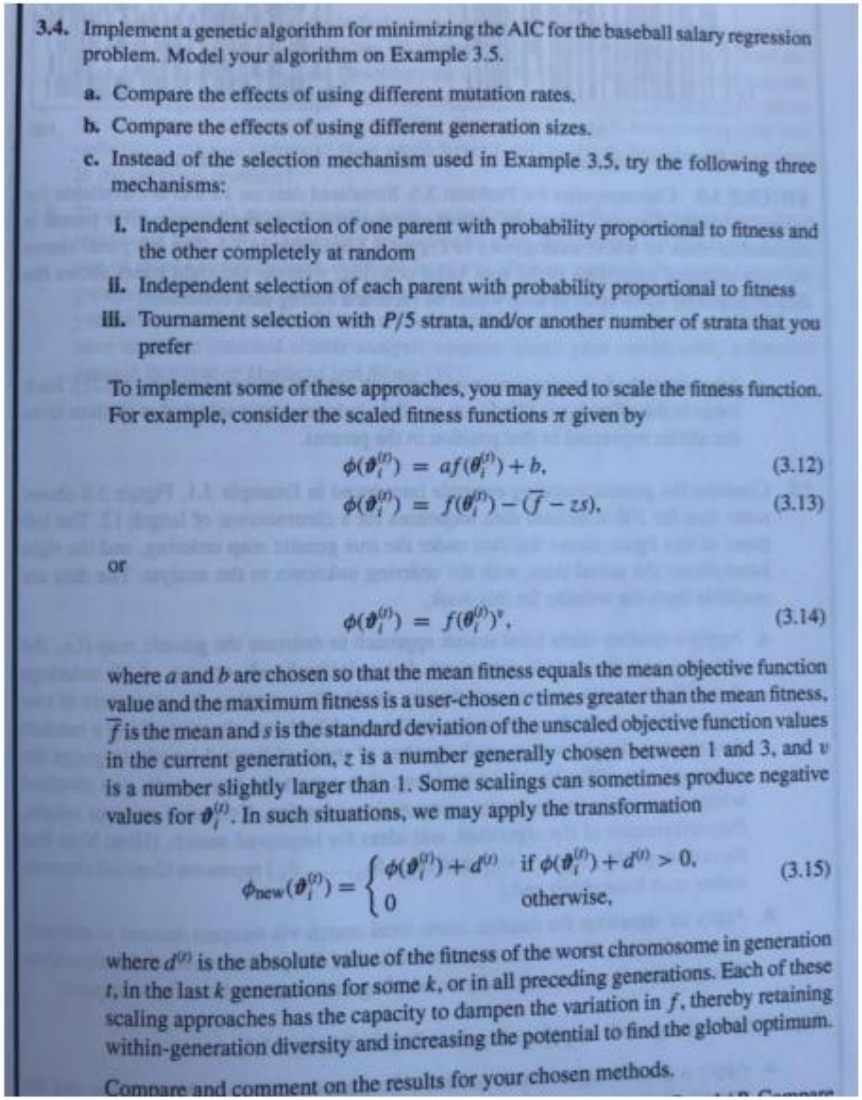
UiO : Matematisk institutt

Det matematisk-naturvitenskapelige fakultet

STK-4051/9051 Computational Statistics Spring 2024 Ex3

Instructor: Odd Kolbjørnsen, oddkol@math.uio.no



3.4. Implement a genetic algorithm for minimizing the AIC for the baseball salary regression problem. Model your algorithm on Example 3.5.

a. Compare the effects of using different mutation rates.

b. Compare the effects of using different generation sizes.

c. Instead of the selection mechanism used in Example 3.5, try the following three mechanisms:

i. Independent selection of one parent with probability proportional to fitness and the other completely at random

ii. Independent selection of each parent with probability proportional to fitness

iii. Tournament selection with $P/5$ strata, and/or another number of strata that you prefer

To implement some of these approaches, you may need to scale the fitness function. For example, consider the scaled fitness functions π given by

$$\phi(\boldsymbol{\vartheta}_i^{(t)}) = af(\boldsymbol{\theta}_i^{(t)}) + b, \tag{3.12}$$

$$\phi(\boldsymbol{\vartheta}_i^{(t)}) = f(\boldsymbol{\theta}_i^{(t)}) - (\overline{f} - zs), \tag{3.13}$$

or

$$\phi(\boldsymbol{\vartheta}_i^{(t)}) = f(\boldsymbol{\theta}_i^{(t)})^v, \tag{3.14}$$

where a and b are chosen so that the mean fitness equals the mean objective function value and the maximum fitness is a user-chosen c times greater than the mean fitness, $\overline{f}$ is the mean and s is the standard deviation of the unscaled objective function values in the current generation, z is a number generally chosen between 1 and 3, and v is a number slightly larger than 1. Some scalings can sometimes produce negative values for $\boldsymbol{\vartheta}_i^{(t)}$. In such situations, we may apply the transformation

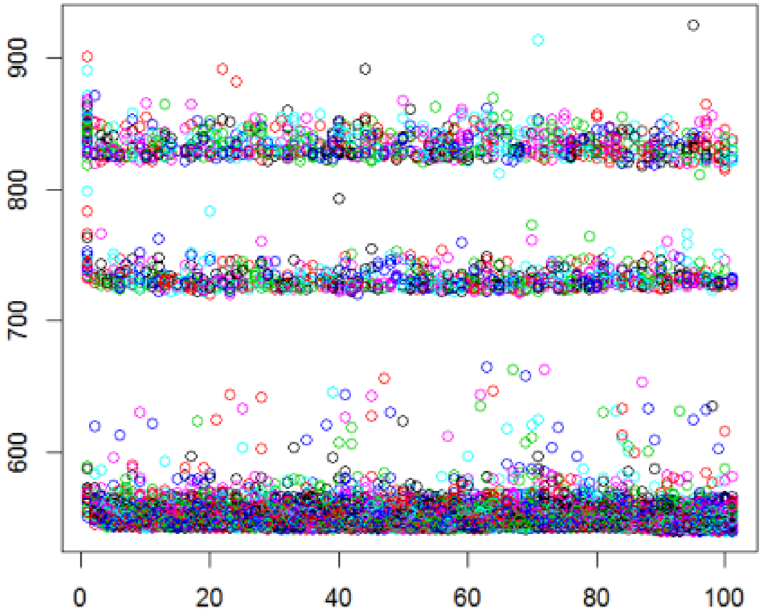
$$\phi_{\text{new}}(\boldsymbol{\vartheta}_i^{(t)}) = \begin{cases} \phi(\boldsymbol{\vartheta}_i^{(t)}) + d^{(t)} & \text{if } \phi(\boldsymbol{\vartheta}_i^{(t)}) + d^{(t)} > 0, \\ 0 & \text{otherwise,} \end{cases} \tag{3.15}$$

where $d^{(t)}$ is the absolute value of the fitness of the worst chromosome in generation t, in the last k generations for some k, or in all preceding generations. Each of these scaling approaches has the capacity to dampen the variation in f, thereby retaining within-generation diversity and increasing the potential to find the global optimum.

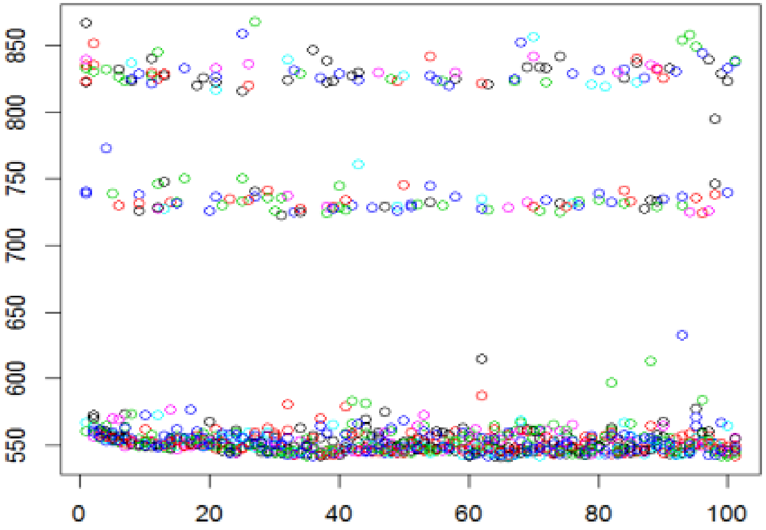
Compare and comment on the results for your chosen methods.

3.4 100 generations p1= AIC fit p2= AIC

Mu=10% P=100

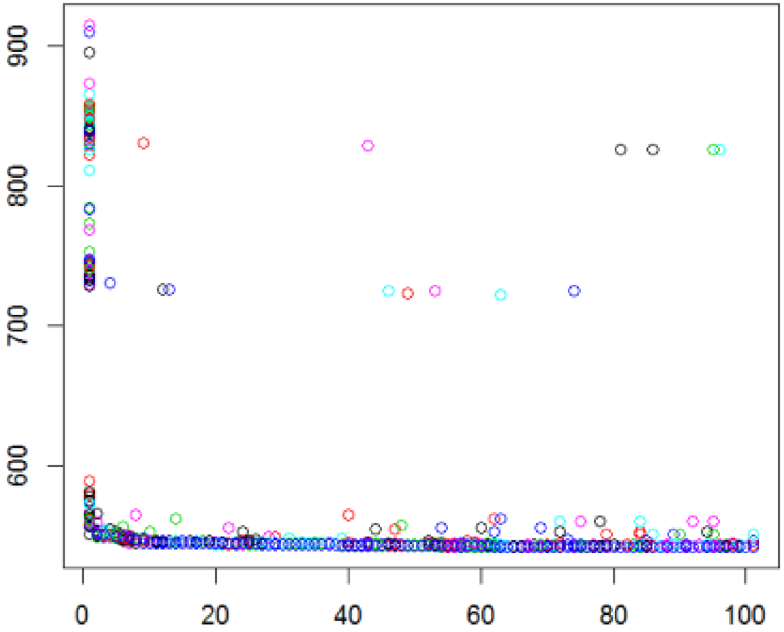


Mu=10% P=10

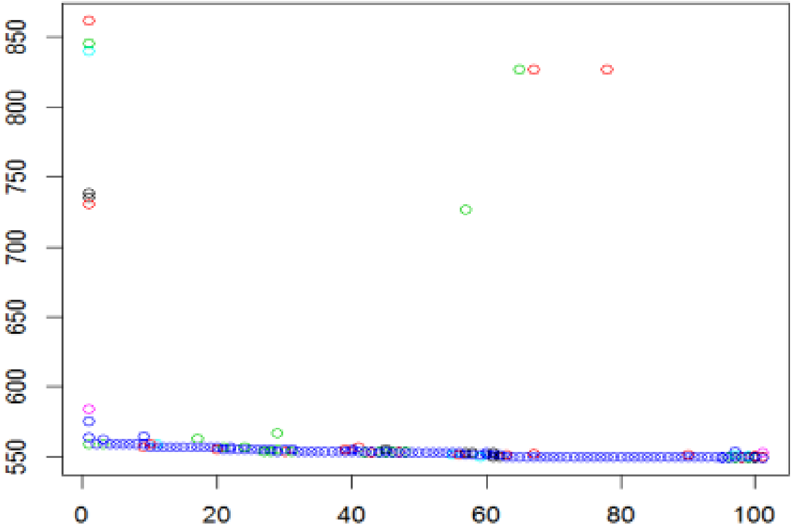


3.4 100 generations p1= AIC fit p2= AIC

Mu=0.1% P=100

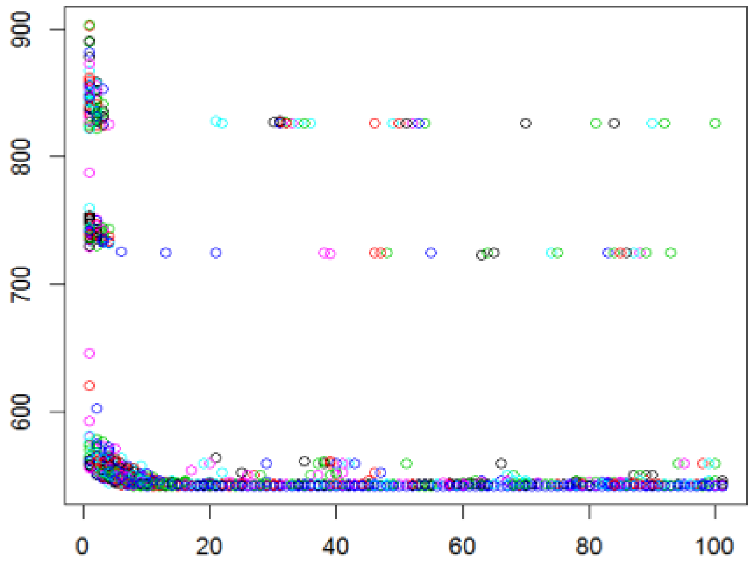


Mu=0.1% P=10

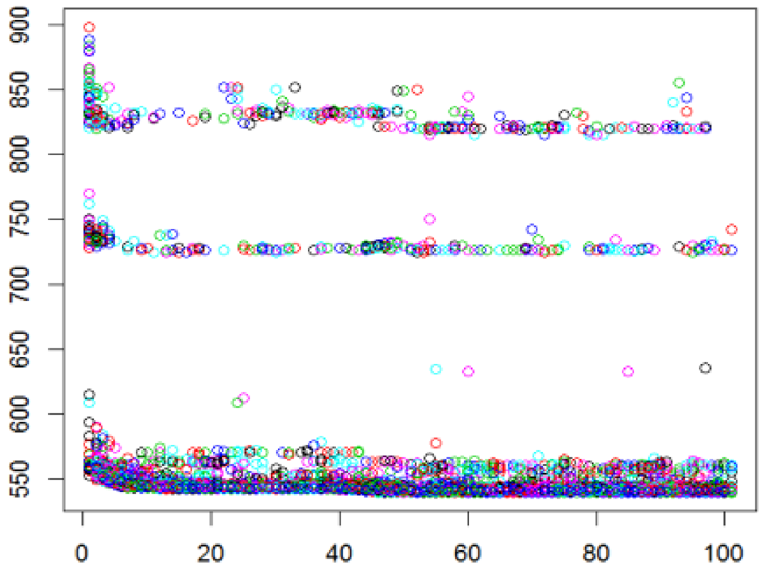


3.4 100 generations p1= AIC fit p2= rand

Mu=0.1% P=100

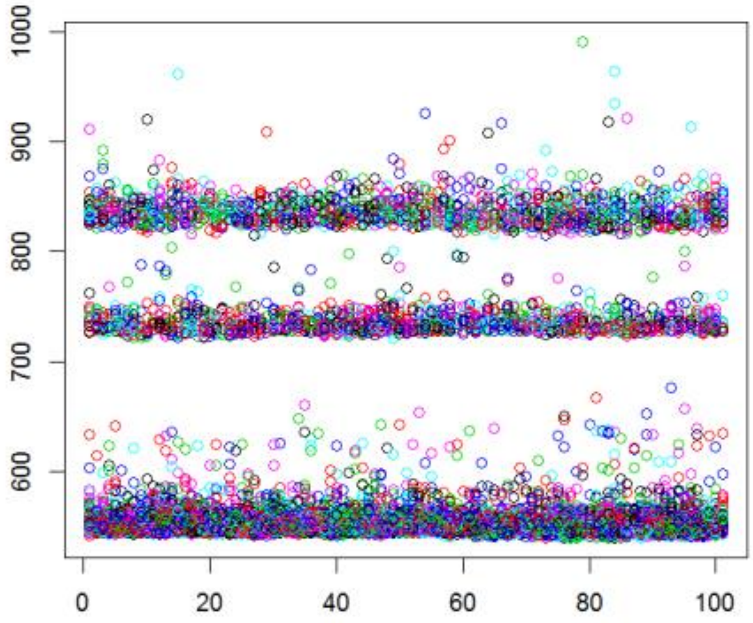


Mu=1% P=100

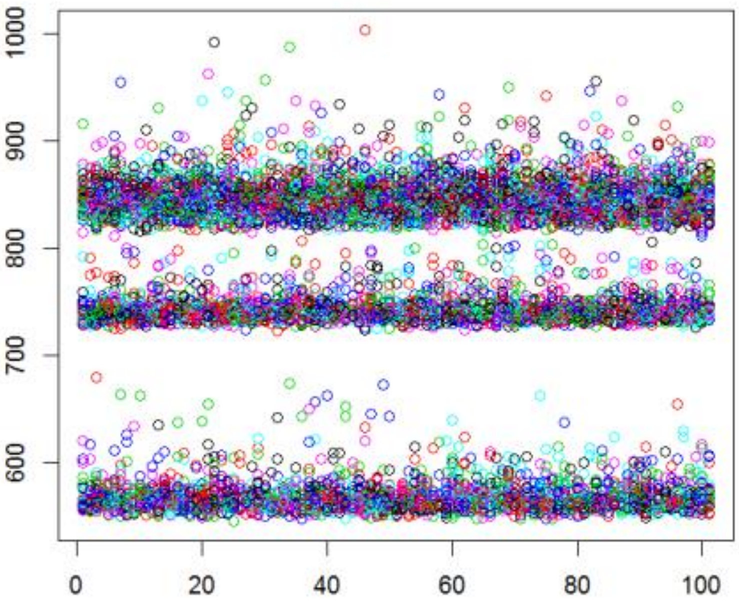


3.4 100 generations p1= AIC fit p2= rand

Mu=10% P=100

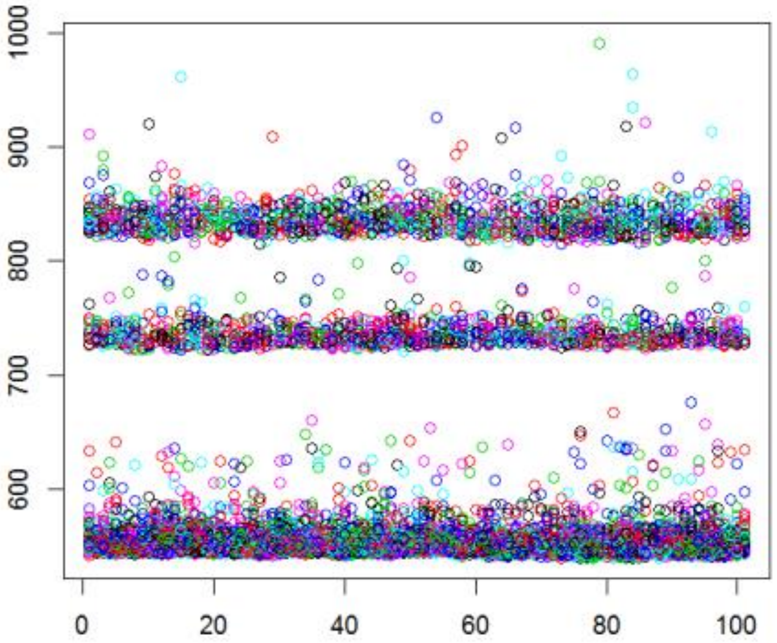


Mu=50% P=100

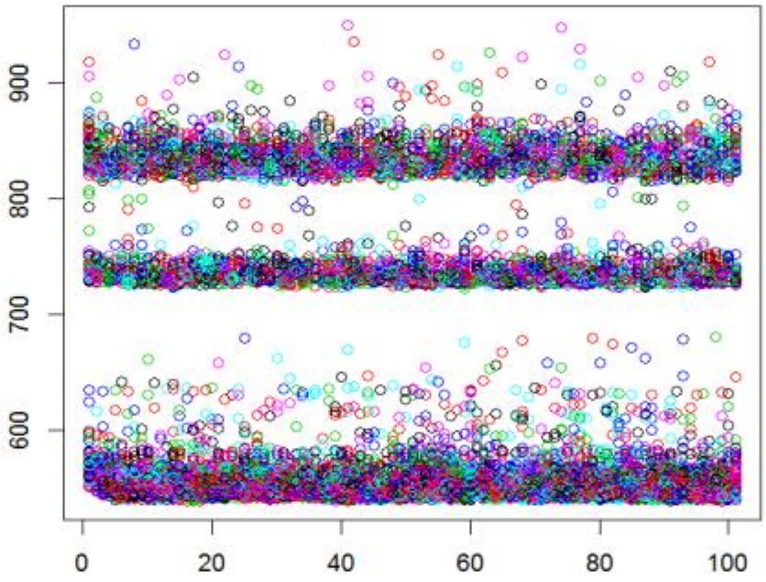


3.4 100 generations p1= AIC fit p2= rand

Mu=10% P=100

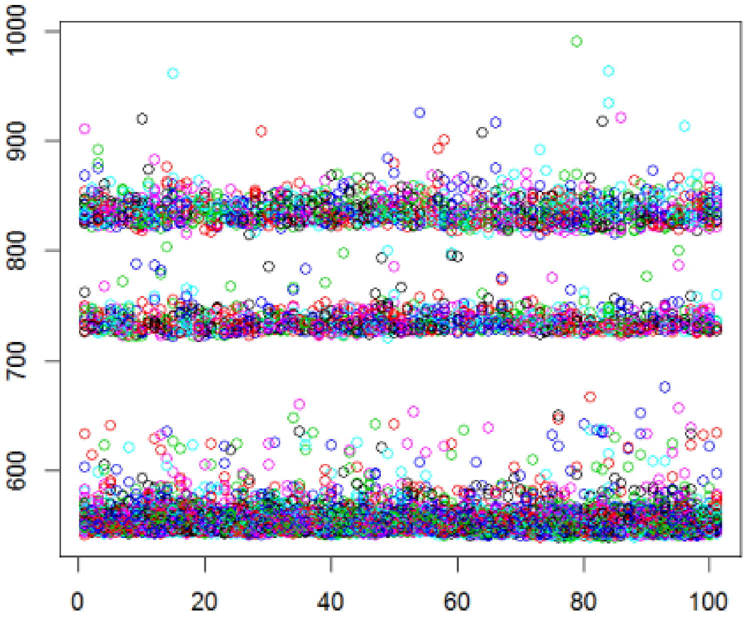


Mu=10% P=200

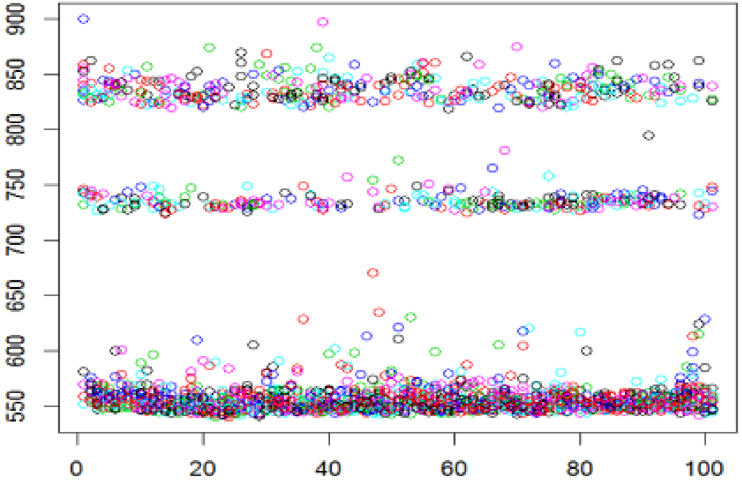


3.4 100 generations p1= AIC fit p2= rand

Mu=10% P=100

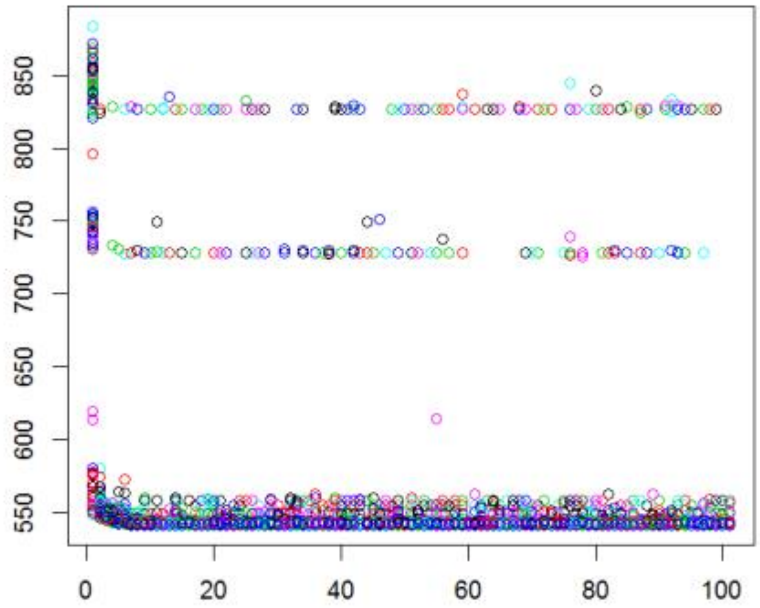


Mu=10% P=20

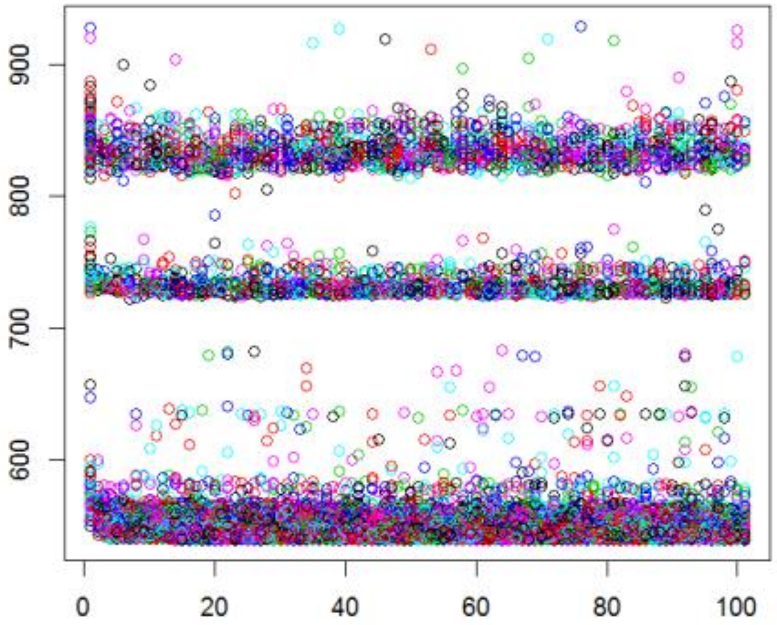


Tournament

Mu=1%   P=100

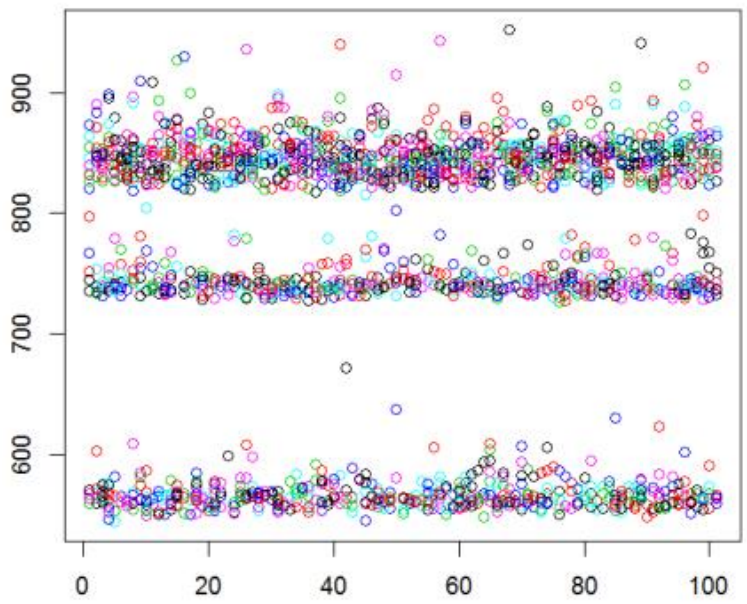


Mu=10% P=200

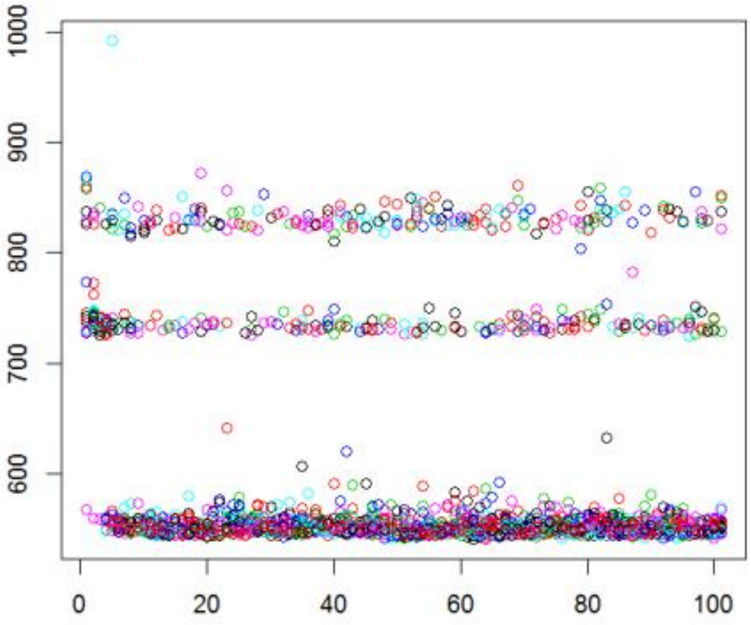


Tournament

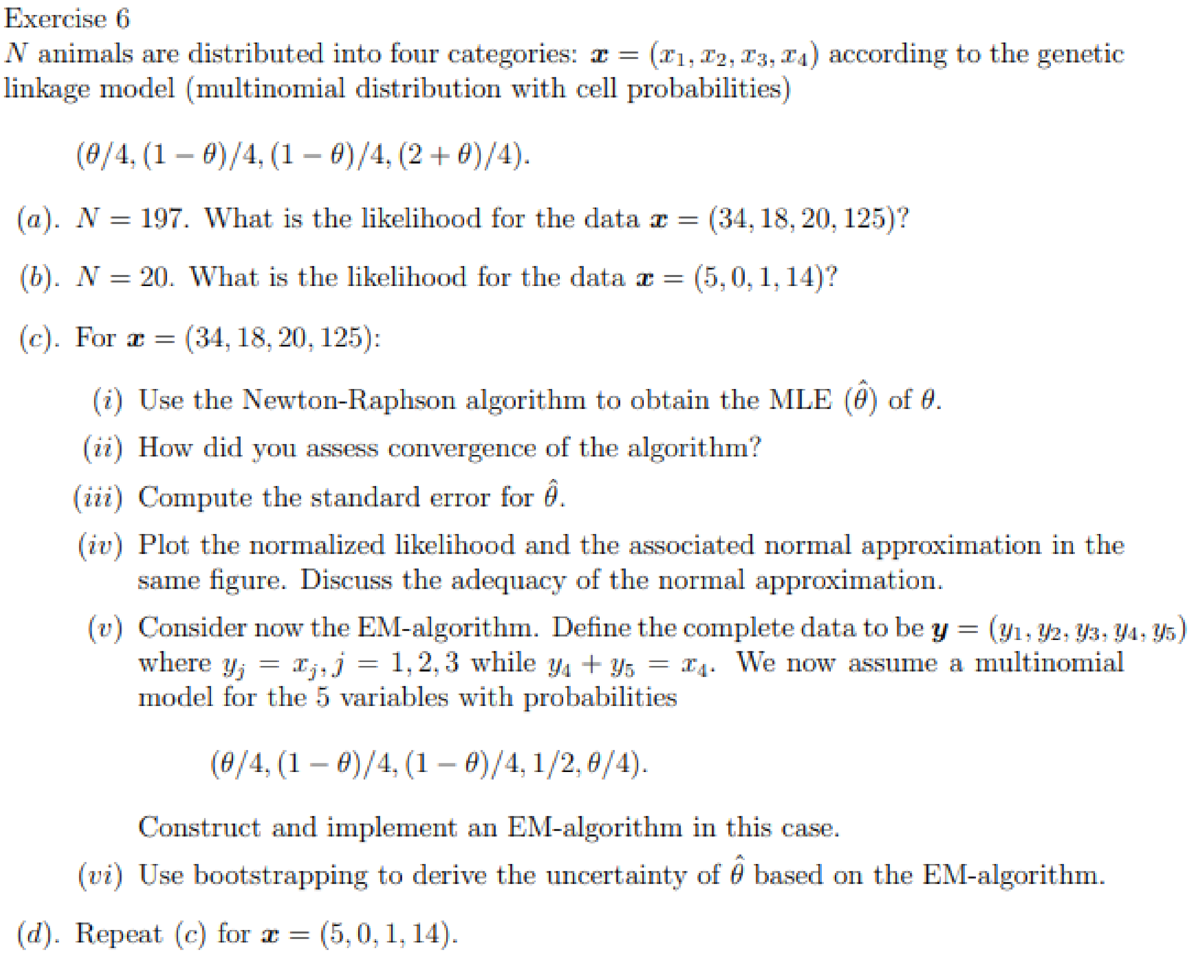
Mu=50% P=20



Mu=10% P=20



Genetic linkage is the tendency of DNA sequences that are close together on a chromosome to be inherited together



Exercise 6

N animals are distributed into four categories: $\boldsymbol{x} = (x_1, x_2, x_3, x_4)$ according to the genetic linkage model (multinomial distribution with cell probabilities)

$$(\theta/4, (1-\theta)/4, (1-\theta)/4, (2+\theta)/4).$$

(a). $N = 197$. What is the likelihood for the data $\boldsymbol{x} = (34, 18, 20, 125)$?

(b). $N = 20$. What is the likelihood for the data $\boldsymbol{x} = (5, 0, 1, 14)$?

(c). For $\boldsymbol{x} = (34, 18, 20, 125)$:

(i) Use the Newton-Raphson algorithm to obtain the MLE $(\hat{\theta})$ of θ.

(ii) How did you assess convergence of the algorithm?

(iii) Compute the standard error for $\hat{\theta}$.

(iv) Plot the normalized likelihood and the associated normal approximation in the same figure. Discuss the adequacy of the normal approximation.

(v) Consider now the EM-algorithm. Define the complete data to be $\boldsymbol{y} = (y_1, y_2, y_3, y_4, y_5)$ where $y_j = x_j, j = 1, 2, 3$ while $y_4 + y_5 = x_4$. We now assume a multinomial model for the 5 variables with probabilities

$$(\theta/4, (1-\theta)/4, (1-\theta)/4, 1/2, \theta/4).$$

Construct and implement an EM-algorithm in this case.

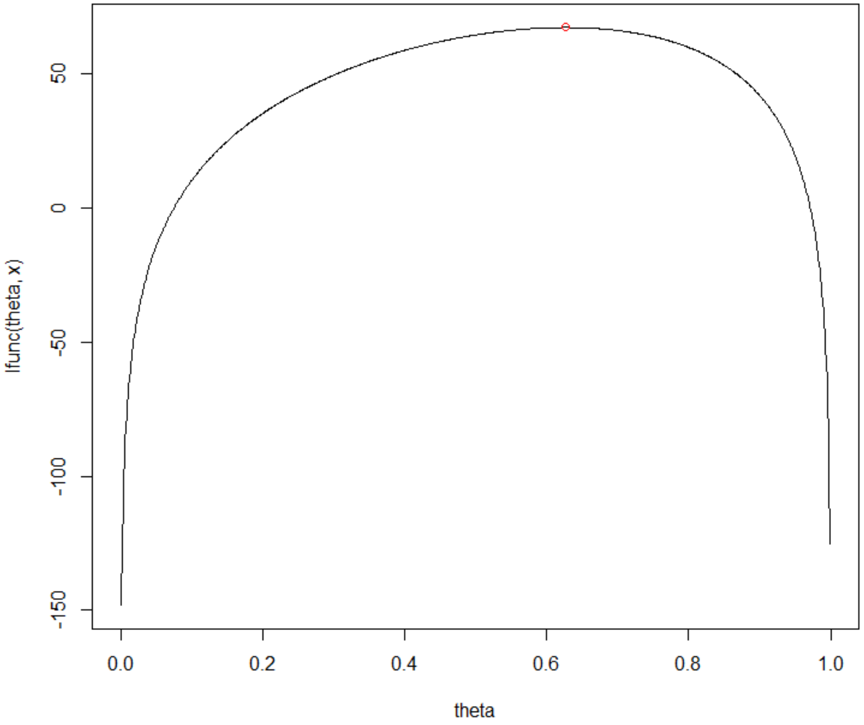
(vi) Use bootstrapping to derive the uncertainty of $\hat{\theta}$ based on the EM-algorithm.

(d). Repeat (c) for $\boldsymbol{x} = (5, 0, 1, 14)$.

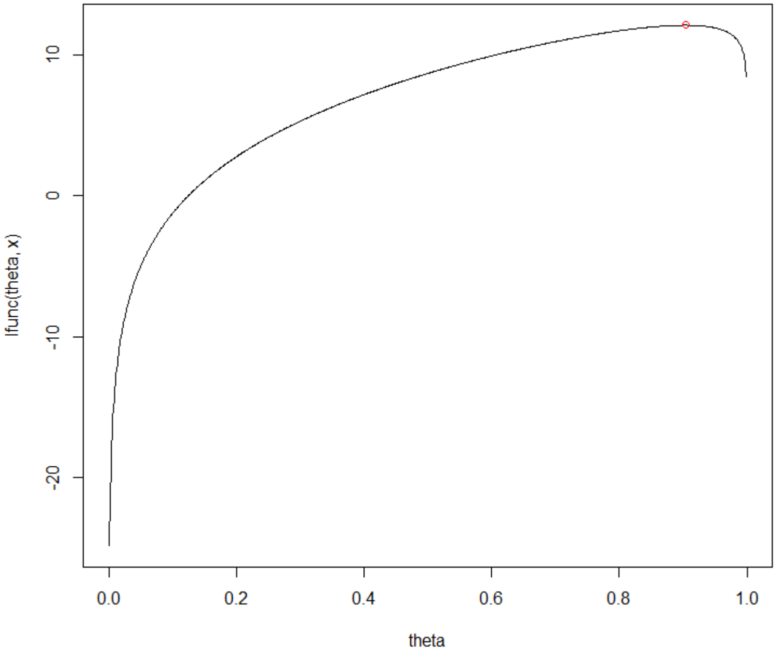
(a). Likelihood-function

$$L(\theta) = \binom{n}{x_1\, x_2\, x_3\, x_4} \frac{1}{4^n} \theta^{x_1} (1-\theta)^{x_2+x_3} (2+\theta)^{x_4}$$
$$\ell(\theta) = \text{Const} + x_1 \log(\theta) + (x_2+x_3)\log(1-\theta) + x_4 \log(2+\theta)$$

For $\boldsymbol{x} = (34, 18, 20, 125)$:

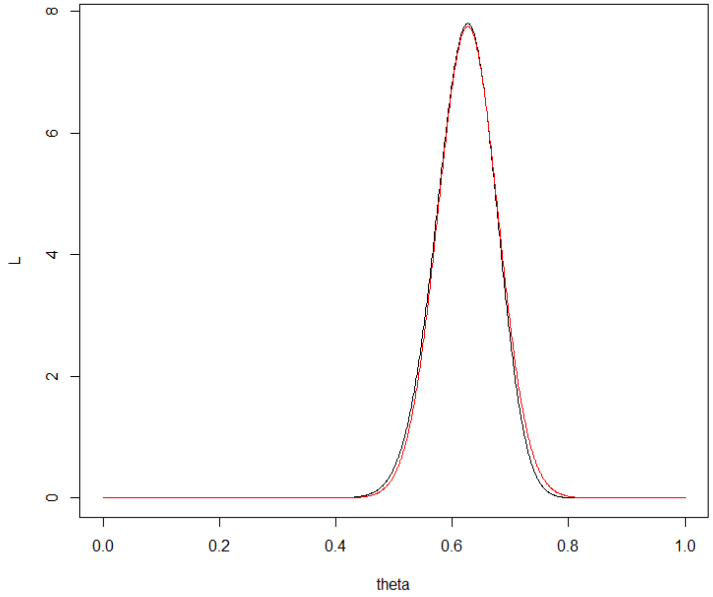


(b). For $\boldsymbol{x} = (5, 0, 1, 14)$:

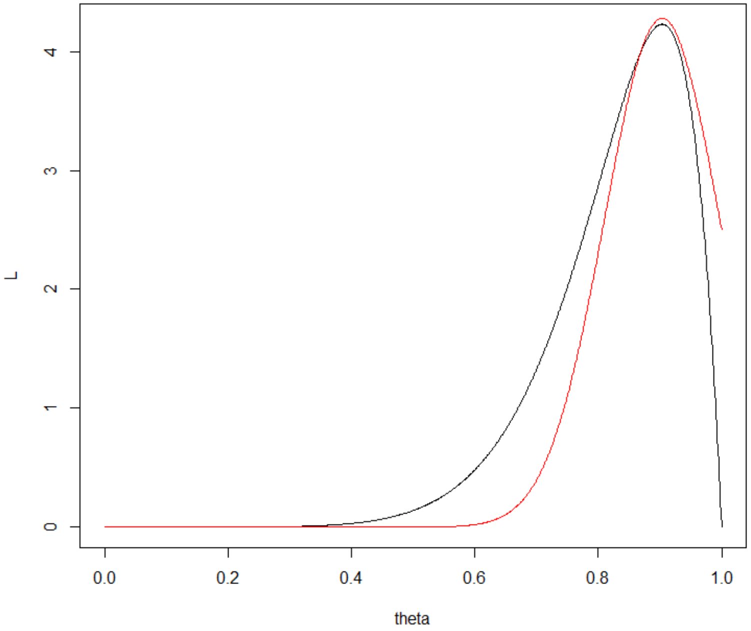


likelihood vs normal approximation

For $\boldsymbol{x} = (34, 18, 20, 125)$:



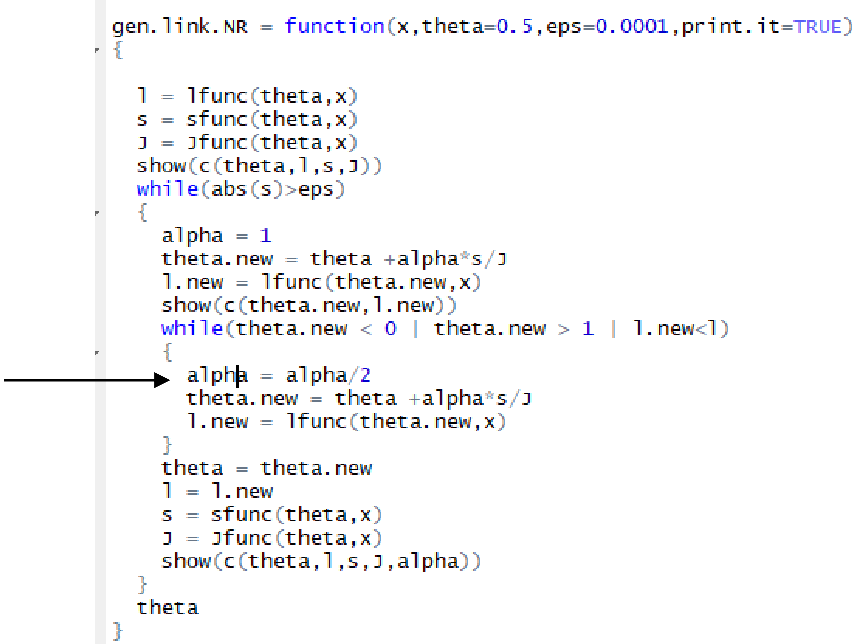
(b). For $\boldsymbol{x} = (5, 0, 1, 14)$:



(*c*). We have

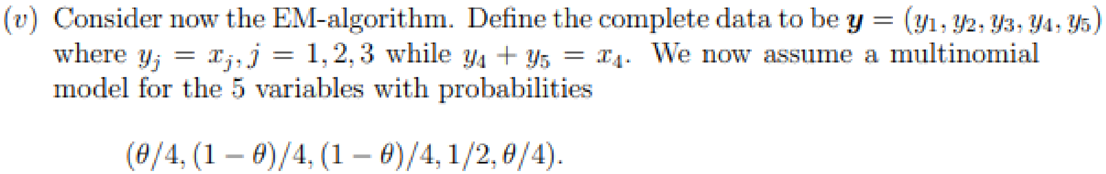
$$s(\theta) = \ell'(\theta) = \tfrac{x_1}{\theta} - \tfrac{x_2+x_3}{1-\theta} + \tfrac{x_4}{2+\theta}$$
$$J(\theta) = -\ell''(\theta) = \tfrac{x_1}{\theta^2} + \tfrac{x_2+x_3}{(1-\theta)^2} + \tfrac{x_4}{(2+\theta)^2}$$

defining the Newton-Raphson method. See the R-script ***genetic_linkage.R*** for implementation. Note that the trick of halving is needed for the second data example.



```
gen.link.NR = function(x,theta=0.5,eps=0.0001,print.it=TRUE)
{

  l = lfunc(theta,x)
  s = sfunc(theta,x)
  J = Jfunc(theta,x)
  show(c(theta,l,s,J))
  while(abs(s)>eps)
  {
    alpha = 1
    theta.new = theta +alpha*s/J
    l.new = lfunc(theta.new,x)
    show(c(theta.new,l.new))
    while(theta.new < 0 | theta.new > 1 | l.new<l)
    {
      alpha = alpha/2
      theta.new = theta +alpha*s/J
      l.new = lfunc(theta.new,x)
    }
    theta = theta.new
    l = l.new
    s = sfunc(theta,x)
    J = Jfunc(theta,x)
    show(c(theta,l,s,J,alpha))
  }
  theta
}
```



(*v*) Consider now the EM-algorithm. Define the complete data to be $\boldsymbol{y} = (y_1, y_2, y_3, y_4, y_5)$ where $y_j = x_j, j = 1,2,3$ while $y_4 + y_5 = x_4$. We now assume a multinomial model for the 5 variables with probabilities

$$(\theta/4, (1-\theta)/4, (1-\theta)/4, 1/2, \theta/4).$$

EM

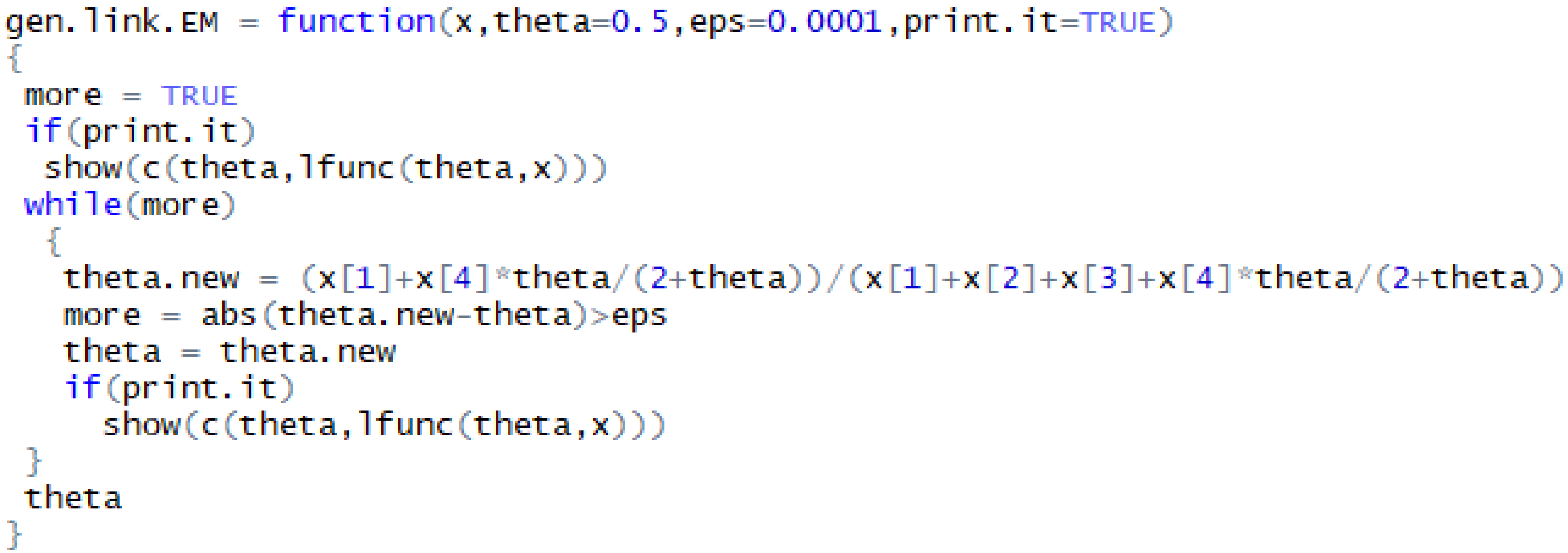
complete log-likelihood is given by

$$\begin{aligned}
\ell(\theta) =& \text{Const} + x_1\log(\theta) + (x_2+x_3)\log(1-\theta) + (x_4-y_5)\log(2) + y_5\log(\theta)\\
Q(\theta|\theta^{(t)}) =& \text{Const} + x_1\log(\theta) + (x_2+x_3)\log(1-\theta) + x_4\log(2)\\
&+ E[Y_5|x_4,\theta^{(t)}][\log(\theta) - \log(2)]\\
Y_5|x_4 \sim& \text{Binom}(x_4, \tfrac{\theta}{2+\theta})\\
Q(\theta|\theta^{(t)}) =& \text{Const} + x_1\log(\theta) + (x_2+x_3)\log(1-\theta) + x_4\log(2) +\\
&\frac{x_4\theta^{(t)}}{2+\theta^{(t)}}[\log(\theta) - \log(2)]\\
\frac{\partial}{\partial\theta}Q(\theta|\theta^{(t)}) =& \frac{x_1}{\theta} - \frac{x_2+x_3}{1-\theta} + \frac{x_4\theta^{(t)}}{2+\theta^{(t)}}\frac{1}{\theta}
\end{aligned}$$

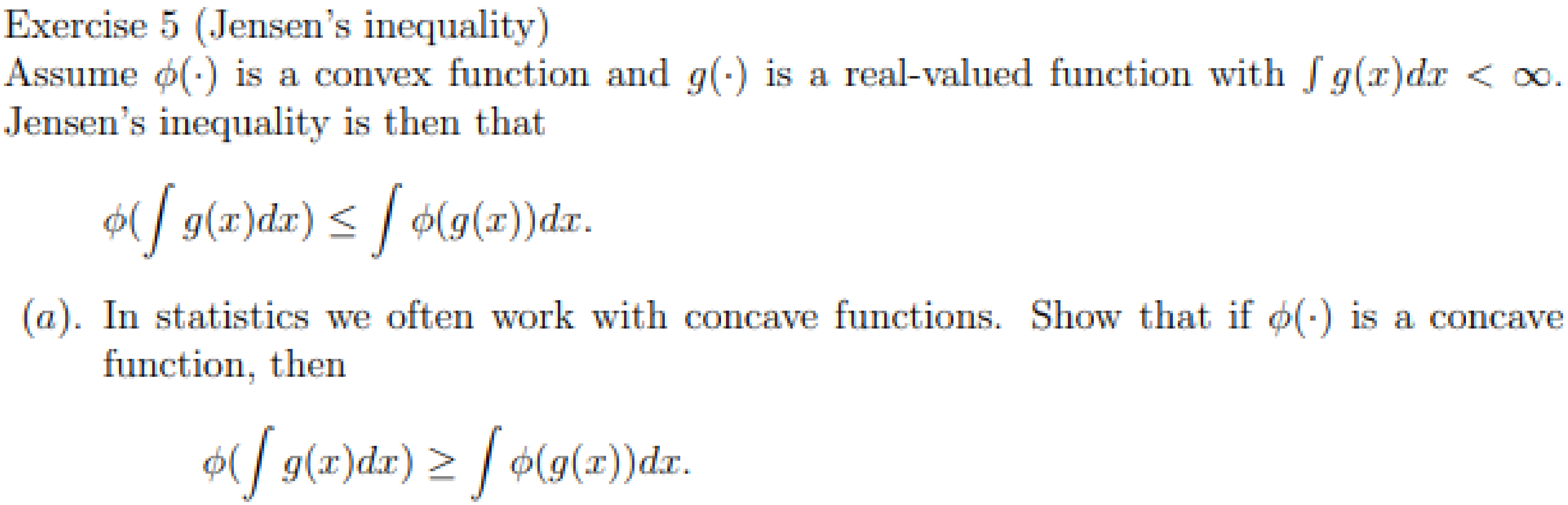
giving

$$\theta^{(t+1)} = \frac{x_1 + x_4\frac{\theta^{(t)}}{2+\theta^{(t)}}}{x_1+x_2+x_3+x_4\frac{\theta^{(t)}}{2+\theta^{(t)}}}$$

See `genetic_linkage.R` for implementation.



```
gen.link.EM = function(x,theta=0.5,eps=0.0001,print.it=TRUE)
{
 more = TRUE
 if(print.it)
  show(c(theta,lfunc(theta,x)))
 while(more)
  {
   theta.new = (x[1]+x[4]*theta/(2+theta))/(x[1]+x[2]+x[3]+x[4]*theta/(2+theta))
   more = abs(theta.new-theta)>eps
   theta = theta.new
   if(print.it)
     show(c(theta,lfunc(theta,x)))
 }
 theta
}
```

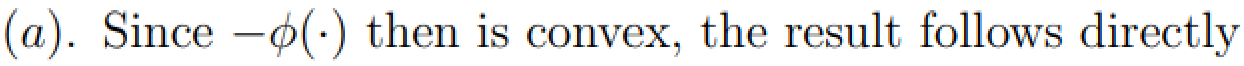


Exercise 5 (Jensen's inequality)
Assume $\phi(\cdot)$ is a convex function and $g(\cdot)$ is a real-valued function with $\int g(x)dx < \infty$. Jensen's inequality is then that

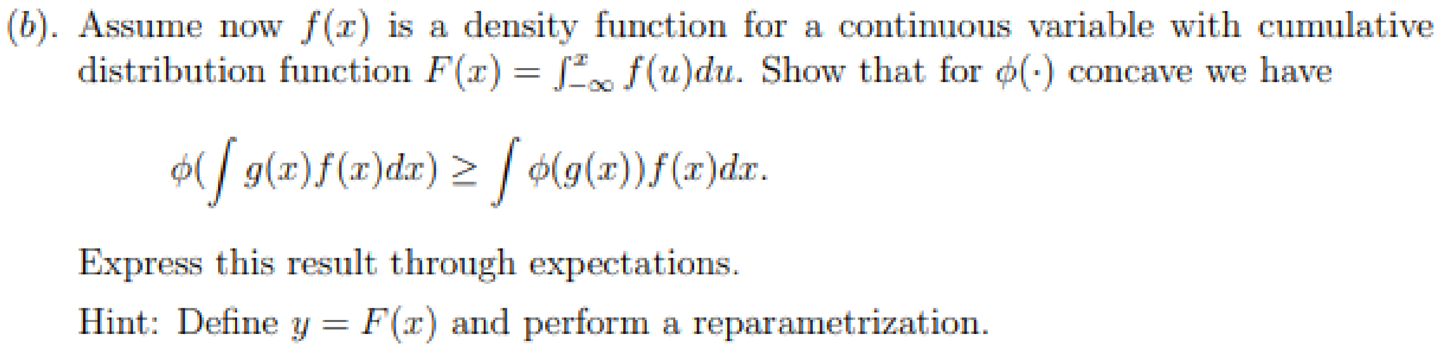
$$\phi(\int g(x)dx) \leq \int \phi(g(x))dx.$$

(a). In statistics we often work with concave functions. Show that if $\phi(\cdot)$ is a concave function, then

$$\phi(\int g(x)dx) \geq \int \phi(g(x))dx.$$



(a). Since $-\phi(\cdot)$ then is convex, the result follows directly

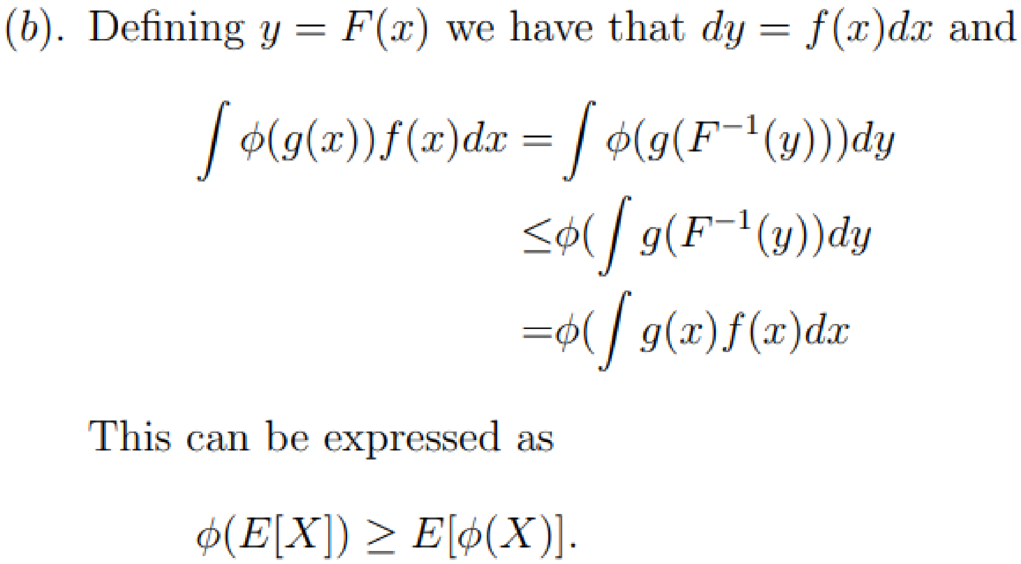


(b). Assume now $f(x)$ is a density function for a continuous variable with cumulative distribution function $F(x) = \int_{-\infty}^{x} f(u)du$. Show that for $\phi(\cdot)$ concave we have

$$\phi(\int g(x)f(x)dx) \geq \int \phi(g(x))f(x)dx.$$

Express this result through expectations.

Hint: Define $y = F(x)$ and perform a reparametrization.

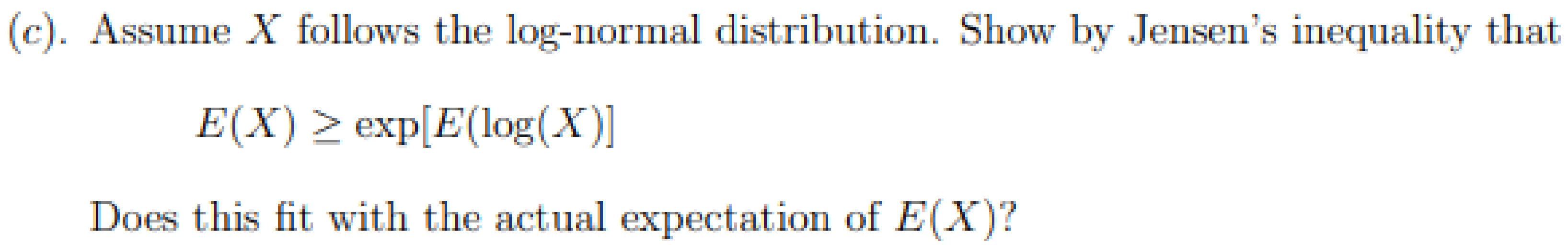


(b). Defining $y = F(x)$ we have that $dy = f(x)dx$ and

$$\begin{aligned}\int \phi(g(x))f(x)dx =& \int \phi(g(F^{-1}(y)))dy \\ \leq& \phi(\int g(F^{-1}(y))dy \\ =& \phi(\int g(x)f(x)dx\end{aligned}$$

This can be expressed as

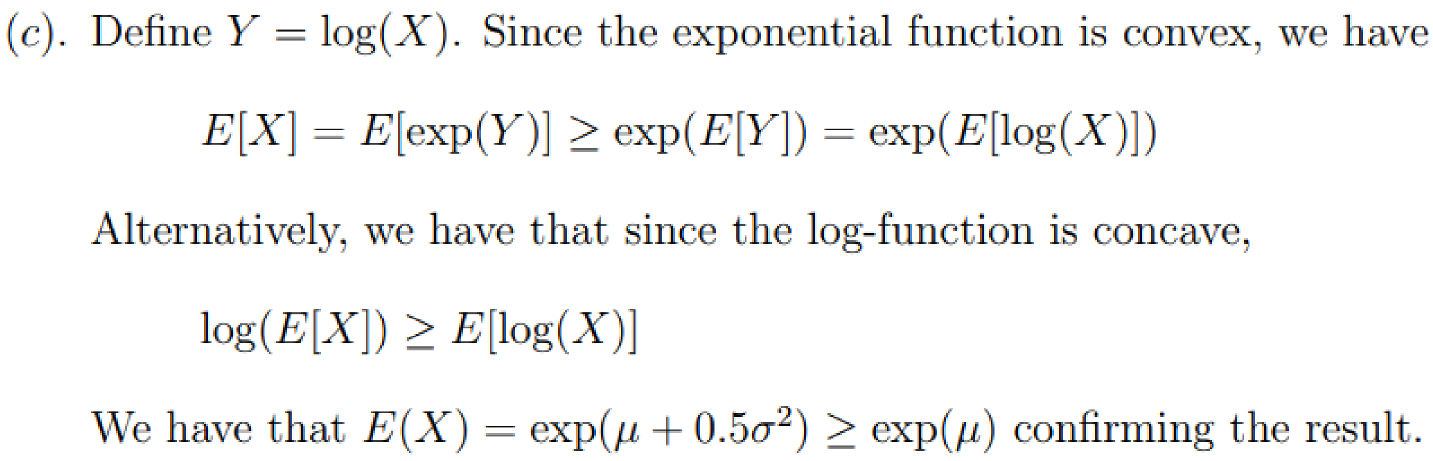
$$\phi(E[X]) \geq E[\phi(X)].$$



(c). Assume X follows the log-normal distribution. Show by Jensen's inequality that

$$E(X) \geq \exp[E(\log(X)]$$

Does this fit with the actual expectation of $E(X)$?



(c). Define $Y = \log(X)$. Since the exponential function is convex, we have that

$$E[X] = E[\exp(Y)] \geq \exp(E[Y]) = \exp(E[\log(X)])$$

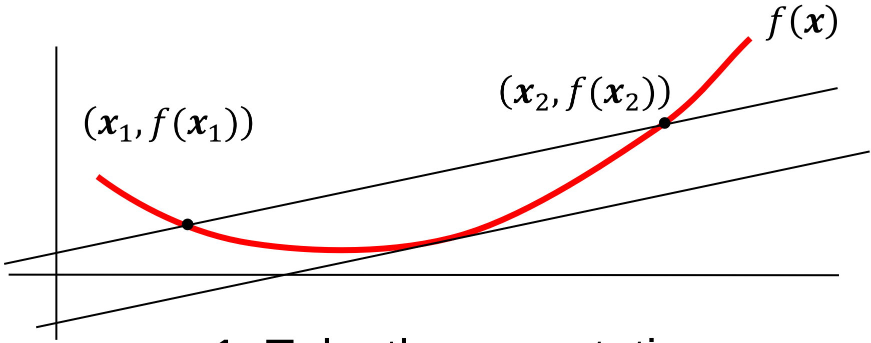
Alternatively, we have that since the log-function is concave,

$$\log(E[X]) \geq E[\log(X)]$$

We have that $E(X) = \exp(\mu + 0.5\sigma^2) \geq \exp(\mu)$ confirming the result.

Jensens inequality

Convex function: $f(\boldsymbol{x})$


$f(\boldsymbol{x})$
$(\boldsymbol{x}_2, f(\boldsymbol{x}_2))$
$(\boldsymbol{x}_1, f(\boldsymbol{x}_1))$

Interpolation always above function

$$f(t\boldsymbol{x}_1 + (1-t)\boldsymbol{x}_2) \leq tf(\boldsymbol{x}_1) + (1-t)f(\boldsymbol{x}_2)$$
$$\text{for } 0 \leq t \leq 1$$

Tangent always below function

$$f(x) \geq f(x_0) + (x - x_0)f'(x_0) \quad \forall\, x_0$$

1. Take the expectation on each side of the inequality

$$E\big(f(X)\big) \geq E[\, f(x_0) + \quad (X - x_0)f'(x_0)\,]$$
$$E\big(f(X)\big) \geq \ f(x_0) \ + \underbrace{(E(X) - x_0)}_{= 0 \text{ when } x_0 = E(X)} f'(x_0)$$

2. Select $x_0 = E(X)$

$$E\big(f(X)\big) \geq \ f(E(X))$$