$$P(\lambda) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}, \quad \lambda > 0$$

$$\Gamma(\alpha) = \int_0^{\infty} \lambda^{\alpha-1} \cdot e^{-\lambda} d\lambda$$

Posterior for $\lambda = \theta$

$$P(\lambda | X^{(n)}) = \frac{\lambda^{\alpha+n-1} \exp\{-\lambda(\beta + \sum_{i=1}^{n} x_i)\}}{\int_0^{\infty} \lambda^{\alpha+n-1} \exp\{-\lambda - \dots\} d\lambda}$$

$$\Rightarrow P(\lambda | X^{(n)}) = \frac{(\beta + \sum_{i=1}^{n} x_i)^{\alpha+n}}{\Gamma(\alpha+n)} \lambda^{\alpha+n-1} \exp\{-\lambda(\beta + \sum_{i=1}^{n} x_i)\}$$

Posterior is $\text{Gamma}(\alpha+n, \beta + \sum_{i=1}^{n} x_i)$

$$\hat{\lambda}_{\text{Pi mean}} = \frac{\alpha+n}{\beta + \sum_{i=1}^{n} x_i} \left\{ \frac{n}{n \cdot \bar{x}} \right\}$$

posterior mode : $\frac{\alpha+n-1}{\beta + \sum_{i=1}^{n} x_i}$

$$\Gamma(\alpha+n) = \int_0^{\infty} \lambda^{\alpha+n-1} e^{-\lambda} d\lambda$$

$$\int_0^{\infty} \lambda^{\alpha+n-1} e^{-\lambda \cdot (\beta + \sum_{i=1}^{n} x_i)} d\lambda$$

$\alpha - m$

$$\lambda = \frac{m}{\beta + \sum_{i=1} x_i}$$

$$\lambda^{\alpha+n-1} = m^{\alpha+n-1} \cdot (\beta + \sum_{i=1} x_i)^{-(\alpha+n-1)}$$

$$(\beta + \sum_{i=1} x_i)^{-\alpha+n\alpha)} \cdot \frac{1}{\beta + \sum x_i} \Gamma(\alpha+n)$$

Lec 3

MATH 459

Recall: $X^{(n)} = \{X_1, \ldots, X_n\}$, $X_i \overset{iid}{\sim} f(x|\theta)$

Likelihood: $L(\theta) \equiv L(\theta; X^{(n)}) = \prod_{i=1}^{n} f(x_i|\theta)$

log likelihood: $\log L(\theta) \equiv \ell(\theta)$

Posterior: $P(\theta|X^{(n)}) = \dfrac{p(\theta)\cdot f(x^{(n)}|\theta)}{\int_\theta p(\tilde{\theta}) f(x^{(n)}|\tilde{\theta})\, d\tilde{\theta}}$

Question: If the $p(\theta)$ is a density, then is $p(\theta) f(x^{(n)}|\theta)$ also a density?

Where does this ~~from~~? posterior ~~co~~ come from?

- Imagine θ is a random variable
  obviously X is a random varible
- Can not assume θ, X are indep
- Let $g(\theta, x)$ be the joint density

$$f(x|\theta) = \frac{g(\theta, x)}{p(\theta)}, \quad p(\theta) > 0$$

the postoriot is $P(\theta|x) = \dfrac{g(\theta, x)}{m(x)}$, $m(x) > 0$

- in general, ~~max~~ $m(x)$ unknown.
- Can find $m(x) = \int_\theta g(\theta, x)\, d\theta$
- and we know

$$g(\theta, x) = f(x|\theta) \cdot p(\theta)$$

- thus $P(\theta|x) = \dfrac{g(\theta, x)}{m(x)} = \dfrac{p(\theta)\cdot f(x|\theta)}{\int_\theta p(\tilde{\theta}) f(x|\tilde{\theta})\, d\tilde{\theta}}$

2 things we learn:

① denominator of posterior is marginal density for data

$$\frac{p(\theta)\cdot p(x^{(n)}\mid\theta)}{\int_\theta p(x^{(n)}\mid\tilde{\theta})\,p(\tilde{\theta})\,d\tilde{\theta}}$$

② $m(x) = \dfrac{p(\theta)f(x\mid\theta)}{p(\theta\mid x)}$

Basic Marginal Likelihood Identity.

of data

② $m(x)$: normalizing constant C

so that $\int_\theta p(\theta\mid x)\,d\theta = 1$

Interpretation:

posterior $\propto$ prior $\times$ likelihood

Multi-parameter Model

θ is a vector

$\theta = (\theta_1, \cdots, \theta_d)$

Then $X^{(n)}$ has $X_i \overset{iid}{\sim} f(x\mid\theta)$

$\theta \in \Theta \subseteq R^d$

Joint prior density $p(\theta_1, \cdots \theta_d)$

$\int_\theta \int p(\theta_1 \cdots \theta_d)\,d\theta_1 \cdots d\theta_d = 1$

not always true

$p(\theta_1 \cdots \theta_d) = p(\theta_1) \cdots p(\theta_d)$

- come up with a joint density
- for a subset of parameters of interest say θ

$p(\theta_1 \mid \theta_2 \cdots \theta_d)$

The likelihood equations:

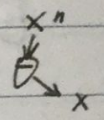
$\nabla_\theta \ell(\theta_1 \cdots \theta_d) = 0$

$\begin{pmatrix} \frac{\partial}{\partial\theta_1} \\ \vdots \\ \frac{\partial}{\partial\theta_2} \end{pmatrix}$

$$m(x|x^n) = \frac{p(\theta|x^n)\, f(x|\theta)}{p(\theta|x)}$$

$$m(x|x^n)\cdot p(\theta|x,x^n) = p(\theta|x^n)\, f(x|\theta,x^n)$$

$$m(x|x^n) = \frac{p(\theta|x^n)\, f(x|\theta,x^n)}{p(\theta|x,x^n)}$$



$$\frac{m(x|x^n)}{p(x^n)} = \frac{\dfrac{p(x^{(n)}|\theta)\cdot p(\theta)}{\int d\theta^n}\cdot f(x|\theta)}{\dfrac{p(\theta|x^n)\cdot p(x|\theta)}{\int d\theta\, d\theta^n}} = \int p(x^n|\theta)\, p(\theta)\, d\theta\, d\theta^n$$

$$\int m\, dx$$

$$m = \int m(x|x^{(n)})\cdot m(x^n)\, dx^n$$

Suppose a Baysian wants to inference for $\theta_1, \cdots \theta_2 \cdots \theta_d$. are nuisance

- Need marginal posterior for θ_1

$$p(\theta_1|x^{(n)}) = \iint\cdots p(\theta_1\cdots\theta_d|x^{(n)})\, d\theta_2\cdots\theta_d.$$

Bayesian estimators.

(1) mean of marg post

$$\int_{\theta_1} \theta_1\, p(\theta_1|x^{(n)})\, d\theta_1$$

(2) median of marg post

$$\hat{\theta}_{1\,med}: \int_{-\infty}^{\hat{\theta}_{1\,med}} p(\theta_1|x^{(n)})\, d\theta$$

$$= 0.5$$

(3) $$E(\theta|x^{(n)}) = \begin{cases} \int \theta_1\, p(\theta_1|x^{(n)})\, d\theta_1 \\ \vdots \\ \int \theta_d \cdots d\theta_d \end{cases}$$

Ex. Nomal model

$X_1 \cdots y_n \overset{iid}{\sim} N(\mu, \sigma^2)$

$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(x-\mu)^2}{2\sigma}\}$

$L(\underset{\sim}{\theta}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(X_i-\mu)^2}{2\sigma^2}$

$l(\mu, \sigma^2) = -\sum_{i=1}^{n} \frac{(X_i-\mu)^2}{2\sigma^2}$

$\hookrightarrow -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2$

likelihood equations

$$\frac{\partial(\mu, \sigma^2)}{\partial\mu} = \frac{-1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu) = 0$$

$$\frac{\partial(\mu, \sigma^2}{\partial\underset{\sim}{\sigma^2}} = \frac{1}{-2\sigma^4}\sum_{i=1}^{n}(X_i-\mu)^2 - \frac{n}{2\sigma^2} = 0$$

$$\mu = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$$

The solution $\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2)$

$$\mu = \bar{X} = \frac{1}{n}\sum_{i=1}(X_i - \bar{X})$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=}^{n}(X_i - \bar{X})^2$$

$$E(\hat{\sigma}^2) \neq \sigma^2 = \frac{n}{n-1}\sigma^2$$

Joint position

$$p(\mu, \sigma^2 | X^{(n)}) \propto p(\mu, \sigma^2) L(\mu, \sigma^2)$$

BDA3 P.64-65

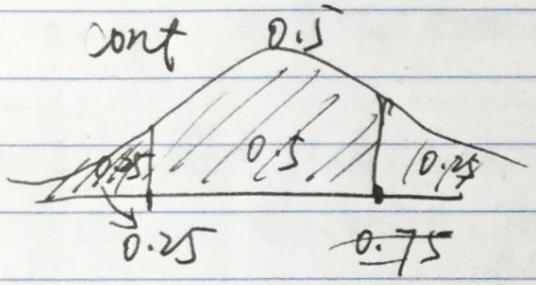
MATH 459

$\hat{\theta} = g(X^{(n)})$

$P(\theta) = 1/\theta \rightarrow$ improper prior (not a density)

↓

may still be possible posterior can be normalized.



$Z_n = \frac{n\bar{X} - n\mu}{\sigma\sqrt{n}} = \frac{n(\bar{X} - \mu)}{\sigma\sqrt{n}}$ (P)

1.96

$\frac{\sqrt{n}\cdot(\bar{X} - E(X))}{\sqrt{Var(X)}} \rightarrow \sim N(0,1)$ (sampled)

$\frac{n(\bar{X} - E(\bar{X}))}{\sigma^2}$

$= \frac{\sigma^2}{n}$

Sampled

$\frac{(\bar{X} - E(\bar{X}))}{\sqrt{Var\,\bar{X}}} \Rightarrow \frac{[\bar{X} - E(\bar{X})]\cdot\sqrt{n}}{\sqrt{Var\,X}}$

$\Rightarrow \frac{\bar{X} - \mu}{S/\sqrt{n}}$

$E[\bar{X}] = E[X]$

$n\,Var\,\bar{X} = n\,Var\,X$

Math I061

What is a good ~~dis~~ decision rule?

Ideal case: $\exists\ d \in D$ s.t.

$R(\theta, d)$ is uniformly smallest $\forall\ \theta \in \Theta$

Not practical.

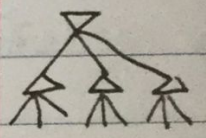
Admissibility: Given two decision rules d, d' d strictly dominates d' if $R(\theta, d) \leq R(\theta, d')$ $\forall\ \theta \in \Theta$ and is strictly less for at least one θ

Any role which is strictly is strictly dominated is inadmissible.

If d is not strictly ~~rode~~ dom. it is admissible

Minimaxity: The max risk of d is

$$MR(d) = \sup_{\theta \in \Theta} R(\theta, d)$$



d is minimax if $MR(d) \leq MR(d')\ \forall d' \in D$

Equivalently, d must satisfy $\sup_\theta R(\theta, d) = \inf_{d' \in \Theta} \sup R(\theta, d')$

If The supremum and infimum are ~~actually~~ attained.

$$\max_{\theta \in \Theta} R(\theta, d) = \min_{d' \in D} \max_{\theta \in \Theta} R(\theta, d')$$

Unbiasedness

d is unbiased if $E_\theta L(\theta', d(x)) \geq E_\theta L(\theta, d(x))$

θ: true

for $\forall\ \theta, \theta' \in \Theta$

d is able to ~~find out~~ different the optimal θ from ~~other~~ θ'

Recall, in estimator:

$d(x)$ is unbiased for θ, if $E_\theta d(x) = \theta$

If $L(\theta, d) = (\theta - d)^2$ these definition are the same

Bayes' decision rule

Problem has a seventh component

corresponding $\longrightarrow$ prior.

Bayes risk of d

$$r(p, d) = \int_{\theta \in \Theta} \underbrace{R(\theta, d)}_{\text{average over data}} p(\theta)\, d\theta$$

Bayes decision rule w.r.t $p(\theta)$ minimizes the Bayes risk, s.t.

$$r(p, d) = \inf_{d' \in D} \underbrace{r(p, d')}_{\text{~~given d~~}} = m_p.$$

Why people like this?

1. always (almost) admissible.
2. every admissible rule is either Bayes or a limit of Bayes rules.

Statistical Decision Problem

① Parameter space

② $\Theta \subseteq \mathbb{R}^d$

set of possible true states of nature

③ Sample space $\mathbb{X}$

where data live

typically n observations

generic element $x \in \mathbb{X}$ is $x = (x_1, \cdots, x_n) \in \mathbb{R}^n$

④ family of prob. dist on $\mathbb{X}$

indexed by values $\theta \in \Theta$

$\{P_\theta(x) : x \in \mathbb{X}, \theta \in \Theta\}$

⑤ Action Space A

Set of all actions available to the experimenter

(a) Hypothesis testing

decide b/w H_0, H_1

$A = \{a_0, a_1\}$ → reject H_0

Fail to reject

(b) Estimation

estimate θ by a func of $X^{(n)}$, e.g

$\bar{X} = n^{-1} \sum_{i=1}^{n} X_i$, or $X_1^3 + X_2 \cos\{\sqrt{X_3}\}$

⑥ Loss function

$L : \Theta \times A \longmapsto \mathbb{R}$

Links action to unknow Parameter. If we take action a when true state of nature is $\theta \in \Theta$, incur loss $L(\theta, a)$

⑦ Set of ~~desi~~ decision rule D

An element $d \in D$, $d : \mathbb{X} \longmapsto A$

s.t each $x \in \mathbb{X}$ is assoc. with $d(x) \in A$

EX. Hypothesis testing

Adopt rule. Fail to reject H_0 if $\bar{x} \le 3.6$, otherwise reject H_0.

$$d(x) = \begin{cases} a_0 & \text{if } \bar{x} \le 3.6 \\ a_1 & \text{if } \bar{x} > 3.6 \end{cases}$$

Risk function

For $\theta \in \Theta$ the risk of decision d based on random X

$R(\theta, d) = E_\theta[L(\theta, d(X))]$ ← ~~given θ~~ average over data. [given θ]

$$= \begin{cases} \int_x L[\theta, d(x)]\, f(x,\theta)\, dx \\ \sum_x L(\theta, d(x))\, f(x;\theta), \text{dx} \end{cases}$$

→ frequentist risk over θ (θ?)

Bayesian on (RL) $\xrightarrow[H(p,d)]{\Rightarrow}$ $E\{E_\theta[L(\theta, d(x))]\}$ (over x)

key notion

Different decision rules should be composed in terms of their risk as function θ

dx dθ

Common loss Functions for ~~estimation~~.

(a) $L(\theta, a) = (\theta - a)^2$

(b) $L(\theta, a) = |\theta - a|$ absolute error

(c) $L(\theta, a) = \begin{cases} 0 & \text{if } |\theta - a| \le \delta \\ 1 & \text{if } |\theta - a| > \delta \end{cases}$

mean square error

common loss func in Testing.

$$L(\theta, a) = \begin{cases} 1 & \text{if } \theta \in H_0 \text{ and } a = a_1 \\ 1 & \text{if } \theta \in H_1 \text{ and } a = a_0 \\ 0 & \text{otherwise} \end{cases}$$

Rist associated with these losses.

Type I and Type II error

Hence $R(\theta, d) = \begin{cases} Pr_\theta\{d(X) = a_1\} & \text{if } \theta \in H_0 \\ Pr_\theta\{d(x) = a_0\} & \text{if } \theta \in H \end{cases}$

Bayes risk

$$r(P_\theta, d) = \int_{\theta \in \Theta} \underbrace{R(\theta, d)}_{\text{given } \theta} p(\theta)\, d\theta$$

Bayes decision rule

$$r(p, d) = \inf_{d' \in D} r(p, d')$$

Bayse rule d satisfies this

Why Bayes is good.

① almost all Bayes rules admissible

② every admissible rule is either Bayes or a limit of Bayes rules

③ a decision rule d is minimax if (a) it is a Bayes for some prior P

(b) $\max_\theta R(\theta, d) \le r(P_\theta, d)$

P: prior, s.t. d is minimax

minimize the worst case ~~cost~~ cost

type 1 error

type 2 error

To minimize $r(p, d)$ sufficient to minimize

$$\int_\theta L(\theta, d(x))\, p(\theta|x)\, d\theta$$

"expected posterior loss"

Bayes Rules

Risk function:

$$R(\theta, d) = E_\theta L(\theta, d)$$

$$= \int_x L(\theta, d(x))\, f(x;\theta)\, dx$$

Bayes risk

$$r(p, d) = \int_\theta R(\theta, d)\, p(\theta)\, d\theta$$

$$= \int_\theta \int_x L(\theta, d(x))\, f(x|\theta)\, p(\theta)\, dx\, d\theta$$

$$= \int_\theta \int_x L(\theta, d(x))\, f(x)\, p(\theta|x)\, dx\, d\theta$$

and $\left(f(x) = \int f(x|\theta)\, p(\theta)\, d\theta\right)$

$f(x;\theta)$ → for frequentist concept

$p(\theta|x)$ → get Bayesian posterior using frequentist concept

↑ key for Bayesian ~~some~~ aspect

$$= \int_x f(x) \underbrace{\left\{\int_\theta L(\theta, d(x))\, p(\theta|x)\, d\theta\right\}}_{\text{w.r.t } d} dx$$

Bayes Point Estimators

Let $L(\theta, d) = (\theta - d)^2$

for observed $X = x$

choose $d(x)$ to minimize

$$\int_\theta (\theta - d)^2 p(\theta | x) d\theta$$

Differentiate w.r.t d

$$\Rightarrow \int_\theta (\theta - d)\, p(\theta | x) d\theta = 0$$

Hint $\int_\theta p(\theta | x)\, d\theta = 1$

$$\Rightarrow \int_\theta \theta\, p(\theta | x)\, d\theta = \int d\, p(\theta | x)\, d\theta$$

$$\Downarrow \qquad\qquad \Downarrow$$

posterior mean $= d$

more generally if $L(\theta, d) = (\theta - d)^2$

consider estimator

$R(\theta, d)$ = Mean squared error.

Now suppose $L(\theta, d) = |\theta - d|$

Bayes rule minimizes

$$\int_{-\infty}^{d} (d - \theta) p(\theta | x)\, d\theta + \int_d^{\infty} (\theta - d) p(\theta | x)\, d\theta$$

$$= \int_\theta |\theta - d|\, p(\theta | x)\, d\theta$$

Differentiate w.r.t d, and set to 0

$$\Rightarrow \int_{-\infty}^{d} p(\theta | x)\, d\theta = \int_d^{\infty} p(\theta | x)\, d\theta$$

when d is posterior median.

$$\int_{-\infty}^{\infty} L(v,\delta) f_{\theta|x}(v|x)\,dv = \int_{-\infty}^{\delta-\varepsilon/2} 1 \cdot f_{\theta|x}(v|x)\,dv + \int_{\delta+\frac{\varepsilon}{2}}^{} 1 \cdot f_{\theta|x}(v|x)\,dx$$

Bayes interval estimator

Suppose $L(\theta,d) = \begin{cases} 0 & \text{if } |\theta - d| \le \delta \\ 1 & \text{if } |\theta - d| > \delta \end{cases}$

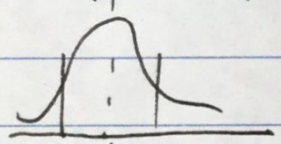
δ is a tolerance level.

Problem: Find the 'best' interval with ~~respect~~ prespecified length 2δ

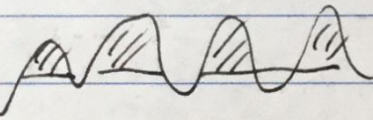
$(d-\delta, d+\delta) \Rightarrow$ LOST $= 0$

'Best' means the interval maximizes the posterior probability that $\theta \in (d-\delta, d+\delta)$

HPDI (Highest posterior desity interval)

If 

easy to construct

If 

tolerance

HPD set union of disjoint interval

simplify problem.

Lecture 1°:

Background: ① Prior Independence

If $p(\theta) = p(\theta_1)p(\theta_2)\cdots p(\theta_d)$

Does not imply a posteriori independence

also need $L(\theta) = L(\theta_1)L(\theta_2)\cdots L(\theta_d)$

① ②

Fisher information

$$I(\theta) = \mathrm{Var}(D_1) = -E(D_2)$$

here, $I(\theta) = E_\theta[\nabla_\theta \log f(X;\theta)\, \nabla_\theta \log f(X;\theta)^T]$

$$I_{ij} = E_\theta\left[\left(\frac{\partial}{\partial\theta_i}\log f(X;\theta)\right)\left(\frac{\partial}{\partial\theta_j}\log f(X;\theta)\right)\right]$$

$$= -E_\theta\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(X;\theta)\right]$$

③ Orthogonality

θ_i, θ_j are orthogonal if $I_{ij} = 0$

Select Prior by Formal Rules

Noninformative Prior & Objective prior

"maximize divergence between pri&post" — little to no influence on the posterior

a func that Jose M. Bernardo ↳ K-L divergence

use flat prior (& not invariant) => variance is large?

Subjective Bayesian

① choose a prior

② apply it to the likelihood the derive the posterior

$p(\eta) = 1$ (noninformative)

$p(\eta) \mid \eta = \log\frac{\theta}{1-\theta}$ is informative and also not invariant under one-to-one reparameterization

- $I(\eta) = I(\theta)\cdot\left(\frac{d\theta}{d\eta}\right)^2$

Invariant $\sqrt{I(\eta)} = \sqrt{I(\theta)}\cdot\left|\frac{d\theta}{d\eta}\right|$ by change-of-variable formula.

this shows Jeffreys prior $\pi_J(\theta) = \sqrt{I(\theta)}$ is invariant to a change of variable: $\eta \leftrightarrow \theta$

ex: $X \sim \exp(\lambda)$

$I(\theta) = -E[D_2] = \lambda^{-2}$

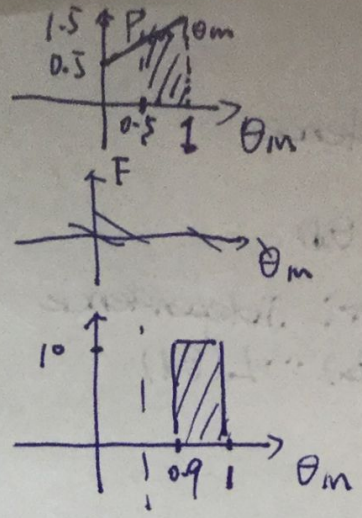
$\pi_J(\lambda) \propto \sqrt{I(\lambda)} = \lambda^{-1}$

$P_M(\theta_M) = \theta_M + 0.5$, $\theta_M \in (0,1)$

$P_{rM}(\theta_M > 0.5) = \int_{0.5}^{1} P_M(\theta) d\theta = 0.625$

$P_N(\theta_S) = 10$, $\theta \in (0.9, 1)$

$P_{rm}(\theta_M > 0.5) = \int_{0.5}^{1} P_M(\theta) d\theta = \infty 1$



prior dist

⇓

posterior dist

⇓

posterior confidence interval

$P(\theta_M | y_M = 0) \propto P(\theta_M) \cdot P(y_M | \theta_M)$

$\propto \frac{\theta_M}{2} - \theta_M^2 + \frac{1}{2}$

$P_{r\theta_M|Y_M}(\theta_M > 0.5) = \int_{0.5}^{1} P_M(\theta_M | y_M) d\theta_M = 0.35$

give up after one try

prior allow us to incorporate common sense

Lecture 9

Lecture 12

Approximate Bayesian Inference = Exact ~~analytic~~ analytic calculation of posterior quantities often not practical

Basics of Parametric Bayesian Asymptotics

Consistency of Posterior

$\Pi f(x|\theta_0) \Rightarrow \Pi(\cdot \mid X^{(n)}) \to 0$

if U near $\theta_0 \in \Theta$

$\pi(U \mid X^{(n)}) \to 1$ as $n \to \infty$

Asymptotic Normality of MLE: 4法則

and $\frac{\theta - \hat{\theta}_n}{\sqrt{n}} \to N(0, I^{-1}(\theta_0))$

MLE → MVB

(S.S.) sufficiency priciple

$\theta = (\mu, \sigma^2)$ $T(x) = (\bar{x}, s^2)$

① inference on θ should be based only on $T(x)$

completeness: unique unbiased S.S

MLE

Some of the priciple vidated by frequentist procedures:

① ad missibility

② unbiasedness

③ probability matching

Lecture 13

Approximate Bayesian Inference II

numerical intergration

inverse $\Rightarrow$ elementary function

~~$P(X \le F(x))$~~

key

$P(X < x) = P(U \le F(x)) = F(x)$

$X = F_X^{-1}(U)$

some special elementary function

Intergation Approximation (Type

① asymptotic expansion

② determinstic numerical approximation (quadrature method) convenge order $O(n^{-2})$ not good for high dim

③ Monte Carlo integration (utilizing randomness) conv order $O(n^{-1/2})$ large sample for high dimension

mc approx for mean: $\mu = E[h(x)]$

$$\hat{\mu}_{mc} = \frac{1}{n}\sum_{i=1}^{n} h(x_i)$$

$$\hat{s^2(f)}_{mc} = \frac{1}{n-1}\sum_{i=1}^{n} h(x_i) - \hat{\mu}_{mc})^2$$

Magic of MH, for every given q, we can construct a M-H kernel k so that f is its stationary distribution.
(more easier than A-R method)

comment:

(1) closely relate to A-R method (Rejection Sampling) but the proposal dist changes over time

(2) the acceptance rate is the averge of all iterations

$$\bar{\rho} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} \rho(X^{(t)}, Y_t)$$

$$= \int \rho(x,y) f(x) q(y|x) \, dy \, dx$$

(3) independent of normalizing constants (M)

(4) in the symmetric case, $\bar{\rho}$ can be reduced to a ratio independent of q. ex. $q(x|y) = q(y|x)$

Detaild Balance: $f(x) k(y|x) = f(y) k(x,y)$ exist a $f(x)$

Independent M-H

genetic MH: q depend on current state of the chain.

If we require q independ of x, i.e. $q(y|x) = g(y)$

We have special case

Given initial $x^{(t)}$

1. Generate $Y_t \sim g(y)$
2. Take

$$X^{(t+1)} = \begin{cases} Y_t & p = \min\left\{\frac{f(Y_t)}{f(X^{(t)})} \cdot \frac{g(X^{(t)})}{g(Y_t)}, 1\right\} \\ X^{(t)} & \text{otherwise} \end{cases}$$

generalization of A-R Sampling

Comparison of Rejection with M-H

① AR is sampled is iid, MH depend on $X^{(t)}$ (not matter how, alway)

② MH involves repeated occurence of the same value reject Y_t ⇒ repeat X_t at time $t+1$

③ no M & M needed

Select of proposal.

① random walk M-H (Generic approc.)
the proposal is random walk,
but the MC is not (A-R involved)

RWMH (long chain, half time revisiting)

Given initial $x^{(t)}$

1. Gen $Y_t \sim g(y - x^{(t)})$

2. Take $X^{t+1} = \begin{cases} Y_t & U \le p = \min\{1, f(Y_t)/f(x^t)\} \\ x^t & \text{other} \end{cases}$

$\bar{p}$ does not depend on g, but different choice of g ⇒ but different ranges for Y_t and ⇒ diff $\bar{p}$

② Langevin algorithm

$Y_t = X^{(t)} + \frac{\delta^2}{2} \nabla \log f(x^t) + \delta \epsilon_t$,

$\epsilon_t \sim g(\epsilon)$

(require a lot of tuning in practice)

Lecture 14

MC method

source of randomness.

① physical process
~~is~~ not fast
need storage.

② Pseudo-random number generators

Linear Congruential Generators LCG/MCG
multiplicative
multiple recursive generator MRG

Mersenne twister RNG. default in R

Definition. for any continous CDF F

$U \sim Unif(0,1)$, then

$X = F^{-1}(U)$

proof. $P(X < x) = P(X < F^{-1}(U))$

$= P(U < F(x))$

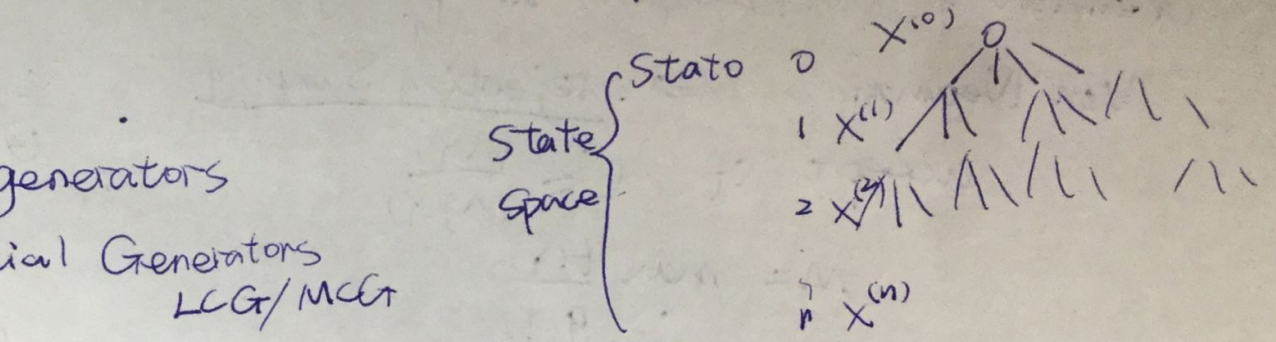
$= F(x)$

problem no high dimension
high cost

revolutionary idea: Markov Chain

Mon Car Markov Chain

MC MC : a random process



$X^{(t+1)} | X^{(0)}, X^{(1)}, X^{(2)} \dots X^{t}$

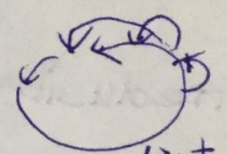
$\sim K(X^{(t)}, X^{(t+1)})$

kernel

- Stationary: if $X^{(t)} \sim f$ then $X^{(t+1)} \sim f$

$\int_x K(x,y) f(x) = f(y)$

- recurrent



the stationary dist is also a limiting dist which means the limiting dist of $X^{(t)}$ is f for almost all initial $X^{(0)}$ ⟹ ergodic ⟹ eventually produce f

Ergodic Theorem → $\frac{1}{T}\sum_{t=1}^{T} h(X^{(t)}) \to E_f[h(x)]$

lecture 15

Metroplis-Hasting. Gibbs sampling
reversible jump, convergence diagnostics

Generating iid samples accept-reject method

Von Neumann's idea: rejection Sampling

accept if $U \le \frac{f(Y)}{Mg(Y)}$

$M = \max_x \frac{f(Y)}{g(Y)}$

$\sup_x$

Simple Random Walk

$X^{(t+1)} = X^{(t)} + \epsilon_t$

$K(X^{(t)}, X^{(t+1)}) \equiv N(X^{(t)}, 1)$

$f(x^{t+1} | x^{(t)})$

irreducible: regardless of $X^{(0)}$

positi prob → reach any part of the state space

an irreducible MC → recurrent if return to arbitary part of the state infinite time

the stationary of is also a limiting dist ⇒ ergodition theorem

Basic idea of MCMC

Given target dist of f

① ⇒ build a Markov kernel

$K(x, y) = P_r(X^{(t)} = y | X^{(t-1)} = x)$

with stationary dist π the same as target f

② → generate a MC $\{X^{(t)}\}$ using the kernel K so that the limiting dist of $\{X^{(t)}\}$ is f

and $T^{-1} \sum_{t=1} h(X^{(t)}) \to E_f[h(x)]$

idea of Metropolis-Hastings

Given initial $X^{(t)}$, and conditional density $q(y|x^{(t)})$ → candidate dist

1. Generate $Y_t \sim q(y | x^{(t)})$
2. Set $x^{(t+1)} \overset{(A)}{=} \begin{cases} Y_t \text{ wit Prob } \rho(x^{(t)}, Y_t) \\ (R)\ x^{(t)} \text{ Prob } 1 - \rho(x^{(t)}, Y_t) \end{cases}$

where $\rho(x, y) = \min\left\{\frac{f(y)}{f(x)} \frac{q(x|y)}{q(y|x)}, 1\right\}$

$\frac{1}{T} \sum_{t=1}^{T} h(x^{(t)}) \to E_f(h(x))$

Gibbs sampling

$\theta = (\theta_1, \cdots \theta_d)^T$ (posterior of θ)

$p(\theta_j \mid \theta_{(-j)}^{t-1}, y)$

↑ expect j

$\theta_{(-j)}^{t-1} = (\theta_1^t, \cdots \theta_{j-1}^t, \theta_{j+1}^{t-1}, \cdots \theta_d^{t-1})^T$

R package

MCMCpack for MCMC

rstan ⟹ output coda mcmc object

or boa

ex. from BDA

(y_1, y_2) (observe)

$\theta = (\theta_1, \theta_2)^T$, $\mathrm{cov}\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$

$\Rightarrow \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \mid y \sim N\left(\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$ (easier)

$\Rightarrow \theta_1 \mid \theta_2, y \sim N(y_1 + \rho(\theta_2 - y_2), 1-\rho^2)$ (Gibbs sample, time consuming)

$\theta_2 \mid \theta_1, y \sim N(y_2 + \rho(\theta_1 - y_1), 1-\rho^2)$

Mixing of Chain

Some application #tuning method

Lec 16

Mixing of Chain. Mixing time

Auto correlation: $\rho_k = \frac{\sum_{i=1}^{n-k}(x_i-\bar{x})(x_{i+k}-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$

expect the k-th to be $\searrow$ as $k\uparrow$

Any Markov chain will have autocorrelation current value depend on the previous one

Lec 17

Linear Regression model

$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{k-1} X_{ik-1} + \varepsilon_i$

$Y = X\beta + \varepsilon$

$E(Y|X) = \beta_0 + \sum_{j=1}^{k-1} \beta_j X_j + E(\varepsilon|X)$, $Var(\varepsilon|X) = \sigma^2 I_{n\times n}$

Least square estimator. $\min_{\beta} (Y-X\beta)^T(Y-X\beta)$

$\Rightarrow \hat{\beta} = \underbrace{(X^TX)^{-1}X^TY}_{\text{provied exsits.}}$, $\hat{\sigma}^2 = \frac{(Y-X\hat{\beta})^T(Y-X\hat{\beta})}{n-k}$

maximize $\underbrace{N(0, \sigma^2 I_{n\times n})}$ $\Rightarrow$ LS problem.

Towards Bayesian Analysis

$f(x,y|\psi,\theta)$, $L(\psi,\theta)$

prior $p(\psi,\theta)$

$\Rightarrow p(\beta,\sigma^2|X) \propto \sigma^{-2}$

$\Rightarrow p(\beta,\sigma^2|X,y) = p(\beta|\sigma^2,X,y)p(\sigma^2|X,y)$

Start by find the posterior of β, conditional on σ, then find marginal distribution of σ^2.

$\Rightarrow \beta|\sigma,X,y \sim N(\hat{\beta}, (X^TX)^{-1}\sigma^2)$

$$p(\sigma^2|y) = \frac{p(\beta,\sigma^2|X,y)}{p(\beta|\sigma^2,X,y)}$$

$\Rightarrow \sigma^2 \sim Inv-\chi^2(n-k, s^2)$

$$s^2 = \frac{1}{n-k}(Y-X\hat{\beta})^T(Y-X\hat{\beta})$$

In practice, we use MCMC by drawing from σ and then drawing $\beta|\sigma$,

The joint posterior $p(\beta,\sigma^2|X,y)$ is proper provided that (X^TX) is invertible

$\Rightarrow$ 1. $n > k$
2. rank $(X) = k$

Posterior Predictive Density

given: $\tilde{y} \mid \beta, \delta^2, \underset{\text{given}}{x^*} \sim N(x^*\beta, \delta^2)$

$$p(\tilde{y} \mid \underset{\text{train}}{y}) = \int p(\tilde{y} \mid \beta, \delta^2)\, p(\beta, \delta^2 \mid y)\, d\beta\, d\delta^2$$

Sample from $p(\beta, \delta^2 \mid X, y)$

1. Compute $\hat{\beta}$ and $(X^T X)^{-1}$
2. Compute S^2
3. Draw δ^2 from inverse χ^2 dist
4. Draw β from multivariate normal dist.

If we know $p(\delta^2 \mid \beta, X, Y)$ and $p(\beta, \delta^2 \mid X, \beta, Y)$
we can sample $p(\beta, \delta^2 \mid X, y)$ using Gibbs sampling

Initialize $\beta^{(1)}$, $\delta^{2(1)}$

For $t = 1 : T$

$\beta_{t+1} \sim p(\beta^{(t)} \mid t_{(t)}, X, Y)$

$\delta^{(t+1)} \sim p(t_t^2 \mid \beta_{t+1}, X, Y)$

END.

The puffin data from LearnBayes package

(Lm) → Frequentist Fit

('blinreg) → Bayesian fit

What MCMCregress does.. Bayesian Fit

~~Simp~~ Simulates from posterior using Gibbs Sampling