CS215 Assignments

Data Analysis and Interpretation

Paarth Jain
190050076

Sahasra Ranjan
190050102

Sibasis Nayak
190050115

September 28, 2020

Assignment 3

1. Solution at the end in handwritten form. In part (c) I have used X_i as the geometric r.v instead of generalised Z. Name of the variable does not matter in context to that (c) part, but for just being rigorous X_i in that part can be considered as Z. The result i.e the expected value and variance of the geometric value remain unchanged though.

2. (a) Distributon function F is reperesented as $F(x) = P(X \leq x)$. Now we know that F is an increasing function for any distribution. In this case however we are given that F is invertible which implies that F(x) is strictly invertible. i.e.

$$F(x_1) > F(x_2) \iff x_1 > x_2$$

Further, by substituting $y1 = F(x_1)$ and $y_2 = F(x_2)$, and using the fact that $F^{-1}(x)$ exists, we can rewrite the above equation as,

$$y_1 > y_2 \iff F^{-1}(y_1) > F^{-1}(y_2)$$

This means that F^{-1} is also an increasing function. Now, let u_i be samples of a random variable U such that $U \sim U(0,1)$ (uniform distribution) (Given in question), and let $v_i = F^{-1}(u_i)$ be the samples of a random variable V whose distribution is to be found.Now,

$$P(U \leq x) = x \qquad \text{(by defn. uniform random variable)}$$

now as F^{-1} is increasing $U \leq x \iff F^{-1}(U) \leq F^{-1}(x)$. Now using this result and substituting $F^{-1}(x) = y$, we have-

$$P(F^{-1}(U) \leq F^{-1}(x)) = x$$
$$P(V \leq F^{-1}(x)) = x$$
$$P(V \leq y) = F(y)$$

Thus V (or equivalently v_i) follow the distribution F.

(b) We know that F is an increasing function (cumulative never decreases) i.e. $x_1 \leq x_2 \iff F(x_1) \leq F(x_2)$

$$P(D \geq d) = P\{max_x|\frac{\sum_i \mathbf{1}(Y_i \leq x)}{n} - F(x)| \geq d\}$$

$$= P\{max_x|\frac{\sum_i \mathbf{1}(F(Y_i) \leq F(x))}{n} - F(x)| \geq d\}$$

$$= P\{max_x|\frac{\sum_i \mathbf{1}(U_i \leq F(x))}{n} - F(x)| \geq d\}$$

substituting $y = F(x)$

$$P(D \geq d) = P\{max_{0\leq y\leq 1}|\frac{\sum_i \mathbf{1}(U_i \leq y)}{n} - y| \geq d\}$$

$$= P(E \geq d)$$

The random variable $max_x|F_e(x) - F(x)|$, in a way represents the error or deviation from the actual distribution during sampling, or max deviation of samples from the actual distribution.

Now, this error/deviation is same(in distribution) for all distributions as each of them are equal (in distribution; to the one where Y = U(0,1)). Thus by transitivity (of sorts) all are equal(in distribution) to each other). Thus the error in inverse sampling method is same as when taking direct samples from a distribution.

3. Solution at the end in handwritten form.

4. (a) Disjoint sets T and V made by taking first 750 elements(T) of Sample and last 250 elements(V) of sample.

   (b) let $\{t_i\}_{i=1}^{750}$ be elements of the set T. Then the pdf estimate can be written as-

$$\hat{p}(x;\sigma) = \frac{\sum_{i=1}^{n} exp(-(x-t_i)^2/2\sigma^2)}{n\sigma\sqrt{2\pi}}$$

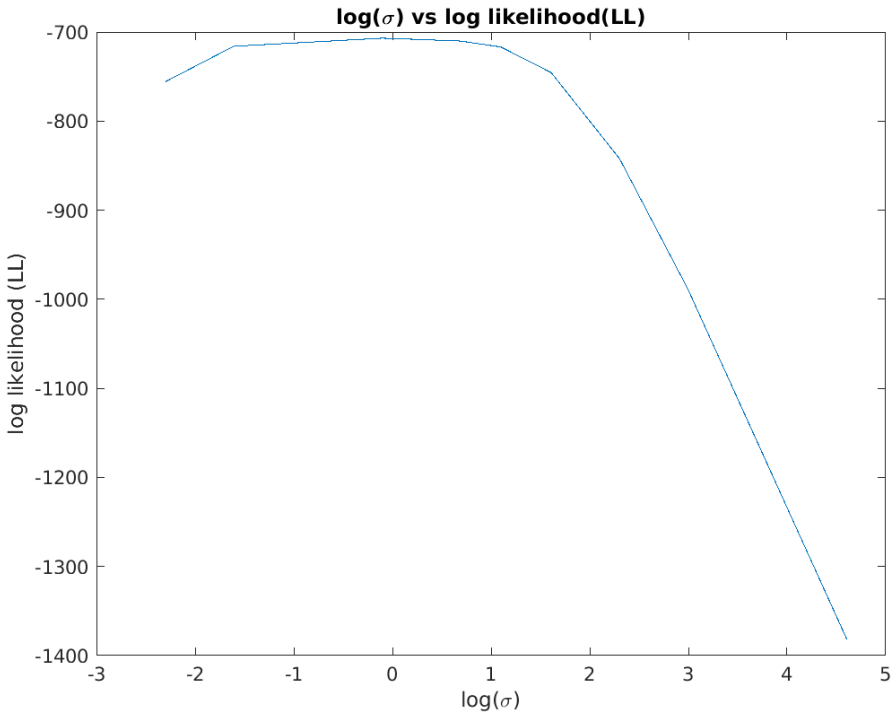
   Thus, for each $v_i \in V$

$$\hat{p}(v_i;\sigma) = \frac{\sum_{i=1}^{n} exp(-(v_i-t_i)^2/2\sigma^2)}{n\sigma\sqrt{2\pi}}$$

   Now, as v_i are independent of each other, the joint pdf

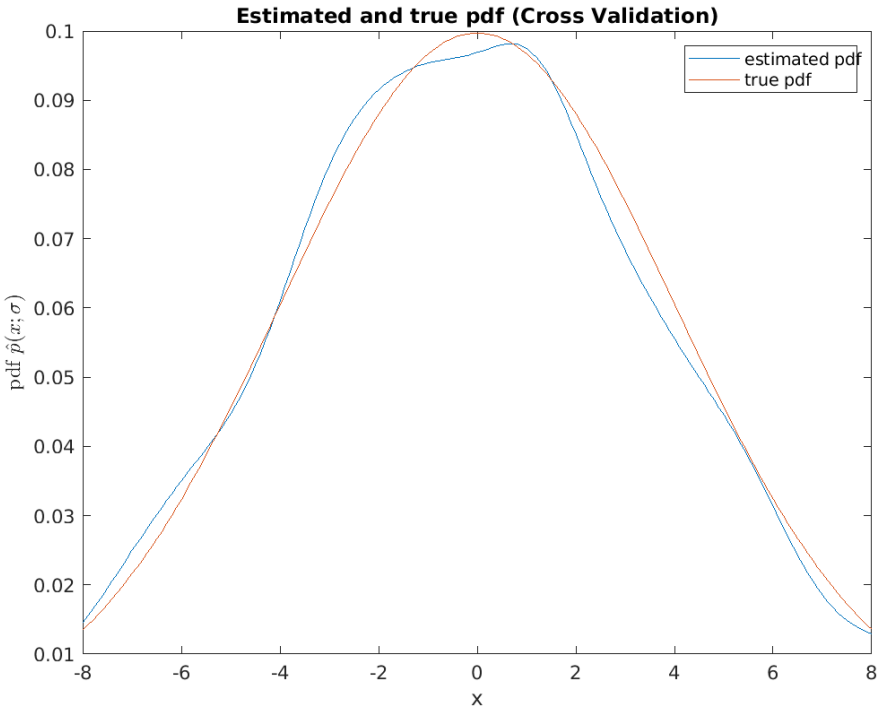
$$\hat{p}(v_1, v_2, ..., v_n;\sigma) = \prod_{i=1}^{n} \hat{p}(v_i;\sigma)$$

   (c) Plotting $log(\sigma)$ vs log likelihood gives-

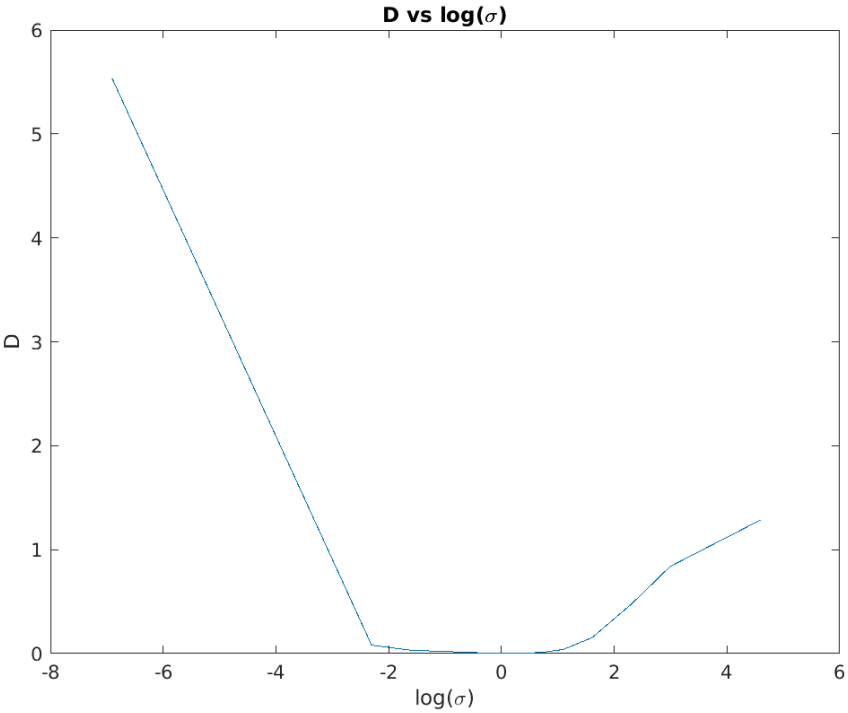


In this case the σ for which log likelihood is maximum comes out to be 0.9, but in general the best sigma assumes 0.9 or 1 value, occasionally deviations are seen (in different runs of the program).(Seeding the program can remove that discrepancy but the question does not ask us to do so.

Now plotting the estimated pdf and overlapping it with true pdf.

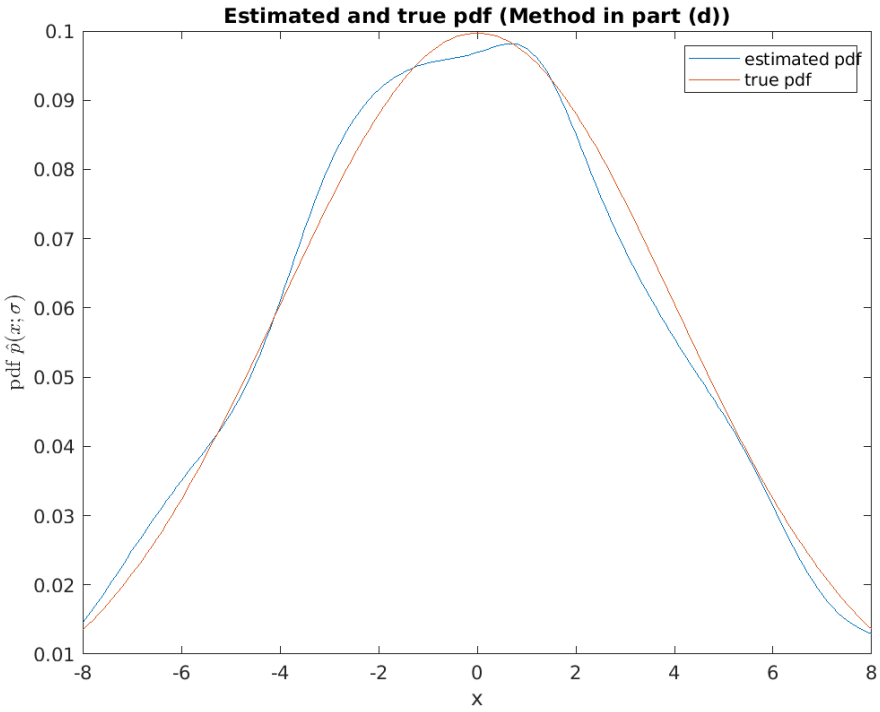


(d) Plotting $log(\sigma)$ vs D gives-

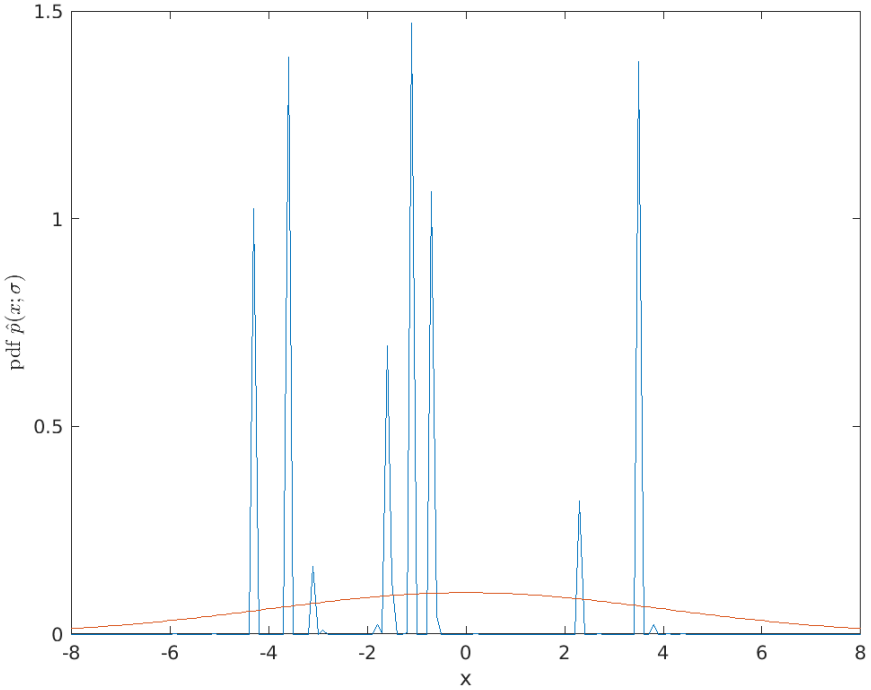


In this case the best sigma (for which D is minimum) comes out to be 1, slightly deviating from the best sigma found in part (c).

D corresponding to this best sigma = 0.0021
D corresponding to best sigma of part c = 0.0024

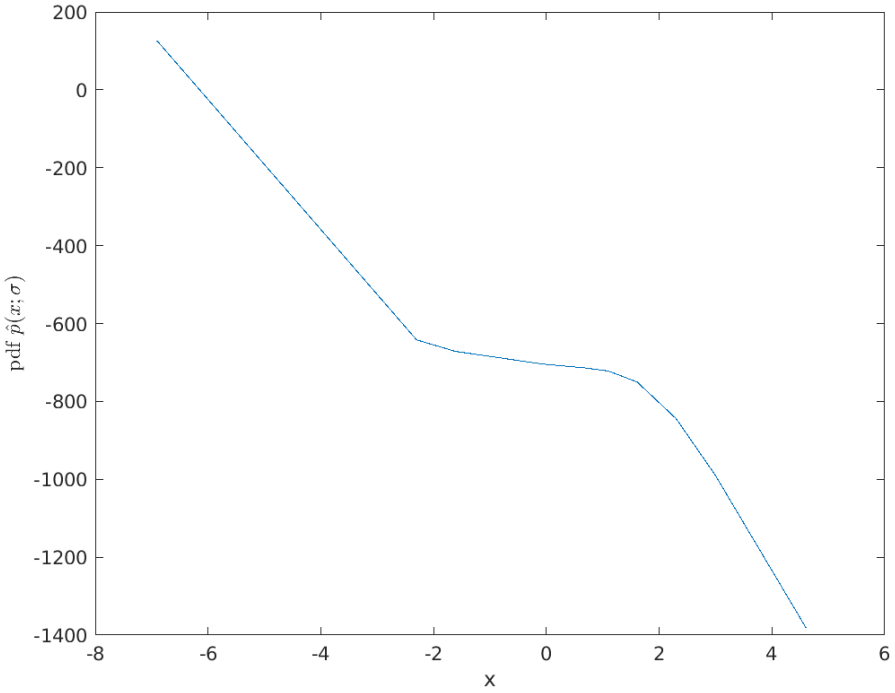


(e) When we take $T = V$, (and use the same joint likelihood function), the CV method, gives $\sigma = 0.001$. And the squared error (D value) for it is very high. The estimated pdf when overlapped with true pdf looks like this-(i have used T = V = first 250 elements of the sample)



Thus the pdf estimation is highly incorrect here.

If we look at the $log(\sigma)$ vs log likelihood graph. Smaller values of sigma, seem to be giving higher log likelihood values.



Here is an explanation to the above observations.
When T equals V. the joint likelihood function looks like-

$$\hat{p}(t_1, t_2, ..., t_n; \sigma) = \prod_{i=1}^{n} \hat{t}(t_i; \sigma)$$

$$\hat{p}(t_1, t_2, ..., t_n; \sigma) = \prod_{i=1}^{n} (\frac{\sum_{j=1}^{n} exp(-(t_i - t_j)^2/2\sigma^2)}{n\sigma\sqrt{2\pi}})$$

Now, for each i (each $\hat{p}(t_i; \sigma)$) there is a term where $t_i = t_j$, thus that term becomes-

$$\frac{1}{n\sigma\sqrt{2\pi}}$$

For large sigma this term is not very big, but for sigma as small as 0.001, this term becomes huge compared to others. Now this comparison of big and small is relative and has a lot of factors in it.

First lets consider the denominator of each term, it is large for large σ and small for small σ.

Second, the numerator is an exponential term. Lets consider the power. (or rather its negative). it is $(t_i - t_j)^2/2\sigma^2$. Now the "behaviour" of difference in values of $t_i and t_j$ remains same in all σ. But we need to keep in mind that the difference varies in accordance with the original pdf. As its std. deviation is 4, the difference can reach order 10 (1-10) mostly. So σ values which are too less, (0.1, 0.001) make the value of the power term higher. Adding the negative sign, results in the exponent term being very small overall.Overshadowing the effect of small denominator.

Thus compared to different V and T, this term where $t_i = t_j$, adds large value to each estimate and

the log of their product (the log likelihood) becomes large. and it is chosen as the best sigma. The correct method to do it when T = V, should be to omit the $t_i = t_j$ term, i.e. the estimate for each t_i is calculated from n-1 variables that are **independent** of t_i. This was verified through a MATLAB program and the results where similar to when $T \cap V = \phi$.

Instructions to run code:

- For q4, run the file q4.m in matlab. The program will print out the required values(these vary with each run as seed is not provided (the question does not ask us to do so)). Also four plots each in different figures are printed out.
- For q1 and q3 run the code in matlab to generate the graph and generate the solutions. q3 prints a column vector corresponding to a,b and c and the noise varince term.