Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

Nonparametrics and Local Methods

Richard L. Sweeney

based on slides by Chris Conlon

Empirical Methods
Spring 2019

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

1 Nonparametric Density Estimation

2 Cross-Validation

3 Example: Auctions

4 Non-parametric Regression

5 Semi-parametrics

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

Section outline

Why nonparametrics?

- sometimes just interested in the distribution
- sometimes this is the first stage and we want to integrate
- sometimes want to do something semiparametric

In this section, we are interested in estimating the **density** $f(x)$ under minimal assumptions.

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

Let's start with the histogram

One of the more successful and popular uses of nonparametric methods is estimating the density or distribution function $f(x)$ or $F(x)$.

$$\hat{f}_{HIST}(x_0) = \frac{1}{N} \sum_{i=1}^{N} \frac{\mathbf{1}(x_0 - h < x_i < x_0 + h)}{2h}$$

- Divide the dataset into bins, count up fraction of observations in each bins

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

Kernel Estimation

Let's rewrite the histogram estimator

$$\hat{f}_{HIST}(x_0) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x_i - x_0}{h}\right)$$

Where $K(z) = \frac{1}{2} \cdot \mathbf{1}(|z| < 1)$

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

Density estimator interpretation

- for each observation, there is probability mass 1 to spread around

- use the function $K(\cdot)$ and smoothing parameter h to choose how to allocate this mass

- then, for any given x_0, sum over these functions that spread out mass, and normalize by dividing by N

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

Smooth Kernels

We call $K(\cdot)$ a Kernel function and h the bandwidth. We usually assume

ⓘ $K(z)$ is symmetric about 0 and continuous.

ⓘⓘ $\int K(z)dz = 1$, $\int zK(z)dz = 0$, $\int |K(z)|dz < \infty$.

ⓘⓘⓘ Either (a) $K(z) = 0$ if $|z| \geq z_0$ for some z_0 or (b) $|z|K(z) \to 0$ as $|z| \to \infty$.

ⓘⓥ $\int z^K(z)dz = \kappa$ where κ is a constant.

Usually we choose a smooth, symmetric K. But a common nonsmooth choice: $K(x) = (|x| < 1/2)$ gives the *histogram* estimate.

Non-
parametrics

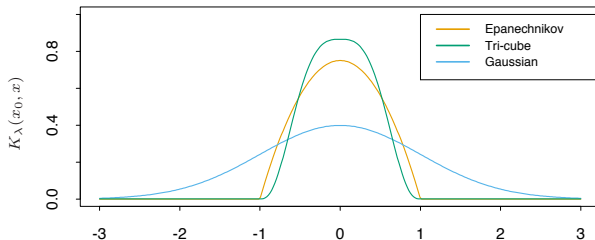Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

Some Common Kernels

Table 9.1. *Kernel Functions: Commonly Used Examples[a]*

Kernel	Kernel Function $K(z)$	δ
Uniform (or box or rectangular)	$\frac{1}{2} \times \mathbf{1}(|z| < 1)$	1.3510
Triangular (or triangle)	$(1 - |z|) \times \mathbf{1}(|z| < 1)$	–
Epanechnikov (or quadratic)	$\frac{3}{4}(1 - z^2) \times \mathbf{1}(|z| < 1)$	1.7188
Quartic (or biweight)	$\frac{15}{16}(1 - z^2)^2 \times \mathbf{1}(|z| < 1)$	2.0362
Triweight	$\frac{35}{32}(1 - z^2)^3 \times \mathbf{1}(|z| < 1)$	2.3122
Tricubic	$\frac{70}{81}(1 - |z|^3)^3 \times \mathbf{1}(|z| < 1)$	–
Gaussian (or normal)	$(2\pi)^{-1/2} \exp(-z^2/2)$	0.7764
Fourth-order Gaussian	$\frac{1}{2}(3 - z)^2 (2\pi)^{-1/2} \exp(-z^2/2)$	–
Fourth-order quartic	$\frac{15}{32}(3 - 10z^2 + 7z^4) \times \mathbf{1}(|z| < 1)$	–

[a] The constant δ is defined in (9.11) and is used to obtain Silverman's plug-in estimate given in (9.13).

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

Kernel Comparison

FIGURE 6.2. *A comparison of three popular kernels for local smoothing. Each has been calibrated to integrate to 1. The tri-cube kernel is compact and has two continuous derivatives at the boundary of its support, while the Epanechnikov kernel has none. The Gaussian kernel is continuously differentiable, but has infinite support.*

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

Mean and Variance of $\hat{f}(x_0)$

Assume that the derivative of $f(x)$ exists and is bounded, and $\int zK(z)dz = 0$

Then the estimator has **bias**

$$b(x_0) = E\left[\hat{f}(x_0)\right] - f(x_0) = \frac{1}{2}h^2 f''(x_0)\int z^2 K(z)dx$$

The **variance** of the estimator is

$$V\left[\hat{f}(x_0)\right] = \frac{1}{Nh}f(x_0)\int K(z)^2 dz \left\{+o(\frac{1}{Nh})\right\}$$

So, unsurprisingly, the bias is *increasing* in h, and the variance is *decreasing* in h.

Non-parametrics

Richard L. Sweeney

Density Estimation

Cross-Validation

Example: Auctions

Non-parametric Regression

Semi-parametrics

References

How to Choose h

- We want both bias and variance to be as small as possible, as usual.

- In parametric estimation, it is not a problem: they both go to zero as sample size increases.

- In nonparametric estimation reducing h reduces bias, but increases variance; how are we to make his trade off?

- Note that how we set h is going to be much more important than the choice of $K(\cdot)$

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

Mean Integrated Square Error

- Start with the *local* performance at x_0

$$MSE\left[\hat{f}(x_0)\right] = E\left[\left(\hat{f}(x_0) - f(x_0)\right)^2\right]$$

- Calculate the *integrated* (as opposed to expected) squared error

$$\int \left(\hat{f}(x) - f(x)\right)^2 dx = \int \text{bias}^2\left(\hat{f}(x)\right) + \text{var}\left(\hat{f}(x)\right) dx$$

- Simple approximate expression (symmetric order 2 kernels):

$$(\text{bias})^2 + \text{variance} = Ah^4 + B/nh$$

with $A = \int \left(f''(x)\right)^2 \left(\int u^2 K\right)^2 /4$ and $B = f(x)\int K^2$

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

Optimal bandwidth

- The AMISE is

$$Ah^4 + B/nh$$

- Minimize by taking the FOC

$$h_n^* = \left(\frac{B}{4An} \right)^{1/5}$$

- bias and standard error are *both* in $n^{-2/5}$
- and the AMISE is $n^{-4/5}$—**not** $1/n$ as it is in parametric models.
- But: A and B both depend on K (known) and $f(y)$ (unknown), and especially "wiggliness" $\int (f'')^2$ (unknown, not easily estimated). Where do we go from here?

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

Optimal bandwidth

Can be shown that the optimal bandwidth is

$$h* = \delta \left(\int f''(x_0)^2 dx_0 \right)^{-0.2} N^{-0.2}$$

where δ depends on the kernel used (Silverman 1986) [these δ's are given in the kernel table]

Note the "optimal" kernel is Epanechnikov, although the difference is small.

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

Silverman's Rule of Thumb

- If f is normal with variance σ^2 (may not be a very appropriate benchmark!), the optimal bandwidth is

$$h_n^* = 1.06\sigma n^{-1/5}$$

- In practice, typically use **Silverman's plug-in estimate**:

$$h_n^* = 0.9 * \min(s, IQ/1.34) * n^{-1/5}$$

  where IQ=interquartile distance

- Investigate changing it by a reasonable multiple.

This tends to work pretty well. But can we do better?

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

Why not search for optimal h in our data?

- Know we want to minimize MISE.
- One option is to find the h that minimizes it *in sample*
  - Loop through increments of h
  - Calculate MISE
- Example: Old Faithful R data
  - Waiting time between eruptions and the duration of the eruption for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA.
  - See R code in this folder.

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

Cross-validation

- General concept in the whole of nonparametrics: choose h to minimize a criterion $CV(h)$ that approximates

$$AMISE(h) = \int E(\hat{f}_n(x) - f(x))^2 dx.$$

- Usually programmed in metrics software. *If you can do it, do it on a subsample, and rescale.*
- CV tries to measure what the expected out of sample (OOS or EPE) prediction error of a new never seen before dataset.
- The main consideration is to prevent overfitting.
  - In sample fit is always going to be maximized by the most complicated model.
  - OOS fit might be a different story.
  - ie 1-NN might do really well in-sample, but with a new sample might perform badly.

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

Sample Splitting/Holdout Method and CV

Cross Validation is actually a more complicated version of sample splitting that is one of the organizing principles in machine learning literature.

Training Set This is where you estimate parameter values.

Validation Set This is where you choose a model- a bandwidth h or tuning parameter λ by computing the error.

Test Set You are only allowed to look at this after you have chosen a model. Only Test Once: compute the error again on fresh data.

- Conventional approach is to allocate 50-80% to training and 10-20% to Validation and Test.
- Sometimes we don't have enough data to do this reliably.

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

Sample Splitting/Holdout Method

FIGURE 5.1. *A schematic display of the validation set approach. A set of n observations are randomly split into a training set (shown in blue, containing observations 7, 22, and 13, among others) and a validation set (shown in beige, and containing observation 91, among others). The statistical learning method is fit on the training set, and its performance is evaluated on the validation set.*

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

Challenge with Sample Splitting
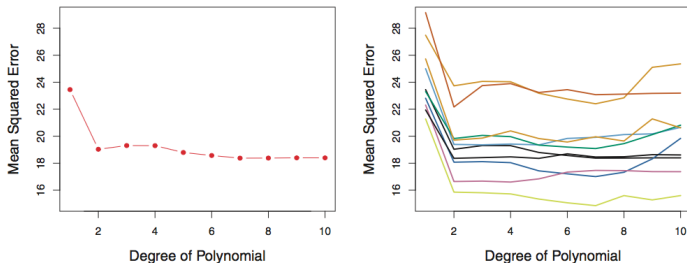
FIGURE 5.2. *The validation set approach was used on the* Auto *data set in order to estimate the test error that results from predicting* mpg *using polynomial functions of* horsepower. *Left: Validation error estimates for a single split into training and validation data sets. Right: The validation method was repeated ten times, each time using a different random split of the observations into a training set and a validation set. This illustrates the variability in the estimated test MSE that results from this approach.*

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

k-fold Cross Validation

- Break the dataset into k equally sized "folds" (at random).
- Withhold $i = 1$ fold
    - Estimate the model parameters $\hat{\theta}^{(-i)}$ on the remaining $k-1$ folds
    - Predict $\hat{y}^{(-i)}$ using $\hat{\theta}^{(-i)}$ estimates for the ith fold (withheld data).
    - Compute $MSE_i = \frac{1}{k \cdot N} \sum_j (y_j^{(-i)} - \hat{y}_j^{(-i)})^2$.
    - Repeat for $i = 1, \ldots, k$.
- Construct $\widehat{MSE}_{k,CV} = \frac{1}{k} \sum_i MSE_i$
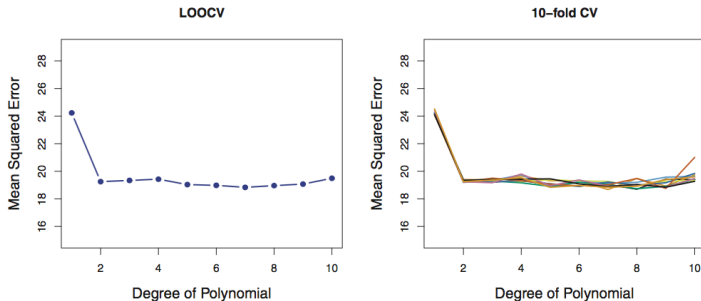
k-fold Cross Validation
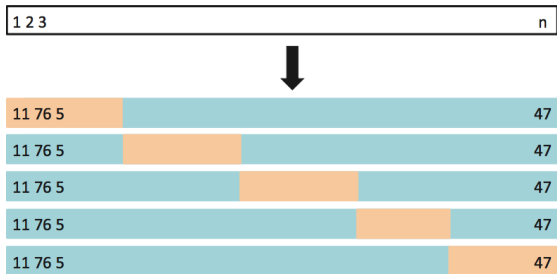
FIGURE 5.5. *A schematic display of* 5*-fold CV. A set of n observations is randomly split into five non-overlapping groups. Each of these fifths acts as a validation set (shown in beige), and the remainder as a training set (shown in blue). The test error is estimated by averaging the five resulting MSE estimates.*

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

Leave One Out Cross Validation (LOOCV)

Same as k-fold but with $k = N$.

- Withhold a single observation i
- Estimate $\hat{\theta}_{(-i)}$.
- Predict $\hat{y}_i$ using $\hat{\theta}^{(-i)}$ estimates
- Compute $MSE_i = \frac{1}{N} \sum_j (y_i - \hat{y}_i(\hat{\theta}^{(-i)}))^2$.

Note: this requires estimating the model N times which can be costly.

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

LOOCV vs k-fold CV

FIGURE 5.4. *Cross-validation was used on the* `Auto` *data set in order to estimate the test error that results from predicting* `mpg` *using polynomial functions of* `horsepower`. *Left: The LOOCV error curve. Right: 10-fold CV was run nine separate times, each with a different random split of the data into ten parts. The figure shows the nine slightly different CV error curves.*

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

Cross Validation

- Main advantage of cross validation is that we use all of the data in both estimation and in validation.
  - For our purposes validation is mostly about choosing the right bandwidth or tuning parameter.
- We have much lower variance in our estimate of the OOS mean squared error.
  - Hopefully our bandwidth choice doesn't depend on randomness of splitting sample.

Non-parametrics

Richard L. Sweeney

Density Estimation

Cross-Validation

Example: Auctions

Non-parametric Regression

Semi-parametrics

References

Test Data

- In Statistics/Machine learning there is a tradition to withhold 10% of the data as Test Data.

- This is completely new data that was not used in the CV procedure.

- The idea is to report the results using this test data because it most accurately simulates true OOS performance.

- We don't do much of this in economics.
  (Should we do more?)

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

Local Bandwidths

If you only care about $f(y)$ at some given point, then

$$A = f''(y)^2 \left(\int u^2 K \right)^2 /4 \text{ and } B = f(y) \int K^2.$$

So in a low-density region, worry about variance and take h larger. In a curvy region, worry about bias and take h small.

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

Higher-Order Kernels

- K of order r iff $\int x^j K(x)dx = 0$ for $j < r$ and $\int x^r K(x)dx \neq 0$. Try $r > 2$?

- The beauty of it: bias in h^r if f is at least C^r...so AMISE can be reduced to $n^{-r/(2r+1)}$, almost $\sqrt{n}$-consistent if r is large.

- But gives wiggly (and sometimes negative) estimates $\rightarrow$ leave them to theorists.

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

Back to the CDF

Since now we have estimated the density with

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right),$$

a natural idea is to integrate; let $\mathcal{K}(x) = \int_{-\infty}^{x} K(t)dt$, try

$$\hat{F}_n(y) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{K}\left(\frac{y - y_i}{h}\right)$$

as a reasonable estimator of the cdf in y.

Very reasonable indeed:

- when $n \longrightarrow \infty$ and h goes to zero (at rate $n^{-1/3}\ldots$) it is consistent at rate $\sqrt{n}$
- it is nicely smooth
- by construction it accords well with the density estimator
- ... it is a much better choice than the empirical cdf.

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

Example application: Auctions

- Why auctions?
- Great introduction to structural approach. Arguably the most successful application.
- Auctions are an example of a game with assymetric information: participants know the primitives of the game, but do not know their rivals exact valuations.
- By imposing rationality/ profit maximization, we can recover the distribution of values.
- Can then run counterfatuals

For more detail, check out Chris Conlon's slides in this folder, or John Asker's PhD lecture notes.

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

Something Familiar

Let's start with something we all know, how to calculate:

$$E[Y_i|X_i] = \beta_0 + \beta_1 X_1 + \ldots + \varepsilon_i$$

- The Gauss-Markov theorem (remember that?) tells us that OLS is best among linear unbiased estimators.

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

Identification

What does it mean for a model to be identified?

$$\hat{\beta}_{OLS} = \arg\max_{\beta}(\mathbf{y} - \mathbf{x}\beta)'(\mathbf{y} - \mathbf{x}\beta)$$

- We need there to be a unique value of β that solves the above equation.
- For OLS this reduces to solving a linear system of $\dim(X) = k$ equations and unknowns.
- The OLS estimator produces $\hat{\beta}_{OLS} = \beta + (X'X)^{-1}X'\varepsilon$.
  - Identification: means that we require $(X'X)$ is invertible (no perfect multicollinearity).
  - When $E[X'\varepsilon] = 0$ then this converges in probability to true β.

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

A Little More Complicated

- For many cases that we care about, $E[X'\varepsilon] \neq 0$. (Endogeneity)

- We know this is not the end of the world if we have an instrument Z:

- Now we need that $Z'X$ is invertible.

$$
\begin{aligned}
E[\mathbf{x}'(\mathbf{y} - \mathbf{x}\beta)] &\neq 0 \\
E[\mathbf{z}'(\mathbf{y} - \mathbf{x}\beta)] &= 0 \\
\rightarrow \hat{\beta}_{TSLS} &= E(Z'X)^{-1}E[Z'Y]
\end{aligned}
$$

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

Some more information

This leads to the common terminology:

- Under-identified: if $\dim(X) > \dim(Z)$.
- Just-identified: if $\dim(X) = \dim(Z)$.
- Over-identified: if $\dim(X) < \dim(Z)$.

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

What does this mean exactly

To understand this we usually construct the GMM form of the moment conditions:

$$
\begin{aligned}
E[\mathbf{z}'(\mathbf{y} - \mathbf{x}\beta)] &= 0 \\
f(\beta) &= (\mathbf{y} - \mathbf{x}\beta)'\mathbf{z} \cdot W \cdot \mathbf{z}'(\mathbf{y} - \mathbf{x}\beta)
\end{aligned}
$$

- Because $f(\beta)$ is a quadratic form, and when W is PSD then $f(\beta) \geq 0$.
- Just-identified: There is exactly one solution to $f(\beta) = 0$ which does not depend on W.
- Under-identified: $f(\beta) = 0$ has many solutions because we have fewer instruments (or moment restrictions) than parameters.
- Over-identified: There are no solutions and $f(\beta) > 0$. Instead we choose β to minimize the distance from zero with distance metric W.

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

The Binary Case

Now let's think about a case where $Y \in \{0, 1\}$:

$$y_i = \mathbf{1}[F(X_i) - \varepsilon_i > 0]$$

- There are different choices of $F(X_i)$ and $f(\varepsilon)$
- Probit: $F(X_i) = \beta X_i$ where $\varepsilon_i \sim N(0, 1)$.
- Logit: same $F(\cdot)$ but different $f(\varepsilon)$ (Type I Extreme Value). $P(Y_i = 1|X_i) = 1/(1 + exp[-\beta X_i])$.
- These choices of F are strong and somewhat arbitrary, we choose them out of convenience because both transformations are monotonic in βX_i, and continuously map $[-\infty, \infty] \rightarrow [0, 1]$

Non-parametrics

Richard L. Sweeney

Density Estimation

Cross-Validation

Example: Auctions

Non-parametric Regression

Semi-parametrics

References

The Binary Case

What is the least I can assume in the binary case and still learn something?

$$y_i = \mathbf{1}[F(X_i) - \varepsilon_i > 0]$$

- Suppose instead we observe $P(x) = Pr(y_i = 1|x_i)$ at many values of x_i. (Often refer to this as the CCP).
- Sometimes CCPs are all we care about. (if we are lucky).

Instead we can try and solve:

$$P(x) = \int \mathbf{1}(F(x) - \varepsilon > 0)dH(\varepsilon|x) = H(F(x)|x)$$

- Identification asks, what is the least we need to assume in order to recover F, H?
- Even with $\varepsilon \perp X$ we still have too many degrees of freedom.

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

Are we stuck?

- We know that Logit and Probit are identified.
- Start with $\varepsilon \perp X$.
- $F(X_i) = G(X_i\beta)$ is a very powerful assumption often called (single) index model
- Can use Marginal Rate of Substitution (MRS) of $P(x)$ to identify β_k/β_l.
  - How do conditional probabilities respond to changes in $X^{(k)}$ versus $X^{(l)}$?
  - Suggests the average derivative estimator which doesn't require normality assumptions of probit
  - More robust, but potentially less efficient if true distribution of $\varepsilon \sim N(0,1)$.
- Can we extend the general intuition to more cases?

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

Single-Index Models

- Consider
$$E\left[y|\mathbf{x}\right] = g(\mathbf{x}'\beta)$$
such as logit and probit.
- With conditional mean of change in x_j
$$\frac{\partial\left[y|\mathbf{x}\right]}{\partial x_j} = g'(\mathbf{x}'\beta)\beta_j$$
- Can use Marginal Rate of Substitution (MRS) to identify β_k/β_l.
$$\frac{\partial\left[y|\mathbf{x}\right]/\partial x_j}{\partial\left[y|\mathbf{x}\right]/\partial x_k} = \frac{\beta_j}{\beta_k}$$
- Suggests the average derivative estimator which doesn't require normality assumptions of probit
- More robust, but potentially less efficient if true distribution of $\varepsilon \sim N(0,1)$.
- Can we extend the general intuition to more cases?

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

A Fake Data Example

Following the THF textbook example, we can generate some fake data and let:
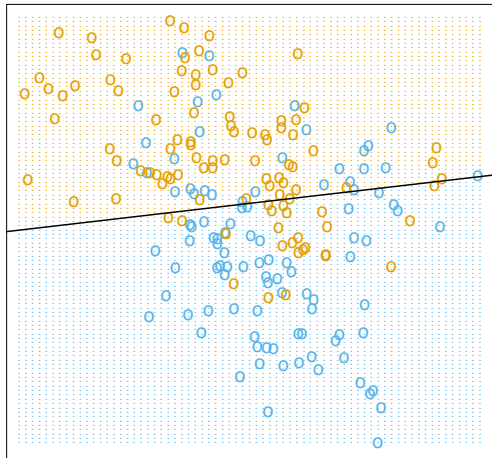
$$Y = ORANGE \text{ if } Y* > 0.5$$
$$Y = BLUE \text{ if } Y* \leq 0.5$$

- Easiest way to recover $Y*$ is by running OLS on the linear probability model.
- Draws from bivariate normal distribution with uncorrelated components but different means (2 overlapping types)
- Mixture of 10 low variance (nearly point mass) normal distributions where the individual means were drawn from another normal distribution. (10 nearly distinct types).

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

Linear Probability Model

Linear Regression of 0/1 Response

FIGURE 2.1. *A classification example in two dimensions. The classes are coded as a binary variable (*BLUE *= 0,* ORANGE *= 1), and then fit by linear regression. The line is the decision boundary defined by $x^T \hat{\beta} = 0.5$. The orange shaded region denotes that part of input space classified as* ORANGE*, while the blue region is classified as* BLUE*.*

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression
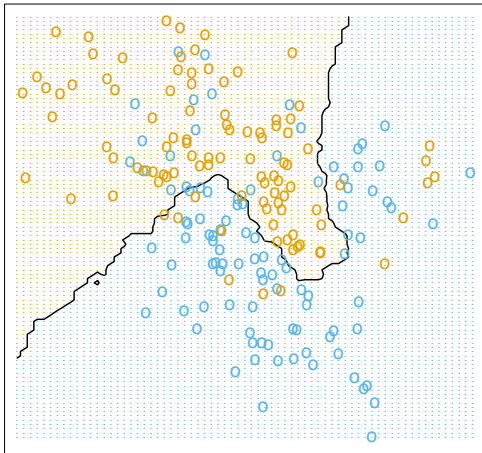
Semi-
parametrics

References

Alternative

- Lots of potential alternatives to our decision rule.
- A simple idea is to hold a majority vote of neighboring points

$$Y^* = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i$$

- To avoid including "yourself" in your neighborhood, we often estimate on one sample and validate on another
- How many parameters does this model have: None? One? k?
- Technically it has something like N/k.
- As $N \to \infty$ this means we have an infinite number of parameters! (This is a defining characteristic of non-parametrics).

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression
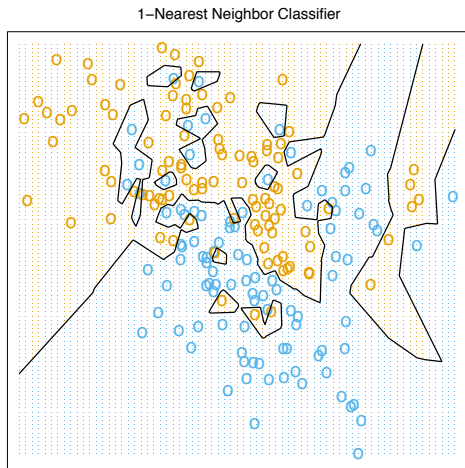
Semi-
parametrics

References

15 Nearest Neighbor

15-Nearest Neighbor Classifier

FIGURE 2.2. *The same classification example in two dimensions as in Figure 2.1. The classes are coded as a binary variable (*BLUE $= 0$, ORANGE $= 1$*) and then fit by 15-nearest-neighbor averaging as in (2.8). The predicted class is hence chosen by majority vote amongst the 15-nearest neighbors.*

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

Extreme: 1 Nearest Neighbor

1–Nearest Neighbor Classifier

FIGURE 2.3. *The same classification example in two dimensions as in Figure 2.1. The classes are coded as a binary variable (BLUE = 0, ORANGE = 1), and then predicted by 1-nearest-neighbor classification.*

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

Comparisons

- What would happen if $K \rightarrow N$?
- The k-NN model is locally constant.
- The k-NN approach tends to be really bumpy which can be undesirable.
- The OLS model is globally linear (is this always true?)

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

What about?

- If we fixed the fact that there are discrete jumps in who is in the neighborhood by smoothly weighting observations and varying those weights instead (Kernels).

- Another drawback of $k - NN$ is that we consider distance in each X dimension on the same scale, perhaps we could rescale the data to improve our "closeness" measure.

- Instead of fitting a constant locally, we fit a linear function locally (Lowess).

- Instead of using a global linear approximation in OLS use a more flexible nonlinear one.

- There is a bias/variance tradeoff. explain.

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

Bias Variance Decomposition

We can decompose any estimator into two components

$$\underbrace{E[(y - \hat{f}(x))^2]}_{MSE} = \underbrace{\left(E[\hat{f}(x) - f(x)]\right)^2}_{Bias^2} + \underbrace{E\left[\left(\hat{f}(x) - E[\hat{f}(x)]\right)^2\right]}_{Variance}$$

- In general we face a tradeoff between bias and variance.
- In k-NN as k gets large we reduce the variance (each point has less influence) but we increase the bias since we start incorporating far away and potentially irrelevant information.
- In OLS we minimize the variance among unbiased estimators assuming that the true f is linear.

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

Bias Variance Decomposition

What minimizes MSE?

$$f(x_i) = E[Y_i | X_i]$$

- Seems simple enough (but we are back where we started).
- How do we compute the expectation ?
- k-NN tries to use local information to estimate conditional mean
- OLS uses entire dataset and adds structure $y = x\beta$ to the problem.

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression
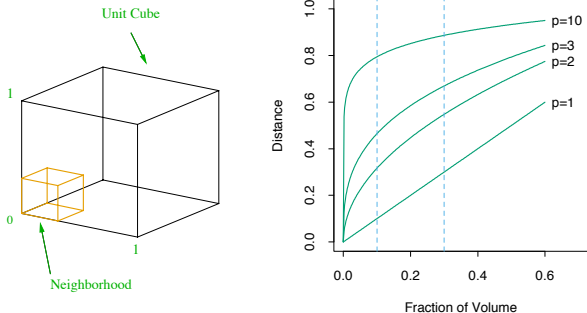
Semi-
parametrics

References

Big Data

- It used to be that if you had $N = 50$ observations then you had a lot of data.

- Those were the days of finite-sample adjusted t-statistics.

- Now we frequently have 1 million observations or more, why can't we use k-NN type methods everywhere?

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

Curse of Dimensionality

Take a unit hypercube in dimension p and we put another hypercube within it that captures a fraction of the observations r within the cube

- Since it corresponds to a fraction of the unit volume, r each edge will be $e_p(r) = r^{1/p}$.

- $e_{10}(0.01) = 0.63$ and $e_{10}(0.1) = 0.80$, so we need almost 80% of the data to cover 10% of the sample!

- If we choose a smaller r (include less in our average) we increase variance quite a bit without really reducing the required interval length substantially.

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

Curse of Dimensionality

FIGURE 2.6. *The curse of dimensionality is well illustrated by a subcubical neighborhood for uniform data in a unit cube. The figure on the right shows the side-length of the subcube needed to capture a fraction r of the volume of the data, for different dimensions p. In ten dimensions we need to cover 80% of the range of each coordinate to capture 10% of the data.*

Curse of Dimensionality

Don't worry, it only gets worse:

$$d(p, N) = \left(1 - \left(\frac{1}{2}\right)^{1/N}\right)^{1/p}$$

- $d(p, N)$ is the distance from the origin to the closest point.
- $N = 500$ and $p = 10$ means $d = 0.52$ or that the closest point is closer to the boundary than the origin!
- Why is this a problem?
- In some dimension nearly every point is the closest point to the boundary – when we average over nearest neighbors we are extrapolating not interpolating.

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

Density/Distribution Estimation

One of the more successful and popular uses of nonparametric methods is estimating the density or distribution function $f(x)$ or $F(x)$.

- Estimating the CDF is easy and something you have already done
- Q-Q plots, etc.

$$\hat{F}_{ECDF}(x_0) = \frac{1}{N} \sum_{i=1}^{N} (x_i \leq x_0)$$

- Differentiating to get density is unhelpful : $F'_{ECDF}(x) = 0$ in most places.

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

What if y is of dimension $p_y > 1$?

"Easy": use p_y-dimensional K (often a p_y-product of 1-dim kernels) and bandwidth h, and do

$$\hat{f}_n(y) = \frac{1}{nh_y^p} \sum_{i=1}^{n} K\left(\frac{y - y_i}{h}\right).$$

- **1st minor pitfall:** the various dimensions may have very different variances, so use $(h_1, \ldots, h_{p_y})$.
- **2nd minor pitfall:** they may be strongly correlated; then sphericize first.
- **Major problem:** next slide. . .

Non-parametrics

Richard L. Sweeney

Density Estimation

Cross-Validation

Example: Auctions

Non-parametric Regression

Semi-parametrics

References

The Curse of Dimensionality

- Computational cost increases exponentially.
- *Much worse:* to achieve precision ϵ in dimension p_y, the number of observations you need increases as

$$n \simeq \epsilon^{-(2+p_y/2)}.$$

- The *empty space* phenomenon: if $(y_1, \ldots, y_{p_y})$ all are iid uniform on $[-1, 1]$, then only $n/(10^{p_y})$ observations on average have all components in $[-0.1, 0.1]$. Bias still in h^2, but variance in $1/nh^{p_y}$ now.

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

Silverman's Table

Silverman (1986 book) provides a table illustrating the difficulty of kernel estimation in high dimensions. To estimate the density at 0 of a $N(0,1)$ with a given accuracy, he reports:

Dimensionality Required	Sample Size
1	4
2	19
5	786
7	10,700
10	842,000

Not to be taken lightly... in any case convergence with the optimal bandwidth is in $n^{-2/(4+p_y)}$ now—and Silverman's rule of thumb for choosing h_n^* must be adapted too.

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

Usually we care about conditional densities

That is: we have covariates x, we want the density $f(y|x)$. Again, "easy":

1. get a kernel estimator of the joint density $f(y, x)$;
2. and one of of the marginal density $f(x)$;
3. then define

$$\hat{f}_n(y|x) = \frac{\hat{f}_n(y, x)}{\hat{f}_n(x)} = \frac{\frac{1}{nh_y^{py}h_x^{px}} \sum_{i=1}^n K\left(\frac{y-y_i}{h_y}\right) K\left(\frac{x-x_i}{h_x}\right)}{\frac{1}{h_y^{py}} \sum_{i=1}^n K\left(\frac{x-x_i}{h_x}\right)}.$$

But the joint density is $(p_x + p_y)$ dimensional... and the curse strikes big time.

What if my distribution is discrete-continuous?

Very often in microeconometrics some covariates only take discrete values (e.g. gender, race, income bracket...). Say the only discrete variable is gender, we care about the density of income of men.

- The kernel approach adapts directly: we separately estimate a density for men (on the corresponding subsample).

- *Better:* mix the two subsamples! Add women, but **with a small weight** w.

- Intuition: by doing so we increase the bias (the density for women is probably different than for men) $\rightarrow$ bad, in w^2 but we reduce the variance, by $O(w)$; and this dominates for small w. (cf Li-Racine).

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

The Seminonparametric Approach

- If we are "pretty sure" that f is almost $f_{m,\sigma}$ for some family of densities indexed by (m, σ), then we can choose a family of positive functions of increasing complexity $P_\theta^1, P_\theta^2, \ldots$

- Choose some M that goes to infinity as n does (more slowly), and maximize over (m, σ, θ) the loglikelihood

$$\sum_{i=1}^n \log f_{m,\sigma}(y_i) P_\theta^M(y_i).$$

It works... but it is hard to constrain it to be a density for large M.

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

Mixtures of Normals

A special case of seminonparametrics, and usually a very good approach: Let $y|x$ be drawn from

$$N(m_1(x,\theta), \sigma_1^2(x,\theta)) \text{ with probability } q_1(x,\theta);$$
$$\ldots$$
$$N(m_K(x,\theta), \sigma_K^2(x,\theta)) \text{ with probability } q_K(x,\theta).$$

where you choose some parameterizations, and the q_k's are positive and sum to 1.

Can be estimated by maximum-likelihood:

$$\max_\theta \sum_{i=1}^n \log \left(\sum_{k=1}^K \frac{q_k(x_i,\theta)}{\sigma_k(x,\theta)} \phi \left(\frac{y_i - m_k(x_i,\theta)}{\sigma_k(x_i,\theta)} \right) \right).$$

Usually works very well with $K \le 3$ (perhaps after transforming y to $\log y$, e.g).

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

Nonparametric Regression

Data $(y_i, x_i)_{i=1}^n$ now, we are after $E(g(y, x)|x) = m(x)$ for some function g.

- Best-fit approach, quite unbiased:
    - if $x = x_i$ then $\hat{m}_n(x) = g(y_i, x_i)$; otherwise . . . whatever.
  But: very jagged estimate; variance independent of n, so not consistent.

- Better and most usual: Nadaraya-Watson, inspired from kernel idea:

$$\hat{m}_n(x) = \frac{\sum_{i=1}^n g(y_i, x_i) K\left(\frac{x-x_i}{h}\right)}{\sum_{i=1}^n K\left(\frac{x-x_i}{h}\right)}.$$

  again, bias in h^2 and variance in $1/nh$ if $p_x = 1$.
  Pitfall 1: very unreliable where $f(x)$ is small.
  Pitfall 2: the formal for the optimal bandwidth h is very ugly.

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

Choosing h

- Plug-in estimates work badly.
- Fortunately, cross-validation amounts to

$$\min_h \sum_{i=1}^{n} \frac{(g(y_i, x_i) - \hat{m}_n(x_i; h))^2}{1 - k_i(h)}$$

  where $k_i(h) = K_h(0) / \sum_{j=1}^{n} K_h(x_i - x_j)$.
- So not that hard, and can be done on a subsample and rescaled.

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression
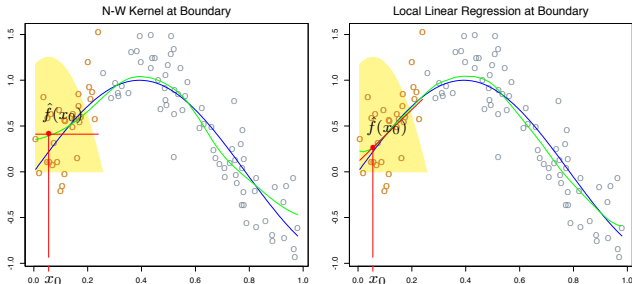
Semi-
parametrics

References

Local Linear Regression

- The Nadaraya-Watson estimator in x can be obtained very simply by regressing $g(y_i, x_i)$ on 1, weighting each observation by $K((x - x_i)/h)$.

- We could also regress on 1 and $(x - x_i)$ (going to higher terms has problems) instead;
  Advantages:
  - the bias becomes 0 if the true $m(x)$ is linear.
  - the coefficient of $(x - x_i)$ estimates $m'(x)$.
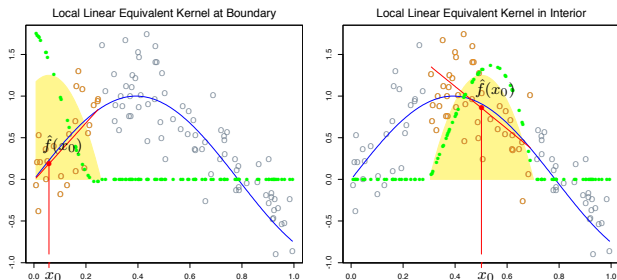 - behaves better in "almost empty" regions.

  Disadvantages: hardly any, just do it!

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics
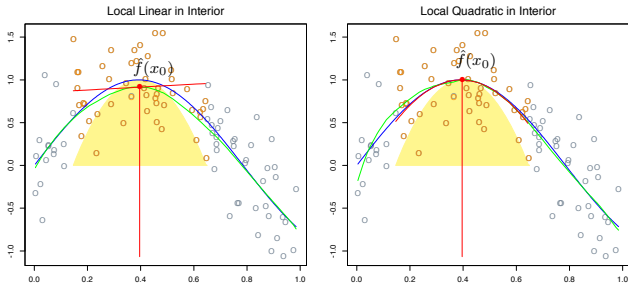
References

Local Linear

FIGURE 6.3. *The locally weighted average has bias problems at or near the boundaries of the domain. The true function is approximately linear here, but most of the observations in the neighborhood have a higher mean than the target point, so despite weighting, their mean will be biased upwards. By fitting a locally weighted linear regression (right panel), this bias is removed to first order.*

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

Local Linear

FIGURE 6.4. *The green points show the equivalent kernel $l_i(x_0)$ for local regression. These are the weights in $\hat{f}(x_0) = \sum_{i=1}^{N} l_i(x_0)y_i$, plotted against their corresponding x_i. For display purposes, these have been rescaled, since in fact they sum to 1. Since the yellow shaded region is the (rescaled) equivalent kernel for the Nadaraya–Watson local average, we see how local regression automatically modifies the weighting kernel to correct for biases due to asymmetry in the smoothing window.*

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

Local Quadratic

FIGURE 6.5. *Local linear fits exhibit bias in regions of curvature of the true function. Local quadratic fits tend to eliminate this bias.*

Nonparametric Regression, summary, 1

Nadaraya–Watson for $E(y|x) = m(x)$

$$\hat{m}(x) = \frac{\sum_i y_i K_h(x - x_i)}{\sum_i K_h(x - x_i)}$$

- bias in $O(h^2)$, variance in $1/(nh^{p_x})$
- optimal h in $n^{-1/(p+4)}$: then bias, standard error and RMSE all converge at rate $n^{-2/(p+4)}$
- to select h, no rule of thumb: cross-validate on a subsample and scale up.

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

Nonparametric Regression, summary, 2

Nadaraya–Watson=**local constant regression**: to get $\hat{m}(x)$,

1. regress y_i on 1 with weight $K_h(x - x_i)$
2. take the estimated coeff as your $\hat{m}(x)$.

Better: **local linear regression**

1. regress y_i on 1 and $(x_i - x)$ with weight $K_h(x - x_i)$
2. take the estimated coeffs as your $\hat{m}(x)$ and $\hat{m}'(x)$.

To estimate the standard errors: bootstrap on an *undersmoothed* estimate (so that bias is negligible.)

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

Seminonparametric (=Flexible) Regression

Idea: we add regressors when we have more data

$\rightarrow$ **series or sieve estimators**: choose a basis of functions $P_k(x_i)$ (x_i^k, or orthogonal polynomials, or sines...)

$\rightarrow$ run *linear regression* $y_i = \sum_{k=1}^{M} P_k(x_i)\theta_k + \epsilon_i$

a reasonable compromise (again, M must go to infinity, more slowly than n).

Still curse of dimensionality, and nonparametric asymptotics.

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

Splines: trading off fit and smoothness

Choose some $0 < \lambda < \infty$ and

$$\min_{m(.)} \sum_i (y_i - m(x_i))^2 + \lambda J(m),$$

Then we "obtain" the natural cubic spline with knots=$(x_1, \ldots, x_n)$:

- m is a cubic polynomial between consecutive x_i's
- it is linear out-of-sample
- it is C^2 everywhere.

"Consecutive" implies one-dimensional... harder to generalize to $p_x > 1$.

Orthogonal polynomials: check out Chebyshev, $1, x, 2x^2 - 1, 4x^3 - 3x \ldots$ (on $[-1, 1]$ here.)

Non-
parametrics

Richard L.
Sweeney

Density
Estimation

Cross-
Validation

Example:
Auctions

Non-
parametric
Regression

Semi-
parametrics

References

Review: What was the point?

- OLS is lowest variance among linear unbiased estimators.
- But there are nonlinear estimators and potentially biased estimators.
  - Everything faces a bias-variance tradeoff.
  - Nearly anything can be written as Kernel.

Acknowledgements/ References