Examination Computational Statistics

Linköpings Universitet, IDA, Statistik

Course:	732A90 Computational Statistics
Date:	2020/08/24, 8–13
Teacher:	Krzysztof Bartoszek
Provided aids:	material in the zip file **exam_material_732A90.zip**

Grades:	A= $[18-20]$ points
	B= $[16-18)$ points
	C= $[14-16)$ points
	D= $[12-14)$ points
	E= $[10-12)$ points
	F= $[0-10)$ points

Instructions:	Provide a detailed report that includes plots, conclusions and interpretations. If you are unable to include a plot in your solution file clearly indicate the section of R code that generates it. Give motivated answers to the questions. If an answer is not motivated, the points are reduced. Provide all necessary codes in an appendix. In a number of questions you are asked to do plots. Make sure that they are informative, have correctly labelled axes, informative axes limits and are correctly described. Points may be deducted for poorly done graphs. Name your solution files as: **[your id]_[own file description].[format]** If you have problems with creating a pdf you may submit your solutions in text files with unambiguous references to graphics and code that are saved in separate files. There are **TWO** assignments (with sub–questions) to solve. Provide a separate solution file for each assignment. Include all R code that was used to obtain your answers in your solution files. Make sure it is clear which code section corresponds to which question. If you also need to provide some hand–written derivations please number each page according to the pattern: Question number . page in question number i.e. Q1.1, Q1.2, Q1.3,. . ., Q2.1, Q2.2, . . ., Q3.1, ... Scan/take photos of such derivations preferably into a single pdf file but if this is not possible multiple pdf or .bmp/.jpg/.png files are fine. Please do not use other formats for scanned/photographed solutions. Please submit all your solutions via LISAM or e–mail. If emailing, please email them to **BOTH** krzysztof.bartoszek@liu.se and KB_LiU_exam@protonmail.ch . During the exam you may ask the examiner questions by emailing them to KB_LiU_exam@protonmail.ch **ONLY**. Other exam procedures in LISAM.

NOTE: If you fail to do a part on which subsequent question(s) depend on describe (maybe using dummy data, partial code e.t.c.) how you would do them given you had done that part. You *might* be eligible for partial points.

Assignment 1 (10p)

The goal is to draw an i.i.d sample from the truncated to the positive half–axis standard normal distribution $\mathcal{N}(0,1)$, whose probability density function is

$$f(x) = \sqrt{\frac{2}{\pi}} \exp\left(-x^2/2\right) \mathbf{1}_{[0,\infty)}(x).$$

The notation $\mathbf{1}_A(x)$, for a set A, means the indicator function of the set A, i.e.

$$\mathbf{1}_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A. \end{cases}$$

The above considered density can be written as a mixture density

$$f(x) = \alpha_1 f_1(x) g_1(x) + \alpha_2 f_1(x) g_1(x),$$

where

$$\begin{array}{ll} \alpha_1 = \sqrt{\frac{2}{\pi}} & \alpha_2 = \frac{1}{\sqrt{2\pi}} \\ f_1(x) = \mathbf{1}_{[0,1]}(x) & f_1(x) = 2\exp\left(-2(x-1)\right)\mathbf{1}_{(1,\infty)}(x) \\ g_1(x) = \exp\left(-x^2/2\right) & g_1(x) = \exp\left(-(x-2)^2/2\right). \end{array}$$

Question 1.1 (6p)

Implement a rejection–sampling algorithm to sample from f.

TIP: The first step is to decide from which component of the mixture to sample from. Notice that $\alpha_1/(\alpha_1+\alpha_2) = 2/3$ and $\alpha_2/(\alpha_1+\alpha_2) = 1/3$. Therefore, use e.g. the uniform random number generator to decide from which part to sample. If you obtain an event with probability 2/3, then you need to dominate the first part and if with probability 1/3, then the second part. You do not need to consider the constants α_1 and α_2 in your majorizing constant, as this was "taken care–off" in the step described in this tip.
TIP: Notice next by what can $g_1(x)$ and $g_2(x)$ be dominated. Plot them, recall properties of the exponential function.
TIP: Next notice that $f_2(x)$ is the exponential distribution with rate 2 shifted by 1 to the right. One takes 1 instead of 0 as the minimal value, i.e. if $Y \sim \exp(2)$, then $Y+1$ is distributed according to $f_2(x)$. You are allowed to use `rexp()`.

Question 1.2 (2p)

In each step of the rejection sampling algorithm a proposed value may be accepted or rejected. Find the rejection rate, i.e. sample many points from $f(x)$ and for each point see how many times a rejection event took place. In particular present the histogram of the amount of rejections per sampled point, the expected value and variance of the amount of rejections per sampled point.

Question 1.3 (2p)

Evaluate how good your rejection sampler is by looking at the expectation, variance and histogram of your generated sample. Compare to the mean, variance and histogram of the considered truncated standard normal. You may use `rnorm()` to simulate from the standard normal and then just throw away all negative values to get a sample from the desired distribution. Use `t.test()` to compare the mean and `var.test()` for the variance.

TIP: Your generated sample should be of size about 10000 to get decently looking histograms. If your implementation is OK, a sample of this size is generated very quickly.

Assignment 2 (10p)

You are asked to estimate the following integral

$$\mathcal{I} = \int\limits_{-\infty}^{\infty} \exp\left(\sin(x)\right) \exp\left(-x^2/2\right) \mathrm{d}x.$$

Question 2.1 (3p)

Implement a Monte Carlo integration method to estimate $\mathcal{I}$. What is your estimate of it?

Question 2.2 (2p)

Rerun your estimator a (large) number of times with different settings to explore how the variance of $\hat{\mathcal{I}}$ behaves.

Question 2.3 (4p)

It is possible to improve a Monte Carlo estimator by dividing the domain of integration and running it separately on each subset. In particular consider the integral

$$\mathcal{J} = \int\limits_{-\infty}^{\infty} h(x) f(x) \mathrm{d}x$$

and its estimator

$$\hat{\mathcal{J}} = \sum_{j=1}^{p} \frac{p_j}{n} \sum_{i=1}^{n} h(Y_i^j),$$

where

$$p_j = \int\limits_{-x_{j-1}}^{x_j} f(x)\mathrm{d}x,\ x_0 = -\infty,\ x_p = \infty, \quad (-\infty, \infty) = \bigcup_{j=1}^{p} \mathcal{X}_j,$$

where for $j = 1, \ldots, p-1$ $\mathcal{X}_j = (x_{j-1}, x_j]$ and $\mathcal{X}_p = (x_{p-1}, \infty)$, and $(Y_1^j, \ldots, Y_n^j)$ is an i.i.d. sample from $f(x)\mathbf{1}_{X_j}(x)$, i.e. $f(x)$ truncated to the interval $\mathcal{X}_j$.

Take $p = 2$, $\mathcal{X}_1 = (-\infty, 0]$ and $\mathcal{X}_2 = (0, \infty)$. Implement such an improved Monte Carlo estimator for $\mathcal{I}$. What estimate of it do you obtain now?

TIP: Remember that if $X \sim \mathcal{N}(0,1)$, then also $-X \sim \mathcal{N}(0,1)$. You may (but do not have to) use `rnorm()`.

Question 2.4 (1p)

Rerun your estimator a (large) number of times with different settings to explore how the variance of the new estimator of $\mathcal{I}$ behaves. How does the variance compare to that of the original estimator?