

*Journées de Statistique et applications en biologie
Rennes Sept. 2005*

Modèles d'arbres généalogiques en estimation non linéaire

Pierre DEL MORAL

Lab J.A. Dieudonné, Dépt. Math., Univ. Nice-Sophia Antipolis

↔ *Feynman-Kac Formulae, Genealogical and Interacting Particle Systems
with Applications, Springer NY. Series : Probability and Applications, (2004)*

↔ (delmoral@math.unice.fr), ↔ [prépubli.+info.] <http://math.unice.fr/delmoral/>

Modèles évolutionnaires

Algo. génétiques	<i>Mutation</i> <i>Sélection/Branchement</i>
Algo. Metropolis-Hastings	<i>Proposition</i> <i>Acceptation/Rejet</i>
Méthodes de Monte Carlo séquentielles	<i>"Sampling"</i> <i>"Resampling (SIR)"</i>
Filtrage/lissage de signaux	<i>Prédiction</i> <i>Mise à jour/Correction</i>
Particule \in Milieu absorbant	<i>Évolution</i> <i>Mort/Création/Anihilation</i>

Autres terminologies : multi-level splitting (Khan-Harris 51), prune enrichment (Rosenbluth 1955), switching algo. (Magill 65), matrix reconfiguration (Hetherington 84), restart (Villen-Altamirano 91), particle filters (Rigal-Salut-DM 92), SIR filters (Gordon-Salmon-Smith 93, Kitagawa 96), go-with-the-winner (Vazirani-Aldous 94), ensemble Kalman-filters (Evensen 1994), quantum Monte Carlo methods (Melik-Nightingale 1999), spawning filters (Fisher-Maybeck 2002), SIR Pilot Exploration Resampling (Liu-Zhang 2002),...

\iff Interprétation particulière d'une formule de Feynman-Kac

Origine : Thèse de R. Feynman's sur les intégrales de chemins, Princeton 1942

Physique \longleftrightarrow Biologie \longleftrightarrow Sciences de l'ingénieur \longleftrightarrow Probabilités-Statistique

- **Physique :**
 - $FK \in$ éq. intégral-diff. non linéaire (\sim modèles de Boltzmann généralisés).
 - Analyse spectrale d'opérateurs de Schrödinger et de grandes matrices à entrées ≥ 0 .
(évolutions de particules dans des milieux désordonnés et absorbants)
 - Problèmes de Dirichlet avec conditions aux bords
 - Interprétations microscopiques et macroscopiques.
- **Biologie :**
 - Chaînes auto-évitantes, polymérisations macro-moléculaire.
 - Processus de branchement, évolutions génétique de populations.
 - Modèles de coalescences et d'arbres généalogiques.

- **Analyse d'évènements rares :**
 - Méthodes “Multi-splitting”, branchement par niveaux (“Restart”).
 - Échantillonnage et lois d'importance, changement de mesures de probabilités.
 - Techniques de simulation d'arbres généalogiques (\sim arbres des défaillances).

- **Nouvelles techniques de traitement du signal :**
 - Dualité filtrage/lissage et régulation optimale, contrôles particuliers en boucle ouverte.
 - Filtres de Kalman-Bucy en interaction (signaux partiellement linéaires-gaussiens).
 - Techniques de maillage aléatoire et adaptatif.
 - Estimation trajectorielle et arbres généalogiques.
 - Intelligence artificielle, apprentissage dynamique (\perp opérateur).

- **Probabilités et Statistique :**
 - Simulation de chaînes restreintes (conditions terminales fixées, régions visitées,...)
 - Analyse de mesures de Boltzmann-Gibbs (simulation, fonctions de partition,...).
 - Algorithmes d'exploration et de recherche évolutionnaires (algo. de Metropolis-Hasting, recuit simulés en interaction,...)

Évolution génétique simple \rightsquigarrow méthode de simulation



uniquement 2 ingrédients !!

(espace temps $n \in \mathbb{N} = \{0, 1, 2, \dots\}$, espace états $E_n (\in \{\mathbb{Z}^d, \mathbb{R}^d, \underbrace{\mathbb{R}^d \times \dots \times \mathbb{R}^d}_{(n+1)\text{-times}}, \text{chemins, excursions, \dots}\})$

- **Mutation/exploration/prédiction/proposition** : \rightarrow Transitions de Markov de E_{n-1} vers E_n .

$$\mathbb{P}(X_n \in dx_n \mid X_{n-1} = x_{n-1}) = M_n(x_{n-1}, dx_n)$$

- **Sélection/absorption/mise à jour/acceptation** : \rightarrow Fonctions potentiel de E_n dans $[0, 1]$.

$$G_n : x_n \in E_n \longrightarrow G_n(x_n) \in [0, 1]$$

Algorithme génétique \Leftrightarrow Chaîne de Markov $\xi_n = (\xi_n^1, \dots, \xi_n^N) \in E_n^N = \underbrace{E_n \times \dots \times E_n}_{N\text{-times}}$

$$\xi_n \in E_n^N \xrightarrow{\text{sélection}} \widehat{\xi}_n \in E_n^N \xrightarrow{\text{mutation}} \xi_{n+1} \in E_{n+1}^N$$

– **Sélection** ($\exists \neq$ types \rightarrow Ex. : accept/reject, branchements,...)

$$\xi_n^i \rightsquigarrow \widehat{\xi}_n^i = \xi_n^i \quad \text{avec proba. } G_n(\xi_n^i) \quad \textbf{[Acceptation]}$$

Sinon, on sélectionne un individu plus “performant” dans la configuration

$$\widehat{\xi}_n^i = \xi_n^j \quad \text{avec proba. } G_n(\xi_n^j) / \sum_{k=1}^N G_n(\xi_n^k) \quad \textbf{[Rejet + Sélection]}$$

– **Mutation**

$$\widehat{\xi}_n^i \rightsquigarrow \xi_{n+1}^i \sim M_{n+1}(\widehat{\xi}_n^i, \bullet) = \mathbb{P}(X_{n+1} \in \bullet \mid X_n = \widehat{\xi}_n^i)$$

Algo. Génétique dans d'espace des chemins



Évolution d'arbres généalogiques

$X_n = (X'_0, \dots, X'_n)$ transitions de Markov M_n et $G_n(X_n) = G'_n(X'_n)$



Modèle génétique trajectoriel

$$\begin{cases} \xi_n^i &= (\xi_{0,n}^i, \xi_{1,n}^i, \dots, \xi_{n,n}^i) \\ \widehat{\xi}_n^i &= (\widehat{\xi}_{0,n}^i, \widehat{\xi}_{1,n}^i, \dots, \widehat{\xi}_{n,n}^i) \in E_n = (E'_0 \times \dots \times E'_n) \end{cases}$$

- Acceptation/(rejet+sélection) de trajectoires.
- Mutations trajectorielles = extensions élémentaires de chemins.

Mesures empiriques ($\forall f_n$, fonction test sur E_n)

Mesures d'occupation

$$\eta_n^N(f_n) = \frac{1}{N} \sum_{i=1}^N f_n(\xi_n^i) = \frac{1}{N} \sum_{i=1}^N f_n \underbrace{(\xi_{0,n}^i, \xi_{1,n}^i, \dots, \xi_{n,n}^i)}_{i\text{-ème ligne ancestrale}}$$

Mesures non biaisées & formules de Feynman-Kac non normalisées

$$\gamma_n^N(f_n) = \eta_n^N(f_n) \times \prod_{0 \leq p < n} \eta_p^N(G_p) \xrightarrow{N \rightarrow \infty} \gamma_n(f_n) = \mathbb{E}(f_n(X_n) \prod_{0 \leq p < n} G_p(X_p))$$

Remarques :

– ($f_n = 1 \implies$) $\eta_n^N(1) = \prod_{0 \leq p < n} \eta_p^N(G_p) \xrightarrow{N \rightarrow \infty} \gamma_n(1) = \mathbb{E}(\prod_{0 \leq p < n} G_p(X_p))$

– *Modèles trajectoriels*

$$[X_n = (X'_0, \dots, X'_n) \text{ et } G_n(X_n) = G'_n(X'_n)] \Rightarrow \gamma_n(f_n) = \mathbb{E}(f_n(X'_0, \dots, X'_n) \prod_{0 \leq p < n} G'_p(X'_p))$$

\implies Mesures d'occupation & Formules de Feynman-Kac normalisées :

$$\eta_n^N(f_n) = \frac{1}{N} \sum_{i=1}^N f_n(\xi_n^i) = \gamma_n^N(f_n) / \gamma_n^N(\mathbf{1}) \xrightarrow{N \rightarrow \infty} \eta_n(f_n) = \gamma_n(f_n) / \gamma_n(\mathbf{1})$$

Remarques :

Formules non normalisées :

$$\gamma_n(f_n) = \eta_n(f_n) \times \prod_{0 \leq p < n} \eta_p(G_p) \quad (\leftarrow \gamma_n^N(f_n) = \eta_n^N(f_n) \times \prod_{0 \leq p < n} \eta_p^N(G_p))$$

Modèles trajectoriels :

$$[X_n = (X'_0, \dots, X'_n) \text{ et } G_n(X_n) = G'_n(X'_n)] \implies \eta_n(f_n) = \frac{\mathbb{E}(f_n(X'_0, \dots, X'_n) \prod_{0 \leq p < n} G'_p(X'_p))}{\mathbb{E}(\prod_{0 \leq p < n} G'_p(X'_p))}$$

Théorie asymptotique “assez complète “

$(n, N) \rightarrow \infty$ (LGN faible+forte, TCL, est. expo., PGD, processus empiriques,...)

QQ résultats :

– Convergence “faible “ [$p \geq 1$ + \mathcal{F}_n pas trop grande + mutations régulières]

$$\sup_{n \geq 0} \mathbb{E}(\sup_{f_n \in \mathcal{F}_n} |\eta_n^N(f_n) - \eta_n(f_n)|^p)^{1/p} \leq c(p)/\sqrt{N}$$

Exemple :

$$E_n = \mathbb{R}, \quad \mathcal{F}_n = \{1_{]-\infty, x]} ; x \in \mathbb{R}\} \Rightarrow \sup_{n \geq 0} \mathbb{E}(\sup_{x \in \mathbb{R}} |\eta_n^N(1_{]-\infty, x]} - \eta_n(1_{]-\infty, x]})|^p)^{1/p} \leq c(p)/\sqrt{N}$$

– Propagations du chaos [tailles blocs finis $q \leq N$]

$$\text{Loi}(\xi_n^1, \dots, \xi_n^q) \simeq \eta_n^{\otimes q} + \frac{1}{N} \mathcal{M}_n^{(q)} \quad \text{avec} \quad \mathcal{M}_n^{(q)} \quad \text{mesure signée t.q.} \quad \sup_{n \geq 0} \sup_{\|F\| \leq 1} |\mathcal{M}_n^{(q)}(F)| \leq c q^2$$

Physique : Markov $X_n \in$ Milieu absorbant, taux $G(x) = e^{-V(x)} \in [0, 1]$

$$X_n^c \in E^c = E \cup \{c\} \xrightarrow{\text{absorption}} \widehat{X}_n^c \xrightarrow{\text{exploration}} X_{n+1}^c$$

Absorption : $\longrightarrow \widehat{X}_n^c = X_n^c$, avec probab. $G(X_n^c)$; sinon la particule est tuée $\widehat{X}_n^c = c$.

\Downarrow

$A = \{x : G(x) = 0\} \longrightarrow$ Obstacles durs

$T = \inf \{n \geq 0 ; \widehat{X}_n^c = c\} \longrightarrow$ Temps d'absorption/durée de vie $X_{T+n}^c = \widehat{X}_{T+n}^c = c$

\implies Modèles de Feynman-Kac (G, M_n) : $\gamma_n = \text{Loi}(X_n^c ; T \geq n)$ et $\gamma_n(1) = \text{Proba}(T \geq n)$

\Downarrow

$$\eta_n = \text{Loi}(X_n^c \mid T \geq n) = \text{Loi}((X_0^{lc}, \dots, X_n^{lc}) \mid T \geq n)$$

Biologie-Chimie : [Macro-molécules + polymères dirigés+branchement]

- Marches auto-évitantes $X'_n \in \mathbb{Z}^d$

$$X_n = (X'_0, \dots, X'_n) \quad \text{et} \quad G_n(X_n) = 1_{\notin\{X'_0, \dots, X'_{n-1}\}}(X'_n)$$

↓

$$\gamma_n(1) = \text{Proba}(\forall 0 \leq p \neq q \leq n, X'_p \neq X'_q) \left(= \frac{|\mathcal{A}_n|}{(2d)^n} \right) \quad \text{et} \quad \eta_n = \text{Loi}(X'_0, \dots, X'_n \mid \forall 0 \leq p \neq q \leq n, X'_p \neq X'_q)$$

- Modèle de polymère d'Edwards

$$X_n = (X'_0, \dots, X'_n) \quad \text{et pénalisation} \quad G_n(X_n) = \exp \left\{ -\beta \sum_{0 \leq p < n} 1_{X'_p}(X'_n) \right\}$$

– **Processus de mutation/branchement** : Markov $\in S = \cup_{p \geq 0} E^p$, avec $E^0 = \{c\}$.

$$\chi_n = (\chi_n^1, \dots, \chi_n^{p_n}) \in E^{p_n} \xrightarrow{\text{branchement}} \widehat{\chi}_n \in \widehat{E}^{p_n} \xrightarrow{\text{mutation} \sim M_n} \chi_{n+1} \in E^{p_{n+1}}$$

Loi de branchement : $g_n(x)$ individus sur chaque site x , t.q. $\mathbb{E}(g_n(x)) = G_n(x) \in \mathbb{R}^+$.

$$\Downarrow [\chi_0 = x_0 \in E]$$

FK=premier moment :

$$\mathbb{E} \left(\sum_{i=1}^{p_n} f(\chi_n^i) \right) = \mathbb{E} \left(f(X_n) \prod_{q=0}^{n-1} G_q(X_q) \right) = \gamma_n(f) \quad (\text{N.b. : } \mathbb{E}(p_n) = \gamma_n(1))$$

Statistique : MCMC séquentiel et modèles de Feynman-Kac-Metropolis

Potentiel/rapport de Metropolis [π loi cible]+[(K, L) paire de transitions Markov]

$$G(y_1, y_2) = \frac{\pi(dy_2)L(y_2, dy_1)}{\pi(dy_1)K(y_1, dy_2)}$$

Exemple : Mesure de Gibbs cible

$$\pi(dy) \propto e^{-V(y)} \lambda(dy) \implies G(y_1, y_2) = e^{(V(y_1)-V(y_2))} \frac{\lambda(dy_2)L(y_2, dy_1)}{\lambda(dy_1)K(y_1, dy_2)}$$

Remarque :

$$(K = L \quad \lambda - \text{réversible}) \quad \text{ou} \quad (\lambda K = \lambda \text{ et } L(y_2, dy_1) = \lambda(dy_1) \frac{dK(y_1, \cdot)}{d\lambda}(y_2))$$

↓

$$G(y_1, y_2) = \exp(V(y_1) - V(y_2))$$

Notation : $\mathbb{E}_\nu^M(\cdot)$ = Opérateur moyenne \sim Markov [transition M , condition initiale ν]

Theorem : (Formule d'inversion du temps), [A. Doucet, P.DM; (Séminaire Probab. 2003)]

$$\mathbb{E}_\pi^L(f_n(Y_n, Y_{n-1}, \dots, Y_0) | Y_n = y) = \frac{\mathbb{E}_y^K(f_n(Y_0, Y_1, \dots, Y_n) \{\prod_{0 \leq p < n} G(Y_p, Y_{p+1})\})}{\mathbb{E}_y^K(\{\prod_{0 \leq p < n} G(Y_p, Y_{p+1})\})}$$

De Plus :

⊕ Loi n -marginale FK-Metropolis : $\lim_{n \rightarrow \infty} \eta_n = \pi$ (vitesse $\perp \pi$)

⊕ Modèles inhomogènes : (π_n, L_n, K_n)

$\pi_n(dy) \propto e^{-\beta_n V(y)} \lambda(dy)$, inverse de température $\beta_n \uparrow \infty$, mutation t.q. $\pi_n = \pi_n K_n$, et $\text{Loi}(X_0) = \pi_0$

↓

$$G_n(y_1, y_2) = \exp[-(\beta_{n+1} - \beta_n)V(y_1)] \quad (\xrightarrow{n \rightarrow \infty} 1)$$

↓

$$\eta_n = \pi_n \quad (\sim \text{recuit simulé "stationnaire"})$$

Analyse d'évt. Rares [2 techniques]

1. Lois d'importance de type Feynman-Kac

$$\mathbb{P}(V_n(X_n) \geq a) = \mathbb{E}(\mathbf{1}_{V_n(X_n) \geq a} e^{-\beta_n V_n(X_n)} e^{+\beta_n V_n(X_n)})$$

Potentiels favorisant la croissance de $V_n(X_n)$: $G_n(X_n, X_{n+1}) = e^{\beta_n(V_{n+1}(X_{n+1}) - V_n(X_n))} \longrightarrow \text{FK}(G_n, X_n)$

\Downarrow

$$\mathbb{P}(V_n(X_n) \geq a) = \gamma_n(\mathbf{1}_{V_n \geq a} e^{-\beta_n V_n})$$

$$\mathbb{E}(f_n(X_n) \mid V_n(X_n) \geq a) = \eta_n(f_n \mathbf{1}_{V_n \geq a} e^{-\beta_n V_n}) / \eta_n(\mathbf{1}_{V_n \geq a} e^{-\beta_n V_n})$$

⊕ Modèles trajectoriels \Rightarrow Arbres généalogiques pondérés

$$X_n = (X'_0, \dots, X'_n) \text{ et } V_n(X_n) = V'_n(X'_n)$$

\Downarrow

$$\mathbb{E}(f_n(X'_0, \dots, X'_n) \mid V'_n(X'_n) \geq a) = \eta_n(f_n \mathbf{1}_{V_n \geq a} e^{-\beta_n V_n}) / \eta_n(\mathbf{1}_{V_n \geq a} e^{-\beta_n V_n})$$

2. Branchements par niveaux (\neq échantillonnage d'importance)

$(E = A \cup A^c)$, Y_n Markov, $Y_0 \in A_0 (\subset A) \rightsquigarrow A^c = (B \cup C)$, $C =$ région absorbante/obstacle dur

Décomposition par niveaux $B = B_m \subset \dots \subset B_1 \subset B_0$ ($A_0 = B_1 - B_0$, $B_0 \cap C = \emptyset$)

\Downarrow

$$\mathbb{P}(Y_n \text{ touche } B \text{ avant } C) = \mathbb{E}\left(\prod_{1 \leq p \leq m} G_p(X_p)\right)$$

Excursions entre niveaux : $T_n = \inf \{p \geq T_{n-1} : Y_p \in B_n \cup C\}$

$$X_n = (Y_p ; T_{n-1} \leq p \leq T_n) \in \text{Espace d'excursions et } G_n(X_n) = \mathbf{1}_{B_n}(Y_{T_n})$$

\Downarrow

Interprétation FK

$$\mathbb{E}(f(Y_0, \dots, Y_{T_m}) \mathbf{1}_{B_m}(X_{T_m})) = \mathbb{E}(f(X_0, \dots, X_m) \prod_{1 \leq p \leq m} G_p(X_p))$$

Traitement du signal [Filtrage/Ch. de Markov cachées/méth. bayésienne]

Signal

$X_n = \text{Chaîne de Markov} \in E_n$

Observation

$Y_n = H_n(X_n, V_n) \in F_n$ avec $\mathbb{P}(H_n(x_n, V_n) \in dy_n) = g_n(x_n, y_n) \lambda_n(dy_n)$

Exemple de capteur : $Y_n = h_n(X_n) + V_n$ ($\in F_n = \mathbb{R}$), avec bruit gaussien $V_n = \mathcal{N}(0, 1)$

↓

$$\mathbb{P}(h_n(x_n) + V_n \in dy_n) = (2\pi)^{-1/2} e^{-\frac{1}{2}(y_n - h_n(x_n))^2} dy_n = \underbrace{\exp[h_n(x_n)y_n - h_n^2(x_n)/2]}_{g_n(x_n, y_n)} \underbrace{\mathcal{N}(0, 1)(dy_n)}_{\lambda_n(dy_n)}$$

Prédiction/filtrage/lissage \Leftrightarrow Formule de Feynman-Kac avec $G_n(x_n) = g_n(x_n, y_n)$

$$\eta_n = \text{Loi}(X_n \mid Y_0 = y_0, \dots, Y_{n-1} = y_{n-1}) = \text{Loi}(X'_0, \dots, X'_n \mid Y_0 = y_0, \dots, Y_{n-1} = y_{n-1})$$

Remarque : Signaux/capteurs plus généraux \rightarrow J.P. Vila et V. Rossi (INRA Montpellier)

Modèles partiellement linéaires-gaussien

$$X_n^1 = \text{Markov} \in E_n \quad + \quad \begin{cases} X_n^2 = A_n(X_n^1) X_{n-1}^2 + a_n(X_n^1) + B_n(X_n^1) W_n \in \mathbb{R}^d \\ Y_n = C_n(X_n^1) X_n^2 + c_n(X_n^1) + D_n(X_n^1) V_n \in \mathbb{R}^d \end{cases}$$

Connaissant une réalisation $X^1 = x \rightarrow$ prédicteur optimal de Kalman-Bucy

$$\hat{X}_{x,n+1}^{2-} = \mathbb{E}(X_{n+1}^2 \mid Y_0, \dots, Y_n, X^1 = x) \quad \text{et} \quad P_{x,n+1}^- = \mathbb{E}([X_{n+1}^2 - \hat{X}_{x,n+1}^{2-}][X_{n+1}^2 - \hat{X}_{x,n+1}^{2-}]')$$

↓

Éq. de Kalman-Bucy = correction + prédiction | $(X^1 = x)$:

$$(\hat{X}_{x,n+1}^{2-}, P_{x,n+1}^-) = \mathcal{B}_{n,(x_n, x_{n+1})}(\hat{X}_{x,n}^{2-}, P_{x,n}^-)$$

Représentation de type Feynman-Kac : $\eta_n \sim (\mathbf{X}_n, \mathbf{G}_n)$ tels que

$$\begin{aligned}
 \mathbf{X}_n &= (X_n^1, (\widehat{X}_{X^1, n+1}^{2-}, P_{X^1, n+1}^-)) \text{ Chaîne de Markov } \in \mathbf{E}_n = (E_n \times \mathbb{R}^d \times \mathbb{R}^{d \times d}) \\
 \mathbf{G}_n(x, m, P) &= \frac{d\mathcal{N}(C_n(x) m + c_n(x), C_n(x) P C_n(x)' + D_n(x) R_n^v D_n(x)')}{d\mathcal{N}(0, D_n(x) R_n^v D_n(x)')} (y_n) \\
 &\propto \text{"Prob}(Y_n \in dy_n \mid X_n^1 = x, \widehat{X}_{X^1, n}^{2-} = m, P_{X^1, n}^- = P)\text{"} \\
 &\sim [\text{capteur virtuel} : Y_n = \{C_n(X_n^1) \widehat{X}_{X^1, n}^{2-} + c_n(X_n^1)\} + \widehat{V}_{X^1, n}]
 \end{aligned}$$

$$\begin{aligned}
 F_n(x, m, P) = f_n(x) &\implies \eta_n(F_n) = \mathbb{E}(f_n(X_n^1) \mid Y_0, \dots, Y_{n-1}) \\
 F_n(x, m, P) = \mathcal{N}(m, P)(f_n) &\implies \eta_n(F_n) = \mathbb{E}(f_n(X_n^2) \mid Y_0, \dots, Y_{n-1})
 \end{aligned}$$

Remarque : \rightsquigarrow Filtres de Kalman-Bucy en interaction, \oplus modèles trajectoriels \rightarrow Arbres génés.

$$X_n^1 = (X_0^1, \dots, X_n^1) \rightsquigarrow \text{Loi}((X_0^1, \dots, X_n^1) \mid Y_0, \dots, Y_{n-1})$$