Summary of statComp

written by Haodong with Math

This is a brief summary for DATA130004, statistcal computing in the school of data science, Fudan University, fall 2021. The summary contains only the logic flow and the esssential parts of the course, and the "importance" is judged by how well I remember. Some details might be ignored. Youu are suppose to refer to this summary for previewing or reviewing for the course or other similar fields. You are not supposed to use it as text book or lecture note record word by word.

The course focus on applying statistical methods on computer, especially for R.

1. generate random variables **2**

1.1. inverse transform 2

1.2. acceptance-rejection method 2

2. Monte Carlo integration and variance reduction **3**

2.1. Simple MC integration 3

2.2. Variance Reduction 4

2.2.1. antithetic variable 4

2.2.2. control variate 4

2.2.3. Antethetic variables as a control variate 5

2.2.4. Control variate and linear regression 5

2.2.5. importance sampling 6

2.2.6. stratified sampling 7

3. MC in statistical inference **9**

3.1. point estimation 9

3.2. confidence interval 9

3.3. hypothesis testing 10

3.3.1. empirical type I error rate 10

3.3.2. Power of a test 10

4. Bootstrap **10**

4.1. point estimation 11

4.2. confidence interval 11

4.2.1. standard normal distribution 11

4.2.2. percentile CI 11

4.2.3. Basic bootstrap CI 11

4.2.4. Bootstrap t CI 12

5. Jacknife **12**

5.1. Bias 12

5.2. Standard Error 13

6. Bayesian statistics and MCMC **13**

6.1. Bayesian problem set-up 13

6.2. Markov chain Monte Carlo 14

6.2.1. Metropolis-Hasting sampler 14
6.2.2. Metropolis sampler 15
6.2.3. Random Walk sampler 15
6.2.4. independent sampler 15
6.2.5. Gibbs sampler 15
6.3. Monitoring the convergence (Gelman-Rubin method) 16
7. EM algorithm 17
8. Variational inference 17
8.1. KL divergence 18
8.2. Evidence lower bound (ELBO) 18
8.3. The mean-field variational family 19

1. generate random variables

In this part, we talk about the methods for generating random numbers. For some certain distributions, like normal distribution, we need to think of how could we use computer to generate numbers from the distributuion we want. This is something you should bare in mind, one of the key idea for this course is to think from the view of a computer programmer.

1.1. inverse transform

You might learned from your probability course one interesting theorem, which you don't know what for in a long time, that is,

Theorem 1 *Probability Integral Transformation*
If X is a continuous random variables with cdf $F_X(x)$*, then* $U = F_X(X) \sim Uniform(0,1)$.

Thus, given a uniform random number generator and a cdf, we can get the random number using the inverse of the cdf.

1. Derive the inverse function $F_X^{-1}(u)$.
2. Generate $u \sim Unif(0,1)$.
3. $x = F^{-1}(u)$

This method requires the F^{-1} easy to compute.

1.2. acceptance-rejection method

In this method, given a generator with pdf $Y \sim g(t)$, we can generate from our target $X \sim f(t)$. We propose the algorithm, then we prove it.

First we assume that c satisfies $\frac{f(t)}{g(t)} \leq c$ for all $t \in \mathbb{R}$, then

1. Generate $y \sim g(t)$, $u \sim Unif(0,1)$.
2. If $u < \frac{f(y)}{cg(y)}$ then accept y and set $x = y$

Now we prove x is actually from target f.

$$\begin{aligned} P(X = y|accept) &= P(Y = y|accept) \\ &= \frac{P(accept|Y = y)P(Y = y)}{P(accept)} \end{aligned}$$

where

$$\begin{aligned} P(accept|Y = y) &= P\left(u < \frac{f(y)}{cg(y)} \middle| Y = y\right) \\ &= \frac{f(y)}{cg(y)} \end{aligned}$$

$$\begin{aligned} P(accept) &= \sum_y P(accept|Y = y)P(Y = y) \\ &= \sum_y \frac{f(y)}{cg(y)} \cdot g(y) = \frac{1}{c} \end{aligned}$$

Thus

$$P(X = y|accept) = \frac{\frac{f(y)}{cg(y)} \cdot g(y)}{\frac{1}{c}} = f(y)$$

The continuous case is shown in the homework.

2. Monte Carlo integration and variance reduction

In this part, we solve the integration problem $\theta = \int g(x)dx$.

For $x_1, \cdots, x_m \overset{iid}{\sim} X$, the empirical averge $\mathbb{E}_M[g(x)] = \frac{1}{m}\sum g(x_i)$ estimates the population mean, i.e., the expectaion, $\mathbb{E}g(x) = \int g(x)f(x)dx,\ X \sim f$. Thus we introduce the Mote Carlo integration.

The Monte Carlo method is a widely used method. When saying MC, we are simulating many times to estimate the target. For example, in this part, for estimate the integral, we simulate many random numbers to approximate the result.

2.1. Simple MC integration

To estimate $\theta = \int_0^1 g(x)dx, X \sim Unif(0, 1)$.

1. Generate $x_1, \cdots, x_m \overset{iid}{\sim} X$

2. $\widehat{\theta} = \frac{1}{m}\sum_{i=1}^{m} g(x_i)$.

Notice that the integration is limited to the range 0 to 1. We can generalize to range $[a, b]$ by change of variable $y = \frac{x-a}{b-a}$, or just generate $X \sim Unif(a, b)$, and $\widehat{\theta} = (b-a)\frac{1}{m}\sum_{i=1}^{m} g(x_i)$.

2.2. Variance Reduction

Although the simple MC integration is unbiased eastimator of θ, we can use better estimators with smaller variance.

2.2.1. antithetic variable

In simple MC, we use independent random variables. However, with usage of dependent variables, we might reduce the variables.

Suppose Y and Z have same distribution with X but dependent.

$Var\left(\frac{1}{2}(Y+Z)\right) = \frac{1}{4}\left\{Var(Y) + Var(Z) + 2Cov(Y, Z)\right\}$, if $Cov(Y, Z) < 0$, the variance can be reduced.

Known that if $U \sim Unif(0, 1)$, then $1 - U \sim Unif(0, 1)$ and U and $1 - U$ are negatively correlated. And we can expect that

Corollary 1

If $g(X) = g(X_1, \cdots, X_n)$ *is monotone, then* $Y = g\left(F_X^{-1}(u_1), \cdots, F_X^{-1}(u_n)\right)$ *and* $Y = g\left(F_X^{-1}(1-u_1), \cdots, F_X^{-1}(1-u_n)\right)$ *are negatively correlated.*

Proof is ignored here.

Then, instead of generating m $Unif(0, 1)$ random varables, we needs $\frac{m}{2}$ generations, and for $j = 1, \cdots, \frac{m}{2}$, we define $Y_j = g\left(F_X^{-1}\left(u_1^{(j)}\right), \cdots, F_X^{-1}\left(u_n^{(j)}\right)\right)$ and $Y_j' = g\left(F_X^{-1}\left(1-u_1^{(j)}\right), \cdots, F_X^{-1}\left(1-u_n^{(j)}\right)\right)$, then $\widehat{\theta} = \frac{1}{m}\sum_{i=1}^{m/2}(Y_j + Y_j')$.

2.2.2. control variate

In this part, we still try to use the benefits of correlation.

Suppose there is a f with $\mu = \mathbb{E}f(x)$ known and f is correlated with g.Then define $\widehat{\theta_c} = g(x) + c(f(x) - \mu)$. $\widehat{\theta_c}$ is still an unbiased estimator of θ and

$$Var\left(\widehat{\theta_c}\right) = Var(g(x)) + 2c \cdot Cov(g(x), f(x)) + c^2 Var(f(x))$$

Let $c^* = -\frac{Cov(g(x), f(x))}{Var(f(x))}$, we minimize the $Var\left(\widehat{\theta_{c^*}}\right) = Var(g(x)) - \frac{Cov^2(g(x), f(x))}{Var(f(x))}$.

The percent of reduction is $\frac{Var(g(x)) - Var(\widehat{\theta}_{c^*})}{Var(g(x))} = Cor^2(g(x), f(x)) \times 100\%$.

2.2.3. Antethetic variables as a control variate

We combine the two methods, formulate control variate as linear combination of two unbiased estimator $\widehat{\theta_c} = c\widehat{\theta_1} + (1-c)\widehat{\theta_2}$. Suppose $\widehat{\theta_1}$ and $\widehat{\theta_2}$ have identical distributions and $r = Cor\left(\widehat{\theta_1}, \widehat{\theta_2}\right) < 0$. Then

$$\begin{aligned} Var\left(\widehat{\theta_c}\right) &= c^2 Var\left(\widehat{\theta_1}\right) + 2c(1-c)Cov\left(\widehat{\theta_1}, \widehat{\theta_2}\right) + (1-c)^2 Var\left(\widehat{\theta_2}\right) \\ &= Var\left(\widehat{\theta_1}\right)\left\{c^2 + 2c(1-c)r + (1-c)^2\right\} \\ &= Var\left(\widehat{\theta_1}\right)\left\{(2-2r)c^2 - (2-2r)c + 1\right\} \end{aligned}$$

$c^* = \frac{1}{2}$.

2.2.4. Control variate and linear regression

In control variate method, suppose we have n samples $(f(x_1), g(x_1)), \cdots, (f(x_n), g(x_n))$. When applying the linear regression $g(x) = \alpha + \beta f(x) + \varepsilon$, we have the following four important properties.

1. the OLS estimators

$$\begin{cases} \widehat{\alpha} = \overline{g(x)} - \widehat{\beta}\overline{f(x)} \\ \widehat{\beta} = \dfrac{\widehat{Cov}(f(x), g(x))}{Var(f(x))} = -\widehat{c^*} \end{cases}$$

2. The predicted value at $\mu = \mathbb{E}f(x)$ is the control variate estimator of the target integration

$$\begin{aligned} \widehat{\alpha} + \widehat{\beta}\mu &= \overline{g(x)} - \widehat{\beta}\left(\overline{f(x)} - \mu\right) \\ &= \overline{g(x)} + \widehat{c^*}\left(\overline{f(x)} - \mu\right) \\ &= \widehat{\theta}_{\widehat{c^*}} \end{aligned}$$

3. The variance of the control variate estimator is the residual mean squared error (MSE).

$$\widehat{Var}\left(\overline{g(x)} + c\left(\overline{f(x)} - \mu\right)\right) = \frac{1}{n}\widehat{Var}(g(x) + c(f(x) - \mu))$$

$$= \frac{1}{n}\widehat{Var}(g(x) - \hat{\beta}f(x) - \hat{\alpha})$$

$$= \frac{\widehat{\sigma_\varepsilon^2}}{n}$$

4. The percetage of improvement is $\{Cor(g(x), f(x))\} \times 100\%$ is the coefficient of determintion.

2.2.5. importance sampling

Simple MC integration, $\frac{b-a}{m}\sum_{i=1}^{m} g(X_i)$, weight the interval $[a, b]$ uniformly. The replicates $X_1, \cdots, X_m$ are uniformly distributed on $[a, b]$. Then we consider other weight functions.

Algorithm 1 *Importance sampling*

1. decide a "envolope" $f(x)$
2. For $i = 1, \cdots, m$
   (a) generate $x_i \sim f$
   (b) record $\frac{g(x_i)}{f(x_i)}$
3. $\widehat{\theta}_{IS} = \frac{1}{m}\sum_{i=1}^{m} \frac{g(x_i)}{f(x_i)}$

Let X be a r.v. with density f such that $f(x) > 0$ on the support of g. Set $Y = \frac{g(x)}{f(x)}$ then

$$\theta = \int g(x)dx = \int \frac{g(x)}{f(x)} f(x)dx = \mathbb{E}_f\left[\frac{g(x)}{f(x)}\right]$$

Thus we can use $\widehat{\theta}_{IS} = \frac{1}{m}\sum_{i=1}^{m} \frac{g(x_i)}{f(x_i)}$ to estimate θ.

Now we analysis the variance.

$$Var(\widehat{\theta}_{IS}) = Var\left(\frac{1}{m}\sum_{i=1}^{m} Y_i\right) = \frac{1}{m}Var(Y)$$

$$= \frac{1}{m}Var\left(\frac{g(x)}{f(x)}\right) = \frac{1}{m}\left\{\mathbb{E}\left(\frac{g(x)}{f(x)}\right)^2 - \left(\mathbb{E}\frac{g(x)}{f(x)}\right)^2\right\}$$

where $\mathbb{E}\frac{g(x)}{f(x)} = \int \frac{g(x)}{f(x)} f(x)dx = \theta$ and $\mathbb{E}\left(\frac{g(x)}{f(x)}\right)^2 = \int \frac{g^2(x)}{f(x)}dx$.

$$Var(\widehat{\theta}_{IS}) = \frac{1}{m}\left\{\int \frac{g^2(x)}{f(x)}dx - \theta^2\right\}$$

where $\int \frac{g^2(x)}{f(x)}dx = \left\{\int \frac{g^2(x)}{f(x)}dx\right\}\left\{\int f(x)dx\right\} \geq \left\{\int g(x)dx\right\}^2$. The equation holds iff $f(x) \propto |g(x)|$.

2.2.6. stratified sampling

Algorithm 2 *Stratified Sampling*

1. divide $[0, 1]$ into k strata where the j-th strata is $I_j = \left(\frac{j-1}{k}, \frac{j}{k}\right)$
2. On each strata, for $i = 1, \cdots m_j$
   (a) generate $x_i^{(j)} \sim Unif(I_j)$ by density $f_j(x) = k \cdot \mathbb{1}(x \in I_j)$
   (b) $\widehat{\theta_j} = \frac{1}{m_j}\sum_{i=1}^{m_j} g\left(x_i^{(j)}\right)$
3. $\widehat{\theta}^S = \frac{1}{k}\sum_{j=1}^{k}\widehat{\theta}_j$

Note that $\mathbb{E}\widehat{\theta_j} = \mathbb{E}g\left(x^{(j)}\right) = \int g(x)k \cdot \mathbb{1}(x \in I_j)dx = k\int_{I_j} g(x)dx$. That's why we need $\frac{1}{k}$.

Now we show that the variance of importance sampling is smaller than simple MC integration. Denote $\widehat{\theta}^M$ be the simple MC estimation. For simplicity, we suppose in importance sampling each strata has equal numer, m, of replicates, and total number $M = mk$. Denote $\theta_j = \mathbb{E}\{g(u)|u \in I_j\}$ and $\sigma_j^2 = Var\{g(u)|u \in I_j\}$.

$$\begin{aligned} Var\left(\widehat{\theta}^S\right) &= Var\left(\frac{1}{k}\sum_{j=1}^{k}\widehat{\theta}_j\right) = \frac{1}{k^2}\sum_{j=1}^{k} Var(\widehat{\theta}_j) \\ &= \frac{1}{k^2}\sum_{j=1}^{k}\frac{\sigma_j^2}{m} = \frac{1}{Mk}\sum_{j=1}^{k}\sigma_j^2 \end{aligned}$$

Suppose a two-step experiment, J is discrete uniform on $\{1, \cdots, k\}$. For $i = 1, \cdots, M$,

1. draw J
2. generate from I_J

$$Var\left(\widehat{\theta}^M\right) = \frac{1}{M}Var(g(u))$$

$$= \frac{1}{M}\left\{Var\left[\mathbb{E}\left(g(u)\middle|J\right)\right] + \mathbb{E}\left[Var\left(g(u)\middle|J\right)\right]\right\}$$

$$= \frac{1}{M}\left\{Var(\theta_J) + \frac{1}{k}\sum_{j=1}^{k}\sigma_j^2\right\}$$

$$= \frac{1}{Mk}\sum_{j=1}^{k}\sigma_j^2 + \frac{1}{M}Var(\theta_J) \geq Var\left(\widehat{\theta}^{S}\right)$$

The equation holds iff $Var(\theta_J) = 0$, i.e., $\theta_1 = \cdots = \theta_k$.

2.2.7. Stratified importance sampling

Algorithm 3 *Stratified importance sampling*

1. choose an importance function f.
2. divide the real line into k strata where the j-th strata is $I_j = (a_{j-1}, a_j)$, where
   $a_0 = -\infty$, $a_j = F^{-1}\left(\frac{j}{k}\right)$, $a_k = +\infty$.
3. On strata j define $g_j(x) = g(x)\mathbb{1}(x \in I_j)$ and
   $f_j(x) = f_{x|I_j}(x|I_j) = \frac{f(x)\mathbb{1}(x \in I_j)}{\int f(x)\mathbb{1}(x \in I_j)dx} = kf(x)\mathbb{1}(x \in I_j)$, for $i = 1, \cdots m$
   (a) generate $x_i^{(j)} \sim f_j(x)$
   (b) $\widehat{\theta_j} = \frac{1}{m}\sum_{i=1}^{m} g\left(x_i^{(j)}\right)$
4. $\widehat{\theta}^{SIS} = \sum_{j=1}^{k}\widehat{\theta}_j$

Note that $\theta_j = \int g_j(x)dx = \int_{I_j} g(x)dx$, thus $\theta = \sum \theta_j$.

Now we prove that $Var\left(\widehat{\theta}^{SIS}\right) \leq Var\left(\widehat{\theta}^{IS}\right)$. In I_j, denote $\theta_j = \int_{I_j} g(x)dx = \mathbb{E}\left\{\frac{g_j(x)}{f_j(x)}\right\}$ and $\sigma_j^2 = Var\left\{\frac{g_j(x)}{f_j(x)}\right\}$, where $x \sim f_j$.

$$Var\left(\widehat{\theta}^{SIS}\right) = Var\left(\sum\widehat{\theta}_j\right) = \sum Var(\widehat{\theta}_j) = \sum\frac{\sigma_j^2}{m} = \frac{k}{M}\sum\sigma_j^2$$

$$Var\left(\widehat{\theta}^{IS}\right) = \frac{1}{M}Var(Y) = \frac{\sigma^2}{M} \text{ where } \sigma^2 = Var\left(\frac{g(x)}{f(x)}\right)$$

Next we show that $\sigma^2 - k\sum \sigma_j^2 \geq 0$. We consider the two-stage experiment, For $J = j$, we geberate x^* from f_j and set $Y^* = \dfrac{g_j(x^*)}{f_j(x^*)} = \dfrac{g(x^*)}{kf(x^*)}$. $x^* \overset{d}{=} x$ and $kY^* \overset{d}{=} Y$.

$$Var(Y^*) = \mathbb{E}\left\{Var(Y^*|J)\right\} + Var\left\{\mathbb{E}(Y^*|J)\right\}$$

where

$$\mathbb{E}\left\{Var(Y^*|J)\right\} = \mathbb{E}(\sigma_J^2) = \frac{1}{k}\sum \sigma_j^2$$
$$Var\left\{\mathbb{E}(Y^*|J)\right\} = Var(\theta_J)$$

Then

$$\sigma^2 = Var(Y) = k^2 Var(Y^*) = k^2 Var(\theta_J) + k\sum \sigma_j^2 \geq k\sum \sigma_j^2$$

3. MC in statistical inference

3.1. point estimation

$$\widehat{\theta} = \frac{1}{m}\sum \widehat{\theta}^{(j)}$$

$$\widehat{se}(\overline{x}) = \begin{cases} \dfrac{1}{n}\left\{\sum (x_i - \overline{x})^2\right\}^{\frac{1}{2}} \\ \dfrac{1}{\sqrt{n}}\left\{\dfrac{1}{n-1}\sum (x_i - \overline{x})^2\right\}^{\frac{1}{2}} \end{cases}$$

$$\widehat{mse} = \frac{1}{m}\sum \left(\widehat{\theta}^{(j)} - \theta\right)^2$$

3.2. confidence interval

Algorithm 4 *Monte Carlo Confindence interval*

1. For each replicate $j = 1, \cdots, m$
   (a) generate the j-th random sample $X_1^{(j)}, \cdots, X_n^{(j)}$
   (b) compute the confidence interval C_j for the j-th sample
   (c) Compute $y_j = \mathbb{1}(\theta \in C_j)$ for the j-the sample
2. Compute the empirical confidence level $\overline{y} = \frac{1}{m}\sum y_j$

3.3. hypothesis testing

3.3.1. empirical type I error rate

Algorithm 5 *MC type I error rate*

1. For each replicate, indexed by $j = 1, \cdots, m$
   (a) Generate the j-th random sample $x_1^{(j)}, \cdots, x_n^{(j)}$ from the null distribution.
   (b) Compute the test statistic T_j from the j-th sample.
   (c) Record the test decision $I_j = 1$ if H_0 is rejected at significance level α and otherwise $I_j = 0$.
2. Compute the proportion of significant tests $\frac{1}{m}\sum I_j$. This proportion is the observed Type I error rate.

3.3.2. Power of a test

Algorithm 6 *MC power of a test*

1. select a particular $\theta_1 \in \Theta_1$
2. For each replicate, indexed by $j = 1, \cdots, m$
   (a) Generate the j-th random sample $x_1^{(j)}, \cdots, x_n^{(j)}$ under θ_1.
   (b) Compute the test statistic T_j from the j-th sample.
   (c) Record the test decision $I_j = 1$ if H_0 is rejected at significance level α and otherwise $I_j = 0$.
3. Compute the proportion of significant tests $\widehat{\pi}(\theta_1) = \frac{1}{m}\sum I_j$.

4. Bootstrap

4.1. bootstrap estimate of distribution

Algorithm 7 *bootstrap estimate of distribution*

1. Foe each Bootstrap replicate $b = 1, \cdots, B$
   (a) Generate samle $x^{*(b)} = \left(x_1^{*(b)}, \cdots, x_m^{*(b)}\right)$ by sample with replacement from the observation $x_1, \cdots, x_n$
   (b) Compute the b-th replicate $\widehat{\theta^{(b)}}$ using $x^{*(b)}$.
2. The bootstrap estimate of $F_{\widehat{\theta}}$ is the empirical distribution of $\widehat{\theta}^{(1)}, \cdots, \widehat{\theta}^{(B)}$

4.2. point estimation

1. se of $\widehat{\theta}$. $\widehat{se}_B \triangleq \widehat{se}\left(\widehat{\theta}^*\right) = \sqrt{\frac{1}{B-1}\Sigma_{b=1}^{B}\left(\widehat{\theta}^{(b)} - \overline{\widehat{\theta}}^*\right)^2}$ where $\overline{\widehat{\theta}}^* = \frac{1}{B}\sum_{b=1}^{B}\widehat{\theta}^{(b)}$.
2. bias of $\widehat{\theta}$. $bias(\widehat{\theta}) = \frac{1}{B}\sum_{b=1}^{B}\widehat{\theta}^{(b)} - \widehat{\theta}$.

4.3. confidence interval

Now we use Bootstrap to estimate confidence intervals.

4.3.1. standard normal distribution

Use approximately normal poperty, $[\widehat{\theta} \pm 1.96\widehat{se}_B(\widehat{\theta})]$

4.3.2. percentile CI

Use the sample quantile $\left[\widehat{\theta}^*_{\alpha/2}, \widehat{\theta}^*_{1-\alpha/2}\right]$.

4.3.3. Basic bootstrap CI

Suppose (L, U) is the confidence interval, i.e., $P(L \geq \theta) = P(U \leq \theta) = \frac{\alpha}{2}$. Then

$$\begin{aligned}\frac{\alpha}{2} = P(L \geq \theta) &= P(L - \widehat{\theta} \geq \theta - \widehat{\theta}) \\ &= P(\widehat{\theta} - \theta \geq \widehat{\theta} - L)\end{aligned}$$

Thus $L - \widehat{\theta}$ is the $1 - \frac{\alpha}{2}$ percentile of $\widehat{\theta} - \theta$. We can estimate the $1 - \frac{\alpha}{2}$ quantiles of $\widehat{\theta}$ using bootstrap replicate $\widehat{\theta}^*_{1-\alpha/2}$, then $\widehat{\theta}^*_{1-\alpha/2} - \widehat{\theta}$ is approximately equal to the $1 - \frac{\alpha}{2}$ quantile of $\widehat{\theta} - \theta$. Set $\widehat{\theta} - L = \widehat{\theta}^*_{1-\alpha/2} - \widehat{\theta}$, we have $L = 2\widehat{\theta} - \widehat{\theta}^*_{1-\alpha/2}$.

Similarly, we have $U = 2\widehat{\theta} - \widehat{\theta}^*_{\alpha/2}$. Thus the CI is $\left[2\widehat{\theta} - \widehat{\theta}^*_{1-\alpha/2}, 2\widehat{\theta} - \widehat{\theta}^*_{\alpha/2}\right]$.

4.3.4. Bootstrap t CI

Algorithm 8 *Bootstrap t confidence interval*

1. Compute $\widehat{\theta}$ from the observed data.
2. For each bootstrap replicate $b = 1, \cdots, B$
   (a) Sample with replacement $x^{(b)} = \left(x_1^{(b)}, \cdots, x_n^{(b)}\right)$
   (b) compute $\widehat{\theta}_{(b)}$ from $x^{(b)}$
   (c) estimate $\widehat{se}\left(\widehat{\theta}^{(b)}\right)$. (another layer of Bootstrap, resample from $x^{(b)}$, not from x)
   (d) compute the b-th replicate of the t^* distribution $t^{(b)} = \dfrac{\widehat{\theta}^{(b)} - \widehat{\theta}}{\widehat{se}\left(\widehat{\theta}^{(b)}\right)}$.
3. find sample quantiles $t^*_{\frac{\alpha}{2}}$ and $t^*_{1-\frac{\alpha}{2}}$ from $\left\{t^{(b)}\right\}_{b=1}^{B}$.
4. compute $\widehat{se}(\widehat{\theta})$ from $\left\{\widehat{\theta}^{(b)}\right\}_{b=1}^{B}$.
5. confidence interval $\left[\widehat{\theta} - t^*_{1-\frac{\alpha}{2}}\widehat{se}(\widehat{\theta}), \widehat{\theta} - t^*_{\frac{\alpha}{2}}\widehat{se}(\widehat{\theta})\right]$

5. Jacknife

Jacknife is similar to the leave-one-out method. In each sample, the Jacknife except one obvervation, i.e., the i-th Jacknife sample is $x_{(i)} = (x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_n)$. The i-th Jacknife estimator is $\widehat{\theta}_{(i)} = f(x_{(i)})$.

5.1. Bias

The jacknife bias is

$$\widehat{bias}_{jack} = (n-1)\left(\overline{\widehat{\theta}}_{(\cdot)} - \widehat{\theta}\right)$$

where $\overline{\widehat{\theta}}_{(\cdot)} = \dfrac{1}{n}\sum_{i=1}^{n}\widehat{\theta}_{(i)}$.

why $n-1$? For example, $\widehat{\theta} = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$,

$bias(\widehat{\theta}) = \mathbb{E}(\widehat{\theta}) - \sigma^2 = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{1}{n}\sigma^2$. while

$$\begin{aligned}\mathbb{E}(\widehat{\theta}_{(i)} - \widehat{\theta}) &= \mathbb{E}(\widehat{\theta}_{(i)} - \theta) - \mathbb{E}(\widehat{\theta} - \theta) \\ &= -\frac{1}{n-1}\sigma^2 + \frac{1}{n}\sigma^2 \\ &= -\frac{\sigma^2}{n(n-1)} \\ &= \frac{bias(\widehat{\theta})}{n-1}\end{aligned}$$

5.2. Standard Error

The jacknife standard error is

$$\widehat{se}_{jack} = \sqrt{\frac{n-1}{n}\Sigma_{i=1}^{n}\left(\widehat{\theta}_{(i)} - \overline{\widehat{\theta}}_{(\cdot)}\right)^2}$$

Why $\frac{n-1}{n}$? For example, if $\widehat{\theta} = \overline{x}$, then $\widehat{\theta}_{(i)} = \frac{n\overline{x} - x_i}{n-1}$, $\overline{\widehat{\theta}}_{(\cdot)} = \overline{x}$.

$$\begin{aligned}\sum_{i=1}^{n}\left(\widehat{\theta}_{(i)} - \overline{\widehat{\theta}}_{(\cdot)}\right)^2 &= \sum_{i=1}^{n}\left(\frac{n\overline{x} - x_i}{n-1} - \overline{x}\right)^2 \\ &= \frac{1}{(n-1)^2}\sum_{i=1}^{n}(\overline{x} - x_i)^2 \\ &= \frac{1}{n-1}s^2\end{aligned}$$

6. Bayesian statistics and MCMC

Bayesian statistics looks problems in a differen way from frenquetists. In frequentist, the parameter is fixed but unknown, and he experiment is repeatable. However, in Bayesian, the parameter is random and unknown, and the experiment is fixed, or to say, the target of Bayesian is the posterior$p(\theta|data)$, rather than the likelihood $p(data|\theta)$.

6.1. Bayesian problem set-up

parameter	θ
data	X
prior distribution of θ	$\pi(\theta)$

sampling model for X	$f(x\|\theta)$
posterior distribution for $\theta\|X$	$p(\theta\|X) = \dfrac{\pi(\theta)f(X\|\theta)}{\int_\Theta \pi(\theta)f(X\|\theta)d\theta} \propto prior \times likelihood$

6.2. Markov chain Monte Carlo

Construct a Markov chain $\{X_t : t = 0, 1, \cdots\}$ whose stationary distribution is target distribution.

6.2.1. Metropolis-Hasting sampler

Now supppose our target distribution is f, to move from one state to another, we have a proposal $Y \sim g(y|X_t)$.

Algorithm 9 *MH sampler*

1. Choose a proper proposal $g(\cdot|X_t)$
2. Initialize X_0 and repeat until the chain converges. At time t,
   (a) Generate Y from $g(\cdot|X_t)$
   (b) Compute the acceptance rate $r(X_t, Y) = \dfrac{f(Y)g(X_t|Y)}{f(X_t)g(Y|X_t)}$
   (c) Let $\alpha(X_t, Y) = \min\{r(X_t, Y), 1\}$ be the acceptance ratio. Set
   $$X_{t+1} = \begin{cases} Y & with\ probability\ \alpha(X_t, Y) \\ X_t & with\ probability\ 1 - \alpha(X_t, Y) \end{cases}$$

Now we show how the MH sampler works.

Let $s,\ t$ be two different state. Without loss of generality, we assume that $f(s)g(r|s) \geq f(r)g(s|r)$. Then $\alpha(r, s) = 1$ and $\alpha(s, r) = \dfrac{f(r)g(s|r)}{f(s)g(r|s)}$, and

$$\begin{aligned} P(X_t = s, X_{t+1} = r) &= P(X_t = s)P(X_{t+1} = r|X_t = s) \\ &= f(s)g(r|s)\alpha(s, r) \\ &= f(r)g(s|r) \\ P(X_t = r, X_{t+1} = s) &= P(X_t = r)P(X_{t+1} = s|X_t = r) \\ &= f(r)g(s|r)\alpha(r, s) \\ &= f(r)g(s|r) \\ \implies P(X_t = s, X_{t+1} = r) &= P(X_t = r, X_{t+1} = s) \end{aligned}$$

Thus

$$\begin{aligned} P(X_t = r) &= \sum_s P(X_t = r, X_{t+1} = s) \\ &= \sum_s P(X_t = s, X_{t+1} = r) \\ &= P(X_{t+1} = r) \end{aligned}$$

The Markov chain is stationary at f. □

We can also prove it in a kernel point of view. Define $K(r,s) = g(s|r)\alpha(r,s)$ be the transition kernel from r to s. We call f is the stationary distriibution if the balance condition holds

$$\int f(r)K(r,s)dr = f(s)$$

Under MH sampler,

$$\begin{aligned} f(r)K(r,s) &= f(r)g(s|r)\alpha(r,s) \\ &= f(r)g(s|r)\min\left\{1, \frac{f(s)g(r|s)}{f(r)g(s|r)}\right\} \\ &= \min\{f(r)g(s|r),\ f(s)g(r|s)\} \\ &= f(s)g(r|s)\alpha(s,r) \\ &= f(s)K(s,r) \end{aligned}$$

We have the detailed balanced consition $f(r)K(r,s) = f(s)K(s,r)$. Then

$$\int f(r)K(r,s)dr = \int f(s)K(s,r)dr = f(s)\int K(s,r)dr = f(s)$$

6.2.2. Metropolis sampler

With symmetric proposal $g(X|Y) = g(Y|X)$. $r(X_t,\ Y) = \dfrac{f(Y)}{f(X_t)}$.

6.2.3. Random Walk sampler

Proposal $g(Y|X_t) = g(|Y - X_t|)$. For example $Y = X_t + N(0, \sigma^2)$. $r(X_t,\ Y) = \dfrac{f(Y)}{f(X_t)}$.

6.2.4. independent sampler

Proposal $g(Y|X_t) = g(Y)$. $r(X_t,\ Y) = \dfrac{f(Y)g(X_t)}{f(X_t)g(Y)}$.

6.2.5. Gibbs sampler

Now suppose we generate samples from a multivariate $f(x)$ where $x \in \chi \subset \mathbb{R}^d$. We partition the d-dimension vector x into K disjoint blocks, denoted by $x = (x_1,\ \cdots, x_K)^T$ where $K \leq d$. Let

$$f_k(x_k|x_{-k}) = f_k(x_k|x_1, \cdots, x_{k-1}, x_{k+1}, \cdots, x_K),\ k = 1, \cdots, K$$

Algorithm 10 *Gibbs sampler*

1. Starting with an arbitrary point $x^{(0)} \in \chi$ with $f\left(x^{(0)}\right) > 0$
2. At time t,
   (1) generate $x_1^{(t)} \sim f_1\left(x_1|x_2^{(t-1)}, \cdots, x_K^{(t-1)}\right)$
   $\vdots$
   (k) generate $x_k^{(t)} \sim f_k\left(x_k|x_1^{(t)}, \cdots, x_{k-1}^{(t)}, x_{k+1}^{(t-1)}, \cdots, x_K^{(t-1)}\right)$
   $\vdots$
   (K) generate $x_K^{(t)} \sim f_K\left(x_K|x_2^{(t-1)}, \cdots, x_{K-1}^{(t-1)}\right)$
3. Set $x^{(t)} = \left(x_1^{(t)}, \cdots, x_K^{(t)}\right)^T$

Gibbs sampler is a special case ofMH sampler, with accept ratio 1. Proof can be found in homework 7.

6.3. Monitoring the convergence (Gelman-Rubin method)

In MCMC, the chain may be traped in some local mode, we can run multiple chains from different initial point to check the convergence. Recall ANOVA in DATA130046 statistics II. Suppose there are J chains, and n is the number in each chain after discarding the burn-in period. ψ is a function of data. Write

$$\psi_i^{(j)} = \psi\left(X_1^{(j)}, \cdots, X_i^{(j)}\right),\ i = 1, \cdots, n; j = 1, \cdots, J$$

$$\overline{\psi}^{(j)} = \frac{1}{n}\sum_{i=1}^{n} \psi_i^{(j)}$$

$$\overline{\psi} = \frac{1}{n}\sum_{i=1}^{n} \overline{\psi}_i^{(j)}$$

Then the Between sequence variance is $B = \frac{n}{J-1}\sum_{j=1}^{J}\left(\overline{\psi}^{(j)} - \overline{\psi}\right)^2$ and the Within sequence variance is $W = \frac{1}{J}\sum_{j=1}^{J} s_j^2$ where $s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\psi_i^{(j)} - \overline{\psi}^{(j)}\right)^2$. The **Gelman-Rubin statistic** is

$$\sqrt{\widehat{R}} = \sqrt{\frac{\frac{n-1}{n}W + \frac{1}{n}B}{W}}$$

which should decrease and converge to 1 if chain converges well. A recommended threshold

is 1.1.

7. EM algorithm

The expectation-maximization algorithm is used to overcome the difficulties in maximizing likelihoods. Suppose except observed data Y, there are unobserved data U. $f(Y|\theta)$ is not easily evaluate, but $f(Y,U|\theta)$ is easy to calculate. Then we write the total log-likelihood as

$$\log f(Y,U|\theta) = \log f(Y|\theta) + \log f(U|Y,\theta)$$

Take expactation with respect to u, and given $Y = y,\ \theta = \theta^*$, we have

$$\mathbb{E}_u\left\{\log f(Y,U|\theta)|Y=y,\theta^*\right\} = \ell(\theta) + \mathbb{E}_u\left\{\log f(U|Y,\theta)|Y=y,\theta^*\right\}$$

Write it as

$$Q(\theta,\theta^*) = \ell(\theta) + C(\theta,\theta^*) \tag{1}$$

Let $\theta = \theta^*$, we have

$$Q(\theta^*,\theta^*) = \ell(\theta^*) + C(\theta^*,\theta^*) \tag{2}$$

(1)-(2) we get $\ell(\theta) - \ell(\theta^*) = Q(\theta,\theta^*) - Q(\theta^*,\theta^*) - \left\{C(\theta,\theta^*) - C(\theta^*,\theta^*)\right\}$.

The right part is

$$\begin{aligned} C(\theta,\theta^*) - C(\theta^*,\theta^*) &= \mathbb{E}\left\{\log\frac{f(U|Y,\theta)}{f(U|Y,\theta^*)}\middle|Y=y,\theta^*\right\} \\ &\le \log\mathbb{E}\left\{\frac{f(U|Y,\theta)}{f(U|Y,\theta^*)}\middle|Y=y,\theta^*\right\} \\ &= \log\left\{\int\frac{f(u|y,\theta)}{f(u|y,\theta^*)}f(u|y,\theta^*)du\right\} \\ &= 0 \end{aligned}$$

Thus $\ell(\theta) - \ell(\theta^*) = Q(\theta,\theta^*) - Q(\theta^*,\theta^*)$. Given θ^*, if we find θ that $Q(\theta,\theta^*) \ge Q(\theta^*,\theta^*)$, then $\ell(\theta) \ge \ell(\theta^*)$.

Algorithm 11 *The EM algorithm*

1. **E-step:** compute $Q(\theta,\theta^{(t)}) = \mathbb{E}_U\left\{\log f(Y,U|\theta)\middle|Y=y,\theta^{(t)}\right\}$
2. **M-step:** $\theta^{(t+1)} = \arg\max_\theta Q(\theta,\theta^{(t)})$

8. Variational inference

To approximate the posterior $p(z|x) = \dfrac{\pi(z)f(x|z)}{f(x)}$ in bayesian statistics, rather than generating random samples using MCMC, we can solve the problem by optimization, i.e., find the closest density to $p(x|z)$. To judge the distance between our density q and the target p, we can use the Kullback-Leibler divergence.

8.1. KL divergence

$$KL(f||g) = \mathbb{E}_f\left\{log\frac{f(x)}{g(x)}\right\} = \int_{-\infty}^{+\infty} log\frac{f(x)}{g(x)}f(x)dx$$

Properties:

1. KL divergence is non-negative and KL=0 iff f=g
2. KL is not a distance, i.e., $KL(f||g) \neq KL(g||f)$

Proof of Property 1:

$$-KL(f||g) = \mathbb{E}_f\left\{log\frac{g(x)}{f(x)}\right\} \overset{Jensen's\ inequality}{\leq} log\left\{\mathbb{E}_f\frac{g(x)}{f(x)}\right\} = log\int_{-\infty}^{+\infty}\frac{g(x)}{f(x)}f(x)dx = 0$$

8.2. Evidence lower bound (ELBO)

Suppose we have a family of densities over latent variables, then we want

$$q^*(z) = \arg\min_{q\in Q} KL(q(z)\ ||\ p(x|z))$$

$$\begin{aligned} KL(q(z)\ ||\ p(x|z)) &= \mathbb{E}_q\left\{log\frac{q(z)}{p(z|x)}\right\} \\ &= \mathbb{E}_q\{log\ q(z)\} - \mathbb{E}_q\{log\ p(z|x)\} \\ &= \mathbb{E}_q\{\log q(z)\} - \mathbb{E}_q\{\log f(x,z)\} + \mathbb{E}_q\{\log f(x)\} \\ &= \log f(x) - ELBO(q) \end{aligned}$$

Define evidence lower bound (ELBO) as

$$ELBO(q) \triangleq \mathbb{E}_q\{\log f(x,z)\} - \mathbb{E}_q\{\log q(z)\}$$

Then

$$\log f(x) = KL(q(z)||p(z|x)) = KL(q(z)\ ||\ p(x|z)) + ELBO(q) \geq ELBO(q)$$

To minimize $KL(q(z)\ ||\ p(x|z))$, we maximize $ELBO(q)$.

$$\begin{aligned} ELBO(q) &= \mathbb{E}_q\{\log f(x,z)\} - \mathbb{E}_q\{\log q(z)\} \\ &= \mathbb{E}_q\{\log f(x|z)\} + \mathbb{E}_q\{\log \pi(z)\} - \mathbb{E}_q\{\log q(z)\} \\ &= \mathbb{E}_q\{\log f(x|z)\} - \mathbb{E}_q\left\{\log\frac{q(z)}{\pi(z)}\right\} \\ &= \mathbb{E}_q\{\log f(x|z)\} - KL(q(z)||\pi(z)) \end{aligned}$$

To maximize $ELBO(q)$, on one hand, we maximize $\mathbb{E}_q\{\log f(x|z)\}$, i.e., find q that fits the data. On the other hand, we minimize $KL(q(z)||\pi(z))$, i.e., find q that fits the prior.

8.3. The mean-field variational family

Now we consider a special case that the family of the q is $Q \triangleq \left\{q:\ q(z) = \prod_{j=1}^{m} q_j(z_j)\right\}$. It can be treated as the elements of the q is "indepdent".

$$\begin{aligned} ELBO(q) &= \mathbb{E}_q\{\log f(x,z)\} - \mathbb{E}_q\{\log q(z)\} \\ &= \int \prod_{j=1}^{m} q_j \left\{ \log f(x,z) - \log q_j \right\} dz \end{aligned}$$

We use a method called coordinate ascend to maximize $ELBO(q)$, i.e., fix all others, climb up on q_k. Then $ELBO(q) = \int q_k \left\{ \int \log f(x,z) \prod_{j\neq k} q_j dz_j \right\} dz_k - \int q_k \log q_k dz_k + c_1$, where c_1 is the constant have nothing with q_k.

Define $\log \widetilde{p}(x, z_k) = \int \log f(x,z) \prod_{j\neq k} q_j dz_j + c_2$, thus

$$\begin{aligned} ELBO(q) &= \int q_k \log \widetilde{p}(x, z_k) dz_k - \int q_k \log q_k dz_k + c_1 \\ &= -KL\left(q_k || \widetilde{p}(x, z_k)\right) + c_3 \\ q_k^* &= \widetilde{p} \propto \exp\left\{\mathbb{E}_{-k} \log f(x,z)\right\} \end{aligned}$$