Nonlinear Regression (Part 3)

Christof Seiler

Stanford University, Spring 2016, STATS 205
Overview

Last time:

▶ Linear Smoothers
  ▶ Local Averages
  ▶ Local Regression
  ▶ Penalized Regression

Today:

▶ Cross-Validation
▶ Variance Estimation
▶ Confidence Bands
▶ Bootstrap Confidence Bands
Nonlinear Regression

- We are given \( n \) pairs of observations \((x_1, Y_1), \ldots, (x_n, Y_n)\).
- The covariates \( x_i \) are fixed.
- The response variable is related to the covariate 
  \[
  Y_i = r(x_i) + \epsilon_i \quad \text{E}(\epsilon_i) = 0, \ i = 1, \ldots, n
  \]
  with \( r \) being the regression function.
- For now, assume that variance \( \text{Var}(\epsilon_i) = \sigma^2 \) is independent of \( x \).
Choosing the Smoothing Parameter

- The choice of kernel is not too important
- Estimates obtained by using different kernels are usually numerically very similar
- Can be confirmed by theoretical calculations showing that risk is insensitive to choice of kernel
- Choice of bandwidth matters which controls the amount of smoothing
- Small bandwidths give very rough estimates while larger bandwidths give smoother estimates
Choosing the Smoothing Parameter

- If the bandwidth is small
  - $\hat{r}_n(x_0)$ is an average of a small number of $Y_i$ close to $x_0$
  - The variance will be relatively large, close to that of an individual $Y_i$
  - The bias will tend to be small, because a close $r(x_i)$ should be similar to $r(x_0)$

- If the bandwidth is large
  - The variance of $\hat{r}_n(x_0)$ will be small relative to the variance of any $Y_i$, because of the effects of averaging
  - The bias will be higher, because we are now using observations $x_i$ further from $x_0$, and there is no guarantee that $r(x_i)$ will be close to $r(x_0)$
Choosing the Smoothing Parameter

▶ The smoothers depend on some smoothing parameter $h$
▶ We define a risk

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

▶ Ideally, we would like to choose $h$ to minimize $R(h)$
▶ But $R(h)$ depends on unknown function $r(x)$
▶ Instead we minimize an estimate $\hat{R}(h)$
▶ As first guess, we might try minimizing the training error

$$\frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{r}_n(x_i))^2$$

▶ This is a poor estimator, because it overfits (undersmoothing)
▶ We use the data twice: to estimate the function and to estimate the risk
Choosing the Smoothing Parameter

- A better idea is to use leave-one-out cross-validation

\[
\text{cv} = \hat{R}(h) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{r}_{(-i)}(x_i))^2
\]

with \(\hat{r}_{(-i)}\) estimator obtained by omitting the \(i\)th pair \((x_i, Y_i)\)

- Define

\[
\hat{r}_{(-i)} = \sum_{j=1}^{n} Y_j l_j,(-i)(x)
\]

- and we set the weight on \(x_i\) to 0 and renormalize the other weights to sum to one

\[
l_j,(-i)(x) = \begin{cases} 
0 & \text{if } j = i \\
\frac{l_j(x)}{\sum_{k \neq i} l_k(x)} & \text{if } j \neq i
\end{cases}
\]

- Cross-validation is approximately the predictive risk (predicting the left-one-out observation)
Choosing the Smoothing Parameter

- We can compute leave-one-out cross-validation without leaving one observation out

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

- This is exactly true, not an approximation!
- After some algebra, we can see that

\[ \hat{r}(x_i) = (1 - L_{ii})\hat{r}_{(-i)}(x_i) + L_{ii} Y_i \]
There are several variance estimators for linear smoothers

Let $\hat{r}_n(x)$ be a linear smoother

A consistent estimator (converges in probability to the true value of the parameter) of $\sigma^2$ is

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

with

$$\nu = \text{tr}(L), \tilde{\nu} = \text{tr}(L^T L) = \sum_{i=1}^{n} \| l(x_i) \|^2$$

and if $r$ is sufficiently smooth
Variance Estimation

- The expected value of our estimator is
  \[
  E(\hat{\sigma}^2) = \frac{E(Y^T \Lambda Y)}{\text{tr}(\Lambda)} = \sigma^2 + \frac{r^T \Lambda r}{n - 2\nu + \tilde{\nu}}
  \]
  with
  \[
  \Lambda = (I - L)^T (I - L)
  \]
  and
  \[
  E(Y^T QY) = \text{tr}(QV) + \mu^T Q\mu
  \]
  where \( V = \text{Var}(Y) \) is covariance matrix of \( Y \) and \( \mu = E(Y) \) is the mean vector

- Assuming that \( \nu \) and \( \hat{\nu} \) do not grow too quickly, and that \( r \) is smooth, the second term is small for large \( n \)

- So \( E(\hat{\sigma}^2) \approx \sigma^2 \)

- and one can show that \( \text{Var}(\hat{\sigma}^2) \to 0 \)
Variance Estimation

- Another variance estimator (order $x_i$’s)

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

- Assuming $r$ is smooth

$$Y_{i+1} - Y_i = [r(x_{i+1}) + \epsilon_{i+1}] - [r(x_i) + \epsilon_i] \approx \epsilon_{i+1} - \epsilon_i$$

- Therefore

$$E(Y_{i+1} - Y_i) \approx E(\epsilon_{i+1}) + E(\epsilon_i) = 2\sigma^2$$
Confidence Bands

- **Variability** bands

\[ \hat{r}_n(x) \pm 2\hat{\sigma}(x) \]

- There is a problem with that

\[ \frac{\hat{r}_n(x) - r(x)}{\hat{\sigma}(x)} = \frac{\hat{r}_n(x) - \bar{r}_n(x)}{\hat{\sigma}(x)} + \frac{\bar{r}_n(x) - r(x)}{\hat{\sigma}(x)} \]

with \( \bar{r}_n(x) \) being the mean

- First term converges to a normal

- If we do a good job trading off bias and variance, the second term doesn’t vanish with large \( n \)

\[ \frac{\bar{r}_n(x) - r(x)}{\hat{\sigma}(x)} = \frac{\text{Bias}(\hat{r}_n(x))}{\sqrt{\text{Variance}(\hat{r}_n(x))}} \]
Confidence Bands

- The result is that the confidence interval will not be centered around the true function $r$ due to the smoothing bias.
- Possible solutions:
  1. Accept the fact that confidence band is for $\bar{r}_n$ not $r$.
  2. Estimate bias (this is difficult because it involves estimating $r''(x)$).
  3. Undersmooth: less smoothing will bias results less, and asymptotically the bias will decrease faster than the variance.

- We will go with the first approach.
Constructing Confidence Bands

- For linear smoother $\hat{r}_n(x)$ with

  $$\bar{r}(x) = \mathbb{E}(\hat{r}_n(x)) = \sum_{i=1}^{n} l_i(x) r(x_i)$$

  and assuming constant variance

  $$\text{Var}(\hat{r}_n(x)) = \sigma^2 \|l(x)\|^2$$

- Consider confidence bands

  $$I(x) = (\hat{r}_n(x) - c\hat{\sigma} \|l(x)\|, \hat{r}_n(x) + c\hat{\sigma} \|l(x)\|)$$

  for some $c$ and $a \leq x \leq b$
Constructing Confidence Bands

- For now, suppose that \( \sigma \) is known, then probability of estimate not in confidence band in at least one position \( x \)

\[
P(\bar{r}(x) \notin \mathcal{I}(x) \text{ for some } x \in [a, b]) = P \left( \max_{x \in [a, b]} \frac{|\hat{r}(x) - \bar{r}|}{\sigma \| l(x) \|} > c \right)
\]

- We are left just with the error term

\[
= P \left( \max_{x \in [a, b]} \frac{|\sum_i \epsilon_i l_i(x)|}{\sigma \| l(x) \|} > c \right) = P \left( \max_{x \in [a, b]} |W(x)| > c \right)
\]

- This is a Gaussian process: a random function such that the vector \((W(x_1), \ldots, W(x_k))\) has a multivariate normal distribution, for any finite set of points \(x_1, \ldots, x_k\)

\[
W(x) = \sum_{i=1}^{n} Z_i T_i(x), \quad Z_i = \epsilon_i / \sigma \sim N(0, 1), \quad T_i(x) = l_i(x) \| l(x) \|
\]
Constructing Confidence Bands

- We want to find $c$ for a fixed probability
- We need to compute the distribution of the maximum of a Gaussian process
- This is a well studied problem
  - Hotelling wrote about in 1939 (Tubes and spheres in $n$-spaces and a class of statistical problems)
  - There is a book treatment on this by Adler and Taylor (Random Fields And Geometry) connecting probability, geometry, and topology
  - In our neuroimaging example, we used permutation test to find maximum voxel clusters
Constructing Confidence Bands

- One can show that (cdf of the standard normal $\Phi$)

$$P \left( \max_x \left| \sum_{i=1}^{n} Z_i T_i(x) \right| > c \right) \approx 2(1 - \Phi(c)) + \frac{\kappa_0}{\pi} e^{-c^2/2}$$

for large $c$, $\kappa_0 = \int_a^b \| T'(x) \| dx$, and $T'(x) = \partial T_i(x) / \partial x$

- Think of $T(x)$ as a curve in $R^n$, and $c$ as defining a tube around it with radius $c$

- Intuition: The task is to calculate the volume of this tube

- We choose $c$ by solving for $\alpha$ (e.g. $\alpha = 0.05$)

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

- So far we assumed that $\sigma$ was known.
- If unknown, we can use an estimate $\hat{\sigma}$.
- In this setting, one replaces the normal distribution with the $t$-distribution, however, for large $n$ the previous approach remains a good approximation.
- For changing variance $\sigma(x)$ as a function of $x$, 

  \[ \text{Var}(\hat{r}_n(x)) = \sum_{i=1}^{n} \sigma^2(x_i) l_i^2(x) \]

- Then this confidence is used

  \[ \mathcal{I}(x) = \hat{r}_n(x) \pm c \sqrt{\sum_{i=1}^{n} \hat{\sigma}^2(x_i) l_i^2(x)} \]

  with $c$ computed the same way.
**Average Coverage**

- So far we required coverage bands to cover the function at all $x$.
- We can relax this requirement a bit.
- Suppose we are estimating $r(x)$ over an interval $[0, 1]$, then the **average coverage** is defined as

$$C = \int_0^1 P(r(x) \in [d(x), u(x)]) \, dx$$
Bootstrap Confidence Bands

- There are at least two different ways to implement the bootstrap for regression problems
- Resample rows:
  - Assume both $Y$ and $X$ are random
  - Rows need to be iid
- Resample residuals:
  - Assume that only $Y$ is random and $x$ is fixed
  - Errors need to be iid
Bootstrap Confidence Bands (Example)

- Experiment with $n = 164$ men to see if the drug cholostyramine lowered blood cholesterol levels
- They were supposed to take six packets of cholostyramine per day, but many actually took much less
Bootstrap Confidence Bands (Example)
Bootstrap Confidence Bands (Example)
References

- Wasserman (2006). All of Nonparametric Statistics
- Efron and Tibshirani (1994). An Introduction to the Bootstrap