Data Analysis and Interpretation
ASSIGNMENT - 3
REPORT

Shreshtha Gupta (Roll No. 24B1033)
Manvi Mehta (Roll No. 24B1059)
Uma Kumari (Roll No. 24B1026)

Contents

Instructions to run our code

Submission Format

```
A2-24B1033-24B1059-24B1026/
├─ Question3/
│  └─ hw3_q3_d.m
├─ Question4/
│  ├─ hw3_q4_a.m
│  ├─ hw3_q4_c.m
│  └─ hw3_q4_d.m
└─ report.pdf
```

Detailed Instructions

- First unzip the submission folder.
- To see the results and solutions open report.pdf.
- To view the codes for question 3 open the folder 'Question3'.
- Similarly, to view the code for question 4, open the folder 'Question4'.
- The plots and comments for question 4 are given in report.pdf.

1 Question 1

$I_k(x, y) = J(x, y) + \omega_k(x, y)$ where $I_k(x, y)$ is the observed intensity at the pixel (x, y) in the k^{th} image where k ranges from 1 to N.

$J(x, y)$ is the true intensity at pixel (x, y) which is a constant value over all N images.

$\omega_k(x, y)$ is the random noise affecting the pixel at (x, y) in the k^{th} image.

Since true intensity is constant over all images, we can estimate it as

$$\hat{J}(x, y) = \frac{1}{N} \sum_{k=1}^{N} I_k(x, y)$$

As $N = 10000$ (i.e. N is large), $\hat{J}(x, y)$ provides a very close approximation to $J(x, y)$.

So Bias$(\hat{J}(x, y)) \approx 0$.

Now the noise sample for each image k at pixel (x, y) can be given as

$$\hat{\omega}_k(x, y) = I_k(x, y) - \hat{J}(x, y)$$

For a single pixel location, we have a set of $N = 10000$ samples of noise $\{\hat{\omega}_1(x, y), \hat{\omega}_2(x, y), \ldots, \hat{\omega}_N(x, y)\}$.

Obtaining noise distribution

The distribution of noise can be obtained by constructing a **histogram** of the above N samples.

This can be done as follows:

- Plot the **frequency** of the calculated noise values $\hat{\omega}_k(x, y)$ against the **noise value**.

The histogram itself is the empirical estimate of PDF of noise that affects a pixel with true intensity $J(x, y)$.

2 Question 2

2.1 2. (a)

$$\begin{aligned}
F_V(v) &= P(V \leq v) \\
&= P(F^{-1}(U) \leq v) \\
&= P(U \leq F(v)) \\
&= F_U(F(v)) \\
&= F(v) \qquad (\because U \text{ is uniform distribution})
\end{aligned}$$

where $U \sim \text{Uniform}[0, 1]$, hence $F_U(u) = u$ so that $P(U \leq u) = u$, where $u \in [0, 1]$.

Thus V has the distribution F.

2.2 2. (b)

$$\begin{aligned}
P(D \geq d) &= P\left\{\max_x \left|\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(Y_i \leq x) - F(x)\right| \geq d\right\} \\
&= P\left\{\max_x \left|\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(F(Y_i) \leq F(x)) - F(x)\right| \geq d\right\} \\
&= P\left\{\max_x \left|\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(U_i \leq F(x)) - F(x)\right| \geq d\right\}
\end{aligned}$$

($\because$ as found in previous part, Y_i follows distribution F, then $Y_i = F^{-1}(U_i) \implies U_i = F(Y_i)$)

If $y = F(x)$, then

$$P\left\{\max_{0\leq y\leq 1}\left|\frac{1}{n}\sum_{i=1}^{n} 1(U_i \leq y) - y\right| \geq d\right\}$$

(The range of y is $[0,1]$ because $y = F(x) = U_i$ and $U_i \in [0,1]$)

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

2.3 2. (c)

$P(D \geq d)$ is thus proved to be independent of the distribution F. It can be used to check whether the distribution of given data matches a known distribution F.

If indeed the data belong to F, then the value of D **should not exceed** the corresponding difference between the empirical CDF of Unif$(0,1)$ random variables and the true Unif$(0,1)$ distribution. (Rather the **probability of this happening will be low**).

3 Question 3

3.1 3. (a)

Log-Likelihood Function

Equation of plane:

$$z = ax + by + c.$$

We know that Noise $(\varepsilon) \sim \mathcal{N}(0, \sigma^2)$. So, for the i-th point,

$$z_i = ax_i + by_i + c + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2).$$

Thus,

$$z_i \sim \mathcal{N}(ax_i + by_i + c,\, \sigma^2).$$

Hence,

$$p(z_i \mid x_i, y_i, a, b, c) = \frac{e^{-\frac{(z_i - ax_i - by_i - c)^2}{2\sigma^2}}}{\sqrt{2\pi}\,\sigma}.$$

Joint likelihood:

$$\mathcal{L}(a,b,c) = \prod_{i=1}^{n} \frac{e^{-\frac{(z_i - ax_i - by_i - c)^2}{2\sigma^2}}}{\sqrt{2\pi}\,\sigma}.$$

Log-likelihood:

$$\ell(a,b,c) = \log \mathcal{L}(a,b,c) = \sum_{i=1}^{n}\left(-\frac{(z_i - ax_i - by_i - c)^2}{2\sigma^2} - \log\sqrt{2\pi} - \log\sigma\right).$$

$$\ell(a,b,c) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(z_i - ax_i - by_i - c)^2 - n\log\sqrt{2\pi} - n\log\sigma.$$

Partial Derivatives

$$\frac{\partial \ell(a,b,c)}{\partial a} = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(-2x_i(z_i - ax_i - by_i - c)) = \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i(z_i - ax_i - by_i - c). \quad \text{(I)}$$

$$\frac{\partial \ell(a,b,c)}{\partial b} = \frac{1}{\sigma^2}\sum_{i=1}^{n} y_i(z_i - ax_i - by_i - c). \quad \text{(II)}$$

$$\frac{\partial \ell(a,b,c)}{\partial c} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(z_i - ax_i - by_i - c). \quad \text{(III)}$$

Linear Equations

For MLE, we solve

$$\frac{\partial \ell}{\partial a} = \frac{\partial \ell}{\partial b} = \frac{\partial \ell}{\partial c} = 0.$$

By (I),

$$\frac{1}{\sigma^2}\sum_{i=1}^{n} x_i(z_i - ax_i - by_i - c) = 0$$

$$\iff \sum_{i=1}^{n} x_i z_i - a\sum_{i=1}^{n} x_i^2 - b\sum_{i=1}^{n} x_i y_i - c\sum_{i=1}^{n} x_i = 0$$

By (II),

$$\frac{1}{\sigma^2}\sum_{i=1}^{n} y_i(z_i - ax_i - by_i - c) = 0$$

$$\iff \sum_{i=1}^{n} y_i z_i - a\sum_{i=1}^{n} x_i y_i - b\sum_{i=1}^{n} y_i^2 - c\sum_{i=1}^{n} y_i = 0$$

By (III),

$$\frac{1}{\sigma^2}\sum_{i=1}^{n} z_i(z_i - ax_i - by_i - c) = 0$$

$$\iff \sum_{i=1}^{n} z_i^2 - a\sum_{i=1}^{n} x_i z_i - b\sum_{i=1}^{n} y_i z_i - nc = 0$$

Matrix and Vector Form

By the equation of the plane,

$$\vec{Z} = \begin{pmatrix}\vec{X} & \vec{Y} & \vec{1}\end{pmatrix}\begin{pmatrix}a\\ b\\ c\end{pmatrix}$$

$$\vec{\theta} = \begin{pmatrix}a\\ b\\ c\end{pmatrix}$$

$$X = \begin{pmatrix} \vec{X} & \vec{Y} & \vec{1} \end{pmatrix}$$

$$\vec{X} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \quad \vec{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

Multiplying by X^T on both sides,

$$X^T \vec{Z} = X^T X \vec{\theta}$$

Multiplying by $(X^T X)^{-1}$ on both sides,

$$(X^T X)^{-1} X^T \vec{Z} = I \vec{\theta}$$

or,

$$\boxed{\vec{\theta} = (X^T X)^{-1} X^T \vec{Z}}$$

3.2 3. (b)

Log-Likelihood

Equation of plane $Z = a_1x^2 + a_2y^2 + a_3xy + a_4x + a_5y + a_6$. We know that Noise($\epsilon$) is $N(0, \sigma^2)$.

So for i^{th} point,

$$z_i = a_1x_i^2 + a_2y_i^2 + a_3x_iy_i + a_4x_i + a_5y_i + a_6 + \epsilon_i$$

$$\epsilon_i \sim N(0, \sigma^2)$$

Thus, $Z_i \sim N\left(a_1x_i^2 + a_2y_i^2 + a_3x_iy_i + a_4x_i + a_5y_i + a_6, \sigma^2\right)$

Hence,

$$p(z_i|x_i, y_i, a_1, \ldots, a_6) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(z_i - a_1x_i^2 - a_2y_i^2 - a_3x_iy_i - a_4x_i - a_5y_i - a_6)^2}{2\sigma^2}}$$

Joint Likelihood:

$$\mathcal{L}(a_1, a_2, a_3, a_4, a_5, a_6) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(z_i - a_1x_i^2 - a_2y_i^2 - a_3x_iy_i - a_4x_i - a_5y_i - a_6)^2}{2\sigma^2}}$$

log-likelihood:

$$\log \mathcal{L}(a_1, \ldots, a_6) = \sum_{i=1}^{n} \left(-\frac{(z_i - a_1x_i^2 - a_2y_i^2 - a_3x_iy_i - a_4x_i - a_5y_i - a_6)^2}{2\sigma^2} - \log\sqrt{2\pi} - \log\sigma \right)$$

$$l(a_1, a_2, a_3, a_4, a_5, a_6) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (z_i - a_1x_i^2 - a_2y_i^2 - a_3x_iy_i - a_4x_i - a_5y_i - a_6)^2 - n\log\sqrt{2\pi} - n\log\sigma$$

Partial Derivatives

$$\frac{\partial l(a_1,a_2,a_3,a_4,a_5,a_6)}{\partial a_1} = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[-2(x_i^2)\left(z_i - a_1x_i^2 - a_2y_i^2 - a_3x_iy_i - a_4x_i - a_5y_i - a_6\right)\right]$$

$$= +\frac{1}{\sigma^2}\sum_{i=1}^{n} x_i^2\left(z_i - a_1x_i^2 - a_2y_i^2 - a_3x_iy_i - a_4x_i - a_5y_i - a_6\right) \quad \text{— (I)}$$

$$\frac{\partial l(a_1,a_2,a_3,a_4,a_5,a_6)}{\partial a_2} = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[-2(y_i^2)\left(z_i - a_1x_i^2 - a_2y_i^2 - a_3x_iy_i - a_4x_i - a_5y_i - a_6\right)\right]$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{n} y_i^2\left(z_i - a_1x_i^2 - a_2y_i^2 - a_3x_iy_i - a_4x_i - a_5y_i - a_6\right) \quad \text{— (II)}$$

$$\frac{\partial l(a_1,a_2,a_3,a_4,a_5,a_6)}{\partial a_3} = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[-2(x_iy_i)\left(z_i - a_1x_i^2 - a_2y_i^2 - a_3x_iy_i - a_4x_i - a_5y_i - a_6\right)\right]$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{n} x_iy_i\left(z_i - a_1x_i^2 - a_2y_i^2 - a_3x_iy_i - a_4x_i - a_5y_i - a_6\right) \quad \text{— (III)}$$

$$\frac{\partial l(a_1,a_2,a_3,a_4,a_5,a_6)}{\partial a_4} = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[-2(x_i)\left(z_i - a_1x_i^2 - a_2y_i^2 - a_3x_iy_i - a_4x_i - a_5y_i - a_6\right)\right]$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i\left(z_i - a_1x_i^2 - a_2y_i^2 - a_3x_iy_i - a_4x_i - a_5y_i - a_6\right) \quad \text{— (IV)}$$

$$\frac{\partial l(a_1,a_2,a_3,a_4,a_5,a_6)}{\partial a_5} = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[-2(y_i)\left(z_i - a_1x_i^2 - a_2y_i^2 - a_3x_iy_i - a_4x_i - a_5y_i - a_6\right)\right]$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{n} y_i\left(z_i - a_1x_i^2 - a_2y_i^2 - a_3x_iy_i - a_4x_i - a_5y_i - a_6\right) \quad \text{— (V)}$$

$$\frac{\partial l(a_1,a_2,a_3,a_4,a_5,a_6)}{\partial a_6} = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[-2(1)\left(z_i - a_1x_i^2 - a_2y_i^2 - a_3x_iy_i - a_4x_i - a_5y_i - a_6\right)\right]$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{n}\left[z_i - a_1x_i^2 - a_2y_i^2 - a_3x_iy_i - a_4x_i - a_5y_i - a_6\right] \quad \text{— (VI)}$$

Linear Equations

For Maximum Likelihood estimate,

$$\frac{\partial l}{\partial a_i} = 0 \quad \text{for } i \in \{1,2,3,4,5,6\}$$

By (I), $\frac{\partial l}{\partial a_1} = 0$

$$\frac{1}{\sigma^2}\sum_{i=1}^{n} x_i^2\left(z_i - a_1x_i^2 - a_2y_i^2 - a_3x_iy_i - a_4x_i - a_5y_i - a_6\right) = 0$$

or,

$$\sum_{i=1}^{n} x_i^2 z_i = \sum_{i=1}^{n}\left(a_1x_i^4 + a_2x_i^2y_i^2 + a_3x_i^3y_i + a_4x_i^3 + a_5x_i^2y_i + a_6x_i^2\right)$$

By (II), $\frac{\partial l}{\partial a_2} = 0$

$$\frac{1}{\sigma^2}\sum_{i=1}^{n} y_i^2\left(z_i - a_1x_i^2 - a_2y_i^2 - a_3x_iy_i - a_4x_i - a_5y_i - a_6\right) = 0$$

or,

$$\sum_{i=1}^{n} y_i^2 z_i = \sum_{i=1}^{n}\left(a_1x_i^2y_i^2 + a_2y_i^4 + a_3x_iy_i^3 + a_4x_iy_i^2 + a_5y_i^3 + a_6y_i^2\right)$$

By (III), $\frac{\partial l}{\partial a_3} = 0$

$$\frac{1}{\sigma^2}\sum_{i=1}^{n} x_iy_i\left(z_i - a_1x_i^2 - a_2y_i^2 - a_3x_iy_i - a_4x_i - a_5y_i - a_6\right) = 0$$

or,

$$\sum_{i=1}^{n} x_iy_iz_i = \sum_{i=1}^{n}\left(a_1x_i^3y_i + a_2x_iy_i^3 + a_3x_i^2y_i^2 + a_4x_i^2y_i + a_5x_iy_i^2 + a_6x_iy_i\right)$$

By (IV), $\frac{\partial l}{\partial a_4} = 0$

$$\frac{1}{\sigma^2}\sum_{i=1}^{n} x_i\left(z_i - a_1x_i^2 - a_2y_i^2 - a_3x_iy_i - a_4x_i - a_5y_i - a_6\right) = 0$$

or,

$$\sum_{i=1}^{n} x_iz_i = \sum_{i=1}^{n}\left(a_1x_i^3 + a_2x_iy_i^2 + a_3x_i^2y_i + a_4x_i^2 + a_5x_iy_i + a_6x_i\right)$$

By (V), $\frac{\partial l}{\partial a_5} = 0$

$$\frac{1}{\sigma^2}\sum_{i=1}^{n} y_i\left(z_i - a_1x_i^2 - a_2y_i^2 - a_3x_iy_i - a_4x_i - a_5y_i - a_6\right) = 0$$

or,

$$\sum_{i=1}^{n} y_iz_i = \sum_{i=1}^{n}\left[a_1x_i^2y_i + a_2y_i^3 + a_3x_iy_i^2 + a_4x_iy_i + a_5y_i^2 + a_6y_i\right]$$

By (VI), $\frac{\partial l}{\partial a_6} = 0$

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}\left(z_i - a_1x_i^2 - a_2y_i^2 - a_3x_iy_i - a_4x_i - a_5y_i - a_6\right) = 0$$

or,

$$\sum_{i=1}^{n} z_i = \sum_{i=1}^{n}\left(a_1x_i^2 + a_2y_i^2 + a_3x_iy_i + a_4x_i + a_5y_i + a_6\right)$$

Matrix and Vector Form

By the equation of the plane,

$$\vec{z} = \begin{pmatrix} \vec{X^2} & \vec{Y^2} & \vec{XY} & \vec{X} & \vec{Y} & \vec{1} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \end{pmatrix}$$

$$X = \begin{pmatrix} \vec{X^2} & \vec{Y^2} & \vec{XY} & \vec{X} & \vec{Y} & \vec{1} \end{pmatrix}$$

Let

$$\vec{\theta} = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \end{pmatrix}$$

$$\vec{X} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \quad \vec{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

—

$$\vec{X^2} = \begin{pmatrix} x_1^2 \\ x_2^2 \\ \vdots \\ x_n^2 \end{pmatrix} \quad \vec{Y^2} = \begin{pmatrix} y_1^2 \\ y_2^2 \\ \vdots \\ y_n^2 \end{pmatrix}$$

—

$$\vec{XY} = \begin{pmatrix} x_1y_1 \\ x_2y_2 \\ \vdots \\ x_ny_n \end{pmatrix}$$

$$\vec{Z} = X\vec{\theta}$$

Multiplying X^T on both sides,

$$X^T\vec{Z} = X^TX\vec{\theta}$$

Multiplying $(X^TX)^{-1}$ on both sides,

$$(X^TX)^{-1}X^T\vec{Z} = I\vec{\theta}$$

or,

$$\boxed{\vec{\theta} = (X^TX)^{-1}X^T\vec{Z}}$$

3.3 3. (c)

Model fitting doesn't require noise variance

No, we generally don't need to know the noise to find the best plane. We just have to find the best plane which minimizes distances from all the points given. The amount of scatter (σ^2) doesn't change the best fit plane because we already know that the distribution is Gaussian. Mathematically, when we try to find the parameters of plane (a,b,c), the noise variance is a common factor and hence will get cancelled out from the equations. So, we can find the best fit plane even without noise variance.

Estimation of Noise Variance

Firstly, we will find the best fit plane, and then will measure how far off our data points are from that plane.

- Finding the best fit plane: As we did in the above parts, we will find out the paramater of plane (a,b,c)
- Error measurement: now, we will compare the actual heights (z-coordinates) to the height predicted by best fit plane.
  ERROR = actual z - predicted z
- Calculations: Variance is the average of squared errors: So,

$$\boxed{\hat{\sigma}^2 = \frac{1}{N-p}\sum_{i=1}^{N}(Z_{\text{actual},i} - Z_{\text{predicted},i})^2}$$

  where p is degrees of freedom i.e, p=3 for part (a) and p=6 for part (b)

3.4 3. (d)

Predicted plane equation: $Z = 10.0022 * X + 19.9980 * Y + 29.9516$
Predicted noise variance: 23.091606

3.5 3. (e)

Algorithm to estimate the plane parameters

To estimate the plane parameters in this case (when we have corrupted data values, many outliers), we will follow the following algorithm recursively:

- We will randomly pick 3 points from the data set because we need 3 points to define a unique plane.
- We will calculate the equation of plane that passes through these 3 points. This will be our candidate model.
- We will iterate over all other points (10000-3=99997) in our data set and will find its distance to our candidate plane. If it is less than some small threshold, we will call it inlier, otherwise outlier.
- We will count the total number of inliers we found for that particular plane.
- Recurse the process again many times(a few hundred or thousand times). Each time, we will get a new candidate plane with its corresponding score.

After all the iterations, the plane with highest score will be the final answer i.e. our best fit plane because that will agree with majority of the data. Also, for a better result, we can take all the inliers from the best-fit plane we just got and perform a final least-square fit using only these points. This will make the plane even more better.

Why it Works?

This algorithm will work because the problem states that only a small number of points are corrupted, so if we pick 3 points at random, its highly probable that all 3 will be good points (non-corrupted). A plane built from 3 good points will be very close to the actual plane. So, it will also be close to majority of the other good points, leading to a high score. But if the plane is built with even one corrupted point(outlier), it will create a garbage candidate plane as it won't align with many of the other points because most of them are non-corrupted, so it will get a very low score and be rejected. Hence, by repeating this process, we will ensure that eventually the highest scorer candidate plane will be the true best-fit plane, by successfully ignoring the outliers.

4 Question 4

4.1 4. (a)

```
clc; clear;
mu=0;
sigma=4;
n=1000;
samples = mu + sigma * randn(n, 1);
shuffle=samples(randperm(n));
T=shuffle(1:750);
V=shuffle(751:end);
disp(['Size of T: ',num2str(length(T))]);
disp(['Size of V: ',num2str(length(V))]);
disp(['Are T and V disjoint? ',num2str(isempty(intersect(T, V)))
    ]);
```

Size of T: 750
Size of V: 250
Are T and V disjoint? 1

4.2 4. (b)

Kernel Density Estimation (KDE) using samples in T is

$$\hat{P}_T(x;\sigma) = \frac{\sum_{i=1}^{n} \exp\left(-\frac{(x-x_i)^2}{2\sigma^2}\right)}{750\,\sigma\sqrt{2\pi}}.$$

This is estimated PDF of training set T.
Validation set $V = \{v_1, v_2, \ldots, v_{250}\}$.
Assuming samples in V to be independent,
So, the joint likelihood is

$$L(\sigma) = \prod_{j=1}^{250} \hat{P}_T(v_j;\sigma).$$

$$L(\sigma) = \prod_{j=1}^{250} \left[\frac{1}{750\,\sigma\sqrt{2\pi}} \sum_{i=1}^{750} \exp\left(-\frac{(v_j-x_i)^2}{2\sigma^2}\right)\right].$$

The expression for joint likelihood of samples in V,
based on estimate of PDF built from T,
with bandwidth parameter σ, is

$$\boxed{L(\sigma) = \prod_{j=1}^{250} \left[\frac{1}{750\,\sigma\sqrt{2\pi}} \sum_{i=1}^{750} \exp\left(-\frac{(v_j-x_i)^2}{2\sigma^2}\right)\right]}$$

4.3 4. (c)

Best sigma = 1.000000
Best Log-Likelihood = -696.757092

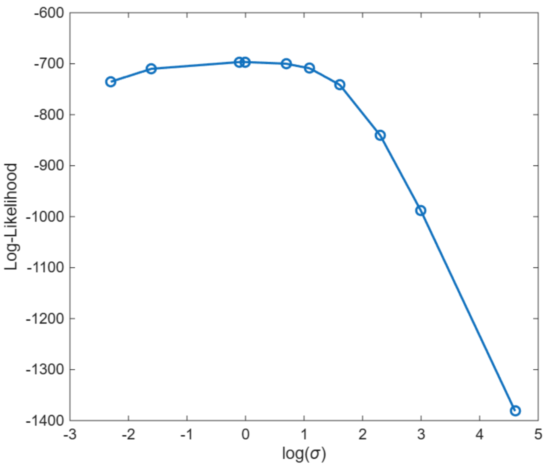

Figure 1: Log-Likelihood of validation set across different bandwidth values (σ)

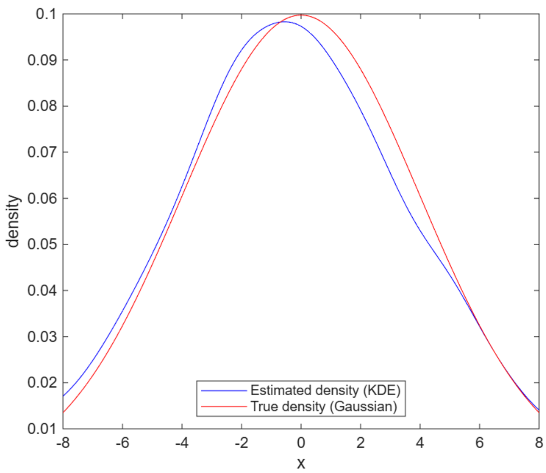

Figure 2: Estimated density (KDE) with optimal bandwidth vs true Gaussian density

4.4 4. (d)

Best sigma by LL = 0.900000 (LL=-703.368722)
Best sigma by D = 1.000000 (D=0.006646)
D value for sigma that maximized LL = 0.006955

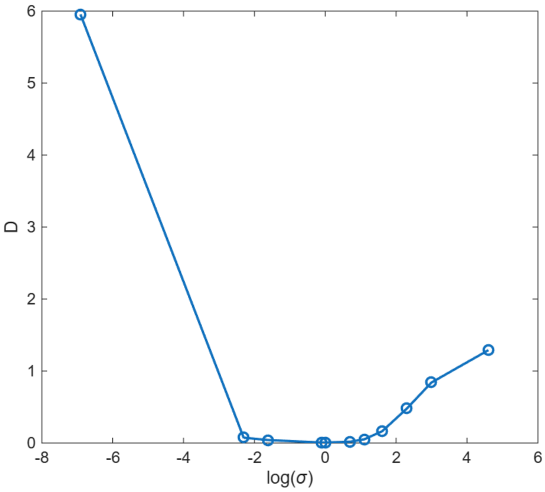

Figure 3: Squared error D between true Gaussian PDF and estimated KDE across different bandwidth values (σ)

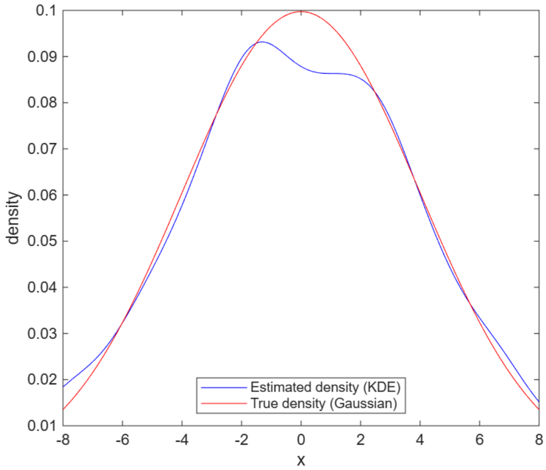

Figure 4: Estimated density (KDE with bandwidth minimizing D) vs true Gaussian density

4.5 4. (e)

$$LL = \sum_{x_i \in V} \log \hat{p}_n(x_i; \sigma)$$

If $T = V$, then the set of samples $\{x_i\}$ used to build the PDF is the exact same set of samples used to evaluate the likelihood.

$$LL = \sum_{x_i \in V} \log \left(\sum_{j=1}^{N} \frac{1}{n\sigma\sqrt{2\pi}} e^{-\frac{(x_i - x_j)^2}{2\sigma^2}} \right)$$

So here, when $x_i = x_j$ the term

$$\frac{1}{n\sigma\sqrt{2\pi}} e^{-\frac{(x_i - x_i)^2}{2\sigma^2}} = \frac{1}{n\sigma\sqrt{2\pi}}$$

So as σ approaches zero ($\sigma \to 0^+$), this term and thus $\hat{p}_n(x_i; \sigma)$ approaches infinity

$$\lim_{\sigma \to 0^+} \hat{p}_n(x_i; \sigma) = \infty$$

Since LL is the sum of log of these estimates, LL will also approach infinity. So this procedure of cross validation will always select the smallest value of σ to be the value maximizing LL, because LL is inversely proportional to σ (because LL is sum of log of $\hat{p}_n(x_i; \sigma)$ and $\hat{p}_n(x_i; \sigma)$ is inversely proportional to σ).