TOPIC:

DATE:

FILE UNDER:

PAGE:

Integration

We often need to compute:

① $\mathbb{E}_f g(X)$ — expectations / moments

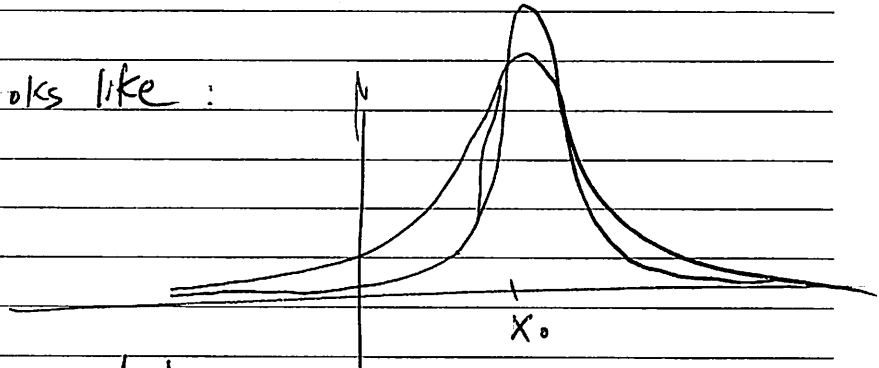
② $\int p(x|\theta)\,dx$ — normalizing constants for densities.

⇒ In general this can be very hard, but there are many special cases.

Suppose we want to calculate:

$$\int g(x)\,dx \qquad \text{and} \quad g \in L^2$$

Suppose g looks like:



So g is highly concentrated about x_0

Then we could say (?)

$$\int g(x)\,dx \doteq g(x_0)\,\varepsilon \quad ?$$

$$\int_{x-a}^{x+a} g(x) =$$

TOPIC:

DATE:

FILE UNDER:

PAGE:

Suppose we want

$$\int_a^b g(x)\,dx = \int_a^b e^{h(x)}\,dx \qquad \left[h(x) = \log g(x)\right]$$

Suppose h achieves its max at x_0

$$\doteq \int_a^b \exp\left(h(x_0) + \underbrace{h'(x_0)(x-x_0)}_{=0} + \frac{h''(x_0)(x-x_0)^2}{2}\right) dx$$

$$= \int_a^b \exp\left(h(x_0) + \frac{h''(x_0)(x-x_0)^2}{2}\right) dx$$

$$= \exp(h(x_0)) \int_a^b \exp\left(\frac{h''(x_0)}{2}(x-x_0)^2\right) dx$$

$$= \exp\left(h(x_0)\right) \int_a^b \exp\left(-\frac{1}{2}\,\frac{(x-x_0)^2}{-h''(x_0)^{-1}}\right) dx$$

$N(x_0, -h''(x_0)^{-1})$ density

$h'' \left(= \left(\frac{g'}{g}\right)'\right)$

$\frac{gg'' - (g')^2}{g^2}$

$g \quad \frac{g''}{g} - \frac{g}{g}\frac{g'}{g}$

$g'' - \frac{g'}{g}\frac{g'}{g}$

$$= \underbrace{\exp(h(x_0))}_{= g(x_0)} \left[\frac{\sqrt{2\pi}}{-h''(x_0)^{1/2}}\right] \underbrace{\left[\Phi(b \mid x_0, -h''(x_0)^{-1}) - \Phi(a \mid x_0, -h''(x_0)^{-1})\right]}_{= 1 \text{ if } b = \infty,\ a = -\infty}$$

$$= g(x_0) \left[\frac{\sqrt{2\pi}}{-h''(x_0)^{1/2}}\right]$$

TOPIC:

DATE

FILE UNDER:

PAGE:

$$\mathbb{E}[\theta] = \frac{\int \theta f(y|\theta)\pi(\theta)}{\int}$$

$$\mathbb{E}\theta = \int \theta\, p(\theta|y)\, d\theta$$

$$= \frac{\int \theta\, f(y|\theta)\pi(\theta)\, d\theta}{\int f(y|\theta)\pi(\theta)\, d\theta}$$

$$= \frac{\int \theta \exp(\log f(y|\theta)\pi(\theta))}{\int \exp(\underbrace{\log f(y|\theta)\pi(\theta)}_{h(\theta|y)})}$$

Need $\hat{\theta}$

$$\doteq \frac{\int \theta \exp(h(\hat{\theta}|y) + h''(\hat{\theta}|y)(\theta-\hat{\theta})^2/2)\, d\theta}{\int \exp(h(\hat{\theta}|y) + h''(\hat{\theta}|y)(\theta-\hat{\theta})^2/2)\, d\theta}$$

$$= \frac{\int \theta \exp\left(\frac{h''(\hat{\theta}|y)}{2}(\theta-\hat{\theta})^2\right) d\theta}{\int \exp\left(\frac{h''(\hat{\theta})}{2}(\theta-\hat{\theta})^2\right) d\theta}$$

Laplace approx to posterior mean is the posterior mode!

$$= \frac{\int \theta \frac{\sqrt{2\pi}}{-h''(\hat{\theta})^{1/2}} \varphi(\theta|\hat{\theta}, -h''(\hat{\theta})^{-1})\, d\theta}{\int \frac{\sqrt{2\pi}}{h''(\hat{\theta})^{1/2}} \varphi(\theta|\hat{\theta}, -h''(\hat{\theta})^{-1})\, d\theta}$$

$$= \hat{\theta}$$

TOPIC:

DATE:

FILE UNDER:

PAGE:

See Lange Chap 4.6 for Laplace Approx.

Monte Carlo

Suppose we want to compute for $h: \mathbb{R}^k \to \mathbb{R}$

$$\mathbb{E}_f[h(X)] = \int h(x) f(x) \, dx$$

If we can simulate $X_1, \ldots X_n \overset{iid}{\sim} f$, then by the LLN,

$$\frac{1}{n} \sum_{i=1}^{n} h(X_i) \longrightarrow \mathbb{E}_f[h(X)]$$

sum — integral

Furthermore, $\mathrm{Var} \, \xi = \frac{1}{n^2} \sum \mathrm{Var}(h(X_i))$

~~$\mathrm{Var}(\frac{1}{n}\sum h(X_i)) = \frac{1}{n}\sqrt{\sum_{i=1}^{n} \mathrm{Var}(h(X_i))}$~~

Notice that the variance doesn't depend on dimension of X. Thus, $\mathrm{Var} \sim \frac{1}{n}$. This is very important, both good and bad.

Applications

(1) Monte Carlo / Simulation studies. We have a method that estimates a parameter $h(X) = \hat{\beta}$ and we want to explore performance of $\hat{\beta}$.

(2) $\mathbb{E}_f h(X)$ might be a posterior mean (f is post. dens)

(3) Might want $\int h(x) \, dx = \int \frac{h(x)}{f(x)} f(x) dx$.

$$\approx \frac{1}{n} \sum \frac{h(x_i)}{f(x_i)}, \; x_1, \ldots x_n \sim f$$

TOPIC

DATE

FILE UNDER

PAGE

We need to be able to simulate numbers, i.e. from some dist. f

Random Number Generation

Most popular (and simplest) are linear congruential generators.

Let X_0 be some starting value, called the "seed". Then generate a sequence (for $n = 0, 1, 2, \ldots$)

$$X_{n+1} = (a X_n + b) \bmod m$$

a = multiplier
b = increment
m = modulus

For uniform RN, just let $U_{n+1} = \frac{X_{n+1}}{m}$

Ideally, $X_1, X_2, \ldots$ will hit every number from 0 to $m-1$ before repeating

# of steps until repeat is "period".
A "maximal period generator" has period m.

Setting a, b, and m is a very tricky business.

For example, this is bad

$$X_{n+1} = (2 X_n + 0) \bmod 2^{32}$$

One good set is

$$a = 106, \quad b = 1283, \quad m = 6075$$

TOPIC:

DATE:

FILE UNDER:

PAGE:

There are quadratic and cubic RNGs, but they require more work and not much better, i.e

$$X_{n+1} = (a X_n^2 + b X_n + c) \bmod m$$

LCG are not useful for things like cryptography. Any polynomial generator can be broken.

RNG are critical for stream ciphers which

Use linear feedback shift registers (LFSR)
"bread + butter"

All PRNGs produce deterministic sequences that "look" random

One can check "randomness" of sequence with a test for uniformity
- Kolmogorov-Smirnov test
- Chi-square test

Marsaglia has "die hard" tests

Also, many PRNGs generate good 1-D sequences but do not look random in $D > 1$ dimension

#s may be stuck on a hyperplane.

TOPIC:

DATE:

FILE UNDER:

PAGE:

Uniforms are good but we need values from some density f. First, assume you can simulate $Unif(0,1)$

(*) Integral transform / rescaling

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

(2) $X \sim F^{-1}(U)$, where $F(x) = \int^{x} f(t)\,dt$ cdf.

exponential(1)

$f(x) = e^{-x/\alpha}$

$F(x) = 1 - e^{-x}$

$F^{-1}(u) = -\log(1-u)$

(*) Transformation

(1) $Z_1 = \sqrt{-2\log U_1}\,\cos(2\pi U_2)$ i.i.d $\sim N(0,1)$

$Z_2 = \sqrt{-2\log U_1}\,\sin(2\pi U_2)$

(2) $\frac{X}{X+Y} \sim Beta(\alpha,\beta)$

where $X \sim gamma(\alpha,1)$

$Y \sim gamma(\beta,1)$

(3) $\mu \sim gamma(\alpha,\beta)$

$Y|\mu \sim Poisson(\mu)$

$Y \sim$ neg. binomial

Most standard dist. are implemented in R.

$$p(y) = \int_0^\infty p(y|\mu)\,p(\mu)\,d\mu$$

Poisson gamma

TOPIC:

DATE:

FILE UNDER:

PAGE:

Multivariate Normal

Want $X \sim N(\mu, \Sigma)$

Let $\Sigma = L^T L$ (cholesky decomp)

(1) Simulate $Z \sim N(0, I)$

(2) $X = \mu + WZ \sim N(\mu, \Sigma)$

~~Sequential sampling~~

Rejection sampling

A way of generating samples/obs. from f by thinning out obs/samples from a candidate density g. ("Random thinning")

Suppose f is our target density and we can evaluate it.

Let g be our candidate density that we can simulate from.

Let $\mathcal{X}_f$ be the support of f and $\mathcal{X}_g$ be support of g and assume $\mathcal{X}_f \subset \mathcal{X}_g$

Assume $C = \sup_{x \in \mathcal{X}_f} \frac{f(x)}{g(x)} < \infty$

and that we can calculate C

TOPIC

DATE

FILE UNDER:

PAGE:

Rejection Sampling

(1) Simulate $u \sim \text{uniform}(0,1)$

(2) Simulate $X \sim g(x)$

(3) If $u \leq \frac{f(x)}{cg(x)}$, "accept" X

otherwise, discard X

go back to (1).

Let $X_1, X_2, \ldots \overset{iid}{\sim} g$

X_1	X_2	X_3	X_4	X_5	X_6	X_7	.....
0	0	1	0	1	0	0	.....
reject	reject	accept		accept			

geometric(p)

iid coin flips Sample size of 2.

# of flips until acceptance is geometric w/ success probability $1/c$.

Pf: $\mathbb{P}(X \text{ accepted}) = \mathbb{P}\left(u \leq \frac{f(x)}{cg(x)}\right)$

$= \int \mathbb{P}\left(u \leq \frac{f(x)}{cg(x)} \mid X = x\right) g(x)\, dx$

$= \int \frac{f(x)}{cg(x)} g(x)\, dx$

$= 1/c$

TOPIC: DATE:

FILE UNDER: PAGE:

The dist. of accepted values is f.

$$\mathbb{P}(X \le t \mid X \text{ accepted}) = \frac{\mathbb{P}(X \le t, X \text{ accepted})}{\mathbb{P}(X \text{ accepted})}$$

$$= \frac{\mathbb{E}_g\left[\mathbb{1}\{X \le t\}\, \mathbb{1}\{X \text{ accepted}\}\right]}{1/c}$$

$$= c\,\mathbb{E}_g\,\mathbb{E}_u\left[\mathbb{1}\{X \le t\}\, \mathbb{1}\left\{u \le \frac{f(x)}{c g(x)}\right\} \,\middle|\, X\right]$$

$$= c\,\mathbb{E}_g\left[\mathbb{1}\{X \le t\}\, \frac{f(x)}{c g(x)}\right]$$

$$= c \int \mathbb{1}\{x \le t\}\, \frac{f(x)}{c g(x)}\, g(x)\, dx$$

$$= \int_{-\infty}^{t} f(x)\, dx = F(t)$$

Notes:

1. Only need to ~~know~~ f or g up to a constant of proportionality
2. Any number $c' > c$ will work, but will be less efficient.
3. Operations can (and should) be performed on log scale

TOPIC:

DATE:

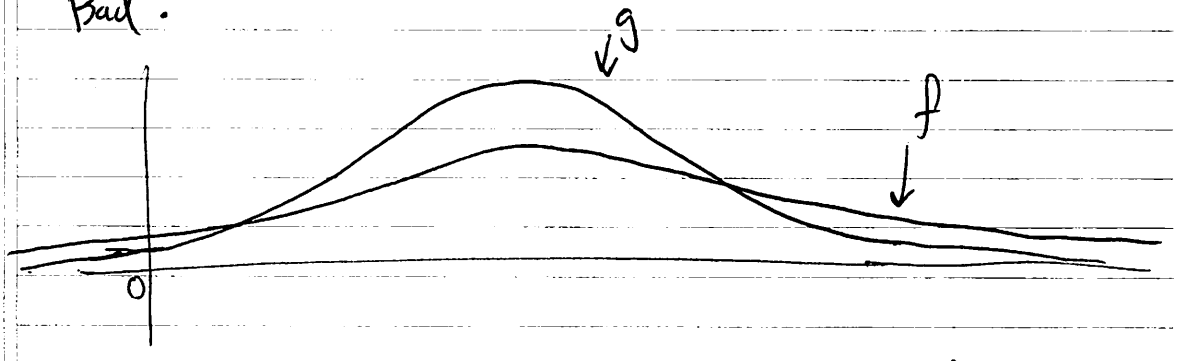
FILE UNDER:

PAGE:

④ The higher the dimension of f, g the more inefficient rejection sampling will be.

⑤ Whether $C = \infty$ depends (usually) on the tails of the candidate, which must be heavier than the target

Bad:



As $x \to \infty$, $g(x) \downarrow 0$ faster than $f(x) \downarrow 0$.

$\Rightarrow f(x)/g(x) \uparrow \infty$.

$\Longrightarrow$ What if we cannot calculate $C = \sup \frac{f(x)}{g(x)}$?

Can we estimate C? Yes!

Empirical sup rejection sampling (Capfo '02) Brewster

① Guess $\hat{C} > 1$

② Generate $U \sim \text{uniform}(0,1)$, $X \sim g$

③ Accept X if $U \leq \frac{f(x)}{\hat{C} g(x)}$

④ Update $\hat{C}^* = \max\left\{\hat{C}, \frac{f(x)}{g(x)}\right\}$ $\Longleftrightarrow$ sup estimator

⑤ Go to step ②, set $\hat{C} = \hat{C}^*$

$h(x) = \frac{f(x)}{g(x)}$

$h(x_{n+1}), h(x_{n-1})$

$x_{n+1} = x_n - \alpha h'(x_n)$

TOPIC:

DATE:

FILE UNDER:

PAGE:

ESUP

Extra assumption is need for ~~ESUP~~:

$C = \frac{f(x_e)}{g(x_e)}$ for some $x_e \in \mathcal{X}_f$

& (sup is achievable)

This is satisfied if g has heavier tails than f.

~~Caffo proved that~~

~~If~~ let $X_1, X_2, X_3, \ldots \overset{iid}{\sim} g$

	X_1	X_2	X_3	X_4	X_5	X_6	...
True $\tilde{Y}_i$	0	0	1 (accept)	0	1 (accept)	0	
$\hat{C}$ Y_i	1 (accept, error)	0	1 (accept)	1 (accept, error)	1 (accept)	0	

Caffo showed that

① $\mathbb{P}(Y_i \neq \tilde{Y}_i \text{ i.o.}) = 0$ if f is discrete

Pf: By assumption 3, $\exists\, x_e \in \mathcal{X}_f$ s.t. $C = \frac{f(x_e)}{g(x_e)}$.

Let $\gamma = \min_i \{X_i = x_e\}$ where $X_i \sim g$.

Then $\gamma \sim \text{geometric}(g(x_e))$. Once $\hat{C} = C$, algorithms are the same.

Coupling lemma $\Rightarrow$ $\mathbb{P}(Y_i \neq \tilde{Y}_i) \leq \mathbb{P}(\gamma \geq i) = (1 - g(x_e))^{i-1}$

$\Rightarrow \sum_{i=1}^{\infty} \mathbb{P}(Y_i \neq \tilde{Y}_i) < \infty$

TOPIC: DATE

FILE UNDER: PAGE:

(2) For continuous f, it's trickier. In general,

$$\mathbb{P}(Y_i \neq \tilde{Y}_i) = O(i^{-1})$$

But if $\log(f/g)$ is smooth (twice differentiable) and unimodal (w/mode at x_c), then

$$\mathbb{P}(Y_i \neq \tilde{Y}_i) = O(i^{-2})$$

Suggests a "burn-in", discard first several variates.

More on burn-in with MCMC

Often we want to simulate from

$$p(\theta|y) \propto \underbrace{f(y|\theta)}_{\text{likelihood}} \underbrace{\pi(\theta)}_{\text{prior}}$$

$\Rightarrow$ use π as the candidate dist.

$$C = \sup_\theta \frac{p(\theta|y)}{\pi(\theta)} = \sup_\theta f(y|\theta) = f(y|\hat{\theta})$$

~~posterior mode~~ MLE

Rejection sampling is

(1) Simulate $U \sim \text{uniform}(0,1)$
(2) Simulate $\theta \sim \pi$
(3) Accept θ if $U \leq \frac{p(\theta|y)}{C\pi(\theta)} = \frac{f(y|\theta)}{f(y|\hat{\theta})}$

TOPIC

DATE

FILE UNDER

PAGE

Ex: $Y_1, \ldots Y_n \overset{iid}{\sim} N(\mu, \sigma^2)$

Prior: $\pi(\mu) = N(0, B)$ → BIG

$$\pi(\sigma^2) = IG(\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} z^{-\alpha-1} e^{-\beta/z} = \beta(\alpha)$$

MLE: $\hat{\mu} = \bar{Y}$, $\hat{\sigma}^2 = S^2 = \frac{1}{n}\sum(Y_i - \bar{Y})^2$

$$C = f(y|\hat{\mu}, \hat{\sigma}^2) = \frac{1}{\sqrt{2\pi}\hat{\sigma}^2} e^{-\frac{1}{2\hat{\sigma}^2}\sum(Y_i - \hat{\mu})^2}$$

$$= \left(\frac{1}{\sqrt{2\pi}}\right)^n \left(\frac{\sum(Y_i - \bar{Y})^2}{n}\right)^{-n/2} e^{-\frac{n}{2}}$$

(1) Simulate $U \sim \text{Uniform}(0,1)$

(2) Simulate $\mu \sim N(0, B)$, $\frac{1}{\sigma^2} \sim \text{gamma}(\alpha, \beta)$

(3) Accept (μ, σ^2) if

$$U \leq \frac{f(y|\mu,\sigma^2)}{f(y|\hat{\mu},\hat{\sigma}^2)} = \left(\frac{\sum(Y_i - \bar{Y})^2}{n\sigma^2}\right)^{n/2} \exp\left(\frac{-1}{2\sigma^2}\sum(Y_i - \mu)^2 + \frac{n}{2}\right)$$

Given a sample from $p(\theta|y)$ we can compute confidence/posterior intervals, mean, variance, etc.

What does ESUP do? It searches for MLE by random search.

TOPIC:

DATE:

FILE UNDER:

PAGE:

(X가 IIF [X|Y])

It is MCMC, but not Gibbs. exactly

Data Augmentation (Tanner + Wong '87)

We have some data $Y_1, \dots, Y_n$, and we propose a model

$$Y \sim p(y|\theta)$$

$$\theta \sim \pi(\theta)$$

So we have a likelihood $p(y|\theta)$ and a prior $\pi(\theta)$.

We want the posterior $p(\theta|y)$.

$$p(\theta|y) \propto p(y|\theta)\pi(\theta).$$

BUT because of some missing data z, it is difficult to evaluate $p(\theta|y)$.

Let (y, z) be the "complete data" and we want to know

$$p(\theta|y) = \int p(\theta|y,z)\, p(z|y)\, dz \qquad (1)$$

$p(\theta|y,z)$: "complete data posterior" → easy to calculate

$p(z|y)$: predictive density for z.

System of integral equations

Notice also that we have

$$p(z|y) = \int p(z|y,\theta)\, p(\theta|y)\, d\theta \qquad (2)$$

($p(z|y,\theta)$: easy to calculate)

Substituting (2) into (1), we get

$$p(\theta|y) = \int p(\theta|y,z) \left[\int p(z|y,\theta')\, p(\theta'|y)\, d\theta' \right] dz$$

$$= \int \left[\int p(\theta|y,z)\, p(z|y,\theta')\, dz \right] p(\theta'|y)\, d\theta'$$

TOPIC

DATE

FILE UNDER:

PAGE

DA algorithm

Roughly: At iteration i

(1) Sample $z_1, \dots z_m \sim p_i(z|y)$

(2) Estimate $p_{i+1}(\theta|y) = \frac{1}{m}\sum_{i=1}^{m} p(\theta|y, z_i)$

(complete data posterior)

Repeat until $\|p_{i+1}(\theta|y) - p_i(\theta|y)\| < \varepsilon$.

---

(1) Pick some initial $p_0(\theta|y)$. & At step i

(2a) Generate $\theta \sim p_i(\theta|y)$

(2b) Generate $z \sim p(z|y, \theta)$

Repeat (2a) – (2b) m times to get $z_1, \dots z_m$.

(3) Let $p_{i+1}(\theta|y) = \frac{1}{m}\sum p(\theta|y, z_i)$

(mixture of conditional densities.)

p_{i+1} ↙

Monte Carlo estimate of

$$\int p(\theta|y, z)\, p(z|y)\, dz$$

DA not so straightforward when there are more than 2 "missing" components.

Use Gibbs Sampling for more complicated problems

TOPIC

DATE:

FILE UNDER

PAGE:

Let $K(\theta, \theta') = \int p(\theta|y,z)\, p(z|y,\theta')\, dz$

$$p(\theta|y) = \int K(\theta,\theta')\, p(\theta'|y)\, d\theta'$$

(1) (2)

(fixed point system) If $p(\theta'|y)$ is the ~~If (1) and (2) are~~ true posterior, then it will map to itself.

Let T be a functional which maps a function $g(\theta)$ to

$$Tg(\theta) = \int K(\theta,\theta')\, g(\theta')\, d\theta'$$

If we take some initial value $p_0(\theta|y)$, what happens when we iterate $T p_0(\theta|y)$ ie.

$$p_1(\theta|y) = T\, p_0(\theta|y)$$

$$= \int K(\theta,\theta')\, p_0(\theta'|y)\, d\theta'.$$

Does the sequence $\{p_i(\theta|y)\}$ converge to anything?

~~(1) ploty) the true posterior is the~~

(1) $\{p_i(\theta|y)\} \longrightarrow p(\theta|y)$ monotonically (always getting closer)
(true posterior)

(2) $p(\theta|y)$ is the unique solution to system of integral equations

(3) $p_i(\theta|y) \longrightarrow p(\theta|y)$ linearly.

TOPIC:

DATE:

FILE UNDER:

PAGE:

Ex. 3-stage hierarchical model

$Y|\alpha \sim N(\alpha, 1)$

$\alpha|\theta \sim N(\theta, 1)$

$\theta \sim \pi(\theta) = N(0, 1)$

$p(Y, \alpha, \theta) = p(Y|\alpha,\theta)\, p(\alpha|\theta)\, \pi(\theta)$

$N(\alpha, 1) \quad N(\theta, 1) \quad N(0, 1)$

$p(\alpha|Y,\theta) \propto p(Y|\alpha,\theta)\, p(\alpha|\theta)$

$= N\left(Y + \frac{1}{2}(\theta - Y),\ \frac{1}{2}\right)$

$p(\theta|Y,\alpha) \propto p(Y|\alpha,\theta)\, p(\alpha|\theta)\, \pi(\theta)$

$N(\quad)$

(1) find some $p_i(\theta|y)$ [Normal?]

(2a) sample $\theta \sim p_i(\theta|y)$

(2b) Sample $\alpha \sim p(\alpha|y,\theta)$

Repeat m times to get $\alpha_1 \rightarrow \alpha_m$

(3) Let $p_{i+1}(\theta|y) = \frac{1}{m}\sum p_i(\theta|y,\alpha)$

TOPIC:

DATE:

FILE UNDER:

PAGE:

	EM	DA
Paradigm	likelihood	Bayesian
Target	MLE	Posterior dist.
Method to Handle Z	Average (conditional)	sample (conditional)
Need (Z|Y)	Compute E	sample from
Std. errors	No, by default	Yes, have entire posterior (need prior)
Rate of convergence	linear	linear
Monotonicity	✓	✓

Both methods "impute" missing data via the specification of the complete data model.

Once $p(Y, Z)$ is specified, everything else is determined.

TOPIC

DATE

FILE UNDER

PAGE

$$\hat{\mu}_n = \frac{\sum \frac{f(X_i)}{g(X_i)} h(X_i)}{\sum \frac{f(X_i)}{g(X_i)}} = \frac{\sum w(X_i) h(X_i)}{\sum w(X_i)}$$

With rejection sampling, we can sample from f given a candidate density g.

What if we want to estimate $\mathbb{E}_f[h(X)]$ for some $h: \mathbb{R}^k \to \mathbb{R}$?

Obvious way: Sample $X_1, \dots X_n \sim f$ and use

$$\mathbb{E}_f[h(X)] \approx \frac{1}{n} \sum_{i=1}^{n} h(X_i) = \hat{\mu}_n$$

$$\sqrt{n}(\hat{\mu}_n - \mu_n) \rightsquigarrow N(0, \sigma^2)$$

But sampling from f is hard, so we use RS with cand. dens. g.

In order to obtain sample of size n, we need on avg, $C \times n$ samples from g, where

$$C = \sup f/g \quad (\text{assumed} < \infty)$$

We reject (on avg.) $(C-1) \times n$ of the samples from g!

Those samples belong in the domain of f, but they may be over/under-represented

eg. if g has heavier tails, there will be too many extreme values — RS rejects those.

But maybe we can down/up weight values to get the right answer.

TOPIC:

DATE:

FILE UNDER:

PAGE:

Note that

$$\mathbb{E}_f[h(x)] = \mathbb{E}_g\left[\frac{f(x)}{g(x)} h(x)\right].$$

If $X_1 \rightarrow X_n \sim g$ then

$$\mathbb{E}_f[h(x)] \approx \frac{1}{n} \sum_{i=1}^{n} \frac{f(x_i)}{g(x_i)} h(x_i)$$

$$= \frac{1}{n} \sum w_i \, h(x_i) = \tilde{\mu}_h$$

$\downarrow$

importance weights

No rejection sampling here

$\Rightarrow$ Assumes we can compute f and g exactly.

Notice that if $f = g$, our estimator is just

$$\frac{1}{n} \sum_{i=1}^{n} h(x_i).$$

If the density of X_i w.r.t g is larger than its density w.r.t f, then $h(x_i)$ is down weighted in the sum (vice versa)

Comparison

Rejection Sampling: Sample directly from f, take averages

Importance Sampling: Sample from g, reweight by f/g

For computing expectations IS is much more efficient because there is no rejection

TOPIC: DATE:

FILE UNDER: PAGE:

~~What if we only Know~~

RS: Let $c = \sup f/g$. Then

sample $X_1 \rightarrow X_n \sim g$ and $U_1 \rightarrow U_n \sim \text{unif}(0,1)$.

$$\hat{\mu}_h = \frac{\sum \mathbb{1}\{U_i \le \frac{f(X_i)}{cg(X_i)}\} h(X_i)}{\sum \mathbb{1}\{U_i \le \frac{f(X_i)}{cg(X_i)}\}}$$

IS: Sample $X_1 \rightarrow X_n \sim g$

$$\tilde{\mu}_h = \frac{1}{n}\sum \frac{f(X_i)}{g(X_i)} h(X_i)$$

(1) IS estimator ($\tilde{\mu}_h$) is "smoother" than $\hat{\mu}_h$ and should have lower variance

(2) RS requires $\sup f/g < \infty$. But IS does not need c.

(3) IS computations cannot be done on log scale

What if we only know $f^* = c_1 f$ and $g^* = c_2 g$. Then use

$$\hat{\mu}_h^* = \frac{\frac{1}{n}\sum_{i=1}^n \frac{f^*(X_i)}{g^*(X_i)} h(X_i)}{\frac{1}{n}\sum_{i=1}^n \frac{f^*(X_i)}{g^*(X_i)}} \quad \text{(ratio estimator)}$$

Since f^* and g^* are in numerator and denominator, constants drop out.

$\hat{\mu}_h^* \rightarrow \mu_h$ by Slutsky theorem

TOPIC:

DATE:

FILE UNDER:

PAGE:

Ex. Importance Sampling for Bayesian sensitivity analysis.

We have data y with a likelihood $L(y|\theta)$ and a prior for θ $\pi(\theta|\psi_0)$ where ψ_0 is a known hyperparameter. The posterior for θ is

$$p(\theta|y,\psi_0) \propto L(y|\theta)\pi(\theta|\psi_0).$$

Suppose we expend much energy obtaining $\theta_1, \dots, \theta_n$, a sample from $p(\theta|y,\psi_0)$, and we can compute posterior mean $\mathbb{E}[\theta] = \frac{1}{n}\sum \theta_i$. What if we want to calculate different values of $\mathbb{E}[\theta|y,\psi]$? We do not need to resample $\theta_1, \dots, \theta_n$, just reweight them. Let ψ be the new hyperparameter, $f = p(\theta|y,\psi)$, $g = p(\theta|y,\psi_0)$. (have a sample from this)

$$\mathbb{E}[\theta|y,\psi] \approx \frac{\sum \theta_i \frac{f(\theta_i)}{g(\theta_i)}}{\sum \frac{f(\theta_i)}{g(\theta_i)}} = \frac{\sum \theta_i \frac{p(\theta_i|y,\psi)}{p(\theta_i|y,\psi_0)}}{\sum \frac{p(\theta_i|y,\psi)}{p(\theta_i|y,\psi_0)}}$$

$$= \frac{\sum \theta_i \frac{L(y|\theta)\pi(\theta|\psi)}{L(y|\theta)\pi(\theta|\psi_0)}}{\sum \frac{L(y|\theta)\pi(\theta|\psi)}{L(y|\theta)\pi(\theta|\psi_0)}} = \frac{\sum \theta_i \frac{\pi(\theta|\psi)}{\pi(\theta|\psi_0)}}{\sum \frac{\pi(\theta|\psi)}{\pi(\theta|\psi_0)}}$$

$\mathbb{E}[\theta|y,\psi] = \int \theta \, p(\theta|y,\psi_0) \, d\theta$ / $p(\theta|y,\psi_0)$ / $p(\theta|y,\psi)$

TOPIC

DATE

FILE UNDER:

PAGE:

Ex. Calculating Marginal likelihoods.

Suppose we have $f(y|u)$, the dist of y given some random effect u and $h(u|\theta)$ the dist. of random effects for parameter θ.

If

$$Y_{ij} \sim N(\mu + u_i, \sigma^2)$$

$$u_i \sim N(0, \theta)$$

We want to maximize

$$L(\theta) = \int f(y|u)\, h(u|\theta)\, du = E_h[f(y|u)]$$

↳ integrate out random effects.

Suppose we simulate $u_1, \ldots, u_n$ from a candidate dist $h(u|\theta_0)$. Then

(wrong dist)

$$\hat{L}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \frac{h(u_i|\theta)}{h(u_i|\theta_0)} f(y|u_i)$$

Could use a more general candidate dist $g(u|\theta_0)$.

$$\hat{L}(\theta) = \frac{1}{n} \sum \frac{h(u_i|\theta)}{g(u_i|\theta_0)} f(y|u_i)$$

$$g(u|\theta) = f(y|u_i)\, h(u|\hat{\theta})$$

Marginal likelihood

Ideal candidate is proportional to $f(y|u)\,h(u|\theta)$ which is $p(u|y,\theta)$. Since we want to maximize $\hat{L}$, the best candidate is $p(u|y,\hat{\theta})$ where $\hat{\theta}$ maximizes $\hat{L}$. But that solves our problem. So try (Geyer, 1990)

(1) Simulate $\{u_i\}$ from $p(u|y,\theta_0)$

(2) Max $\hat{L}$ to get θ_1

(3) Set $\theta_0 = \theta_1$, Goto (1)

MCEM

MCEM ("Ascent-based") (Caffo et al 2005)

Given Y, z, the E-step computes

$$Q_t = \mathbb{E}_h\left[\log g(Y, z|\theta) \,\middle|\, Y, \theta_0\right]$$

$h(z|Y,\theta_0) \longrightarrow$ simulate from using rejection sampling

$$\tilde{Q}_t \approx \frac{1}{M_t}\sum_{i=1}^{M_t} \log g(Y, z_i|\theta) \xrightarrow{LLN} Q_t$$

as $M_t \to \infty$

How do we know that θ^* which max $\tilde{Q}$ has the ascent property, $Q(\theta^*|\theta_0^*) \geq Q(\theta_0^*|\theta_0^*)$?

Let $\Delta Q = Q(\theta^*|\theta_0^*) - Q(\theta_0^*|\theta_0^*)$

$$\approx \Delta\tilde{Q} = \tilde{Q}(\theta^*|\theta_0^*) - \tilde{Q}(\theta^*|\theta_0^*)$$

$\sqrt{M_t}(\Delta\tilde{Q} - \Delta Q) \sim N(0, \sigma^2)$ [Caffo shows]

$\hat{\sigma}^2$ depends on IS or RS.

Updating the Sample Size

$$\Delta \tilde{Q}\left(\hat{\theta}^{*}_{t+1}, \theta^{*}_{t}\right) \sim N\left(\Delta Q(\theta_{t+1} | \theta^{0}_{1}) \frac{\hat{\tau}^{2}}{M_{t+1}}\right)$$

$$M_{t+1} = \frac{\hat{\sigma}^{2}(z_{\alpha} + z_{\beta})^{2}}{\Delta \tilde{Q}\left(\theta^{*}_{t} | \theta^{*}_{t-1}\right)^{2}}$$

TOPIC:

DATE:

FILE UNDER:

PAGE:

For IS we do not need $\sup f/g < \infty$ but what is required?

Recall Cramer's Theorem (Delta Method)

Let $Y_1 \to Y_n$ be s.t. $\sqrt{n}(\bar{Y} - \mu) \to N(0, \Sigma)$

Where $\mu = \mathbb{E}[\bar{Y}_i]$. Let $g: \mathbb{R}^k \to \mathbb{R}$ be a differentiable map, then

$$\sqrt{n}(g(\bar{Y}) - g(\mu)) \to N(0, g'(\mu)^T \Sigma g'(\mu))$$

Importance Sampling

Let $Y_i = \begin{pmatrix} h(X_i) w_i \\ w_i \end{pmatrix}$ where $X_i \sim g$, $w_i = \frac{f(X_i)}{g(X_i)}$

So $\mathbb{E}_g Y_i = \begin{pmatrix} \mathbb{E} h(X_i) \\ 1 \end{pmatrix} = \begin{pmatrix} \mu_h \\ 1 \end{pmatrix}$.

Note that we can estimate $\mathrm{Var}(Y_i)$ Consistently with

$$\hat{\Sigma} = \begin{bmatrix} \text{Sample Var}\{h(X_i)w_i\} & \text{Sample Cov}\{h(X_i)w_i, w_i\} \\ \text{Sample Cov}\{h(X_i)w_i, w_i\} & \text{Sample Var}\{w_i\} \end{bmatrix}$$

Let $V(a,b) = \frac{a}{b}$, Then $\hat{\mu}_h = V\left(\frac{1}{n}\sum h(X_i)w_i, \frac{1}{n}\sum w_i\right)$.

Also $V'(a,b) = \left(\frac{1}{b}, -\frac{a}{b^2}\right)$

TOPIC

DATE

FILE UNDER

PAGE

Variance estimates for IS

Cramer's Theorem: Suppose $Y_1, \dots Y_n$ are s.t. $\sqrt{n}(\bar{Y} - \mu) \longrightarrow N(0, \Sigma)$ where $\mu = \mathbb{E}\bar{Y}$. Let $g: \mathbb{R}^K \to \mathbb{R}$ be a differentiable map. Then

$$\sqrt{n}(g(\bar{Y}) - g(\mu)) \longrightarrow N(0, g'(\mu)^T \Sigma g'(\mu))$$

f is target density

g is proposal density

$$\mu_h = \mathbb{E}_f h(X)$$

Let $Y_i = \begin{pmatrix} h(x_i) w_i \\ w_i \end{pmatrix}$ where $X_i \sim g$ and $w_i = \frac{f(x_i)}{g(x_i)}$

$$\mathbb{E}_g Y_i = \begin{pmatrix} \mathbb{E}_g\left[h(x_i)\frac{f(x_i)}{g(x_i)}\right] \\ \mathbb{E}_g\left[\frac{f(x_i)}{g(x_i)}\right] \end{pmatrix} = \begin{pmatrix} \mu_h \\ 1 \end{pmatrix}$$

TOPIC:

DATE:

FILE UNDER:

PAGE:

Let $m\begin{pmatrix} a \\ b \end{pmatrix} = \frac{a}{b}$. $m'\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 1/b \\ -a/b^2 \end{pmatrix}$

$$m(\bar{Y}) = m\begin{pmatrix} \frac{1}{n}\sum h(x_i)w_i \\ \frac{1}{n}\sum w_i \end{pmatrix} = \frac{\frac{1}{n}\sum h(x_i)w_i}{\frac{1}{n}\sum w_i}$$

Note $\mathbb{E}_g Y_i = \begin{pmatrix} \mathbb{E}_g\left[h(x_i)\frac{f(x_i)}{g(x_i)}\right] \\ \mathbb{E}_g\left[\frac{f(x_i)}{g(x_i)}\right] \end{pmatrix} = \begin{pmatrix} \mathbb{E}_f[h(x)] \\ 1 \end{pmatrix}$

$$= \begin{pmatrix} \mu_h \\ 1 \end{pmatrix}$$

$\mathbb{E}\bar{Y} = \begin{pmatrix} \mu_h \\ 1 \end{pmatrix}$

$m(\mathbb{E}\bar{Y}) = \mu_h$

We can estimate $\mathrm{Var}(Y_i)$ constantly w.M

$$\hat{\Sigma} = \begin{pmatrix} \hat{\mathrm{Var}}\{h(x_i)w_i\} & \hat{\mathrm{Cov}}\{h(x_i)w_i, w_i\} \\ \hat{\mathrm{Cov}}\{h(x_i)w_i, w_i\} & \hat{\mathrm{Var}}\{w_i\} \end{pmatrix}$$

Cramer's Thm says:

$$\sqrt{n}\left(\frac{m(\bar{Y}) - \mu_h}{m'(\bar{Y})^T \hat{\Sigma}\, m'(\bar{Y})}\right) \longrightarrow N(0,1)$$

↳

TOPIC:

DATE:

FILE UNDER:

PAGE:

$\mathbb{E}X^2 - (\mathbb{E}X)^2 > 0$

$(\mathbb{E}X)^2 \leq \mathbb{E}X^2$

Cramer's Theorem says:

$$\sqrt{n}\left(\frac{V(\bar{Y}) - \mu_h}{V'(\bar{Y})^T \hat{\Sigma} V'(\bar{Y})}\right) \longrightarrow N(0,1)$$

And

$$V'(\bar{Y})^T \hat{\Sigma} V'(\bar{Y}) =$$

$$n\left(\frac{\sum h(x_i)w_i}{\sum w_i}\right)^2 \left(\frac{\sum h(x_i)^2 w_i^2}{(\sum h(x_i)w_i)^2} - 2\frac{\sum h(x_i)w_i^2}{(\sum h(x_i)w_i)(\sum w_i)} + \frac{\sum w_i^2}{(\sum w_i)^2}\right)$$

$\Rightarrow$ We need

$$\mathbb{E}_g[h(x)^2 w^2] = \mathbb{E}\left[h(x)^2 \frac{f(x)^2}{g(x)^2}\right] < \infty$$

$$\mathbb{E}[w^2] = \mathbb{E}\left[\left(\frac{f(x)}{g(x)}\right)^2\right] < \infty$$

$$\mathbb{E}[h(x)w^2] = \mathbb{E}\left[h(x)\left(\frac{f(x)}{g(x)}\right)^2\right] < \infty$$

All true if $\frac{f(x)}{g(x)}$ is bounded, which is required for RS.

Close to $\hat{\theta}$ $\sum_{i=0}^{\infty} a^i = \frac{1}{1-a}$ $\quad Q'(\theta|\theta_t) = 0$

$\theta_{t+1} = g(\theta_t) = \arg\max_{\theta} Q(\theta|\theta_t)$

$g(\theta_t) \doteq g(\theta_{t-1}) + J(\theta_{t-1})(\theta_t - \theta_{t-1})$

$\theta_{t+1} \doteq \theta_t + J(\theta_{t-1})(\theta_t - \theta_{t-1})$

$d_{t+1} \doteq J(\theta_{t-1}) d_t$

$d_{t+2} \doteq J(\theta_t) d_{t+1} = J(\theta_t) J(\theta_{t+1}) d_t$

$d_{t+j+1} \doteq J_t^t(\hat{\theta}) d_{j+1}$

$\hat{\theta} = \theta_j + \sum_{l=1}^{\infty} d_{l+j} = \theta_j + (\theta_{j+1} - \theta_j) + (\theta_{j+2} - \theta_{j+1}) + \cdots$

$= \theta_j + \left(\sum_{l=0}^{\infty} J_l(\hat{\theta})\right) d_{j+1}$ → Eigenvalues of $J(\theta)$ are all < 1

$\theta_{j+1}^* = \theta_j + (1 - J(\hat{\theta}))^{-1} d_{j+1}$ (Aitken acceleration)

$$(1 - J(\hat{\theta}))^{-1} = I_X I_Y^{-1}$$

$$\hookrightarrow -Q''(\hat{\theta}|\hat{\theta})$$

~~I_X I_Y~~

$$I_Y = I_X - \left(\mathbb{E}\left[s(x|\hat{\theta})^* s(x|\hat{\theta})'\right] - s(y|\hat{\theta}) s(y|\hat{\theta})'\right)$$

Non-monotone, numerical instability, difficult?

SQUAREM

Super linear Convergence?

1. $\theta_1 = g(\theta_0)$
2. $\theta_2 = g(\theta_1)$
3. $r = \theta_1 - \theta_0$
4. $v = \theta_2 - \theta_1 - r$
5. Compute $\alpha = -\frac{\|r\|}{\|v\|}$
6. $\theta' = \theta_0 - 2\alpha r + \alpha^2 v$
7. $\theta_0 = g(\theta')$
8. Converge? Goto 1