Agence Nationale de la Recherche ANR	Domaines émergents	Réservé à l'organisme gestionnaire du programme N° de dossier : ANR-08-XXXX-00 Date de révision :
	Document de soumission B	Edition 2008

Acronyme	**ClasSel**
Titre du projet	**Classification croisée et sélection de modèle**
Proposal title	***Co-clustering and model selection***

Résumé

ClasSel est un projet de recherche académique qui vise à développer des méthodes de transformation de données en connaissances. Les données en question se présentant sous la forme d'une matrice individus–variables, nous cherchons à produire de la connaissance sous la forme de groupes homogènes de données associant conjointement les individus et les variables. C'est le problème de classification croisée. Nous envisageons d'attaquer ce problème formellement à travers une modélisation probabiliste. Notre projet vise à adapter cette modélisation aux problèmes spécifiques de la classification croisée pour les données de grande taille, une attention particulière étant mise sur le problème, fondamental, du choix du nombre de groupes. C'est la question de la sélection de modèle. À cette fin, nous comptons nous placer dans un cadre statistique nouveau et particulièrement bien adapté. Nous nous proposons aussi de mettre en œuvre nos solutions sur des exemples concrets, comme le challenge Netflix sur les systèmes de recommandation, et de traiter des applications en analyse automatique de texte et en marketing.

Notre stratégie scientifique consiste à attaquer de front les questions de fond de la modélisation en apprentissage et de la sélection de modèle pour trouver des solutions en rupture avec l'existant. Pour atteindre cet objectif, nous proposons de mettre en œuvre une approche décloisonnée mobilisant des chercheurs de différentes communautés STIC (statistiques, analyse de données, apprentissage et informatique) sur des applications concrètes liées à de grandes masses de données.

Table des matières

1 Programme scientifique et technique/Description du projet

1.1 Problème posé

ClasSel est un projet de recherche académique qui vise à développer des méthodes qui permettent de transformer des données en connaissances. Les données en question sont sous la forme d'une matrice individus–variables. Nous cherchons à comprendre comment construire de manière automatique, à partir de ces données, des groupes ou des hiérarchies d'« individus » et de « variables » définies simultanément. Ces hiérarchies associant individus et variables sont ensuite exploitées pour compléter les données ou pour servir de base à la définition de terminologies ou de « contextes ». C'est le problème de classification croisée. Lorsque qu'on attaque ce type de problème, une attention toute particulière doit être portée au problème fondamental du choix du nombre de groupe : c'est la question de la sélection de modèle. Nous abordons ces questions formellement dans un cadre statistique nouveau et particulièrement bien adapté.

Nous avons structuré notre projet autour de quatre tâches :

1. l'étude de la classification croisée,
2. l'étude du problème spécifique de la sélection de modèle,
3. les questions algorithmiques liées notamment à notre volonté d'attaquer des grandes masses de données,
4. les applications.

Nous proposons d'attaquer le problème de classification croisée formellement en utilisant une modélisation probabiliste. Notre projet vise à adapter ce type de modèle aux problèmes spécifiques de la classification croisée pour les données de grande taille et à ajuster en conséquence les algorithmes d'estimation du type EM.

Le but principal de l'aspect « Sélection de modèles » de notre projet est le développement de nouvelles méthodes de mise en œuvre de sélection de modèles appliquées à l'apprentissage statistique. Nous comptons intégrer les points de vue issus des domaines de la fouille de données et d'apprentissage à la Statistique paramétrique classique à fin d'explorer de très grands ensembles de données. Nous visons à déterminer, au travers des idées de la sélection de modèle, quels prédicteurs ont le plus d'influence sur les résultats et à évaluer le degré d'incertitude de nos prévisions.

La partie algorithmique a pour but d'adapter les solutions proposées aux contraintes liées au passage à l'échelle. Elle a aussi pour objectif de tester les différentes solutions envisagées et de fournir à la communauté des composants logiciels réutilisables.

Les applications sont vues à la fois comme moteurs et démonstrateurs de nos recherches. Il sont en effet moteurs à travers les problèmes spécifiques posés. Ces applications concernent la segmentation marketing en collaboration avec l'université de Vienne (Autriche), les systèmes de recommandation à travers le challenge Netflix et la fouille de texte.

Notre stratégie scientifique consiste à attaquer de front les questions fondamentales de la modélisation en apprentissage et de la sélection de modèle pour trouver des solutions en rupture avec l'existant. Pour atteindre cet objectif, nous proposons de mettre en œuvre une approche décloisonnée mobilisant des chercheurs de différentes communautés STIC (statistiques, analyse de données, apprentissage et informatique) sur des applications concrètes liées à de grandes masses de données. Le groupe projet associe logiquement des statisticiens (D. Fourdrinier du LITIS), des spécialistes de l'analyse des données (G. Govaert d'Heudiasyc et M. Nadif du CRIP 5), des chercheurs de la communauté apprentissage (S. Canu, A. Rakotomamonjy, G. Gasso du LITIS et Y. Grandvalet de Heudiasyc) et des informaticiens (F.X. Jollois du CRIP 5).

1.2 Contexte et enjeux du projet

La théorie et les algorithmes issus de l'apprentissage statistique jouent un rôle crucial dans la perspective de percées technologiques pour les ordinateurs. Ils interviennent dans de nombreux domaines tels la bio-informatique, l'extraction d'information, le filtrage collaboratif, l'analyse d'image et les interfaces cerveau-machines. Une raison pour laquelle les progrès espérés ont été reportés dans ce domaine réside en la difficulté d'appréhender les fondations théoriques de l'apprentissage. D'où l'importance d'aborder de front les questions fondamentales tels la modélisation et la sélection de modèle pour aboutir à des avancées significatives. Il s'agit, dans ce projet, de les aborder à travers un problème particulier : celui de la classification croisée. Le regain d'intérêt pour ce type de méthode est lié à de nombreuses applications potentielles notamment pour les systèmes de recommandations. En effet l'approche de type classification croisée semble la plus pertinente pour résoudre ce type de problème.

Le cas du challenge Netflix Netflix est une société de location de vidéo qui cherche se doter d'un système de recommandation performant. Les clients de Netflix sont invités à donner une note de 1 à 5 sur les films qu'ils ont vus. Le prix Netflix est une compétition en cours qui vise à récompenser le premier algorithme capable d'améliorer le système de prédiction de note maison de 10% au sens d'un critère quadratique. Doté d'un million de dollars, cette

compétition rencontre un certain succès, ce qui signifie déjà en soi que le problème reste ouvert et que les solutions actuelles ne sont pas satisfaisantes. Il est aussi symptomatique de noter que, parmi les solutions performantes proposées, il ne s'en trouve aucune clairement identifiée comme issue de l'hexagone. Cela peut se justifier par notre peu de goût pour ce genre d'exercice, dont par ailleurs l'intérêt est discutable, mais il n'en reste pas moins qu'aborder ce genre d'exercice permet de se poser des questions fondamentales intéressantes et difficiles soit en l'occurrence : comment développer une technique de classification croisée permettant le passage à l'échelle. Pour développer une telle méthode, il apparait fondamental de régler la question centrale de la sélection de modèle (combien de groupes et quel modèle pour chaque groupe), toujours critique dans les problèmes d'apprentissage, et ce d'autant plus que les problèmes sont de grande taille. Notre projet attaque ces questions et vise à proposer des solutions génériques, applicables à un grand nombre de problèmes importants aujourd'hui, notamment en filtrage collaboratif, bioinformatique, traitement d'information textuelle et en marketing.

La classification croisée La Classification Automatique (*Clustering*) ou classification non supervisée dans la terminologie de la reconnaissance des formes a pour but d'obtenir une représentation simplifiée des données initiales. Elle consiste à organiser un ensemble d'objets (individus) décrits par un ensemble de caractères (variables) en classes homogènes ou classes naturelles. Il s'agit d'une démarche très courante qui permet de mieux comprendre l'ensemble analysé. Le concept récent d'extraction de connaissances à partir de données (*Knowledge Discovery and Data Mining*) cherche à répondre à cette préoccupation en combinant différentes techniques d'analyse de données et d'apprentissage dont la classification automatique. Le problème posé par les méthodes de classification non supervisée est un problème difficile d'optimisation : il s'agit de rechercher dans l'ensemble des partitions possibles celle qui optimise un critère donné assurant une homogénéité des classes. Citons, par exemple, la méthode des centres mobiles connue sous le nom de méthode réallocation-centrage ou *k-means* et reposant sur un critère d'inertie défini à partir d'une métrique. Ce type d'algorithme et ses variantes sont conçus d'un point de vue heuristique et convergent itérativement, à partir d'une situation initiale, vers un optimum local. Les inconvénients de cette approche sont, d'une part, la justification du choix de la métrique et du critère utilisés et, d'autre part, la dépendance entre la solution fournie et l'initialisation de l'algorithme.

Les problèmes cités ci-dessus ont conduit depuis quelques années à une évolution de l'approche algorithmique, heuristique et géométrique vers une approche statistique. L'introduction de modèles de mélanges (voir par exemple les deux ouvrages (McLachlan and Basford, 1988; McLachlan and Peel, 2000)), a permis de formaliser l'idée intuitive de la notion de classe naturelle et de donner une interprétation statistique à certains critères métriques. De plus, cette approche permet de proposer de nouveaux critères répondant à des hypothèses précises qui faisaient défaut aux algorithmes traditionnellement utilisés et diffusés dans les logiciels. Cette démarche permet de répondre à plusieurs interrogations concernant la classification dans le cas où les données sont de différents types, continu, catégoriel, binaire ou encore des données structurées sous forme de table de contingence.

Pour aborder le problème de la classification, l'approche mélange repose sur l'idée suivante : étant donné que les classes diffèrent entre elles, on suppose que les variables, pour chaque classe, suivent une loi de probabilité dont les paramètres sont en général différents d'une classe à l'autre ; on parle alors de modèle de mélange de lois de probabilité. Sous cette hypothèse, les données initiales sont considérées comme un échantillon de taille n d'une variable aléatoire p-dimensionnelle dont la densité est un mélange des k distributions de probabilité spécifiques à chaque classe. Pour déterminer les g classes (g connu ou à estimer) dans le cadre de la classification, il suffit alors d'identifier les g distributions en estimant leurs paramètres. Le problème de la classification peut alors être traité sous deux approches différentes. L'approche « Estimation » (*Maximum Likelihood*) qui s'attaque directement à l'estimation des paramètres à partir desquels se déduit une partition et l'approche « Classification » (*Classification Maximum Likelihood*) qui recherche directement la partition à partir de laquelle on peut estimer les paramètres. Les algorithmes utilisés généralement sont de type EM (Dempster et al., 1977) maximisant la vraisemblance des données dans l'approche « Estimation » et la vraisemblance classifiante, dite aussi vraisemblance des données complétées, dans l'approche « Classification ».

La classification automatique, comme la plupart des méthodes d'analyse de données, peut être considérée comme une méthode de réduction et de simplification des données. Dans le cas où les données mettent en jeu deux ensembles I et J, ce qui est le cas le plus fréquent, la classification automatique en ne faisant porter la structure recherchée que sur un seul des deux ensembles, agit de façon dissymétrique et privilégie un des deux ensembles, contrairement par exemple à l'analyse factorielle des correspondances qui obtient simultanément des résultats sur les deux ensembles ; il est alors intéressant de rechercher *simultanément* une partition des deux ensembles. Ce type d'approche a suscité récemment beaucoup d'intérêt dans divers domaines tels celui des biopuces où l'objectif est de caractériser des groupes de gènes par des groupes de conditions expérimentales (Madeira and Oliveira, 2004), (Cheng and Church, 2000) ou encore celui de l'analyse textuelle où l'objectif est de caractériser des classes de documents par des classes de mots (Dhillon, 2001; Dhillon et al., 2003).

Dans ce projet, nous nous focaliserons sur ce dernier type de méthodes. Plus précisément, celui-ci consiste à rechercher de blocs homogènes par une classification simultanée des individus et des variables. On parlera dans ce cas de classification croisée.

La classification croisée ou classification par blocs est connue dans la littérature anglaise sous différents noms. Souvent on parle de *two-mode, two-side and two-way clustering, block clustering, biclustering* ou encore *co-clustering*. Ces approches diffèrent souvent dans les algorithmes employés et la nature des blocs recherchés qui peuvent être isolés ou imbriqués. Dans ce projet nous nous concentrons sur la recherche de blocs isolés qui, notons le, peuvent être obtenus par une simple permutation des lignes et des colonnes d'un tableau de données. Autrement dit, dans cette situation, la classification croisée consiste à chercher une paire de partitions $(\mathbf{z}, \mathbf{w})$, où $\mathbf{z}$ est une partition de l'ensemble I des n individus et $\mathbf{w}$ est une partition de l'ensemble J des p variables. Le problème de recherche de $\mathbf{z}$ et $\mathbf{w}$ est habituellement résolu dans la littérature de manière itérative par une optimisation alternée de la partition des individus en fixant celle des variables puis de la partition des variables en fixant celle des individus. La classification par blocs s'avère très utile et plus riche que la classification simple dans certaines situations. Parmi les premiers travaux témoignant déjà de l'intérêt de ce type de méthodes, nous citerons Hartigan (1975), Govaert (1977, 1983), Bock (1979) et Marchotorchino (1987). Plus récemment plusieurs auteurs se sont intéressés à ce domaine sous différentes approches, nous retiendrons les travaux de Vichi (2000), Vichi and A.L. (2001), Bock (2003) et de Van Mechelen et al. (2004).

Dans le contexte actuel où les données sont de plus en plus de grande taille, un des premiers avantages de ce type de classifications est de pouvoir résumer rapidement et efficacement un tableau de données de grande taille. Bien entendu, la recherche de blocs homogènes peut être réalisée à l'aide des méthodes classiques en procédant par classification séparée sur l'ensemble des individus et l'ensemble des variables puis en croisant les partitions optimales afin de construire des blocs homogènes. Un tel procédé privilégie malheureusement un des deux ensembles en ne faisant porter la structure recherchée que sur l'ensemble des individus ou l'ensemble des variables, et ne traduit pas convenablement les relations existantes entre un groupe d'individus et un groupe de variables.

Suivant la nature des données, plusieurs modèles de mélange croisés ont été proposés (Govaert and Nadif, 2003, 2005, 2006). En considérant les deux approches « Estimation » et « Classification », plusieurs algorithmes de type EM ont été développés par ces même auteurs (Govaert and Nadif, 2008). En plus de leur parcimonie et de leur flexibilité, les modèles étudiés permettent, sous l'approche "Classification", de retrouver des critères déjà utilisés dans la littérature. Les comportements et les performances des algorithmes ont été largement étudiés mettant en évidence leurs intérêts : rapidité, efficacité et capacité de gestion des matrices creuses (*sparse data*). Cependant ces algorithmes présentent quelques inconvénients. Comme tous les algorithmes itératifs, ils reposent sur la donnée d'une situation initiale qui peut conduire à des optimums locaux non satisfaisants. D'autre part, la maximisation des critères par les différents algorithmes proposés n'est pas directe mais alternée ce qui implique la gestion des conditions d'arrêt dans chaque étape de l'algorithme. Ces algorithmes nécessitent une adaptation pour classifier des données comportant des valeurs manquantes. Enfin, et pour une meilleure exploitation des données, ces algorithmes ont besoin d'un outil de visualisation efficace. Tous ces problèmes sont d'actualité et peuvent être abordés dans un cadre unifié, celui du modèle de mélange croisé. Nous proposons dans ce projet de traiter les différentes tâches, en tenant compte du type des données qui peuvent être binaires, continues ou encore se présentant sous forme de table de contingence.

Le problème de sélection de modèle La question de la sélection de modèle est double dans le cas de la classification croisée. Il s'agit de trouver combien de classes par lignes et combien de classes par colomme il faut considérer. Notre stratégie consiste à adapter les techniques existantes les plus prometteuses à ce cadre spécifique, ce qui, à notre connaissance, n'a pas encore été fait.

Massart (2007) donne un exposé complet d'une théorie non asymptotique de la sélection de modèle avec un certain nombre d'applications à la sélection de variables et à l'apprentissage statistique. Dans ce cadre, nous avons en vue de proposer de nouvelles procédures de sélection de modèle contruites elles aussi sur la base de critères non asymptotiques. Pour introduire notre approche, nous considèrerons le contexte simple d'un modèle de régression linéaire.

Un cadre particulier de la sélection de modèle où des critères non asymptotiques de comparaison de procédures de sélection peuvent être exprimés facilement est celui de la sélection de variables. Exprimons ce dernier contexte par le modèle de régression linéaire dans $\mathbb{R}^n$, soit $y = X\beta + \epsilon$ où y et ϵ sont les vecteurs-colonnes n-dimensionnels respectivement constitués des observations et des erreurs aléatoires, X est la matrice de plein rang $n \times \pi$ dont les colonnes forment les π variables explicatives et où β est le vecteur de $\mathbb{R}^\pi$ composé des coefficients inconnus de la régression. Lorsque le nombre π de variables est trop grand pour fournir une explication claire des observations, on est naturellement intéressé par la recherche de la sélection, parmi elles, d'un nombre p de variables qui restent suffisamment significatives dans cette régression (p étant, si possible, bien plus petit que π). Dans cette hypothèse, $\pi - p$ composantes de β sont nulles et le "vrai" modèle est $y = X_I\beta_I + \epsilon$ où I est un sous-ensemble de p nombres entiers positifs distincts inférieurs à π, β_I est un vecteur de $\mathbb{R}^p$ contenant les composantes de β indexées par I et X_I est la matrice $n \times p$ qui contient les colonnes de X indexées par I.

Dans la littérature, de nombreux critères de sélection ont été proposés. Ils peuvent être rangés en deux classes, chacune correspondant à un principe qui stipule ce que doit prendre en compte une règle de choix de p variables. Tout le monde s'accorde sur le fait qu'un bon critère doit contenir un terme qui traduit l'adéquation des variables

au modèle. Usuellement, sur la base de l'estimation de β_I, par l'estimateur des moindes carrés $\hat{\beta}_I = (X_I^t X_I)^{-1} X_I$, on utilise la somme des carrés résiduels $||y - X\hat{\beta}_I||^2$. La différence d'appréciation dans la conception d'une bonne sélection vient de ce que certains exigent d'y ajouter un terme de complexité alors que d'autres préconisent de la pondérer par un terme de pénalité de sur ou sous-représentation des variables dans la régression. Ainsi le premier type de règles de sélection est fondé sur

$$||y - X\hat{\beta}_I||^2 + \delta(I)$$

tandis que le second se base sur

$$\alpha(I)||y - X\hat{\beta}_I||^2$$

où $\delta(I)$ et $\alpha(I) > 0$ sont des constantes indépendantes de l'observation y.

Parmi les règles du premier type, on reconnait le critère de Mallows (1970) et le critère d'information d'Akaike (Akaike, 1970) où $\delta(I) = 2\,p$. Le critère d'information bayésien de Schwarz (1978) correspond à $\delta(I) = p \operatorname{Log} n$. La procédure utilisée par Foster and George (1994) utilise $\delta(I) = p \operatorname{Log} p$. Le critère de prédiction finale proposé par Rissanen (1986) est de cette forme avec $\delta(I) = \psi\, p$ où ψ est une constante strictement positive, le choix $\psi = \frac{2\,n-d}{n-d}$ fournissant une procédure de sélection asymptotiquement équivalente à la validation croisée "d-fold" (Zhang, 1993).

La classe des règles de la deuxième forme contient le critère de validation croisée généralisée de Craven and Wahba (1978) avec $\alpha(I) = n\,(n-p)^{-2}$. Une procédure voisine est celle proposée par Hocking (1976) et Thomson (1978a,b) où $\alpha(I) = (n-1)\,(n-p)^{-1}\,(n-p-1)^{-1}$. Miller (1990) montre que $\alpha(I) = (n-2)\,(n-p)^{-2}$ fournit une approximation du critère de Allen (1971).

Cette liste de critères est loin d'être exhaustive et il est intéressant de noter que chacun d'eux provient d'une motivation qui lui est propre. Ainsi, originellement, les procédures de Mallows et Akaike furent respectivement déduites d'une estimation du carré de l'erreur de prédiction moyenne et de la comparaison de la vraisemblance approchée d'un modèle donné à un modèle de référence. Il n'était pas invoqué de motifs en termes d'adéquation et de complexité.

1.3 Objectifs et caractère ambitieux/novateur du projet

Les objectifs scientifiques du projet sont d'ordre théorique, algorithmes et pratiques. En résumé s'agit de développer une approche formelle générique de la classification croisée, permettant de bien figurer dans le challenge Netflix.

Cela signifie que nous allons considérer de grandes masses de données avec énormément de valeurs manquantes. Dans le cas des données Netflix par exemple il s'agit de traiter les appréciations de 17500 spectateurs sur 450 000 films avec près de 95% de valeurs manquantes.

Pour arriver à un tel résultat, il nous faut nous attaquer à deux verrous scientifiques : celui de la modélisation en classification croisé et celui de la sélection de modèle dans ce cadre spécifique.

Bien que le problème de classification croisée ait déjà abordé dans les années 70, ce n'est que récemment qu'il a suscité un nouvel engouement motivé par les applications (notamment systèmes de recommandation et bioinformatique). Les questions fondamentales à résoudre concernent le type de modèle que l'on recherche, le choix des critères à optimiser et l'algorithmique à mettre en œuvre dans un souci d'efficacité. Une autre point concerne l'importance d'arriver à établir des ponts entre des approches jusque là déconnectés comme les méthodes spectrales, les méthodes de factorisation et les méthodes à base de modèle. Ces différentes approches devraient ainsi connectées pourvoir d'enrichir mutuellement. Le moyen pour y arriver consiste à considérer les modèles de mélange pour la classification et de les adapter au cas de classification croisée. Si la validation des modèles de mélange pose déjà des problème difficiles dans la pratique, la validation d'une classification croisée est encore plus difficile à formaliser. Cette question sera aussi abordée dans le projet.

Considérant le cas des données manquantes, le problème est apparenté à celui de la complétion de matrice connu pour être d'une complexité non polynômiale. Il nous faudra donc considérer des approximations et des hypothèses fortes sur la régularité de la solution recherchée.

Ces dernières décennies, de multiples contributions ont été apportées en méthodologie statistique. De nombreuses méthodes et des algorithmes divers sont disponibles dans les logiciels actuels d'apprentissage statistique. Aussi l'utilisateur de ces méthodes doit-il faire face au problème d'un choix pertinent et adapté à ces données et aux divers objectifs. Aujourd'hui la modélisation statistique procède généralement en deux étapes : utiliser une famille de méthodes plutôt qu'une seule et, ensuite, essayer de choisir celle qui s'accorde le mieux avec les données. La sélection de modèle est par conséquent un problème central, mais difficile, à la fois du point de vue théorique et pratique.

Les critères classiques de sélection de modèle, fondés sur des hypothèses souvent irréalistes, sont des critères de pénalisation de contraste minimum à pénalités fixées. Un de nos objectifs principaux est de fournir des critères de sélection de modèle efficaces dont les termes de pénalité sont liés aux données. Dans ce contexte, notre proposition a pour but d'améliorer la panoplie des critères de la sélection statistique de modèle dans ses aspects à la fois théorique et pratique. Du point de vue théorique, elle vise à fournir des critères de sélection bien fondés et, du

point de vue pratique, elle a pour dessein de traiter des problèmes réels et complexes tel que le choix du modèle de mélange le plus adapté aux données.

Pour chaque type de donnée, plusieurs modèles ont été proposées, la sélection de celui qui s'ajuste au mieux aux données est un problème difficile. En effet, les critères utilisés qui dépendent généralement de la taille des données, du nombre de paramètres dont celui des nombres des classes sont généralement sensibles aux degrés de mélanges des classes. Dans le cadre de ce projet, nous nous attelons à traiter ce problème.

Le produit final du projet sera une ensemble de programmes mettant en œuvre sur de brande masses de données les méthodes développées au long du projet, associé a trois démonstrateurs sur des problèmes réels : les données Netflix, une application de segmentation de clientelle en marketing avec nos collègues de Vienne et une application à l'analyse de texte.

1.4 Positionnement du projet

1.4.1 Positionnement par rapport au contexte scientifique

Filtrage collaboratif et systèmes de recommandation Ce domaine de recherche concerne la réalisation de systèmes de filtrage d'informations en fonction de l'avis d'utilisateurs. La première génération opérationnelle est liée aux travaux de sociétés comme Amazon, GroupLens ou Netflix. Mais les systèmes existant comme celui de la société Netflix peuvent encore être améliorés. Cette dernière a lancé un défi qui fédère aujourd'hui la recherche active dans le domaine. Parmi les équipes les plus en pointe on retrouve l'équipe de Toronto qui utilise une des machines de Boltzman, une équipe hongroise de l'Académie des Sciences qui travaille sur la factorisation de matrice, l'approche bayésienne utilisée à Stanford, une approche statistique à IBM et une méthode basée sur les plus proches voisins développée par une équipe de AT&T Labs. Ces équipes se sont retrouvées lors de la dernière édition de la *KDD cup* dans une session consacré au challenge Netflix. Aucune équipe française n'était représentée lors de ce workshop.

En France, certaines équipes travaillent sur une approche distribuée (multi agent) comme le projet MAIA du LORIA. Balázs Kégl du LRI utilise Adaboost pour un système de recommandation pour la musique. Enfin, le projet CADI (ANR 2007) avec lequel notre projet s'articule vise à évaluer les techniques de recommandation les plus avancées aujourd'hui. Notre projet est clairement lié à CADI car le LITIS participe aussi à ce projet. La démarche de ClasSel vient en appui à CADI pour mener une recherche plus fondamentale, plus à moyen terme sur ce type de problème, avec une approche pluridisciplinaire associant des partenaires spécialistes des statistiques, de l'analyse de données, de l'apprentissage et de l'informatique.

Classification croisée Selon la nature de l'application, on trouve différentes traductions pour le terme « classification croisée ». Dans le domaine du texte et des systèmes de recommandations, on parle de *coclustering*. En bioinformatique, on évoque plutôt le terme de *biclustering*. Enfin certains auteurs parlent de *two mode or two way clustering* de *bi dimensional clustering* ou encore de *subspace clustering*. Concernant le coclustering, parmi les équipes les plus actives, il faut citer celle de Inderjit S. Dhillon (département d'informatique de l'université du Texas à Austin) qui travaille sur des critères d'information et sur la divergence de Bregman. Dans le domaine de la bioinformatique, il existe des algorithmes disponibles comme Samba (Statistical-Algorithmic Method for Bicluster Analysis) du *Computational Genomics Lab* de l'institut Weizmann et RoBA (Robust Biclustering Algorithm) développé à l'université du Minnesota. Le groupe de bioinformatique de l'Université de Leuven utilise des réseaux bayésiens. Dans le domaine de la modélisation probabiliste, les travaux les plus avancés sont ceux du groupe de Milios (Faculty of Computer Science de l'université de Dalhousie).

Sélection de modèle Il n'existe pas, à notre connaissance, d'équipe travaillant spécifiquement sur le problème de la sélection de modèle pour la classification croisée. En revanche, certaines techniques usuelles ont été adaptées. Du coté des « errements de la pratique » les méthodes de rééchantillonnage sont très populaires dans la communauté *machine learning*. Mais elles ne sont pas satisfaisantes leur passage à l'échelle nécessitant des approximations les rendant sous-optimales (voir à ce sujet le numéro spécial de JMLR sur la sélection de variables).

L'approche bayésienne permet le calcul analytique d'un critère. Elle a été utilisée pour la classification dans les travaux de Marc Boullé de France Télecom et par de nombreuses équipes dans le monde : en Amérique de Nord à l'université de Toronto et au département de statistiques de l'université de Carnegie Mellon. Ce n'est pas notre point de vue.

En France, le problème de sélection de modèle est étudié par des statisticiens dont le représentant emblématique est le projet Select de l'INRIA dirigé par P. Massart. Cette école étudie ce problème à partir d'inégalités de concentration qui ont une portée très générale car elles s'appliquent à toute distribution. L'inconvénient de leur grande généralité est leur tendance à être pessimistes et, partant, à sélectionner des modèles trop simples. Notre approche est différente lorsque nous spécifions la famille des distributions comme étant à symétrie sphérique. Ainsi nous prenons un risque inverse, mais, dans les cas nombreux où notre approche est adaptée, cette manière

d'aborder le problème sera optimale. Même dans le cas où ces hypothèses ne sont pas vérifiées, on observe souvent une certaine robustesse des outils analytiques déduits.

Enfin, une équipe en pointe dans le domaine est celle de M. Wells à Cornell avec laquelle D. Fourdrinier collabore depuis plus de quinze ans (Fourdrinier and Wells, 1994) et qui interviendra en appui de ClasSel.

1.4.2 Positionnement par rapport à l'appel à projet

Notre projet est de type académique. Il vise à trouver comment structurer des données en travaillant simultanément sur les individus et les variables. Il s'agit de transformer des données en connaissances pour de nouvelles applications (marketing, bio informatique, systèmes de recommandation). Nous nous situons donc dans le sous-thème « des données aux connaissances » de l'axe 2 « du signal à l'information, des données aux connaissances ». Mais on retrouve aussi de nombreux éléments de positionnement dans le sous-thème « du signal à l'information ». Ces éléments concernent des problématiques communes aux deux sous-thèmes : une approche formelle de l'« apprentissage », le traitement de masses de données et la prise en compte des contextes. Nous nous situons donc à l'intersection des deux sous-thèmes avec un centre de gravité plutôt du coté du sous-thème « des données aux connaissances ».

Notre stratégie scientifique consiste à attaquer de front les questions fondamentales de la modélisation en apprentissage et de la sélection de modèle pour trouver des solutions en rupture avec l'existant. Pour atteindre cet objectif, conformément à l'appel d'offre, nous proposons de mettre en œuvre une approche décloisonnée mobilisant des chercheurs de différentes communautés STIC (statistiques, analyse de données, apprentissage et informatique).

Dans le détail, les caractéristiques de notre défi en accord avec l'appel d'offre sont les suivantes.

- Nous proposons une approche formelle pour la modélisation et la **sélection de modèle** dans un cadre unifié.
- **L'apprentissage** est une forme de « programmation par l'exemple ». Il s'agit d'une approche intelligente, adaptative et évolutive qui vise à acquérir de manière dynamique et incrémentale de la connaissance. Il est symptomatique de souligner que parfois l'apprentissage est présenté comme une forme de compression des données. Ainsi des problématiques analogues ont vu le jour dans les deux communautés (traitement du signal et informatique) dont celle de la parcimonie qui nous intéresse dans ce projet. En effet, la parcimonie est un moyen permettant de traiter les grandes masses de données.
- Les problèmes de classification croisée et de sélection de modèles sont abordés avec le souci du passage à l'échelle. En effet, les applications visées traitent d'un **gros volume de données** (le *challenge* Netflix concerne les notes de 450 000 spectateurs pour 17500 films). En conséquence, la complexité des solutions proposées doit évoluer linéairement par rapport au volume de données à traiter. Cette contrainte pèse très fortement sur la manière d'aborder le problème.
- Un moyen permettant de faire face à ce volume de données est d'utiliser la notion de **contexte** (qui cependant reste à définir formellement). Dans le cadre de la classification croisée, nous proposons de développer des méthodes de découverte automatique de contexte dans le sens « groupes homogènes de données ».

1.5 Description des travaux : programme scientifique et technique

Tâche 0.1 : Outils et communication Les objectifs de cette tâche sont dans les trois premiers mois la mise en place du site web et le l'intranet du projet, des outils collaboratifs et de communication adaptés. Cette tâche sera consacré ensuite à la politique de publication et de valorisation des résultats. Il est notamment envisagé d'organiser un workshop international sur la question de la classification croisée et de la sélection de modèle.

Tâche 0.2 : Gestion du projet Cette tâche assure la gestion du projet. La structure opérationnelle d'animation prévue est un comité de pilotage chargé de définir les grands axes de travail et valider les étapes principales. Il aura pour mission le lancement du projet, son suivi et le reporting. Le comité de pilotage sera composé d'un responsable par partenaire et des responsables des tâches actives du projet.

1.5.1 Modèles pour la classification croisée

Tâche 1.1 : État de l'art Dans cette partie, nous allons faire une étude bibliographique du problème de classification croisée.

La classification croisée consiste à chercher une paire de partitions $(\mathbf{z}, \mathbf{w})$, où $\mathbf{z}$ est une partition de l'ensemble I des n individus en g classes et $\mathbf{w}$ est une partition de l'ensemble J des p variables en m classes, g et m étant connus ou inconnus. Une partition $\mathbf{z}$ est représentée par la matrice de classification $(z_{ik}; i = 1, \ldots, n; k = 1, \ldots, g)$ où $z_{ik} = 1$ si i appartient au k^{e} classe et 0 sinon. On adopte la même notation pour la partition des colonnes $\mathbf{w}$ représentée par $(w_{j\ell}; j = 1, \ldots, p; \ell = 1, \ldots, m)$. La classification croisée consiste à chercher une paire de partitions $(\mathbf{z}, \mathbf{w})$, où $\mathbf{z}$ est une partition de l'ensemble I des n individus en g classes et $\mathbf{w}$ est une partition de l'ensemble J des p variables en m classes, g et m étant connus ou inconnus. Une partition $\mathbf{z}$ est représentée par la matrice de

classification $(z_{ik}; i = 1, \ldots, n; k = 1, \ldots, g)$ où $z_{ik} = 1$ si i appartient au k^{e} classe et 0 sinon. On adopte la même notation pour la partition des colonnes $\mathbf{w}$ représentée par $(w_{j\ell}; j = 1, \ldots, p; \ell = 1, \ldots, m)$.

Pour aborder le problème de la classification croisée, en utilisant l'approche modèle de mélange, Govaert and Nadif (2003) ont proposé un modèle de mélange croisé dont la densité des données observées $\mathbf{x}$ s'écrit sous la forme $f(\mathbf{x}; \boldsymbol{\theta}) = \sum_{(\mathbf{z},\mathbf{w}) \in \mathcal{Z} \times \mathcal{W}} p(\mathbf{z}; \boldsymbol{\theta}) p(\mathbf{w}; \boldsymbol{\theta}) f(\mathbf{x}|\mathbf{z}, \mathbf{w}; \boldsymbol{\theta})$. Les ensembles $\mathcal{Z}$ et $\mathcal{W}$ représentent les ensembles de toutes les partitions possibles de I et de J. De plus, les probabilités associées aux partitions $\mathbf{z}$ et $\mathbf{w}$ sont supposées prendre les formes suivantes $p(\mathbf{z}; \boldsymbol{\theta}) = \prod_{i,k} \pi_k^{z_{ik}}$ et $p(\mathbf{w}; \boldsymbol{\theta}) = \prod_{j,\ell} \rho_\ell^{w_{j\ell}}$. Les $n \times p$ variables aléatoires x_{ij} sont supposées indépendantes sachant $\mathbf{z}$ et $\mathbf{w}$ fixés ; autrement dit nous avons $f(\mathbf{x}|\mathbf{z}, \mathbf{w}; \boldsymbol{\theta}) = \prod_{i,j,k,\ell} \varphi(x_{ij}; \alpha_{k\ell})^{z_{ik} w_{j\ell}}$ où $\varphi(., \alpha_{k\ell})$ une densité definie sur $\mathbb{R}$. Les paramètres $\pi = (\pi_1, \ldots, \pi_g)$ et $\rho = (\rho_1, \ldots, \rho_m)$ sont les proportions des classes et α est un paramètre qui dépendra de la situation étudiée.

Ce type de modèles peut être employé pour différents type de données. Par exemple, lorsque les données sont binaires, en considérant des mélanges de Bernoulli (Govaert and Nadif, 2003) plusieurs algorithmes peuvent être proposés. Une synthèse sur ces algorithmes issus de différentes approches (Estimation, classification et floue) est disponible dans (Govaert and Nadif, 2008). D'autre part lorsque l'information se présente sous forme de tableaux de contingence ou tableaux des co-occurrences. Le modèle de mélange dans ce cas repose sur l'hypothèse que les variables aléatoires x_{ij} sont distribuées suivant une loi de Poisson $\mathcal{P}(\mu_i \nu_j \alpha_{k\ell})$ avec μ_i et ν_j représentant les effets de de la ligne i et de la colonne j, $\alpha_{k\ell}$ désigne l'effet du bloc $k\ell$ (Nadif and Govaert, 2005). $\varphi(x_{ij}; \mu_i, \nu_j, \alpha_{k\ell}) = \frac{e^{-\mu_i \nu_j \alpha_{k\ell}} (\mu_i \nu_j \alpha_{k\ell})^{x_{ij}}}{x_{ij}!}$, $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{\alpha})$, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_g)$ et $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_m)$ sont les vecteurs des proportions des classes de $\mathbf{z}$ et $\mathbf{w}$, $\boldsymbol{\mu}$ et $\boldsymbol{\nu}$ sont les vecteurs composés des effets lignes et colonnes. Sous l'approche classification, les auteurs ont montré que lorsque les proportions sont égales, la maximisation de la vraisemblance classifiante est équivalente à la maximisation de l'information mutuelle et approximativement à la maximisation du critère $\chi^2(\mathbf{z}, \mathbf{w})$. Notons que, dans ce cas, l'analyse des correspondances binaire, reposant sur la métrique du χ^2, constitue un outil de visualisation à utiliser de manière complémentaire à la classification croisée.

Tâche 1.2. Modèles pour les données continues Lorsque les données sont de type continu, comme par exemple le degré d'expression des gènes dans la bioinformatique ou les données issues d'enquêtes de marketing, le modèle additif a souvent été utilisé $\mathbf{x} = \mathbf{zaw}' + \mathbf{e}$ où $\mathbf{a}$ est la matrice des centres de dimension $g \times m$ et $\mathbf{e}$ est l'erreur du modèle de dimension $n \times p$. La fonction objectif à minimiser s'écrit dans ce cas, si on considère une métrique euclidienne, $W(\mathbf{z}, \mathbf{w}, \mathbf{a}) = ||\mathbf{x} - \mathbf{zaw}'||^2$. Typiquement, ce critère a été employé par de nombreux auteurs en considérant différents algorithmes pour l'optimisation de W (Bock, 1979; Gaul and Schader, 1996; Rocci and Vichi, 2008). Par exemple, dans (Gaul and Schader, 1996) a été proposé l'algorithme l'algorithme *Alternating Exchanges* qui consiste à alterner la mise à jour de $\mathbf{a}$ après chaque transfert d'une ligne ou d'une colonne dans une classe minimisant la fonction Objectif. Par contre, dans l'algorithme *Croeuc* par Govaert (1983), la mise à jour est effectuée uniquement après le transfert des toutes les lignes ou de toutes les colonnes. Notons que *Croeuc*, contrairement à *Alternating Exchanges* et à *Two-mode k-Means*, travaille sur des matrices intermédiaires de tailles réduites, et par conséquent plus rapide. Contrairement aux données binaires ou aux tables de contingence où l'utilisation de notre modèle de mélange croisé qui est symétrique apparaît naturelle, dans le cas continu ce modèle ne semble pas approprié. En effet, pour les données binaires et de contingence, le modèle de mélange croisé proposé s'appuie sur l'hypothèse que les deux ensembles mis en correspondance dans le tableau de données peuvent être considérés comme des échantillons issus de deux populations. Cette hypothèse ne peut plus être en général maintenue pour les tableaux de variables continues où il est plus difficile de considérer l'ensemble des variables disponibles comme un échantillon.

Les objectifs de cette tâche seront donc les suivants :

- Proposer un modèle tel que, dans la situation la plus simple et pour la version classifiante associée, on retrouve exactement l'algorithme *Croeuc*.
- Développer l'approche « Estimation » associée à ce modèle.
- Etendre ce modèle à des situations plus générales prenant en compte par exemple des classes d'effectifs variables ou des variances pouvant être différentes suivant les classes d'individus ou de variables.
- Enfin, la dernière étape sera d'adapter ces modèles aux données de recommandation qui peuvent être vues comme des données continues un peu particulières.

Tâche 1.3. Méthodes factorielles Étant donnée une matrice de données X de dimension (n, p), l'analyse en composantes principales (ACP) peut être présentée comme une solution à trois problèmes de décomposition matricielle :

1. Recherche de la décomposition $\mathbf{cu}'$, où $\mathbf{c}$ est une matrice quelconque de dimension (n, ℓ) et $\mathbf{u}$ est une matrice formée de vecteurs orthonormés de dimension (p, ℓ), minimisant $||\mathbf{x} - \mathbf{cu}'||^2$;
2. Recherche de la décomposition $\mathbf{vd}$, où $\mathbf{d}$ est une matrice quelconque de dimension (g, p) et $\mathbf{u}$ est une matrice formée de vecteurs orthonormés de dimension (n, g), minimisant $||\mathbf{x} - \mathbf{vd}||^2$;

3. Recherche de la décomposition $\mathbf{vau'}$, où $\mathbf{a}$ est une matrice quelconque de dimension (g, ℓ) et $\mathbf{u}$ et $\mathbf{v}$ sont des matrices vérifiant les conditions indiquées précédemment, minimisant $||\mathbf{x} - \mathbf{vau'}||^2$.

En réalité, ces 3 problèmes n'en forment qu'un et l'ACP fournit une solution à ces trois problèmes : les matrices recherchées sont obtenues à l'aide de l'analyse spectrale de la matrice des données X. Ce résultat est souvent cité en ACP sous le nom de « formule de reconstitution des données ».

Par ailleurs, il est possible d'associer de manière canonique à une partition des variables un ensemble de vecteurs orthonormés ce qui permet de considérer que la classification des variables en m classes revient à faire une ACP sous contrainte de l'espace des individus et, de manière symétrique, que la classification des lignes revient à faire une ACP sous contrainte de l'espace des variables.

Si on ajoute ces contraintes aux matrices $\mathbf{u}$ et $\mathbf{v}$, les deux premiers problèmes reviennent alors à la recherche de partitions respectivement des variables et des individus minimisant le critère d'inertie intraclasse (Howard, 1969) et le troisième revient à minimiser, comme il a été dit précédemment, le critère utilisé dans la classification croisée (Govaert, 1983). Ainsi, alors que pour l'ACP, les trois problèmes étaient finalement équivalents, l'ajout des contraintes conduit à trois problèmes différents : classification des individus, classification des variables et classification croisée et la classification droisée optimalene correspondant pas nécessairement au produit de la classification optimale des individus par la classification optimale des variables.

Tâche 1.4. Classification croisée hiérarchique Il existe deux grands familles de méthodes de classification hiérarchique : une descendante, dite divisive, et une ascendante, dite agglomérative. La première, moins utilisée, consiste à partir d'une seule classe regroupant tous les objets, à partager celle-ci en deux. Cette opération est répétée à chaque itération jusqu'à ce que toutes les classes soient réduites à des singletons. La seconde qui est la plus couramment utilisée consiste, à partir des objets (chacun est dans sa propre classe), à agglomérer les classes les plus proches, afin de n'en obtenir plus qu'une seule contenant tous les objets. S'il est assez aisé de calculer une distance entre deux points, il est moins évident de calculer une distance entre une classe et un point, ou encore entre deux classes. Suivant le choix de cette distance dite critère d'agrégation on obtient plusieurs critères dont les plus couramment utilisés sont le critère du lien minimum, le critère du lien maximum, le critère du lien moyen et le critère de Ward qui résulte de la perte d'inertie en regroupant deux classes z_k et $z_{k'}$ et qui s'écrit : $\delta_{ward}(z_k, z_{k'}) = \frac{Card(z_k) \times card(z_{k'})}{Card(z_k) + Card(z_{k'})} \mathbf{d}^2(z_k, z_{k'})$. A une hiérarchie est associé un indice qui est une fonction strictement croissante et tel que pour toute classe singleton son indice est nul. Ainsi, pour les classes du bas de la hiérarchie l'indice vaut 0 et pour les autres classes, cet indice est défini en associant à chacune des classes construites au cours de la méthode la distance δ qui séparaient les deux classes fusionnées pour former cette nouvelle classe. Les critères d'agrégation cités précédemment conduisent bien à un indice, d'autres critères par contre présentent quelques difficultés. La complexité d'un tel algorithme est quadratique. Ceci nous restreint donc à l'application de cette méthode sur des tableaux de taille raisonnable. Dans un contexte de données de grande taille, un tel algorithme est assez peu utilisé ; on se contente souvent de l'appliquer sur des échantillons de l'ensemble des données ou encore sur des résumés des données obtenus avec une méthode de partitionnement. Cette méthodologie simple et efficace permet d'une part de classifier des données de grande taille et d'autre part de répondre au problème du nombre de classes. Une méthode mixte combinant $k$means et la hierarchie obtenue à l'aide du critère de Ward (Wong, 1982) est souvent utilisée.

En se plaçant dans un cadre de mélange, le résumé recherché correspond aux paramètres des densités des classes (composents du mélange). En effet, dans ce cas nous pouvons définir une distance entre deux classes z_k et z'_k à partir des vraisemblances classifiantes associées aux classes : $L_c(z_k, \theta)$, $L_c(z_{k'}, \theta)$ et $L_c(z_k \cup z_{k'}, \theta)$. Dans le cas des données continues, en assumant que le mélange est gaussien et que les proportions sont égales, nous pouvons retrouver le critère de Ward à partir de $d(z_k, z_{k'}) = L_c(z_k, \theta) + L_c(z_{k'}, \theta) - L_c(z_k \cup z_{k'}, \theta)$. En fait, ici les vraisemblances sont proportionnelles aux inerties intraclasses. En étendant ces travaux au cas de la classification croisée, nous avons proposé un algorithme de classification hiérarchique croisé appelé HBCM (Nadif et al., 2002) à partir du critère $Croeuc$ cité précédemment. L'algorithme HBCM alterne des classifications sur les lignes en fixant la partition des colonnes et puis celle des colonnes en fixant la partition des lignes. Un seul indice est asocié aux deux arbres hiérarchiques obtenues. Dans ce projet, nous proposons une extension de ce travail.

Les objectifs de cette tâche seront donc les suivants :

- Proposer des nouveaux critères d'agrégation en s'appuyant sur les vraisemblances classifiantes issues des modèles proposés : on traitera le cas des données continues, binaires et les tables de contingence.
- Développer des formules de recurrences ou à défaut des startégies pour l'accélération de ce type d'algorithmes.
- Evaluer le nombre de classes, décrire et visualiser des regroupements.
- Enfin, étendre la méthode mixte au cas croisé.

Tâche 1.5. Traitement des données manquantes Les données envisagées dans ce projet comportent généralement des données manquantes. Il s'agit d'une situation habituelle en statistique mais la caractéristique dans ce projet est que le nombre de données manquantes peut être très grand. La technique la plus souvent employée en

présence de données manquantes consiste à supprimer les lignes, ou les colonnes, du tableau de données de façon à travailler avec un tableau de données complet. Cette solution, simple et qui peut se justifier lorsqu'il y a peu de données manquantes est inutilisable dans les situations envisagées dans ce projet. Une autre technique heuristique consiste à remplacer les donnée manquantes par une valeur supposée « raisonnable » (*simple imputation* en anglais). La plus simple est de remplacer la valeur manquante par la moyenne. En classification, une procédure équivalente est de remplacer par la moyenne de la classe.

L'intérêt d'utiliser des méthodes de classification s'appuyant sur des modèles probabilistes est que ces modèles sont capable de prendre en compte ces données manquantes sans avoir besoin de les remplacer. Par ailleurs, ce type d'approche est très bien traitée si l'on utilise l'algorithme EM dont la caractéristique principale est de s'appuyer sur la notion de données manquantes. Les développements d'algorithmes utilisant ce type d'approche pour la classification simple s'appuyant sur les modèles de mélange ont montré une très grande efficacité. Nous proposons dans cette tâche d'étendre ces travaux en s'appuyant sur le modèle de mélange croisé.

Les objectifs de cette tâche seront donc les suivants :

- Faire une étude théorique sur la prise en compte des données manquantes dans la classification croisée.
- Étudier la stabilité des résultats et mettre en évidence l'intérêt de s'appuyer sur la classification croisée en comparaison à des méthodes de classification simple lorsque les données sont de grande taille.
- Étudier l'influence du taux de données manquantes, qui peut atteindre jusqu'à 95 %, sur la stabilité des résultats.
- Étudier l'influence des données manquantes sur les critères de sélection de modèle.

1.5.2 Sélection de modèle

Tâche 2.1 : État de l'art Dans cette partie, nous allons faire une étude bibliographique du problème de la sélection de modèle.

Tâche 2.2 : Critères asymptotiques Cette tâche consiste à adapter les solutions existantes au cas de la classification croisiée et d'en étudier les propriétés.

La classification automatique et, en particulier la classification croisée, conduit au choix difficile mais fondamental du critère de classification et du nombre de classes. Placer la classification automatique sous l'angle des modèles probabilistes et, en particulier des modèles de mélanges, permet de proposer des solutions à ce type de problèmes en s'appuyant sur les méthodes et les outils développés dans un cadre statistique très général pour choisir la dimension d'un modèle. Une des approches les plus utilisées est alors l'utilisation de critères de vraisemblance pénalisée qui font un compromis entre la qualité d'ajustement et la complexité du modèle pour obtenir des modèles parcimonieux. On peut citer, par exemple, le critère AIC *(Akaike Information criterion)* qui s'appuie sur mesure de la distance de Kullback-Leibler et le critère BIC *(Bayesian Information criterion)* de Schwarz qui s'appuie sur la maximisation a posteriori de la probabilité du modèle. Toutefois, ces deux critères ne tiennent pas compte l'aspect classificatoire de la solution recherchée lorsque les modèles de mélanges sont utilisés dans un contexte de classification et un troisième critère, appelé ICL *(Integrated Completed Likelihood* reposant sur la vraisemblance intégrée permet en ajoutant un terme d'entropie de prendre en compte cet aspect. En pratique, pour les modèles de mélange classiques, les critères BIC et ICL se sont révélés beaucoup plus performants que le critère AIC, par ailleurs très largement utilisé pour d'autres types de modèle. En outre, le critère ICL s'est révélé plus robustes face à des écarts de modèle.

Le critère BIC et sa variante classificatoire ICL, reposent des résultats asymptotiques dépendant de la taille de l'échantillon. L'application aux modèles croisés posent donc un problème. En effet, pour ces modèles, la taille du problème est caractérisée à la fois par le nombre de lignes et le nombre de colonnes et la notion de taille de l'échantillon n'est pas directe. Les premiers résultats tentant de mettre en œuvre ce type de critères avec une taille d'échantillon défini de manière empirique n'ont pas été convaincants. L'objectif de cette tâche est donc d'étudier de manière théorique et expérimentale la mis en place de ces deux critères, d'en proposer éventuellement de nouveaux et de les comparer à d'autres approches et en particulier au critère AIC.

Tâche 2.3 : Sélection à partir d'un critère non asymptotique Cette tâche consiste à élaborer une méthode de séléction de modèles adapté et performante à partir d'une approche décisionelle.

Pour comparer ces divers critères, nous nous proposons de suivre (puis d'étendre, voir plus loin) l'approche décisionnelle proposée par Fourdrinier and Wells (1994). L'idée est d'interpréter chaque procédure de sélection comme un estimateur du coût de l'estimateur des moindres carrés $\hat{\beta}_I$ de β_I ; autrement dit, si λ est cet estimateur de coût, $\lambda(y)$ est une estimation de $||\hat{\beta}_I - \beta||^2$. On définit alors la procédure de sélection comme suit : le modèle I sélectionné sera celui qui minimise cette estimation. Il reste enfin à choisir un "bon" estimateur de coût. Une évaluation simple est donnée par le risque quadratique

$$\mathcal{R}(\lambda, \beta_I, \hat{\beta}_I) = E_{\beta_I}[(\lambda - ||\hat{\beta}_I - \beta||^2)^2]$$

où E_{β_I} désigne l'espérance par rapport à la loi. Un estimateur λ' est alors dit dominé par λ si, pour toute valeur de β_I, on a $\mathcal{R}(\lambda, \beta, \hat{\beta}_I) \leq \mathcal{R}(\lambda', \beta, \hat{\beta}_I)$, l'inégalité étant stricte pour au moins un valeur de β_I.

Le principe de cette approche de la sélection d'un modèle au travers de l'estimation de coût est alors simple. Si l'on connaissait le coût $||\hat{\beta}_I - \beta||^2$, on choisirait le modèle I de plus petit coût. Puisque, dépendant du paramètre β_I inconnu, il ne peut être appréhendé, on range les modèles au vu de l'estimation de leur coût : un bon modèle I est celui correspondant au plus petit coût estimé.

À cette étape, il nous faut préciser ce qu'est la loi sous-jacente à l'espérance E_{β_I}. Dans les problèmes de sélection de modèles, on souhaite généralement s'affranchir de toute hypothèse distributionnelle restrictive afin de ne pas être contraint par l'impossibilité de vérifier si la loi retenue est bien la "bonne" loi qui régit les observations. Cependant, dans de nombreuses modélisations, la loi normale est utilisée car elle fournit un cadre où des méthodes statistiques peuvent être élaborées (cette loi peut aussi avoir des justifications au travers du théorème de limite centrale). Ce qui importe alors ce sont les outils d'analyse créés et, même si l'on n'adhère pas vraiment à l'hypothèse gaussienne, on espère tout du moins que l'inférence qui s'en déduit est robuste (et quelquefois on le démontre).

Pour élaborer nos procédures de sélection, nous partirons d'un contexte distributionnel plus large que le cadre gaussien en supposant que la loi de l'erreur $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ est radiale (soit invariante par transformation orthogonale ; autrement dit la loi de y est à symétrie sphérique autour de $X_I\beta_I$). Cette hypothèse induit une possibilité de dépendance entre les composantes ϵ_i (qui ne sont indépendantes dans ce cadre que si la loi est normale). Surtout, en travaillant conditionnellement au rayon $R = ||y - X_I\beta_I||$, elle permet de mettre en évidence des propriétés qui se trouvent masquées dès lors que l'on se restreint à la loi normale. On est ainsi à même d'exhiber des estimateurs de coût nouveaux améliorant les estimateurs classiques, et donc de bonnes procédures de sélection. Bien entendu les réflexions faites ci-dessus en rapport au cadre gaussien restent valables dans notre approche sphérique ; on souhaite mettre en œuvre ces procédures de sélection sans condition distributionnelle particulière.

Cette approche a eu une première mise en œuvre par Fourdrinier and Wells (1994) qui ont montré que, par rapport aux estimateurs de coût de la forme $||y - X\hat{\beta}_I||^2 + \delta(I)$, les estimateurs $\alpha(I)||y - X\hat{\beta}_I||^2$ sont robustes dans la classe des lois à symétrie sphérique et que, de plus, un estimateur optimal dans cette classe (c'est-à-dire de risque minimum) est donné par

$$\lambda^*(Y) = \frac{p}{n-p+2}||y - X\hat{\beta}_I||^2 .$$

Surtout ils ont mis en évidence que cet estimateur optimal peut être lui-même amélioré en terme de risque par des estimateurs de la forme

$$\lambda(Y) = \lambda^*(Y) - ||y - X\hat{\beta}_I||^4 \gamma(X\hat{\beta}_I)$$

où γ est une fonction positive indiquant, relativement au carré de la somme des carrés résiduels, comment doit être corrigé $\lambda^*(Y)$. Un exemple typique d'une telle amélioration est obtenu pour

$$\gamma(t) = \frac{2(p-4)}{(n-p+4)(n-p+6)} \frac{1}{||t||^2} .$$

On voit que le facteur de correction

$$||y - X\hat{\beta}_I||^4 \frac{2(p-4)}{(n-p+4)(n-p+6)} \frac{1}{||X\hat{\beta}_I||^2}$$

est construit de telle sorte que, si le modèle I est inadapté, alors cette correction est grande et, qu'au contraire, s'il y a une bonne adéquation de ce modèle, $\lambda^*(Y)$ est peu modifié. En outre, ce nouvel estimateur λ, vu comme une procédure de sélection, est un sélecteur pénalisé dont la fonction de pénalisation dépend aussi des données et pas seulement de la dimension du modèle.

Comme élément d'appréciation de la qualité de la règle de sélection fondée sur λ, les simulations faites dans (Fourdrinier and Wells, 1994) montrent que celle-ci l'emporte sur les critères de Mallows, de validation croisée et de validation croisée généralisée du point de vue des probabilités empiriques de bonne sélection de sous-modèles. C'est ce bon comportement qui donne une justification ultime à l'approche décisionnelle de la sélection de variables.

Cependant un inconvénient évident de l'application ci-dessus est que tous les sous-modèles (c'est-à-dire tous les sous-groupes de variables) doivent être examinés pour être comparés du point de vue de leur estimation de coût. Ce passage obligé est dû à l'utilisation brute de l'estimateur des moindres carrés $\hat{\beta}_I$ qui n'induit pas de direction privilégiée à la recherche des bonnes variables à sélectionner. Or Efron et al. (2004) proposent, au travers de l'estimateur LAR (pour Least Angle Regression), de sélectionner variable après variable suivant un chemin prédéterminé algébriquement. Ainsi tous les sous-ensembles de variables ne sont-ils pas visités. On peut donc naturellement envisager d'adapter l'approche décisionnelle décrite plus haut à l'estimateur LAR pour fournir un critère d'arrêt sur son chemin. Outre cet algorithme efficace sélectionnant chaque bonne variable l'une après l'autre, nous disposerions, par l'intermédiaire d'un minimum de l'estimation de son coût, d'un moyen d'appréhender quand il ne faut plus introduire de nouvelles variables.

La régression linéaire que nous avons envisagée est un cadre simple qui permet d'introduire notre approche décisionnelle de construction de procédure de sélection de variable. Notre but est de l'étendre aux problèmes plus généraux de sélection de modèle dans le cadre de la classification croisée. Pour ce faire nous étudierons d'abord le cas de la régression parcimonieuse et de la classification.

1.5.3 Algorithmes pour la classification croisée

Tâche 3.1. Choix des méthodes d'optimisation Le but de cette tâche est de faire le bilan des méthodes d'optimisation existantes pour extraire les techniques nous permettant une mise en œuvre efficace des méthodes proposées dans les tâches de classification croisées et de sélection de modèle. En quelque sorte, une fois que l'on a montré que la solution recherchée est donnée en calculant les p premières valeurs singulière de la matrice des données correctement pré traitées, il reste à trouver quel algorithme permet de calculer une approximation de ces valeurs singulières en un temps raisonnable sur des matrices creuses (*sparse*).

Pour ce faire nous envisageons d'étudier les approches de type « gradient stochastique » appliqué à des critères convexes conçus pour permettre de calculer efficacement une bonne approximation de la solution. Les méthodes de décomposition stochastique seront aussi étudiées car elles sont très efficaces, simples à mettre en œuvre et il existe maintenant un cadre formelle à leur utilisation.

Tâche 3.2. Mise en œuvre des méthodes Parmi les différentes difficultés liées à la mise en œuvre des méthodes développées dans ce projet, les objectifs de cette tâche seront d'étudier de manière détaillée les problèmes liés à l'initialisation des algorithmes et à celui de la gestion des classes vides.

- Initialisation : l'optimisation des critères de vraisemblance ou de de vraisemblance classificante effectuée dans les méthodes de classification croisée s'appuyant sur des modèles probablistes nécessite le recours à des algorithmes itératifs et l'existence de nombreux maxima locaux des fonctions de vraisemblance peuvent entraver leur performance. L'algorithme EM ou ses variantes fournissent généralement un maximum local fortement dépendant de la position initiale. Pour limiter cette dépendance, il est courant d'effectuer plusieurs essais de EM au hasard pour ne conserver que le meilleur d'entre eux. Diverses stratégies (Biernacki et al., 2001) permettent d'améliorer cette stratégie simple mais restent tout de même coûteuses en temps de calcul et seront difficilement exploitables pour de gros tableaux de données. Pour résoudre cette difficulté, plusieurs solutions seront envisagées parmi lesquelles on peut citer les suivantes :
  - Sélectionner le tirage initial en utilisant par exemple les axes factoriels de l'ACP ;
  - Utiliser une étape stochastique (recuit simulé, algorithme SEM,...).

- Problème des classes vides : les algorithmes de classification automatique sont souvent confrontés au problème des classes qui se vident lors du déroulement de l'algorithme. Ce problème est d'autant plus important ici que les méthodes de classification croisée ont une tendance encore plus marquée à vider des classes. La solution actuellement retenue consiste à arrêter l'algorithme et à recommencer un nouvel essai. Cette stratégie simple est efficace mais coûteuse en temps de calcul ce qui la rend irréaliste pour les gros tableaux de données.
  Une première piste envisagée est d'étudier si les méthodes d'initialisation de l'algorithme à partir d'axes factorielles d'une ACP ne pourraient pas être une solution aussi à ce problème. Une seconde serait de s'appuyer sur des algorithmes de classification avec contrainte topologique comme les cartes de Kohonen. Il est en effet connu que, dans le cas de la classification simple, les cartes de Kohonen permettent d'obtenir un grand nombre de classes non vides contrairement, par exemple, à l'agorithme des *k-means*.

Tâche 3.3. Intégration logicielle Le développement des méthodes de classification s'appuyant sur les modèles de mélange s'est traduit par l'existence de plusieurs logiciels libres et performants parmi lesquels on peut citer C.A.MAN, EMMIX, FLEXMIX, MClust, MULTIMIX, SNOB, Autoclass et MIXMOD. Ce dernier (http ://www-math.univ-fcomte.fr/mixmod/fr/index.php), dont l'un des participants de cette ANR est l'un des co-auteurs, est le résultat d'un partenariat entre l'INRIA (projet Select), le laboratoire de mathématiques de Besançon (UMR CNRS 6623), le laboratoire Heudiasyc de Compiègne (UMR CNRS 6599) et le laboratoire Paul Painlevé de Lille (UMR CNRS 8524). MIXMOD est un logiciel permettant de traiter des problèmes de classification (supervisée ou non) sur un ensemble de données par un modèle de mélange de lois. Ses usages sont multiples (fouille de données, reconnaissance des formes, décision statistique, ...) et les domaines d'utilisation très divers (biologie, analyse d'images, sciences sociales, ...) MIXMOD propose une grande variété d'algorithmes (EM, CEM, SEM, ...) pour estimer les paramètres d'un mélange. Il intègre plusieurs critères (BIC, ICL, ...) permettant de sélectionner le meilleur modèle parmi un large choix. Enfin, la possibilité d'enchaîner différents algorithmes et de choisir parmi plusieurs méthodes d'initialisation en font un outil à la fois souple et puissant pour traiter des problématiques de classification automatique et d'analyse discriminante. L'ajout de la composante classification croisée serait donc une contribution importante.

Afin d'assurer la diffusion des programmes qui seront développés dans ce projet, l'option retenue serait à terme de développer une version finale en C++ qui permettrait de les intégrer aussi bien dans le logiciel libre MIXMOD que dans le logiciel de statistiques R ou encore dans Matlab.

L'option minimum sera de proposer une bibliothèque de fonctions écrites en Matlab.

Tâche 3.4. Validation et test des algorithmes Il s'agit d'une tâche de benchmarking dont le but est de valider empiriquement les solutions proposées sur des problèmes artificiels montrant :

- leur capacité à passer à l'échelle et leur complexité empirique,
- leur capacité à traiter les valeurs manquantes,
- leur capacité à sélectionner les bon modèles,
- leur robustesse et leur résistance aux bruits.

Pour ce faire, des problèmes jouets spécifiques seront développées sur lesquels ont pourra jouer sur la taille des données, le taux de valeurs manquantes, le nombre de groupes et différents types de perturbations. Pour construire ce simulateur on utilisera un modèle génératif construit de manière interactive avec l'expérimentateur.

Le livrable de cette tâche comporte deux parties : le simulateur ainsi spécifié (les code source) et les résultats des différentes expérimentations.

Tâche 3.5. Visualisation La visualisation des résultats est un aspect important et souvent négligé des méthodes de classification. Pour la classification simple d'un tableau (n, p), la description des classes d'une partition des individus nécessite d'associer à chaque classe un vecteur de dimension p où p peut être très grand. Bien que les méthodes d'analyse factorielle soient très puissantes et contribuent efficacement à la visualisation des données, les grands échantillons nécessitent de nouvelles méthodes mieux adaptées. En effet, les algorithmes de décomposition matricielle rencontrent leurs limites sur les grands tableaux numériques ; en outre, la construction de nombreux plans de projection, du fait des grandes dimensions, rend la tâche d'interprétation difficile pour recouper les informations disséminées sur ces plans. Finalement une grande quantité de données implique une grande quantité d'informations à synthétiser et des relations complexes entre individus et/ou variables étudiés. Il est alors possible, dans ce contexte, d'utiliser les cartes de Kohonen ou cartes auto-organisatrices (SOM) (Kohonen, 1997) qui sont des méthodes de classification automatique utilisant une contrainte de voisinage sur les classes pour conférer un sens topologique aux partitions obtenues. La carte auto-organisatrice originelle peut être vue comme une variante de l'algorithme des *k-means* intégrant une contrainte d'ordre topologique sur les centres. Malgré sa popularité, SOM présente des inconvénients d'ordre théorique (justification de la convergence) et pratique (absence de la fonction objectif à optimiser). Comme la taille croissante des ensembles de données permet une estimation pertinente de variables cachées synthétisant de manière interprétable l'information, les modèles génératifs sont devenus très utiles. L'algorithme GTM (*Generative Topographic Model*) (Bishop et al., 1998), qui n'a pas les inconvénients de SOM, est une méthode auto-organisatrice probabiliste basée sur un modèle gaussien. Une première extension de cet algorithme utilisant les modèles de mélanges de Bernoulli a donné des résultats encourageants pour la visualisation de données binaires (Priam and Nadif, 2006).

Les objectifs de cette tâche seront donc :

- Étendre l'utilisation de GTM mais en s'appuyant sur les modèles de mélanges croisés qui sont plus parcimonieux que les modèles de mélange classiques.
- Traiter les différents types de données binaires, continues et tables de contingence.
- Enfin, fournir une sortie simple et compréhensible des résultats obtenus.

1.5.4 Applications

Tâche 4.1. Marketing L'université de Vienne (WU Wien) possède un département qui travaille sur la problématique du tourisme, notamment en Autriche. Le problème est d'arriver à caractériser des groupes de touristes afin de mieux répondre à leurs attentes. C'est typiquement une tâche de filtrage collaboratif. Les données considérées sont les avis de touristes à propose de différents sites de la capitale.

Tâche 4.2. Netflix Le but de cette tâche est d'appliquer les algorithmes et les méthodes développées par le projet sur les données du challenge NetFlix.

Netflix est une société de location de vidéo qui cherche se doter d'un système de recommandation performant. Les clients de Netflix sont invités à donner une note de 1 à 5 sur les films qu'ils ont vus. Le prix Netflix est une compétition en cours qui vise à récompenser le premier algorithme capable d'améliorer le système de prédiction de note maison de 10% au sens d'un critère quadratique.

Tâche 4.3. Analyse de textes Dans l'analyse textuelle, on cherche à déceler des blocs homogènes lors du traitement d'un ensemble de documents croisant un ensemble de mots clefs. La classification croisée est nettement plus appropriée que la classification séparée. En effet sur des tables de données généralement de très grande

	Partenaires/Partners				Chronogramme / chemin critique (Timing diagram/ critical path)											
					Année 1 / Year 1				Année 2 / Year 2				Année 3 / Year 3			
	LITIS	Heudiasyc	CRIP 5	WU Wien[1]	t+3	t+6	t+9	t+12	t+15	t+18	t+21	t+24	t+27	t+30	t+33	t+36
Tache 0.1																
Tache 0.2																
Tache 1.1																
Tache 1.2																
Tache 1.3																
Tache 1.4																
Tache 1.5																
Tache 2.1																
Tache 2.2																
Tache 2.3																
Tache 3.1																
Tache 3.2																
Tache 3.3																
Tache 3.4																
Tache 3.5																
Tache 4.1																
Tache 4.2																
Tache 4.3																
Livrables /Jalons *Deliverables/Milestones*								J1				J2				J3
Rapport d'avancement / état des dépenses																
Accord de consortium																
Rapport de synthèse																

TAB. 1 – Chronogramme / chemin critique. [(1)]WU Wien est un partenaire associé au projet et non un partenaire explicite. Les informations le concernant sont données à titre indicatif.

taille, la caractérisation *simultanée* des classes de documents par des classes de mots clefs est plus riche. Ces données peuvent être de type binaire (présence ou absence d'un mot clef dans un document) ou encore chaque valeur correspond à occurrence d'un mot clef dans un document (tableau de contingence). Ce type de données présente plusieurs centres d'intérêt pour ce projet : grande taille de données, matrices creuses, présence de données manquantes et le besoin de la visualisation des documents et des mots. En considérant l'approche mélange croisé, le modèle possonien pour les tables de contingence et le modèle de Bernoulli pour les données binaires apparaissent bien adaptés aux données textuelles. Les algorithmes de classification et de visualisation développés à partir de ces modèles, serviront pour des expérimentations sur des données réelles souvent utilisés dans la littérature et disponibles sur les suivants :

http ://www.cs.utexas.edu/users/dml/Software/gmeans.html
http ://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html
http ://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.html
http ://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/
http ://dblp.uni-trier.de/xml/

1.6 Résultats escomptés et Retombées attendues

La valorisation des résultats attendus et les connaissances à diffuser sera effectuée à travers des publication scientifiques et l'organisation d'un workshop.

Les retombées scientifiques pourront être évaluées à travers le nombre de publications liées au projet et la position de notre proposition dans le challenge Netflix.

Pour finir une remarque sur la création d'emploi. A court terme, ce n'est pas dans les objectifs du projet mais souvenons nous que Google a commencé par une thèse. . . C'est la rupture technologique proposée dans cette thèse qui a été à l'origine de l'entreprise que l'on connaît aujourd'hui. Dans le domaine des technologies de l'information, une méthode pour créer la rupture consiste à s'attaquer aux questions fondamentales qui font la difficulté des problèmes. C'est l'ambition principale de notre projet.

1.7 Organisation du projet

TABLEAU des LIVRABLES et des JALONS (le cas échéant)/ Deliverables and milestones			
Tâche/ Task	Intitulé et nature des livrables et des jalons/ Title and substance of the deliverables and milestones	Date de fourniture nombre de mois à compter de T0 / Delivery date, in months starting from T0	Partenaire responsable du livrable/jalon/ Partner in charge of the deliverable/ milestone
0 - Administration du projet			
	0.1 Site WEB et outils collaboratifs	T0+3	LITIS
	0.2 Gestion du projet	T0+36	LITIS
1 - Modèles pour la classification croisée			
	1.1 Etat de l'art	T0+3	LITIS
	1.2 Modèles pour les données continues	T0+12	Heudiasyc
	1.3 Méthodes factorielles	T0+18	LITIS
	1.4 Classification croisée hiérarchique	T0+24	Heudiasyc
	1.5 Traitement des données manquantes	T0+18	CRIP5
2 - Sélection de modèle			
	2.1 Etat de l'art	T0+3	LITIS
	2.2 Aspects asymptotiques	T0+18	Heudiasyc
	2.3 Méthodes non asymptotiques	T0+24	LITIS
3 - Algorithmes pour la classification croisée			
	3.1 Choix des méthodes d'optimisation	T0+18	LITIS
	3.2 Mise en oeuvre des méthodes	T0+24	CRIP5
	3.3 Intégration logicielle	T0+30	Heudiasyc
	3.4 Validation et test des algorithmes	T0+36	LITIS
	3.5 Visualisation	T0+30	CRIP5
4 - Démonstrateurs			
	4.1 Marketing	T0+30	LITIS avec (WU Wien[2])
	4.2 Netflix	T0+30	LITIS
	4.3 Textes	T0+36	CRIP5

TAB. 2 – TABLEAU des LIVRABLES. [(2)]WU Vien est un partenaire associé au projet et non un partenaire explicite. Les informations le concernant sont données à titre indicatif.

Le tableau 1 récapitule les responsables de chaque tâche et les partenaires impliqués. Les jalons scientifiques et/ou techniques, les principaux points de rendez-vous ainsi que les revues de projet prévues sont synthétisé dans le tableau 1. Le tableau 2 présente l'ensemble des livrables du projet.

1.8 Organisation du partenariat

Nous présentons un groupe projet composé de trois partenaires : le LITIS qui coordonne le projet, le laboratoire de l'UTC Heudiasyc et le CRIP5 de l'université de Paris 5. Nous avons explicitement associé à ce projet l'université de Vienne (WU Wien) avec laquelle LITIS collabore déjà sur ce type de problème. Un projet bilatéral sera déposé sur ce sujet en Mai 2008. L'université de Vienne ne demande aucun financement dans le projet et apparaît plus pour mémoire. La personne clé à Vienne est le professeur Kurt Hornik. Il possède une grande expérience notamment des approches de type combinaison estimateur (schéma de vote) dans le domaine de la classification. Il est aussi *R core member*et a ce titre il est à même d'aider le projet dans touts ces développements en R. Mais l'apport principal de WU Wien reste le problème de segmentation de marché évoqué dans la tache 4.1.

1.8.1 Pertinence du/des partenaires

coordinateur : LITIS (Participants au projet : Dominique Fourdrinier-PR, Alakin Rakotomamonjy-PR, Stéphane Canu-PR et Gilles Gasso-MC) Dominique Froudrinier est coordinateur du Projet. C'est un statisticien qui a une grande expérience du problème de sélection de modèle. Alakin Rakotomamonjy, Stéphane Canu et Gilles Gasso effectuent leurs recherches dans le domaine de la théorie de l'apprentissage. A ce titre, S. Canu est notamment coordinateur inter réseaux dans le réseaux d'excellence Pascal 2 qui à pour objet la recherche en reconnaissance des formes et l'apprentissage.

Partenaire 1 : Heudiasyc (Participants au projet : Gérard Govaert-PR et Yves Grandvalet-CR)
Responsable : Govaert Gérard, 59 ans
Formation :
- Doctorat d'État es-sciences de l'Université Paris 6 (1983),
- Doctorat de 3e cycle de l'Université Paris 6 (1975),
- Maîtrise et DEA d'informatique à l'Université Paris 6 (1972),
- École Normale Supérieure de Cachan (1968),
- École Normale d'instituteurs (1963).

Situation :
- Professeur classe exceptionnelle à l'université de Technologie de Compiègne
- Titulaire de la prime d'encadrement doctoral et de recherche
- Membre de l'UMR CNRS 6599 Heudiasyc
- Membre du conseil de branche du département Génie Informatique
- Membre du réseau d'excellence PASCAL

Coordonnées :
- Adresse personnelle : 1, avenue du maréchal Juin, 60200 Compiègne
- Adresse professionnelle : UTC, BP 20529, 60205 Compiègne cedex
- gerard.govaert@utc.fr, http://www.hds.utc.fr/~ggovaert

Coopération actuelle :
- Convention de recherche avec l'INERIS (projet SIGFRIED) : étude de l'exposition des populations et des risques liés à l'environnement extérieur et intérieur sur les déces par cancer.
- Co-auteur du logiciel Mixmod, logiciel d'estimation de modèles de mélange de lois de probabilité pouvant être utilisé dans un objectif de classification automatique ou de discrimination. Ce logiciel, distribué sous la licence GPL et disponible sous plusieurs systèmes (Linux, Unix, Windows), a été développé conjointement par l'INRIA futurs (action SELECT), le laboratoire de mathématiques de Besançon et le laboratoire Heudiasyc de Compiègne (www-math.univ-fcomte.fr/mixmod).

Liste des cinq publications les plus significatives des cinq dernières années :

1. Govaert, G. and Nadif, M., An EM algorithm for the Block Mixture Model, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27, 4, pp. 643-647, 2005.
2. Govaert, G. and Nadif, M., Fuzzy Clustering to estimate the parameters of block mixture models, Soft Computing, 10, 5, 415-422, 2005
3. Biernacki, C. and Celeux, G. and Govaert, G. and Langrognet, F., Model-based cluster and discriminant analysis with the MIXMOD software, Computational Statistics and Data Analysis, 51, 587-600, 2006
4. Samé, A. and Ambroise, C. and Govaert, G., A Classification EM Algorithm for Binned Data, Computational Statistics and Data Analysis, 51, 466-480, 2006
5. Govaert, G. and Nadif, M., Clustering of contingency table and mixture model, European Journal of Operational Research, 183, 1055-1066, 2007
6. Samé, A. and Ambroise, C. and Govaert, G., An online Classification EM algorithm based on the mixture model, Statistics and Computing, 17, 3 209-218, 2007
7. Govaert, G. and Nadif, M., Block clustering with Bernoulli mixture models : Comparison of different approaches, Computational Statistics and Data Analysis , 52, 3233-3245, 2008

Partenaire 2 : CRIP5 (Participants au projet : Mohamed Nadif-PR et François-Xavier Jollois-MCF)
Responsable : Mohamed Nadif, 44 ans
Formation :
- Habilitation à diriger des recherches, Université Paul Verlaine-Metz (2004),
- Doctorat en en Informatique, Université Paul Verlaine-Metz (1991),
- DEA en Mathématiques appliquées, Université Paul Verlaine-Metz (1986),
- DEUG-Licence en mathématiques-Maîtrise en ingénierie de mathématiques (1982-1986),

Situation actuelle :
- Professeur des Universités seconde classe à l'université Paris Descartes
- Responsable de l'équipe Apprentissage du CRIP5
- Membre de la mission d'évaluation de la recherche à l'université Paris Descartes
- Membre de la commission de sélection à l'université Paris Descartes

Coordonnées :
- Adresse personnelle : 5, rue du JURA, 75013 Paris
- Adresse professionnelle : Université Paris Descartes, CRIP5, UFR mathématiques-Informatique, 45 rue des Saints Pères, 75260 Paris Cedex 06

– mohamed.nadif@univ-paris5.fr, http://www.math-info.univ-paris5.fr/~nadifmoh/

Liste des cinq publications les plus significatives des cinq dernières années :

1. Govaert, G. and Nadif, M., Block clustering with Bernoulli mixture models : Comparison of different approaches, Computational Statistics and Data Analysis , 52, 3233-3245, 2008
2. Jollois, F-X. and Nadif, M., Speed up EM algorithm for categorical data, Journal of Global Optimization, 37, 513-525, 2007
3. Govaert, G. and Nadif, M., Clustering of contingency table and mixture model, European Journal of Operational Research, 183, 1055-1066, 2007
4. Govaert, G. and Nadif, M., An EM algorithm for the Block Mixture Model, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27, 4, pp. 643-647, 2005.
5. Govaert, G. and Nadif, M., Fuzzy Clustering to estimate the parameters of block mixture models, Soft Computing, 10, 5, 415-422, 2005

Liste des invitations dans des conférences internationales des cinq dernières années :

1. M. Nadif and G. Govaert. Block Clustering and Statistical modeling. Symposium on mixture modeling with special interest to applications in educational measurement and bioinformatics, Leuven, Belgium, 2007.
2. M. Nadif and G. Govaert. Dimensionality reduction via block clustering. 21st European Conference on Operational Research, Reykjavik, Iceland, july 2-5, 2006.
3. M. Nadif and G. Govaert. A review on block clustering under the mixture approach. IFCS Conference data science and Classification, Ljubljana, Slovenia, july 25-29, 2006.
4. G. Govaert and M. Nadif. Clustering of contingency table with Poisson mixture model. In Classification and Data analysis, editors Sergio Zani and Andrea Cerioli, Parme, Italy, pages 101-104, 2005.

Prix Vedior Bis

– M. Nadif et F.X. Jollois. Accélération de EM pour données qualitatives : étude comparative de différentes versions, présenté au congrès EGC 2004. http ://www.polytech.univ-nantes.fr/associationEGC/vedior.html

1.8.2 Complémentarité des partenaires

Le LITIS apporte des compétence en statistique (sélection de modèle), apprentissage et reconnaissance des formes. Il participe au projet Cadi qui lui confère une bonne vision pratique des problèmes. Heudiasyc apporte une compétence en analyse des données et classification croisée partagée avec le CRIP5. Heudiasyc apporte aussi avec Y. Grandvalet une expertise dans le domaine de l'apprentissage. Le CRIP 5 apporte en plus le poids de son expertise informatique.

Au sein du LITIS, la collaboration entre statistiques et apprentissage est récente. Elle s'exerce dans sur des problèmes réels de sélection de variables dans les modèles de régression linéaire où les deux compétences informatique et statistique forment une synergie. Cette complémentarité entre équipes apprentissage machine et statistique est manifeste : l'équipe statistique développe de nouveaux estimateurs au travers de méthodes de « shrinkage » qui sont mis en œuvre sur des applications réelles par l'équipe apprentissage machine.

La collaboration entre le CRIP5 et Heudiasyc, à travers celle de M. Nadif et G. Govaert, est de longue date. Ces dernières années, elle a été fructueuse et a permis de placer les travaux de Govaert (1983) sur la classification croisée dans un cadre probabiliste. Ainsi, les travaux menés, à partir de nouveaux modèles de mélanges, ont permis d'une part de donner un sens aux différents critères métriques proposés, d'autre part de proposer de nouveaux critères et de nouveaux algorithmes plus efficaces qui ont donné des résultats très encourageants. Cette modélisation a suscité beaucoup d'intérêts dans la communauté scientifique et a été traduite par plusieurs invitations dans des conférences internationales.

Les Liens entre le LITIS et Heudiasyc sont de longue date et attesté par les publications jointes de Y. Grandvalet, A. Rakotomamonjy et S. Canu.

1.8.3 Qualification du coordinateur du projet

Coordinateur du projet : Dominique Fourdrinier, 56 ans

Habilitation à diriger des recherches, Université de Rouen, 1994.
Statistique Mathématique « Sur l'estimation de coût »
Jury :

– James O. Berger, Professeur, Purdue University (rapporteur)
– Denis Bosq, Professeur, Université Paris 6, (rapporteur)
– William E. Strawderman, Professeur, Rutgers University (rapporteur)
– Marc Hallin, Professeur, Université libre de Bruxelles

– Claude Dellacherie, Directeur de Recherche CNRS, Université de Rouen

Situation actuelle : Professeur des Universités 1ère classe, Université de Rouen et Adjonct Professor, Cornell University, depuis novembre 2006

Expérience professionnelle :
– 01/01/2000-15/01/2006 : Directeur du Laboratoire de Mathématiques Raphaël Salem, Université de Rouen
– 01/09/1996-31/08/2003 : Professeur 2ème classe, Département de Mathématiques, Université de Rouen
– 01/08/1992-31/08/1996 : Maître de Conférences, Département de Mathématiques, Université de Rouen
– 01/02/1992-31/07/1992 : Professeur invité, Cornell University
– 01/01/1988-31/01/1992 : Maître de Conférences, Département de Mathématiques, Université de Rouen
– 01/06/1983-31/12/1987 : Assistant, Département de Mathématiques, Université de Rouen
– 01/09/1980-31/05/1983 : Chercheur sur contrats avec des entreprises privées, Département de Mathématiques, Université de Rouen
– 01/01/1980-31/08/1980 : Maître-Assistant, Département de Mathématiques, Université de Oran, Algérie
– 01/09/1978-31/08/79 : Assistant, Département de Mathématiques, Université de Oran, Algérie

S. Canu et

Liste des cinq publications les plus significatives des cinq dernières années :

1. Fourdrinier, D., Strawderman, W. E. and Wells W. T. Estimation of a location parameter with restrictions or "vague information" for spherically symmetric distributions. The Annals of the Institute of Statistical Mathematics, 58, 73–92, 2006.
2. Fourdrinier, D. and Pergamenchtchikov S. M. Improved model selection method for a regression function with dependent noise. The Annals of the Institute of Statistical Mathematics, 59, 435–464, 2007.
3. Fourdrinier, D., Kortbi, O. and Strawderman, W. E. Bayes minimax estimators of the mean of a scale mixture of multivariate normal distributions. Journal of Multivariate Analysis, 99, 74–93, 2008.
4. Fourdrinier, D. and Lepelletier, P. Estimating a general function of a quadratic function. The Annals of the Institute of Statistical Mathematics, 60, 85–119, 2008.
5. Fourdrinier, D. and Strawderman, W. E., Generalized Bayes minimax estimators of location vector for spherically symmetric distributions. Journal of Multivariate Analysis, à paraître, 2008.

Nombre total de publications dans les revues internationales et actes de congrès à comité de lecture : 42

Expérience de coordination de projet : Dominique Fourdrinier à été coordinateur des deux projets suivants :

1. Projet CNRS/NSF, « Sélection de Variables » avec Martin T. Wells (Cornell University), 01/01/1998 - 31/12/2000 (20 k€).
2. Projet Université de Rouen-Université de Tomsk (Russie) « Les Méthodes Séquentielles et les Méthodes d'Estimation Améliorée », 01/10/2003 - 31/12/ 2003 (2 k €).

Enfin, nous nous proposons d'établir un accord de consortium.

2 Justification scientifique des moyens demandés

2.1 Coordinateur : Le LITIS

2.1.1 Equipement

Pas d'achat d'un coût supérieur à 4000 €

2.1.2 Personnel

Personnels non permanents : Un ingénieur de recherche financé sur 26 mois (26 × 3470 €= 90,2 k€) permettra d'assurer une partie des recherches, et notamment les aspects applicatifs sur les données Netflix. Les stagiaires auront pour mission la réalisation des documentation et la mise en oeuvre des composants logiciels prévus par le projet.

– Un ingénieur de recherche — Coût prévisionel : 90,2 k€
– deux stagiaitres — Coût prévisionnel : 12 k€

Coût prévisionel total : 102,2 k€

2.1.3 Prestation de service externe

Pas de recours prévu à des services externes

2.1.4 Missions

Les frais de mission incluent les voyages en france pour rencnttrer les partenaires et la participation à deux conférences internationales pour effectuer la dissémination des résultats du projet. Coût prévisionnel : 5 k€

2.1.5 Autres dépenses de fonctionnement

Petit matériel, consomables, reprographie, affranchissement, téléphone : Coût prévisionnel : 5 k€
matériel informatique pour les chercheurs : Trois ordinateurs portables et logiciels Coût prévisionnel : 6 k€

Coût prévisionnel total : 11 k€

2.2 Partenaire 1 : Heudiasyc

2.2.1 Equipement

Pas d'achat d'un coût supérieur à 4000 €

2.2.2 Personnel

Personnel non permanent : Un ingénieur de recherche financé sur 15 mois permettra d'assurer une partie des recherches, et notamment la mise en œuvre des méthodes prévues par le projet et leur intégration logicielle. Les stagiaires auront essentiellement pour mission la mise en place des expérimentations et des applications.

- Un ingénieur de recherche cout prévisionel : 52,05k €
- deux stagiaires coût prévisionnel : 10k €

Coût prévisionnel total : 62,05 k€

2.2.3 Prestation de service externe

Pas de recours prévu à des services externes

2.2.4 Missions

Les frais de mission incluent les voyages en france pour rencontrer les partenaires et la participation à plusieurs conférences internationales pour effectuer la dissémination des résultats du projet. Coût prévisionnel : 13 k€

2.2.5 Autres dépenses de fonctionnement

Petit matériel, consommables, reprographie, livres, affranchissement, téléphone : Coût prévisionnel : 5 k€
Matériel informatique pour les chercheurs : deux stations de travail, un portable et logiciels Coût prévisionnel : 12 k€

Coût prévisionnel total : 17 k€

2.3 Partenaire 2 : CRIP5

2.3.1 Equipement

Pas d'achat d'un coût supérieur à 4000 €

2.3.2 Personnel

Personnels non permanents : Un ingénieur de recherche financé sur 12 mois (12×3470 €= 41,64 k€) permettra d'assurer une partie des recherches, et notamment la mise en œuvre et l'intégration logicielle des méthodes prévues dans le projet ainsi que l'aspect visualisation des données et des résultats. Les stagiaires auront essentiellement pour mission la mise en place des expérimentations sur les différentes applications de la tâche 4. L'évaluation des algorithmes et l'étude expérimentale à travers des simulations de Monte Carlo nécessiteront des stages de master 2 d'une durée moyenne de 6 mois.

- Un ingénieur de recherche coût prévisionnel : 41,64 k€
- Trois stagiaires (1 stagiaire par an) coût prévisionnel : 15 k€

Coût prévisionnel total : 56,64 k€

2.3.3 Prestation de service externe

Pas de recours prévu à des services externes

2.3.4 Missions

Les frais de mission incluent les voyages en france pour rencontrer les partenaires et la participation à des conférences internationales afin de pouvoir assurer la présentation des résultats du projet.

Coût prévisionnel : 12 k€

2.3.5 Autres dépenses de fonctionnement

Petit matériel, consommables, reprographie, livres, affranchissement, téléphone : Coût prévisionel : 5 k€
Matériel informatique : 2 PC, 1 imprimante, 2 portables et logiciels : Coût prévisionnel : 12 k€

Coût prévisionnel total : 17 k€

2.3.6 Organisation de *Workshop*

Organisation de 2 workshops internationaux sur le thème du projet afin de permettre une large diffusion.

Coût prévisionnel total : 10 k€

3 Annexes : description des partenaires

3.1 Le LITIS

Le Laboratoire d'informatique, du traitement de l'information et des systèmes (LITIS) est l'unité de recherche dans le domaine des sciences et technologies de l'information de Haute Normandie. Il implique les trois principaux établissements d'enseignement supérieur de la région : l'Université de Rouen, l'Université du Havre et l'Institut National des Sciences Appliquées (INSA).

Le laboratoire développe des démarches cohérentes pour mieux comprendre et maîtriser la nature de « l'information » et de son utilisation contextuelle. Les recherche protent à la fois sur des aspects théoriques, algorithmiques et sur la mise en oeuvre systèmes sensibles au contexte, du capteur à la base de donnée. Le LITIS structure ses recherches autour de trois axes qui organisent sept équipes de recherche : l'axe « Combinatoire et algorithmes » qui aborde les aspects formels de l'informatrion, l'axe « Traitement des masses de données » qui associe les équipes « Document et apprentissage », « Traitement de l'information et vivant », « Imagerie médicale » et « Systèmes de transport intelligents », et l'axe « Interaction et systèmes complexes » composé des équipes « Modélisation, interactions et usages » et « Réseaux d'interactions et intelligence collective ».

La démarche du LITIS est résolument pluridisciplinaire, associant praticiens et théoriciens à la jonction de l'informatique, de la reconnaissance des formes, du traitement du signal et des images de la médecine et des mathématiques, tous associés dans de nombreux projets.

3.2 Heudiasyc

Heudiasyc est une U.M.R. C.N.R.S. de l'Université de technologie de Compiègne. Un des deux axes du domaine DI (Décision et Image) concerne le développement de cadres théoriques et d'outils pour l'analyse de données, l'apprentissage, le raisonnement et la décision. L'équipe rassemblé autour de cet axe se distingue par la volonté d'aborder les cadres probabiliste et non probabiliste, dont elle s'efforce de démontrer la complémentarité - dans des contextes pratiques difficiles (données multidimensionnelles, hétérogènes, incertaines, incomplètes, structurées). Elle s'est fait connaître par sa contribution à l'utilisation des modèles de mélanges en classification automatique et par l'extension au cadre probabiliste des techniques de classification croisée dont elle avait été un des promoteurs les plus précoces.

3.3 Le CRIP 5

Le Crip5 (Centre de Recherche en Informatique de Paris 5) est né du regroupement de plusieurs équipes de recherche. Il a deux principaux thèmes de recherche : Intelligence Artificielle (IA) et Signal Parole Image Réseaux (SPIR). Le Crip5 fait partie du Pôle de compétitivité Cap Digital de l'île de France. Dans le pôle IA et dans l'équipe Apprentissage, l'approche Modèle de mélange à la fois pour la classification automatique simple et la classification croisée offre un cadre probabiliste intéressant pour aborder différents problèmes d'apprentissage. Dans l'équipe, cette approche est utilisée principalement pour la fouille de données et l'extraction des connaissances, la réduction de la dimension et la visualisation à partir d'un ensemble de données de grande taille.

Références

Akaike, H. (1970). Statistical predictor identification. *Annals of the Institute of Statistical Mathematics*, 22 :207–217.

Allen, D. M. (1971). Mean square error of prediction as a criterion for selecting variables. *Technometrics*, 13 :469–475.

Biernacki, C., Celeux, C., and Govaert, G. (2001). Assessing a mixture model for clustering with the integrated classification likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7) :719–725.

Bishop, C. M., Svensén, M., and Williams, C. K. I. (1998). Developpements of generative topographic mapping. *Neurocomputing*, 21 :203–224.

Bock, H. (1979). Simultaneous clustering of objects and variables. In , E., editor, *Analyse des Données et Informatique*, pages 187–203. INRIA.

Bock, H. (2003). Two-way clustering for contingency tables maximizing a dependence measure. In Schader, M., Gaul, W., and Vichi, M., editors, *Between Data Science and Applied Data Analysis*, pages 143–155. Springer Heidelberg.

Cheng, Y. and Church, G. (2000). Biclustering of expression data. In *ISMB2000, 8th International Conference on Intelligent Systems for Molecular Biology*, pages 93–103, San Diego, California.

Craven, P. and Wahba, G. (1978). Smoothing noisy data with spline functions : estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, 31 :377–403.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm (with discussion). *Journal of the Royal Statistical Society*, B 39 :1–38.

Dhillon, I. (2001). Co-clustering documents and words using bipartite spectral graph partitioning. In *Seventh ACM SIGKDD Conference*, pages 269–274, San Francisco, California, USA.

Dhillon, I., Mallela, S., and Modha, D. (2003). Information-theoretic co-clustering. In *Proceedings of The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD-2003)*, pages 89–98.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32 :407–451.

Foster, D. P. and George, E. I. (1994). The risk inflation criterion for multiple regression. *The Annals of Statistics*, 22(4)(4) :1947–1975.

Fourdrinier, D. and Wells, M. T. (1994). Comparaisons de procédures de sélection d'un modèle de régression : une approche décisionnelle. *Comptes Rendus de l'Académie des Sciences*, 319(Série I) :865–870.

Gaul, W. and Schader, M. (1996). *Data Analysis and Information Systems : Statistical and Conceptual Approaches*, chapter A New Algorithm for Two-Mode Clustering. Springer.

Govaert, G. (1977). Algorithme de classification d'un tableau de contingence. In *First international symposium on data analysis and informatics*, pages 487–500, Versailles. INRIA.

Govaert, G. (1983). *Classification croisée*. Thèse d'état, Université Paris 6, France.

Govaert, G. and Nadif, M. (2003). Clustering with block mixture models. *Pattern Recognition*, 36 :463–473.

Govaert, G. and Nadif, M. (2005). Classification d'un tableau de contingence et modèle probabiliste. *RNTI-E-3, Revue des Nouvelles Technologies de l'Information*, 1 :213–218.

Govaert, G. and Nadif, M. (2006). Fuzzy clustering to estimate the parameters of block mixture models. *Soft Computing*, 10(5) :415–422.

Govaert, G. and Nadif, M. (2008). Block clustering with bernoulli mixture models : Comparison of different approaches. *Computational Statistics and Data Analysis*, 52 :3233–3245.

Hartigan, J. A. (1975). *Clustering Algorithms*. Wiley, New York.

Hocking, R. (1976). The analysis and selection of variables in linear regression. *Biometrics*, 32 :1–49.

Howard, N. (1969). *Quantitative ecological analysis in the social sciences*, chapter Least squares classification and principal component analysis : a. comparison. M.I.T Press.

Kohonen, T. (1997). *Self-organizing maps*. Springer.

Madeira, S. C. and Oliveira, A. L. (2004). Biclustering algorithms for biological data analysis : A survey. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 1(1) :24–45.

Mallows, C. (1970). Some comments on cp. *Technometrics*, 15(4) :661–675.

Marchotorchino, F. (1987). Block seriation problems : A unified approach. *Applied Stochastic Models and Data Analysis*, 3 :73–91.

Massart, P. (2007). *Concentration inequalities and model selection*. Lecture notes in mathematics, Vol. 1896, École d'été de probabilités de Saint-Flour XXXIII.

McLachlan, G. J. and Basford, K. E. (1988). *Mixture Models, Inference and applications to clustering*. Marcel Dekker, New York.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. Wiley, New York.

Miller, A. J. (1990). *Subset selection in regression*. Chapman and Hall, New York.

Nadif, M. and Govaert, G. (2005). Block clustering of contingency table and mixture model. In *LNCS 3646, Advances in Intelligent Data Analysis VI*, pages 249–259. Springer.

Nadif, M., Govaert, G., and Jollois, F.-X. (2002). A hybrid system for identifying homogenous blocks in large data sets. In *Second Euro-Japanese Workshop on Stochastic Risk Modelling for Finance, Insurance, Production and Reliability, 16-19 septembre, Chamonix, France*, pages 324–333.

Priam, R. and Nadif, M. (2006). Carte auto-organisatrice probabiliste sur données binaires. *RNTI (EGC'2006 proceedings)*, pages 445–456.

Rissanen, J. (1986). A predictive least squares principle. *IMA Journal of Mathematical Control and Information*, 3 :211–222.

Rocci, R. and Vichi, M. (2008). Two-mode multi-partitioning. *Computational Statistics and Data Analysis*, 52 :1984–2003.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2) :461–464.

Thomson, M. L. (1978a). Selection of variables in multiple regression : Part i. a review and evaluation. *International Statistical Review*, 46 :1–19.

Thomson, M. L. (1978b). Selection of variables in multiple regression : Part ii. choosen procedures, computations and examples. *International Statistical Review*, 46 :129–146.

Van Mechelen, I., Bock, H., and De Boeck, P. (2004). Two-mode clustering methods : a structured overview. *Statist. Methods Medical Res.*, 13(5) :363–394.

Vichi, M. (2000). Double k-means clustering for simultaneous of objects and variables. In Borra, S. e. a., editor, *Advances in Classification and Data Analysis*. Springer Berlin.

Vichi, M. and A.L., K. H. (2001). Factoriel k-means analysis for two-way data. *Computational Statistics and Data Analysis*, 37 :49–64.

Wong, M. A. (1982). A hybrid clustering method for identifying high-density clusters. *J. Am. Statis. Assoc.*, 77 :841–847.

Zhang, P. (1993). Model selection via multifold cross validation. *The Annals of Statistics*, 21 :299–313.