Pierre Pudlo — Curriculum vitæ

Maître de conférences à l'Université Montpellier 2, Faculté des Sciences

I3M — Institut de Mathématiques et Modélisation de Montpellier
UMR CNRS 5149
Place Eugène Bataillon ; 34095 Montpellier CEDEX, France
Tél. : 04 67 14 42 11 / 06 85 17 78 46

Email : pierre.pudlo@univ-montp2.fr
URL : http://www.math.univ-montp2.fr/~pudlo

Né le 20 septembre 1977 à Villers-Semeuse (08 – Ardennes) Nationalité française.

Études et diplômes

2014 HABILITATION À DIRIGER DES RECHERCHES, université Montpellier 2 décernée après les rapports de Mark BEAUMONT (U. Bristol), Gérard BIAU (U. P. & M. Curie), Chris HOLMES (U. Oxford) et Adrian E. RAFTERY (U. Washington) et la soutenance devant le jury présidé par Gilles CELEUX (INRIA), composé de Mark BEAUMONT (U. Bristol), Gérard BIAU (U. P. & M. Curie), Arnaud ESTOUP (INRA), Jean-Michel MARIN (U. Montpellier 2) et Didier PIAU (U. J. Fourier)

2001–2004 THÈSE au laboratoire de Probabilités, Combinatoire et Statistique, université Claude Bernard Lyon 1 sous la direction de Didier PIAU *Estimations précises de grandes déviations et applications à la statistique des séquences biologiques*

1998–2002 Élève à l'École Normale Supérieure de Lyon : MAGISTÈRE mathématiques et applications (MMA)
LICENCE, MAÎTRISE ET DEA à l'université Lyon 1 ; AGRÉGATION de mathématiques (reçu 52ème)

Postes occupés

septembre 2011–août 2013 Délégation INRA au Centre de Biologie pour la Gestion des Populations

septembre 2006–... Maître de conférences à l'université Montpellier 2 (UM2)

2005–2006 ATER, université de Franche-Comté, Laboratoire de Mathématiques de Besançon (UMR 6623)

2002–2005 Allocataire-moniteur, université Lyon 1

1998–2002 Fonctionnaire stagiaire, École Normale Supérieure de Lyon

Thèmes de recherche

Classification : Clustering spectral • Clustering basé sur la densité • Machine learning • Théorèmes asymptotiques • constante de Cheeger • Graphes aléatoires.
Probabilités numériques et statistique bayésienne : méthodes de Monte Carlo • Génétique des populations • ABC (Approximate Bayesian Computation) • échantillonnage préférentiel • Vraisemblance empirique.
Étudiants en co-encadrement de thèse : Mohammed SEDKI (2009-2012, actuellement MC à la faculté de Médecine de l'université Paris Sud) ; Julien STOEHR (2012-..., ancien élève de l'ENS Rennes) ; Coralie MERLE (2013-...) ; Paul-Marie GROLLEMUND (2014-...)

Principaux collaborateurs

Communauté mathématique : Ery ARIAS-CASTRO (University of Californie San Diego), Benoît CADRE (Pr, École Normale Supérieure de Cachan, antenne de Rennes), Jean-Michel MARIN (Pr, université Montpellier 2), Kerrie MENGERSEN (Pr, Queensland University of Technology, Brisbane, Australie), Bruno PELLETIER (Pr, université Rennes 2), Didier PIAU (Pr, université Joseph Fourier Grenoble 1) Christian P. ROBERT (Pr, université Paris-Dauphine & IUF).
Communauté biologique : Jean-Marie CORNUET (DR, INRA CBGP), Arnaud ESTOUP (DR, INRA CBGP), Mathieu GAUTIER (CR, INRA CBGP), Raphaël LEBLOIS (CR, INRA CBGP) et François ROUSSET (DR, CNRS ISE-M).

Responsabilités administratives et scientifiques

2013–... Responsable de l'axe « *Algorithmes et calculs* » du Labex NUMEV

2010-... Membre élu du Conseil de l'UMR I3M

2008–2014 Responsable du séminaire de probabilités et statistique de Montpellier commun à l'UM2 et à l'UMR MISTEA (INRA, SupAgro)

2012, 2013,2014 Membre des comités d'organisation des Écoles-Ateliers « *Mathematical and Computational evolutionary biology* », juin 2012, 2013 et 2014.

2012 Membre du comité de sélection à l'université Lyon 1 pour recrutement d'un maître de conférences en statistique.

2010 Membre du comité d'organisation des Journées de Statistiques du Sud à Mèze (juin 2010): « *Modelling and Statistics in System Biology* ».

2009–2010 Membre de comités de sélection à Montpellier 2.

2008 Membre élu de la commission de spécialistes CNU 26 à Montpellier 2.

2004–2005 Administrateur du serveur informatique du LaPCS (Laboratoire de Probabilités, Combinatoire et Statistique, Lyon 1)

Liste de publications

Articles publiés

(A1) P. Pudlo (2009) Large deviations and full Edgeworth expansions for finite Markov chains with applications to the analysis of genomic sequences. *ESAIM: Probab. and Statis.* 14, pp. 435–455.

(A2) B. Pelletier and P. Pudlo (2011) Operator norm convergence of spectral clustering on level sets. *Journal of Machine Learning Research*, 12, pp. 349–380.

(A3) E. Arias-Castro, B. Pelletier and P. Pudlo (2012) The Normalized Graph Cut and Cheeger Constant: from Discrete to Continuous. *Advances in Applied Probability*, 44(4), dec 2012.

(A4) B. Cadre, B. Pelletier and P. Pudlo (2013) Estimation of density level sets with a given probability content. *Journal of Nonparametric Statistics* 25(1), pp. 261–272.

(A5) J.–M. Marin, P. Pudlo , C. P. Robert and R. Ryder (2012) Approximate Bayesian Computational methods. *Statistics and Computing* 22(6), pp. 1167–1180.

(A6) A. Estoup, E. Lombaert, J.–M. Marin, T. Guillemaud, P. Pudlo, C. P. Robert and J.–M. Cornuet (2012) Estimation of demo-genetic model probabilities with Approximate Bayesian Computation using linear discriminant analysis on summary statistics. *Molecular Ecology Resources* 12(5), pp. 846–855.

(A7) Mengerson, K.L., Pudlo, P. and Robert, C. P. (2013) Bayesian computation via empirical likelihood. *Proc. Natl. Acad. Sci. USA* 110(4), pp. 1321–1326.

(A8) Gautier, M., Foucaud, J., Gharbi, K., Cezard, T., Galan, M., Loiseau, A., Thomson, M., Pudlo, P., Kerdelhué, C., Estoup, A. (2013) Estimation of population allele frequencies from next-generation sequencing data: pooled versus individual genotyping. *Molecular Ecology* 22(14), pp. 3766–3779.

(A9) Gautier, M., Gharbi, K., Cezard, T., Foucaud, J., Kerdelhué, C., Pudlo, P., Cornuet, J.-M., Estoup, A. (2013) The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Molecular Ecology* 22(11), pp. 3165–3178.

(A12) Cornuet J.-M., Pudlo P., Veyssier J., Dehne-Garcia A., Gautier M., Leblois R., Marin J.-M., Estoup A. (2014) DIYABC v2.0: a software to make Approximate Bayesian Computation inferences about population history using Single Nucleotide Polymorphism, DNA sequence and microsatellite data. *Bioinformatics* 30(8), pp. 1187–1189. Voir http://www1.montpellier.inra.fr/CBGP/diyabc/

(A13) Baragatti, M. and Pudlo, P. (2014) An overview on Approximate Bayesian computation. *ESAIM: Proceedings* 44, pp. 291–299.

(A14) Leblois, R., Pudlo, P., Néron, J., Bertaux, F., Beeravolu, C. R., Vitalis, R. and Rousset, F. (2014) Maximum likelihood inference of population size contractions from microsatellite data. *Molecular Biology and Evolution*, 31(10), pp. 2805–2823.

(A15) Stoehr, J., Pudlo, P. and Cucala, L. (2014) Adaptive ABC model choice and geometric summary statistics for hidden Gibbs random fields. Accepté dans *Statistics and Computing*. Voir Arxiv:1402.1380

Preprints et articles soumis

(A10) Ratmann, O., Pudlo, P., Richardson, S. and Robert, C. P. (2011) *Monte Carlo algorithms for model assessment via conflicting summaries*. Voir Arxiv:1106.5919

(A11) Sedki, M., Pudlo, P., J.–M. Marin, C. P. Robert and J.–M. Cornuet (2013) *Efficient learning in ABC algorithms*. Soumis. Voir Arxiv:1210.1388

(A16) Marin, J.-M., Pudlo, P. and Sedki, M. (2012 ; 2014) *Consistency of the Adaptive Multiple Importance Sampling*. Soumis. Voir Arxiv:1211.2548.

(A17) Pudlo, P., Marin, J.-M., Estoup, A., Gautier, M., Cornuet, J.-M. and Robert, C. P. *ABC model choice via random forests*. Soumis. Voir Arxiv:1406.6288

Voici quelques éléments concernant la qualité des revues dans lesquelles j'ai publié.

TABLE 1 – Référentiel de notoriété

Revue	Facteur d'impact (FI à 5 ans)	Notoriété
Advances in Applied Probability	0.900 (0.841)	**
ESAIM: Probab. and Statis.	0.408 (–)	*
Journal of Machine Learning Research	3.420 (4.284)	***
Journal of Nonparametric Statistics	0.533 (0.652)	*
Molecular Biology and Evolution	10.353 (11.221)	****
Molecular Ecology	6.275 (6.792)	***
Molecular Ecology Resources	7.432 (4.150)	***
Proc. Natl. Acad. Sci. USA	9.737 (10.583)	****
Statistics and Computing	1.977 (2.663)	***

**** = Exceptionnelle, *** = Excellente, ** = Correcte, * = Médiocre

La notoriété des revues est tirée du *Référentiel de notoriété 2012*, Erist de Jouy-en-Josas – Crebi, M. Désiré, M.-H. Magri et A. Solari. Elle provient d'une étude de la distribution des facteurs d'impact par discipline.

Workshop, conférences et communications orales

(T1) 46èmes Journées de Statistique, Rennes 2014.
(T2) Advances in Scalable Bayesian Computation (invitation) à Banff (Alberta, Canada)
(T3) MCMSki IV à Chamonix, janvier 2014 (Poster avec Julien STOEHR)
(T4) ABC in Rome, Mai 2013 (Poster avec Julien STOEHR)
(T5) Séminaire de génétique des populations Vienne, Autriche, Avril 2013 (invitation)
(T6) Journées du groupe Modélisation Aléatoire et Statistique de la SMAI, Clermont-Ferrand, Août 2012 (invitation)
(T7) Mathematical and Computational Evolutionary Biology, Montpellier, Juin 2012.
(T8) International Workshop on Applied Probability, Jérusalem, Juin 2012. (invitation)
(T9) Séminaires Statistique Mathématique et Applications, Fréjus, Août-Septembre 2011.
(T10) 5èmes Journées Statistiques du Sud, Nice, Juin 2011 (invitation)
(T11) Approximate Bayesian Computation in London, Mai 2011.
(T12) 3rd conference of the International Biometric Society Channel Network, avril 2011.
(T13) 42èmes Journées de Statistique, Marseille 2010.
(T14) 41èmes Journées de Statistique, Bordeaux 2009.
(T15) XXXIVème École d'Été de Probabilités de Saint-Flour, 2004.
(T16) Journées de probabilités, La Rochelle, 2002.

DIVERS SÉMINAIRES EN FRANCE : entre autre, Toulouse (2012), AgroParisTech (2012), Grenoble au LECA (2011), Avignon (2011), Besançon (2010, 2013), Rennes (2010), Grenoble (2008)…

Brevet international

(B1) Demande PCT no. EP13153512.2 : *Process for identifying rare events* (2014).

Pierre Pudlo — Activités d'enseignement, de recherche, ...

Enseignements

Depuis mon recrutement comme maître de conférences en 2006, j'ai eu l'occasion d'effectuer de nombreuses heures d'enseignement à l'UFR faculté des Sciences de l'université Montpellier 2 (UM2), parmis lesquelles on peut citer :

PhD Responsable du module doctoral *Programmation orientée objet : modélisation probabiliste & calcul numérique en statistique pour la biologie* (30h/an)

M2R Responsable du module de *Classification Supervisée et Non-Supervisée* (20h/an)

M1 Responsable du module *Processus stochastique* / *Réseaux et files d'attente* (50h/an)

L3 Responsable du module *Traitement de données* (50h/an) pour les licences de *Biologie* et de *Géologie-Biologie-Environnement*

J'ai ainsi pu diversifier mes enseignements. Les modules dont j'ai eu la responsabilité vont de cours théoriques et techniques, par exemple une UE de probabilités pour des M1 de mathématiques (mode de convergence, espérance et loi conditionnelle, chaînes de Markov et martingales discrètes), à des cours pratiques, soit au niveau M2 du master de mathématiques, soit pour un public d'étudiants en biologie.

En outre, j'ai encadré une dizaine de projets d'étudiants de M1 sur diverses questions de probabilités ou de statistique.

Encadrement

Outre une dizaine d'étudiants en Master 1, j'ai encadré les travaux ci-dessous.

sept 2014 – ... Thèse de Paul GROLLEMUND en co-direction avec Christophe ABRAHAM et Meïli BARAGATTI (UMR MISTEA, INRA SupAgro) : *Régression linéaire bayésienne fonctionnelle interprétable.*

sept 2013 – ... Thèse de Coralie MERLE en co-direction effective avec Raphaël LEBLOIS (CR INRA, CBGP) : *Nouvelles méthodes d'inférence de l'histoire démographique à partir de données génétiques.*

avril – sept 2013 Stage de M2 de Coralie MERLE (Master MathSV, Université Paris Sud – École Polytechnique) avec Raphaël LEBLOIS (CR INRA, CBGP).

2012 – ... Thèse de Julien STOEHR co-encadrée avec Jean-Michel MARIN (PR, Montpellier 2) et Lionel CUCALA (MCF, Montpellier 2) : *Choix de modèles pour les champs de Gibbs* (en particulier via ABC)

mars – juin 2012 Stage de M2 de Julien STOEHR, élève de l'École Normale Supérieure de Cachan

2009 – 2012 Thèse de Mohammed SEDKI avec J.-M. MARIN (PR, Montpellier 2) : *Échantillonnage préférentiel adaptatif et méthodes bayésiennes approchées appliquées à la génétique des populations.* (soutenue le 31 octobre 2012). M. SEDKI est MAÎTRE DE CONFÉRENCES à l'université d'Orsay (faculté de médecine – Le Kremlin-Bicetre) depuis septembre 2013.

Travaux de recherche

Mes travaux de recherche ont touché à des thématiques diversifiées, allant de résultats théoriques en probabilités au développement d'algorithmes d'inférence et à leurs mises en œuvre. Deux caractéristiques dominantes se dégagent : (1) l'omniprésence d'algorithmes (depuis l'algorithme glouton de ma thèse aux algorithmes de Monte-Carlo dans mes derniers travaux), et (2) leur adaptation en biologie, principalement en génétique des populations. Du fait de la taille croissante des données produites notamment en génomique, les méthodes statistiques doivent gagner en efficacité sans perdre le détail de l'information incluse dans ses grandes bases de données. Mes travaux détaillés ci-dessous ont fourni aussi bien des contributions importantes à l'analyse et la compréhension des performances de ces algorithmes, qu'à la conception de nouveaux algorithmes gagnant en précision ou efficacité d'estimation.

Classification non supervisée

Publications. (A2), (A3) et (A4), voir CV. Un brevet international en cours de dépôt.

Financements. Projet ANR CLARA (2009–2013): *Clustering in High Dimension: Algorithms and Applications*, porté par B. PELLETIER; projet de maturation avec la société d'accélération du transfert de technologies de Montpellier (SATT AxLR).

Mots clés. *Machine learning, classification spectrale, théorèmes asymptotiques, constante de Cheeger, graph-cut, graphes de voisinages, théorie spectrale d'opérateurs.*

Les techniques de classifications spectrales qui m'ont intéressées occupent une place importante dans le champ de recherche de méthodes algorithmiquement efficaces sur jeu de données de grande taille. Elles permettent de détecter des groupes d'observations de forme quelconque, contrairement aux k-means ou méthodes de mélange, qui ne détectent que des groupes convexes.

En modifiant l'algorithme de classification spectrale, nous avons montré dans (A2) la consistance de l'algorithme et que le partitionnement limite coïncide avec la définition géométrique d'un cluster proposé par Hartigan. Cette démonstration repose sur un mode de convergence fort d'opérateurs associés aux matrices de similarité de l'échantillon. Et nous avons étudié une re-paramétrisation plus intuitive de ces clusters définis par Hartigan dans (A4)

L'algorithme de classification spectrale peut se voir comme une approximation du problème NP-difficile de graph-cut ou de détection de goulets d'étranglement dans des graphes de similarité ou des réseaux d'interaction. Dans (A3), nous avons obtenu sur le problème d'optimisation NP-difficile des résultats asymptotiques donnant la limite continue lorsque la taille de l'échantillon grandit en adaptant la fonction de similarité binaire à cette taille.

Récemment, nous avons été contacté par le CHU avec André Mas (PR UM2, I3M) pour des traiter des jeux de données d'analyse sanguine par cytomètre en flux. La question initiale était de détecter un petit cluster de cellules rares (des cellules circulantes à cause d'un cancer par exemple) parmi un très grand nombre d'observations. L'algorithme que nous avons développé est en cours de protection par le dépôt d'un brevet international. Avec la société d'accélération du transfert de technologies de Montpellier (SATT AxLR), nous venons d'obtenir un financement de maturation autour de cet algorithme et d'autres questions de transferts pour les données de cytométrie en flux.

Statistique computationnelle

Publications. (A5), (A6), (A7), (A8), (A9), (A12), (A13), (A14) et (A15), ainsi que les pré-publications (A10), (A11), (A16) et (A17) voir CV.

Documents et logiciels à vocation de transfert. Cornuet J-M, Pudlo P, Veyssier J, Dehne-Garcia A, Estoup A (2013) DIYABC V2.0. a user-friendly package for inferring population history through Approximate Bayesian Computation using microsatellites, DNA sequence and SNP data. Logiciel et notice détaillée d'utilisation de 91 pages disponibles sur le site http://www1.montpellier.inra.fr/CBGP/diyabc/

Financements.

- une délégation INRA (département SPE) de deux ans au Centre de Biologie pour la Gestion des Populations (CBGP, UMR INRA SupAgro Cirad IRD, Montpellier)
- le projet ANR EMILE (2009–2013): *Études de Méthodes Inférentielles et Logiciels pour l'Évolution*, porté par J.M. CORNUET, puis R. VITALIS (DR INRA, CBGP) ;
- le Labex NUMEV, Montpellier ;
- l'Institut de Biologie Computationnelle (projet Investissement d'Avenir, Montpellier) : l'axe dont je suis membre regroupe des statisticiens, des bioinformaticiens et des généticiens des populations ;
- le projet PEPS (CNRS) « Comprendre les maladies émergentes et les épidémies : modélisation, évolution, histoire et société », que je porte avec Raphaël LEBLOIS (CR INRA, CBGP).

Mots clés. *Méthodes de Monte Carlo, statistique computationnelle, méthodes ABC, échantillonnage préférentiel, vraisemblance empirique, génétique des populations.*

Les problèmes que j'ai regardé en statistique appliquée se situent dans des cas où il est difficile voire impossible de calculer numériquement la vraisemblance des données. Donnons deux cas particuliers : *(i)* les données s'expliquent à l'aide d'un processus latent vivant dans un espace de grande dimension (génétique des populations sous neutralité) ou *(ii)* la loi du jeu de données est connue à une constante de normalisation dont la valeur dépend du paramètre (champs de Markov).

Les méthodes bayésiennes approchées (ABC ou *approximate Bayesian computation*) contournent le calcul de la vraisemblance en comparant des jeux de données simulées aux données observées au travers de quantités numériques (statistiques résumées) supposées informatives, voir les deux articles de review (A5) et (A13) qui se complètent. Elles permettent donc de mener une analyse bayésienne dans le contexte où la vraisemblance est incalculable, mais où l'on peut simuler des jeux de données de façon efficace. Avec mon premier étudiant en thèse, Mohammed SEDKI, nous avons développé un algorithme séquentiel d'inférence ABC permettant d'économiser en nombre de simulations nécessaires (A11). Comparé à l'état de l'art, cet algorithme est auto-calibré (auto-tuning) et plus efficace, d'où un gain de temps pour obtenir une réponse de même qualité que l'algorithme standard.

Au Centre de Biologie pour la Gestion des Populations (UMR INRA SupAgro Cirad IRD, Montpellier), je me suis fortement impliqué dans le codage de la seconde version de DIYABC, qui vient de sortir (A12). Ce logiciel condense toute l'expérience acquise sur les méthodes ABC pour la génétique des populations. En particulier, pour gérer les situations où le nombre de statistiques résumées est important, nous avons proposé de faire du choix de modèle en utilisant comme statistiques résumées des axes discriminants issus d'un pré-traitement (A6).

Je co-encadre actuellement les travaux de thèse de Julien STOEHR, qui porte sur la sélection de modèle pour des champs de Markov latents. Dans (A15), nous avons mis en place une procédure ABC de choix de modèle, qui renonce à l'approximation de la probabilité *a posteriori* de chacun des modèles (qui représentent différentes structures de dépendance) pour améliorer le taux de mauvaise classification (c'est-à-dire de mauvais choix de modèle), via une procédure des k plus proches voisins parmis les simulations ABC. Nous avons prolongé ces travaux en remplaçant la méthode des k plus proches voisins par des forêts aléatoires de Breiman, entrainées sur les simulations ABC. Ce type de classifieur, qui prédit un modèle en fonction des statistiques résumées, est bien moins sensible à la dimension que la méthode des k plus proches voisins et donne de bien meilleurs résultats en génétique des populations où le nombre de statistiques résumées est de l'ordre de la centaine (à comparer avec la dizaine de statistiques résumées dans les questions de champs markoviens cachés des travaux de Julien STOEHR), voir (A17).

En revanche, la taille du jeu de données observés s'accroit, on ne peut plus simuler efficacement des jeux de données de taille identifique, ni donc recourir à des méthodes ABC. Il faut alors utiliser des pseudo-vraisemblances pour inférer le paramètre d'intérêt. Par exemple, le maximum de la vraisemblance composite par paire fournit un estimateur raisonnable. Mais cette pseudo-vraisemblance, plus étroite que la véritable vraisemblance, ne peut être utilisée directement dans une méthode bayésienne comme ersatz de vraisemblance. Dans (A7), nous proposons une utilisation originale de la vraisemblance empirique d'Owen (1998, 2910) pour construire un algorithme de calcul bayésien (BC_{el}, Bayesian Computation via empirical likelihood). Initialement conçue pour s'affranchir d'hypothèses paramétriques, la vraisemblance empirique peut être vue comme une pseudo-vraisemblance, reconstruite à partir des données, qui possède de bonnes propriétés. En particulier, elle permet de construire un test de rapport de vraisemblance, donc des intervalles de confiances de largeurs correctes, ce qui n'est pas le cas d'autres pseudo-vraisemblance comme la vraisemblance composite par paires. En outre, BC_{el} permet de réduire grandement les temps de calcul (plusieurs heures en ABC deviennent ici autant de minutes), donc de traiter des jeux de données de dimension plus grande. Cette piste est donc prometteuse pour l'avenir.

Les données de polymorphisme génétique collectées par séquençage ultra-haut débit (*Next Generation Sequencing data* ou NGS) fournissent des jeux de données de telle dimension. Mais, comme toute nouvelle technique, celle-ci introduit différents problèmes. Dans (A8), nous étudions l'information perdue par des schémas de génotypage par lots (ou *pool*) d'individus plutôt qu'un séquençage individuel. Et dans (A9), nous étudions le biais introduit par le séquençage RAD (Restriction site associated DNA), qui s'avère relativement faible dans la plupart des cas concrets. Nous proposons en oure une méthode pour filtrer les sites où le biais est important.