MATH 441 – Robust and Nonparametric Statistics

Pr. Stephan Morgenthaler

Course notes by
Léo Belzile
leo.belzile@epfl.ch

THE CURRENT VERSION IS THAT OF JANUARY 27, 2016
FALL 2015, EPFL

Please signal the author by email if you find any typo.
These notes should be read carefully, as they almost surely contain typos and mistakes.
The content and the notation may differ with that presented in class.

Contents

1 Review of Likelihood Methods **4**

2 Robustness **8**

2.1 Critique of the likelihood method 8

2.2 Basic types of estimates 11

2.2.1 M-Estimators 11

2.3 L-estimators and R-estimators 18

2.4 Influence function 19

2.4.1 Functional interpretation 20

2.4.2 Asymptotic analysis 22

2.4.3 Interpretation of the influence function 23

2.4.4 Qualitative indicators 24

2.5 Optimal B-robust estimates 25

2.5.1 B-robust estimates 25

2.5.2 Minimum variance B-robust estimation 28

2.6 Robust minimax theory 31

2.7 Robust regression 36

2.7.1 M-estimate 41

2.7.2 Alternative estimates 41

3 Rank-based statistical procedures **44**

3.1 Introduction 44

3.1.1 Order statistics, ranks and their properties 44

3.2 Examples of rank statistics 46

3.2.1 One sample statistics 46

3.2.2 Two sample statistics 48

3.3 Locally most powerful rank tests 50

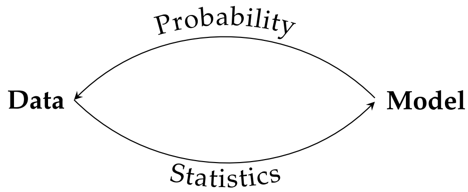
3.4 Asymptotic distribution of rank tests 54
3.5 Asymptotic power and Pitman efficacy 59
3.5.1 General formula for the Pitman efficacy for test statistics with Gaussian limits 61
3.5.2 Asymptotic relative efficiency of tests 63

4 Nonparametric regression **67**
4.1 Smoothing 67
4.1.1 Bias and variance of nearest neighbours smoothers 68
4.2 Smoothing splines 70
4.2.1 Splines 70
4.2.2 Optimality properties of splines 72
4.2.3 Natural smoothing splines 72
4.2.4 Natural splines determined by their values and second derivatives 73
4.2.5 Summary on linear smoothers 76
4.2.6 Estimation of the variance of the random errors 78
4.2.7 Cross-validation 79
4.3 Kernel smoothers and local regression 82
4.3.1 Kernel smoothers 82
4.3.2 Local regression 83
4.4 Orthonormal basis 84
4.5 Shrinkage 86

This course gives an overview of the theory of robustness and distribution-free tests. It will also give an introduction to estimation in a function space. After the course, the student will know the most important results in these areas and will have learned to use robust and non-parametric methods.

Part 1
Review of Likelihood Methods

The basis of modeling is a stochastic process which describes the data generation (simulating from the model should generate output resembling the data).



Typically, a model contains parameters. The data is used to **estimate** (identify, learn) the true value of the parameter or to **test questions** about the parameters.

Example 1.1 (Regression models)

(a) The simple linear regression has $Y_i = \alpha + \beta x_i + \sigma E_i$ for $i = 1, \ldots, n$. We can assume $E_1, \ldots, E_n \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, but also $E_1, \ldots, E_n \overset{\text{iid}}{\sim} G$ where $G \overset{\cdot}{\sim} \mathcal{N}(0,1)$; a method which works well with this model is robust. On the contrary, we can also consider lifting restrictions as $E_1, \ldots, E_n \overset{\text{iid}}{\sim} F \subset \mathcal{F}$, typically a class of absolutely continuous distributions. This problem is rank-based, and non-parametric, meaning that the parameter belongs to an infinite dimensional space.

(b) Multiple regression, of the form $Y_i = \theta_0 + \boldsymbol{x}_i^\top \boldsymbol{\theta} + \sigma E_i$, with $\dim(\boldsymbol{\theta}) = p$.

(c) Another non-parametric problem is smoothing: for $x_i \in \mathbb{R}$, we consider $Y_i = f(x_i) + \sigma E_i$, where $f \in \mathcal{C}^2(\mathbb{R})$

Definition 1.1 (Likelihood)

The joint density of the data, viewed as a function of the parameter(s), is called the **likelihood** (function) and denoted $L_n(\theta)$, n indicating the sample size.

Example 1.2 (Likelihood of Bernoulli variates)

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{B}(p)$. Then, the density for one observation x is

$$f(x \mid p) = p^x (1-p)^{1-x} = \begin{cases} p & \text{if } x = 1, \\ 1-p & \text{if } x = 0 \end{cases}$$

and thus

$$L_n(p) = p^{\Sigma_n} (1-p)^{n-\Sigma_n}, \qquad \Sigma_n = \sum_{i=1}^n X_i$$

To find the maximum of this equation as a function of p, we compute

$$\frac{\partial}{\partial p} L_n(p) = \Sigma_n p^{\Sigma_n - 1}(1-p)^{n-\Sigma_n} - p^{\Sigma_n}(1-p)^{n-\Sigma_n-1}(n-\Sigma_n)$$

and the root of $\partial L_n / \partial p$ is given by

$$\frac{\Sigma_n}{\widehat{p}_n} = \frac{n-\Sigma_n}{1-\widehat{p}_n} \qquad \Leftrightarrow \qquad \widehat{p}_n = \frac{\Sigma_n}{n} = \bar{X}_n$$

If $L_n(p_1) > L_n(p_2)$, the Bernoulli model $\mathcal{B}(p_1)$ is more probable that $\mathcal{B}(p_2)$ to have generated the data.

Definition 1.2 (Maximum likelihood estimator)
The value $\widehat{\boldsymbol{\theta}}$ of the parameter which satisfies $L_n(\widehat{\boldsymbol{\theta}}_n) \geq L_n(\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \Theta$ is called the **maximum likelihood estimator** (MLE). It is often found as a root of the **likelihood equation**

$$\frac{\partial}{\partial \boldsymbol{\theta}} \log(L_n(\boldsymbol{\theta})) = \mathbf{0}.$$

The function $\ell_n(\boldsymbol{\theta}) = \log(L_n(\boldsymbol{\theta}))$ is called **log-likelihood** and the score equation for a sample of size n is

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ell_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \log(f(X_i \mid \boldsymbol{\theta})) = \mathbf{0}.$$

Example 1.3 (Likelihood of Bernoulli variates (continued))
The log-likelihood for an n-sample of Bernoulli variates is

$$\ell_n(p) = \Sigma_n \log(p) + (n-\Sigma_n)\log(1-p).$$

Example 1.4 (Independent and identically distributed random variables)
Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F(x \mid \theta)$ with density $f(x \mid \theta)$ for some unknown $\theta = \theta_0$. We have $L_n(\theta) = \prod_{i=1}^{n} f(X_i \mid \theta)$ and

$$\frac{1}{n}\ell_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell_n(f(X_i \mid \theta)) \xrightarrow{\text{a.s.}} \mathsf{E}\left(\log(f(X \mid \theta))\right)$$

where

$$\begin{aligned} \mathsf{E}\left(\log(f(X \mid \theta))\right) &= \int \left[\log\left(\frac{f(x \mid \theta)}{f(x \mid \theta_0)}\right) + \log(f(x \mid \theta))\right] f(x \mid \theta_0)\, \mathrm{d}x \\ &= -\mathsf{I}_{\theta:\theta_0} + \int \log(f(x \mid \theta_0)) f(x \mid \theta_0)\, \mathrm{d}x. \end{aligned}$$

$\mathsf{I}_{\theta:\theta_0}$ is the **divergence**. We stick with the original definition of Kullback & Leibler (1951).

Definition 1.3 (Kullback–Leibler divergence)
Let f_1 and f_2 be two densities. The quantity

$$\mathsf{I}_{1:2} := \int \log\left(\frac{f_1(x)}{f_2(x)}\right) f_1(x)\,\mathrm{d}x \geq 0$$

is called the **mean information** in an observation from f_1 for discriminating f_1 from f_2.[1] We have that $\mathsf{I}_{1:2} = 0$ if and only if $f_1 = f_2$ almost surely.

The quantity

$$\mathsf{J}(f_1, f_2) := \mathsf{I}_{1:2} + \mathsf{I}_{2:1} = \int (f_1(x) - f_2(x)) \log\left(\frac{f_1(x)}{f_2(x)}\right)\,\mathrm{d}x$$

is the **Kullback–Leibler divergence**. Jeffreys noted the "symmetry, positive definiteness and additivity" of the divergence, as well as its "invariance to non-singular transformations".

This shows that $\mathsf{E}_{\theta_0}\left(\log(f(X \mid \theta))\right)$ has a unique maximum at $\theta = \theta_0$ under the condition of identifiability (if $\theta_1 \neq \theta_2$, then $f(x \mid \theta_1) \neq f(x \mid \theta_2)$).

Usually, one has $\widehat{\theta}_n \to \theta_0$ almost surely or in probability as $n \to \infty$. Is θ_0 an asymptotic root of the likelihood equation? That is the case if the MLE is consistent. We investigate the **Fisher consistency** of the score. First,

$$\frac{1}{n}\frac{\partial}{\partial\theta}\ell_n(\theta_0) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log(f(X_i \mid \theta_0)) \to \mathsf{E}_{\theta_0}\left(\frac{\partial}{\partial\theta}\log(f(X \mid \theta_0))\right)$$

as $n \to \infty$, where

$$\begin{aligned}
\mathsf{E}_{\theta_0}\left(\frac{\partial}{\partial\theta}\log(f(X \mid \theta_0))\right) &= \int \frac{\partial}{\partial\theta}\log(f(x \mid \theta_0)) f(x \mid \theta_0)\,\mathrm{d}x \\
&= \int \frac{\frac{\partial}{\partial\theta} f(x \mid \theta_0)}{f(x \mid \theta_0)} f(x \mid \theta_0)\,\mathrm{d}x \\
&= \frac{\partial}{\partial\theta}\int f(x \mid \theta_0)\,\mathrm{d}x = \frac{\partial}{\partial\theta}1 = 0
\end{aligned}$$

For large values of n and for scalar θ:

$$0 = \frac{\partial}{\partial\theta}\ell_n(\widehat{\theta}_n) \cong \frac{\partial}{\partial\theta}\ell_n(\theta)\bigg|_{\theta=\theta_0} + (\widehat{\theta}_n - \theta_0)\,\frac{\partial^2}{\partial\theta^2}\ell_n(\theta)\bigg|_{\theta=\theta_0} \tag{1.1}$$

[1] This is what is often nowadays termed the Kullback–Leibler divergence of f_2 from f_1.

and thus

$$\widehat{\theta}_n - \theta_0 \cong \left. \frac{\sum_{i=1}^n \frac{\partial}{\partial \theta} \log(f(x_i \mid \theta))}{-\sum_{i=1}^n \frac{\partial^2}{\partial \theta^2} \log(f(x_i \mid \theta))} \right|_{\theta=\theta_0}.$$

We therefore have

Theorem 1.4 (Asymptotic normality of maximum likelihood estimator)
The asymptotic behaviour of the MLE is as follows:

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}\left(0, \mathcal{I}(\theta_0)^{-1}\right),$$

where the Fisher information is

$$\mathcal{I}(\theta) := \mathsf{E}_\theta\left(\left[\frac{\partial}{\partial \theta} \log(f(X \mid \theta))\right]^2\right) \geq 0.$$

Under additional regularity conditions, we also have equality of the Fisher information with the negative Hessian matrix

$$\mathcal{I}(\theta) = -\mathsf{E}_\theta\left(\frac{\partial^2}{\partial \theta^2} \log(f(X \mid \theta))\right).$$

An immediate consequence of Theorem 1.4 is that $\widehat{\theta}_n \dot{\sim} \mathcal{N}(\theta_0, (n\mathcal{I}(\theta_0))^{-1})$.

Often, the MLE has to be computed by numerical optimization. Equation (1.1) shows a possible avenue, taking

$$\widehat{\theta}_{i+1} = \widehat{\theta}_i + \frac{\frac{\partial}{\partial \theta}\ell_n(\widehat{\theta}_i)}{-\frac{\partial^2}{\partial \theta^2}\ell_n(\widehat{\theta}_i)}$$

which is in a nutshell the **Newton-Raphson algorithm**. The quantity

$$\mathcal{J}(\widehat{\theta}) = -\frac{1}{n}\frac{\partial^2}{\partial \theta^2}\ell_n(\widehat{\theta}_i)$$

is called the **observed information**. Another possible algorithm for calculations of the MLE is the Fisher's scoring algorithm,

$$\widehat{\theta}_{i+1} = \widehat{\theta}_i + \frac{\frac{\partial}{\partial \theta}\ell_n(\widehat{\theta}_i)}{\mathcal{I}(\widehat{\theta}_i)},$$

which uses the **expected information**.

Part 2
Robustness

2.1. Critique of the likelihood method

We can show that if the model is accurate, then the MLE will be asymptotically the most efficient consistent estimator (under regularity conditions).

What is the behaviour of the MLE if the model only holds approximately? Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} G(x)$, which we approximate by $F(X \mid \theta)$.

Example 2.1
Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} (1-\varepsilon)\mathcal{N}(\mu_0, 1) + \varepsilon H$. What will be the impact of ε and H on the MLE? Suppose $H = \delta_k$ for $k \in \mathbb{R}^+$. We can show that the expectation and the variance of the MLE are strongly perturbated. Express X_i as

$$X_i = (1 - B_i)Z_i + B_i k$$

where $B_i \overset{\text{iid}}{\sim} \mathcal{B}(\varepsilon)$ and $Z_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu_0, 1)$. Then

$$\begin{aligned}
\mathsf{E}\left(\widehat{\mu}_0\right) &= \mathsf{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \mathsf{E}\left(X_i\right) \\
&= \mathsf{E}\left((1 - B_i)Z_i + kB_i\right) \\
&= \mathsf{E}_{B_i}\left(\mathsf{E}_{Z_i|B_i}\left((1 - B_i)Z_i + kB_i\right)\right) \\
&= (1 - \varepsilon)\mu_0 + k\varepsilon
\end{aligned}$$

and

$$\mathsf{Var}\left(\widehat{\mu}_n\right) = \frac{1}{n}\left(\varepsilon(1 - \varepsilon)(\mu_0^2 + k^2)\right) + (1 - \varepsilon)$$

If $\varepsilon \approx 0.05$, $\mu_0 = 0$, $n = 100$ and k is very large, we are easily convinced that the median will be a "better" estimator in terms of the MSE.

The following material is extracted from Huber & Ronchetti (2009).

In order to estimate the standard error of Normal data $\mathcal{N}(\mu, \sigma^2)$, Fisher proposed to use the MLE

$$S_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

If the model is correct, this estimator is the most efficient (among consistent estimators).

Despite this argument, the physicist Eddington observed that

$$\widetilde{S}_n = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|$$

performs better.

An eye opening example has been given by Tukey (1960). If we suppose a mixture of good observations $\mathcal{N}(\mu, \sigma^2)$ and bad ones $\mathcal{N}(\mu, 9\sigma^2)$ in proportions $1-\varepsilon, \varepsilon$, then even for small values such as $\varepsilon \approx 0.02$, the robust estimator of Eddington performs "better" than the MLE in terms of the asymptotic relative efficiency (ARE) of $\widetilde{S}_n$ to S_n, given by

$$\mathsf{ARE} = \lim_{n\to\infty} \frac{\mathsf{Var}\,(S_n)\,/\,[\mathsf{E}\,(S_n)]^2}{\mathsf{Var}\big(\widetilde{S}_n\big)\,/\,\big[\mathsf{E}\big(\widetilde{S}_n\big)\big]^2} = \frac{\frac{1}{4}\left[\frac{3(1+80\varepsilon)}{(1+8\varepsilon)^2} - 1\right]}{\frac{\pi(1+8\varepsilon)}{2(1+2\varepsilon)^2} - 1}.$$

Why use the ARE? Note that $\mathsf{E}\,(S_n) \to \sigma$, while $\mathsf{E}\big(\widetilde{S}_n\big) = \sqrt{2/\pi}\sigma \cong 0.8\sigma$. We should compare

$$\lim_{n\to\infty} \frac{\mathsf{Var}\,(c_{1,n}S_n)}{\mathsf{Var}\left(c_{2,n}\widetilde{S}_n\right)},$$

with $\mathsf{E}\,(c_{1,n}S_n) \to \sigma$, $\mathsf{E}\,(c_{2,n}S_n) \to \sigma$ and $c_{1,n} = \sigma/\mathsf{E}\,(S_n)$ and $c_{2,n} = \sigma/\mathsf{E}\big(\widetilde{S}_n\big)$.

What **robustness** is trying to do is to broaden the model. The model assumptions may fail to hold, and so this leads to a question, namely what should a robust procedure achieve?

- it should have a nearly optimal performance at the assumed model
- small deviations from the model assumptions should impair the performance only slightly
- somewhat larger deviations from the model should not cause an immediate breakdown (catastrophe, absurd values).

Gross errors (termed outliers) occurring in a small proportion are considered a small deviation. Similarly, small errors (through for example rounding) in a large proportion of the data are also small deviations.

Remark 2.1

Let $\mathcal{M}$ be the set of all probability measures on a sample space Ω, usually a finite dimensional Euclidean space. Then, the above ideas of "small deviations" are captured by the **Prohorov metric**:

$$d_P(F,G) := \inf\left\{\varepsilon > 0 : F\{A\} \leq G\{A^\varepsilon\} + \varepsilon, \text{ for all Borel sets } A \subset \mathbb{B}(\Omega)\right\},$$

for $F, G \in \mathcal{M}$, where

$$A^{\varepsilon} = \left\{ x \in \Omega : \inf_{y \in A} \|x - y\| < \varepsilon \right\}.$$

the closed ε-neighbourhood of A.

Theorem 2.1 (von Strassen)
This theorem is extracted from von Strassen (1965), pp. 423-439. Let $F, G \in \mathcal{M}$; then

$$F\{A\} \leq G\{A^{\delta}\} + \varepsilon, \qquad \forall\ A \in \mathbb{B}(\Omega) \text{ and } \delta, \varepsilon > 0$$

if and only if there exists two random variables X, Y with values in Ω and laws F and G, respectively, such that

$$\mathsf{P}\left(\|X - Y\| \leq \delta\right) \geq 1 - \varepsilon.$$

The first paper using the term **robust** was due to G.E.P. Box, in the '30s, on behaviour of test statistics. The near optimality of performance is usually judged by asymptotic indicators (with $n \to \infty$), such as the asymptotic variance, the asymptotic efficiency or the asymptotic relative efficiency. But unless the convergence is uniform in the neighbourhood of the assumed model, there are difficulties with this approach because the results cannot guarantee anything for finite samples.

The theory of robust statistics laid out hereafter is based on the work of Peter Huber and Frank Hampel, both professors at ETHZ, and is mostly asymptotic in nature. It is also possible to stay within the given sample and to study robustness with regard to small changes of the sample. This approach is due to John Tukey, and was termed by him as "resistance".

Definition 2.2 (Sensitivity curve of an estimate)
A function of the statistic $T_{n-1}(x_1, \ldots, x_{n-1})$, the subscript indicating the number of observations. The sensitivity curve is

$$\mathrm{SC}_{n-1}(x) = n[T_n(x_1, \ldots, x_{n-1}, x) - T_{n-1}(x_1, \ldots, x_{n-1})].$$

The asymptotic version of the sensitivity curve is Frank Hampel's influence function. The scaling by the sample size preserves the scale.

Example 2.2 (Sensitivity curve of mean and median)
For the arithmetic mean, we have $\bar{x} = (x_1 + \cdots + x_{n-1})/(n-1)$, so that

$$\begin{aligned} \mathrm{SC}_{n-1}(x) &= n\left(\frac{1}{n}\left[(n-1)\bar{x} + x\right)\right] - \bar{x}\right) \\ &= x - \bar{x} \end{aligned}$$

which goes to ∞ as $x \to \infty$, certainly an unbounded function.

In contrast, if T is the median, then

$$\begin{aligned}
\mathrm{SC}_{n-1}(x) =& n\left(\operatorname{med}(x_1,\dots,x_{n-1},x) - \operatorname{med}(x_1,\dots,x_{n-1})\right) \\
=& n\left(\frac{1}{2}\left(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right) - x_{(\frac{n}{2})}\right) \\
=& \frac{n}{2}\left(x_{(\frac{n}{2}+1)} - x_{(\frac{n}{2})}\right).
\end{aligned}$$

for n even, $n \geq 4$, you can write this and the distance is $\mathrm{O}_p(n)$. We note that the mean has an unbounded sensitivity, whereas the median has a bounded sensitivity.

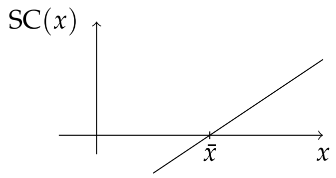


Figure 1: Sensitivity curve for the mean

2.2. Basic types of estimates

In order to construct optimal robust estimates, we need to have a large set of candidate estimates to choose from, such as the class of all unbiased estimators, admissible estimators, etc.

2.2.1. M-Estimators

These were introduced by Peter Huber, who also established their asymptotic properties. M-estimator stands for "maximum likelihood like" estimators.

Definition 2.3 (M-estimators)
An M-estimator is any estimate T_n defined by a minimization problem of the form $\sum_{i=1}^{n} \varrho(x_i; T_n)$ for data x_i and T_n which minimizes this sum, where ϱ is an arbitary function, ψ is the derivative with respect to θ of ϱ, that is $\psi(x;\theta) := \pm\frac{\partial}{\partial\theta}\varrho(x;\theta)$, and

$$T_n = \arg\min_{\theta} \sum_{i=1}^{n} \varrho(x_i;\theta).$$

It can be defined by an implicit equation of the form

$$\sum_{i=1}^{n} \psi(x_i; T_n) = 0. \tag{2.2}$$

Equation (2.2), is nowadays called **estimating equation**. Note that choosing $\varrho(x;\theta) = -\log f(x;\theta)$ leads to the MLE.

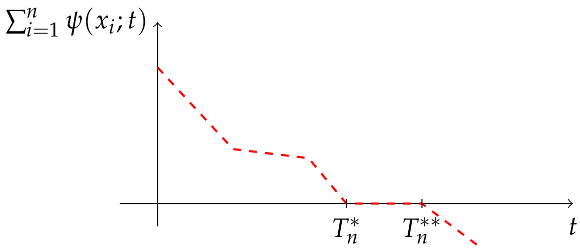
Example 2.3 (Location estimates)
Often, we will encounter $T_n = \arg\min_\theta \sum \varrho(x_i - \theta)$ and $\sum_{i=1}^n \psi(x_i - T_n) = 0$. In this case, we only need to specify a function ϱ and $\psi \equiv \varrho'$.

Let $\theta \in \mathbb{R}$ be a location parameter for an absolutely continuous distribution function F. A location family is such that $f_\theta(x) = f_0(x - \theta)$ for a probability density function with $\theta = 0$. Oftentimes, it is possible to express the results for the asymptotic variance in terms of $\mathrm{d}F_0(0)$.

Asymptotic properties of M-estimates

The mathematical development is similar to that for maximum likelihood estimators. A simple theory is possible if ϱ is convex in θ or if $\psi(x;\theta)$ is monotone in θ. Assume that $\psi(x;\theta)$ is measurable in x and non-increasing in θ.

In the most general case, this may look like



To be precise, let $T_n^* = \sup\{t : \sum_{i=1}^n \psi(x_i;t) > 0\}$ and $T_n^{**} = \inf\{t : \sum_{i=1}^n \psi(x_i;t) < 0\}$. Any $T_n^* \leq T_n \leq T_n^{**}$ is an M-estimate under weak monotonicity. We have

$$\{t : T_n^* < t\} \subset \left\{t : \sum_{i=1}^n \psi(x_i;t) \leq 0\right\} \subset \{t : T_n^* \leq t\}$$
$$\{t : T_n^{**} < t\} \subset \left\{t : \sum_{i=1}^n \psi(x_i;t) < 0\right\} \subset \{t : T_n^{**} \leq t\}.$$

Thus, at a continuity point of the distribution of T_n^* and T_n^{**},

$$\begin{aligned} \mathsf{P}\left(T_n^* < t\right) &= \mathsf{P}\left(\sum_{i=1}^n \psi(X_i;t) \leq 0\right) \\ \mathsf{P}\left(T_n^{**} < t\right) &= \mathsf{P}\left(\sum_{i=1}^n \psi(X_i;t) < 0\right) \end{aligned} \tag{2.3}$$

The exact distribution of T_n^* and T_n^{**} can be worked out via convolution of the law of

$\psi(X;t)$. To investigate the limiting distribution of T_n, put

$$\lambda(t) = \lambda(t,F) = \mathsf{E}_F\left(\psi(X;t)\right).$$

If $\lambda(t)$ exists and is finite for at least one t, then it exists and is monotone for all t, because of the assumption of monotonicity of ψ, meaning $\psi(x;t) - \psi(x;s) \geq 0$ if $t \leq s$. Its expectation exists (but is possibly infinite).

Theorem 2.4 (Uniqueness of minimizer)
Assume the existence of t_0 such that $\lambda(t) > 0$ for $t < t_0$ and $\lambda(t) < 0$ for $t > t_0$. Then, both T_n^* and T_n^{**} converge in probability and almost surely to t_0.

Remark 2.2
The existence of t_0 is related to **Fisher consistency**.

Definition 2.5 (Fisher consistency)
An M-estimate is called Fisher consistent for the parametric model $\{F(x;\theta);\theta\in\Theta\}$ if

$$\mathsf{E}_{F_{\theta_0}}\left(\psi(X;\theta_0)\right) = \int \psi(x;\theta_0)\,\mathrm{d}F_{\theta_0}(x) = 0$$

for any $\theta_0 \in \Theta$ and if

$$\mathsf{E}_{F_{\theta_0}}\left(\psi(X;\theta_1)\right) = \int \psi(x;\theta_1)\,\mathrm{d}F_{\theta_0}(x) \neq 0, \text{ for all } \theta_1 \neq \theta_0.$$

Proof of Theorem 2.4 This follows from applying the strong (weak) law of large numbers to $n^{-1}\sum_{i=1}^n \psi(x_i;t_0 \pm \varepsilon)$ and the exact distributions in eq. (2.3). ∎

One can view M-estimators under a functional point of view. Let

$$F_n(x) = n^{-1}\sum_{i=1}^n \mathbb{I}_{X_i \leq x}$$

be the empirical distribution function. Then, the estimating equation can be written as an expectation:

$$0 = \frac{1}{n}\sum_{i=1}^n \psi(x_i;T_n) = \int \psi(x;T_n)\,\mathrm{d}F_n(x) = \mathsf{E}_{F_n}\left(\psi(X;T_n)\right).$$

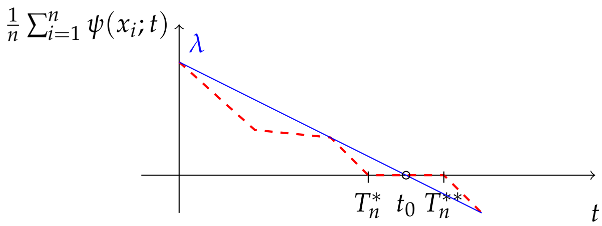
We can extend this functional to a subset $\mathcal{D} \subset \mathcal{M}$, so that $T(F)$ is the solution of $\mathsf{E}_F\left(\psi(X;T(F))\right)$. Using the empirical distribution from the sample, with $T(F_n) \equiv T_n$, we recover the previous result.

Corollary 2.6
If $T(F)$ is uniquely defined, then T_n is consistent for $T(F)$, meaning that $T_n \to T(F)$ where $\to$ is either almost sure convergence, $\xrightarrow{\text{a.s.}}$, or convergence in probability, $\xrightarrow{\mathsf{p}}$.

To derive the asymptotic distribution of T_n, it is convenient to first study $\lambda(T_n)$. By our theorem, this is a consistent estimate of zero, as $\lambda(t_0) = 0$. Since λ is non-increasing,

$$\{t : -\lambda(T_n) < -\lambda(t)\} \subset \{t : T_n < t\} \subset \{t : T_n \leq t\} \subset \{t : -\lambda(T_n) \leq -\lambda(t)\}$$

We will show that $-\sqrt{n}\lambda(T_n)$ is asymptotically normal. If this holds, an expansion of λ around t_0 will lead to the asymptotic distribution of T_n. This requires that the derivative $\lambda'(t_0) < 0$ exists.



Assumptions:

(A1) $\psi(x;t)$ is measurable in x and nonincreasing in t.
(A2) There is at least one t_0 with $\lambda(t_0) = 0$.
(A3) Let $\Gamma_0 := \{t : \lambda(t) = 0\}$. Then λ is continuous in a neighbourhood of Γ_0.
(A4) $\sigma^2(t) = \mathsf{E}\left(\psi^2(X;t)\right) - \lambda(t,F)^2$ is finite, non-zero and continuous in a neighbourhood of Γ_0.

Theorem 2.7
Under assumptions (A1)-(A4),

$$\mathsf{P}\left(-\sqrt{n}\lambda(T_n) < y\right) \to \Phi\left(\frac{y}{\sigma_0}\right)$$

uniformly in y, where $\sigma_0 = \sigma(t_0)$.

Corollary 2.8
If λ has a derivative $\lambda'(t_0) < 0$, then

$$\sqrt{n}(T_n - t_0) \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma_0^2}{(\lambda'(t_0))^2}\right)$$

with the variance being given by

$$\sigma_{\text{asy}}^2 = \frac{\sigma_0^2}{(\lambda'(t_0))^2} = \frac{\mathsf{E}_F\left(\psi^2(X;t_0)\right)}{\left[-\frac{\partial}{\partial\theta}\mathsf{E}_F\left(\psi(X;\theta)\right)\big|_{\theta=t_0}\right]^2}.$$

Example 2.4
The arithmetic mean corresponds to $\psi(x,\theta) = (x-\theta)$ and $-\frac{\partial}{\partial\theta}\psi(x,\theta) = \pm 1$. Then $\sigma_{\text{asy}}^2 = \mathsf{E}\left((X - T(F))^2\right)$.

Recall we assumed ψ to be monotone. One can look for the root of the equation $\sum_{i=1}^{n} \psi(x_i, T_n) = 0$, and the result stays the same if we multiply this by n^{-1}. This converge in probability to $\lambda(T) = \int \psi(x, T(F))\,\mathrm{d}F(x)$ if $X_i \overset{\text{iid}}{\sim} F$. Even if this is on the asymptotic level, we can look at this in terms of the empirical distribution for finite sample, namely $\lambda(T_n) = \int \psi(x, T_n)\,\mathrm{d}F_n(x)$.

Proof of Theorem 2.7 Let $y \in \mathbb{R}$. It follows

$$\mathsf{P}\left(-\sqrt{n}\lambda(T_n^*) < y\right) = \mathsf{P}\left(T_n^* < t_n\right)$$

where $y = -\sqrt{n}\lambda(t_n)$ and the sequence $\{t_n\}$ exists for sufficiently large n because of the continuity of λ in a neighbourhood of Γ_0.

$$\begin{aligned}
\mathsf{P}\left(T_n^* < t_n\right) &= \mathsf{P}\left(\sum_{i=1}^{n} \psi(X_i, t_n) \leq 0\right) \\
&= \mathsf{P}\left(\sum_{i=1}^{n} (\psi(X_i, t_n) - \lambda(t_n)) \leq -n\lambda(t_n)\right) \\
&= \mathsf{P}\left(\sum_{i=1}^{n} (\psi(X_i, t_n) - \lambda(t_n)) \leq \sqrt{n}y\right) \\
&= \mathsf{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{\psi(X_i, t_n) - \lambda(t_n)}{\sigma(t_n)}\right) \leq \frac{y}{\sigma(t_n)}\right)
\end{aligned}$$

The sequence $Y_{i,n} = [\psi(X_i, t_n) - \lambda(t_n)]/\sigma(t_n)$ for $i = 1, \ldots, n$ contains iid random variables with expectation zero and variance 1 (because of the standardisation). We need to verify a central limit theorem applies.

Lemma 2.9 (Central limit theorem)

$$\mathsf{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_{n,i} \leq z\right) \to \Phi(z)$$

uniformly in z.

Proof of Lemma 2.9 The convergence to normality follows from the Lindeberg condition: for every $\varepsilon > 0$,

$$\mathsf{E}\left(Y_{n,i}^2 \ ; |Y_{n,i}| > \sqrt{n}\varepsilon\right) \to 0$$

as $n \to \infty$. Since λ and σ are continuous, this follows from

$$\mathsf{E}\left(\psi^2(X, t_n) \ ; |\psi(X, t_n)| > \sqrt{n}\varepsilon\right) \to 0$$

as $n \to \infty$ because of the monotonicity assumption. ■

We have thus

$$\mathsf{P}\left(-\sqrt{n}\lambda(T_n) < y\right) \to \Phi\left(\frac{y}{\sigma_0}\right).$$

uniformly in y. Note that any T_n such that $T_n^* \le T_n \le T_n^{**}$ has the same asymptotic behavior. ■

Proof of Corollary 2.8 If $\lambda'(t_0) < 0$, then

$$t_n = t_0 - \frac{y}{\sqrt{n}\lambda'(t_0)} + \mathrm{o}(n^{-1/2})$$

which comes from a Taylor series expansion $\lambda(t_n) = \lambda(t_0) + (t_n - t_0)\lambda'(t_0) + \mathrm{o}(n^{-1/2})$. Thus, taking $y = -\sqrt{n}\lambda(t_n)$ and using the above, we get

$$\begin{aligned}
\mathsf{P}\left(T_n^* < t_n\right) &= \mathsf{P}\left(T_n^* < t_0 - \frac{y}{\sqrt{n}\lambda'(t_0)} + \mathrm{o}(n^{-1/2})\right) \\
&= \mathsf{P}\left(\sqrt{n}(T_n^* - t_0) \le -\frac{y}{\lambda'(t_0)} + \mathrm{o}(1)\right) \\
&\to \Phi\left(\frac{y}{\sigma_0}\right)
\end{aligned}$$

as $n \to \infty$. Using Theorem 2.7, we get

$$\mathsf{P}\left(\sqrt{n}\left(\frac{-\lambda'(t_0)}{\sigma_0}\right)(T_n^* - t_0) \le y\right) \to \Phi(y)$$

as $n \to \infty$. ■

Remark 2.3

1. The trick of using $\lambda(T_n)$ instead of T_n is useful. For example, in the location case, $\psi(x,t) = \psi(x-t)$, we have $\lambda_F(t) = \int \psi(x-t) f(x)\,\mathrm{d}x = \int \psi(x) f(x+t)\,\mathrm{d}x$. [2]
2. The general result and a longer list of conditions is in Huber (1967).
3. (a) The result says $\sqrt{n}(T_n - T(F)) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{\text{asy}})$ where

$$\sigma^2_{\text{asy}} = \frac{\sigma_0^2}{(\lambda'(t_0))^2}$$

   (b) This type of limit allows us to compute confidence intervals with the help of the limit and apply them to T_n.

[2]We can then differentiate under the integral sign without problems; λ is smoother as we integrate over a smooth density.

(c) We can also say

$$T_n \dot\sim \mathcal{N}\left(T(F), \frac{\sigma^2_{\text{asy}}}{n}\right)$$

for n large enough, but it does not mean that $\mathsf{Var}\,(T_n) = \sigma^2_{\text{asy}}/n$, nor that $\mathsf{E}\,(T_n) = T(F)$.

(d) A more intuitive proof in the case of a consistent M-estimate can be based on a Taylor series expansion of the estimating equation. T_n is in a neighbourhood of $T(F)$, in an interval which is of length $\mathrm{O}_\mathsf{p}(n^{-1/2})$, meaning

$$\sum_{i=1}^n \psi(x_i, T_n) = 0 = \sum_{i=1}^n \psi(x_i, T(F)) + \left(\frac{\partial}{\partial\theta}\sum_{i=1}^n \psi(x_i,\theta)\bigg|_{T(F)}\right)(T_n - T(F)) + \mathrm{o}(n^{-1/2}). \tag{2.4}$$

This is hand-wavy, but one can look analysed rigorously. Rewriting this is

$$\sqrt{n}(T_n - T(F)) = \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n \psi(x_i, T(F))}{-\frac{\partial}{\partial\theta}\frac{1}{n}\sum_{i=1}^n \psi(x_i,\theta)\Big|_{T(F)}}$$

after multiplying by $\sqrt{n}$ on each side of eq. (2.4). Using the central limit theorem, the numerator converges in law, under mild regularity conditions (given in Huber), to

$$\mathcal{N}\left(\mathsf{E}_F\,(\psi(X,T(F))), \mathsf{E}\left(\psi^2(X,T(F))\right)\right);$$

given that $\mathsf{E}_F\,(\psi(X,T(F))$=0, the expression for the variance simplify. Using a law of large number, the denominator converges to

$$\int -\frac{\partial}{\partial\theta}\psi(x,\theta)\bigg|_{T(F)} \mathrm{d}F(x).$$

This leads to $\sigma_0^2/(\lambda'(t_0))^2$, but only if

$$\frac{\partial}{\partial\theta}\int \psi(x,\theta)\bigg|_{T(F)} \mathrm{d}F(x) = \int \frac{\partial}{\partial\theta}\psi(x,\theta)\bigg|_{T(F)} \mathrm{d}F(x).$$

The asymptotic variance of an M-estimate (if it exists), is of the form

$$\frac{\mathsf{E}\,(\psi^2(X,T(F)))}{\left(\mathsf{E}\left(-\frac{\partial}{\partial\theta}\psi(x,\theta)\Big|_{\theta=T(F)}\right)\right)^2} = \frac{\int \psi^2(x,T(F))\,\mathrm{d}F(x)}{\left(\int -\frac{\partial}{\partial\theta}\psi(x,\theta)\Big|_{\theta=T(F)}\,\mathrm{d}F(x)\right)^2}.$$

4. Multivariate case: if we take $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top$, then $\boldsymbol{\psi}(x, \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \varrho(x, \boldsymbol{\theta}) \in \mathbb{R}^p$ is the gradient of ϱ. The Taylor expansion leads to

$$\left(- \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\psi}(x, \boldsymbol{\theta}) \bigg|_{\boldsymbol{\theta}=T(F)} \right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{\psi}(x_i, T(F)) \right) + \mathrm{o}(n^{-1/2})$$

where $\sum_{i=1}^{n} \boldsymbol{\psi}(x_i, T(F)) \in \mathbb{R}_p$ and the first term is an $p \times p$ matrix. This yields

$$\sqrt{n}(T_n - T(F)) = \left(- \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\psi}(x, \boldsymbol{\theta}) \bigg|_{\boldsymbol{\theta}=T(F)} \right)^{-1} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{\psi}(x_i, T(F)) \right) + \mathrm{o}(1)$$

This time, the asymptotic variance is $\mathbf{B}^{-1}\mathbf{A}\mathbf{B}^{-1}$ where the symmetric $p \times p$ matrices $\mathbf{A}, \mathbf{B}$ are given by

$$\mathbf{A} = \int \boldsymbol{\psi}(x, T(F)) \boldsymbol{\psi}^\top(x, T(F)) \, \mathrm{d}F(x)$$
$$\mathbf{B} = \int - \frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\psi}(x, \boldsymbol{\theta}) \bigg|_{T(F)} \mathrm{d}F(x)$$

2.3. L-estimators and R-estimators

L-estimates are linear combinations of order statistics, $(x_1, \ldots, x_n) \mapsto (x_{(1)}, \ldots, x_{(n)})$ where $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ are ordered. An example of such estimator is the trimmed mean, for which $\alpha \cdot 100\%$ of the observations at each end are deleted and the mean of the remaining sample is taken. This can be written as

$$T_n = \frac{\sum_{i=1}^{n} a_{ni} x_{(i)}}{\sum_{i=1}^{n} a_{ni}}$$

where $a_{ni} = \mathbb{I}_{\alpha n < i < (1-\alpha)n}$, that is 1 if $\alpha n < i < (1-\alpha)n$ and zero otherwise.

R-estimates are estimates derived from rank tests. To test the point null hypothesis $\mathscr{H}_0$: {the median is θ}, one can use for example the sign test,

$$s_\theta(x_1, \ldots, x_n) = \sum_{i=1}^{n} \mathbb{I}_{x_i - \theta},$$

thus s_θ is the number of x_i greater than θ. If $\theta \to -\infty, S_\theta \to n$, while if $\theta \to \infty, S_\theta \to 0$. In the neighbourhood of the median, $S_\theta \dot{\sim} n/2$.

An R-estimate could be such that $s_\theta(x_1, \ldots, x_n)$ is close to $n/2$.

2.4. Influence function

This concept is due to Hampel (1974), who published material from his thesis in JASA.

Definition 2.10 (Influence function)
Suppose $T(F)$ is a **statistical functional**, that is $T : \mathcal{D} \subseteq \mathcal{M} \to \mathbb{R}$. The **influence function** $\mathsf{IF}_T(x, F)$ at $x \in \mathbb{R}$ of T evaluated at the model F, provided it exists, is

$$\mathsf{IF}(x, T, F) = \frac{\mathrm{d}}{\mathrm{d}t} T(F_t)\bigg|_{t=0}$$

where F_t is a mixture law, $F_t = (1-t)F + t\Delta_x$ and Δ_x is the Heavyside function, whose derivative is the Dirac $\delta(x)$, a point mass at x, given by

$$\Delta_x = \begin{cases} 0, & \text{if } u < x \\ 1, & \text{if } u \geq x \end{cases}$$

or $H(u-x)$, where $H(u)$ is the Heavyside function

$$H(u) = \begin{cases} 0 & \text{if } u < 0 \\ 1 & \text{if } u \geq 0 \end{cases}$$

which is a jump function of height 1 at $u = 0$.

The distribution F_t mixes the model F and the point mass at x. A random variable $Y \sim F_t$ can be simulated in two steps:

(i) Generate $u \sim \mathcal{U}(0, 1)$
(ii) If $u \leq 1 - t$, sample $y \sim F$ and if $u > 1 - t$, set $y = x$.

F_t is a perturbed version of F or more precisely a contaminated version of F.

Example 2.5 (Influence function of M-estimates)
$T(F)$ is implicitly defined as a root of

$$\begin{aligned}
0 = \lambda_{F_t}(T(F_t)) &= \int \psi(u, T(F_t))\,\mathrm{d}F_t \\
&= (1-t)\int \psi(u, T(F_t))\,\mathrm{d}F(u) + t\int \psi(u, T(F_t))\delta_x(u) \\
&= (1-t)\lambda_F(T(F_t)) + t\psi(x, T(F_t)).
\end{aligned}$$

We need the derivative of $\lambda_{F_t}(T(F_t))$ with respect to t:

$$-\lambda_F(T(F_t)) + (1-t)\lambda'_F(T(F_t))\frac{\mathrm{d}}{\mathrm{d}t}T(F_t) + \psi(x, T(F_t)) + t\,\frac{\partial}{\partial\theta}\psi(x,\theta)\bigg|_{T(F_t)}\frac{\mathrm{d}}{\mathrm{d}t}T(F_t) = 0$$

At $t = 0$, we get

$$\lambda_F'(T(F))\mathsf{IF}(x) + \psi(x, T(F))$$

since $\lambda_F(T(F)) = 0$. The influence function, it it exists, is thus equal to

$$\mathsf{IF}(x) = \frac{\psi(x, T(F))}{-\lambda_F'(T(F))}.$$

If we can interchange the integral and the differential operators, the denominator becomes

$$-\lambda_F'(T(F)) = -\int \frac{\partial}{\partial\theta}\psi(x,\theta)\bigg|_{\theta=T(F)} \mathrm{d}F(x),$$

which may be more easily computed.

Example 2.6 (ψ functions of common M-estimates)

- The arithmetic mean has $\psi(x, \theta) = x - \theta$, and each new observation will pull the mean on either side of θ.
- The median is $\psi(x, \theta) = 2H(x - \theta) - 1$
- Huber function is $\psi_k(x, \theta) = \min(\max(x - \theta, -k), k)$

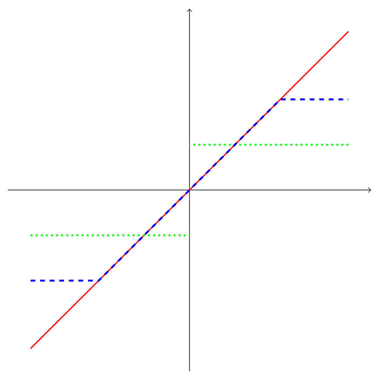


Figure 2: Plot of $\psi(x, \theta)$ functions for the mean (full red), the median (dotted green) and Huber's function with $k = 2$ (dashed blue).

2.4.1. Functional interpretation

The operator appearing in the definition of the influence function,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}T(F_t)\bigg|_{t=0} &= \lim_{t\to 0^+}\frac{T(F_t) - T(F)}{t}\\ &= \lim_{t\downarrow 0}\frac{T(F + t(\Delta_x - F)) - T(F)}{t},\end{aligned}$$

is the **Gâteaux derivative** of the functional T evaluated at F in the direction of $\Delta_x - F$.

$$\mathrm{d}T(F; G) := \lim_{t \to 0} \frac{T((1-t)F + tG) - T(F)}{t}.$$

what is itself a function in G.

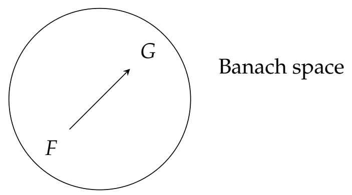


Figure 3: Schematic representation of Gâteau derivative in a Banach space

Fréchet differentiability

The familiar meaning of differentiability demands more than the mere existence of the Gâteaux limit

$$\lim_{t \to 0} \frac{T((1-t)F + tG) - T(F)}{t}.$$

For this reason, we introduce the **Fréchet differential** of a functional.

Definition 2.11 (Fréchet differential)

A statistical function $\theta(F)$ is Fréchet differentiable at $G \in \mathcal{M}$ if there exists a bounded linear function $L_G(F)$ such that

$$|\theta(F) - \theta(G) - L_G(F - G)| = \mathrm{o}(d(F, G))$$

where d is a distance.

Remark 2.4

1. If θ is Fréchet differentiable at G, then θ is Gâteaux differentiable at G.
2. A bounded linear functional defined on a subset of $\mathcal{M}$ has the form

$$L_F(G) = \int \psi_F(x)\, \mathrm{d}G(x),$$

where ψ is bounded.

3. The distance function can be chosen as the Prohorov distance (see remark 2.1).
4. We provide an analogy for $\mathcal{C}^1(\mathbb{R})$ assuming $f : \mathbb{R} \to \mathbb{R}$. The derivative $f'(x)$ exists if and only if

$$\lim_{y \to x} \frac{f(y) - f(x)}{y - x} = f'(x) \;\Leftrightarrow\; |f(y) - f(x) - (y - x)f'(x)| = \mathrm{o}(|x - y|).$$

5. We accept that if $F_t = (1-t)F + t\Delta x$, then $d_P(F, F_t) = \mathrm{o}(t)$.

2.4.2. Asymptotic analysis

With the Fréchet differential, the asymptotic analysis is simplified. Suppose F_n, the empirical distribution based on a sample $x_1, \ldots, x_n$ where $X_i \sim F$. We can express F_n as $F_n = n^{-1} \sum_{i=1}^{n} \Delta_{x_i}$

Lemma 2.12
Assuming that θ is Fréchet differentiable with respect to the Prokhorov distance,

$$\theta(F_n) - \theta(F) = \frac{1}{n} \sum_{i=1}^{n} \mathsf{IF}\,(x_i) + \mathrm{o}(d_P(F_n, F)).$$

Remark 2.5
By Donsker's theorem, $\sqrt{n}(F_n - F)$ converges weakly to a Brownian bridge. From this, it can be shown that $d_P(F_n, F) = \mathrm{O}_p(n^{-1/2})$ and therefore

$$\theta(F_n) - \theta(F) = \frac{1}{n} \sum_{i=1}^{n} \mathsf{IF}\,(x_i) + \mathrm{o}_p(n^{-1/2}).$$

Proof of Lemma 2.12 Because θ is Fréchet differentiable by assumption,

$$\begin{aligned}\theta(F_n) - \theta(F) &= L_F(F_n - F) + \mathrm{o}(d_P(F_n, F)) \\ &= \frac{1}{n} \sum_{i=1}^{n} L_F(\Delta_{x_i} - F) + \mathrm{o}(d_P(F_n, F))\end{aligned}$$

We need to show that $L_F(\Delta_x - F) = \mathsf{IF}_F\,(x)$. As θ is Gâteaux differentiable,

$$\mathsf{IF}_F\,(x) = \lim_{t \to 0} \frac{(\theta((1-t)F + t\Delta_x) - \theta(F))}{t}.$$

θ is also Fréchet differentiable, thus

$$\begin{aligned}\frac{\theta((1-t)F + t\Delta_x) - \theta(F)}{t} &= \frac{1}{t} L_F(t\Delta_x - tF) + \frac{1}{t}\mathrm{o}(t) \\ &= L_F(\Delta_x - F) + \mathrm{o}(1)\end{aligned}$$

using the linearity of L_F and remark 2.4 (5). This implies $L_F(\Delta_{X_i} - F) = \mathsf{IF}_F\,(X_i)$. ■

Lemma 2.13
The expectation of the influence function vanishes over the domain, meaning

$$\int \mathsf{IF}_F\,(x)\ \mathrm{d}F(x) = 0.$$

Proof We know that $\mathsf{IF}_F(X) = L_F(\Delta_X - F)$ and, from remark 2.4 (2),

$$\begin{aligned} L_F(\Delta_x - F) &= \int \psi_F(u)\,\mathrm{d}(\Delta_x - F)(u) \\ &= \psi_F(x) - \int \psi_F(u)\,\mathrm{d}F(u). \end{aligned}$$

implying that

$$\int \mathsf{IF}_F(x)\,\mathrm{d}F(x) = \int \psi_F(x)\,\mathrm{d}F(x) - \int \psi_F(u)\,\mathrm{d}F(u) = 0.$$

■

We are now ready to provide the asymptotic normality of the estimator θ.

Theorem 2.14 (Asymptotic normality of $\theta(F_n)$ assuming Fréchet differentiability) Under the assumption of Fréchet differentiability of $\theta(\cdot)$,

$$\sqrt{n}(\theta(F_n) - \theta(F)) \xrightarrow{d} \mathcal{N}\left(0, \mathsf{E}\left([\mathsf{IF}_F(X)]^2\right)\right).$$

Proof In Lemma 2.12, it can be shown that the error is $\mathrm{o}_p(n^{-1/2})$, so that the asymptotic distribution of $\theta(F_n)$ is the same as the limiting distribution of

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathsf{IF}_F(X_i).$$

We know that $\mathsf{E}(\mathsf{IF}_F(X)) = 0$ by Lemma 2.13 and that $\mathsf{Var}(\mathsf{IF}_F(X)) = \mathsf{E}\left([\mathsf{IF}_F(X)]^2\right)$. Therefore, by the central limit theorem,

$$\sqrt{n}(\theta(F_n) - \theta(F)) = \sqrt{n}\frac{1}{n}\sum_{i=1}^{n} \mathsf{IF}_F(X_i) + o$$

where $o \xrightarrow{\mathsf{P}} 0$ and the statement follows. ■

2.4.3. Interpretation of the influence function

We now tackle the asymptotic approximation of estimates. Any function $T(F)$ such that the corresponding estimator $T(F_n)$ has an asymptotically normal distribution can be locally approximated by an M-estimate, whose asymptotic properties remain the same. We simply have to choose

$$\psi(x, \theta(F)) \propto \mathsf{IF}_f(x).$$

Example 2.7
The trimmed mean can be asymptotically approximated by a Huber estimate.

2.4.4. Qualitative indicators

Various qualitative features are of interest.

1. The boundedness of the **gross-error sensitivity**,

$$\gamma^* = \sup_{x \in \mathbb{R}} |\mathsf{IF}_F(x)|,$$

which is finite for all Fréchet differentiable estimates. If γ^* is infinite, a few outliers can lead to wrong estimates. Estimates with finite γ^* are called B**-robust**.
2. The boundedness of the **local-shift sensitivity**,

$$\lambda^* = \sup_{x \neq y} \frac{|\mathsf{IF}(x) - \mathsf{IF}(y)|}{|x-y|}.$$

If λ^* is large, a small change in an observation may have a large impact.
3. The existence of a **rejection point**

$$\rho^* = \inf\{r > 0 : \mathsf{IF}(x) = 0 \text{ if } |x| > r\}.$$

This concerns estimates that completely reject selected observations.

It is possible to design estimators for which the qualitative indicators are good (heuristically giving ψ that are continuous, bounded and vanish for $|x| > k$ for some $k \in \mathbb{R}$).

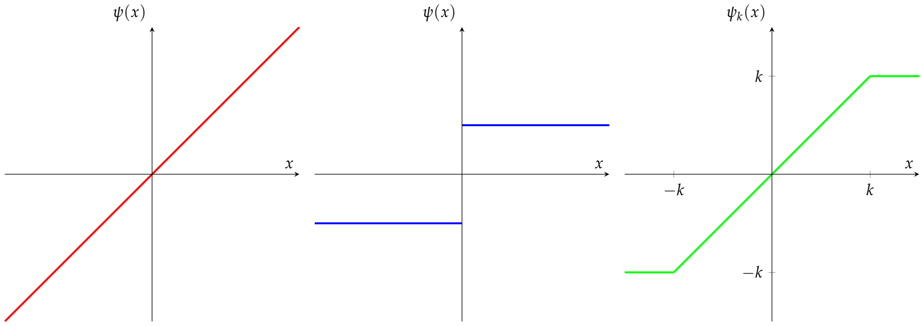
Other than the previously defined function, we introduce the

- skipped mean: $\psi_k(x;\theta) = x\mathbb{I}_{-k \leq x \leq k}$
- Hampel's three points redescending estimator, for which

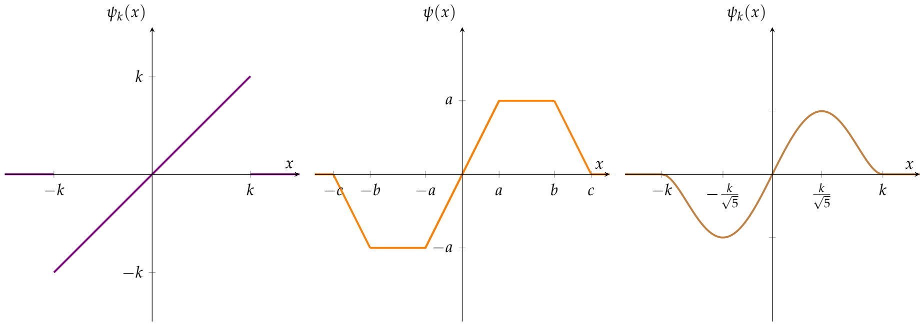
$$\psi_{a,b,c}(x;\theta) = \begin{cases} \frac{a}{c-b}(c+x) & \text{if } -c \leq x \leq b \\ -a & \text{if } -b < x < -a \\ x - \theta & \text{if } -a \leq x - \theta \leq a \\ a & \text{if } a < x < b \\ -\frac{a}{c-b}(c+x) & \text{if } b \leq x \leq c \\ 0 & \text{otherwise} \end{cases}$$

- Tukey's biweight function, given by $\psi_k(x;\theta) = x(1-(x/k)^2)^2\mathbb{I}_{|x|\leq k}$.

In terms of our qualitative indicators, the mean only satisfies (2), the median (1), Huber satisfies (1) and (2), but not (3). Note that the median does not have low local-shift sensitivity, in contrast with Huber function. In contrast, the skipped mean only satisfies



(a) mean *(b)* median *(c)* Huber



(d) skipped mean *(e)* Hampel's three-parts *(f)* Tukey's bisquare

(2) and (3), since it vanishes and thus is bounded. Tukey and Hampel estimators satisfy all the qualitative requirements.

The concept of robustness goes back to precursors such as Daniel Bernoulli (1769) who proposed, rather than using equal weights for the sample, to cover points by a semi-circle and weight the observations accordingly. He did not address how broad the semi-circle should be. This is an example of redescending M-estimator.

The median is an old idea as well, since $\min_\theta \sum_{i=1}^n |x_i - \theta|$ corresponds to minimization of the residuals. In the case of simple linear regression, this leads to $\sum_{i=1}^n |(y_i - a - bx_i)|$. This is undoable, and a partial strategy was proposed by Besikovic.

2.5. Optimal B-robust estimates

2.5.1. B-robust estimates

B-robust means that the estimator has bounded influence, $|\mathsf{IF}_F(x, T)| \leq b < \infty$. Let $F(x;\theta)$ be the model of interest and assume that $\mathrm{d}F(x;\theta) = f(x;\theta)\,\mathrm{d}x$, $\theta \in \Theta = \mathbb{R}$ in the subsequent examples.

A natural question is: find $\psi(x, \theta)$ which satisfies the following:

- Fisher consistency: for all θ, $\int \psi(x,\theta) f(x;\theta)\,\mathrm{d}x = 0$;
- bounded influence: $|\mathsf{IF}_{F(x;\theta)}(x, \psi)| \leq b(\theta)$.

Which ψ minimizes the asymptotic variance subject to these two constraints?

If the boundedness requirement is dropped, then

$$\psi_{\text{MLE}}(x,\theta) \propto \frac{\partial}{\partial\theta}\log f(x;\theta)$$

solves the equation. If ψ_{MLE} is not bounded, then what?

Example 2.8 (Gaussian scale problem)
The model is $F(x;\theta) = \Phi(x/\theta)$, referred to as the normal scale problem. The likelihood score is

$$\begin{aligned}\frac{\partial}{\partial\theta}\log\left(\frac{1}{\theta}\phi\left(\frac{x}{\theta}\right)\right) = -\frac{1}{\theta} - \frac{x}{\theta^2}\frac{\phi'(x/\theta)}{\phi(x/\theta)} = -\frac{1}{\theta} + \frac{x^2}{\theta^3} \\ \propto x^2 - \theta^2.\end{aligned}$$

This is bounded from below, but not from above. We can truncate at k, and this idea does work somehow. There are however two problems with this solution.

First, the influence function

$$\frac{\psi_{\text{MLE}_k}(x,\theta)}{\int -\frac{\partial}{\partial\theta}\psi_{\text{MLE}_k}(x,\theta)f(x;\theta)\,\mathrm{d}x} = \mathsf{IF}_{F(x;\theta)}\left(x,\psi_{\text{MLE}_k}\right)$$

is bounded by a constant b rather than k, since

$$|\mathsf{IF}_\theta\left(x,\psi_{\text{MLE}_k}\right)| \leq \frac{k}{\left|\int -\frac{\partial}{\partial\theta}\psi_{\text{MLE}_k}(x,\theta)f(x;\theta)\,\mathrm{d}x\right|} = b.$$

Second, the estimate $\psi(x,\theta)$ is not automatically Fisher consistent.

$$\int \psi_{\text{MLE}}(x,\theta)\,\mathrm{d}F(x;\theta) = 0 \not\Rightarrow \int \psi_{\text{MLE}_k}(x,\theta)\,\mathrm{d}F(x;\theta) = 0$$

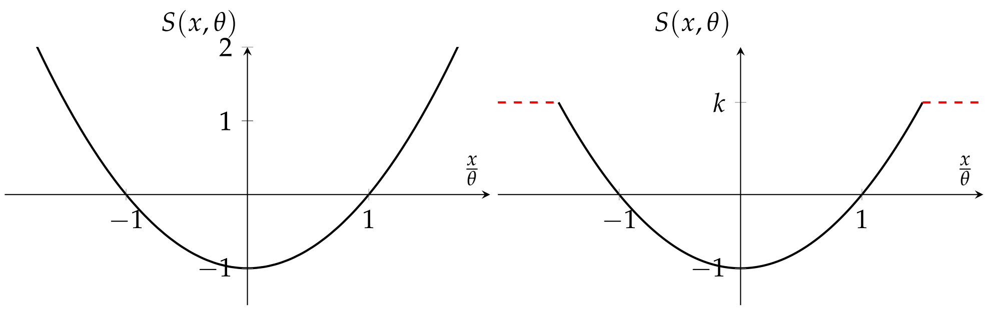


Figure 5: (Truncated) score function for the Gaussian scale problem

Standardisation: from now on, all the functions $\psi(x, \theta)$ we considered will be multiplied by a constant such that

$$\int -\frac{\partial}{\partial \theta} \psi(x, \theta) f(x; \theta) \, \mathrm{d}x = 1.$$

As a consequence, we have

- $\mathsf{IF}_\theta(x, \psi) = \psi(x, \theta)$
- the asymptotic variance of the M-estimator is $\int \psi^2(x, \theta) f(x; \theta) \, \mathrm{d}x$.

From Fisher consistency,

$$\int \psi(x, \theta) f(x; \theta) \, \mathrm{d}x = 0$$

for any ψ. Thus,

$$\frac{\partial}{\partial \theta} \int \psi(x, \theta) f(x; \theta) \, \mathrm{d}x = 0.$$

Under regularity conditions, we can interchange the derivative and the integral signs so as to get

$$\int \frac{\partial}{\partial \theta} \psi(x, \theta) f(x; \theta) \, \mathrm{d}x + \int \psi(x, \theta) \frac{\partial}{\partial \theta} f(x; \theta) \, \mathrm{d}x = 0,$$

which implies that

$$-\int \frac{\partial}{\partial \theta} \psi(x, \theta) f(x; \theta) \, \mathrm{d}x = \int \psi(x, \theta) S(x, \theta) f(x; \theta) \, \mathrm{d}x$$

since $S(x, \theta) = \frac{\partial}{\partial \theta} \log f(x; \theta) = \frac{1}{f(x;\theta)} \frac{\partial f(x;\theta)}{\partial \theta}$. If $\psi(x, \theta) = S(x, \theta)$, then

$$\int -\frac{\partial}{\partial \theta} S(x, \theta) f(x; \theta) \, \mathrm{d}x = \int S^2(x, \theta) f(x; \theta) \, \mathrm{d}x.$$

The numerator and the denominator appearing in the expression for the asymptotic variance of the M-estimator both agree in the case of maximum likelihood if we take ψ to be the score function. It is not true in general for M-estimators that the asymptotic variance will be equal to 1, that is

$$\int -\frac{\partial}{\partial \theta} \psi(x, \theta) f(x; \theta) \, \mathrm{d}x \neq \int \psi^2(x, \theta) f(x; \theta) \, \mathrm{d}x.$$

For standardized ψ-functions, we have

$$1 = \int \psi(x, \theta) S(x, \theta) f(x; \theta) \, \mathrm{d}x.$$

Remark 2.6 (Fisher consistency)
Suppose that $\int \psi(x,\theta) f(x;\theta)\,\mathrm{d}x = a(\theta) \neq 0$. How could one modify $\psi(x,\theta)$ to create a Fisher-consistent version of it? If we can calculate $a(\theta)$, then $\psi_{\text{modif}}(x,\theta) = \psi(x,\theta) - a(\theta)$ will be Fisher consistent. This trick is also used for weighted likelihoods.

2.5.2. Minimum variance B-robust estimation

Lemma 2.15 (Hampel)
Let $F(x;\theta)$ be a regular model with maximum likelihood estimator given by $\psi_{\text{MLE}}(x,\theta)$. Then,

$$\psi_b(x,\theta) \propto \Big[\psi_{\text{MLE}}(x,\theta) - a(\theta)\Big]_{-b}^{b} = \begin{cases} \psi_{\text{MLE}}(x,\theta) - a(\theta) & \text{if } |\psi_{\text{MLE}} - a(\theta)| < b \\ \operatorname{sign}\left(\psi_{\text{MLE}} - a(\theta)\right) b, & \text{otherwise}\,. \end{cases}$$

is the optimal B-robust estimator.

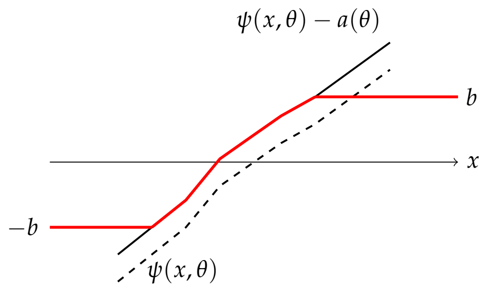


Figure 6: Optimal B-robust estimate.

Proof We construct $\psi(x,\theta)$ for a fixed value of θ.

Let $\psi(x,\theta)$ be an M-estimate such that

$$\begin{aligned} \int -\frac{\partial}{\partial\theta}\psi(x,\theta) f(x;\theta)\,\mathrm{d}x &= 1, \\ \int \psi(x,\theta) f(x;\theta)\,\mathrm{d}x &= 0. \end{aligned}$$

Then

$$\begin{aligned} &\int \left(\psi_{\text{MLE}}(x,\theta) - a(\theta) - \psi(x,\theta)\right)^2 f(x;\theta)\,\mathrm{d}x \\ &\quad= \int \left(\psi_{\text{MLE}}(x,\theta) - a(\theta)\right)^2 f(x;\theta)\,\mathrm{d}x + \int \psi^2(x,\theta) f(x;\theta)\,\mathrm{d}x \\ &\qquad - 2\int \left(\psi_{\text{MLE}}(x,\theta) - a(\theta)\right)\psi(x,\theta) f(x;\theta)\,\mathrm{d}x \\ &\quad= k(\theta) + \int \left(\psi_{\text{MLE}}(x,\theta) - a(\theta)\right)^2 f(x;\theta)\,\mathrm{d}x + \int \psi^2(x,\theta) f(x;\theta)\,\mathrm{d}x, \end{aligned}$$

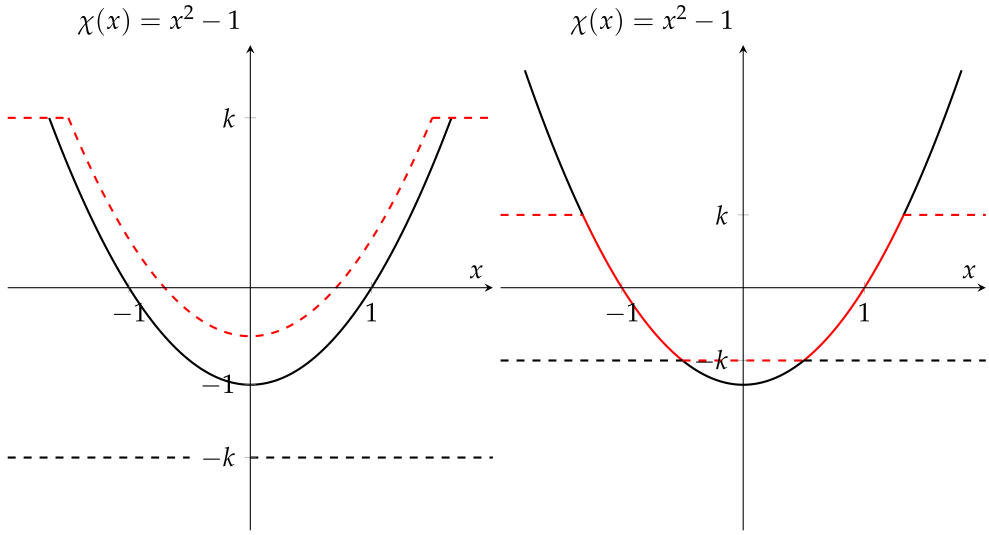


Figure 7: Estimate for the normal scale problem with different values of k. In the left plot, ψ_{MLE_k} is shifted by $a(\theta)$ to preserve Fisher consistency.

a constant in θ plus the asymptotic variance of the ψ-estimates. The cross term is constant because one can eliminate $a(\theta)$ and one can show that

$$\int \psi(x,\theta)\psi_{\mathrm{MLE}}(x,\theta) f(x;\theta)\,\mathrm{d}x = c_{\mathrm{MLE}}(\theta),$$

where $S(x,\theta)c_{\mathrm{MLE}}(\theta) = \psi_{\mathrm{MLE}}(x,\theta)$ and $-2k(\theta) = c_{\mathrm{MLE}}$. Minimizing the asymptotic variance requires minimizing the left hand side. ■

Example 2.9 (Normal location and normal scale problems)

The normal location model is $F(x;\theta) = \Phi(x-\theta)$ and $\psi(x,\theta) = \psi(x-\theta)$. It is enough to construct ψ for $\theta = 0$, as only one function $\chi(u)$ needs to be determined. For the normal scale,

$$\psi(x,\theta) = \left(\frac{x}{\theta}\right)^2 - 1 = \chi\left(\frac{x}{\theta}\right).$$

The corresponding estimate $\sum_{i=1}^{n} \chi(x_i/\widehat{\theta}_n) = 0$. The a has a multiplicative effect on this estimate. With a smaller bound, the negative values of $\chi(x)$ are also affected (see the right panel of fig. 7). Note that the estimate in fig. 7 is linked to Huber's second proposal, which looks like the square of the Huber estimator.

If b tends to zero, the corresponding estimate satisfies $\sum_{i=1}^{n} \chi(x_i/\widehat{\theta}) = 0$. $\widehat{\theta}_n$ is such that half of the x_i satisfy $|x_i| \leq \widehat{\theta}_n$. Thus, $\widehat{\theta}_n = \mathrm{med}(|x_1|,\ldots,|x_n|)$.

This has been generalized to the model where both the location and the scale are un-

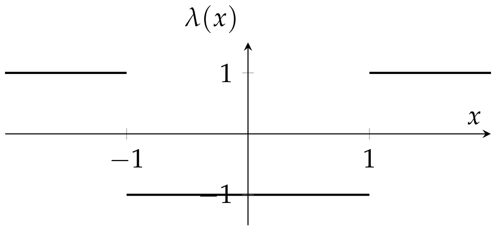


Figure 8: Plot of the limiting estimate when $b \to 0$

known. The estimator

$$\widehat{\theta}_n = \text{median}(|x_1 - \text{med}|, \dots, |x_n - \text{med}|),$$

where med $= \text{median}(x_1, \dots, x_n)$, is the median absolute deviation (from the median) estimator, also known as MAD. The MAD is such that half of the observations are between $\pm$ MAD. This can also be achieved with quantiles. The interquartile range $q_{75\%} - q_{25\%}$ also covers the middle half of the distribution. For a symmetric distribution, $\frac{1}{2}(q_{75\%} - q_{25\%}) \cong$ MAD (they have the same asymptotic variance). The problem now boils down to the estimation of $q_{75\%}$. The choice is not trivial.

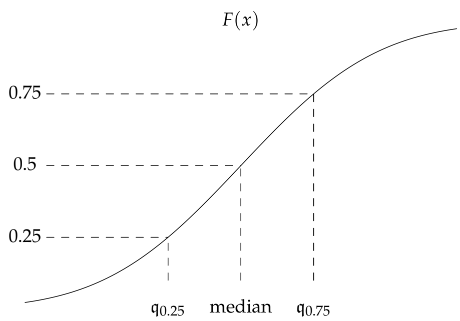


Figure 9: Quartiles of a distribution function $F(x)$

Breakdown point

The following part was covered solely in the exercise session and is included merely for completeness.

Definition 2.16 (Breakdown point)

Let T be a statistical functional. The **breakdown point** BP of T for the distribution F is defined as

$$\mathsf{BP} := \min\left\{\varepsilon : \sup_H \{|T[(1-\varepsilon)F + \varepsilon H]|\} = \infty\right\},$$

where H is any distribution function. The term $\sup_H \{|T[(1-\varepsilon)F + \varepsilon H]|\}$ is the maximum (asymptotic) bias of θ.

The finite sample equivalent notion for an n-sample is defined analogously. We consider the notion in the case of ε-replacement, which gives

$$\mathsf{BP}_n := \frac{1}{n} \min \left\{ k : \sup_{\mathcal{X}_{n-k}, \mathcal{Y}_k} \{|T(\mathcal{X}_{n-k}, \mathcal{Y}_k) - T(x_1, \ldots, x_n)|\} = \infty \right\}$$

for $\mathcal{X}_{n-k}$ a subset of size $n-k$ from $\{x_1, \ldots, x_n\}$ and $\mathcal{Y}_k$ subset of k observations from an arbitrary distribution.

High breakdown is achieved by the median and the LMS, notably, which have $\varepsilon = 0.5$. High breakdown estimates are difficult to compute because of the non-convexity of the optimisation problem.

2.6. Robust minimax theory

Huber considered the following game with two players, one of which chooses a distribution, while the other chooses an estimator. The models player 2, Nature, is allowed to choose are $F(x;\theta) = F(x-\theta)$ with $F \approx \Phi$. Player 1, the statistician, chooses an estimator $\psi(x-\theta) = \psi(x,\theta)$. There is a payoff from player 1 to player 2 given by the asymptotic variance of your function under the chosen model, $\mathsf{Var}_F(\psi)$. On the other hand, player 2 must pay an entry cost to play the game.

As we will see, it turns out there is a minimax solution to this problem and an optimal choice of ψ_0 and F_0. The payoff matrix has a saddle point if the distribution $F \in \mathcal{F}$ has special properties, namely $\mathcal{F}$ is convex and compact. This means $\mathsf{Var}_{F_0}(\psi_0) \leq \mathsf{Var}_{F_0}(\psi)$ for all ψ and $\mathsf{Var}_{F_0}(\psi_0) \geq \mathsf{Var}_F(\psi_0)$ for all F. For couples (ψ, F) we therefore have

$$\mathsf{Var}_F(\psi_0) \leq \mathsf{Var}_{F_0}(\psi_0) \leq \mathsf{Var}_{F_0}(\psi). \tag{2.5}$$

Player 2 maximizes his gain by choosing F_0, while player 1 minimizes his loss by taking ψ_0. Thus $\mathsf{Var}_{F_0}(\psi_0) = \min_\psi \max_F \mathsf{Var}_F(\psi)$. The optimal choice of ψ_0 is ψ_{MLE} under F_0, thus the likelihood score for F_0, that is $-\frac{\mathrm{d}}{\mathrm{d}x}\log f(x) = -f'(x)/f(x)$. The other one is not so easy. We want $\mathsf{Var}_F(\psi_0) \leq \mathsf{Var}_{F_0}(\psi_0)$. Let ψ_F be the MLE for the distribution F. Then

$$\frac{1}{\mathcal{I}(F)} = \mathsf{Var}_F(\psi_F) \leq \mathsf{Var}_F(\psi_0) \leq \mathsf{Var}_{F_0}(\psi_0) = \frac{1}{\mathcal{I}(F_0)}$$

or $\mathcal{I}(F_0) \leq \mathcal{I}(F)$ for all F. The saddlepoint is determined by the least-informative distribution F_0, that is, the one minimizing $\mathcal{I}(F)$ over the set $\mathcal{F}$.

It is not difficult to show that $\mathcal{I}(F)$ is a convex functional. So, if $\mathcal{F}$ is convex and

$F, G \in \mathcal{F}$, then $\mathcal{I}(F_t) = \mathcal{I}((1-t)F + tG)$ is a convex function in t.

Theorem 2.17 (Huber's minimax solution under ε-contaminated normal distributions) Let

$$\mathcal{F} = \{(1-\varepsilon)\Phi + \varepsilon H : H \text{ absolutely continuous and symmetric around } 0\}.$$

The optimal robust estimate ψ_0 for the least informative ε-contaminated distribution $F_0 \in \mathcal{F}$ is

$$\psi_0(x) = \min(k, \max(x, -k))$$

where

$$2\frac{\phi(k)}{k} - 2\Phi(-k) = \frac{\varepsilon}{1-\varepsilon}. \tag{2.6}$$

Proof of Theorem 2.17 The optimal f_0 turns out to be

$$f_0(x) = \begin{cases} (1-\varepsilon)\phi(k)\exp(-k(|x|-k)) & \text{if } |x| > k, \\ (1-\varepsilon)\phi(x) & \text{if } |x| \leq k, \end{cases}$$

and accordingly, since ψ_0 should be bounded for the variance to be bounded,

$$\psi_0(x) = \frac{\frac{\mathrm{d}}{\mathrm{d}x} f_0(x)}{f_0(x)} = \begin{cases} -\frac{\mathrm{d}}{\mathrm{d}x}\log f_0(x) = -k & \text{if } x < -k, \\ \phi'(x)/\phi(x) & \text{if } |x| \leq k, \\ \frac{\mathrm{d}}{\mathrm{d}x}\log f_0(x) = k & \text{if } x > k. \end{cases}$$

We want

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{I}((1-t)F_0 + tG)\bigg|_{t=t_0} > 0$$

for all G. This means $\int (2\psi_0' - \psi_0^2)(g - f_0)\,\mathrm{d}x \geq 0$. This inequality holds true for the following f_0:

Conditions

(a) there is a $k > 0$ such that $f_0(x) = (1-\varepsilon)\phi(x)$ on $[-k, k]$ and $(g - f_0)(x) \geq 0$ for $-k \leq x \leq k$.

(b) $2\psi_0' - \psi_0^2$ is constant for $x \in (-\infty, -k) \sqcup (k, \infty)$.

For this choice,

$$\int (2\psi_0' - \psi_0^2 + k^2)(g - f_0)\,\mathrm{d}x \geq 0,$$

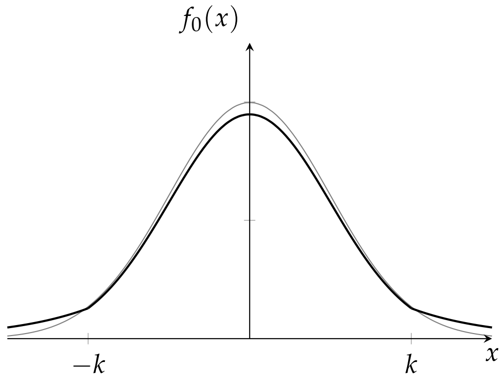


Figure 10: Optimal choice of f_0

since integrating a constant against a difference of densities is zero. ψ_0' is zero because the function is constant, $\psi_0^2 = k^2$. While the right hand side of eq. (2.5) is easy to prove if ψ_0 is the maximum likelihood estimator under F_0 (exercises), the second inequality is more difficult to prove for a given F. We need to introduce a lemma.

Lemma 2.18
If $\mathcal{F}$ is convex and $F, F_0 \in \mathcal{F}$ and $\psi_0 \equiv \psi_{\text{MLE}}(F_0)$, then

1. We have
$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{\mathsf{Var}_{F_t}(\psi_0)}\bigg|_{t=0} = \int (2\psi_0' - \psi_0^2)(f - f_0)(x)\,\mathrm{d}x$$
where $F_t(x) = (1-t)F_0(x) + tF(x)$ and similarly $f_t(x) = (1-t)f_0(x) + tf(x)$.
2. $1/\mathsf{Var}_{F_t}(\psi_0)$ is convex.

Proof of Lemma 2.18 We have using the quotient rule that

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{\mathsf{Var}_{F_t}(\psi_0)} &= \frac{\mathrm{d}}{\mathrm{d}t}\frac{\left(\int \psi_0' f_t\right)^2}{\int \psi_0^2 f_t}\bigg|_{t=0} \\
&= \frac{2\left(\int \psi_0' f_t\right)\left(\int \psi_0'(f - f_0)\right)\left(\int \psi_0^2 f_t\right) - \left(\int \psi_0' f_t\right)^2 \int \psi_0^2 (f - f_0)}{\left(\int \psi_0^2 f_t\right)^2}\Bigg|_{t=0} \\
&= \frac{2\left(\int \psi_0' f_0\right)\left(\int \psi_0'(f - f_0)\right)\left(\int \psi_0^2 f_0\right)}{\left(\int \psi_0^2 f_0\right)^2} - \frac{\left(\int \psi_0' f_0\right)^2 \int \psi_0^2 (f - f_0)}{\left(\int \psi_0^2 f_0\right)^2} \\
&= \int (2\psi_0{}' - \psi_0^2)(f - f_0)
\end{aligned}$$

with a slight abuse of notation. In the last step, we used that fact that $\psi_0 = \psi_{\text{MLE}}(F_0)$

and thus $\int \psi_0' f_0 = \int \psi_0^2 f_0$. We get

$$\frac{1}{\mathsf{Var}_{F_t}(\psi_0)} = \frac{u(t)^2}{v(t)} = w(t)$$

where u, v are linear in t, $v > 0$ and one can easily show that

$$w''(t) = \frac{2(u'v - uv)^2}{v^3} \geq 0.$$

■

Remark 2.7

1. If $\mathcal{F}$ is convex, then $1/\mathsf{Var}_F(\psi_0)$ is convex. If f_0 is such that for all f such that for all $F \in \mathcal{F}$, then

$$\int (2\psi_0' - \psi_0^2)(f - f_0) \geq 0, \tag{2.7}$$

then f is optimal. Therefore, $\mathsf{Var}_F(\psi_0) \leq \mathsf{Var}_{F_t}(\psi_0)$.

2. One can show that maximizing $\mathsf{Var}_F(\psi_0)$ is equivalent to maximizing $\mathsf{Var}_F(\psi_F)$, which in turn is equivalent to minimizing the information $\mathcal{I}$. Indeed, since it holds true that $\mathcal{I}(F) = 1/\mathsf{Var}_F(\psi_F)$ and if $\mathcal{F}$ is convex, then

$$\frac{\mathrm{d}}{\mathrm{d}t} \frac{1}{\mathsf{Var}_{F_t}(\psi_0)}\bigg|_{t=0} = \frac{\mathrm{d}}{\mathrm{d}t}\mathcal{I}(F)\bigg|_{t=0}, \qquad \text{with } \mathcal{I}(F) \text{ convex.}$$

We want to show eq. (2.5). The left hand side follows from an exercise, in which it is shown that $\psi_0(x) = \psi_{\mathrm{MLE}}(x; F_0) = \log f_0(x)$. For the right hand side, we use Lemma 2.18. We look for f_0 such that for all f such that $F \in \mathcal{F}$, eq. (2.7) is satisfied, and then remark 2.7 (1) leads to the result. We see that

$$2\psi_0'(x) - \psi_0^2(x) = -k^2, \quad \text{if } |x| \geq k \tag{2.8a}$$

$$f(x) - f_0(x) \geq 0, \quad \text{if } |x| \leq k \tag{2.8b}$$

Then

$$\begin{aligned}\int \left(2\psi_0'(x) - \psi_0^2(x)\right)(f - f_0) &= \int \left(2\psi_0'(x) - \psi_0^2(x) + k^2\right)(f - f_0) \\ &= \int_{-k}^{k} \left(2\psi_0'(x) - \psi_0^2(x) + k^2\right)(f - f_0) \geq 0\end{aligned}$$

using eq. (2.8b) as both $2 - x^2 + k^2 \geq 0$ and $f - f_0 \geq 0$ if $|x| < k$.

ε	k	$1/\mathcal{I}(F_0)$
0	∞	1
0.01	1.945	1.065
0.05	1.399	1.256
0.10	1.140	1.490

Table 1: ε contaminated normal distributions least informative for location.

It remains to prove eq. (2.6). Since f_0 is a valid density,

$$\begin{aligned}
& 1 = \int_{-k}^{k} (1-\varepsilon)\phi(x)\,\mathrm{d}x + 2\int_{k}^{\infty} (1-\varepsilon)\phi(k)\exp(-k(x-k))\,\mathrm{d}x \\
\Leftrightarrow\quad & 1 = (1-\varepsilon)\left\{\Phi(x) - \Phi(-k)\right\} + 2(1-\varepsilon)\phi(k)\left[-\frac{1}{k}\exp(-k(x-k))\right]_{k}^{\infty} \\
\Leftrightarrow\quad & 1 = (1-\varepsilon)\left(1 - 2\Phi(-k) + 2\frac{\phi(k)}{k}\right) \\
\Leftrightarrow\quad & \frac{\varepsilon}{1-\varepsilon} = -2\Phi(-k) + 2\frac{\phi(k)}{k}
\end{aligned}$$

■

Remark 2.8
Huber guessed the result. He supposed that in the middle, $f_0(x) = (1-\varepsilon)\psi(x)$ for $|x| \leq k$ and that the tails are such that the difference $2\psi_0'(x) - \psi_0^2(x)$ is constant for all $|x| > k$. We only need to verify this is indeed the case.

Remark 2.9
We need to choose ε, which readily gives by eq. (2.6) a value of k. The rightmost column of table 1 gives the asymptotic variance when choosing the optimal estimator for F_0. Table 1 is extracted from Huber & Ronchetti (2009), Exhibit 4.3.

Theorem 2.19 (Minimax solution under bounded variance)
Let

$$\mathcal{F} = \left\{ F \text{ symmetric, absolutely continuous with } \int x^2 f(x)\,\mathrm{d}x \leq b \right\}.$$

Then, the minimax solution (ψ_0, F_0) comprises the optimal robust estimate $\psi_0(x) = x$ and the least informative distribution given by $f_0(x) = \phi(x/\sqrt{b})/\sqrt{b}$, the density of $\mathcal{N}(0, b)$.

Proof We want to show eq. (2.5). Again, the right hand side follows from the proof that $\psi_0(x)$ is the MLE for F_0, which is $S(x)$. For the left hand side of eq. (2.5), we know that

$\int f(x)\,\mathrm{d}x = 1$, $\int xf(x)\,\mathrm{d}x = 0$ and $\int x^2 f(x)\,\mathrm{d}x \leq b$. We want eq. (2.7) and suppose that

$$\begin{aligned} 2\psi_0'(x) - \psi_0^2(x) &= \alpha + \beta x - \gamma^2 x^2, \\ \int x^2 f_0(x)\,\mathrm{d}x &= b. \end{aligned}$$

Then, we have

$$\begin{aligned} &\int (2\psi_0'(x) - \psi_0^2)(x)(f - f_0)(x)\,\mathrm{d}x \\ &\quad = \int (\alpha + \beta x - \gamma^2 x^2)(f - f_0)(x)\,\mathrm{d}x \\ &\quad = \int \alpha(f - f_0)(x)\,\mathrm{d}x + \int \beta x(f - f_0)(x)\,\mathrm{d}x + \gamma^2 b - \gamma^2 \int x^2 f(x)\,\mathrm{d}x \geq 0 \end{aligned}$$

since the first two terms are zero, $\gamma^2 \int x^2 f_0(x)\,\mathrm{d}x = \gamma^2 b$ and $\gamma^2 \int x^2 f(x)\,\mathrm{d}x \leq \gamma^2 b$.

We need to find the corresponding f_0 of ψ_0. We solve $2\psi_0' - \psi_0^2 = \alpha + \beta x - \gamma^2 x$ and guess that $\psi_0(x) = \gamma x + \widetilde{\alpha}$. Finally, we know that

$$\psi_0(x) = \gamma x + \widetilde{\alpha} = \psi_{\mathrm{MLE}}(x; F_0) = S(x),$$

which implies that

$$\begin{aligned} \log(f_0(x)) &= \frac{\gamma}{2} x^2 + \widetilde{\alpha} x + c \\ f_0(x) &\propto \exp\left\{ \frac{\gamma}{2} \left(x - \frac{\widetilde{\alpha}}{\gamma} \right)^2 \right\} \end{aligned}$$

where c is a constant. This is nothing but the kernel of a $\mathcal{N}(\widetilde{\alpha}/\gamma, -1/\gamma)$ distribution. Because of the symmetry requirement of f_0, we have $\widetilde{\alpha} = 0$ and $b = -1/\gamma$ for the variance. This shows that the minimax solution is

$$f_0(x) = \frac{1}{\sqrt{b}} \phi\left(\frac{x}{\sqrt{b}} \right)$$

and $\psi_0(x) = x$. ■

2.7. Robust regression

The classical least squares goes back to Legendre and Gauss. The problem setup is

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$$

for $\boldsymbol{\varepsilon} \in \mathbb{R}^n$, $\boldsymbol{\theta} \in \mathbb{R}^p$ and

$$\widehat{\boldsymbol{\theta}} = \arg\min \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \mathrm{X}_{ij} \theta_j \right)^2 .$$

The solution

$$\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \mathrm{X}_{ij} \widehat{\theta}_j \right) \mathrm{X}_{ik} = 0, \quad k = 1, \ldots, p$$

or $\mathbf{X}^\top \mathbf{X} \widehat{\boldsymbol{\theta}} = \mathbf{X}^\top \boldsymbol{y}$. If $\mathbf{X}$ has full rank, then

$$\widehat{\boldsymbol{\theta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{y}$$

and

$$\widehat{\boldsymbol{y}} = \mathbf{H} \boldsymbol{y} = \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{y}$$

and $\mathbf{H} \in \mathbb{R}^{n \times n}$ is a projection matrix, $\mathbf{HH} = \mathbf{H}$. In particular, $\mathbf{H}$ has p eigenvalues equal to 1 and $(n - p)$ equal to zero, thus $\operatorname{tr}(\mathbf{H}) = p$. This estimator is optimal if $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$, but however non robust: it has breakdown point 0 and an unbounded influence function (exercise). Note that the diagonal elements of $\mathbf{H}$, $h_i := \mathrm{H}_{ii}$, satisfy $0 \leq h_i \leq 1$ and "corresponds to the self-influence of y_i on its own fitted value $\widehat{y}_i$".

What happens if we imagine a sequence of regression problems such that n tends to infinity and p may also grow? The answer is that asymptotic normality and a formula for the asymptotic variance given by $\sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$ both hold, provided we have independent errors from a distribution with finite variance.

Proposition 2.20
Assume that $\varepsilon_1, \varepsilon_2, \ldots$ are independent with mean zero and variance σ^2. Then

$$\widehat{y}_i = \mathbf{X}_i \widehat{\boldsymbol{\theta}} \to \mathsf{E}\left(y_i\right), \quad \text{as } n \to \infty$$

if and only if

$$h = \max_{1 \leq i \leq n} h_i \to 0, \qquad \text{as } n \to \infty.$$

Note that $\mathbf{X}_i$, $1 \leq i \leq n$ denote hereafter $1 \times p$ row vectors.

Proof Sufficiency of $h \to 0$. If $\mathsf{E}\left(\boldsymbol{\varepsilon}\right) = \mathbf{0}$, then $\mathsf{E}\left(\widehat{\boldsymbol{y}}\right) = \mathsf{E}\left(\boldsymbol{y}\right) = \mathbf{X}\boldsymbol{\theta}$. It follows that

$$\mathsf{Var}\left(\widehat{\boldsymbol{y}}\right) = \mathsf{E}\left(\mathbf{H} \boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^\top \mathbf{H}^\top\right) = \mathbf{H} \mathsf{E}\left(\boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^\top\right) \mathbf{H} = \sigma^2 \mathbf{HIH} = \sigma^2 \mathbf{H}.$$

In particular,

$$\widehat{y}_i = \sum_{k=1}^{p} \mathrm{H}_{ik} y_k$$

which implies $\mathsf{Var}\left(\widehat{y}_i\right) = \sum_{k=1}^{p}(\mathrm{H}_{ik})^2\sigma^2 = h_i\sigma_i^2$ using independence of the ε_i and since $\mathbf{H}$ is symmetric and idempotent. Then, by Chebyshev's inequality,

$$\mathsf{P}\left(\left|\widehat{y}_i - \mathsf{E}\left(\widehat{y}_i\right)\right| \geq \delta\right) \leq \frac{h_i\sigma^2}{\delta^2}.$$

The necessity of $h \to 0$ is more difficult. Since the fitted values are such that

$$\widehat{\boldsymbol{y}} = \mathbf{H}\boldsymbol{y} = \mathbf{H}(\mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}) = \mathbf{X}\boldsymbol{\theta} + \mathbf{H}\boldsymbol{\varepsilon},$$

we look at

$$\begin{aligned}\widehat{y}_i - \mathsf{E}\left(\widehat{y}_i\right) &= \sum_{k=1}^{p} \mathrm{H}_{ik}\varepsilon_k \\ &= h_i\varepsilon_i + \sum_{k\neq i} \mathrm{H}_{ik}\varepsilon_k.\end{aligned}$$

Then, since both terms are independent, we have the lower bound

$$\begin{aligned}\mathsf{P}\left(\left|\widehat{y}_i - \mathsf{E}\left(\widehat{y}_i\right)\right| \geq \delta\right) &\geq \mathsf{P}\left(h_i\varepsilon_i \geq \delta, \sum_{k\neq i} \mathrm{H}_{ik}\varepsilon_k \geq 0\right) + \mathsf{P}\left(h_i\varepsilon_i \leq -\delta, \sum_{k\neq i} \mathrm{H}_{ik}\varepsilon_k \leq 0\right) \\ &= \mathsf{P}\left(h_i\varepsilon_i \geq \delta\right)\mathsf{P}\left(\sum_{k\neq i} \mathrm{H}_{ik}\varepsilon_k \geq 0\right) + \mathsf{P}\left(h_i\varepsilon_i \leq -\delta\right)\left(1 - \mathsf{P}\left(\sum_{k\neq i} \mathrm{H}_{ik}\varepsilon_k \geq 0\right)\right) \\ &\geq \min\left\{\mathsf{P}\left(\varepsilon_i \geq \frac{\delta}{h_i}\right), \mathsf{P}\left(\varepsilon_i \leq -\frac{\delta}{h_i}\right)\right\},\end{aligned}$$

which completes the proof. ■

Remark 2.10

Note that since the maximum is greater than the average, $\max(h_i) \geq \overline{h_i} = p/n$. This implies that unless p/n goes to zero, consistency of all $\widehat{y}_i$ cannot happen.

One can even show more: if $\boldsymbol{a} \in \mathbb{R}^p$ and such that $\boldsymbol{a}^\top\boldsymbol{a} = 1$, then $\boldsymbol{a}^\top\widehat{\boldsymbol{\theta}}$ is asymptotically normal if and only if $\max(h_i) \to 0$. If it is not the case, there is at least one residual value that does not have expectation zero.

We define the residuals as $r_i = y_i - \widehat{y}_i = (1 - h_i)y_i - \sum_{k\neq i} \mathrm{H}_{ik}y_k$. This is called the i^{th} residual. If h_i is close to 1, then y_i has little influence on the estimate r_i and a gross error in y_i may not be visible in r_i. The error in y_i may however show up in another residual r_k, if H_{ik} happens to be sizeable. There is a large literature on procedures for finding outliers exists (regression diagnostics). Many are based on leaving out one observation and observing the effect.

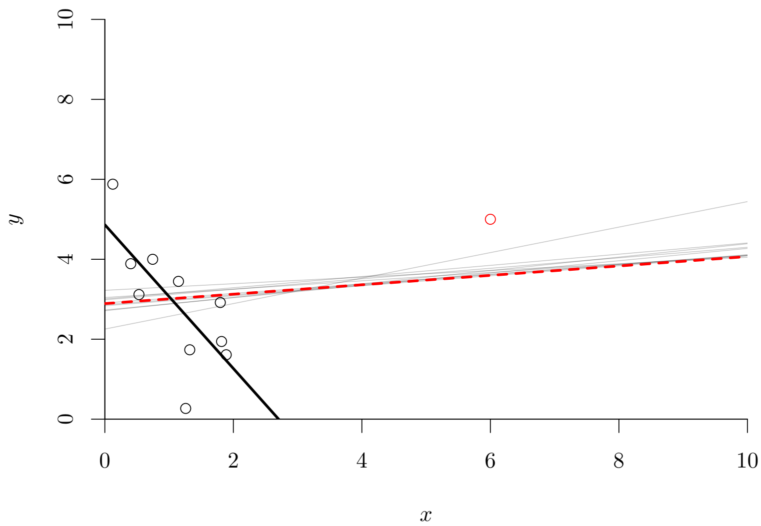


Figure 11: Plot of least square fit with (dashed red) and without (full black). The outlying observations is flagged in red. Grey lines correspond to fit from leave-one-out cross validation.

Without loss of generality, assume that $\mathbf{X}^\top\mathbf{X} = \mathbf{I}$, in which case $H_{ik} = \mathbf{X}_i(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}_k^\top = \mathbf{X}_i\mathbf{X}_k^\top$. We now delete one of the observations: set $\mathbf{X} \to \mathbf{X}_{-i} = \widetilde{\mathbf{X}}$ and $\boldsymbol{y} \to \boldsymbol{y}_{-i}$; we are after $\boldsymbol{\theta} \to \widetilde{\boldsymbol{\theta}}$. Then

$$(\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}) = \sum_{k\neq i}\mathbf{X}_k\mathbf{X}_k^\top = \mathbf{X}^\top\mathbf{X} - \mathbf{X}_i^\top\mathbf{X}_i$$

and

$$(\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}})^{-1} = \mathbf{I} + \frac{\mathbf{X}_i^\top\mathbf{X}_i}{1 - \mathbf{X}_i\mathbf{X}_i^\top}$$

by the Sherman–Morrison formula. It follows that

$$\begin{aligned}\widetilde{\boldsymbol{\theta}} &= \left(\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}\right)^{-1}\widetilde{\mathbf{X}}^\top\widetilde{\boldsymbol{y}} \\ &= \left(\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}\right)^{-1}\sum_{k\neq i} y_k\mathbf{X}_k^\top\end{aligned}$$

$$
\begin{aligned}
&= \left(\mathbf{I} + \frac{\mathbf{X}_i^\top \mathbf{X}_i}{1 - \mathbf{X}_i \mathbf{X}_i^\top}\right)\left(\mathbf{X}^\top \boldsymbol{y} - y_i \mathbf{X}_i^\top\right) \\
&= \widehat{\boldsymbol{\theta}} - y_i \mathbf{X}_i^\top + \mathbf{X}_i^\top \mathbf{X}_i \mathbf{X}^\top \boldsymbol{y} \left(\frac{1}{1-h_i}\right) - \frac{y_i h_i}{1-h_i}\mathbf{X}_i^\top \\
&= \widehat{\boldsymbol{\theta}} + (\widehat{y}_i - y_i)\mathbf{X}_i^\top \frac{1}{1-h_i}.
\end{aligned}
$$

Note that

$$
\mathbf{X}_i \mathbf{X}^\top \boldsymbol{y} = \mathbf{X}_i \begin{pmatrix} \mathbf{X}_1^\top & \mathbf{X}_2^\top & \cdots & \mathbf{X}_p^\top \end{pmatrix} \boldsymbol{y} = \begin{pmatrix} \mathrm{H}_{i1} & \mathrm{H}_{i2} & \cdots & \mathrm{H}_{in} \end{pmatrix} \boldsymbol{y} = \widehat{y}_i,
$$

which implies $\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}} = r_i \mathbf{X}_i^\top / (1 - h_i) \in \mathbb{R}^p$; this is the **sensitivity**. The fitted values satisfy

$$
\mathbf{X}_i \widetilde{\boldsymbol{\theta}} - \mathbf{X}_i \widehat{\boldsymbol{\theta}} = \frac{r_i}{1-h_i} \mathbf{X}_i \mathbf{X}_i^\top = \frac{h_i r_i}{1-h_i}.
$$

Leave-out-one residuals are then

$$
\begin{aligned}
y_i - \mathbf{X}_i \widetilde{\boldsymbol{\theta}} &= y_i - \mathbf{X}_i \left(\widehat{\boldsymbol{\theta}} + \frac{r_i}{1-h_i}\mathbf{X}_i^\top\right) \\
&= r_i + r_i \frac{h_i}{1-h_i} = r_i \frac{1}{1-h_i}.
\end{aligned}
$$

Another diagnostic is Cook's distance, defined as

$$
D_i := \frac{1}{pS^2} \|\widetilde{\boldsymbol{y}} - \widehat{\boldsymbol{y}}\|^2
$$

where $S^2 = (n-p)^{-1} \sum_{i=1}^n r_i^2$ is the empirical variance estimate. Then

$$
\widetilde{\boldsymbol{y}} = \mathbf{X}\widetilde{\boldsymbol{\theta}} = \mathbf{X}\widehat{\boldsymbol{\theta}} + (\widehat{y}_i - y_i)\mathbf{X}\mathbf{X}_i^\top \frac{1}{1-h_i}
$$

and the difference

$$
\widetilde{\boldsymbol{y}} - \widehat{\boldsymbol{y}} = \frac{r_i}{1-h_i} \begin{pmatrix} \mathrm{H}_{1i} \\ \vdots \\ \mathrm{H}_{pi} \end{pmatrix}
$$

with

$$
\begin{aligned}
D_i &= \frac{r_i^2}{(1-h_i)^2} \frac{1}{pS^2} \sum_{k=1}^p \mathrm{H}_{ik}^2 \\
&= \frac{r_i^2 h_i}{(1-h_i)^2} \frac{1}{pS^2}
\end{aligned}
$$

All these efforts are easy to fool through masking! One can think of several outliers occurring in the same region.

2.7.1. *M*-estimate

We can simply replace the least square criterion $\sum_i r_i^2$ by $\sum_i \varrho(r_i)$, where ϱ is convex and exhibits subquadratic growth. This is a proposal of Huber.

The estimating equations become

$$\sum \psi(r_i)\mathbf{X}_i^\top = \sum_{i=1}^n \psi(y_i - \mathbf{X}_i\widehat{\boldsymbol{\theta}})\mathbf{X}_i^\top = \mathbf{0}$$

where $\psi = \varrho'$. In order to achieve the correct scale behaviour, the residuals need to be standardised.

$$\sum_i \psi\left(\frac{r_i}{\widehat{\sigma}}\right)\mathbf{X}_i^\top = \mathbf{0}.$$

The error scale $\widehat{\sigma}$ is either estimated simultaneously or by a preliminary robust estimation! The influence function is

$$\mathsf{IF}(\mathbf{Z}_0, y_0) = \frac{\sigma}{\mathsf{E}\left(\psi'\left(\frac{\varepsilon}{\sigma}\right)\right)}\psi\left(\frac{y_0 - \mathbf{Z}_0^\top\boldsymbol{\theta}}{\sigma}\right)\mathrm{V}_{\mathbf{X}}^{-1}\mathbf{Z}_0, \qquad \mathbf{Z}_0 \in \mathbb{R}^p, y_0 \in \mathbb{R}$$

where $\boldsymbol{\theta}, \sigma$ are the values under the model for which we compute the influence function and $\mathrm{V}_{\mathbf{X}} = \int \mathbf{X}\mathbf{X}^\top \mathrm{d}H(\mathbf{X})$ where $H(\mathbf{X})$ is the distribution of the covariates. Note that $\|\mathsf{IF}(\mathbf{Z}_0, y_0)\|$ is unbounded! It is indeed easy to drive to $\mathrm{V}_{\mathbf{X}}^{-1}\mathbf{Z}_0$ to infinity.

In terms of asymptotics, if $n \to \infty$ and p is constant,

$$\widehat{\boldsymbol{\theta}} \dot{\sim} \mathcal{N}_p\left(\boldsymbol{\theta}, v \cdot (\mathbf{X}^\top\mathbf{X})^{-1}\right)$$

where v is a modifier that depends on the scaled variance,

$$v = \sigma^2 \frac{\mathsf{E}\left(\psi^2\left(\frac{\varepsilon}{\sigma}\right)\right)}{\left[\mathsf{E}\left(\psi'\left(\frac{\varepsilon}{\sigma}\right)\right)\right]^2}.$$

2.7.2. Alternative estimates

Other than *M*-estimators and ordinary least squares (OLS), there are other alternatives for regression. These include ℓ_1 penalties such as least absolute deviation (LAD), which consists of minimizing $\sum_{i=1}^n |r_i|$. The least median of squares (LMS) estimate was the first working proposal in the robust literature and is due to P. Rousseeuw. This estimator minimizes

$$\text{median}\left(r_i^2(\boldsymbol{\theta}), \ldots, r_n^2(\boldsymbol{\theta})\right)$$

where $r_i = r_i(\boldsymbol{\theta}) = y_i - \mathbf{X}_i\boldsymbol{\theta}$. Under some conditions, this is consistent and we have $n^{1/3}(\widehat{\boldsymbol{\theta}}_{\text{LMS}} - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}_p(\mathbf{0}, \boldsymbol{Y})$; note the slower rate of convergence.

MM-estimates

In order to eliminate the influence of leverage points, it seems natural to use redescending ψ-functions.

Example 2.10 (Tukey's bisquare)
Tukey's bisquare function is defined as

$$\psi_{\text{bi}} = x(1-x^2)^2 \mathbb{I}_{|x|\leq 1}.$$

Integrating out ψ_{bi}, we recover

$$\varrho_{\text{bi}}(x) = \begin{cases} \frac{1}{6}\left(1-(1-x^2)^3\right) & \text{if } |x| \leq 1 \\ \frac{1}{6} & \text{otherwise .} \end{cases}$$

The idea is thus to use bounded ϱ-functions and consider for example

$$\sum_{i=1}^{n} \psi_{\text{bi}}\left(\frac{r_i}{\widehat{\sigma}}\right) \mathbf{X}_i^\top = \mathbf{0}.$$

This will have multiple solutions and the difficulty is to select a right one.

Regression estimates based on robust scales are analogous to estimators which minimize the estimated variance of the errors. Using an M-estimator, such that $\varrho(u)$ is a bounded function, the corresponding ψ is redescending. $\widehat{\sigma}_{\text{robust}} = \widehat{\sigma}_{\varrho}$ is such that

$$\frac{1}{n}\sum_{i=1}^{n} \varrho\left(\frac{r_i}{\widehat{\sigma}_{\varrho}}\right) = \delta,$$

where δ can be for example is $1/2$. If we minimize $\widehat{\sigma}_{\varrho}$ with regard to θ, this leads to an M-estimator that is computable and unique. This idea bears the name S-estimator .

Historical notes

M-estimators with unbounded ϱ-function have unbounded influence function. To correct for this, proposals were made involving the weights of the observation depending on their position! Colin Mallows suggested taking

$$\sum_{i=1}^{n} \psi\left(\frac{r_i(\theta)}{\widehat{\sigma}}\right) \mathbf{X}_i^\top \cdot W(\,\mathrm{d}(\mathbf{X}_i)) = 0,$$

where $d(\mathbf{X}_i)$ measures the outlyingness of the observation $\mathbf{X}_i$. For example, one can take the Mahalanobis distance, $(\mathbf{X}_i - \bar{\mathbf{X}})\widehat{\boldsymbol{\Sigma}}^{-1}(\mathbf{X}_i - \bar{\mathbf{X}})^\top$. Another proposal using weights was given by Schweppe, namely

$$\sum_{i=1}^{n} \frac{1}{d(\mathbf{X}_i)} \psi \left(d(\mathbf{X}_i) \frac{r_i(\theta)}{\widehat{\sigma}} \right).$$

Both can have bounded influence. Kraster and Welsch showed that the Schweppe estimator with the Huber ψ-function is the optimal bounded influence estimator.

Part 3
Rank-based statistical procedures

3.1. Introduction

3.1.1. Order statistics, ranks and their properties

Given a vector $x_1, \ldots, x_N = \boldsymbol{x}$, we denote by $o_i(\boldsymbol{x})$ the i^{th} smallest coordinate of $\boldsymbol{x}$. We will write $x_{(i)} = o_i(x)$, meaning that $x_{(1)} \leq \cdots \leq x_{(N)}$. Consider a collection of random variables $X_1, \ldots, X_N$, usually independent and identically distributed. Then $X_{(i)} = o_i(X)$ is the i^{th} is the order statistics. The gap is the distance between two order statistics. For $\boldsymbol{x} = (x_1, \ldots, x_N)$ such that $x_i \neq x_j$ for all i, j with $i \neq j$. Let $r_i(x)$ be the rank of x_i among the x_i's. The $r_i(x)$ is the number of x's that are less than or equal to x_i and $x_i = x_{(r_i)}$. The order statistics and the ranks correspond to a decomposition of the data into ancillary and sufficient statistics and having both is sufficient to recover the sample. We denote by $r_i(X) = R_i$ the rank of X_i among $X_1, \ldots, X_N$. The collection of ranks is $\boldsymbol{R} = (R_1, \ldots, R_N)$ and we denote by $\mathcal{X}$ the set $\mathcal{X} = \{x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(N)}\}$.

Ranks are well-defined only if a tie or a coincidence between two values has probability zero. We will thus assume absolute continuity hereafter.

Theorem 3.1

The joint density of $\boldsymbol{X} = (X_1, \ldots, X_N)$ is $f(x_1, \ldots, x_N)$. It then follows that the joint density of the order statistics $X_{(1)}, \ldots, X_{(N)}$ is equal to

$$f^*(x_{(1)}, \ldots, x_{(N)}) = \sum_{\pi \in \mathcal{S}_N} f(x_{(\pi_1)}, \ldots, x_{(\pi_N)}) \mathbb{I}_{\boldsymbol{x}_{(\cdot)} \in \mathcal{X}}$$

where $\mathcal{S}_N$ is the symmetric group of all permutations of $\{1, \ldots, N\}$, thus $|\mathcal{S}_n| = N!$

Corollary 3.2

If $f(x_1, \ldots, x_N)$ is invariant under permutations of the arguments, then

$$f^*(x_{(1)}, \ldots, x_{(N)}) = N! f(x_{(1)}, \ldots, x_{(N)}) \mathbb{I}_{\boldsymbol{x}_{(\cdot)} \in \mathcal{X}}.$$

Proof of Theorem 3.1

$$\begin{aligned}
&\underset{(X_{(1)}, \ldots, X_{(N)}) \in \mathcal{A}}{\int \cdots \int} f(x_1, \ldots, x_N) \,\mathrm{d}x_1 \cdots \mathrm{d}x_N \\
&\quad= \sum_{\pi \in \mathcal{S}_N} \underset{\substack{(X_{(1)}, \ldots, X_{(N)}) \in \mathcal{A}, \\ \boldsymbol{R} = \boldsymbol{\pi}}}{\int \cdots \int} f(x_1, \ldots, x_N) \,\mathrm{d}x_1 \cdots \mathrm{d}x_N \\
&\quad= \sum_{\pi \in \mathcal{S}_N} \int \cdots \int_{\mathcal{A}} f(x_{(\pi_1)}, \ldots, x_{(\pi_N)}) \,\mathrm{d}x_{(1)} \cdots \mathrm{d}x_{(N)}
\end{aligned}$$

$$= \int \cdots \int_{\mathcal{A}} f^*(x_{(\pi_1)}, \ldots, x_{(\pi_N)})\, \mathrm{d}x_{(1)} \cdots \mathrm{d}x_{(N)}$$

as the Jacobian of the linear transformation of $(x_1, \ldots, x_N) \to (x_{(\pi_1)}, \ldots, x_{(\pi_N)})$ is ± 1 for all events $\{\boldsymbol{R} = \boldsymbol{\pi}\}$. It remains to associate the corresponding densities ■

Theorem 3.3
If $\boldsymbol{X} = (X_1, \ldots, X_N)$ has joint density $f(x_1, \ldots, x_N)$, then

$$\mathsf{P}\left(\boldsymbol{R} = \boldsymbol{\pi} \mid X_{(1)} = x_{(1)}, \ldots, X_{(N)} = x_{(N)}\right) = \frac{f(x_{(\pi_1)}, \ldots, x_{(\pi_N)})}{f^*(x_{(1)}, \ldots, x_{(N)})}.$$

Proof First note that the joint density of $(\boldsymbol{R}, \boldsymbol{X}_{(\cdot)})$ is $f(x_1, \ldots, x_N)$. It follows that

$$\begin{aligned}
&\mathsf{P}\left(\boldsymbol{R} = \boldsymbol{\pi}, (X_{(1)}, \ldots, X_{(N)}) \in \mathcal{A}\right) \\
&\quad = \underset{\substack{(X_{(1)}, \ldots, X_{(N)}) \in \mathcal{A}, \\ \boldsymbol{R} = \boldsymbol{\pi}}}{\int \cdots \int} f(x_1, \ldots, x_N)\, \mathrm{d}x_1 \cdots \mathrm{d}x_N \\
&\quad = \int \cdots \int_{\mathcal{A}} f(x_{(\pi_1)}, \ldots, x_{(\pi_N)})\, \mathrm{d}x_{(1)} \cdots \mathrm{d}x_{(N)} \\
&\quad = \int \cdots \int_{\mathcal{A}} \frac{f(x_{(\pi_1)}, \ldots, x_{(\pi_N)})}{f^*(x_{(1)}, \ldots, x_{(N)})} f^*(x_{(1)}, \ldots, x_{(N)})\, \mathrm{d}x_{(1)} \cdots \mathrm{d}x_{(N)}
\end{aligned}$$

by a change of measure. ■

Corollary 3.4
If $f(x_1, \ldots, x_N)$ is invariant under permutations of the coordinates (this holds true if $f(x_1, \ldots, x_N) = \prod_{i=1}^N f(x_i)$), then the quantities $\boldsymbol{R} = (R_1, \ldots, R_N)$ and $(X_{(1)}, \ldots, X_{(N)})$ are independent with $\mathsf{P}(R = \pi) = \frac{1}{N!}$ for all $\pi \in \mathcal{S}_N$.

In the case of a sample from a symmetric distribution function, $X_i \sim F$, with $f(x)$ is such that $f(-x) = f(x)$ (so that $F(x) + F(-x) = 1$), then $|X_i|$ and

$$\operatorname{sign}(X_i) = \begin{cases} 1 & \text{if } X_i > 0, \\ -1 & \text{if } X_i < 0, \\ 0 & \text{if } X_i = 0, \end{cases}$$

are independent. Indeed, $\mathsf{P}(\operatorname{sign}(X_i) = 1) = \frac{1}{2}$ and

$$\begin{aligned}
\mathsf{P}(|X_i| \leq t, \operatorname{sign}(X_i) = 1) &= \mathsf{P}(0 \leq X_i \leq t) = F(t) - F(0) \\
&= F(t) - \frac{1}{2}.
\end{aligned}$$

On the other hand,

$$\begin{aligned}\mathsf{P}\left(|X_i| \le t, \operatorname{sign}(X_i) = -1\right) &= \mathsf{P}\left(-t \le X_i \le 0\right) = F(0) - F(-t)\\ &= \frac{1}{2} - (1 - F(t)) = F(t) - \frac{1}{2}.\end{aligned}$$

3.2. Examples of rank statistics

A statistic $T(X_1, \ldots, X_N)$ whose value only depends on the ranks is called a **rank statistic**. Note that the ranks $(R_1, \ldots, R_N)$ do not change if we apply a monotone increasing transformation to $X_1, \ldots, X_N$.

3.2.1. One sample statistics

In this case, the signs are used along with the ranks.

Proposition 3.5 (Sign test)
Consider the null hypothesis

$$\begin{aligned}\mathscr{H}_0 &: X_1, \ldots, X_N \overset{\text{iid}}{\sim} F, \qquad \operatorname{med}(\boldsymbol{X}) = \theta_0\\ \mathscr{H}_1 &: X_1, \ldots, X_N \overset{\text{iid}}{\sim} G, \qquad \operatorname{med}(\boldsymbol{X}) \ne \theta_0.\end{aligned}$$

Another popular null hypothesis is that of symmetry, which has

$$\begin{aligned}\mathscr{H}_0 &: X_1, \ldots, X_N \overset{\text{iid}}{\sim} F, \qquad f(x) = f(-x), \quad (\operatorname{med}(\boldsymbol{X}) = 0),\\ \mathscr{H}_1 &: X_1, \ldots, X_N \overset{\text{iid}}{\sim} G, \qquad g(x) \text{ not symmetric.}\end{aligned}$$

The statistic is

$$S(\theta_0) = \sum_{i=1}^{N} \operatorname{sign}(X_i - \theta_0)$$

or

$$\begin{aligned}S(\theta_0) &= \text{number of } X_i > \theta_0\\ &= \sum_{i=1}^{N} H(x_i - \theta_0).\end{aligned}$$

Note that the Heavyside function can be expressed as $H(y) = \mathbb{I}_{y \ge 0} = \frac{1}{2}(\operatorname{sign}(y) + 1)$.

Under $\mathscr{H}_0$, the number of $X_i > \theta_0$ follows a Binomial distribution $\mathcal{B}(N, 1/2)$. In using this as a test, we reject $\mathscr{H}_0 : \theta = \theta_0$ in favor of $\mathscr{H}_1 : \theta < \theta_0$ if $S(\theta_0)$ is sufficiently small.

Table 2: Calculation of the null distribution of the sign test

values of S	0	1	2	$\cdots$	N
$\mathsf{P}_{\mathscr{H}_0}(\cdot)$	$(1/2)^N$	$N(1/2)^N$	$\binom{N}{2}(1/2)^N$	$\cdots$	$(1/2)^N$

For a given $\alpha \in (0,1)$, we look for k_α such that

$$\alpha_0 = \sum_{i=0}^{k_\alpha} \binom{N}{i} \left(\frac{1}{2}\right)^N \leq \alpha.$$

This implies that $\alpha_0 + \binom{n}{k_\alpha+1}(1/2)^N > \alpha$. We then reject $\mathscr{H}_0$ if $S \leq k_\alpha$. This is a test of size $\mathsf{P}_{\mathscr{H}_0}(\text{false rejection}) = \alpha_0$. To get exactly level α, randomization is necessary.

Any null hypothesis $\mathscr{H}_0 : \theta = \theta_0$ can be tested. If we collect all θ_0 not rejected, the two-sided test will generate a confidence interval with confidence (coverage probability) $1 - 2\alpha_0$.

Proposition 3.6 (Wilcoxon signed rank test)
We consider

$$\mathscr{H}_0 : X_1, \ldots, X_N \overset{\text{iid}}{\sim} F, \quad f(\theta_0 - x) = f(\theta_0 + x)$$

the hypothesis of symmetry of f around θ_0. We look at $R_i^+ := r_i(|X_i - \theta_0|)$, the rank of $|X_i - \theta_0|$ among $|X_1 - \theta_0|, \ldots, |X_N - \theta_0|$. Then

$$W^+(\theta_0) = \sum_{i=1}^{N} \mathbb{I}_{X_i > \theta_0} R_i^+.$$

This remarkably simply idea is almost as powerful as the Student's t-test. From the general theory of ranks and signs under symmetry, we know that under $\mathscr{H}_0$, W^+ has a distribution that does not depend on the particular f.

Table 3: Calculation of the null distribution of the signed rank test

values of N	0	1	2	3	4	5	6
$N = 1$	1/2	1/2	0	0	0	0	0
$N = 2$	1/4	1/4	1/4	1/4	0	0	0
$N = 3$	1/8	1/8	1/8	2/8	1/8	1/8	1/8

For general N, there is mass on the first $N(N+1)/2$ observations.

We wish to test the same null hypothesis as that of the Wilcoxon test, namely symmetry. A broader class of statistics can be obtained by assigning a score to the ranks.

Definition 3.7 (Linear rank statistic for symmetry)
A linear rank statistic for symmetry is of the form

$$L_N(\theta_0) = \sum_{i=1}^{N} a(R_i^+)\,\mathrm{sign}(X_i - \theta_0)$$

where $(a(1), \ldots, a(N))$ is a vector of **rank scores**. You can easily prove that under $\mathscr{H}_0$, $\mathsf{E}\,(L(\theta_0)) = 0$ and $\mathsf{Var}\,(L(\theta_0)) = \sum_{i=1}^{N}(a(i))^2$.

3.2.2. Two sample statistics

We mostly consider the setting where the null consists of equality in distribution for the sequences $\boldsymbol{X}, \boldsymbol{Y}$, namely $X_1, \ldots, X_m \overset{\text{iid}}{\sim} F$ and $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} F$. We think often of the alternative $\mathscr{H}_1$ where the density of $\boldsymbol{Y}$ is $f(y_i - \Delta)$ rather than f, but the framework is much more general.

Proposition 3.8 (Wilcoxon two sample test)
This statistic, proposed in Wilcoxon (1945), consists of the sum of the X ranks among the pooled observations, that is

$$W = \sum_{i=1}^{m} R_i$$

where R_i is the rank of X_i among $\{X_1, \ldots, X_m, Y_1, \ldots, Y_n\}$. If F_X and F_Y are continuous (meaning there are no ties), then

$$\mathsf{P}_{\mathscr{H}_0}\,(W = k) = \frac{\pi_{m,n}(k)}{\binom{n+m}{m}}$$

where $\pi_{m,n}(k)$ is the number of ways to select m among $\{1, 2, \ldots, m+n\}$ such that their sum equals k.

Lemma 3.9 (Mann–Whitney recurrence relation)
The recurrence relation gives the distribution of $\pi_{m,n}(k)$ as

$$\pi_{m,n}(k) = \pi_{m,n-1}(k) + \pi_{m-1,n}(k - m - n).$$

The interpretation is as follows; consider without loss of generality the largest value of the combined sample, with rank $m+n$. Either this observation is $Y_{(n)}$, in which case it does not contribute to the value of the statistic and the sum k must be obtained from the remaining observations. Otherwise, if the $m+n$ rank corresponds to $X_{(m)}$, its contribution to W is $m+n$, meaning the other $m-1$ observations $\{X_{(j)}\}_{j=1}^{m-1}$ contribute exactly $k - (m+n)$.

Proof On the right hand side, the selections are among $\{1, \ldots, m+n-1\}$. The initial conditions are

$$\pi_{m,0}(k) = \begin{cases} 1 & \text{if } k = 1 + \cdots + m \\ 0 & \text{otherwise ;} \end{cases}$$

$$\pi_{0,n}(k) = \begin{cases} 1 & \text{if } k = 0 \\ 0 & \text{otherwise .} \end{cases}$$

and the recurrence is given by

$$\pi_{1,1}(k) = \pi_{1,0}(k) + \pi_{0,1}(k-2) = \begin{cases} 1 & \text{if } k = 1, 2 \\ 0 & \text{otherwise} \end{cases}$$

since $\pi_{1,0} = \mathbb{I}_{k=1}$ and $\pi_{0,1} = \mathbb{I}_{k=2}$, and

$$\begin{aligned} \pi_{2,2}(k) &= \pi_{2,1}(k) + \pi_{1,2}(k-4) \\ &= \pi_{2,0}(k) + \pi_{1,1}(k-3) + \pi_{1,1}(k-4) + \pi_{0,2}(k-7) \end{aligned}$$

If the sample has ties, then the resulting distribution cannot be applied directly. ■

Proposition 3.10 (Mann–Whitney test)
This test, due to Mann & Whitney (1947) counts among all $m \cdot n$ pairs (x-value, y-values) the number of times that x exceeds y and is given by

$$\begin{aligned} M &= \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbb{I}_{X_i > Y_j} = \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbb{I}_{X_{(i)} > Y_{(j)}} \\ &= \sum_{i=1}^{m} (R_i - i) \\ &= W - \frac{m(m+1)}{2} \end{aligned}$$

if there are no ties. Indeed, there are $R_1 - 1$ values of Y less than $X_{(1)}$, $R_2 - 1$ values less than $X_{(2)}$, of which 1 is $X_{(1)}$, etc.

Sometimes, rank statistics are in the form of what is called U-statistics.

Definition 3.11 (U-statistic)
A U-statistic of a single sample $(X_1, \ldots, X_N)$ is based on a kernel $h(X_1, \ldots, X_N)$ and equals

$$\frac{1}{\binom{n}{r}} \sum_{\Gamma_{n,r}} h(X_{i_1}, \ldots, X_{i_r})$$

where the sum runs over $\Gamma_{n,r}$, all possible combinations $\binom{n}{r}$ of index sets of size r.

The expectation of a U-statistics clearly is equal to $\mathsf{E}\left(h(X_1,\ldots,X_r)\right)$

Example 3.1 (Examples of U-statistic)
An example of kernel of rank $r=1$ is the mean, with kernel $h(x)=x$, while for $r=2$, we have $h(x_1,x_2)=\frac{1}{2}(x_1-x_2)^2$

In the two-sample case, one considers kernels of the form $h(X_{i_1},\ldots,X_{i_r},Y_{j_1},\ldots,Y_{j_s})$. The U-statistic is then

$$\binom{m}{r}^{-1}\binom{n}{s}^{-1}\sum_{\Gamma_{m,r}}\sum_{\Gamma_{n,s}}h(X_{i_1},\ldots,X_{i_r},Y_{j_1},\ldots,Y_{j_s}).$$

The Mann–Whitney statistic M is a two-sample U-statistic with $r=s=1$.

Definition 3.12 (Linear rank tests)
In the bivariate setting, we define linear rank tests as

$$L=\sum_{i=1}^{m}a(R_i)=\sum_{i=1}^{n+m}a(R_i)c_i$$

where c_i, termed **regression constant**, is a binary indicator

$$c_i=\begin{cases}1 & \text{if } 1\leq i\leq m\\ 0 & \text{otherwise .}\end{cases}$$

and R_i is the rank of $X_1,\ldots,X_m$ for $1\leq i\leq m$ and the rank of $Y_1,\ldots,Y_n$ for $m+1\leq i\leq m+n$.

3.3. Locally most powerful rank tests

The whole theory will be done with the two-sample Wilcoxon test. Consider the location shift case where $X_1,\ldots,X_m\overset{\text{iid}}{\sim}f(x-\Delta)$ and $Y_1,\ldots,Y_n\overset{\text{iid}}{\sim}f(y)$ and $\mathscr{H}_0:\Delta=0$ against the alternative $\mathscr{H}_1:\Delta>0$. We denote by $a(1),\ldots,a(m+n)$ the score given to the ranks (an increasing function), a mapping $a:\{1,\ldots,m+n\}\to\mathbb{R}$. We reject $\mathscr{H}_0$ if $\sum_{i=1}^{m}a(R_i)$ is greater than the critical value. How to choose $a(1),\ldots,a(m+n)$?

This test will reject if the ranks $(R_1,\ldots,R_n)\in\mathcal{R}$ where $\mathcal{R}$ is the rejection region. Of course,

$$\mathcal{R}=\left\{r_1,\ldots,r_m,r_{m+1},\ldots,r_{m+n};\sum_{i=1}^{m}a(r_i)\geq\text{ critical value}\right\}.$$

Because of the uniform distribution of the ranks,

$$\mathsf{P}_{\mathscr{H}_0}(\text{rejection}) = \frac{|\mathcal{R}|}{(n+m)!}.$$

The power of the test is equal to

$$\mathsf{P}_{\Delta}(\text{rejection}) = \sum_{\substack{(r_1,\dots,r_m) \\ \text{from } \mathcal{R}}} \mathsf{P}_{\Delta}(R_1 = r_1, \dots, R_m = r_m).$$

So, we need to compute the distribution of the ranks under the alternative.

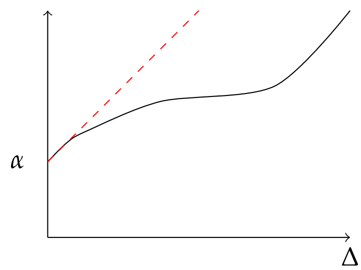


Figure 12: Power curve and local power at zero for location-shift test.

The size α is linked to the critical value. The slope of the tangent is called the **local power**, which is defined as

$$\text{local power} := \frac{\mathrm{d}}{\mathrm{d}\Delta}\mathsf{P}_{\Delta}(\text{rejection})\bigg|_{\Delta=0}.$$

To find the distribution of the order statistics, we present the following general result.

Theorem 3.13
Let $(X_1, \dots, X_N)$ have a joint distribution $f(\boldsymbol{x})$ invariant under permutations of the arguments. Let $T(X_1, \dots, X_N)$ be some statistic. The expected value conditional on the ranks,

$$\mathsf{E}(T(\boldsymbol{X}) \mid R_1 = r_1, \dots, R_N = r_N) = \mathsf{E}\left(T(X_{(r_1)}, \dots, X_{(r_N)})\right)$$

Proof The proof relies merely on relabelling. The independence between ranks and the order statistics readily gives

$$\begin{aligned}
\mathsf{E}(T(\boldsymbol{X}) \mid R_1 = r_1, \dots, R_N = r_N) &= \mathsf{E}\left(T(X_{(r_1)}, \dots, X_{(r_N)}) \mid R_1 = r_1, \dots, R_N = r_N\right) \\
&= \mathsf{E}\left(T(X_{(r_1)}, \dots, X_{(r_N)})\right).
\end{aligned}$$

■

Choose $T(X_1,\dots,X_N) = g(\boldsymbol{X})/f(\boldsymbol{X})$ where $g(\boldsymbol{X})$ is an alternative density absolutely continuous with respect to $\mathrm{d}F(x)$, meaning $f(\boldsymbol{x}) = 0$ implies $g(\boldsymbol{x}) = 0$. Then

$$\begin{aligned}
&\mathsf{E}_f\left(\frac{g(X_1,\dots,X_N)}{f(X_1,\dots,X_N)}\,\middle|\, R_1 = r_1,\dots,R_N = r_N\right) \\
&= \underset{\{R_1=r_1,\dots,R_N=r_N\}}{\int\cdots\int} \frac{g(x_1,\dots,x_N)}{f(x_1,\dots,x_N)} N! f(x_1,\dots,x_N)\,\mathrm{d}x_1\dots\mathrm{d}x_N \\
&= N!\mathsf{P}_g\left(R_1 = r_1,\dots,R_N = r_N\right).
\end{aligned} \tag{3.9}$$

To get the conditional probability with respect to only the first variables, we can integrate out the remaining ones. We present another proof instead.

Lemma 3.14

Let $X_1,\dots,X_m \overset{\text{iid}}{\sim} F_X$ (with $f_X = f(x-\Delta)$, the alternative) and $Y_1,\dots,Y_n \overset{\text{iid}}{\sim} F_Y$ (with $f_Y = f(x)$) and assume that $f_X(t) > 0$ implies $f_Y(t) > 0$. Further denote $\boldsymbol{V} = (\boldsymbol{X},\boldsymbol{Y})$. Then,

$$\mathsf{P}\left(\operatorname{rank}(X_1) = r_1,\dots,\operatorname{rank}(X_m) = r_m\right) = \binom{n+m}{m}^{-1} \mathsf{E}\left(\prod_{i=1}^m \frac{f_X(V_{(r_i)})}{f_Y(V_{(r_i)})}\right)$$

where $V_{(r)}$ are the order statistics of $V_1,\dots,V_{n+m} \overset{\text{iid}}{\sim} F_Y$.

Proof

$$\begin{aligned}
&\mathsf{P}\left(\operatorname{rank}(X_{(1)}) = r_1,\dots,\operatorname{rank}(X_{(m)}) = r_m \mid X_{(1)} = t_1,\dots,X_{(m)} = t_m\right) \\
&\quad= \frac{n!(F_Y(t_1))^{r_1-1}(F_Y(t_2)-F_Y(t_1))^{r_2-r_1-1}\cdots(1-F_Y(t_m))^{m+n-r_m}}{(r_1-1)!(r_2-r_1-1)!\cdots(n+m-r_m)!}.
\end{aligned}$$

This needs to be integrated with regard to the joint density of $X_{(1)},\dots,X_{(m)}$, given by $m!f_X(t_1)f_X(t_2)\cdots f_X(t_m)$. We have

$$\begin{aligned}
&\mathsf{P}\left(\operatorname{rank}(X_{(1)}) = r_1,\dots,\operatorname{rank}(X_{(m)}) = r_m\right) \\
&= \underset{t_1\le t_2\le\dots\le t_m}{\int\cdots\int} \frac{n!(F_Y(t_1))^{r_1-1}(F_Y(t_2)-F_Y(t_1))^{r_2-r_1-1}\cdots(1-F_Y(t_m))^{m+n-r_m}}{(r_1-1)!(r_2-r_1-1)!\cdots(n+m-r_m)!} \\
&\qquad\qquad \times m!\left(\prod_{i=1}^m f_X(t_i)\right)\mathrm{d}t_1\cdots\mathrm{d}t_m \\
&= \frac{n!m!}{(n+m)!}\underset{t_1\le t_2\le\dots\le t_m}{\int\cdots\int} g_{r_1,\dots,r_m}(t_1,\dots,t_m)\left(\prod_{i=1}^m f_Y(t_i)\right)^{-1}\left(\prod_{i=1}^m f_X(t_i)\right)\mathrm{d}t_1\cdots\mathrm{d}t_m
\end{aligned}$$

where the g-density, the joint density of $(V_{(r_1)},\dots,V_{(r_2)},\dots,V_{(r_m)})$, is given by

$$g_{\boldsymbol{r}}(\boldsymbol{t}) = \frac{(n+m)!(F_Y(t_1))^{r_1-1}f_Y(t_1)\cdots(1-F_Y(t_m))^{m+n-r_m}f_Y(t_m)}{(r_1-1)!(r_2-r_1-1)!\cdots(n+m-r_m)!},$$

yielding

$$\mathsf{P}\left(\operatorname{rank}(X_{(1)}) = r_1, \dots, \operatorname{rank}(X_{(m)}) = r_m\right) = \frac{m!n!}{(n+m)!}\mathsf{E}_g\left(\prod_{i=1}^{m}\frac{f_X(V_{(r_i)})}{f_Y(V_{(r_i)})}\right)$$

making use in the last step of eq. (3.9) and theorem 3.13 to conclude. ■

We now return to the power at the alternative Δ, which is

$$\beta(\Delta) = \sum_{\substack{(r_1,\dots,r_m)\\ \text{reject}}} \binom{n+m}{m}^{-1} \mathsf{E}\left(\prod_{i=1}^{m}\frac{f(V_{(r_i)}-\Delta)}{f(V_{(r_i)})}\right)$$

where $V_1, \dots, V_{n+m} \overset{\text{iid}}{\sim} f$. The local power, given by the derivative of the power $\beta(\cdot)$ at Δ evaluated at $\Delta = 0$, is

$$\left.\frac{\mathrm{d}}{\mathrm{d}\Delta}\beta(\Delta)\right|_{\Delta=0} = \sum_{\substack{(r_1,\dots,r_m)\\ \text{reject}}} \binom{n+m}{m}^{-1} \sum_{i=1}^{m}\mathsf{E}\left(-\frac{f'}{f}(V_{(r_i)})\right).$$

This holds because

$$\frac{\mathrm{d}}{\mathrm{d}\Delta}\prod_{i=1}^{m}\frac{f(V_{(r_i)}-\Delta)}{f(V_{(r_i)})} = \sum_{i=1}^{m} -f'(V_{(r_i)}-\Delta)\prod_{j\neq i} f(V_{(r_j)}-\Delta)\left(\prod_{i=1}^{m} f(V_{(r_i)})\right)^{-1}$$

using the product rule. We now know how the locally most powerful test for a shift in location behaves. It has to be based on the score

$$a_{\text{opt}}(r) = \mathsf{E}\left(-\frac{f'}{f}(V_{(r)})\right).$$

Example 3.2

One can now ask many questions, for example

- for which f do we find $\mathsf{E}\left(-f'(V_{(r)})/f(V_{(r)})\right) = A + Br$?
- what score is best for the normal density $f(x) = \phi(x)$?

Definition 3.15 (Score-generating functions)

We can always represent $V_{(r)}$ as $F^{-1}(U_{(r)})$ where $U_1, \dots, U_{n+m} \overset{\text{iid}}{\sim} \mathcal{U}(0,1)$ and $F(X) = \int_{-\infty}^{x} f(u)\,\mathrm{d}u$, which implies the expected value F scores are given by

$$a^N_{\text{opt}}(r) = \mathsf{E}\left(-\frac{f'}{f}\left(F^{-1}(U_{(r)})\right)\right) = \mathsf{E}\left(\varphi_f(V_{(r)})\right),$$

where $N = n + m$. The function $\varphi_f(u) = -f'(F^{-1}(u))/f(F^{-1}(u))$, for $0 \leq u \leq 1$, is termed **score generating function** for the expected value F scores $a_{\text{opt}} = \mathsf{E}\left(\varphi_f(V_{(r)})\right)$. If the latter are hard to calculate, we can resort to the quantile F scores,

$$\widetilde{a}_{\text{opt}}^N = \varphi_f\left(\frac{r}{n+m+1}\right) = \varphi_f\left(\mathsf{E}\left(U_{(r)}\right)\right).$$

Remark 3.1
The optimal score for the signed rank test

$$\sum_{i=1}^{N} \operatorname{sign}(X_i) a^+(R_i^+)$$

is

$$a_{\text{opt}}^{+,N}(r) = \mathsf{E}\left(-\frac{f'}{f}(|X|_{(r)})\right).$$

What is $\varphi_f^+(u)$?

The distribution of $|X|$ is

$$\mathsf{P}\left(|X| \le t\right) = \mathsf{P}\left(-t \le X \le t\right) = F(t) - F(-t) = 2F(t) - 1$$

assuming symmetry of X about 0. The inverse of $G(t) = 2F(t) - 1$ for $t > 0$ is $F^{-1}((1+u)/2)$ and so

$$|X|_{(r)} = F^{-1}\left(\frac{1+U_{(r)}}{2}\right).$$

For a sample $X_1, \ldots, X_N \overset{\text{iid}}{\sim} F$ and $|X|_{(1)} \le |X|_{(2)} \le \cdots \le |X|_{(N)}$, we consider an equivalent sample from the uniform distribution with $U_1, \ldots, U_N \overset{\text{iid}}{\sim} \mathcal{U}(0,1)$ and

$$\varphi_F^+(u) = -\frac{f'}{f}\left(F^{-1}\left(\frac{1+u}{2}\right)\right)$$

3.4. Asymptotic distribution of rank tests

Asymptotic distribution under the null hypothesis

Consider for example the Mann–Whitney test. The mean and the variance of the test are

$$\mathsf{E}\left(M\right) = \frac{1}{2}m(m+n+1), \qquad \mathsf{Var}\left(M\right) = \frac{1}{12}mn(m+n+1);$$

see the exercises. With the same method, one can solve the general linear rank test. Assume that $a_N(r) = \varphi(r/(N+1))$ for $r = 1, \dots, n+m$.

Lemma 3.16 (Asymptotic mean and variance of linear rank tests)
Consider sequences of experiments indexed by $N = m + n$ such that $N \to \infty$ and $m/N \to \lambda \in (0,1)$. The asymptotic mean and variance of

$$T_N := \frac{L_N}{N} := \frac{1}{N} \sum_{i=1}^{m} \varphi\left(\frac{R_i}{N+1}\right),$$

where R_i is the rank of X_i in the combined sample, are given respectively by

$$\lim_{N\to\infty} \mathsf{E}\left(T_N\right) = \lambda\bar{\varphi} = \lambda \int_0^1 \varphi(u)\,\mathrm{d}u$$

and

$$\lim_{N\to\infty} N\mathsf{Var}\left(T_N\right) = \lambda(1-\lambda) \int_0^1 (\varphi(u) - \bar{\varphi})^2\,\mathrm{d}u.$$

The variance we obtain only makes sense if $0 < \lambda < 1$ and $\int \varphi^2(u)\,\mathrm{d}u < \infty$.

Consider for example $\varphi(u) = \varphi_f(u) = -f'/f(F^{-1}(u))$ and

$$\int_0^1 \varphi_f^2(u)\,\mathrm{d}u = \int_0^1 \left[-\frac{f'}{f}\left(F^{-1}(u)\right)\right]^2 \mathrm{d}u.$$

It is natural to make a change of variable $F^{-1}(u) = x$ meaning $F(x) = u$ and $f(x)\,\mathrm{d}x = \mathrm{d}u$. Thus,

$$\begin{aligned}\int_0^1 \varphi_f^2(u)\,\mathrm{d}u &= \int_{F^{-1}(0)}^{F^{-1}(1)} \left[-\frac{f'}{f}(x)\right]^2 f(x)\,\mathrm{d}x \\ &= \mathcal{I}(F).\end{aligned}$$

For this particular φ, we need a finite Fisher information.

Proof We will replace

$$NT_N = L_N = \sum_{i=1}^{m} \varphi\left(\frac{R_i}{N+1}\right)$$

by

$$\sum_{i=1}^{N} c_N(i)\varphi\left(\frac{R_i}{N+1}\right),$$

where $c_N(i) = \mathbb{I}_{\mathcal{A}_i}$ where $\mathcal{A}_i$ is the the event that the i^{th} observation is from the X-sample. The vector $\boldsymbol{c}_N = (c_N(1), \ldots, c_N(N))$ is the vector of **regression constants**. Let

$$T_N = \frac{1}{N} \sum_{i=1}^{N} B_i \varphi\left(\frac{i}{N+1}\right)$$

where $B_1, \ldots, B_N \sim \mathcal{B}(p)$ where $p = m/N$. Thus,

$$\mathsf{E}\left(T_N\right) = \frac{1}{N} \sum_{i=1}^{N} \frac{m}{N} \varphi\left(\frac{i}{N+1}\right) \to \lambda \int_0^1 \varphi(u)\, \mathrm{d}u$$

as $N \to \infty$. If $\varphi(u)$ is Riemann integrable, then

$$N\mathsf{Var}\left(T_N\right) = \frac{1}{N} \sum_{i=1}^{N} \frac{m}{N}\left(1 - \frac{m}{N}\right) \varphi^2\left(\frac{i}{N+1}\right) \tag{3.10}$$

$$+ \frac{1}{N} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ i \neq j}}^{N} \mathsf{Cov}\left(B_i, B_j\right) \varphi\left(\frac{i}{N+1}\right) \varphi\left(\frac{j}{N+1}\right) \tag{3.11}$$

where

$$\begin{aligned}
\mathsf{Cov}\left(B_i, B_j\right) &= \mathsf{P}\left(B_i = 1, B_j = 1\right) - \left(\frac{m}{N}\right)^2 \\
&= \frac{\binom{2}{2}\binom{N-2}{m-2}}{\binom{N}{m}} - \left(\frac{m}{N}\right)^2 \\
&= \frac{(N-2)!m!(N-m)!}{(m-2)!(N-m)!N!} - \left(\frac{m}{N}\right)^2 \\
&= \frac{m(m-1)}{N(N-1)} - \frac{m^2}{N^2} \\
&= -\frac{m(N-m)}{N^2(N-1)}.
\end{aligned}$$

Another easy proof for the covariance is derived as follows: since $\sum_{i=1}^{N} B_i = m$,

$$\mathsf{Var}\left(\sum_{i=1}^{N} B_i\right) = 0 = N\frac{m}{N}\left(1 - \frac{m}{N}\right) + 2\binom{N}{2}x$$

and solving for x, we get

$$\mathsf{Cov}\left(B_i, B_j\right) = -m\left(1 - \frac{m}{N}\right)\left(N(N-1)\right)^{-1}.$$

The behaviour as $N \to \infty$ of the right hand side of eq. (3.10) is $\lambda(1-\lambda)\int_0^1 \varphi^2(u)\, \mathrm{d}u$. Analogously, we recover $-\lambda(1-\lambda)\bar{\varphi}^2$ for eq. (3.11). This requires a few more manipulations to show that the diagonal terms (with $i = j$) are of lower order and negligible. Together, this gives

$$\lambda(1-\lambda)\left[\int_0^1 \varphi^2(u)\,\mathrm{d}u - \left(\int_0^1 \varphi(u)\,\mathrm{d}u\right)^2\right] = \lambda(1-\lambda)\int_0^1 (\varphi(u)-\bar{\varphi})^2\,\mathrm{d}u.$$

■

Remark 3.2
In the one-sample case, $L_N = \sum_{i=1}^N \operatorname{sign}(X_i)\varphi(R_i^+/(N+1))$.

We now aim to derive the asymptotic normality under the null hypothesis. Historically, this question has been instrumental for developing asymptotic.

Under the null hypothesis,

$$\mathsf{E}\,(L_N) = \mathsf{E}\left(\sum_{i=1}^N c_N(i)a_N(R_i)\right) = N\bar{c}_N\bar{a}_N$$

because $\mathsf{E}\,(a_N) = \bar{a}_N$ for a rank R, where $a_N(r) = \varphi(r/(N+1))$, $\bar{a}_N = N^{-1}\sum_{i=1}^N a_N(i)$ and $\bar{c}_N = N^{-1}\sum_{i=1}^N c_N(i)$.

It will prove easier to consider another score for derivations.

Theorem 3.17 (Projection method)
Let $\boldsymbol{R}_N = (R_N(1),\ldots,R_N(N))$ be the vector of ranks derived under the null hypothesis from a vector $\boldsymbol{Z} = (Z_1,\ldots,Z_N)$ with iid coordinates distributed according to F, absolutely continuous with density f. Put $U_i = F(Z_i)$ and let $\boldsymbol{a}_N = (a_N(1),\ldots,a_N(N))$ be a vector with entries $a_N(r) = \mathsf{E}\big(\varphi(U_{(r)})\big)$, a measurable score generating function which is not constant and such that $\int_0^1 \varphi^2(u)\,\mathrm{d}u < \infty$. Then,

$$L_N = \sum_{i=1}^N c_N(i)a(R_N(i))$$

has the same asymptotic distribution as $\tilde{L}_N = N\bar{c}_N\bar{a}_N + \sum_{i=1}^N (c_N(i)-\bar{c}_N)\varphi(U_i)$.

Using Theorem 3.13, we have

$$\begin{aligned} a_N(R_N(i)) &= \mathsf{E}\,(\varphi(U_i) \mid R_N) \\ &= \mathsf{E}\left(\varphi(F(Z_{R_N(i)})) \mid R_N\right) \\ &= \mathsf{E}\left(\varphi(U_{R_N(i)})\right) \end{aligned}$$

If we project $\tilde{L}_N$ onto the subspace of square integrable functions defined by the span of R_N, we get the conditional expectation

$$\mathsf{E}\left(\tilde{L}_N \mid R_N\right)$$

and this turns out to be L_N. Indeed,

$$\mathsf{E}\left(\tilde{L}_N \mid R_N\right) = N\bar{c}_N\bar{a}_N + \sum_{i=1}^{N}(c_N(i) - \bar{c}_N)a_N(R_N(i))$$

because $a_N(R_N(i)) = \mathsf{E}\left(\varphi(U_i) \mid R_N\right)$ and because

$$\sum_{i=1}^{N}\bar{c}_N a_N(R_N(i)) = \bar{c}_N\sum_{i=1}^{N}a_N(i) = N\bar{c}_N\bar{a}_N.$$

Remark 3.3 (on the projection method)
Consider a sample $Z_1, \ldots, Z_N$. If we are dealing with a complicated statistic U_N, we would like to approximate it by a statistic of the form

$$\tilde{U}_N = \sum_{i=1}^{N}h_i(Z_i),$$

where $h_i(Z_i)$ can be found as the projection $\mathsf{E}\left(U_N \mid Z_i\right)$.

The proof of the following fact will be omitted:

$$\frac{\mathsf{Var}\left(L_N\right)}{\mathsf{Var}\left(\tilde{L}_N\right)} \to 1$$

as $N \to \infty$; this is a sufficient condition for asymptotic equivalence (i.e. L_N and $\tilde{L}_N$ have the same asymptotic distribution). [3] In our case,

$$\frac{\mathsf{Var}\left(L_N\right)}{\mathsf{Var}\left(\tilde{L}_N\right)} = \frac{\frac{1}{N-1}\sum_{i=1}^{N}(c_N(i) - \bar{c}_N)^2\sum_{i=1}^{N}(a_N(i) - \bar{a}_N)^2}{\frac{1}{N}\sum_{i=1}^{N}(c_N(i) - \bar{c}_N)^2\mathsf{Var}\left(\varphi(U)\right)} \mathrel{\dot\sim} \frac{\mathsf{Var}\left(a_N(\text{ranks})\right)}{\mathsf{Var}\left(\varphi(U)\right)}$$

which can be proved in various ways!

Corollary 3.18 (Asymptotic normality of linear rank tests)
If $\max_i(c_N(i) - \bar{c}_N)^2 / \sum_{i=1}^{N}(c_N(i) - \bar{c}_N)^2 \to 0$ as $N \to \infty$, then it holds that

$$\frac{L_N - \mathsf{E}\left(L_N\right)}{\sqrt{\mathsf{Var}\left(L_N\right)}} \xrightarrow{d} \mathcal{N}(0, 1)$$

as $N \to \infty$.

[3] See for example Van der Vaart, *Asymptotic Statistics*, p.176–177.

In the two-sample case, m of the $c_N(i) = 1$ and n of the $c_N(i) = 0$, which imply $\bar{c}_N = m/N$, so

$$\bigvee_{i=1}^{N} \left\{ (c_N(i) - \bar{c}_N)^2 \right\} = \max\left\{ \left(\frac{m}{N}\right)^2, \left(1 - \frac{m}{N}\right)^2 \right\}$$

and

$$\begin{aligned}\sum_{i=1}^{N} (c_N(i) - \bar{c}_N)^2 &= m\left(1 - \frac{m}{N}\right)^2 + n\left(\frac{m}{N}\right)^2 \\ &= m\left(1 - \frac{m}{N}\right)\left(1 - \frac{m}{N} + \frac{m}{N}\right) = N\frac{m}{N}\left(1 - \frac{m}{N}\right)\end{aligned}$$

since $c_N(i)$ is binary, each term is either $1 - m/N$ or $-m/N$ and there are respectively m and n of them. Does

$$\frac{1}{N} \frac{\max\left(\frac{m}{N}, \frac{n}{N}\right)^2}{\frac{m}{N}\frac{n}{N}} \to 0$$

as $N \to \infty$? Yes, unless $m \equiv N$ or $m \equiv 0$, since the denominator converges to $\lambda(1 - \lambda)$ and the numerator to $\max(\lambda, 1 - \lambda)^2$.

Asymptotic distribution under the alternative

Let $Z_1, \ldots, Z_n$ be observations and let $R_N(i)$ denote the rank of Z_i, the number of Z values that are less than or equal to Z_i. Alternatively, we can express $R_N(i)$ as $R_N(i) = NF_N(Z_i)$ where $F_N(x)$, the number of $Z_i \leq x$ upon N, is the empirical distribution. The latter can be analysed as a stochastic process indexed by x. The representation

$$\frac{1}{N}\sum_{i=1}^{N} c_N(i) a_N(R_N(i)) = \int a_N(NF_N(x)) c_N(x) \, \mathrm{d}F_N(x).$$

can be used to get asymptotic results. Another way to obtain the power is by means of asymptotic arguments, and this is what is used in practice.

3.5. Asymptotic power and Pitman efficacy

We begin by illustrating the concept with the Student's t-test.

Example 3.3 (Power of test for Student's t-test)
Let $X_1, \ldots, X_N \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, where $\mathscr{H}_0 : \mu = 0$ versus $\mathscr{H}_1 : \mu > 0$. Under the null hypothesis,

$$T_N = \sqrt{N}\frac{\bar{X}_N}{S_N} \sim t_{N-1},$$

where $\bar{X}_N = N^{-1}\sum_{i=1}^{N} X_i$ and $S_N^2 = (N-1)^{-1}\sum_{i=1}^{N}(X_i - \bar{X}_N)^2$.

To have $\mathsf{P}_{\mathscr{H}_0}(\text{rejection}) = \alpha$, we reject if $T_N > \mathfrak{t}_{N-1}(1-\alpha) \cong \mathfrak{z}_{1-\alpha} = \Phi^{-1}(1-\alpha)$, for large N. At $\mu > 0$, this test has power

$$\mathsf{P}_\mu\,(\text{rejection}) = \mathsf{P}_\mu\,(T_N > \mathfrak{z}_{1-\alpha})\,.$$

and it is not easy to work out, since T_N follows a non-central Student-t distribution. It is apparent that as $N \to \infty$, the power goes to 1, except at the null hypothesis, provided that the test is reasonably behaved. Hence, this situation is not terribly interesting.

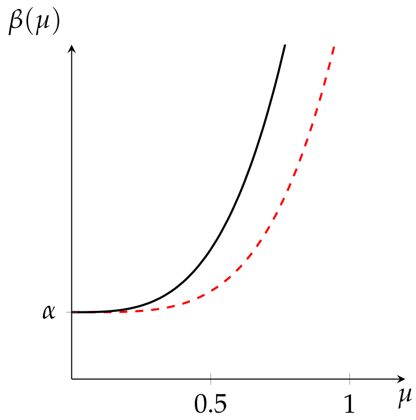


Figure 13: Power as a function of N_i, for N_1 (dashed red) and N_2 (black) where $N_1 < N_2$ for the $\mathscr{H}_0 : \mu = 0$

Pitman's idea was to compute the power at a local alternative $\mu_N \to 0$ so that the power at μ_N converges to a number $\beta \in (0,1)$ as $N \to \infty$.

Example 3.4 (Location shift for Normal distributions)
This is an example of a Student's t-test for known σ. We reject if

$$\frac{\sqrt{N}\bar{X}_N}{\sigma} > \mathfrak{z}_{1-\alpha}$$

since $\bar{X}_N/\sigma \sim \mathcal{N}(0, N^{-1})$ under $\mathscr{H}_0$. The power of this test statistic is

$$\mathsf{P}_\mu\left(\frac{\sqrt{N}\bar{X}_N}{\sigma} > \mathfrak{z}_{1-\alpha}\right)$$

since $\bar{X}_N/\sigma \sim \mathcal{N}(\mu/\sigma, N^{-1})$ under the alternative. Thus,

$$\mathsf{P}_\mu\left(\frac{\sqrt{N}\bar{X}_N}{\sigma} - \frac{\sqrt{N}\mu}{\sigma} > \mathfrak{z}_{1-\alpha} - \frac{\sqrt{N}\mu}{\sigma}\right)$$

$$
\begin{aligned}
&= 1 - \Phi\left(\mathfrak{z}_{1-\alpha} - \frac{\sqrt{N}\mu}{\sigma}\right)\\
&= \Phi\left(\frac{\sqrt{N}\mu}{\sigma} - \mathfrak{z}_{1-\alpha}\right)
\end{aligned}
$$

converges to 1 as $N \to \infty$. What should μ_N be in order for $\Phi(\sqrt{N}\mu/\sigma - \mathfrak{z}_{1-\alpha}) \to \beta$ for $\beta \in (0,1)$? Taking $\mu_N = h/\sqrt{N}$ will work! In general, the power at the alternative $h/\sqrt{N}$ approximately

$$\Phi\left(h\mathrm{e}_{\mathrm{P}} - \mathfrak{z}_{1-\alpha}\right)$$

where e_{P} is the **Pitman efficacy**.

The power at the alternative μ_N is $\beta(\mu_N) = \Phi(h/\sigma - \mathfrak{z}_{1-\alpha})$

3.5.1. General formula for the Pitman efficacy for test statistics with Gaussian limits

Consider the test statistic $T_N = T(X_1, \ldots, X_N)$ with $X_i \overset{\text{iid}}{\sim} F_\theta$ for $\theta \in \mathbb{R}$. We wish to test an hypothesis of the form $\mathscr{H}_0 : \theta = 0$ versus $\mathscr{H}_1 : \theta > 0$ for an arbitrary family. Then, for statistics that converge in law to a Gaussian distribution, we have

$$\sqrt{N}\frac{T_N - \mu(\theta_N)}{\sigma(\theta_N)} \xrightarrow{d} \mathcal{N}(0,1) \tag{3.12}$$

where $\mu(\cdot)$, $\sigma(\cdot)$ are appropriate location and scale functions, respectively. We term $\theta_N = h/\sqrt{N}$ the **Pitman alternative**. Similar expressions can be derived for χ^2 or F limits.

Example 3.5 (continuation of example 3.3)
In the setting of example 3.3, we had under $\mathscr{H}_0$, $T_N \sim t_{N-1} \mathrel{\dot{\sim}} \mathcal{N}(0,1)$ for N large. Taking $\mu(\mu) = \mu/\sigma$ and $\sigma(\mu) = 1$, we have

$$
\begin{aligned}
\frac{\sqrt{N}\bar{X}_N}{S_N} - \frac{\sqrt{N}\mu_N}{\sigma} &= \sqrt{N}\left(\frac{\bar{X}_N - \mu_N}{S_N} + \frac{\mu_N}{S_N} - \frac{\mu_N}{\sigma}\right)\\
&= \sqrt{N}\frac{\bar{X}_N - \mu_N}{S_N} + \sqrt{N}\mu_N\left(\frac{1}{S_N} - \frac{1}{\sigma}\right)\\
&\xrightarrow{d} \mathcal{N}(0,1)
\end{aligned}
$$

since the term $h(1/S_N - 1/\sigma) \to 0$ by Slutsky's theorem. By the central limit theorem, the first term is asymptotically normally distributed and so eq. (3.12) holds for the Student's t-test. Whether or not eq. (3.12) holds needs to be checked in each case.

Theorem 3.19 (Power at Pitman alternative for test statistics with Gaussian limit)
Assume that eq. (3.12) holds, $\mu(\theta)$ is differentiable at zero and $\sigma(\theta)$ is continuous. Then,

the power at θ_n converges to

$$1 - \Phi\left(\mathfrak{z}_{1-\alpha} - h\frac{\mu'(0)}{\sigma(0)}\right) = \Phi\left(h\frac{\mu'(0)}{\sigma(0)} - \mathfrak{z}_{1-\alpha}\right) \tag{3.13}$$

where the test has asymptotic level α and leads to rejection of $\mathscr{H}_0$ rejects for large values of T_n.

Proof To achieve asymptotic level α, we reject the null hypothesis if

$$\sqrt{n}\frac{T_n - \mu(0)}{\sigma(0)} > \mathfrak{z}_{1-\alpha} \quad \Leftrightarrow \quad T_n > \mu(0) + \frac{\sigma(0)}{\sqrt{n}}\mathfrak{z}_{1-\alpha}$$

The power at θ_n is asymptotically equal to

$$\begin{aligned}
\beta(\theta_n) &= \mathsf{P}_{\theta_n}\left(T_n > \mu(0) + \frac{\sigma(0)}{\sqrt{n}}\mathfrak{z}_{1-\alpha}\right) \\
&= \mathsf{P}_{\theta_n}\left(\frac{\sqrt{n}(T_n - \mu(\theta_n))}{\sigma(\theta_n)} > \frac{\sqrt{n}\mu(0)}{\sigma(\theta_n)} + \frac{\sqrt{n}\sigma(0)}{\sqrt{n}\sigma(\theta_n)}\mathfrak{z}_{1-\alpha} - \frac{\sqrt{n}\mu(\theta_n)}{\sigma(\theta_n)}\right) \\
&= \mathsf{P}_{\theta_n}\left(\frac{\sqrt{n}(T_n - \mu(\theta_n))}{\sigma(\theta_n)} > \frac{\sqrt{n}(\mu(0) - \mu(\theta_n))}{\sigma(\theta_n)} + \frac{\sigma(0)}{\sigma(\theta_n)}\mathfrak{z}_{1-\alpha}\right).
\end{aligned}$$

Since $\sigma(0)/\sigma(\theta_n) \to 1$ as $n \to \infty$, $\mu(\theta_n) = \mu(0) + \theta_n\mu'(0) + \mathrm{o}(n^{-1/2})$ and thus

$$\frac{\sqrt{n}(\mu(0) - \mu(\theta_n))}{\sigma(\theta_n)} = \frac{\sqrt{n}\left(-\frac{h}{\sqrt{n}}\mu'(0) + \mathrm{o}(n^{-1/2})\right)}{\sigma(\theta_n)}$$

converges to $-h\mu'(0)/\sigma(0)$. The asymptotic power is thus equal to eq. (3.13). ■

Definition 3.20 (Pitman efficacy)
We define the **Pitman efficacy** as

$$\mathrm{e_P}(\text{test}, F_\theta) = \frac{\mu'(0)}{\sigma(0)}.$$

For positive location shift, $\mathrm{e_P} > 0$. One is interested in local alternatives $\theta > 0$, but $\theta \approx 0$. If $\theta = h/\sqrt{n}$, the power at θ is

$$\beta(\theta) \approx \Phi(\sqrt{n}\theta\mathrm{e_P} - \mathfrak{z}_{1-\alpha}).$$

The bigger $\mathrm{e_P}$, the better the test.

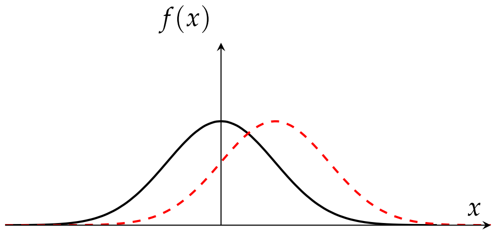


Figure 14: Location shift of $\theta > 0$ (red dashed curve) against postulated null distribution with $\theta = 0$.

Example 3.6
For Student's t-test,

$$e_P = \frac{\frac{d}{d\theta}\left(\frac{\theta}{\sigma}\right)\Big|_{\theta=0}}{1} = \frac{1}{\sigma}$$

and the same for the Z-test.

3.5.2. Asymptotic relative efficiency of tests

Definition 3.21 (Asymptotic relative efficiency)
Let Test A T_A and Test B T_B be two competing tests. The ratio

$$\mathsf{ARE}(\mathrm{A}, \mathrm{B}) := \left(\frac{e_P(\text{Test A}, F)}{e_P(\text{Test B}, F)}\right)^2$$

is called the asymptotic relative efficiency of Test A relative to Test B, also denoted ARE(Test A relative to Test B). At a local alternative θ and assuming both tests have level α, the two tests have equal power if $\sqrt{n_A}e_P(\text{Test A}) = \sqrt{n_B}e_P(\text{Test B})$ when

$$\mathsf{ARE}(A, B) = \frac{n_B}{n_A}$$

If Test A is more efficient that Test B, it reaches the same power with fewer observations.

Note
Historically, Pitman studied ratios of sample sizes. The asymptotic power allows relatively easy comparisons with existing tests. Note in passing that Theorem 3.19 is due to Noether!

Example 3.7 (ARE for Wilcoxon's signed rank test)
To compute $\mu'(0)/\sigma(0)$, we need to know $\mu(\theta)$ for small values of θ and $\sigma(0)$ (where $\sigma^2(0)$ is the asymptotic variance of the test statistic under the null).

Recall the statistic is given by

$$
\begin{aligned}
\widetilde{T}_N &= \sum_{i=1}^{N} \operatorname{sign}(X_i) R_i^+ \\
&= 2 \sum_{i=1}^{N} \mathbb{I}_{X_i > 0} R_i^+ - \frac{N(N+1)}{2}.
\end{aligned}
$$

where R_i^+ is the number of $|X_j|$ less than or equal to $|X_i|$, $\#\{|X_j| < |X_i| + 1\}$. Then

$$
T_N = \frac{\widetilde{T}_N}{N(N+1)} = \frac{N-1}{N+1} U_N^{(1)} + \frac{2}{N+1} U_N^{(2)},
$$

where

$$
\begin{aligned}
U_N^{(1)} &= \binom{N}{2}^{-1} \sum_{j=1}^{n} \sum_{i<j} \operatorname{sign}(X_i + X_i), \\
U_N^{(2)} &= \frac{1}{N} \sum_{i=1}^{N} \operatorname{sign}(X_i)
\end{aligned}
$$

are proportional to respectively a two-sample U-statistic of degree (1, 1) and a U-statistic of degree 1.

Lef F be continuous around 0. Put $G(x) = F(x - \theta)$. The expectation of T_N at G is equal to

$$
\begin{aligned}
&\frac{N-1}{N+1} \mathsf{E}\left(\operatorname{sign}(X_1 + X_2)\right) + \frac{2}{N+1} \mathsf{E}\left(\operatorname{sign}(X_1)\right) \\
&\quad = \frac{N-1}{N+1} \left(\mathsf{P}\left(X_1 > -X_2\right) - \mathsf{P}\left(X_2 < -X_1\right)\right) + \frac{2}{N+1} \left(\mathsf{P}\left(X_1 > 0\right) - \mathsf{P}\left(X_1 < 0\right)\right)
\end{aligned}
$$

since it suffices to calculate the expectation of the kernel of the U-statistics, where

$$
\begin{aligned}
\mathsf{P}\left(X_1 > 0\right) &= 1 - G(0) = 1 - F(-\theta) \\
\mathsf{P}\left(X_1 > -X_2\right) &= \int \mathsf{P}\left(X_1 > -X_2 \mid X_2 = x\right) f(x - \theta) \, \mathrm{d}x \\
&= \int (1 - F(-x - \theta)) f(x - \theta) \, \mathrm{d}x
\end{aligned}
$$

For large N, only

$$
\mathsf{P}\left(X_1 > -X_2\right) - \mathsf{P}\left(X_2 < -X_1\right) = \mu(\theta)
$$

matters. For the variance under $\mathscr{H}_0$, recall the formula

$$
\mathsf{Var}_{\mathscr{H}_0}\left(T_N\right) = \frac{1}{N^2 (N+1)^2} \sum_{i=1}^{N} i^2
$$

$$
\begin{aligned}
&= \frac{N(N+1)(2N+1)}{6N^2(N+1)^2} \\
&\sim \frac{1}{3N}
\end{aligned}
$$

and $\sigma(0) = 1/\sqrt{3}$. If one compares the Wilcoxon test to the t-test, one finds (see exercises)

$$\left(\frac{e_P(\text{Wilcoxon})}{e_P(t\text{-test})}\right)^2 = 0$$

if the variance of F (and G) is infinite. Actually, the Wilcoxon test is never too bad a choice when compared to the t test meaning (there is a universal lower bound of 0.864 for the ARE).

Remark 3.4

1. A similar theory can be developed for tests with an asymptotic χ^2-distribution.
2. The Pitman alternatives and the behavior of the likelihood ratio of the Pitman alternative versus the null has been studied extensively by Lucien LeCam, who investigated **contiguous sequences** of laws, under the name **local asymptotic normality**. Equation (3.12) can be checked using this theory.
3. Other asymptotic comparisons of tests have been developed. The most popular is **Bahadur's slope**. This is also linked to Chernoff's results on large deviations.

Example 3.8 (Normal shift)

For $x \to \infty$, we have

$$1 - \Phi(x) \sim \frac{\phi(x)}{x} = \frac{1}{\sqrt{2\pi}x} \exp\left(-\frac{x^2}{2}\right)$$

which goes quickly to zero. We had for the power the formula $\Phi(\sqrt{n}\mu - \mathfrak{z}_{1-\alpha})$. Thus

$$1 - \beta(\mu) \sim \frac{\exp\left(-\frac{1}{2}(\sqrt{n}\mu - \mathfrak{z}_{1-\alpha})^2\right)}{\sqrt{2\pi}(\sqrt{n}\mu - \mathfrak{z}_{1-\alpha})}$$

and

$$\frac{1}{n}\log(1-\beta(\mu)) \sim -\frac{1}{2n}\left(\sqrt{n}\mu - \mathfrak{z}_{1-\alpha}\right)^2 - \frac{\log(\sqrt{2\pi}(\sqrt{n}\mu - \mathfrak{z}_{1-\alpha}))}{n},$$

meaning $\log(1-\beta(\mu)) \sim -\frac{1}{2}n\mu^2$. We also have 1-$\beta(\mu) \sim \exp(-n\mu^2/2)$, which resembles the Chernoff bounds for large deviations.

For many tests, $n^{-1}\log(1-\beta(\mu)) \sim -\frac{1}{2}c(\mu)$ at a fixed alternative μ. $c(\mu)$ is called the **exact Bahadur slope**.

We left many additional topics untouched, in particular regression. Others we only did in the exercises, including notably the Hodges–Lehmann estimators.

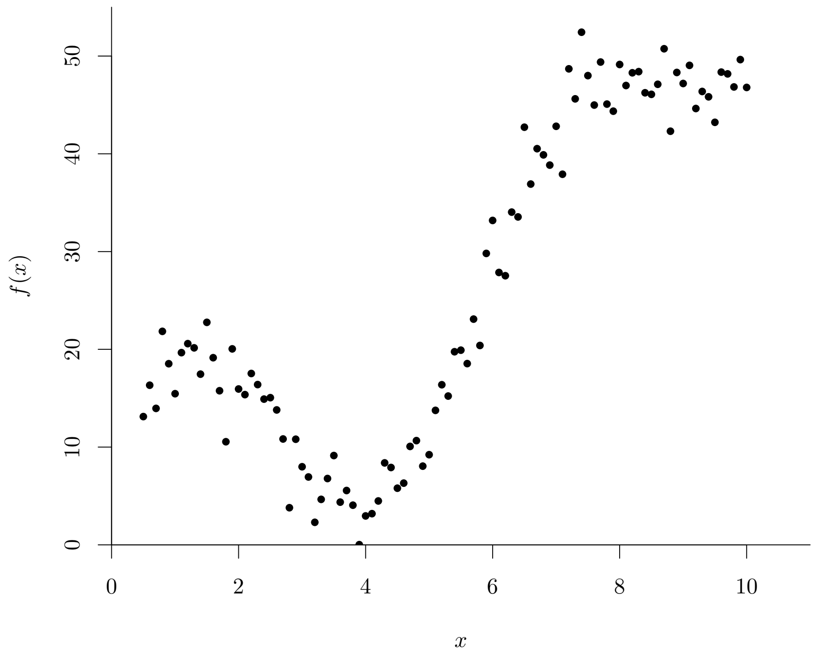
A collection of nonparametric tests

Test	Statistic	Expectation	Variance
Tests for a shift between two samples $X_1,\dots,X_m$ and $Y_1,\dots,Y_n$. The ranks are computed on the pooled sample and $N=n+m$.			
Wilcoxon	$\sum_{i=1}^n R(Y_i)$	$n(N+1)/2$	$nm(N+1)/12$
van der Waerden	$\sum_{i=1}^n \Phi^{-1}[R(Y_i)/(N+1)]$	0	$(nm/N(N-1))\sum_{i=1}^N (\Phi^{-1}[i/(N+1)])^2$
Normal Scores	$\sum_{i=1}^n \mathbb{E}[\Phi^{-1}[U_{(R(Y_i))}]]$	0	$(nm/N(N-1))\sum_{i=1}^N (\mathbb{E}[\Phi^{-1}[U_{(i)}]])^2$
Instead of $\mathbb{E}[\Phi^{-1}[U_{(i)}]]$, we could also write $\mathbb{E}[Z_{(i)}]$, where $U_{(i)}$ and $Z_{(i)}$ denote uniform and normal order statistics.			
Tests of symmetry for one sample $X_1,\dots,X_n$. The ranks R_i^+ are computed on the absolute deviations from μ_0, the center of symmetry.			
Wilcoxon	$\sum_{i:X_i>0} R_i^+$	$n(n+1)/4$	$n(n+1)(2n+1)/24$
van der Waerden	$\sum_{i:X_i>0} \Phi^{-1}\left[\frac{1}{2}+\frac{R_i^+}{2(n+1)}\right]$	$\frac{1}{2}\sum_{i=1}^n \Phi^{-1}\left[\frac{1}{2}+\frac{i}{2(n+1)}\right]$	$\frac{1}{4}\sum_{i=1}^n \Phi^{-1}\left[\frac{1}{2}+\frac{i}{2(n+1)}\right]^2$
Normal Scores	$\sum_{i:X_i>0} \mathbb{E}[\|Z\|_{(R_i^+)}]$	$n/\sqrt{2\pi}$	$\frac{1}{4}\sum_{i=1}^n \left(\mathbb{E}[\|Z\|_{(i)}]\right)^2$
Two sample tests of scale. The ranks are computed on the pooled sample.			
Ansari-Bradley	$\sum_{i=1}^n \left(\frac{N+1}{2}-\|R(Y_i)-\frac{N+1}{2}\|\right)$	$\frac{n(N+2)}{4}$	$\frac{nm(N-2)(N+2)}{48(N-1)}$
log-rank, Savage	$\sum_{i=1}^n \sum_{j=N-R(Y_i)+1}^N \frac{1}{j}$	n	$\frac{nm}{N-1}\left(1-\frac{1}{N}\sum_{j=1}^N \frac{1}{j}\right)$
The rank score used by Savage's test is $a(r)=\sum_{j=N-r+1}^N \frac{1}{j} \approx -\log\left(1-\frac{r}{N+1}\right)$.			
Tests for I samples X_{ij} for $i=1,\dots,I$ and $j=1,\dots,n_i$. The ranks are computed on the pooled sample. Here $N=n_1+\cdots+n_I$.			
Kruskal-Wallis	$\frac{12}{N(N+1)}\sum_{i=1}^I \frac{1}{n_i}\left[\sum_{j=1}^{n_i} R(X_{ij})-n_i\frac{N+1}{2}\right]^2$	has an approximate χ^2_{I-1} null distribution.	
Tests of independence based on a sample $(X_1,Y_1),\dots,(X_n,Y_n)$ of a bivariate random vector. The components are ranked individually and result in $R_1,\dots,R_n$ and $S_1,\dots,S_n$. Under the null, the two components are independent.			
Spearman's ρ	$\frac{12}{n^3-n}\sum_{i=1}^n \left(R_i-\frac{n+1}{2}\right)\left(S_i-\frac{n+1}{2}\right)$	0	$\frac{1}{n-1}$
Kendall's τ	$\frac{1}{n(n-1)}\sum\sum_{i\neq j} \operatorname{sign}(R_i-R_j)\operatorname{sign}(S_i-S_j)$	0	$\frac{2(2n+5)}{9n(n-1)}$
Test for regression coefficients. The data are (x_i,Y_i) for $(i=1,\dots,n)$, where $Y_i=x_i^T\beta+\mu+\varepsilon_i$. We test $H_0:\beta=\beta_0$. The ranks are based on the residuals $Y_i-x_i^T\beta_0$ and the rank scores are $a(1)\leq\ldots\leq a(n)$, which satisfy $a(1)+\cdots+a(n)=0$.			
Rank Regression	$\sum_{i=1}^n a(R_i)(Y_i-x_i^T\beta_0)$	the rank estimator $\hat\beta$ is the value of β_0, which minimizes this statistic.	

Part 4
Nonparametric regression

4.1. Smoothing

We consider regression model where the response variable y varies smoothly with the covariate x. Suppose we have measurements $y_i = f(x_i) + \varepsilon_i$ where $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ at equally spaced nodes $x_i = x_0 + (i-1)\Delta$ for $i = 1, \ldots, n$.



We wish to estimate f from the observations. Our goal is to estimate $f(x)$ by $\widehat{f}(x)$; a simple linear estimator is of the form

$$\widehat{f}(x) = \frac{1}{W}\sum_{i=1}^{n} W_i y_i = \frac{1}{W(x)}\sum_{i=1}^{n} W_i(x) y_i$$

for weights $W, \{W_i\}$. This leads naturally, if one averages over values close to unobserved x's, to

$$\widehat{f}(x_i) = \frac{1}{2k+1}\sum_{j=-k}^{k} Y_{i+j}$$

a moving-average estimator akin to those used in time-series. This estimator, the mean of nearest neighbours, does not work at the boundary. Its mean and variance are

$$\begin{aligned}
\widehat{f}(x_i) &= \frac{1}{2k+1}\sum_{j=-k}^{k}\left(f(x_{i+j})+\varepsilon_{i+j}\right) \\
&= \frac{1}{2k+1}\sum_{j=-k}^{k} f(x_{i+j}) + \sum_{j=-k}^{k}\frac{\varepsilon_{i+j}}{2k+1}
\end{aligned}$$

which means

$$\begin{aligned}
\mathsf{E}\left(\widehat{f}(x_i)\right) &= \frac{1}{2k+1}\sum_{j=-k}^{k} f(x_{i+j}) \\
\mathsf{Var}\left(\widehat{f}(x_i)\right) &= \frac{\sigma^2}{2k+1}.
\end{aligned}$$

4.1.1. Bias and variance of nearest neighbours smoothers

We begin with a definition.

Definition 4.1 (Hölder continuity)
If $f(x) \in \mathcal{C}^1([0,1])$, namely it has a continuous first derivative, then $f(x)$ has a **Hölder continuous** first derivative if

$$|f'(x) - f'(y)| \leq C|x-y|^{\beta}, \quad 0 < \beta \leq 1$$

for all $x, y \in [0,1]$. If f had a second derivative at x, then $|f'(x) - f'(y)| = \mathrm{o}(|x-y|)$ $(\beta \geq 1)$. In the case $\beta = 1$, f' has a Lipschitz-continuous first derivative.

We usually consider $\mathcal{C}^p$ functions for $p \geq 2$ or assume Hölder continuity of the second derivative. Suppose $\mathsf{E}(\varepsilon_i) = 0, \varepsilon_i \overset{\text{iid}}{\sim} F$ and $\mathsf{Var}(\varepsilon_i) = \sigma^2 < \infty$. We consider fitted values

$$\widehat{Y}_i = \frac{1}{2k+1}\sum_{j=-k}^{k} Y_{i+j}$$

Theorem 4.2 (Optimal bandwidth for nearest neighbours smoother)
Suppose $f : [0,1] \to \mathbb{R}$ is twice continuous differentiable and $|f''(x)| \leq M$ for all $0 \leq x \leq 1$ since we are looking at a compact set. In this case, $k = \mathrm{O}(n^{4/5})$ is the optimal spacing choice and leads to a mean squared error of $\mathrm{O}(n^{4/5})$.

Proof We already have $\mathsf{Var}\left(\widehat{Y}_i\right) = \sigma^2/(2k+1) = \mathrm{O}(k^{-1})$. A Taylor's expansion of $f(x)$ around x_i gives

$$f(x) = f(x_i) + f'(x_i)(x-x_i) + f''(\xi_x)\frac{(x-x_i)^2}{2}.$$

The bias is

$$
\begin{aligned}
\mathsf{bias}(\widehat{Y}_i) &= \mathsf{E}\left(\widehat{Y}_i - f(x_i)\right) \\
&= \frac{1}{2k+1}\sum_{j=-k}^{k}\left(f'(x_i)(x_{i+j}-x_i) + f''(\xi_{x_{i+j}})\frac{(x_{i+j}-x_i)^2}{2}\right) \\
&= \frac{1}{2k+1}\frac{f'(x_i)}{n-1}\sum_{j=-k}^{k} j + \frac{1}{2k+1}\sum_{j=-k}^{k}\frac{1}{2}f''(\xi_{x_{i+j}})\left(\frac{j}{n-1}\right)^2
\end{aligned}
$$

since our points x_i are equally spaced, the distance between any two adjacent points between $x_1 = 0$ and $x_n = 1$ is $1/(n-1)$. The first sum vanishes since it involves an odd function. Thus

$$
\begin{aligned}
\left|\mathsf{bias}(\widehat{Y}_i)\right| &\leq \frac{M}{2(2k+1)}\left(\frac{1}{n-1}\right)^2\sum_{j=-k}^{k} j^2 \\
&= \frac{M}{2(2k+1)(n-1)^2}\frac{k(k+1)(2k+1)}{3} \\
&= \mathrm{O}\left(\frac{k^2}{n^2}\right)
\end{aligned}
$$

If we put $k(n) = n^\alpha = nh(n)$, then $h(n) = n^{\alpha-1}$ is called the **bandwidth**.

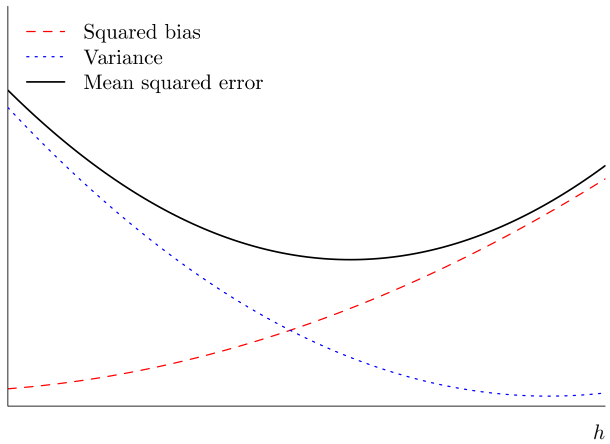


Figure 15: Squared bias and variance trade-off curve

The mean squared error of $\widehat{Y}_i$ or the sum over all $\widehat{Y}_i$ is equal to

$$
\mathsf{bias}(\widehat{Y}_i)^2 + \mathsf{Var}(\widehat{Y}_i) = \mathrm{O}\left(\frac{k^4}{n^4}\right) + \mathrm{O}\left(\frac{1}{k}\right) = \mathrm{O}(h^4) + \mathrm{O}\left(\frac{1}{nh}\right)
$$

and the bias-variance tradeoff is minimized when both are of the same order. If we want $h^4 = 1/nh$, this is equivalent to $h^5 = n^{-1}$ and so choosing $h = n^{-1/5}$ leads to

$k = hn = n^{4/5}$. Thus, the mean-squared error is of order $\mathrm{O}(n^{-4/5})$, since it is the order of both the squared bias and the variance. ■

In parametric and robust statistics, we are used to $\mathrm{O}(n^{-1})$. As the dimension of the smoothing problems increases from $d = 1$, the rate of the the mean squared error becomes $\mathrm{O}(n^{-4/(4+d)})$. This is sometimes known as the curse of dimensionality.

In determining the optimal value of the bandwidth h, both σ^2 and f'' play a role.

If we assume Hölder continuity requirement instead of boundedness of the second derivative, we get from the Hölder bound

$$\mathsf{bias}(\widehat{f}(x_i)) \leq C \sum_{|j|\leq k} \left|\frac{j}{n-1}\right| \left|\frac{j}{n-1}\right|^{\beta}.$$

If $\beta = 1$, we recover the result of our theorem. One cans show that with $f \in \mathcal{C}^s([0,1])$ and Hölder continuous with coefficient β, the optimal mean squared error is of order $\mathrm{O}(n^{-r})$ where $r = 2(s+\beta)/[2(s+\beta)+1]$. As s increases, r approaches 1. If $\beta = s = 1$, we get the rate $r = 4/5$.

4.2. Smoothing splines

4.2.1. Splines

A spline of degree p with knots $t_1 < t_2 < \cdots < t_k$ is a function $s : [t_1, t_k] \to \mathbb{R}$ such that

- it has $p-1$ continuous derivatives $s', s'', \ldots, s^{(p-1)}$;
- on each interval $[t_i, t_{i+1}]$, it is a polynomial of degree p.

We will call the space of these functions $\mathcal{S}^p(x_1, \ldots, x_n)$.

Proposition 4.3 (Dimension of the spline space)
$\mathcal{S}^p(x_1, \ldots, x_k)$ is a vector space of dimension $k + p - 1$

Proof For $s_1, s_2 \in \mathcal{S}^p(x_1, \ldots, x_n)$, it is obvious that $a_1 s_1 + a_2 s_2 \in \mathcal{S}^p(x_1, \ldots, x_n)$ if $a_1, a_2 \in \mathbb{R}$. The dimension is $(k-1)(p+1) - (k-2)p = k + p - 1$ because of the restrictions: there are $k-2$ interior points, each with p restrictions due to the requirement of continuity of the derivatives. ■

In order to fit splines, choose a basis $b_1(x), \ldots, b_{k+p-1}(x)$ and find $\widehat{\theta}_1, \ldots, \widehat{\theta}_{k+p-1}$ such that the observations $y_1, \ldots, y_n$ are close to $\widehat{y}_i = \widehat{\theta}_1 b_1(x_i) + \cdots + \widehat{\theta}_{k+p-1} b_{k+p-1}(x_i)$. If

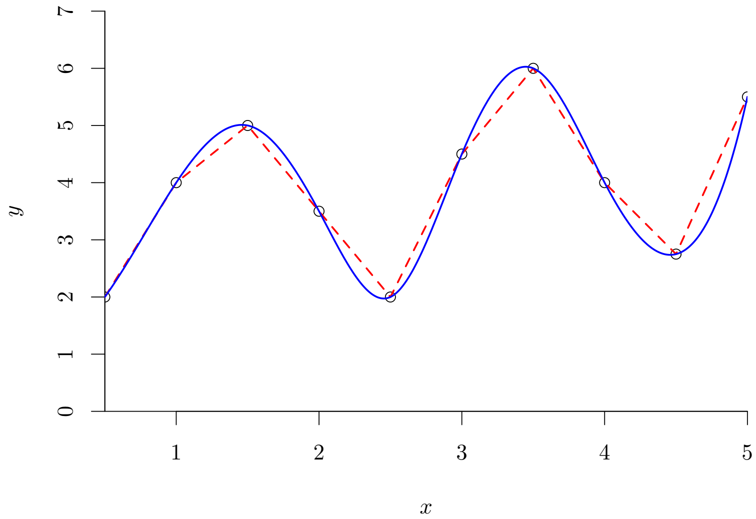


Figure 16: Cubic and linear spline fit with knots at the observations.

we fit by least squares, we obtain

$$\widehat{\boldsymbol{\theta}} = \begin{pmatrix} \widehat{\theta}_1 \\ \vdots \\ \widehat{\theta}_{k+p-1} \end{pmatrix} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \boldsymbol{y}$$

with

$$\mathbf{X} = \begin{pmatrix} b_1(x_1) & \cdots & b_{k+p-1}(x_1) \\ \vdots & \ddots & \vdots \\ b_1(x_n) & \cdots & b_{k+p-1}(x_n) \end{pmatrix}.$$

The fitted values will be

$$\widehat{\boldsymbol{y}} = \mathbf{X}\widehat{\boldsymbol{\theta}} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{y} = \mathbf{H}\boldsymbol{y}$$

where $\mathbf{H}$ is a linear application and thus the smoother is linear. $\mathbf{H}$ is a projection matrix (it projects onto the span of $\mathbf{X}$) and thus all its eigenvalues are either 0 or 1. Other interesting things to look at are the eigenvectors of $\mathbf{H}$, filtering the signal if the eigenvalue is 0 and letting it pass if the eigenvalue is 1. The columns of $\mathbf{H}$ are impulse responses.

4.2.2. Optimality properties of splines

Proposition 4.4
Let $f(x)$ be a differentiable such that $\int_{x_1}^{x_n}(f'(x))^2\,\mathrm{d}x < \infty$. Let $s(x) \in \mathcal{S}^1(x_1,\dots,x_n)$ with $s(x_i) = f(x_i)$ for $i = 1,\dots,n$. Then

$$\int \left(f'(x)\right)^2\,\mathrm{d}x \geq \int (s'(x))^2\,\mathrm{d}x,$$

This property has analogues, as for example cubic splines minimize $\int_{x_1}^{x_n}(s''(x))^2\,\mathrm{d}x$.

Proof It is enough to consider a single interval $[x_i, x_{i+1})$. Write

$$\begin{aligned}\int_{x_i}^{x_{i+1}} \left(f'(x) - s'(x)\right)^2\,\mathrm{d}x = \int_{x_i}^{x_{i+1}} \left(f'(x)\right)^2\,\mathrm{d}x - \int_{x_i}^{x_{i+1}} \left(s'(x)\right)^2\,\mathrm{d}x \\ - 2\int_{x_i}^{x_{i+1}} s'(x)\left(f'(x) - s'(x)\right)\,\mathrm{d}x.\end{aligned}$$

The integral of a square is positive, so it remains to prove that the last term is zero. Using integration by part

$$\int_{x_i}^{x_{i+1}} s'(x)\left(f'(x) - s'(x)\right)\,\mathrm{d}x = s'(x)\left(f(x) - s(x)\right)\Big|_{x_i}^{x_{i+1}} - \int_{x_i}^{x_{i+1}} s''(x)\left(f(x) - s(x)\right)\,\mathrm{d}x.$$

The first term equals zero because $f(x) = s(x)$ at the data points x_i, x_{i+1} and the second term vanishes because $s''(x) = 0$. ■

In more generality, the odd degree splines have an interesting property. For a spline of degree $2m-1$, $\int s^{(m)}(x)^2\,\mathrm{d}x$ is a minimum. The difficulty with this approach is the choice of the knots location. You want many knots where the curvature of f is high and less knots where the function is flat or constant.

4.2.3. Natural smoothing splines

Take $(x_1,\dots,x_n)$ as knots and consider the space of **natural splines**

$$\mathcal{S}^3_{\mathrm{n}} := \left\{ s \in \mathcal{S}^3(x_1,\dots,x_n) : s''(x_1) = s''(x_n) = 0 \right\}.$$

Note that $\dim(\mathcal{S}^3(x_1,\dots,x_n)) = n + p - 1 = n + 2$. The 2 additional conditions reduce the dimension of $\mathcal{S}^3_{\mathrm{n}}$ to n.

Problem 4.1 (Rensch's problem)
Find $s(x)$, a natural cubic spline, such that $\sum_{i=1}^{n}(Y_i - s(x_i))^2 \leq M$ for $M \geq 0$ and $s(x)$ should minimize $\int_{x_1}^{x_n}(s''(x))^2\,\mathrm{d}x$. If $M = 0$, this yields the **interpolating spline**, which passes through all the points. If M is large, $s(x)$ is smooth.

The cubic spline $s(x)$ should minimize $\sum_{i=1}^{n}(Y_i - s(x_i))^2 + \lambda \int_{x_1}^{x_n} (s''(x))^2 \, \mathrm{d}x$.[4] This is equal to the goodness-of-fit plus λ times a smoothness penalty. This is the same problem as Rensch's, except that it is formulated using Lagrange multipliers. The case of $\lambda = 0$ gives the interpolating spline (rough), while $\lambda \to \infty$ is the least-square fit (smooth). The best choice in terms of variance-bias trade-off lies in between. We are now going to study the computations of $s(x)$.

For computational purposes there are two convenient parametrizations. The first is obtained by writing $s(x) \in \mathcal{S}^3(x_1, \dots, x_n)$ as a linear combination of the B-spline basis. B-splines have minimal support and can be computed efficiently. The second is by way of the second derivatives at the interior knots, where we work with natural splines for which the second derivative is zero at the two extreme knots. When using the B-spline basis, these two constraints have to be explicitly added.

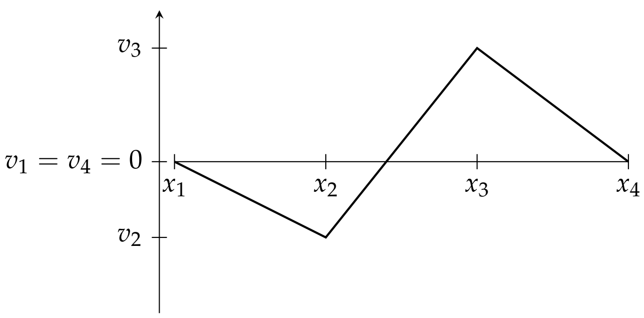
4.2.4. Natural splines determined by their values and second derivatives

Proposition 4.5 (Solution to Rensch's problem with the natural splines formulation) We denote the node vector by $\boldsymbol{x} = (x_1, \dots, x_n)$ and by $d_i = x_{i+1} - x_i$, $1 \le i < n$, the gap between successive knots. Consider the vectors $\boldsymbol{s} = (s(x_1), \dots, s(x_n))^\top \in \mathbb{R}^n$ and $\boldsymbol{v} = (s''(x_2), \dots, s''(x_{n-1}))^\top \in \mathbb{R}^{n-2}$. Indeed, because the spline is natural, we do not need to record values $v_1 = s''(x_1) = v_n = s''(x_n) = 0$.

If $s \in \mathcal{S}^3(x_1, \dots, x_n)$, then $s'' \in \mathcal{S}^1(x_1, \dots, x_n)$. The second derivative $s''(x)$ is a continuous piecewise linear function, with value v_i at the interior knot x_i and zero at the end knots. It thus is for $x_i \le x \le x_{i+1}$ equal to

$$s''(x) = \frac{v_{i+1} - v_i}{x_{i+1} - x_i}(x - x_i) + v_i = \frac{(x - x_i)v_{i+1} + (x_{i+1} - x)v_i}{d_i},$$

which is a linear function in x with the required values at the boundaries of the interval.



Slope from v_i to v_{i+1}:

$$\frac{v_{i+1} - v_i}{x_{i+1} - x_i}$$

If we integrate it twice and choose the constants of integration in order to obtain the

[4] For a spline of degree p, it should minimize $\sum_{i=1}^{n}(Y_i - s(x_i))^2 + \lambda \int_{x_1}^{x_n} s^{\left(\frac{p+1}{2}\right)}(x)^2 \, \mathrm{d}x$

required values s_i at the knots, we obtain

$$s(x) = \frac{(x - x_i)s_{i+1} + (x_{i+1} - x)s_i}{d_i} - \frac{1}{6}(x - x_i)(x_{i+1} - x)\Xi \tag{4.14}$$

where

$$\Xi = \left\{ \left(1 + \frac{x - x_i}{d_i}\right) v_{i+1} + \left(1 + \frac{x_{i+1} - x}{d_i}\right) v_i \right\}.$$

The last equation requires checking. Clearly, $s(x)$ is a piecewise cubic and its values at the knots are as required, $s(x_i) = s_i$. This is evident, because the cubic part is equal to zero at the knots and the linear first term interpolates the values s_i. The first derivative is easily calculated as

$$s'(x) = \frac{s_{i+1} - s_i}{d_i} - \frac{1}{6}(x_{i+1} - x)\Xi + \frac{1}{6}(x - x_i)\Xi - \frac{1}{6}(x - x_i)(x_{i+1} - x)\left[\frac{v_{i+1} - v_i}{d_i}\right],$$

while the second derivative is

$$s''(x) = \frac{2}{6}\Xi - \frac{2}{6}(x_{i+1} - x)\left[\frac{v_{i+1} - v_i}{d_i}\right] + \frac{2}{6}(x - x_i)\left[\frac{v_{i+1} - v_i}{d_i}\right].$$

The last equation shows that the cubic polynomial in eq. (4.14) does have indeed the correct derivative values at the knots, because

$$s''(x_i) = \frac{1}{3}\{v_{i+1} + (1+1)v_i\} - \frac{1}{3}(v_{i+1} - v_i) = v_i$$

and analogously $s''(x_{i+1}) = v_{i+1}$. The only way in which the function in eq. (4.14) can fail to be a spline is by discontinuity of the first derivative. Take any interior knot x_i for $1 < i < n$. There are two ways in which we can compute the first derivative. Once at the right boundary of the interval $[x_{i-1}, x_i]$, which leads to

$$s'(x_i) = \frac{s_i - s_{i-1}}{d_{i-1}} + \frac{1}{6}(x_i - x_{i-1})(2v_i + v_{i-1}),$$

and once at the left-hand boundary of $[x_i, x_{i+1}]$, which gives

$$s'(x_i) = \frac{s_{i+1} - s_i}{d_i} + \frac{1}{6}(x_{i+1} - x_i)(v_{i+1} + 2v_i).$$

These two values must be equal, that is,

$$\begin{aligned}\frac{s_{i+1} - s_i}{d_i} - \frac{s_i - s_{i-1}}{d_{i-1}} &= \frac{s_{i-1}}{d_{i-1}} - \left(\frac{1}{d_{i-1}} + \frac{1}{d_i}\right)s_i + \frac{s_{i+1}}{d_i} \\ &= \frac{1}{6}d_{i-1}v_{i-1} + \frac{1}{3}(d_i + d_{i-1})v_i + \frac{1}{6}d_i v_{i+1}\end{aligned}$$

which can be written as a system of $n-2$ linear equations linking the vectors $\boldsymbol{s}$ and $\boldsymbol{v}$, namely $\boldsymbol{Q}^\top \boldsymbol{s} = \boldsymbol{R}\boldsymbol{v}$ where $\boldsymbol{Q} \in \mathbb{R}^{n\times(n-2)}$ is a tri-band matrix and $\boldsymbol{R} \in \mathbb{R}^{(n-2)\times(n-2)}$ is a square, invertible and tri-band matrix with diagonal values $(d_i - d_{i-1})/3$.

The smoothness penalty is equal to the integral of the square of the second derivative of the fitted function, $(s''(x))^2\,\mathrm{d}x$. We have

$$\begin{aligned}
\int_{x_1}^{x_n} (s''(x))^2\,\mathrm{d}x &= s''(x)s'(x)\Big|_{x_1}^{x_n} - \sum_{i=1}^{n-1}\int_{x_i}^{x_{i+1}} s'''(x)s'(x)\,\mathrm{d}x \\
&= 0 - \sum_{i=1}^{n-1}\left(\frac{v_{i+1}-v_i}{d_i}\right)(s_{i+1}-s_i) \\
&= \sum_{i=2}^{n-1} v_i \left(\frac{s_{i-1}-s_i}{d_{i-1}} - \frac{s_i - s_{i+1}}{d_i}\right) \\
&= \boldsymbol{v}^\top \boldsymbol{Q}^\top \boldsymbol{s}
\end{aligned}$$

where we made use of the fact that $s''(x_1) = s''(x_n) = 0$ and that $s'''(x)$ is constant on each of the intervals. The smoothness penalty can be re-expressed as

$$\boldsymbol{v}^\top \boldsymbol{Q}^\top \boldsymbol{s} = \boldsymbol{v}^\top \boldsymbol{R}\boldsymbol{v} = \boldsymbol{v}^\top \boldsymbol{R}\boldsymbol{R}^{-1}\boldsymbol{Q}^\top \boldsymbol{s} = \boldsymbol{s}^\top \boldsymbol{Q}\boldsymbol{R}^{-1}\boldsymbol{Q}^\top \boldsymbol{s}.$$

since $\boldsymbol{R} = \boldsymbol{R}^\top$. For a natural cubic spline $s(x)$, we have in summary to minimize

$$\begin{aligned}
&\sum_{i=1}^{n}(Y_i - s(x_i))^2 + \lambda \int_{x_1}^{x_n} \left(s''(x)\right)^2\,\mathrm{d}x \qquad (4.15)\\
&= (\boldsymbol{Y}-\boldsymbol{s})^\top(\boldsymbol{Y}-\boldsymbol{s}) + \lambda \boldsymbol{v}^\top \boldsymbol{R}\boldsymbol{v} \\
&= \boldsymbol{Y}^\top\boldsymbol{Y} - 2\boldsymbol{Y}^\top\boldsymbol{s} + \boldsymbol{s}^\top\boldsymbol{s} + \lambda \boldsymbol{s}^\top \boldsymbol{Q}\boldsymbol{R}^{-1}\boldsymbol{Q}^\top \boldsymbol{s}
\end{aligned}$$

From this, the criterion that $\boldsymbol{s}$ should minimize is

$$\boldsymbol{s}^\top \left(\mathbf{I} + \lambda \boldsymbol{Q}\boldsymbol{R}^{-1}\boldsymbol{Q}^\top\right)\boldsymbol{s} - 2\boldsymbol{s}^\top\boldsymbol{Y} + \boldsymbol{Y}^\top\boldsymbol{Y}.$$

This leads to

$$2\left(\mathbf{I} + \lambda \boldsymbol{Q}\boldsymbol{R}^{-1}\boldsymbol{Q}^\top\right)\boldsymbol{s}_{\text{opt}} - 2\boldsymbol{Y} = 0.$$

The optimal solution can also be written in terms of $\boldsymbol{v}$:

$$\boldsymbol{s}_{\text{opt}} = \boldsymbol{Y} - \lambda \boldsymbol{Q}\boldsymbol{R}^{-1}\boldsymbol{Q}^\top \boldsymbol{s}_{\text{opt}} = \boldsymbol{Y} - \lambda \boldsymbol{Q}\boldsymbol{v}_{\text{opt}}$$

meaning

$$\boldsymbol{Q}^\top \boldsymbol{s}_{\text{opt}} = \boldsymbol{R}\boldsymbol{v}_{\text{opt}} = \boldsymbol{Q}^\top \boldsymbol{Y} - \lambda \boldsymbol{Q}^\top \boldsymbol{Q}\boldsymbol{v}_{\text{opt}}$$

and thus

$$v_{\mathrm{opt}} = \left(\boldsymbol{R} + \lambda \boldsymbol{Q}^\top \boldsymbol{Q}\right)^{-1} \boldsymbol{Q}^\top \boldsymbol{Y}.$$

Computationally it is easiest to compute first $\boldsymbol{v}$ and then $\boldsymbol{s}$, because it avoids the computation of $\boldsymbol{R}^{-1}$. It turns out that $\boldsymbol{R} + \lambda \boldsymbol{Q}^\top \boldsymbol{Q}$ is banded of width 5 (with a main diagonal, and two diagonals above and two below), and that

$$\boldsymbol{Q}^\top \boldsymbol{Y} = \frac{Y_{i+1} - Y_i}{d_i} - \frac{Y_i - Y_{i-1}}{d_{i-1}}.$$

Proposition 4.6
The cubic smoothing spline is **the** natural spline that minimizes the criterion (4.15) among all functions defined on $[x_1, x_n]$ with absolutely continuous first derivative.[5]

Proof Suppose $f(x)$ is a good smooth element of this space close to the observations $\boldsymbol{Y}$. Let $s(x)$ be the natural cubic spline such that $f(x_i) = s(x_i)$, meaning the goodness-of-fit criterion in eq. (4.15) is the same for $f(x)$ and $s(x)$. But from the optimality properties of splines, which states the the squared second derivative is minimized among all such functions interpolating a given vector $\boldsymbol{s}$, we have

$$\int \left(s''(x)\right)^2 \mathrm{d}x \le \int \left(f''(x)\right)^2 \mathrm{d}x,$$

which proves the claim. ■

The reason for the popularity of smoothing splines is the low dimension of the spline space.

$$s(x) = \sum_{j=1}^{n} b_j(x)\theta_j,$$

where $(b_1(x), b_2(x), \ldots, b_n(x))$ is a basis of the natural cubic spline space.

4.2.5. Summary on linear smoothers

The data comes in pairs $(x_1, Y_1), \ldots, (x_n, Y_n)$, such that the observations $Y_i = f(x_i) + \varepsilon_i$ where ε_i is a random error. Thus, we have a noisy observation of a smooth function and attempt to retrieve the signal, as $\widehat{\boldsymbol{Y}} = \mathbf{H}\boldsymbol{Y}$. Thus,

$$\widehat{f}(x) = \sum_{j=1}^{n} Y_j h_j(x), \qquad x \in \{x_1, \ldots, x_n\}$$

and $h_j(x_i)$ returns the i^{th} element of the column vector $\mathbf{h}_j$, namely H_{ij}.

[5] $f'(x)$ absolutely continuous means there exists $f'(x)$ is such that $\int_a^b f''(x)\,\mathrm{d}x = f'(b) - f'(a)$ for all $x_1 \le a \le b \le x_n$.

Example 4.1 (Linear regression)
If $f(x) = \theta_0 + \theta_1 b_1(x) + \cdots + \theta_p b_p(x)$ is in the $(p+1)$ dimensional space, the

$$\mathbf{H} = \mathbf{B}(\mathbf{B}^\top \mathbf{B})^{-1}\mathbf{B}^\top$$

where $\mathrm{B}_{ij} = (b_j(x_i)) \in \mathbb{R}^{n\times(p+1)}$. The eigenvalues of $\mathbf{H}$ are equal to either 0 or 1. The trace $\mathrm{tr}(\mathbf{H}) = p+1$, corresponding to the number of coefficients, or equivalently the number of degrees of freedom of the model.

It is often true that $\mathbf{H}\mathbf{1} = \mathbf{1}$, given the sum of the rows is 1.

Example 4.2 (Nearest neighbour smoothing)
For $k = 1, n = 7$, we have

$$\mathbf{H} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

Generally, each row contains the value $1/(2k+1)$ in position $i-k, \ldots, i+k$. You need to adapt ad-hoc the values near the edges. This Toeplitz structure gives rise to interesting patterns, since the eigenfunctions behave like sinusoids.

The eigenvalues are no longer 0 or 1 and may even be negative.

Example 4.3 (Smoothing splines)
The matrix $\mathbf{H}$ remains a bit hidden, unless we work with basis functions. Consider a basis consisting of $b_1(x), \ldots b_n(x)$. The smoothing spline is then of the form $\sum_{j=1}^n \theta_j b_j(x)$ and $\widehat{\boldsymbol{Y}} = \mathbf{B}\boldsymbol{\theta}$, with again $\mathrm{B}_{ij} = b_j(x_i)$. We can express the sum of squared residuals as

$$\begin{aligned}\sum_{i=1}^n r_i^2 &= (\boldsymbol{Y} - \widehat{\boldsymbol{Y}})^\top (\boldsymbol{Y} - \widehat{\boldsymbol{Y}}) \\ &= (\boldsymbol{Y} - \mathbf{B}\boldsymbol{\theta})^\top (\boldsymbol{Y} - \mathbf{B}\boldsymbol{\theta})\end{aligned}$$

while

$$\begin{aligned}\int_{x_1}^{x_n} \left(s''(x)\right)^2 \mathrm{d}x &= \int_{x_1}^{x_n} \left(\sum_{j=1}^n \theta_j b_j''(x)\right)^2 \mathrm{d}x \\ &= \sum_{j=1}^n \sum_{k=1}^n \theta_j \theta_k \int_{x_1}^{x_n} b_j''(x) b_k''(x)\, \mathrm{d}x \\ &= \boldsymbol{\theta}^\top \boldsymbol{\Omega} \boldsymbol{\theta}\end{aligned}$$

where the entries of $\mathbf{\Omega}$ are $\Omega_{jk} = \int_{x_1}^{x_n} b_j''(x) b_k''(x) \, \mathrm{d}x$. The resulting criterion is

$$\mathrm{Crit}(s) = (\boldsymbol{Y} - \mathbf{B}\boldsymbol{\theta})^\top (\boldsymbol{Y} - \mathbf{B}\boldsymbol{\theta}) + \lambda \boldsymbol{\theta}^\top \mathbf{\Omega} \boldsymbol{\theta}$$

and

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}} \mathrm{Crit}(s) \bigg|_{\widehat{\boldsymbol{\theta}}} = -2\mathbf{B}^\top \boldsymbol{Y} + 2\mathbf{B}^\top \mathbf{B}\widehat{\boldsymbol{\theta}} + 2\lambda \mathbf{\Omega} \widehat{\boldsymbol{\theta}} = 0$$

yielding

$$(\mathbf{B}^\top \mathbf{B} + \lambda \mathbf{\Omega})\widehat{\boldsymbol{\theta}} = \mathbf{B}^\top \boldsymbol{Y}.$$

The matrix $\mathbf{H}$ is such that

$$\begin{aligned}
\widehat{\boldsymbol{Y}} &= \mathbf{B}\widehat{\boldsymbol{\theta}} \\
&= \mathbf{B}(\mathbf{B}^\top \mathbf{B} + \lambda \mathbf{\Omega})^{-1} \mathbf{B}^\top \boldsymbol{Y} \\
&= \mathbf{H}\boldsymbol{Y}
\end{aligned}$$

and thus

$$\mathbf{H} = \mathbf{B}(\mathbf{B}^\top \mathbf{B} + \lambda \mathbf{\Omega})^{-1} \mathbf{B}^\top$$

4.2.6. Estimation of the variance of the random errors

We consider the natural estimator of the variance, the residuals sum of squares

$$\mathsf{RSS} = \sum_{i=1}^n (Y_i - \widehat{Y}_i)^2 = \sum_{i=1}^n (Y_i - \widehat{f}(x_i))^2 = \sum_{i=1}^n (Y_i - s(x_i))^2,$$

which is the basis for estimation of $\mathsf{Var}\,(\varepsilon_i) = \mathsf{Var}\,(Y_i - f(x_i))$. For linear smoothers,

$$\begin{aligned}
\mathsf{E}\,(\mathsf{RSS}) &= \mathsf{E}\left((\boldsymbol{Y} - \widehat{\boldsymbol{Y}})^\top (\boldsymbol{Y} - \widehat{\boldsymbol{Y}})\right) \\
&= \mathsf{E}\left(((\mathbf{I} - \mathbf{H})\boldsymbol{Y})^\top ((\mathbf{I} - \mathbf{H})\boldsymbol{Y})\right) \\
&= \mathsf{E}\left((\boldsymbol{Y} - \boldsymbol{f})^\top (\mathbf{I} - \mathbf{H})^\top (\mathbf{I} - \mathbf{H})(\boldsymbol{Y} - \boldsymbol{f})\right) - \mathsf{E}\left(\boldsymbol{f}^\top (\mathbf{I} - \mathbf{H})^\top (\mathbf{I} - \mathbf{H})\boldsymbol{f}\right) \\
&\quad + 2\mathsf{E}\left(\boldsymbol{f}^\top (\mathbf{I} - \mathbf{H})^\top (\mathbf{I} - \mathbf{H})\boldsymbol{Y}\right) \\
&= \mathsf{E}\left(\mathrm{tr}\left((\mathbf{I} - \mathbf{H})^\top (\mathbf{I} - \mathbf{H})(\boldsymbol{Y} - \boldsymbol{f})(\boldsymbol{Y} - \boldsymbol{f})^\top\right)\right) - \boldsymbol{f}^\top (\mathbf{I} - \mathbf{H})^\top (\mathbf{I} - \mathbf{H})\boldsymbol{f} \\
&\quad + 2\boldsymbol{f}^\top (\mathbf{I} - \mathbf{H})^\top (\mathbf{I} - \mathbf{H})\mathsf{E}\,(\boldsymbol{Y}) \\
&= \mathsf{E}\left(\mathrm{tr}\left((\mathbf{I} - \mathbf{H})^\top (\mathbf{I} - \mathbf{H})(\boldsymbol{Y} - \boldsymbol{f})(\boldsymbol{Y} - \boldsymbol{f})^\top\right)\right) + \boldsymbol{f}^\top (\mathbf{I} + \mathbf{H})^\top (\mathbf{I} - \mathbf{H})\boldsymbol{f}
\end{aligned}$$

and the term $(\mathbf{I}-\mathbf{H})\boldsymbol{f}$, the vector of residuals obtained by smoothing $\boldsymbol{f}$, ought to be small for smooth functions $f(x)$. We can thus get a good approximation by neglecting this quadratic form and focusing on

$$\begin{aligned}\mathsf{E}\left(\mathsf{RSS}\right) &\cong \operatorname{tr}\left((\mathbf{I}-\mathbf{H})^\top(\mathbf{I}-\mathbf{H})\mathsf{E}\left((\boldsymbol{Y}-\boldsymbol{f})(\boldsymbol{Y}-\boldsymbol{f})^\top\right)\right)\\ &\cong \sigma^2\operatorname{tr}\left((\mathbf{I}-\mathbf{H})^\top(\mathbf{I}-\mathbf{H})\right)\\ &\cong \sigma^2\left(n-2\operatorname{tr}(\mathbf{H})+\operatorname{tr}(\mathbf{H}^\top\mathbf{H})\right)\end{aligned}$$

where $\boldsymbol{f}=(f(x_1),\ldots,f(x_n))^\top=\mathsf{E}\left(\boldsymbol{Y}\right)$. An appropriate estimator is

$$\widehat{\sigma^2}=\frac{\mathsf{RSS}}{n-2\operatorname{tr}(\mathbf{H})+\operatorname{tr}(\mathbf{H}^\top\mathbf{H})}$$

If $\mathbf{H}$ is an orthogonal projection, $\mathbf{H}=\mathbf{H}^\top=\mathbf{H}^2$ and then

$$\widehat{\sigma^2}=\frac{\mathsf{RSS}}{n-\mathrm{df}_{\text{model}}}=\frac{\mathsf{RSS}}{n-\operatorname{tr}(\mathbf{H})}$$

Note that $\operatorname{tr}(\mathbf{H}^\top\mathbf{H})=\sum_{i=1}^n\mathbf{h}_i^\top\mathbf{h}_i$, where $\mathbf{h}_1,\ldots\mathbf{h}_n$ are the columns of $\mathbf{H}$.

4.2.7. Cross-validation

In prediction problems, for example in predicting $f(x)$ or class(x), the class of individuals with attributes x_j, one often has **training data** to assist in estimating $\widehat{f}(x)$ on $\widehat{\text{class}}(x)$. But, from the fact that these predictions work well on the training data, one cannot conclude that they are good predictors. An independent **validation** on new data is called for. The cross validation is a way out of this.

Estimating the prediction error

Define the mean predictive squared error, MPSE, as

$$\mathsf{MPSE}=\frac{1}{n}\sum_{i=1}^n\mathsf{E}\left(\left[f(x_i)-\widehat{f}(x_i)\right]^2\right).$$

For linear smoothers, $\widehat{\boldsymbol{Y}}=\mathbf{H}\boldsymbol{Y}$ and thus

$$\widehat{Y}_i=(\mathbf{H}\boldsymbol{Y})_i=\sum_{j=1}^n Y_j h_j(x_i)$$

with $\mathbf{H1} = \mathbf{1}$ and as before $h_j(x_i) = \mathrm{H}_{ij}$. One can show that if we leave out the i^{th} observation to estimate f at x_i, $\widehat{f}(x_i)$, we have

$$\widehat{f}_i(x_i) = \sum_{j=1}^{n} h_j^{-i}(x_i) Y_j$$

where

$$h_j^{-i}(x_i) := \begin{cases} 0 & \text{if } j = i, \\ \dfrac{h_j(x_i)}{\left(\sum_{k \neq i} h_k(x_i)\right)} = \dfrac{h_j(x_i)}{1 - \mathrm{H}_{ii}} & \text{otherwise .} \end{cases}$$

The estimate of the MPSE based on leave-one-out cross-validation (CV) is

$$\begin{aligned} \widehat{\mathsf{MPSE}}_{\mathsf{CV}} &= \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \widehat{f}_i(x_i) \right)^2 \\ &= \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \sum_{j \neq i} Y_j \frac{h_j(x_i)}{1 - \mathrm{H}_{ii}} \right)^2 \\ &= \frac{1}{n} \sum_{i=1}^{n} \left(\left(\frac{1}{1 - \mathrm{H}_{ii}} \right) \left((1 - \mathrm{H}_{ii}) Y_i - \sum_{j \neq i} Y_j h_j(x_i) \right) \right)^2 \\ &= \frac{1}{n} \sum_{i=1}^{n} \frac{(Y_i - \widehat{Y}_i)^2}{(1 - \mathrm{H}_{ii})^2} \end{aligned}$$

We have seen this formula in the regression case.

Generalized cross-validation

Replace H_{ii} by $\mathrm{tr}(\mathbf{H})/n$, the average diagonal value. This yields

$$\begin{aligned} \widehat{\mathsf{MPSE}}_{\mathsf{GCV}} &= \frac{1}{n} \sum_{i=1}^{n} \frac{(Y_i - \widehat{Y}_i)^2}{\left(1 - \mathrm{tr}\left(\frac{\mathbf{H}}{n}\right)\right)^2} \\ &= \frac{n\mathsf{RSS}}{(n - \mathrm{tr}(\mathbf{H}))^2} \\ &\cong \frac{1}{n} \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2 \left(1 + \frac{2}{n} \mathrm{tr}(\mathbf{H})\right) \end{aligned} \tag{4.16}$$

for n large. The last approximation comes from a Taylor expansion (twice), for $\mathrm{tr}(\mathbf{H})$ is small compared to n. The criterion in eq. (4.16) is known as Mallow's C_p . The parameter $p \approx \mathrm{tr}(\mathbf{H})$ and C_p yields approximately the AIC.

Estimating the smoothing parameter λ

How to choose the smoothing parameter (which could be the bandwidth λ, k/n, the smoothing spline penalty, etc.)? A possible principle for estimation is the prediction error: we choose λ such that $\mathsf{MPSE}(f_\lambda)$ is small! That is, in practice $\widehat{\mathsf{MPSE}}_{\mathsf{CV}}(\widehat{f}_\lambda)$ is small; see Figure 15. In order to apply this idea, one must make a sequence of fits for different values of λ, compute $\widehat{f}_\lambda(x_i)$ and $\widehat{\mathsf{MPSE}}_{\mathsf{GCV}}$ and minimize it.

For the sequel, we consider the approach of Sun & Loader (1994). Note that we choose to find a band for $\bar{f}$ rather than f, thus avoiding the problem of bias.

For linear smoothers,

$$\widehat{f}(x) = \sum_{j=1}^{n} Y_j h_j(x).$$

If we base ourselves on this formula, we get

$$\begin{aligned}\mathsf{E}\left(\widehat{f}(x)\right) &= \sum_{j=1}^{n} f(x_j) h_j(x),\\ \mathsf{Var}\left(\widehat{f}(x)\right) &= \sum_{j=1}^{n} \sigma^2 h_j^2(x).\end{aligned}$$

The confidence band can be obtained as

$$\left[\widehat{f}(x) \pm c_\alpha \sigma \sqrt{\sum_{j=1}^{n} h_j^2(x)}\right].$$

The aim is to choose c_α such that $\bar{f}(x)$ is inside the band for all $x \in [a, b]$ with probability $1 - \alpha$. We want

$$\left|\widehat{f}(x) - \bar{f}(x)\right| \leq c_\alpha \sigma \sqrt{\sum_{j=1}^{n} h_j^2(x)}$$

for all $x \in [a, b]$. This is equivalent to

$$\max_{a \leq x \leq b} \frac{\sum_{j=1}^{n} Y_j h_j(x) - \sum_{j=1}^{n} f(x_j) h_j(x)}{\sigma \sqrt{\sum_{j=1}^{n} h_j^2(x)}} \leq c_\alpha.$$

If we work on the numerator, we can make it in a single sum, which combined with σ yields a normalized quantity $Z_j = (Y_j - f(x_j))/\sigma = \varepsilon_j/\sigma$. We assume normality of the errors ε_j, meaning $Z_j \sim \mathcal{N}(0, 1)$ are independent standard normal variables. The problem reduces to the quantity

$$\max_{a \leq x \leq b} \sum_{j=1}^{n} \left| Z_j \frac{h_j}{\sqrt{\sum_{j=1}^{n} \leq c_\alpha h_j^2(x)}} \right|.$$

Finding the tail probabilities of this process and subsequently the constant c_α such that it holds with probability $1-\alpha$ is a well-studied problem.

4.3. Kernel smoothers and local regression

4.3.1. Kernel smoothers

A kernel $k(w) \geq 0$ is a positive function used to give weights to the different observations when estimating $f(x)$, estimating $Y_i = f(x_i) + \varepsilon_i$ where the errors ε_i have mean zero and variance σ^2. The **Nadaraja–Watson estimator**, which leads to the nearest-neighbour smoothers, is

$$\widehat{f}(x) = \frac{\sum_{i=1}^n w_i Y_i}{\sum_{i=1}^n w_i} = \frac{\sum_{i=1}^n k(x_i - x) Y_i}{\sum_{i=1}^n k(x_i - x)}.$$

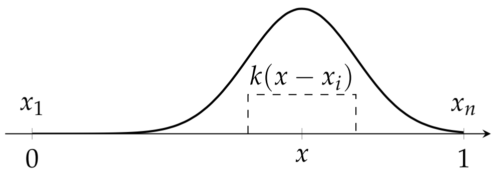


Figure 17: Kernel smoother window

It was published in Watson (1964) and in Nadaraja (1964). When $n \to \infty$, the kernel should become more concentrated. It is therefore natural to consider a scaled version of the kernel,

$$k_\lambda(w) = \frac{1}{\lambda} k\left(\frac{w}{\lambda}\right), \quad \lambda > 0.$$

As n increases, λ goes to zero.

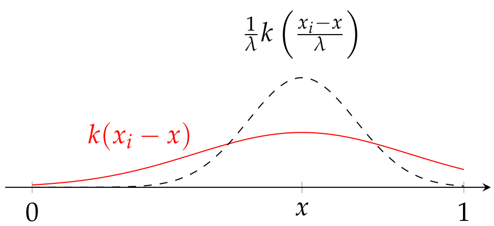


Figure 18: Scaled version of kernel smoother for $\lambda < 1$.

If n is large, $|x_2 - x_1|$, the gap between two consecutive x-values, is $1/(n-1) \approx 1/n$

and

$$\sum_{i=1}^{n} k(x_i - x)|x_2 - x_1| \approx \int_{\mathbb{R}} k(u)\,\mathrm{d}u,$$

where the right-hand side of the equation is the Riemann integral. The convention is to take $\int k(u)\,\mathrm{d}u = 1$, meaning $k(u)$ is a probability density. It is also useful to add the constraints

$$\int uk(u)\,\mathrm{d}u = 0$$
$$\int u^2 k(u)\,\mathrm{d}u < \infty.$$

The Nadaraja–Watson estimator for large n is

$$\widehat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} k\left(\frac{x_i - x}{\lambda}\right)\frac{1}{\lambda}Y_i$$

and the bias depends on $\int k^2(u)\,\mathrm{d}u$ and $\lambda^2 \int u^2 k(u)\,\mathrm{d}u$, while the variance is of order $\int k^2(u)\,\mathrm{d}u/(n\lambda)$. The choice of kernel has a small influence on the asymptotic properties. The kernel smoother is asymptotically unbiased if $\lambda(u) \to 0$ as $n \to \infty$. There is a huge literature on kernel smoothers: they are a useful tool and easy to analyze. The main difficulties are "border effects": how should one modify the kernel?

4.3.2. Local regression

This notion, proposed by Cleveland (1979), bears the acronym loess, which stands for local weighted scatterplot smoothing. The estimate $\widehat{f}(x)$ is the fitted local regression at x (estimated via a weighted least squares fit).

Example 4.4 (Constant and linear local regression)

1. constant regression: consider $\widehat{f}(x) = \theta$ over the domain $x \in [0,1]$ with weights $w_i = k((x_i - x)/\lambda)/\lambda$. We fit θ by $\widehat{\theta}$ which minimizes $\sum_{i=1}^{n} w_i(Y_i - \theta)^2$. This is equivalent to minimizing the least square criterion

$$-2\sum_{i=1}^{n} w_i(Y_i - \widehat{\theta}) = 0,$$

leading to the Nadaraja–Watson estimate

$$\widehat{\theta} = \widehat{f}(x) = \frac{\sum_{i=1}^{n} w_i Y_i}{\sum_{i=1}^{n} w_i}.$$

2. linear regression: consider the model

$$\widehat{f}(u) = \tilde{\theta}_0 + \theta_1 u = \theta_0 + \theta_1(u - x)$$

where $\tilde{\theta}_0 = \theta_0 - \theta_1 x$. Our weighting function $w_i^x = \frac{1}{\lambda} k((x_i - x)/\lambda)$ and we seek to minimize

$$\sum_{i=1}^{n} w_i \left[Y_i - \theta_0^x - \theta_1^x (x_i - x)\right]^2 .$$

The solution is

$$\begin{pmatrix} \widehat{\theta}_0^x \\ \widehat{\theta}_1^x \end{pmatrix} = \left(\mathbf{X}_x{}^\top \mathbf{W}_x \mathbf{X}_x\right)^{-1} \mathbf{X}_x{}^\top \mathbf{W}_x \boldsymbol{y}$$

where

$$\mathbf{W}_x = \operatorname{diag}\left(w_1^x, \dots, w_n^x\right)$$
$$\mathbf{X}_x = \begin{pmatrix} 1 & x_1 - x \\ \vdots & \vdots \\ 1 & x_n - x \end{pmatrix}$$

We thus have that the intercept is given by

$$\widehat{f}(x) = \widehat{\theta}_0^x = \sum_{i=1}^{n} Y_i h_i(x)$$

where $(h_1(x), \dots, h_n(x))$ is the first row of $(\mathbf{X}_x{}^\top \mathbf{W}_x \mathbf{X}_x)^{-1} \mathbf{X}_x{}^\top \mathbf{W}_x$.
The weights $(h_1(x), \dots, h_n(x))$ adapt to x and thus the linear loess estimator is **not** a kernel estimator. It behaves well at the borders.

Remark 4.1

The smoothing spline can be approximated by a kernel estimator.

4.4. Orthonormal basis

Consider $f(x) \in \mathcal{L}^2([0,1])$ and data (x_i, Y_i), where $Y_i = f(x_i) + \varepsilon_i$. The space $\mathcal{L}^2([0,1])$ has an (orthonormal) basis $\Phi_1(\lambda), \Phi_2(\lambda), \dots$ such that

$$\int_0^1 \Phi_i(x)\Phi_j(x)\,\mathrm{d}x = \delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j. \end{cases}$$

We can thus express $f(x)$ as

$$f(x) = \sum_{i=1}^{\infty} \theta_i \Phi_i(x).$$

If n is large,

$$\theta_j = \int_0^1 f(x)\Phi_j(x)\,\mathrm{d}x$$

leading to an estimator

$$\widehat{\theta}_j = \frac{1}{n}\sum_{i=1}^{n} Y_i \Phi_j(x_i).$$

Among the most famous examples of orthonormal basis are the following.

1. Haar wavelet (1907): consider

$$\Psi_{jk} = 2^{j/2}\Psi_0\left(2^j x - k\right)$$

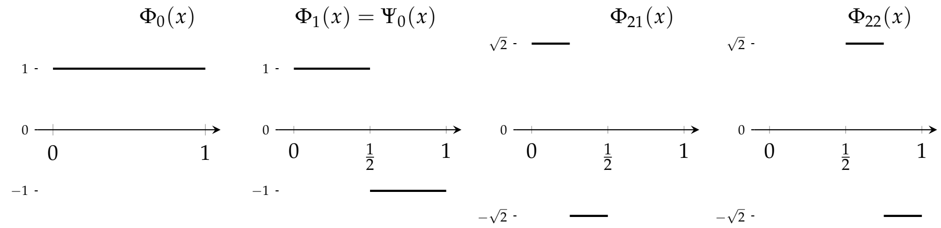


Figure 19: First four Haar wavelet functions

The function Φ_0 is the **father wavelet**, while $\Phi_1(x) = \Psi_0(x)$ is the **mother wavelet**.

2. Fourier (trigonometric basis)
On $[-\pi, \pi]$, take the functions $\Phi_1(x) = 1/\sqrt{2\pi}$, $\Phi_2(x) = \cos(x)/\sqrt{\pi}$, $\Phi_3(x) = \sin(x)/\sqrt{\pi}$, $\Phi_4(x) = \cos(2x)/\sqrt{\pi}, \ldots$ $\{\Phi\}$ is a periodic basis, but is not localized.

3. Orthogonal polynomials

$$\widehat{f}(x) = \sum_{i=1}^{n} \widehat{\theta}_i \Phi_i(x)$$

Define

$$Z_j = \widehat{\theta}_j = \frac{1}{n}\sum_{i=1}^{n} Y_i \Phi_j(x_i)$$

The random variables $\widehat{\theta}_j$ have interesting properties, as $\mathsf{E}\left(Z_j\right) = \theta_j + \mathrm{O}\left(\frac{1}{n}\right)$

$$\begin{aligned}
\mathsf{E}\left(Z_j\right) &= \frac{1}{n}\sum_{i=1}^{n} \mathsf{E}\left(Y_i\right)\Phi_j(x_i) \\
&\cong \int_0^1 f(x)\Phi_j(x)\,\mathrm{d}x \\
&\cong \theta_j.
\end{aligned}$$

In a similar fashion,

$$\mathsf{Cov}\left(Z_j, Z_k\right) \cong \mathsf{E}\left(Z_j Z_k\right) - \theta_j\theta_k$$

$$
\begin{aligned}
&\cong \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2\Phi_j(x_i)\Phi_k(x_i) + \frac{1}{n^2}\theta_j\theta_k\sum_{i=1}^{n}\Phi_j(x_i)\sum_{\substack{l=1\\ l\neq i}}^{n}\Phi_k(x_l) - \theta_j\theta_k \\
&\cong \frac{\sigma^2}{n}\int_0^1 \Phi_j(x)\Phi_k(x)\,\mathrm{d}x + \theta_j\theta_k\int_0^1\int_0^1 \Phi_j(x)\Phi_k(y)\,\mathrm{d}x\,\mathrm{d}y - \theta_j\theta_k \\
&= \frac{\sigma^2}{n}\int_0^1 \Phi_j(x)\Phi_k(x)\,\mathrm{d}x
\end{aligned}
$$

and thus $\mathsf{Var}\left(Z_j\right) \cong \sigma^2/n$ and $\mathsf{Cov}\left(Z_j, Z_k\right) \cong 0$ approximately for $k \neq j$ and n large.

4.5. Shrinkage

Assume now that $\mathbf{Z} = (Z_1, \ldots, Z_n)^\top \sim \mathcal{N}_n(\boldsymbol{\theta}, \sigma^2\mathbf{I}/n)$. We consider a paradox due to Charles Stein: suppose we consider arbitrary estimates $\widehat{\boldsymbol{\theta}}(\mathbf{Z}) = (\widehat{\theta}_1(\mathbf{Z}), \ldots, \widehat{\theta}_n(\mathbf{Z}))^\top$ and we wish to minimize the risk

$$
\mathcal{R}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \sum_{j=1}^{n} \mathsf{E}\Big(\big(\widehat{\theta}_j - \theta_j\big)^2\Big). \tag{4.17}
$$

In the context of the smoothing problem, this is the summed mean squared error. The "naive estimator" of $\boldsymbol{\theta}$ is $\widehat{\boldsymbol{\theta}} = \mathbf{Z}$; if $\widehat{\theta}_j(z) = z_j$, the MLE for a one-sample from the n-dimensional distribution, then the risk is $\mathcal{R}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = n \cdot (\sigma^2/n) = \sigma^2$. Stein (1964) showed that an estimator that shrinks $\mathbf{Z}$ towards zero,

$$
\widehat{\boldsymbol{\theta}}_{\mathrm{JS}} = \left(1 - \frac{(n-2)\widehat{\sigma}^2}{\|\mathbf{Z}\|^2}\right)\mathbf{Z},
$$

has lower risk than the sample observations, meaning

$$
\mathcal{R}(\widehat{\boldsymbol{\theta}}_{\mathrm{JS}}, \boldsymbol{\theta}) < \mathcal{R}(\mathbf{Z}, \boldsymbol{\theta})
$$

for dimension $n > 3$. This case was further generalized by Stein (1981).

Proposition 4.7 (Stein's unbiased risk estimate)
Assume that $\mathbf{Z} \sim \mathcal{N}_n(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ and that $\widehat{\boldsymbol{\theta}}(z)$ is an estimator of $\boldsymbol{\theta}$ such that $g(z) = \widehat{\boldsymbol{\theta}}(z) - z$ is differentiable where, as usual, $g(z) = (g_1(z), \ldots, g_n(z))^\top$. If we consider the risk function defined in eq. (4.17), the statistic

$$
\widehat{R}(z) = \mathrm{tr}(\boldsymbol{\Sigma}) + 2\,\mathrm{tr}(\boldsymbol{\Sigma}\mathbf{D}) + \sum_{i=1}^{n} g_i(z)^2
$$

with $\mathbf{D} = (\mathrm{D}_{ij})_{i,j=1}^{n}$ where $\mathrm{D}_{ij} = \partial g_i/\partial z_i$, is unbiased, that is $\mathsf{E}\big(\widehat{R}(\mathbf{Z})\big) = \mathcal{R}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta})$.

This is called Stein's unbiased estimate of risk. If one applies this to the smoothing problem, one first estimates $\widehat{\theta}_j$ naively and subsequently shrinks the small values (those satisfying $|\widehat{\theta}_j| \leq \lambda$) to zero.

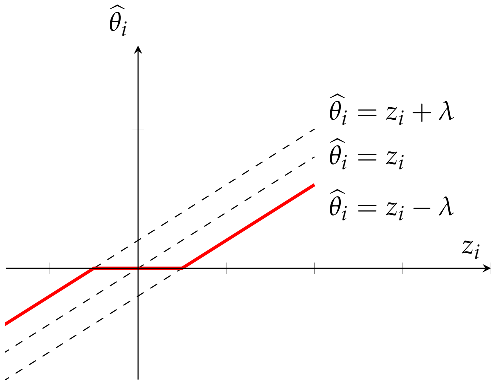


Figure 20: Shrinkage estimate

Proof For simplicity, consider the case $\mathbf{\Sigma} = \sigma^2 \mathbf{I}_n$. We first need a lemma.

Lemma 4.8
Let $u(x)$ be a differentiable function of $X \sim \mathcal{N}(\mu, \sigma^2)$. It follows that

$$\sigma^2 \mathsf{E}\left(u'(X)\right) = \mathsf{E}\left(u(X)(X-\mu)\right).$$

Proof of Lemma 4.8 This is true because

$$\sigma^2 \int u'(x) \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right) \mathrm{d}x = \sigma^2 u(x)\frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)\Big|_{-\infty}^{\infty} + \sigma^2 \int u(x)\frac{1}{\sigma}\left(\frac{x-\mu}{\sigma}\right)\phi\left(\frac{x-\mu}{\sigma}\right) \mathrm{d}x$$

using integration by part. The first term vanishes upon evaluating $\phi(x)$ at $\pm\infty$. ■

In our simplified setting, $\mathrm{tr}(\mathbf{\Sigma}) = n\sigma^2$ and

$$\mathrm{tr}(\mathbf{\Sigma D}) = \sigma^2 \sum_{i=1}^{n} \frac{\partial g_i}{\partial z_i}(z_i).$$

Applying the lemma to

$$\sigma^2 \mathsf{E}\left(\frac{\partial g_i}{\partial Z_i}\right) = \mathsf{E}\left(g_i(Z_i - \theta_i)\right)$$

$$\mathsf{E}\left(\widehat{R}(\mathbf{Z})\right) = \sum_{i=1}^{n} \mathsf{E}\left((Z_i-\theta_i)^2\right) + 2\sum_{i=1}^{n} \mathsf{E}\left((\widehat{\theta}_i - Z_i)(Z_i-\theta_i)\right) + \sum_{i=1}^{n} \mathsf{E}\left((\widehat{\theta}_i-\theta_i)^2\right)$$

We leave as exercise to calculate $\widehat{R}(z)$. One finds that, for a well chosen λ,

$$\widehat{R}(z) = \sum_{i=1}^{n} \left(\sigma^2 - Z\sigma^2 \mathbb{I}_{|Z_i|<\lambda} + \min\{Z_i^2, \lambda^2\} \right) < n\sigma^2.$$

$\blacksquare$

References

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.*, 74(368), 829–836.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69, 383–393.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust statistics. The approach based on influence functions*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York.

Hettmansperger, T. P. (1984). *Statistical inference based on ranks*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66), Vol. I: Statistics* (pp. 221–233). Univ. California Press, Berkeley, Calif.

Huber, P. J. (1996). *Robust statistical procedures*, volume 68 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, second edition.

Huber, P. J. & Ronchetti, E. M. (2009). *Robust statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, second edition.

Kullback, S. & Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Statist.*, 22(1), 79–86.

Mann, H. B. & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Statistics*, 18, 50–60.

Nadaraja, È. A. (1964). On a regression estimate. *Teor. Verojatnost. i Primenen.*, 9, 157–159.

Randles, R. H. & Wolfe, D. A. (1979). *Introduction to the theory of nonparametric statistics*. John Wiley & Sons, New York-Chichester-Brisbane. Wiley Series in Probability and Mathematical Statistics.

Stein, C. (1964). Inadmissibility of the usual estimator for the variance of a normal distribution with unknown mean. *Ann. Inst. Statist. Math.*, 16, 155–160.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, 9(6), 1135–1151.

Strassen, V. (1965). The existence of probability measures with given marginals. *Ann. Math. Statist.*, 36, 423–439.

Sun, J. & Loader, C. R. (1994). Simultaneous confidence bands for linear regression and smoothing. *Ann. Statist.*, 22(3), 1328–1345.

Tukey, J. W. (1960). A survey of sampling from contaminated distributions. In *Contributions to probability and statistics* (pp. 448–485). Stanford Univ. Press, Stanford, Calif.

Wasserman, L. (2006). *All of nonparametric statistics*. Springer Texts in Statistics. Springer, New York.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā Ser. A*, 26, 359–372.

Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6), 80–83.

Index

asymptotic relative efficiency, 9, 63

B-robust estimate, 24, 25
Bahadur slope, 65
bandwidth, 69
basis
  Fourier, 85
  Haar wavelets, 85
breakdown point, 30

cross-validation, 79, 80

derivative
  Fréchet differential, 21
  Gâteaux differential, 21
Dirac-δ distribution, 19

estimating equation, 11

Fisher consistency, 13
Fisher information, 7
  expected, 7
  observed, 7
Fisher's scoring algorithm, 7

gross-error sensitivity, 24

Hampel's
  3 point estimator, 24
  lemma, 28
Heavyside function, 19
Hölder continuity, 68
Huber's
  minimax solution, 31
  regression proposal, 41

influence function, 19
interquartile range, 30

kernel smoother, 82
Kullback–Leibler divergence, 6

L-estimator, 18
least absolute deviation, 41
least median of squares (LMS), 41
linear rank statistic, 48
local power, 51
local-shift sensitivity, 24
locally MP test, 50
loess, 83

M-estimator, 11
  asymptotic normality, 14
  asymptotic variance, 17
  Fisher consistent version, 28
  influence function, 19
Mallow's C_p, 80
Mann–Whitney test, 49
  mean and variance, 54
maximum likelihood estimator, 5
  asymptotic normality, 7
mean square prediction error, 79
median absolute deviation (MAD), 30
minimax theory, 31

Nadaraja–Watson estimator, 82
Newton-Raphson algorithm, 7

orthonormal basis, 84

Pitman alternative, 61
Pitman efficacy, 61, 62
Prohorov metric, 9
ψ function, 11
  Huber function, 20
  mean, 20
  median, 20

R-estimator, 18
rank test, 46
  asymptotic distribution, 54
  linear rank test, 50

scores, 48
redescending estimator, 24, 42
regression constant, 56
regression diagnostics, 37
Cook's distance, 40
LOO residuals, 40, 80
rejection point, 24
ϱ function, 11
robustness, 9

S-estimator, 42
score generating function, 53
sensitivity, 40
sensitivity curve, 10
shrinkage, 86
sign test, 46
skipped mean, 24
spline, 70
natural, 72
optimality, 72
smoothing, 72
space, 70
Stein's estimate, 86
Stein's paradox, 86

Tukey's bisquare, 24, 42

U-statistics, 50

Wilcoxon's
signed rank test, 47
two-sample test, 48