Statistical Models

Daniele Zago

March 21, 2022

Contents

Part I

Nonparametric statistics

Nonparametric statistics can and should be broadly defined to include all methodology that does not use a model based on a single parametric family.

(Handles, Hettmansperger and Casella)

The basic idea of nonparametric inference is to use data to infer an unknown quantity while making as few assumptions as possible.

(Wasserman, 2005)

This part of the course will focus on nonparametric statistics, and with particular focus on nonparametric estimation and regression. The topics of ranks, rank test and permutation methods will be left for other, more in-depth, courses.

References Wasserman (2005)

Lecture 4: Nonparametric statistics

2022-03-09

The main idea of nonparametric statistics is to make inferences about unknown quantities without resorting to simple parametric reductions of the problem.

4.1 Introduction to nonparametric statistics

Example (Parametric approach)
Suppose $Y \sim F$ and we wish to estimate $\mathbb{E}[Y]$ or $\mathbb{P}(Y > 1)$; the approach taken by parametric statistic is to assume F to belong to a family of distributions

$$\mathcal{F} = \left\{F(\cdot, \vartheta), \vartheta \in \Theta \subseteq \mathbb{R}^d\right\},$$

which can be described by a finite number of parameters, $d < \infty$. Inference about the quantities we were originally interested in ($\mathbb{E}[Y]$ or $\mathbb{P}(Y > 1)$) are carried out based on assuming $Y \sim F(\cdot; \widehat{\vartheta})$ for an estimated set of parameters $\widehat{\vartheta}$.

Another example is to assume a linear model

$$\mathbb{E}[Y|X] = \beta_0 + \beta_1 X,$$

and we use the estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ to carry out inference about $\mathbb{E}[Y|X]$.

Both of the aforementioned parametric approach rely on a reduction of the original problem. They assume that all uncertainty regarding F, or $\mathbb{E}[Y|X]$, can be reduced to just two unknown numbers (i.e. parameters).

The perspective of nonparametric statistics is to make as few assumptions as possible about the data:

- we allow $F(y)$ to be any function that satisfies the cumulative distribution function properties;
- we allow $\mathbb{E}[Y|X]$ to be any continuous function of X.

Obviously, this requires the development of a whole new set of tools, as instead of estimating parameters, we will be estimating functions – which are much more complex.

4.2 Estimating the CDF and functionals

Let $Y_1, Y_2, \ldots, Y_n \sim F$ be a random sample from Y with cumulative distribution function

$$F(y) = \mathbb{P}(Y \leq y).$$

Def. (Empirical CDF)

We define the empirical cumulative distribution function as the nonparametric estimator

$$\widehat{F}_n(y) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(-\infty,y]}(Y_i).$$

Remark. The above function is an average of random variables, $T(Y) = \mathbb{1}_{(-\infty,y]}(Y)$.

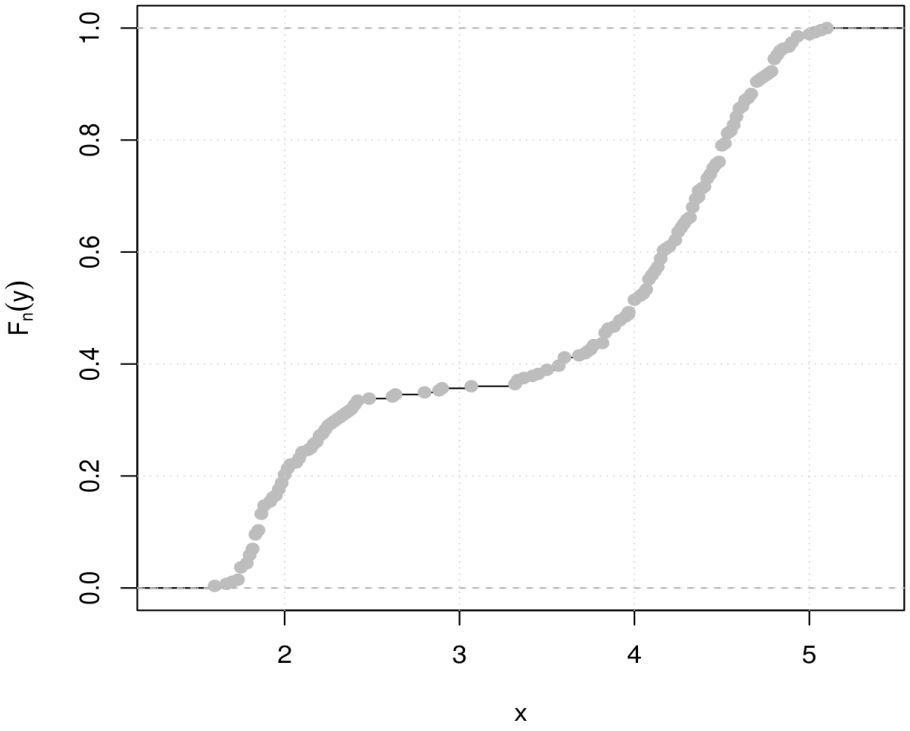

Figure 1: Estimated cumulative distribution function for the Old Faithful geyser dataset.

Prop. 1 (Properties of the ecdf)

Let $Y_1, Y_2, \ldots, Y_n \sim F$ and $\widehat{F}_n$ be the empirical cumulative distribution function, then for any fixed value y we have

1. $\mathbb{E}[\widehat{F}_n(y)] = F(y)$ *and* $\mathbb{V}[\widehat{F}_n(y)] = \frac{F(y)(1-F(y))}{n}$
2. $\widehat{F}_n(y) \xrightarrow{P} F(y)$
3. $\sqrt{n}\big(\widehat{F}_n(y) - F(y)\big) \xrightarrow{d} \mathcal{N}\big(0, F(y)(1-F(y))\big)$.

Proof.

Homework.

□

Theorem 1 (Glivenko-Cantelli)
If $\widehat{F}_n$ is the empirical cumulative distribution function, then

$$\sup_y |\widehat{F}_n(y) - F(y)| \xrightarrow{a.s.} 0.$$

Remark. This result is much more powerful, since it states a functional approximation result.

Theorem 2 (Dvoretzky-Kiefer-Wolfowitz)
For any $\varepsilon > 0$, we have

$$\mathbb{P}(\sup_y |\widehat{F}_n(y) - F(y)| > \varepsilon) \leq 2e^{-2n\varepsilon^2}.$$

Remark. The DKW inequality specifies the rate of convergence of the Glivenko-Cantelli theorem.

We discuss the notion of a confidence interval for a CDF F and compare its notions with those of confidence bands.

› A pointwise confidence interval finds a region $C(y)$ such that, for any F and fixed y,

$$\mathbb{P}\big(F(y) \in C(y)\big) \geq 1 - \alpha$$

› A different approach to inference is to find a confidence band $C(y)$ such that, for any CDF F

$$\mathbb{P}\big(F(y) \in C(y), \forall y\big) \geq 1 - \alpha$$

Prop. 2 (Confidence band)
For any distribution function F and all n,

$$\mathbb{P}\big(L_n(y) \leq F(y) \leq U_n(y), \forall y\big) \geq 1 - \alpha,$$

where

$$L_n(y) = \max\left\{\widehat{F}_n(y) - \sqrt{\frac{1}{2n}\log(2/\alpha)}, 0\right\}$$

$$U_n(y) = \min\left\{\widehat{F}_n(y) + \sqrt{\frac{1}{2n}\log(2/\alpha)}, 1\right\}$$

Proof.
Homework.

□

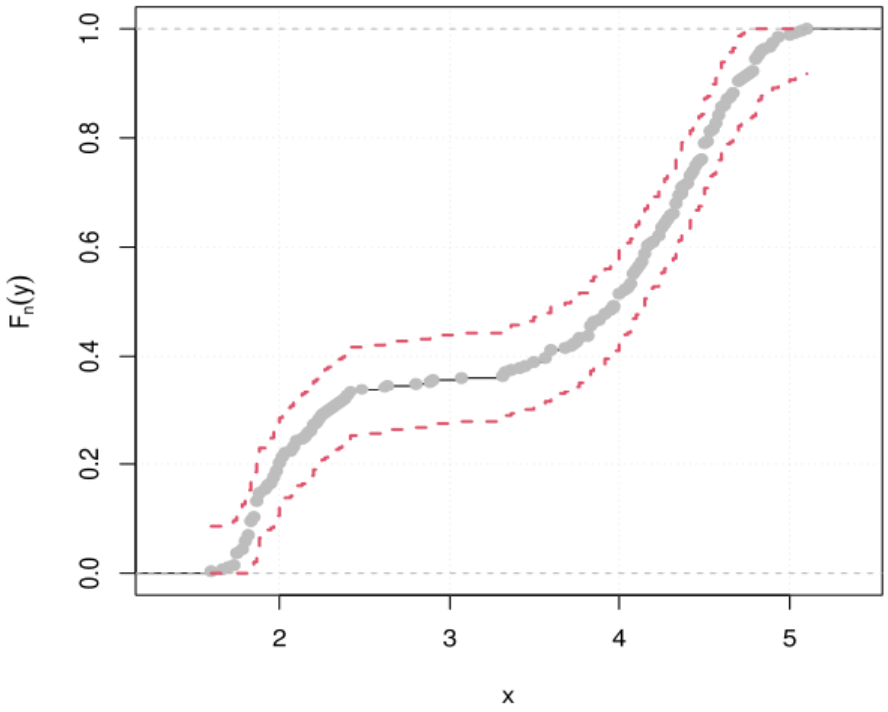

Figure 2: Nonparametric confidence band for the function $\widehat{F}_n$.

This important result now shows that empirical cumulative distribution function is a maximum likelihood estimator. Assuming that

$$f_n(y) = \sum_{i=1}^{n} p_i \mathbb{1}_{\{y\}}(Y_i),$$

then the nonparametric or empirical likelihood for $(p_1, \ldots, p_n)$ is

$$L(p_1, p_2, \ldots, p_n) = \prod_{i=1}^{n} f_n(Y_i) = \prod_{i=1}^{n} p_i.$$

Prop. 3 (Geometric mean is bounded by the arithmetic mean)
It holds that

$$\left(\prod_{i=1}^{n} p_i \right)^{1/n} \leq \frac{1}{n} \sum_{i=1}^{n} p_i = \frac{1}{n},$$

and the equality holds $\iff p_1 = \ldots = p_n = 1/n$.

Theorem 3 (Nonparametric maximum likelihood)
The empirical cumulative distribution function maximizes the empirical likelihood

$$L(p_1, \dots, p_n) = \prod_{i=1}^{n} p_i.$$

Proof.
Using Prop. 3 and setting $\widehat{p}_i = 1/n$ to get the maximum, we have that

$$L(p_1, \dots, p_n) \leq L(\widehat{p}_1, \dots, \widehat{p}_n),$$

and the empirical cumulative distribution function $\widehat{F}_n$ corresponds to the cumulative distribution function of the density

$$\widehat{f}_n(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{y_i\}}(y).$$

Hence, it maximizes the empirical likelihood.

□

4.3 Statistical functionals

Why do we put so much emphasis on estimating F? The reason is that $\widehat{F}_n$ will play the same role in nonparametric estimation played in parametric estimation from the MLE $\widehat{\vartheta}$. In general we are interested in transformations of $\widehat{F}_n$, but nonparametric statistics does not provide such estimates.

Def. (Statistical functional)
A statistical functional $T(F)$ is any function of F.

Example (Functionals)
Some functionals can be expressed as functionals

- *Mean*: $T(F) = \int y \, \mathrm{d}F(y)$.
- *Mean*: $T(F) = \int y^2 \, \mathrm{d}F(y) - \left(\int y \, \mathrm{d}F(y) \right)^2$.
- *Quantile*: $T(F) = F^{-1}(p)$, where

$$F^{-1}(p) = \inf\{y : F(y) \geq p\}, \quad p \in (0, 1).$$

Estimation of quantities such as the above is based on the nonparametric mle of F,

$$\widehat{T}(F) = T(\widehat{F}_n).$$

Example
The estimated functionals in the above example correspond to $\overline{Y}$, $\widehat{\sigma}^2$ and the sample quantile, respectively.

Is the plug-in estimator a good estimator in all cases? Not always: consider the estimation of a density, where the functional is

$$T(F) = \frac{\partial}{\partial y} F(y),$$

then we would like to characterize the convergence of $T(\widehat{F}_n) \to T(F)$ using the fact that from theorem 1 we have $\widehat{F}_n \xrightarrow{\text{a.s.}} F$.

Def. (Gâteaux derivative)
The Gâteaux derivative of T at F in the direction G (G is a CDF) is defined by

$$\begin{aligned} L_F(T; G) &= \lim_{\varepsilon \to 0} \frac{T\big((1-\varepsilon)F + \varepsilon G\big) - T(F)}{\varepsilon} \\ &= \lim_{\varepsilon \to 0} \frac{T\big(F + \varepsilon D\big) - T(F)}{\varepsilon}, \end{aligned}$$

if we define $D = G - F$.

Remark. The Gâteaux derivative is a generalization of the concept of a directional derivative to the functional analysis setting.

Remark. From a statistical perspective, it represents the rate of change in a statistical functional upon a small amount (ε) of contamination by another distribution G (mixture of distributions). This has a long history in the robust inference literature.

Statisticians usually prefer to work with a particular Gâteaux derivative, which is a special case of the above definition when G places a point mass of 1 at the point y, i.e.

$$G_y(u) = \mathbb{1}_{(-\infty, y]}(u).$$

This yield the so-called influence function definition as a specialization of the Gâteaux derivative.

Def. (Influence function)
The influence function of T at F is defined by

$$L_F(y) = \lim_{\varepsilon \to 0} \frac{T\big((1-\varepsilon)F + \varepsilon G_y\big) - T(F)}{\varepsilon},$$

where $G_y(u) = \mathbb{1}_{(-\infty, y]}(u)$.

Def. (Empirical influence function)

The empirical influence function is the estimate of $L_F(y)$ using the empirical cumulative distribution function,

$$\widehat{L}_n(y) = L_{\widehat{F}_n}(y) = \lim_{\varepsilon \to 0} \frac{T\big((1-\varepsilon)\widehat{F}_n + \varepsilon G_y\big) - T(\widehat{F}_n)}{\varepsilon}.$$

Example (Influence function)

Consider the functional $T(F) = \mu = \int y \, \mathrm{d}F(y)$, then the perturbed functional is

$$T\big((1-\varepsilon)F + \varepsilon G_y\big) = (1-\varepsilon)\mu + \varepsilon y,$$

so that

$$L_F(y) = y - \mu.$$

An important consequence is that, for $T(F) = \mu$ we have $\mathbb{E}[L_F(Y)] = 0$. Estimating it using the empirical influence function we have $T(\widehat{F}_n) = \bar{y}$ and so

$$\widehat{L}_n(y) = y - \bar{y}.$$

Linear functionals are particularly easy to work with, since their linearity allows the simple characterization of the influence function.

Def. (Linear functional)

A functional T is a linear functional if it is defined as

$$T(F) = \int a(y) \, \mathrm{d}F(y),$$

for some function $a(y)$.

Remark. We have that the influence function has a very simple representation,

$$L_F(y) = a(y) - T(F)$$

$$\widehat{L}_n(y) = a(y) - T(\widehat{F}_n)$$

The Gâteaux derivative has many of the same properties as ordinary derivatives, in particular we have that:

Theorem 4 (Chain rule)

Suppose that a functional can be written as $T(F) = h\big(T_1(F), \ldots, T_m(F)\big)$, for some derivable function $h : \mathcal{F} \longrightarrow \mathbb{R}$, then

$$L_F(y) = \sum_{i=1}^{m} \frac{\partial h}{\partial t_i} L_i(y),$$

where $L_i(y)$ is the influence function of $T_i(F)$.

4.4 Functional delta method

In parametric statistics, we estimate ϑ by $\widehat{\vartheta}_n$ and we can use the delta method to obtain approximate distributional results for $g(\widehat{\vartheta}_n)$, for instance using a first-order representation with $g'(\vartheta) \neq 0$

$$\sqrt{n}\big(g(\widehat{\vartheta}_n) - g(\vartheta)\big) \xrightarrow{d} \mathcal{N}\big(0, \sigma^2 g'(\vartheta)^2\big).$$

In nonparametric statistics we can use the functional delta method to obtain distributional results for $T(\widehat{F}_n)$.

Lemma 1 ("Functional theorem of calculus")
Assume that $T(F)$ is a linear functional, then for any function G it holds that

$$T(G) = T(F) + \int L_F(y)\, \mathrm{d}G(y). \tag{1}$$

Corollary 1 (Expectation of a linear functional)
It follows from the above lemma that, by setting $G = F$,

$$\int L_F(y)\, \mathrm{d}F(y) = 0. \tag{2}$$

Remark. Equation (2) states that the influence function – which is a sort of "derivative" – behaves like the score function in parametric estimation. Indeed, we know from Bartlett's identities that

$$\mathbb{E}_\vartheta\Big[\frac{\partial}{\partial\vartheta}\log \ell(\vartheta; Y)\Big] = 0.$$

Suppose that our plug in estimate of $T(F)$ is $T(\widehat{F}_n)$, then this plug-in estimate in the linear case can be written using (1) and setting $G = \widehat{F}_n$ as

$$T(\widehat{F}_n) = T(F) + \underbrace{\frac{1}{n}\sum_{i=1}^{n} L_F(Y_i)}_{\Longrightarrow \text{CLT}}.$$

Proof.
Homework.

□

Lemma 2 (Central limit theorem)
Let $\tau^2 = \int L_F^2(y)\, \mathrm{d}F(y)$, then if $\tau^2 < \infty$ we have that

$$\sqrt{n}\frac{T(\widehat{F}_n) - T(F)}{\tau} \xrightarrow{d} \mathcal{N}(0, 1).$$

Lemma 3 (estimator of τ)
Let $\widehat{\tau}_n^2 = n^{-1}\sum_{i=1}^n \widehat{L}_n^2(Y_i)$, then we have that

$$\widehat{\tau}_n^2 \xrightarrow{P} \tau^2$$

$$\frac{\widehat{SE}_n}{SE} \xrightarrow{P} 1,$$

where $\widehat{SE}_n = \widehat{\tau}_n/\sqrt{n}$ and $SE = \sqrt{\mathbb{V}[T(\widehat{F}_n)]}$.

Theorem 5 (functional delta method for a linear functional)

$$\sqrt{n}\frac{T(\widehat{F}_n) - T(F)}{\widehat{\tau}_n} \xrightarrow{d} \mathcal{N}(0,1).$$

Remark. In the general case of a non-linear functional T, the above theorem still holds by writing

$$T(\widehat{F}_n) = T(F) + \frac{1}{n}\sum_{i=1}^n L_F(Y_i) + o_p(1).$$

We can try to extend Taylor's theorem to the functional case. We recall that

Theorem 6 (Taylor's theorem)
Suppose f is a real function on $[a,b]$, $f^{(K-1)}$ is continuous on $[a,b]$, $f^{(K)}(x)$ is bounded for $y \in (a,b)$ then for any two distinct points $y_0 < y_1$ in $[a,b]$ there exists a point y between $y_0 < y < y_1$ such that

$$f(y_1) = f(y_0) + \sum_{k=1}^{K-1}\frac{f^{(k)}(y_0)}{k!}(y_1 - y_0)^k + \frac{f^{(k)}(y)}{K!}(y_1 - y_0)^K.$$

The question is whether or not there exists a functional extension to the above Taylor theorem. The answer is that yes, there exists, under the condition that T is Hadamard differentiable at F. Define $D = G - F$ and the Gâteaux derivative $L_F(D)$ of T as

$$\lim_{\varepsilon \to 0}\left(\frac{T(F + \varepsilon D) - T(F)}{\varepsilon} - L_F(D)\right) = 0.$$

Def. (Hadamard differentiability)
A functional T is Hadamard differentiable at F if, for any sequence $\varepsilon_n \to 0$ and D_n satisfying $\sup_y |D_n(y) - D(y)| \to 0$, we have

$$\lim_{\varepsilon_n \to 0}\left(\frac{T(F + \varepsilon_n D_n) - T(F)}{\varepsilon_n} - L_F(D_n)\right) = 0.$$

Prop. 4 (Properties)
The following properties hold:

- *if T is Hadamard differentiable, then $T(\widehat{F}_n) \xrightarrow{P} T(F)$;*
- *if T is Hadamard differentiable at F, then*

$$\sqrt{n}\frac{T(\widehat{F}_n) - T(F)}{\tau} \xrightarrow{d} \mathcal{N}(0,1),$$

  where $\tau^2 = \int L_F(y)^2 \, \mathrm{d}F(y)$;
- *also,*

$$\sqrt{n}\frac{T(\widehat{F}_n) - T(F)}{\widehat{\tau}} \xrightarrow{d} \mathcal{N}(0,1),$$

  where $\widehat{\tau}^2 = n^{-1}\sum_{i=1}^n \widehat{L}_n^2(Y_i)$.

Thus, under appropriate regularity conditions, a $1-\alpha$ confidence interval for $\vartheta = T(F)$ is

$$T(\widehat{F}_n) \pm z_{\alpha/2} n^{-1/2} \widehat{\tau}.$$

Example (Sample mean)
Using $\widehat{\vartheta} = T(\widehat{F}_n) = \bar{y}$, then

$$\widehat{L}_n(y_i) = y_i - \bar{y},$$

from which we obtain

$$\widehat{\tau}_n^2 = n^{-1}\sum_{i=1}^n \widehat{L}_n^2(y_i) = n^{-1}\sum_{i=1}^n (y_i - \bar{y})^2,$$

and the asymptotic confidence interval for ϑ is

$$y \pm z_{1-\alpha/2} n^{-1/2} \widehat{\tau}_n.$$

This is the usual confidence interval when the variance estimator is the biased version.

4.4.1 Score function and influence function

Let $\ell(\vartheta|y)$ be the log-likelihood for $\vartheta \in \mathbb{R}$ and $U_\vartheta(y) = \frac{\partial}{\partial\vartheta}\ell(\vartheta|y)$ be the score function, then we have for the maximum likelihood estimator $\widehat{\vartheta}$

$$\mathbb{E}_\vartheta[U_\vartheta(Y)] = 0$$

$$\mathbb{V}_\vartheta[\widehat{\vartheta}] \approx \frac{1}{n\mathbb{V}_\vartheta[U_\vartheta(Y)]} = \frac{1}{n\mathbb{E}_\vartheta[U_\vartheta^2(Y)]}$$

Whereas for the influence function, if we define $\vartheta = T(F)$ then

$$\mathbb{E}_\vartheta[L_F(Y)] = 0$$

$$\mathbb{V}_\vartheta[\widehat{\vartheta}] \approx \frac{\mathbb{V}[L_F(Y)]}{n} = \frac{\mathbb{E}[L_F^2(Y)]}{n}$$

where the approximation is exact for a linear functional T. This relationship is satisfied by deriving he influence function of a parametric model,

$$L_\vartheta(y) = I(\vartheta)^{-1} U_\vartheta(y),$$

where $I(\vartheta)$ is the Fisher information.

4.4.2 Misspecified models

Note that $\mathbb{E}_\vartheta[U_\vartheta^2(Y)] = I(\vartheta)$ is correct only if the parametric model is not misspecified. In general, we can estimate $\mathbb{V}_\vartheta[\widehat{\vartheta}]$ using the nonparametric delta method,

$$\widehat{\mathbb{V}}[\widehat{\vartheta}] = \frac{1}{n}\sum_{i=1}^n \widehat{L}(y_i)^2,$$

and our estimate would be

$$\frac{\frac{1}{n}\sum_{i=1}^n \widehat{L}(y_i)^2}{n} = \frac{\frac{1}{n}\sum_{i=1}^n U_{\widehat{\vartheta}}(y_i)^2}{nI(\vartheta)^2},$$

to get the so-called sandwich estimator,

$$\mathbb{V}[\widehat{\vartheta}] = \frac{1}{n} I(\widehat{\vartheta})^{-1} \left[\frac{1}{n}\sum_{i=1}^n U_{\widehat{\vartheta}}(y_i)^2\right] I(\widehat{\vartheta})^{-1}.$$

Lecture 5: Simulation-based inference

2022-03-11

We described influence functions as a tool for assessing the standard error of statistical functionals and obtaining nonparametric confidence intervals. Today, we will introduce other ideas for estimating standard errors in a nonparametric way, as well as estimates of the bias.

5.1 Jackknife

We will see that the jackknife is built on essentially the same idea as the influence function, although the jackknife was proposed much earlier (1949).
Suppose we have an estimator T_n of $\vartheta = T(F)$ which can be computed from a sample $(Y_1, \dots, Y_n)$.

Def. (Jackknife estimator)
The Jackknife estimator of T is

$$\overline{T}_n = \frac{1}{n}\sum_{i=1}^{n} T_{(-i)},$$

where $T_{(i)}$ is T computed on $(Y_1, \dots, Y_{i-1}, Y_{i+1}, \dots, Y_n)$.

Remark. If T_n is unbiased, then

$$\mathbb{E}[\overline{T}_n] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[T_{(-i)}] = 0.$$

On the other hand, if T_n is asymptotically unbiased so that $\lim_{n\to\infty} \mathbb{E}[T_n] = \vartheta$ with

$$\mathbb{E}[T_n] = \vartheta + \frac{a}{n} + \frac{b}{n^2} + O(n^{-3}), \tag{3}$$

then

$$\mathbb{E}[T_{(-i)}] = \vartheta + \frac{a}{n-1} + \frac{b}{(n-1)^2} + O(n^{-3}),$$

and we can estimate the bias using a quantity defined as

$$b_{\text{jack}} = (n-1)(\overline{T}_n - T_n).$$

We have that

$$\mathbb{E}[b_{\text{jack}}] = \frac{a}{n} + \frac{(2n-1)b}{n^2(n-1)} + O(n^{-2}).$$

Def. (Bias-corrected jacknife estimator)
The bias-corrected jacknife estimator of T is defined as

$$T_{\text{jack}} = T_n - b_{\text{jack}}. \tag{4}$$

Remark. This is an unbiased estimate of ϑ up to *second order*, which is an improved from the first-order of (3),

$$\text{Bias}(T_{\text{jack}}) = -\frac{b}{n(n-1)} + O(n^{-2}).$$

Example (Estimator of variance)
We can do even more: consider the plug-in estimate of the variance ϑ, one possibility would be to use

$$T(F) = \vartheta = \int y^2 dF(y) - \left(\int y dF(y) \right)^2,$$

which has sample analogue

$$T(\widehat{F}_n) = n^{-1} \sum_{i=1}^{n} (Y_i - \overline{Y})^2,$$

and the expected value of $T(F_n)$

$$\mathbb{E}[T(\widehat{F}_n)] = \frac{n-1}{n} \vartheta.$$

Hence, the expected value of the jackknife estimate of bias is

$$\mathbb{E}[b_{\text{jack}}] = -\frac{\vartheta}{n} = \text{Bias}\big(T(\widehat{F}_n)\big),$$

and the bias-corrected estimate is

$$T_{\text{jack}} = s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \overline{Y})^2.$$

Another way to think about the jackknife is in terms of the pseudo-values,

$$\tilde{T}_i = nT_n - (n-1)T_{(-i)},$$

then we have that

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} \tilde{T}_i &= nT_n - (n-1)\overline{T}_n \\ &= T_n - b_{\text{jack}} \\ &= T_{\text{jack}}. \end{aligned}$$

The idea behind pseudo-values is that it allows us to think of the bias-corrected estimate as simply the mean of n "independent" data values. This allows us to study properties of T_{jack} in terms of central limit theorems, although we need some care since these random variables are not independent.

Remark. he pseudo-values $\tilde{T}_i$ are not, in general, independent, although note that for the special case of a linear statistic

$$T_n = \frac{1}{n} \sum_{i=1}^{n} a(Y_i) \implies \tilde{T}_i = a(Y_i) \implies \tilde{T}_j \perp\!\!\!\perp T_i, \quad i \neq j.$$

A reasonable idea, therefore, is to treat $\tilde{T}_i$ as linear approximations to i.i.d observations and approach inference for T_{jack} as we would the sample mean. Thus, we can calculate the sample variance of the

pseudo-values,

$$\tilde{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\tilde{T}_i - T_{\text{jack}})^2 \implies \widehat{\mathbb{V}}[T_n] = v_{\text{jack}} = \frac{\tilde{s}^2}{n}.$$

Def. (Jackknife variance estimator)
The jackknife variance estimator is defined as

$$\tilde{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\tilde{T}_i - T_{\text{jack}})^2 \implies \widehat{\mathbb{V}}[T_n] = v_{\text{jack}} = \frac{\tilde{s}^2}{n}. \tag{5}$$

Hence, we have

- unbiased estimate up to second order;
- an estimate of the variance of the estimator;
- confidence intervals which are similar to those obtained by the functional delta method (later), using

  $$T_{\text{jack}} \pm t_{1-\frac{\alpha}{2},n-1}\sqrt{v_{\text{jack}}}.$$

  The above CI is approximately correct since observations are neither Gaussian nor independent.

Consistency. Until now there are no distributional assumptions about Y_i, but we must investigate the conditions under which v_{jack} is a good estimator of $\mathbb{V}[T_n]$.
In general, if g is a continuously differentiable function, then v_{jack} is a consistent estimator of $g(\overline{Y})$,

$$\frac{v_{\text{jack}}}{\mathbb{V}[g(\overline{Y})]} \xrightarrow{P} 1.$$

In particular, v_{jack} can be shown to perform poorly when the estimator is not a smooth function of the data, for example the *median*. It can be shown (Efron, 1982) that the jackknife variance estimate is inconsistent for all F and all quantiles of order p, and for the median, $p = 1/2$ and

$$\frac{v_{\text{jack}}}{\mathbb{V}[T_n]} \xrightarrow{d} \left(\frac{1}{2}\xi^2\right)^2.$$

Influence function. There is a close connection between the jackknife and influence functions. In particular, the jackknife can be seen as a plug-in estimator, which calculates n estimates based on a perturbed version of the empirical distribution function and compares these altered estimates to the plug-in estimate in order to assess variability of the estimate.

Removing a single value, we obtain $T_{(-i)} = T\big(F_n^{(-i)}(y)\big)$, where

$$\begin{aligned} F_n^{(-i)} = \frac{1}{n-1}\sum_{j\neq i}^{n}\mathbb{1}_{(-\infty,y]}(Y_j) &= \frac{n}{n-1}\widehat{F}_n(y) - \frac{1}{n-1}\mathbb{1}_{(-\infty,y]}(Y_i) \\ &= \frac{n}{n-1}\widehat{F}_n(y) - \frac{1}{n-1}G_{Y_i}(y). \end{aligned} \tag{6}$$

And here, $G_{Y_i}(y)$ is the cumulative distribution function that puts unitary mass at $y = Y_i$. In fact, we can use this quantity to approximate the influence function, by setting $F = \widehat{F}_n$ and $\varepsilon = -\frac{1}{n-1}$,

$$L_F(Y_i) \approx \frac{\overbrace{T\left(\frac{n}{n-1}\widehat{F}_n - \frac{1}{n-1}G_{Y_i}\right)}^{T_{(-i)}} - \overbrace{T(\widehat{F}_n)}^{T_n}}{-1/(n-1)} = (n-1)(T_n - T_{(-i)}).$$

In a sense, then, the jackknife is a numerical approximation to the functional delta, and this justifies the alternative name of infinitesimal jackknife for the functional delta method.

Differences. There is an important difference, however, between the jackknife and the functional delta method: the delta method adds point mass to Y_i, while the jackknife in (6) takes point mass away.

Positive jackknife. Another take on the jackknife, then, is to compute n estimates $T(i)$ by adding an observation at Y_i instead of taking one away (i.e., $\varepsilon = 1/(n+1)$)). This method is called the positive jackknife, which however is not commonly used.

Delete-*d* jackknife. Another variation on the jackknife that has been proposed is called the delete-d jackknife, which leaves out d observations for each estimate. This can be an improvement for reducing dependence between the Jackknife estimates, and if d is appropriately chosen then the estimate of the variance is consistent for the median. However, it has the drawback that instead of calculating n leave-one-out estimates, we now have to calculate $\binom{n}{d}$ leave-d-out estimates - a much larger number, often bordering on the computationally infeasible

5.2 Bootstrap

The bootstrap was introduced by Efron (1979) as a computer-based method to estimate the variance and the distribution of an estimate and more generally of a statistic $T_n = T(Y_1, Y_2, \ldots, Y_n)$. In general, we can use it to construct confidence intervals.

The advantages of the bootstrap are multiple:

- Completely automatic.
- Requires no theoretical calculations.
- Not based on asymptotic results.
- Available no matter how complicated the statistic is.

Suppose that we want to calculate the variance of an estimator,

$$\mathbb{V}_F[T_n] = \int T_n^2 dF(y_1)\cdots dF(y_n) - \left(\int T_n dF(y_1)\cdots dF(y_n)\right)^2,$$

and the ideal bootstrap estimates $\mathbb{V}_F[T_n]$ with $\mathbb{V}_{\widehat{F}_n}[T_n]$ using a plug-in estimator of the variance.

Problem. A possible drawback is that $\mathbb{V}_{\widehat{F}_n}[T_n]$ might be difficult to compute. Indeed, the above integrals have to be performed over $\mathbb{R}^n$ and, by plugin, the sums become $\underbrace{\sum_{i=1}^{n} \cdots \sum_{i=1}^{n}}_{n \text{ times}}$, hence it is computationally $O(n^n)$.

Solution. The idea is to approximate it using a simulation, i.e. by using a subset of size B of the n^n terms of the summation.

Algorithm 1 Bootstrap

1: Sample $\boldsymbol{Y}_1^*, \dots, \boldsymbol{Y}_B^*$ with replacement from $\widehat{F}_n$, $Y_{b,i}^* \overset{\text{iid}}{\sim} \widehat{F}_n, \quad b = 1, \dots, B$.
2: Calculate the bootstrap replication $T^* = g(\boldsymbol{Y}_b^*)$ for $b = 1, \dots, B$
3: Estimate the variance using

$$\mathbb{V}[T_n] = \frac{1}{B-1} \sum_{b=1}^{B} (T_b^* - \overline{T}^*)^2,$$

with $\overline{T}^* = B^{-1} \sum_{b=1}^{B} T_b^*$.

By the law of large numbers, then

$$\mathbb{V}_{\text{boot}}[T_n] \xrightarrow{B \to \infty} \mathbb{V}_{\widehat{F}_n}[T_n].$$

We can also use it to estimate the bias, or any aspect of T_n. Indeed, we can write the bias of ϑ as

$$\text{bias}_{\text{boot}} = \overline{\vartheta}^* - \widehat{\vartheta} = \frac{1}{B} \sum_{B=1}^{B} \vartheta_b^* - \widehat{\vartheta},$$

or by setting $G = \mathbb{P}(T_n \leq t)$, then the bootstrap approximation to G is

$$\widehat{G}_n^*(t) = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}_{(-\infty, t]}(T_b^*)$$

Theorem 7 (Consistent estimator of G)
If $T_n = T(F)$ is Hadamard differentiable, then $\widehat{G}_n$ is a consistent estimator of G.

Proof.
Wasserman (2005)

□

Example (Failure of the bootstrap)
Let $Y_1, Y_2, \dots, Y_n$ be a random sample from F and that $\mathbb{E}[Y] = \mu$, $\mathbb{V}[Y] = 1$. Consider

$\vartheta = |\int y dF(y)|$ with plug-in estimator $\widehat{\vartheta}_n = |\overline{Y}|$.

If $\vartheta = 0$, then the bootstrap is not consistent for estimating the distribution of

$$E_n = \sqrt{n}(|\overline{Y}_n| - |\mu|) \xrightarrow[\mu=0]{d} |Z|, \quad Z \sim \mathcal{N}(0,1)$$

It can be shown however that

$$\left(\sqrt{n}(\overline{Y}_n - \mu), \sqrt{n}(\overline{Y}_n^* - \overline{Y})\right) \xrightarrow{d} (Z_1, Z_2),$$

where Z_1 and Z_2 are independent $\mathcal{N}(0,1)$ random variables. In practice,

$$E_n^* = \sqrt{n}(\overline{Y}^* - \overline{Y}) \dots$$

5.2.1 Confidence intervals

There are several ways of constructing confidence intervals, and the bootstrap allows us to do so for a general statistic.

Def. (Confidence interval)
A confidence interval of level $1-\alpha$ is a random interval which contains ϑ_0 with probability $1-\alpha$, i.e.

$$\mathbb{P}(\vartheta \in C) = 1 - \alpha.$$

Def. (Pivotal interval)
Let $\vartheta = T(F)$, $\widehat{\vartheta}_n = T(\widehat{F}_n)$, $R_n = \widehat{\vartheta}_n - \vartheta$ be the pivot and $H(r) = \mathbb{P}(R_n \leq r)$. Then, the pivotal interval is defined as the interval such that

$$\mathbb{P}(\widehat{\vartheta}_n - H^{-1}(1 - \alpha/2) \leq \vartheta \leq \widehat{\vartheta}_n - H^{-1}(\alpha/2)) = 1 - \alpha.$$

Remark. Since H is unknown, the bootstrap estimate of H is

$$\widehat{H}(r) = \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}_{(-\infty, r]}(R_b^*), \quad \text{where } R_b^* = \widehat{\vartheta}_b^* - \widehat{\vartheta}_n,$$

from which we obtain

$$1 - \alpha = \mathbb{P}(\widehat{\vartheta}_n - H^{-1}(1 - \alpha/2) \leq \vartheta \leq \widehat{\vartheta}_n - H^{-1}(\alpha/2))$$
$$\overset{\text{boot}}{=} \dots$$

Theorem 8 (Pivot interval)
If $T(F)$ is Hadamard differentiable, then

$$C_n = \big(2\widehat{\vartheta}_n - \vartheta^*_{1-\frac{\alpha}{2}}, 2\widehat{\vartheta}_n - \vartheta^*_{\frac{\alpha}{2}}\big)$$

is such that

$$\mathbb{P}\big(T(F) \in C_n\big) \xrightarrow{n\to\infty} 1-\alpha.$$

Def. (Studentized pivotal interval)
Let $Z_n = (\widehat{\vartheta}_n - \vartheta)/\widehat{\text{se}}_{\text{boot}}$, and

$$Z^*_b = \frac{\widehat{\vartheta}^*_b - \widehat{\vartheta}_n}{\widehat{\text{se}}^*_b},$$

where $\widehat{\text{se}}^*_b$ is an estimate of the standard error of $\widehat{\vartheta}^*_b$. The studentized pivotal interval is defined as

$$C_n = \big(\widehat{\vartheta}_n - z^*_{1-\frac{\alpha}{2}}\widehat{\text{se}}_{\text{boot}}, \widehat{\vartheta}_n + z^*_{1-\frac{\alpha}{2}}\widehat{\text{se}}_{\text{boot}}\big).$$

Remark. In order to compute $\widehat{\text{se}}^*_{\text{boot}}$ we need to nest the bootstraps.

Def. (Percentile interval)
The percentile interval is defined as the interval

$$C = \big(G^{-1}(\alpha/2), G^{-1}(1-\alpha/2)\big),$$

which is estimated using the α quantiles of ϑ^*_b,

$$C_n = \big(t^*_{\frac{\alpha}{2}}, t^*_{1-\frac{\alpha}{2}}\big).$$

Suppose that we have a one-sided interval of the form $[\widehat{\vartheta}_\alpha, \infty)$, then we would like our confidence interval to be such

$$\mathbb{P}(\vartheta > \widehat{\vartheta}_\alpha) = 1-\alpha \iff \mathbb{P}(\vartheta \leq \widehat{\vartheta}_\alpha) = \alpha.$$

› If $\mathbb{P}(\vartheta \leq \widehat{\vartheta}_\alpha) = \alpha + O(n^{-1/2})$, then the interval is first-order accurate.

› If $\mathbb{P}(\vartheta \leq \widehat{\vartheta}_\alpha) = \alpha + O(n^{-1})$, then the interval is second-order accurate.

Prop. 5 (Accuracy)
For the approximated confidence intervals, we have that

› *Normal interval:* $\mathbb{P}(\vartheta \leq \widehat{\vartheta}_\alpha) = \alpha + O(n^{-1/2})$

› *Pivotal interval:* $\mathbb{P}(\vartheta \leq \widehat{\vartheta}_\alpha) = \alpha + O(n^{-1/2})$

› *Studentized interval:* $\mathbb{P}(\vartheta \leq \widehat{\vartheta}_\alpha) = \alpha + O(n^{-1})$

› *Percentile interval:* $\mathbb{P}(\vartheta \leq \widehat{\vartheta}_\alpha) = \alpha + O(n^{-1/2})$

Remark. The percentile interval has more justifications in terms of invariance with respect to reparametrization, hence a popular extension is the bias-corrected and accelerated percentile interval.

Lecture 6: Nonparametric density estimation

2022-03-16

Suppose we observe $Y_1, Y_2, \ldots, Y_n$ where F has density $f = dF/dy$ with respect to Lebesgue measure on the real line. Some observations about the estimation of f is that, although $\widehat{F}_n(y)$ is often a good estimator of F, $d\widehat{F}_n(y)/dy$ is usually not a good estimator of f. The estimator $d\widehat{F}(n)/dy$ can be represented simply as the empirical histogram using a set of contiguous intervals B_k, $k = 1, \ldots, K$.

6.1 Kernel density estimator

Def. (Histogram)
Formally, we can define the histogram as the estimator

$$\widehat{f}_n(y) = \sum_{k=1}^{K} \frac{1}{w_k} \mathbb{1}_{B_k}(y) \widehat{p}_k,$$

where $w_k = \operatorname{diam}(B_k)$ and $\widehat{p}_k = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{B_k}(Y_i)$.

Remark. The estimator is not continuous and depends heavily on the choice of bins. In general, it is a rudimentary estimator which gives basic information about the true population.

In general, we can define an estimator of f as the limit for $\operatorname{diam}(B_k) \to 0$.

Def. (Naive estimator)
Given a sample of n observations $Y_1, Y_2, \ldots, Y_n$, the naive estimator for f is

$$\widehat{f}_n(y) = \frac{\widehat{F}_n(y+h) - \widehat{F}_n(y-h)}{2h}$$

Remark. The estimator is simply the plug-in estimator using the definition of a differentiable F,

$$f(y) = \frac{\partial}{\partial y} F(y) = \lim_{h \to 0} \frac{F(y+h) - F(y-h)}{2h}$$

Remark. Starting from what we defined before, we can rewrite it as

$$\begin{aligned}
\widehat{f}_n(y) &= \frac{\sum_{i=1}^{n} \mathbb{1}_{(-\infty, y+h}(Y_i) - \sum_{i=1}^{n} \mathbb{1}_{(-\infty, y-h)}(Y_i)}{2nh} \\
&= \frac{1}{nh} \sum_{i=1}^{n} \frac{1}{2} \mathbb{1}_{(-1,1]} \left(\frac{y - Y_i}{h} \right) \\
&= \frac{1}{nh} \sum_{i=1}^{n} K \left(\frac{y - Y_i}{h} \right),
\end{aligned}$$

where $K(u) = \frac{\mathbb{1}_{(-1,1]}(u)}{2}$ is the density function of $U \sim \operatorname{Unif}(-1, 1)$. Hence, the naive estimator is the average of density functions scaled by some width h.

Properties. In general, we can state the following properties about the naive estimator:

- $\widehat{f}$ is more flexible than the simple histogram.
- $\widehat{f}$ depends heavily on the value of the bandwidth h, although is still rough for large h.
- It is a density for any value of h.

The idea of the general kernel density estimator is to place a small density around each observation.

Def. (Kernel density estimator)

The kernel density estimator of f is

$$\widehat{f}_n(y) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{y - Y_i}{h}\right), \tag{7}$$

where K is a symmetric density centered in zero.

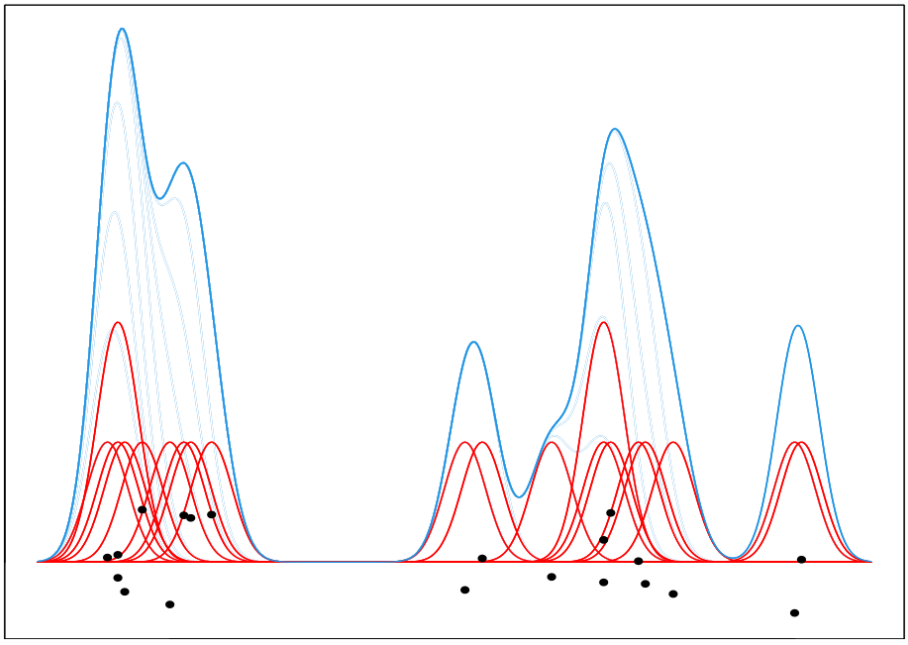

Figure 3: General idea for the kernel density estimator of the density of f.

Kernel	$K(u)$
Uniform (Rectangular)	$\frac{1}{2} I_{[-1,1]}(\lvert u\rvert)$
Triangle	$(1-\lvert u\rvert) I_{[-1,1]}(\lvert u\rvert)$
Triweight	$\frac{35}{32}(1-u^2)^3 I_{[-1,1]}(\lvert u\rvert)$
Quartic (Biweight)	$\frac{15}{16}(1-u^2)^2 I_{[-1,1]}(\lvert u\rvert)$
Gaussian	$\frac{1}{\sqrt{2\pi}} e^{-u^2/2}$
Epanechnikov	$\frac{3}{4}(1-u^2) I_{[-1,1]}(\lvert u\rvert)$
Cosine	$\frac{\pi}{4} \cos\left(\frac{\pi}{2} u\right) I_{[-1,1]}(\lvert u\rvert)$

Figure 4: Most common kernels in kernel density estimates for f, see `density` function in `R`. Note that most kernels are bounded density, whereas the Gaussian is unbounded.

Remark. The shape of the kernel doesn't really affect the asymptotic properties of $\widehat{f}$. In smoothing in general there is a fundamental trade-off between the bias and variance of the estimate $\widehat{f}_n$, and this trade-off is governed by the smoothing parameter h.

Def. (Mean-squared error)
We define the mean-squared error of an estimator $\widehat{f}_n$ as the risk

$$\mathrm{MSE}(\widehat{f}_n) = \mathbb{E}[L(\widehat{f}_n, f(y))] = \mathbb{E}\left[\left(\widehat{f}_n(y) - f(y)\right)^2\right].$$

Bias-variance. In general, we can highlight the mean-squared error as a combination of bias and variance of the estimator, since

$$\begin{aligned}\mathrm{MSE}(\widehat{f}_n) &= \Big(\mathbb{E}[\widehat{f}_n(Y)] - f(Y)\Big)^2 + \mathbb{V}[\widehat{f}_n(Y)] \\ &= \mathrm{Bias}_Y^2\big(\widehat{f}_n(Y)\big) + \mathbb{V}_Y[\widehat{f}_n(Y)].\end{aligned}$$

Most of the time we instead consider averaged versions of the mean-squared error.

Def. (Integrated mean-squared error)
The mean-integrated -squared error (MISE) is

$$\mathrm{MISE}(\widehat{f}_n, f) = \int R(\widehat{f}_n(y), f(y)) dy. \tag{8}$$

Def. (Average mean-squared error)
The average mean-squared error (AMSE) is

$$\mathrm{AMSE}(\widehat{f}_n, f) = \frac{1}{n}\sum_{i=1}^{n} R(\widehat{f}_n(y_i), f(y_i)). \tag{9}$$

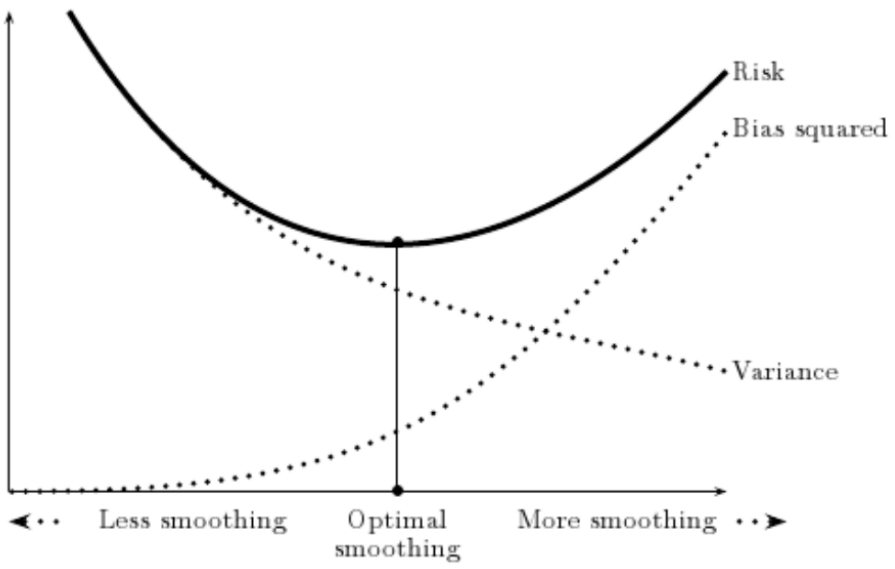

Figure 5: Example of bias-variance trade-off in kernel density estimation.

We could use other loss functions apart from the L_2 loss, such as:

› *L^p loss*, using
$$\left\{\int |\widehat{f}_n(y) - f(y)|^p dy\right\}^{1/p},$$
for which L_2 yields results that are easier to achieve.

› *Kullback-Leibler loss*, used especially in the machine learning community,
$$L(\widehat{f}_n, f) = \int f(y) \log\left(\frac{f(y)}{\widehat{f}_n(y)}\right) dy,$$
which is sensitive to the tails of the distribution (Hall, 1987).

6.1.1 Bias of the estimator

We have that the estimator in (7) has bias

$$\begin{aligned}
\mathbb{E}[\widehat{f}_n(y)] &= \frac{1}{nh}\sum_{i=1}^{n}\mathbb{E}\left[K\Big(\frac{y - Y_i}{h}\Big)\right] \\
&= \mathbb{E}\left[\frac{1}{h}K\Big(\frac{y - Y}{h}\Big)\right] \\
&= \int K_h(y-u)f(u)du \\
&= K_h * f(y),
\end{aligned}$$

which is the convolution between $K_h(u) = K(u/h)/h$ and f. In particular, we have that

$$\text{Bias}\big(\widehat{f}_n(y)\big) = K_h * f(y) - f(y).$$

6.1.2 Variance of the estimator

As for the variance, we have that in the case the data is i.i.d,

$$\begin{aligned}\mathbb{V}[\widehat{f}_n(y)] &= \frac{1}{n}\,\mathbb{V}\left[\frac{1}{h}K\Big(\frac{y-Y_i}{h}\Big)\right] \\ &= \frac{1}{n}\Big(\mathbb{E}[K_h(h-Y_i)^2] - \mathbb{E}[K_h(h-Y_i) - f(y)]^2\Big) \\ &= \frac{1}{n}\Big(K_h^2 * f(y) - [K_h * f(y)]^2\Big)\end{aligned}$$

6.1.3 Mean-squared error of the estimator

In general, we can conclude that

$$\mathrm{MSE}(\widehat{f}_n(y)) = \frac{1}{n}\Big(K_h^2 * f(y) - [K_h * f(y)]^2\Big) + \left(K_h * f(y) - f(y)\right)^2,$$

hence the bias does not tend to zero as as n increases. The only way to do so if we work on the bandwidth h of the kernel, for instance by choosing the kernel K_h such that

$$K_h * f(y) \xrightarrow{n\to 0} f(y),$$

and this can be done if both $n \to \infty$ and $h \to 0$. We will consider the mean integrated squared error (**??**) to study the asymptotic properties of the kernel density estimator.

Suppose that we are under the following assumptions for the unknown function f and the kernel K:

1. Conditions on f:
   - f is three-times differentiable with $|f^{(j)}| \leq C$ for $j = 0, \ldots, 3$.
2. Conditions on K:
   - K is positive and symmetric.
   - $\int K(u)du = 1$ and $\int uK(u)du$
   - $\int u^2K(u)du < \infty$
   - $\int |u|^3K(u)du < \infty$
   - $\int K^2(u)du < \infty$
3. Conditions on h
   - $h \xrightarrow{n\to\infty} 0$ and $nh \xrightarrow{n\to\infty} \infty$.

Using a Taylor expansion of order 2 of f around y, we find

$$\begin{aligned}\mathbb{E}[\widehat{f}_n(y)] &= \int K(y)[f(y) - huf'(y) + \frac{(hu)^2}{2}f''(y) + o\big((hu)^2\big)]du \\ &= f(y)\underbrace{\int K(u)du}_{=1} - hf'(y)\underbrace{\int uK(u)du}_{=0} + \frac{h^2}{2}f''(y)\underbrace{\int u^2K(u)du}_{=1} + o(h^2) \\ &= f(y) + \frac{h^2}{2}f''(y)\mu_{K,2} + o(h^2),\end{aligned}$$

and it is clear that as $h \to 0$ the asymptotic bias disappears. In general, the asymptotic bias depends on $f''(y)$, hence it is more pronounced in the peaks and valleys of f.

Using a Taylor expansion of order 1 for $\mathbb{E}[K_h^2(y - Y_i)]$ allows us to write that

$$\begin{aligned}\mathbb{E}[K_h^2(y - Y_i)] &= h^{-1}\int K^2(u)f(y - hu)du \\ &= k^{-1}\int K^2(y)[f(y) - huf'(y) + o(h, u)]du \\ &= h^{-1}\dots \\ &= \frac{1}{h}f(y)\int K^2(u)du + o(1),\end{aligned}$$

hence

$$\begin{aligned}\mathbb{V}[\widehat{f}_n(y)] &= \frac{1}{nh}f(y)\int K^2(u)du + O(n^{-1}) \\ &= \frac{1}{nh}f(y)\int R^2(u)du + O(n^{-1})\end{aligned}$$

Remark. The variance increases depending on $R(K) = \int K^2(u)du$, and this is the only main contribution given by the kernel. In general, we can find bandwidths $h_1, h_2, \dots$ such that different kernels $K_1, K_2, \dots$ yield practically the same result in terms of mean-squared error.

Combining the two results above yields

$$\mathrm{AMSE}\big(\widehat{f}_n(y), h\big) = \underbrace{\frac{h^4}{4}\mu_{K,2}^2 f''(y)^2}_{\mathrm{Bias}^2} + \underbrace{\frac{R(K)}{nh}f(y)}_{\mathrm{Variance}}, \tag{10}$$

and we observe that the estimator is consistent if $nh \xrightarrow{n\to\infty} 0$. If we integrate the mean-squared error we obtain

$$\begin{aligned}\mathrm{MISE}(\widehat{f}_n, h) &= \int \mathrm{MSE}(\widehat{f}_n(y))dy \\ &= \dots \\ &= \mathrm{AMSE}(\widehat{f}_n, h) + O(1/n) + o(h^4).\end{aligned}$$

6.1.4 Optimal global bandwidth

Although we have an expression for the asymptotic mean-squared error, we are interested in the optimal bandwidth h_{opt} for a finite value of n. From (10), we can focus on the leading terms to write

$$h_{\text{opt}} \approx \operatorname*{argmin}_h \text{AMISE}(\widehat{f}_n, h) = \left(\frac{R(K)}{\mu_{K,2}^2 R(f'')} \right)^{1/5} n^{-1/5},$$

although this depends in practice by f'' which is unknown. We can replace it by an estimate based on the sample using some empirical rules.

Normal reference rule Consider f to be a Gaussian density, then

$$R(f'') = \int f''(y)^2 dy = \frac{3}{8\sigma^5\sqrt{\pi}},$$

hence we can estimate $\widehat{\sigma} = \min\{s, r\}$, where s is the sample estimate of the standard error and

$$r = \frac{\widehat{F}_n^{-1}(0.75) - \widehat{F}_n^{-1}(0.25)}{\Phi^{-1}(0.75) - \Phi^{-1}(0.25)} \approx \sigma \quad \text{if } f = \mathcal{N}(\cdot|\mu, \sigma^2).$$

Another reference rule could be using the plug-in bandwidth using an estimator

$$\widehat{R}(f'') = \int \widehat{f}_n''(y)^2 dy,$$

where $\widehat{f}_n''$ is estimated based on the kernel.

Another estimator is based on the leave-one-out cross-validation procedure. Starting from MISE, we have

$$\begin{aligned} \text{MISE}(\widehat{f}_n, h) &= \int \text{MSE}(y; h) dy \\ &= \mathbb{E} \int \widehat{f}_n(y)^2 dy - 2\,\mathbb{E} \int \widehat{f}_n f(y) dy + \int f^2(y) dy, \end{aligned}$$

which can be minimized only by considering the terms that depend on $\widehat{f}_n$,

$$\mathbb{E} \int \widehat{f}_n(y)^2 dy - 2\,\mathbb{E} \int \widehat{f}_n(y) f(y) dy,$$

and the first item can be estimated using the unbiased estimate $\int \widehat{f}_n(y)^2 dy$. The second term is more complicated, and can be estimated using the cross-validation procedure to obtain

$$\int \widehat{f}_n(y) f(y) dy \approx \frac{1}{n} \sum_{i=1}^n \widehat{f}_n^{(-i)}(Y_i),$$

where $\widehat{f}_n^{(-i)}(y)$ is the kernel density estimator with the i^{th} observation removed. We have that

$$\mathbb{E}[\widehat{f}_n^{(-i)}(Y_i)] = \mathbb{E}\Big[\int \widehat{f}_n(y) f(u) du \Big],$$

hence

Theorem 9 (Cross-validation criterion)
We have that

$$CV(\widehat{f}_n, h) = \int \widehat{f}_n(y)^2 dy - \frac{2}{n} \sum_{i=1}^{n} \widehat{f}_n^{(-i)}(Y_i)$$

is an unbiased estimator of

$$MISE(\widehat{f}_n, h) - \int f(u)^2 dy.$$

Def. (Data-driven bandwidth)
The data-driven bandwidth criterion for h is

$$\widehat{h}_{\text{opt}} = \underset{h}{\operatorname{argmin}}\, \text{CV}(\widehat{f}_n, h).$$

Remarks.

› Since the procedure is computationally intensive, we only use a grid of bandwidths.

› The quality of the resulting estimate is very variable.

› The optimal solution might not be unique.

Remark. The bandwidth rule usually 335G

Boundary problems
The results we have seen are only valid when the density f is continuous on its support. If instead f is the density of $Y \sim \text{Exp}(\vartheta)$, we have that f is supported on $[0, \infty)$ with finite right limit as $x \to 0$.

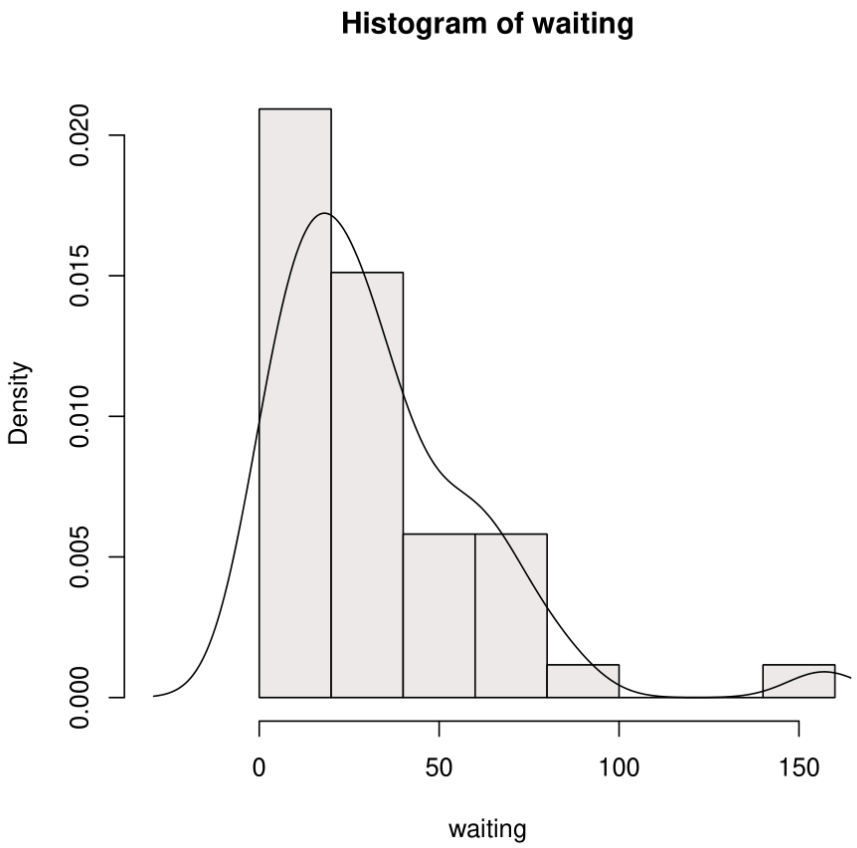

Figure 6: Example of a wrong kernel density estimator output when applied to positive data.

Remarks. Some naive solutions would be truncating $\widehat{f}_n$ for $y < 0$, but the density would not integrate to 1. Rescaling the truncation yields an insufficient estimator.

6.1.5 Kernel density estimator with finite support

Suppose that supp $f = [0, \infty]$ and that f is two-times continuously differentiable. If K is symmetric with support $[-1, 1]$, then consider $y = ph$ with $p < 1$. For $p \geq 1$ we are in the interior and the usual properties apply, whereas if $p < 1$ we have that

$$\mathbb{E}[\widehat{f}_n(y)] = a_0(p)f(y) - ha_1(p)f'(y) + o(h),$$

so we are not able to remove the first part of the bias. The kernel density estimator is not consistent at the boundary because

$$\mathbb{E}[\widehat{f}_n(0)] \approx a_0(p)f(0) = f(0)\int_{-1}^{0} K(u)du = \frac{1}{2}f(0).$$

Suppose that we have an estimate $\widehat{a}_0(p)$, then if we consider a rescaled version we have that

$$\mathbb{E}\left[\frac{\widehat{f}(y)}{a_0(p)}\right] = 1.$$

Def. (Boundary correction via renormalization)
The boundary correction of the kernel density estimator is the renormalized version of $\widehat{f}_n$,

$$\tilde{f}_n(y) = \frac{\widehat{f}_n(y)}{a_0(p)},$$

and $a_0(p) = 1$ for $p \geq 1$, hence it is valid in the interior.

Remark. Calculating the bias and variance (slides), we find that $\tilde{f}_n$ is consistent but the bias is of order $O(h)$ near the boundary.

Remark. The optimal MSE is $O(n^{-2/3})$ at the boundary and $O(n^{-4/5})$ elsewhere.

Consider instead the augmented dataset

$$Y_1, -Y_1, Y_2, -Y_2, \ldots, Y_n, -Y_n,$$

then we could construct a consistent kernel density estimator for $f(y)$ based on the augmented dataset.

Def. (Boundary correction by reflection)
The boundary correction by reflection is given by

$$\widehat{f}_n^R(y) = \begin{cases} 2\tilde{f}(y) & \text{if } y \geq 0 \\ 0 & \text{otherwise,} \end{cases}$$

where $\tilde{f}(y)$ is the simple kernel density estimator based on the augmented dataset, $Y_1, -Y_1, Y_2, -Y_2, \ldots, Y_n, -Y_n$.

Remark. The estimate corresponds to replacing the kernel with a modified kernel,

$$\tilde{K}_h(y - Y_i) = K_h(y - Y_i) + K_h(-y - Y_i) \implies \widehat{f}_n^R(y) = \widehat{f}_n(y) + \widehat{f}_n(-y).$$

Hence, we have explicit formulas for the bias and variance (slides).

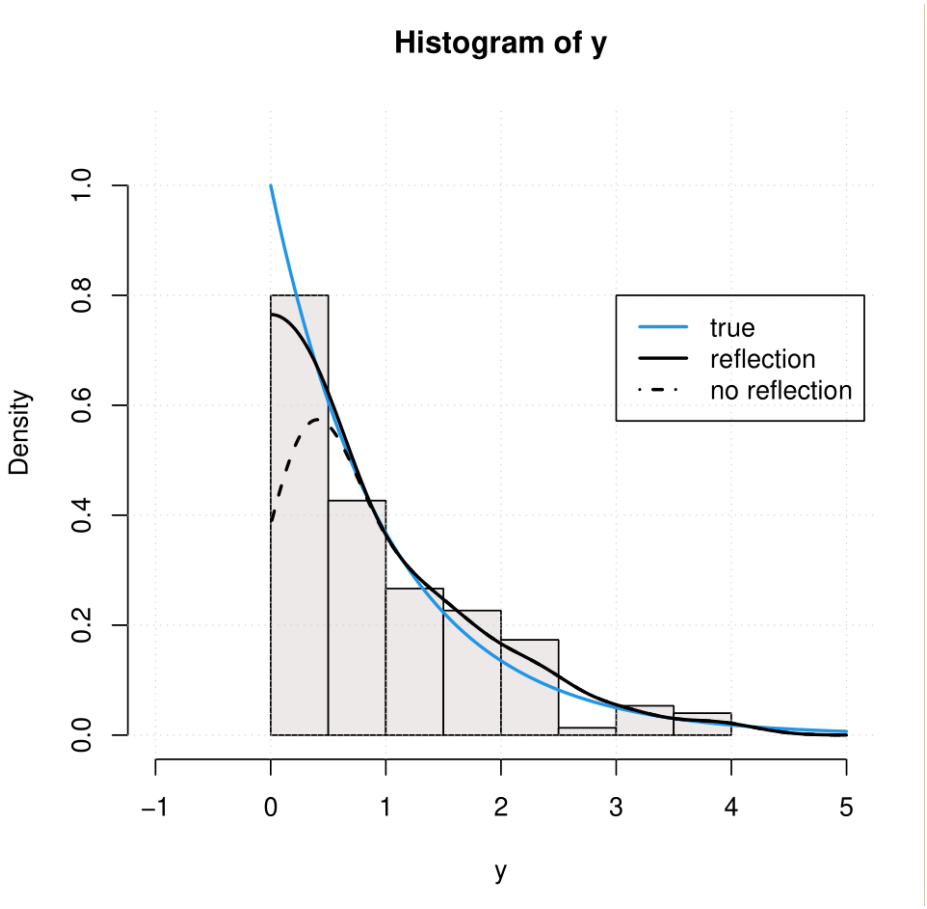

Figure 7: modifiedDensityEstimators

A third possibility is transforming the data and then moving them back to the original scale. Using the properties of transformation of the density, $Z = g(Y)$, we have that

$$f_Y(y) = f_Z\big(g(y)\big)g'(y).$$

Def. (Boundary correction by transformation)
The boundary correction by transformation uses the estimate

$$\widehat{f}_n(y) = \frac{g'(y)}{nh_Z}\sum_{i=1}^{n} K\left(\frac{g(y) - g(Y_i)}{h_Z}\right),$$

where h_Z is chosen based on the Z scale.

6.2 Local polynomials

Other methods of estimating the density can for example be based on local polynomials by using

$$\ell(f) = \sum_{i=1}^{n} \log f(y_i),$$

or by using the more general definition

$$\mathcal{L}(f) = \sum_{i=1}^{n} \log f(Y_i) - n\left(\int f(y)dy - 1\right),$$

where the second term is zero when f integrates to one. Including this term allows us to maximize over all non-negative functions f while imposing the constraint that f is a density.

Def. (Local likelihood)
The local likelihood is a weighted likelihood such that

$$\mathcal{L}_y(f) = \sum_{i=1}^{n} K\left(\frac{Y_i - y}{h}\right) \log f(Y_i) - n\left(\int K\left(\frac{u-y}{h}\right) f(u)du - 1\right).$$

Remark. Since f is unknown, we replace $\log f(u)$ by an approximation

$$\log f(u) = P_Y(\alpha, u; q) = \alpha_0 + \alpha_1(y-u) + \ldots + \frac{\alpha_q}{q!}(y-u)^q.$$

Note that

$$f(y) \in (0,1) \implies \log f(y) \in \mathbb{R}.$$

Def. (Local polynomial likelihood)
The local polynomial likelihood is defined as the approximated local likelihood using P_Y,

$$\mathcal{L}_y(f) = \sum_{i=1}^{n} K\left(\frac{Y_i - y}{h}\right) P_Y(\alpha, u; q) - n\left(\int K\left(\frac{u-y}{h}\right) \exp\left\{P_Y(\alpha, u; q)\right\} du - 1\right).$$

Let $\widehat{\alpha} = \text{argmax}_\alpha \, \mathcal{L}_Y(\alpha)$, then the local likelihood density estimate is

$$\widehat{f}_n(y) = e^{\widehat{\alpha}_0(y)}.$$

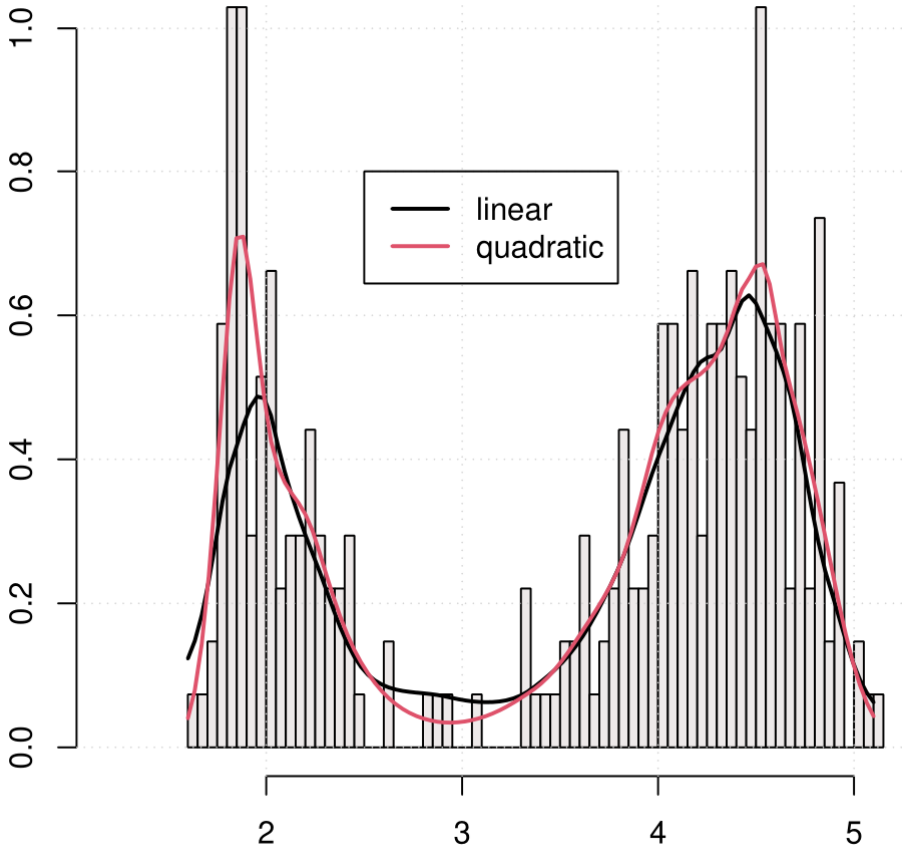

Figure 8: Local linear smoothing using linear and quadratic approximations.

6.3 Gaussian mixture models

Using a Gaussian mixture model we can approximate a wide range of densities simply by controlling the number of components K of the mixture.

Def. (Gaussian mixture model)
We define the gaussian mixture model for the density f of Y as

$$f(y;\vartheta) = \sum_{k=1}^{K} p_k \varphi(y;\mu_k,\sigma_k^2),$$

where $\varphi(\cdot;\mu,\sigma^2)$ is the Gaussian density with mean μ and variance σ^2.

Remark. The parameters are $\vartheta = (p_1, p_2, \ldots, p_k, \mu_1, \mu_2, \ldots, \mu_k, \sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2)$.

Estimation. We can find the maximum likelihood estimate of the parameters using the expectation-maximization algorithm (EM), which is useful for maximizing likelihoods in the presence of unobserved latent variables Z, in this case the indicators of the mixture.

Define the complete-data log-likelihood for (Y, Z) as

$$\mathcal{L}_{Y,Z}(\vartheta) = \sum_{i:Z_i=1}^{n} \log\big(p\varphi(Y_i;\mu_1,\sigma_1^2)\big) + \sum_{i:Z_i=0} (\tilde{\vartheta}) \log\big((1-p)\varphi(Y_i;\mu_2,\sigma_2^2)\big),$$

$$= \sum_{i=1}^{n} \log\big(p\varphi(Y_i;\mu_1,\sigma_1^2)\big) + (1-Z_i)\log\big((1-p)\varphi(Y_i;\mu_2,\sigma_2^2)\big),$$

Algorithm 2 EM algorithm for Gaussian mixture, $k = 2$

1: E-step: given a current value $\tilde{\vartheta}$, evaluate $Q(\vartheta, \tilde{\vartheta}) = \mathbb{E}_{\tilde{\vartheta}}[\mathcal{L}_{Y,Z}(\vartheta)|Y]$, i.e.

$$Q(\vartheta, \tilde{\vartheta}) = \sum_{i=1}^{n} \log\big(p\varphi(Y_i;\mu_1,\sigma_1^2)\big) + (1-w_i(\tilde{\vartheta}))\log\big((1-p)\varphi(Y_i;\mu_2,\sigma_2^2)\big),$$

where

$$w_i(\tilde{\vartheta}) = \mathbb{E}_{\tilde{\vartheta}}[Z_i|Y_i] = \frac{\cdots}{\cdots}.$$

2: M-step: maximize $Q(\vartheta, \tilde{\vartheta})$ to get the updated value of $\tilde{\vartheta}$. This is especially convenient for mixture models, since we have explicit forms for the updated parameters.

Remark. In practice we need to estimate the number of components K. In general, we can minimize the following criteria:

- AIC: $k^* = \text{argmin}_k \, \text{AIC}(k) = \text{argmin}_k \, -2\mathcal{L} + 2k$.
- BIC: $k^* = \text{argmin}_k \, \text{BIC}(k) = \text{argmin}_k \, -2\mathcal{L} + k \log n$.

Sensitivity Since the Gaussian is sensitive to outliers, we could replace it by a t distribution. We exchange more robustness by paying the price of not having explicit estimates for the parameters, hence a slower estimation routine.

Bandwidth. If $\sigma_k^2 = \sigma^2 > 0$ is fixed and $K \to n$, then $\widehat{p}_k \to \frac{1}{n}$ and the MLE converges to the kernel density estimate.

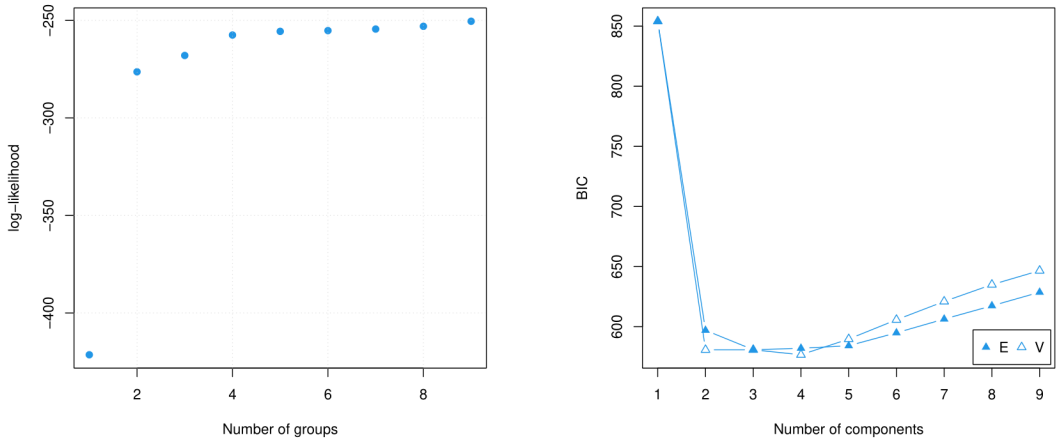

Figure 9: Mean-squared error for the Gaussian mixture model using equal variance "E" and different variances "V".

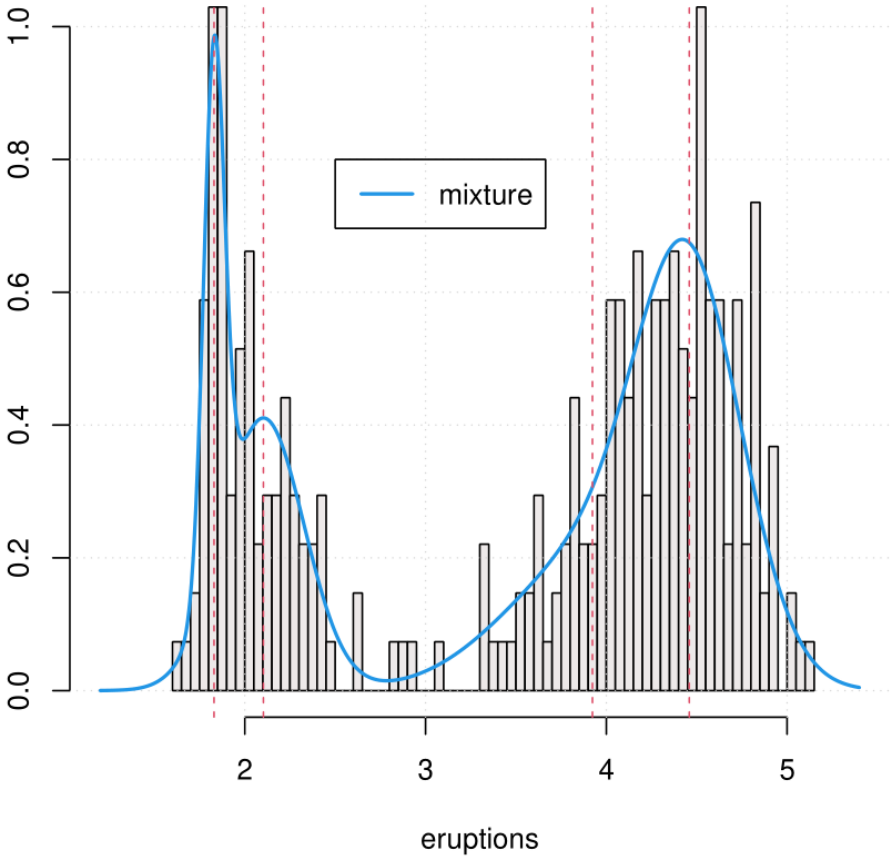

Figure 10: Example of an estimated density using four groups.

Extensions.

- Nearest neighbours
- Orthogonal series estimators
- Maximum penalized likelihood estimators
- Wavelet estimators

Extensions to higher dimension are straightforward generalizations of the univariate case.

Lecture 7: Nonparametric regression

2022-03-18

A common problem in applied statistics is that one has an dependent variable or outcome Y and various independent variable or covariates $X_1, X_2, \ldots, X_p$. The goal of this lecture is to provide some models for estimating a regression model when the response is nonparametrically dependent on $\boldsymbol{X}$.

1. Y and X can be random variables, i.e. $m(x_1, \ldots, x_p) = \mathbb{E}[Y|X_1, \ldots, X_p]$ is the regression function.

   Def. (Random design model)
   The random design model for the regression of Y on X is defined as

   $$Y = m(X) + \varepsilon,$$

   where $\mathbb{E}[Y|X = x] = m(x)$.

2. For some designed experiments we have complete control over the values of X, hence there is no reason to assume it to be a random variable, but rather we assume that $x_i = (x_{i1}, \ldots, x_{ip})$ are fixed design points.

   Def. (Fixed design model)
   The fixed design model instead assumes X to be known and fixed to the observed covariates x, i.e.

   $$Y = m(x) + \varepsilon,$$

   where $\mathbb{E}[Y] = m(x)$.

7.1 Linear regression

A common procedure for statistical modelling is to use a linear regression, i.e.

$$\mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}] = \sum_{i=1}^{p} \beta_j x_j,$$

under the assumption that $Y|\boldsymbol{X}$ follows a normal distribution. On he other hand, if $Y|\boldsymbol{X}$ comes from a general dispersion family (Pace and Salvan, 1997), we can specify

$$g\big(\mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}]\big) = \sum_{j=1}^{p} \beta_j x_j,$$

and $g : \operatorname{conv} \mathcal{Y} \longrightarrow \mathbb{R}$ is link function which is usually chosen for convenience reasons.

Remark 1. The parameters β usually have a direct scientific interpretation.

Remark 2. If the model is appropriate the estimates have many desirable statistical properties.

Problem. Linearity and additivity are two very strong assumptions, and sometimes we might want to relax them at the cost of less efficiency in the estimate, by specifying a more complex regression function

$$m(x) = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}], \quad x \in \mathbb{R}.$$

7.2 Smoothing

Suppose that we have a relationship such as that in Figure 11, then we would like to estimate a model of the form

$$Y_i = m(\boldsymbol{X}_i) + \varepsilon_i,$$

hence a simple estimator would be to take expectations under the empirical cumulative distribution function $\widehat{F}_n$, from which we have the plug-in estimate

$$\mathbb{E}_{\widehat{F}_n}[Y|\boldsymbol{X} = \boldsymbol{x}] = \frac{\sum_{i=1}^n Y_i \mathbb{1}_x(x_i)}{\sum_{j=1}^n \mathbb{1}_x(x_j)}. \tag{11}$$

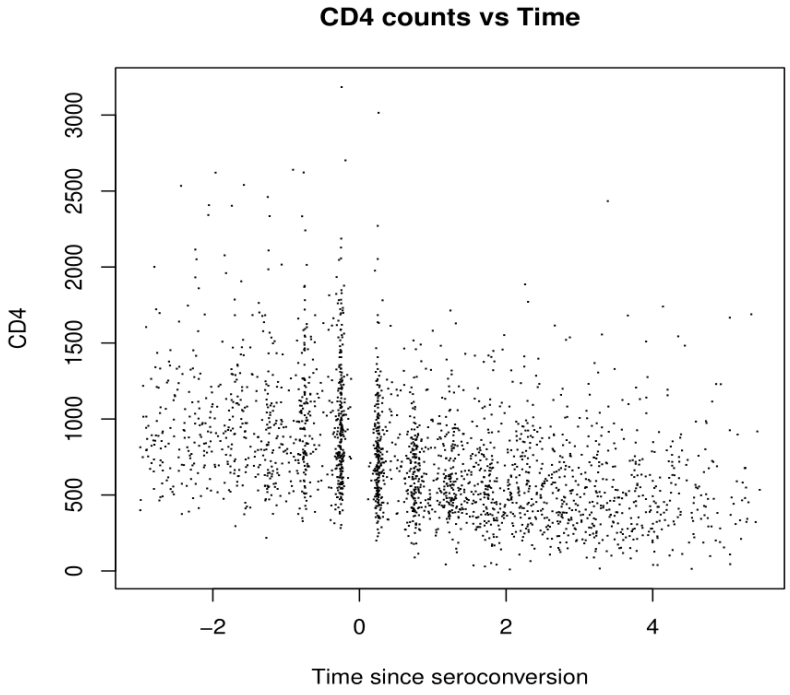

Figure 11: Scatterplot of a bivariate relationship between two continuous variables.

Some problems with the estimator in (11) is that the estimates are not smooth, and if we do not observe the particular value of x we are interested in, then we cannot calculate this expectation. Therefore, we need some approaches which allow extrapolating the nonparametric function $m(x)$ to unobserved values of $\boldsymbol{x}$, much like linear regression.

7.2.1 Parametric smoother

The parametric smoother is simply a linear regression using a polynomial basis $X = (1, x, x^2, \ldots, x^p)$ as covariates. We define a function defined by "few" parameters on the data and use least squares to find the most appropriate estimates for the parameters,

$$\widehat{m}(x) = X(X^\top X)^{-1} X^\top Y = \sum_{i=1}^n W_i Y_i,$$

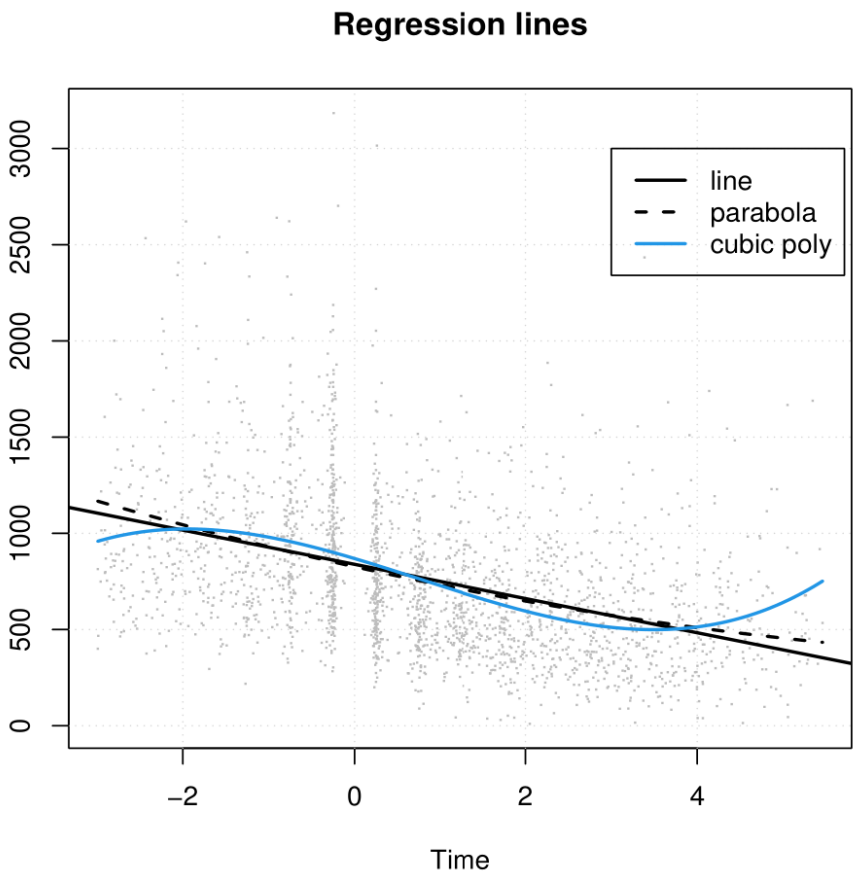

Figure 12: Example of parametric smoothing using polynomials of order 1, 2, 3. Note the problematic behaviour at the extremes of the covariate space.

7.2.2 Bin smoothers

A bin smoother, also known as a regressogram, mimics a categorical smoother by partitioning the predicted value into disjoint regions, $R_k = \{i : c_k \leq x_i < c_{k+1}\}$ for $k = 0, \ldots, K$, and then averaging the response Y in each region.

Def. (Regressogram estimator)

The regressogram estimator is defined as

$$\widehat{m}(x) = \frac{\sum_{i=1}^{n} y_i \mathbb{1}_{R_k}(x_i)}{\sum_{j=1}^{n} \mathbb{1}_{R_k}(x_j)}, \quad x \in R_k = \{i : c_k \leq x_i < c_{k+1}\}.$$

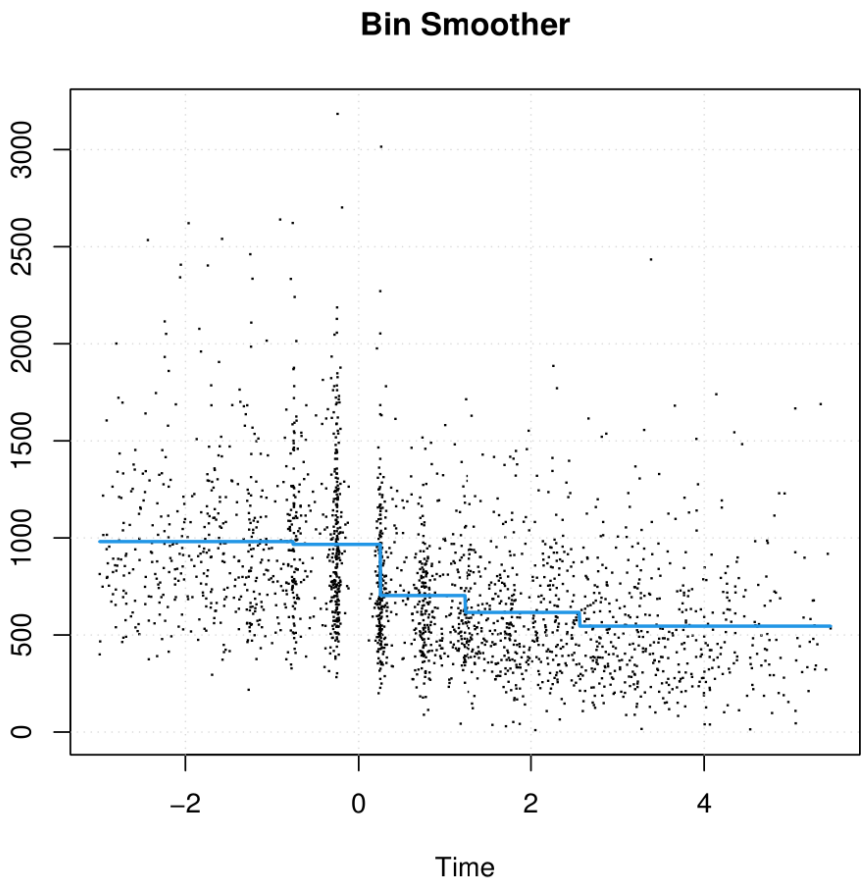

Figure 13: The bin smoother is a better-behaving smoother at the extremes of the covariate space, but shows a discontinuous behaviour in the interior.

7.2.3 Moving average

Since in general we expect the regression function $\widehat{m}(x)$ to be smooth, we can define a smoother with respect to a neighbourhood

$$N_\delta(x) = \{x_i : \|x - x_i\| \leq \delta\},$$

and estimate the δ-neighbour estimator as

$$\widehat{m}(x) = \frac{\sum_{i=1}^n y_i \mathbb{1}_{N_\delta(x)}(x_i)}{\sum_{j=1}^n \mathbb{1}_{N_\delta(x)(x_j)}}.$$

Problem. A δ-neighbour might be empty, since we might have that all x are at distance $\delta + \varepsilon$ from x_i. Therefore, we prefer a neighbourhood which groups a fixed number of observations, the k-neighbourhood.

Def. (Moving average estimator)
If $d_i = \|x - x_i\|$ and $N_k(x) = \{x_i : d_i \leq d_{(2k+1)}\}$, then the moving average estimator of $m(x)$ is

$$\widehat{m}(x) = \frac{\sum_{i=1}^{n} Y_i \mathbb{1}_{N_k(x)}(x_i)}{2k+1}.$$

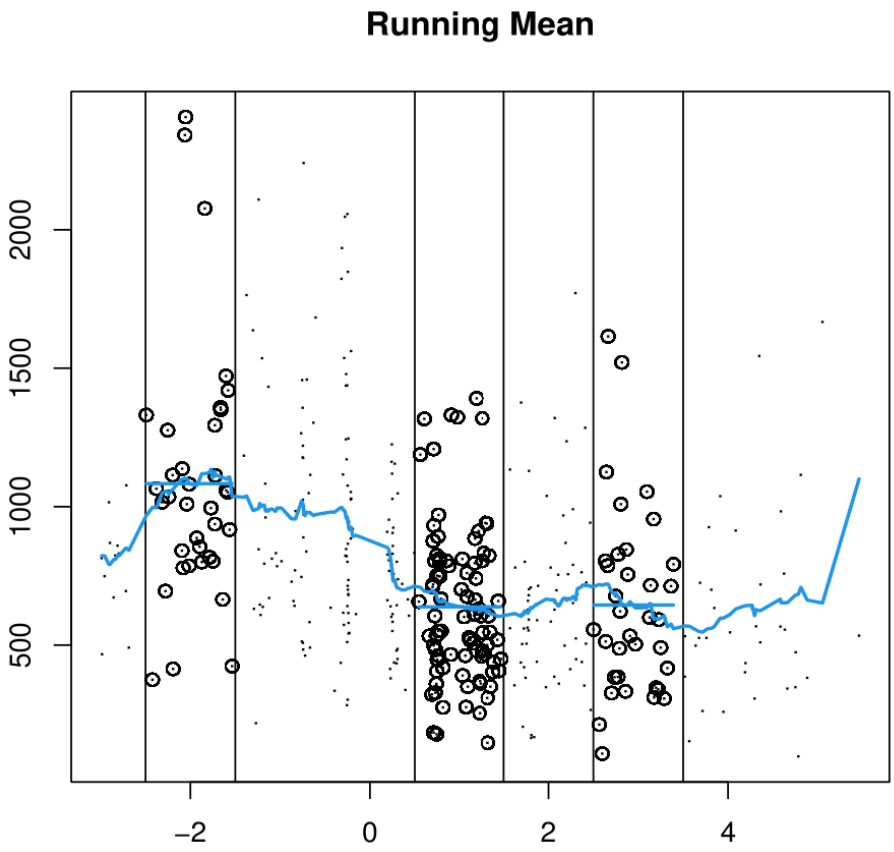

Figure 14: Moving average estimator for a particular choice of k. Note that the number k controls the smoothness of the resulting estimate.

Remark. The estimate such as in Figure 14 is usually too wiggly to be considered useful. Notice we can also fit a line instead of a constant, and the procedure is called running-line.

7.3 Kernel smoothers

One of the reasons why the previous smoother is wiggly is because when we move from x_i to x_{i+1}, two points are usually exchanged in the group we average about. If the exchanged points are highly dissimilar, then $\widehat{m}(x_i)$ and $\widehat{m}(x_{i+1})$ may be quite far from each other.

Idea. One way to try and fix this is by making the transition $x_i \to x_{i+1}$ smoother, for instance by using a kernel smoother.

7.3.1 Random design

Def. (Nadaraya-Watson estimator)

The Nadaraya-Watson estimator is defined as

$$\widehat{m}(x) = \sum_{i=1}^{n} W_i(x) Y_i, \tag{12}$$

where

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

and K is a positive integrable function called kernel and $h > 0$ is a scale parameter.

Remark. $\sum_{i=1}^{n} W_i(x_i) = 1$, hence $\widehat{m}(x)$ is a weighted average.

Remark. We do not require $\int_{\mathbb{R}} K(u)du = 1$, therefore the kernels might not be normalized.

Remark. This strategy makes sense for the estimation of the regression function,

$$m(x) = \mathbb{E}[Y|X = x] = \frac{\int_{\mathbb{R}} y f_{X,Y}(x, y)dxdy}{f_X(x)}, \tag{13}$$

and the problem is that we do not know neither $f_{X,Y}$ nor f_X. However, assuming that the kernel function is such that

$$\int K_a(u)du = 1, \quad \int_{\mathbb{R}} uK_a(y)du = 0,$$

then we estimate $\widehat{f}_{X,Y}(x, y)$ in (13) using two separate kernels (product kernel) for X and Y,

$$\widehat{f}_{X,Y}(x, y) = \frac{1}{nh_x h_y} \sum_{i=1}^{n} K_x(\frac{x - x_i}{h_x}) K_y(\frac{y - y_i}{h_y}),$$

from which we obtain

$$\begin{aligned} \widehat{m}(x) &= \frac{\int_{\mathbb{R}} y \widehat{f}_{X,Y}(x, y)dxdy}{\widehat{f}_X(x)}, \\ &= \frac{\sum_{i=1}^{n} K_X\left(\frac{x-x_i}{h_x}\right) Y_i}{\sum_{i=1}^{n} K_X\left(\frac{x-x_i}{h_x}\right)} \end{aligned}$$

Remark. The Gaussian kernel is not computationally efficient, since it requires all observations (x_i, y_i) for estimating the function at any point x. Since the kernel has a low impact on the resulting regression function, we usually prefer a bounded kernel for complexity reasons.

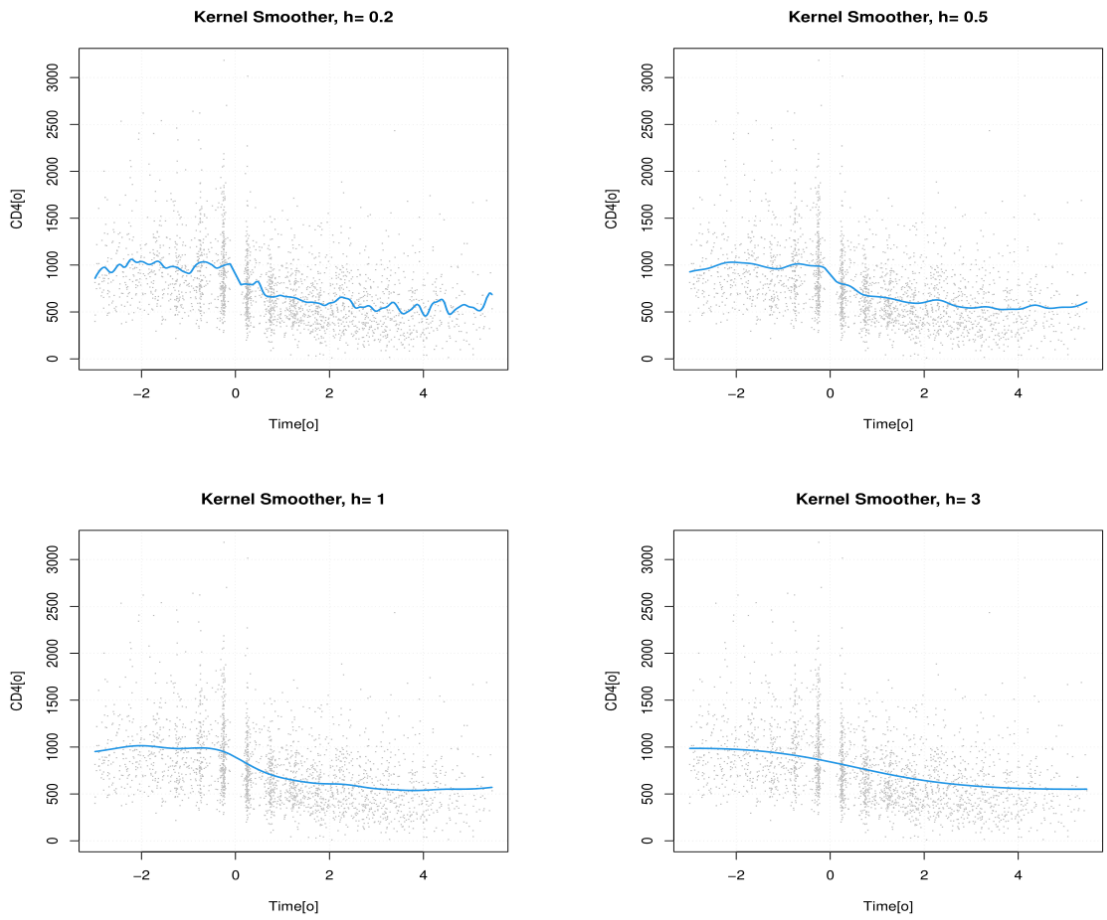

Figure 15: Example of kernel smoothed regression functions, for various choices of bandwidth h.

7.3.2 Fixed design

Suppose now that X is known, e.g. when we have a fixed design procedure and there is no distribution for X. In that case,

$$W_i(x) = \frac{1}{nh} \frac{K\left(\frac{x-x_i}{h}\right)}{f_X(x)},$$

1. *Fixed design*: $x_1 < \ldots < x_n$ and $x_i \in [a, b]$, then the "density" at the chosen points is equal to

$$\widehat{f}_X(x_i) = \frac{1}{n(x_i - x_{i-1})} \tag{14}$$

Def. (Priestley-Chao kernel estimator)

The Priestley-Chao kernel estimator for a fixed design uses the "fixed density" in (14),

$$\widehat{m}_{\mathrm{PC}}(x) = \sum_{i=1}^{n} (x_i - x_{i-1}) \frac{1}{h} K\left(\frac{x - x_i}{h}\right) Y_i.$$

2. On the other hand, we can use the weights given by

$$W_i^{\mathrm{GM}}(x) = \int_{s_{i-1}}^{s_i} \frac{1}{h} K\left(\frac{x-u}{h}\right) du, \tag{15}$$

for a choice of $x_{i-1} \leq s_{i-1} \leq x_i$. Usually, the default choice is to set $s_i = (x_i + x_{i+1})/2$, $s_0 = a$, $s_{n+1} = b$.

Def. (Gasser-Müller estimator)

The Gasser-Müller estimator uses the weights in (15) to write

$$\widehat{m}_{\mathrm{GM}}(x) = \sum_{i=1}^{n} \left[\int_{s_{i-1}}^{s_i} \frac{1}{h} K\left(\frac{x-u}{h}\right) du \right] Y_i.$$

Problem. A common problem when using nonparametric regression estimators is the boundary bias, due to the asymmetric contribution of observations near the boundary (Figure 12).

Solution. One possible idea is to constrain the estimator to be locally linear, such as the LOESS regression.

7.4 Consistency of the kernel regression estimator

Assume that we have a single covariate under the random design (for simplicity), with a response variable

$$Y_i = m(X_i) + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2),$$

and we apply a kernel density estimator with kernel K and bandwidth h. We want to characterize the behaviour of $\widehat{m}(x)$ as $n \to \infty$, and in order to do that we have to assume that:

1. $\int |K(u)| du < \infty$;
2. $\lim_{|u| \to \infty} uK(u) = 0$;
3. $\mathbb{E}[Y^2] < \infty$;
4. $h \to 0$ and $nh \to \infty$.

Prop. 6 (Consistency of the KDE)

With the above assumptions, at every point of continuity of $m(x)$, we have that

$$\frac{\sum_{i=1}^n K\left(\frac{x-x_i}{h}\right) Y_i}{\sum_{i=1}^n K\left(\frac{x-x_i}{h}\right)} \xrightarrow{P} m(x).$$

Remark. Under these assumptions, the Nadaraya-Watson estimator in (12) has asymptotic mean-squared error given by

$$\mathrm{AMSE}(\widehat{m}_n) = \frac{h^4}{4} \mu_2^2(K) \left\{ m''(x) + 2 \frac{m'(x) f_X'(x)}{f_X(x)} \right\}^2 + \frac{1}{nh} \frac{\sigma^2}{f_X(x)} \|K\|_2^2. \tag{16}$$

Assume instead the univariate fixed design model and the following assumptions:

› K has support $[-1,1]$ and $K(-1) = K(1) = 0$;

› m is twice continuously differentiable;

› $\max_i |x_i - x_{i-1}| = O(n^{-1})$;

› $\mathbb{V}[\varepsilon_i] = \sigma^2 < \infty$.

Then, for $n \to \infty$ and $h \to 0$ with $nh \to \infty$ we have that

$$\text{Bias}^2(\widehat{m}_n) = \frac{h^4}{4}\mu_2^2(K)m''(x)^2 + o(h^4),$$

$$\mathbb{V}[\widehat{m}_n] = \frac{1}{nh}\sigma^2\|K\|_2^2 + o\big((nh)^{-1}\big).$$

Remark. The resulting variance is not influenced in the first order by the shape of the regression function $m(x)$, whereas the bias is affected. The asymptotic mean-squared error is therefore

$$\text{AMSE}(\widehat{m}_n) = \frac{h^4}{4}\mu_2^2(K)\underbrace{m''(x)^2}_{\text{unknown}} + \frac{1}{nh}\sigma^2\|K\|_2^2. \tag{17}$$

Remark. The asymptotic MSE of the random design (16) and of the fixed design (17) can be written in both cases as

$$\text{AMSE}(n,h) = \frac{C_1}{nh} + h^4 C_2,$$

and minimizing with respect to h gives the asymptotic behaviour of the optimal bandwidth,

$$h_{\text{opt}} = O(n^{-1/5}), \tag{18}$$

with asymptotic mean-squared error of order $O(n^{-4/5})$, which is less efficient than the $O(n^{-1})$ that we obtain using OLS. This loss of performance is the drawback that we face for using a flexible mean function estimator.

7.5 Local linear regression

Starting from Taylor's theorem, we can say that any smooth function can be approximated with a sufficiently high-degree polynomial. Assume

$$Y_i = m(x_i) + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} WN(0,\sigma^2)$$

Theorem 10 (Taylor's theorem — original)
Suppose f is a real function on $[a,b]$, $f^{(K-1)}$ is continuous on $[a,b]$, $f^{(K)}(x)$ is bounded for $x \in (a,b)$, then for any distinct points $x_0 < x_1$ in $[a,b]$ there exists a point $x \in (x_0,x_1)$ such that

$$f(x_1) = f(x_0) + \sum_{k=1}^{K-1}\frac{f^{(k)}(x_0)}{k!}(x_1-x_0)^k + \frac{f^{(K)}(x)}{(K)!}(x_1-x_0)^K.$$

Theorem 11 (Taylor's theorem — Young's version)
Let f be such that $f^{(K)}(x_0)$ is bounded for x_0, then

$$f(x) = f(x_0) + \sum_{k=1}^{K} \frac{f^{(k)(x_0)}}{k!}(x - x_0)^k + o(|x - x_0|^K),$$

as $|x - x_0| \to 0$.

Remark. Another refinement of Taylor's theorem called Jackson's Inequality. Suppose f is a real function on $[a, b]$ with K is continuous derivatives, then if $\mathcal{P}_k$ is the space of polynomials of degree k we have that

$$\min_{g \in \mathcal{P}_k} \sup_{x \in [a,b]} |f(x) - g(x)| \leq C \left(\frac{b-a}{2k} \right)^K,$$

where $\mathcal{P}_k$ is the linear space of polynomials of degree k,

$$\mathcal{P}_k = \{a_0 + a_1 x + \ldots + a_k x^k,\ (a_0, \ldots, a_k) \in \mathbb{R}^{k+1}\}.$$

7.5.1 Local linear regression (LOESS)

We will now define the recipe to obtain a loess (local regression) smoother for a target covariate x_0. For computational and theoretical purposes we will define this weight function so that only values within a smoothing window $[x_0 - h(x_0), x_0 + h(x_0)]$ will be considered in the estimate of $m(x_0)$.

We define $h(x_0)$ so that we include a fixed $\alpha \times 100\%$ of the data, so that we have approximately regular variance at all points of the estimates for both fixed as well as random designs. Within the smoothing window $[x_0 - h(x_0), x_0 + h(x_0)]$, $m(x)$ is approximated by a polynomial, typically linear or quadratic,

$$m(x) \approx \beta_0 + \beta_1(x - x_0) + \frac{1}{2}\beta_2(x - x_0)^2,$$

Def. (General local linear regression)
The general local linear regression estimator of degree p for $m(x)$ at $x = x_0$ is the solution to

$$\sum_{i=1}^{n} \left(Y_i - \beta_0 + \beta_1(x_i - x_0) + \beta_2(x_i - x_0)^2 - \ldots - \beta_p(x_i - x)^p\right)^2 w_i(x_0), \tag{19}$$

where $w_i(x_0) = K\left(\frac{x_i - x_0}{h(x_0)}\right)$ is the weight for the i^{th} observation.

Remarks.

1. The kernel smoother is local linear regression when $p = 0$.
2. This estimator varies with x (in contrast to parametric least squares);
3. We need a careful choice of h in random design framework, since the AMSE in (16) is strongly influenced by $f(x)$.

Remark. The local linear regression allows the estimation of the derivative by using $\widehat{m}$ and calculating

$$\widehat{m}_p^{(\nu)}(x) = \nu!\widehat{\beta}_\nu(x),$$

and we usually have the order of the polynomial about $p = \nu + 1$ or $p = \nu + 3$.

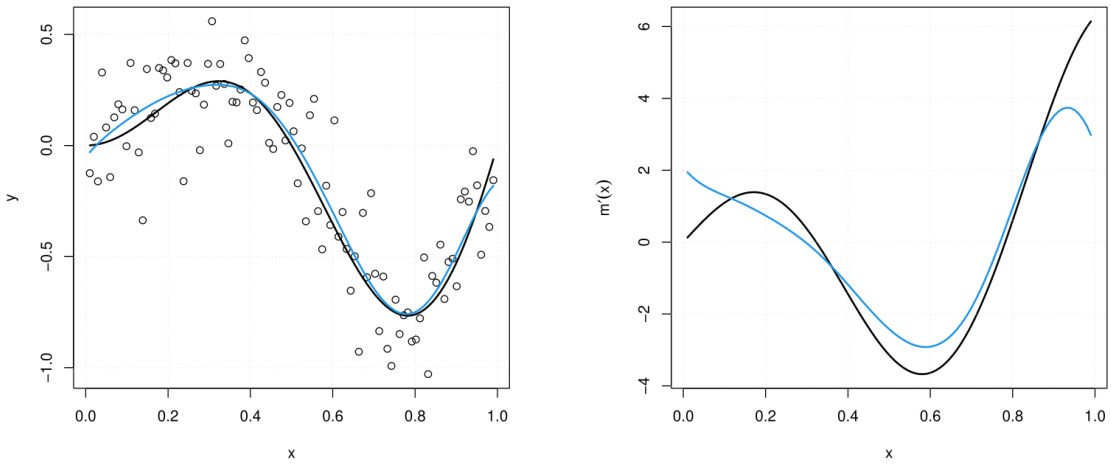

Figure 16: Estimator of the underlying function (*left*) and first derivative (*right*) for a local linear regression. We still observe a problem in the boundary of the support of x.

7.5.2 Estimator for the derivative of $m(x)$

Assume that K has support $[-1, 1]$ and $x \in [a, b]$ and that we want to estimate the k^{the} derivative $m^{(k)}(x)$ using the derivative of the local linear regression estimator in (19). Given some particular functions v and B_1, B_2, B_3 which depend on p, k, m, K (Heckman's notes), we have that

a) *Variance of the estimator*:

$$\mathbb{V}\left[\widehat{m}^{(k)}(x)\right] \sim \frac{v(x)}{nh^{2k+1}},$$

hence we need $nh^{2k+1} \to 0$.

b) *Bias of the estimator*:

$$p - k \text{ EVEN}: \quad \text{Bias}\big(\widehat{m}^{(k)}(x)\big) \sim \begin{cases} B_1(x)h^{p-k+2} & x \in (a, b) \\ B_2(x)h^{p-k+1} & x \in \{a, b\} \end{cases}$$

$$p - k \text{ ODD}: \quad \text{Bias}\big(\widehat{m}^{(k)}(x)\big) \sim B_3(x)h^{p-k+1}$$

p	$x \in [a+h, b-h]$	$x \notin [a+h, b-h]$
0	h^2	h
1	h^2	h^2
2	h^4	h^3
3	h^4	h^4
p odd	h^{p+1}	h^{p+1}
p even	h^{p+2}	h^{p+1}

Figure 17: Asymptotic order of the bias for the derivative estimator.

Remark. Seeing the results above, many researchers usually choose $p-k$ odd, for instance $p = k+1$, so that the bias for the k^{th} derivative is on the same order for both the interior and the boundary of the covariate space.

Remark. When we want to estimate m, i.e. $k = 0$, then using a local linear estimate with $p = 1$ yields asymptotically similar results to a local quadratic estimator but saves lots of computational time.

7.5.3 Robust fitting

If the errors have a symmetric distribution (heavy tails), or if there appears to be outliers we can use a robust extension of the loess. Consider the residuals from the application of the local linear regression estimator,

$$\widehat{\varepsilon}_i = Y_i - \widehat{m}(x_i),$$

then we can consider a bisquare weight function,

$$B(u; b) = \begin{cases} \left[1 - (u/b)^2\right]^2 & \text{if } |u| < b \\ 0 & \text{if } |u| > b \end{cases}$$

and apply a second smoothing on the Y_i's which takes into account the first smoothing iteration. Specifically, we can define the robust weights as

$$r_i = B(\widehat{\varepsilon}_i; 6m), \quad m = \text{median}(|\widehat{\varepsilon}_i|),$$

and the local linear regression is repeated by replacing the weights $w_i(x)$ with the new weights $r_i w_i(x)$.

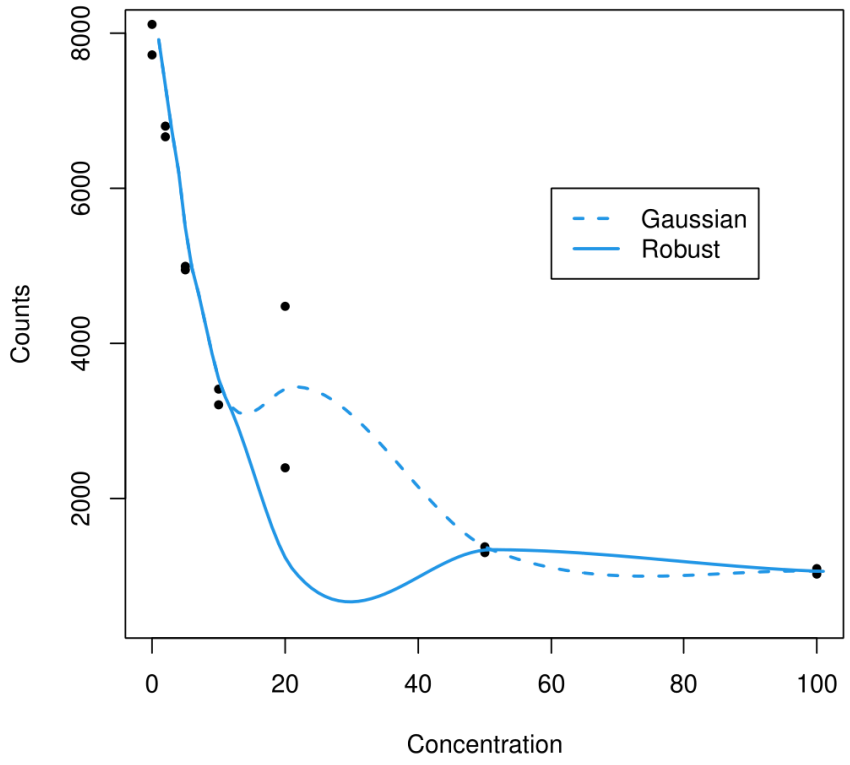

Figure 18: Comparison between robust local linear regression and LOESS.

Remark. The robust estimate is the result of repeating the above procedure several times (e.g. three times), which is how the function `lowess` is implemented in `R`.

7.5.4 Autocorrelated data

Suppose that we observe a times series with equi-spaced times $t_1, t_2, \ldots$, i.e.

$$Y_i = m(t_i) + \varepsilon_i,$$

where ε is a stationary zero-mean process (White Noise process),

$$\mathbb{E}[\varepsilon_i] = 0$$

$$\mathrm{Cov}(\varepsilon_i, \varepsilon_{i+k}) = \gamma(k)$$

with a mixing condition $\sum_{k=1}^{\infty} k|\gamma(k)| < \infty$. Then, the asymptotic mean-squared error of the smoothed estimate takes into account the autocovariance of the process,

$$\mathrm{AMISE}(\widehat{m}_n) = \underbrace{\frac{h^4 \mu_2^2(K) \int m''(u)^2 du}{4}}_{\mathrm{Bias}^2} + \underbrace{\frac{\sigma^2 + 2\sum_{k=1}^{\infty} |\gamma(k)|}{nh} \|K\|_2^2}_{\mathbb{V}},$$

which inflates the variance of the estimator.

7.5.5 Local likelihood model

The local regression framework can be generalized to a general local likelihood regression model, by assuming that the response variable is distributed according to a dispersion family (Pace and Salvan, 1997),

$$Y_i \sim f(y, \vartheta(x_i)),$$

where $\vartheta_i = \vartheta(x_i)$ is the regression parameter. Then, the usual log-likelihood for the parameter ϑ is simply

$$\ell(\vartheta) = \sum_{i=1}^{n} \ell\big(Y_i, \vartheta(x_i)\big).$$

Suppose that $\vartheta(x) = \beta_0 + \beta_1 x$, then we can generalized the above quantity to a local likelihood for ϑ by weighting each contribution with a kernel, analogously to the weighted sum of squares,

$$\mathcal{L}_x(\vartheta) = \sum_{i=1}^{n} \ell\big(Y_i, \beta_0 + \beta_1(x_i - x)\big) K\left(\frac{x_i - x}{h(x)}\right).$$

Def. (Local likelihood estimator)
Maximizing $\mathcal{L}_x$ over the unknown β's defines the local likelihood estimator,

$$\widehat{\vartheta}(x) = \widehat{\beta}_0(x).$$

Remark. Extending to the GLM case, if $f(y, \vartheta)$ is a parametric family of distributions with mean $\mathbb{E}_\vartheta[Y_i] = \mu(\vartheta)$, then $g = \mu^{-1}$ is the link function such that $\vartheta = g(\mu)$.

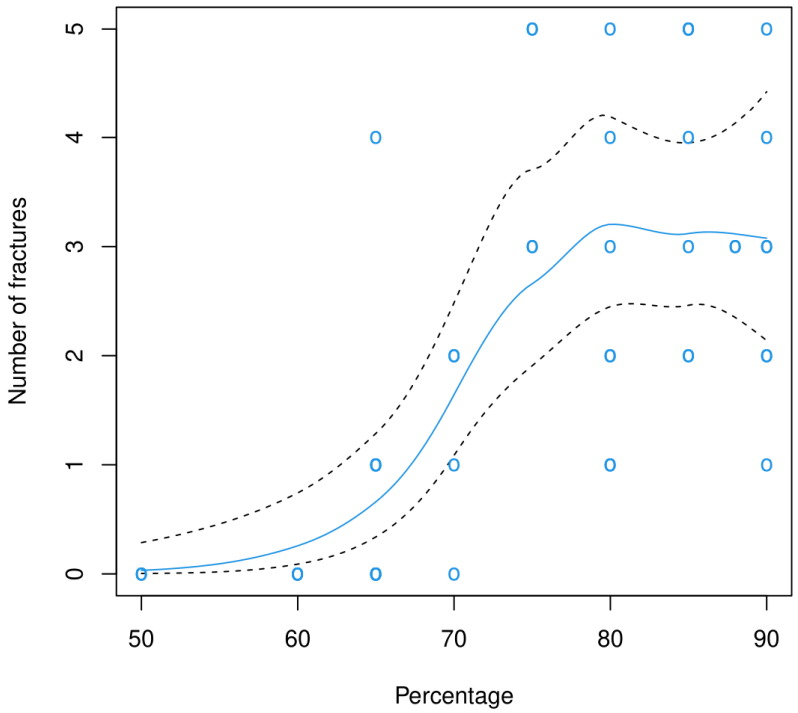

Figure 19: Local log-linear model with $Y \sim \text{Pois}(\mu)$ and $\vartheta = \log \mu$ for the number of fractures as a function of percentage of extraction.

7.6 Orthogonal series estimator

Suppose m is supported on a compact interval, i.e. $[0,1]$, and that we are in a fixed design case,

$$Y_i = m(x_i) + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \text{WN}(0, \sigma^2).$$

where $x_i = i/n$.

Remark. The fixed design assumption is critical, since the estimators are not consistent in the random design case. Consider $L_2[0,1] = \{f : \int_0^1 f(x)^2 dx < \infty\}$, then we have multiple choices for defining a complete orthonormal basis $f_1, f_2, \ldots$, of functions in $L_2[0,1]$ such that

$$\int_0^1 f_i(x) f_j(x) dx = \delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \tag{20}$$

For instance, standard results from functional analysis can be used to show that that any function $m \in L_2[0,1]$ can be represented as

$$m(x) = \sum_{k=1}^{\infty} m_k f_k(x) dx, \quad m_k = \int_0^1 m(x) f_k(x) dx,$$

where $f_1, f_2, \ldots$ are such that (20) is satisfied.

Example (Fourier series)
Any function $m \in L_2[0,1]$ can be represented as

$$m(x) = m_1 + \sum_{k=1}^{\infty} m_{2k} \sqrt{2} \cos(2\pi k x) + \sum_{k=1}^{\infty} m_{2k+1} \sqrt{2} \sin(2\pi k x),$$

where

$$m_1 = \int_0^1 m(x) dx$$

$$m_{2k} = \sqrt{2} \int_0^1 m(x) \cos(2^k \pi x) dx$$

$$m_{2k+1} = \sqrt{2} \int_0^1 m(x) \sin(2^k \pi x) dx$$

Example (Legendre polynomials)
Wikipedia.

Example (Wavelets)
Wikipedia.

Problem. The basis functions f_k are known (you choose your favourite basis of $L_2[0,1]$), but the m_k's are unknown, since we do not know the true underlying function m.

Solution. A good estimator of the unknown weights m_k's can be shown to be

$$\widehat{m}_k = \frac{1}{n}\sum_{i=1}^{n} Y_i f_k(x_i),$$

and the evaluation is especially fast using the Fast Fourier Transform (FFT).

Remark. Can we really obtain an infinite number of coefficients m_k? If we take a look at

$$\mathbb{E}[\widehat{m}_k] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[Y_i]f_k(x_i) = \frac{1}{n}\sum_{i=1}^{n} m(x_i)f_k(x_i) = \frac{1}{n}\sum_{i=1}^{n} m(i/n)f_k(i/n) \approx \int_0^1 m(x)f_k(x)dx,$$

then we see that $\widehat{m}_k$ is simply an asymptotic unbiased estimator of m_k, hence if we use too many coefficients $\widehat{m}_k$ it will be too variable.

Def. (Truncated orthogonal series)

We can define the truncated orthogonal series estimator as

$$\widehat{m}_K(x) = \sum_{k=1}^{K} \widehat{m}_k f_k(x).$$

Remark. K plays the role of a smoothing parameter, i.e.

- K small $\implies$ $\widehat{m}_K$ is biased
- K large $\implies$ $\widehat{m}_K$ is too variable

In practice, we choose K using cross-validation and if m is smooth then m_k will only be large for a few initial elements. However, we can have some bias issues if the largest coefficients arise late in the series and we truncate the series too early.

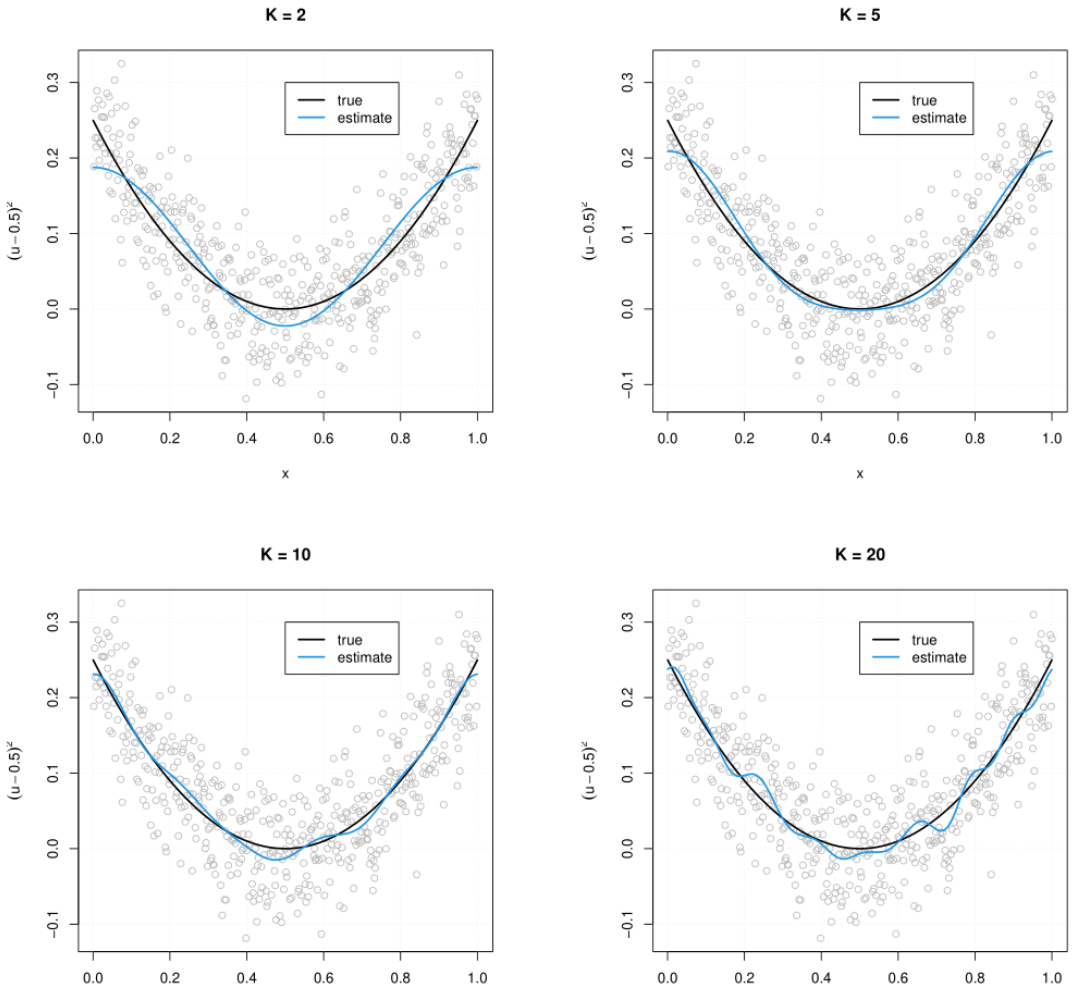

Figure 20: Simulated example of an orthogonal estimate using a Fourier basis.

References

Hall, P. (1987). «On Kullback-Leibler Loss and Density Estimation». In: *The Annals of Statistics* 15.4, 1491–1519.

Pace, L. and Salvan, A. (1997). *Principles of Statistical Inference: From a Neo-Fisherian Perspective*. Singapore: World Scientific Pub.

Wasserman, L. (2005). *All of Nonparametric Statistics*. New York: Springer.