聚类分析与生物分子动力学

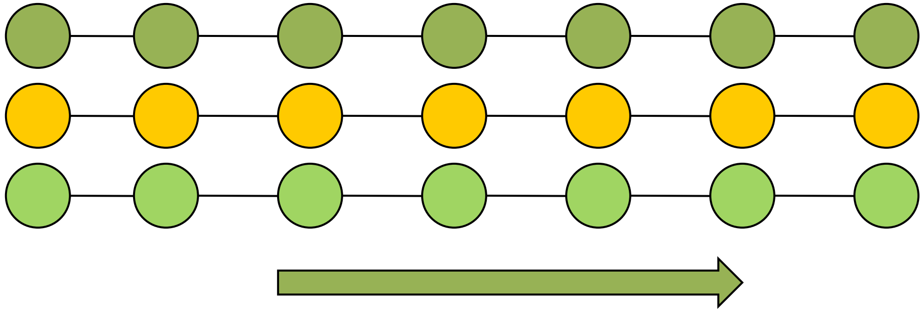
五

姚 远

2011.3.22

Time Series Analysis in Molecular Dynamics

Dataset: Multiple trajectories with a lot of conformations.



Trajectory 1

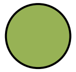
Trajectory 2

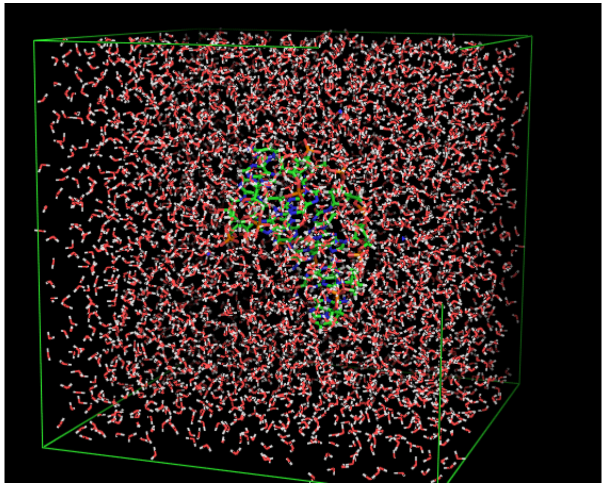
Trajectory 3

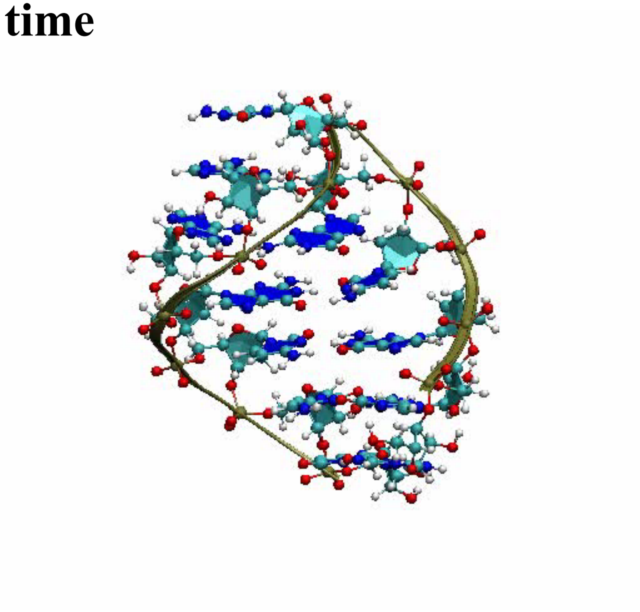
Trajectory

time

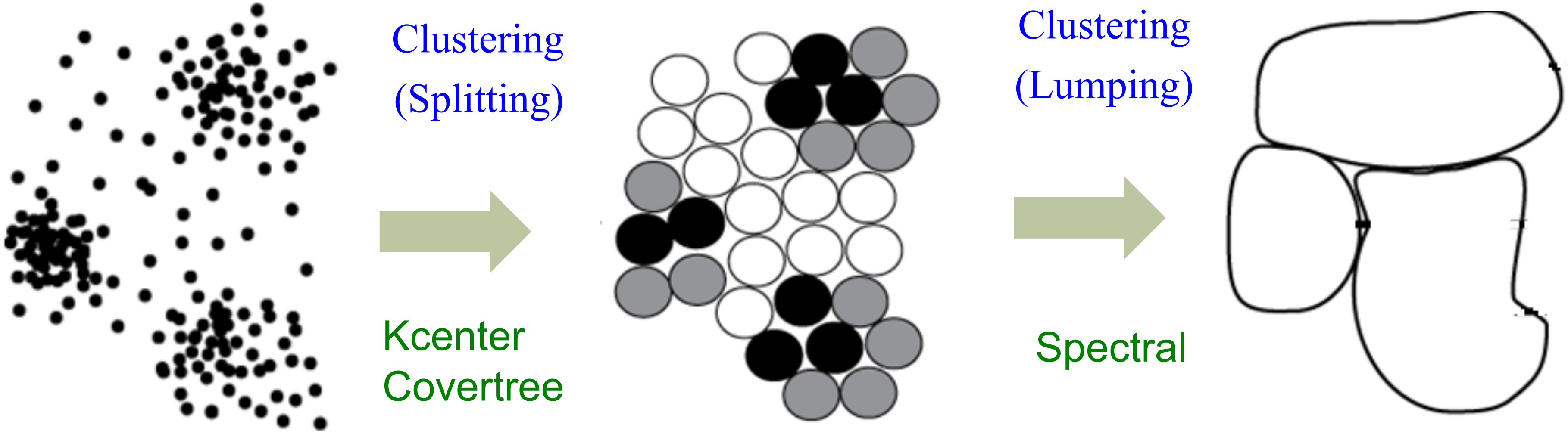
Conformation







RoadMap



Geometric Clustering (Splitting)

Kcenter
Covertree

Spectral Clustering (Lumping)

Spectral

Conformations

Microstates

Macrostates

Bayesian Inference for MSM

General
Reversible

$$T(\tau) = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{15} \\ p_{21} & p_{22} & & \\ \vdots & & \ddots & \\ p_{51} & & & p_{55} \end{bmatrix}$$

Chodera. et. al. *J. Chem. Phys.* 2007
Noé. et.al. *J. Chem. Phys.* 2007
Deuflhard and Weber, *ZIB-report,* 2003
Weber, *ZIB-report,* 2004
Bowman, Huang, and Pande. *Methods* 2009.
Barcalado, et al. *J. Chem. Phys.* 2009

Recall

- Build up Microstates:
  - k-center
  - cover-tree (CHEN, Ying: last lecture)
- Build up Macrostates:
  - Lumpability of Markov chains
  - Spectral clustering for lumping
  - Nystrom method for denoising
- Bayesian Inference for MSM (this lecture)

Bayesian Inference for (reversible) Markov Models

Reference

- Diaconis and Rolle, *Bayesian Analysis for Reversible Markov Chains*, Ann. Stat. 34(3): 1270-1292, 2006.
- Sergio Bacallado, John D. Chodera, and Vijay Pande, *Bayesian comparison of Markov models of molecular dynamics with detailed balance constraint*. J. of Chem. Phys. 131:045106, 2009
- Bacallado, *Bayesian Analysis of Variable-order, Reversible Markov Chains*, submitted to Ann. Stat., 2010

Acknowledgement

- Part of the slides are from Sergio Bacallado in Pande group meeting at Stanford University, July 6, 2009

Outline

- 1 Bayesian Inference in a nutshell
- 2 Hidden Markov Models in MD
- 3 Polya Urn and Dirichlet Prior for Multinomial Models
- 4 Edge Reinforcement Random Walk and a Conjugate Prior for reversible Markov models
- 5 Projects

Bayesian Inference in a nutshell

- Let
  - M: model
  - D: data
  - P(D): probability of data D
  - P(M): prior probability of model M
  - P(D|M): likelihood to generate D by M
  - P(M|D): posterior probability of M given Data (evidence)
- Bayesian Theorem:

$$P(M \mid D) = \frac{P(D \mid M)P(M)}{P(D)}$$

Bayesian Model Comparison

- Given 2 models
  - M_1, M_2: model
  - D: data
- Which model better explains the data?
- Bayesian factor:

$$\frac{P(M_1 | D)}{P(M_2 | D)} = \frac{P(D | M_1)P(M_1)}{P(D | M_2)P(M_2)}$$

Bayesian Model Comparison

- If the model defines a distribution on {M} parametrized by a random variable θ, we can write this

$$\frac{P(M_1 \mid D)}{P(M_2 \mid D)} = \frac{\int d\theta P(D \mid \theta, M_1) P(\theta \mid M_1)}{\int d\theta \underbrace{P(D \mid \theta, M_2)}_{likelihood} \underbrace{P(\theta \mid M_2)}_{prior}}$$

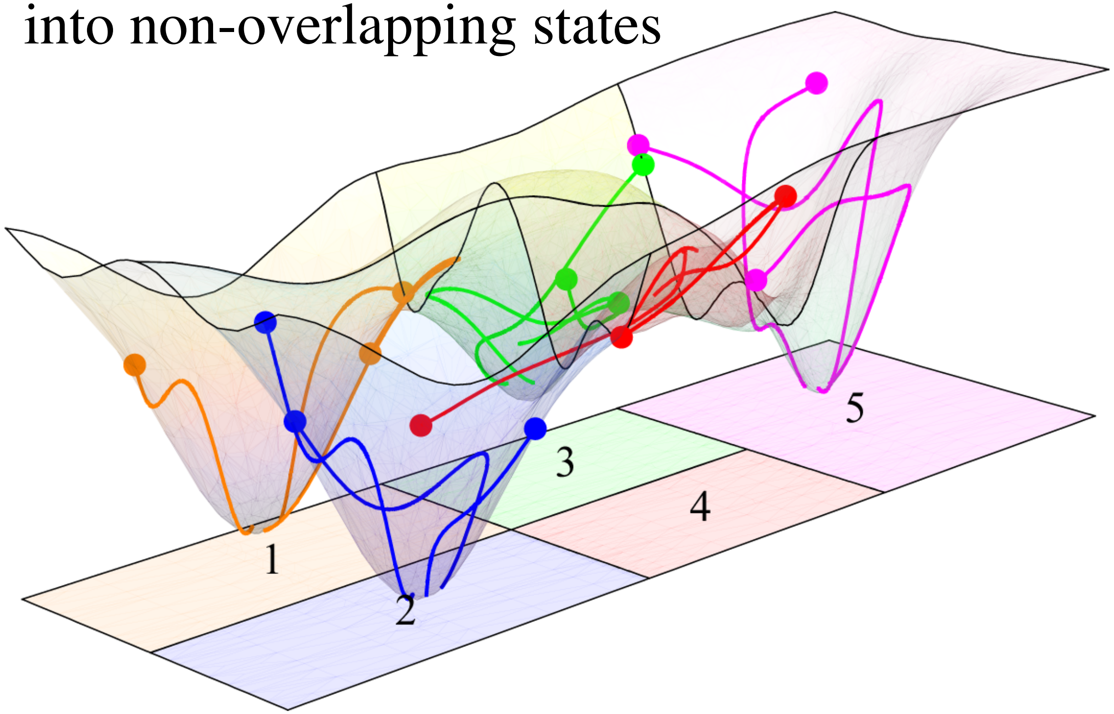
All we need is generative models and *good* priors for M!

Examples

- D: observe a time series
  - 0 1 1 1 0 0 0 1 0 1 0 0 0 1 1 0 1
  - A T A C G G C T A G C A T C G
- From what model are the sequences generated?
  - M_1: Bernoulli model
  - M_2: a general Markov model $P(x_t{:}x_{t-1})$
  - M_3: a reversible Markov model $P(x_t{:}x_{t-1})$, $\pi_i P_{ij} = \pi_j P_{ji}$
  - M_4: high order Markov models…

Markov State Models in Molecular Dynamics

The configuration space is decomposed into non-overlapping states



Define transition probabilities between states

$$T(\tau) = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{15} \\ p_{21} & p_{22} & & \\ \vdots & & \ddots & \\ p_{51} & & & p_{55} \end{bmatrix}$$

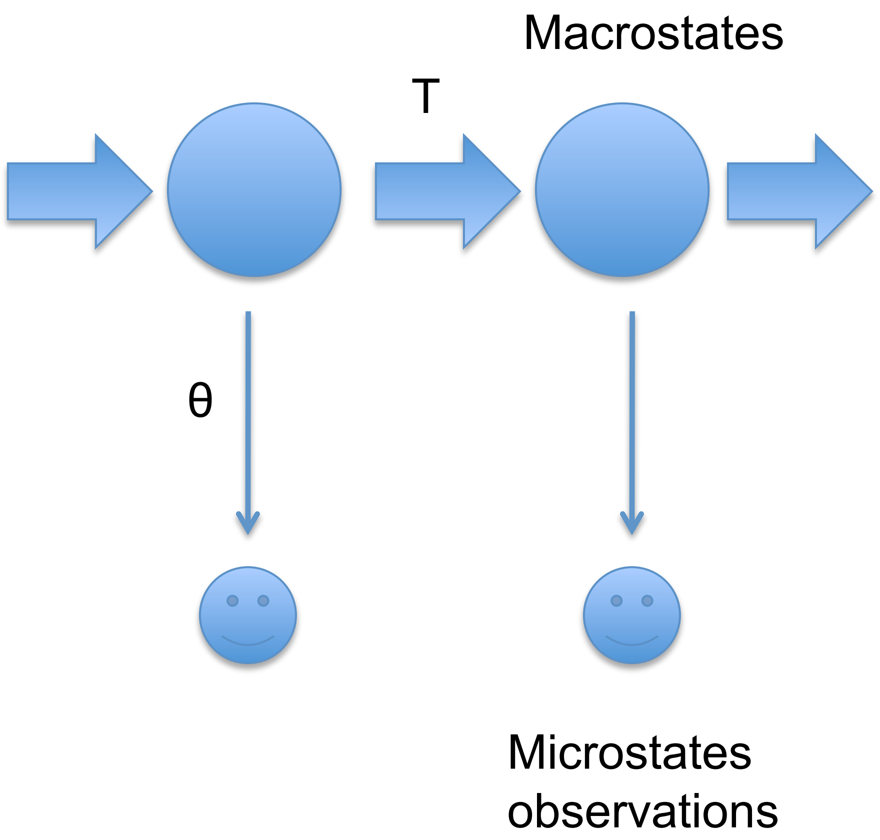
We can extract long time dynamics from MSMs built from short simulations

$$P(n\tau) = [T(\tau)]^n P(0)$$

The time is coarse-grained in τ

Hidden Markov Models in Molecular Dynamics

- Define
  - X: configuration space
  - Y: microstate space
  - Z: macrostate space
  - *S*: X -> Y splitting map of X into microstate space
  - *L*: Y -> Z lumpping map of Y into macrostate space
  - T: ZxZ -> R, Markov transition matrix on Z, T_{ij} is transition probability from macrostate i to j
  - Θ: Z -> Y, a multinomial choice of microstate within macrostate



1. Given a microstate sequence, which lumping *L* is the best?
2. Are high order Markov Chains better in explaining the data?

Models Parameters

- For an observed microstate sequence:
  - $y_1, y_2, y_3, \ldots, y_t, \ldots$
- The generative model parameters are
  - $T(z_{t-1}=L(y_{t-1}), z_t=L(y_t))$: markov transition matrix
  - Θ: multinomial choice of y_t in $z_t=L(y_t)$
- They are independent:

$$P(M \mid y_n) = \int dT d\theta \underbrace{P(y_n \mid T,\theta,M)}_{likelihood} \underbrace{P(T,\theta \mid M)}_{prior}$$

$$= \underbrace{\int dT P(z_n \mid T,M) P(T \mid M)}_{MacrostateMarkovTransition} \bullet \underbrace{\int d\theta P(y_n \mid z_n,\theta,M) P(\theta \mid M)}_{MicrostateSelection}$$

What kind of Priors?

- A good prior should be
  - – Reflect our prior knowledge about the model
    - Every transition is possible
    - Reversibility (detailed balance, i.e. , $\pi_i T_{ij} = \pi_j T_{ji}$)
  - – Easy to compute posterior

$$P(\theta \mid D, M) \propto P(D \mid \theta, M) P(\theta, M)$$

  - – Analytic form for P(D|M), e.g.

$$P(M \mid y_n) = \underbrace{\int dT P(z_n \mid T, M) P(T \mid M)}_{Macrostate Markov Transition} \bullet \underbrace{\int d\theta P(y_n \mid z_n, \theta, M) P(\theta \mid M)}_{Microstate Selection}$$

Dirichlet Prior for Multinomial Models

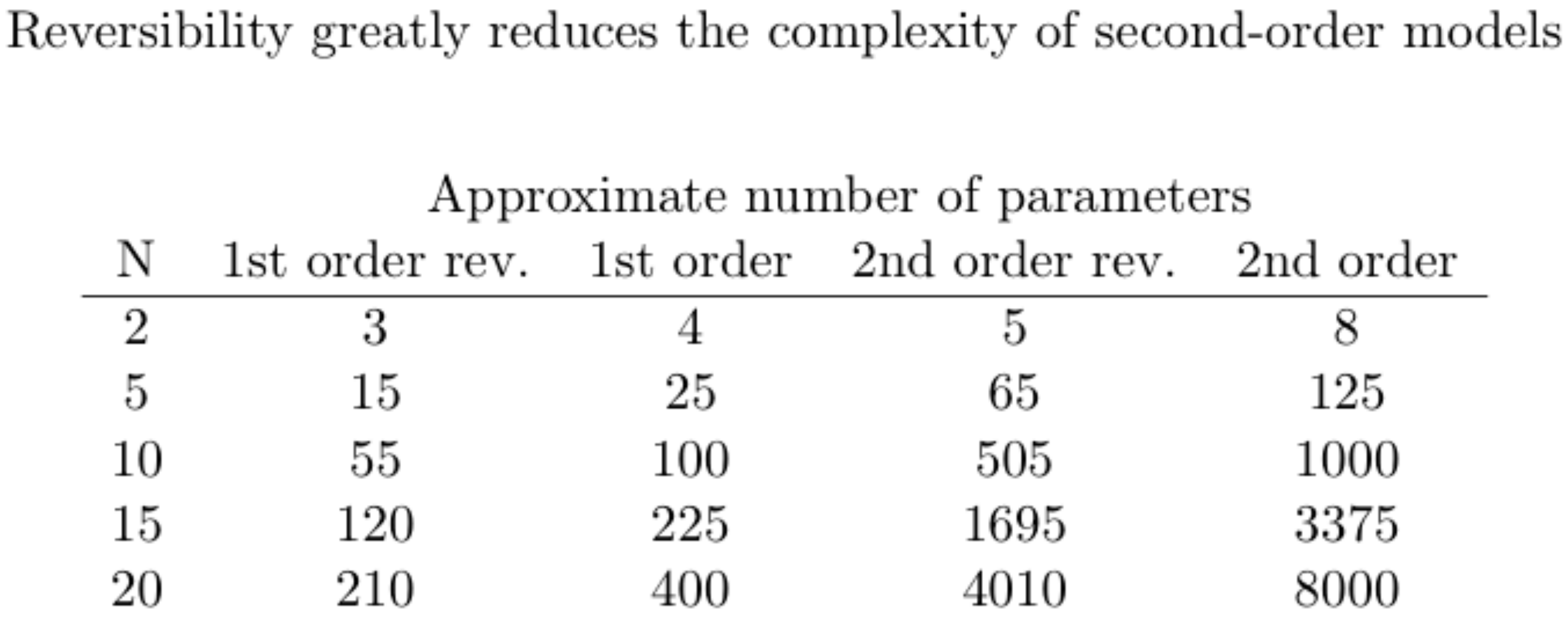
- For the multinomial models θ, Dirichlet Prior is the perfect prior,
  - Conjugacy: The posterior P(θ|D,M) is also a Dirichlet distribution, straightforward to obtain from prior and data.
  - Analytical integration: The integral for P(D|M) is a simple function of the data and the prior, e.g.

$$P(y_n \mid z_n, M) = \int d\theta P(y_n \mid z_n, \theta, M) P(\theta \mid M)$$

Markov Models

- For simplicity, below we only discuss
  - first order Markov models
  - General Markov models (M_1)
  - Reversible Markov models (M_2)

Why Reversible?



Reversibility greatly reduces the complexity of second-order models

	Approximate number of parameters			
N	1st order rev.	1st order	2nd order rev.	2nd order
2	3	4	5	8
5	15	25	65	125
10	55	100	505	1000
15	120	225	1695	3375
20	210	400	4010	8000

General Markov Models

- General Markov Models
  - For each row, T(i,*) is a multinomial distribution
  - Independent Dirichlet Prior for each row
  - However, this may violates the reversible Markov chain assumption
  - So we need a prior restricted on reversible Markov chains

P. Diaconis and S. Rolles. Bayesian Analysis of Reversible Markov Chains. Annals of Statistics, no. 3, 1270-1292, 2006.

Reversible Markov Models

- Apply the conjugate prior for reversible Markov chains to compute Bayesian factors between different models {M1, M2, ...}.
- Numerical trick when the data are parallel trajectories.
- Pros:
  - Can compare models with different numbers of macrostates
  - Easier to interpret than heuristics.
  - Can find best model among bad models (adaptive sampling).
- Cons:
  - Not an absolute validation of Markovity.
  - Can find the best model in {M1, M2, ...}, but all models might be poor.

S. Bacallado, J. Chodera, and V. Pande. Bayesian comparison of Markov models of molecular dynamics with detailed balance constraint. J. of Chem. Phys. 131:045106, 2009.

de Finetti's Theorem

Bruno de Finetti (1906-1985).

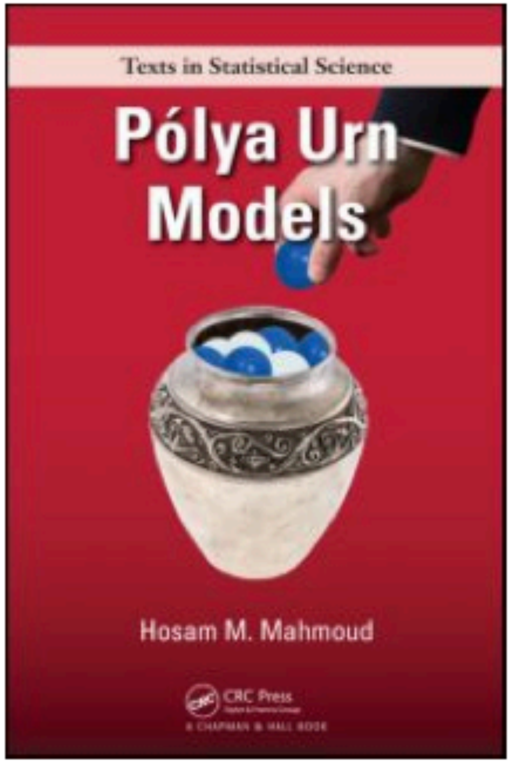
Let $x = \{x_1, x_2, ...\}$ be an infinite sequence of random variables valued 0 or 1. Let two sequences be equivalen, $x \sim y$, if one is a finite permutation of the other. Then the probability of a sequence is invariant for all equivalence classes:

$$x \sim y \implies P(x) = P(y) \quad : \quad \textit{Exchangeability}$$

if and only if $P(x_1, ..., x_n)$ is a mixture of i.i.d. coin tosses, or

$$P(x_1, ..., x_n) = \int dq \, \underbrace{(q^{|\{i:x_i=1\}|}(1-q)^{|\{i:x_i=0\}|})}_{\text{Bernoulli likelihood}} \psi(q)$$

Polya Urn Scheme


Texts in Statistical Science

Pólya Urn Models

Hosam M. Mahmoud

CRC Press

A CHAPMAN & HALL BOOK

- We have an urn with n blue balls, and m white balls. We pick a ball from the urn at random, and then return it along with another ball of the same color. What is the distribution of this process?
- It is easy to show that the process is *exchangeable*. By de Finetti's theorem, it is a mixture of Bernoulli distributions. If there are more than 2 colors, it is a mixture of multinomial distributions. But what is the mixing density?

The mixing density is Dirichelet Distribution

- The probability P_{Polya} is a mixture of multinomials with a Dirichlet density on the parameters:

$$P_{\text{Polya}}(x_1, ..., x_n) = \int d\theta \quad \underbrace{f(x_1, ..., x_n|\theta)}_{\text{multinomial likelihood}} \quad \text{Dir}(\theta).$$

- If we are doing Bayesian inference with a multinomial likelihood, and we choose a Dirichlet prior, then the evidence is of the form of P_{Polya}.
- Conclusion: The Dirichlet distribution is convenient for computing the evidence *because* it is the mixing measure of an exchangeable process.

de Finetti Theorem for Markov Chain (Diaconis-Freeman 1980)

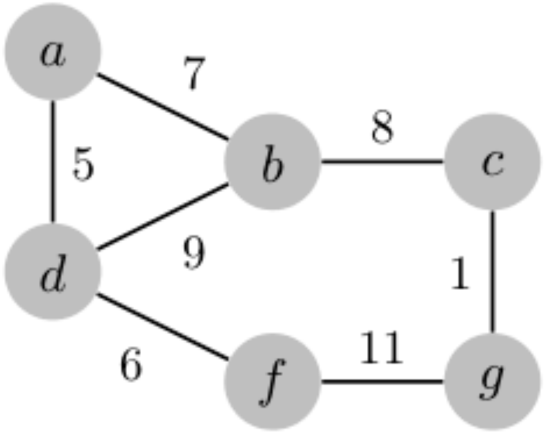
Let x be a recurrent process. Two sequences are equivalent, $x \sim y$, if they have the same transition count matrix and initial state. Then partial exchangeability:

$$x \sim y \implies P(x) = P(y)$$

holds if and only if the process is a mixture of Markov chains:

$$P(x_1, ..., x_n) = \int dT \underbrace{f(x_1, ..., x_n|T)}_{\text{Markov likelihood}} \phi(T).$$

Random Walk on Undirected Graphs: Reversible Mark Chains

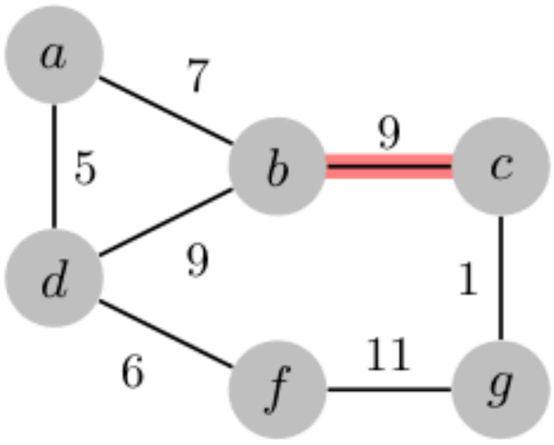


P. Diaconis and S. Rolles. Bayesian Analysis of Reversible Markov Chains. *Annals of Statistics*, no. 3, 1270-1292, 2006.

- A reversible Markov chain is equivalent to a random walk on an undirected, edge-weighted graph.
- The normalized edge-weights, x, can be related to the transition matrix by:

$$T_{i,j} = \frac{x_{i,j}}{x_i}.$$

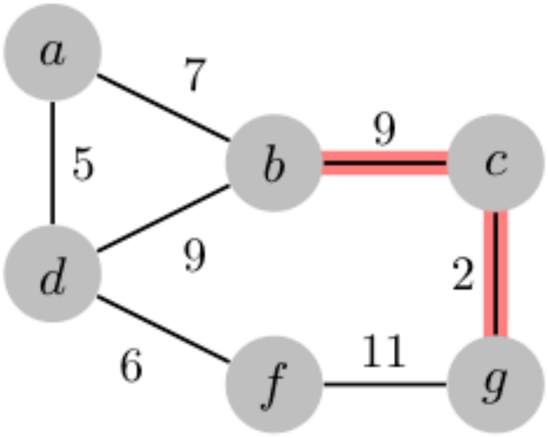
Random Walk on Undirected Graphs: Reversible Mark Chains



P. Diaconis and S. Rolles. Bayesian Analysis of Reversible Markov Chains. *Annals of Statistics*, no. 3, 1270-1292, 2006.

- An *edge-reinforced random walk* (ERW) proceeds as follows: We take random walk on a weighted graph, but every time we cross an edge, we increase its weight by 1.
- Generalization of a Polya-urn scheme.

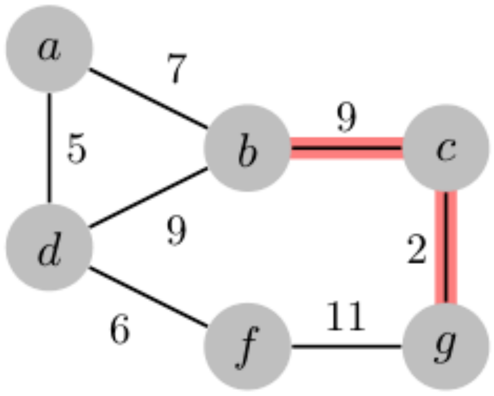
Random Walk on Undirected Graphs: Reversible Mark Chains



P. Diaconis and S. Rolles. Bayesian Analysis of Reversible Markov Chains. *Annals of Statistics*, no. 3, 1270-1292, 2006.

- An *edge-reinforced random walk* (ERW) proceeds as follows: We take random walk on a weighted graph, but every time we cross an edge, we increase its weight by 1.
- Generalization of a Polya-urn scheme.

de Finetti's Theorem



P. Diaconis and S. Rolles. Bayesian Analysis of Reversible Markov Chains. *Annals of Statistics*, no. 3, 1270-1292, 2006.

- It is easy to show that the probability P_{ERW} of an edge-reinforced random walk is partially exchangeable.
- By de Finetti's theorem, it is a mixture of Markov chains:

$$P_{\mathrm{ERW}}(x_1, ..., x_n) = \int dT\, \underbrace{f(x_1, ..., x_n|T)}_{\text{Markov likelihood}}\, \phi(T).$$

- If we do inference on a reversible Markov chain, with a prior $\phi(T)$ on the transition matrix, then the evidence is of the form P_{ERW}.

A Conjugate Prior for Reversible Markov Chains

- ϕ(T) is a conjugate prior over reversible Markov Chains
- ϕ(T) is characterized by a initial vertex v_0, and initial weight assignment on G=(V,E), a:E->[0,∞)

$$\phi_{v_0,a}(x) := Z_{v_0,a}^{-1} \frac{\prod_{e \in E \setminus E_{\mathrm{loop}}} x_e^{a_e - 1/2} \prod_{e \in E_{\mathrm{loop}}} x_e^{(a_e/2)-1}}{x_{v_0}^{a_{v_0}/2} \prod_{v \in V \setminus \{v_0\}} x_v^{(a_v+1)/2}} \sqrt{\det(A(x))}$$

  - det(A(x)) and normalization factor $Z_{v0,a}$ both have closed forms

A Conjugate Prior for Reversible Markov Chains

- Posterior of $\phi_{v0,a}(x)$ after observing $(X_0=v_0,X_1,\ldots,X_n=v_n)$ will be

$$\phi_{vn,a+k(e)}(x)$$

  - where k(e) counts the number of visits on e

- We could extend this to higher order Markov chains

Python codes: msmBayes

- Sergio Bacallado, Stanford
- Not publically available
- For class purpose, at our subdirectory
  - ./msmBayes/
- I translate a few matlab codes
  - ./matlab/MultinomialLogLik.m
  - ./matlab/MarkovChainLogLik.m
  - ./matlab/RevChainLogLik.m

Software: MSMBuilder

- Free Python toolbox
- Greg Bowman, et. al.
- https://simtk.org/home/msmbuilder
- Implements
  - Kcenter, Kmeans splitting
  - Spectral lumping
  - Bayesian Factor

MSMBuilder

Project Overview

Description: Understanding a molecule's conformational dynamics requires mapping out the dominant metastable, or long lived, states that it occupies and then determining the rates for transitioning between these states. Markov State Models (MSMs) provide a natural framework for accomplishing this objective. To facilitate more widespread use of MSMs we have developed the MSMBuilder package. Besides building MSMs, the code includes tools for verifying that the resulting model is Markovian as well as analyzing and visualizing the model. For example, it is possible to determine the populations of each state with error-bars and to extract representative conformations for each state so that the systems dynamics may be visualized. The code is written in object oriented C++ and Python so new developments may be incorporated rapidly.

The video embedded below provides a more thorough introduction to the motivation for and use of MSMBuilder.

Downloads & Source Code

msmbuilder

This project uses Simtk's **Subversion code repository**, but has restricted access to project members.

News

MSMBuilder 1.0 released

Greg Bowman awarded the 2010 Kuhn Paradigm Shift Award

MSMBuilder for analyzing dynamics released

Project Lead

Greg Bowman
Contact

Kyle Beauchamp
Contact

Driving Biological Problems

This project is part of

Protein Folding

RNA Folding

Projects

- Design a time series
- Molecular Dynamics
  - ./data/alanine_dipeptide_phi-psi.mat
  - ./data/confs_3D.txt
  - ./data/alanine_dipeptide_traj_coords.mat
  - ./data/T5000.mat
- Molecular Motor
  - ./data/motor_stepsize.txt

Molecular Motor

- Reference
  - Sun et al. Single-molecule stepping and structural dynamics of myosin X. *Nature Structural & Molecular Biology*, 17(4): 485-491, 2010. (./Myosin X nsmb.pdf)
  - ./data/motor_stepsize.txt records the step sizes in Figure 3a, distributed approximately by 2-mode Gaussian mixture. Is the data from 2-state Markov Model, reversible Markov model, or just purely random?