Finding slow modes and accessing very long timescales in molecular dynamics

Frank Noé (FU Berlin)
frank.noe@fu-berlin.de
Proteins

McGufee and Elcock, PloS Comput Biol 2010
Protein-Protein binding
Protein-Protein binding

Plattner, Doerr, De Fabritiis, Noé

0.1 microseconds
Part I

Direct approximation of slow modes / processes


Propagator

$$\rho_{t+\tau} = \mathcal{P}_\tau \rho_t = \int p_\tau(y \mid x) \rho_t(x) \, dx$$

Transfer operator / Perron-Frobenius operator

= propagator for densities $u(x) = \frac{\rho(x)}{\pi(x)}$ with stationary density $\pi(x)$.

$$u_{t+\tau} = \mathcal{T}_\tau u_t = \int \frac{\pi(x)}{\pi(y)} p_\tau(y \mid x) \rho_t(x) \, dx$$

Koopman operator

Adjoint to $\mathcal{P}$, adjoint to $\mathcal{T}$ with respect to $\pi$

$$f_{t+\tau} = \mathcal{K}_\tau f_t = \int p_\tau(y \mid x) f_t(y) \, dy = \mathbb{E}[f_{t+\tau}(x)]$$

with detailed balance: $\pi(x) p_\tau(y \mid x) = \pi(y) p_\tau(x \mid y)$ we have $\mathcal{K}_\tau \equiv \mathcal{T}_\tau$

Slow processes

Backward propagator

\[ \rho_\tau = \mathcal{T}(\tau)\rho_0 \]

Spectral decomposition

\[ \rho_\tau = \sum_{i=1}^{\infty} e^{-\tau \kappa_i} \langle \psi_i | \rho_0 \rangle \psi_i \]

Processes:

1. Eigenvalues / timescales \( \kappa_i \)
2. Backward propagator
3. Spectral decomposition

Collective variables: Slow modes versus reaction coordinates

Identify the slowest processes

\[
\rho_\tau = \sum_{i=1}^{\infty} e^{-\tau \kappa_i} \langle \psi_i | \rho_0 \rangle \psi_i
\]

Identify the “best” / “slowest” coordinate connecting A and B

\[
\tau_A, \tau_B \quad \text{time to hit } A, B \text{ next}
q_{AB}(x) = \text{Prob}(\tau_B < \tau_A)
\]

Approximate Markov operator **eigenfunctions**

Approximate **comittor**

Eigenfunctions and committor are closely related when considering reactions between metastable states

Schütte, Noé, Lu, Sarich, Vanden-Eijnden JCP 134, 204105
Molecular order parameters

Candidates $x_1 \ldots x_n$

- Cartesian coordinates
- Distances between heavy atoms or Ca’s
- $\cos/\sin$ of dihedrals
- Gaussian densities around sampled configurations (e.g. in RMSD space)
- Field variables or coordination numbers
- Your favorite order parameter
- All of them

Noé and Nüske, **MMS** 11, 635-655 (2013)
Nüske et al, **JCTC** 10, 1739-1752 (2014)
Slow processes

Backward propagator

\[ \rho_\tau = \mathcal{T}(\tau) \rho_0 \]

Eigenvalues / timescales \( \kappa_i^{-1} \)

Spectral decomposition

\[ \rho_\tau = \sum_{i=1}^{\infty} e^{-\tau \kappa_i} \langle \psi_i | \rho_0 | \psi_i \rangle \]

Processes:

(a) Energy U(s)
(b) Transition from A to D
(c) \( \phi_1, \phi_2, \phi_3, \phi_4 \)
(d) \( \psi_1, \psi_2, \psi_3, \psi_4 \)
Variational approach for Markov processes

The first $m$ eigenfunctions $\psi_1, \ldots, \psi_m$ are the solution to the problem

$$\max_{f_1, \ldots, f_m} \sum_{i=1}^{m} \mathbb{E} \left[ f_i(x_t) f_i(x_{t+\tau}) \right]$$

s.t. $\mathbb{E} \left[ f_i(x_t)^2 \right] = 1$
$$\mathbb{E} \left[ f_i(x_t) f_j(x_{t+\tau}) \right] = 0, \text{ for } i \neq j$$

and the maximum value is the sum of $\lambda_1, \ldots, \lambda_m$

Properties:

- $\psi_i$ and $\psi_j$ are uncorrelated for $i \neq j$.
- $\psi_i$ are the directions of slow kinetics with maximal autocorrelations $\mathbb{E}_\mu \left[ \psi_i(x_t) \psi_i(x_{t+\tau}) \right] = \lambda_i(\tau)$.
- Population changes along $\psi_i$ coordinates decay with $\lambda_i(\tau) = e^{-\frac{\tau}{\xi_i}}$.
- For every other set of functions, the eigenvalues will be underestimated $\hat{\lambda}_i(\tau) \leq \lambda_i(\tau)$.

Noé and Nüske, MMS 11, 635-655 (2013)
Nüske et al, JCTC 10, 1739-1752 (2014)
Variational approach for Markov processes: Linear variation

**Ansatz:** Define basis set $\chi = [\chi_1(x), ..., \chi_n(x)]^T$ and seek the linear expansions:

$$\hat{\psi}_i(x) = \sum_j r_{ij} \chi_j(x)$$

**Algorithm:**

1. Estimate $C^0$ and $C^\tau$ with:

   $$c_{ij}^0 = \mathbb{E}_t [\chi_i(x_t)\chi_j(x_t)]$$
   $$c_{ij}^\tau = \mathbb{E}_t [\chi_i(x_t)\chi_j(x_{t+\tau})]$$

2. Solve

   $$C^\tau r_i = C^0 \hat{\lambda}_i r_i$$

3. Expand

   $$\hat{\psi}_i(x) = \sum_j r_{ij} \chi_j(x)$$

Variational approach for Markov processes: Empirical estimation

1. Define

\[ X_0 = \begin{bmatrix} \chi_1(x_0) & \cdots & \chi_n(x_0) \\ \vdots & \ddots & \vdots \\ \chi_1(x_{T-\tau}) & \cdots & \chi_n(x_{T-\tau}) \end{bmatrix} \quad X_\tau = \begin{bmatrix} \chi_1(x_\tau) & \cdots & \chi_n(x_\tau) \\ \vdots & \ddots & \vdots \\ \chi_1(x_T) & \cdots & \chi_n(x_T) \end{bmatrix} \]

2. Empirical covariance matrices: \( C^0 \) and \( C^\tau \) with:

\[ C^0 = X_0^T X_0 \]
\[ C^\tau = X_0^T X_\tau \]

3. Solve

\[ C^\tau r_i = C^0 \hat{\lambda}_i r_i \]

4. The projections

\[ \Psi = XR \]

approximate the transfer operator eigenfunctions on the sampled configurations \( x_t \).

Noé and Nüske, **MMS** 11, 635-655 (2013)
Nüske et al, **JCTC** 10, 1739-1752 (2014)
Linear combination of basis functions

\[ \psi_i \text{ are the directions of slow kinetics with maximal autocorrelations} \]
\[ \mathbb{E}_\mu [\psi_i (x_t) \psi_i (x_{t+\tau})] = \lambda_i (\tau) = e^{-\frac{\tau}{\tau_i}}. \]

Optimal for variational principle of conformation dynamics (maximum autocorrelation)

maximum variance

Simple Ansatz:

\[ \psi_2(x, y) = ax + by \]

Noé and Nüske, MMS 11, 635-655 (2013)
Nüske et al, JCTC 10, 1739-1752 (2014)
Comparison with other algorithms

Extended dynamic mode decomposition (EDMD)

Time-lagged independent component analysis TICA / Blind source separation

Dynamic mode decomposition (DMD)

Markov state models (MSM)
see also papers by: Noé, Pande, Hummer, Weber, Swope, …
Extended dynamic mode decomposition

Relationship to Extended dynamic mode decomposition:
The VAMP eigenvalue problem can be written

\[ Kr_i = r_i \lambda_i \]

with Koopman matrix

\[ K = (C^0)^{-1} C^\tau \]
\[ = (X_0^T X_0)^{-1} X_0^T X_{\tau} = X_0^+ X_{\tau}^T \]

That is the optimal solution of the regression problem

\[ Y = X K \]

Remarks:

- The VAMP algorithm is equivalent to EDMD.
- VAMP/EDMD are also applicable to nonreversible and nonstationary dynamics.
- VAMP/EDMD results are variationally optimal and we have an eigenvalue bound in the reversible case.
- EDMD and DMD are usually implemented using singular value decomposition of the data matrix instead of computing a covariance matrix.

Williams, Kevrikidis, Rowley
Special case: Time-lagged independent component analysis (TICA)

Also known as: Blind source separation, Molgedey-Schuster transform

**Original paper:** Molgedey and Schuster, *PRL* 1994

TICA is the variationally optimal approximation to the Markov operator eigenfunctions when using **linear combinations of mean-free molecular coordinates**

Applications of TICA

Molecular Dynamics

Identification of slow molecular order parameters for Markov model construction

Naritomi and Fuchigami, *JCP* (2011)
Slow dynamics in protein fluctuations …

Improvements in Markov state modeling …
Special case: Markov state model

Choice of basis set: Characteristic functions on sets $S_1, \ldots, S_n$ that partition phase space

$$\chi_i(x_t) = \begin{cases} 1 & x_t \in S_i \\ 0 & \text{else.} \end{cases}$$

where $q$ are the position coordinates of the system.

Covariance matrices:

$$C^0 = X_0^\top X_0 = \text{diag}(\sum_k N_{1k}(\tau), \ldots, \sum_k N_{nk}(\tau))$$

$$C^\tau = X_0^\top X_\tau = [N_{ij}(\tau)]_{i,j=1,\ldots,n}$$

where $N_{ij}(\tau)$ are the number of transitions from state $i$ to $j$ at lag time $\tau$.

GEV problem:

$$C^\tau R = C^0 R \hat{A}$$

$$PR = R \hat{A}$$

with

$$P = [p_{ij}(\tau)]$$

$$p_{ij}(\tau) = \frac{N_{ij}(\tau)}{\sum_k N_{ik}(\tau)}$$
Related: Dynamic mode decomposition (DMD)


DMD: computes left eigenvectors of $K$ $\implies$ dynamic modes
TICA/VAMP compute right eigenvectors of $K$ $\implies$ direct approximation of eigenfunctions

Examples
Variational Approach
Noé and Nüske, MMS 11, 635-655 (2013)
Nüske et al, JCTC 10, 1739-1752 (2014)

Perez-Hernandez et al, JCP, 139, 1502 (2013)
Identification of slow molecular order parameters for Markov model construction
Variational Approach


Variational Approach

Kinetic map:
Expansion in molecular order parameters

Candidates $x_1 \ldots x_n$

- Cartesian coordinates
- Distances between heavy atoms or Ca’s
- cos/sin of dihedrals
- Gaussian densities around sampled configurations (e.g. in RMSD space)
- Field variables or coordination numbers
- Your favorite order parameter
- All of them

Simple Ansatz:

$$\psi_i(x) = b_{i,1} x_1 + \cdots + b_{i,n} x_n$$

Noé and Nüske, MMS 11, 635-655 (2013)
Nüske et al, JCTC 10, 1739-1752 (2014)
1FME peptide - Simulation data from DESRES, Lindorff-Larsen et al, Science 2011
from pyemma import coordinates, plots

# define input
feat = coordinates.featurizer("protein.pdb")
feat.add_residue_mindist()
data = coordinates.source("protein.xtc", feat)

# TICA
tica = coordinates.tica(data, lag=100, dim=2)
Y = tica.get_output()[0]

# visualize
plots.plot_free_energy(Y[:, 0], Y[:, 1])
Estimation
Estimation problems

1) Nonreversible estimate

\[
\hat{C}(0) = \frac{1}{N}X^TX
\]

\[
\hat{C}(\tau) = \frac{1}{N}X^TY
\]

2) Standard reversible estimate

\[
\hat{C}_{\text{sym}}(0) \approx \frac{1}{2N}(X^TX + Y^TY)
\]

\[
\hat{C}_{\text{sym}}(\tau) \approx \frac{1}{2N}(X^TY + Y^TX)
\]

This leads to a strong bias for out-of equilibrium data!
Equilibrium estimation from out-of-equilibrium data

1. Extend data
\[ X^\dagger = [X \ 1] \]

2. Compute (nonreversible) Koopman matrix \( K \) from data

3. Compute reweighting vector \( u \) using the eigenvalue problem
\[ \hat{u} = \hat{u}K \]
and normalization \( u = \frac{1}{\hat{u}_{n+1}} \hat{u} \)

4. Equilibrium expectations:
\[ \mathbb{E}_\pi [f(x_t)] = \frac{1}{N} \sum_{t=1}^{T-\tau} u \chi(x_t)^T f(x_t) \]
\[ \mathbb{E}_\pi [\chi_i(x_t) \chi_j(x_{t+k\tau})] = \frac{1}{N} \sum_{t=1}^{T-\tau} u \chi(x_t) \chi(x_t)^T \chi^T(x_t) K^k \]

Wu et al, \textbf{arXiv} 1610.06773 (2016)
Equilibrium estimation from out-of-equilibrium data

Wu et al, arXiv 1610.06773 (2016)
Equilibrium estimation from out-of-equilibrium data

Wu et al, \textbf{arXiv} 1610.06773 (2016)
Acknowledgements

Collaborations
Cecilia Clementi (Rice University)
Christof Schütte (FU Berlin)
Eric Vanden-Eijnden (Courant NY)
Thomas Weikl (MPI Potsdam)
Edina Rosta (King’s College London)
Bettina Keller (FU Berlin)
Vijay Pande (Stanford)
Volker Haucke (FMP Berlin)
Stephan Sigrist (FU Berlin)
Oliver Daumke (MDC)
John Chodera (MSKCC NY)
Gianni de Fabritiis (Barcelona)

Funding