Lecture 23

MCMC
Some Applications

- HW7 OPTIONAL
- Project !!!
- BE PRESENT @ 6/6 LECTURE (mandatory)

Lecture Notes VIII – Markov Chain Monte Carlo

Marina Meila
mmp@stat.washington.edu

Department of Statistics
University of Washington

April 2025

Applications

1. Bayesian estimation

...

Notation ✓

Gibbs sampling ✓

The detailed balance ✓

Metropolis-Hastings sampling ✓

tractable conjugate prior

$N(\mu, \sigma^2)$, $\mu = \theta$

closed form

$\mu \sim N(0, \sigma_0^2)$

σ_0 large

Parameter $\theta \in \Theta$

- $f^0(\theta)$ = prior distribution of θ
- Data $\mathcal{D}$
- Model class

$L(\theta) = \Pr[\mathcal{D} / \theta]$ likelihood

wanted $f(\theta / \mathcal{D})$ = posterior of θ given $\mathcal{D}$

Bayes' formula

$$f(\theta / \mathcal{D}) = \frac{f^0(\theta) L(\theta)}{\underbrace{\int_\Theta f^0(\theta) L(\theta) \, d\theta}_{\text{evidence}}}$$

→ Z intractable

MH for Bayesian inference

$\mathcal{D}$

$$f(\theta|\mathcal{D}) \propto f^0(\theta)\, L(\theta)$$

MH ~~proposal~~ $s(\theta, \theta')$

O(n) calculations

acceptance prob $$a(\theta^t, \theta') = \min\left[1, \frac{f^0(\theta')\, L(\theta')}{f^0(\theta^t)\, L(\theta^t)} \cdot \frac{s(\theta^t, \theta')}{s(\theta', \theta^t)}\right]$$

Bayesian Inference with hidden Variables $\longrightarrow$ 2.

2. Lecture II – Clustering – Part II: Non-parametric clustering

Marina Meilă
mmp@stat.washington.edu

Department of Statistics
University of Washington

CSE 547/STAT 548
Winter 2022

1. Paradigms for clustering

2. Methods based on non-parametric density estimation

3. Model-based: Dirichlet process mixture models

What is clustering? Problem and Notation

- **Informal definition** **Clustering** = Finding groups in data
- **Notation**

$$\begin{aligned}
\mathcal{D} &= \{x_1, x_2, \ldots x_n\} \text{ a } \textbf{data set} \\
n &= \text{number of } \textbf{data points} \\
K &= \text{number of } \textbf{clusters } (K << n) \\
\Delta &= \{C_1, C_2, \ldots, C_K\} \text{ a partition of } \mathcal{D} \text{ into disjoint subsets} \\
k(i) &= \text{the } \textbf{label} \text{ of point } i \\
\mathcal{L}(\Delta) &= \text{cost (loss) of } \Delta \text{ (to be minimized)}
\end{aligned}$$

- **Second informal definition** **Clustering** = given n **data points**, separate them into K **clusters**
- Hard vs. soft clusterings
  - **Hard** clustering Δ: an item belongs to only 1 cluster
  - **Soft** clustering $\gamma = \{\gamma_{ki}\}_{k=1:K}^{i=1:n}$

    γ_{ki} = the **degree of membership** of point i to cluster k

$$\sum_k \gamma_{ki} = 1 \quad \text{for all } i$$

(usually associated with a probabilistic model)

Clustering Paradigms

Depend on type of data, type of clustering, type of cost (probabilistic or not), and constraints (about K, shape of clusters)

- Data = vectors $\{x_i\}$ in $\mathbb{R}^d$

Parametric (K known)	Cost based [hard] Model based [soft]
Non-parametric (K determined by algorithm)	Dirichlet process mixtures [soft] Information bottleneck [soft] Modes of distribution [hard] Gaussian blurring mean shift**?** [hard] Level sets of distribution [hard]

- Data = similarities between pairs of points $[S_{ij}]_{i,j=1:n}$, $S_{ij} = S_{ji} \geq 0$ **Similarity based clustering**

Graph partitioning	spectral clustering [hard, K fixed, cost based] typical cuts [hard non-parametric, cost based]
Affinity propagation	[hard/soft non-parametric]

The Dirichlet distribution

- $Z \in \{1:r\}$ a discrete random variable, let $\theta_j = P_z(j)$, $j = 1, \ldots r$.
- **Multinomial distribution** Probability of i.i.d. sample of size N from P_z

$$P(z^{1,\ldots n}) \;=\; \prod_{j=1}^{r} \theta_j^{n_j}$$

where $n_j = \#$the value j is observed, $j = 1, \ldots r$

- $n_{1:r}$ are the **sufficient statistics** of the data.
- The **Dirichlet distribution** is defined over domain of $\theta_{1,\ldots r}$, with real parameters $N'_{1,\ldots r} > 0$ by

$$D(\theta_{1,\ldots r}; n'_{1,\ldots r}) \;=\; \frac{\Gamma(\sum_j n'_j)}{\prod_j \Gamma(n'_j)} \prod_j \theta_j^{n'_j - 1}$$

where $\Gamma(p) \;=\; \int_0^\infty t^{p-1} e^{-t} dt$.

Dirichlet process mixtures $\in$ Non-parametric clustering

Model-based $\in$ Clustering

- Model-based
- generalization of mixture models to
  - "infinite" K = K is a random variable that can grow with n
  - Bayesian framework

NOT meaningful for the model understanding

- denote θ_k = parameters for component f_k
- assume $f_k(x) \equiv f(x, \theta_k) \in \{f(x, \theta)\}$
- assume prior distributions for parameters $g_0(\theta)$
- prior with hyperparameter $\alpha > 0$ on the number of clusters
- very flexible model

"clusters"

model $f(x) = \sum_{k=1}^{K} \pi_k \underbrace{N(\mu_k, \sigma^2 I)}$

$\sum_{k=1}^{K} \pi_k = 1$

point $x_i \in C_1$ cluster k

$k(i) = 1$ $\quad k(i) \in 1:K$

data $X_{1:n}$

A sampling model for the data

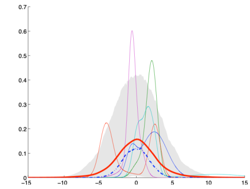
- **Example: Gaussian mixtures**, $d = 1$, $\sigma_k = \sigma$ fixed
- $\theta = \mu$
- prior for μ is $Normal(0, \sigma_0^2 I_d)$
- Sampling process = Model we are fitting
  - for $i = 1:n$ sample x_i, $k(i)$ as follows

    denote $\{1:K\}$ the clusters after step $i-1$
    define n_k the size of cluster k after step $i-1$

    1.
$$k(i) = \begin{cases} k & \text{w.p } \frac{n_k}{i-1+\alpha}, \quad k = 1:K \\ K+1 & \text{w.p } \frac{\alpha}{i-1+\alpha} \end{cases} \tag{1}$$

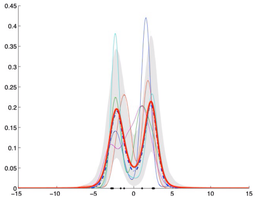
    2. if $k(i) = K+1$ sample $\mu_i \equiv \mu_{K+1}$ from $Normal(0, \sigma_0^2)$
    3. sample x_i from $Normal(\mu_{k(i)}, \sigma^2)$
- can be shown that the distribution of $x_{1:n}$ is **interchangeable** (does not depend on data permutation)

Prior:



Red: mean density. Blue: median density. Grey: 5-95 quantile. Others: draws.

Posterior:



Red: mean density. Blue: median density. Grey: 5-95 quantile. Black: data. Others: draws.

The hyperparameters

- σ_0 controls spread of centers
  - should be large
- α controls number of cluster centers
  - α large $\Rightarrow$ many clusters
- cluster sizes non-uniform (larger clusters attract more new points)
- many single point clusters possible

General Dirichlet mixture model

- cluster densities $\{f(x, \theta)\}$
- parameters θ sampled from prior $g_0(\theta, \beta)$
- cluster membership $k(i)$ sampled as in (1)
- x_i sampled from $f(x, \theta_{k(i)})$
- **Model Hyperparameters** α, β

Clustering with Dirichlet mixtures

The clustering problem

- $\alpha, g_0, \beta, \{f\}$ given
- $\mathcal{D}$ given
- wanted $\theta_{1:n}$ (not all distinct!)
- note:
  - $\theta_{1:n}$ determines a hard clustering Δ
  - the posterior of $\theta_{1:n}$ given the data determines a soft clustering via $P(x_i \mid k) \propto \int f(x_i|\theta_k)g_k(\theta_k)d\theta_k$

Estimating $\theta_{1:n}$ cannot be solved in closed form
Usually solved by MCMC (Markov Chain Monte Carlo) sampling

Clustering with Dirichlet mixtures via MCMC

Gibbs

μ_k prior $N(\mu, \sigma^2 I)$

α = smoothness propensity for new clusters

MCMC estimation for Dirichlet mixture

Input $\alpha, g_0, \beta, \{f\}, \mathcal{D}$

State cluster assignments $k(i)$, $i = 1:n$, parameters θ_k for all distinct k

Iterate
1. for $i = 1:n$ (reassign data to clusters)
   1. remove i from its cluster (hence $\sum_k n_k = n - 1$)
   2. resample $k(i)$ by

size of C_k

likelihood of $x_i | \theta_k$

$$k(i) = \begin{cases} \text{existing } k & \text{w.p } \propto \frac{n_k}{n-1+\alpha} f(x_i, \theta_k) \\ \text{new cluster} & \text{w.p } \frac{\alpha}{n-1+\alpha} \int f(x_i, \theta) g_0(\theta) d\theta \end{cases} \quad (2)$$

   3. if $k(i)$ is new label, sample a new $\theta_{k(i)} \propto g_0 f(x_i, \theta)$
2. for $k \in \{k(1:n)\}$ (resample cluster parameters)   from posterior
   1. sample θ_k from posterior $g_k(\theta) \propto g_0(\theta, \beta) \prod_{i \in C_k} f(x_i, \theta)$

g_k can be computed in closed form if g_0 is conjugate prior

Output a state with high posterior

n points x_i out $\Rightarrow$ $n-1$ total

$n_k = |C_k|$ after x_i out

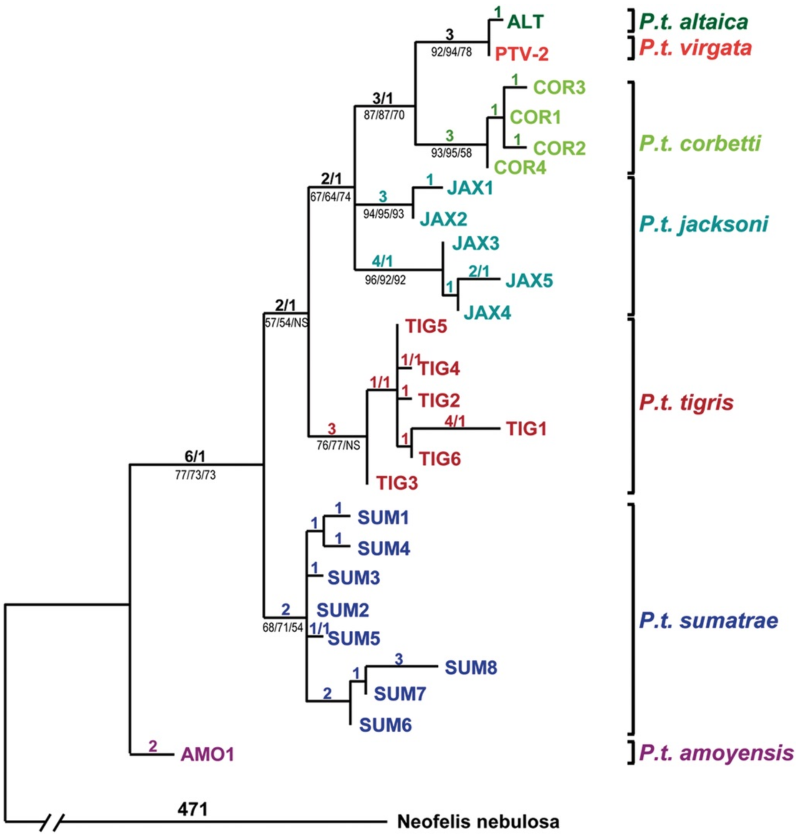
$\sum n_k = n - 1$

No new clusters:

$$\frac{n_k f(x_i | \theta_k)}{n-1}$$

3. Sampling (Bayesian estimation) for Phylogenetic Trees

Pantera tigris (P.t.)



Tiger family Phylogeny