

Proposition de Projet INRIA

Mistis

Modélisation et inférence de systèmes stochastiques
complexes et structurés

présentée par
Florence Forbes

Composition de l'équipe

- **Chef de projet:**

F. Forbes [CR1, Inria]

- **Membres permanents:**

S. Girard [CR1, HDR, Inria]

L. Gardes [MdC, UPMF, Grenoble]

J-B. Durand [MdC, INPG Grenoble, 50%]

- **Collaborateurs extérieurs:**

H. Bertholon [MdC, CNAM, Paris, 20%]

G. Fort [CR, CNRS, Paris, 20%]

- **Thésards:**

J. Blanchet [UJF, co-encadrement Lear, oct. 2004]

L. Donini [CIFRE, Xerox, dec. 2006]

V. Khalidov [UJF, nov. 2006]

M. Vignes [UJF, co-encadrement Select, oct. 2003]

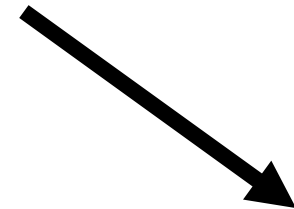
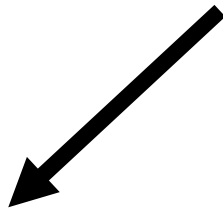
- **Post-doctorants:**

C. Bernard-Michel [décembre 2006-décembre 2007]

P. Doc [2007, ARC SeLMIC]

Bases scientifiques
Savoir faire de Mistis

Modèles de mélange
Modèles markoviens, graphiques
Méthodes semi et non paramétriques



Biologie/Médecine

Neurosciences:
intégration sensorielle multimodale

Génomique:
classification de gènes

Génétique des populations:
identification de structure

Angio-IRM (temporel)

Analyse d'images

Imagerie médicale, **IRM**

Vision par ordinateur:
eg. Reconnaissance d'objets

Imagerie hyperspectrale:
télédétection

Visualisation de données

Optimisation des procédés

Gestion du risque

Optimisation d'infrastructures

Fiabilité

Détection de pannes

Objectif

Aller au-delà des modélisations statistiques traditionnelles

Proposer des méthodes statistiques adaptées à une classe de problèmes:

Sources de complexité des données:

- Mesures répétées pour un même individu
- Effet de la méthode d'échantillonnage
- Effet des instruments de mesures
- Données de grande dimension, hétérogènes
- Données manquantes, ...



Beaucoup de variables, d'interactions, de mesures

Stratégie de modélisation statistique:

propriétés simples (locales ou de régularité), associées de manière cohérente



Exhiber et gérer la complexité

Structure: variables liées, graphe hiérarchique, séquentiel

- Interprétation claire
- Analyse statistique difficile (dépendances)

Savoir Faire de Mistis

I. Modèles markoviens, graphiques

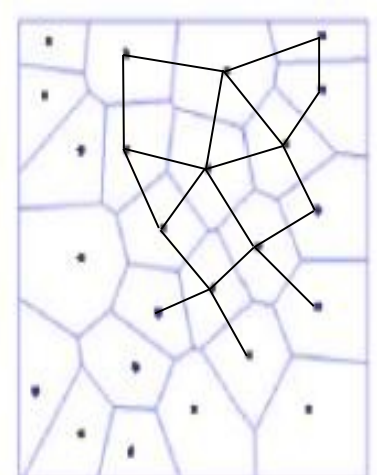
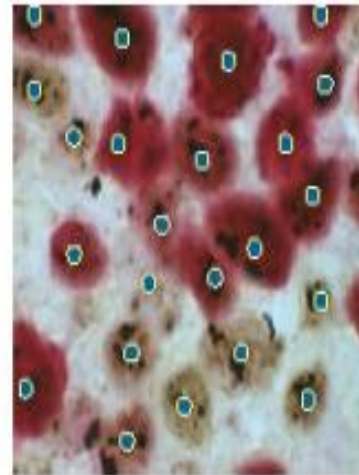
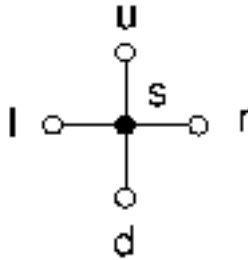
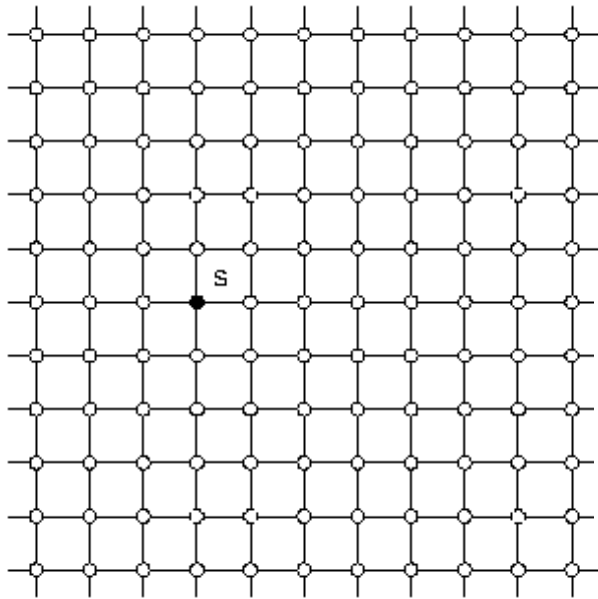
II. Modèles de mélange

**III. Méthodes semi et non
paramétriques**

I. Modèles markoviens, graphiques

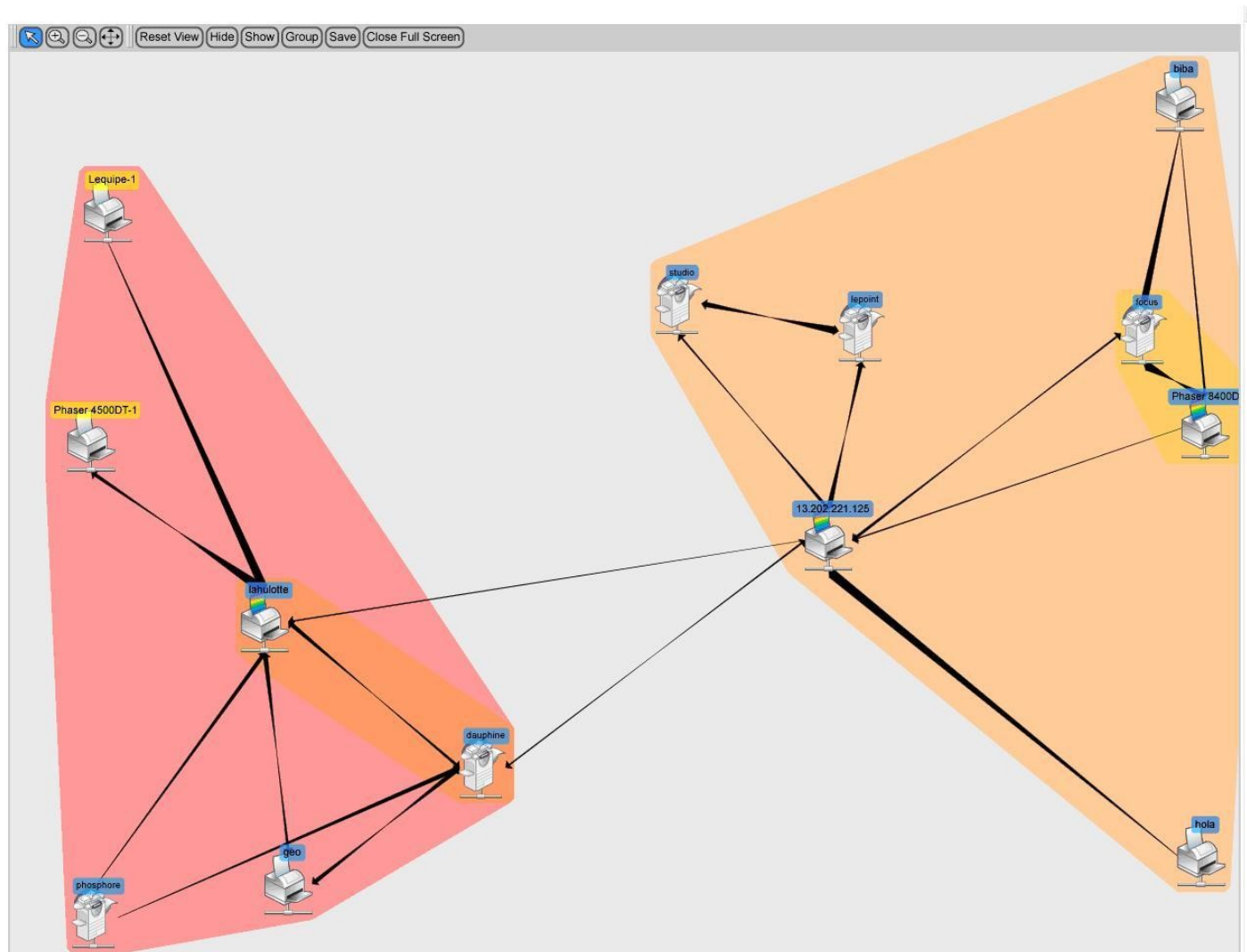
indépendance conditionnelle, dépendance et localité

exemple: analyse d'images, grilles de pixels ou graphes non réguliers



structure cachée ou non (réseaux bayésiens, HMM), hiérarchiques, fixes ou dynamiques

Exemple: infrastructure d'imprimantes



II. Modèles de mélanges:

Données manquantes, hétérogènes



Modèles à structure cachée

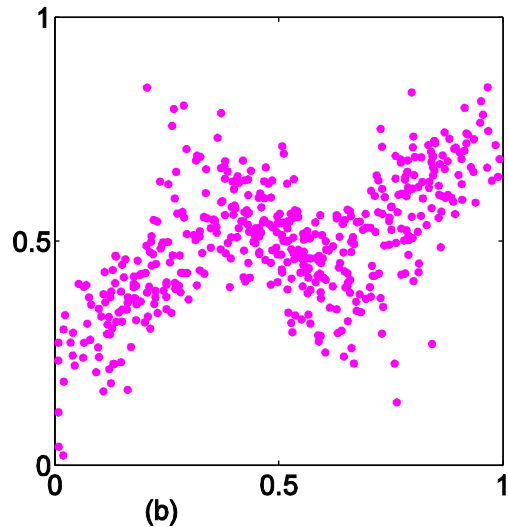


Algorithme EM:

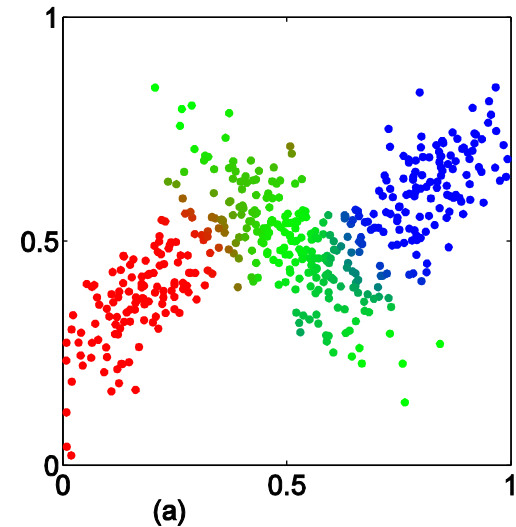
estimation des paramètres

restauration des données manquantes

Clustering (classes manquantes) et estimation de densités



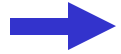
EM
→



Problème d'étiquetage (coloriage, configuration)

III. Méthodes semi et non paramétriques

Absence de modèles paramétriques



Inférence fonctionnelle (Méthodes de projection ou méthodes à noyaux)

Exemple: estimation de frontières, regression non lineaire

$$f_n(x_1; \dots; x_n)$$

- **Grand nombre de variables:**

$$f_n(x_1; \dots; x_n) \gg f_1(a_1 x_1 + \dots + a_n x_n)$$

réduction de dimension



$$f_n \longrightarrow f_1; (a_1; \dots; a_n)$$

Modèle semi paramétrique

- **Analyse d'évènements extrêmes:**



Semi paramétrique: partie fonctionnelle + indice des valeurs extrêmes

Projets collaboratifs

- **ACI Masses de données MOVISTAR (2003-2006)**

Modèles visuels et statistiques en reconnaissance et description d'images

Partenaires: Lear, SMS, Heudiasyc Compiègne
2 thèses, nombreuses publications

- **ARC IBN (2005-2006)**

Integrated biological networks

Partenaires: Helix, Symbiose, Inst. Pasteur, INRA

Projets collaboratifs

- **STREP Européen POP (2006-2009): Perception On Purpose**

Partenaires: Perception, U. Osnabruck, U. Hospital Hamburg-Eppendorf, U. Coimbra, U. Sheffield

Axes de la participation de Mistis:

- **Intégration d'indices visuels monoculaires et binoculaires**
collaboration Perception, doctorat de R. Narasimha
- **Perception binaurale**
collaboration U. Sheffield, doctorat V. Khalidov
- **Intégration audio-visuelle (eg. tracking)**
collaboration U. Sheffield et Perception

Méthodologie: Modèles Bayesiens, modèles graphiques, Markoviens

- **ARC SeLMIC (2007-2008) (F. Forbes, Mistis)**

Segmentation longitudinale et multimodale par IRM
d'anomalies cérébrales (AVC et SEP)

Partenaires: Visages, U594 INSERM, Magma/LIG
1 post-doc

- **ARC ChromoNet (2007-2008) (MF. Sagot, Helix)**

Computational Analysis of Inter-Chromosomal
Interaction Networks

Partenaires: Helix, INRA, MRC Imperial College London
1 post-doc

Programme de recherche

	Modèles de mélanges	Modèles markoviens	Méthodes semi et non paramétriques
Apprentissage et classification	Extensions, Classes chevauchantes	Dépendances et interactions	
Techniques d'approximations		Déterministes/par simulation couplage MCMC variationnel	
Propriétés théoriques	Des algorithmes: convergence	Des approximations: qualité	Des estimateurs: tests, correction de biais
Réduction de dimension	Mélanges adaptés aux grandes dimensions	Dépendances et interactions	ACP généralisée
Modèles intégrés	Interprétation des paramètres	choix de la structure (voisinage)	
Extrêmes			estimateurs, Tests d'adéquation
Estimation de frontières			inférence fonctionnelle

Réduction de dimension

Modèles pour données de grande dimension

EX: Reconnaissance d'objets et de classes d'images [Lear]

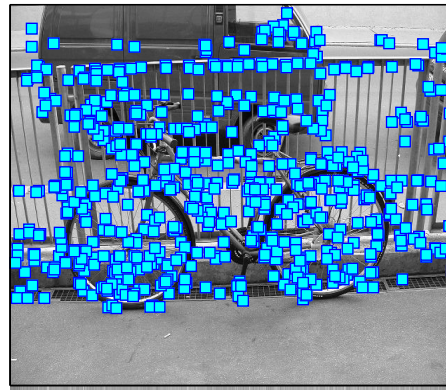
Détection de points d'intérêt

Extraction de descripteurs (**dimension 128**)

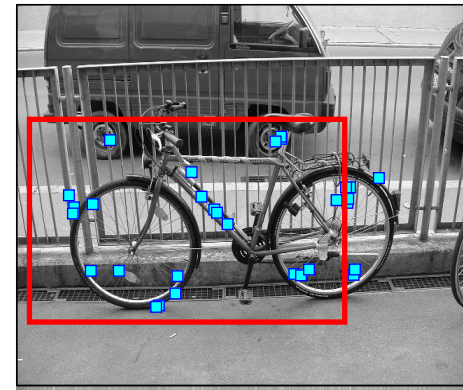
Classification Supervisée (HDDA) ou non supervisée (HDDC)



Image originale



detections

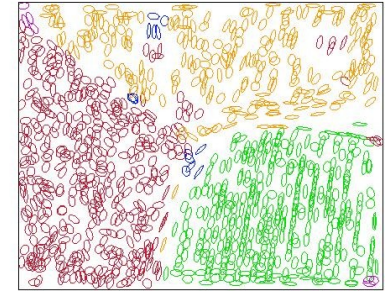
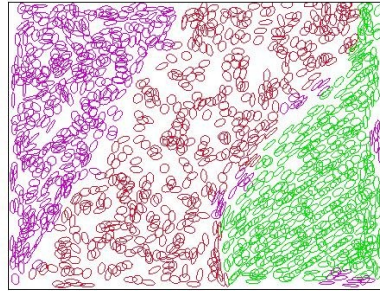
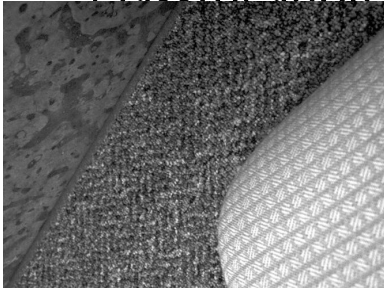


localisation finale

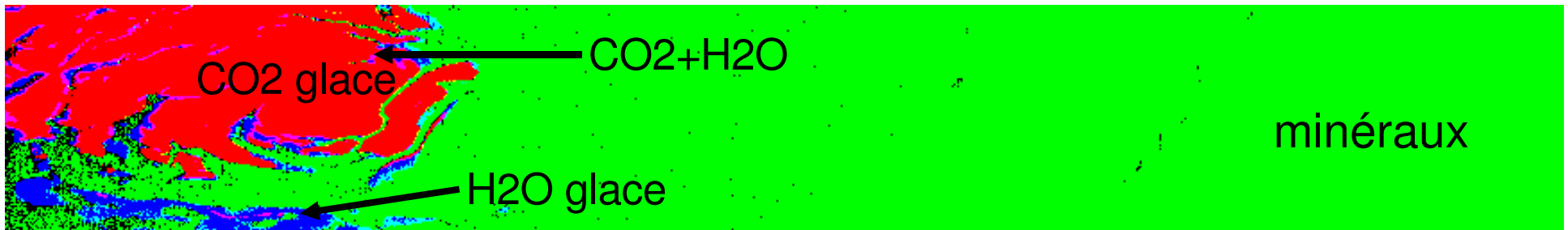
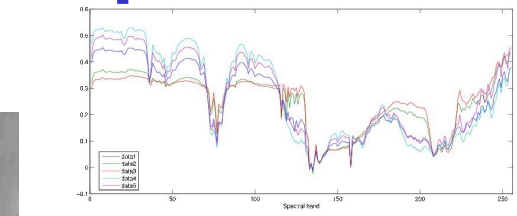
Combiner réduction de dimension et interactions

EX: Reconnaissance de textures [Lear] apprentissage +

classification



Ex: Analyse d'images hyperspectrales [Mars-Express,LPG]



Aspect non paramétrique de la réduction de dimension

[U. Lille I]

Généralisation de l'ACP linéaire

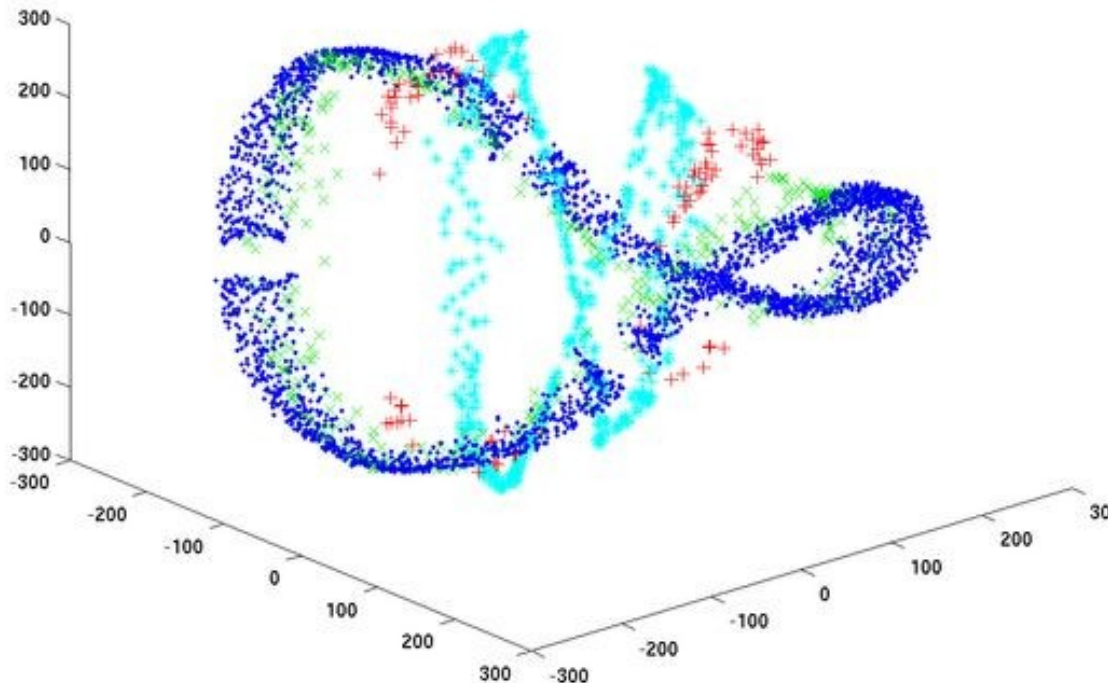
$$\sum_{i=1}^p a_i x_i = 0$$

Approximation d'un nuage de points par une variété

$$f_n(x_1; \dots; x_n) = 0$$



variables



: nuage de points ou l'ACP

lineaire est inefficace

Modèles intégrés, fusion d'informations

Prise en compte de différentes sources, de connaissances a priori

- **Modélisation des connaissances (souvent haut-niveau)**
- **Intégration dans les modèles (à travers les paramètres, la structure)**
- **Interprétation des paramètres**
- **Choix/influence de la paramétrisation, de la structure (voisinage)**

Exemples: **IRM:** intégration de connaissances anatomiques

Génomique: intégration de données réseaux

Perception: mélanger stimulus visuel et auditif

Segmentation markovienne distribuée et coopérative

d'IRM [B.Scherrer, INSERM/LIG]

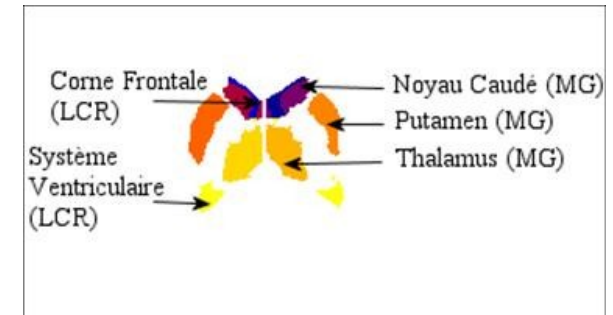
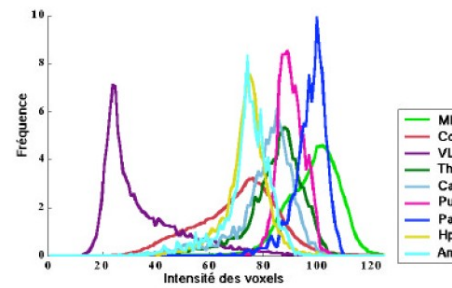
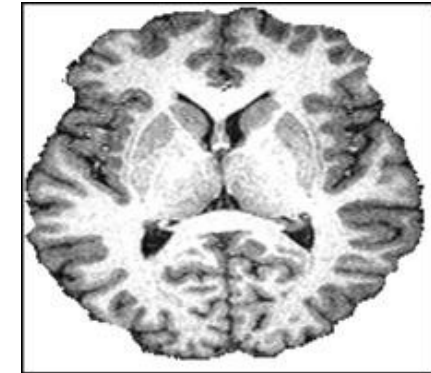
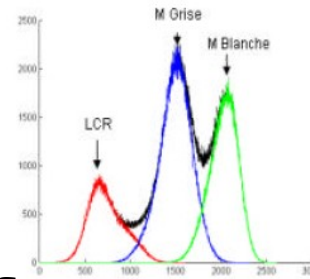
Segmentation des tissus (LCR, MB, MG):

se base sur la radiométrie des tissus

Segmentation des structures:

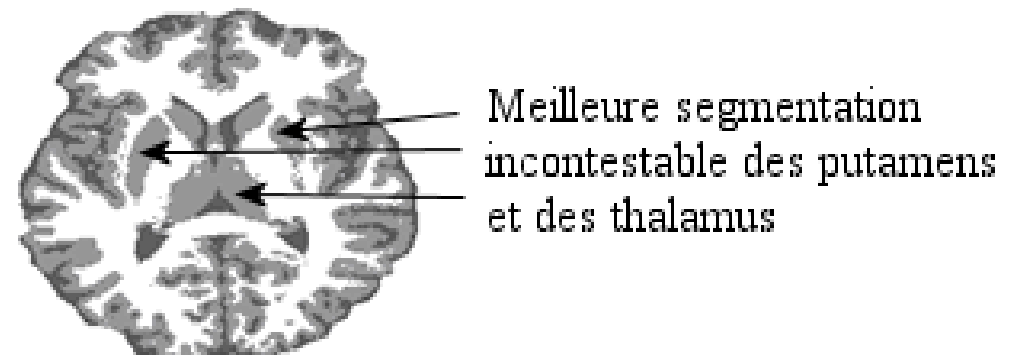
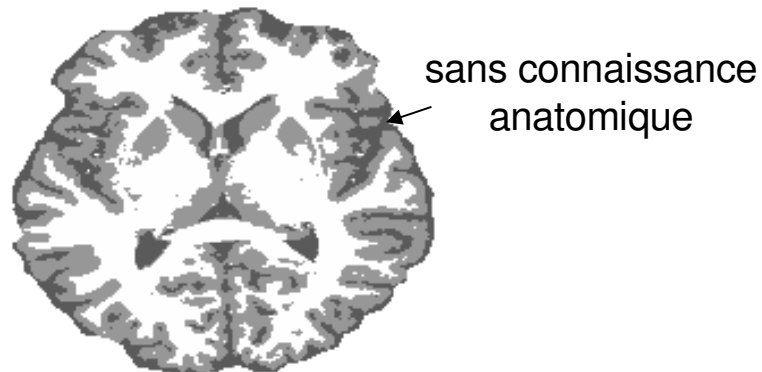
Informations radiométriques insuffisantes

- **Apporter de l'information *a priori***
- **Peut améliorer la segmentation des tissus**



Intégration de connaissances anatomiques (distance, orientation, symétrie): ex. “le putamen est

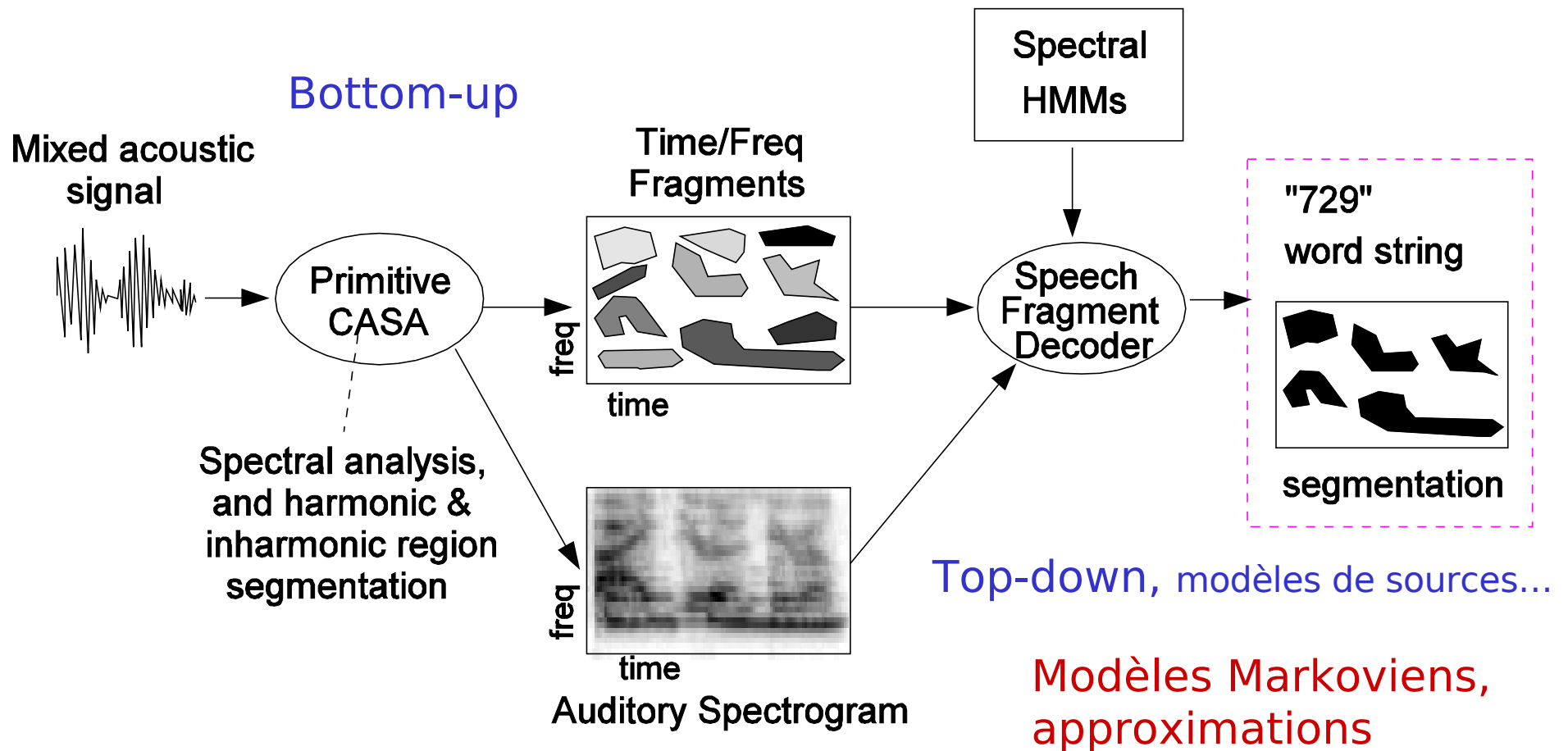
à environ 20mm du plan interhémisphérique” → **cartes floues**



Intégration de données audio-visuelles [projet POP]

Décomposition du son en fragments (U. Sheffield, Barker et al.)

- Combinaison au niveau du “frame”: trop ambiguë
- Associer indices visuels et **fragments** audio



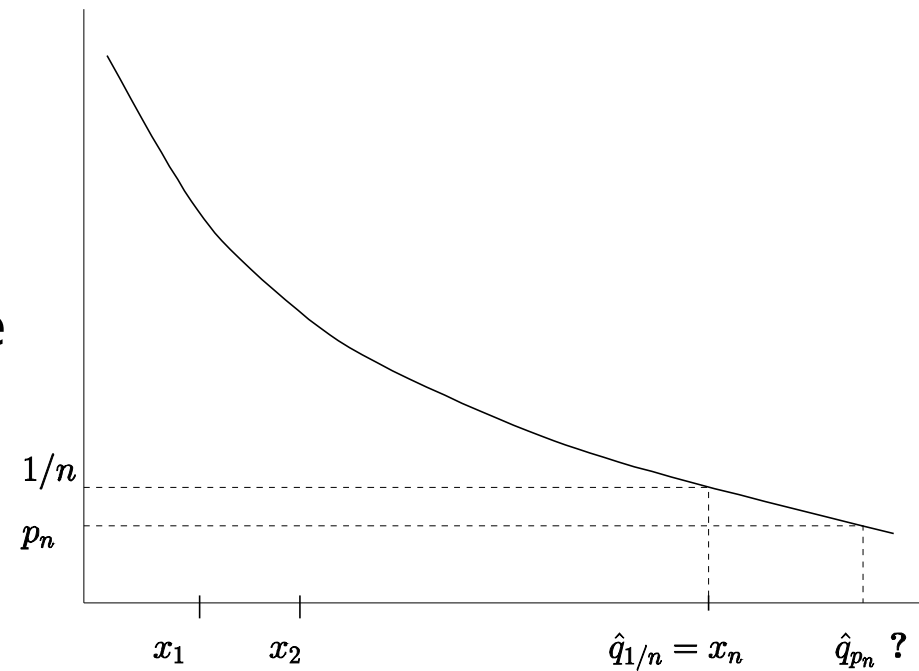
Modélisation d'évènements extrêmes [TimB, INRA, U. Strasbourg, U. Marne-La-Vallee]

Estimer le risque d'évènements extrêmes (gestion du risque, fiabilité)

Ex: fissures de cuves de centrales nucléaires (CEA Cadarache)

Estimer des p-quantiles avec $p < 1/n$

Ex: estimer hauteur de la crue centennale



Développer des estimateurs semi-paramétriques

Développer des tests d'adéquation aux queues de distributions (vs partie centrale)

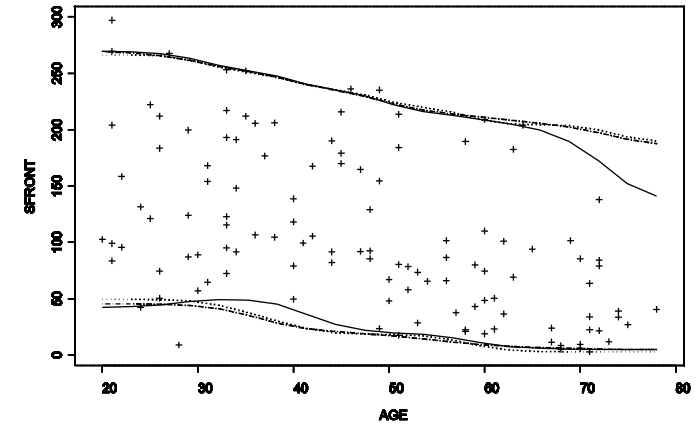
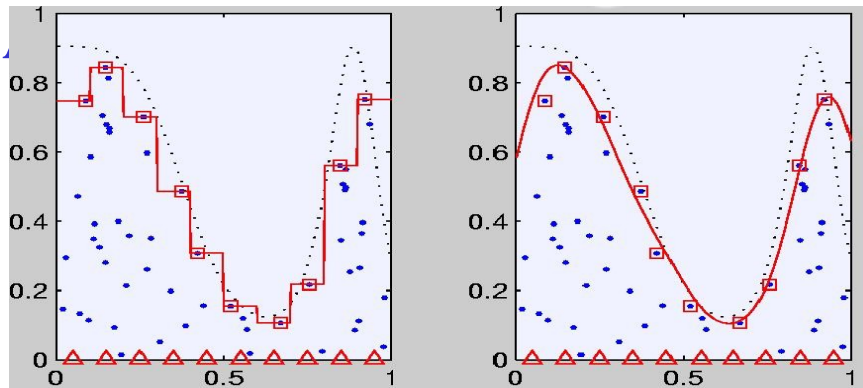
Valeurs extrêmes multivariées

Estimation de frontières [U. Moscou, U. Montpellier 2, SMS]

Détection de outliers

Courbes de référence, ex: courbes de croissance

Estimation de supports, de contours



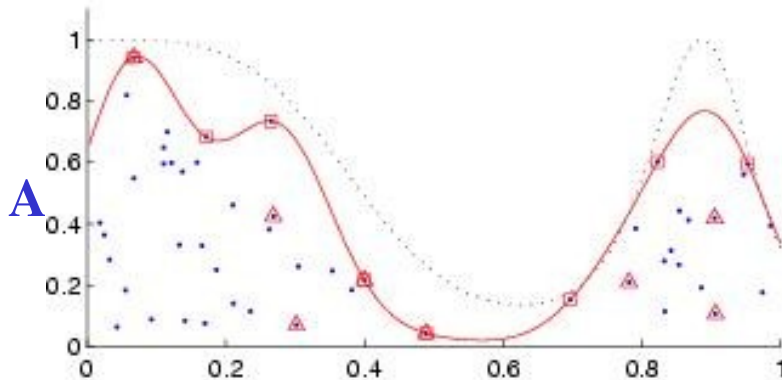
Propriétés théoriques:
IC et tests, correction possible du biais

Choix du découpage en pratique



Lien SVM, pas de découpage

Propriétés théoriques à étudier



Développement logiciel

Extrêmes: [SG, J Diebolt (CNRS), M. Garrido (INRA), J.Ecarnot, EDF R&D] dépôt APP

Théorie des valeurs extrêmes, événements rares

- code C++, interface Matlab

Toolboxes HDDA et HDDC: [SG, CB]

Modèles gaussiens pour données de grande dimension

- Code C/interface Matlab, implémentation dans MIXMOD prévue

SpaCEM3 (Spatial Clustering with EM and Markov Models): [FF, JB] dépôt APP

Classification de données spatiales à partir de modèles markoviens

- Code C++

Fastruct : [FF, CC, O. Francois TimB] dépôt APP

Modélisation et inférence en génétique des populations

- Code C, Interface utilisateurs et guide

TESS : [FF, CC, O. Francois TimB] dépôt APP

Clustering Bayésien de données spatiales de génétique des populations

- Code C , Interface utilisateurs et guide

Positionnement thématique:

approche et spécificités

- **Originalité:**

Stratégie de modélisation statistique qui repose sur la notion de **structure** dans

- les systèmes/phénomènes complexes: modélisation, interprétation
- les modèles: estimation, sélection, techniques d'approximation
- les données: réduction de dimension, analyse de données non linéaire

Etude théorique et pratique des méthodes:

- qualité des approximations,
- comportement asymptotique des estimateurs,
- convergence des algorithmes

- **Pertinence:** Couplage modèles/données, **au-delà des modèles statistiques traditionnels**
- **Intérêt au sein de l'Inria:** **apport d'une rigueur et formalisme statistiques**
 - Liens avec plusieurs projets/ équipes: Lear, Perception, Helix, Visages, Inserm, TimB, LPG
 - Participation à des projets: ACI Movistar, ARC IBN, ChromoNet, SeLMIC, Strep POP
 - Co-encadrements de thèses (4) et post-docs (3)

Positionnement à l'INRIA

- Mots-clefs: systèmes complexes, information, applications médecine et biologie
- Equipes de culture statistique:
 - Select (G. Celeux): sélection de modèles et apprentissage/ modèles markoviens
 - Aspi (F. Legland): Chaines de Markov cachées / Champs (2D), approximations déterministes
 - Sisthem (M. Basseville): détection de ruptures, méthodes de rejet/ outils différents
- Thèmes et intérêts communs:
 - Ariana (J. Zerubia): modèles markoviens/ au delà de l'image
 - Vista (P. Bouthemy): modèles markoviens en image
 - Lear (C. Schmid): apprentissage en vision
 - Perception (R. Horaud): modèles cachés et bayésiens
 - Prima (J. Crowley): modèles de mélanges, analyse de la parole
 - e-Motion (C. Laugier): réseaux bayésiens
 - Helix (A. Viari): traitement statistique de données génomiques
 - Mescal (B. Gaujal): modèles structurels, méthodes de simulation
 - Moise (E. Blayo): mot-clefs communs/ approches différentes
 - Visages (C. Barillot): IRM
 - Virtual Plants (C. Godin): modèles markoviens de plantes

Positionnement national

Interactions possibles au sein du dept Stat. du LJK (44 permanents)

3 anciens membres (SG, LG, JBD) de SMS/ex Labsad
séminaires, co-organisations, enseignement

Principaux thèmes d'interactions:

- Inférence fonctionnelle, analyse multi-résolution
- Traitement de données de grande dimension
- Modèles markoviens et graphiques
- Fiabilité

• Apports et spécificités Mistis:

- Modèles markoviens graphiques pour l'analyse de données spatiales
- Modèles de mélanges et interactions spatiales pour les données de grande dimension
- Dynamisme, visibilité et opportunités INRIA
- Ouverture supplémentaire vers les applications

• Apports et spécificités SMS/ex Labsad:

- Approches complémentaires en inférence fonctionnelle et multi résolution
- Outils pointus pour l'étude théorique d'algorithmes et méthodes de simulation
- Graphes aléatoires
- Richesse, variété et solidité des compétences statistiques

Positionnement national (suite)

- **INRA, Avignon, unité de biométrie** (6 permanents): statistiques spatiales, 1 ancien membre, Réseau INRA, co-publications
- **Heudiasyc, UT Compiègne, équipe Astrid** (14 permanents): reconnaissance de formes statistique: ACI masses de données
- **Département TSI, ENST, équipe TSAC** (15 permanents): modèles de Markov cachés, algorithmes stochastiques, 1 collaborateur extérieur de Mistis, co-publication
- **Groupe SSB [INRA, Génopole Evry, INA-PG, U. Rouen, U. Toulouse,** 30 permanents]: statistique appliquée aux données génomiques, ARC IBN, ChromoNet
- **Montpellier 2, équipe proba-stat** (18 permanents): estimation fonctionnelle, 2 anciens membres, 2 collaborateurs, co-publications
- **Toulouse 3, groupe MAFIA** (12 permanents): stat. non paramétrique, algorithmes stochastiques, dépendance, aspects théoriques.
- **Paris 6, LSTA** (20 permanents): stat. non paramétrique, stat. bayésiennes, aspects fondamentaux, 1 collaborateur, co-publications

Positionnement international

Europe:

Département de statistique, Glasgow, Ecosse (15 permanents): Statistiques spatiales, modèles markoviens, méthodes variationnelles, **prof. invité, co-publications**

Centre de Statistique et Applications (CEAUL), Lisbonne, Portugal (30 permanents): statistiques spatiales, classification, extrêmes

Etats-unis:

Département de statistique, U. Washington, Seattle, USA (30 permanents): classification, **étroite collaboration, co-publications**

Départments de statistique Purdue et Berkeley (40 permanents): moins focalisés

Australie:

Programme Statistique du Mathematical Sciences Institut à Camberra

(10 permanents): modèles de mélanges, statistiques spatiales, estimation de frontières, réduction de dimension

Collaborations industrielles

CEA

- Ancien contrat avec le CEA de Cadarache, Laboratoire de Conduite et Fiabilité des Réacteurs (C. Lavergne et thèse de J.Jacques)
- Projet de nouvelle collaboration avec N. Devictor: Post-doc en [Fiabilité et traitement statistique des incertitudes](#)
- Contact avec le CEA de Grenoble, A. de Crecy: [algorithmes EM dans les processus physiques](#)

TRIXELL (groupe Thales)

- Première réunion avec J.M. Vignolle autour de [l'analyse statistique d'images](#)

Xerox

- Contrat CIFRE pour l'encadrement de L. Donini (S. Girard, J.B. Durand). [Analyse de logs d'imprimantes](#)