Problem 1

a) To get exact samples we can use either rejection sampling or probability inversion.

Details of probability inversion For the Laplace distribution we have: [obtained by integrating the pdf provided]

$$F(x|\mu,\sigma) = \begin{cases} \frac{1}{2}\exp\left(\frac{x-\mu}{\sigma}\right) & x \le \mu \\ 1 - \frac{1}{2}\exp\left(-\frac{x-\mu}{\sigma}\right) & x > \mu \end{cases}$$

We want to invert $u = F(x|\mu,\sigma)$, to find x we further have $F(\mu|\mu,\sigma) = \frac{1}{2}$. Such that the change of function description happens for $u = 1/2$.

$$F^{-1}(x|\mu,\sigma) = \begin{cases} \mu + \sigma \cdot \log\ 2u & u \le 1/2 \\ \mu - \sigma \log 2(1-u) & u > 1/2 \end{cases}$$

##

An alternative is to sample a standard exponential distribution $z = -\log u$, and a sample s from a uniform distribution on $\{-1,1\}$. $P(s=-1) = P(s=1) = 1/2$, and set:

$$x = \mu + s \cdot \sigma \cdot z$$

This approach uses the symmetry of the distribution abound the median, and that the left and right tails decays like the exponential distribution.

Problem 2

In the problem of model selection, the parameter is $\gamma_k \in \{0,1\}, k = 1, \ldots p$. There are 2^p, possibilities thus the problem is NP hard. To evaluate the target function (AIC), we find $m = \sum_{k=1}^{p} \gamma_k$, and we find the likelihood by performing maximum likelihood estimation for the linear regression model:

$$y_i|\gamma\ = \beta_0 + \sum_{\{k:\gamma_k=1\}} \beta_k x_{ki} + \varepsilon_i$$

and evaluate the likelihood at the estimate obtained.

Fitness function: The fitness function should be such that the optimal fitness corresponds to the optimal AIC. The desired values should have high fitness. Thus chose a fitness function which is monotone decreasing function of AIC.

Population: The population size should be selected according to a the maximum number of explanatory variables desired. For binary problems [such as the model selection] we usually use:

$$p < \text{Population size} < 2p$$

Selection: There should be an element of survival of the fittest, such that the individuals with the best fit are selected more often. One option is to sample the population with weights proportional to the fit, or a monotone function of the fit: One example could be:

$$P(i) \propto \exp(-AIC(i) \cdot\ b - c)$$

In the above expression i refer to one individual in the population. It is possible to choose just one of the parents according to the probability and select the other at random.

Crossover: This is a way to combine the parents. For the binary selection we have. It is common to fix the values that both parents have in common and sample the remaining randomly.

Mutation: After the crossover it completed, we change a limited number of the gamma values at random. This should not bee a too large fraction as then the inheritance gets too little effect. .The probability of making a change could be 1/p, thus there is on average one change in each offspring.

The feature which makes the genetic algorithm different from local methods based on neighborhood, is that we consider multiple solutions at the same time and let these interact to provide a better solution together.

Problem 3

a) The likelihood:

$$L(\beta, \sigma) = \prod_{i=1}^{n} \frac{1}{2\sigma} \exp\left(-\frac{|y_i - \beta^T x_i|}{\sigma}\right)$$

The log likelihood:

$$l(\beta, \sigma) = \ -n \cdot \log 2\ - n \cdot \log \sigma - \sum_{i=1}^{n} \frac{|y_i - \beta^T x_i|}{\sigma}$$

Since $l(\beta, \sigma)$ depend on β only through the sum, and $1/\sigma$ is a positive constant, the maximum with respect to $l(\beta, \sigma)$ is the minimum of $\sum_{i=1}^{n} |y_i - \beta^T x_i|$

$$\frac{\partial l(\beta, \sigma)}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^2} \sum_{i=1}^{n} |y_i - \beta^T x_i|$$

$$\frac{\partial l(\beta, \sigma)}{\partial \beta} = \frac{1}{\sigma} \sum_{i=1}^{n} \text{sign}(\beta^T x_i - y_i)\ \cdot\ x_i$$

b) Newton methods in general requires the second derivative of the loglikelihood, which is not well-defined in the problem(since it would involve the derivative of a discontinuous

function), Gauss-Newton is a variant suited for nonlinear least-squares problems which do not require the second derivative, but our problem is not least-squares, thus it is not suited for this problem. Nelder-Mead is a method which do not require the derivative of any sort thus it is well suited for this case.

In IRLS, we use weighted least squares, where the weight is dependent on the parameter estimate. Thus, we can rewrite the problem as:

$$\sum_{i=1}^{n} w_i(\beta) \cdot |y_i - \beta^T x_i|^2$$

with $w_i(\beta) = 1/|y_i - \beta^T x_i|$, The algorithm is:

1) Initiate $W^0 = I_{n\times n}$ (the identity)
2) Solve weighted least squares $\beta^{k+1} = \left(XW^kX\right)^{-1}XW^ky$
3) Set W^{k+1} to be a diagonal matrix with $W_{ii}^{k+1} = w_i(\beta^{k+1})$
4) Increment k and go to 2)

End loop when [relative error] $\left\|\beta^{k+1} - \beta^k\right\| < \epsilon \cdot \left\|\beta^k\right\|$, with ϵ being a small number or

when [absolute error] $\left\|\beta^{k+1} - \beta^k\right\| < \epsilon$ with ϵ being a small number

c) The call optim(c(0,0), sad, sadGrad ,y,X), computed the MLE for β, sig is the MLE for σ. The for loop perform nonparametric Bootstrap of the parameters. It samples pairs of y_i and x_i , and run the exact same algorithm as is done for obtaining the estimates.

If there is a large deviation between the original estimates and the average of the bootstrap, there is problems with bias in the method. By comparison we see that these correspond well. The square root of the diagonal of the covariance matrix is the standard deviation. Thus the final line illustrate the uncertainty in the estimate [in terms of the standard deviation]

Problem 4

The complete log likelihood is given below and is subject to the constraint $\sum_{k=1}^{K}\pi_k = 1$

$$l(\theta|\boldsymbol{x},\boldsymbol{c}) = \sum_{i=1}^{n}\sum_{k=1}^{K} I(c_i = k)\left(\log\pi_k + \log\phi(x_i;\mu_k,\sigma^2)\right)$$

$$= \sum_{i=1}^{n}\sum_{k=1}^{K} I(c_i = k)\left(\log\pi_k - \frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2}\frac{(x_i-\mu_k)^2}{\sigma^2}\right)$$

$$Q\left(\theta\middle|\theta^{(t)}\right) = E\left(l(\theta|\boldsymbol{x},\boldsymbol{c})\middle|\boldsymbol{x},\theta^{(t)}\right)$$

$$= \text{Const} + \sum_{i=1}^{n}\sum_{k=1}^{K} P\left(C_i = k|\boldsymbol{x},\theta^{(t)}\right)\left(\log\pi_k - \frac{1}{2}\log(\sigma^2) - \frac{1}{2}\frac{(x_i-\mu_k)^2}{\sigma^2}\right)$$

$$Q_L\left(\theta\middle|\theta^{(t)}\right) = Q\left(\theta\middle|\theta^{(t)}\right) + \lambda\cdot\left(1 - \sum_{k=1}^{K}\pi_k\right)$$

$$\frac{\partial Q\left(\theta\middle|\theta^{(t)}\right)}{\partial \pi_k} = \sum_{i=1}^{n} P\left(C_i = k|\boldsymbol{x}, \theta^{(t)}\right) \cdot \frac{1}{\pi_k} - \lambda = 0$$

$$\pi_k = \frac{1}{\lambda}\sum_{i=1}^{n} P\left(C_i = k|\boldsymbol{x}, \theta^{(t)}\right)$$

$$1 = \sum_{k=1}^{K} \pi_k = \frac{1}{\lambda}\sum_{i=1}^{n}\sum_{k=1}^{K} P\left(C_i = k|\boldsymbol{x}, \theta^{(t)}\right) = \frac{1}{\lambda} \cdot n \Rightarrow \lambda = n$$

$$\frac{\partial Q_L\left(\theta\middle|\theta^{(t)}\right)}{\partial \mu_k} = \sum_{i=1}^{n} P\left(C_i = k|\boldsymbol{x}, \theta^{(t)}\right)\left(\frac{(x_i - \mu_k)}{\sigma^2}\right) = 0$$

$$\mu_k \sum_{i=1}^{n} P\left(C_i = k|\boldsymbol{x}, \theta^{(t)}\right) = \sum_{i=1}^{n} P\left(C_i = k|\boldsymbol{x}, \theta^{(t)}\right) x_i$$

$$\mu_k = \frac{\sum_{i=1}^{n} P\left(C_i = k|\boldsymbol{x}, \theta^{(t)}\right) x_i}{\sum_{i=1}^{n} P(C_i = k|\boldsymbol{x}, \theta^{(t)})}$$

$$\frac{\partial Q_L\left(\theta\middle|\theta^{(t)}\right)}{\partial \sigma^2} = \sum_{i=1}^{n}\sum_{k=1}^{K} P\left(C_i = k|\boldsymbol{x}, \theta^{(t)}\right)\left(-\frac{1}{2\sigma^2} + \frac{1}{2}\frac{(x_i - \mu_k)^2}{(\sigma^2)^2}\right) = 0$$

$$\sum_{i=1}^{n}\sum_{k=1}^{K} P\left(C_i = k|\boldsymbol{x}, \theta^{(t)}\right) = \frac{1}{\sigma^2}\sum_{i=1}^{n}\sum_{k=1}^{K} P\left(C_i = k|\boldsymbol{x}, \theta^{(t)}\right)(x_i - \mu_k)^2$$

$$\sigma^2 = \frac{\sum_{i=1}^{n}\sum_{k=1}^{K} P\left(C_i = k|\boldsymbol{x}, \theta^{(t)}\right)(x_i - \mu_k)^2}{n}$$

$$P\left(C_i = k|\boldsymbol{x}, \theta^{(t)}\right) = \frac{p(C_i = k, X_i = x_i|\, \theta^{(t)})}{p(X_i = x_i|\theta^{(t)})} = \frac{\pi_k \phi\left(x_i; \mu_k, \sigma_k^2\right)}{\sum_{m=1}^{K} \pi_k \phi(x_i; \mu_m, \sigma_m^2)} = \frac{\pi_k \phi\left(x_i; \mu_k, \sigma_k^2\right)}{\sum_{m=1}^{K} \pi_k \phi(x_i; \mu_m, \sigma_m^2)}$$

b) The marginal likelihood:

$$L(\theta|\boldsymbol{x}) = \prod_{i=1}^{n}\left(\sum_{k=1}^{K} \pi_k\, \phi(x_i; \mu_k, \sigma^2)\right)$$

The observed information matrix is the negative Hessian matrix of the log likelihood evaluated around the MLE estimate, it can be approximated by a numerical computation of the Hessian. The important ting to be aware off is that our computation (of the negative hessian) should provide a symmetric positive definite matrix. Let $\Delta\theta_i$ be a vector being zero

except for element component *i*,. Further let the change be positive and have size $||\Delta\theta_i||$. The diagonal is then:

$$\frac{\partial^2 \log L(\theta|x)}{\partial \theta_i^2} \approx \frac{-2 \cdot \log L(\theta|x) \ + \log L(\theta + \Delta\theta_i|x) \ + \log L(\theta - \Delta\theta_i|x)}{||\Delta\theta_i||^2}$$

The off diagonal is e.g:

$$\frac{\partial^2 \log L(\theta|x)}{\partial \theta_i \partial \theta_j} \approx \frac{\log L\big(\theta + \Delta\theta_i + \Delta\theta_j\big|x\big) \ + \log L(\theta|x) \ - \log L(\theta + \Delta\theta_i|x) \ - \log L\big(\theta + \Delta\theta_j\big|x\big)}{\big|\big|\Delta\theta_{i|}\big|\big|\Delta\theta_j||}$$

The parameter $||\Delta\theta_i||$ is important to select to provide a stable estimate. Not too small to divide by a small number. Not too large since the error in the derivative becomes too large.

The Fisher information can also be approximated as n times the covariance of the individual score functions [derivative of the marginal log likelihood] evaluated around the MLE estimate. So it is possible to compute the individual contributions numerically as well.

c) The change is that the probability weight in the EM estimator also becomes conditioned to the partially erroneous label.

$$P(C = k|F = f, X = x\ ) \propto P(C = k) \cdot P(F = f|C = k) \cdot p(x|C = c)$$
$$\propto \pi_k[p_F + (p_T - p_F)I(k = f)]\phi(x|\mu_k, \sigma\ )$$

Where:

$$P(F = f|C = k) = p_F + (p_T - p_F)I(k = f)$$

d) The intuitive way to update the estimate is just to take the average probability of being right:

$$p_T = \frac{1}{n}\sum_{i=1}^{n} P(C_i = f_i|x_i, f_i, \theta^t)$$

For one case we have we have the joint distribution:

$$P(C_i, f_i, x_i) \propto \prod_{k=1}^{K}[\ \pi_k \phi_k(x_i|\mu_k, \sigma)p_k(f_i)\ ]^{I(c_i=k)}$$

$$p_k(f_i) = p_T^{I(k=f_i)} p_F^{I(k\neq f_i)} = p_T^{I(k=f_i)} \left[\frac{1 - p_T}{K - 1}\right]^{I(k\neq f_i)}$$

This gives the complete log likelihood:

$$l(\theta|\boldsymbol{x}, \boldsymbol{c}\ ) = \sum_{i=1}^{n}\sum_{k=1}^{K} I(c_i = k)\ (\log \pi_k + \log \phi(x_i; \mu_k, \sigma^2\ ) + I(k = f_i)\log(p_T)$$
$$+I(k \neq f_i)\log(1 - p_T) - I(k \neq f_i)\log\ (K - 1))$$

Which will give the result.

Problem 5

a) We skip the first part of the samples since we do not have a guarantee that the initial sample is representative. The distribution will converge towards the true distribution

[given sufficient regularity]. Thus we exclude the sample in order to avoid unnecessarily bias in our computations. The period to we skip is called burn-in.

The GR-statistic measure how similar averages are in the parallel chains. If there is a full match the statistics is 1. It is somewhat strange that the GR-statistic increases up until 10^5 samples for the variance parameters. This indicates that the four chains are more similar than in the initial part than after a while. This might be indications of apparent convergence, and it is certainly a sign that we need longer chains. To assure convergence we need at least 10^6 samples. In this case we see that there has been a stabilizing effect and that the chains have a uniform decay from $5x10^5$ to 10^6,also a rule of thumb is that the Gelman-Rubin statistics should be below 1.05 for all parameter this occurs first time for 10^6.

$$p(x,y|z,\beta_0,\beta,\sigma^2,\tau^2) \propto p(x,y,z,\beta_o,\beta,\sigma^2,\tau^2)$$

$$= p(\beta_0)p(\beta)p(\sigma^2)p(\tau^2)\prod_{i=1}^{n}\phi\left(y_i|\beta_0+\beta^T x_{i'},\sigma^2\right)\cdot\prod_{j=1}^{q}\phi\left(x_{ji}|z_{ji},\tau^2\right)$$

$$\propto \frac{1}{\sigma^{2(1+\frac{n}{2})}}\cdot\frac{1}{\tau^{2(1+n/2)}}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0-\beta x_i)^2\right)\exp\left(-\frac{1}{2\tau^2}\sum_{i=1}^{n}(z_i-x_i)^2\right)$$

The individual becomes:

By ignoring all factors not including sigma, we recognize the parameters in the inverse gamma

$$p(\sigma^2|\beta_0,\beta,\tau^2,x,z,y) = \text{invGam}(\text{shape}=\frac{\text{n}}{2},\text{scale}=\frac{1}{2}\sum_{i=1}^{n}(y_i-\beta_0-\beta x_i)^2$$

By ignoring all factors not including tau, we recognize the parameters in the inverse gamma

$$p(\tau^2|\beta_0,\beta,\sigma^2,x,z,y) = \text{invGam}(\text{shape}=\frac{\text{n}}{2},\text{scale}=\frac{1}{2}\sum_{i=1}^{n}(x_i-z_i)^2$$

We just get the likelihood part here, this is like for standard regression, we can compute the betas jointly or individually, this gives normal distributions:

$$p(\beta_0|\beta,\sigma^2,\tau^2,x,z,y) \propto \exp\left(-\frac{n}{2\sigma^2}\left(\beta_0^2-2\beta_0\cdot\frac{1}{n}\sum_{i=1}^{n}(y_i-\beta x_i)\right)\right)$$

$$p(\beta|\beta_0,\sigma^2,\tau^2,x,z,y) \propto \exp\left(-\frac{\sum_{i=1}^{n}x_i^2}{2\sigma^2}\left(\beta^2-2\beta\cdot\frac{1}{\sum_{i=1}^{n}x_i^2}\sum_{i=1}^{n}x_i(y_i-\beta_o)\right)\right)$$

the precision is the factor multiplying ½ in front of the brackets, the mean is the factor multiplying 2 times the parameter.

For the explanatory variable, we get one factor from each product (or one term from each sum in the exponential)

$$p(x_i|\beta_0,\beta,\sigma^2,\tau^2,x_{-i},z,y\,) \propto \exp\left(-\frac{1}{2\sigma^2}(\beta x_i-[y_i-\beta_0])^2\ -\frac{1}{2\tau^2}(x_i-z_i)^2\right)$$

Getting this on quadratic form: $\exp\left(-\frac{1}{2}Q\cdot(x_i-b)^2\right)$, we find the precision and the mean in the normal distribution

$$Q=\frac{\beta^2}{\sigma^2}+\frac{1}{\tau^2}\quad,\quad b=\left(\frac{\beta[y_i-\beta_0]}{\sigma^2}-\frac{z_i}{\tau^2}\right)/Q$$