. A partir de là, il devient possible d'ajouter à la donnée brute (position de l'image dans l'espace des pixels) une information topologique codée ici sous forme de classe d'homotopie pour aider le classifieur au sens des K-plus-proches-voisins à décider de l'étiquette à attribuer à un nouveau chiffre.

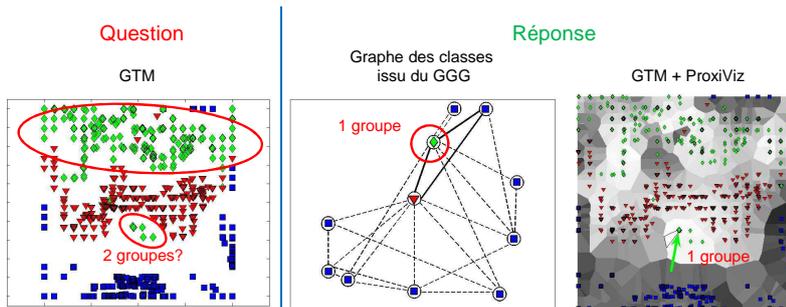
Figure 25a Analyse exploratoire de données étiquetées avec le GGG



Les données Oil Flow (1000 individus, 3 classes, 12 variables) sont projetées (à droite) sur les deux premiers axes principaux fournies par l'ACP. Un seul groupe de points de la classe bleue semble se détacher nettement en bas de la représentation (cadre rouge). Est-ce vrai?

Chaque sommet du graphe des classes obtenu à partir du GGG (au centre) est positionné sur le centre de gravité des individus qu'il représente (Maximum A Posteriori) puis projeté sur les deux premières composantes principales fournies par l'ACP. En bas de la représentation, deux sommets non connectés mais voisins apparaissent. Une ACP locale sur les seuls individus encadrés en rouge montre clairement deux composantes connexes, confirmant les conclusions obtenues à partir du graphe des classes : le groupe de points bleus en bas sur la vue de gauche est issu du chevauchement de deux groupes distincts dans l'espace d'origine.

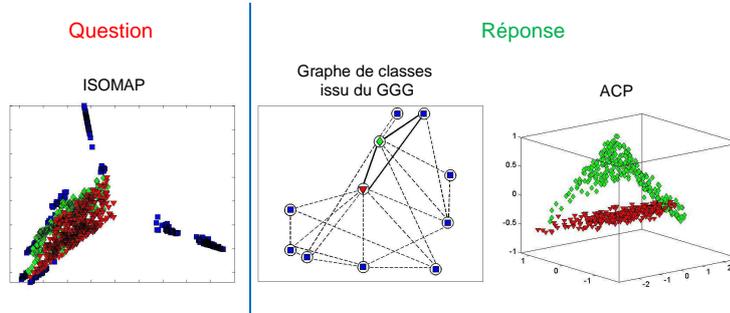
Figure 25b Analyse exploratoire de données étiquetées avec le GGG



Les données Oil Flow (1000 individus, 3 classes, 12 variables) sont projetées (à droite) par une carte topographique générative (GTM). La classe verte semble formée de deux composantes (ellipses rouges). Est-ce vrai?

Chaque sommet du graphe des classes obtenu à partir du GGG (au centre) est positionné sur le centre de gravité des individus qu'il représente (Maximum A Posteriori) puis projeté dans le même espace de projection que celui de gauche obtenu par GTM. Dans le graphe des classes, la classe verte est représentée par un unique sommet, signe qu'elle est d'un seul tenant. La vue de droite montre le GTM avec ProxiViz. L'un des individus projeté dans la composante verte du bas, est sélectionné comme point de référence (flèche verte). De nombreux points de la composante verte du haut s'éclaircissent fortement confirmant que les deux composantes n'en forment qu'une.

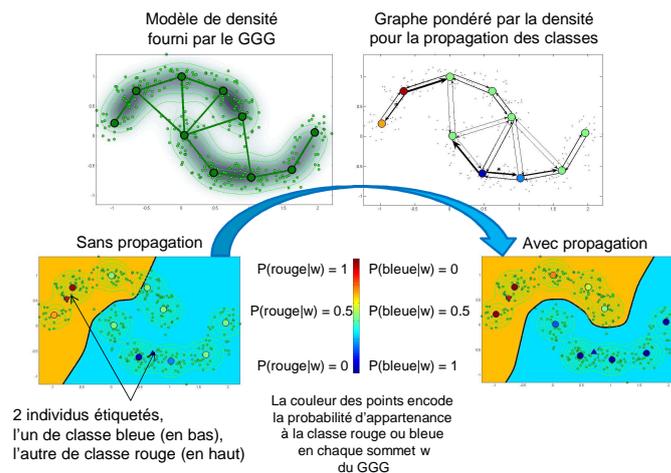
Figure 25c Analyse exploratoire de données étiquetées avec le GGG



Les données Oil Flow (1000 individus, 3 classes, 12 variables) sont projetées (à droite) par la méthode ISOMAP. La classe verte et la classe rouge semblent se chevaucher et former une seule composante mixte. Est-ce vrai?

Chaque sommet du graphe des classes obtenu à partir du GGG (au centre) est positionné sur le centre de gravité des individus qu'il représente (Maximum A Posteriori) puis projeté dans le même espace de projection que celui de gauche obtenu par ISOMAP. Dans le graphe des classes, la classe verte et la classe rouge sont représentées chacune par un unique sommet, et ces deux sommets sont connectés. La projection sur les trois premières composantes principales fournies par l'ACP (à droite) des individus de classes verte et rouge uniquement, confirme que les deux classes sont clairement séparées et chacune d'un seul tenant.

Figure 26 Apprentissage semi-supervisé avec le GGG



5 Conclusion et perspectives

5.1 En résumé

Les systèmes de mesure génèrent des masses de données à la fois en termes de nombre de variables (capteurs) et de nombre d'individus (taille de l'échantillon). Ces données doivent être analysées interactivement par un analyste. Les outils d'aide à la décision ont un rôle crucial : ils doivent appuyer la décision de l'analyste et pour cela lui rendre interprétables les informations contenues dans les données brutes pour lui permettre de construire un modèle du phénomène physique mesuré. Ce modèle peut être élaboré mentalement *a posteriori* à partir de l'analyse descriptive ou doit l'être mentalement *a priori* pour formuler une hypothèse prédictive. Dans tous les cas, les représentations graphiques, qu'elles soient conventionnelles ou perceptuelles, sont privilégiées pour transmettre l'information de l'outil vers l'analyste et lui permettre par l'exploration interactive de comprendre les données et leur modèle.

La Topologie est le domaine des mathématiques qui s'intéresse aux structures et à leurs invariants. Elle fournit le substrat sur lequel se greffent toutes les métriques dont les probabilités. La chaîne de mesure et de traitement transfère l'information du signal à l'écran qui lui-même la transmet aux systèmes visuel puis cognitif de l'analyste. L'ensemble des distorsions accumulées le long de cette chaîne ne permet pas de préserver complètement l'information initiale. Cependant, l'information de nature topologique est la plus robuste à ces distorsions car elle est préservée dans ce qu'elle a d'essentiel (la connexité au sens large), par homotopie, une classe très large de transformations contenant les homéomorphismes qui eux-mêmes contiennent les similitudes. Par ailleurs, la théorie psychologique de la Gestalt nous indique que l'information topologique telle que la connexité est perçue de manière pré-attentive par notre système visuel, sans surcharge de la mémoire de travail dont on sait que son empan est limité à moins d'une dizaine d'éléments. De plus, nous savons que les neurones du cortex cérébral s'organisent de manière à préserver au moins en partie la topologie des stimuli perçus. Il est donc probable que la préservation de cette information topologique joue un rôle significatif dans nos processus cognitifs puisqu'elle a résisté à la sélection Darwinienne. Enfin, les modèles que nous cherchons à produire doivent être interprétables par l'analyste, ils doivent faire sens pour lui, i.e. relier les grandeurs qu'ils produisent aux variables que celui-ci a introduit en entrée du modèle. Ces liaisons entre variables d'entrée et de sortie sont en premier lieu de

nature topologique et leur manifestation graphique est le meilleur moyen de les transmettre à l'analyste. Je fais donc l'hypothèse que la topologie est le substrat essentiel de l'interprétabilité.

Mes travaux ont consisté à développer des modèles permettant d'extraire cette information topologique à partir d'un ensemble d'individus décrits dans un espace multi-varié, et de transmettre cette information à l'analyste. Différents cas d'application ont été expérimentés en fouille de données et apprentissage statistique, notamment : la recherche de classes d'objets aux formes complexes dans des espace multidimensionnels (GGG); la compréhension des relations de connexité intra et inter classes de données étiquetées à travers leur projection fiabilisée (CheckViz et ProxiViz) ou automatiquement in situ par un modèle génératif (SGGG); ou encore la propagation d'étiquettes de classes le long des structures sous-jacentes aux données en apprentissage semi-supervisé (SSGGG).

Ces outils et les possibilités d'analyse et de traitement qu'ils offrent posent de nombreux problèmes encore non résolus et ouvrent plusieurs voies de recherches que je présente maintenant.

5.2 Visualisation d'information

5.2.1 Topologie et distortions

J'ai introduit dans [10] les distortions "topologiques" que j'ai appelées "déchirure" et "recollement". En pratique pourtant l'on ne peut accéder qu'à une information géométrique d'écart entre similarité désirée et similarité obtenue entre deux points. Les distortions topologiques ne sont en fait pas encore formalisées. Il faudrait pour cela définir une variété génératrice du nuage de points originel et une autre génératrice du nuage de point projeté, pour pouvoir définir précisément ces distortions et en donner la mesure objective. On peut envisager cette formalisation à partir d'un graphe encodant la connexité de ces variétés. Si deux sommets voisins du graphes originel ne sont plus voisins dans le graphe projeté, il y a déchirure. Si deux sommets voisins du graphe projeté ne sont pas voisins dans le graphe originel, il y a recollement. Cette définition n'est cependant pas suffisante car si l'on permute deux sommets voisins du graphe originel dans le graphe projeté, il apparait à la fois une déchirure et un recollement. Comment préciser cette définition ? Les mesures topologiques proposées pour évaluer le degré de préservation topologique des carte auto-organisées de Kohonen [118] pourraient peut-être être utilisées dans ce cas.

L'une des hypothèses de mes travaux est que l'information topo-

logique étant perçue pré-attentivement, il est préférable de la préserver au mieux dans la représentation graphique afin de décharger l'analyste de la tâche de la reconstituer, et qu'il concentre ses ressources cognitives sur d'autres tâches. Cette hypothèse implique que des points perçus comme proches dans la représentation graphique seront supposés pré-attentivement comme similaires dans l'espace originel, donc il est préférable que dans la majeure partie de la représentation les points proches soient effectivement similaires, et donc à choisir il vaut mieux que la représentation induise des déchirures plutôt que des recollements. Les méthodes de projection permettent de privilégier l'une ou l'autre des distorsions suivant la fonction de "stress" qu'elles minimisent :

- Soit la méthode tente au mieux et prioritairement de faire en sorte que les points voisins à l'origine demeurent voisins après projection quoiqu'il advienne par ailleurs (e.g. les projections linéaires ou la projection non linéaire de Sammon). Dans ce cas, des points éloignés originellement (non contraints par la fonction de stress) peuvent être projetés voisins et engendrer ainsi des recollements.
- Soit la méthode tente au mieux et prioritairement de faire en sorte que les points voisins après projection le soient aussi avant quoiqu'il advienne par ailleurs (e.g. l'Analyse en Composantes Curvilignes de Demartines et Hérault). Dans ce cas, des points éloignés après projection (non contraints par la fonction de stress) peuvent être voisins originellement et engendrer ainsi des déchirures.

Nicolas Heulot, dans le cadre de sa thèse [59], a mené une expérience auprès d'analystes dans une tâche d'inférence structurelle (comptage du nombre de classes de points originelles à partir du nuage de points projeté) pour vérifier que cette hypothèse est bien valide. Il apparaît que les méthodes privilégiant les déchirures sont significativement plus efficaces que les autres pour inférer la structure topologique du nuage de point originel. Nous avons aussi profité de cette expérience pour évaluer l'efficacité dans ProxiViz, de la représentation graphique des distances originelles par la coloration des cellules de Voronoï des points associés, comparée à la méthode d'interpolation des distances inverses de Shepard. L'interpolation de Shepard apparaît plus efficace que la coloration des cellules de Voronoï, elle-même plus efficace que la coloration des seuls points. Nous avons aussi montré que l'affichage des distances d'origine est significativement plus utile que celui du stress local pour inférer la structure topologique du nuage de points originel.

5.2.2 Paradigme WinSitu et critères de fiabilité et d'authenticité

Toute représentation graphique devrait être évaluée à l'aune du critère d'authenticité, afin d'assurer qu'elle montre bien une information exploitable²¹. Cela paraît une évidence car tel est le cas dans de nombreuses représentations graphiques, comme par exemple lors de la représentation classique d'une variable sur un axe linéaire gradué (repères orthogonaux ou parallèles) où le nom des variables et les unités de mesure sont affichées, ou par construction dans le cas de représentations par arbres, graphes, dendrogrammes dont la topologie tracée reflète exactement celle de ces structures. Pourtant ce critère n'a jamais été appliqué pour les cartes auto-organisées de Kohonen qui ont fait pourtant l'objet de milliers de publications, ni pour les projections non linéaires sans quantification vectorielle (ISOMAP, LLE, Sammon, ACC...). Notons qu'il existe des usages de la représentation par projection où le critère de fiabilité n'est que secondaire, en particulier lorsque la représentation est utilisée pour guider visuellement une recherche d'information (les noeuds de la carte représentent des sites internet par exemple [70]) qui sera simplement d'autant moins efficace qu'il y aura des distorsions. La disposition spatiale des mots d'un TagCloud n'a elle aussi pas d'autre objectif que de remplir un espace donné avec des mots de taille plus ou moins grande, représentant par exemple les thématiques principales d'un document. Les mots eux-mêmes portent l'information authentique, mais leur disposition spatiale n'est pas liée à une mesure de similarité entre les mots, la fiabilité n'est pas essentielle. Evaluer le caractère authentique de l'information montrée demeure néanmoins primordiale s'il s'agit de réaliser une tâche d'inférence structurelle comme en classification automatique.

Le critère d'authenticité ne précise pas quelle information doit être montrée parmi l'ensemble de celles disponibles. Plus généralement, bien qu'il existe de nombreuses heuristiques pour sélectionner la bonne variable graphique à utiliser suivant la nature de la variable originelle à visualiser, et même une grammaire décrivant toutes les combinaisons possibles de variables graphiques (la forme) [121], il n'existe pas de représentation graphique synthétique du contenu sémantique (le fond) de la représentation graphique choisie. Une telle représentation permettrait une comparaison de deux représentations graphiques non sur la forme mais sur le fond de ce qui est montré. En l'état actuel de l'art, nous disposons de représentations graphiques d'un ensemble de variables, accompagnées d'une repré-

21. Que ce qui est montré graphiquement soit perçu visuellement est qualifié par le critère d'efficacité.

tation conventionnelle (texte, symboles) des informations montrées sur cette représentation graphique. Deux représentations graphiques différentes (la forme) d'un même ensemble de variables (le fond) permettent en les comparant de juger de l'efficacité et de l'expressivité de ces représentations, mais ne disent rien de l'information authentique qu'elles représentent. La représentation intermédiaire que j'envisage permettrait d'exprimer graphiquement le contenu sémantique des représentations graphiques, et donc de faciliter la comparaison du contenu informationnel brut objectif de deux représentations graphiques en supprimant la part subjective perturbante liée à leur réalisation particulière dans l'espace des variables graphiques, et ainsi de faciliter la recherche de nouvelles représentations graphiques.

Par définition cette représentation graphique du fond, devrait être sa propre représentation canonique. Au-delà, une telle représentation canonique pourrait servir à représenter les interactions avec l'analyste et fournir une sorte d'algèbre permettant d'explorer systématiquement l'ensemble des représentations possibles en amont de leur projection sur les variables graphiques.

5.2.3 ProxiViz et au-delà

Dans ProxiViz, la représentation des distances originelles de chaque point au point sélectionné est statique et utilise la couleur des cellules de Voronoï de ces points. On peut envisager de visualiser par ce biais d'autres informations authentiques à corrélérer visuellement à la disposition spatiale des points, par exemple la dimension intrinsèque locale. Pour accroître la richesse des configurations géométriques et topologiques discernables (structures emboîtées, entrelacées, chevauchantes...), on pourrait envisager une représentation spatiale à trois dimensions perçue grâce à un système d'immersion (lunettes 3D par exemple).

Mieux qu'une vue statique, l'interactivité de la représentation est aussi primordiale pour mieux appréhender la structure globale du nuage de points originel. On souhaiterait naviguer dans l'espace multi-dimensionnel d'origine au travers de la représentation ProxiViz, à la manière du GrandTour [6] ou de GraphDice [21] par exemple. Naviguer signifie aussi constituer un parcours exploratoire des données, en conserver une trace en tout point de laquelle on puisse relancer une nouvelle branche exploratoire, voire être guidé dans le choix des vues les plus prometteuses à explorer. Outre la mémorisation des vues traversées et du chemin les reliant, il faut aussi envisager des outils d'assistance à la modélisation structurelle qui permettent de construire à la volée une structure topologique

vraisemblable sous-jacente au nuage de points, telle le graphe des classes par exemple.

Des problèmes de représentation graphique se posent pour ProxiViz, lorsque le nombre de points devient trop important pour qu'ils soient tous représentés sans occlusion. D'une part, un trop grand nombre de points signifie un trop grand nombre de vues différentes à explorer (une vue possible par point sélectionné). Un outil de guidage automatique vers les vues prototypiques (zones de forte confiance où les vues voisines sont similaires, les distorsions sont faibles et les structure géométriques montrées sont fidèles à celle originelles) et les vues atypiques (zones de fortes distorsions nécessitant une analyse détaillée) serait nécessaire pour cela. Les méthodes d'extraction de la topologie *in situ* permettraient de détecter les points associés à des vues pertinentes : loin des bords d'une variété d'origine pour les vues prototypiques, ou au niveau des points de raccord entre variétés d'origine pour les vues atypiques. D'autre part, les approches classiques d'agrandissement et de translation pourraient être adaptées à ProxiViz.

Est-il envisageable d'utiliser un même paradigme de représentation (ici les projections non linéaires) pour représenter à la fois les données que l'on explore et l'historique des vues de ces données générées par l'exploration ? J'étudie ces différentes pistes avec Nicolas Heulot.

5.2.4 Formaliser l'interprétabilité

L'authenticité et la fiabilité paraissent être des critères essentiels à l'interprétabilité d'une représentation. Les critères constitutifs de l'interprétabilité sont des critères de forme induits par nos limites physiologiques (empan mnésique, capacité du système visuel et cognitif...). La fiabilité est un facteur d'amélioration de l'efficacité de la représentation. L'authenticité est le seul critère de fond qui requiert que ce qui est montré soit lié le plus "directement" possible au phénomène que l'on veut analyser. Formaliser l'interprétabilité nécessiterait ainsi de préciser pour des fonctions de transfert élémentaires leur degré de préservation de la sémantique introduite dans les variables d'entrée. Quels sont les critères que doit respecter une fonction de transfert pour que sa sortie soit interprétable par l'analyste ? Pouvons-nous estimer ce degré d'interprétabilité a priori par composition des degrés d'interprétabilité élémentaires des sous-fonctions qui composent la fonction de transfert ? Ou devons-nous nous contenter de ressentir *a posteriori* ce degré d'interprétabilité ? L'interprétabilité est-elle objectivable ou restera-t-elle subjective ?

Dans quelle mesure l'implication de l'analyste dans la conception du modèle (la fonction de transfert) est-elle nécessaire pour lui rendre interprétable le résultat obtenu ?

5.2.5 Visualiser *in situ* d'autres types de données

La variété des données que l'on peut observer est sans fin : nuages de points, arbres, graphes, hypergraphes, multi-signaux, champs scalaires ou vectoriels, attracteurs dynamiques multidimensionnels... Comment l'approche ProxiViz et plus généralement le paradigme WinSitu peuvent-ils se décliner sur ces types de données ? Sont-ils génériques ou restreints au cas des nuages de points ?

L'information topologique peut-elle être visualisée autrement que par un graphe (graphe des classes par exemple) ? Il existe en fait des représentations de hiérarchies qui ne se basent pas sur le paradigme noeud-lien, mais sur l'adjacence de régions colorées [103] comme dans les TreeMaps [120]. Quel intérêt peut présenter l'approche ProxiViz appliquée à ce type de représentations ?

Les caractéristiques topologiques vont au-delà de l'arc-connexité encodable dans un graphe que l'on peut ensuite visualiser. Les nombres de Betti [26] représentent une information topologique plus complexe que l'arc-connexité (tunnels, cavités...), la dimension intrinsèque peut aussi être extraite. Quelles représentations graphiques permettraient d'appréhender ces caractéristiques ? Les codes-barres ou le diagramme de persistance topologique sont des moyens encore rudimentaires de représentation graphique, ils peuvent certainement être améliorés. On pourrait par exemple imaginer un graphe des classes situant les structures topologiques locales décrites par leurs nombres de Betti, les unes par rapport aux autres, ou intégrant la persistance topologique ou les structures multi-échelles.

5.3 Modélisation *in situ*

5.3.1 Au-delà des graphes

Nous avons montré comment construire un graphe génératif. Les invariants topologiques associés aux graphes sont limités : connexité et genre par exemple. Si l'on veut extraire des invariants topologiques plus fins pour caractériser la population sous-jacente au nuage de points, en particulier les nombres de Betti de rang supérieur à 0 (β_0 est le nombre de composantes connexes), il faut considérer l'extension aux complexes simpliciaux.

Dans ce cas des problèmes de complexité algorithmique et mémoire se posent. L'approche que nous suivons est celle de l'élagage,

elle consiste à construire une structure riche (le graphe de Delaunay des sommets) puis à l'élaguer grâce au critère BIC. Une telle approche n'est pas envisageable si l'on considère le complexe simplicial de Delaunay dont le nombre de simplexes de dimension D est en $O\left(N^{\lceil \frac{D}{2} \rceil}\right)$.

A l'inverse, l'approche constructive employée par la méthode des Witness Complex évite cet écueil puisque seuls les simplexes de Delaunay "utiles" sont construits et ils le sont itérativement par dimension croissante : d'abord les paires de sommets "témoins" plus proches voisins de chaque donnée, puis les triplets de sommets témoins déjà deux à deux témoins et ainsi de suite.

On peut envisager une voie similaire pour construire un complexe simplicial génératif par dimension croissante en élaguant au fur et à mesure les simplexes inutiles via un critère de vraisemblance pénalisée de type BIC : d'abord on construit toutes les arêtes du graphe de Delaunay ($O(N^3)$), on calcule les paramètres du modèle qui optimisent la vraisemblance pénalisée menant à l'élagage des arêtes inutiles pour expliquer les observations, puis on construit les triangles dont les arêtes sont encore présentes et l'on traite les paramètres associés de la même manière menant à l'élagage des triangles inutiles, puis on passe aux tétraèdres construits à partir des triangles restants et ainsi de suite jusqu'à ce qu'il ne puisse plus être construit de k -simplexe à partir des $(k - 1)$ -simplexes présents après élagage. Cette voie soulève cependant plusieurs problèmes : les ensembles fermés de k -simplexes de Delaunay ne forment pas nécessairement un $(k + 1)$ -simplexe de Delaunay, par exemple dans le plan, trois arêtes de Delaunay peuvent former un triangle non-Delaunay car contenant d'autres sommets du graphe de Delaunay dans son cercle circonscrit ; Par ailleurs, ce procédé reste une heuristique pour optimiser le critère de vraisemblance pénalisée qu'optimise l'approche directe qui considère le complexe simplicial de Delaunay dès le départ. Comment résoudre ces problèmes ?

5.3.2 Quel paradigme d'apprentissage ?

L'information topologique est extraite de façon non supervisée, le problème de sélection de modèle se pose naturellement dans ce contexte. Nous employons le critère BIC pour déterminer le "bon" nombre de sommets et le "bon" seuil d'élagage des arêtes. Ce critère pose un problème technique d'optimisation hybride mêlant des variables continues (les paramètres du modèle de mélange) et une variable discrète (le nombre de composants) qui ne permet pas d'utiliser les méthodes efficaces d'optimisation continue basées sur le gra-

dient comme la méthode du gradient conjugué ou celle de Levenberg-Marquardt par exemple. Peut-on envisager une version continue du critère BIC avec une entrée en jeu progressive des composants du modèle plutôt qu'en tout-ou-rien ? Il ne s'agirait plus du critère BIC à proprement parler mais d'un critère de pénalisation "à la" BIC. Sur quels fondements théoriques s'appuierait un tel critère de sélection de modèle ?

La méthode EM est adaptée à l'apprentissage d'un modèle de densité, tandis que l'élagage que nous imposons pour découvrir la structure topologique est similaire à une opération de classification (chaque composante connexe par exemple peut-être assimilée à une classe) à laquelle conviendrait mieux des proportions π_i "franches" pour les composants du modèle. La méthode CEM proposée dans [27] serait-elle plus adaptée ?

L'utilisation du graphe de Delaunay et plus généralement du complexe simplicial de Delaunay restreint la famille des variétés apprenables, mais permet d'assurer que la structure obtenue est un complexe simplicial, structure à partir de laquelle on sait extraire les invariants topologiques. La construction du graphe et davantage du complexe simplicial de Delaunay est aussi coûteuse, et devrait idéalement être mise à jour à chaque déplacement des sommets (les μ_i), ce qui pose par ailleurs la question de la stabilité de la structure et donc des invariants qui en sont extraits. En effet, un déplacement continu d'un sommet, entraîne un changement discret de la structure du complexe simplicial (apparition/disparition d'arêtes, et de simplexes). De plus la fonction de vraisemblance admet de multiples optima locaux sauf à ne considérer que les proportions π_i comme paramètres du modèle de mélange, la recherche de l'optimum global rend donc l'apprentissage coûteux en temp de calcul.

Aussi pourrait-on envisager d'utiliser comme structure de départ, le graphe complet ou le complexe simplicial complet (complexe simplicial abstrait non-homéomorphe à l'espace des données contenant $\sum_{k=0}^{N_0} \mathcal{C}_{N_0+1}^{k+1} = 2^{N_0+1} - 1$ faces où N_0 est le nombre de sommets), non plus basé sur des prototypes, mais sur les données elles-mêmes. Ainsi on pourrait procéder à la manière des méthodes parcimonieuses de type Machines à Vecteurs Supports [102] qui, à partir de l'optimisation d'un critère convexe, font émerger quelques données "clefs" (les vecteurs supports) suffisantes à décrire l'ensemble des données et des classes.

Dans les paramètres de ce modèle ne subsisteraient que les proportions π_i et la variance σ , et le critère d'optimisation basé sur la vraisemblance pourrait n'avoir qu'un optimum global. Cela éliminerait les problèmes d'optima locaux, et de construction initiale du

complexe simplicial, de sa mise à jour et donc de sa stabilité. Cependant cela renforcerait le problème de complexité mémoire lié au nombre de paramètres du modèle²² et poserait le problème d'assurer l'émergence d'une structure de complexe simplicial plongé (homéomorphe à un sous-espace de l'espace des données) en fin d'apprentissage. Ce critère idéal mêlant vraisemblance, parcimonie, optimum global et émergence d'un complexe simplicial plongé reste à découvrir. C'est cette piste que je poursuis à court terme avec les travaux de thèse de Maxime Maillot.

5.3.3 Les structures multi-échelles

Les méthodes de l'état de l'art sur le problème d'inférence topologique utilisent pour la plupart le critère de persistance topologique pour identifier les structures saillantes dans les données [38, 106, 28]. L'idée est de définir une filtration du complexe simplicial, c'est-à-dire un ensemble de complexes simpliciaux emboîtés décrivant la structure du nuage de points à différentes "échelles" d'observation. Les invariants topologiques qui persistent le plus longtemps dans cette filtration sont jugés les plus saillants. Par exemple, considérons un nuage de points au voisinage d'un cercle. Si l'on centre un disque de rayon r sur chaque point, et que l'on fait croître ce rayon, on agrège progressivement les points les uns aux autres, formant un nombre décroissant de composantes connexes dont la taille s'accroît. A partir d'un certain rayon r , la structure circulaire apparaît, puis elle finit par disparaître au-delà d'un rayon $R > r$ lorsque plus aucun trou ne subsiste au milieu du nuage de points. L'écart $R - r$ indique la persistance de cette structure en anneau, qui apparaît ainsi la plus saillante. Si l'on considère un tore autour duquel s'enroule une hélice échantillonnée uniformément, alors suivant l'échelle d'observation, on peut considérer comme saillante la structure en hélice ou bien celle du tore [53]. Le diagramme de persistance permet de montrer cette structure multi-échelles. Comment intégrer cette approche de persistance à un cadre statistique ?

Les travaux se multiplient autour de cette problématique. Une approche de Bubenik [23] propose de donner une masse probabiliste à chaque donnée et d'en dériver des nombres de Betti probabilistes. Une récente approche de Chazal [28] s'appuie sur un modèle en lien avec la méthode MeanShift [44] de classification automatique. Les approches bayésiennes qui fournissent non plus un modèle unique à son maximum de vraisemblance, mais une distribution de modèles,

²². Quoique les formes additives du modèle de mélange et du critère de vraisemblance permettent d'envisager une implémentation distribuée des calculs.

pourraient être considérées comme un point de départ. En effet, les modes de la distribution des modèles a posteriori pourraient correspondre aux structures saillantes recherchées. Tennenbaum [66] propose déjà une approche Bayésienne pour l'apprentissage de graphes orientés ou d'arbres à partir d'une matrice de similarité entre les points, mais il ne traite pas le cas des complexes simpliciaux et des nombres de Betti. La piste Bayésienne me paraît intéressante à explorer à moyen terme.

5.3.4 Gérer la complexité algorithmique

La multiplication des capteurs disséminés dans l'environnement ou les masses d'informations échangées chaque seconde sur internet génèrent des flux continus de données que les méthodes que nous avons présentées ne permettent pas de traiter en temps réel.

Les modèles de mélanges additifs peuvent facilement être parallélisés, ainsi que les méthodes de cubature par simulation Monte-Carlo pour évaluer les simplexes Gaussiens de notre modèle. Mais l'évaluation des fonctions exponentielle (loi normale) et logarithme (log-vraisemblance) reste coûteuse. Les calculateurs graphiques (GPU) sont de plus en plus répandus et offrent une alternative nouvelle pour la parallélisation. Ils permettent des calculs d'algèbre linéaire, et de recherche d'extrema parallélisés très rapides car ils sont spécifiquement dédiés à cela. Une implémentation des modèles génératifs sur GPU est un axe de recherche prometteur à court et moyen terme. En particulier avec ce que j'ai nommé la divergence de Hausdorff [11], mesure d'écart entre deux fonctions densité de probabilité p et q que j'ai proposée. J'ai démontré dans le cas univarié, que minimiser la distance moyenne de Hausdorff entre deux échantillons de densité de probabilité p et q correspond asymptotiquement à minimiser la divergence de Hausdorff entre ces deux densités.

Si l'on considère que les données sont l'un des échantillons et que l'autre échantillon est généré par le modèle génératif, alors on peut en trouvant les paramètres du modèle qui minimisent la distance moyenne de Hausdorff entre les deux échantillons obtenir un modèle génératif asymptotiquement "vraisemblable" des données du point de vue de la divergence de Hausdorff. Or la distance moyenne de Hausdorff est implémentable simplement sur GPU, elle pourrait donc être l'une des clefs du passage sur GPU des modèles de mélange gaussien et de l'estimation de leurs paramètres : on remplacerait la représentation analytique de la densité gaussienne et le calcul de la vraisemblance des données pour cette loi, par un échantillon de cette loi et le calcul de la distance moyenne de Hausdorff entre les

données et cet échantillon. Des travaux récents présentent d'autres méthodes de comparaison de distributions sous forme de nuages de points [64] qui pourraient aussi être utilisées dans ce cadre.

5.3.5 Complexité et topologie

Dans sa thèse [45], Pierre Gaillard a proposé d'utiliser le critère BIC pour sélectionner le modèle dont la topologie est censée être la plus proche de celle de la population génératrice des données. Cela nous a paru un critère pertinent mais aucune justification théorique ne valide complètement ce choix.

Le critère BIC est un critère déterminant un compromis particulier entre une mesure d'adéquation aux données à son optimum (la log-vraisemblance à son maximum) et un facteur de pénalisation (le demi produit du nombre de paramètres et du logarithme du nombre de données) encourageant la parcimonie suivant le principe du rasoir d'Occam. Cependant la topologie du modèle n'intervient pas directement dans la formulation de ce critère. Comment intégrer la topologie au critère de sélection de modèle ?

En suivant toujours le principe d'Occam, nous pouvons postuler qu'un bon modèle "topologique" pour une même adéquation optimale aux données, est un modèle dont la topologie est la plus "simple" possible. On doit alors aussi assurer l'identifiabilité des paramètres du modèles par l'unicité de la solution : à une même adéquation optimale aux données et une même "simplicité" topologique doit correspondre un unique jeu de paramètres. Sans être encore en mesure de formaliser complètement ces propositions, le modèle de complexe simplicial génératif semble en première analyse fournir une solution élégante vérifiant ces deux propositions. En effet, un segment gaussien a le "pouvoir explicatif" d'un mélange d'une infinité de gaussiennes uniformément réparties le long du segment, donc à vraisemblance égale, la complexité d'un mélange de K gaussiennes réparties uniformément sur un segment $(K * D + 1)$ est toujours plus grande que celle d'un segment gaussien $(2D + 1)$. Il est donc toujours plus avantageux du point de vue de la complexité et donc d'un critère parcimonieux "à la" BIC, d'expliquer localement des données issues d'une variété linéique uniforme perturbée par un bruit gaussien, avec un segment gaussien (même topologie) qu'avec un mélange de K points gaussiens (K variétés de dimension 0). De même, expliquer ces données avec un d -simplexe gaussien ($d > 1$) dont les sommets seraient alignés (donc une variété modèle dont la dimension est trop grande par rapport à celle de la variété à retrouver) augmenterait la complexité du modèle $((d + 1)D + 1)$ sans améliorer

sa vraisemblance, donc dégraderait aussi la parcimonie. Le critère de parcimonie appliqué à un complexe simplicial génératif semble donc fournir un modèle topologique optimal au sens de l'identifiabilité. L'idée d'utiliser le critère BIC pour élaguer le GGG est donc confortée par cette analyse. Nous cherchons à vérifier, approfondir et formaliser cette proposition dans la thèse de Maxime Maillot.

Le GGG peut être vu comme une généralisation des modèles de mélange classiques à des points et des segments : un mélange de gaussiennes est un GGG sans arêtes. La variance des composants du modèle est volontairement choisie comme un scalaire identique pour tous. En effet, la complexité topologique de la population génératrice ne peut être capturée par ce seul paramètre de variance, et se voit donc expliquée par l'ensemble des autres paramètres liés à la structure topologique du complexe simplicial. L'information topologique (espaces tangents et dimension locaux donnés par les simplexes, connexités) peut être extraite du modèle car il est assez flexible pour la capturer et la restituer. A l'inverse, si nous utilisons une matrice de covariance pleine et indépendante pour la loi normale convoluée à chaque composant du modèle génératif, l'information topologique serait en grande partie capturée par cet ensemble de paramètres de variance mais ne pourrait être que partiellement restituée : les matrices de covariance indiqueraient les directions principales et la dimension significative des sous-espaces tangents locaux de chaque composant, mais ne restitueraient pas la connexité simple entre les composants ni les connexités d'ordre supérieur (nombres de Betti). Les autres modèles génératifs ne fournissent aucun indice sur cette connexité, excepté le Generative Topographic Mapping [22] et les courbes principales probabilistes [112], mais dans ces deux cas, la connexité du modèle est contrainte *a priori* et non apprise des données.

Du point de vue modèle de densité, le graphe génératif et le complexe simplicial génératif que j'envisage, en faisant abstraction du bruit, sont uniformes par morceaux. La fonction densité de probabilité qu'ils génèrent est donc constante par morceaux. Compte tenu de la structure de complexe simplicial, on peut envisager de définir la fonction densité de probabilité globale comme l'interpolation linéaire des densités associées aux sommets de chaque simplexe. Cela aurait pour effet de réduire le nombre de paramètres, puisque seuls les sommets seraient pondérés, et une pondération binaire suffirait à déterminer la présence ou l'absence d'une facette de dimension 1 ou plus du complexe simplicial. Comment réaliser l'estimation des paramètres et l'élagage de ce modèle ? Comment l'élaguer pour en extraire les invariants topologiques ou définir une filtration pour cal-

culer la persistance topologique ?

On pourrait aussi réduire le nombre de paramètres du modèle génératif en considérant que la présence d'un simplexe signifie que seule subsiste la pondération de sa facette principale, *i.e.* il ne pourrait y avoir dans le modèle de simplexes "creux". Là encore, quid de l'apprentissage et de l'extraction des invariants topologiques d'un tel modèle ?

Enfin, dans un cadre non supervisé, il est intéressant de détecter automatiquement les sous-groupes indépendants de variables dépendantes, et pour chacun d'utiliser un complexe simplicial pour modéliser la population sous-jacente au nuage de points dans ce sous-espace. L'approche des complexes simpliciaux génératifs détecte automatiquement les variables dépendantes en sculptant dans l'espace joint le complexe initial pour en extraire un sous-complexe dont la dimension locale correspond au nombre de variables latentes indépendantes nécessaire et suffisant pour décrire localement le nuage de points. Cependant elle ne permet pas de détecter automatiquement les sous-groupes indépendants de variables dépendantes. Ainsi si V_1 et V_2 sont deux variables liées (*e.g.* $V_1 = \sin(V_2)$) ainsi que V_3 et V_4 (*e.g.* $V_3 = \cos(V_4)$), mais qu'aucune relation de dépendance n'existe entre V_1 et V_3 , V_1 et V_4 , V_2 et V_3 et V_2 et V_4 , alors le nuage de points dans l'espace joint (V_1, V_2, V_3, V_4) est situé sur une surface à deux dimensions et sera modélisé par un complexe simplicial constitué de triangles. Un modèle plus parcimonieux consisterait à modéliser la variété linéique dans l'espace (V_1, V_2) et l'autre variété linéique dans l'espace (V_3, V_4) , ce qui utiliserait de l'ordre de $M_{12} + M_{34}$ paramètres au lieu de $M_{12} \times M_{34}$ paramètres, où M_{12} et M_{34} sont les nombres de paramètres nécessaires au modèle dans les sous-espace (V_1, V_2) et (V_3, V_4) respectivement. La topologie de la variété dans l'espace joint s'obtiendrait comme le produit cartésien des variétés linéiques marginales, on ne perdrait donc pas d'information topologique. Comment doter les modèles de complexes simpliciaux génératifs de cette propriété de détection automatique de sous-espaces indépendants ?

5.3.6 Modèles interprétables

Les modèles génératifs topologiques sont construits directement dans l'espace des données. Les nombres de Betti peuvent alors être utilisés comme attributs supplémentaires pour les systèmes de décision aval. Cependant ce gain en précision obtenu en évitant les distorsions que produiraient une réduction de dimension préalable, se fait au détriment de l'interprétabilité. En effet, il est difficile d'inter-

prêter le vecteur de paramètres autrement que par le résumé fournit par les nombre de Betti ou les diagrammes de persistance topologique. Je propose donc d'explorer de nouvelles pistes pour rendre interprétables ces modèles.

Les systèmes d'inférences flous (SIF) sont des systèmes de décision automatique basés sur un ensemble de règles floues. Ils servent habituellement à encoder dans les paramètres d'un modèle une expertise humaine décrite par des termes linguistiques plutôt que numériques. La thèse de Laurence Cornez dont une publication résume l'essentiel [30] a montré comment une règle de décision floue peut être implémentée sous la forme d'un composant d'un modèle de mélange gaussien et un SIF sous forme de la règle de décision du Maximum A Posteriori appliquée au modèle de mélange. Dans ce cas, les règles flous ne sont plus dictées par un expert mais extraites automatiquement des données, et sous une forme qui reste interprétable par l'expert. Puisque les modèles génératifs basés sur les graphes et les complexes simpliciaux sont des modèles de mélange, on peut se demander comment interpréter ces modèles en tant que SIF.

Une autre extension des graphes et complexes simpliciaux génératifs concerne leur visualisation sous la forme synthétique du graphe des classes. Le graphe des classes porte seulement l'information de connexité simple en indiquant quelles classes sont voisines de quelles autres. Pourrait-on l'étendre pour synthétiser l'information portée par les nombres de Betti ? Par exemple, comment indiquer que deux 1-cycles (boucles) sont entrelacés, ou qu'une composante est à l'intérieur du 3-cycle (cavité) d'une autre ?

On pourrait aussi envisager d'utiliser les méthodes de visualisation in situ dans les graphes des classes pour compléter la synthèse topologique par une information géométrique sur les distances relatives entre composantes connexes.

5.3.7 Autres types de données

Nous avons traité des données de type nuages de points, mais les données peuvent prendre de nombreuses autres formes, comme par exemple les graphes issus de modèles de molécules, les arbres phylogénétiques issus de données génomiques, les hypergraphes issus de modèles sociologiques, les signaux multivariés issus de réseaux de capteurs, les graphes ou multigraphes orientés dont sommets et arêtes sont munis d'attributs vectoriels issus de la modélisation de réseaux sociaux. Si l'on considère des données de tels types, les mêmes questions topologiques font sens : quelles sont les données voisines ? Existe-t-il des structures topologique particulières ? Ont-

elles un sens pour l'analyste ? Peuvent-elles permettre de distinguer différents ensembles de données ?

On sait par exemple construire les cellules de Voronoï d'un sous-ensemble de sommets prototypes d'un graphe connexe non orienté : il suffit de déterminer pour chaque sommet prototype, l'ensemble des sommets dont il est le plus proche au sens par exemple de la longueur en nombre d'arêtes du chemin les reliant. Il devient alors possible de définir le complexe simplicial de Delaunay structure topologique duale du complexe cellulaire de Voronoï. Tous les outils d'analyse des complexes simpliciaux sont alors utilisables. On pourrait alors s'intéresser à la définition d'un modèle génératif dans ce cadre particulier qui permettrait d'élaguer ou de pondérer les simplexes de ce complexe de Delaunay et d'analyser plus finement la topologie sous-jacente au graphe initial.

5.3.8 Autres applications

Nous avons expérimenté le GGG dans différents cadres. En classification non supervisée et semi-supervisée ainsi que pour l'analyse exploratoire de données étiquetées. D'autres champs d'applications sont possibles.

Le graphe pondéré obtenu fournit un moyen de calculer une approximation de la distance géodésique entre deux points de la variété sous-jacente au nuage de points, une arête de poids nul signifiant l'absence de points pour supporter un chemin, donc une distance infinie. On peut alors calculer la matrice des distances géodésiques entre les points du nuage et l'utiliser pour projeter ce nuage dans le plan comme dans ISOTOP [75] où le Topology Representing Network est utilisé. La structure de graphe permet aussi de naviguer dans l'espace sous-jacent au nuage de points : si ces points représentent par exemple une base d'images, alors on peut se déplacer d'images en images dont les contenus sont proches, avec un voisinage dépendant du degré des sommets du graphe. Remplaçons les images par n'importe quel type de données pouvant se mettre sous une forme vectorielle, et nous disposons d'un moyen de naviguer de proche en proche dans une base de données.

Les caractéristiques topologiques (e.g. nombres de Betti) peuvent compléter les attributs numériques des données dans le cadre d'une tâche de discrimination. Nous avons fait des tests préliminaires avec des caractères manuscrits dans la thèse de Pierre Gaillard. Nous explorerons davantage cette voie dans la thèse de Maxime Maillot avec des données de types signaux d'Electro-Encéphalogrammes.

Le complexe simplicial génératif que l'on étudie dans la thèse de

Maxime Maillot permettrait de projeter l'ensemble des données étiquetées sur le complexe simplicial (débruitage) et de réaliser une discrimination de ces données directement dans l'espace généré par ce complexe. Ainsi on opèrerait une réduction de dimension locale adaptée à la dimension effective des données plutôt qu'une réduction de dimension globale de l'espace de dimension D vers un espace de dimension Q trop réductrice pour les régions de dimension intrinsèque supérieure à Q , et pas assez réductrice pour celles de dimension intrinsèque inférieure à Q .

Enfin l'apprentissage automatique de règles floues spatiales déjà abordé, paraît une autre voie prometteuse pour les applications à l'interface avec l'homme. L'extension à l'apprentissage d'ontologies paraît aussi possible, les sommets du graphe ou du complexe simplicial (hyper-graphe particulier) représentant des concepts prototypes reliés entre eux par des hyper-arêtes (ensembles de sommets) représentant l'apparition concomitante de ces concepts. L'apparition d'une nouvelle donnée activerait le concept le plus proche qui à son tour activerait les concepts voisins dans le graphe et ainsi de suite, à l'image des enchaînements d'idées que chacun peut observer par introspection. Cette voie plus hypothétique formerait la jonction entre les données numériques fournies par les capteurs et les concepts symboliques sur lesquels s'appliquent l'inférence et les raisonnements et qui sont plus interprétables par l'homme, une réponse possible à l'un des problèmes fondamentaux de l'Intelligence Artificielle de l'ancrage des symboles dans les perceptions²³ (symbol grounding problem [55][96][36]). Ce dernier point nourrit les perspectives de travaux que j'envisage à plus long terme sur les relations entre hommes et machines et sur la Conscience Artificielle.

5.4 Futurs possibles

5.4.1 Des données à la connaissance

Un objectif à long terme de mes recherches est de rendre interprétables aux décideurs les masses de données accumulées par les entreprises dans l'espoir qu'une pépite d'information stratégique s'y trouve cachée.

La visualisation est le médium le plus approprié pour assister l'analyste humain, mais encore faut-il visualiser la bonne information (authenticité), uniquement elle (expressivité) avec toute la pré-

23. Comment les symboles ou concepts par nature discrets et le raisonnement symbolique (enchaînement de relations entre entités discrètes) émergent-ils dans notre cerveau à partir des nos perceptions qui par nature sont des signaux continus? Comment les symboles acquièrent-ils un sens?

cision requise (vérité) en utilisant les variables graphiques les plus simplement interprétables sans artefact visuel ni surcharge cognitive (efficacité et fiabilité). Au-delà de la simple représentation graphique, il faut engager l'analyste dans une interaction avec cette représentation, lui permettre d'expérimenter, de manipuler les objets représentés et leur représentation pour les appréhender complètement. La thèse de Nicolas Heulot est un premier pas dans cette direction. Les invariants topologiques sont au premier rang des informations à visualiser car ils sont les plus susceptibles de résister aux distorsions imposées par la chaîne de mesure et de visualisation.

Face à la taille gigantesque des bases de données ou de l'espace des paramètres des modèles que les analystes devront explorer, traiter et synthétiser, ils n'auront d'autre solution que d'être accompagnés d'une multitude d'agents intelligents et autonomes parcourant en parallèle ces bases à la recherche d'indices spécifiques, l'analyste ne visualisera pas en premier lieu les données elles-mêmes mais l'état des milliers d'agents scrutant ces données, il pourra visuellement prendre la mesure de l'avancement de l'exploration et interroger les agents signalant une pépite, il pourra les coordonner, focaliser une partie d'entre eux dans une région de la base de donnée ou de l'espace des paramètres à explorer. Les agents eux-mêmes apprendront de l'analyste ce qu'ils doivent rechercher, soit par un paramétrage explicite de celui-ci soit par observation et apprentissage de ses comportements.

Le défi est bien sûr de rendre utilisable ce type d'outils par le grand public, qu'il s'agisse d'acheter une pizza ou un véhicule, de rechercher un circuit touristique, ou de piloter sa maison intelligente. Les agents virtuels d'aujourd'hui en état de veille permanente ou recherchant automatiquement des informations sur internet n'en sont encore qu'une préfiguration sommaire essentiellement focalisés sur des indices textuels fournis par les utilisateurs humains. Les interfaces graphiques seront primordiales pour mettre en relation ces agents et leurs maîtres, que les uns et les autres se comprennent, que les hommes puissent graphiquement et dynamiquement (gestuelle) autant que textuellement spécifier explicitement ou rendre tangible implicitement ce qu'ils recherchent, et que les agents puissent rendre compte efficacement de leurs missions et conseiller de nouvelles pistes, cela implique tout autant de représenter la trajectoire des agents et les points saillants de celle-ci, mais aussi le paysage de recherche et son état d'exploration dans lequel ces trajectoires s'inscrivent. La topologie et la visualisation *in situ* y auront une place de choix puisqu'ils sont des moyens élémentaires de préserver l'information primordiale dans les représentations graphiques en

générale. On peut imaginer qu'existera une grammaire des compositions graphiques interprétables issues des recherches expérimentales sur les capacités visuelles et cognitives de l'être humain. Pour le moment il s'agit d'un ensemble de règles empiriques élémentaires mais il reste à réaliser un travail d'organisation et de formalisation de ces connaissances afin d'automatiser la conception de représentations graphiques systématiquement efficaces. Les plus gros efforts devront aussi se porter sur l'étude et la formalisation des aspects dynamiques et interactifs que devront mettre en oeuvre ces représentations, ainsi que prendre en compte les facultés de mémorisation et de focalisation que l'homme possède naturellement, pour *in fine* accroître la surface de contact de l'interface afin qu'elle devienne transparente et que la communication directe entre la machine et l'homme ait lieu sans effort pour ce dernier.

Au-delà des bases de données ou des espaces de paramètres d'un modèle, d'autres objets devront pouvoir être explorés. La topologie est au coeur des réseaux : réseaux sociaux, réseaux de transport, réseaux énergétiques, réseaux sémantiques (ontologies), réseaux du système nerveux, réseaux biologiques, réseaux de nano-machines... L'analyse et le pilotage de ces réseaux devra se faire par des outils de navigation et de visualisation adaptés, capable d'ingérer leur complexité et de ne montrer que l'essentiel en accord avec les objectifs explicites ou implicites de l'analyste. La visualisation des réseaux est pour l'essentiel réalisée par des diagrammes noeuds-liens. Il faut dépasser les limites de ce paradigme de représentation, proposer de nouveaux modèles focalisant l'attention sur les variables d'intérêt : à quoi sert-il de montrer l'ensemble des milliers, millions ou milliards de liens d'un réseau ? Les artefacts de ces représentations sont tels qu'une vue d'ensemble ne peut apporter aucune information authentique donc exploitable. Les modèles génératifs topologiques peuvent être un moyen de simplifier ces représentations et de guider la navigation. Le paradigme WinSitu est une autre approche que l'on peut envisager d'utiliser pour proposer de nouvelles représentations graphiques interprétables de ces réseaux.

5.4.2 Des machines conscientes

Dans un futur proche, les objets communicants qui envahissent déjà notre quotidien, mais aussi les bâtiments, ou les véhicules deviendront de plus en plus intelligents et autonomes. Ils seront probablement dotés d'une conscience d'eux-même et d'états émotionnels ²⁴

24. La Conscience Artificielle est un champ de recherche en plein essor. Le site <http://www.conscious-robots.com> répertorie les initiatives de recherches dans ce domaine.

constituant ainsi des êtres artificiels avec lesquels l'homme sera en relation plus qu'en interaction, comme il l'est avec des êtres vivants²⁵. A notre service, ces machines observeront, agiront et penseront de manière plus ou moins autonome pour nous assister ou pour nous divertir.

Outre les questions philosophiques et éthiques que cela soulève, il reste bien difficile de prévoir a priori la réaction des hommes face à la diffusion en masse de telles machines dans leur environnement quotidien. Nous pouvons cependant imaginer des barrières qui limiteront l'intégration des machines intelligentes à la communauté humaine si elles ne sont pas surmontées. Une barrière légale exigera très probablement que ces machines aussi autonomes qu'elles puissent être, demeurent sous le contrôle de leurs propriétaires humains *a priori* seuls responsables des conséquences de leurs actions. Mais le champ d'action de ces machines versatiles sera beaucoup plus large que celui d'une cafetière programmable, ainsi leurs actions dans le monde physique seront sources de multiples dangers potentiels que l'utilisateur voudra maîtriser ou éviter. Alors les utilisateurs demanderont que ces machines soient sous leur contrôle total comme l'est leur machine à laver ou leur voiture. Il ne suffira pas pour cela de les équiper d'un bouton marche-arrêt mais il faudra fournir à l'utilisateur les moyens de ce contrôle. De plus, leur valeur ajoutée résidera dans les comportements complexes qu'elles auront acquis par leurs capacités d'observation et d'apprentissage, si bien que les capteurs sensoriels nécessaires à l'acquisition de ces comportements les rendront témoins de scènes de vie que l'utilisateur voudra pouvoir contrôler ou supprimer de leur mémoire, sans parler de la mise en réseau possible de ces informations²⁶. Enfin, il est aussi probable qu'apparaissent chez ces machines des pathologies liées à la conscience de soi et la présence d'états émotifs, que l'on observe d'ordinaire chez les humains (dépression, psychoses, névroses...), et qu'il faudra savoir traiter.

Il sera donc nécessaire de concevoir le système de contrôle de ces

La conscience de soi permet notamment de distinguer dans ce que nous percevons, ce qui dépend de soi de ce qui n'en dépend pas, afin d'optimiser nos prises de décisions. Par ailleurs, le dernier ouvrage d'Antonio Damasio [33] décrit avec précision comment la conscience de soi (de son corps, de ses sensations et actions et de son histoire) émerge dans le cerveau humain rendant ainsi envisageable l'implémentation de tels processus dans une machine. Des projets sont en cours pour réaliser de tels artefacts [25, 5]. Le livre de Claude Touzet [115] décrit une approche neuronale supposée mener à un tel artefact.

25. Dominique Sciamma, créateur et directeur du département des systèmes et produits interactifs au Strate College de Sèvres [105], et ses étudiants, imaginent et étudient de tels objets-robots qu'ils nomment "Robjets".

26. Le site <http://www.roboearth.org/> propose déjà un réseau internet dédié à la mise en commun des connaissances acquises par les robots, que les robots pourront consulter pour accroître leurs connaissances.

machines de façon à pouvoir intégrer *ab initio* des connaissances factuelles et procédurales élémentaires comme l'identité de leur propriétaire, la capacité à communiquer ou à se déplacer, ou des principes moraux, mais aussi de permettre à tout moment une intervention par les personnes autorisées, pour analyser, vérifier, supprimer ou corriger les connaissances initiales ou acquises empiriquement durant leur fonctionnement quotidien. Des interfaces de visualisation et d'intervention spécifiques devront être conçues conjointement aux systèmes cognitifs artificiels de ces machines pour conserver la maîtrise de leur fonctionnement présent et futur, et pouvoir assurer et rassurer leur propriétaire de ce qu'elles savent et savent faire. Il est probable que ces systèmes ne dépendront pas uniquement d'une base de règles et d'une base de faits, par nature symboliques, mais reposeront aussi sur le traitement distribué et parallélisé du flux ininterrompu de signaux issus des nombreux capteurs dont ils seront dotés pour percevoir leur environnement et leur propre corps, à l'instar du cerveau des êtres vivants dont ils seront inspirés.

Pour autant, ce qui fait sens pour la machine devra toujours faire sens pour son propriétaire. Alors il faudra pouvoir naviguer dans les souvenirs factuels ou procéduraux accumulés par la machine, comprendre le rôle des inhibitions ou excitations progressivement ajoutées par le système aux règles initiales, visualiser et analyser la topologie des liens tissés entre différentes connaissances, comprendre le sens assigné par la machine à certains processus ou motifs d'activation de ses composants cognitifs, détecter la présence d'un virus, ou encore supprimer un processus anormal ou dangereux. Ainsi ces machines intelligentes voire conscientes poseront inévitablement des problèmes qui chez l'homme relèveraient de l'intervention d'un psychologue. Il faudra donc les concevoir de telle sorte que ces interventions sur leur système cognitif puissent être aussi précises et efficaces que l'on peut l'attendre des interventions d'un psychologue, en d'autres termes, il nous faudra pouvoir lire dans leurs pensées²⁷.

Une première étape consistera à concevoir une machine élémentaire consciente d'elle-même et dont les états mentaux demeureront interprétables et modifiables par l'observateur extérieur. Puis à partir d'une telle base, nous pourrons envisager de développer les systèmes de contrôle de ces "robjets" ainsi que les outils logiciels associés de diagnostic et d'intervention dont auront besoin les futurs "robopsychologues".

Les méthodes et modèles présentés dans cette synthèse pourraient

27. Isaac Asimov dans sa nouvelle "Menteur!" parue en 1941 avait imaginé le personnage de Susan Calvin dont la profession de "robopsychologue" consiste à intervenir sur le système cognitif artificiel des robots afin de les comprendre et de les réparer.

contribuer à ce que cette possibilité devienne réalité. En effet, je suis persuadé que la topologie est au coeur du substrat fondamental à l'émergence de nos processus cognitifs humains, et que la structure de ce substrat est acquise par apprentissage. Il me paraît donc pertinent d'utiliser ce même substrat et son processus de génération automatique dans les machines afin d'une part de les rendre aussi intelligentes que nous jugeons l'être, et d'autre part de nous permettre de les comprendre et d'en garder la maîtrise, et *in fine* de mieux nous comprendre. Aussi mon programme de recherche s'attache-t-il à explorer cette hypothèse et à en exploiter les résultats.

Références

- [1] M. Aanjaneya, F. Chazal, D. Chen, M. Glisse, L. Guibas, and D. Morozov. Metric graph reconstruction from noisy data. *Proceedings of the Annual Symposium on Computational Geometry*, pages 37–46, 2011.
- [2] E. Agrell. A method for examining vector quantizer structures. *Proceedings of IEEE International Symposium on Information Theory, San Antonio, TX*, page 394, 1993.
- [3] A. Ahalt, D.M. Krishnamurthy, and P. Chen. Competitive learning algorithms for vector quantization. *Neural Networks*, 3, 1990.
- [4] J. M. Alonso, L. Magdalena, and G. González-Rodríguez. Looking for a good fuzzy system interpretability index : An experimental approach. *International Journal of Approximate Reasoning*, 51(1) :115–134, December 2009.
- [5] R. Arrabales Moreno. Conscious robots. <http://www.conscious-robots.com>.
- [6] D. Asimov. The grand tour : A tool for viewing multidimensional data. *SIAM J. Scientific and Statistical Computing*, 6 :128–143, 1985.
- [7] I. Asimov. *I, Robot*. Doubleday, Garden City, N.Y., 1963.
- [8] M. Aupetit. Robust topology representing networks. *European Symp. on Artificial Neural Networks, Bruges (Belgium), d-side eds.*, pages 45–50, 2003.
- [9] M. Aupetit. Learning topology with the generative gaussian graph and the em algorithm. *Advances in Neural Information Processing Systems*, 18 :83–90, 2006.

- [10] M. Aupetit. Visualizing distortions and recovering topology in continuous projection techniques. *Neurocomputing*, 70(7-9) :1304–1330, 2007.
- [11] M. Aupetit. Nearly homogeneous multi-partitioning with a deterministic generator. *Neurocomputing*, 72(7-9) :1379–1389, 2009.
- [12] M. Aupetit. Winsitu, un nouveau paradigme pour l’analyse exploratoire de données basée sur des projections. *Revue des Nouvelles Technologies de l’Information*, A.4 Apprentissage et Visualisation(1) :79–98, 2010.
- [13] M. Aupetit. Winsitu : a new information visualization paradigm for visual mining of multidimensional data. *Submitted to Data Mining and Knowledge Discovery special issue on Intelligent Interactive Data Visualization*, 2012.
- [14] M. Aupetit, L. Allano, I. Espagnon, and G. Sannie. Visual analytics to check marine containers in the eritr@c project. In *Proc. of the International Symposium on Visual Analytics Science and Technology (EuroVAST)*, pages 57–60. J. Kohlhammer and D. Keim, Bordeaux, 2010.
- [15] M. Aupetit and T. Catz. High-dimensional labeled data analysis with topology representing graphs. *Neurocomputing, Elsevier*, 63 :139–169, 2005.
- [16] M. Aupetit, F. Chazal, G. Gasso, D. Cohen-Steiner, and P. Gaillard. Nips workshop on topology learning : New challenges at the crossing of machine learning, computational geometry and topology. <http://topolearnnips2007.insa-rouen.fr/index.html>, 2007.
- [17] M. Aupetit and P. Gaillard. Mesurer et visualiser les distortions dans les techniques de projection continues. *Revue Intelligence Artificielle, Visualisation et extraction des connaissances*, 22 :443–472, 2008.
- [18] C. B. Barber, D. P. Dobkin, and H. Huhdanpaa. The quickhull algorithm for convex hulls. *ACM Trans. Math. Software*, 22 :469–483, 1996.
- [19] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization : A geometric framework for learning from examples. *Journal of Machine Learning Research*, 7 :2399–2434, 2006.
- [20] Y. Bengio. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1) :1–127, 2009.

- [21] A. Bezerianos, F. Chevalier, P. Dragicevic, N. Elmqvist, and J.-D. Fekete. Graphdice : A system for exploring multivariate social networks. *Comput. Graph. Forum*, 29(3) :863–872, 2010.
- [22] C. M. Bishop, M. Svensén, and C. K. I. Williams. Gtm : The generative topographic mapping. *Neural Computation*, 10(1) :215–234, 1998.
- [23] P. Bubenik and P. T. Kim. A statistical approach to persistent homology. *Homology, homotopy and Applications*, 9 :337–362, 2007.
- [24] Ghaoui C. *Encyclopedia of Human Computer Interaction*. Idea group reference edition, 2005.
- [25] A. Cardon. Conscience artificielle et systèmes adaptatifs. <http://www.artificial-brain-project.com/>, 1999.
- [26] G. Carlsson. Topology and data. *American Mathematical Society*, 46.2 :255–308, 2009.
- [27] G. Celeux and G. Govaert. A classification em algorithm for clustering and two stochastic versions. *Comput. Stat. Data Anal.*, 14 :315–332, October 1992.
- [28] F. Chazal, L. J. Guibas, S. Y. Oudot, and P. Skraba. Persistence-based clustering in riemannian manifolds. In *Proceedings of the 27th annual ACM symposium on Computational geometry*, SoCG '11, pages 97–106, New York, NY, USA, 2011. ACM.
- [29] E. H. Chi. A taxonomy of visualization techniques using the data state reference model. In *IEEE Symposium on Information Visualization (InfoVis '00)*, page 69–75. IEEE Computer Society Press, 2000.
- [30] L. Cornez, M. Samuelides, and J.-D. Muller. Neuro-fuzzy inference system to learn expert decision : Between performance and intelligibility. In Lipo Wang and Yaochu Jin, editors, *Fuzzy Systems and Knowledge Discovery (FSKD'05)*, volume 3614 of *Lecture Notes in Computer Science*, pages 1281–1293. Springer, 2005.
- [31] P. J. Crutzen. Geology of mankind. *Nature*, 415(6867) :23, 2002.
- [32] G. Cybenko. Approximations by superpositions of sigmoidal functions. *Mathematics of Control, Signals, and Systems*, 2(4) :303–314, 1989.
- [33] A. Damasio. *Self comes to mind*. Pantheon edition, 2010.

- [34] P. Demartines and J. Hérault. Curvilinear component analysis : a self-organising neural network for non-linear mapping of data sets. *IEEE Trans. on Neural Networks*, 8(1) :148–154, 1997.
- [35] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1) :1–38, 1977.
- [36] R. Doursat and J. Petitot. Dynamical systems and cognitive linguistics : toward an active morphodynamical semantics. *Neural Networks*, 18 :628–638, 2005.
- [37] R. Dyer, H. Zhang, and T. Möller. Gabriel meshes and delaunay edge flips. *SIAM/ACM Joint Conference on Geometric and Physical Modeling (SPM '09)*, pages 295–300, 2009.
- [38] H. Edelsbrunner, D. Letscher, and A. Zomorodian. Topological persistence and simplification. In *IEEE Symp. on Found. of Comp. Sci.*, pages 454–463. IEEE Computer Society, 2000.
- [39] H. Edelsbrunner and N.R. Shah. Triangulating Topological Spaces. *International Journal of Computational Geometry and Applications*, 7(4) :365–378, 1997.
- [40] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17(3) :37–54, 1996.
- [41] C. Fraley and A. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97 :611–631, 2002.
- [42] B. Fritzke. Growing cell structures-a self-organizing network in k dimensions. *Artificial Neural Networks*, 2 :1051–1056, 1992.
- [43] B. Fritzke. A growing neural gas network learns topologies. *Advances in Neural Information Processing Systems*, 7, 1995.
- [44] K. Fukunaga and L. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *Information Theory, IEEE Transactions on*, 21(1) :32–40, 1975.
- [45] P. Gaillard. Apprentissage de la connexité d’un nuage de points par modèle génératif. applications à l’analyse exploratoire de données et à la classification semi-supervisée. *Thèse de l’Université de Technologie de Compiègne - Commissariat à l’Energie Atomique*, 2008.
- [46] P. Gaillard, M. Aupetit, and G. Govaert. Learning topology of a labeled data set with the supervised generative gaussian graph. *Neurocomputing*, 71(7-9) :1283–1299, 2008.

- [47] P. Gaillard, M. Aupetit, and G. Govaert. Learning topology of a labeled data set with the supervised generative gaussian graph. *Neurocomputing*, 71(7-9) :1283–1299, 2008.
- [48] P. Gaillard, M. Aupetit, and G. Govaert. Un graphe génératif pour la classification semi-supervisée. *Ingénierie des systèmes d’information*, 15(2) :97–119, 2010.
- [49] S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. *Neural Comput.*, 4 :1–58, January 1992.
- [50] J. Geusebroek, G. Burghouts, and A. Smeulders. The amsterdam library of object images. *International Journal of Computer Vision*, 61(1) :103–112, 2005.
- [51] James J. Gibson. *The Ecological Approach to Visual Perception*. Houghton Mifflin, Boston, 1979.
- [52] R. Gilmore. Topological analysis of chaotic dynamical systems. *Rev. Mod. Phys.*, 70 :1455–1530, 1998.
- [53] L.J. Guibas and S.Y. Oudot. Reconstruction using witness complexes. *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms (SODA ’07)*, pages 1076–1085, 2007.
- [54] F. van Ham and J. J. van Wijk. Interactive visualization of small world graphs. In Matthew O. Ward and Tamara Munzner, editors, *INFOVIS*, pages 199–206. IEEE Computer Society, 2004.
- [55] S. Harnad. The symbol grounding problem. *Physica D*, 42 :335–346, 1990.
- [56] T. Hastie and W. Stuetzle. Principal curves. *Journal of the American Statistical Association*, 84 :502–516, 1989.
- [57] A. Hatcher. *Algebraic Topology*. Cambridge university press edition, 2001.
- [58] S. Haykin. *Neural Networks : A Comprehensive Foundation (2nd ed.)*. Prentice hall edition, 1998.
- [59] N. Heulot, M. Aupetit, and J.-D. Fekete. Evaluation of proxiviz for the visual analysis of multidimensional data. *Submitted to IEEE Information Visualizatin conference*, 2012.
- [60] G.E. Hinton, P. Dayan, and M. Revow. Modeling the manifolds of images of handwritten digits. *Neural Networks, IEEE Transactions on*, 8(1) :65–74, 1997.
- [61] A. Inselberg. The plane with parallel coordinates. *The Visual Computer*, 1(2) :69–91, 1985.

- [62] J.W. Jaromczyk and G.T. Toussaint. Relative neighborhood graphs and their relatives. *Proceedings of the IEEE*, 80(9) :1502–1517, 1992.
- [63] I.T. Jolliffe. *Principal Component Analysis*. Springer Verlag, NY, 1986.
- [64] S. Joshi, R. Varma Kommaraji, J. M. Phillips, and S. Venkatasubramanian. Comparing distributions and shapes using the kernel distance. In *Proceedings of the 27th annual ACM symposium on Computational geometry*, SoCG '11, pages 47–56, New York, NY, USA, 2011. ACM.
- [65] M. Kaufmann and D. Wagner. Drawing graphs - methods and models. *Lecture Notes in Computer Science*, 2025, 2001.
- [66] C. Kemp and J. B. B. Tenenbaum. The discovery of structural form. *Proceedings of the National Academy of Sciences of the United States of America*, 105(31) :10687–10692, 2008.
- [67] J. Kielman and J. Thomas. Special issue : Foundations and frontiers of visual analytics. *Information Visualization*, 8(4) :239–314, 2009.
- [68] T. Kohonen. *Self-Organization and Associative Memory Formation*. Springer Verlag, 1988.
- [69] T. Kohonen. *Self-Organizing Maps*. Berlin, Heidelberg, New York : Springer Series in Information Sciences, 2001.
- [70] T. Kohonen, S. Kaski, K. Lagus, J. Salojärvi, J. Honkela, V. Paatero, and A. Saarela. Organization of a massive document collection. *IEEE Transactions on Neural Networks, Special Issue on Neural Networks for Data Mining and Knowledge Discovery*, 11(3) :574–585, May 2000.
- [71] J.B. Kruskal. Multidimensional scaling : A numerical method. *Psychometrika*, 29 :115–129, 1964.
- [72] R. Kurzweil. The singularity is near : when humans transcend biology. <http://books.google.fr/books?id=88U6hdUi6D0C>, 2005.
- [73] J. Lee, A. Lendasse, and M. Verleysen. Curvilinear distance analysis versus isomap. *Proceedings of the European Symposium on Artificial Neural Networks, Bruges (Belgium)*, pages 185–192, 2002.
- [74] J. A. Lee and M. Verleysen. *Nonlinear Dimensionality Reduction*. Springer edition, 2007.

- [75] J.A. Lee, C. Archambeau, and M. Verleysen. Locally linear embedding versus isotop. *Proceedings of the European Symposium on Artificial Neural Network 2003*, pages 527–534, 2003.
- [76] T. Leonova. L’approche écologique de la cognition social et son impact sur la conception des traits de personnalité. *L’année psychologique*, 104 :249–294, 2004.
- [77] D. Lepetz, M. Némoz-Gaillard, and M. Aupetit. Concerning the differentiability of the energy function in vector quantization algorithms. *Neural Networks*, 20 :621–630, 2007.
- [78] S. Lespinats, Meyer-Bayer A., and M. Aupetit. Classimap : a supervised non-linear mapping which preserves the topology of the classes. *Submitted to IEEE Transaction on Visualization and Computer Graphics*, 2012.
- [79] S. Lespinats and M. Aupetit. False neighbourhoods and tears are the main mapping defaults. how to avoid it? how to exhibit remaining ones? *Proceeding of Quality Issues, Measures of Interestingness and Evaluation of data mining models (QI-MIE’09)*, pages 55–65, 2009.
- [80] S. Lespinats and M. Aupetit. CheckViz : Sanity Check and Topological Clues for Linear and Non-Linear Mappings. *Computer Graphics Forum*, 30(1) :113–125, 2011.
- [81] S. Lespinats, M. Verleysen, A. Giron, and G. Fertil. Dd-hds : A method for visualization and exploration of high-dimensional data. *Neural Networks, IEEE Transactions on*, 18(5) :1265–1279, 2007.
- [82] L.J.P. van der Maaten and G.E. Hinton. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9 :2579–2605, 2008.
- [83] G. Mac Lachlan and D. Peel. *Finite Mixture Models*. New York : John Wiley & Sons, 2000.
- [84] J. Mac Queen. Some methods of classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.
- [85] J. Mackinlay. Automating the design of graphical presentations of relational information. *ACM Transactions on Graphics TOG*, 5(2) :110–141, 1986.
- [86] P. C. Mahalanobis. On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India*, 2(1) :49–55, 1936.

- [87] T. M. Martinetz, S. G. Berkovitch, and K. J. Schulten. “neural-gas” network for vector quantization and its application to time-series prediction. *IEEE Trans. on NN*, 4(4) :558–569, 1993.
- [88] T. M. Martinetz and K. J. Schulten. Topology representing networks. *Neural Networks, Elsevier London*, 7 :507–522, 1994.
- [89] J. Mendez and J. Lorenzo. Computing voronoi adjacencies in high dimensional spaces by using linear programming. *Mathematical Methodologies in Pattern Recognition and Machine Learning*, 30 :507–522, 2013.
- [90] D. Mercier, P. Gaillard, M. Aupetit, C. Maillard, R. Quach, and J.-D. Muller. How to help seismic analysts to verify the french seismic bulletin? *Engineering Applications of Artificial Intelligence*, 19(7) :797–806, 2006.
- [91] G. A. Miller. The magical number seven, plus or minus two : Some limits on our capacity for processing information. *The Psychological Review*, 63(2) :81–97, 1956.
- [92] R. E. Moustafa. Andrews curves. *Wiley Interdisciplinary Reviews : Computational Statistics*, 3(4) :373–382, 2011.
- [93] J. Munkres. *Elements of Algebraic Topology*. Westview Press, 1993.
- [94] D.J. Newman, S. Hettich, C.L. Blake, and C.J. Merz. Uci repository of machine learning databases. *Irvine, CA : Dept. of Information and Computer Science, University of California at Irvine*. <http://www.ics.uci.edu/mlearn/MLRepository.html>, 1998.
- [95] A. Pang, C. Wittenbrink, and S. Lodha. Approaches to uncertainty visualization. *The Visual Computer*, 13(8) :370–390, 1997.
- [96] J. Petitot. Syntaxe topologique et grammaire cognitive. *Languages*, 1991.
- [97] K. Roeder and L. Wasserman. Practical bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association*, 92(439) :894–902, 1997.
- [98] N. le Roux, Y. Bengio, P. Lamblin, M. Joliveau, and B. Kégl. Learning the 2-d topology of images. In John C. Platt, Daphne Koller, Yoram Singer, and Sam T. Roweis, editors, *NIPS*. Curran Associates, Inc., 2007.
- [99] S. Rueping. *Learning Interpretable Models*. University Dortmund edition, 2006.

- [100] J. W. Sammon. A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers - C*, 18(5) :401–409, 1969.
- [101] G. Saporta. *Probabilités, analyse des données et statistique*. Technip, 1990.
- [102] B. Scholkopf, C. J. C. Burges, and A. J. Smola. *Advances in Kernel Methods : Support Vector Learning*. Mit press, cambridge, ma edition, 1999.
- [103] H.-J. Schulz, S. Hadlak, and H. Schumann. The design space of implicit hierarchy visualization : A survey. *IEEE Trans. Vis. Comput. Graph.*, 17(4) :393–411, 2011.
- [104] G. Schwartz. Estimating the dimension of a model. *The Annals of Statistics*, 6 :461–464, 1978.
- [105] D. Sciamma. <http://www.millenaire3.com/>, 2011.
- [106] V. de Silva and G. Carlsson. Topological estimation using witness complexes. In *M. Alexa and S. Rusinkiewicz (Eds) Eurographics Symposium on Point-Based Graphics, ETH, Zürich, Switzerland, June 2-4*, pages 157–166, 2004.
- [107] V. de Silva and P. Perry. Plex : Simplicial complexes in matlab. <http://comptop.stanford.edu/u/programs/plex/>, 2003.
- [108] V. de Silva and J. B. Tenenbaum. Global versus local methods for nonlinear dimensionality reduction. In *S. Becker, S. Thrun, K. Obermayer (Eds) Advances in Neural Information Processing Systems, MIT Press, Cambridge, MA*, 15 :705–712, 2003.
- [109] B.W. Silverman. *Density Estimation for Statistics and Data Analysis*. London : Chapman & Hall/CRC, 1998.
- [110] A. J. Smola and B. Scholkopf. A tutorial on support vector regression. *Stat. Comput.*, 14 :199–222, 2004.
- [111] A. Sparkes, R. D. King, W. Aubrey, M. Benway, E. Byrne, A. Clare, M. Liakata, M. Markham, K. E. Whelan, and M. Young. An integrated laboratory robotic system for autonomous discovery of gene function. *Journal of the Association for Laboratory Automation*, 15(1) :33–40, 2010.
- [112] R. Tibshirani. Principal curves revisited. *Statistics and Computing*, (2) :183–190, 1992.
- [113] M. E. Tipping. Sparse bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.*, 1 :211–244, September 2001.

- [114] W.S. Torgerson. Multidimensional scaling i - theory and methods. *Psychometrika*, 17 :401–419, 1952.
- [115] C. Touzet. Conscience, intelligence, libre-arbitre ? les réponses de la théorie neuronale de la cognition. <http://www.machotte.fr/>, 2010.
- [116] J. Tukey and J. Wilder. *Exploratory Data Analysis*. Addison-Wesley, 1977.
- [117] J. Venna, J. Peltonen, K. Nybo, H. Aidos, and S. Kaski. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *Journal of Machine Learning Research*, 11 :451–490, 2010.
- [118] T. Villmann, R. Der, and T. Martinetz. A new quantitative measure of topology preservation in kohonen’s feature maps. In *International Conference on Neural Networks*, volume 2, pages 645–648, 1994.
- [119] C. Ware. *Information Visualization : Perception for Design (2nd ed.)*. Morgan kaufman edition, 2004.
- [120] J. J. van Wijk and H. van de Wetering. Cushion treemaps : Visualization of hierarchical information. In *INFOVIS*, pages 73–78. IEEE Computer Society, 1999.
- [121] L. Wilkinson. *The Grammar of Graphics (2nd ed.)*. Springer edition, 2005.
- [122] M. Zeller, R. Sharma, and K. Schulten. Topology representing network for sensor-based robot motion planning. *World Congress on Neural Networks, INNS Press*, pages 100–103, 1996.
- [123] D.A. Zighed, S. Lallich, and F. Muhlenbach. A statistical approach to class separability : Research articles. *Appl. Stoch. Model. Bus. Ind.*, 21 :187–197, March 2005.
- [124] A. J. Zomorodian. *Topology for Computing*. Cambridge university press edition, 2005.