Mar 20

Bayesian Updates

Conjugate Priors

Data Analysis

Static or confirmatory

Method → data $y_1 \cdots y_n$ ↓ output

Adaptive

Method → data ↓ output (Go in a cycle)

↑ (data keep coming)

Statistical Inference

- Frequentist Inference
- Bayesian Inference

Is a kind of statistical inference that the entire conclusion is based on data by focusing on frequency of it

No prior assumption on its distribution

Is a kind of statistical inference that a prior distribution is assumed and gets updated as data arrives ⇒ (posterior)

Bayesian Approach :

Pros: model flexibility (Monte Carlo Markov chain)

accurate estimates of parameters

Cons: Subjectivity (being subjective due to choice of prior)

costs (computational costs for cycling process)

Bayesian theorem:

Ω sample space $[B_1, \dots B_d \Rightarrow \bigcup B_i = \Omega, B_i \cap B_j = \emptyset \; (j \neq i)]$

$$P(B_j|A) = \frac{P(A|B_j) \cdot P(B_j)}{P(A)} = \frac{P(A|B_j) \cdot P(B_j)}{\sum_{\forall A} P(A|B_i) P(B_i)}$$

A: data, B_j: parameter

continuous version

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x) f_X(x)}{f_Y(y)} = \frac{f_{Y|X}(y|x) f_X(x)}{\int f_{Y|X}(y|x) f_X(x) dx}$$

(conditional: $f_{Y|X}(y|x)$; marginal: $f_X(x)$)

Prior distribution + new information $\Rightarrow$ posterior distribution

Start with a parameter set (θ)

Prior knowledge: $\pi(\theta)$: prior distribution on that parameter set

Bayesian Approach :

- $\pi(\theta)$
- $f(y|\theta)$: likelihood function (likelihood)

Posterior distribution

$$\pi(\theta|y) = \frac{f_{Y|\theta}(y|\theta)\,\pi(\theta)}{\int_\theta f_{Y|\theta}(y|\theta)\,\pi(\theta)\,d\theta}$$

($f_{Y|\theta}(y|\theta)$: likelihood; $\pi(\theta)$: prior; $\pi(\theta|y)$: posterior; denominator: difficult to compute, but a constant)

$$\Rightarrow \quad \pi(\theta|y) \propto f_{Y|\theta}(y|\theta)\cdot\pi(\theta)$$

($\propto$: proportional)

posterior $\propto$ prior $\cdot$ likelihood

MCMC (Monte Carlo Markov chain)

example :

- likelihood function

$$Y|\theta \sim \text{Binomial}(n,\theta)$$

(θ: probability of success)

$$P(Y|\theta) = P_r(Y=y) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$$

- prior distribution (subjective)

$$\pi(\theta) = \text{Beta}(\theta;\alpha,\beta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Posterior $\propto$ prior $\times$ likelihood : $\pi(\theta|Y) \propto \pi(\theta)\cdot P(Y|\theta)$

$$\propto \frac{1}{B(\alpha,\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \cdot \binom{n}{y} \theta^{y}(1-\theta)^{n-y}$$

$$\propto \frac{1}{B(\alpha,\beta)} \binom{n}{y} \theta^{\alpha+y-1}(1-\theta)^{\beta+n-y-1}$$

$$\propto \text{Beta}(\theta; \alpha+y; \beta+n-y)$$

Conjugate Priors

convenient choice for prior are conjugate priors

The posteriors belong to the same family as the prior with different parameters

$$\boxed{\begin{matrix} \alpha_0 \\ \alpha(y) \end{matrix}} \qquad \begin{matrix} \alpha & \beta \\ \downarrow & \downarrow \\ \alpha+y & \beta+n-y \end{matrix}$$

Observations

New data influence posterior through a parameter change

The shape or type of distribution is unchanged

Beta distribution is conjugate for the Binomial Likelihood

$f(y\|\theta)$	$\pi(\theta)$	$\pi(\theta\|y)$
Likelihood	prior	
$N(\theta, \sigma^2)$	$N(\mu, \tau^2)$	$N\left(\frac{\sigma^2\mu + \tau^2 y}{\sigma^2+\tau^2}, \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}\right)$
$N(\mu, \frac{1}{\theta})$ precision $\theta = \frac{1}{\sigma^2}$	$\Gamma(\alpha, \beta)$	$\Gamma\left(\alpha + \frac{1}{2}, \beta + \frac{1}{2}(\mu - y_i)^2\right)$
$\Gamma(\nu, \theta)$	$\Gamma(\alpha, \beta)$	$\Gamma(\alpha+\nu, \beta+y)$
$Bin(n, \theta)$	$Beta(\alpha, \beta)$	$Beta(\alpha+y, \beta+n-y)$
$Poisson(\theta)$	$\Gamma(\alpha, \beta)$	$\Gamma(\alpha+y, \beta+1)$

Problem / Issues in choice of Prior

- For non-conjugate priors
  the posterior distribution is not available in analytical form
- Conjugate priors might not reflect
  uncertainty about θ in a right way

The conjugate prior for normal distribution

Likelihood $p(y_1 \cdots y_n \mid \mu, \sigma^2) = \prod^n p(y_i \mid \mu, \sigma^2)$ i.i.d.

$$= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{1}{2\sigma^2}\cdot\sum_i^n (y_i-\mu)^2}$$

$$\propto \frac{1}{\sigma^n} e^{-\frac{1}{2\sigma^2}\sum_i^n (y_i-\mu)^2}$$

$\bar{y} = \frac{1}{n}\sum_i y_i$

$s^2 = \frac{1}{n}\sum_i (y_i-\bar{y})^2$

exercise $\propto \frac{1}{\sigma^n} e^{-\frac{n}{\sigma^2}(\bar{y}-\mu)^2 - \frac{ns^2}{2\sigma^2}}$

Goal: Find conjugate prior distribution for μ, σ^2 $(\lambda = \frac{1}{\sigma^2})$

Case ①: σ^2 known, μ unknown

Case ②: σ^2 unknown, μ known

Case ③: both unknown (assignment 5 explain detailed)

① σ^2 is known, μ is unknown

$\pi(\mu) \sim N(\mu_0, \sigma_0^2)$

$$\propto \frac{1}{\sigma_0^2} e^{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}}$$

$\pi(\mu \mid y_1, \cdots, y_n)$

multiply two distributions & complete the square in the component

$$e^{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}} \quad e^{-\frac{1}{2\sigma^2}\sum_i (y_i-\mu)^2}$$

$$= \exp\left(-\frac{\sigma^2(\mu^2+\mu_0^2-2\mu\mu_0) + \sigma_0^2\sum_i(y_i^2+\mu^2-2y_i\mu)}{2\sigma^2\sigma_0^2}\right)$$

$$= \exp\left(-\frac{(\sigma^2+n\sigma_0^2)\mu^2 + \cdots - 2(\sigma^2\mu_0+\sigma_0^2\Sigma y_j)\mu}{2\sigma^2\sigma_0^2}\right)$$

$$= \exp\left(-\frac{\mu^2 + \cdots - \frac{2(\sigma^2\mu_0+\sigma_0^2\Sigma y_j)\mu}{(\sigma^2+n\sigma_0^2)}}{\frac{2\sigma^2\sigma_0^2}{\sigma^2+n\sigma_0^2}}\right)$$

$$\propto \exp\left(-\frac{\left(\mu - \frac{\sigma^2\mu_0+\sigma_0^2\Sigma y_j}{\sigma^2+n\sigma_0^2}\right)^2}{\frac{2\sigma_0^2\sigma^2}{\sigma^2+n\sigma_0^2}}\right)$$

$$K^2 = \frac{2\sigma_0^2\sigma^2}{\sigma_0^2+n\sigma_0^2} \qquad \eta =$$

(?) $\propto e^{-(\quad)}$

$$e^{-\frac{(\mu-\eta)^2}{K^2}}$$

(2) μ known, $\lambda = \frac{1}{\sigma^2}$ unknown

conjugate prior for λ is gamma distribution

$$p(\lambda;\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha}\lambda^{\alpha-1}e^{-\lambda/\beta}$$

$$p(\lambda;\alpha,\tfrac{1}{\beta}) = \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\lambda\beta}$$

$$\frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\lambda\beta}\cdot\lambda^{\frac{n}{2}}e^{-\frac{1}{2}\Sigma_i(y_j-\mu)^2}$$

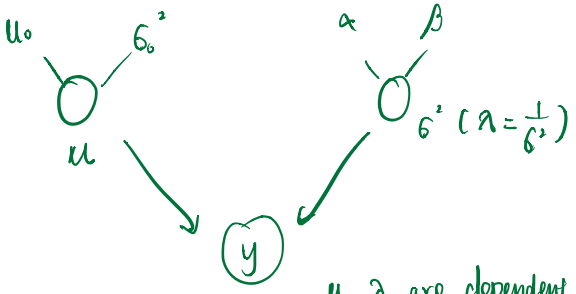
$$\pi(\lambda \mid y_1 \cdots y_n) \propto \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot \lambda^{\alpha-1} \cdot e^{-\lambda\beta} \cdot \lambda^{\frac{n}{2}} \cdot e^{-\frac{1}{2}\lambda\sum (y_i-\mu)^2}$$

$$\propto \lambda^{\alpha-1+\frac{n}{2}} e^{-\lambda\left(\beta+\frac{1}{2}\sum (y_i-\mu)^2\right)}$$

$$\propto Ga\left(\alpha+\frac{n}{2}, \frac{1}{\beta+\frac{1}{2}\sum (y_i-\mu)^2}\right)$$

compare $Ga(\alpha, \frac{1}{\beta})$

③ both μ, $\lambda = \frac{1}{\sigma^2}$ unknown



μ, λ are dependent conditioned on y

Selecting a Prior

choosing an appropriate prior is a key part of Bayesian modeling

Conjugate Priors:

tractability

$$\pi(\theta \mid y) \propto \pi(\theta) \cdot p(y \mid \theta)$$

(not consider constant)

↓ Sampling problem

Assume $p(x) = \frac{1}{x_p} \tilde{p}(x)$, hard to compute x_p (drop)

$$\tilde{p}(\theta) = p(y|\theta) \cdot \pi(\theta)$$

$$x_p = p(y) = \int_\theta \pi(\theta)\, p(y|\theta)\, d\theta$$

Q: How to generate samples from $p(x)$ just having $\tilde{p}(x)$

Recall Acceptance Rejection

Mar 22

$$p(x) = \frac{1}{x_p}\,\tilde{p}(x)$$

$\frac{1}{x_p}$: hard to compute

$\tilde{p}(x)$: easy to compute

$\tilde{p}(x) \geq 0$

Acceptance - Rejection

- choose $q(x)$ easy to sample from

$$\tilde{p}(x) \leq c\,q(x) \quad \forall x$$

- generate $X \sim q(\cdot)$ & $U \sim u(0,1)$
- Accept X if $U \leq \frac{\tilde{p}(x)}{cq(x)}$

otherwise start over

Q: How many iterations on average to get a sample

- probability of success on each iteration is $\frac{x_p}{c}$
- n iterations

$n-1$ failures until success

$$E(n) = \frac{1}{p} = \frac{c}{x_p}$$

$q(x)$ closer to p

High-dimensional case is Markov Chain Monte Carlo

Simulating Markov Chains

Many SPs used for modelling are Markovian & it makes it pretty easy to simulate from them

Def of a Markov Chain

A stochastic process $\{X_n : n \geq 0\}$ is a Markov Chain if for all n and for all states $i_0, i_1, \dots, i_{n-1}, i, j \in S$

$$P(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0)$$

(X_{n+1}: very next state; X_n: current)

$$= P(X_{n+1} = j \mid X_n = i)$$

$$= P_{ij}$$

P_{ij}: one-step transition probability

$P = (P_{ij})_{i,j \in S}$: one-step transition matrix

For each i $\sum_{j \in S} P_{ij} = 1$ (probability distribution)

Assuming transition probabilities Do Not depend on time

$$P_{ij} = P(X_{n+1} = j \mid X_n = i) = P(X_1 = j \mid X_0 = i)$$

↑ Time-homogeneous MC (time stationary)

Future is independent of the past information given the current state

Trading: local stationarity & pattern recognition

$$P(X_{n+1} = i \mid X_n = j_1, X_{n-1} = j_2, \dots, X_{n-l} = j_{l+1})$$

not the entire history

Voting & Learning

Portfolio Construction
+
Asset Allocation

Example: Simulation of a 2-state MC

$$\{X_n : n \geq 0\} \quad S = \{0, 1\}$$

$$P = \begin{pmatrix} 0.2 & 0.8 \\ 0.6 & 0.4 \end{pmatrix} \qquad P(X_1 = 0 \mid X_0 = 0) = 0.2$$

$X_0 = 1$
↑ time 0 ↑ state 1 (without loss of generality)

Generate $U \sim U(0,1)$

set $X = 0$ if $U \le 0.6$

set $X = 1$ if $U > 0.6$

Set $X_1 = X$

$X_1 = 1$

$X_2 = 0$

Generate $U \sim U(0,1)$

set $X = 0$ if $U \le 0.2$

otherwise set $X = 1$

Set $X_3 = X$

General Algorithm

For a MC with $P = (P_{ij})$, $i, j \in S$

Let Y_i be a generic r.v distributed as the i^{th} row of P

$P(Y_i = j) = P_{ij}$, $j \in S$

$S = \{0, 1, \dots, d\}$

$U \sim U(0,1)$

if $U \le P_{i,0}$:

set $Y_i = 0$

else if $P_{i,0} < U \le P_{i,0} + P_{i,1}$: Set $Y_i = 1$

$\vdots$

else if $\sum_{k=0}^{i-1} P_{i,k} < U \le \sum_{k=0}^{i} P_{i,k}$: Set $Y_i = j$

end if

Examples of Markov Chains

1. Random Walk $(\Delta_n : n \geq 1)$

$$X_n = \Delta_1 + \Delta_2 + \cdots + \Delta_n \quad , \quad X_0 = 0$$

$$X_{n+1} = \Delta_1 + \cdots + \Delta_n + \Delta_{n+1} = X_n + \Delta_{n+1}$$

Markov Chains as recursions

$f(x, v)$ real-valued function

$\{V_n : n \geq 0\}$ iid sequence of random variable

The recursion

$$X_{n+1} = f(X_n, V_n) \quad n \geq 0$$

defines a Markov Chain

X_0 independent of $\{V_n : n \geq 0\}$

Recursion makes it easy to simulate from

choose X_0

sequentially generate a V_n

Set $X_{n+1} = f(X_n, V_n)$

In discrete case :

The transition probabilities from a recursively defined MC

are determined by $P_{ij} = P(f(i, V) = j)$

Proposition : all MC can be represented as

$$X_{n+1} = f(X_n, V_n) \quad n \geq 0$$

for some function $f(x,v)$ & iid sequence $\{v_n\}$

Mar 27

Markov Chains as recursions

$f(x, v)$

$\{v_n : n \geq 0\}$ i.i.d. segence of r.v

$X_{n+1} = f(X_n, v_n) \quad n \geq 0$

defines a Markov chain

Proposition :

Every MC can be represented as

$X_{n+1} = f(X_n, v_n) \quad n \geq 0$ for some function $f(x, v)$ & i.i.d. sequence of $\{v_n\}$

(exam of deriving f)

In discrete case :

Let F_i be cdf of the i^{th} row of transition probability matrix

$F_i^{-1}(y) = \inf\{x : F(x) \geq y\}$

Define

$f(i, u) = F_i^{-1}(u) \quad i \in Z, \ u \in (0, 1)$

we have our desired f

$i^{th} \rightarrow \begin{bmatrix} 0.25 & 0.2 & \cdots & 0.15 & 0.35 \end{bmatrix}_{5\times 5}$

$f_i = 0.25 \quad 0.2 \quad 0.05 \quad 0.15 \quad 0.35$

$F_i = 0.25 \quad 0.45 \quad 0.5 \quad 0.65 \quad 1$

$F_i^{-1}(y) = \inf\{x : F(x) \geq y\}$

$u = 0.7 \qquad F_i^{-1}(0.7) = 4$

$U \sim u(0,1)$

$f(i,u) = F_i^{-1}(u) = F_i^{-1}(0.7) = 4$

Examples

1. Random Walk

$X_{n+1} = X_n + \Delta_{n+1}$ i.i.d. increments $\begin{cases} +1 \\ -1 \end{cases}$

let $\nu_n = \Delta_{n+1}$, $n \geq 0$

$f(x, \nu) = x + \nu$ is the desired function for the recursion

Thus $P_{ij} = P(i + \Delta = j) = P(\Delta = j - i)$

Let F^{-1} be the inverse of

$$F(x) = P(\Delta \leq x)$$

this would allow us to use $U \sim u(0,1)$ to obtain recursion

$X_{n+1} = X_n + F^{-1}(U_n)$ if $U_n \leq 0.5 : F^{-1}(U_n) = -1$

$f(x, U) = x + F^{-1}(u)$ else if $U_n > 0.5 : F^{-1}(U_n) = +1$

2. Binomial Lattice model

$S_0 \quad S_1 = S_0 \times Y_1 \quad S_2 = S_1 Y_2$

$S_n = S_0 Y_1 Y_2 \cdots Y_n$

Y_i i.i.d. distributed as $P(Y=u) = p$

$P(Y=d) = q = 1-p$

$d < 1+r < u$

↑ interest rate

In recursive form

$S_{n+1} = S_n Y_{n+1}$

which you would let $V_n = Y_{n+1}$

leads to $S_{n+1} = S_n \cdot V_n$

$\boxed{f(x,v) = xv}$

$S = \{ S_0 u^k d^m : k \geq 0 \text{ & } m \geq 0 \}$

$Y = u \mathbb{1}_{\{U \leq p\}} + d \mathbb{1}_{\{U > p\}}$

$U \sim u(0,1)$

$\Rightarrow f(x,v) = x \cdot (u \mathbb{1}_{\{U \leq p\}} + d \mathbb{1}_{\{U > p\}})$

Markov Chains in continuous time

Def

A Stochastic process $\{X_t : t \geq 0\}$

is called CTMC if for all $t \geq 0$, $s \geq 0$, $i \in S$, $j \in S$

$$P(X(s+t) = j \mid X(s) = i, \{X(u) : 0 \leq u < s\}) = P(X(s+t) = j \mid X(s) = i)$$

$$= P_{ij}(t) \quad \Rightarrow \text{(s is not important : time-homogeneous)}$$

$P_{ij}(t)$: t time units from now, the chain will be at state j, given it is at state i now

In continuous time, being at $i \in S$

the chain spends some time before moving to state $j \in S$

holding time H_i

waiting time W_i

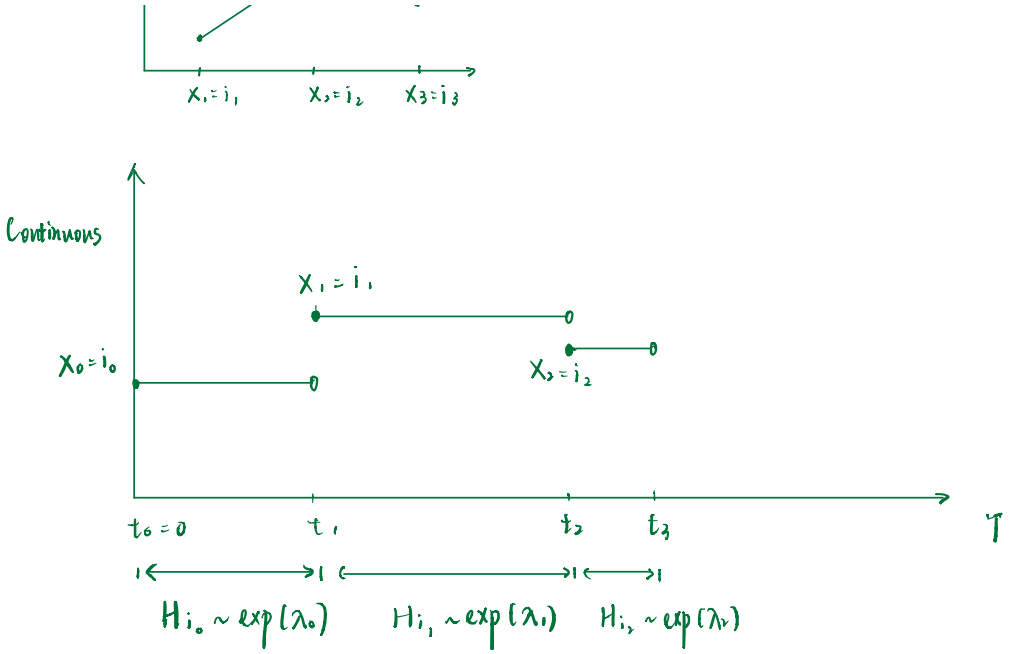
Assume H_i has an exponential distribution at rate λ_i (↓ depend on state i)

$\lambda_0, \lambda_1, \cdots, \lambda_n$

when holding time ends the chain makes a transition

into another state G $i \to j$ according to P_{ij}

Discrete ↑



(keep track of time and state)

$t = t + t_i$ $\quad \{t_n\}$

$X_i \to X_j$ $\quad \{X_n\}$

Def

* A markov chain is ergodic if it is possible to go from every state to every state
  - not necessarily in one shot
  - must be in finite time

* Ergodic MCs are also irreducible (difference) ?

  Regardless of the present state, we can reach any other state in finite time

$$\forall i,j \in S \quad \exists m < \infty : \quad P(X_{n+m} = j \mid X_n = i) > 0$$

A markov chain is called regular

if some power of the transition matrix has only positive matrix

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

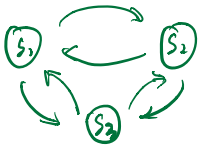
ergodic not regular

Aperiodic MCs

period = 2

S_1 S_2 S_1 S_2 S_1

pattern



aperiodic

S_1 S_2 S_3 S_1

S_1 S_3 S_1

aperiodic

Def

A stationary distribution of a MC is a distribution π on Ω

s.t

(continuous) $\pi(y) = \sum_{x \in \Omega} P(y \mid x)\, \pi(x)$

(discrete) $\pi_j = \sum_{i \in S} \pi_i P_{ij} \quad \forall j \in S$

$\pi = (\pi_1, \pi_2, \dots, \pi_n) \quad \sum_i \pi_i = 1$

Theorem: A finite ergodic MC has a unique stationary distribution

$\pi = \pi P \quad \sum_i \pi_i = 1$

Reversible Markov Chains

Def: A MC is <u>reversible</u> if there exists a probability measure π s.t

(continuous) $P(y|x)\pi(x) = P(x|y)\pi(y)$

(discrete) $\pi_i P_{ij} = \pi_j P_{ji} \quad \forall i,j \in S$

which means

$$\sum_x P(y|x)\pi(x) = \sum_y P(x|y)\pi(y)$$

$$\sum_i \pi_i P_{ij} = \sum_j \pi_j P_{ij} = \pi_j \underbrace{\sum_j P_{ij}}_{1} = \pi_j \quad (?)$$

It means that probability of seeing a transition from j to i is the same as seeing a transition from i to j

$$P = \begin{pmatrix} 0.5 & 0.2 & 0.3 \\ 0.15 & 0.7 & 0.15 \\ 0.2 & 0.25 & 0.55 \end{pmatrix}$$

$$= \begin{pmatrix} 0.274 & 0.4319 & 0.3137 \end{pmatrix}$$

$\pi P = \pi$

$\pi_1 + \pi_2 + \pi_3 = 1$

$\pi_i P_{ij} = \pi_j P_{ji}$

$\pi_1 P_{12} \neq \pi_2 P_{21} \Rightarrow$ not reversible

Example: Random Walk with states $0, 1, \cdots, M$ & $(P_{ij})_{M+1 \times M+1}$

$$P_{i,i+1} = \alpha_i = 1 - P_{i,i-1}, \quad i = 1, \dots, M-1$$

$$P_{0,1} = \alpha_0 = 1 - P_{0,0}$$

$$P_{M,M} = \alpha_M = 1 - P_{M,M-1}$$

Claim: This MC is (time) Reversible

$$P = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ M-1 \\ M \end{array} \begin{array}{c} \begin{array}{ccccccc} 0 & 1 & 2 & \cdots & & M-1 & M \end{array} \\ \begin{pmatrix} 1-\alpha_0 & \alpha_0 & & & & & \\ 1-\alpha_1 & 0 & \alpha_1 & & & & \\ & 1-\alpha_2 & 0 & \alpha_2 & & & \\ & & & & & & \\ & & & & & & \\ & & & & & 1-\alpha_M & \alpha_M \end{pmatrix} \end{array}$$

$$\pi = \pi P$$

π ← limiting probabilities

$$\pi = (\pi_0, \pi_1, \dots, \pi_M)$$

$$\pi_0 = (1-\alpha_0)\pi_0 + (1-\alpha_1)\pi_1$$

$$\left.\begin{aligned} \pi_0 \alpha_0 &= \pi_1 (1-\alpha_1) \\ \pi_1 \alpha_1 &= \pi_2 (1-\alpha_2) \\ &\vdots \\ \pi_i \alpha_i &= \pi_{i+1} (1-\alpha_{i+1}) \end{aligned}\right\} \; i = 0, \dots, M-1$$

Solve in terms of π_0

$$\pi_1 = \frac{\alpha_0}{1-\alpha_0}\pi_0$$

$$\pi_2 = \frac{\alpha_1}{1-\alpha_2}\pi_1 = \frac{\alpha_1\alpha_0}{(1-\alpha_2)(1-\alpha_1)}\pi_0$$

$$\vdots$$

$$\pi_i = \frac{\alpha_{i-1}\alpha_{i-2}\cdots\alpha_0}{(1-\alpha_i)(1-\alpha_{i-1})\cdots(1-\alpha_1)}\pi_0 \qquad i = 1, \cdots, M$$

subject to $\sum_{j=0}^{M} \pi_j = 1$

$$\pi_0 = \frac{1}{1 + \sum_{j=1}^{M} \frac{\prod_{i=0}^{j-1} \alpha_i}{\prod_{i=1}^{j} (1-\alpha_i)}}$$

Verify that $\pi_i P_{i,i+1} = \pi_{i+1} P_{i+1,i}$ ✓

If $|j-i| > 1$, $P_{ij} = P_{ji} = 0$

$$\pi_i P_{ij} = \pi_j P_{ji} = 0$$

Markov Chain Monte Carlo

- Early inception in the late 1940s
- Metropolis algorithm published by Metropolis et al 1953 (Los Alamos)
- It was generalized by Hastings in 1970 and later Peskun 1973

- Geman & Geman 1984 (Gibbs)

- Besag & Clifford (Gibbs Sampling)

The Metropolis-Hastings algorithm

$$p(x) = \frac{\tilde{p}(x)}{x_p}$$

Initialization

- set $t = 0$
- $x_t = x$

Iteration

- Generate $Y \sim g(\cdot|x)$
- Calculate acceptance probability

$$q(y|x) = \min\left(1, \frac{\tilde{p}(y)}{\tilde{p}(x)} \times \frac{g(x|y)}{g(y|x)}\right)$$

- Accept or reject
  - $U \sim u(0,1)$
  - if $U \leq q(y|x)$ accept

$$x_{t+1} = y$$

  - else reject

$$x_{t+1} = x_t = x$$

Claim: The resulting MC is reversible with stationary distribution $p(x) = \frac{\hat{p}(x)}{x_p}$

* Important (observation)

x_p is NOT required for the M-H algorithm

What we do is:

We run the M-H algorithm

until we reach (Q) stationarity & use the

simulated points as our samples

Q: How do we know where to stop

Check convergence

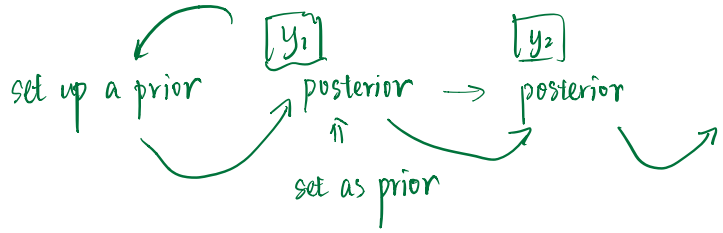
Q: Choice of $g(\cdot|\cdot)$

Q: Are the samples that we have obtained through M-H algorithm independent

Mar 29

Likelihood

{ prior
  posterior }



expression $y_1, y_2, \cdots$

prior → posterior

Conjugate priors

Case 1

Case 2

Case 3 update 2 priors 2 posteriors

⇒ derivation in detailed

Claim: Resulting MC is reversible

Proof:

$$q(y|x) = \min\left(1, \frac{\tilde{P}(y)}{\tilde{P}(x)} \cdot \frac{g(x|y)}{g(y|x)}\right)$$

$$q(y|x)\, g(y|x)\, \tilde{p}(x) = q(x|y)\, g(x|y)\, \tilde{P}(y)$$

$p(y|x)$ $\quad$ $p(x|y)$

$$q(y|x)\ g(y|x)\ \tilde{p}(x) =$$

$$\min\left(1, \frac{\tilde{p}(y)}{\tilde{p}(x)} \cdot \frac{g(x|y)}{g(y|x)}\right) \cdot g(y|x)\ \tilde{p}(x)$$

$$= \min\left[g(y|x)\ \tilde{p}(x),\ \tilde{p}(y)\ g(x|y) \right]$$

$$= \tilde{p}(y)\ g(x|y) \underbrace{\min\left(\frac{\tilde{p}(x)}{\tilde{p}(y)} \cdot \frac{q(y|x)}{q(x|y)}, 1 \right)}_{q(x|y)}$$

$$= q(x|y)\ g(x|y)\ \tilde{p}(y)$$

Example Bivariate Normal distribution

$$\mu = (\mu_1, \mu_2)$$

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix} \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix}$$

$$\Sigma_{d \times d} = \begin{pmatrix} \sigma_1 & & \sigma_d \end{pmatrix}^T \Lambda \begin{pmatrix} \sigma_1 & & \sigma_d \end{pmatrix}$$

$$\Lambda \uparrow \begin{pmatrix} 1 & & 1 \end{pmatrix}$$

recursively $\sigma_{j,t+1} = f(\sigma_j, t)$

pdf $\pi(\theta) \propto \frac{1}{|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\theta-\mu)^T \Sigma^{-1}(\theta-\mu)}$

$\theta = (\theta_1, \theta_2)$

Marginal Distributions

$\theta_1 \sim N(\mu, \sigma_{11}^2)$ $\quad$ $\theta_2 \sim N(\mu_2, \sigma_{22}^2)$

$\sim N(\mu, S_{11})$ $\quad$ $\sim N(\mu_2, S_{22})$

Example (code)

$\mu = (2, 3)$

$\Lambda = \begin{pmatrix} 1 & -0.6 \\ -0.6 & 1 \end{pmatrix}$ $\quad$ $\Sigma = \sigma^T \Lambda \sigma$

$\sigma = \begin{pmatrix} 1.2 & 0 \\ 0 & 1.1 \end{pmatrix}$

candidate density $\quad g(\theta|\theta')$

$g(\theta'|\theta)$

A good candidate density should be also bivariate normal (but independent)

$\theta' = \theta + \sigma Z$ $\quad Z \sim N(0, I)$ $\quad I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

$\theta_1' = \theta_1 + \eta_1 Z$

$\theta_2' = \theta_2 + \eta_2 Z$

$g(\theta|\theta') = g(\theta'|\theta)$

$$g(\theta'|\theta) = \min\left(1, \frac{\hat{p}(\theta')}{\hat{p}(\theta)} \cdot \frac{g(\theta|\theta')}{g(\theta'|\theta)}\right)$$

$$= \min\left(1, \frac{e^{-\frac{1}{2}(\theta'-\mu)^T \Sigma^{-1}(\theta'-\mu)}}{e^{-\frac{1}{2}(\theta-\mu)^T \Sigma^{-1}(\theta-\mu)}}\right)$$

Psendo-code

From an initial state $\theta^{(0)}$

(? difference q & g)

For $i = 1, 2, \ldots$

$\theta' = \theta^{(i-1)} + \eta Z$

compute $q(\theta' | \theta^{(i-1)})$

$U \sim u(0,1)$

if $U < q(\theta' | \theta^{(i-1)})$

$\theta^{(i)} = \theta'$

else

$\theta^{(i)} = \theta^{(i-1)}$

end

Ideal properties of MCMC

- $(X_1^{(0)}, X_2^{(0)})$

To be chosen to be in the region of high probability

unde $p(x_1, x_2)$ (In general, hard to Do)

- We run the chain for n iterations

  we dump the first B samples

  burn-in

- If we run long enough ($n \uparrow$), the choices of

  B doesn't matter

- Performance of MCMC

  $\frac{1}{N} \sum_{i}^{N} h(X_1^{(i)}, X_2^{(i)})$

  how quickly the samples averages converge

  mixing rate

- An algorithm with good performance is called to

  have good mixing

Gibbs Sampling

It is an MCMC sampler which was introduced by German Geman

- Suppose $p(x_1, x_2)$ is a pdf that is difficult to sample from it directly

- Assumption, we can easily sample from the conditional distribution

$p(x_1 | x_2)$ & $p(x_2 | x_1)$

The Gibbs sampling would be as follows:

1. set x_1 & x_2 to some initial values
2. Sample $x_1 | x_2$ then sample $x_2 | x_1$

   then $x_1 | x_2$ & so on

Gibbs Sampler (WLOG)

1. set $(x_1^{(t)}, x_2^{(t)})$
2. Sample $x_1^{(t+1)} \sim p(x_1 | x_2^{(t)})$

   current state: $(x_1^{(t+1)}, x_2^{(t)})$

   Sample $x_2^{(t+1)} \sim p(x_2 | x_1^{(t+1)})$

   current state: $(x_1^{(t+1)}, x_2^{(t+1)})$
3. Sample $x_1^{(t+2)} \sim p(x_1 | x_2^{(t+1)})$

   ⋮

n times

This scheme produces a sequence of pairs of r.v

$(x_1^{(t+1)}, x_2^{(t+1)})$, $(x_1^{(t+2)}, x_2^{(t+2)})$ ...

Markov Chain & dependence

The pairs satisfy the property of being a Markov chain (Exercise)

The conditional distribution of $(X_1^{(t+\tau)}, X_2^{(t+\tau)})$ given all previous ones only depends on $(X_1^{(t+\tau-1)}, X_2^{(t+\tau-1)})$ (obvious)

The pairs ARE NOT actually i.i.d. samples

Why $\Rightarrow$ (Exercise)

Example: Binomial - Beta

One observation $X \sim Bin(n, \theta)$ (trials) $\binom{n}{k}\theta^k(1-\theta)^{n-k}$

$$\pi(\theta) = Beta(\alpha, \beta)$$

$$\frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Bayes theorem:

$$\pi(\theta|x) \propto \theta^x(1-\theta)^{n-x}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

(get rid of the constants)

(exercise: verify)

Joint density for (X, θ)

$$\binom{n}{x}\theta^x(1-\theta)^{n-x} \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Marginal distribution of X (exercise)

✓ $$f(x) = \binom{n}{x} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \times \frac{\Gamma(\alpha+x)\Gamma(\beta+n-x)}{\Gamma(\alpha+\beta+n)}$$

Gibbs Sampling

- initial value $(X^{(0)}, \theta^{(0)})$
- For $i = 1, 2, \dots, m$
  - $X^{(i)} \sim f(X \mid \theta^{(i-1)}) = Bin(n, \theta^{(i-1)})$
  - $\theta^{(i)} \sim \pi(\theta \mid X^{(i)})$
    $= Beta(\alpha + X^{(i)}, \beta + n - X^{(i)})$
- The procedure produces samples of (X, θ)
- Approximate $f(x)$ with value of $X^{(i)}$

☆ $$f(x) = P(X = x) \approx \frac{\#\text{ of } X^{(i)} = x}{m}$$

  - sample $\theta^{(i)} \sim \pi(\theta \mid X^{(i-1)})$
  - sample $X^{(i)} \sim f(x \mid \theta^{(i)})$

(Code: need freq (count) to build distribution)

We are interested in $f(x)$, but impossible to find it analytically (close-form)

Example Bivariate Normal distribution

$u\ (u_1, u_2)$

$\Sigma = \delta^T \Lambda \delta$

pdf $f(x \mid u, \Sigma) \propto \frac{1}{|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-u)^T \Sigma^{-1} (x-u)}$

- Marginal distribution

$X_1 \sim N(u_1, \sigma_1^2)$

$X_2 \sim N(u_2, \sigma_2^2)$

For Gibbs samplings, need to find

$f(x_1 \mid x_2)$

$f(x_2 \mid x_1)$

$f(x_1 \mid x_2) = N\left(u_1 + \frac{\sigma_{12}}{\sigma_2^2}(x_2 - u_2),\ \sigma_1^2 - \left(\frac{\sigma_{12}}{\sigma_2}\right)^2\right)$

$f(x_2 \mid x_1) = \cdots$

$X_1 = u_1 + \sigma_1 Z_1$

$X_2 = u_2 + \sigma_2 (\rho Z_1 + \sqrt{1-\rho^2}\, Z_2)$ ($Z_1, Z_2 \sim N(0,1)$)

$f(x_1 \mid x_2) = N\left(u_1 + \rho \frac{\sigma_1}{\sigma_2}(x_2 - u_2),\ \sigma_1^2 (1-\rho^2)\right)$

$\mu = (2, 1)$

σ_1 & $\sigma_2 = 1$ To be completed

$\sigma_{12} = 0.7$

Multi-stage Gibbs Sampler

$X^{(0)} \in R^D$

Initialize $X^{(0)} \sim q(x)$

for $i = 1, 2, 3, \cdots$

$$X_1^{(i)} \sim p(x_1 = x_1 \mid x_2 = x_2^{(i-1)}, x_3 = x_3^{(i-1)}, \cdots, x_D = x_d^{(i-1)})$$

$$X_2^{(i)} \sim p(x_2 = x_2 \mid x_1 = x_1^{(i-1)}, x_3 = x_3^{(i-1)}, \cdots)$$

$$\vdots$$

$$X_D^{(i)} \sim p(x_D = x_d \mid x_1 = x_1^{(i-1)}, \cdots, x_{D-1} = x_{d-1}^{(i-1)})$$

end for

- Always using the most recent values of all the other variables
- The conditional distribution of a variable given all the others is referred as the full conditional
- If full conditional simulation is not possible

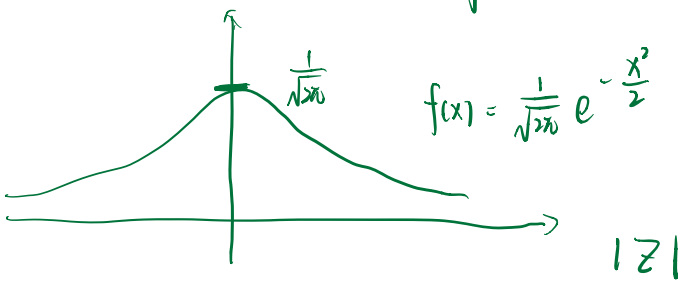
  we would use Metropolis-Hastings

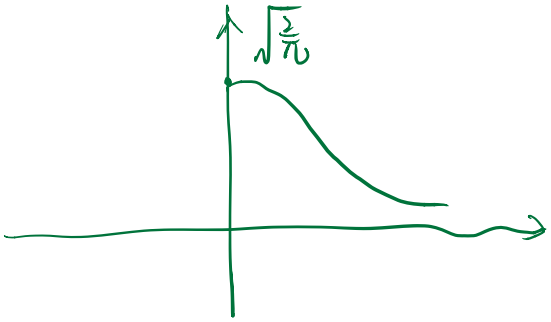
Apr 3

Apr 3.

Sampling from Standard Normal via Acceptance - Rejection

visualization of A-R



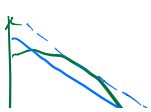


By symmetry, we can obtain Z if we have a sample from $|Z|$ by independently generating a random variable S which determines the side (positive or negative) with probability $p = \frac{1}{2}$

and $Z = S|Z|$

$U \sim U(0,1)$ if $U < 0.5$ $S = +1$

else $S = -1$



$f(|Z|) = \sqrt{\frac{2}{\pi}} e^{-\frac{z^2}{2}}$

target distribution $g(x)$

$$g(x) = e^{-x}$$

$\exists\, c > 1$ $\quad f(x) \le c\, g(x) \quad \forall x$

find optimal c

$$c^* = \max_{x \in (0,\infty)} \left(\frac{f(x)}{g(x)} \right)$$

$$x^* = \operatorname{argmax} \frac{f(x)}{g(x)} = \frac{\sqrt{\frac{2}{\pi}} e^{-\frac{x^2}{2}}}{e^{-x}}$$

$$= \sqrt{\frac{2}{\pi}} e^{-\frac{x^2}{2} + x}$$

$$(-x+1) = 0 \qquad x^* = 1$$

$$c^* = \frac{\sqrt{\frac{2}{\pi}} e^{-\frac{1}{2}}}{e^{-1}} = \sqrt{\frac{2}{\pi}} e^{\frac{1}{2}}$$

Pseudo-code

- generate X from g

$$\begin{cases} U_1 \sim u(0,1) \\ X = -\ln U_1 \end{cases}$$

(matlab: $X =$ exprnd ;)

- generate $U_2 \sim u(0,1)$

- if $U_2 \le \frac{f(x)}{cg(x)} = e^{-\frac{(x-1)^2}{2}}$

set $|Z| = X$

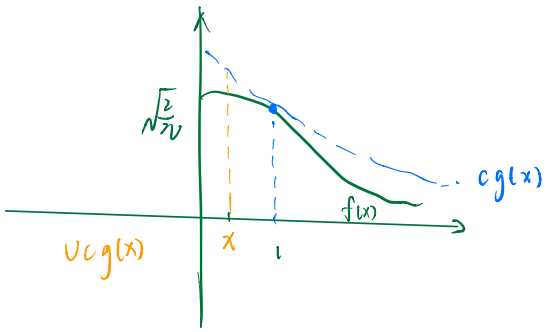
else

reject & start over

- generate $U_3 \sim u(0,1)$

set $Z = |Z|$ if $U_3 < a_5$

otherwise $Z = -|Z|$



Gibbs Sampling for multi-level Models

Example: Binomial - Beta - Poisson

- n unknown

$\pi(n) = Poisson(\lambda)$ ← known

- $X \sim Binomial(n, \theta)$

$\pi(\theta) = Beta(\alpha, \beta)$

- Joint distribution for (X, θ, n) is

$$\binom{n}{x}\theta^x(1-\theta)^{n-x}\,\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,\theta^{\alpha-1}(1-\theta)^{\beta-1}\;e^{-\lambda}\,\frac{\lambda^n}{(n-1)!}$$

$x = 0, 1, \cdots, n$; $0 < \theta < 1$, $n = 1, 2, \cdots$

↑ bounded by n

we are interested in $f(x)$ But impossible to find it analytically

What we do is : (usually get rid of constants)

$$f(\theta, x, n) \propto \binom{n}{x}\theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}\cdot\frac{\lambda^n}{n!}$$

(previously : n is known)

To perform Gibbs sampling,

we must find the conditionals :

$$f(x \mid \theta, n) \propto \binom{n}{x}\theta^x(1-\theta)^{n-x} \propto \text{Binomial}(n, \theta)$$

$$\pi(\theta \mid x, n) \propto \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1} \propto \text{Beta}(\alpha+x, \beta+n-x)$$

$$\pi(n \mid \theta, x) \propto \binom{n}{x}\frac{\lambda^n}{n!}(1-\theta)^{n-x} \propto \frac{(\lambda(1-\theta))^{n-x}}{(n-x)!}$$

$n = x, x+1, \cdots$

if we set $\eta = n - x$

then one can see $\eta \sim \text{Poisson}(\lambda(1-\theta))$

Algorithm

Set initial values $(x^{(0)}, \theta^{(0)}, n^{(0)})$

For $i = 1, 2, 3, \cdots$

- $X^{(i)} \sim \text{Bin}(n^{(i-1)}, \theta^{(i-1)})$
- $\theta^{(i)} \sim \text{Beta}(\alpha + X^{(i)}, \beta + n^{(i-1)} - X^{(i)})$
- $n^{(i)} = X^{(i)} + \eta$

  $\eta = \text{Poisson}(\lambda(1-\theta^{(i)}))$

(always use the most updated ovailable one)

Example : Poisson process with a regime change (switch to new parameter set)

- Multi-level models

  Heirarchical models

  1992 by Carlin, Gelfand & Smith

They enable one to use Gibbs Sampling and priors which learn from data

X_i $\{x_i\}$ $i = 1, 2, \cdots, n$

- $X_i \sim \text{Poisson}(\mu)$ $i = 1, 2, \cdots, k$
- $X_i \sim \text{Poisson}(\lambda)$ $i = k+1, \cdots, n$

Parameters of interest are (μ, λ, k)

Independent priors on (μ, λ, k)

- k discrete uniform on $\{1, 2, \cdots, m\}$ m = sample size

$\mu \sim \text{gamma}(\alpha_1, \beta_1)$

$\lambda \sim \text{gamma}(\alpha_2, \beta_2)$

- $\alpha_1, \beta_1, \alpha_2, \beta_2$ are fixed

The joint posterior distribution for (μ, λ, k) would be of the following form

$$\pi(\mu, \lambda, k \mid x_1, \cdots, x_m) \propto f(x \mid \mu, \lambda, k) \cdot \pi_1(\mu)\pi_2(\lambda)\pi_3(k)$$

The likelihood function

$$f(x \mid \mu, \lambda, k) = \prod_{i=1}^{k} f(x_i \mid \mu, k) \times \prod_{i=k+1}^{m} f(x_i \mid \lambda, k)$$

$$= \prod_{i=1}^{k} \frac{\mu^{x_i} e^{-\mu}}{x_i!} \prod_{i=k+1}^{m} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$$

Therefore

$$\pi(\mu, \lambda, k \mid x) \propto \prod_{i=1}^{k} \frac{\mu^{x_i} e^{-\mu}}{x_i!} \prod_{i=k+1}^{m} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$$

$$\times \left(\mu^{\alpha_1 - 1} e^{-\mu\beta_1}\right)\left(\lambda^{\alpha_2 - 1} e^{-\lambda\beta_2}\right) \cdot \frac{1}{m}$$

Or equivalently

$$\pi(\mu, \lambda, k \mid x) \propto \mu^{\alpha_1 + \sum_{i=1}^{k} x_i - 1} e^{-\mu(k+\beta_1)}$$

(exercise) $$\times \lambda^{\alpha_2 + \sum_{i=k+1}^{m} x_i - 1} e^{-\lambda(m-k+\beta_2)}$$

Full Conditional Distributions:

$$\cdot\ \pi_1(\mu \mid \lambda, x, k) \propto \mu^{\alpha_1 + \sum_{i=1}^{k} x_i - 1} e^{-\mu(k+\beta_1)}$$

$$\propto \text{gamma}\left(\alpha_1 + \sum_{i=1}^{k} x_i,\ k + \beta_1\right)$$

(notice the update of data)

$$\cdot\ \pi_2(\lambda \mid \mu, k, x) \propto \lambda^{\alpha_2 + \sum_{i=k+1}^{m} x_i - 1} e^{-\lambda(m-k+\beta_2)}$$

$$\propto \text{gamma}\left(\alpha_2 + \sum_{i=k+1}^{m} x_i,\ m-k+\beta_2\right)$$

$$\cdot\ \pi_3(k \mid \mu, \lambda, x) \propto \mu^{\sum_{i=1}^{k} x_i} \lambda^{\sum_{i=k+1}^{m} x_i} e^{-k(\mu-\lambda)}$$

$$= \ell(k)$$

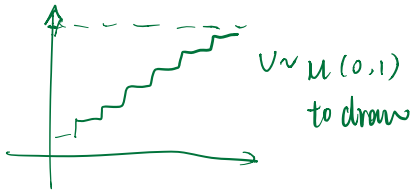
so $$p(k \mid \mu, \lambda, x) = \frac{\ell(k)}{\sum_{k=1}^{m} \ell(k)}$$

build a probability mass function

$k = 1, 2, \cdots, m$

Code:

nIters : num of samplings



Apr 05

Example 1 Normal with Semi-conjugate prior

Consider $X_1, \cdots, X_n \mid \mu, \lambda \sim N(\mu, \frac{1}{\lambda})$

The independently consider

$\mu \sim N(\mu_0, \frac{1}{\lambda_0})$

$\lambda \sim \text{gamma}(\alpha, \beta)$

This is called a semi-conjugate case

i.e prior on μ is conjugate for each fixed value of λ

and prior on λ is conjugate for each fixed value of μ_0

For any fixed λ in this case

$\mu \mid \lambda, x_1, \cdots, x_n \sim N(M_\lambda, \frac{1}{L_\lambda})$

(λ given)

where

$L_\lambda = \lambda_0 + n\lambda$

$M_\lambda = \dfrac{\lambda_0 \mu_0 + \lambda \sum_{i=1}^{n} x_i}{\lambda_0 + n\lambda}$

Update as data coming in

For any fixed μ

$\lambda \mid \mu, x_1, \cdots, x_n \sim \text{gamma}(\alpha_\mu, \beta_\mu)$

where

$\alpha_\mu = \alpha + \frac{n}{2}$

$\beta_\mu = \beta + \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^2$

To implement Gibbs Sampling each iteration would be as follows:

$$\mu \mid \lambda, x_1, \cdots, x_n \sim N\left(M_\lambda, \frac{1}{L_\lambda}\right)$$

$$\lambda \mid \mu, x_1, \cdots, x_n \sim \text{gamma}(\alpha_\mu, \beta_\mu)$$

Example 2: The multi-level Normal model

Data y_{ij} $\quad i = 1, \cdots, n_j$
$\quad j = 1, \cdots, J$

Total # of observations $\quad n = \sum_{j=1}^{J} n_j$

assumed to be independently normally distributed ✱

with each of J groups

with different means θ_j

but common σ^2

which implies

$$y_{ij} \mid \theta_j \sim N(\theta_j, \sigma^2)$$

θ_j follows a normal distribution with unknown μ, τ^2

i.e $\theta_j \sim N(\mu, \tau^2)$

We assume uniform prior

$$P(\mu, \log\sigma, \log\tau) \propto \tau \quad \epsilon \text{ (assumed)}$$

Then posterior is

$$p(\theta, \log\sigma, \log\tau \mid y_{ij}) \propto \tau \prod_{j=1}^{J} N(\theta_j \mid \mu, \tau^2) \cdot \prod_{j=1}^{J}\prod_{i=1}^{n_j} N(y_{ij} \mid \theta_j, \sigma^2)$$

For Gibbs Sample

what we need is all the conditional distribution

(a) Conditional Posterior Distribution of each θ_j

$$\theta_j \mid \underbrace{\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_J}_{\theta_{-j}}$$

$$\theta_j \mid \theta_{-j}, \mu, \sigma, \tau, \underbrace{y}_{\text{all of } y_{ij}} \sim N(\hat{\theta}_j, V_{\theta_j})$$

$$\hat{\theta}_j = \frac{\frac{1}{\tau^2}\mu + \frac{n_j}{\sigma^2}\bar{y}_j}{\frac{1}{\tau^2} + \frac{n_j}{\sigma^2}}$$

$$V_{\theta_j} = \frac{1}{\frac{1}{\tau^2} + \frac{n_j}{\sigma^2}}$$

J of these conditionals

(b) Conditional Posterior Distribution of μ

$$\mu \mid \underbrace{\theta}_{\theta_1, \ldots, \theta_J}, \sigma, \tau, \underbrace{y}_{\text{all } y_{ij} \text{ (observations)}} \sim N(\hat{\mu}, \frac{\tau^2}{J})$$

$$\hat{\mu} = \frac{1}{J}\sum_{j=1}^{J}\theta_j$$

(c) Conditional Posterior Distribution of σ^2

$$\sigma^2 \mid \theta, \mu, \tau, y \sim \text{Inv-}\chi^2(n, \hat{\sigma}^2) \quad \text{(inverse Chi}^2\text{)}$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{j=1}^{J} \sum_{i=1}^{n_j} (y_{ij} - \theta_j)^2$$

$$\begin{cases} X \sim \chi^2(\cdot) \\ Y = \frac{1}{X} \cdot \quad Y \sim \text{Inv-}\chi^2(\cdot) \end{cases}$$

(d) Conditional Posterior of τ^2

$$\tau^2 \mid \theta, \mu, \sigma, y \sim \text{Inv-}\chi^2(J-1, \hat{\tau}^2)$$

$$\hat{\tau}^2 = \frac{1}{J-1} \sum_{j=1}^{J} (\theta_j - \mu)^2$$

Gibbs Sampler

⋮

Example 3 : Pareto Principle

1/0.b Italian — civil engineer, sociologist, economist

80% of Italy's land was owned by 20% of the people

Def : Pareto Principle

Is a prediction that 80% of effects come from 20% of causes

Distributions of frequencies & sizes or the lik often tend to follow a power-law distribution

- wealth
- sales
- city size
- return
- word
- general

Power-law Distribution

Pareto distribution with shape $\alpha > 0$

& scale $\gamma > 0$ has the following pdf

$$\text{Pareto}(x \mid \alpha, \gamma) = \frac{\alpha \gamma^{\alpha}}{x^{1+\alpha}} 1\{x > \gamma\}$$

$$\propto \frac{1}{x^{1+\alpha}} 1\{x > \gamma\}$$

referred to as a Power Law Distribution

because the pdf is proportional to x raised to a power

γ is a lower bound on the observed values

Gibbs Sampling to perform inference for α & γ

Parameters of Pareto distribution

α : Scaling relationship between sizes of cities, returns on stocks & probability of occurring

γ : cutoff point :

any cities or returns on stocks smaller than γ are excluded from the dataset

Selection of the Prior :

- Proper Priors

- An Improper / default prior is a non-negative function of parameters which integrate to infinity

$$P(\alpha, \gamma) \propto 1_{\{\alpha, \gamma > 0\}}$$

- The resulting posterior (OFTEN) will be proper
- If not the case then the entire framework would break down

Recall $P(x \mid \alpha, \gamma) = \frac{\alpha \gamma^{\alpha}}{x^{\alpha+1}} 1_{\{x > \gamma\}}$

$1_{\{\alpha, \gamma > 0\}}$ ← improper one

Posterior

$$P(\alpha, \gamma \mid x_1, \cdots, x_n) \propto P(x_1, \cdots, x_n \mid \alpha, \gamma) \cdot P(\alpha, \gamma)$$

$$\propto 1_{\{\alpha, \gamma > 0\}} \prod_{i=1}^{n} \left(\frac{\alpha \gamma^{\alpha}}{x_i^{\alpha+1}} 1_{\{x_i > \gamma\}} \right)$$

$$= \frac{\alpha^n \gamma^{n\alpha}}{(\prod x_i)^{\alpha+1}} 1_{\{x_{min} > \gamma\}} 1_{\{\alpha, \gamma > 0\}}$$

$x_{min} = \min\{x_1, \cdots, x_n\}$

Gibbs Sampling

To use Gibbs sampler we need to sample from $\alpha \mid \gamma, x_1, \cdots, x_n$ and $\gamma \mid \alpha, x_1, \cdots, x_n$

what we have is $P(\alpha, \gamma \mid x_1, \cdots, x_n) \propto \frac{\alpha^n \gamma^{n\alpha}}{(\prod x_i)^{\alpha+1}} 1_{\{x_{min} > \gamma\}} 1_{\{\alpha, \gamma > 0\}}$

- $P(\alpha \mid \gamma, x_1, \cdots, x_n) \propto \frac{\alpha^n \gamma^{n\alpha}}{(\prod x_i)^{\alpha+1}} 1\{\alpha > 0\}$ ← ↓ drop since no α

$$= \alpha^n e^{-\alpha(\sum \log x_i - n\log\gamma)} 1\{\alpha > 0\}$$

$$f(x; a, b) = \frac{1}{\Gamma(a)} b^a x^{a-1} e^{-bx}$$

vs

$$\alpha^n e^{-\alpha(\sum \log x_i - n\log\gamma)} 1\{\alpha > 0\}$$

$$\propto \text{gamma}(n+1, \sum \log x_i - n\log\gamma)$$

- $P(\gamma \mid \alpha, x_1, \cdots, x_n) \propto \gamma^{n\alpha} 1\{x_{min} > \gamma > 0\}$

$$P(\gamma \mid \alpha, x_1, \cdots, x_n) \propto \gamma^{n\alpha} 1\{x_{min} > \gamma > 0\}$$

$$\text{Mono}(\alpha, x_{min})$$

Def Mono Distribution

For $\alpha > 0$ & $\beta > 0$

Mono (α, β) with probability distribution

$$\text{Mono}(x \mid \alpha, \beta) \propto x^{\alpha-1} 1\{0 < x < \beta\}$$

since $\int_0^\beta x^{\alpha-1} dx = \frac{\beta^\alpha}{\alpha}$

we have pdf: $\text{Mono}(x \mid \alpha, \beta) = \frac{\alpha}{\beta^\alpha} x^{\alpha-1} 1\{0 < x < \beta\}$

its cdf : $F(x \mid \alpha, \beta) = \int_0^x \text{Mono}(u \mid \alpha, \beta)\, du$

$$= \frac{\alpha}{\beta^\alpha} \cdot \frac{x^\alpha}{\alpha} = \left(\frac{x}{\beta}\right)^\alpha$$

To sample from $\text{Mono}(\alpha, \beta)$, utilize inverse transform

$U \sim u(0,1)$

$U = \left(\frac{x}{\beta}\right)^\alpha$ solve for x

$\beta^\alpha U = x^\alpha \Rightarrow x = \beta U^{\frac{1}{\alpha}}$

Exercise :

If $X \sim \text{Pareto}(\alpha, \gamma)$

$\frac{1}{X} \sim \text{Mono}(\alpha, \frac{1}{\gamma})$

Gibbs Sampling (Pareto Example)

- Initialize $\alpha^{(0)}$, $\gamma^{(0)}$
- update

for $i = 1, \dots, n$ → (# of sampling)

$\alpha^{(i)} \mid \gamma^{(i-1)}, x_{1:n} \sim \text{gamma}(n+1, \sum \log x_i - n \log \gamma^{(i-1)})$

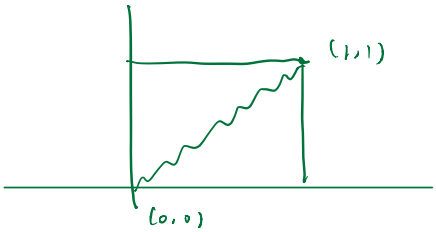
$\gamma^{(i)} \mid \alpha^{(i)}, x_{1:n} \sim \text{Mono}(n\alpha^{(i)} + 1, x_{\min})$

(Need more samples to stablize / converge)

Survival functions

$$S(x) = P(X > x) = 1 - P(X \leq x)$$

$$S(x \mid x_1, \cdots, x_n) = P(X_{n+1} > x \mid x_1, \cdots, x_n)$$



v.s

$$\hat{F}(x) = \frac{1}{n} \sum_{j=1}^{n} 1\{x \geq x_j\}$$

(Go through the code for approximation of integration)