Assignment 3

Sagnik Nandi (23B0905)

Jainendrajeet (23B1008)

Sumedh S S (23B1079)

October 10, 2024

1 Finding optimal bandwidth

1.1 A

To prove:

$$\int \hat{f}(x)^2 dx = \frac{1}{n^2 h} \sum_{j=1}^{m} v_j^2 \tag{1}$$

$$\hat{f}(x) = \sum_{j=1}^{m} \frac{\hat{p}_j}{h} 1[x \in B_j] \tag{2}$$

Where $\hat{p_j}$ is the estimated probability that a points falls in the j^{th} bin i.e. $\hat{p_j} = \frac{v_j}{n}$, v_j is the number of points that fall in the j^{th} bin and n is the total number of points.

$$\begin{aligned}
\int \hat{f}(x)^2 dx &= \int \sum_{j=1}^{m} \left(\frac{v_j 1[x \in B_j]}{nh} \right)^2 dx \\
&= \int \sum_{i=1}^{m} \frac{v_i^2}{n^2 h^2} \, dx + \int \sum_{1 \le i < j \le m} \frac{v_i 1[x \in B_i] v_j 1[x \in B_j]}{n^2 h^2} \, dx \\
&= \left(\sum_{i=1}^{m} \frac{v_i^2}{n^2 h^2} \right) h, \text{ (since } 1[x \in B_i] 1[x \in B_j] = 0) \\
&= \sum_{i=1}^{m} \frac{v_i^2}{n^2 h}
\end{aligned}$$

B To prove:

$$\sum_{i=1}^{n} \hat{f}(X_i) = \frac{1}{(n-1)h} \sum_{j=1}^{m} v_j^2 - v_j \tag{3}$$

$$\hat{f}(x) = \sum_{j=1}^{m} \frac{\hat{p}_j}{h} I[x \in B_j] \tag{4}$$

$$\hat{f}_{-i}(X) = \sum_{j=1}^{m} \frac{v_{j-1}}{h(n-1)} I[x \in B_j] \tag{5}$$

Where $\hat{p_j}$ is the estimated probability that a points falls in the j^{th} bin i.e. $\hat{p_j} = \frac{v_j}{n}$, v_j is the number of points that fall in the j^{th} bin and n is the total number of points.

$$\sum_{i=1}^{n} \hat{f}_{-1}(X_i) = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{v_{j-1} \, I[X_i \in B_j]}{h(n-1)} \tag{6}$$

For $X_i \in B_k$,

$$\sum_{j=1}^{m} \frac{v_{j-1}\, I[X_i \in B_j]}{h(n-1)} = \frac{v_{k-1}}{h(n-1)} \tag{7}$$

Thus, the above summation(equation 6) becomes

$$\sum_{k=1}^{m} v_k \left(\frac{v_{k-1}}{h(n-1)} \right) = \frac{1}{h(n-1)} \sum_{k=1}^{m} v_k^2 - v_k. \tag{8}$$

1.2 A

The Estimated probability for different bins are following :
Estimated probability for bin 1: 0.2059
Estimated probability for bin 2: 0.4882
Estimated probability for bin 3: 0.0471
Estimated probability for bin 4: 0.0412
Estimated probability for bin 5: 0.1353
Estimated probability for bin 6: 0.0588
Estimated probability for bin 7: 0.0059
Estimated probability for bin 8: 0.0000
Estimated probability for bin 9: 0.0118
Estimated probability for bin 10: 0.0059

B The distributions seems oversmoothed.

D The optimal value of bin width h is **0.0684 Mpc**

E It has more number of bins than number of bins in part(a) and, It seems just-fit for the optimal h.

2 Detecting Anomalous Transactions using KDE

2.1 Designing a custom KDE Class

2.2 Estimating Distribution of Transactions

The resulting distribution contains **2** modes.

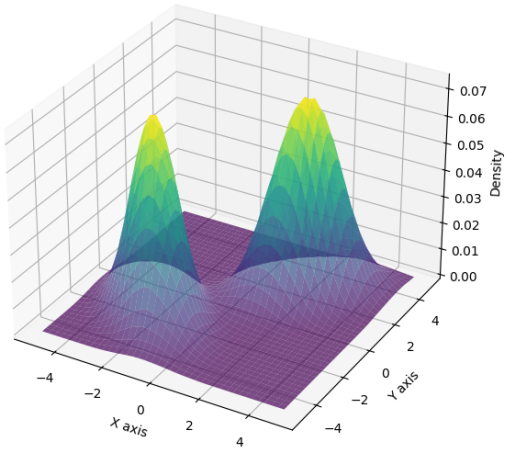

Figure 1: Density plot

3 Higher-Order Regression

3.1 Show that the point $(\bar{x}, \bar{y})$ lies exactly on the least squares regression line

Let regression model be Y= β_0 +Xβ_1 +ϵ where $\epsilon \sim N(0, \sigma^2)$.
Then the least square line is $Y = B_0 + B_1 X$ where $B_0 = \bar{y} - B_1\bar{x}$. Therefore the point $(\bar{x}, \bar{y})$ lies exactly on the least squares regression line.

3.2 Suppose that you as an analyst decide to use z = $x - \bar{x}$ as the regressor variable. So the new model becomes Y= β_0* +zβ_1* +ϵ. Find least squares estimates of β_0* and β_1*

$$
\begin{aligned}
Y &= \beta_0 * + z\beta_1 * + \epsilon = (\beta_0 * - \bar{x}\beta_1 *) + x\beta_1 * + \epsilon \\
Y &= \beta_0 + X\beta_1 + \epsilon \\
\beta_1 * &= \beta_1 \\
\beta_0 * - \bar{x}\beta_1 * &= \beta_0 \implies \beta_0 * = \beta_0 + \bar{x}\beta_1 = \bar{y}
\end{aligned}
$$

Relationship between models:

- Slope (β_1*) remains unchanged.

- Intercept (β_1*) shifts to mean of Y values.

Difference between models:

- Original model's intercept is at (0, β_0) centered model's intercept is at (0, $\bar{y}$).
- Centered predictors can offer better numerical properties in more complex models.

4 Non- Parametric Regression

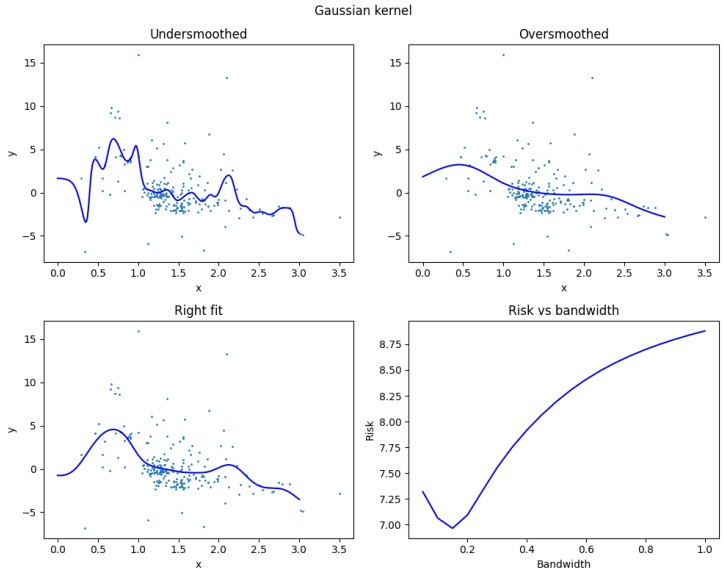

Figure 2: Gaussian kernel

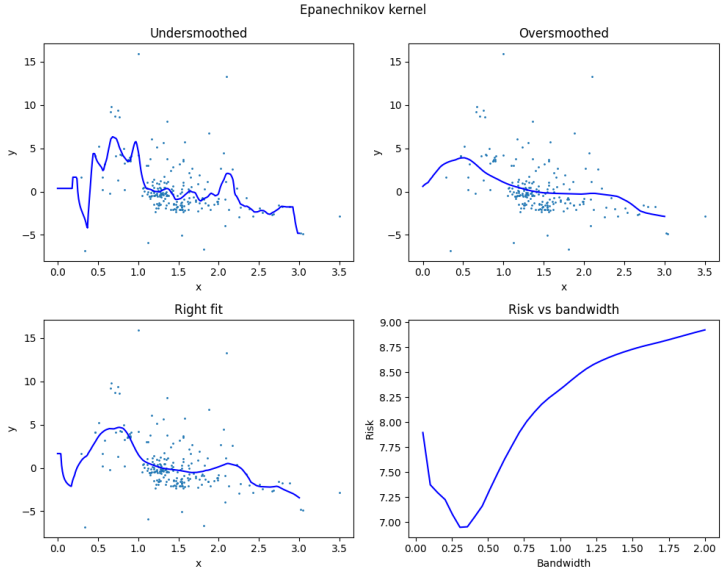

Figure 3: Epanechnikov kernel

Optimum Bandwidth

- For the case of gaussian kernel, the optimum bandwidth was obtained to be 0.15.
- For the case of epanechnikov kernel, the optimum bandwidth was obtained to be 0.3.

Effect of using different kernels

Similarities

- The choice of bandwidth is more important than the kernel as we found the optimum risk for both the cases to be about 6.9. Therefore, the choice of kernel doesnt affect performance in a significant manner.

Differences

- The optimal bandwidth varies on the choice of kernel used for ex, it was 0.15 for gaussian kernel and 0.3 for epanechnikov kernel
- Different kernels (e.g., Gaussian, Epanechnikov) offer varying levels of smoothness. A smoother kernel might generalize better to the underlying function, while a less smooth kernel could capture more local variations
- Some kernels have finite support (e.g., Epanechnikov), while others (e.g., Gaussian) do not. This affects how the influence of data points decays with distance. Finite support kernels can be computationally more efficient.

5 Wild Blueberry Yield Prediction Challenge!

Team Name- Sumedh1024

Email-IDs of teammates

- sumedh.ss37@gmail.com
- backup1.sagnik@gmail.com