StatsM254 Statistical Methods in Computational Biology Lecture 17 - 5/31/2017

Lecture 17 - Sampling Methods

Lecturer: Jingyi Jessica Li *Scribe: Harry Yang, Christina Burghard*

1 Canonical Correlation Analysis (CCA)

Dimentionality Reduction for two sets of variables. e.g. n = 40 mice, r = 120 gene expression levels associated with nutrition, s = 21 fatty acids' concentration

X = Gene expression matrix (40 x 120)
Y = Concentration matrix (40 x 21)

CCA seeks linear transformation:

$$g_{(120\times 1)} \in R^r$$
$$h_{(21\times 1)} \in R^s$$

so that $Cor(X-g, Y-h)$ is maximized.
$(g_1, h_1) = 1^{st}$ canonical correlation vector ... $(g_k, h_k) = k^{th}$ canonical correlation vector

1.1 Algirothm

$S_{11} = \hat{Cov}(x), S_{22} = \hat{Cov}(y) = \frac{1}{n}Y^TY$ if every column of Y is centered, $S_{12} = \hat{Cov}(x,y), S_{21} = Cov(\hat{Y}, X)$

$$(g_k, h_k) = argmax_{g,h}(g^TS_{12}h/\sqrt{g^TS_{11}g \cdot h^TS_{22}h})$$
where $g^TS_{11}g_j = 0, h^TS_{22}h_j = 0, j = 0 \dots k-1$

2 Simulation Methods

2.1 Monte Carlo Method

Goal: to evaluate a parameter $\theta = E[f(x)]$ for $X \sim P$, where P is the target distribution.

For example,

$$P = N(\mu, \sigma^2) \tag{1}$$

where $\mu = E[x]$ and $\sigma^2 = E[(x-\mu)^2]$.

For example, how to calculate the posterior mean if we choose a prior such that the posterior doesn't have a nice distribution function?

1. Direct (naive) Monte Carlo

   Sample x_i as independent and identically distributed from P

   Take the average $\hat{\theta} = \frac{1}{n}\sum_{i=1}^n f(x_i)$

   Find the 95% conifdence interval, eg $E[x]$

   The indicator function is $E[I(x > c)]$

   This is useful for:

- Bayesian inference (from the posterior)
- High-dimensional parameter space
- When there is no analytic form of the target distribution P

Example: X and Y are independent random variables from $Unif(0,1)$. What is $P(X^2+Y^2 \geq 1)$?

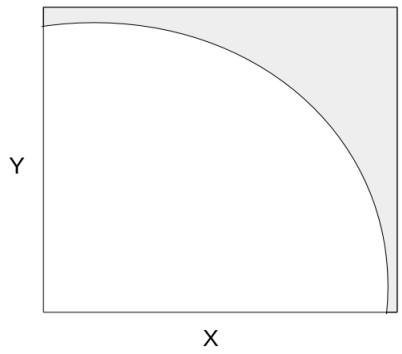

Figure 1: The white area is $\frac{\pi}{4}$, the grey region is $1-\frac{\pi}{4}$, which is equal to the probability.

We draw $X_i \sim U(0,1), y$ and $Y_i \sim i = 1, ..., n$
$I(X_i^2 + Y_i^2 \geq 1) \longrightarrow 0$ or 1
As n increases, see if the estimate approaches $1-\frac{\pi}{4}$.

2. How to simulate from a distribution

   Theorem: let $U \sim Unif(0,1)$, and a distribution has a known cumulative distribution function (CDF) with inverse.
   Let $X = F^{-1}(U)$, then $X \sim F$

$$\mathbb{P}(X \leq x) = \mathbb{P}(F^{-}1(U) \leq x) = \mathbb{P}(U \leq F(x)) \longrightarrow X \sim F = F(x) \tag{2}$$

   e.g.: (using R function rnorm) if $X \sim F, F(x) = \mathbb{P}(X \leq x)$. We can use U to project x.
   Example: To sample $X_1, ..., X_n \sim Exponential(1)$:

   - In R: rexp(n,1) (use set.seed(0) in order to make this replicable)
   - Manually: sample $U_1, ..., U_n \sim Unif(0,1)$
     CDF of Exp(1):
     $F(x) = 1 - e^{-x}, x \geq 0$
     $\longrightarrow F^{-1}(x) = -\log(1-x), x \in [0,1]$
     Let $X_i = -\log(1-U_i)$
     Then $X_1, ..., X_n$ $Exp(1)$

3. Rejection Method [1] Setting:
   (a) We want to sample from a target distribution with density $\pi(x)$.
   (b) The density is known to a constant $l(x) = C\pi(x)$, where l is known, and C and $\pi(x)$ are unknown.
   (c) We can construct

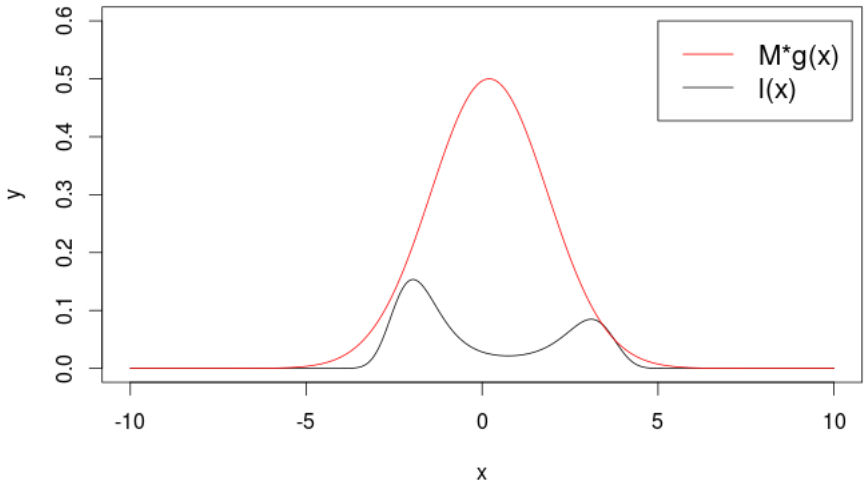

Figure 2: Here the grey line is $l(x)$ and the black line is $m * g(x)$.

i. an envelope function, $g(x)$
ii. a constant m such that $mg(x) \geq l(x) \forall x...$

Procedure:

i. Draw a sample point from $g(x)$ and compute the ratio $r = \frac{l(x)}{mg(x)} \in [0, 1]$. (Because $g(x)$ is normal, it's easy to draw a sample.)
ii. Flip a coin with probability of success $= r$, or draw $y \sim Bernoulli(r)$.
If $y = 1$, keep x. Otherwise, discard x.
iii. Repeat step 1 until the n^{th} sample is accepted.

The result of this is that if the sample point from the envelope function is further from the target distribution, it will have a higher ratio and is more likely to be discarded.

3 Rejection Method

Goal: To sample from a target distribution with density $\pi(x)$ which is unknown. Sampling from a cdf requires the inverse function, which is often difficult to obtain, so for those cases, the rejection method is convenient. The target function can be represented as:

$$l(x) = c \cdot \pi(x)$$

where $\pi(x)$ and c are unknown and $l(x)$ is known but difficult to integrate. The approach used by this method is to construct an envelope function $g(x)$ which is known and has a simple, known distribution form (e.g. normal density) such that:

$$m \cdot g(x) \geqslant l(x) \qquad \forall x$$

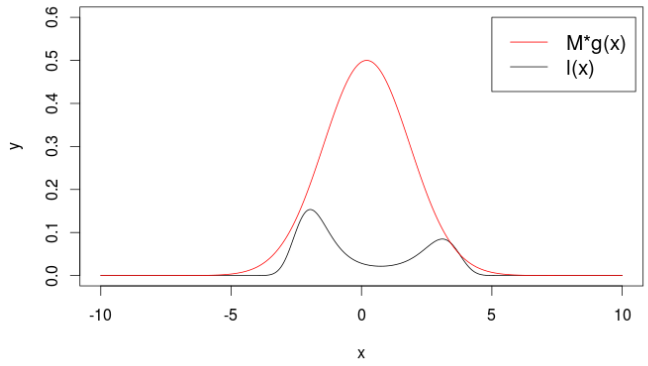

Figure 3: Target and Envelope function

3.1 Algorithm

To obtain a sample of n data points.

- Draw a point x_i from the envelope function $g(x)$ so that $x_i \sim g(x)$.
- Compute the ratio r:

$$r = \frac{l(x_i)}{m \cdot g(x_i)}$$

- Draw z as a Bernoulli trial with probability r, i.e. $z \sim bernoulli(r)$ (rbinom function in R). If $z = 1$ keep x_i, else discard x_i.
- Repeat until you get n data points.

3.2 Why does it work?

$$\text{Let } z = \begin{cases} 1 & \text{if } X \sim g(x) \text{ is accepted} \\ 0 & \text{otherwise} \end{cases}$$

Then

$$\begin{aligned} p(z=1) &= \int p(z=1, X=x)dx \\ &= \int p(z=1 \mid X=x) \cdot g(x)dx \\ &= \int \frac{l(x)}{m \cdot \cancel{g(x)}} \cdot \cancel{g(x)}dx \\ &= \int \frac{c \cdot \pi(x)}{m}dx = \frac{c}{m} \end{aligned}$$

Therefore, the probability of a value being kept for the sample is equal to the target distribution $\pi(x)$

$$
\begin{aligned}
p(X = x \mid z = 1) &= \frac{p(X = x, z = 1)}{p(z = 1)} \\
&= \frac{p(X = x \mid z = 1) \cdot g(x)}{p(z = 1)} \\
&= \frac{\frac{l(x)}{m \cdot g(x)} \cdot g(x)}{\frac{c}{m}} = \pi(x)
\end{aligned}
$$

3.3 How to choose a good envelope

Consider the distribution $\pi(x)$

$$\pi(x) \propto \Phi(x) \cdot I(x > c)$$

$$\text{where } \Phi \sim \mathcal{N}(0, 1)$$

- **Method 1**
  For the first method we choose the Normal Gaussian distribution as the envelope function $g(x) = \Phi(x)$ (Figure 2).

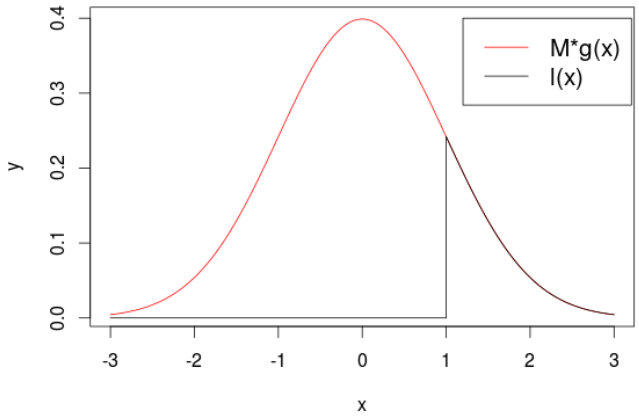

Figure 4: Target and Envelope function(Normal Distribution), $c = 1$.

$$r = \begin{cases} 0 & \text{if } x \leq c \\ 1 & \text{if } x > c \end{cases}$$

For this approach, for high values of c, the acceptance rate is very low. This implies that a lot of points need to be rejected before completing the sample with n points. This approach is not very effective.

- **Method 2**
For the second approach, we consider the exponential density $g(x) = \lambda e^{-\lambda(x-c)}$ where $x \in [c, \infty]$. This

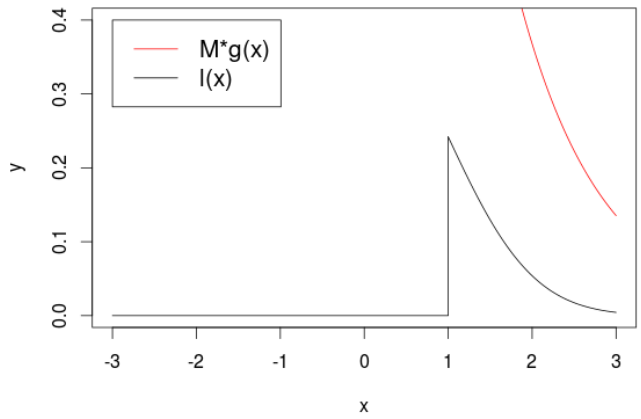

Figure 5: Target and Envelope function(Exponential), lambda = 1.

approach is more effective since we can draw samples from the defined range of x. and for larger c values, it has a higher acceptance rate (Figure 3).

Ok, but what is λ ?
Find the smallest M such that:

$$M \geqslant \frac{\Phi(x)}{g(x)} \qquad \forall x \geqslant c$$

$$M \geqslant \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}}{\lambda e^{-\lambda(x-c)}} = \frac{1}{\lambda\sqrt{2\pi}} e^{-\left(\frac{x^2}{2} - \lambda x - \lambda c\right)}$$

For the expression above, we need to find the values of x that maximizes the whole expression to guarantee the inequality holds. To find this value, we need to find the x that minimizes $\frac{x^2}{2} - \lambda x - \lambda c$ and is equal or greater than c which is $x = \lambda$

$$M = \max_{x>c} \frac{1}{\lambda\sqrt{2\pi}} e^{-\left(\frac{x^2}{2} - \lambda x - \lambda c\right)} = \frac{1}{\lambda\sqrt{2\pi}} e^{\left(\frac{\lambda^2}{2} - \lambda c\right)}$$

Now we have M as a function of λ. To find the smallest M we need to find the value of λ that minimizes

that function.

$$\lambda = \underset{\lambda}{\operatorname{argmin}}\, M(\lambda) = \underset{\lambda}{\operatorname{argmin}}\, \frac{1}{\lambda\sqrt{2\pi}} e^{\left(\frac{\lambda^2}{2} - \lambda c\right)}$$

$$\lambda = c$$

Acceptance rate:

Method	$c = -1$	$c = 2$	$c = 3$
1	0.84	0.02	0.0009
2	0.57	0.88	0.96

From the table above we can see that the exponential envelope function has a higher acceptance rate for larger values of c, therefore it's a more efficient approach.
A good envelope function have the following properties:

- It should be easy to sample from
- It should be easy to construct (find M)

References

[1] J. von Neumann, "Various techniques used in connection with random digits. Monte Carlo methods", *Nat. Bureau Standards*, 12 , pp. 3638, 1951.