

**Curriculum Vitæ**  
**Travaux et programme de recherche**

**Pierre Vandekerkhove**

**Université Paris-Est Marne-la-Vallée**  
Laboratoire d'Analyse et de Mathématiques Appliquées  
5, Bd. Descartes, Champs-sur-Marne  
77454 Marne-la-Vallée CEDEX 2.

Tel. 01 60 95 75 25

Fax 01 60 95 75 45

Courriel : [pierre.vandek@univ-mlv.fr](mailto:pierre.vandek@univ-mlv.fr)

Web : <http://perso-math.univ-mlv.fr/users/vandekerkhove.pierre>

# Table des matières

<b>1</b>	<b>Curriculum Vitæ</b>	<b>2</b>
1.1	État civil et situation actuelle . . . . .	2
1.2	Cursus et diplômes . . . . .	2
1.3	Domaines de recherche . . . . .	3
1.4	Production scientifique . . . . .	3
1.5	Encadrement de stages . . . . .	3
1.6	Diffusion scientifique et reconnaissances . . . . .	4
1.7	Projets ANR blancs : Mixture, BigMC . . . . .	4
1.8	Collaborations et visites de laboratoires étrangers . . . . .	5
1.9	Responsabilités administratives . . . . .	6
<b>2</b>	<b>Enseignements</b>	<b>6</b>
<b>3</b>	<b>Publications et conférences</b>	<b>7</b>
3.1	Travaux publiés . . . . .	7
3.2	Articles soumis ou en cours de rédaction . . . . .	7
3.3	Articles non publiés . . . . .	8
3.4	Conférences nationales et internationales . . . . .	8
<b>4</b>	<b>Travaux de recherche</b>	<b>8</b>
4.1	Modèles de Markov cachés . . . . .	9
4.2	Modèles de mélange semi-paramétriques . . . . .	10
4.3	Méthodes de Monte Carlo . . . . .	11
4.4	Algorithmes stochastiques à pas décroissant . . . . .	12
<b>5</b>	<b>Programme de recherche</b>	<b>13</b>
5.1	Algorithme EM semi-paramétrique . . . . .	13
5.2	Compression de structure pour les lois en grande dimension . . . . .	13
5.3	Mélange semi-paramétrique de régressions . . . . .	13
5.4	Robustesse des approches semi-paramétriques dans le cadre des modèles à données manquantes . . . . .	14

# 1 Curriculum Vitæ

## 1.1 État civil et situation actuelle

Pierre VANDEKERKHOVE  
39 ans, né le 28 novembre 1969 à Melilla (Espagne)  
Maître de conférences  
Français, marié, 1 enfant.

Université Paris-Est Marne-la-Vallée  
Laboratoire d'Analyse et de Mathématiques Appliquées  
5, Bd. Descartes, Champs-sur-Marne  
77454 Marne-la-Vallée CEDEX 2.

Tel. 01 60 95 75 25

Portable 06 22 65 02 87

Fax 01 60 95 75 45

Courriel : pierre.vandek@univ-mlv.fr

Web : <http://perso-math.univ-mlv.fr/users/vandekerkhove.pierre>

## 1.2 Cursus et diplômes

Maître de conférences (section 26 du CNU) depuis septembre 1998 au Laboratoire d'Analyse et de Mathématiques Appliquées (UMR-CNRS 8050) de l'université Paris-Est Marne-la-Vallée.

### FORMATION ET TITRES

**2007** Habilitation à diriger des recherches de l'université Paris-Est Marne-la-Vallée.

Titre : *Contribution à l'étude statistique des modèles à données manquantes et apprentissage statistique autour des modèles markoviens.*

Rapporteurs : P. Bertail (Nanterre), B. G. Lindsay (université de Penn State), E. Moulines (ENST).

Jury : P. Bertail (Nanterre), J.-F. Delmas (ENPC), P. Del Moral (INRIA Bordeaux), E. Gassiat (Orsay), M. Hoffmann (Marne-la-Vallée), D. Lambertson (Marne-la-Vallée), E. Moulines (ENST).

**1997-98** Post-doc à l'université d'Économie de Pavie (Italie) dans le cadre du réseau européen Training and Mobility Researcher.

**1997** Doctorat en Statistique de l'université Montpellier II, sous la direction de X. Milhaud.

Titre : *Identification de l'ordre des processus ARMA stables et contribution à l'étude statistique des chaînes de Markov cachées.*

Rapporteurs : M. Duflo (Marne-la-Vallée), E. Gassiat (Évry).

Jury : D. Bakry (Toulouse), A. Berlinet (Montpellier), D. Dacunha-Castelle (Orsay), G. Ducharme (Montpellier), M. Duflo (Marne-la-Vallée), E. Gassiat (Évry), L. Miclo (Toulouse), X. Milhaud (Montpellier).

**1992** DEA de Biostatistique de l'université Montpellier II, sous la direction de X. Milhaud et H. Friedman (Stanford).

Titre : *Sensibilité de l'algorithme ACE (Alternating Conditional Expectation).*

**1991** Maîtrise de Mathématiques Appliquées, option Topologie Algébrique, de l'université Montpellier II.

## EXPÉRIENCES PROFESSIONNELLES

**1996-97** ATER (demi-poste) à l'université Paul Sabatier de Toulouse.

**1994-96** Allocataire-Moniteur à l'université Paul Sabatier de Toulouse.

**1993-94** Allocataire-Moniteur à l'université Montpellier II.

**1992** Stage de DEA de 3 mois effectué à l'INRA de Montpellier.

### 1.3 Domaines de recherche

Mes travaux de recherche portent sur trois domaines distincts :

- les algorithmes stochastiques : critères de convergence, ergodicité.
- les méthodes de Monte Carlo : chaînes de Markov, contrôle et optimisation des vitesses de convergence, choix de stratégie, réduction de la variance, détection automatique de forme.
- les modèles à données manquantes : modèles de Markov cachés, mélanges semi-paramétriques, problèmes inverses, tests, applications en Biologie et à la Fiabilité.

### 1.4 Production scientifique

- 12 publications ;
- 4 articles soumis ou en cours de rédaction ;
- 2 preprints non publiés ;
- 2 conférences internationales sur invitation ;
- 3 séminaires dans des laboratoires étrangers ;
- 9 conférences nationales ;
- 22 séminaires effectués dans diverses universités françaises depuis 1997 : Séminaire Parisien de Statistique (IHP), CREST-INSEE, université de Marne-la-Vallée, université Paul Sabatier (Toulouse), université Paris 11 (Orsay), université Montpellier II, ENSAI Rennes, UTC de Compiègne, université Paris 13 (Villetaneuse), Génopôle d'Evry, université Lille 1, université de Pau Pays de l'Adour, Agro Paris-Tech.

### 1.5 Encadrement de stages

**2007** Encadrement doctoral du stage de Master 2 Mathématiques et Applications de M. Diouf Kora Sally.

Titre : *Filtres particuliers (lecture du chapitre 7 du livre de Cappé, Moulines et Rydén, 2005).*

**2007** Encadrement du stage de Master 1 Mathématiques et Applications de Mlle Cécilia Lavannant.

Titre : *Réduction de la variance pour l'Importance Sampling au moyen de lois instrumentales corrigées non-paramétriquement par la loi cible.*

**2006** Encadrement du stage de Master 1 Mathématiques et Applications de Mlle Marieme Ba.

Titre : *Modèles de mélange et algorithme EM.*

**2005** Encadrement du stage de Master 1 Mathématiques et Applications de M. Thimothée Mbogtjama.

Titre : *Convergence de l'algorithme de Robbins-Monro (sur la base d'un devoir en temps libre proposé par Dominique Bakry).*

**2003** Encadrement du stage de Maîtrise de Mathématiques de Mlle Fatiha Benakli.

Titre : *Théorème Central limite pour des variables aléatoires  $\alpha$ -mélangeantes (d'après le livre de Billingsley, 1995).*

**2002** Encadrement du stage de Maîtrise de Mathématiques de Mlle Audrey Leeman.  
Titre : *Convergence de l'algorithme de Hastings-Metropolis. Etude numérique de la sensibilité à la loi instrumentale.*

## 1.6 Diffusion scientifique et reconnaissances

- Organisation d'un workshop à Pau les 23-24 juin 2008 sur les approches semi-ou non-paramétriques pour les modèles à données manquantes.  
Site web <http://lma-umr5142.univ-pau.fr/live/actualites/00-01+-+Archives>
- Organisation des Journées de Statistique sur les Modèles à Données Manquantes du 13-14 janvier 2005 à l'université de Marne-la-Vallée, en collaboration avec Laurent Bordes et Didier Chauveau.  
Site web <http://congres-math.univ-mlv.fr/a14fe87d/index.html>
- Prime d'Encadrement Doctoral et de Recherche : depuis septembre 2006.
- Délégation CNRS de 6 mois en 2005.
- Commissions de spécialistes : université de Technologie de Compiègne (2005-07), université de Marne-la-Vallée (Vice président jusqu'en 2006).
- Lecture ou arbitrage d'articles pour : *J. Statist. Plann. Inference, Scand. J. Statist.; Bernoulli; ESAIM P&S; Statistica Sinica; Comm. Statist. Theory and Method; CRAS Paris; Maghreb Math. Rev.*

## 1.7 Projets ANR blancs : Mixture, BigMC

### PORTEUR DU PROJET MIXTURE (PROJET BLANC 2009)

- L'objectif de ce projet ANR blanc d'une durée de 4 ans, dont je suis le principal coordinateur, est de traiter la question des modèles de mélanges, ou plus généralement des modèles à données manquantes, sous l'angle semi- ou non-paramétrique.
- Ce projet regroupe 4 partenaires principaux : l'université Paris-Est Marne-la-Vallée, le Génomètre d'Evry, l'université d'Orléans, l'université de Pau, et l'université de Penn State (États-Unis) au titre de partenaire associé.
- Ce projet ANR implique 3 professeurs, 5 maîtres de conférences, 1 chargé de recherche CNRS.

Le programme scientifique de cette ANR se découpe en 3 tâches :

**Tâche n°1** (coordinateur P. Vandekerckhove) : exploration des fondements mathématiques (identifiabilité, choix de modèle, vitesses minimax, efficacité, convergence de l'algorithme EM semi-paramétrique).

**Tâche n°2** (coordinatrice C. Matias) : perspectives et extensions (modèles de Markov cachés, mélanges de régression, mélanges de graphes aléatoires).

**Tâche n°3** (coordinateur D. Chauveau) : applications et valorisation informatique (puces ADN, données écologiques, conception de packages R).

La demande financière pour ce projet s'élève à 267 487euros, cette somme incluant le financement d'une thèse et de 6 mois de contrat pour un ingénieur d'étude.

### PARTICIPATION AU PROJET BIGMC (PROJET BLANC 2008 FINANCÉ)

#### COORGANISATEUR DU SÉMINAIRE

- Ce projet fait suite à un projet ANR blanc qui vient de s'achever sur les méthodes de Monte Carlo adaptatives. Il s'agit dans ce nouveau projet, tout en continuant l'exploration des méthodes adaptatives, de mettre l'accent sur les problèmes de simulation en grande dimension (lois de probabilité, fonctionnelles de processus stochastiques, problèmes issus de la Physique).

- Ce projet regroupe 3 partenaires : Institut TELECOM, CEREMADE (Dauphine), CERMICS (ENPC, université Paris-Est).
- Ce projet ANR implique 4 professeurs, 4 maîtres de conférences, 2 chargés de recherche (CNRS, INRIA).

Le programme scientifique de cette ANR se découpe en 4 tâches :

**Tâche n°1** (coordinateur R. Douc) : proposition et étude de méthodes dites "Population Monte Carlo" dans des problèmes en grande dimension.

**Tâche n°2** (coordinatrice G. Fort) : proposition et étude de méthodes de Monte Carlo par chaînes de Markov (MCMC) pour des problèmes en grande dimension.

**Tâche n°3** (coordinateur C. Robert) : étude des problèmes liés à la simulation à partir de lois approchées.

**Tâche n°4** (coordinateur B. Jourdain) : applications aux diffusions et à la dynamique moléculaire.

La demande financière pour ce projet s'élève à 237 536 euros, cette somme incluant le financement de 3 post-doc.

## 1.8 Collaborations et visites de laboratoires étrangers

### COLLABORATIONS NATIONALES

- Laurent Bordes (université de Pau) : modèles de mélanges semi-paramétriques et Chaînes de Markov Cachées (CMC).
- Didier Chauveau (université d'Orléans) : algorithmes MCMC et algorithmes de type EM pour les mélanges semi-paramétriques.
- Céline Delmas (INRA) : modèles de mélange avec application aux puces ADN.
- Stéphane Mottelet (université de technologie de Compiègne) : modèles de mélanges semi-paramétriques.
- Nadia Oudjane (EDF-Clamart, Dauphine, Villetaneuse) : méthodes de Monte Carlo.

### COLLABORATIONS INTERNATIONALES

- Paolo Giudici (université de Pavie, Italie) : modèles graphiques et CMC.
- David Hunter (Penn State university, États-Unis) : mélanges semi-paramétrique.
- Tobias Rydén (université de Lund, Suède) : modèles graphiques et CMC.
- Pierre Tarrès (université d'Oxford, Grande-Bretagne) : algorithme du bandit à deux bras.

### VISITE DE LABORATOIRES ÉTRANGERS

- **Penn State** university (États-Unis) : visite de 2 semaines prévue été 2009, sur invitation de Thomas Hettmansperger et David Hunter.
- **Oxford** : institut de Mathématiques d'Oxford (Grande-Bretagne), sur invitation de Pierre Tarrès (2 semaines en janvier 2006).
- **Stanford** : département de Statistique de l'université de Stanford (Sequoia Hall, États Unis) sur invitation de Susan Holmes (2 semaines sur septembre et octobre 2005).
- **Cambridge** : département du Signal de Cambridge (Grande Bretagne) sur invitation d'Arnaud Doucet et de Christophe Andrieu (2 semaines en septembre 1999).
- **Lund** : centre des sciences Mathématiques de Lund (Suède) sur invitation de Tobias Rydén (2 semaines en janvier 1998).
- **Pavie** : département d'Économie et d'études quantitatives de l'université de Pavie (Italie) dans le cadre de mon post-doc (1 an 1997-98).

## 1.9 Responsabilités administratives

**2001-05** Membre élu du bureau de Laboratoire Marne-la-Vallée/Créteil.

**2000-06** Co-responsable de la bibliothèque.

**2000-06** Correspondant de la SMAI.

## 2 Enseignements

**1993-94** Allocataire-Moniteur à l'université Montpellier II.

**1994-95** Allocataire-Moniteur à l'université Montpellier II.

**1996** ATER (mi-temps) de 1996-97 à l'université Paul Sabatier de Toulouse.

### ENSEIGNEMENTS ACTUELS

**2006-09** Cours, en partenariat avec J.F. Delmas (ENPC), sur les *Modèles Aléatoires et Applications* en **Master 2** Mathématiques et Applications (20 heures).

Thèmes :

- analyse des zones codantes dans l'ADN ;
- chaînes de Markov cachées ;
- algorithme EM ;
- contrôle optimal ;
- processus de sauts, file d'attente ;
- recuit simulé, algorithme de Hastings Metropolis.

**2006-09** TD de *Processus Stochastiques* en **Master 1** Mathématiques et Applications (24 heures).

**2002-09** TD de *Statistique mathématique* en **Master 1** IMIS et Mathématiques et Applications (18 heures).

**2003-09** Cours (amphithéâtre) et TD de *Calculus* en **L1** (60 heures).

**2006-09** TD d'*Algèbre Linéaire* en **L1** (36 heures).

**2008** Cours et TD de *Probabilité* en **L2** option Maths-Info (45 heures).

### ENSEIGNEMENTS DÉJÀ EFFECTUÉS

- Cours et TD de *Probabilité* (45 heures) en **L3**-Maths durant 6 années.
- TD et TP de *Statistique Empirique et Analyse des Données* (40 heures) sous **SAS** et **MATLAB** en **Master 1** IMIS et Mathématiques et Applications durant 5 années.
- Cours et TD de *Probabilité* (20 heures) pour la filière **Ingénieur 2000** durant 2 années.
- TD de *Probabilité* (18 heures) en **L2** option Physique durant 3 années.
- TD de *Calcul différentiel* (36 heures) en **L2** option Physique durant 2 années.
- TD d'*Analyse* (36 heures) en **L2** option Physique durant 2 années.
- TD d'*Algèbre linéaire* en **L2** option Maths durant 4 années.

### ENSEIGNEMENTS EN ÉCOLE D'INGÉNIEUR

**2002-06** Cours et TP (**Scilab**) de *Statistique* (28 heures) en deuxième année à l'École Nationale des Ponts et Chaussées.

**2007-09** Cours (amphithéâtre) et TD d'*Analyse* ( $2 \times 30$  heures) en première année de prépa intégrée à l'École Supérieure d'Ingénierie en Électronique et Électrotechnique (ESIEE).

**2007-08** Cours et TD de mise à niveau en *Analyse* (45 heures) pour les intégrants en 3-ème année à l'ESIEE.

### 3 Publications et conférences

#### 3.1 Travaux publiés

- [A1] Bakry, D., Milhaud, X. et Vandekerkhove, P. (1997). Statistics of Hidden Markov chains with finite state space. The nonstationary case. *C. R. Acad. Sci. Paris*, Série I, 203–206.
- [A2] Vandekerkhove, P. (1998). Simulated annealing with a sequential estimator of the energy. *C. R. Acad. Sci. Paris*, Série I, 1003–1006.
- [A3] Chauveau, D. et Vandekerkhove, P. (1999). Un algorithme de Hastings-Metropolis avec apprentissage séquentiel. *C. R. Acad. Sci. Paris*, Série I, 173–176.
- [A4] Giudici, P., Rydén, T. et Vandekerkhove, P. (2000). Likelihood-Ratio Tests for Hidden Markov Models. *Biometrics*, **56**, 742–747.
- [A5] Chauveau, D. et Vandekerkhove, P. (2001). Algorithmes de Hastings-Metropolis en interaction. *C. R. Acad. Sci. Paris*, Série I, 881–884.
- [A6] Chauveau, D. et Vandekerkhove, P. (2002). Improving convergence of the Hastings-Metropolis Algorithm with a learning proposal. *Scand. J. Statist.*, **28**, 13–29.
- [A7] Bordes, L. et Vandekerkhove, P. (2005). Statistical inference for Partially Hidden Markov Models. *Communications in Statistics*, **34**, 1081–1104.
- [A8] Vandekerkhove, P. (2005). Consistent et asymptotically normal estimates for hidden Markov mixtures of Markov models. *Bernoulli*, **11**, 103–129.
- [A9] Bordes, L., Mottelet, S. et Vandekerkhove, P. (2006). Semiparametric estimation of a two component mixture model. *Ann. Statist.*, **34**, 1204–1232.
- [A10] Bordes, L., Delmas, C. et P. Vandekerkhove. (2006). Semiparametric estimation of a two-component mixture model where a component is known. *Scand. J. Statist.*, **33**, 733–752.
- [A11] Chauveau, D. et Vandekerkhove, P. (2007). A Monte Carlo estimation of the entropy for Markov chains. *Methodology et Computing in Applied Probability*, **9**, 133–149.
- [A12] Bordes, L., Chauveau, D. et Vandekerkhove, P. (2007). A stochastic EM algorithm for a semiparametric mixture model. *Computational Statistics and Data Analysis*, **51**, 5429–5443.

#### 3.2 Articles soumis ou en cours de rédaction

- [B1] Bordes, L. et Vandekerkhove, P. (2009). Semiparametric two-component mixture model with a known component : a class of asymptotically normal estimators. *Soumis à JSPI*.
- [B2] Tarrès, P. et Vandekerkhove, P. (2009). On the ergodic two-armed bandit. *Soumis à Ann. Appl. Probab.*
- [B3] Chauveau, D. et Vandekerkhove, P. (2009). Simulation-based entropy estimation for comparing MCMC algorithms. *Soumis à Appl. Math. and Optim.*

[B4] Chauveau, D., Oudjane, N., et Vandekerkhove, P. (2008). Weighted kernel resampling method : super CLT and applications. *En préparation*.

**N.B.** : les pièces [A1-12], ainsi que [B1-3] sont jointes au dossier.

### 3.3 Articles non publiés

[C1] Vandekerkhove, P. (1998). A sequential Metropolis-Hastings algorithm. *universita di Pavia, Quaderni di Dipartimento # 80*.

[C2] Vandekerkhove, P. (1998). A new simulated annealing method. *universita di Pavia, Quaderni di Dipartimento # 79*.

### 3.4 Conférences nationales et internationales

- XXVII-èmes Journées de Statistique, Laval (Québec), 25-29 mai 1996. *Consistance du critère ODQ dans le cas d'un ARMA stable*.
- Congrès Franco-Tunisien, Toulouse, mai 1996. *Consistance du critère ODQ dans le cas d'un ARMA stable*.
- SMAI, Toulouse, septembre 1996. *Statistique de chaînes de Markov cachées à états finis. Le cas non-stationnaire*.
- XXVIII-èmes Journées de Statistique, Carcassonne, mai 1997. *Estimation des paramètres d'une chaîne de Markov cachée par recuit simulé*.
- Spatial Computational Statistics Workshop (Aussois, janvier 1998). *On hidden Multivariate Markov Models identification*.
- XXXI-èmes Journées de Statistique, Grenoble, mai 1999. *Un algorithme de Hastings-Metropolis avec apprentissage séquentiel*.
- TMR Second on Spatial and Computational Statistics (Crète, juin 1999). *A Hastings-Metropolis algorithm with learning proposal*.
- XXXII-èmes Journées de Statistique, Fès (Maroc), mai 2000. *Comparaison de vitesse de convergence d'algorithmes basée sur l'entropie*.
- XXXV-èmes Journées de Statistique, Lyon, 2-6 juin 2003. *Inférence statistique pour des modèles de Markov partiellement caché*.
- XXXVI-èmes Journées de Statistique, Montpellier, 24-28 mai 2004. *Estimation semi-paramétrique d'un mélange*.
- XXXVII-èmes Journées de Statistique, Pau, 6-10 juin 2005. *Estimation semi-paramétrique d'un mélange à deux composantes dont une composante est connue*.

## 4 Travaux de recherche

Je présente dans ce qui suit les travaux académiques réalisés durant mon parcours de recherche. Dans un souci de clarté, je tiens à préciser que les travaux [A1] et [A2] découlent directement de ma thèse de doctorat et que [A4] a été obtenu durant mon post-doc.

Mes travaux s'articulent autour de quatre thèmes principaux : les modèles de Markov cachés ; les modèles de mélange semi-paramétriques ; l'optimisation des méthodes de Monte Carlo (et leur utilisation pour l'étude des chaînes de Markov) ; et enfin les algorithmes stochastiques à pas décroissant.

## 4.1 Modèles de Markov cachés

**Mots clés :** maximum de vraisemblance, (non)-stationnarité, test du rapport de vraisemblance, mélanges de processus.

**Applications :** données de pollution, Fiabilité, analyse des signaux EEG, canaux d'ions.

Dans cette première partie, je m'intéresse à diverses questions en lien avec les modèles de Markov cachés (MMC). La première question concerne l'hypothèse de stationnarité dans l'étude de l'estimateur du maximum de vraisemblance (EMV) pour les chaînes de Markov cachées (CMC). Baum et Petrie (1966) étudient l'EMV pour les CMC en montrant que la log-vraisemblance normalisée (par la taille de l'échantillon) d'une CMC a le même comportement asymptotique que la moyenne empirique du log des probabilités conditionnelles saturées (probabilité conditionnelle de l'observation sachant le passé infini de la chaîne observée). Ces auteurs utilisent alors l'hypothèse de stationnarité pour en déduire que les probabilités conditionnelles saturées constituent une famille de variables aléatoires invariantes en loi sous l'opérateur retard. Cette remarque leur permet en effet d'établir que, pour toute valeur du paramètre, la log-vraisemblance normalisée des CMC vérifie le théorème ergodique et converge presque sûrement vers une quantité (fonction du paramètre) appelée *entropie* et satisfaisant la propriété dite de *contraste*. L'objet du travail, fait durant ma thèse de doctorat en collaboration avec Dominique Bakry et Xavier Milhaud [A1], consiste essentiellement à montrer que l'étude menée par Baum et Petrie peut s'étendre au cas des CMC partant d'une condition initiale déterministe. L'argument clé de la preuve pour ce travail est l'utilisation d'une technique de couplage permettant de comparer efficacement le comportement de la log-vraisemblance d'une CMC sous une condition initiale arbitraire, avec celle d'une CMC démarrant sous le régime stationnaire. Nous montrons aussi dans [A1] la propriété LAN (local asymptotic normality) pour les CMC.

À la suite de ce travail, j'ai eu l'opportunité de travailler (dans le cadre mon post-doc) avec Paolo Giudici et Tobias Rydén [A4], sur des problèmes de test pour les MMC. En utilisant les travaux de Bickel *et al.* (1998) et certains résultats d'identifiabilité pour les familles de mélanges multivariés, nous établissons des tests de type rapport de vraisemblance permettant de traiter des hypothèses ponctuelles ou composites sur les paramètres du modèle. Nous appliquons de plus notre travail sur des données réelles de pollution. Les modèles utilisés sont alors des MMC graphiques sur lesquels il s'agit de tester des zéros dans l'inverse de la matrice de corrélation des vecteurs observés (supposés gaussiens conditionnellement à la chaîne sous-jacente).

Avec Laurent Bordes [A7], je me suis intéressé à un modèle de Markov partiellement caché trouvant un intérêt naturel dans le domaine de la Fiabilité. Il s'agit d'un MMC pour lequel l'information portant sur l'atteinte d'un état fixé de la chaîne sous-jacente est systématiquement connue. Une telle situation se rencontre par exemple lorsqu'un système est en fonctionnement et que l'on souhaite faire de l'inférence sur son modèle de dégradation (supposé markovien) au travers de certaines variables observables (température, niveau vibratoire, etc.). En cas de panne nous pouvons considérer que nous avons atteint de manière sûre le pire état de dégradation du système. Nous montrons pour ce modèle la consistance et la normalité asymptotique de l'EMV sous des conditions plus faibles que celles exigées pour les MMC classiques. Notons en particulier que l'hypothèse d'apériodicité, cruciale dans l'étude de Baum et Petrie (1966) et dans tous les travaux qui ont suivi sur l'inférence des MMC, n'est ici plus requise.

Je conclus enfin cette partie avec l'introduction et l'étude, faite dans [A8], d'une nouvelle

classe de processus à données manquantes. Il s’agit des mélanges Markoviens de processus de Markov dont la définition consiste en la mise bout à bout de trajectoires de processus de Markov mutuellement indépendants, le choix des trajectoires s’opérant au moyen d’une chaîne de Markov non observée. En raison de la grande complexité de la vraisemblance associée à ce modèle et des nombreuses impasses techniques qu’elle engendre, je me suis intéressé à l’estimateur du maximum de vraisemblance des données tronquées (EMVT) introduit par Rydén (1994). Je montre, sous des conditions standards d’identifiabilité, de régularité et de mélangeance des processus, la consistance et la normalité asymptotique de l’EMVT. Une des principales difficultés associées à ce type de modèle étant la paramétrisation des densités des mesures invariantes associées à chaque processus, j’indique une procédure de Monte Carlo permettant de les estimer ponctuellement. Cette étape cruciale permet de calculer en pratique la vraisemblance des données tronquées pour toute valeur du paramètre. Je montre d’autre part que les hypothèses assurant la consistance et la normalité asymptotique de l’EMVT sont satisfaites dans le cadre de mélanges de processus autoregressifs d’ordre 1 gaussiens.

## 4.2 Modèles de mélange semi-paramétriques

**Mots clés :** mélange, identifiabilité semi-paramétrique, contraste, algorithme EM, TCL, tests.

**Applications :** puces ADN, éruption de geyser.

Dans cette deuxième partie, je présente diverses contributions à l’étude des modèles de mélanges semi-paramétriques. Les modèles auxquels nous nous intéressons ont été inspirés par Hall et Zhou (2004). Le premier modèle, étudié en collaboration avec Stéphane Mottelet et Laurent Bordes [A9], est un modèle de mélange à deux composantes symétriques égales à un paramètre de localisation près. Le deuxième (utilisé dans l’analyse des puces ADN), étudié en collaboration avec Céline Delmas et Laurent Bordes [A10], correspond à un mélange à deux composantes dont l’une est connue et l’autre est simplement supposée symétrique autour d’un paramètre de localisation inconnu. Nous abordons en détail le problème de l’identifiabilité et proposons des méthodes d’estimation des paramètres euclidiens pour ces deux modèles. L’existence de formules d’inversion permettant d’isoler la loi des composantes inconnues, couplée avec l’hypothèse de symétrie, nous permettent d’exhiber des mesures de discrédance sur l’espace des paramètres, induisant ainsi des procédures d’estimation naturelles de type *minimum de contraste*. Une utilisation ”plug-in” des formules d’inversion permettent alors de reconstituer les paramètres fonctionnels (fonction de répartition et densité) inconnus des modèles. Nous montrons la consistance de nos procédures d’estimation ainsi que certaines vitesses de convergence presque sûres. Dans un travail en cours [B1], nous montrons pour le deuxième modèle, un théorème central limite (TCL) fonctionnel associé à l’estimateur du vecteur des paramètres constitué par : la proportion du mélange, le paramètre de localisation, et la fonction de répartition de la composante inconnue. Ce dernier résultat nous permet de développer des procédures de test concernant des hypothèses ponctuelles sur les paramètres euclidiens ainsi que l’hypothèse de symétrie sur la composante inconnue. Nous appliquons respectivement les modèles précédents à des données de pluviométrie et à un problème réel de comparaison de gestation chez les bovins issus de l’insémination *artificielle* ou *in vitro*.

À la fin de cette partie, je présente un travail concernant les aspects algorithmiques de l’estimation du premier modèle. Ce travail, réalisé en collaboration avec Didier Chauveau et Laurent Bordes [A12], propose d’adapter l’algorithme EM (Expectation/Maximisation) classique en y ajoutant une étape d’estimation de la densité inconnue. Nous donnons une expli-

cation heuristique à cette approche et montrons par des simulations que cette méthodologie donne de très bons résultats avec des temps de calculs beaucoup moins longs que ceux générés par des méthodes d'optimisation classiques.

### 4.3 Méthodes de Monte Carlo

**Mots clés :** accélération de convergence, réduction de variance, entropie, estimation non-paramétrique, ergodicité.

**Applications :** Hastings Metropolis, échantillonneur de Gibbs, échantillonnage d'importance, stabilité markovienne, vitesse dans le TCL, analyse de forme.

Dans cette troisième partie je présente une série de travaux, tous effectués en collaboration avec Didier Chauveau, concernant l'amélioration de certaines méthodes de Monte Carlo par chaîne de Markov (MCMC). La première approche que nous avons envisagée dans [A6], avait pour but d'accélérer la vitesse de convergence de l'algorithme dit de Hastings-Metropolis (HM). L'algorithme de HM génère, au moyen d'un mécanisme d'acceptation/rejet sur des données simulées au moyen d'une loi instrumentale, une chaîne de Markov dont la loi invariante a pour densité une densité souhaitée. Divers auteurs, comme Menegersen et Tweedie (1996) ou Holden (1998), ont mis en évidence les liens entre la vitesse de convergence de cet algorithme et la proximité entre la loi cible et la loi instrumentale. Étant donné la convergence (même lente) des densités de l'algorithme de HM vers la densité cible, il nous est apparu intéressant d'estimer non-paramétriquement ces densités en générant plusieurs algorithmes de HM en parallèle (i.i.d.) et de les exploiter à leur tour comme densités de nouvelles lois instrumentales (facilement simulables en raison de la structure de mélange des estimateurs à noyaux). Nous montrons, dans le cadre de densités à support compact, et pour un nombre à priori aussi grand que l'on veut d'algorithmes en parallèle, que notre procédure génère des algorithmes asymptotiquement plus rapides que tout algorithme de HM utilisant une loi instrumentale arbitraire. Nous présentons des simulations réalisées en dimension 1 et 2 illustrant le rendement important de ce type d'approche face à des algorithmes de HM standards même convenablement calibrés.

La deuxième contribution au domaine des MCMC, réalisée dans [B3], porte sur une méthode destinée à hiérarchiser l'efficacité de diverses approches/stratégies face à un problème de simulation par chaîne de Markov. Considérons par exemple deux algorithmes de MCMC ayant pour but de simuler la même loi. Il est alors intéressant de connaître l'algorithme qui converge le plus rapidement vers sa loi stationnaire. Pour répondre à ce type de question nous devrions être en mesure d'estimer une distance entre la densité des algorithmes et la densité cible, or cette dernière n'est en général connue qu'à une constante de normalisation près (cadre bayésien). On peut cependant remarquer que la différence des distances de Kullback entre la densité des algorithmes et la densité cible est indépendante de cette constante de normalisation ; ainsi l'estimation de ces différences et l'analyse de leur comportement au cours des premières itérations pourrait s'avérer un outil synthétique permettant de mieux appréhender la qualité relative de chaque algorithme. Partant de ce constat, nous avons choisi d'estimer ces différences de distance de Kullback au moyen de plusieurs algorithmes lancés en parallèle. La méthode d'estimation utilisée repose sur l'estimateur de l'entropie introduit par Györfi et Van Der Meulen (1989). La principale difficulté de notre travail a été de montrer que les conditions techniques assurant la consistance des estimateurs étaient satisfaites pour les densités successives de l'algorithme de Hastings-Metropolis au prix de certaines hypothèses (toutes vérifiées dans le cas gaussien).

Nous avons enfin un travail en cours (en collaboration avec Nadia Oudjane) sur une

procédure permettant d'améliorer les performances de la méthode d'*échantillonnage d'importance* (traduction de *importance sampling*). L'échantillonnage d'importance permet de calculer des espérances de fonctions test au sens d'une densité connue à une constante de normalisation près. Le principe de cette méthode utilise un rapport de loi forte des grands nombres portant sur des fonctionnelles d'échantillons instrumentaux judicieusement choisies. Comme pour l'algorithme de HM, il est admis que la qualité de l'estimation (variance) par échantillonnage d'importance est liée à la ressemblance entre la loi instrumentale et la loi cible. Afin d'augmenter cette ressemblance, nous proposons un estimateur à noyau de la loi cible utilisant l'échantillon instrumental, la connaissance de la densité instrumentale, et le numérateur de la densité cible. L'étape de rééchantillonnage, au sens de cet estimateur à noyau repondéré, peut s'apparenter à une version lissée de la méthode de rééchantillonnage par poids d'importance (voir, *e.g.* Cappé *et al.*, 2005), couramment utilisée en filtrage particulaire ou dans les approches bayésiennes. Nous montrons que cette méthode converge à une vitesse (exprimée par un TCL) plus rapide que  $\sqrt{N}$ , dont la normalisation indexée par la dimension tend cependant vers  $\sqrt{N}$  lorsque la dimension tend vers l'infini.

Dans le cadre de ce travail une méthodologie très efficace a été mise à jour permettant de détecter automatiquement les "zones de masse" de la loi cible (en chaînant plusieurs fois les étapes d'estimation et de rééchantillonnage décrites plus haut). Nous pensons que cette dernière approche devrait trouver des applications naturelles en analyse et compression d'image (les *niveaux de gris* pouvant clairement s'apparenter à des *niveaux de vraisemblance* de la loi cible).

Nous concluons cette partie consacrée aux méthodes de Monte Carlo, par un travail [A11] concernant l'estimation de l'entropie des chaînes de Markov. Étant donné le noyau de transition d'une chaîne de Markov (supposé simulable), on s'intéresse à la possibilité éventuelle de pouvoir représenter de manière consistante l'évolution dans le temps de la distance de Kullback entre les lois de deux chaînes partant de conditions initiales différentes, ou entre la loi d'une chaîne et sa loi stationnaire, lorsque celle-ci est connue (et simulable). Nous proposons pour cela un estimateur de type double Monte Carlo permettant d'estimer l'entropie de la chaîne contre sa concurrente ou bien sa loi stationnaire. Nous montrons la consistance et la normalité de nos estimateurs sous des conditions faibles de moments. Nous mettons enfin en oeuvre notre méthode pour tester la stabilité de processus autoregressifs d'ordre 1, et pour évaluer la vitesse de convergence (au sens de Kullback) dans le TCL pour des échantillons i.i.d. suivant diverses lois (Student, Uniforme, etc.).

## 4.4 Algorithmes stochastiques à pas décroissant

**Mots clés :** recuit simulé, apprentissage, source ergodique.

**Applications :** décision séquentielle, Finance.

Cette dernière partie est consacrée à la convergence de deux algorithmes stochastiques à pas décroissant. Dans [A2] je m'intéresse au comportement en temps long de l'algorithme du recuit simulé lorsque le potentiel n'est pas "parfaitement connu", mais peut être approché par une suite d'estimateurs uniformément convergents admettant une vitesse de convergence presque sûre suffisamment rapide. Avec Pierre Tarrès [B2] nous étudions un critère de convergence pour l'algorithme du bandit à deux bras dans le cas où les bras sont simplement supposés ergodiques. Notons que cette extension non triviale du cas i.i.d. laisse entrevoir des applications intéressantes de cet algorithme à la Finance (allocation de portefeuilles d'actions).

## 5 Programme de recherche

### 5.1 Algorithme EM semi-paramétrique

Parmi les problèmes ouverts en lien avec les modèles de mélange semi-paramétriques, il semble que l'étude de la convergence de l'algorithme EM, proposé par Bordes *et al.* (2007), soit l'un des plus urgents à traiter en raison de la compétition internationale qu'il génère autour, notamment, de l'analyse des puces ADN (voir Robin *et al.*, 2007). L'essentiel du problème réside ici en le fait que l'étape d'estimation de la densité inconnue ne permet pas d'assurer la maximisation de l'analogue de l'opérateur (E)spérance impliqué dans l'algorithme EM paramétrique standard. Cet algorithme, qui donne en pratique de très bons résultats, doit selon nous être réinterprété de manière plus générale en considérant sans doute l'étape d'estimation de la densité comme un problème de minimisation au sens d'une norme qui reste à définir.

### 5.2 Compression de structure pour les lois en grande dimension

Dans le prolongement du travail fait en collaboration avec Nadia Oudjane et Didier Chauveau sur l'échantillonnage d'importance suivant une loi instrumentale approchant non-paramétriquement la loi cible, nous souhaiterions (les méthodes non-paramétriques souffrant du fléau de la dimension) transposer cette dernière idée à des situations en grande dimension. L'idée générale de la méthode que nous proposons, consiste à visiter dans un premier temps le support de la loi d'intérêt, dont la densité  $f$  est connue à un facteur de normalisation près, au moyen de particules  $(X_1, \dots, X_N)$ , i.i.d suivant une densité instrumentale  $g$  couvrant suffisamment le support de  $f$ ; d'affecter à chaque particule un poids proportionnel à sa vraisemblance, *i.e.*  $f(X_i) / \sum_{j=1}^N f(X_j)$  pour  $i = 1, \dots, N$ ; et d'effectuer une Analyse en Composante Principale (ACP) sur ce tableau de particules pondérées. Le résultat attendu est, si la densité  $f$  est suffisamment structurée, de découvrir le sous-espace qui portent principalement la masse de  $f$  (critère d'inertie) afin d'estimer non-paramétriquement la loi marginale de l'échantillon de départ sur ce sous-espace. L'étape suivante consiste alors à rééchantillonner essentiellement suivant la loi estimée sur le sous-espace principal et plus modérément sur son supplémentaire (le ratio pouvant être donné par le rapport d'inertie de chacun des sous-espaces). Nous pensons que cette approche devrait permettre de réduire la variance des estimateurs de type échantillonnage d'importance en grande dimension (dans le cas de lois suffisamment structurées). A titre d'information, il semble que le travail de Francis Bach (voir Bach, 2008) puisse, de par la sophistication des outils qu'il utilise, nous être d'une grande utilité lors de l'étude de cette méthode.

### 5.3 Mélange semi-paramétrique de régressions

On suppose ici que pour une variable d'entrée  $X$  on puisse avoir des réponses de nature diverse

$$Y = a_j X + b_j + \varepsilon, \tag{1}$$

conditionnellement au fait que  $\{Z = j\}$ ,  $j = 1, \dots, K$ , où

$$Z \sim \text{Mult}(p) \text{ avec } p = \left\{ (p_1, \dots, p_K); \sum_{j=1}^K p_j = 1 \right\}$$

et  $\varepsilon$  est une variable aléatoire de densité  $f$  inconnue (indépendante de  $j$ ). Notre objectif est alors d'estimer le paramètre Euclidien  $\theta = \{(a_j, b_j, p_j); j = 1, \dots, K\}$  ainsi que le paramètre fonctionnel  $f$ , sur la base d'un échantillon i.i.d.  $(X_i, Y_i)$ , pour  $i = 1, \dots, n$ . Un travail

préalable, fait en collaboration avec Laurent Bordes, Didier Chauveau, et David Hunter, a permis de montrer l'identifiabilité de ce modèle sous une simple condition de symétrie sur  $f$ . Nous étudions actuellement, parmi les techniques d'estimation qui s'offrent à nous (Bordes *et al.*, 2006 ou Hunter *et. al* 2007), laquelle est la plus adaptée (modulo une certaine généralisation) à ce type de modèle. Notons que dans le cas où  $(X, Y) \in \mathbb{R}^2$ , un tel modèle correspond à un modèle de mélange semi-paramétrique en dimension 2 ce qui était considéré comme un problème ouvert dans Hall et Zhou (2003).

#### 5.4 Robustesse des approches semi-paramétriques dans le cadre des modèles à données manquantes

La grande difficulté des approches semi-paramétriques dans le cadre des modèles à données manquantes, réside en la proposition d'hypothèses minimales permettant d'assurer à la fois l'identifiabilité et l'estimation du modèle. À ce jour, les méthodes proposées en dimension 1 utilisent toujours la symétrie d'une des lois composant le modèle (le contraste semi-paramétrique testant cette symétrie sous la valeur du paramètre Euclidien). Dans le cadre du modèle de contamination semi-paramétrique rappelé ci-dessous :

$$g(x) = pf_0(x) + (1 - p)f(x - \mu), \quad x \in \mathbb{R} \tag{2}$$

où  $p \in ]0, 1[$ ,  $\mu \in \mathbb{R}$ , et  $(f_0, f)$  sont deux densités paires sur  $\mathbb{R}$  avec  $f_0$  connue et  $f$  inconnue, nous avons pu observer de bonnes performances en terme d'estimation (pour de grandes tailles d'échantillon) lorsque  $p$  était proche de 1 avec des profils de densité non-nécessairement symétriques sur  $f$ . Afin d'expliquer plus précisément le bon comportement (et l'avantage) de l'approche semi-paramétrique dans le cadre du modèle (2), nous souhaiterions mener une étude sur la robustesse de nos méthodes d'estimation (voir Bordes *et al.*, 2006 et [B1]). Notons que l'étape la plus importante pour ce travail sera sans doute l'écriture de l'analogue de la *fonction d'influence*, communément utilisée en étude de robustesse paramétrique, dans le cas semi-paramétrique.