# Arbres Généalogiques et Estimation Non Linéaire

Pierre Del Moral

*Lab. Stat. Proba., CNRS UMR 5583, Toulouse*

**Résumé**

Dans cette communication, nous introduisons une méthodologie particulaire moderne pour l'analyse et l'estimation de divers problèmes non linéaires issus de disciplines scientifiques connexes : statistique des processus, physique des particules, biologie macromoléculaire, traitement du signal et de l'information. Ces algorithmes sont fondés sur l'évolution d'arbres généalogiques. Ces derniers sont définis en terme de lignes ancestrales associées à des processus en interaction de type génétique. Dans une seconde partie, nous présentons très brièvement les fondements d'une nouvelle théorie asymptotique sur ces processus en interaction.

**Résumé en anglais**

In this report, we present a modern particle methodology for the analysis and the numerical solving of a class of non linear problems arising in a variety of scientific disciplines, including statistics, particle physics, macromolecular biology and signal processing. These genealogical tree models represent the ancestral lines evolution associated to genetic type algorithms. In the second part, we will briefly introduce the fondations of a new asymptotic theory on these interacting processes.

**Mots clés :** algorithmes génétiques, processus d'évolution, lignes ancestrales, arbres généalogiques, estimation et filtrage non linéaire, formules de Feynman-Kac.

# 1 Introduction aux processus d'évolution

Les arbres généalogiques représentent l'évolution des lignes ancestrales d'une population d'individus. Ces évolutions sont essentiellement fondées sur deux transitions génétiques classiques : *la mutation et la sélection.* Pour les modèles à temps discret, la mutation des individus correspond à la donnée d'une chaîne de Markov $X_n$ sur des espaces mesurables $(E_n, \mathcal{E}_n)$. Le choix de ces espaces d'état dépend bien entendu du modèle biologique ou physique étudié. A un instant $n \in \mathbb{N}$, l'état $X_n \in E_n$ peut autant représenter un code génétique sur un espace fini, que la position d'un individu dans l'espace, ou encore sa trajectoire d'évolution complète depuis son origine. La loi initiale $\eta_0(dx_0)$, et les transitions élémentaires $M_n(x_{n-1}, dx_n)$, de cette chaîne décrivent respectivement, la distribution à l'origine de chaque individu, et leurs mutations entre deux étapes de sélection.

L'étape de sélection est associée à la donnée d'une classe de fonctions mesurables et bornées $G_n : x_n \in E_n \to G_n(x_n) \in [0, \infty)$. Dans chaque domaine d'application, ces fonctions ont à nouveau des interprétations différentes. En biologie, elles expriment la dominance ou la récession de séquences génétiques; ou dans un autre contexte, l'énergie associée à des polymérisations de macro-molécules dans des solvants chimiques. En physique des particules, ces fonctions représentent plutôt des puits de potentiels d'énergie. Dans ce contexte, elles expriment la durée de vie de particules évoluant dans des milieux absorbants. En traitement du signal, et plus particulièrement en filtrage non linéaire, ces fonctions de sélection servent à mesurer la vraisemblance d'une position spatio-temporelle, par rapport aux observations délivrées par les capteurs. En résumé, ces fonctions de selection $G_n$ décrivent les qualités d'attraction ou de répulsion de chaque région de l'espace $E_n$. La sélection favorisera ainsi la reproduction des individus ayant atteint des régions attractives à fort potentiel.

Ces modèles d'évolution sont le plus souvent décrits mathématiquement par la donnée d'une chaîne de Markov dans des espaces produits :

$$\xi_n = (\xi_n^i)_{1 \le i \le N} \in E_n^N \xrightarrow{\text{selection}} \widehat{\xi}_n = (\widehat{\xi}_n^i)_{1 \le i \le N} \in E_n^N \xrightarrow{\text{mutation}} \xi_{n+1}$$

Le paramètre $N \ge 1$ représente la taille des populations, que nous avons ici considérée constante dans le temps pour simplifier la présentation. La description suivante est définie jusqu'au premier instant $\tau^N$ où tous les individus ont un potentiel nul $G_{\tau^N}(\xi_{\tau^N}^i) = 0$. Dans ce cas, l'évolution s'arrête, et les populations sont placées dans un état cimetière, $\widehat{\xi}_n = \xi_{n+1} = c$ pour tout $n \ge \tau^N$.

Initialement, les individus $\xi_0^i$ sont simulés indépendamment, et selon la même loi $\eta_0$. Durant l'étape de sélection, chaque individu $\widehat{\xi}_n^i$ est choisi selon la distribution suivante

$$\epsilon_n^N \ G_n(\xi_n^i) \ \delta_{\xi_n^i} + (1 - \epsilon_n^N \ G_n(\xi_n^i)) \ \sum_{j=1}^N \frac{G_n(\xi_n^j)}{\sum_{k=1}^N G_n(\xi_n^k)} \ \delta_{\xi_n^j}$$

Dans la formule ci-dessus, $\delta_{x_n}$ représente la mesure de Dirac usuelle au point $x_n \in E_n$, et $\epsilon_n^N \ge 0$ est un paramètre auxiliaire, arbitrairement choisi de sorte à avoir $\epsilon_n^N \ G_n(\xi_n^j) \in [0, 1]$, pour chaque $1 \le j \le N$. On peut ainsi choisir $\epsilon_n^N \in \{0, 1/\max_j G_n(\xi_n^j), 1/\|G_n\|\}$, ou tout simplement $\epsilon_n^N = 1$ lorsque $G_n$ est à valeurs dans $[0, 1]$. Ces transitions sélectives $\xi_n^i \rightsquigarrow \widehat{\xi}_n^i$ s'expriment donc de la façon suivante. Avec la probabilité $\epsilon_n^N \ G_n(\xi_n^i)$, l'individu $\xi_n^i$ reste à sa place. Dans le cas contraire, il est remplacé par un nouvel individu $\xi_n^j$, aléatoirement choisi avec une probabilité proportionelle à son potentiel $G_n(\xi_n^j)$. Cette dernière situation peut aussi s'interpréter comme l'extinction du $i^{\text{ème}}$ individu, et la naissance instantanée d'un individu sur le $j^{\text{ème}}$ site. Les mutations $\widehat{\xi}_n^i \rightsquigarrow \xi_{n+1}^i$ consistent simplement à faire évoluer chacun des individus sélectionnés selon les transitions $M_{n+1}$. Plus formellement, chaque $\xi_{n+1}^i$ est une variable aléatoire de loi $M_{n+1}(\widehat{\xi}_n^i, .)$.

En terme de processus de naissance et mort, chacun des individus $\xi^i_{n,n} = \xi^i_n$ s'interprète comme l'enfant de $\widehat{\xi}^i_{n-1}$. Si ce dernier $\widehat{\xi}^i_{n-1} = \xi^j_{n-1}$ est le fruit d'une naissance au site $j$, le premier aura donc pour ancêtre $\xi^i_{n-1,n} = \xi^j_{n-1}$, etc. Les lignes ancestrales des individus $\xi^i_n$ s'expriment ainsi *récursivement* en terme d'évolutions de trajectoires depuis l'origine

$$\zeta^i_n = (\xi^i_{0,n}, \xi^i_{1,n}, \ldots, \xi^i_{n,n}) \in E_{[0,n]} = E_0 \times E_1 \times \ldots \times E_n$$

Les lignes ancestrales après sélection $\widehat{\zeta}^i_n = (\widehat{\xi}^i_{p,n})_{0 \leq p \leq n}$ se définissent par un raisonnement analogue.

# 2 Estimation non linéaire

Les processus d'évolution que nous venons d'introduire sont clairement liés aux algorithmes génétiques. L'introduction de ces modèles est le plus souvent attribuée aux travaux de J.H. Holland (1975). Néanmoins, une recherche bibliographique plus approfondie souligne l'utilisation de tels algorithmes en physique, en biologie et dans les sciences de l'ingénieur avec les travaux de T.E. Harris et H. Kahn (1951) sur le transport de particules, puis ceux de N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, E. Teller et A.H. Teller (1953) sur la simulation de marches aléatoires sans intersection, ou encore ceux de D.D. Magill (1965) sur l'estimation de signaux dans des problèmes de filtrage à multi-modèles, et plus ultérieurement ceux de J.H. Hetherington (1984) sur le calcul de valeurs propres de matrices positives.

Les domaines d'applications de ces algorithmes sont de nos jours en constant développement. Au cours de ces dernières années, cette méthodologie particulaire a notamment permis des avancées très significatives dans des domaines aussi variés que le traitement du signal et de l'image, le contrôle et régulation de processus, l'estimation d'évènements rares, l'analyse macromoléculaire, ou encore la physique des particules. Afin d'aiguiser la curiosité du lecteur, on notera que chacune de ces disciplines a adopté un vocabulaire différent pour décrire le *même* algorithme : méthodes de Monte Carlo séquentielles, filtres particulaires, filtres par bootstrap, filtre de Kalman d'ensembles. D'autres terminologies, plus ésotériques et issues de la littérature anglo-saxonne, sont pour ma part plus difficiles à traduire : "spawning filters", "auxiliary particle and condensation filters" "sampling-importance-resampling", "matrix reconfiguration", "switching algorithms", "go-with-the-winner", "multisplitting models", "quantum and diffusion Monte Carlo methods".

Ces divers algorithmes sont généralement décrits dans la littérature sous forme de schémas heuristiques simples, naturels et performants, mais sans aucune analyse rigoureuse, ni référence à leurs fondements physiques, biologiques ou mathématiques. Pour une théorie asymptotique complète, une présentation assez exhaustive des divers domaines d'applications, ainsi qu'une bibliographie plus détaillée, je renvoie le lecteur à mon récent ouvrage de recherche, Del Moral (2004).

Afin d'apprécier l'impact en pratique et les fondations de ces méthodes, il est essentiel de noter que la variété des champs d'application que nous venons d'exposer, provient tout simplement des différentes interprétations que l'on peut donner à ces modèles d'évolution.

D'un point de vue statistique, les mécanismes de mutation/sélection peuvent s'interpréter comme une technique de simulation de lois trajectorielles de type acceptation/rejet. Sous cet angle, ces modèles d'interaction induisent une nouvelle méthodologie de simulation particulaire de lois complexes dans des espaces de grandes dimensions. Sur ce thème, de récentes études de P. Del Moral et A. Doucet (2003). ont mis en évidence une nouvelle classe d'algorithmes de Metropolis en interaction. A la différence des techniques traditionnelles de Monte Carlo par chaînes de Markov, la vitesse de convergence de ces modèles en interaction ne dépend plus de la nature des lois cibles, mais plutôt des propriétés de mélange du processus de proposition/mutation.

D'un point de vue physique, ces modèles traduisent une interprétation particulaire de type champ moyen de certaines quantités, telles les valeurs spectrales fondamentales d'opérateurs de Shrödinger, ou encore les solutions de problèmes de Dirichlet avec conditions au bord. D'un point de vue plus probabiliste, ces modèles peuvent aussi s'interpréter comme une technique de linéarisation stochastique locale d'équations de conservation à valeurs mesures et non linéaires.

Enfin, du point de vue de l'ingénieur, ces algorithmes particulaires permettent une exploration aléatoire et adaptative d'un espace d'état. Ces stratégies d'exploration sélective permettent ainsi de résoudre divers problèmes d'estimation trajectorielle, ou d'optimisation globale non linéaires. Par exemple, en traitement du signal, les transitions de mutation/sélection coïncident avec les deux étapes classiques de prédiction/correction du filtre optimal. Dans ce contexte, les potentiels de sélection correspondent aux fonctions de vraisemblance, et les mutations représentent les transitions élémentaires du signal observé. Dans l'analyse de problèmes de filtrage de multi-modèles et multi-capteurs (partiellement linéaires et gaussiens), un autre choix judicieux du couple sélection/mutation conduit au développement d'une nouvelle classe de filtres de Kalman-Bucy en interaction.

# 3 Analyse d'arbres généalogiques

Les difficultés d'analyse des modèles d'arbres généalogiques décrits ci-dessus, sont liées à la non linéarité des processus d'interaction, ainsi qu'à la complexité des espaces trajectoriels sous jacents. Ces deux aspects ont empêché, pendant une cinquantaine d'années, toute tentative d'élaboration d'un cadre d'étude mathématique rigoureux justifiant le bien fondé de ces modèles, et permettant d'étudier avec précision leur comportement asymptotique.

L'étude de ces modèles en interaction, s'est déroulée, au fil de ces dix dernières années, suivant des recherches de précisions numériques de plus en plus fines : inégalités précises de type Burkholder-Davis-Gundy, propagation forte, croissante et uniforme du chaos par rapport à l'horizon temporel ou selon la taille de blocs de particules, convergence de processus empiriques, inégalités de concentration, théorèmes limites sur les fluctuations trajectorielles, théorèmes de type Donsker et inégalités de type Berry-Esseen, principes de grandes déviations, trajectoriels ou sur l'espace des distributions muni de topologies fortes, estimations uniformes par rapport au paramètre temporel des fonctionnelles de taux. La majorité de ces résultats est présentée dans mon livre, Del Moral (2004). Afin de présenter très brièvement les fondements de cette nouvelle théorie asymptotique, nous noterons

$$\mathbb{Q}_n^N = \frac{1}{N}\sum_{i=1}^N \delta_{(\xi^i_{0,n},\xi^i_{1,n},\ldots,\xi^i_{n,n})} \quad \text{et} \quad \widehat{\mathbb{Q}}_n^N = \frac{1}{N}\sum_{i=1}^N \delta_{(\widehat{\xi}^i_{0,n},\widehat{\xi}^i_{1,n},\ldots,\widehat{\xi}^i_{n,n})},$$

les mesures d'occupation des lignes ancestrales associées aux modèles d'arbres généalogiques décrits précédemment (lorsque $\tau^N \leq n$, on prendra la convention $\widehat{\mathbb{Q}}_n^N = \mathbb{Q}_{n+1}^N = 0$). Leurs marginales respectives par rapport aux instants terminaux, notées $\eta_n^N = \frac{1}{N}\sum_{i=1}^N \delta_{\xi^i_n}$ et $\widehat{\eta}_n^N = \frac{1}{N}\sum_{i=1}^N \delta_{\widehat{\xi}^i_n}$. Ces dernières correspondent tout simplement aux mesures d'occupation des populations après mutation et sélection. Avec des degrés de précision variés, on peut décrire les convergences faibles

$$\lim_{N\to\infty} (\mathbb{Q}_n^N, \widehat{\mathbb{Q}}_n^N) = (\mathbb{Q}_n, \widehat{\mathbb{Q}}_n),$$

ou estimer finement l'indépendance par $q$-blocs et l'adéquation des modèles

$$\lim_{N\to\infty} \|\mathrm{Loi}((\xi^i_{0,n},\ldots,\xi^i_{n,n})_{i=1,\ldots,q}) - \mathbb{Q}_n^{\otimes q}\|_{\mathrm{tv}} = 0,$$

en termes de mesures déterministes définies par les formules suivantes

$$\mathbb{Q}_n(d(x_0,\ldots,x_n)) = \frac{1}{\mathcal{Z}_n}\,[\prod_{0\leq p<n} G_p(x_p)]\ \mathbb{P}_n(d(x_0,\ldots,x_n)) \tag{1}$$

Les quantités $\mathcal{Z}_n$ et $\mathbb{P}_n$ désignent respectivement une constante de normalisation et la distribution de la trajectoire de la chaîne $X_p$ depuis son origine $p = 0$ jusqu'à l'instant $p = n$,

$$\mathbb{P}_n(d(x_0,\ldots,x_n)) = \eta_0(dx_0)M_1(x_0,dx_1)\ldots M_n(x_{n-1},dx_n).$$

Les mesures $\widehat{\mathbb{Q}}_n$ sont définies de façon analogue, en prenant dans (1) les produits des potentiels jusqu'à l'instant $n$ inclus. En physique mathématique, les constantes de normalisation $\mathcal{Z}_n$ sont parfois appelées les fonctions de partitions ou encore d'énergie libre. Ces dernières correspondent aussi aux fonctions de log-vraissemblance utiles à l'étude de chaînes de Markov cachées. Enfin, elles permettent aussi de décrire des probabilités de temps de sortie, ou d'évènements rares, rencontrés dans l'analyse de problèmes de fiabilité ou de milieux absorbants.

Ces quantités sont difficiles à calculer analytiquement. Cependant, elles peuvent être estimées finement et sans biais selon la formule suivante

$$\lim_{N\to\infty} \prod_{0\leq p<n} \frac{1}{N}\sum_{i=1}^N G_p(\xi^i_p) = \int [\prod_{0\leq p<n} G_p(x_p)]\ \mathbb{P}_n(d(x_0,\ldots,x_n)).$$

# 4 Mesures de Feynman-Kac

Les versions à temps continu de ces distributions trajectorielles (1) ont été introduites en 1942 par le physicien R. Feynman, pour analyser l'évolution de particules quantiques évoluant dans un puit de potentiel. Ces formules "de changement de probabilités" sont désormais d'un usage courant dans diverses disciplines scientifiques. Plus précisément, elles offrent de nos jours un modèle mathématique moderne, et une interprétation probabiliste puissante, pour l'analyse de nombreux problèmes non linéaires de la physique, de la biologie, et des sciences de l'ingénieur. En physique théorique, elles représentent les distributions de particules évoluant dans des milieux absorbants liés à des potentiels, ou encore les lois de temps de sortie ou d'atteinte de niveaux rares. Elles expriment aussi les solutions de certains problèmes de Dirichlet avec conditions aux bords. Dans un autre registre, ces mesures permettent des interprétations probabilistes de quantités spectrales d'opérateurs de Schrödinger. Elles correspondent aussi à des exemples non triviaux d'équations de Boltzmann généralisées, ou d'évolutions génétiques de type Fleming-Viot. En filtrage, elles représentent les lois conditionnelles de signaux sachant des observations partielles et bruitées. Enfin, en statistiques et probabilité, elles modélisent les distributions de trajectoires markoviennes restreintes, telles les lois de chaînes sans intersection, ou encore à conditions terminales fixées, les premiers moments de processus de branchement,etc. Pour conclure cette communication, on notera que ces techniques particulaires sont au coeur des méthodes tant de la statistique fréquentiste, que de la statistique bayesienne. Elles ouvrent par ailleurs de nouvelles avenues de recherches pluridisciplinaires tant sur le plan théorique qu'au niveau des applications.

## Bibliographie

[1] P. Del Moral. *Feynman-Kac Formulae, Genealogical and Interacting Particle Systems with Applications.* Springer Verlag, New York. Series: Probability and Applications, (555 pages), 2004. http://www.lsp.ups-tlse.fr/Delmoral/gips.html
[2] P. Del Moral et A. Doucet. On a class of genealogical and interacting metropolis models. *Séminaire de Probabilités XXXVII*, J. Azéma, M. Emery, M. Ledoux, and M. Yor, eds., Lecture Notes in Mathematics no. 1832, pp.415–446, Springer-Verlag, Berlin, 2003.
[3] T.E. Harris et H. Kahn. Estimation of particle transmission by random sampling. *Natl. Bur. Stand. Appl. Math. Ser.*, 12:27–30, 1951.
[4] J.H. Holland. *Adaptation in Natural and Artificial Systems.* University of Michigan Press, Ann Arbor, 1975.
[5] J.H. Hetherington, Observations on the statistical iteration of matrices, *Physical Review A*, vol. 30, no. 5, pp. 2713–2719, 1984.
[6] D.D. Magill. Optimal adaptive estimation of sampled stochastic processes. *IEEE Trans. on Autom. Control*, 10(4):434–439, 1965.
[7] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, E.Teller, A.H. Teller. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 90:233–241, 1953.