Lecture 14: Smoothing

Pratheepa Jeganathan

05/03/2019

Recall

- One sample sign test, Wilcoxon signed rank test, large-sample approximation, median, Hodges-Lehman estimator, distribution-free confidence interval.
- Jackknife for bias and standard error of an estimator.
- Bootstrap samples, bootstrap replicates.
- Bootstrap standard error of an estimator.
- Bootstrap percentile confidence interval.
- Hypothesis testing with the bootstrap (one-sample problem.)
- Assessing the error in bootstrap estimates.
- Example: inference on ratio of heart attack rates in the aspirin-intake group to the placebo group.
- The exhaustive bootstrap distribution.

- Discrete data problems (one-sample, two-sample proportion tests, test of homogeneity, test of independence).
- Two-sample problems (location problem - equal variance, unequal variance, exact test or Monte Carlo, large-sample approximation, H-L estimator, dispersion problem, general distribution).
- Permutation tests (permutation test for continuous data, different test statistic, accuracy of permutation tests).
- Permutation tests (discrete data problems, exchangeability.)
- Rank-based correlation analysis (Kendall and Spearman correlation coefficients.)
- Rank-based regression (straight line, multiple linear regression, statistical inference about the unknown parameters, nonparametric procedures - does not depend on the distribution of error term.)

Smoothing

Introduction

- Smoothing or estimating curves.
  - Density estimation.
  - Nonparametric regression.

Density estimation

Introduction

- A curve of interest can be a probability density function f.
- $X_1, X_2, \cdots, X_n$ are a random sample from a continuous population with cumulative distribution function F and density function f.
- Goal is to estimate f.

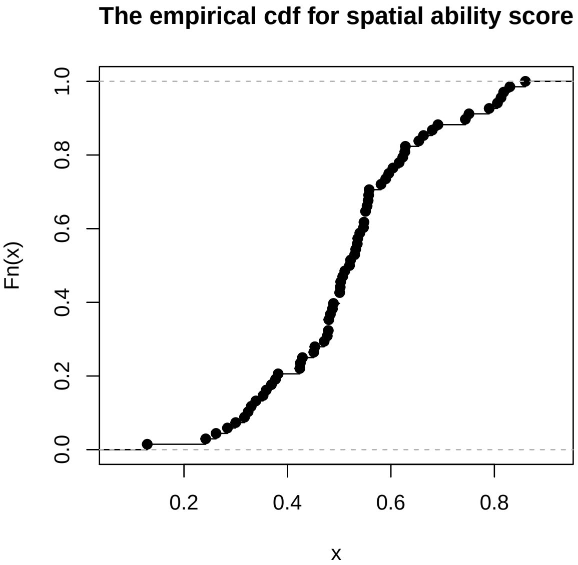
Empirical cumulative distribution function

- A study examining the relation between student mathematical performance and their preference for solving problem using average of four test scores.

```
library(NSM3)
data(discrepancy.scores)
discrepancy.scores = discrepancy.scores
```

Empirical cumulative distribution function

```r
plot(ecdf(discrepancy.scores),
  main = "The empirical cdf for spatial ability score")
```



The empirical cdf for spatial ability score
Fn(x)
x
0.0
0.2
0.4
0.6
0.8
1.0
0.2
0.4
0.6
0.8

Histogram (density estimation)

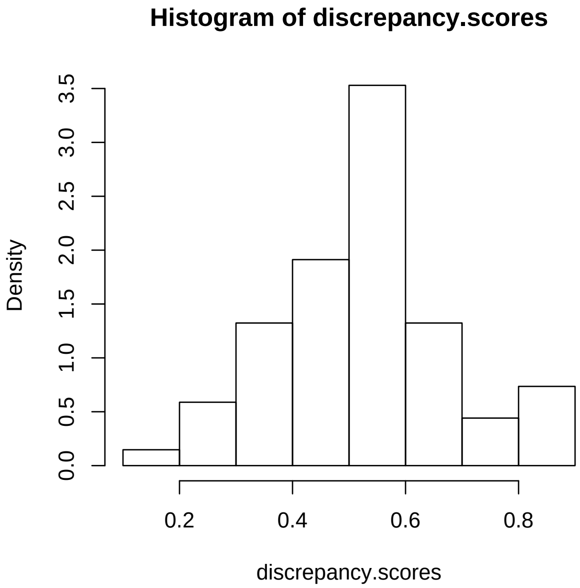
- Let $c_j, j = 1, \cdots, m$ centering points, $I_j = (c_j - h/2, c_j + h/2]$ overlapping intervals, where h is width of the interval.

$$\hat{f}(x) = \frac{\#\ \text{ of }\ X_i\ \text{ in }\ I_j}{nh}.$$

- Bin-width $h = 2 \cdot \mathrm{IQR} \cdot n^{-1/3}$ Freedman and Diaconis (1981)

Histogram

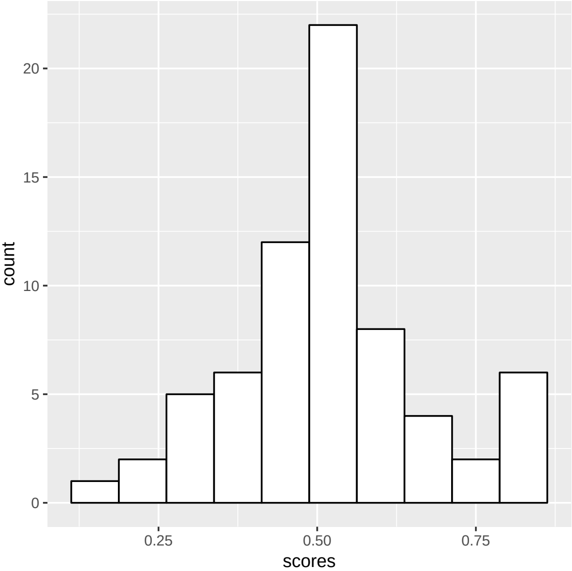
```
hist(discrepancy.scores,
  freq = FALSE, breaks = "FD")
```



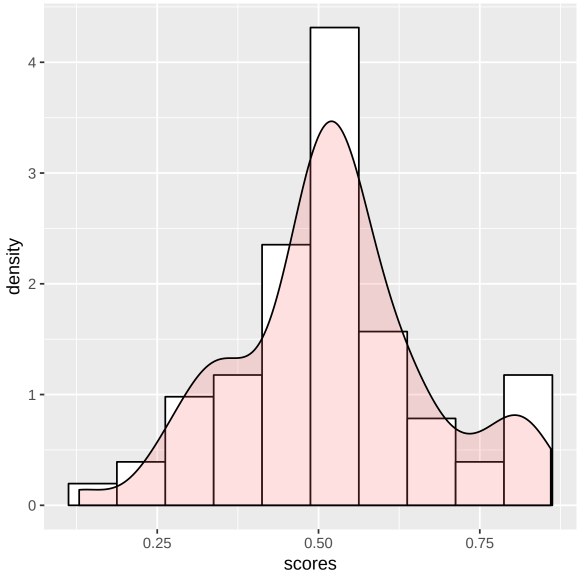
Histogram of discrepancy.scores
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0.2
0.4
0.6
0.8
discrepancy.scores

```
library(ggplot2)
cal.binwidth = function(x){
  2*(as.numeric(quantile(x, probs = .75)) - as.numeric(quan
}
binwidth = round(cal.binwidth(discrepancy.scores), digits =

ggplot(data = data.frame(scores = discrepancy.scores),
  aes(x = scores)) +
  geom_histogram(binwidth = binwidth, stat = "bin",
    fill = "white", color = "black")
```


count
20
15
10
5
0
0.25
0.50
0.75
scores

```r
ggplot(data = data.frame(scores = discrepancy.scores),
  aes(x = scores)) +
  geom_histogram(aes(y = ..density..),
    binwidth = binwidth, stat = "bin",
    fill = "white", color = "black")  +
  geom_density(alpha=.2, fill="#FF6666")
```


density
4
3
2
1
0
0.25
0.50
0.75
scores

Kernel density estimation

- Kernel K is a function such that
  - $K(x) \geq 0, -\infty < x < \infty$.
  - $K(x) = K(-x)$.
  - $\int_{-\infty}^{\infty} K(x)\, dx = 1$.
- $$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right),$$

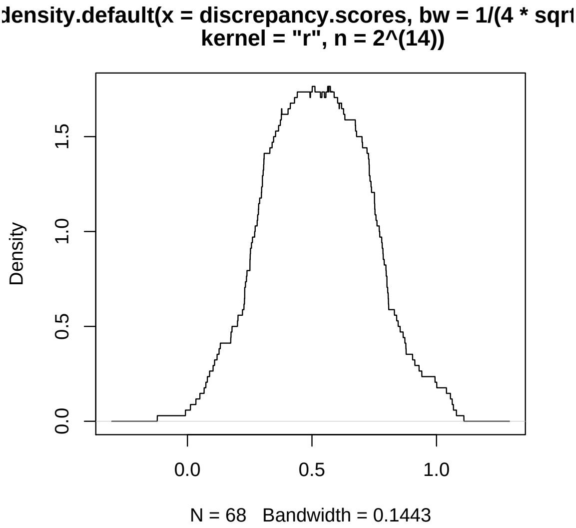
where h is bandwidth.

- kernel = “r” specifies the rectangular kernel.

```
density(discrepancy.scores,
  kernel="r", bw=1/(4 * sqrt(3)), n=2^(14))
```

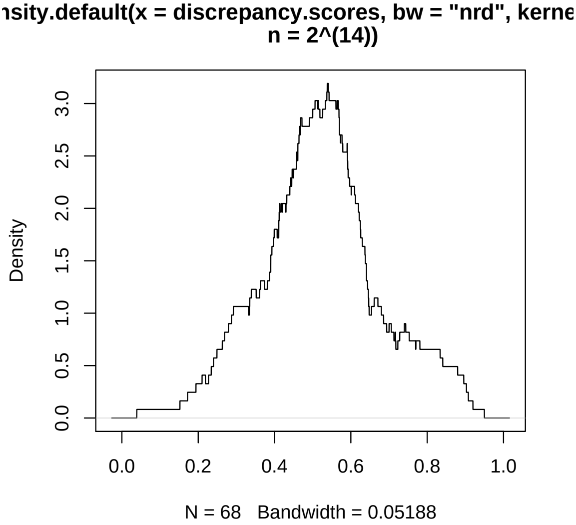
```
##
## Call:
##  density.default(x = discrepancy.scores, bw = 1/(4 * sq
##
## Data: discrepancy.scores (68 obs.);  Bandwidth 'bw' = 0
##
##        x                    y
##  Min.   :-0.30401   Min.   :0.00000
##  1st Qu.: 0.09524   1st Qu.:0.02941
##  Median : 0.49450   Median :0.32353
##  Mean   : 0.49450   Mean   :0.62616
##  3rd Qu.: 0.89376   3rd Qu.:1.41176
##  Max.   : 1.29301   Max.   :1.76471
```

```r
plot(density(discrepancy.scores,
  kernel="r", bw=1/(4 * sqrt(3)), n=2^(14)))
```



Change the bandwitdh of the Kernel

```
plot(density(discrepancy.scores,
  kernel="r", bw="nrd", n=2^(14)))
```



Nonparametric regression

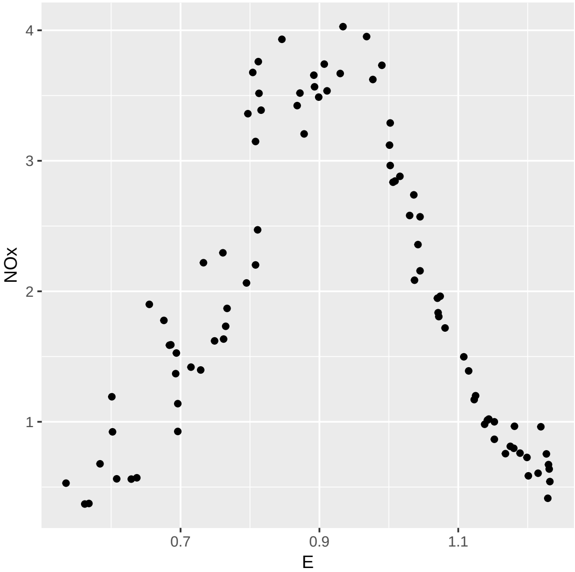
Introduction

- Follow notation in **W(2006)** Chapter 4 and 5.
- There are n pairs of observations $(x_1, Y_1), (x_2, Y_2), \cdots, (x_n, Y_n)$.
- Regression is $Y_i = r(x_i) + \epsilon_i$, where $\mathbb{E}(\epsilon_i) = 0$.
- A curve of interest is the regression function r.

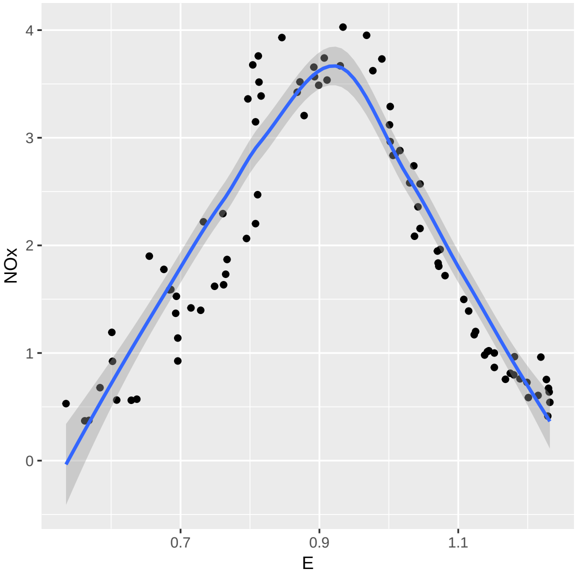
Nonparametric regression

- Example 14.2 (Page 662) Nitrogen Oxide Concentrations
  - Brinkman (1981) collected data on the nitrogen oxide concentrations ($\boldsymbol{Y}$) found in engine exhaust for ethanol engines with various equivalence ratios ($\boldsymbol{x}$).

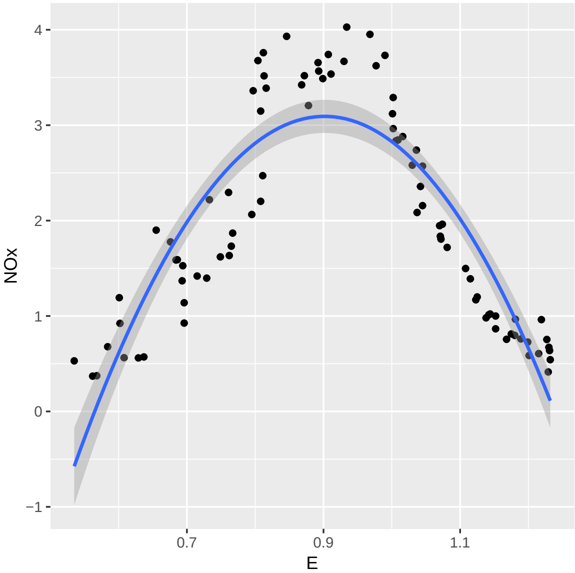
Nonparametric regression



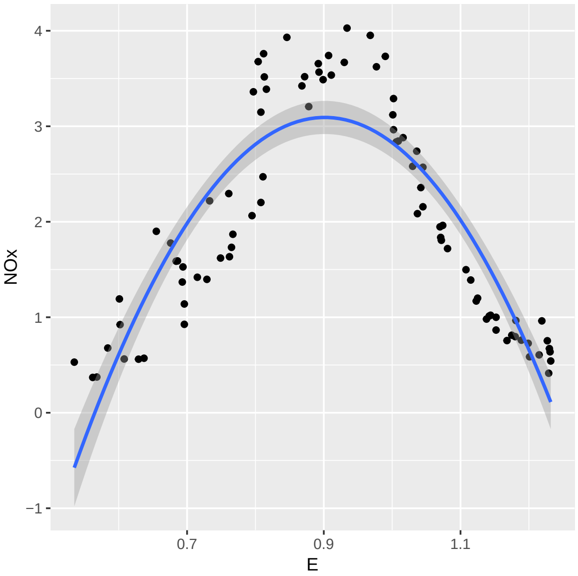
- locally weighted regression



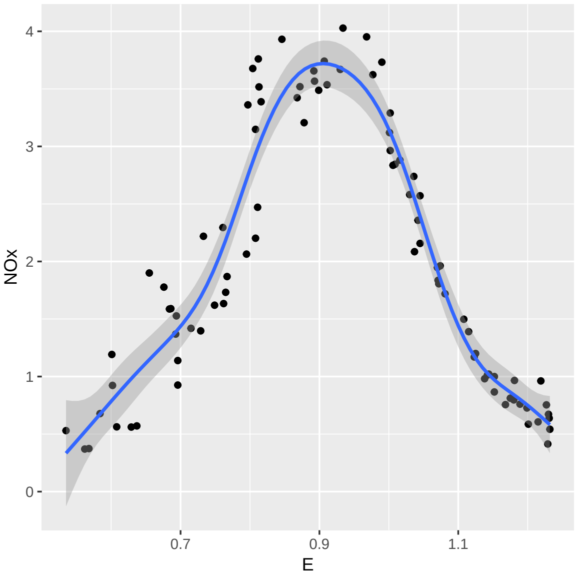
- Include squared term of E in the regression model



- Regression function with a second order (quadratic) polynomial



- generalized additive model (GAM) from the `mgcv` package



- **HWC** Chapter 9.7
  - Example 1: Running Line Smoother
  - Example 2: Kernel Regression Smoother.
  - Example 3: Local Regression Smoother.
  - Example 4: Spline Regression Smoother.
  - Example 5: Wavelet Smoother
- Determine which approach to choose (HWC Chapter 9.7).
  - Consider relative importance of minimizing bias versus minimizing variance (and computing cost).
  - Bias-variance trade-off.

Bias-variance trade-off

- Let $\hat{f}_n(x)$ be an estimate of a function $f(x)$.
- Define the **squared error** or (L_2) loss function is

$$L\left(f(x), \hat{f}_n(x)\right) = \left(f(x) - \hat{f}_n(x)\right)^2.$$

- Define average of this loss as **risk** or **mean squared error** (MSE)

$$\text{MSE} = R\left(f(x), \hat{f}_n(x)\right) = \mathbb{E}\left(L\left(f(x), \hat{f}_n(x)\right)\right).$$

- The random variable in the MSE is $\hat{f}_n(x)$ which implicitly depends on the observed data.
- The MSE can be decomposed into a bias and variance term:

$$\text{Risk} = \text{MSE} = \text{Bias}^2 + \text{Variance}.$$

Decomposition of MSE

- Bias $= f(x) - \mathbb{E}\left(\hat{f}_n(x)\right)$.
- *Variance* $= \mathbb{E}\left(\hat{f}(x) - \mathbb{E}\left(\hat{f}_n(x)\right)\right)^2$.

$$
\begin{aligned}
\mathbb{E}\left(f(x) - \hat{f}_n(x)\right)^2 =& \mathbb{E}\left(f(x) - \mathbb{E}\,\hat{f}_n(x) + \mathbb{E}\,\hat{f}_n(x) - \hat{f}_n(x)\right)^2 \\
=& \mathbb{E}\left(f - \mathbb{E}\,\hat{f}_n(x)\right)^2 + \mathbb{E}\left(\mathbb{E}\,\hat{f}_n(x) - \hat{f}_n(x)\right)^2 + \\
& 2\,\mathbb{E}\left(f - \mathbb{E}\,\hat{f}_n(x)\right)\left(\mathbb{E}\,\hat{f}_n(x) - \hat{f}_n(x)\right) \\
=& \left(\mathbb{E}\,\hat{f}_n(x) - f\right)^2 + \mathbb{E}\left(\hat{f}_n(x) - \mathbb{E}\,\hat{f}_n(x)\right)^2.
\end{aligned}
$$

Bias-variance trade-off

- Above definitions refer to the risk at point x.
- In density estimation problem, the **integrated risk** or **integrated mean squared error** is

$$R\left(f, \hat{f}_n\right) = \int R\left(f(x), \hat{f}_n(x)\right) dx.$$

- For regression problems, the **integrated MSE** or **average MSE** is

$$R(r, \hat{r}_n) = \frac{1}{n} \sum_{i=1}^{n} R(r(x_i), \hat{r}_n(x_i)).$$

Bias-variance trade-off

- Predictive risk
  - Nonparametric regression model is $Y_i = r(x_i) + \epsilon_i$.
  - Suppose we draw a new observation $Y_i^* = r(x_i) + \epsilon_i^*$ at each x_i.
  - Predict Y_i^* with $\hat{r}(x_i)$.
  - **Predictive risk**

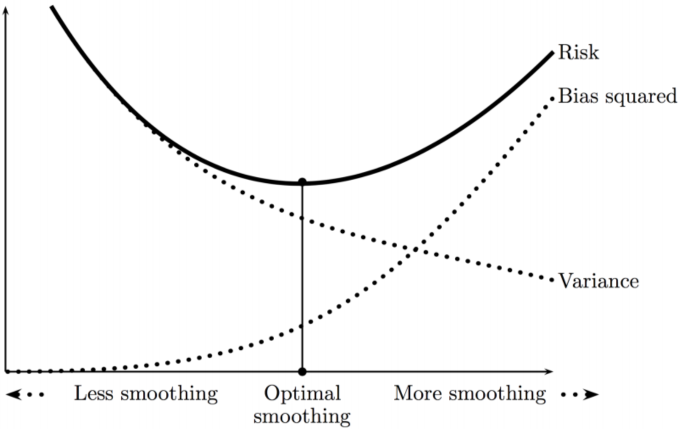
$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\left(Y_i^* - \hat{r}(x_i)\right)^2\right).$$

- Average risk and predictive risk
  - predictive risk $= R(r, \hat{r}_n) + \sigma^2$, where σ^2 is variance of ϵ_i.

Bias-variance trade-off

- Minimizing risk corresponds to balancing between bias and variance.
- Smoothing is chosen based on the bias-variance trade-off.
- Oversmooth data have a large bias term and small variance.
- Under smooth data have a small bias term and large variance.

The Bias–Variance Tradeoff



Source: Wassermann (2006)

Example(Bias-variance trade-off)

- Let $X \sim f(x)$, where f be a pdf.
- Consider estimating $f(0)$.
- Let h be a small and positive number.
- Define

$$p_h := P\left(-\frac{h}{2} < X < \frac{h}{2}\right) = \int_{-h/2}^{h/2} f(x)\,dx \approx hf(0).$$

- Hence

$$f(0) \approx \frac{p_h}{h}.$$

- Let Y be the number of observations in the interval $\left(-\frac{h}{2}, -\frac{h}{2}\right)$.
  - $Y \sim \text{Binomial}\left(n, p_h\right)$.
- An estimate of p_h is $\hat{p}_h = \frac{Y}{n}$. Thus, an estimate of $f\left(0\right)$ is

$$\hat{f}_n\left(0\right) = \frac{\hat{p}_h}{n} = \frac{Y}{nh}.$$

- How do we choose h?
- We will show that for some constant A and B,

$$\text{MSE}\left(\hat{f}_n\left(0\right)\right) \approx Ah^4 + \frac{B}{nh} = \text{Bias}^2 + \text{Variance}$$

.

  - Then, we can minimize $\text{MSE}\left(\hat{f}_n\left(0\right)\right)$ to find the optimal h.

- Show that Bias $= \mathbb{E}\left(\hat{f}_n(0)\right) - f(0) \approx \dfrac{f^{''}(0)\,h^2}{24}$.
- $$\mathbb{E}\left(\hat{f}_n(0)\right) = \frac{\mathbb{E}\,Y}{nh} = \frac{p_h}{h}. \tag{1}$$
- Taylor expansion of f at 0,
$$f(x) \approx f(0) + xf^{'}(0) + \frac{x^2}{2} f^{''}(0).$$
- Plug-in

$$\begin{aligned} p_h &= \int_{-h/2}^{h/2} f(x)\,dx \\ &\approx \int_{-h/2}^{h/2} \left(f(0) + xf^{'}(0) + \frac{x^2}{2} f^{''}(0) \right) dx \\ &= hf(0) + \frac{f^{''}(0)}{24} h^3. \end{aligned} \tag{2}$$

- Now plug-in (2) to (1)
  - $\mathbb{E}\left(\hat{f}_n(0)\right) \approx f(0) + \frac{f^{''}(0)}{24} h^2.$
- Thus, Bias $= \mathbb{E}\left(\hat{f}_n(0)\right) - f(0) \approx \frac{f^{''}(0) h^2}{24}.$
- $\mathbb{V}\left(\hat{f}_n(0)\right) = \frac{\mathbb{V}(Y)}{n^2 h^2} = \frac{p_h(1-p_h)}{nh^2}.$
  - $1 - p_h \approx 1$ for small h.
  - Thus, $\mathbb{V}\left(\hat{f}_n(0)\right) \approx \frac{p_h}{nh^2}.$
  - By plug-in (2), we can show that $\mathbb{V}\left(\hat{f}_n(0)\right) \approx \frac{f(0)}{nh}.$

- Now

$$\mathrm{MSE}\left(\hat{f}_n(0)\right) = \mathrm{Bias}^2 + \mathrm{Variance} = \left(\frac{f''(0)\,h^2}{24}\right)^2 + \frac{f(0)}{nh} = Ah^4 + \frac{B}{nh}.$$

  - If we smooth less (decrease h), bias term decreases and the variance term increases.
  - If we oversmooth (increase h), bias term increases and the variance term decreases.
- We should balance between bias and variance to find the optimal h.

Choosing other loss functions

- L_p loss function

$$\left\{\int \left| f(x) - \hat{f}_n(x) \right| \right\}^{1/p}.$$

- In parametric context (and in machine learning community) - Kullback-Leibler loss

$$L\left(f, \hat{f}_n\right) = \int f(x) \left(log \frac{f(x)}{\hat{f}_n(x)} \right) dx.$$

  - This loss function is not appropriate for smoothing problems due to sensitivity in the tails of the distribution (Hall 1987).

The curse of dimensionality

- Estimation in smoothing getting harder with dimensionality - curse of dimensionality or computationally expensive.
- Curse
  - Computational curse: computational cost increase exponentially with dimension d.
  - Statistical curse of dimensionality: sample size n needs to increase exponentially with dimension d.
- The MSE of any nonparametric estimator of a smooth curve (twice differentiable) has the form

$$\text{MSE} \approx \frac{c}{n^{4/(4+d)}}.$$

- If we fixed $\text{MSE} = \delta$ to a small number, then,

$$n \propto \left(\frac{c}{\delta}\right)^{d/4}$$

which grows exponentially with d.

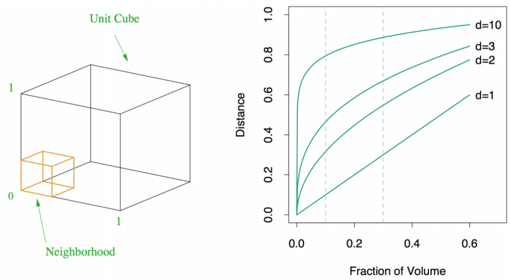
Why this phenomenon in smoothing?

- Smoothing involves estimating $f(x)$ using data points in a local neighborhood of x.
  - When d is large (data are sparse), local neighborhood contains very few points.

(Example) The curse of dimensionality

- Suppose n data points uniformly distributed on the interval $[0, 1]$.
  - Number of points in the interval $[-.1, .1] \approx \frac{n}{10}$ points.
- Suppose n data points on the 10-dimensional unit cube $[0, 1]^{10} = [0, 1] \times \cdots \times [0, 1]$.
  - Number of data points in the cube $[-.1, .1]^{10} \approx n \times \left(\frac{.2}{2}\right) = \frac{n}{10,000,000,000}$.
  - n should be large enough to ensure that small neighborhood have any data.
- Smoothing methods can be used in high-dimensional problems. Due to statistical curse of dimensionality
  - Estimator may not be accurate.
  - Confidence interval around the estimate may be large.
  - doesn't mean the smoothing method is wrong.

(Example) The curse of dimensionality



Source: Hastie, Tibshirani, and Friedman (2009)

- When $d = 10$, 10% of data are in the 80% range.
- When $d \leq 3$, 10% of data are in the less than 40% range.

How to deal with the curse of dimensionality

- Dimension reduction: find a low-dimension approximation to the data (principal component analysis, independent component analysis projection pursuit.)
- Variable selection: covariates that do not predict Y are removed from the regression.

References for this lecture

HWC Chapter 9.7 (an introduction to nonparametric regression.)

HWC Chapter 12 (density estimation)

W Chapter 4 (smoothing: general concepts)

Seiler2016: Lecture notes.