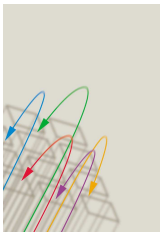
BELGIAN SCIENCE POLICY



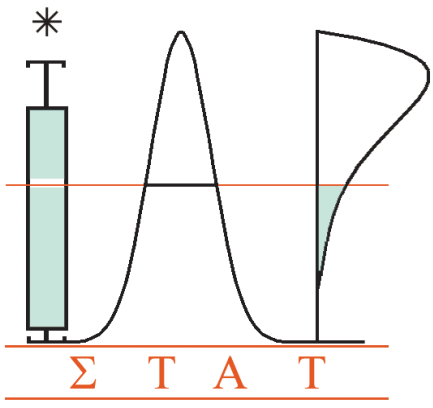
Eindrapport 2007-2011

IUAP-netwerk in Statistiek

Contract P6/03

10 mei 2012

http://www.stat.ucl.ac.be/IAP/PhaseVI



Inhoudsopgave

1 Inleiding

Het onderzoeksnetwerk bestaat uit 5 Belgische partners en 4 Europese partners, die vermeld staan in Tabel 1.

Afkorting	Partner
UCL	Université catholique de Louvain
KUL-1	Katholieke Universiteit Leuven 1
KUL-2	Katholieke Universiteit Leuven 2
UG	Universiteit Gent
UH	Universiteit Hasselt
UJF	Université Joseph Fourier
EMC	Erasmus Medical Center
USC	Universidad de Santiago de Compostela
LSHTM	London School of Hygiene and Tropical Medicine

Tabel 1: *Belgische en Europese partners in het netwerk.*

Het onderzoeksproject is opgebouwd rond 5 werkpakketten. Tabel 2 vermeldt de voornaamste partners voor elk werkpakket en geeft per werkpakket aan welke partner het werk coördineert.

Werkpakket	Partners
WP1: Multivariate data met kwalitatieve beperkingen	UCL, KUL-1*, UH, UJF, EMC, USC
WP2: Tijd- en ruimte-gerelateerde data	UCL*, KUL-1, UG, UJF, EMC, USC
WP3: Onvolledige data	UCL, KUL-1, KUL-2, UG, UH*, USC, LSHTM
WP4: Data met latente heterogeneteit	UCL, KUL-1, KUL-2*, UH, UJF, EMC, USC
WP5: Hoog-dimensionele en samengestelde data	UCL, KUL-1, KUL-2, UG*, UH, UJF, EMC, USC

Tabel 2: *Voornaamste partners per werkpakket, en de coördinerende partner per werkpakket (aangeduid met een *).*

De belangrijkste doelstellingen van de vijf werkpakketten kunnen als volgt samengevat worden :

- WP1 : Het ontwikkelen van performante statistische technieken voor het schatten van ongekende (mogelijks multidimensionele) functionele grootheden en dit in een flexibele modeleringscontext, met bijzondere aandacht voor functionele grootheden waaraan kwalitatieve eisen worden opgelegd.
- WP2 : Het bestuderen en modelleren van tijd- en ruimte-gerelateerde data met nadruk op niet- en semiparametrische efficiente methoden, online schatting en automatische model selectie onder niet-standaard condities zoals niet-stationariteit en heterogeneiteit.
- WP3 : Het ontwikkelen van methodologie voor ontbrekende en/of gecensureerde waarnemingen, vaak in de context van complexe datastructuren. De propagatie van dergelijke methodologie in een aantal toepassingsgebieden.
- WP4 : Het modelleren van heterogeneiteit voor complexe datastructuren, en het op punt stellen van inferentie methoden voor deze modellen. De nadruk ligt op het bestuderen van gemengde modellen

met partieel gespecifieerde afhankelijkheid van de residu's, en gegeneraliseerde lineaire gemengde modellen.

- WP5 : Het ontwikkelen van specifieke statistische technieken om de essentiele informatie op een wetenschappelijk verantwoorde manier uit hoog-dimensionele gegevens te distilleren. Deze technieken worden ontwikkeld in de context van bio-informatica en chemometrie.

In de hierna volgende subsecties beschrijven we kort de resultaten die bekomen zijn in de verschillende werkpakketten. De eerste doelstelling van de onderzoeksactiviteiten van het netwerk was het verwezenlijken van de hierboven beschreven objectieven. Naast deze objectieven hebben leden van het netwerk ook belangrijke bijdragen geleverd in andere gebieden, nauw verwant met de objectieven van het project.

Voor meer details betreffende de onderzoeksresultaten en publicaties per werkpakket en per jaar, verwijzen we naar de jaarrapporten en het overzichtsrapport van het netwerk. Deze kunnen gedownload worden van de webpagina van het netwerk :

`http://www.stat.ucl.ac.be/IAP/PhaseVI`

De webpagina bevat eveneens gedetailleerde informatie betreffende andere netwerkactiviteiten, zoals workshops, meetings, short courses, werkgroepen, seminaries, enz.

2 Module de travail 1: Données multivariées sous des contraintes qualitatives

2.1 Frontières et analyse d'efficacité et de productivité

D'intenses activités de recherche ont été développées dans les domaines de l'estimation du support d'une fonction de densité (univariée ou multivariée), de l'estimation d'un support plus général, ainsi que de l'estimation de la frontière du support.

L'estimation de supports généraux et l'estimation de la densité d'un support univarié (basée sur des données (non-) contaminées ont été étudiées dans plusieurs travaux. L'estimation d'une fonction de frontière a été effectuée dans deux cadres: de modèles de frontière déterministe et de modèles de frontière stochastique.

2.1.1 Modèles de frontière déterministe

La théorie asymptotique des estimateurs par enveloppement des données (Data Envelopment Analysis-DEA) a été établie et des procédures de bootstrap ont été développées, ceci incluant des résultats théoriques aussi bien que des algorithmes. Une théorie complète du cas multivarié dans l'estimation de frontière est maintenant disponible.

L'utilisation des estimateurs par enveloppement soulève un autre défi, celui de la robustesse aux valeurs extrêmes et aux observations atypiques (ou aberrantes- outliers). Deux concepts de frontières partielles ont été introduits et étudiés (tant sur le plan théorique et que sur le plan de l'utilisation pratique) par des membres du réseau. Dans les années récentes on a obtenu de nouvelles perspectives sur les relations entre ces différents concepts de frontières partielles. Ces perspectives ont été exploitées plus avant dans l'estimation de modèles de frontières de coût. Des membres du réseau ont aussi trouvé des liens prometteurs entre l'estimation de frontières et la théorie des valeurs extrêmes et les ont exploités davantage.

2.1.2 Modèles de frontière stochastique

Dans cette classe de modèles, le défi consiste à être capable d'identifier le bruit de l'inefficacité lorsqu'on évalue la performance des firmes. Dans un cadre complètement paramétrique (fonction de frontière paramétrique et hypothèses paramétriques fortes pour les distributions du bruit et de l'inefficacité) on a montré comment des techniques basées sur le bootstrap sont utiles pour améliorer l'inférence relative aux performances d'efficacité individuelle. Dans un cadre complètement nonparamétrique, les modèles stochastiques ne sont pas identifiés. Dès lors, un ensemble minimal d'hypothèses est à rechercher en vue d'obtenir des modèles plus flexibles que les modèles paramétriques restreints. Une première tentative dans cette direction était basée sur des méthodes de maximum de vraisemblance local. L'introduction de bruit dans des méthodologies DEA/FDH a maintenant été réalisée dans un cadre très général.

Le problème des modèles de frontières stochastiques peut en fait se voir comme un problème de déconvolution. Un grand nombre de recherches ont été réalisées par des membres de ce réseau (et de l'antérieur) sur la déconvolution de densité et de régression. Plusieurs avancées importantes ont été réalisées dans ce domaine de recherche.

2.2 Estimation nonparamétrique et semiparamétrique de courbes et de surfaces (irrégulières), et estimation sous contraintes qualitatives

En utilisant, comme ingrédient de base, un ajustement linéaire local, on a développé des techniques d'estimation simples et cependant puissantes en vue de montrer des irrégularités de différents types pour des courbes et/ou des surfaces. Plusieurs thèses de doctorat ont été rédigées sur ce sujet. Les propriétés théoriques de ces méthodes ont été établies, et leur performances ont été illustrées dans différents domaines d'application.

Une autre approche pour traiter de l'inférence pour des courbes avec irrégularités consiste à utiliser des techniques de régression avec pénalisation au moyen d'un choix approprié de la fonction de pénalité. On a proposé un cadre unifié pour l'estimation nonparamétrique de régression dans des modèles linéaires généralisés en utilisant des fonctions splines pénalisées avec pénalité non-quadratique. Un type attrayant de pénalités dans le contexte des B-splines est donné par le type différence de pénalités. Des études théoriques détaillées ont révélé l'impact de l'ordre de l'opérateur de différenciation (une partie de ce travail a aussi reçu une distinction de meilleur article). La puissante technique des P-splines a aussi été utilisée comme ingrédient de base dans des procédures de sélection de variables pour des modèles additifs et des modèles à coefficients variables; elle a été aussi utilisée dans de nombreux autres problèmes tels que celui de l'estimation d'une densité bivariée. Dans ce contexte, assez bien d'attention a aussi été consacrée à l'analyse Bayésienne.

A part l'estimation d'une fonction de moyenne, on s'est aussi intéressé à l'estimation d'une fonction de variance, ou plus généralement d'une fonction de dispersion. Plus particulièrement, dans l'analyse de données sur-dispersées ou sous-dispersées l'estimation d'une fonction de dispersion est importante. Plusieurs contributions relatives à l'estimation de fonctions de dispersion ont été réalisées. Une procédure développée pour l'estimation de moyenne et de dispersion et basée sur les approximations P-splines a été utilisée dans l'analyse de plusieurs données, incluant l'analyse de données d'avortement en Italie ; cette expérience a aussi révélé le besoin de procédures robustes. La recherche dans le réseau s'est aussi intéressée à des techniques de régularisation telles que le lasso, le lasso adaptatif et les techniques de régularisation groupée.

Des techniques de décomposition par ondelettes combinées avec une technique de pénalisation ont été utilisées pour estimer une fonction de régression monotone.

2.3 Procédures d'inférence nonparamétrique et semiparamétrique

Les modèles de régression localisation-échelle fournissent une jolie classe de modèles pour décrire l'influence de co-variables sur une variable d'intérêt. Plusieurs papiers de recherche ont exploité l'utilisation de modèles localisation-échelle afin de répondre à des questions spécifiques. Par exemple, dans des études médicales, les courbes ROC sont des outils utiles pour analyser la capacité de discrimination d'une variable de diagnostic. L'effet d'une co-variable sur une variable de diagnostic a été modélisé au moyen de modèles de régression localisation-échelle non-paramétrique. On a aussi analysé différentes applications pour tester les hypothèses d'une régression multiple non-paramétrique.

Les modèles à indice simple (single index) sont un autre outil utile pour modéliser l'influence de nombreuses co-variables sur une variable d'intérêt. Les modèles à indice simple ont été introduits dans l'analyse des tables de contingence.

On a aussi étudié les modèles du type transformation semi-paramétrique qui sont un autre outil de modélisation . Deux méthodes pour l'estimation du paramètre de transformation ont été proposées: maximiser une fonction de profil de vraisemblance ou minimiser l'erreur quadratique moyenne par rapport à l'indépendence.

Diverses contributions importantes dans le domaine du test de la qualité d'ajustement pour des modèles de régression paramétrique ont été fournies. Une première contribution est l'étude du choix d'un paramètre pour la largeur de bande qui soit approprié dans un tel cadre de test. Une seconde contribution consiste en une nouvelle approche de test en vue de tester si une fonction de régression appartient à une famille paramétrique de fonctions de régression; dans cette approche on mesure la distance entre les fonctions de distributions empirique des résidus paramétriques et non-paramétriques. Une troisème contribution concerne les tests de qualité d'ajustement pour des fonctions de régression paramétriques en présence d'erreurs de séries chronologiques. Un quatrième ensemble de contributions est relatif à des tests de vraisemblance empirique qui sont capables de tester la qualité d'ajustement pour une classe de modèles de régression paramétriques et semi-paramétriques. Un problème qui y est associé et qui a été aussi étudié est celui de la comparaison de fonctions de régression provenant de plusieurs populations et, en régression, du développement de tests de qualité d'ajustement pour la forme de la fonction de variance .

Au sujet de la vraisemblance empirique (EL: empirical likelihood) plusieurs contributions fondatrices ont été réalisées: (i) revue des méthodes EL pour la régression, en y incluant des modèles paramétriques, semi-paramétriques et non-paramétriques; (ii) prise en compte de données manquantes et de données censurées; (iii) méthode du jacknife EL dans l'inférence pour les fonctions copule; (iv) le champ de la méthodologie EL a été étendu dans plusieurs directions.

2.4 Données multivariées, analyse robuste et inférence statistique

La recherche développée dans le réseau a contribué à de nombreux domaines importants de l'analyse robuste multivariée: (i) des méthodes de classification pour des données asymétriques de petites et de grandes dimensions; (ii) des techniques de calcul efficace afin d'examiner en profondeur et de manière robuste les prédicteurs les plus pertinents dans une large classe de régresseurs candidats; (iii) des versions robustes d'algorithmes de sélection de variables selon des méthodes en-avant ou par-étape; (iv) l'étude de procédés de bootstrap rapides et robustes; (v) des procédés robustes pour l'estimation fonctions de moyenne et de dispersion dans des modèles de régression flexible.

2.5 Décompositions par ondelettes, analyse d'image et données fonctionnelles

La décomposition par ondelettes a été appliquée dans des cadres variés. Dans l'analyse statistique d'images, on a présenté une approche d'échelle multidimensionnelle pour la caractérisation statistique d'images fonctionnelles hétérogènes dans le temps et dans l'espace, telles qu'il s'en présente p.ex. dans des données médicales- MRI ou NMR- ou dans des données de satellites. Des techniques d'ondelettes ont aussi été appliquées pour estimer une densité de probabilité et une fonction cumulative de probabilité.

De par leur nature propre, les données fonctionnelles exigent une attention spécifique dans l'analyse statistique. On a établi la validité asymptotique d'un procédé de bootstrap naïf et sauvage dans le cas d'un modèle de régression fonctionnelle nonparamétrique avec réponse univariée et covariable fonctionnelle. De plus, on a introduit la généralisation de mesures d'association, telle le tau de Kendall, pour le cas de mesures d'association conditionnelle, étant données des covariables fonctionnelles.

Les méthodes développées pour des données fonctionnelles ont été appliquées dans une variété de domaines d'application: pour analyser des observations atmosphériques, pour analyser des données médicales fonctionnelles, pour une recherche efficace de protéines et de combinaisons de protéines qui peuvent être utilisées comme biomarqueurs dans un diagnostic de cancer ou dans des buts de pronostic.

2.6 Modélisation des dépendences, fonctions de copules et théorie des valeurs extrêmes

Parmi les contributions majeures du réseau dans la modélisation des dépendances ` travers les copules, il y a: (i) le développement de méthodes semi- et non-paramétriques pour des copules non-conditionnelles et conditionnelles; (ii) l'introduction de différentes mesures d'associations non-conditionnelles et conditionnelles; (iii) le développement de procédés de test pour tester des structures spécifiques de dépendance; (iv) modéliser des événements extrêmes au moyen de copules de valeurs extrêmes.

De nombreuses contributions au domaine de la théorie des valeurs extrêmes ont été fournies par des membres du réseau: (i) de nouvelles méthodes de réduction de biais pour l'estimation de l'indice de queue; (ii) estimation de l'indice de valeur extrême sous censure aléatoire; (iii) estimation basée sur les rangs des fonctions de dépendance de Pickand; (iv) un répertoire complet et convivial des queues de copules Archimédiennes, sous la forme d'un arbre de décision; (v) l'inférence statistique sur la dépendance de queue au moyen d'un attracteur de valeur extrême; (vi) l'utilisation de variation régulière et son extension à l'étude de comportement extrême, p.ex. quand on étudie des mesures de risque relatives à la queue, telles que la valeur-de-risque (value-at-risk).

2.7 Interaction avec d'autres Modules de travail

Le Module de Travail 1 (Workpackage 1 – WP 1) se concentre surtout sur le développement de techniques statistiques puissantes d'inférence pour différentes situations de données complexes rencontrées dans des applications. La plupart des sujets traités dans ce module de travail sont, comme tels, directement en relation avec des sujets d'autres modules. Une sélection de mots-clés à partir du résumé et liens avec les autres modules de travail: données censurées et/ou données dépendantes (WP 3 et 2); hétérogénéité (WP 4); sélection de variables (WP 5).

3 Module de travail 2: Données temporelles et spatiales

La plupart des partenaires belges (KUL-1, KUL-2, UCL, UG) et la plupart des partenaires européens (EMC, UJF, USC) ont contribué au WP 2.

Dans le cadre des données univariées corrélées, les principales contributions ont été réalisées pour le cas d'une non-stationarité due à des ruptures structurelles ou à des spectres variant dans le temps. Dans le cas de la stationarité, les nouveaux développements incluent l'inférence statistique pour des séries chronologiques avec spectres irréguliers, pour leurs valeurs extrêmes, pour des données censurées, pour des processus non linéaires et pour des processus en temps continu. Un aspect important en a été le développement et la recherche de modèles opérationnels pour des données complexes reliées temporellement dans les domaines financiers ou psychologiques. On a étudié des extensions à des données multivariées reliées dans le temps. De nouvelles méthodes sont développées dans le domaine de modèles multi-niveaux en temps continu, en y incluant l'inférence bayésienne. On a modélisé et analysé au point de vue statistique la corrélation au sein de séries multivariées. Les nouveaux développements concernent des modèles dynamiques de correélation conditionnelle, des extensions multivariées de modèles de volatilité et la modélisation de valeurs extrêmes dans des séries multivariées. De nouveaux procédés d'estimation ont été développés pour des cas de violation des hypothèses de stationnarité et d'homogénéité. Finalement, la détection et la modélisation de correélation spatiale ont été investiguées en utilisant, par exemple, des techniques spectrales. De nouveaux procédés d'estimation ont aussi été proposés en régression non-linéaire avec des données spatialement correélées.

Dans de nombreuses sciences, on rencontre des donnes reliées temporellement et spatialement et leur analyse statistique a une longue histoire. Et pourtant on rencontre encore des problèmes importants et non-résolus dans le développement et l'analyse de modèles pour des données corrélées temporelles ou spatio-temporelles sous des hypothèses non-standard, en temps discret ou continu, en une ou plusieurs dimensions. Les principaux résultats des recherches du réseau pour des données reliées temporellement et/ou spatialement peuvent être répartis dans les catégories suivantes: chroniques univariées complexes, chroniques multivariées, modèles en temps continu et données reliées dans l'espace.

3.1 Chroniques univariées complexes

Une source de complexité très souvent rencontrée en sciences appliquées est le non-respect de la stationnarité. Dans Van Bellegem et von Sachs (2008), un estimateur adaptatif ponctuel d'un spectre variable dans le temps a été proposé pour un modèle explicite basé sur des ondelettes; ce modèle de stationnarité locale permet, pour le spectre, de très soudains changements dans le temps. En présence de stationnarité locale, l'estimation du paramètre de mémoire longue variant dans le temps a été considérée dans Roueff et von Sachs (2011). D'autre part, en utilisant un ajustement linéaire/polynomial, Croux *et al.* dérivent des méthodes de prédiction robuste pour des chroniques non-stationnaires. Des chroniques à variations régulières dans des espaces de Banach ont été étudiées dans Meinguet et Segers (2010). Une caractéristique commune et source potentielle de non-stationnarité souvent rencontrées dans l'analyse des chroniques sont les ruptures de structure dans le temps. Des procédures non-paramétriques de test sont développées dans Gao *et al.* (2008) qui testent simultanément des ruptures structurelles dans les espérances conditionnelles et les variances conditionnelles. D'autre part, dans les données réelles, la complexité est souvent due à l'irrégularité de la densité spectrale d'une série temporelle stationnaire. Dans Fryzlewicz *et al.* (2008), on étudie une nouvelle approche pour l'estimation de seuil, par ondelettes, des densités spectrales de séries temporelles stationnaires. Desmet et Gijbels (2010) traite de l'estimation non-paramétrique d'une densité spectrale avec une amélioration de l'estimation des pics. Des techniques

d'ajustement linéaire local constituent aussi l'ingrédient de base pour l'estimation non-paramétrique de la fonction de volatilité dans Casas et Gijbels (2010), ce qui le relie au sujet étudié dans Gao *et al.* (2008). L'ajustement polynomial local (dans un contexte différent de celui des séries chronologiques) est parmi les méthodes étudiées en détail et appliquées dans différents contextes dans le WP1. Une technique alternative très populaire est l'estimation par ondelettes munies de seuil, considérée par exemple dans Freyermuth *et al.* (2010). Les séries chronologiques peuvent aussi être étudiées du point de vue de leurs plus grandes valeurs, les pointes, et de l'aggrégation temporelles de ces pointes en grappes. Robert *et al.* propose un nouvel estimanter pour l'indice d'extrémité, une mesure du degré de concentration des extrêmes. Johannes et Subba (2011) ont discuté de l'estimation de la densité et de problèmes associés en régression non-paramétrique pour des données dépendantes lorsque les observations ne proviennent pas nécessairement d'un processus linéaire. Antoniadis (2009) a étudié des séquences de variables aléatoires corrrélées avec périodicité, tandis que Guillotin-Plantard et Prieur (2009a, 2009b) ont donné des théorèmes de limite centrale sous des conditions de dépendance faible. Pour valider des modèles sur des données, la construction d'intervalles de confiance et de tests d'ajustement est également importante. Dette *et al.* (2009) présente un procédé pour tester la proportionalité entre la fonction de régression et la fonction d'échelle dans un modèle de régression non-paramétrique avec données dépendantes. El Ghouch *et al.* (2010) développe trois types d'intervalle de confiance pour une classe générale de fonctionnelles d'une fonction de survie basée sur des données dépendantes censurées. La censure de données de séries chronologiques est un autre problème parfois rencontré avec des données réelles. El Ghouch et Van Keilegom (2008,2009) considère un modèle de régression non-paramétrique pour lequel les données sont dépendantes et la réponse est sujette à une censure aléatoire à droite. Teodorescu *et al.* (2010) étudie des modèles conditionnels linéaires dépendant du temps avec troncature à gauche et censure à droite. Dans des modèles de régression non-paramétrique du type localisation-échelle, Heuchenne et Van Keilegom (2010) considère l'estimation basée sur des données censurées tandis que Lambert (2010) suppose des données censurées par intervalle. Antoniadis *et al.* (2010) étudie le sélecteur de Dantzig pour une régression avec des données censurées à droite. Pour une distribution relative à des données tronquées à gauche et censurées à droite, Molanez-Lopez *et al.* (2010) propose des intervalles de confiance pour la vraisemblance empirique lissée. Ces articles sont en relation étroite avec les sujets de recherche étudié dans WP1 et WP3.

Applications. De nombreux travaux réalisés dans le réseau se sont attachés à développer des modèles pour des données financières. Les modèles GARCH et à volatilité stochastique (SV)sont deux modèles concurrents bien-connus et d'utilisation fréquente pour expliquer la volatilité des séries financières. Pour un modèle de volatilité stochastique, Hafner et Preminger (2009, 2010) considèrent un estimateur en forme close, dérivent ses propriétés asymptotiques et proposent, en vue de comparer la capacité des modèles GARCH et SV, un ensemble de règles de décision simples et fortement consistantes. En utilisant des technique d'ajustement polynomial, Casas et Gijbels (2010) estime non-paramétriquement la fonction de volatilité. Une approche alternative a été développée par Monsalve-Cobis *et al.* (2011). Reboredo *et al.* (2010) développe et évalue de nouveaux algorithmes basés sur des modèles GARCH, des réseaux de neurones et des techniques de forçage (boosting techniques) élaborés en vue de modéliser et de prédire des séries chronologiques hétéroscédastiques. Un domaine où les séries chronologiques en temps discret surgissent naturellement est celui du test de capacité. Goegebeur *et al.* (2009,2010) développent un modèle probabiliste en vue de tester l'agilité (speediness) et investiguent s'il est possible de détecter, en utilisant des diagnostics d'influence locale, les personnes qui sont particulièrement sensibles à la pression du temps. Une autre ligne de travail a développé un algorithme qui surmonte la charge de calcul associée aux modèles markoviens latents. Rijmen *et al.* propose une méthode qui associe les modèles en question à

un graphe dirigé acyclique et qui applique à ce graphe des transformations afin d'arriver à un algorithme efficace d'inférence (voir aussi WP5). Dans des essais pré-cliniques et cliniques, on développe des études pharmacocinétiques pour analyser l'évolution, dans le temps, de la concentration du médicament dans le plasma. Jullion *et al.* (2009) propose un modèle non-paramétrique bayésien basé sur des P-splines, pour le cas de données raréfiées et sujettes à des bruits; ce cas comporte typiquement deux difficultés: le choix d'un modèle adéquat de comportement, étant donnés le faible nombre des données et le manque de convergence des méthodes non-linéaires.

3.2 Séries chronologiques multivariées

Un aspect important du projet de recherche est l'extension de l'analyse univariée au cas des séries chronologiques multivariées . Hafner et Reznikova (2010) considère l'estimation efficace d'un modèle dynamique de copule semi-paramétrique. Segers (2010) dérive la convergence faible du processus empirique de copule sous des hypothèses non restrictives concernant leur propriété d'être plus ou moins lisse. On s'est aussi intéressé au développement de méthodes de travail pour l'analyse de corrélation entre séries multivariées ainsi qu'à leur application à des données financières. Par exemple, dans le contexte des données d'assurance, Manner et Segers (2010) caractérise les queues de mélanges de corrélations de copules elliptiques. Genest et Segers (2010) étudie asymptotiquement la covariance du processus empirique de copule. Dans Hafner et Franses (2009), Hafner et Reznikova (2010) ou Hafner *et al.* (2010), on suggère une généralisation du modèle de corrélation conditionnelle dynamique de Engle (2002), ce qui permet d'introduire des corrélations spécifiques aux titres financiers et qui est utile, en particulier, si on a pour objectif de synthétiser un grand nombre de rendements de titres. On a aussi considéré des généralisations de résultats connus de la théorie asymptotique des modèles GARCH et de leur agrégation temporelle (voir Hafner (2009), Hafner et Preminger (2009a, 2009b), Hafner et Herwartz (2009)) . Les valeurs extrêmes de séries chronologiques multivariées peuvent révéler une dépendance entre coordonnées et dans le temps. Basrak et Segers (2009) introduit un processus de queue en vue de décrire et de modéliser de telles valeurs extrêmes et montre que la théorie est très facilement applicable aux solutions stationnaires de processus stochastiques autoégressifs avec matrice aléatoire de coefficients, un intéressant cas spécial étant un modèle GARCH à facteurs, récemment proposé. Dans le contexte des séries chronologiques multivariées, la violation des hypothèses de stationnarité et d'homogénéité pose d'importants problèmes. Böhm et von Sachs (2008, 2009) ou Böhm *et al.* (2010) considèrent des modèles de facteurs pour des données de panel dont le dimension de coupe transversale et de temps est grande et auquel on ajuste des séries chronologiques localement stationnaires pour lesquelles les composantes communes sont estimées par les vecteurs propres d'une matrice de densité spectrale estimée non-paramétriquement. Hafner et Reznikova (2010) considèrent la stationnarité locale dans le contexte des modèles multivariés de volatilité et des modèles non-linéaires avec dépendance et Omrane et Hafner (2009) introduit une nouvelle méthodologie d'impulsion de réponse pour analyser les effets, sur la volatilité des trois des taux de change les plus importants (Euro, Pound Sterling et Yen), des annonces de nouvelles macro-économiques des Etats-Unis.

3.3 Modèles en temps continu

Concernant les modèles en temps continu à un niveau, il y a des contributions traitant de la prise de décision. Les modèles dans ce secteur sont basés sur des processus de diffusion, qui sont des processus markoviens à valeur réelle et en temps continu. Vandekerckhove, J. et F. Tuerlinckx (2007, 2008) ont plongé le modèle de diffusion de Ratcliff dans un cadre statistique de régression et, conjointement avec ce travail théorique, ils ont développé une boite à outils Matlab librement déchargeable qui permet, aux

psychologues expérimentaux, de réaliser de façon commode des analyses du processus de diffusion de Ratcliff. Alors que les contributions précédentes se basaient sur un cadre fréquentiste, Vandekerckhove *et al.* (2008) ont aussi développé des outils d'inférence bayésienne pour le processus de diffusion de Ratcliff. Tuerlinckx (2010) a contribué à une analyse théorique de modèles basés sur la distribution de Weibull pour des données d'événements instantanés (time to event data) tandis que Oravecz et Tuerlinckx (2010) ont développé une extension à un modèle de processus d'Ornstein-Uhlenbeck multi-niveau, ou hiérarchique, ainsi qu'un algorithme d'inférence bayésienne. Vandekerckhove *et al.* (2010) présente un modèle hiérarchique de diffusion de Ratcliff et l'explique d'un point de vue psychométrique. Ceci a conduit à un cadre de modélisation qui, pour les chercheurs appliqués, est extrêmement flexible et commode à développer. Vandekerckhove *et al.* (2010) présente son application à une tâche accélérée de catégorisation sémantique. Cependant, pour la modélisation simultanée des différences entre individus et entre populations, une extensionn hiérarchique du modèle de base à espace d'états est nécessaire. Lodewyckx et al. (2011) introduit un modèle flexible hiérarchique bayésien avec des effets aléatoires pour les paramètres du système. Souvent un chercheur ne collecte pas les données avec un seul objectif, mais bien avec plusieurs objectifs. Il y a des liens clairs entre ce travail et WP4. Oravecz *et al.* (2009) a développé des modèles d'un processus d'Ornstein-Uhlenbeck multi-niveau, ou hiérarchique, ainsi qu'un algorithme d'inférence bayésienne, afin d'analyser simultanément des séries chronologiques pour différents sujets, multivariées et espacées irrégulièrement .

3.4 Données reliées dans l'espace

L'analyse de données spatiales a été un autre objectif important de ce module de travail. Crujeiras et Van Keilegom (2010) étudie les propriétés asymptotiques et d'échantillons finis d'un estimateur d'un modèle de régression non-linéaire lorsque les erreurs sont spatialement corrélées et lorsque la structure de dépendance spatiale est inconnue. Il existe un intérêt croissant pour améliorer le niveau des connaissances de processus spatiaux et spatio-temporels utilisant des techniques spectrales, voir Gonzàlez-Manteiga and Crujeiras (2011). Crujeiras *et al.* (2010) suggère d'étendre, à la densité spectrale spatiale, deux techniques différentes de tests d'ajustement. Dans Teodorescu *et al.* (2010), les auteurs étudient les propriétés asymptotiques et d'échantillons finis d'un estimateur d'un modèle de régression non-linéaire lorsque les erreurs sont spatialement corrélées et lorsque la structure de dépendance spatiale est inconnue. Crujeiras et Fernàndez-Casal (2010) étudient les propriétés asymptotiques d'estimateurs de densité spectrale par noyau non-paramétrique lissé pour la densité spectrale spatiale d'un processus spatial continu stationnaire dans un cadre asymptotique de rétrécissement. Gonzàlez-Manteiga *et al.* (2009) étudie les propriétés du périodogramme multidimensionnel aussi bien dans le cas de pincement (tapering) que de non-pincement, et sous l'hypothèse de dimension finie d'une grille (lattice) régulière où le processus spatio-temporel est observé. Crujeiras *et al.* (2010) étudie différentes techniques de tests d'ajustement pour la densité spectrale spatiale tandis que Crujeiras *et al.* (2009) considère la détection de la séparabilité dans la structure de la dépendance spatio-temporelle. Ruiz-Medina et Crujeiras (2011) et Crujeiras et Ruiz-Medina (2011) étudient des champs aléatoires gaussiens de fractal semi-paramétrique. Antoniadis *et al.* (2011) ou Gonzàlez-Quintela *et al.* (2011) fournissent des applications sur des données réelles.

4 Werkpakket 3: Onvolledige data

Onvolledige gegevens of, meer algemeen, ruwe gegevens, verwijst naar het probleem dat in ongeveer alle empirisch onderzoek bestaat: de waarnemingen zijn minder fijn dan wat gepland werd in het proefopzet.

Ruwe gegevens nemen verschillende vormen aan: (1) ontbrekende gegevens: herhaalde metingsreeksen worden vroegtijdig afgebroken; in surveys worden bepaalde items blanco gelaten of weigeren enkele gezinsleden hun medewerking; (2) stoptijden: de studie wordt afgebroken op een moment dat voor een aantal subjecten nog geen stoptijd werd waargenomen; dit is zogenaamde censurering; (3) groepering: gegevens worden gegroepeerd, waardoor geen eenheden maar afgeronde groepen worden waargenomen (bijv. het aantal pakjes gerookte sigaretten i.p.v. het aantal sigaretten; klassieke afronding); (4) niveaus gaan beneden de detectielimiet en kunnen dus niet waargenomen worden. Bij uitbreiding kunnen ook latente structuren als onvolledige gegevens gezien worden; dit omvat random effecten, latent klassen, latente variabelen, mengverdelingen, enz.

Het is evident dat deze situaties leiden tot problemen bij het analyseren van gegevens en dan aangepaste opzetten en analysemethoden moeten voorgesteld worden. Opzet verwijst dan voornamelijk naar de poging om onvolledige gegevens te voorkomen. Het is niet mogelijk om onvolledige gegevens volledig te vermijden, maar reduceren door goed gekozen steekproefopzet lukt meestal wel. Analyse verwijst naar het ontwikkelen van methoden die onvolledige gegevens desondanks toch kunnen analyseren, zij het onder aanname van assumpties. Tenslotte is er ook nog sensitiviteitsanalyse, een klasse van methoden die dient om na te gaan wat de impact van de onvolledige gegevens op de conclusies is.

Het consortium is actief geweest in al deze gebieden. We geven een overzicht.

4.1 Goede praktijken bij onvolledige gegevens

In heel wat toepassingsgebieden zijn te lang sub-optimale methoden gebruikt. Een voorbeeld is onvolledige gegevens in klinische studies. Hierbij zijn drie spelers betrokken: academici, de biofarmaceutische industrie en de regelatoren (Food and Drug Administration, European Medicines Agency). Er dient uiteraard een strikt kader te zijn, doch veranderingen aanbrengen is verre van evident door de inertie van het systeem. Leden van het consortium hebben uitgebreid bijgedragen, via papers, maar ook via voordrachten, cursussen, en een rapport in opdracht van de National Academy of Sciences (National Research Council) in de Verenigde Staten. Een lid van het consortium was het enige niet VS-lid van de werkgroep. De aanbevelingen worden ondertussen breed geïmplementeerd in studies. Een sterk punt is dat de klassieke tegenstellingen tussen de parametrische en niet-parametrische scholen overstegen werden in het rapport, wat er de authoriteit van verhoogde. Gelijkaardige beschouwingen gelden voor toepassingen in andere gebieden.

Het consortium is zeer actief geweest zowel op het vlak van de ontbrekende gegevens als de gecensureerde waarnemingen. Daarnaast is er ook heel wat werk geweest in data met latente (dus niet-geobserveerde) structuren. Voor dat laatste verwijzen we naar WP4.

4.2 Methodologische ontwikkelingen voor ontbrekende gegevens

In de context van ontbrekende gegevens werd uitgebreid gewerkt op het voorstellen van likelihood en Bayesiaanse methoden. Tegelijk werd er heel wat aandacht besteed aan semi-parametrische modellen. Deze laatste komt vaak samen voor met wegingsmethoden. De eerste klasse is flexibel en vrij makkelijk op het vlak van implementatie. Ze leidt ook tot efficiënte schatters. Het nadeel is dat ze gevoelig is aan modelstoringen. De semi-parametrische klasse is moeilijker uit theoretisch oogpunt, is minder efficiënt en niet altijd eenvoudig om te implementeren. De klasse is echter voordelig op het vlak van vertekening. Een derde klasse methoden is deze van multipele imputatie. Ook op dat laatse is het team actief. Er zijn bijvoorbeeld methoden voorgesteld voor optimale gewichten bij semi-parametrische methoden. Daarnaast is er aandacht gegaan naar de zogenaamde dubbele robuste methoden, niet enkel op het vlak van

generalized estimating equations (GEE), de meest frequent voorkomende methode semi-parametrische methode, maar ook voor pseudo-likelihood. Naast het gebruik van gewichten bij GEE, is het ook mogelijk de methode te combineren met multipele imputatie. Zo ontstaat een hybride methode, tussen de parametrische en de semi-parametrische in. Via uitgebreide simulaties en theoretische beschouwingen werd aangetoond dat de methode praktisch goed werkt. Deze methodologie werd voornamelijk ontwikkeld door UH, KUL-2, UG en LSHTM.

Een overzichstekst werd geschreven door Molenberghs (UH, KUL-2) en Kenward (LSHTM), en uitgegeven bij Wiley. Heel wat werk van de eerste helft van de voorbije periode omtrent het thema van ontbrekende gegevens komen in het boek aan bod.

4.3 Methodologische ontwikkelingen voor gecensureerde waarnemingen

Op dit vlak bestaat er een lange en internationaal erkende onderzoekstraditie binnen het consortium, verspreid over UCL, KUL-1, UH, UG, en USC. Deze laatste werkt nauw samen met de groepen uit Vigo en La Coruña. Heel wat aandacht is uitgegaan naar het voorstellen van efficiënte schatters met goede eigenschappen in het geval van censurering. Deze laatste kan niet-informatief zijn, maar ook het meer complexe informatief censureren is toegelaten. Tegelijk is er bijzonder veel werk besteed aan gecorreleerde stoptijden. Hier onderscheiden we twee grote scholen. De eerste vertrekt van het copula-principe, een klasse van flexibele associatiestructuren tussen stochastische veranderlijken. Hun toepassing is niet beperkt tot stoptijden, maar kent er wel vele en belangrijke toepassingen. De partners binnen het consortium hadden reeds enige bekendheid op dit terrein, doch samen hebben ze een sterke reputatie verworven op het vlak van copula's. Naast belangrijke publicaties op het vlak van copula-modellen voor gecensureerde waarnemingen, werden ook heel wat voordrachten op uitnodiging gegeven. Het andere prevalente paradigma is dat van de zogenaamde frailties, essentieel random effecten voor stoptijden. Op het vlak van frailties is evenees heel wat werk gebeurd. Duchateau (UG) en Janssen (UH) schreven een boek over dit thema; het is het eerste in zijn soort en bijzonder goed ontvangen.

Ook al zijn frailties en copula's onderling verbonden, in een aantal gevallen via analytische overgangen, toch kunnen we van twee scholen gewagen. Dit is niet verschillend van de hoger aangehaalde schoolvorming rond aan de ene kant parametrische en aan de andere kant niet-parametrische methoden voor ontbrekende gegevens. De kracht van het consortium is geweest dat het geen partij kiest in deze dichotomieën, doch eerder actief is binnen beide polen ervan. Meer nog, bruggen werden geslagen tussen beide, een extra reden waarom het werk internationaal goed ontvangen werd.

4.4 Complexe modellen voor ontbrekende gegevens en sensitiviteitsanalyse

Alle hogerstaand werk valt onder de categorie gekend als *missing at random* (MAR). Dit betekent dat het mechanisme dat uitval van gegevens stuurt kan afhangen van geobserveerde waarnemingen maar niet verder van wat niet werd geobserveerd. Het is een flexibel mechanisme en wordt nu standaard naar voren geschoven als datgene waar primaire analyses van bijvoorbeeld klinische studies best uit gekozen worden. Echter, de meer algemene *missing not at random* (MNAR) mechanismen kunnen niet helemaal uitgesloten worden. Daartoe werden dan ook modellen voorgesteld, onder meer gebaseerd op mengverdelingen. Ook zeer algemene *shared parameter modellen* (SPM) werden voorgesteld, waar de gegevens en het mechanisme dat uitval stuurt verbonden worden via random effecten. Het consortium is erin geslaagd MAR te beschrijven binnen het SPM mechanisme, wat voorheen nog niet gebeurde. Bovendien is het gelukt van aan te tonen dat voor een willekeurig MNAR model er een equivalent model bestaat dat MAR is, in de zin dat beide dezelfde aanpassing (fit) hebben aan de waargenomen gegevens.

De bovenstaande beschouwen hebben implicaties voor zogenaamde sensitiviteitsanalyse. Daarmee verstaan we het feit dat modellen, hetzij parametrisch, hetzij semi-parametrisch, altijd gebaseerd zijn op assumpties die niet te verifiïen zijn zonder de ontbrekende gegevens te kennen, wat per definitie niet kan. Het is daarom belangrijk niet alles te zetten op één bepaald MNAR model, maar de sensitiviteit te exploreren. Dit kan gebeuren op verschillende manieren: (a) bestuderen hoe conclusies veranderen binnen een klasse modellen, vaak modellen met gelijk(w)aardige aanpassingen aan de geobserveerde gegevens; (b) bestuderen hoe individuele waarnemingen in een onvolledige studie impact hebben op de conclusies. Het consortium verwierf een sterke reputatie op dit vlak, niet alleen voor ontbrekende gegevens, maar ook voor gecensureerde waarnemingen. Het werk situeert zich op drie vlakken. Ten eerste werden een reeks technieken ontwikkeld om sensitiviteitsanalyse uit te voeren. Ten tweede werden een aantal technische en meer algemeen toegankelijke manuscripten geschreven om het voorkomen van sensitiviteit te illustreren. Ten derde werd in heel wat werk ingegaan op het feit dat sensitiviteit niet beperkt is tot onbrekende waarnemingen of gecensureerde gegevens, maar voorkomt in alle vormen van onvolledige gegevens, zoals geschetst aan het begin van deze sectie. We vermelden, bij wijze van voorbeeld, het aantonen van het feit dat de random effect verdeling in hiërarchische modellen in grote mate arbitrair is en kan vervangen worden door een andere, zonder de aanpassing van het model aan de gegevens te verstoren.

4.5 Gezamenlijk modeleren van herhaalde metingen en stoptijden

Het gezamenlijk voorkomen van beide is niet ongebruikelijk. We denken hierbij aan twee typische situaties. Ten eerste komt het vaak voor dat een longitudinale covariaat gemeten wordt, samen met een stoptijd. Omdat beide stochastisch zijn kan men niet zomaar de ene conditioneren op de andere, doch is gezamenlijk modeleren aan de orde. Ten tweede kan dropout bij longitudinale waarnemingen gezien worden als de actie van een stoptijd. Met andere woorden, uitval kan gezien worden als een gediscretiseerde manifestatie van een stoptijd. Ook hier is het gemeenschappelijk modeleren noodzakelijk. Werk hogerop vermeld, zoals het SPM werk, is ook relevant voor deze categorie. Daarnaast verhoogt de sensitiviteit omdat beide processen eraan onderhevig zijn. Naast het puur formuleren van modeleren werd heel wat aandacht besteed aan het in kaart brengen en aanpakken van deze sensitiviteit.

5 Module de travail 4: Données avec hétérogénéité latente

Dans de nombreuses situations, on utilise des modèles statistiques qui supposent la présence de structures latentes, non-observables, pour expliquer la variabilité observée dans les données. Précisément, le fait que ces structures, par définition, ne peuvent jamais être observées pose des probleèmes particuliers par rapport à l'identifiabilité et à l'interprétation des résultats. Plusieurs publications se sont concentrées sur le thème de l'identifiabilité et sur la question de comment interpréter correctement les résultats issus de l'ajustement de tels modèles. En dépit des problèmes associés à ces modèles, ceux-ci ont prouvé leur grande utilité dans de nombreux modèles, et quelques illustrations en seront résumées dans la Section 5.1. Dans de nombreuses applications, des structures latentes sont requises à différents niveaux, ce qui implique des modèles avec effets appelés aléatoires croisés. Quelques exemples seront discutés dans la Section 5.2. Beaucoup de modèles statistiques utilisent des structures latentes pour engendrer des structures d'association dans des structures de données corrélées. Une application particulière est la construction de modèles pour l'analyse conjointe de résultats multiples. De nombreux exemples seront seront discutés dans la Section 5.3. Beaucoup de modèles statistiques standard sont basés sur des hypothèses sous-jacentes très strictes. Tels sont de nombreux modèles mixtes. Dans des applications spécifiques, les

hypothèses du modèle doivent être relachées pour que les modèles soient réalistes pour des ensembles spécifiques de données sous la main. Des exemples en seront présentés dans la Section 5.4. Bien que les modèles étudiés dans ce module de travail aient été dans l'air depuis plusieurs décades, ils posent encore des problèmes spécifiques par rapport à l'ajustement de modèle, et il y a souvent un manque de logiciel pour ajuster des modèles particuliers qui sont nécessaires pour répondre à des questions de recherche propres à certains domaines. Des contributions dans ce domaines seront discutées dans a Section 5.5.

5.1 Applications à des modèles avec structures latentes

Lorsque les structures latentes sont supposées être d'une nature continue, on se réfère traditionnellement à des modèles comme modèles à effets mixtes. Souvent les modèles sont des modèles de régression linéaire, linéaire généralisé ou non-linéaire dans lesquels certains paramètres sont supposés être échantillonnés d'une distribution continue, souvent une distribution multivariée normale. Ces effets aléatoires peuvent avoir différentes interprétations. Dans le contexte de notre projet de recherche, les exemples en sont des niveaux de médecins généralistes dans une étude avec des centres multiples, des niveaux d'hôpitaux dans une étude comparant différents hôpitaux par rapport à la sécurité des patients, des niveaux de patients suivis longitudinalement, des niveaux de fermes dans une étude du bien-être animal, des niveaux de sujets dans une expérience basée sur des micro-groupes (micro-array), *etc.* Les modèles à effets aléatoires peuvent aussi être appliqués dans de nombreuses situations où des points terminaux de substitution doivent être évalués dans un cadre d'essai clinique en recherche médicale, mais aussi, en recherche psychométrique, dans des modèles de réponse à des questions, de diffusion ou de Rasch. Souvent, les effets aléatoires sont utilisés pour modéliser implicitement des structures de corrélation dans les données. Un exemple typique est l'analyse des mesures répétées, où la corrélation entre les mesures répétées sur un même sujet est modélisée par des coefficients aléatoires spécifiques à l'individu, partageées par ces mesures répétées. Beaucoup d'exemples dans notre module de travail tombent dans le cadre de ce contexte longitudinal.

Parfois, les variables latentes sont supposées discrètes. On se réfère alors à ces modèles comme des modèles de mélange et ils posent souvent des hypothèses moins strictes que les autres modèles paramétriques. Plusieurs exemples ont été développés, par exemple dans le contexte de données longitudinales sujettes à des abandons ou la détection de groupes, pour des buts de classification, ou pour l'étude du fonctionnement différentiel des rubriques (Differential Item Functioning- DIF) dans un contexte psychométrique.

Les modèles mixtes ont aussi été appliqués de nombreuses fois dans des domaines où l'intérêt se trouvait dans la validation psychométrique (fiabilité et généralisabilité). L'avantage d'une telle approche est que les données existantes peuvent être utilisées, évitant ainsi le besoin d'études à objectif spécial. De plus, on n'est pas confiné à une planification typique pré-post, mais par contre on peut utiliser des séquences entières de mesures répétées. Des données aussi bien continues que non-continues peuvent être traitées de cette façon.

Les modèles mixtes ont aussi été très utiles pour estimer la force d'une infection à partir de données de prévalence sérologique et l'utilisation de données d'électroencéphalogramme pour discriminer entre des composés psychotoniques potentiels. D'autres applications incluent l'analyse de données longitudinales sur les feux neuronaux (neuronfiring) aussi bien que l'analyse des données de taux de couverture de vaccins.

Finalement, un domaine dans lequel les modèles mixtes ont été largement appliqués est l'analyse de données d'expérience sur micro-groupes. Dans une application, une méthode de calibration a été proposée pour un traitement préliminaire des pics dans des expriences par micro-groupes, basé sur des modèles

non-linéaires à effets mixtes. Cette méthode utilisait un pic dans une courbe de calibration pour estimer les valeurs normalisées d'expression absolue; de plus, des intervalles asymptotiques de confiance pour les valeurs d'expression estimées ont été construits. D'autres contextes dans lesquels les modèles mixtes non-linéaires ont été appliqués ont été la modélisation de la régénération des vaisseaux sanguins, ou la modélisation des niveaux de concentration des médicaments dans le plasma, pour des études pharmacocinétiques .

5.2 Effets aléatoires croisés

Des effets aléatoires croisés se rencontrent lorsqu'une variation aléatoire doit, dans un modèle, être introduite à différents niveaux. Un exemple, au sein de notre module de travail, a été l'utilisation de modèles à effets aléatoires croisés pour l'analyse de planifications complexes avec des modèles de type ANOVA contenant des effets aléatoires multivariés en vue de l'analyse des interactions entre des facteurs fixes et des facteurs aléatoires emboités. Une application en a été dans le contexte d'un modèle de diffusion dans l'analyse des temps de réaction au choix, avec des effets aléatoires spcifiques à la rubrique et à l'individu.

5.3 Structures latentes pour modélisation conjointe

On ne peut répondre à de nombreuses questions de recherche orientées sujet que par une modélisation conjointe de plusieurs résultats. Supposer un modèle mixte ou de mélange pour chaque résultat séparément et permettre à des structures latentes de différents résultats d'être communes ou corrélées impliquent, de façon bien évidente, des modèles conjoints. Cette idée a été appliquée dans l'analyse conjointe de 4 marqueurs différents, mesurés longitudinalement sur des patients ayant subi une transplantation rhénale. Le but était d'anticiper un échec de la greffe rhénale. Il a été montré que le modèle conjoint permettait une prédiction bien meilleure que ce qui aurait pu être obtenu par des modèles univariés séparés. D'autres applications incluent la modélisation conjointe d'un résultat de survie avec un résultat (binaire) mesuré longitudinalement et la modélisation de données binaires multivariées pour des anomalies dans des études de tératologie. Dans cette dernière application, l'hypothèse usuelle d'indépendance conditionnelle, c'est-à -dire l'hypothèse que les résultats sont indépendants conditionnellement aux variables latentes, était relachée au moyen de fonctions de copule et la distribution des variables latentes était maintenue flexible au moyen de mélanges finis de distributions normales.

5.4 Extensions de modèles et/ou modèles flexibles

Beaucoup de modèles standard mixtes ou de mélange sont basés sur des hypothèses très strictes. Dans des appliquations spécifiques, ces hypothèses doivent être relachées afin que ces modèles soient réalistes pour des ensembles spécifiques de données disponibles. Dans un exemple, un modèle pénalisé de mélange gaussien a été utilisé pour analyser des données longitudinales d'effets aléatoires avec un accent mis sur la modélisation flexible de la distribution des effets aléatoires. Ceci a permis d'examiner l'impact de la distribution gaussienne sur l'estimation de la partie d'effets fixes du modèle . La conclusion en a été que l'impact du choix correct de la distribution des effets aléatoires est variable d'une situation à l'autre. Un modèle semblable a été utilisé pour l'estimation de l'association entre deux résultats censurés par intervalles; il y a été montré que cette approche fonctionne bien dans beaucoup de situations réalistes et qu'elle est supérieure aux autres approches existantes. En vue de garder le modèle très flexible, tous ces modèles permettent un (très) grand nombre de composantes du mélange, et l'ajustement du modèle est basé sur une approche de vraisemblance pénalisée. Dans un modèle semblable mais avec un contexte

différent, l'approche de la vraisemblance pénalisée a été remplacée par une méthode d'estimation basée sur une méthode d'échange de sommets, un algorithme emprunté à la littérature des mélanges finis.

De nombreuses applications ont été données pour des modèles avec des structures latentes flexibles. Dans un exemple dans le contexte des données dentaires, on observait des comptages de non-gonflement (zero-inflated counts) pour l'indice dmft qui est la somme des dents de lait avec expériences de caries. Le non-gonflement se rencontrait à cause de la corrélation entre les expériences de caries d'une dent dans une même bouche, mais ceci impliquait aussi une distribution sur-dispersée des nombres de non-gonflements (distribution binomiale négative). Un autre application était la combinaison d'un processus de diffusion de Wiener pour les temps de réponse au choix avec des techniques de la psychométrie en vue de construire un modèle hiérarchique de diffusion. Ceci a conduit à un cadre de modélisation qui est hautement flexible et très commode à opérer.

5.5 Développement de méthodes d'estimation et logiciels

Comme des extensions de modèle sont souvent requises pour considérer adéquatement des questions de recherche liées à un domaine particulier, les méthodes standard d'estimation et/ou les outils standard de logiciel ne sont souvent pas suffisants pour traiter les modèles disponibles. Dans ce module de travail, plusieurs équipes ont contribué au développement d'une nouvelle théorie de l'estimation, d' algorithmes d'ajustement et/ou de logiciels.

Pour l'ajustement des modèles mixtes, avec des variables latentes continues, le logiciel standard se base sur des méthodes d'approximation du modèle ou des données. Des méthodes telles que la quasi-vraisemblance marginale (MQL) et la quasi-vraisemblance pénalisée (PQL) se mettent facilement en place, mais conduisent souvent à des estimations biaisées de paramètres. Alternativement, des méthodes de quadrature peuvent être utilisées, mais elles sont souvent très instables et consommatrices de temps, particulièrement dans des modèles avec de nombreux effets aléatoires. Une équipe a comparé différentes méthodes pour l'estimation de la distribution des variables latentes, dans le contexte des modèles linéaires mixtes tandis qu'une autre équipe proposait d'utiliser des approximations de Laplace complètement exponentielles, qu'on a montré rapides et stables. Une alternative était d'utiliser une estimation appelée variationnelle, qui ne requière pas d'évaluation numérique d'intégrales et est basée sur une approximation de la borne inférieure du modèle logistique. Dans le contexte de modèles de fragilité (frailty) pour les données de survie, une méthode innovante d'estimation a été proposée; elle utilise une transformation basée sur la fonction de hasard cumulée pour transformer les données, de telle sorte que les données peuvent être analysées en utilisant un modèle linéaire à effets mixtes. Les propriétés de la méthode ont été analysées en utilisant une étude par simulation et une étude d'un cas de vie réelle. Une autre équipe a proposé un estimateur itératif du mode a posteriori en théorie de réponse de rubrique (item response theory- IRT) comme une technique améliorée pour estimer les niveaux d'habilité.

Pour ce qui concerne le développement de logiciels, une méthode d'estimation a été proposée, réalisable avec des logiciels facilement disponibles, pour ajuster un modèle uni- ou bi-dimensionnel de mélange pour DIF. D'autres ont rendu disponibles des logiciels dans le but de rendre plus accessibles aux psychologues expérimentaux le modèle de diffusion de Ratcliff pour les données de temps de réaction et de précision. Finalement, une boite à outils MATLAB pour les modèles IRT (IRTm) a été développée: elle permet d'ajuster une grande variété de modèles IRT, avec l'inclusion de modèles IRT de copule, récemment développés, afin de traiter les dépendances locales entre rubriques.

6 Module de travail 5: Données en grandes dimensions et données composées

On rencontre des données de grande dimension et des données composées dans différentes disciplines. La plupart des méthodes développées dans ce module de travail sont concernées avec une discipline spécifique; la production de recherche dans ce module de travail est donc divisée en trois disciplines: bioinformatique, analyse exploratoire et psychométrie. Evidemment, beaucoup des méthodes développées dans une discipline pourraient être fortement pertinentes dans une autre discipline.

6.1 Bioinformatique

Une premier domaine important de recherche en bioinformatique concerne la *génétique des populations*. Dans ce contexte, on a développé des tests pour l'interaction statistique et suffisante des causes, avec un accent sur le test d'interaction de gene-environnement, sur l'analyse de médiation (c'est-à -dire de séparation des effets d'exposition directe et indirecte) et sur l'ajustement pour la confusion variable dans le temps. On a proposé, dans des études de familles, des tests d'interaction de gene-environnement sur la base de particularités (traits) complexes. Parmi les caractéristiques désirables de cette approche il y a le fait qu'elle permet, dans des études de familles, l'estimation d'effets génétiques (alors que l'accent était mis précédemment sur les tests), qu'elle ajuste pour des facteurs de confusion non-mesurés et dûs à des mélanges de populations, et qu'elle exige seulement de modéliser l'effet génétique principal. Ces techniques innovantes sont basées sur de multiples tests statistiques robustes d'interaction qui restent sans biais sous des erreurs de spécification de modèles, pourvu que l'information sur la distribution du temps d'exposition soit disponible, et peuvent donc être utilisés même lorsqu'un ajustement pour des covariables de grande dimension est requis.

Un autre secteur de recherche concerne des modèles évolutifs de substitution qui permettent des dépendances au voisin le plus proche. Sur la base d'informations nucléotides pour bon nombre d'espèces pour lesquelles l'arbre génétique est connu, nous avons développé une inférence pour les paramètres d'évolution basée sur un enrichissement de données et nous avons trouvé une forte confirmation de la présence de dépendances aux voisins. Étant données la haute dimensionalité des modèles et la complexité des calculs, des stratégies heuristiques de sélection de modèles ont été proposées; elles sont basées sur une distribution a posteriori à utiliser lorsque les facteurs de Bayes sont trop intensifs en calcul. Finalement, la méthodologie pour la génétique des populations a aussi été étudiée pour la culture des plantes en utilisant des techniques de modèles mixtes pour traiter des sources de variabilité dans l'analyse des données d'expérience de phénotype avec du riz transgénique. La variabilité somaclonale et la variabilité d'insertion sont séparées l'une de l'autre afin d'évaluer correctement l'effet du gène greffé et sa variabilité. Comme extension de ce travail, on a aussi étudié des aspects de planification de tels essais dans le but d'évaluer l'effet spécifique du gène et de produire un mutant productif.

Un autre sujet d'intérêt en bioinformatique est le problème des *tests multiples*. On a bien documenté le problème de résultats faussement positifs dûs à la répétition de tests lorsqu'on examine les gènes avec un effet sur la particularité. Dans celui-ci, et dans d'autres cadres de tests répétés, la nécessité de contrôler a conduit à un contrôle beaucoup plus strict des erreurs de type 1, avec, comme conséquence, un affaiblissement de la puissance. Lorsque des cultivateurs de plantes scrutent pour des gènes prometteurs, ils sont cependant fortement concernés par l'erreur de type 2 parce qu'ils ne souhaitent pas perdre de vue trop tôt des gènes vraiment prometteurs. Afin de répondre à ces besoins, nous avons développé un test équilibré pour un examen approfondi où on maximise une moyenne pondérée entre la probabilité de tirer

une conclusion correcte sous l'hypothèse nulle et sous l'hypothèse alternative. Un classement fondamentalement différent des gènes prometteurs résulte de cette approche avec due considération de l'hypothèse alternative. On en dérive une mesure correspondante de l'évidence relative en faveur de l'hypothèse alternative par rapport à l'hypothèse nulle. Cette approche a été de plus adaptée aux planifications à deux étapes et on a développé plusieurs adaptations à des structures particulières de données. Un problème semblable de sur-optimisme se révèle lorsqu'un modèle de régression est ajusté en vue de la prédiction et nous avons pris pour objectif d'évaluer la validité de l'ajustement du modèle de prédiction. Dans une approche de régression qui cible la prédiction directement sur l'échelle des données originales, en utilisant la norme L1, nous considérons, comme critère d'évaluation de modèle, l'erreur absolue de prédiction, la valeur attendue de la différence absolue entre les réponses futures et prédites.

On a aussi développé différentes méthodes d'analyse pour des données de *micro-cellules* (micro-array). Plusieurs procédés de tests pour une tendance monotone dans des expériences de réponse à des doses ont fait l'objet de recherches et ont été appliquées dans des contexte de données de micro-cellules. Nous combinons, en particulier, des procédés avec ajustements pour des tests multiples entre gènes, et adressons le thème des petites tailles d'échantillon en ayant recours à des inférences basées sur un ré-échantillonnage. De plus, nous considérons une gamme de méthodes de classification, et les appliquons au problème de classifier des échantillons basés sur des données de micro-cellules. Nous comparons les méthodes en utilisant une étude de simulation. Il en ressort que les méthodes classiques, telles que la LDA diagonale, semblent jouir d'un très bon comportement en comparaison avec d'autres techniques plus compliquées. Nous avons, de plus, réalisé une étude par simulation afin d'analyser la performance de plusieurs méthodes de sélection de gènes en combinaison avec plusieurs techniques de classification dans le contexte des micro-cellules. La stabilité des méthodes par rapport aux hypothèsse de distribution a ét examinée en considérant aussi des données simulées à partir de distribution de Laplace symétrique et non-symétrique, en plus de données de micro-cellules distribuées normalement. En utilisant une étude de simulation et une étude de cas, nous avons analysé la performance du SAM dans de nombreuses expériences sur des micro-cellules. On a évalué, en particulier, l'influence de la présence de gènes avec une faible variabilité des caractéristiques opérationnelles du SAM. De plus, on a développé un algorithme innovant pour le déploiement multidimensionnel et qui maîtrise à la fois des problèmes généraux et des problèmes spécifiques à l'analyse de données d'expression génétique. Cette méthode offre un outil utile pour des explorations préliminaires de données de micro-cellules. Au niveau de la sélection de modèles/ test d'hypothèses, nous avons aussi recherché un procédé approprié de test statistique pour détecter des gènes qui sont sur-représentés dans un seul tissus en comparaison avec tous les membres du panel d'autres tissus(le problème du test d'hypothèses avec une hypothèse alternative qui implique une intersection).

6.2 Analyse exploratoire (data mining)

Dans l'analyse exploratoire, on est souvent confronté à un grand ensemble de données, avec souvent une collection énorme de variables différentes. Pour des données d'aussi grande dimension, la réduction de dimension est souvent appliquée au moyen de la méthode de l'analyse en composantes principales (Principle Component Analysis- PCA). Les propriétés théoriques d'une méthode PCA robuste, appelée ROBPCA, ont été étudiées en dérivant sa fonction d'influence. De plus, on a développé une méthode PCA robuste pour des données unimodales non-symétriques et des cartes appropriées d'observations atypiques afin de visualiser les éventuelles observations atypiques. On a aussi proposé une nouvelle méthode PCA à noyau ainsi que des outils de diagnostic afin de détecter les observations influentes. Les techniques de machine à support vectoriel comptent parmi les méthodes utilisées pour des données

de grande dimension. On a développé des techniques générales pour un cadre typique dans l'analyse exploratoire avec un grand nombre de variables observées et l'exigence de sélectionner les variables les plus importantes. La sélection de variables dans une structure de modèle additif a été réalisée en utilisant une combinaison de la technique du garrot non-négatif et de l'estimation par P-splines. Ceci pour résultat une puissante méthode de sélection de variables, dont on a illustré la performance en comparaison avec d'autres procédures récentes de sélection de variables.

On a entrepris une recherche substantielle concernant le développement de méthodes robustes qui puissent supporter la présence d'observations atypiques dans un cadre d'analyse exploratoire. L'algorithme LARS robustifié en est juste un exemple.

On a aussi considéré des applications de l'analyse exploratoire en chemiométrie, telles que l'identification de biomarqueurs métobonomiques. Les approches de découvertes métobonomiques à base NMR exigent des méthodes statistiques pour extraire, à partir de bases de données spectrales complexes et de grande dimension, des biomarqueurs ou des variables biologiquement signifiantes qui représentent au mieux des conditions biologiques définies. Nous avons exploré l'efficacité respective de six méthodes multivariées: les tests multiples d'hypothèses, des extensions supervisées de l'analyse en composantes principales (PCA) et indépendantes (ICA), des moindres carrés partiels discriminants, la régression linéaire logistique et les arbres de classification. Une des difficultés majeures dans le contexte du développement des méthodes chromatographiques consiste en la détection automatique des pics provenant de matrices chimiques complexes. Une méthodologie intégrée basée sur l'analyse en composantes indédendantes (Independent Components Analysis -ICA) et la mise en grappes (clustering) a été proposée pour solutionner ce problème et a été appliquée au chromatogrammes HPLC-UV-DAD.

6.3 Psychométrie

Des données composées et de grande dimension se rencontrent aussi fréquemment dans la littérature psychologique. Pour traiter de grands ensembles de données (d'objets de grande dimension par type de données variables), on a investigué une grande gamme de méthodes qui impliquent une réduction de variables (et éventuellement aussi d'objets), avec cette réduction étant soit catégorique soit de nature dimensionnelle (regroupement en grappes, respectivement réduction de dimension). Les contributions en ce domaine incluent le développement de modèles innovants, tels que un modèle qui regroupe des objets dans des grappes se recouvrant et réduit simultanément l'espace des variables, et un modèle qui regroupe simultanément à la fois des objets et des variables dans des grappes se recouvrant; ce dernier modèle vient avec une extension particulièrement utile qui permet de découvrir la nature des interactions telles que présentes dans les données. Les méthodes qui ont été développées sont de nature aussi bien déterministe que stochastique, avec une méthode générique proposée pour étendre les modèles déterministes en des contreparties stochastiques immédiates.

Pour traiter des données en classification multiple, on a développé des modèles qui captent la structure de données à trois classifications en utilisant des effets aléatoires. On a de plus investigué la relation entre deux modèles précédemment développés pour le regroupement simultané de classifications multiples. En outre, au delà de modèles spécifiques, on a développé un modèle unificateur qui implique une réduction de catégories et/ou de dimensions d'une ou plusieurs modalités dans un ensemble de données à classification multiple; ceci englobe une large classe de modèles existants de réduction comme cas spéciaux et peut ainsi agir comme un outil puissant pour la création de modèles innovants.