Estimation de dynamique effective à mémoire pour des systèmes complexes

Pierre Monmarché

Rencontres chercheur·euse·s et ingénieur·e·s

Collaboration avec Hadrien Vroylandt, Ludovic Goudenège, Fabio Pietrucci et Benjamin Rotenberg.

Simulation moléculaire “tous atomes”

Système : N atomes repérés par les coordonnées $x \in \mathbb{R}^{3N}$ des noyaux.

Dynamique : hamiltonienne (M la masse, U l'énergie) :

$$M\ddot{x}_t = -\nabla U(x_t)$$

ou diffusion de Langevin ($\gamma > 0$ la friction, $(\zeta_t)_{t \geqslant 0}$ un bruit blanc) :

$$M\ddot{x}_t = -\nabla U(x_t) - \gamma \dot{x}_t + \zeta_t$$

Simulation moléculaire "tous atomes"

Système : N atomes repérés par les coordonnées $x \in \mathbb{R}^{3N}$ des noyaux.

Dynamique : hamiltonienne (M la masse, U l'énergie) :

$$M\ddot{x}_t = -\nabla U(x_t)$$

ou diffusion de Langevin ($\gamma > 0$ la friction, $(\zeta_t)_{t \geqslant 0}$ un bruit blanc) :

$$M\ddot{x}_t = -\nabla U(x_t) - \gamma \dot{x}_t + \zeta_t$$

Caractéristiques :

- Grande dimension ($N \simeq 10^5$), calcul de ∇U coûteux.
- Multi-échelle : période d'oscillation d'un atome d'hydrogène (1 fs$= 10^{-15}$s) contre temps de changement de conformation macroscopique (1ms, voir 1s).
- Multi-modal/métastable : U non convexe, les transitions d'un minimum local d'énergie à un autre sont des événements rares.

Coordonnées de réaction, variables collectives

On s'intéresse à $\xi(x)$ une description plus grossière du système, où $\xi : \mathbb{R}^{3N} \to \mathbb{R}^d$, $d < 3N$.

Exemples :

- Modèles gros grain, $\xi(x) =$ coordonnées du barycentre de groupes de molécules (ex : molécule d'eau).
- Une distance entre deux atomes particuliers (ou le centre d'un ligand et le centre du site d'une protéine)
- Un angle formé par trois atomes

Coordonnées de réaction, variables collectives

On s'intéresse à $\xi(x)$ une description plus grossière du système, où $\xi : \mathbb{R}^{3N} \to \mathbb{R}^d$, $d < 3N$.

Exemples :

- Modèles gros grain, $\xi(x) =$ coordonnées du barycentre de groupes de molécules (ex : molécule d'eau).
- Une distance entre deux atomes particuliers (ou le centre d'un ligand et le centre du site d'une protéine)
- Un angle formé par trois atomes

Objectif : décrire (approximativement) la dynamique de $z_t = \xi(x_t)$ à l'aide d'une équation fermée en z_t (une *dynamique effective*).

- Description intelligible d'un phénomène d'intérêt.
- Simulations moins coûteuses.

Équations de Langevin généralisées

Modèle :

$$\ddot{z}_t = F(z_t) - \int_0^t K(t-s)\dot{z}_s \mathrm{d}s + \zeta_t$$

avec

- F une force effective.
- $K(t)$ un noyau de mémoire/friction.
- $(\zeta_t)_{t\geqslant 0}$ un bruit aléatoire.

Motivation : on peut écrire exactement cette équation pour $\xi(x_t)$, mais avec ζ_t remplacé par une quantité dépendant de la condition initiale x_0 et, si x_0 est aléatoire (distribué selon la loi de Gibbs associé à l'énergie U), décorrélée pour tout $t \geqslant 0$ de $\xi(x_0)$ (et certaines fonctions de $\xi(x_0)$).

Approche par variables auxiliaires

Quand $K(t) = Ae^{tB}C$ avec des matrices A, B, C constantes, on peut se ramener à un processus étendu Markovien (sans mémoire) :

$$(\star) \qquad \begin{cases} \dot{z}_t &= v_t \\ \dot{v}_t &= F(z_t) - M_{11}v_t - M_{12}h_t + \Sigma_1\zeta_t \\ \dot{h}_t &= -M_{21}v_t - M_{22}h_t + \Sigma_2\zeta_t \end{cases}$$

avec M_{ij}, Σ_i des matrices constantes et ζ un bruit blanc.

Approche par variables auxiliaires

Quand $K(t) = Ae^{tB}C$ avec des matrices A, B, C constantes, on peut se ramener à un processus étendu Markovien (sans mémoire) :

$$(\star) \qquad \begin{cases} \dot{z}_t &= v_t \\ \dot{v}_t &= F(z_t) - M_{11}v_t - M_{12}h_t + \Sigma_1\zeta_t \\ \dot{h}_t &= -M_{21}v_t - M_{22}h_t + \Sigma_2\zeta_t \end{cases}$$

avec M_{ij}, Σ_i des matrices constantes et ζ un bruit blanc.

Problème statistique :

- Données : réalisation $(z_k)_{k\in[\![1,N]\!]}$ avec $z_k = \xi(x_{k\Delta t})$.
- Modèle : un schéma d'Euler de $(\star)$ (variables h cachées)

$$\begin{cases} z_{k+1} &= z_k + \Delta t v_t \\ v_{k+1} &= v_k + \Delta t\left(F(z_k)t - M_{11}v_k - M_{12}h_k\right) + \sqrt{\Delta t}\Sigma_1 G_k \\ h_{k+1} &= h_k - \Delta t\left(M_{21}v_k + M_{22}h_k\right) + \sqrt{\Delta t}\Sigma_2 G_k \end{cases}$$

- Paramètres : nombre de variables auxiliaires, F matrices M, Σ (et F)
- Approche : maximiser la vraisemblance de la trajectoire observée.

Algorithme EM (espérance/maximisation) en général

Observations Y, variables cachées W. Modèle : (Y, W) est une variable aléatoire de densité $f_\theta(y, w)$, de paramètre θ inconnu. On a observé une réalisation y.

Estimateur du maximum de vraisemblance :

$$\theta = \operatorname{argmax} \int f_\theta(y, w) \mathrm{d}w$$

Problème : contrairement à la densité jointe, la densité marginale $\int f_\theta(y, w) \mathrm{d}w$ de Y n'est pas explicite ou ne donne pas un problème d'optimisation facile à résoudre.

Algorithme EM (espérance/maximisation) en général

Observations Y, variables cachées W. Modèle : (Y, W) est une variable aléatoire de densité $f_\theta(y, w)$, de paramètre θ inconnu. On a observé une réalisation y.

Estimateur du maximum de vraisemblance :

$$\theta = \operatorname{argmax} \int f_\theta(y, w) \mathrm{d}w$$

Problème : contrairement à la densité jointe, la densité marginale $\int f_\theta(y, w) \mathrm{d}w$ de Y n'est pas explicite ou ne donne pas un problème d'optimisation facile à résoudre.

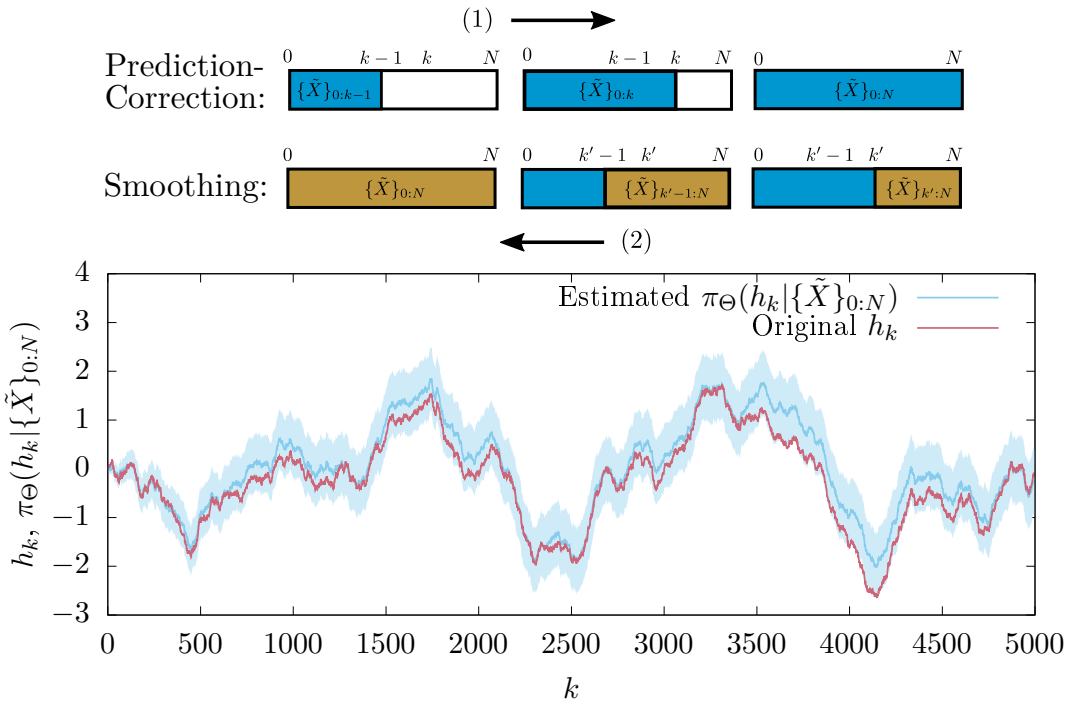
Algorithme EM : itératif (paramètres $(\theta_k)_{k \in \mathbb{N}}$), on alterne deux étapes :

- Estimation des variables cachées (si le paramètre était la valeur actuelle θ_k) sachant qu'on a observé y.
- Optimisation en θ (détermine θ_{k+1}) de la densité jointe f_θ moyennée contre la loi conditionnelle des variables cachées.

Application à notre cas

On a $Y = (z_k, v_k)_{k \in [\![1,N]\!]}$ et $W = (h_k)_{k \in [\![1,N]\!]}$.

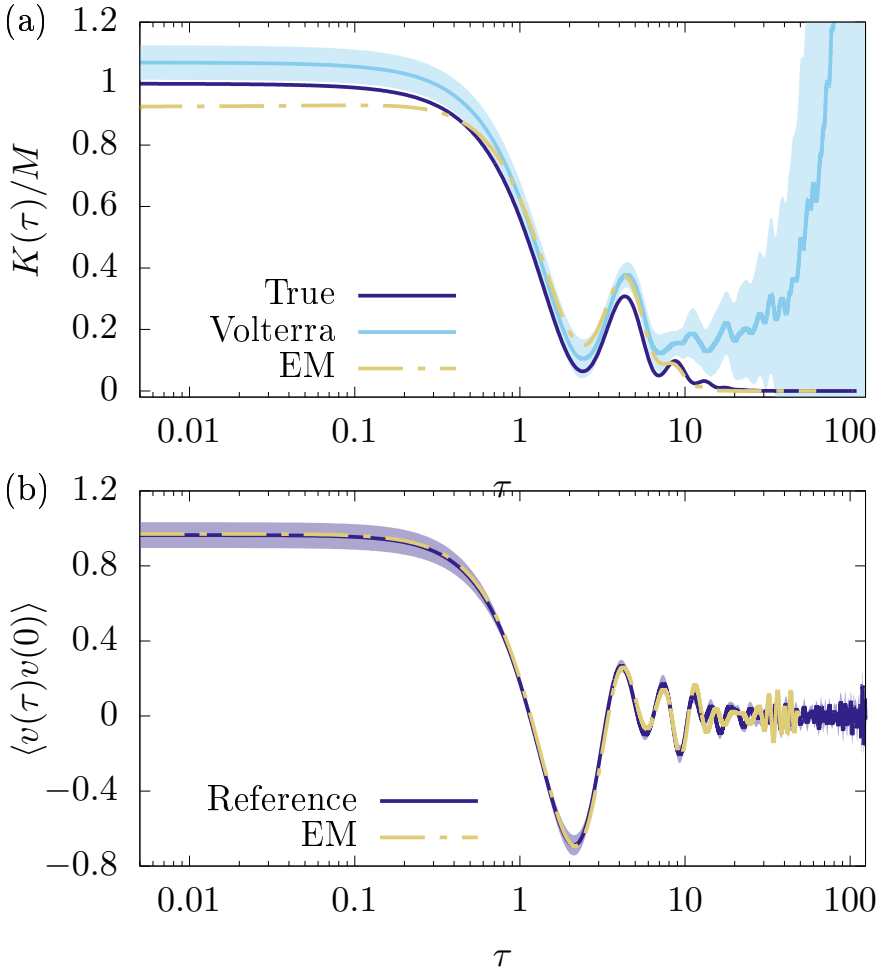
- La densité jointe est explicite (transition schéma d'Euler + Markov).
- La densité conditionnelle de W sachant Y est gaussienne, calculable :



- Le problème d'optimisation sur θ obtenu après estimation des variables cachées est similaire au problème du maximum de vraisemblance pour des transitions gaussiennes sans variables cachées.

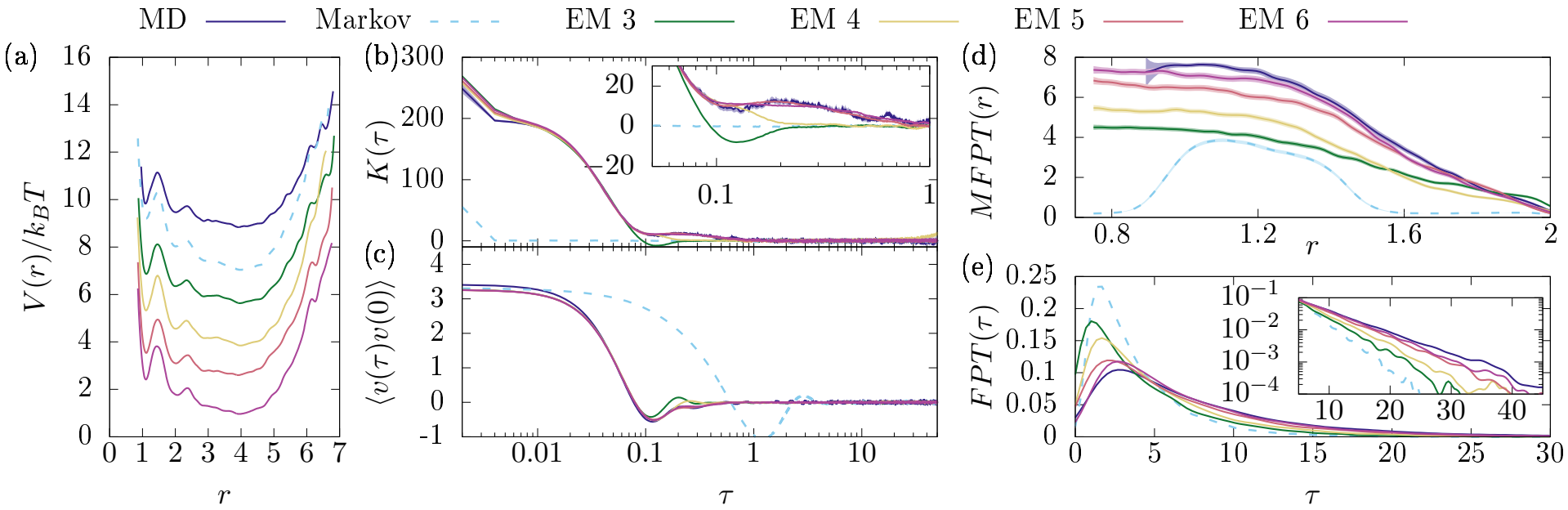
Résultats numériques

Problème jouet (reconstruction d'une équation de Langevin généralisée connue, dimension 1, 5 variables cachées)



Résultats numériques

Dimère parmi 512 particules de Lennard-Jones (force effective, noyau de friction, covariance des vitesses, temps moyen de premiers passages et distribution des temps de premiers passages entre états *contact pair* et *solvent shared pair*).



Perspectives

- Applications :
  - En cours, notamment pour des problèmes de nucléation de cristaux ioniques pour des batteries au lithium.
  - Autres (réactions chimiques en solution, changements conformationnels pour les biomolécules, transition de phase dans la matière condensée...).
- Développements méthodologiques :
  - Robustesse au choix du pas de temps et de l'intégrateur (Euler-Maruyama $\Rightarrow$ schéma de splitting).
  - Utilisation des GLE pour sélectionner des coordonnées de réaction.
  - Bruit non gaussien.