# DAI Assignment-3

Aditya Neeraje, Balaji Karedla, Moulik Jindal

October 8, 2024

## Contents

# 1 Finding optimal bandwidth

## 1.1 Part 1

### 1.1.1 Part (a)

We know that the estimator for distribution function $\hat{f}$ is given by

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

Also given, $v_j$ is the number of points falling in the $j^{th}$ bin and $\hat{p}_j = \frac{v_j}{n}$.
From this,

$$\begin{aligned}
\int \hat{f}(x)^2 dx &= \int \left( \sum_{j=1}^{m} \frac{\hat{p}_j}{h} \mathbb{1}\left[x \in B_j\right] \right)^2 dx \\
&= \sum_{j=1}^{m} \int_{B_j} \left( \sum_{j=1}^{m} \frac{\hat{p}_j}{h} \mathbb{1}\left[x \in B_j\right] \right)^2 dx \\
&= \sum_{j=1}^{m} \int_{B_j} \left( \frac{\hat{p}_j}{h} \right)^2 dx \\
&= \sum_{j=1}^{m} \left( \frac{\hat{p}_j}{h} \right)^2 \times h \\
&= \frac{1}{n^2 h} \sum_{j=1}^{m} v_j^2
\end{aligned}$$

### 1.1.2 Part (b)

The histogram estimator after removing the $i^{th}$ observation is given by

$$\hat{f}_{(-i)}(x) = \sum_{j=1}^{m} \frac{\hat{p}_{j,-i}}{h} \mathbb{1}\left[x \in B_j\right] \tag{2}$$

where $\hat{p}_{j,-i} = \frac{v_{j,-i}}{n-1}$ and $v_{j,-i}$ is the number of points falling in the $j^{th}$ bin after removing the $i^{th}$ observation.
Now,

$$
\begin{aligned}
\sum_{i=1}^{n} \hat{f}_{(-1)}(X_i) &= \sum_{1}^{n} \sum_{j=1}^{m} \frac{\hat{p}_{j,-i}}{h} \mathbb{1}\left[X_i \in B_j\right] \\
&= \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{v_{j,-i}}{(n-1)h} \mathbb{1}\left[X_i \in B_j\right] \\
&= \sum_{j=1}^{m} \sum_{i=1}^{n} \frac{v_{j,-i}}{(n-1)h} \mathbb{1}\left[X_i \in B_j\right] \\
&= \sum_{j=1}^{m} \frac{v_j - 1}{(n-1)h} \mathbb{I}\left[X_i \in B_j\right] && \text{(since whenever } \mathbb{1}[X_i \in B_j] = 1, v_{j,-i} = v_j - 1\text{)} \\
&= \sum_{j=1}^{m} \frac{v_j - 1}{(n-1)h} v_j \\
&= \frac{1}{(n-1)h} \sum_{j=1}^{m} v_j^2 - v_j
\end{aligned}
$$

## 1.2 Part 2

### 1.2.1 Part (a)

The histogram of the filtered data for number of bins = 10 is shown in Figure 1.

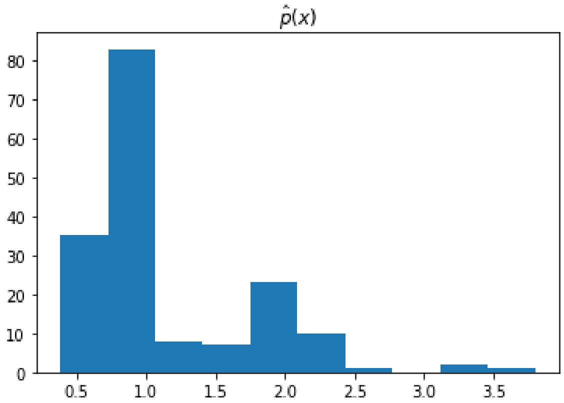

Figure 1: Histogram of filtered data with 10 bins

The values of $\hat{p}_j$ are as follows:

| Bin | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $\hat{p}_j$ | 0.602 | 1.428 | 0.138 | 0.12 | 0.396 | 0.172 | 0.017 | 0.0 | 0.034 | 0.017 |

### 1.2.2 Part (b)

The probablility distribution is underfit as the number of bins is too low. The distribution is not smooth and the bins are too wide. This is evident from the histogram in Figure 1.

### 1.2.3 Part (c)

The graph of $\hat{J}(h)$ vs $h$ for the number of bins ranging from 1 to 1000 is shown in Figure 2.

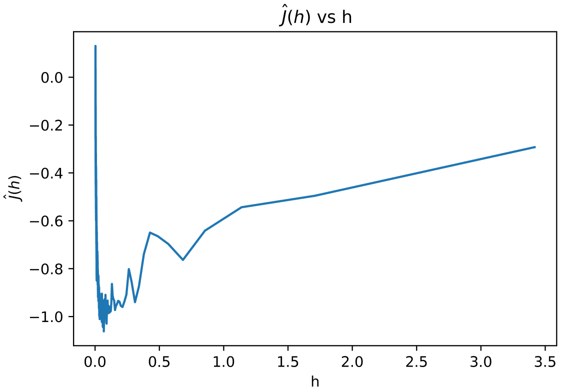

Figure 2: $\hat{J}(h)$ vs $h$

### 1.2.4 Part (d)

Based on the Cross Validation plot in Figure 2, the optimal number of bins is 50 with a value of $h^* = 0.06836$.

### 1.2.5 Part (e)

The histogram of the filtered data for number of bins = 50 is shown in Figure 3.

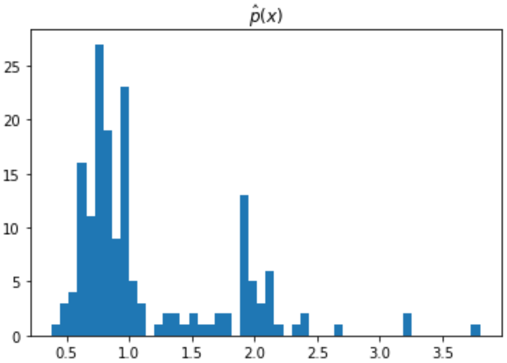

Figure 3: Histogram of filtered data with 50 bins

# 2 Detecting Anomalous Transactions using KDE

## 2.1 Designing a custom KDE Class

The KDE class with the Epachnikov kernel

$$K(x) = \frac{3}{4}(1 - ||x||_2^2) \quad \text{if} \quad ||x||_2 \leq 1 \tag{3}$$

The code was implemented in the file `2.py`.

## 2.2 Estimating Distribution of Transactions

After estimating the distribution of transactions using the Epachnikov kernel, the estimated probability density graph is shown in Figure 4.

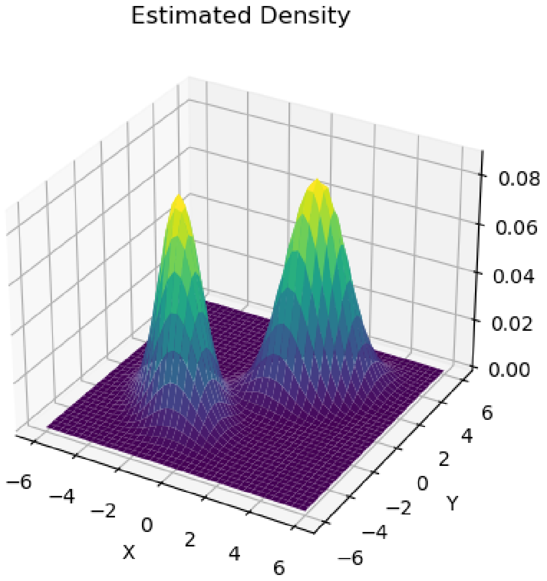

Figure 4: Estimated Probability Density of Transactions