HW3 Q4

Kavya Gupta

September 2023

Instructions to Run the Code

The MATLAB Code for **part (c) and (d)** is contained in the file `A3Q4.m` present in the main zip directory. Running that file will give 4 plots **in the same order** :-

- `Q4_Pic1.png` : Contains Plot of Log Likelihood (LL) vs log(sigma)
- `Q4_Pic2.png` : Contains Plot of D Value (D) vs log(sigma)
- `Q4_Pic3.png` : Contains graph for plot of density with best sigma for **maximum LL** and true density.
- `Q4_Pic4.png` : Contains graph for plot of density with best sigma for **minimum D** and true density.

The code will also successfully output the required print statements in the command line output, please see that.
The MATLAB Code for **part (e)** is contained in the file `A3Q4_2.m` present in the main zip directory. Running that file will give 4 plots **in the same order** :-

- `Q4_Pic5.png` : Contains Plot of Log Likelihood (LL) vs log(sigma) (T = V)
- `Q4_Pic6.png` : Contains Plot of D Value (D) vs log(sigma) (T = V)
- `Q4_Pic7.png` : Contains graph for plot of density with best sigma for **maximum LL** and true density (T = V)
- `Q4_Pic8.png` : Contains graph for plot of density with best sigma for **minimum D** and true density (T = V)

The code will also successfully output the required print statements in the command line output, please see that.
Note that all these eight plots are already included in this pdf and are also present in folder `Q4`. This folder is in main zip directory.

Part (a)

I have used `randperm` function to get permutation of the indices from 1 to n and took first 750 indices for Sample space `T` and rest for Validation space `V`. Also I have set the seed to 0 using `rng(0);`, so that I can reproduce same result every time.

Part (b)

We are given the Likelihood of one point x : $\hat{p_n}(x, \sigma)$. So, let's say that i^{th} entry of V as v_i and j^{th} entry of T as t_j, hence likelihood for 1 point of V :-

$$\hat{p_n}(v_i, \sigma) = \frac{\sum_{j=1}^{|T|} e^{-\frac{(v_i - t_j)^2}{2\sigma^2}}}{|T|\sigma\sqrt{2\pi}}$$

Here $|T|$ represents cardinality of set T and hence its size.
So the Joint Likelihood for $\{v_i\}_{i=1}^{|V|}$ will be :-

$$\hat{p_n}(\{v_i\}_{i=1}^{|V|}, \sigma) = \prod_{i=1}^{|V|} \hat{p_n}(v_i, \sigma) = \prod_{i=1}^{|V|} \frac{\sum_{j=1}^{|T|} e^{-\frac{(v_i - t_j)^2}{2\sigma^2}}}{|T|\sigma\sqrt{2\pi}}$$

We were able to split $\hat{p_n}(\{v_i\}_{i=1}^{|V|}, \sigma)$ into products as $\{v_i\}_{i=1}^{|V|}$ sampling is considered **independent**.

Part (c)

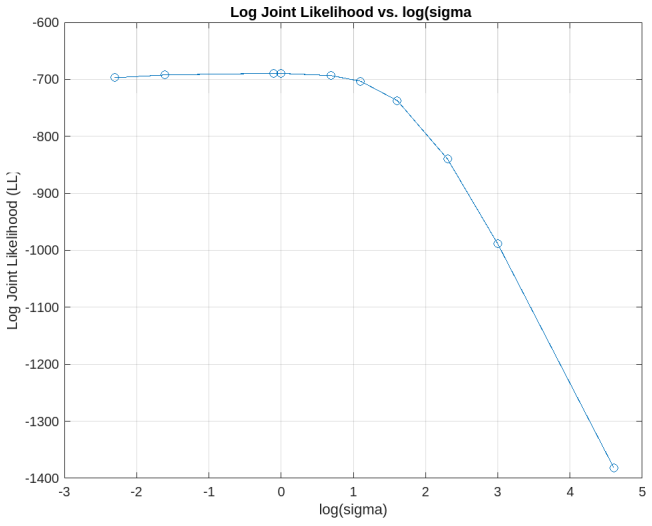

Figure 1: Plot of Log Likelihood (LL) vs log(sigma)

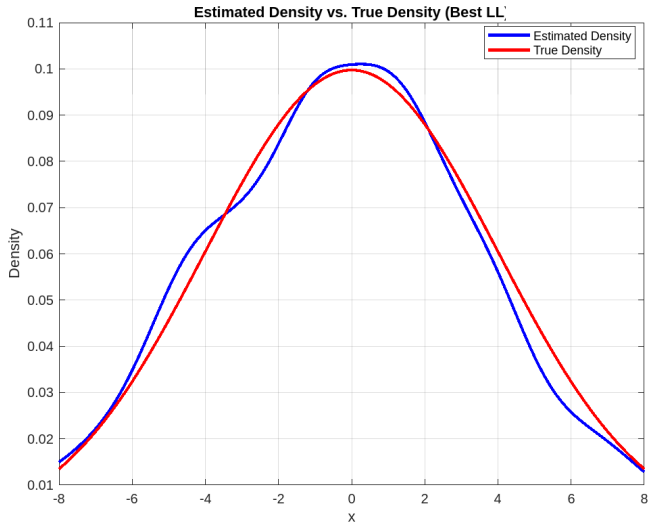

Figure 2: Graph for plot of density with best sigma for maximum LL = **0.9000** and true density

The value of σ for best LL : **0.9000**
Corresponding LL Value : **-689.3216**

Part (d)

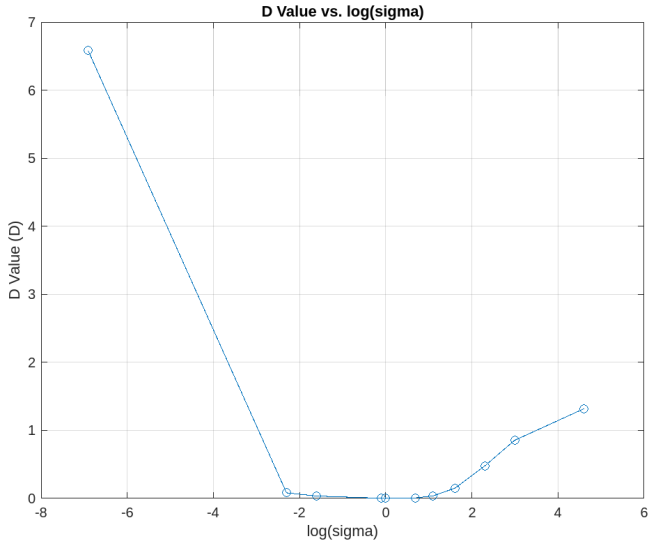

Figure 3: Plot of D value (D) vs log(sigma)

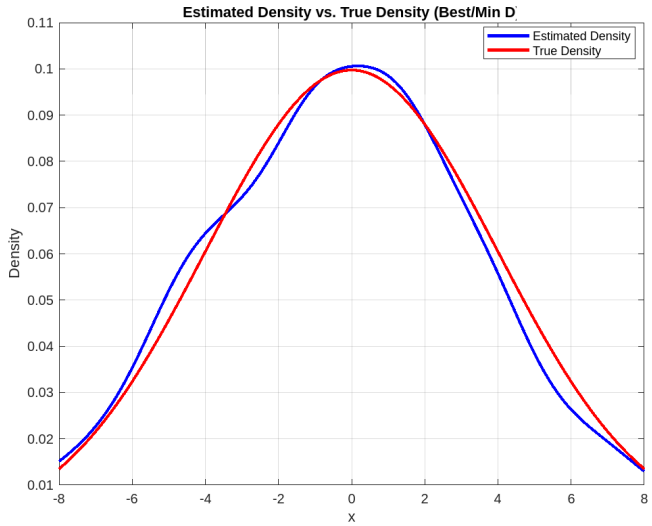

Figure 4: Graph for plot of density with best sigma for minimum D = **1.0000** and true density

The value of σ for best/min D : **1.0000**
Corresponding D Value : **0.0029**
D Value for Sigma with Max LL (1.0000) : **0.0034**

Worth noticing that the estimate graphs found by best LL and best D methods are both **quite close to each other**.

Part (e)

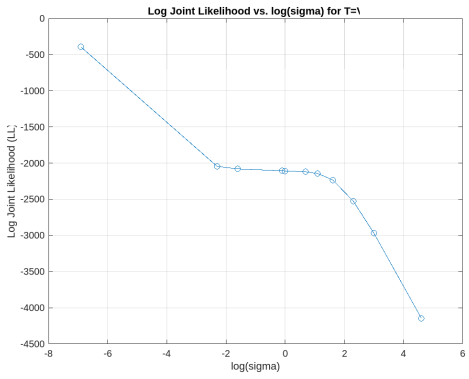

Figure 5: Plot of Log Likelihood (LL) vs log(sigma) (T=V)

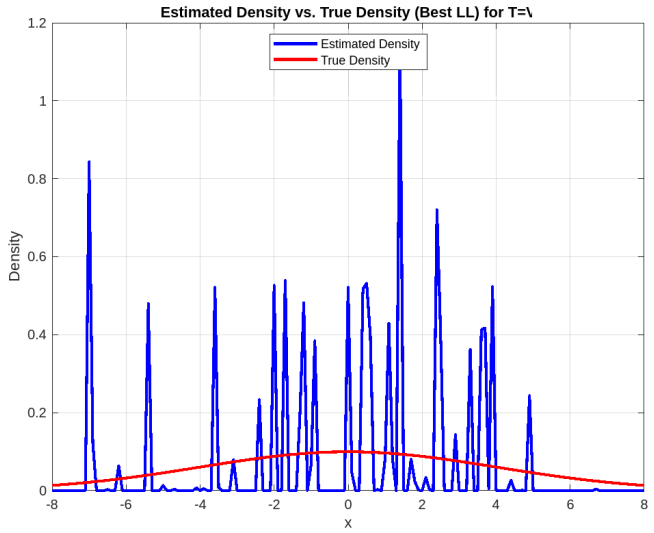

Figure 6: Graph for plot of density with best sigma for maximum LL = **0.001** and true density (T=V)

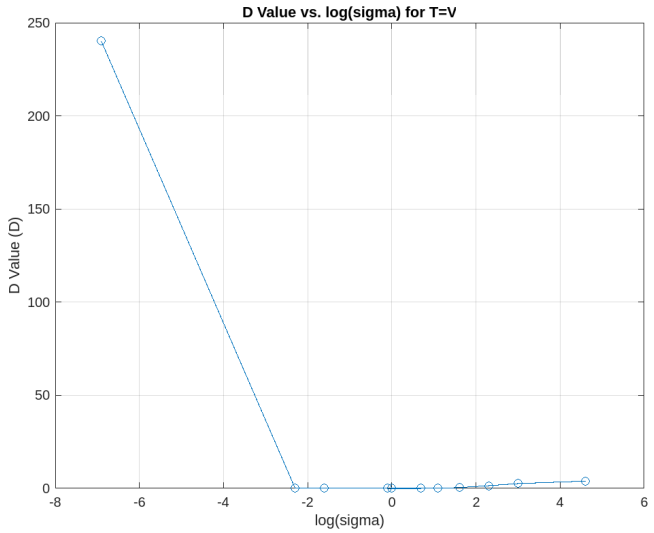

Figure 7: Plot of D value (D) vs log(sigma) (T=V)

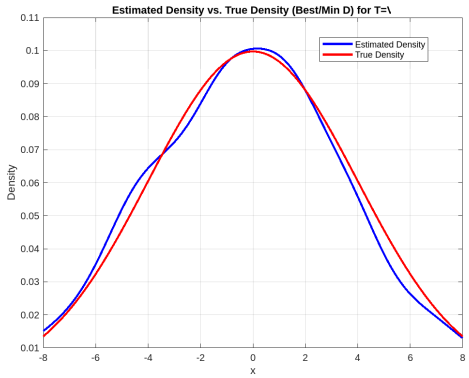

Figure 8: Graph for plot of density with best sigma for minimum D = **1.0000** and true density (T=V)

Best Sigma for Max. LL for T=V: **0.0010**
Best Sigma for Min. D for T=V: **1.0000**

So if the sets T and V became equal to each other, in other words, training sample and validation set became same; then what I observed was from the code that σ for best LL is the **smallest one** = 0.001 from the given set.
Meaning for smaller σ, likelihood is maximum.
Also the graph of LL vs log(sigma) seems to **greatly increase as log(sigma) tends to** $-\infty$**.**

Reason

When T and V become equal, the density is calculated at the same points from where PDF was estimated. So density of a term t_i will become :-

$$\hat{p_n}(t_i, \sigma) = \frac{\sum_{j=1}^{|T|} e^{-\frac{(t_i - t_j)^2}{2\sigma^2}}}{|T|\sigma\sqrt{2\pi}} = \frac{1 + \sum_{j \neq i} \cdots}{|T|\sigma\sqrt{2\pi}} = \frac{1}{|T|\sigma\sqrt{2\pi}} + \ldots$$

The above term becomes independent from the data and rather depends on σ**.**
This term $\frac{1}{|T|\sigma\sqrt{2\pi}}$ tends to ∞ as σ tends to 0, hence joint likelihood which is product of individual likelihoods also attains max near 0.
Hence max LL will be found at $\sigma \to 0^+$. **Hence we see spikes in data...**
Such a term exists as one term t_i exactly cancels out t_j for $i = j$.

Conclusion

When T and V match then our cross validation procedure tends to fail. **Even for max LL, we got the worst possible matching graph.**