Nonparametric and Semiparametric Data Modeling

MA750 Lecture Notes 2

Ashis Gangopadhyay

Boston University

Density Estimation

Initially, we will focus on the estimation of univariate density functions. There exist straightforward extensions of all of the methods discussed in this section to the multivariate setting, which we will discuss later.

- Let $X_1, ..., X_n$ i.i.d from unknown pdf $f(x)$.
- How to estimate $f(x)$?
- How about Maximum Likelihood Estimate(MLE)?

$$\sup_{\{f | f \geq 0, \int f = 1\}} \Pi_{i=1}^{n} f(X_i) = \infty$$

- Thus MLE doesn't exist!
- Reason: Maximizing over too large a class.
- New approach: Think about a smaller class of densities.
- One possibility: Restriction to step functions yields histogram estimator

Histogram

- The oldest and most commonly used density estimator is the histogram.
- Given an origin x_0 and a bandwidth (bin width) of h, define the bins of the histogram as the intervals $B_j = [x_0 + (j-1)h, x_0 + jh)$, for positive and negative integers j.
- The histogram estimator restricts the density to be constant within each interval. Suppose $x \in B_j$,
- Then the area under the curve $f(x)$ in bin B_j

$$P(X \in B_j) \approx \frac{\#X_i \text{ in the same bin as } x}{n} \tag{1}$$

- However, the same area under the curve $f(x)$ in bin B_j is also

$$\int_a^b \widehat{f}(x)dx = (b-a)\widehat{f}(x) = h\widehat{f}(x) \tag{2}$$

the first eqauality follows as the density is a constant in B_j.

- Therefore, from (1) and (2) the estimate for the density at x is

$$\widehat{f}(x) = \frac{\#X_i \text{ in the same bin as } x}{nh}$$

where n is the total number of observations.

- Formally, for $x \in B_j$,

$$\widehat{f}(x) = \frac{\#X_i \text{ in the same bin as } x}{nh} = \frac{f_j}{nh}$$

where f_j is the frequency of the bin B_j,i.e., $f_j = \sum_{i=1}^{n} I(X_i \in B_j)$.

Remark

For small bandwidth, (1) is bad, but (2) is good. For large bandwidth (2) is bad, but (1) is good. Thus (1) and (2) illustrates the two sides of a smoothing problem.

Some Statistical Properties of Histogram

Bias and Variance

- Without loss of generality, assume $x_0 = 0$. For $x \in B_j = [(j-1)h, jh)$

$$Bias(\widehat{f}(x)) = ((j - \frac{1}{2})h - x)f'((j - \frac{1}{2})h) + o(h)$$

$$Var(\widehat{f}(x)) = \frac{f(x)}{nh} + o(\frac{1}{nh})$$

$$\begin{aligned} MSE(\widehat{f}(x)) &= Var(\widehat{f}(x)) + Bias(\widehat{f}(x))^2 = \frac{f(x)}{nh} \\ &\quad + ((j - \frac{1}{2})h - x)^2 f'((j - \frac{1}{2})h)^2 + o(h^2) + o(\frac{1}{nh}) \end{aligned}$$

- It follows that as $h \to 0$ and $nh \to \infty$, $MSE(\widehat{f}(x)) \to 0$.
- Hence we conclude that for each x, $\widehat{f}(x) \xrightarrow{p} f(x)$.

Some Statistical Properties of Histogram

Remark

The conditions on h have an intuitive explanation. We have to decrease h ($h \to 0$) to keep bias low. We also have to ensure that enough observation fall into the bin of x ($nh \to \infty$), which keeps variance low. The big question is how to use these features to choose the bandwidth h?

- A global measure of performance: MISE

$$MISE = E \int (\widehat{f}(x) - f(x))^2 dx = \int E(\widehat{f}(x) - f(x))^2 dx = \int MSE(x) dx$$

- For Histogram

$$MISE(\widehat{f}(x)) \approx \frac{1}{nh} + \frac{h^2}{12} \int (f'(x))^2 dx$$

- We can use this expression for choosing optimal bandwidth h by minimizing MISE.

Some Statistical Properties of Histogram

- Note that

$$\frac{d}{dh} MISE(h) = 0 \Longrightarrow h_{opt} = \left[\frac{6}{n \int (f')^2} \right]^{1/3} \sim n^{-1/3}$$

- This is a theoretically elegant choice of h, but not very useful in practice since $\int (f')^2$ is unknown.
- For $h \sim n^{-1/3}$, $MISE \sim n^{-2/3}$, optimal rate of convergence achieved by taking $h = h_{opt}$.
- Consider a parametric estimation problem. Suppose we fit a $N(\theta, 1)$ density to the data. MSE$=E(\overline{X} - \theta)^2 = \frac{1}{n}$. The price of nonparametric flexibility shows up as $n^{-2/3} > n^{-1}$.

Problems with Histogram

- Sharp Corners: Very biased away from the bin centers.
  Possible solution: Use histogram only at the center of the bin. "Connect dots" by piecewise linear or by a spline functions. (Scott (1985) JASA).
- Bin edge effect: What origin should we choose? It makes a difference. In general histogram is very sensitive to the grid location.
  Possible solution: Average Shifted Histogram (Scott (1985) Annals of Statistics).
- Bin-width (bandwidth) selection.
  Possible solution: data-dependent optimal choice of bandwidth.

Example: Old Faithful Geyser Data (Duration)

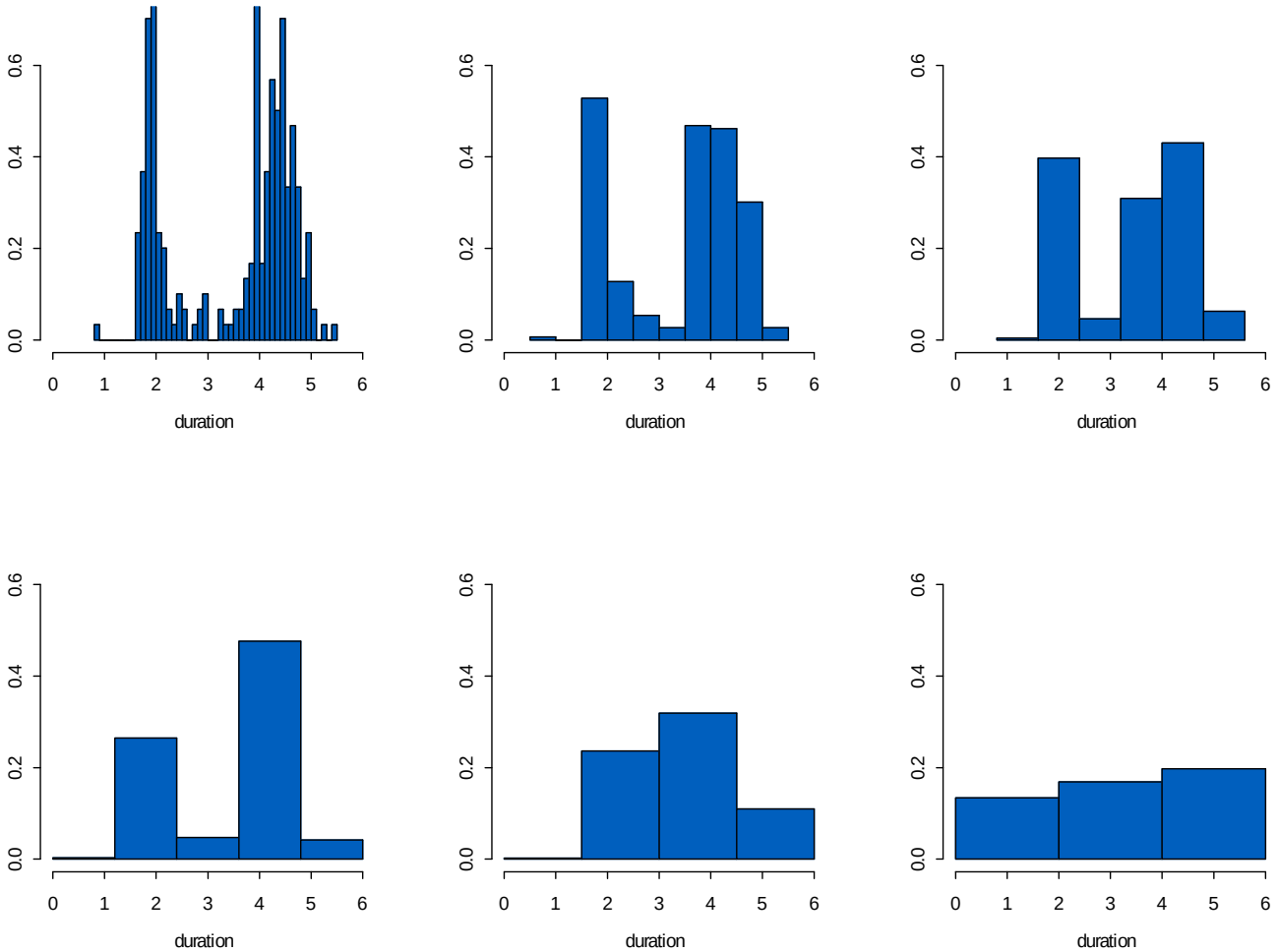

Fig1: Histograms with different bin widths

Example: Old Faithful Geyser Data (Duration)

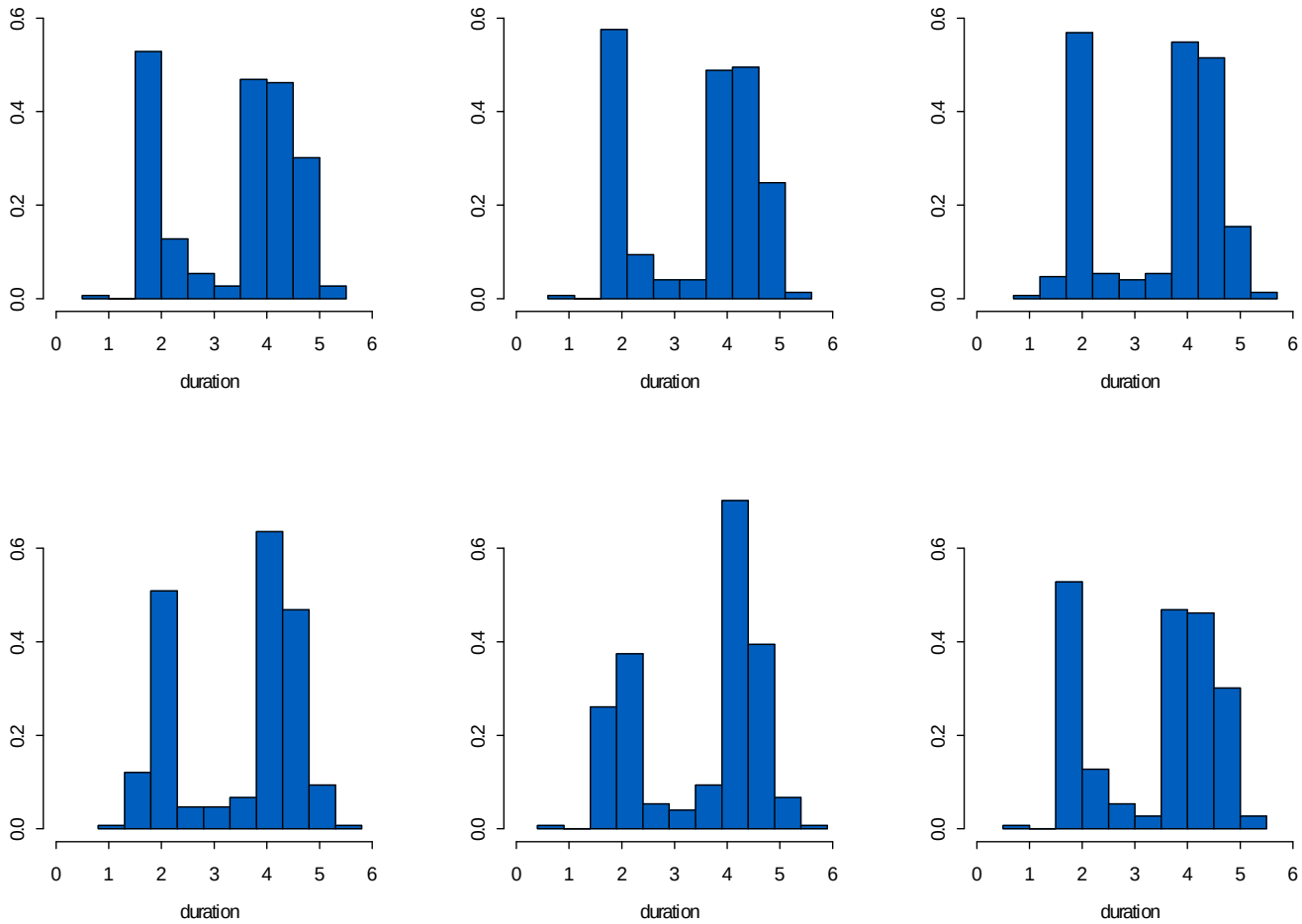

Fig 2: Histograms with differents shifts (0-0.5)

Average Shifted Histogram (ASH)

- First note that we can write an ordinary histogram in the following way:

$$\widehat{f}(x) = \frac{f_j}{nh} = \frac{1}{nh} \sum_{i=1}^{n} \sum_{j \in Z} 1_{[(j-1)h, jh)}(x) 1_{[(j-1)h, jh)}(X_i)$$

- This ordinary bin boundaries are: $.....(j-1)h, jh, (j+1)h....$
- Shifted bins indexed by $l = 0, 1, ..., (k-1)$, bin boundaries are $.....(j-1+\frac{l}{k})h, (j+\frac{l}{k})h, (j+1+\frac{l}{k})h....$
- Thus

$$\widehat{f_l}(x) = \frac{1}{nh} \sum_{i=1}^{n} \sum_{j \in Z} 1_{[(j-1+\frac{l}{k})h, (j+\frac{l}{k})h)}(x) 1_{[(j-1+\frac{l}{k})h, (j+\frac{l}{k})h)}(X_i)$$

and ASH is given by

$$\widehat{f}(x) = \frac{1}{k} \sum_{l=0}^{k-1} \widehat{f_l}(x)$$

Average Shifted Histogram (ASH)

ASH solves the problem of origin in histogram, but still have rough edges. Also, it turns out that the ASH is approximately the same as a kernel estimator (to be discussed next). To see this, note that we can write the ASH estimator as:

$$\begin{aligned}\widehat{f}(x) &= \frac{1}{k}\sum_{l=0}^{k-1}\widehat{f}_l(x) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{kh}\sum_{l=0}^{k-1}\sum_{j\in Z}1_{[(j-1+\frac{l}{k})h,(j+\frac{l}{k})h)}(x)1_{[(j-1+\frac{l}{k})h,(j+\frac{l}{k})h)}(X_i)\right]\end{aligned}$$

Note that given x and X_i, the sum $\sum_l \sum_{j\in Z} 1 * 1 = \#\{(l,j) : x \text{ and } X_i \in [(j-1+\frac{l}{k})h, (j+\frac{l}{k})h)\}$.
If $|x - X_i| > h$, then $\#$ is 0
If $|x - X_i| \leq h$, then $\#$ is ≥ 1.
If $|x - X_i| \approx h$, then $\#$ is ≈ 0
If $|x - X_i| \approx 0$, then $\#$ is $\approx k$

Average Shifted Histogram (ASH)

Suppose the count behaves approximately linearly in between, then we can say that

$$\begin{aligned}\sum_l \sum_j 1 \times 1 &= \begin{cases} \frac{k}{h}(h - |x - X_i|), & |x - X_i| \leq h \\ 0, & |x - X_i| > h \end{cases} \\ &= \frac{k}{h}(h - |x - X_i|) \times 1_{[-h,h)}(x - X_i)\end{aligned}$$

Thus

$$\begin{aligned}\widehat{f}(x) &= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h^2}(h - |x - X_i|) \times 1_{[-h,h)}(x - X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h}(1 - \frac{|x - X_i|}{h}) \times 1_{[-h,h)}(x - X_i)\end{aligned}$$

which is a kernel estimator with $K(x) = \frac{1}{h}(1 - \frac{|x|}{h}) \times 1_{[-h,h)}(x)$, called a triangular kernel.

Computation Note

R has a built in function *hist* to construct a histogram. However, the function *truehist* in the MASS library is a more flexible alternative. For example, the figure 2 was constructed using the following code:

```
library(MASS)
attach(geyser)
par(mfrow=c(2,3))
truehist(duration, h=0.5, x0=0.0, xlim=c(0, 6), ymax=0.7)
truehist(duration, h=0.5, x0=0.1, xlim=c(0, 6), ymax=0.7)
truehist(duration, h=0.5, x0=0.2, xlim=c(0, 6), ymax=0.7)
truehist(duration, h=0.5, x0=0.3, xlim=c(0, 6), ymax=0.7)
truehist(duration, h=0.5, x0=0.4, xlim=c(0, 6), ymax=0.7)
truehist(duration, h=0.5, x0=0.5, xlim=c(0, 6), ymax=0.7)
```