UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Exam in: STK4051/STK9051 — Computational statistics

Day of examination: Monday June 7th 2021

Examination hours: 09.00 – 13.00.

This problem set consists 6 pages, final page for STK9051 only

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 (Rejection sampling)

The Pareto distribution was originally applied to describing the distribution of wealth in a society. The distribution is characterized by two parameters (α, x_{m}), where $\alpha > 0$ and $x_{\mathrm{m}} > 0$. The probability density is given by the formula:

$$f(x) = \begin{cases} \dfrac{\alpha x_{\mathrm{m}}^{\alpha}}{x^{\alpha+1}} & \text{for } x > x_{\mathrm{m}} \\ 0 & \text{otherwise} \end{cases}$$

a) How can you use probability inversion to transform a set of samples, $u_1, u_2, \ldots, u_n$ from a uniform distribution on the unit interval to a set of samples from the Pareto distribution? State the principles and derive a formula for the transformation.

The Cauchy distribution is defined by the density:

$$f(x) = \frac{1}{\pi(1+x^2)}, \qquad -\infty < x < +\infty$$

b) As a rejection sampling algorithm to sample from the Cauchy distribution a proposal distribution $g(x)$ is suggested. This distribution has the form:

$$g(x) = \begin{cases} \dfrac{1}{4} & \text{if } |x| \le 1 \\ \dfrac{\alpha}{4x^{\alpha+1}} & \text{if } |x| > 1 \end{cases}$$

Which values of $\alpha > 0$ can be used to perform rejection sampling? For those values of α which can be used for rejection sampling of the Cauchy distribution, define an envelope and compute the acceptance rate in terms of α.

Problem 2 (EM algorithm)

Consider the mixture model for clustering:

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

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

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 μ and variance σ^2. Our aim is to obtain maximum likelihood estimates of $\theta = \{\mu_k, \sigma_k^2, k = 1, \dots, K\}$ based on observations $x = (x_1, \dots x_n)$. The class labels are missing.

a) In the context of the EM algorithm write down the expression for the complete log-likelihood, derive the expression for $Q(\theta | \theta^{(t)})$, and show that the update on μ_k , and σ_k^2, is given by:

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

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

Derive also the expression for $P(C_i = k | x_i, \theta^{(t)})$.

b) In semi supervised learning it is possible to enhance learning by actively observing the class membership of some of the observations, thus we get the additional information that $C_i = c_i$ for $i = 1, \dots, m$ where $m < n$. Given this additional information how would you change the updating rule for μ_k and σ_k^2 above. You do not need to show the full derivation of the new update, but comment on how the new information changes $Q(\theta | \theta^{(t)})$. Relate this change to the to the definition of $Q(\theta | \theta^{(t)})$.

c) To assess the uncertainty in the EM estimator it is possible to use a bootstrap procedure. In the setting of the semi supervised learning from 2b describe both a parametric and a nonparametric bootstrap for assessing the uncertainty in the EM estimator. Discuss strengths and weaknesses in these two different approaches when applied to the problem of semi supervised learning in 2b.

In the remainder of the problem we will assume that $K = 2$ and that $(\sigma_1, \sigma_2) = (1,2)$. Thus, the unknown parameters in this problem is $\theta = (\mu_1, \mu_2)$ with $\mu_1 < \mu_2$. The histogram from one such model is shown in figure 1.

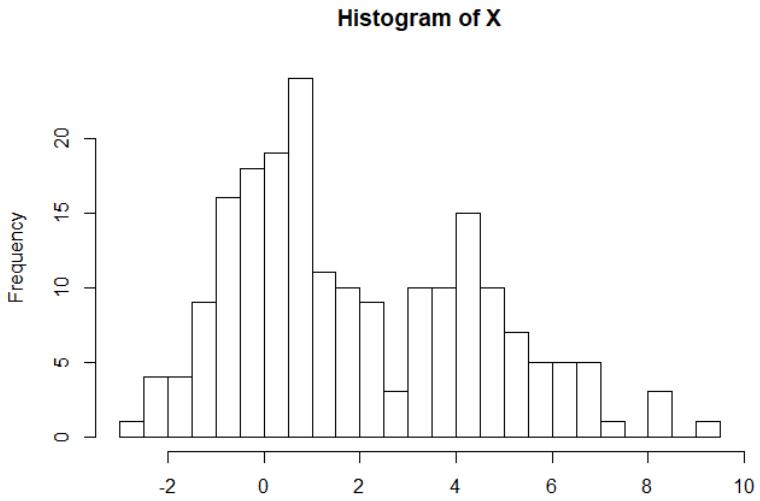

Figure 1: Histogram of the observations $x_1, \ldots, x_n$ used in problem 1d.

We will now consider the question about how to collect labels. Three different strategies are proposed:

A) Collect 10% of the labels at random
B) Collect the label from the 5% highest and 5% lowest values
C) Collect data from the 10% of the data closest to the median

d) Table 1 and 2 gives the values for the observed information matrix and its inverse, given the observations in figure 1. The cases shown are the unsupervised case from 1a, the three strategies A, B and C, and the complete likelihood where we know all class labels. Why is the inverse of the observed information matrix relevant? Based on the two tables, discuss which of the three models A, B and C that provide the most information, comment also on the result by comparing with the unsupervised and complete case. Which strategy for label selection would you use?

	Unsupervised		Strategy A		Strategy B		Strategy C		Complete	
	μ_1	μ_2	μ_1	μ_2	μ_1	μ_2	μ_1	μ_2	μ_1	μ_2
μ_1	74.26	-6.04	77.09	-5.34	76.54	-7.47	82.74	-2.55	99.00	0.00
μ_2	-6.04	16.29	-5.34	17.51	-7.47	17.04	-2.55	18.00	0.00	25.25

Table 1: Observed information matrix for the data in figure 1.

	Unsupervised		Strategy A		Strategy B		Strategy C		Complete	
	μ_1	μ_2	μ_1	μ_2	μ_1	μ_2	μ_1	μ_2	μ_1	μ_2
μ_1	0.014	0.005	0.013	0.004	0.014	0.006	0.012	0.002	0.010	0.000
μ_2	0.005	0.063	0.004	0.058	0.006	0.061	0.002	0.056	0.000	0.040

Table 2: The inverse of the observed information matrix for the data in figure 1.

Problem 3 (Importance sampling)

A random pair (X, W) is properly weighted with respect to a target distribution $\pi(x)$ if - for any square integrable function h, we have that $E(Wh(X)) = c \cdot E_\pi(h(X))$.

a) In the setting of the Bayesian inference, the probability model is defined by the prior distribution $p(x)$, and the likelihood defined by $p(y|x)$. The target distribution is the posterior distribution $p(x|y)$. We will solve the problem of Bayesian inference using importance sampling. Show that if the probability density of X is $g(x)$ and we define $W = p(X)p(X|y)/g(X)$ then the random pair (X, W) will be properly weighted with respect to $p(x|y)$. What is the value of c? Given M samples of the random pair (X, W), say $(x_1, w_1), (x_2, w_2), \dots, (x_M, w_M)$, provide an estimate of the integral:

$$\int_{-\infty}^{\infty} h(x)p(x|y)dx$$

Consider now a generic state space model described by: $p(x_1)$ and $p(x_{i+1}|x_i)$, $i = 1, \dots, (n-1)$, i.e. and the likelihood $p(y_i|x_i)$, for $i = 1, \dots, n$. The samples are generated by another state space model defined by $g(x_1)$ and $g(x_{i+1}|x_i)$, for $i = 1, \dots, (n-1)$.

b) For the model above derive an expression for weights W_i which will be properly weigthed with respect to $p(x_i|y_1, \dots, y_i)$. Show also that this expression can be given the recursive form:

$$W_i = W_{i-1}\frac{p(X_i|X_{i-1})p(y_i|X_i)}{g(X_i|X_{i-1})}, \qquad \text{for } i = 2, \dots, n$$

Problem 4 (Markov chain Monte Carlo, McMC)

In this problem, we will analyze a variance component model using Bayesian inference. In a variance component model with one layer we have three parameters $(\mu, \sigma_X^2, \sigma_R^2)$. The statistical model is defined as $y_{ij} = x_i + \varepsilon_{ij}$, where x_i, $i = 1, \dots n$, is iid with distribution $N(\mu, \sigma_X^2)$, and ε_{ij} for $i = 1, \dots, n$ and $j = 1, 2$ is iid with distribution $N(0, \sigma_R^2)$.

The likelihood can be given as:

$$L(\mu, \sigma_x^2, \sigma_R^2; \boldsymbol{y}) = \prod_{i=1}^{n} \phi_2\left(\begin{bmatrix} y_{i1} \\ y_{i,2} \end{bmatrix}; \begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} \sigma_R^2 + \sigma_X^2 & \sigma_X^2 \\ \sigma_X^2 & \sigma_R^2 + \sigma_X^2 \end{bmatrix}\right)$$

Where $\phi_2(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a bivariate gaussian distribution. The prior distributions for the parameters are improper, with $p(\mu) \propto 1$, $p(\sigma_R^2) \propto 1/\sigma_R^2$ and $p(\sigma_X^2) \propto 1/\sigma_X^2$.

a) A sampling path for a random walk McMC method is shown in figure 2. Characterize the different parts of the path, and comment on the characteristics of the sample path. Would you expect the chain to have a high or low acceptance rate? In general, what is the requirements for a Markov chain to converge to the target distribution, and how can you improve your confidence in the results obtained by McMC methods?

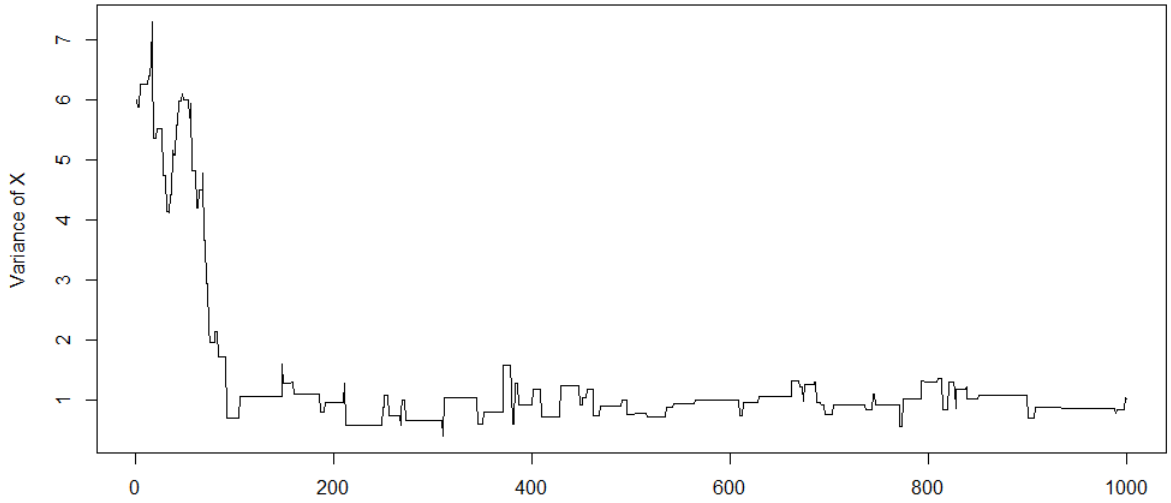

Figure 2: Sample path of σ_X^2 for a random walk Markov chain

b) A key parameter in the random walk algorithm is the magnitude of the permutation. Table 3 lists the acceptance rate and the integral of the autocorrelation for 5 values of the magnitude of the permutation. Relate the numbers given in the list to the effective number of samples. Comment on how the random walk algorithm responds to the magnitude of the permutation. Among the five magnitudes which would you select?

Magnitude of scale	**0.02**	**0.05**	**0.10**	**0.20**	**0.5**
Acceptance rate	0.91	0.78	0.59	0.34	0.06
$\sum_{h\geq 1}\rho_{\mu}(h)$	82.5	9.9	3.3	3.8	7.4
$\sum_{h\geq 1}\rho_{\sigma_R^2}(h)$	113.6	34.7	8.6	7.6	15.3
$\sum_{h\geq 1}\rho_{\sigma_X^2}(h)$	24.8	14.3	2.9	4.4	13.2
$\sum_{h\geq 1}\rho_{l}(h)$	76.1	11.8	6.6	3.8	17.4

Table 3: Summary numbers for 5 different runs of a random walk Markov chain. The summary numbers from top down: are acceptance rate, integral of autocorrelation function for parameters μ , σ_R^2 , σ_X^2, and the log likelihood.

To preserve the positivity of the variance, we use a transformed variable in the problem definition, define $\eta_R = \log \sigma_R^2$ and $\eta_X = \log \sigma_X^2$.

c) We want to get samples from the posterior distribution of $\mu, \sigma_x^2, \sigma_R^2$, by a random walk in the transformed domain, i.e. using μ, η_X, η_R. Define the random walk and derive the metropolis Hastings ratio in the transformed domain. You can write the algorithm and formulas needed or provide a code which does the job, you are free to use whatever programming language you like. (Hint: If you chose to provide R-code you can use the function dmvnorm from the library mvtnorm to compute $\phi_2(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$.)

d) For STK 9051 only. An alternative to the random walk is the Gibbs sampler. To reduce the complexity of the likelihood function, introduce also the unobserved observed group average $X_i, i = 1, \ldots, n$. Set up the joint distribution $p(\mu, \sigma_R^2, \sigma_X^2, \boldsymbol{x}, \boldsymbol{y})$, and compute the conditional distributions required for the Gibbs sampler. You might need the density for the inverse gamma distribution:

$$p(z; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\frac{1}{z}\right)^{\alpha+1} \exp(-\beta/z)$$