

Curriculum Vitæ de Christine Tuleau incluant une description des travaux de recherche et d'enseignement

Outre un descriptif succinct de ma situation personnelle et professionnelle, ce document présente de manière analytique d'une part mes travaux de thèse, et plus généralement de recherche et d'autre part ceux d'enseignement. Il constitue un élément majeur de mon dossier de candidature au poste de maître de conférence.

Table des matières

Curriculum Vitæ	3
Fonctions exercées	4
Thèse	8
Activités d'enseignement	10
Activités de recherche	12
Perspectives	21
Publications	23
Communications orales	24

Curriculum Vitæ

Christine Tuleau (épouse Malot)

Nationalité française

Née le 13 décembre 1977

Mariée, 1 enfant

Adresse universitaire :

Laboratoire Jean-Alexandre Dieudonné

Université Nice-Sophia Antipolis

06100 Nice

Tél. : 04 92 07 62 07

email : malot@unice.fr

Page Web : <http://math.unice.fr/~malot>

Situation actuelle : Maître de conférence

- Déc. 2007 Maître de conférence à l'Université de Nice-Sophia Antipolis
Déc. 2006 Maître de conférence stagiaire à l'Université de Nice-Sophia Antipolis
Déc. 2006 Membre de l'IREM
Sept. 2008 Responsable scientifique de la bibliothèque de Mathématiques du Laboratoire

Qualifiée en 26ème section

Cursus universitaire

- Déc. 2005 Obtention du Doctorat de mathématiques, mention TRÈS HONORABLE
Sept. 2002 DEA de Modélisation stochastique et Statistique, Université Paris-Sud, mention BIEN
Juil. 2001 Agrégée de mathématiques, option probabilités et statistique, rang 111
Sept. 2000 Maîtrise de mathématiques, Université Paris-XI, mention BIEN
Sept. 1999 Licence de mathématiques, Université Paris-XI, mention ASSEZ BIEN
1995/1998 Classes préparatoires scientifiques au lycée Marcelin Berthelot (Val de Marne)

Langues étrangères et Informatique

- Anglais Écrit et parlé [5 ans d'étude et séjours linguistiques].
Italien Écrit et parlé [10 ans d'étude et séjours linguistiques].
Allemand Notions [7 ans d'étude et séjours linguistiques].
Informatique Logiciels mathématiques : Matlab, Maple, Mathematica, R, SAS.
 Logiciels de présentation : L^AT_EX, Word, Excel, PowerPoint.
 Environnements : Windows, Linux.

Fonctions exercées

- Maître de conférence et à l'université de Nice-Sophia Antipolis.

2007/2008 **T.D. de Probabilité en L2 Economie et Gestion**, associés au cours magistral de Christian Séguret.

Cet enseignement fait suite à celui suivi par les étudiants lors de leur première année. Nous abordons alors directement les variables aléatoires puis les couples de variables aléatoires continues. On s'intéresse aussi tout particulièrement à la loi Normale standard qui sera très utilisée dans la suite du cours.

Cours et TD relatifs aux Méthodes CART en M2 MASS.

L'objectif du cours est de donner aux étudiants une méthode qui leur permet de construire, rapidement et facilement, des estimateurs tant dans un contexte de régression que de classification. Cette méthode, qui repose sur la construction d'arbres binaires, a le mérite de fournir des résultats facilement interprétables et exploitables. En ce qui concerne les TD, le but est d'illustrer les différentes notions introduites dans le cours, et surtout de s'attarder davantage sur l'aspect pratique, à savoir le côté interprétation et utilisation des sorties. Pour cela, nous avons recours au logiciel R.

Atelier en M2 MASS.

Cet atelier a pour objectif de donner aux étudiants une certaine autonomie de façon à les préparer à ce qui pourrait être leur tâche par la suite. A partir de données réelles, ils doivent mener, par groupes de 3 ou 4, une étude complète de ces données en se plaçant en qualité de personnes mandatées par une institution par exemple. Ils doivent ainsi mettre en application différentes méthodes de statistique, tant descriptive qu'inférentielle, et surtout réaliser une interprétation des résultats obtenus.

Cours de statistique en L2 MI.

En 6 semaines de temps, les étudiants abordent de manière théorique d'une part la statistique descriptive et la simulation et d'autre part la statistique inférentielle en survolant la partie estimation et tests. L'objectif étant de leur faire entrevoir les différents aspects de la statistique.

- Maître de conférence et A.T.E.R à l'université de Nice-Sophia Antipolis (192 heures annuelles).

2006/2007 **T.D. de Probabilité en L2 Economie et Gestion**, associés au cours magistral de Christian Séguret.

Les T.D. permettent aux étudiants d'illustrer les notions abordées en cours. D'une part, il s'agit de revenir sur les calculs de probabilités, puis d'introduire la notion de variables aléatoires, tant dans le contexte discret et continu.

Cours et TD relatifs aux Méthodes CART en M2 MASS.

Le programme est identique à celui précédemment exposé.

Atelier en M2 MASS.

Même chose que précédemment.

TP de statistique en L1 MASS.

A l'aide du logiciel Excel, les différentes notions abordées de manière théorique dans le cours sont visualisées, à l'image de la loi des Grands Nombres ou du Théorème Central Limit entre autre. La simulation de variables aléatoires est mise en œuvre, de même que la statistique descriptive est explorée.

Cours et TP de statistique en L2 MI.

Le programme est identique à celui présenté précédemment simplement, j'ai également participé aux TP associés qui avaient pour vertu d'aborder l'aspect pratique car ils étaient réalisés sur machine.

- A.T.E.R à l'université de Paris X - Nanterre (192 heures annuelles).

2005/2006 **T.D. d'Inférence Statistique en L3 MMIA**, associés au cours magistral de Christian Léonard.

Les T.D. permettent aux étudiants d'illustrer les notions abordées en cours. D'une part, il s'agit de mettre en œuvre les techniques d'estimation (méthodes des moments et du maximum de vraisemblance) et d'évaluer la qualité des estimateurs obtenus (biais, convergence, consistance, intervalle de confiance...). D'autre part un accent est mis sur l'application de la théorie des tests (identification des hypothèses, règle de décision, erreurs, puissance, tests sur les moyennes, variances et proportions).

J'ai construit les différentes feuilles de T.D. en prenant conseil auprès d'autres professeurs et en tenant compte des demandes et attentes de Christian Léonard.

T.P. d'Analyse de données sous R en M1 ISIFAR, associés au cours magistral de Karine Triboulet.

L'objectif de ces T.P. est de permettre aux étudiants d'être "à l'aise" avec le logiciel R, en maîtrisant notamment les fonctions statistiques de base de R. Plus précisément, on souhaite que les étudiants soient capables d'utiliser ce logiciel afin de traiter et analyser statistiquement des données, autrement dit être capables de commenter et exploiter les sorties de commande.

Les feuilles de T.P. ont été élaborées en collaboration avec Mathilde Mougeot. Elles suivent la progression du cours magistral et étaient orientées vers l'application.

T.D. de Probabilités en L2 Licence Économie et Gestion, associés au cours magistral de Philippe Soulier.

Le programme porte sur les bases des probabilités : combinatoire, indépendance, probabilités conditionnelles, modélisation (espace de probabilité, variables aléatoires, lois, couple de variables discrètes,...), application du théorème de la limite centrale (calcul de probabilité, intervalle de confiance).

Cours/TD de Statistiques descriptives en L1 SSA, responsable Pierre-Luc Morien.

La majorité des sciences, qu'il s'agisse des sciences expérimentales ou des sciences humaines font appel à des données, souvent nombreuses, qu'il convient de traiter à l'aide d'une méthodologie appropriée. La statistique constitue une méthode consistant dans le recueil, le traitement et l'interprétation d'ensembles de données. Le but de ces cours/TD est de définir les outils statistiques usuels permettant la description d'un caractère ou d'un couple de caractères.

- Monitrice dans le département STID de l'IUT de Paris 5 (64 heures annuelles).

Tuteur : Jean-Michel Poggi.

2002/2005 **Cours et TD de Probabilités en 1ère année.**

Le programme porte sur des notions de bases de probabilités qui viennent compléter celles acquises au cours des deux premiers trimestres. Ainsi sont abordés : les lois continues particulières (chi-deux, fisher, student), les couples de variables aléatoires discrètes et continues.

J'ai eu la responsabilité totale du cours, de l'élaboration des feuilles d'exercices et de l'évaluation des étudiants dans le cadre du contrôle continu.

Cours, T.D. et T.P. d'introduction à la simulation à l'aide du logiciel Matlab en 1ère année.

Cet enseignement a pour objectif d'apprendre aux étudiants les techniques de base des simulation afin d'illustrer les notions de statistiques fondamentales (théorème central limite, loi des grands nombres,...) et descriptives (histogrammes, descripteurs,...) et de leur permettre de mieux comprendre, par exemple, la notion "d'aléatoire" et les problèmes d'échantillonnage.

J'ai eu la responsabilité totale de cet enseignement, tant sur le contenu que sur l'évaluation. Cependant, une concertation entre les différents professeurs du département a permis de délimiter son contenu.

Projets tutorés de fin de 1ère année : élaboration de sujets, encadrement et soutenance.

Les objectifs de ces projets de fin de 1ère année sont de favoriser la pluridisciplinarité du travail et de donner aux étudiants une première expérience d'un véritable travail de groupe. Les sujets portent sur un problème de type statistique ou probabiliste et comprennent un travail mathématique qui donne lieu ensuite à une mise en œuvre informatique. Le travail mathématique permet aux étudiants de valoriser les acquis de l'année et d'aborder des thèmes et notions à la limite de leur programme.

- Allocataire de recherche à l'Université Paris XI Orsay.

2002/2005 Doctorat de Mathématiques.

- Colleuse au lycée Marcelin Berthelot (94) et au lycée Louis le Grand (75).

2001/2002 colles en classes de mathématiques supérieures (MPSI, HEC) et en mathématiques spéciales (PSI*).

- Aide scolaire.

2000/2005 soutien scolaire pour des étudiants allant de la 6ème aux classes post-bac (HEC, filière GEA professionnelle).

2004/2005 stage de vacances organisé par Prepamath.

Contrat

2004 - 2005 Contrat avec la Direction de la Recherche de Renault
“Méthodologie de hiérarchisation de mesures et d’identification de plages pertinentes pour objectiver une prestation. Application au décollage BVR (Boîtes de Vitesses Robotisée)”

Stages

- stage de maîtrise dans un laboratoire de physique : étude mathématique et expérimentale des équations différentielles régissant les langues d’Arnold [4 mois]
- stage de DEA à la Direction de la Recherche de Renault : élaboration de typologies de conducteurs par classification hiérarchique [6 mois]

Thèse

Sélection de variables pour la discrimination en grande dimension et classification de données fonctionnelles

Directeur de thèse

M. Jean-Michel POGGI, professeur à l'IUT de Paris V.

Rapporteurs de thèse

M. Philippe BESSE, professeur à l'Université Paul Sabatier de Toulouse,
M. Gérard BIAU, professeur à l'Université de Montpellier II.

Composition du jury

M. Philippe BESSE, professeur à l'Université Paul Sabatier de Toulouse
M. Gérard BIAU, professeur à l'Université de Montpellier II,
M. Jean-Jacques DAUDIN, professeur à l'INAPG,
M. Pascal MASSART, professeur à l'université de Paris XI Orsay, (président du jury)
M. Jean-Michel POGGI, professeur à l'IUT de Paris V,

Mme Nadine ANSALDI, Direction de la Recherche de Renault (invitée).

Résumé

Cette thèse s'inscrit dans le cadre de la statistique non paramétrique et porte sur la classification et la discrimination en grande dimension et plus particulièrement la sélection de variables. Elle comporte à la fois des aspects théoriques et des aspects appliqués.

Une première partie traite du problème de la sélection de variables au moyen de l'algorithme CART, tant dans un contexte de régression que de classification binaire. L'objectif est de fournir une procédure alternative à celle basée sur l'importance des variables, proposée par Breiman *et al.* Cette nouvelle procédure permet de déterminer automatiquement un paquet de variables explicatives qui intervient, de façon essentielle, dans l'explication de la réponse Y . Concrètement, nous fouillons dans une famille finie, mais typiquement grande, de paquets de variables explicatives, et nous déterminons celui qui satisfait "au mieux" notre objectif. Ainsi, nous transformons notre problème de sélection de variables en un problème de sélection de modèle. Afin de procéder à la sélection attendue, nous utilisons d'une part l'algorithme CART et d'autre part, nous nous basons sur la sélection de modèle par pénalisation développée par Birgé et Massart.

Une seconde partie est motivée par un problème réel émanant de la Direction de la Recherche de Renault qui consiste à objectiver la prestation évaluée, en l'occurrence le décollage à plat. Autrement dit, à partir de signaux temporels, mesurés au cours d'essais, nous souhaitons déterminer les signaux pertinents pour expliquer l'agrément de conduite, à savoir le ressenti de confort du conducteur lors de l'évaluation de la prestation. D'autre part, on souhaite identifier les plages temporelles responsables de cette pertinence. Par ailleurs, le caractère fonctionnel des variables explicatives fait que le problème est mal posé dans le sens où le nombre de variables explicatives

est nettement supérieur au nombre d'observations. La démarche de résolution s'articule en trois points : un prétraitement des signaux, une réduction de la taille des signaux par compression dans une base d'ondelettes commune et enfin, l'extraction des variables utiles au moyen d'une stratégie incluant des applications successives de la méthode CART.

Enfin, une dernière partie aborde le thème de la classification de données fonctionnelles au moyen de la procédure des k -plus proches voisins, méthode largement étudiée et utilisée dans le cadre de données à valeurs dans un espace fini-dimensionnel. Pour des données de type fonctionnel, on commence par les projeter dans une base de dimension d sur laquelle on utilise alors une procédure des k -plus proches voisins pour sélectionner simultanément la dimension d et la règle de classification. Nous nous intéressons, théoriquement et pratiquement, à cette phase de sélection. Tout d'abord, nous considérons la procédure classique des k -plus proches voisins puis une version légèrement pénalisée, l'idée de la pénalisation ayant été introduite par Biau *et al.*

Activités d'enseignement

Mon expérience de l'enseignement, je l'ai acquise, essentiellement, à travers mes postes de monitrice à l'IUT de Paris 5 et d'A.T.E.R. à l'Université Paris X - Nanterre, **équipe Modal'X**. J'ai en effet eu la possibilité de dispenser des enseignements de Probabilités et Statistique divers et variés, qui m'ont permis d'entrevoir le comportement et la réaction des étudiants face à cette discipline qui constitue le cadre de mes recherches. J'ai ainsi pu constater que les étudiants sont souvent demandeurs du "pourquoi et comment", à savoir qu'ils apprécient l'introduction d'exemples concrets, la mise en relation avec leurs activités futures, l'ouverture sur leurs cours à venir et un rapprochement avec mes propres travaux. Notamment, ils attendent qu'on leur explique davantage notre métier de chercheur de façon à être plus à même de comprendre en quoi cela consiste.

Tout d'abord, j'ai exercé pendant trois ans comme monitrice au sein du département **Statistique et Traitement Informatique des Données de l'IUT de Paris 5**. J'ai été chargée de dispenser, **en 1ère année**, des Cours et TD de Probabilités qui m'ont appris d'une part la structuration et la conception d'un cours, et d'autre part sa mise en application à travers d'exercices de difficulté progressive, mêlant théorie et application. Par ailleurs, les projets m'ont initiée à l'encadrement des étudiants en alliant suivis pédagogique et mathématique sur des sujets conçus par nos soins. Ces derniers avaient vocation à mettre en lumière les notions abordées dans l'année et à appréhender partiellement celles à venir. Enfin, les TP sur **Matlab**, en salle informatique, m'ont montré la difficulté à gérer de tels enseignements de part la caractère plus "ludique" de cet enseignement. Mais surtout, cela m'a permis de constater que les étudiants ont, en majorité, des difficultés à mettre en pratique les outils mathématiques et que les TP constituent un des outils à même de leur fournir une meilleure compréhension et interprétation des résultats. Par ailleurs, j'ai constaté, contre toute attente, que bien que l'informatique soit un outil connu de la majorité des étudiants, il n'en demeure pas moins que les logiciels mathématiques et la programmation restent des obstacles pour eux.

En parallèle à cet apprentissage, j'ai reçu une formation pédagogique et théorique par l'intermédiaire des différents stages que j'ai suivi dans le cadre du **C.I.E.S. Sorbonne**. A cette occasion, j'ai découvert la diversité et la complexité des systèmes universitaires en Europe, la multiplicité des méthodes d'évaluation des étudiants comme des enseignants. J'ai également recueilli des informations relatives aux droits et devoirs d'un enseignement, ainsi qu'à sa carrière. En sus, des séances de discussions entre moniteurs, encadrées par des "professionnels", ont été enrichissantes notamment pour évoquer les problèmes concrets rencontrés par chacun ou tout simplement échanger des idées générales relatives à l'Education Nationale. Enfin, des ateliers m'ont donné la possibilité de faire un travail sur moi-même afin de m'améliorer dans ce métier d'enseignant-chercheur qui me passionne (prise de parole en public, travail sur la portée de la voix, valorisation de sa recherche et de soi,...).

En tant qu'A.T.E.R., j'ai assuré des enseignements pour des étudiants allant de la **1ère année de DEUG** à la **1ère année de Maîtrise**, dans des filière mathématiques (**ISIFAR**) et économique (**AES, Économie et Gestion**). Cette diversité du public étudiant m'a permis d'aborder les mathématiques, et plus précisément les Probabilités et Statistiques, sous divers aspects. En effet, il faut tantôt insister sur la teneur théorique et tantôt sur l'aspect appliqué et l'interprétation. J'ai aussi dû apprendre à adapter mes enseignements en fonction des étudiants présents de façon à permettre à chacun de comprendre les notions dispensées. Cette expérience une cette université à caractère "sciences sociales" a été riche d'enseignements pour moi.

A l'issue de ces quatre premières années, j'ai énormément appris, tant sur le plan pédagogique que sur celui théorique. Et surtout, ce que je retiens par dessus tout, c'est le formidable accueil et l'accompagnement que j'ai pu recevoir, tout au long de ces quatre années, de la part des différentes équipes pédagogiques que j'ai pu rencontrer et côtoyer. Elles m'ont aidée lors des difficultés qui sont survenues, et m'ont fait partager leurs expériences personnelles que ce soit en matière d'enseignement ou de recherche. J'ai ainsi pu mesurer l'importance de la présence d'une équipe pédagogique. De même, ces années m'ont initiée au métier d'enseignant-chercheur puisque simultanément à cette expérience de l'enseignement, j'ai mené en parallèle mes travaux de thèse. J'ai ainsi appris à mener et gérer deux carrières, a priori distinctes (celle de chercheur d'une part et celle d'enseignant d'autre part), et qui pourtant s'imbriquent et se complètent très bien.

A présent, je poursuis ma fonction d'enseignant en qualité de maître de conférence à l'université de Nice-Sophia Antipolis. De nouvelles expériences s'offrent, notamment avec la responsabilité de certains enseignements dans leur globalité, à l'image du cours relatif aux **Méthodes CART** que je dispense à des étudiants de **2ème année de Master MASS**. Il s'agit d'une expérience nouvelle qui va m'apporter de nouvelles "connaissances" sur le métier mais aussi ma capacité d'autonomie et à enseigner.

Activités de recherche

Mes travaux de recherche se situent en statistique et portent sur la classification et la discrimination en grande dimension. Ils sont plus particulièrement axés sur le problème de la sélection de variables. Mes travaux comportent à la fois des aspects théoriques et des aspects appliqués. Ces derniers se retrouvent notamment dans le travail industriel, mené en partenariat avec la Direction de la Recherche de Renault, qui a donné lieu à contrat de recherche avec cette industrie.

Voici succinctement le problème que j'ai considéré. On dispose d'une variable traduisant le phénomène étudié et l'on désire l'expliquer à l'aide de variables explicatives. Une difficulté réside dans le fait qu'en général, on dispose d'un nombre conséquent de ces variables puisque les technologies d'aujourd'hui le permettent. Cependant, afin d'expliquer le phénomène étudié, on ne souhaite n'en conserver qu'un petit nombre. Comment faire pour exhiber ce faible nombre de variables ?

Présentation des travaux réalisés

- Méthode de Sélection de Variables utilisant CART.

Ce travail est le fruit d'une collaboration avec Marie Sauvé (en thèse à Orsay, sous la direction de Pascal Massart), et aborde le thème de la sélection de variables.

Le problème :

Soit $\mathcal{L} = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$, n copies indépendantes d'un couple de variables aléatoires (X, Y) où Y est la variable réponse et $X = (X^1, \dots, X^p)$ un vecteur de p variables explicatives à valeurs dans \mathbb{R} . On souhaite déterminer, parmi ces p variables, le plus petit paquet de variables capable, à lui seul, d'expliquer la variable Y . Autrement dit, on cherche à mettre en œuvre de la sélection de variables.

De nombreuses méthodes de sélection de variables existent, notamment dans le cadre des modèles linéaires. On peut, par exemple, citer la "Subset Selection", Lasso ou encore LARS qui sont des méthodes exhaustives ou pénalisées qui font chacune intervenir le critère des moindres carrés.

Pour notre part, nous privilégions CART et une méthode pénalisée en recourant à une approche de sélection de modèle par minimisation d'un contraste empirique pénalisé. Voici succinctement une description de notre procédure.

Soit $\Lambda = \{X^1, \dots, X^p\}$. Pour tout sous-ensemble ou paquet M de Λ , on construit l'arbre CART maximal $T_{max}^{(M)}$ en ne faisant intervenir dans les divisions de l'arbre que les variables du paquet M . Ensuite, pour tout sous-arbre M de $T_{max}^{(M)}$ noté $T \preceq T_{max}^{(M)}$, on considère le modèle $S_{M,T}$ constitué des fonctions constantes par morceaux sur la partition induite par T . Pour finir, on procède à la sélection de modèle dans la collection $\{S_{M,T}, M \in \mathcal{P}(\Lambda), T \preceq T_{max}^{(M)}\}$, en minimisant un contraste empirique pénalisé.

La question naturelle qui se pose est : Comment choisir le terme de pénalité afin que la procédure soit théoriquement valide ?

Ceci constitue le but majeur de notre travail.

Dans un second temps, nous nous sommes intéressées à l'application de la procédure proposée lorsque la valeur de p est grande.

La question qui survient est : Comment déterminer une famille \mathcal{P}^* qui serait plus petite que $\mathcal{P}(\Lambda)$ et qui pourrait se substituer à $\mathcal{P}(\Lambda)$ dans la procédure ?

Le cadre :

Afin d'apporter une réponse à la première interrogation, nous considérons les deux cadres d'étude que sont :

- la régression définie par $Y = s(X) + \varepsilon$ avec :
 - $\mathbb{E}[\varepsilon|X] = 0$;
 - il existe $\rho \geq 0$ et $\sigma > 0$ tels que pour tout $\lambda \in (-1/\rho, 1/\rho)$, $\log \mathbb{E} [e^{\lambda \varepsilon_i} | X_i] \leq \frac{\sigma^2 \lambda^2}{2(1-\rho|\lambda|)}$, avec la convention $1/0 = \infty$;
 - $\|s\|_\infty \leq R$ avec $R > 0$.
- la classification binaire, $Y \in \{0; 1\}$ avec :
 - s est le classifieur de Bayes défini par $s(x) = \mathbb{1}_{\eta(x) \geq 1/2}$ où $\eta(x) = P(Y = 1 | X = x)$;
 - une hypothèse de marge du type : $\exists h > 0, \forall x \in \mathcal{X}, |2\eta(x) - 1| > h$.

Par ailleurs, dans chacun de ces deux contextes, on considère deux situations :

(M1) : L'échantillon \mathcal{L} est scindé en trois parties indépendantes $\mathcal{L}_1, \mathcal{L}_2$ et \mathcal{L}_3 de tailles respectives n_1, n_2 et n_3 et respectivement appelées échantillon d'apprentissage, échantillon de validation et échantillon test.

L'échantillon \mathcal{L}_1 sert à la construction d'un arbre maximal tandis que \mathcal{L}_2 est utilisé dans la phase de sélection de modèle qui produit une suite de sous-arbres $\{(T_k)_{1 \leq k \leq K}\}$. L'échantillon \mathcal{L}_3 permet de procéder à la phase de sélection finale d'un paquet ainsi que d'un estimateur.

(M2) : L'échantillon \mathcal{L} est divisé en deux parties indépendantes \mathcal{L}_1 et \mathcal{L}_3 . La construction des arbres maximaux et la phase de sélection de modèle s'opèrent toutes les deux avec \mathcal{L}_1 tandis que \mathcal{L}_3 est utilisé, ici encore, comme échantillon témoin pour la phase de sélection finale.

Afin de déterminer les fonctions de pénalité bien adaptées à chacune des situations considérées, nous utilisons des méthodes de sélection de modèle et plus précisément les résultats de Birgé et Massart ([4],[5]) dans le cadre de la régression et de Massart et Nédélec [10] dans celui de la classification.

Les résultats :

Dans ce paragraphe, nous utilisons les notations suivantes.

Soit M un élément de $\mathcal{P}(\Lambda)$, autrement dit un paquet de variables explicatives, de cardinal noté $|M|$. A chaque sous-arbre T de $T_{max}^{(M)}$, ce que l'on note $T \preceq T_{max}^{(M)}$, on associe le sous-espace $S_{M,T}$ des fonctions constantes par morceaux sur la partition \tilde{T} où \tilde{T} représente l'ensemble des feuilles de l'arbre T ; $|T|$ désigne le cardinal de \tilde{T} . Pour finir, $\hat{s}_{M,T}$ désigne l'estimateur par histogramme associé à l'arbre T .

Lors de la phase de sélection de modèle, on procède à la minimisation du critère

$$crit_{\alpha,\beta}(M, T) = \gamma_{n_2}(\hat{s}_{M,T}) + pen(M, T) \quad (1)$$

à α et β fixés, α et β étant des paramètres intervenant dans l'expression de la fonction de pénalité pen .

On obtient alors, pour chaque valeur du couple (α, β) un estimateur $\tilde{s} = \widehat{s}_{\widehat{M}, \widehat{T}}$ où

$$(\widehat{M}, \widehat{T}) = \underset{M \in \mathcal{P}(\Lambda); T \preceq T_{max}^{(M)}}{\operatorname{argmin}} \operatorname{crit}_{\alpha, \beta}(M, T).$$

A l'aide de l'échantillon témoin, on sélectionne parmi la collection d'estimateurs $\{\tilde{s}; \alpha > \alpha_0 \text{ et } \beta > \beta_0\}$ l'estimateur $\tilde{\tilde{s}}$, ainsi que le paquet optimal associé.

$$\tilde{\tilde{s}} = \underset{\alpha > \alpha_0; \beta > \beta_0}{\operatorname{argmin}} \gamma_{n_3}(\tilde{s}).$$

Le but est de déterminer une fonction de pénalité pen pour laquelle on puisse obtenir des inégalités de type "oracle".

Les résultats sont obtenus conditionnellement à la construction des arbres maximaux $T_{max}^{(M)}$ et les performances des différents estimateurs sont évaluées par leurs risques ou leurs normes empiriques conditionnels.

Dans le cadre de la régression, les résultats font intervenir les normes $\|\cdot\|_{n_1}$, $\|\cdot\|_{n_2}$ et $\|\cdot\|_{n_3}$ qui sont respectivement les normes empiriques sur les grilles $\{X_i; (X_i, Y_i) \in \mathcal{L}_1\}$, $\{X_i; (X_i, Y_i) \in \mathcal{L}_2\}$ et $\{X_i; (X_i, Y_i) \in \mathcal{L}_3\}$.

Voici, deux résultats obtenus dans le cadre de la régression. Ils font intervenir les notations introduites auparavant. Le premier concerne l'estimateur \tilde{s} et le second l'estimateur final $\tilde{\tilde{s}}$.

Si \tilde{s} est obtenu par (M1) :

Si la fonction de pénalité est telle que :

$\forall M \in \mathcal{P}(\Lambda)$ et $\forall T \preceq T_{max}^{(M)}$

$$pen(M, T) = \alpha (\sigma^2 + \rho R) \frac{|T|}{n_2} + \beta (\sigma^2 + \rho R) \frac{|M|}{n_2} \left(1 + \log \left(\frac{p}{|M|} \right) \right)$$

alors, pour α et β suffisamment grands et sous certaines conditions, on a

$$\begin{aligned} \mathbb{E} [\|s - \tilde{s}\|_{n_2}^2 \mid \mathcal{L}_1] &\leq C_1 \inf_{(M, T)} \left\{ \inf_{u \in S_{M, T}} \|s - u\|_{\mu}^2 + pen(M, T) \right\} + C_2 \frac{(\sigma^2 + \rho R)}{n_2} \\ &\quad + C(\rho, \sigma, R) \frac{\mathbb{1}_{\rho \neq 0}}{n_2 \log(n_2)}. \end{aligned}$$

Ce premier résultat valide le choix des fonctions de pénalisation puisqu'une inégalité de type "oracle" est obtenue.

Quand on analyse ce résultat, on constate que la fonction de pénalité pen est la somme d'un terme proportionnel à $|T|$, le nombre de feuilles et d'un terme proportionnel à $|M|$ la taille du paquet considéré. Le premier terme n'est autre que la pénalité proposée par Breiman *et al.* [8], il sert à pénaliser les arbres de trop grande taille. Le second terme est propre à la sélection de variables, en effet il pénalise les modèles impliquant un trop

grand nombre de variables.

Des résultats similaires sont obtenus dans la situation (M2) ou dans le contexte de la classification, seule la forme des pénalités se trouve modifiée.

Si \tilde{s} est obtenu par (M1) :

Pour $\xi > 0$, avec probabilité $\geq 1 - h(\xi)$,
 $\forall \eta \in (0, 1)$,

$$\|s - \tilde{s}\|_{n_3}^2 \leq \frac{(1 + \eta^{-1} - \eta)}{\eta^2} \inf_{\tilde{s}(\alpha, \beta) \in \mathcal{G}} \|s - \tilde{s}(\alpha, \beta)\|_{n_3}^2 + \frac{C(\sigma, \rho, R, \eta) (2 \log K + \xi)}{\eta^2 n_3}$$

L'intérêt de ce résultat réside dans le fait qu'il permet de valider la phase de sélection finale de notre procédure. En effet, il montre, avec grande probabilité, que la sélection finale par échantillon test, n'altère pas, de façon notable la qualité de l'estimateur final. Dans la situation (M2), un résultat identique est obtenu.

En ce qui concerne le contexte de la classification, le résultat obtenu est analogue sauf que l'inégalité en grande probabilité est remplacée par une inégalité en espérance qui autorise la comparaison entre l'estimateur final et tout autre.

Il apparaît donc que dans chacune des situations envisagées, la procédure est validée théoriquement par l'obtention de fonctions de pénalité conduisant à des estimateurs convenables en termes de performance.

En ce qui concerne la seconde interrogation, à savoir la mise en œuvre pratique de cette procédure, nous proposons d'appliquer la procédure à une famille restreinte de paquets de variables, déterminée à partir des données et d'une idée initiée par Poggi et Tuleau dans [11] qui consiste à associer l'importance des variables, définie par Breiman *et al.* [8] et une procédure de sélection ascendante, classique en régression linéaire.

La mise en œuvre, sur un exemple simulé, de la procédure pratique, autrement dit restreinte à une famille convenablement choisie, nous permet de constater que nous obtenons alors une procédure efficace en termes de sélection de variables et de temps de calculs.

Par ailleurs, le fait que la famille soit déterminée à l'aide d'une stratégie basée sur les données, et non de façon déterministe, justifie l'emploi de la pénalité théorique définie en considérant l'ensemble des 2^p paquets possibles.

- L'objectivation de l'agrément de conduite automobile.

Ce travail a été effectué dans le cadre d'un contrat de collaboration de recherche avec la Direction de la Recherche de Renault. Il a été mené avec Jean-Michel Poggi et porte sur un problème pratique de sélection de variables.

Le problème :

L'industrie automobile, comme par exemple Renault, souhaite satisfaire sa clientèle. Dans ce but, des sondages sont réalisés afin de déterminer les prestations à améliorer. Une fois

ces dernières identifiées, il s’agit de les quantifier ou objectiver afin pouvoir intégrer les résultats dans le cahier des charges relatif à la conception du véhicule.

Concrètement, cela signifie qu’il faut déterminer des critères véhicule, encore appelés critères “physiques”, responsables de la satisfaction du conducteur (ou agrément de conduite) liée à la prestation évaluée.

L’étude qui nous occupe est relative à la boîte de vitesses et au confort ressenti par le conducteur lors de la mise en mouvement du véhicule.

Le cadre :

Afin de pouvoir mener à bien cette étude, une campagne d’essais a été réalisée. Celle-ci a requis 7 pilotes essayeurs et a mêlé différentes conditions de roulage afin de traduire diverses situations (route, autoroute, ...) et façons de conduire (brusque, douce, ...) ainsi que la charge du véhicule. De même, elle a impliqué différents réglages de la boîte de vitesses, organe évalué lors de la prestation considérée, en l’occurrence le décollage à plat pour un groupe moto-propulseur à boîte de vitesses robotisée.

En raison du caractère purement subjectif de la prestation, puisqu’il s’agit du confort ressenti, et donc de la difficulté inhérente à l’évaluation des essais, ces derniers ont été réalisés par des comparaisons par paires. Plus précisément, les pilotes testaient successivement deux produits distincts, un produit étant une combinaison d’un réglage de boîte et de conditions de roulage spécifiques. A l’issue de chaque paire testée, le pilote donnait alors l’essai préféré, sachant que, bien entendu, il n’avait pas connaissance des réglages de la boîte de vitesses afin de ne pas influencer son jugement.

Par ailleurs, lors de chacun des essais, divers signaux “physiques”, tels les accélérations ou les couples, ont été relevés à l’aide de capteurs disposés dans le véhicule utilisé pour tous les essais.

Après l’élimination des essais associés à des mesures erronées, la sélection d’un ensemble de signaux mesurés et le traitement de la réponse pilote, conduisant à l’obtention d’un ordre sur les produits qui est à la fois total et indépendant des pilotes, les données qui nous ont été communiquées, et qui constituent nos données d’étude, sont les suivantes :

$$\left\{ \begin{array}{l} X_i = (X_i^1, \dots, X_i^{21}) \text{ avec } X_i^j = X_i^j(t) \text{ la } j^{\text{ème}} \text{ variable fonctionnelle (signal)} \\ \quad \text{mesurée lors de l’essai } i, \\ Y_i = \text{ le rang, dans l’ordre total, attribué au produit testé au cours de l’essai } i. \end{array} \right.$$

Une étude menée par Ansaldi [1] établit une méthodologie d’objectivation, autrement dit une méthodologie à même de définir les grandeurs physiques représentatives de la prestation étudiée et leurs plages de valeurs optimales. Elle peut se résumer en trois grandes étapes que sont :

1. l’association d’un agrément à chacun des produits :

A l’issue des essais, on obtient une préférence individuelle (pilote par pilote) donnée sur des paires de produits. A partir de ces données “locales”, un ordre “global” est obtenu sur les produits. Il permet leur classement, du plus au moins apprécié, indépendamment du pilote.

2. l'extraction de critères et l'Analyse discriminante :

Après la génération d'une liste très importante de critères candidats grâce notamment à des règles d'expertise, ceux utiles à l'explication de l'agrément sont identifiés par une méthode d'analyse discriminante arborescente par moindres écarts. Cette étape vise à extraire des signaux la quantité minimum d'information suffisante pour expliquer l'agrément.

3. le calcul d'intervalles de tolérance :

Les critères pertinents étant déterminés, on cherche pour chacun d'eux la plage de valeurs la plus grande compatible avec la maximisation de l'agrément sous contraintes.

Pour notre part, nous nous sommes intéressés à la seconde phase en essayant de s'affranchir au maximum des connaissances "métier" et en tenant davantage compte de l'aspect fonctionnel du problème.

Les résultats :

Afin de déterminer les critères "physiques" pertinents pour l'explication de l'agrément de conduite automobile, nous devons réaliser une double phase de sélection de variables. La première consiste à identifier les variables fonctionnelles pertinentes et la seconde, à déterminer pour chacune d'elles les plages temporelles responsables de cette pertinence. En raison de la nécessité d'interpréter physiquement les résultats, chacune des deux sélections de variables doit s'effectuer dans le paquet des variables d'origine et non au sein d'un ensemble de variables construit à partir des données d'origine. Par exemple, toute combinaison linéaire est proscrite. Ceci exclut notamment des approches de type PLS ou de régression sur composantes principales.

Par ailleurs, on note que le caractère fonctionnel des données nous empêche d'appliquer la procédure de sélection de variables décrite précédemment.

Alors, afin de procéder à la double phase de sélection, nous avons adopté une démarche essentiellement basée sur deux outils : les ondelettes d'une part et CART d'autre part. Les ondelettes permettent de concentrer, en un petit nombre de coefficients, l'information contenue dans les signaux, tout en préservant une interprétation dans la grille temporelle d'origine.

L'algorithme CART intervient quant à lui dans la double phase de sélection puisque par son intermédiaire et l'importance des variables qui lui est étroitement associée, nous déterminons non seulement les variables fonctionnelles pertinentes, mais également leurs plages temporelles.

La démarche est donnée constitué des trois phases :

1. prétraitements :

Cette phase applique aux signaux mesurés des étapes de synchronisation, de débruitage par ondelettes et de recalage, visant à rendre les signaux "homogènes".

2. compression :

En raison de la dimension élevée (chacun des 114 essais est résumé par 21 signaux d'environ 1000 points), on souhaite mettre à profit la redondance de l'information dans une courbe et réduire la dimension de l'espace des variables explicative afin de

réduire le fléau de la dimension. La stratégie adoptée consiste à travailler variable fonctionnelle par variable fonctionnelle, et à projeter, dans un espace d'approximation commun, chacun des signaux.

3. sélection :

Après compression, chacun des signaux est résumé par un petit nombre de coefficients d'approximation. Cependant, cette réduction de la dimension s'avère encore insuffisante. On procède alors à une sélection pas à pas qui s'articule en cinq étapes. La première consiste à sélectionner, variable fonctionnelle par variable fonctionnelle, les coefficients discriminants en recourant à l'importance des variables. La seconde étape hiérarchise les variables fonctionnelles au moyen du coût de fausse classification. Les deux étapes suivantes procèdent à une première sélection de variables fonctionnelles en construisant une suite de modèles emboîtés et en choisissant celui de coût minimal. L'ultime étape réalise la sélection finale à l'aide de l'importance des variables.

Les résultats obtenus, à l'issue de cette stratégie, recourent pour l'essentiel ceux de l'étude menée par Ansaldi [1] et en termes d'erreur les résultats sont comparables, ce qui est intéressant en raison du caractère totalement "non informé" de notre démarche. Mais ils apportent également des informations complémentaires en retenant d'autres variables fonctionnelles cohérentes avec l'application.

- Les k -plus proches voisins pour des données fonctionnelles.

Ce travail est le fruit d'une collaboration avec Magalie Fromont (maître de conférence à l'Université de Rennes 2) et porte sur le thème de la classification de données fonctionnelles.

Le problème :

La classification binaire consiste à déterminer, au moyen d'un jeu de données $\mathcal{L} = \{(X_i, Y_i)_{1 \leq i \leq n}\}$ où $(X_i, Y_i) \in \mathcal{X} \times \{0; 1\}$, une fonction appelée classifieur qui permet d'associer à chaque observation de l'espace \mathcal{X} une réponse dans $\{0; 1\}$.

Une méthode classique et usuelle pour déterminer des classifieurs consiste à utiliser la méthode des k -plus proches voisins. Cette méthode a largement été étudiée dans le cas de données multivariées, autrement dit lorsque $\mathcal{X} = \mathbb{R}^d$.

Cependant, aujourd'hui de nombreuses applications font appel à des données de type fonctionnel auxquelles on souhaite pouvoir appliquer la méthode des k -plus proches voisins. Comment faire pour obtenir une procédure efficace ?

Le cadre :

On dispose d'un échantillon $\mathcal{L} = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ tel que les observations X_i appartiennent à un espace fonctionnel \mathcal{X} supposé séparable. Le but est de procéder à la classification supervisée de ces données fonctionnelles.

Notre travail repose sur l'approche développée par Biau, Bunea et Wegkamp dans [3]. Il s'agit de projeter ces données sur une base de l'espace \mathcal{X} et, pour $d \in \mathbb{N}^*$, à considérer les variables explicatives X_i^d qui sont les d premiers coefficients de la projection de la variables X_i .

Par ce biais, nous nous ramenons au cadre multivarié, dans lequel il est alors envisageable

de procéder à une classification par la règle des k -plus proches voisins qui doit intégrer la sélection simultanée de la dimension d de l'espace de projection et du nombre k de voisins.

Afin de procéder à cette double phase de sélection et lutter contre le “fléau de la dimension”, Biau *et al.* ont proposé de pénaliser la procédure des k -plus proches voisins par un terme en $\log(d)/m$ où m est le nombre d'observations utilisées lors de la phase de validation de la procédure des k -plus proches voisins.

Ils ont alors obtenu une inégalité de type “oracle”. Cependant, dans la pratique, ils ont utilisé une version non pénalisée de cette procédure.

Notre but est de justifier théoriquement le fait que considérer une version non pénalisée est efficace et que l'ajout d'un léger terme de pénalité, qui permet de stabiliser la procédure, n'altère pas les performances. En outre, un travail sur des données réelles et simulées permet de donner un ordre de grandeur de la pénalisation à mettre en œuvre.

Les résultats :

Dans ce paragraphe, nous utilisons les notations suivantes.

L'échantillon \mathcal{L} est scindé en un échantillon d'apprentissage $\mathcal{L}_a = \{(X_i, Y_i), i \in \mathcal{T}_l\}$ de taille l et un échantillon de validation $\mathcal{L}_v = \{(X_i, Y_i), i \in \mathcal{V}_m\}$ de taille m tel que $m + l = n$. Pour chaque $k \in \{1, \dots, l\}$ et chaque $d \in \mathcal{D} \subset \mathbb{N}^*$, on note $\hat{p}_{l,k,d}$ la règle des k -plus proches voisins construite à partir de l'échantillon $\{(X_i^d, Y_i), i \in \mathcal{T}_l\}$. Autrement dit, $\hat{p}_{l,k,d}$ est défini, pour $x \in \mathbb{R}^d$ par :

$$\hat{p}_{l,k,d}(x) = \begin{cases} 0 & \text{si } \sum_{i=1}^k \mathbb{I}_{Y_{(i)}(x)=0} \geq \sum_{i=1}^k \mathbb{I}_{Y_{(i)}(x)=1}, \\ 1 & \text{sinon} \end{cases}$$

avec $(X_{(1)}^d, Y_{(1)}), \dots, (X_{(l)}^d, Y_{(l)})$ les variables de l'échantillon \mathcal{L}_a réordonnées selon l'ordre croissant des distances euclidiennes $\|X_i^d - x\|$.

Par ailleurs, pour une observation x de l'espace fonctionnel \mathcal{X} , on introduit le classifieur $\hat{\phi}_{l,k,d}(x) = \hat{p}_{l,k,d}(x^d)$.

On définit alors la dimension finale \hat{d} et le nombre de voisins associé \hat{k} par :

$$(\hat{d}, \hat{k}) = \underset{k \in \{1, \dots, l\}, d \in \mathcal{D}}{\operatorname{argmin}} \left(\frac{1}{m} \sum_{i \in \mathcal{T}_m} \mathbb{I}_{\hat{\phi}_{l,k,d}(X_i) \neq Y_i} + \operatorname{pen}(d) \right),$$

où pen est une fonction de pénalité, positive ou nulle, à ajuster.

L'estimateur final est alors pour $x \in \mathcal{X}$, $\hat{\phi}_n(x) = \hat{\phi}_{l,\hat{k},\hat{d}}(x)$.

On obtient un résultat théorique sur les performances de l'estimateur final $\hat{\phi}_n$ relativement au classifieur de Bayes noté ϕ^* :

Soit $n \geq 2$, $\theta \geq 1$ et $\operatorname{pen}(d)$ un terme de pénalité positif ou nul, alors pour $\beta > 0$, et sous une hypothèse de marge, on a :

$$(1-\beta)\mathbb{E}[L(\hat{\phi}_n) - L(\phi^*) | \mathcal{L}_a] \leq (1+\beta) \inf_{k \in \{1, \dots, l\}, d \in \mathcal{D}} \left\{ \left(L(\hat{\phi}_{l,k,d}) - L(\phi^*) \right) + \operatorname{pen}(d) \right\} + R. \quad (2)$$

où R est un terme de reste.

Si dans ce résultat on pose $pen(d) = 0$, ceci permet de prouver l'efficacité de la procédure non pénalisée des k -plus proches voisins lorsque l'on dispose de données fonctionnelles. Par ailleurs, on note que l'ajout d'une pénalité ne vient pas altérer les performances de l'estimateur final dans la mesure où cette dernière ne perturbe pas le terme $L(\hat{\phi}_{l,k,d}) - L(\phi^*)$ dont on ne connaît qu'une évaluation partielle grâce aux travaux de Györfi.

De plus, la procédure des k -plus proches voisins présente une instabilité qui résulte d'une part de la sélection de la dimension \hat{d} et d'autre part du découpage en deux de l'échantillon initial. On envisage ensuite une pénalité qui, correctement calibrée, peut gommer cette variabilité comme l'illustre le travail effectué tant sur des données réelles que sur des données simulées.

Ainsi, bien que théoriquement la pénalisation n'engendre pas d'amélioration en termes de risque, en pratique la prise en compte d'un terme de pénalité peut s'avérer appréciable dans le sens où ce dernier apporte une stabilisation des résultats dans tous les exemples étudiés et en améliore parfois les performances.

Références

- [1] N. Ansaldi. *Contributions des méthodes statistiques à la quantification de l'agrément de conduite*. PhD thesis, Marne-la-Vallée, (2002).
- [2] A. Berlinet, G. Biau, and L. Rouvière. Functional learning with wavelets. soumis à IEEE Trans. Inf. Theory, (2005).
- [3] G. Biau, F. Bunea, and M. Wegkamp. Functional classification in Hilbert spaces. *IEEE Trans. Inf. Theory*, 51(6) :2163–2172, 2005.
- [4] L. Birgé and P. Massart. Gaussian model selection. *Journal of the European Mathematical Society*, 3(3) :203–268, (2001).
- [5] L. Birgé and P. Massart. Minimal penalties for gaussian model selection. à paraître dans Probability Theory and Related Fields, (2005).
- [6] L. Breiman. Bagging predictors. *Machine Learning*, 24(2) :123–140, (1996).
- [7] L. Breiman. Random forests. *Machine Learning*, 45(1) :5–32, (2001).
- [8] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification And Regression Trees*. Chapman et Hall, (1984).
- [9] E. Lebarbier. *Quelques approches pour la détection de ruptures à horizon fini*. PhD thesis, Université Paris XI Orsay, (2002).
- [10] P. Massart and E. Nédélec. Risk bounds for statistical learning. accepted to the Annals of Statistics, (2005).
- [11] J.M. Poggi and C. Tuleau. Classification supervisée en grande dimension. Application à l'agrément de conduite automobile. *Preprint Université Paris XI Orsay*, pages 1–16, (2005).
- [12] F. Rossi and B. Conan-Guez. Functional multi-layer perceptron : a non-linear tool for functional data analysis. *Neural networks*, 18(1) :45–60, (2005).
- [13] F. Rossi and N. Villa. Classification in Hilbert spaces with support vector machines. In *Proceedings of ASMDA 2005*, pages 635–642, Brest, France, (2005).

Perspectives

Les travaux menés peuvent être saisis au travers de deux axes. Le premier est thématique, on peut dégager deux directions :

- la sélection de variables dans un contexte de régression et de classification supervisée ;
- le traitement de données fonctionnelles dans le cadre de la classification supervisée.

D'autre part, un second axe concerne le type de ces travaux qui sont tant théoriques qu'appliqués et méthodologiques.

Les prolongements et les perspectives liés à ces travaux sont, par conséquent, variés. Quelques-uns sont esquissés ci-dessous :

- Les travaux relatifs à la “sélection de variable à travers CART” font intervenir une fonction de pénalité dépendant de deux paramètres α et β . Ils sont déterminés, dans l'étude proposée, au moyen d'un échantillon test. Une perspective consiste à “calibrer”, à partir des données, ces deux paramètres en utilisant une méthode basée sur la détection de changement de pente du contraste empirique, à l'image de l'“heuristique de pente” (cf. Lebarbier [9]).
- Dans ce même travail, une stratégie de mise en œuvre pratique de la procédure de sélection de variables est proposée et son application est illustrée sur un exemple. Cette voie est à explorer plus en profondeur en testant d'autres stratégies ou des variantes et en tentant de la justifier théoriquement dans des contextes restreints.
- Le travail consacré à “l'objectivation de l'agrément de conduite” se situe dans un contexte industriel et concerne le thème de la sélection de variables tout en traitant des données fonctionnelles. Il s'inscrit dans la continuité du travail, mené par Ansaldi [1], sur l'objectivation d'une prestation. Il se poursuit par une thèse CIFRE, qui débute actuellement chez Renault, sur les questions d'objectivation simultanée de plusieurs prestations. Ainsi, le travail réalisé est une contribution qui s'inscrit dans le programme de recherche de l'industriel automobile.
- Par ailleurs, dans cette étude industrielle, on tente de faire un lien entre des parties théoriques et appliquées de l'étude relative au premier thème : la sélection de variables. On peut penser à développer des liens semblables pour le second thème : le traitement de données fonctionnelles, notamment en considérant le problème du choix de la base permettant de représenter les objets d'intérêt. En effet, dans l'étude à teneur industrielle, pour chacune des variables fonctionnelles, on détermine la base de représentation commune en la choisissant, parmi les bases des espaces d'approximation, à l'aide d'un critère indépendant de l'objectif de discrimination. Dans celle relative à la classification de données fonctionnelles au moyen des k -plus proches voisins, la base de représentation privilégiée est celle de Fourier, mais de récents travaux dûs à Berlinet *et al.* [2] montrent que l'on peut également considérer les bases d'ondelettes. Mais quelle que soit la base considérée, l'introduction d'un “léger” terme de pénalité a le mérite de stabiliser les résultats tout en préservant de bonnes performances. Des prolongements consistent d'une part à affiner la détermination de la fonction de pénalité à mettre en jeu, et d'autre part, à envisager le recours à d'autres types de base.

- Concernant les travaux consacrés à la méthode de classification des k -plus proches voisins, voici deux pistes un peu plus spéculatives.
Sur le plan appliqué, on peut envisager de mettre en compétition diverses méthodes comme les Support Vector Machines (cf. Rossi et Villa [13]) ou les Réseaux de Neurones (cf. Rossi et Conan-Guez [12]), qui sont des méthodes disponibles pour la classification de données fonctionnelles. Bien que plus délicates à étudier théoriquement, on peut également penser à des méthodes basées sur le rééchantillonnage, à l'instar du bagging (cf. Breiman [6]) et des Random Forests (cf. Breiman [7]), par exemple.
Par ailleurs, sur le plan théorique, on peut s'intéresser à l'étude de l'estimateur des k -plus proches voisins pour des données fonctionnelles, en examinant son optimalité au sens minimax.

Références

- [1] N. Ansaldi. *Contributions des méthodes statistiques à la quantification de l'agrément de conduite*. PhD thesis, Marne-la-Vallée, (2002).
- [2] A. Berlinet, G. Biau, and L. Rouvière. Functional learning with wavelets. soumis à *IEEE Trans. Inf. Theory*, (2005).
- [3] G. Biau, F. Bunea, and M. Wegkamp. Functional classification in Hilbert spaces. *IEEE Trans. Inf. Theory*, 51(6) :2163–2172, 2005.
- [4] L. Birgé and P. Massart. Gaussian model selection. *Journal of the European Mathematical Society*, 3(3) :203–268, (2001).
- [5] L. Birgé and P. Massart. Minimal penalties for gaussian model selection. à paraître dans *Probability Theory and Related Fields*, (2005).
- [6] L. Breiman. Bagging predictors. *Machine Learning*, 24(2) :123–140, (1996).
- [7] L. Breiman. Random forests. *Machine Learning*, 45(1) :5–32, (2001).
- [8] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification And Regression Trees*. Chapman et Hall, (1984).
- [9] E. Lebarbier. *Quelques approches pour la détection de ruptures à horizon fini*. PhD thesis, Université Paris XI Orsay, (2002).
- [10] P. Massart and E. Nédélec. Risk bounds for statistical learning. accepted to the *Annals of Statistics*, (2005).
- [11] J.M. Poggi and C. Tuleau. Classification supervisée en grande dimension. Application à l'agrément de conduite automobile. *Preprint Université Paris XI Orsay*, pages 1–16, (2005).
- [12] F. Rossi and B. Conan-Guez. Functional multi-layer perceptron : a non-linear tool for functional data analysis. *Neural networks*, 18(1) :45–60, (2005).
- [13] F. Rossi and N. Villa. Classification in Hilbert spaces with support vector machines. In *Proceedings of ASMDA 2005*, pages 635–642, Brest, France, (2005).

Publications

- (P1.1) J.M Poggi et C. Tuleau, *Classification supervisée en grande dimension. Application à l'agrément de conduite automobile*. Preprint avril 2005 et article accepté pour publication par la Revue de Statistique Appliquée
- (P1.2) J.M Poggi et C. Tuleau, *Classification supervisée en grande dimension. Application à l'agrément de conduite automobile*, Actes de conférence des 36^{ème} journées de Statistique, mai 2004
- (P2.1) M. Fromont et C. Tuleau, *Functional classification with margin conditions*. Article accepté à COLT'06 (19th Annual Conference on Learning Theory, Pittsburgh)
- (P2.2) M. Fromont et C. Tuleau, *Les k-plus proches voisins pour des données fonctionnelles*. Article accepté aux 38^{ème} journées de Statistique, mai-juin 2006
- (P3.1) M. Sauvé et C. Tuleau, *Sélection de variables avec CART*. Actes de conférence des 37^{ème} journées de Statistique, juin 2005
- (P3.2) M. Sauvé et C. Tuleau, *Variables selection with CART*. Préprint pour mai 2006

Groupes de travail et congrès

- Groupe de travail MODAL'X de l'Université Paris X-Nanterre (2005/2006) ;
- Groupe de travail de l'ENS Ulm organisé par P. Massart, P. Reynault et G. Stolz (Apprentissage, 2005/2006 et Support Vector Machine, 2004/2005) ;
- Groupe de travail INAPG-SELECT (Analyse de données de biopuces, 2004/2006) ;
- Journées de Statistique de la Sfds (mai 2006, juin 2005, mai 2004) ;
- Workshop : Statistical methods for post-genomic data (Toulouse, mars 2006) ;
- Séminaire de statistique mathématique et application organisé par S. Huet et P. Massart au CIRM (novembre 2004, septembre 2002) ;
- Congrès des Jeunes Probabilistes et Statisticiens de France (Aussois, avril 2006 et septembre 2005) ;
- Journées de Statistique de Rennes2 (“application en biologie”, septembre 2005 et “Données fonctionnelles”, septembre 2004).

Communications orales

Thème théorique (P2 et P3) :

- Communication à COLT'06 (19th Annual Conference on Learning Theory, Pittsburgh, Juin 2006)
- Communication orale aux 38^{ème} Journées de Statistique, Clamart (Mai-Juin 2006)
- Groupe de travail Modal'X (Mars 2006)
- Groupe de travail “Apprentissage” de l'ENS Ulm (Janvier 2005)
- Communication aux premières rencontres des jeunes statisticiens, Aussois (Septembre 2005)
- Communication aux Rencontres mathématiques (“Séminaire de statistique mathématique et applications”) organisées par S. Huet et P. Massart, Luminy (Novembre 2004)

Thème appliqué (P1) :

- Séminaire appliqué de la Haute Ecole Robert Schuman à Arlon (Belgique) (Mars 2005)
- Poster aux journées de “Données fonctionnelles” de l'Université Rennes 2 (Septembre 2004)
- Communication orale aux 36^{ème} Journées de Statistique, Montpellier (Mai 2004)
- Groupe de travail INAPG-SELECT “analyse de données de biopuce” (Février 2004)
- Communications orales à la Direction de la Recherche de Renault

Séminaires :

- Université de Nice, équipe de Probabilités et Statistiques, (Avril 2006)
- Université de Lille 1, laboratoire de Probabilités et Statistique Paul Painlevé, (Mars 2006)
- Université Paul Sabatier de Toulouse, équipe de Statistique et Probabilité, (Février 2006)
- Université de Montpellier II, équipe de Probabilités et Statistique, (Février 2006)
- Université de Rennes 2, équipe de statistique de l'IRMAR, (Février 2006)
- Université Paris - Dauphine, (Janvier 2006)