Advanced Statistical Modeling

Part 2. Nonparametric Modeling

Session 2:

Nonparametric regression model II

Pedro Delicado

Departament d'Estadística i Investigació Operativa

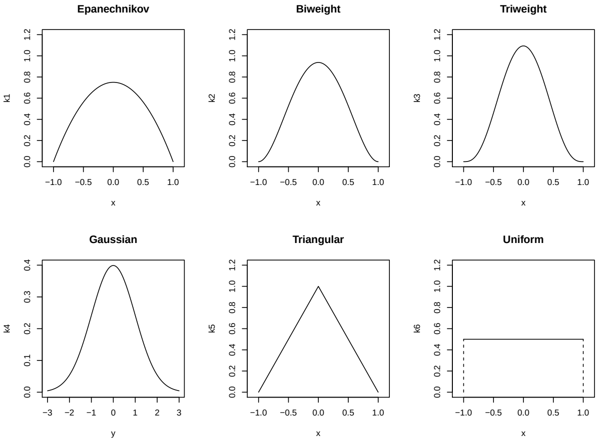
Universitat Politècnica de Catalunya

Kernel functions

Theoretical properties. The bias-variance trade-off
- Local properties of local polynomial estimator
- The bias-variance trade-off

Linear smoothers
- Effective number of parameters
- Two estimators of σ^2

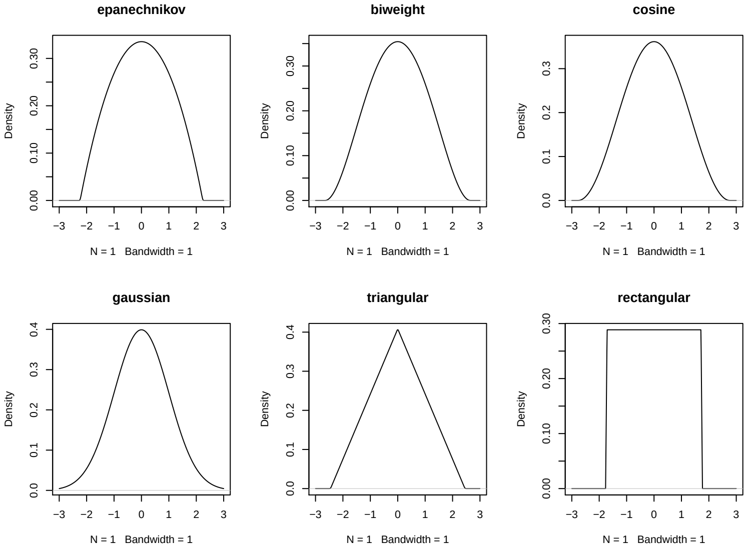
Kernel functions



Examples of Kernel functions used in nonparametric estimation.

Kernel K	Expression	Efficiency
Epanechnikov (K^*)	$(3/4)(1-x^2)I_{[-1,1]}(x)$	1
Biweight	$(15/16)(1-x^2)^2 I_{[-1,1]}(x)$	0.994
Triweight	$(35/32)(1-x^2)^3 I_{[-1,1]}(x)$	0.987
Gaussian	$(1/\sqrt{2\pi})\exp(-x^2/2)$	0.951
Triangular	$(1-\lvert x\rvert)I_{[-1,1]}(x)$	0.986
Uniform	$(1/2)I_{[-1,1]}(x)$	0.930

Kernel	Original expression	Original variance	Rescaled expression
Epanechnikov	$(3/4)(1-x^2)I_{[-1,1]}(x)$	$1/5$	$(3/4\sqrt{5})(1-x^2/5)I_{[-\sqrt{5},\sqrt{5}]}(x)$
Biweight	$(15/16)(1-x^2)^2 I_{[-1,1]}(x)$	$1/7$	$(15/16\sqrt{7})(1-x^2/7)^2 I_{[-\sqrt{7},\sqrt{7}]}(x)$
Triweight	$(35/32)(1-x^2)^3 I_{[-1,1]}(x)$	$1/9$	$(35/96)(1-x^2/9)^3 I_{[-3,3]}(x)$
Gaussian	$(1/\sqrt{2\pi})\exp(-x^2/2)$	1	$(1/\sqrt{2\pi})\exp(-x^2/2)$
Triangular	$(1-\lvert x\rvert)I_{[-1,1]}(x)$	$1/6$	$(1/\sqrt{6})(1-\lvert x\rvert/\sqrt{6})I_{[-\sqrt{6},\sqrt{6}]}(x)$
Uniform	$(1/2)I_{[-1,1]}(x)$	$1/3$	$(1/2\sqrt{3})I_{[-\sqrt{3},\sqrt{3}]}(x)$



Examples of rescaled kernel functions.

Kernel density estimation

- Let $x_1, \dots, x_n$ be independent observation of a random variable X having probability density function f.
- Let $t \in \mathbb{R}$. The goal is to estimate the density value at t: $f(t)$.
- Observe that

$$f(t) = F'(t) = \lim_{h \longrightarrow 0} \frac{F(t+h) - F(t-h)}{2h} = \lim_{h \longrightarrow 0} \frac{\Pr(t-h \leq X \leq t+h)}{2h} \approx$$

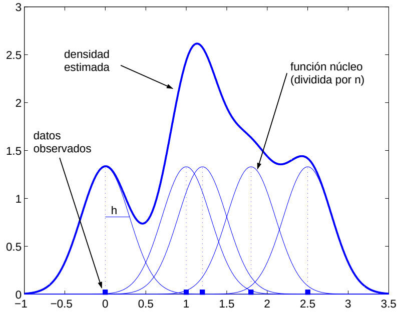
$$\lim_{h \longrightarrow 0} \frac{\#\{x_i : t-h \leq x_i \leq t+h\}/n}{2h} \overset{h \text{ small}}{\approx} \frac{\#\{x_i : t-h \leq x_i \leq t+h\}/n}{2h} =$$

$$\frac{1}{2nh} \sum_{i=1}^{n} I_{[t-h,t+h]}(x_i) = \frac{1}{nh} \sum_{i=1}^{n} \frac{1}{2} I_{[-1,1]} \left(\frac{t-x_i}{h} \right) = \frac{1}{nh} \sum_{i=1}^{n} K_U \left(\frac{t-x_i}{h} \right)$$

being K_U the uniform kernel.

For a kernel K and a bandwidth h, the kernel density estimate of $f(t)$ is

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



The weight $1/n$, corresponding to each observed data x_i, is continuously spread around x_i.

Nadaraya-Watson directly from density estimation

Given that

$$m(x) = E(Y|X = x) = \int_{\mathbb{R}} y f_Y(y|X = x) dy = \int_{\mathbb{R}} y \frac{f(x, y)}{f_X(x)} dy,$$

an estimator of $m(x)$ can be obtained replacing the unknown densities $f(x, y)$ and $f_X(x)$ by its kernel estimators

$$\hat{f}(x, y) = \frac{1}{n h_X h_Y} \sum_{i=1}^{n} K_X\left(\frac{x - x_i}{h_X}\right) K_Y\left(\frac{y - y_i}{h_Y}\right),$$

$$\hat{f}_X(x) = \frac{1}{n h_X} \sum_{i=1}^{n} K_X\left(\frac{x - x_i}{h_X}\right) = \int_R \hat{f}(x, y) dy.$$

Doing these replacements:

$$\hat{m}(x) = \int_{\mathbb{R}} y \frac{\hat{f}(x,y)}{\hat{f}_X(x)} dy = \int_{\mathbb{R}} y \frac{\frac{1}{nh_X h_Y} \sum_{i=1}^n K_X\left(\frac{x-x_i}{h_X}\right) K_Y\left(\frac{y-y_i}{h_Y}\right)}{\frac{1}{nh_X} \sum_{i=1}^n K_X\left(\frac{x-x_i}{h_X}\right)} dy =$$

$$\frac{\sum_{i=1}^n K_X\left(\frac{x-x_i}{h_X}\right) \int_{\mathbb{R}} y \frac{1}{h_Y} K_Y\left(\frac{y-y_i}{h_Y}\right) dy}{\sum_{i=1}^n K_X\left(\frac{x-x_i}{h_X}\right)}.$$

Doing the change of variable $u = (y - y_i)/h_Y$ ($y = y_i + h_Y u$) the numerator integral is equal to $\int_{\mathbb{R}} (y_i + h_Y u) K_Y(u) du = y_i$, because kernel K_Y integrates 1 and has first moment equal to zero.

Doing $h = h_X$ and $K = K_X$, we obtain this estimator of $m(x)$,

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

that is precisely the Nadaraya-Watson estimator.

Practice:

Kernel functions

Local properties of local polynomial estimator

- The term *local behavior* refers to the statistical properties of a nonparametric estimate $\hat{m}(t)$ as estimator of the unknown value $m(t)$, for a fixed value t.
- Is $\hat{m}(t)$ an unbiased estimator of $m(t)$?
  Is $E(\hat{m}(t)) = m(t)$?
- Is $\hat{m}(t)$ a consistent estimator of $m(t)$?
  Does $\hat{m}(t)$ converge to $m(t)$ in some sense?
- We talk about global properties when our interest is on $\hat{m}(t)$ as estimator of $m(t)$ for all $t \in [a, b]$, being $[a, b]$ the interval where explanatory variable takes values.
- Global properties: Does the estimated function $\hat{m}$ converge to the unknown function m in some sense appropriated for functions?

Bias and variance of $\hat{m}_0(t)$ and $\hat{m}_1(t)$

Theorem. *Consider the nonparametric regression model*

$$Y_i = m(X_i) + \varepsilon_i,\ i = 1 \dots n$$

where $\varepsilon_1, \dots, \varepsilon_n$ *are independent r.v. with* $E(\varepsilon_i) = 0$ *and* $V(\varepsilon_i) = \sigma^2(x_i)$, $X_1, \dots, X_n$ *are independent r.v. with density* f, *with* $\Pr(a \leq X_i \leq b) = 1$, *for some* $a, b \in \mathbb{R}$. *Assume the following regularity conditions:*

1. $f(t) > 0$.
2. $f(t)$, $m''(t)$ *y* $\sigma^2(t)$ *are continuous in a neighborhood of* t.
3. K *is symmetric with support on* $[-1, 1]$, $\int_R K(u)du = 1$, $\int_R uK(u)du = 0$.
4. $t \in (a, b)$.
5. $h \longrightarrow 0$ *and* $nh \longrightarrow \infty$ *when* $n \longrightarrow \infty$.

In this context, and conditioning on $X_1, \ldots, X_n$, we have the following:

- *The Nadaraya-Watson estimator and the local linear estimator both have variance*

$$\frac{\sigma^2(t)}{nhf(t)} \int_{-1}^{1} K^2(u)du + o\left(\frac{1}{nh}\right).$$

- *The Nadaraya-Watson estimator has bias*

$$\left(\frac{m'(t)f'(t)}{f(t)} + \frac{m''(t)}{2}\right) h^2 \int_{-1}^{1} u^2 K(u)du + o(h^2).$$

- *The local linear regression estimator has bias*

$$\frac{m''(t)}{2} h^2 \int_{-1}^{1} u^2 K(u)du + o(h^2).$$

- *The Mean Squared Error (MSE) of $\hat{m}(t)$ as an estimator of $m(t)$,*

$$E[(\hat{m}(t) - m(t))^2] = Bias(\hat{m}(t))^2 + V(\hat{m}(t))$$

is $O(h^4) + O(1/(nh))$ for both estimators. Then both converge to $m(t)$ in quadratic mean and in probability.

Bias and variance of $\hat{m}_q(t)$

- Local polynomial estimators with degrees $p = 2k$ and $p = 2k + 1$ give similar asymptotic results:

$$\text{MSE}(\hat{m}_p(t)) = O(h^{4k+4}) + O\left(1/(nh)\right).$$

- The bias asymptotic expression is simpler for p odd. They do not depend on the density function of X_i.
- A general recommendation is to use the degree $p = 2k + 1$ instead of using $p = 2k$.

Asymptotic Mean Squared Error (I)

- The Asymptotic Mean Squared Error (AMSE) is the main part of the MSE (ignoring the infinitesimal terms).
- The AMSE can be interpreted as a function of bandwidth h.
- For the local linear estimator:

$$\text{AMSE}(h) = \frac{(m''(t))^2}{4} h^4 \left(\int_{-1}^{1} u^2 K(u) du \right)^2 + \frac{\sigma^2(t)}{nhf(t)} \int_{-1}^{1} K^2(u) du$$

- Minimizing AMSE(h) in h, the optimal value is

$$h_{\text{AMSE}} = O\left(n^{-\frac{1}{5}}\right), \ \text{AMSE}^* = \text{AMSE}(h_{\text{AMSE}}) = O\left(n^{-\frac{4}{5}}\right).$$

The same applies for the Nadaraya-Watson estimator.

Asymptotic Mean Squared Error (II)

- For local polynomials with degree $p = 2k$ or $p = 2k + 1$,

$$V(\hat{m}_p(t)) = O\left(\frac{1}{nh}\right), \text{ (the same order for all } p\text{)},$$

$$\text{Bias}(\hat{m}_p(t)) = E(\hat{m}_p(t)) - m(t) = O\left(h^{2k+2}\right).$$

- The optimal bandwidth and the corresponding AMSE are

$$h_{\text{AMSE}} = O\left(n^{-\frac{1}{4k+5}}\right), \text{ AMSE}^* = O\left(n^{-\frac{4k+4}{4k+5}}\right), \text{ for } p = 2k \text{ or } p = 2k+1.$$

- Observe that the bias decreases when the local polynomial degree increases. But this is at the cost of an increasing in the constants appearing in the variance term.

The bias-variance trade-off

- Let us consider the AMSE for the local linear estimator:

$$\text{AMSE}(h) = \frac{(m''(t))^2}{4} h^4 \left(\int_{-1}^{1} u^2 K(u) du \right)^2 + \frac{\sigma^2(t)}{nhf(t)} \int_{-1}^{1} K^2(u) du$$

- The first term, the squared bias, increases with h.
- The second term, the variance, decreases with h.
- The optimal value h_{AMSE} represents a compromise between bias and variance.

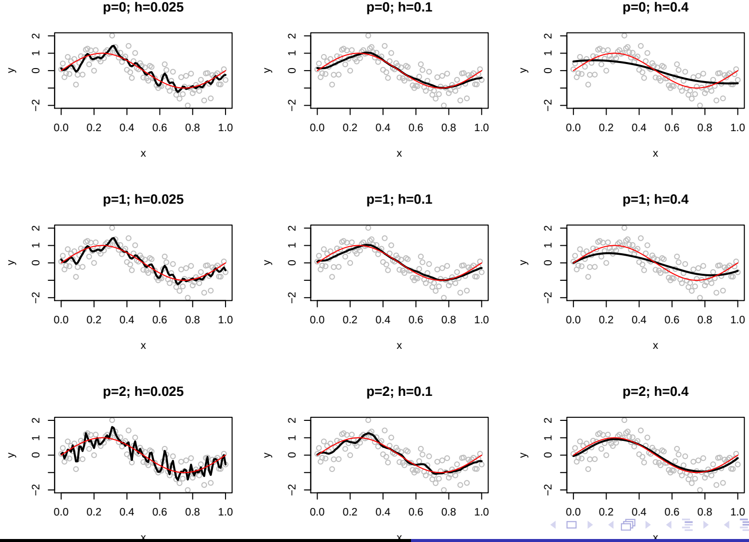
- Let $g(h) = \text{AMSE}(h)$. It has the expression $g(h) = ah^4 + b/h$. Doing $g'(h) = 0$ it follows that the minimum of g is at $h^* = (b/4a)^{1/5}$ and $g(h^*) = 5a(h^*)^4$.
- Therefore,

$$h_{\text{AMSE}} = \left(\frac{\sigma^2(t)}{nf(t)(m''(t))^2} \right)^{1/5} \left(\frac{\int_{-1}^{1} K^2(u)du}{\left(\int_{-1}^{1} u^2 K(u)du \right)^2} \right)^{1/5} n^{-1/5},$$

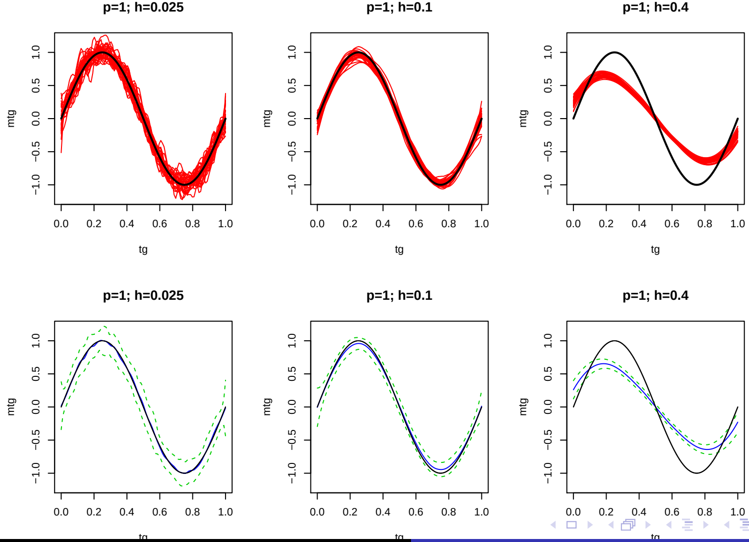
$$AMSE(h_{\text{AMSE}}) = \frac{5}{4^{4/5}} \frac{(\sigma^2(t))^{4/5}((m''(t))^2)^{1/5}}{f(t)^{4/5}}$$

$$\left(\int_{-1}^{1} K^2(u)du \right)^{4/5} \left(\int_{-1}^{1} u^2 K(u)du \right)^{2/5} n^{-4/5}.$$

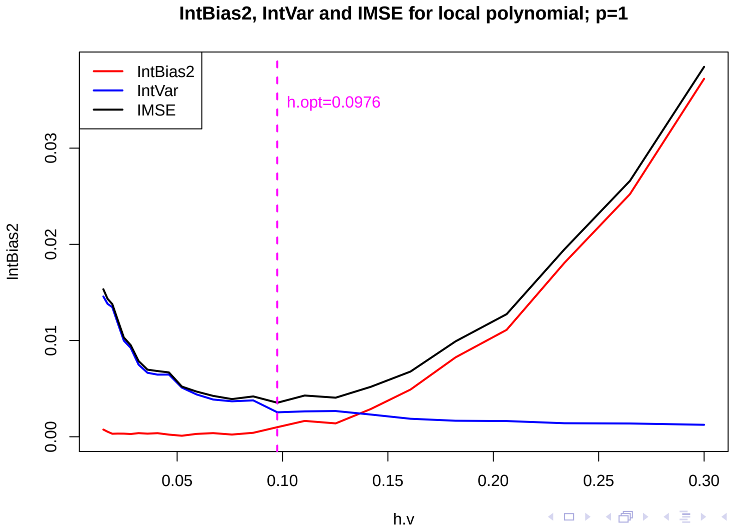
Effect of bandwidth *h* and degree *p* on a single sample



Effect of bandwidth h on many samples



Global Variance, Bias and IMSE as a function of *h*



Practice:

Bias-variance trade-off

Linear smoothers

- A nonparametric regression estimator $\hat{m}(\cdot)$ is said to be a linear estimator when for any fix t, $\hat{m}(t)$ is a linear function of $y_1, \ldots, y_n$:

$$\hat{m}(t) = \sum_{i=1}^{n} w(t, x_i) y_i.$$

for some weight function $w(\cdot, \cdot)$.

- Let

$$\hat{y}_i = \hat{m}(x_i) = \sum_{j=1}^{n} w(x_i, x_j) y_j$$

be the fitted values for the n observed values x_i of the explanatory variable.

- In matrix format,

$$\hat{Y} = SY,$$

where the column vectors Y and $\hat{Y}$ have elements y_i and $\hat{y}_i$, respectively, and the matrix S has generic (i,j) element

$$s_{ij} = w(x_i, x_j).$$

- Matrix S is called the smoothing matrix, because its effect on the observed data (x_i, y_i), $i = 1, \ldots, n$, is to transform them into $(x_i, \hat{y}_i)$, $i = 1, \ldots, n$, that is a much smoother data configuration.
- The smoothing matrix S is analogous to the hat matrix $H = X(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}$ in multiple linear regression:

$$\hat{Y}_L = X(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}Y = HY.$$

- Consider the multiple linear regression with k regressors, including the constant term:

$$Y = X\beta + \varepsilon,$$

  X being a $n \times k$ matrix.

- It is known that

$$\text{Trace}(H) = \text{Trace}(X(X^{\text{T}}X)^{-1}X^{\text{T}}) = \text{Trace}((X^{\text{T}}X)^{-1}X^{\text{T}}X) = \text{Trace}(I_k) = k,$$

  that is the number of parameters in the model.

- For a linear smoother with smoothing matrix S ($\hat{Y} = SY$) we define

$$\nu = \text{Trace}(S) = \sum_{i=1}^{n} s_{ii},$$

  the sum of diagonal elements of S.

- $\nu = \text{Trace}(S)$ is called the effective number of parameters of the nonparametric estimator corresponding to smoothing matrix S.

- In the case of local polynomial regression $\nu = \nu(h)$ is a decreasing function of smoothing parameter h:
  - Small values of h correspond to large numbers ν of effective parameters, that is, to highly complex and very flexible nonparametric models.
  - Large values of h correspond to small numbers ν of effective parameters, that is, to nonparametric models with low complexity and flexibility.
- The interpretation of ν as the effective number of parameters is valid for any linear nonparametric estimator.
- Then we can compare the degree of smoothing of two linear nonparametric estimators just comparing their effective numbers of parameters.

Kernel functions

Theoretical properties. The bias-variance trade-off

- Local properties of local polynomial estimator
- The bias-variance trade-off

Linear smoothers

- Effective number of parameters
- Two estimators of σ^2

Two estimators of σ^2

- The analogy with multiple linear regression suggests how the residual variance, $\sigma^2 = V(\varepsilon_i)$, can be estimated.
- In linear regression with k regressors,

$$Y = X\beta + \varepsilon,\ \hat{Y}_L = HY,\ \hat{\varepsilon} = Y - \hat{Y}_L = (I - H)Y.$$

  If $\varepsilon \sim N(0, \sigma^2 I)$ then $\frac{1}{\sigma^2}\hat{\varepsilon}^{\mathsf{T}}\hat{\varepsilon} = \varepsilon^{\mathsf{T}}(I - H)^{\mathsf{T}}(I - H)\varepsilon \sim \chi^2_{n-k}$.
- The value $(n - k)$ is called degrees of freedom of the model.
- Given that the expected value of a χ^2 is its degrees of freedom,

$$\hat{\sigma}^2 = \frac{1}{n-k}\hat{\varepsilon}^{\mathsf{T}}\hat{\varepsilon} = \frac{1}{n-k}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

  is an unbiased estimator of σ^2, the residual variance.
- A first estimator of σ^2 in nonparametric estimation:

$$\hat{\sigma}^2 = \frac{1}{n-\nu}\sum_{i=1}^{n}(y_i - \hat{m}(x_i))^2.$$

- Observe that $(n-k)$ is the trace of $(I-H)^{\mathsf{T}}(I-H)$, the matrix that defines the above quadratic form on ε:

$$\text{Trace}((I-H)^{\mathsf{T}}(I-H)) = \text{Trace}((I-H)(I-H)) =$$

$$\text{Trace}(I-H) = \text{Trace}(I) - \text{Trace}(H) = n-k.$$

- It has been used that H is a symmetric and idempotent matrix: $H = H^{\mathsf{T}}$ and $H^2 = H$. This is also true for matrix $(I-H)$.
- In linear nonparametric estimators, the smoothing matrix S plays a similar role to hat matrix H in multiple regression.
- The effective degrees of freedom of a linear smoother is defined as

$$\eta = \text{Trace}((I-S)^{\mathsf{T}}(I-S)) = \text{Trace}(I - S^{\mathsf{T}} - S + S^{\mathsf{T}}S) =$$

$$n - 2\,\text{Trace}(S) + \text{Trace}(S^{\mathsf{T}}S).$$

- Observe that S is not necessarily symmetric neither idempotent. Then $\text{Trace}(S) \neq \text{Trace}(S^{\mathsf{T}}S)$ in general.

- We have defined $\nu = \text{Trace}(S)$, the effective number of parameters.
- Similarly, we define $\tilde{\nu} = \text{Trace}(S^\mathsf{T} S)$.
- Then the effective degrees of freedom is

$$\eta = n - 2\nu + \tilde{\nu}.$$

- An alternative estimator of σ^2, the residual variance in the nonparametric regression model, is defined as

$$\hat{\sigma}^2 = \frac{1}{n - 2\nu + \tilde{\nu}} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

- It can be proved that $\hat{\sigma}^2$ is unbiased when the regression function $m(x)$ is linear.
- Moreover, $\hat{\sigma}^2$ is a consistent estimator of σ^2 under certain regularity conditions on $m(x)$.

Practice:

Linear smoothers

Alternative residual variance estimators

Fan, J. and I. Gijbels (1996).
Local polynomial modelling and its applications.
London: Chapman & Hall.

Hastie, T. J. and R. J. Tibshirani (1990).
Generalized additive models.
Monographs on Statistics and Applied Probability. London: Chapman and Hall Ltd.

Wand, M. P. and M. C. Jones (1995).
Kernel smoothing.
London: Chapman and Hall.

Wasserman, L. (2006).
All of Nonparametric Statistics.
New York: Springer.