STAT440 Computational Inference

Austin Xia

April 10, 2021

Contents

List of Figures

List of Tables

Instructor: Shojaeddin Chenouri
Grading: 15% assignments * 4 + 40% final

1 Chapter 1: Statistical Inference, a brief review

1.1 Introduction

Probability distribution of the phenomenon X can be assumed to be

- parametric
- nonparametric

In parametric approach, we assume pdf of X is known but depends on unknown finite-dimensional parameter θ

$$X \sim f(x|\theta)$$

In nonparametric approach, we assume functional form of pdf of X is not known. In other words, the family of such distribution is infinite-dimensional as one needs to estimate f(x) for any x in the sample space
For parametric modelling, we can evaluate the inferential tool because sample size is finite
non parameter modelling is only justified asymptotically
question regarding page3
main purpose of Statistical analysis is making inference about θ i.e. estimate a function of θ, test a claim about it, predict a event which base on it

1.2 Statistical paradigms

Statistical inference can be classified into 3 major schools: Bayesian, Fisherian and Neyman-Pearson-Wald (NPW)
In Bayesian statistics, we assume θ is random and has a distribution(*prior distribution* $\pi(\theta)$)
then use Bayes's rule to combine prior distribution with data model $f(x|\theta)$
to come up with an updated distribution of θ taking into account the information contained in x about θ
the updated pdf is called the *posterior*.

Theorem 1.2.1 (the Bayes theorem)

$$g(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int_\theta f(x|\theta)\pi(\theta)d\theta}$$

to continue the experimentation, current posterior $g(\theta|x)$ can be treated as future prior
we use it to accumulate information about θ
there is also controversy on chosing prior distribution $\pi(\theta)$ (to get information for θ)
Fisherian and Neyman-Pearson-Wald schools are called frequentist school, they are practical for solving problem but lack a unified framework
Bayesians excel at combining information from different sources *coherently*
in comparision, a common frequentist tactic is to pull problem apart, focusing for sake of objectivity on subset of data, which can be analyzed optimally

1.3 Statistical Principle

Theorem 1.3.1 (estimator)
An estimator is a function that maps the sample space to a set of sample estimate

Definition 1.3.1 (sufficient)
$X \sim f(x|\theta)$, T(X) is sufficient if distribution of X given T(X) does not depend on θ, it means T(X) contains all information in X about θ

$$P(\theta|t(x)) = P(\theta|x)$$

Theorem 1.3.2 (factorization theorem)
T(X) is sufficient iff density of X can be written as

$$f(x|\theta) = g(T(x)|\theta)h(x)$$

Theorem 1.3.3 (sufficiency Principle)
consider a sufficient statistics T, if two observations x, y are such that T(x)=T(y) then x and y must lead to the same inference about θ

Theorem 1.3.4

$$E(X) = E(E(X|Y))$$
$$Var(X) = E(Var(X|Y)) + Var(E(X|Y)) \geq VAR(E(X|Y))$$

from this, we have

$$E(U(X)) = E(E(U(X)|T(X)))$$

Example 1.3.5
$X_1 ... X_k$ be random sample from binomial $Bin(n_i, \theta)$, the joint pdf of $X = X_1 ... X_k$ is

$$f(x|\theta) = \prod_{i=1}^{k} \left(\frac{\theta}{1-\theta}\right)^{\sum_{i=1}^{k} x_i} (1-\theta)^{\sum_{i=1}^{k} n_i}$$

we see $g(T(x)|\theta) = f(x|\theta)$ and h(x)=1
so $T(X) = \sum_{i=1}^{k} X_i$ is a sufficient statistic for θ in this model

likelihood Principle is a consequence of sufficiency Principle.
recall

$$L(\theta|x) = \prod_{i=1}^{n} f(x_i|\theta)$$

Theorem 1.3.6 (Strong likelihood Principle)
the likelihood function $L(\theta|x)$ carries all information x has about θ if x and y are two observations from possibly different models with same parameter θ, and there is constant c

$$L_1(\theta|x) = cL_2(\theta|y)$$

for every θ, then they carry same information about θ and lead to identical inference

Theorem 1.3.7
suppose we are interested in $\nu = g(\theta)$, MLE of ν is

$$g(\hat{\theta}_{MLE})$$

Example 1.3.8
$X \sim Bin(12, \theta)$ with X=9, likelihood of observing x=9 is

$$L_f(\theta|9) = \binom{12}{9}\theta^9(1-\theta)^3, \theta \in (0,1)$$

this likelihood function could be derived easily from pdf $f(x|\theta)$
if there is another likelihood function which is also $c\theta^9(1-\theta)^3$, then the two models has the same θ
questions to implementing likelihood methods:

- what are the observables
- what is the parameter space

Theorem 1.3.9 (conditionality principle)
if 2 experiments E_1 and E_2 on parameter θ are available and if one of these 2 experiments is selected with probability p. the resulting inference on θ should only depend on the selected experiment

1.4 Statistical Inference

let $X_1...X_n$ be generated from distribution $F(x)$, we are interested in parameter $\theta = \theta(F)$
the aim of statistical inference is to use observed data $x_1...x_n$ to make inference about unknow value of θ
3 main approach:

- Neyman-Pearson-Wald inference (pre-experimental)
- Fisherian inference (mostly post-experimental)
- Bayesian inference (post-experimental)

1.4.1 Neyman-Pearson-Wald inference

Usually an estimating function $T = T(X, \theta)$ is constructed
NPW is based on *sampling distribution* of T under repeated sampling, procedures are made before observing data. sampling distribution may be known or approximated using asymptotic argument
examples are

- Unbiased estimation: An estimator U(x) is unbasied for parameter $y(\theta)$ if for all θ

$$E_\theta(U(X)) = \int U(x) f_\theta(x) dx = y(\theta)$$

- Minimum Risk estimator and UMVE's
- Empirical Bayes Inference
- Neyman-Pearson Hypothesis Testing
- Robustness

1.4.2 Fiesherian inference

a point estimate $\hat{\theta}$ of θ is the value of θ which maimizes $L(\theta|x)$
this is equivalent to maxmization of loglikelihood

$$l(\theta|x) = log(L(\theta|x))$$

under some regularity conditions, for large n

$$(\hat{\theta} - \theta) \rightarrow^D N(0, J^{-1}(\theta))$$

where $J(\theta)$ is matrix

$$J(\theta) = E\left(-\frac{d^2 l(\theta|X)}{d\theta^2}\right)$$

likelihood probability statements are based on sampling distribution of maximum likelihood estimator under repeated sampling

Definition 1.4.1
let $x_1 ... x_n$ be random sample with distribution $f(x|\theta)$ If $\theta = (\theta_1, \theta_2)$ and we are only interested in θ_1, one useful method is to get profile-likelihood

$$L_p(\theta_1|x) = L(\theta_1, \hat{\theta_2}(\theta_1))$$

where $\hat{\theta_2}(\theta_1))$ maximizes likelihood for θ_1

1.4.3 Bayesian Inference

here θ is considered to be a random variable on some space Θ with distribution pdf $\pi(\theta)$ called *prior distribution*
this prior distribution represents an expression of belif about an unknown quantity before data are available
the data distribution $F(x|\theta)$ is now treated as conditional distribution of X given θ
when data x are available and pdf is given by $f(x|\theta)$, one use Bayes therem to update their belief

$$g(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d\theta} \propto f(x|\theta)\pi(\theta)$$

question: why Propto?
$g(\theta|x)$ is called *posterior distribution* Bayesian statistics claims that all inference should be based on *posterior distribution*

obviously when $\pi(\theta)$ is constant in θ, Bayesian posterior and likelihood function are identical
for point estimator, Bayesian statistician may either use mean or mode of posterior distribution

$$\hat{\theta} = E_\pi(\theta|x) = \int \theta g(\theta|x) d\theta$$

and

$$\tilde{\theta} = argmax_{\theta \in \Theta} \pi(\theta|x)$$

$\hat{\theta}$ is called Bayes estimate, $\hat{\theta}_B$. $\tilde{\theta}$ is maximum a posterior estimate$\hat{\theta}_M AP$
for confidence interval

$$P(\theta \in C|x) = \int_a^b \pi(\theta|x) d\theta = 1 - \alpha$$

- a prior $\pi(\theta)$ is called *conjugate* if posterior $g(\theta|x)$ belong to same family of distribution as $\pi(\theta)$
  For example: beta distribution is conjugate prior for θ when $X \sim Bin(n, \theta)$
  in other words, prior distribution is same family as posterior distribution
- a uniform prior distribution on bounded interval express an indeifference among all values of the parameter (question can there be unbounded interval)
- a prior is *improper* if $\sum \pi(\theta) d\theta = \infty$
- for improper prior, posterior is still a probability distribution
- Bayesian Inference of posterior is sometimes weighted sum of sample mean and prior mean

2 monte carlo

provide solution to problems by performing statistical experiment on computer
applied to problems with probabilistic and deterministic nature.

Example 1 parallel lines t unit apart, drop a needle length l, probability needle cross a line is
let X be distance between midpoint of needle to nearest line, θ be angle with the line, the needle crosses a line iff

$$X \leq \frac{l * sin(\theta)}{2}$$

$X \sim uniform(0, t/2) \quad \theta \sim uniform(0, \pi)$

$$f(x, \theta) = \frac{2}{\pi t} \text{ if } 0 \leq x \leq t/2, 0 \leq \theta \leq \pi$$

Therefor pr(needle cross line)$\int_0^\pi \int_0^{lsin(\theta)/2} \frac{2}{\pi t} dx d\theta = \frac{2l}{\pi t}$
people used this experiment to approximated π

$$\frac{n_0}{n} \approx\rightarrow \pi \approx \frac{2ln}{tn_0}$$

2.1 Simple MonteCarlo

we are interested in evaluating integral of a sufficiently complicated function g(x)

$$\theta = \int_x g(x)dx$$

to do it with MonteCarlo, we assume X is a rv with range $\mathbf{x}$ and density f, then

$$\theta = E[\delta(X)] = \int_{\mathbf{x}} \delta(x)f(x)dx$$

where $g(x) = \delta(x)f(x)$ monte carlo evalution of the integral consists of
simulating a random sample of $x_1...x_n$ from f and averaging the observed values $\delta(x_1)...\delta(x_n)$
question: why averaging

$$\hat{\theta} = \frac{\sum_{i=1} \delta(X_i)}{n}$$

this is unbiased estimate of θ
however variance of $\hat{\theta}$ is complicate and depend on θ

Theorem 2.1.1 (strong law of large numbers)
if $X_1...X_n$ are independent copies of rv with pdf f, then for any function δ

$$lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \delta(X_i) = E(\delta(X)) = \theta$$

we try to find a useful bound for approximation error

$$\left| \frac{1}{n}\sum_{i=1}^{n} \delta(X_i) - \theta \right|$$

we can use CLT to find a distribution of it.

Theorem 2.1.2 (CLT)
$X_1...X_n$ be independent sample from mean μ var σ^2 then

$$\frac{\sum_{i=1} X_i - n\mu}{\sqrt{n\sigma^2}}$$

has a limiting distribution N(0,1) as n approaches infinity

therefor

$$P(|\frac{1}{n}\sum_{i=1}^{n} \delta(X_i) - \theta| < \epsilon\sqrt{\frac{Var[\delta(X_i)]}{n}}) \approx \phi(\epsilon) - \phi(-\epsilon)$$

we can then construct a 0.95 precent CI

$$\frac{1}{n}\sum \delta(X_i) \pm 1.96\sqrt{(\frac{var[\delta(Xi)]}{n})}$$

well this depend on value of θ as variance of $\hat{\theta}$ is a function of θ however good news:

$$\hat{Var}(\hat{\theta}) = \frac{1}{n}[\frac{1}{n}\sum \delta^2(x_i) - \hat{\theta^2}]$$

when reporting this, we need to include size of the estimation, standard error and confidence interval

Example 2.1.3
esitmate π

$$\pi = \int_0^1 \int_0^1 4I\{x^2 + y^2 \leq 1\}dydx$$

$$\delta(x, y) = 4I\{x^2 + y^2 \leq 1\}$$

and $f(x, y) = 1$ if $x \in (0, 1), y \in (0, 1)$

Theorem 2.1.4 (MonteCarlo algorithm)

- generate n independent X_iY_i with random number from uniform distribution
- for each pair, find sqr distance from (0,0)
- count how many are less than 1 and multiply answer by 4/n

and we get

$$\hat{pi} = \frac{4}{n}\sum_{i=1}^{n} \mathbf{I}\{X_i^2 + Y_i^2 \leq 1\}$$

note

$$I\{X_i^2 + Y_i^2 \leq 1\} \sim Bernoulli(\pi/4)$$

$$\sum_{i-1}^{n} I\{X_i^2 + Y_i^2 \leq 1\} \sim Bin(n, \pi/4)$$

2.2 method of composition

$f(y|x)$
we interested in obtain random sample $Y_1...Y_m$ from marginal distribution

$$k(y) = \int_X f(y|x)g(x)dx = E_g[f(y|X)]$$

to do so we use method of composition as follows

- draw x* from g(x)
- draw y* from $f(y|x^*)$

repeat m times we obtain x_m, y_m
form joint pdf $h(x, y) = f(y|x)g(x)$

similarly,

$$k(y) = \int f(y|x, s, t)g(x|s, t)h(s|t)p(t)dxdsdt$$

to sample from marginal probability density function k(y),

- draw t* from p(t)
- Draw s* from $h|t^*$
- Draw x* from $g(x|s^*t^*)$
- draw y* from $f(x^*s^*t^*)$

while y* is an observation from marginal k(y)

Example 2.2.1 (predictive distribution)
$y_{obs} = (y_1...y_n)^T$ we interested in y^f, a future observation. suppose $\pi(\theta|y_{obs})$ is posterior distribution, then predictive distribution is

$$p(y^f|y_{obs}) = \int p(y^f|y_{obs},\theta) * \pi(\theta|y_{obs})d\theta$$

3 pseudo random number generation

$x \in \{a_1...a_k...\}$ with probability p_k partition interval (0,1] into subintervals, $I_k = (F_{k-1}, F_k), F_0 = 0, f_k = p_1 + ...p_k$
each subinterval coresponds to a single value for X and after observing generated value from U(0,1), one verifies which interval U belong so

$$X = a_i \text{if } U \in I_i$$

3.1 inverse method for gnerating continous random number

background:

$$P(F(X) \leq F(t)) = P(x \leq t) = F(t)$$

Theorem 3.1.1

- if F(x) continous, U=F(X) is uniform on [0,1]
  proof: F(t)=u,
  $$P(F(X) \leq u) = P(F(X) \leq F(t)) = u$$
- $F^{-1}(U)$ has distribution F(x)
  proof $u \leq F(t) \equiv F^{-1}(u) \leq t$
- even if F(x) not continous, $P(F(x) \leq t) \leq t$ still hold

Example 3.1.2

$$F(x) = 1 - e^{-x} \leftrightarrow F^{-1}(u) = -ln(1-u)$$

then $Y = -\lambda ln(U)$ is exponentially distributed with mean λ

3.2 rejection-acceptance method

$$e(x) = Mg(x) \geq f(x) \quad \text{e(x) is envelop function}, M \geq 1$$

The rejection acceptance algorithm:

- Sample $Y \sim g$

- Sample $U \sim U(0,1)$
- Reject Y if $U \geq f(Y)/e(Y)$ go to step 1
- Accept if otherwise, X=Y

we will have $P(X \leq x) = P(Y \leq x | X = Y) = P(Y \leq x | U \leq f(Y)/e(Y)) = F(x)$
Remarks:

- acceptance probabiilty is 1/M, when M is large, inefficient
- g(x) resembles f(x) , then M small.
- if g(x) shorter tail than f(x). may not exist suitable M

4 Importance Sampling

weighted version of simple Monte Carlo

5 control variate

if we want

$$\theta = E_f(\delta(X)) = \int \delta(x) f(x) dx$$

it's easy to obtain

$$\theta^* = E_f(\mu(X)) = \int \mu(x) f(x) dx$$

then for any constant $\alpha, \hat{\theta} + \alpha(\hat{\theta}^* - \theta^*)$ is unbiased estimator for θ
to get best α, we minimize

$$Var(\hat{\theta} + \alpha(\hat{\theta}^* - \theta^*))$$

we have variance reduction over $\hat{\theta}_{MC}$ as

$$minVar(\hat{\theta}_{CV}) = Var(\hat{\theta}_{MC})[1 - Corr^2(\hat{\theta}_{MC}, \hat{\theta}^*_{MC})]$$

6 MCMC

Definition 6.0.1 (reversibility of MC)
$q_{ij} = pr(Z_n = j | Z_{n-1} = i) = p_{ji} \frac{\pi_j}{\pi_i}$
if X_n and Z_n are from $(-\infty, \infty)$ then

6.1 metropolis algorithum

- the proposal probability q_{ij} are symmetric
- Hasting extended Metropolis algorithm to more general without sysmmetric requirement
- acceptance probability is

$$\alpha_{ij} = min\{\frac{\pi_j}{\pi_i}, 1\}$$

the algorithm:

- set $x_{n-1} = i, i \in S$
- generate j from $q_{ij}, j \in S$
- set $r = \frac{\pi_j}{\pi_i}$
- if $r \geq 1$ set $x_n = j$ otherwise generate $u \sim U(0,1)$
- if $u < r$ set $x_n = j$ else set $x_n = x_{n-1}$
- set n=n+1 back to step 1

6.2 metropolis algorithum

- the proposal probability q_{ij} are symmetric
- Hasting extended Metropolis algorithm to more general without sysmmetric requirement
- acceptance probability is

$$\alpha_{ij} = min\{\frac{\pi_j q_{ji}}{\pi_i q_{ij}}\}$$

the algorithm:

- set $x_{n-1} = i, i \in S$
- generate j from $q_{ij}, j \in S$
- set $r = \frac{\pi_j q_{ji}}{\pi_i q_{ij}}$
- if $r \geq 1$ set $x_n = j$ otherwise generate $u \sim U(0,1)$
- if $u < r$ set $x_n = j$ else set $x_n = x_{n-1}$
- set n=n+1

the detailed balance condition is satisfied $\pi_i p_{ij} = \pi_j p_{ji}$, so πj is stationary distribution

6.2.1 metropolis hastings for bayesian

$$\pi(\theta|y) \propto f(y|\theta)\pi(\theta)$$

so acceptance probability is

$$\alpha(\theta_n, \theta^*) = min\{\frac{\pi(\theta^*|y)q(\theta^*, \theta_n)}{\pi(\theta_n|y)q(\theta_n, \theta^*)}, 1\}$$

6.2.2 independent metropolis hastings

q(x,y) depends only on y
acceptance probability is $\alpha(x,y) = min\{\frac{\pi(y)g(x)}{\pi(x)g(y)}, 1\}$

6.2.3 random walk metropolis hastings

algorithm

- n=0, start with x_n
- generate ϵ_n from $g()$ which is usually $N(0, \sigma^2 I)$, and $u \sim U(0,1)$
- set $y = x_n + \epsilon_n$
- if $u \leq \alpha(x_n, y)$ set $x_{n+1} = y$. else set $x_{n+1} = x_n$
- set n=n+1, go to step 1
- and we have $\{x_n ... x_N\}$

here $\alpha(x,y) = min\{\frac{\pi(y)}{\pi(x)}, 1\}$

6.3 single-component metropolis hastings

- $X = (X_1, ..., X_d) \sim \pi(x)$
- if we use metropolis-hasting, we generate y from x by q(x,y)
- alternative is generate sample component-wise
- each component is a single dimension in this case

$$x = (x_1 ... x_d)$$

- each component is updated one by one by seperate metropolis-hasting steps
- at ith step, yi is generated from $q_i(x_i, y_i)$ which means q() changes
- $\alpha_i(x_i, y_i) = min\{\frac{\pi_i(y_i) q_i(y_i, x_i)}{\pi_i(x_i) q_i(x_i, y_i)}, 1\}$
- if accepted, $x_i^{n+1} = y_i$ otherwise $x_i^{n+1} = x_i^n$
- the remaining components of x_n not changed in step i
- after repeating for i=1,2..d, the entire x_n is updated

some details: $\pi_i(x_i) = \pi(x_i | x_{-i}) = \pi(X) / \int \pi(X) dx_i$
$\pi(x)$ is uniquely determined by $\pi_i(x_i)$ i from 1 to d

6.4 Gibbs

- gibbs sampler is special kind of single component metropolis-hasting sampling
- $\alpha(x,y) = 1$
- one only consider univariate conditional distribution
- easier to simulate than joint distriubtion
- that is one simulate d random variate sequentailly from the d univate conditions

the proposal distribution is $q_i(x_i, y_i) = \pi_i(y_i)$
$\pi_i(y_i) = \pi(y_i|x_{-i})$
algorithm:

- initailize $x^0 = x_1^0 ... x_d^0$ n=0
- $$x_1^{n+1} \sim \pi(x_1|x_2^n, ...x_d^n)$$
  $$x_2^{n+1} \sim \pi(x_2|x_1^n, ...x_d^n)$$
- this create x^{n+1}
- increment n, go to step 2

6.4.1 metropolis within gibbs

- in gibbs we assume we know how to draw from $\pi_i(y_i|x_{-i}), i = 1...d$
- if we do not know how, we use a Metropolis hasting step
- suppose (y,x) is proposal distribution for jth component
- see section 3.4.6

7 chapter 5

Definition 7.0.1 (global optimizer)
$\theta_0 \in R^p$ is global minimizer of h if for all $\theta \in R^p h(\theta_0 \leq h(\theta)$

Definition 7.0.2 (local optimizer)
$\theta_0 \in R^p$ is local minimizer of h if for all $\theta \in N(\theta_0) h(\theta_0) < h(\theta)$

Definition 7.0.3
$\delta h(\theta) = (\frac{dh}{d\theta_1} ... \frac{dh}{d\theta_p})$

Theorem 7.0.1 (first order necessary condition)
if θ_0 is local minimizer, h is continously differentiable in neibourhood of θ_0 then $\delta h(\theta_0) = 0$

Theorem 7.0.2 (second order neccessary)
θ_0 is local minimizer of h , $H_h(\theta)$ is continous, then $\delta h(\theta_0) = 0, H_h(\theta_0)$ is positive

Theorem 7.0.3 (second order sufficient)
$H_h(\theta)$ is continous,$\delta h(\theta_0) = 0, H_h(\theta_0)$ is positive definite then θ_0 is local minimizer

Definition 7.0.4 (convex set)
a set C is convex if for any 2 points in C,

$$\lambda\theta + (1-\lambda)\mu \in C$$

Definition 7.0.5 (convex function)
a function is convex if for all θ, μ

$$h(\lambda\theta + (1-\lambda)\mu) \leq \lambda h(\theta) + (1-\lambda)h(\mu)$$

Theorem 7.0.4
for a convex function, any local minimizer of h is global minimizer
in addition if h is differentiable then any stationary point θ of h is a global minimizer

Theorem 7.0.5
let h be twice differentiable function on open convex set, if second derivative is positive semi-definite, then h() is convex, If H() is positive definite, h() is strictly convex

7.1 direct search method

do not require explicit evaluation of any partial derivatives

7.1.1 goden search method

to apply golden search we require h is well defined on [a,b] and unimodel

Definition 7.1.1 (unimodal function)
H() is unimodal on [a,b] if there is unique p s.t. h() is decerasing on[a,p], h() is increasing on [p,b]

7.2 nelder mead or simplex

Definition 7.2.1
a simplex in p dim is a feometrical object consisting of p+1 points(or vertices) and all their interconnecting line segments, polygonal faces etc.
1d simplex is interval 2d is triangle each step we reject one vertex with highest value

algrithum:

- initial triangle BGW: three initial vertices B, G, W; $h(B) \leq h(G) \leq h(w)$
- $M = \frac{B+G}{2} = (\frac{\lambda_1+\mu_1}{2}, \frac{\lambda_2+\mu_2}{2})$
- reflection using point R, by finding M=(G+B)/2, R=2M-W
- if $h(R) \leq h(W)$ we succeed, if so we extend R to E. if $h(E) \leq h(R)$ then $E = 2R - M$

- if h(R)=h(W), test another point C1, C2
- if $h(C) \geq h(W)$ the point G W shrink toward B replace G with M, W with S

7.3 mimization using derivatives

7.3.1 bracketing method

identify 3 pionts $\theta_0, \theta_1 = \theta_0 + \alpha, \theta_2 = \theta_0 + 2\alpha$ s.t. $h(\theta_0) > h(\theta_1), h(\theta_1) < h(\theta_2)$

7.4 quadratic appocimation

if we know function value at 3 points we approximate by $g(\theta) = A\theta^2 + B\theta + C$ solve it and find $\theta_m in = -B/2A$

7.5 linear search

find a direction p and decide how far to move along the direction

7.6 steepest descend

evaluate gradient, compute direction, perform single parameter minimiazation, construct next point, perform test termination

7.7 newton raphson

$$\theta_{k+1} = \theta_k - \frac{e(\theta_k)}{e'(\theta_k)}$$

algorithm:

- start from θ^0
- suppose after kth iteration we have θ^k
- compute $H_h^{-1}(\theta^k)$ and $\delta h(\theta^k)$
- new update $\theta^{k+1} = \theta^k - H^{-1}\theta^k \delta h(\theta^k)$

7.8 conjugate gradient

does not require 2nd derivatives
go in direction of the 1st derivative

8 stochastic optimazation

stochastic optimazation methods randomly search θ
for example generate θ uniformly and get minimum of $h\theta$

8.1 simulated annealing

used for function with multiple peaks. it will recieve a reasonable downhill move prob to explore the entire space
as process proceed, we decrease pr of downhill, related to thermal dynamics
if temperature decrease slowly, the cooling process is annueing
algorithm

- start with θ_0 and an anealing schedule $T_1 \geq ...T_K$
- for k = 1 to K j=1 to N_k process θ from $q(\theta_{j-1}, y)$
- generate u, if $u < min\{\frac{\pi_k(\theta)}{\pi_k(\theta_{j-1})}, 1\}$ then set θ_j to be
- one of θ or θ_{j-1}
- set $\theta_0 = \theta_{N_k}$

8.2 GA genetic algorithum

based on principle of survial of fittest
GA works with *population* of condidate solution

9 EM Algorithum

$f_{com}(y_{com}; \theta) = f_{obs}(y_{obs}; \theta) * f_{mis|obs}(y_{obs}; \theta)$
add expectation $E_{mis|obs;\theta_0}$ to the zbove formula

$$l_{obs}(\theta; y_{obs}) = Q(\theta; \theta_0) - H(\theta; \theta_0)$$

Lemma 1 *$H(\theta; \theta_0)$ is maximized with resepct to θ if $\theta = \theta_0$*
for $\theta \neq \theta_0$ $H(\theta; \theta_0) < H(\theta_0; \theta_0)$
if we let $\theta_1 = max_\theta Q(\theta; \theta_0)$

EM algorithm:

- start from θ_0
- in step k=0,1... compute $Q(\theta; \theta^{(k)})$
  which is $E_{mis|obs}; \theta^k(l_{com}(\theta; Y_{com}|y_{obs}))$
- M-step : Find $\theta^{k+1}; \theta^k > Q(\theta; \theta^k)$ for all θ
  continue untill difference of $l_bos(\theta^{k+1}; y_{obs} - L_obs(\theta^k; y_{obs}))$ is low