Lecture 16: Nonparametric regression II

Pratheepa Jeganathan

05/08/2019

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 (density estimation, bias-variance trade-off, curse of dimensionality)
- Nonparamteric regression (Local averaging, local regression, kernel smoothing, local polynomial, penalized regression)

Nonparametric regression II

Introduction

- Cross-Validation
- Variance Estimation
- Confidence Bands
- Bootstrap Confidence Bands

Choosing smoothing parameter

Choosing smoothing parameter

- Risk depends on unknown function $r(x)$.

$$R(h) = \mathbb{E}\left(\frac{1}{n}\left(\hat{r}_n(x_i) - r(x_i)\right)^2\right).$$

1) Training error
   - $\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{r}_n(x_i)\right)^2$.
   - Using data twice.
     - to estimate r.
     - to estimate the risk R.
   - Function estimate is chosen to make $\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{r}_n(x_i)\right)^2$ small so risk is underestimated.

Choosing smoothing parameter

2) Leave-one-out cross-validation score

$$\mathrm{CV} = \hat{R}(h) = \frac{1}{n}\sum_{i=1}^{n}\Big(Y_i - \hat{r}_{(-i)}(x_i)\Big)^2$$

- $\hat{r}_{(-i)}$ is the estimator obtained by omitting i-th pair (x_i, Y_i).
- $\hat{r}_{(-i)}(x) = \sum_{j=1}^{n} Y_j l_{j,(-i)}(x)$, where

$$l_{j,(-i)}(x) = \begin{cases} 0 & \text{if } \; j = i \\ \dfrac{l_j(x)}{\sum_{k\neq i} l_k(x)} & \text{if } \; j \neq i. \end{cases} \qquad (1)$$

- Set weight on x_i to 0 and renormalize the other weights to sum to one.
- Do this for different h.

Choosing smoothing parameter

2) Leave-one-out cross-validation

- Intuition: $\mathbb{E}\left(Y_i - \hat{r}_{(-i)}(x_i)\right)^2 \approx \sigma^2 + \mathbb{E}\left(r(x_i) - \hat{r}_n(x_i)\right)^2 =$ predictive error. $\hat{R}$ score is nearly unbiased estimate of the risk.
- Shortcut formula to compute $\hat{R}$

$$\hat{R}(h) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \hat{r}_n(x_i)}{1 - L_{ii}}\right)^2,$$

where $L_{ii} = l_i(x_i)$ is the i-th diagonal element if the smoothing matrix L.

Choosing smoothing parameter

3) Generalized cross-validation

$$\mathrm{GCV}\left(h\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \hat{r}_n\left(x_i\right)}{1-\nu/n}\right)^2,$$

where $\nu = \mathrm{tr}\left(L\right)$ is the effective degrees of freedom.

- a formula similar to Colin Mallows C_p statistic.

Variance estimation

Variance estimation

- We assume $\mathbb{V}\left(\epsilon_i\right) = \sigma^2$.
  - constant variance

1) For linear smoother $\boldsymbol{r} = \mathbf{L}\boldsymbol{Y}$, an unbiased estimate of σ^2 is

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{r}\left(x_i\right)\right)^2}{n - 2\nu + \tilde{\nu}},$$

where $\nu = \operatorname{tr}\left(L\right)$ and $\tilde{\nu} = \operatorname{tr}\left(L^T L\right) = \sum_{i=1}^{n} \left|\left|l\left(x_i\right)\right|\right|^2.$

- If r is sufficiently smooth, then $\hat{\sigma}^2$ is a consistent estimator of σ^2.

Variance estimation

2) Alternative formula (Rice 1984).

- Suppose x_is are ordered.

$$\hat{\sigma}^2 = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} (Y_{i+1} - Y_i)^2 .$$

- Intuition: an average of the residuals that results from fitting a line to the first and third point of each consecutive triple of design points.

Variance estimation (Spatially inhomogeneous functions)

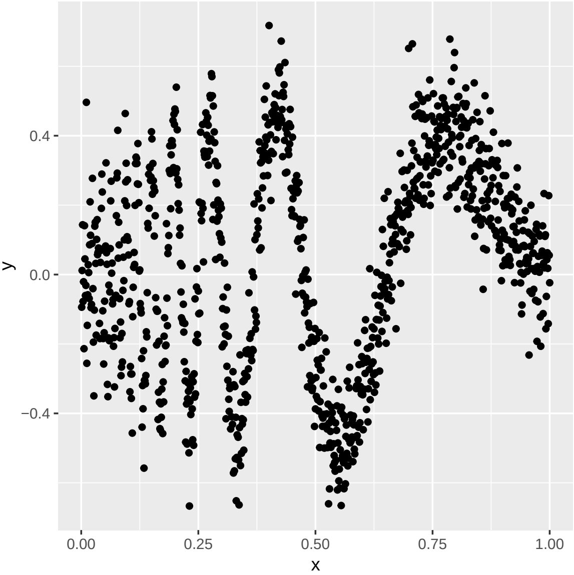
- Inhomegenity of variance.
- $\hat{r}_n(x)$ is relatively insensitive to heteroscedastic.
- We need to account for the unconstant variance when making confidence bands.

Example

- Doppler function

```
library(ggplot2)
r = function(x){
  sqrt(x*(1-x))*sin(2.1*pi/(x+.05))
}
ep = rnorm(1000)
y = r(seq(1, 1000, by = 1)/1000) + .1 * ep
df = data.frame(x = seq(1, 1000, by = 1)/1000, y = y)
ggplot(df) +
  geom_point(aes(x = x, y = y))
```

Example



Example

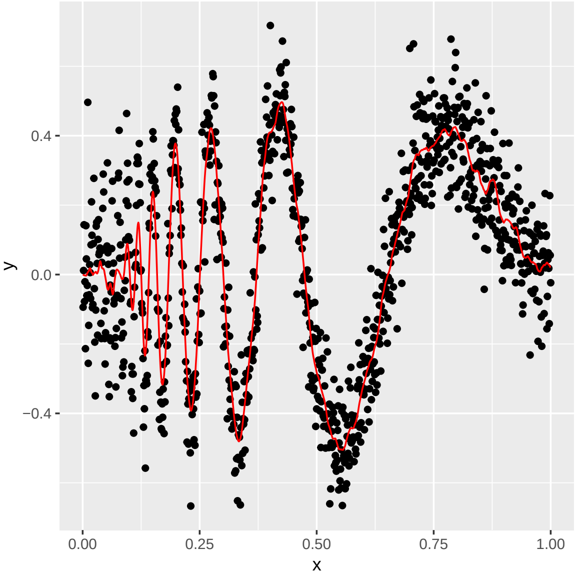
- Doppler function is spatially inhomogeneous (smoothness varies over x).
- Estimate by local linear regression

```
library(np)
doppler.npreg <- npreg(bws=.005,
  txdat=df$x,
      tydat=df$y,
  ckertype="epanechnikov")

doppler.npreg.fit = data.frame(x = df$x,
  y = df$y,
  kernel.fit = fitted(doppler.npreg))

p = ggplot(doppler.npreg.fit) +
  geom_point(aes(x = x, y = y)) +
  geom_line(aes(x = x, y= kernel.fit), color = "red")
```

Example



Example

- Doppler function fit using local linear regression.
  - Effective degrees of freedom 166.
  - Fitted function is very wiggly.
  - If we smooth more, right-hand side of the fit would look better at the cost of missing structure near $x = 0$.
- Wavelets

Variance estimation

- Estimate $r(x)$ with any nonparamteric method to get $\hat{r}_n(x)$.
- Compute the squared residuals $Z_i = (Y_i - \hat{r}_n(x_i))^2$.
- Regress Z_i on x_i to get an estimate $\hat{q}(x)$.
- $\hat{\sigma}(x) = \hat{q}(x)$.

Confidence Bands

Confidence Bands

- Can we get confidence bands for $r(x)$?
- Let mean and standard deviation of $\hat{r}_n(x)$ is $\bar{r}_n(x)$ and $\hat{s}_n(x)$, respectively.
- Bias Problem:

$$
\begin{aligned}
\frac{\hat{r}_n(x) - r(x)}{\hat{s}_n(x)} &= \frac{\hat{r}_n(x) - \bar{r}_n(x)}{\hat{s}_n(x)} + \frac{\bar{r}_n(x) - r(x)}{\hat{s}_n(x)} \\
&= Z_n(x) + \frac{\operatorname{bias}(\hat{r}_n(x))}{\sqrt{\operatorname{variance}(\hat{r}_n(x))}}.
\end{aligned} \tag{2}
$$

- Typically $Z_n(x) = \dfrac{\hat{r}_n(x) - \bar{r}_n(x)}{\hat{s}_n(x)}$ follows a standard normal and used to derive confidence bands
- In nonparametric regression, the second term in (2) does not vanish.
  - Optimal smoothing balance between bias and the standard deviation.

Confidence Bands

- Confidence bands for $\bar{r}_n(x)$ is

$$\hat{r}_n(x) \pm c \times \text{se}(x),$$

  where $c > 0$ some constant.
- $\bar{r}_n(x) = \mathbb{E}(\hat{r}_n(x))$.
  - We don't get a confidence band for $r(x)$.
- c is computed from the distribution of the maximum of a Gaussian process. Choose c by solving

$$2(1 - \Phi(c)) + \frac{\kappa_0}{\pi} e^{-c^2/2} = \alpha,$$

  where $\kappa_0 = \int_a^b \left|\left|T'(x)\right|\right|$ and $T_i(x) = \frac{l_i(x)}{||l_i(x)||}$.

Confidence Bands

- To get simultaneous confidence band, compute c such that

$$2\left(1-\phi\left(c\right)\right)+\frac{\kappa_0}{\pi}e^{c^2/2}=\alpha.$$

- The variance of $\hat{r}_n\left(x\right)$ is

$$\mathbb{V}\left(\hat{r}_n\left(x\right)\right)=\sum_{i=1}^{n}\sigma^2\left(x_i\right)l_i^2\left(x_i\right).$$

- The approximate confidence band is

$$\mathbb{I}\left(x\right)=\hat{r}_n\left(x\right)\pm c\sqrt{\sum_{i=1}^{n}\hat{\sigma}^2\left(x_i\right)l_i^2\left(x_i\right)}.$$

Bootstrap Confidence Bands

- Reference: [link here](https://www.stat.cmu.edu/~cshalizi/402/lectures/08-bootstrap/lecture-08.pdf#page20).

1) Resample rows:
   - Resample (x, y) pair.
2) Resample residuals:
   - Hold the x fixed, but make T equal to $\hat{r}(x)$ plus a randomly re-sampled ϵ_i.
   - Errors need to be iid.

Bootstrap Confidence Bands (Example)

- Resample rows

```
library(NSM3)
library(dplyr)
data("ethanol")
ethanol.df = select(ethanol,
  c(E, NOx))

resample.data = function(df) {
sample.rows = sample(1:nrow(df),
  replace = TRUE)
return(df[sample.rows,])
}
```

Bootstrap Confidence Bands (Example)

```r
# use kernel smoothing
library(np)
npr.nox.on.E = function(df.star) {
  bw = npregbw(NOx ~ E,
    data = df.star)
  fit =  npreg(bw)
  return(fit)
}
```

Bootstrap Confidence Bands (Example)

```
# Use uniform grid points to predict the values.
evaluation.points = seq((min(ethanol.df$E) -.1),
  (max(ethanol.df$E)+.1), by =.01)

eval.npr = function(npr) {
   return(predict(npr,
     exdat = evaluation.points))
}

ethanol.npr = npr.nox.on.E(ethanol.df)
```

```
##
Multistart 1 of 1 |
Multistart 1 of 1 |
Multistart 1 of 1 |
Multistart 1 of 1 /
Multistart 1 of 1 |
Multistart 1 of 1 |
```

Bootstrap Confidence Bands (Example)

```
npr.cis = function(B,alpha, df, obs.curve) {
  tboot= replicate(B,
    eval.npr(npr.nox.on.E(resample.data(df))))
  low.quantiles = apply(tboot, 1,
    quantile,
    probs = alpha/2)
  high.quantiles = apply(tboot, 1,
    quantile,
    probs = (1-alpha/2))
  low.cis = 2*obs.curve - high.quantiles
  high.cis = 2*obs.curve - low.quantiles
  cis <- rbind(low.cis, high.cis)
  return(list(cis=cis, tboot= t(tboot)))
}
```

Bootstrap Confidence Bands (Example)

```
ethanol.npr.cis = npr.cis(B = 100,
  alpha = 0.05,
  df = ethanol.df,
  obs.curve = obs.curve)
```

```
##
Multistart 1 of 1 |
Multistart 1 of 1 |
Multistart 1 of 1 |
Multistart 1 of 1 /
Multistart 1 of 1 -
Multistart 1 of 1 |
Multistart 1 of 1 |


Multistart 1 of 1 |
Multistart 1 of 1 |
Multistart 1 of 1 |
```

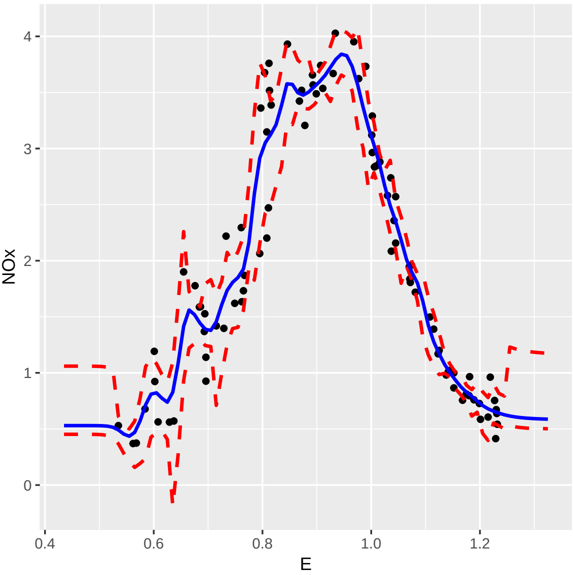
Bootstrap Confidence Bands (Example)

```r
df.plot.ci = data.frame(x = evaluation.points,
  obs.curve = obs.curve,
  low.cis = ethanol.npr.cis$cis[1,],
  upper.cis = ethanol.npr.cis$cis[2,])

p = ggplot() +
  geom_point(data = ethanol.df,
    aes(x = E, y = NOx)
    ) +
  geom_line(data = df.plot.ci,
    aes(x = evaluation.points, y = low.cis),
    color = "red", linetype = "dashed",
    size = 1)
```

Bootstrap Confidence Bands (Example)

```
p = p +
  geom_line(data = df.plot.ci,
    aes(x = evaluation.points, y = upper.cis),
    color = "red", linetype = "dashed",
    size = 1) +
  geom_line(data = df.plot.ci,
    aes(x = evaluation.points, y = obs.curve),
    color = "blue",
    size = 1)
```



4
3
2
1
0
NOx
0.4
0.6
0.8
1.0
1.2
E

- Notes
  - Confidence bands get wider where there is less data.
  - If variance is not constant, use resampling residuals with heteroskedasticity method describe in the following [link 4.4](https://www.stat.cmu.edu/~cshalizi/402/lectures/08-bootstrap/lecture-08.pdf#page20).

References for this lecture

W Chapter 5

Reference for bootstrap confidence bands: [link here](https://www.stat.cmu.edu/~cshalizi/402/lectures/08-bootstrap/lecture-08.pdf#page20).