# UNIVERSITY OF OSLO

## Faculty of mathematics and natural sciences

Exam in: STK4051/STK9051 — Computational statistics

Day of examination: Monday June 13th 2022

Examination hours: 09.00 – 13.00.

This problem set consists of 5 pages, problem 4d is only for STK 9051

Appendices: None

Permitted aids: All examination aids are allowed (e.g. books, online resources, WolframAlpha, scientific programming tools, etc.).

It is not allowed to collaborate or communicate with others during the exam about the assignments.

### Problem 1

The Laplace distribution with parameters $\mu$, $\sigma$ has the density:

$$f(x|\mu,\sigma) = \frac{1}{2\sigma} \cdot \exp\left(-\frac{|x-\mu|}{\sigma}\right). \quad (1)$$

The parameters $\mu$ and $\sigma$ are denoted location and scale parameters respectively. For a centered Laplace distribution $\mu$ =0.

a) Describe one method to derive exact samples from this distribution. Comment on the method chosen, and provide all expressions needed to perform the sampling (as code or formulas).

### Problem 2

In the context of model selection in linear regression, we can use a parameterization of the form:

$$y_i = \beta_0 + \sum_{k=1}^{p} \gamma_k \beta_k x_{ki} + \varepsilon_i, \;\; i = 1, \dots, n, \quad (2)$$

where $y_i$ is the dependent variable corresponding to the explanatory variables $x_i = [x_{1i}, \dots, x_{pi}]$, $\varepsilon_i$ is an error term, $\beta_k, k = 0, \dots p$, are scalar coefficients; and $\gamma_k \in \{0,1\}, k = 1, \dots p$ are indicators. The parameter $\gamma_k$ is thus 1 if the explanatory variable $x_k$ is a part of the linear regression model. Akaike information criterion is defined as:

$$AIC \;=\; 2m - 2\log L\big(\hat{\theta}|y, X\big), \tag{3}$$

where $m$ is the number of parameters in the model and $\log L\big(\hat{\theta}|y, X\big)$, is the log likelihood evaluated for the maximum likelihood estimator of the parameters. We want the model which minimize the AIC.

a) Formulate this problem of model selection as a problem of combinatorial optimization and give details about how to use a genetic algorithm to solve the problem. In particular discuss specific choices of, fitness function, population, selection, crossover and mutation in the context of the problem of model selection above. What is the feature which makes the genetic algorithm different from local methods based on neighborhood, e.g. steepest decent, simulated annealing etc.

## Problem 3

Consider a case of liner regression

$$y_i = \beta^T x_i + \varepsilon_i, \;\; i = 1, \dots, n, \tag{4}$$

where $\varepsilon_i$ are independent, identically distributed according to a centered Laplace distribution with scale parameter $\sigma$, (see problem 1 a).

a) Write down an expression for the likelihood of the parameters $\beta$ and $\sigma$ in this problem. Show that the maximum likelihood estimator for $\beta$ can be found as the solution to the minimization problem:

$$\hat{\beta}_{ML} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} |y_i - \beta^T x_i|, \tag{5}$$

and that the maximum likelihood estimator for $\sigma$ is given by:

$$\hat{\sigma}_{ML} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \beta^T x_i|. \tag{6}$$

b) In the context of solving problem (5) consider the methods Gauss-Newton, and Nelder-Mead. Which of these two methods would you recommend? Argue the case for your selection. An alternative method, is the iteratively reweighted least squares (IRLS), give details on how you would set up IRLS for problem (5) in terms of pseudocode.

c) Figure 1 shows a sample of the regression model from a) having n=100 datapoints.

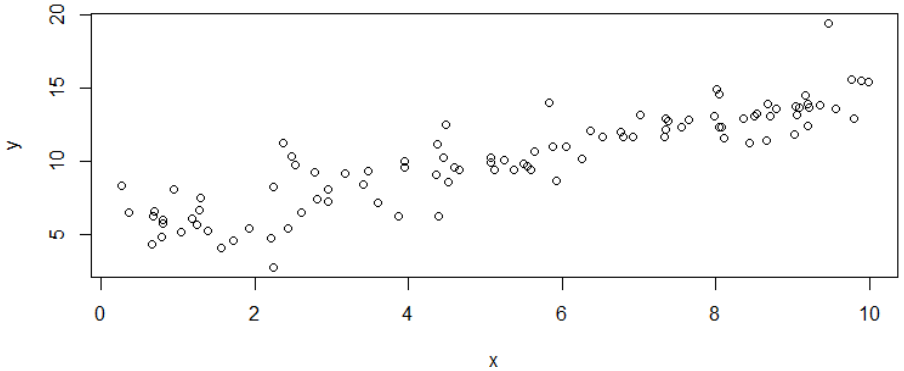


*Figure1: Scatterplot of dependent and independent variables in Problem 3*

Assume the dependent variable is `y` where `dim(y)` is 100×1, and that the design matrix is `X`, where `dim(X)` is 100×2. The last column in X is just ones. Consider the R-code:

```
sad<-function(beta,sigma,y,X){
      sum( abs( X%*%beta - y ) )
}

sadGrad<-function(beta,y,X){
  t(X)%*%sign( X%*%beta-y )
}

N= 100
beta=optim(c(0,0), sad, sadGrad ,ySample,XSample)
sig=sad(resultoptim$par,ySample,XSample/N


B=1000
est=matrix(0,B,3)
colnames(est)<-c("slope","intercept",'MAD')

for(i in 1:B){
  sampleCase  = sample(1:N,N,replace=TRUE)
  ySample = y[sampleCase]
  XSample = X[sampleCase,]
  resultoptim=optim(c(0,0), sad, sadGrad ,ySample,XSample)
  sig=sad(resultoptim$par,ySample,XSample/N
  est[i,]=c(resultoptim$par[1],resultoptim$par[2],sig)
}
```

After running the code, we get the output:

```
> cat(beta$par,sig)
0.9422203 5.2172924 1.050212

> colMeans(est)
    slope intercept       MAD
0.949150  5.142126  1.043432

> sqrt(diag(cov(est)))
   slope  intercept        MAD
0.05304239 0.32169059 0.13221354
```

What is the purpose of the code above, i.e. what does it do? What is the interpretation of the results?

## Problem 4 (EM algorithm)

Consider the mixture model for clustering:

$$P(C_i = k) = \pi_k, \quad k = 1, \dots, K, i = 1, \dots, n \tag{7}$$

$$p(x_i | C_i = k) = \phi(x;\ \mu_k, \sigma^2), \qquad k = 1, .. K, i = 1, \dots, n \tag{8}$$

Where $x = (x_1, \dots x_n)$ is the observations, $C = (C_1, \dots C_n)$ is the class labels, $\phi(x; \mu, \sigma^2)$ is the normal density with mean $\mu$ and variance $\sigma^2$. Our aim is to obtain maximum likelihood estimates of $\theta = \{\sigma^2, \pi_k, \mu_k, , k = 1, \dots, K\}$ based on observations $x = (x_1, \dots x_n)$. The class labels $C$ are missing, and there is a common standard deviation for all classes.

a) In the context of the EM algorithm write down the expression for the complete log-likelihood, derive the expression for $Q\left(\theta\middle|\theta^{(t)}\right)$, and derive the update on, $\pi_k$, $\mu_k$ , and $\sigma^2$. Derive also the expression for $P\left(C_i = k|x_i, \theta^{(t)}\right)$.

b) Write down the expression for the marginal log-likelihood, i.e. the log-likelihood of the observed data. After running the code from a to convergence we obtain a point estimate of the maximum likelihood estimator. We now want to assess the uncertainty in the estimator. Describe how you would use numerical differentiation to assess the uncertainty. Give details and highlight choices you make for the evaluation.

c) We are offered a set of labels $F = (F_1, \ldots F_n)$, which are known to contain some errors. The errors are randomly distributed, conditionally independent of $X_i$ given the class, i.e. $p(x_i, F_i|C_i) = p(x_i|C_i)p(F_i|C_i)$. Given the true class $C_i$, the distribution of labels are:

$$P(F_i = j|C_i = k) = \begin{cases} p_T & j = k, \\ p_F & j \neq k, \end{cases} \tag{9}$$

where we have the relation $p_T + (K-1)p_F = 1$. How would you modify the expressions in part a) to include this information in the estimation? In this task consider $p_T$ to be known in advance.

d) *For STK9051 only*. How would you include estimation of $p_T$ in the framework above? Give formulas for the basic setup and/or propose a formula for estimation.

## Problem 5 (Markov chain Monte Carlo, McMC)

Consider a regression model where there are errors in the explanatory variables. The model can be defined as,

$$y_i = \beta_0 + \beta^T x_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{10}$$

$$z_i = x_i + \eta_i, \quad i = 1, \ldots, n, \tag{11}$$

where $y_i$ is the dependent variable, $x_i = [x_{1i}, \ldots x_{q,i}]$ is the unobserved explanatory variable of dimension $(q+1) \times 1$, and $z_i$ is an observation of the explanatory variable with error. The error terms $\varepsilon_i$ and $\eta_i$ are independent normally distributed with mean zero and variance $\sigma^2$ and $\tau^2$ respectively. The joint probability distribution for $X$ and $Y$ given Z is:

$$p(x, y|z, \beta_0, \beta, \sigma, \tau) = \prod_{i=1}^{n} \phi(y_i|\beta_0 + \beta^T x_i, \sigma^2) \cdot \prod_{j=1}^{q} \phi\left(x_{ji}\middle|z_{ji}, \tau^2\right), \tag{12}$$

where:

$$\phi(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right). \tag{13}$$

The prior distributions of the parameters are improper and given as:

$p(\beta_0) \propto 1, \quad p(\beta) \propto 1\,, \quad p(\sigma^2) \propto 1/\sigma^2, \quad p(\tau^2) \propto 1/\tau^2$

A McMC - algorithm for sampling this distribution is run in four parallels for slightly more than $10^6$ iterations. The Gelman-Rubin statistics (GR) is computed based on the sample path from all four parallel chains, for all parameters. Based on the GR-Statistics we want to determine the appropriate length of the chain. The statistics has therefore been computed using fractions of the four chains. The statistics are computed for using $L$ consecutive numbers starting at sample number $D = 1001$, and data from all four parallel chains. Table 1 summarizes this outcome of the experiment for important parameters. For the explanatory variables, we report only the minimum and maximum over all cases.

*Table 1: Gelman-Rubin statistics based on four parallel chains for five different lengths of the chains.*

| $L$ | $10^4$ | $5 \times 10^4$ | $10^5$ | $5 \times 10^5$ | $10^6$ |
|---|---|---|---|---|---|
| $GR(\beta_0)$ | 1.201987 | 1.080185 | 1.030826 | 1.008041 | 1.005866 |
| $GR(\beta)$ | 1.135730 | 1.052199 | 1.027198 | 1.006974 | 1.002051 |
| $GR(\sigma)$ | 2.799689 | 3.688193 | 4.568641 | 1.344375 | 1.005035 |
| $GR(\tau)$ | 1.749298 | 2.247959 | 2.532234 | 1.182309 | 1.004613 |
| $\min_k(GR(x_k))$ | 1.007988 | 1.001624 | 1.002567 | 1.000183 | 1.000056 |
| $\max_k(GR(x_k))$ | 1.713188 | 2.244938 | 2.585777 | 1.226586 | 1.006909 |

a) Why is it common to skip the first part of the sample path when computing summary statistics in MCMC methods? Are any of the numbers in table 1 surprising? Based on the results in Table 1, what would be your recommendation in terms of sample length? Which other numbers could be useful to evaluate number of samples desired?

b) For the case $q = 1$, derive the expressions that are needed for performing Gibbs-sampling, i.e., compute:
- $p(\sigma^2 \,|\beta_0, \beta, \tau^2, x, z, y)$
- $p(\tau^2|\beta_0, \beta, \sigma^2, x, z, y)$
- $p(\beta_0|\beta, \sigma^2, \tau^2, x, z, y)$
- $p(\beta|\beta_0, \sigma^2, \tau^2, x, z, y)$
- $p(x_i|\beta_0, \beta, \sigma^2, \tau^2, x_{-i}, z, y)$

In 5b) it might be helpful to know the inverse gamma distribution. with parameters $(\alpha, \beta)$ being shape and scale parameter respectively is defined over the support $x > 0$:

$$f(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot \frac{1}{x^{\alpha+1}} \cdot \exp\left(-\frac{\beta}{x}\right) \tag{14}$$

where $\frac{\beta^\alpha}{\Gamma(\alpha)}$, is a normalizing factor.