ch 2.

- what is conjugate prior.
- change of variable · { why not uniform
- Jeffreys prior : det(I(θ))^{1/2}

  what it connects to data

- Non-informative · Reference Prior.
- Expert prior ·
- KL distance

· { · Direct integration
  · with (hyper) prior

why? Integration is hard.

- Bayesian Regression
- Bayesian variable selection

ch 4.

- Laplace approximation
- Independent MC methods

Bayesian Testing

H_0: θ belongs to some interval.

$\theta \in \Theta_0$

$\pi(\Theta_0|x) = \int_{\Theta_0} \pi(\theta|x)$

- Bayes Factor $\frac{\pi(\Theta_0|x)}{\pi(\Theta_1|x)} \frac{\pi(\Theta_1)}{\pi(\Theta_0)}$ kinds of the (hyper)priors.

"Posterior odds" × "prior odds"

- A subset of Lesbegue Measure zero.

e.g. $\Theta_0 = \{\mu = \mu_0\}$ zero mean

$\Rightarrow \pi(d\theta) = p_0 \pi_0(d\theta) + (1-p_0)\pi_1(\theta)\, d\theta$

π_0 : dist. concentrated on Θ_0

p_0 : prior prob.

p-values : Frequent: evidence against null hypothesis

$\pi(\theta_0|x) = \frac{f(x|\theta_0)}{f(x|\theta_0) + \int \pi_1}$ $\Theta_0 = \{\theta_0\}$

$\geq \frac{f(x|\theta_0)}{f(x|\theta_0) + \sup_\theta f(x|\theta)}$

A - Bayesian confidence set with level $(1-\alpha)$.

$(1-\alpha)$ creditable set.

$P(\theta \in C_x | X = x) = \pi(C_x|x) \geq 1-\alpha$

depend on x

HPD, the one minimizing the volume

$L_k = \{\theta; \pi(\theta|x) \geq k\}$ $k_\alpha = \sup\{k; \pi(L_k|x) \geq 1-\alpha\}$

why.

. Δ

Bayesian asymptotics

Fisher Information

$\hat{\theta}_n$ MLE $\sim N(\theta_0, \frac{1}{I_n})$

Bayesian $\theta | x_n \sim N(\hat{\theta}_n, \frac{1}{n} I^{-1})$

influence of prior vanish.

Bayes Formula 1.1

i countable set

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\Sigma_k P(B|A_i)P(A_i)} \quad P(A_i)>0,\ P(B)>0$$

continuous case.

$$f_{x|y}(x|y) = \frac{f_{xy}(x,y)}{f_y(y)} =$$

$$f(y|x) = \frac{f(x,y)}{f(x)} = \frac{f(x|y)f(y)}{f(x)} = \frac{}{f_{R}}$$

- Frequentist: Relatively frequency of A_i among those happening with B
- Bayesian: describing the uncertainty after we saw the data.

- How to choose a prior?
  1. Cannot depend on data.
  2.

prior and posterior predictive distribution

$f(x) = \int f(x|\theta)\pi(\theta)d\theta$ is called marginal density.
prior predictor density

After X observed:
Could observe distribution of Y from the same model

Posterior predictive distribution

$$f(y|x) = \int f(y,\theta|x)d\theta$$
$$= \int f(y|x;\theta)\pi(\theta|x)d\theta$$
$$\Delta \quad = \int f(y|\theta)\pi(\theta|x)d\theta \quad Y \perp X|\theta$$

Example All the observation IN

得3: $\exp\left(-\frac{1}{2}\left(\Sigma x_i - \mu\right)^2 - \frac{n\tau^2}{1+n\tau^2}(\bar{x}-\mu)^2\right)$

-

frequest risk
Bayes
weighted

Point Estimation

Summarize Baye dist. in a number.

Loss function $L: \Theta^2 \longrightarrow [0,\infty)$

Find the minimize of risk

$$\hat{\theta} = \arg\min_T \rho(T(x), \pi) \qquad \rho(T(x),\pi) = \mathbb{E}[L(T(x),\theta|x)] = \int_\Theta L(T(x),\theta)\pi(\theta|x)d\theta$$

有了 T_1, I_2 are good for a particular estimator but not all $R(T_1,\theta_1) < R(T_2,\theta_1)$ but $R(T_1,\theta_2) >$ ...

$\frac{1}{2}$ To look for estimator which minimizes the max risk

$\sup_\theta R(T,\theta)$

$$R(T,w) = \int_\Theta R(T,\theta)w(\theta)d\theta$$

If $\int w(\theta)d\theta = 1$ w as prior for θ.

$T(x) = \arg\min_T \rho(T(x),w) = \arg\min_T \mathbb{E}[L(T,\theta)|x]$

is well defined for $f(x) = \int f(x|\theta)w(\theta)d\theta$

Then any minimizer T' is almost surely T

Admissible No other estimator is better
$R(T',\theta) \le R(T,\theta)$

Bayes estimators are often biased.

I_{θ}

I_r

2.1 conjugate prior

Def $P_\Xi = \{\pi_\xi(\theta), \xi \in \Xi\}$ of prior densities is called conjugate

if $\forall \pi \in P_\Xi$, $\pi(\theta|x) \in P_\Xi$

Example: Normal → Normal
Gamma → Gamma..

- choosing a prior's hyperparameters is even harder than choosing parameters itself. Similar in the case of Normal mean, it tells how much we want to rely on the $\pi(\xi)$

2.2 Non-informative Priors

Problems with const. uniform/flat prior.

- Θ must be finite which is often not the case
- under c.o.v. the density is NOT invariant.

e.g. $\eta = g(\theta)$ be the new parameter

$\lambda(\eta) := \pi(g^{-1}(\theta)) |\det Dg^{-1}(\eta)|$ when g is NOT linear, $\lambda(\eta)$ is NOT const.

Improper priors: a measure that is NOT a probability measure.

If $\pi(\theta) f(x|\theta)$ has FINITE mass, then inference can be made easily as before. and this is NOT a easy task to check

Jeffrey's Prior $\pi(\theta) \propto \det(I(\theta))^{\frac{1}{2}}$

eg. $X \sim N(\theta, 1)$ $I(\theta) = 1$ — improper on $\mathbb{R}$.

$X \sim N(0, \theta^2)$ $I(\theta) = 2/\theta^2$ ~ $\frac{1}{\theta}$

- Reasoning behind Jeffrey's priors:

since $I(\theta)^{-1}$ is asymptotic variance of MLE. thus it is a indicator of How much info. from data about θ.

It's also invariant under change of variables so. Good!

eg. $\eta = g(\theta)$ $I_\eta(\eta) = Dg^{-1}(\eta)^T I_\theta(g^{-1}(\eta)) Dg^{-1}(\eta)$

$\Rightarrow \pi_\eta(\eta) \propto \det I_\eta(\eta)^{\frac{1}{2}} = |\det Dg^{-1}| \det I_\theta(g^{-1}(\eta))$

same results with c.o.v. formula.

Reference Priors

If X (data) has the largest impact then the impact of the prior is minimal.

KL-distance $KL(f, g) = \int f(x) \log \frac{f(x)}{g(x)} dx$

Idea is to maximize

$$I(X, \theta) = \int_X f(x) \int_\Theta \pi(\theta|x) \log \frac{\pi(\theta|x)}{\pi(\theta)} d\theta dx$$

$$= \int_\Theta \pi(\theta) \int_X f(x|\theta) \log \frac{\pi(\theta) f(x|\theta)}{\pi(\theta) f(x)} d\theta dx$$

$$= \int_\Theta \pi(\theta) \int_X f(x|\theta) \log \pi(\theta|x) dx d\theta - \int_\Theta \pi(\theta) \log \pi(\theta) d\theta$$

is also unfeasible to find the maximum value.

How to solve

$I_\infty(X, \theta) = \lim_n I((x_1, \dots x_n), \theta)$, still infinite

Approximating.

we again have Jeffrey's prior in the limit

The hierarchical Bayesian Model

hyperparameters that $\pi(\theta)$ depends on also has a prior distribution.

$$\pi(\xi)\,\pi(\theta|\xi)\,f(x|\theta)$$

The primary interest is posterior $\pi(\theta|x)$

Way 1. $\pi(\theta) = \int \pi(\theta|\xi)\pi(\xi)\,d\xi$

Way 2 $\pi(\theta|x) = \int \pi(\theta,\xi|x)\,d\xi = \int \pi(\theta|x,\xi)\pi(\xi|x)\,d\xi$

$$\propto \int \pi(\theta|x,\xi)\pi(\xi)f(x|\xi)\,d\xi$$

There leaves $\underset{①}{\pi(\theta|x,\xi)}$ and $f(x|\xi)$ / $\underset{②}{\pi(\xi|x)}$ to determine

$$\pi(\xi|x) = \frac{\pi(\theta,\xi|x)}{\pi(\theta|x,\xi)} \quad \text{or} = \int \pi(\theta,\xi|x)\,d\theta$$

Conjugate priors have explicit expression for $\pi(\theta|x,\xi)$ also $f(x|\xi)$

1. Normal means

Setting $f(x|\theta) \sim N(\theta, 1)$

prior: $\pi(\theta|\mu,\tau^2) \sim N(\mu,\tau^2)$ with μ fixed

$\tau^{-2} \sim Gamma(\gamma,\lambda)$

The posterior $\pi(\theta|\mu,\tau^2,x) \sim N\left(\frac{(\mu+n\tau^2\bar{x})}{1+n\tau^2}, \frac{\tau^2}{1+n\tau^2}\right)$

Using way 1 Direct integration

$$\pi(\theta) \propto \int (\tau^{-2})^{\frac{1}{2}}\exp\left\{-\tau^{-2}\frac{(\theta-\mu)^2}{2}\right\}(\tau^{-2})^{\gamma-1}\exp(-\lambda\tau^{-2})\,d\tau^{-2}$$

$$= \int u^{\gamma-1/2}\exp\left(-\left(\lambda+\frac{(\theta-\mu)^2}{2}\right)u\right)du$$

$$\propto \frac{1}{\left(\lambda+\frac{(\theta-\mu)^2}{2}\right)^{\gamma+1/2}} \quad \text{scaled and shifted}$$

t-dist. w/ 2γ d.o.f.

$\pi(\theta|x) \propto \pi(\theta)f(x|\theta)$ does NOT belong to standard family but we know sth about the posterior mode.

is closed to $\bar{x}$ if prior and data are in conflict,

Using Way 2

$\pi(\theta|\tau^2,x)$ is already known

$f(x|\tau^2)$ is also kind of known by marginaling

$$f(x|\tau^2) \propto (1+n\tau^2)^{-\frac{1}{2}}\exp\left\{-\frac{n}{2(1+n\tau^2)}(\bar{x}-\mu)^2\right\}$$

$\Rightarrow$ marginal posterior of τ^2 has little mass for small values. $\Rightarrow$ It favors large values of τ^2

prior and data are in conflict

Empirical Bayes for Normal means.

$\hat{\xi}(x) = \underset{\xi}{\arg\max}\, f(x|\xi)$

For $f(x|\tau^2) \Rightarrow \hat{\tau}^2 = \max\left(0, (\bar{x}-\mu)^2 - \frac{1}{n}\right)$

Means: we choose a wider prior when $\bar{x}$ and μ are in conflict.

$$E[\theta|x,\hat{\tau}^2] = \begin{cases} \mu \\ \bar{x} + \frac{1}{n(\mu-\bar{x})^2} \end{cases}$$

converges to μ as $n\uparrow\infty$

$\Rightarrow$ EB methods tends to underestimate uncertainty

2. Hierarchical Poisson Model

likelihood $f(x|\theta_j)$ — Poisson $\dfrac{e^{-\theta_j}\,\theta_j^{x_j}}{x_j!}$

Prior Gam: $\theta_j \sim Gam(\gamma,\lambda) = \dfrac{\lambda^\gamma}{\Gamma(\gamma)}\theta_j^{\gamma-1}\exp(-\lambda\theta_j)$

先不管 $\pi(\gamma,\lambda)$

Way 1 $\pi(\theta_1\cdots\theta_J) = \int \frac{\lambda^{J\gamma}}{\Gamma(\gamma)^J}\prod\theta_j^{\gamma-1}\exp\{-\lambda\sum\theta_j\}\pi(\lambda,\gamma)\,d\lambda\,d\gamma$

This is NOT a standard distribution.

Way 2 $\pi(\theta_1\cdots\theta_J|\gamma,\lambda,x) = \prod^J \pi(\theta_j|\gamma,\lambda,x_j)$

$f(x_1\ldots x_J|\gamma,\lambda) = \prod^J f(x_j|\gamma,\lambda)$

$$\pi(\theta_j|\gamma,\lambda,x_j) = \frac{(\lambda+1)^{\gamma+x_j}}{\Gamma(\gamma+x_j)}\cdot\theta_j^{\gamma+x_j-1}\exp(-(\lambda+1)\theta_j)$$

~~Gamma~~

To solve $f(x_j|\gamma,\lambda) = \int f(x_j|\theta_j)\pi(\theta|\gamma,\lambda,x_j)\,d\theta_j$

$$= \int Pois(\theta_j)\cdot Gamma\ d\theta_j$$

$$= \frac{\Gamma(\gamma+x_j)}{x_j!\,\Gamma(\gamma)}\frac{\lambda^\gamma}{(\lambda+1)^{\gamma+x_j}}$$

$$= \binom{\gamma+x_j}{x_j}\left(\frac{\lambda}{\lambda+1}\right)^\gamma\left(\frac{1}{\lambda+1}\right)^{x_j}$$

Negative binomial

$$\Rightarrow \pi(\theta_j|x_1\ldots x_n) = \int \pi(\theta_j|\gamma,\lambda,x_j)\,\pi(\gamma,\lambda|x)\,d\gamma\,d\lambda$$

$$\propto \int \prod^J f(x_j|\gamma,\lambda)\,\pi(\gamma,\lambda)\,d\gamma\,d\lambda$$

Cannot be Integrated in close form

Empirical Bayes.

$f(x|\gamma,\lambda) \sim$ Negative Binomial

$$\Rightarrow J\gamma\log\left(\frac{\lambda}{1+\lambda}\right) - \sum^J x_j\log(1+\lambda) + \sum\sum\log(\gamma+k)$$

$$\Rightarrow \frac{\hat{\gamma}}{\hat{\lambda}} = \bar{x}$$

Regression and model selection

$y = \alpha 1 + X_\gamma \beta_\gamma + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2)$

Assumption: γ fixed α independent

$\pi(\alpha, \beta_\gamma, \sigma^2) \propto \pi(\sigma^2) \cdot \pi(\beta_\gamma | \sigma^2$ since α is flat

$= \sigma^{-2} \pi(\beta_\gamma | \sigma^2)$

Popular choice of g-prior of zellner

$\pi(\beta_\gamma | \sigma^2) \sim N(0, g\sigma^2 (X_\gamma^T X_\gamma)^{-1})$

Posterior: $\beta_\gamma | \sigma^2, y \sim N(\frac{g}{g+1} \hat{\beta}_\gamma, \frac{g\sigma^2}{\sigma^2+1} (X_\gamma^T X_\gamma)^{-1})$

$\alpha | \sigma^2, y \sim N(\hat{\alpha}, \frac{\sigma^2}{n})$

$\hat{\beta}, \hat{\alpha}$ is the MLE.

posterior mean $\frac{g}{g+1} \hat{\beta}_\gamma + \frac{1}{g+1} \beta_\gamma^0$

a convex combination of prior mean and MLE

shrinkage towards zero

- $g \to \infty$ non-informative prior

$B(\gamma, 0) \to 0$ in the limit we will always choose empty model.

x k compont.
we have to know $\pi_i = \pi(x_i | x_{-i})$
Every π_i needs to be known.

KWM $Y^t \sim N(X^{t-1}, \Sigma)$

MH algorithm.
- Reversible distribution:
$$\int_A \pi(x) P(x, B) dx = \int_B \pi(x) P(x, A) dx$$
ie $\mathbb{P}[X^t \in A, X^{t+1} \in B) = \mathbb{P}[X^{t+1} \in A, X^t \in B]$
- π is also invariant
- $\pi(x) p(x,y) = \pi(y) p(y,x) \quad \forall x, y.$
reversibly

MH assumption
- generate a chain which has reversible distribution
可以随意转移吗？ No, $\int p(x,y) dy = 1$
usually does NOT hold for x

Algorithm
1. Simulate X^0
2. For t=1...
ⓐ Generate $Y^t \sim q(X^{t-1}, x) dx$ and U^t
ⓑ $X^t = \begin{cases} Y^t & \text{if } U^t \le a(X^{t-1}, Y^t) \\ X^{t-1} & \text{else} \end{cases}$

How to solve the problem of reversibility
$p(x,y) = q(x,y) \qquad p(y,x) = \frac{\pi(x) q(x,y)}{\pi(y)}$
↗ Arbitary Transition Kernel.
$\Rightarrow p(x,y) = \frac{\pi(y) q(y,x)}{\pi(x)} \qquad p(y,x) = q(y,x)$
折中一下 $\quad p(x,y) \le q$
$p(y,x) \le q \qquad \forall x \ne y$
$$a(x,y) = \min\left(1, \frac{\pi(y)}{\pi(x)} \frac{q(y,x)}{q(x,y)}\right)$$

Hamiltonian MC
- Drawbacks of Gibbs sampler / MH
step-wise is small.
Assume: we can evaluate the gradient efficiently
How: 1 consider a new target $\tilde{\pi}$
$\tilde{\pi}(x,u) \propto \pi(x) \exp\{-\frac{1}{2} u^T M^{-1} u\} \quad M = \mathrm{diag}(m_i)$
u_i is called monetum variable.
Based on a determistics, invertible map $G(x,u)$

The transformation $G(x,u)$ is given by ODE
$$\frac{dx_i}{dt_i} = \frac{\partial H(x,u)}{\partial u_i} = \frac{u_i}{m_i}$$
$$\frac{du_i}{dt_i} = \frac{\partial H(x,u)}{\partial x_i} = \frac{\partial \log \pi(x)}{\partial x_i} \qquad 0 \le t_i \le T$$

Remarks of MCMC.
. Biasness $\quad \mathbb{E}\, h_{N,r} \ne \int h(x) \pi(x) dx$
2. $\mathrm{Var}(h_{N,r})$ is complicated as it includes the covariance term.

4 Bayesian Computation

1 Laplace Approximation

$$\int h(\theta) q(\theta)\, d\theta$$ where • q has max • h arbitary smooth

• $\log q(\theta) = \log q(\theta_0) - \frac{1}{2}(\theta-\theta_0)^T J (\theta-\theta_0)$

$\Rightarrow \int h(\theta) q(\theta) d\theta \approx h(\theta_0) q(\theta_0) (\det J_0)^{-\frac{1}{2}} (2\pi)^{\frac{p}{2}}$

$q(\frac{1}{N})$

Bayes Factor and BIC

$B_{12} = \frac{f(\ \ |M_1)}{f(\ \ |M_2)}$

$f(x_1 \dots x_n | M_i) = \int \pi \prod f(\ |\theta)\, d\theta_i$

$\approx \pi_i \hat{\theta}_i \prod f(x_i | \hat{\theta}_i) (\det n I(\theta))^{-\frac{1}{2}} (2\pi)^{\frac{p_i}{2}}$

$\Rightarrow \log f(x_n | M_i) \approx \sum \log \quad - \frac{p_i}{2} \log n + O(1)$

BIC this

2. Independent Monte Carlo Methods

on the basis of drawing independent U^t

Basic idea is that we can use quantile function and $(0,1)$ to generate RV following certain distribution

$F^{-1}(u) = \inf\{x, \quad \}$

But this is often intractable in high-dim setting.

Rejection sampling

• Simulate with a diff. dist γ (Proposal)

• Connect to target π

$$\alpha(x) = \frac{\pi(x)}{M\gamma(x)} \le 1 \quad \text{envelope}$$

Proof: $\mathbb{P}[X \in dx] = \mathbb{P}[Y \in dx \mid U \le \frac{\pi(Y)}{M\gamma(Y)}]$

$\propto \mathbb{P}[Y \in dx]\, \mathbb{P}[U \le \ \mid Y \in dx]$

$= \frac{\pi(x)}{M\gamma(x)} \gamma(x)\, dx \propto \pi(x) dx$

Requirement know:

π up to a normalizing const.

Rejection is high unless π is close to γ

However, It is hard to find a really close one

Importance Sampling

Goal: $\mathbb{E}_\pi(h(x)) = \int h(x) \pi(x)$

Instead of rejection $\Rightarrow$ weight them with an appropriate weighting function.

$\frac{1}{N} \sum_{t=1}^{N} h(Y^t) w(Y^t)$ where $w(x) = \frac{\pi(x)}{\gamma(x)}$

No need to bound this

These are methods to generate independent

MCMC generate

$X^t = G(X^{t-1}, U^t)$ can be more in $\mathbb{R}^d$

$P(X^t \in A \mid \text{—}) = \mathbb{P}[X^t \in A \mid X^{t-1}] = \mathbb{P}[X^{t-1}, A]$

Determined by G

$P(x, A) = P(G(x,u) \in A) = \mathbb{P}[u \in \{u; G(x,u) \in A\}]$

How to specify G?

Some properties about MC.

Positivity: All state can be reached $\pi(B) > 0$

• Invariance / Stationary properties

$\pi(A) = \int \pi(x) P(x, A) dx \quad \forall A$