CHAPTER 13

ROBUST MEAN

One from the most influential method of science is averaging. Result of averaging provides a convenient smooth representation of studied quantities and also naturally suppresses potentially unrelated effects.

An arithmetic mean is commonly used method of averaging characterised by simple formulas and widely known statistical properties. The arithmetical mean can be introduced by definition (as in Section 13.1) or derived from the principle of maximum likelihood (Section 13.2). The subsequent matter of this chapter, develops methods for determination of the robust mean, the averaging method "insensitive to small deviation from assumption" as has been introduced in Chapter 3, from the maximum likelihood.

This introduced chapter is considered as the detailed description of methods suitable for computation of robust mean including its deep explanation. If reader is not interested in details, Section 13.11 gives short summary of reliable algorithm for estimation of robust mean.

See Section 13.11 for the algorithm.

There are two important books giving background for this chapter. The general introduction to statistic in data processing – the book Brandt (2014) including developing of the maximum likelihood method. Robust parts of this chapter are evolved on base of ideas of the book Huber (1981).

Although, the averaging, realised by the robust mean estimation, looks trivially on the first sight, it demonstrates, in plain basic form, all important ideas of robust algorithms. This chapter opens gate of a robust land. The abstract land where butterflies can fly without fear that theirs wings will be broke by storm drops raised by oneself.

13.1 ARITHMETIC MEAN

The arithmetic mean is a standard way to estimate of average of a data-sample. An general practise, to compute of the arithmetic mean, of a set of N values

$$\{x_1, x_2, \dots, x_N\}, \tag{13.1}$$

is, by definition, the formula

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i. \tag{13.2}$$

The mean $\bar{x}$ gives estimation of location of a centre of the input set. The scatter of points x_i around the mean $\bar{x}$ is represented by the standard (quadratic) deviation

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2. \tag{13.3}$$

The mean $\bar{x}$ itself is localised more precisely than s indicates. To estimate statistical uncertainty σ of $\bar{x}$, we will suppose that all points x_i has approximately the same $\sigma_i \approx s$ deviation and the points are statistically independent. By using of the assumptions, we can use the model for the error propagation (Brandt (2014))

$$\sigma^2 = \sum_{i=1}^{N}\left(\frac{\partial f}{\partial x_i}\right)^2 \sigma_i^2 \tag{13.4}$$

on a function of arithmetical mean defined as $f(\{x_i\}) = (x_1 + x_2 + \cdots + x_N)/N$ which is for every point $\partial f/\partial x_i = 1/N$. Putting all the terms together, and summation of a constant term, gives

$$\sigma^2 = \sum_{i=1}^{N}\frac{1}{N^2}\,s^2 = \frac{s^2}{N}, \tag{13.5}$$

or $\sigma = s/\sqrt{N}$.

Result of arithmetic meaning are usually presented in the form of the confidence interval

$$\bar{x} \pm \sigma \tag{13.6}$$

which optimistically estimates that the true value X of a quantity lies inside interval $\bar{x} - \sigma \leq X \leq \bar{x} + \sigma$ with probability of 68%.

Arithmetic mean is commonly used method. One is easy to use, numerically stable and gives smooth results. Arithmetic mean is ideal for use in computing by hand. The formula (13.2) is very simple, data can be easy inspected and potentially false data suppressed. On the contrary, the arithmetic mean can not be recommended for machine processing. One is sensitive on deviated data. In case of its presence in a sample, results will be scattered or, much worse, completely random.

13.2 ARITHMETIC MEAN BY MAXIMUM LIKELIHOOD

The method of maximum likelihood presents impressive point of view on the full field of arithmetic mean. The method brings also a very effective framework which provides optimal estimates of the average for data with an arbitrary statistical distribution. So important for generalisation.

This section briefly summarises important steps used to estimate arithmetic mean and its statistical properties by maximum likelihood method. The description is summary of equivalent chapter of Brandt (2014) which provides more detailed and exact description.

A basic principle of maximum likelihood method is maximisation of probability which describes measured quantities. In case of mean, we are looking for probability, common to all points. Supposing of statistical independence of single point with the probability density function $\phi(x_i|\bar{x}, s)$, the join probability of of fitness of all data points can be composed as its products

$$\phi(x_1|\bar{x}, s) \cdot \phi(x_2|\bar{x}, s) \cdots \phi(x_N|\bar{x}, s). \tag{13.7}$$

For use of the method on a data, one *a priory* assumes a statistical distribution of every data point. In case of arithmetical mean, we will suppose Normal distribution $\mathcal{N}(\bar{x}, s)$ with the probability density function

$$\phi(x|\bar{x}, s) = \frac{1}{\sqrt{2\pi}s} \mathrm{e}^{-(x-\bar{x})^2/2s^2}. \tag{13.8}$$

The analytic form of the probability is given by our assumption. The function represents a "distance" (measure) that every single point with x and how much the point belong to the distribution set.

Lets define the likelihood function

$$L(\bar{x}, s) = \prod_{i=1}^{N} \phi(x_i|\bar{x}, s) \tag{13.9}$$

as function of parameters (independent variables) $\bar{x}, s$. The parameters can be estimated by the way: The goal of our effort is set parameters $\bar{x}, s$ such way to maximise probability which is a priory known.

The methods for searching of maximum of such products are indirect. We use property of logarithm which is converts products on sums and one is a monotone function so the maximum is has unchanged point. Logarithm of maximum likelihood is

$$\ln L = \sum_{i=1}^{N} \ln \phi(x_i|\bar{x}, s) \tag{13.10}$$

which depends on the distribution function. With substituting of $\phi(.)$ the function and summing of constant elements gives

$$\ln L = -\sum_{i=1}^{N} \frac{(x_i - \bar{x})^2}{2s^2} - N \ln s - \frac{N}{2} \ln 2\pi. \tag{13.11}$$

The second term is result of sum of constant function over all data. The extreme of the function is in a point where derivations of the function by both parameters are vanishing

$$\frac{\partial \ln L}{\partial \bar{x}} = \sum_{i=1}^{N} \frac{x_i - \bar{x}}{s^2} = 0, \tag{13.12a}$$

$$\frac{\partial \ln L}{\partial s} = \sum_{i=1}^{N} \frac{(x_i - \bar{x})^2}{s^3} - \frac{N}{s} = 0. \tag{13.12b}$$

The solution illustrates how the arithmetical mean can be derived from use of general principle of maximum likelihood and Normal distribution. The solution of (13.12a) against to $\bar{x}$ is

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{13.13}$$

and one is identical to (**??**). The solution of (13.12b) for s[1] gives

$$s^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2. \tag{13.14}$$

The solution is no more equal to (13.3) because we get estimate with maximum probability, but estimation of the mean value of s is biased. A possible convenience (Bessel's correction) way to estimate s for the right centre is replace it as (Brandt (2014))

$$\frac{N}{N-1} s^2 \to s^2 \tag{13.15}$$

which reduces (13.12b) to (13.3) whilst the difference is negligible for larger data sets.

The scatter of the parameters is commonly estimated from the hessian in extreme. Lets suppose that a function f can be in a point $\mathbf{r}$ as

$$\mathbf{r} = \begin{pmatrix} x, \\ s \end{pmatrix} \tag{13.16}$$

expanded to Taylor series

$$f(\mathbf{r} + \Delta\mathbf{r}) \approx f(\mathbf{r}) + \mathbf{J}(\mathbf{r})\Delta\mathbf{r} + \frac{1}{2}\Delta\mathbf{r}^T \hat{H}(\mathbf{r})\Delta\mathbf{r} + \dots \tag{13.17}$$

[1] Interpretation of s is a parameter of distribution $\phi(.)$. The parameter is numerically equivalent to s defined by (13.3) (the same symbol is used for different quantities with the same meaning).

where we introduce Jacobian matrix J as the gradient

$$\mathbf{J} = \nabla f = \begin{pmatrix} \dfrac{\partial f}{\partial \bar{x}} \\ \dfrac{\partial f}{\partial s} \end{pmatrix}. \tag{13.18}$$

Elements of J are given by formulas (13.12). Hessian $\hat{H}$ is the matrix of second derivatives (in maximum for us)

$$\hat{H} = \begin{pmatrix} \dfrac{\partial^2 f}{\partial \bar{x}^2} & \dfrac{\partial^2 f}{\partial \bar{x} \partial \sigma} \\ \dfrac{\partial^2 f}{\partial \sigma \partial \bar{x}} & \dfrac{\partial^2 f}{\partial \sigma^2} \end{pmatrix} \tag{13.19}$$

which gives for us in a general point

$$-\frac{1}{s^2} \begin{pmatrix} N & (2/s)\sum_i (x_i - \bar{x}) \\ (2/s)\sum_i (x_i - \bar{x}) & N - (3/s^2)\sum_i (x_i - \bar{x})^2 \end{pmatrix} \tag{13.20}$$

and in correctly determined extreme where $\sum_i (x_i - \bar{x}) \to 0$, $\sum_i (x_i - \bar{x})^2 \to -2N/s^2$ (see (13.12)), we get

$$\hat{H} = -\frac{N}{s^2} \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}. \tag{13.21}$$

The covariance matrix is inverse of hessian in minimum and gives in this case with (nearly) diagonal matrix

$$\hat{H}^{-1} = -\frac{s^2}{N} \begin{pmatrix} 1 & 0 \\ 0 & 1/2 \end{pmatrix}. \tag{13.22}$$

Diagonal matrix means orthogonal choice of parameters, which means, that changes in the coordinates are independent.

The statistical error of $\bar{x}$ can be estimated as

$$\sigma^2 = |h_{11}^{-1}|. \tag{13.23}$$

or $\sigma = s/\sqrt{N}$ which reproduces result (13.5).

While estimation of statistical error of arithmetical mean is highly appreciated, the error of s is leaved unnoticed. While it can be easy estimated on $s/\sqrt{N/2}$.

In a minimum of probability, the function $\ln L$ can be approximated as

$$\ln L(x, s) = \ln L(\bar{x}, s) + \frac{1}{2} \Delta \mathbf{r}^T \hat{H}(\mathbf{r}) \Delta \mathbf{r} + \dots \tag{13.24}$$

which gives

$$L(x|\bar{x},\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-(x-\bar{x})^2/2\sigma^2}. \tag{13.25}$$

And the join distribution function can be approximated again as Normal distribution with more narrow width σ.

The fact can be illustrated on test data in Section 13.13. Graph 13.3 shows lilelihood function of this section as "Normal". The generated points has distribution $\mathcal{N}(1, 0.1)$ so individual functions are wider than graph itself. But the products of all the funcstions has width about 0.026. The shape of the product "Normal" and $\mathcal{N}(1, 0.026)$ is approximatelly the same, as was expected. This is one from demonstration of law of large numbers.

The parameter $\bar{x}$ has meaning of centre of distribution of measured values. The values is usually the goal of our measures and usually one is a description of a real system which we are interested in. Opposite with this, the s only seldom is related to the measured system and one is more property of a measuring device as its precision. Seldom, there is requirements for using of measurements with devices with different precision. In the case, see Section 13.8.

13.3 ROBUST MEAN BY MAXIMUM LIKELIHOOD

The maximim likelihood framework sumarised in previsous section can be used also on robust estimation of mean as the central moment of a general, as well as robust, distribution. We can see the application in robust case with robust distribution but one can be used in general case.

The robust mean can be described by the general distribution f and parameters in the fashion

$$\frac{1}{\Gamma}\frac{1}{s} f\left(\frac{x_i - \tilde{x}}{s}\right). \tag{13.26}$$

The function is designed for estimation of central moment $\tilde{x}$ and the scale s together with analogy to Normal distribution. Note the s which scales all the function. The factor is necessory to be able for its estimate. The transformation of x_i is importnat and because distribution functions are centerd on origin and unit mean scatter. The normalisatrion factor Γ (like $\sqrt{2\pi}$ for Normal distribution) is defined by property with $s =$

$$\frac{1}{\Gamma}\int_{-\infty}^{\infty} f(x - \tilde{x})\,\mathrm{d}x = 1.$$

The maximum likelihood principle is for the case is as one expected

$$L = \prod_{i=1}^{N} \frac{1}{\Gamma}\frac{1}{s} f\left(\frac{x_i - \tilde{x}}{s}\right) \tag{13.27}$$

and its logarithm version

$$\ln L = \sum_{i=1}^{N} \ln f\left(\frac{x_i - \tilde{x}}{s}\right) - N \ln s - N \ln \Gamma. \tag{13.28}$$

The derivations with use of the substitution (note choice of sign)

$$\psi = -(\ln f)' = -\frac{f'}{f} \tag{13.29}$$

where ψ is a suitable (robust) function. To simplify notation, we will use the substitution normalised residuals as

$$r_i = \frac{x_i - \tilde{x}}{s}. \tag{13.30}$$

The solution leads to the system of equations (analogy of (13.12))

$$\frac{\partial \ln L}{\partial \tilde{x}} = \frac{1}{s} \sum_{i=1}^{N} \psi(r_i) = 0, \tag{13.31a}$$

$$\frac{\partial \ln L}{\partial s} = \frac{1}{s} \sum_{i=1}^{N} \psi(r_i) \cdot r_i - \frac{N}{s} = 0. \tag{13.31b}$$

The Hessian is a symetric matrix and has these elements

$$\frac{\partial^2 \ln L}{\partial \tilde{x}^2} = -\frac{1}{s^2} \sum_{i=1}^{N} \psi'(r_i), \tag{13.32a}$$

$$\frac{\partial^2 \ln L}{\partial \tilde{x}\, \partial s} = -\frac{1}{s^2} \sum_{i=1}^{N} \left[\psi(r_i) + \psi'(r_i) \cdot r_i\right], \tag{13.32b}$$

$$\frac{\partial^2 \ln L}{\partial s^2} = -\frac{1}{s^2} \sum_{i=1}^{N} \left[2\psi(r_i) \cdot r_i + \psi'(r_i) \cdot r_i\right] + N. \tag{13.32c}$$

The Hessian near of minimum will be (using of (13.31))

$$\hat{H} = -\frac{1}{s^2} \begin{pmatrix} \sum \psi'(r_i) & \sum \psi'(r_i) r_i \\ \sum \psi'(r_i) r_i & N - \sum \psi'(r_i) r_i^2 \end{pmatrix}. \tag{13.33}$$

The off-diagonal terms will vanish because ψ' is supposed to be a constant term and r_i will distributed to give a minimal sum.

Statistical errors of the parameters are estimated by using of the robust version of covarience matrix (Huber (1981))

$$\sigma^2 = s^2 \frac{N}{N-1} \frac{\sum \psi^2(r_i)}{[\sum \psi'(r_i)]^2} \tag{13.34}$$

where we identify the bias correction of scale.[2]

The equations looks unfamiliar. Fortunately, ones can be easy understaned with the special choice which assympoticaly uses the least squares $f(x) \propto \exp(-x^2/2)$ and by (13.29) we get

$$\psi(x) = x \tag{13.38}$$

which reduces the general functions on the aready known set of equations of arithmerical mean case. Really, lets look on the relations

$$\psi'(x) = 1, \tag{13.39a}$$
$$\psi'(r_i) \to 1, \tag{13.39b}$$
$$\psi(r_i) \to r_i. \tag{13.39c}$$

which reduces the system (13.32) to (13.20).

The meaning of individual terms in (13.34) can be easy understand with analogy with Normal distributiion case. The

$$\sigma^2 \asymp s^2 \sum \psi(r_i)^2 \tag{13.40}$$

is practically the robust analogy of residual sum S_0. Also

$$N \asymp \sum \psi'(r_i) \tag{13.41}$$

is estimation of number of acceptable data.

13.4 SOLUTION OF THE NON-LINEAR SYSTEM

There are complication in computation of system of equations for robust mean (13.31) against to the same system for arithmetic mean (13.12). While the unknown values can be easy separated in case of (13.12), the system (13.31) is solvable only numerical way.

[2] Huber (1981) suggest multiply the term under square by K-correction. One is for p parameters

$$K = 1 + \frac{p}{N}\frac{\mathrm{var}(\psi')}{(E\psi')^2} \tag{13.35}$$

with

$$E\psi' \approx m = \frac{1}{n}\sum_i \psi'(r_i), \tag{13.36}$$

and

$$\mathrm{var}(\psi') \approx \frac{1}{n}\sum_i [\psi'(r_i) - m]^2. \tag{13.37}$$

The corrections has are due to $1/n$ dependency neglibible except very noisly data and small datasetds.

The basic method with fixes s is described in Subsection 13.4. Join estimates of $\tilde{x}$ and s are in 13.4, but the method can not be reccoemned as text of this secrion descriobes. The reccomended method is use of standard minimalisation provedures.

The estimation of robust mean in the unkind numerical environment of modern machines is extremely delicate job. The estimation of arithmetical mean is also little bit delicated, but practically is limited by precision of represetnation of float numbers.

The robust mean is more compilcated, but due to scaling of values, more numericaly stable. The posssible comlications arises from general non-lineariry of common robust functions. The solution is more time consuming and drastically depends on initial estimates.

There is many of methods for solution of non-linear systems of equations like (13.31). The methods needs are

Lets we know (good) initial estimates of solution $\tilde{x}^{(0)}, s^{(0)}$, Hubber Huber (1981) in section 6.7 "The computation of M-Estimates", reccomends four variants of estimates. I tested extensive (many years of testing, possible 10^{12} or more computations has been performed) these only two principial variants:

MODIFIED RESIDUALS

The Newtons method for solution of our problem

$$f(\tilde{x}) = \sum_{i=1}^{N} \psi\left(\frac{x_i - \tilde{x}}{s}\right) = 0. \tag{13.42}$$

The derivation is

$$f'(\tilde{x}) = \frac{1}{s}\sum_{i=1}^{N} \psi'\left(\frac{x_i - \tilde{x}}{s}\right) \neq 0. \tag{13.43}$$

and we supppose that is non-zero. The general form of the Newton's method is

$$\tilde{x}^{(i+1)} = \tilde{x}^{(i)} + \frac{f(r_i)}{f'(r_i)}. \tag{13.44}$$

The initial estimate of scale $s^{(0)}$ is fixed and Newton's method is used to estimate of better approximation of robust mean

$$\tilde{x}^{(i+1)} = \tilde{x}^{(i)} + s^{(0)}\frac{\sum \psi(r_i)}{\sum \psi'(r_i)}. \tag{13.45}$$

The method is very reliable. One converges very quickly with averadge data. The number of iterates is up to 10 for single precision (10^{-7}) and up to 15 for double precision (10^{-16}). Therefore I am limiting it to seventeen.

The precision is tested agains to machine precision ε and also to relative precision of expresion

$$\delta = \left| \frac{\sum \psi(r_i)}{\sum \psi'(r_i)} \right| \tag{13.46}$$

as the conditions

$$\delta / |\tilde{x}^{(i+1)}| < \varepsilon \wedge \delta < \varepsilon. \tag{13.47}$$

It may be important check the doneminator $\sum \psi'(r_i)$ because a bad initial estimate of $s^{(0)}$ or a bad data will lead to nullify of it. With the normaly distributed data, the term will practically constant with meaning of amont of data. The property is also noticed by Huber (1981).

JOINT M-ESTIMATES OF LOCATION AND SCALE

The previous method is rely on proper initial estimation of scale. The assumption is only partially true and the estimation can be imprecise on level of order of ten percent. Because the estimation has fluence on estimation of mean, the difference can leads to differnce in mean on level of a few percent which is inadequte for precisse reculs. Therefore, the similtenoust estimation is equired. Huber (1981) reccomends[3] replace of equations (13.31) by the

$$\sum_{i=1}^{N} \psi \left(\frac{x_i - \tilde{x}}{s} \right) = 0, \tag{13.48a}$$

$$\sum_{i=1}^{N} \psi^2 \left(\frac{x_i - \tilde{x}}{s} \right) = N - 1, \tag{13.48b}$$

where the we replaced N by $(N-1)$ and we are supposse that the expectation value of the second moment of the distribution $E\psi^2$ (deviation) is exacltly one.

The system is reccomended to by solved as

$$s^{(i+1)} = \sqrt{\frac{[s^{(i)}]^2}{N-1} \sum_i \psi^2 \left(\frac{x_i - \tilde{x}^{(i)}}{s^{(i)}} \right)} \tag{13.49a}$$

$$\tilde{x}^{(i+1)} = \tilde{x}^{(i)} + s^{(i)} \frac{\sum \psi(r_i)}{\sum \psi'(r_i)}. \tag{13.49b}$$

The first equation is the iteration method while the second is Newton's method. The separating solution on the parts is only possible when the initial estimation is very good and both the parameters are near-orthogonal

[3] ψ^2 is insensitive to outliers, if winsorisation is applied on data, the original (13.31) will be so good as well as.

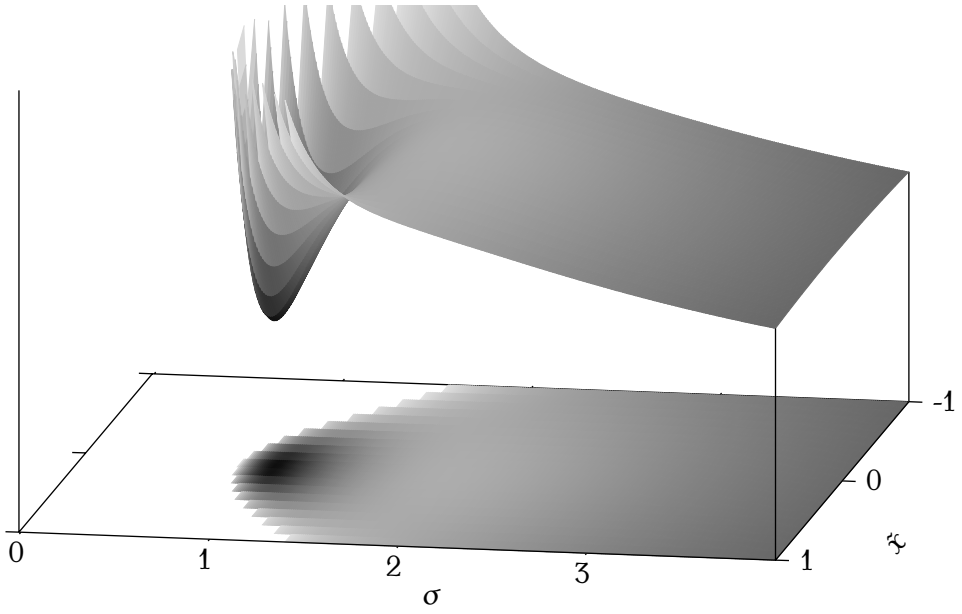

Figure 13.1: Minimum of gradient function

each to other. The orthogonality means independence of mean on scatter. It is true only partially in real situations (see Section 13.13). The use of general method for computation of the set of non-linear equation provided by Minpack as `lmder` (or `lmdif`) is reccomended.

There is a few important thinks. All methods belives in the satisfy of condition

$$s > 0 \tag{13.50}$$

which is very important. Both the hessian and gradinet are symeptric under $s \to -s$ and depends on absolute value only.

13.5 REGION OF CONVERGENCE OF GRADIENT

Function (13.28) and its equivalents has one global minimum because this is sum of functions with one global minimum. There are unique location of the minimum.

The fast gradient methods – Levendberg - Marquart method – does not uses the function. The methods locates minimimum of the gradient $\varphi = \nabla f$ as

$$\sum_i \varphi_i \cdot \varphi_i. \tag{13.51}$$

For functions which we are restricting, the minimum is in the same point. But the function itself can be different far from the minimum. For robust functions, the The heel shoes like shape see Figure 13.1.

Figure 13.1 is derived from 666 data points with 95% of $N(0,1)$ and 5% of $N(0,5)$. The grah shows value of function f^2 for various values of

$\tilde{x}, \sigma$. The function is a surface with lightness of gray protportional to the function value. The heel has minimum at $0, 1$.

There are an important implication. The initial estimate of local minimum must be relative near the right mininimum. The local turnabout is approximate about 2 in Figure 13.1.

As the general reccomendation, the optimalisation strategy has two ways:

a) We can directly minimizes (13.28). The method can not use gradient (except near neigberhood of minimum). Simplex method (Nealder-Mead) can be used. The convergence is slow. Estimation of statistical errors from Hessian is complicated.

b) The minimuzation of gradient (13.31) which reqiures reliable initial estimation (ideally by a non-gradient method). The convergence is fast and precise. The estimation of Hessian in minimum is a side effect of the minimization.

13.6 INITIAL ESTIMATES

The initial estimation of $\tilde{x}^{(0)}, s^{(0)}$ is crutial for success of everything. Huber (1981) reccomends, on base study of statistical properties of varisou distributions, the median (med) and mean absolute deviation MAD

$$\mu = \text{median}\{x_1, x_2, \ldots, x_N\} \tag{13.52}$$

the MAD is related to standard devaition as MAD $= \Phi^{-1}(3/4) \approx 0.6745$ of inversion cumulative function to Normal distribution and so

$$s^{(0)} = \frac{\text{median}\,|x_i - \mu|}{0.6745}. \tag{13.53}$$

My expriences with this recommended estimators are exclenent for large datasets of normally distributed data with insignificant number of outliers (up to 1/4 of full data).

For a few points (up to ten), the estimatios are unrealible and the sometimes random. For a few data, it is better use of quantiles which are more robust and less sensitive to non-uniform (normal) distribnution of data. To prevent the problems, the quantiles of the empirocal distributiuons are used. The prepared the following an algorithm is used. The algorithm can replace also estimate by median, unfortunatelly it is significantly slower.

Lets we define the empricial distribution function from input data as

$$F_n = \frac{1}{N} \sum_{i=1}^{n} 1\{x_i < n/N\} \qquad \text{for } n = 1, N \tag{13.54}$$

what symetrizes covering of an interval of points $1/N$ to

$$(1/N)/2, (2/N)/2 - (1/N)/2, \ldots, 1 - (1/N)/2 \tag{13.55}$$

and the parameters can be estimated as quantiles $q_{1/2} = 1/2$ which is lineary interpolated in the quantile function where for n such $F_i < q_{1/2} \leq F_{i+1}$

$$\mu = \frac{x_{n+1} - x_n}{F_{n+1} - F_n}(q_{1/2} - F_n) + x_n \tag{13.56}$$

and analogicaly for $s^{(0)}$ the $|x - \mu|$ is used.

The use of the cuimuulative distribution function is fine version of median. The median uses sorted values, as CDF, but the center is roughly estimated as middle point. There an interpolation on $q = 0.5$ is used.

13.7 STUDENTIZING

13.8 QUESTION OF WEIGHTS

The presented approach can be generalized on data with different σ_i of each observation x_i. The situation can be encountered when the obseravtions has been acquired by devices with different internal precision or under differnent conditions. The data with significantly different σ_i are for data with Normal distribution meet very rarely.[4]

The key change of the use of already known dispersion σ_i is transformation

$$\frac{x_i - \bar{x}}{\sigma_i}$$

of all data to normal-like distribution $N(0, 1)$. The preassumption is very optimistics. Therefore, it is better to suppose that values σ_i are estimated only roughly and the transformation on the normal distribution can be assumed as

$$\frac{x_i - \bar{x}}{s\sigma_i} \tag{13.57}$$

where s is an effective scale of observed scatter. When estimates of σ_i are correct, one can expect $s \approx 1$. This is important espetially for non-least information distributions.

The generalisation of equation (13.9) is strightforward

$$L = \prod_{i=1}^{N} f(x_i|\bar{x}, s\sigma_i) \tag{13.58}$$

[4]Metaphoricaly, one assignes a weight w_i for every point. The choice of individual weights depends on "experience". From statistical point of view, the weights will $w_i = 1/\sigma_i^2$. Unfortunatelly, the weights has been choosed randomly or, more worstly, to remove inconvenient points. This kind of manipulation can not be reccomended by any way. Primarily, the robust methods offers better way.

and for Normal distribution (non-robust) leads to set of equations

$$\frac{\partial \ln L}{\partial \bar{x}} = \frac{1}{s^2} \sum_{i=1}^{N} \frac{x_i - \bar{x}}{\sigma_i^2} = 0, \tag{13.59a}$$

$$\frac{\partial \ln L}{\partial s} = \frac{1}{s^2} \sum_{i=1}^{N} \frac{(x_i - \bar{x})^2}{s\sigma_i^2} - N = 0. \tag{13.59b}$$

which has the solution

$$\bar{x} = \frac{\sum_{i=1}^{N} \frac{x_i}{\sigma_i^2}}{\sum_{i=1}^{N} \frac{1}{\sigma_i^2}} \tag{13.60}$$

and

$$s^2 = \frac{1}{N} \sum_{i=1}^{N} \frac{(x_i - \bar{x})^2}{\sigma_i^2}. \tag{13.61}$$

The solution for s^2 is χ^2 distribution divided by total count of data.

Another important aspect is change of meaning of σ. It no more means scatter of data but it is a scale of data scater.

Robust approach is similar. We also starts from (13.9) where f is an robust function. Logarith of likelikood fucntion is

$$\ln L = \sum_{i=1}^{N} \ln f\left(\frac{x_i - \tilde{x}}{s\sigma_i}\right) - \sum_{i=1}^{N} \ln \sigma_i - N \ln s. \tag{13.62}$$

with solution given by a set of non-linear equations (subtitution (13.29))

$$\frac{\partial \ln L}{\partial \tilde{x}} = \sum_{i=1}^{N} \psi\left(\frac{x_i - \tilde{x}}{s\sigma_i}\right) \frac{1}{s\sigma_i} = 0, \tag{13.63a}$$

$$\frac{\partial \ln L}{\partial s} = \sum_{i=1}^{N} \psi\left(\frac{x_i - \tilde{x}}{s\sigma_i}\right) \frac{x_i - \tilde{x}}{s^2\sigma_i} - \frac{N}{s} = 0. \tag{13.63b}$$

which must be performed numericaly. Note that the equation (13.63b) for s should be rewrited to equivalent form

$$\sum_{i=1}^{N} \psi^2\left(\frac{x_i - \tilde{x}}{s\sigma_i}\right) = (N-1)s. \tag{13.64}$$

The simultaneous estimation of both $\tilde{x}$ and s is highly reccomended because estimates of σ_i are rarely correct which can significantly degrade robust estimates.

Just for information, robust mean can be rewriten in the form of weight mean as ....

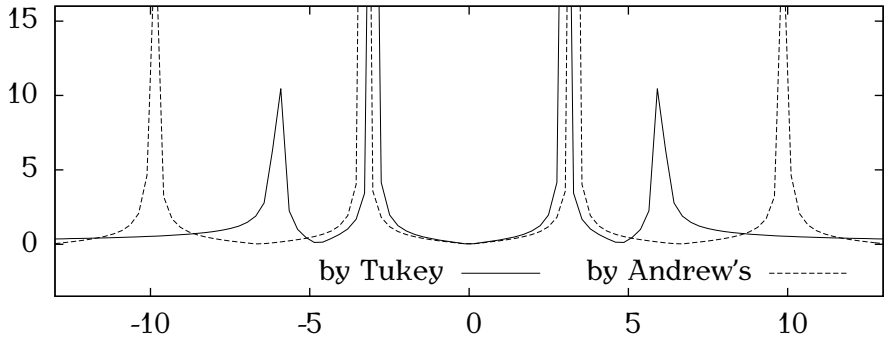

Figure 13.2: Convergence intervals for descending functions

13.9 QUESTION OF DESCENDING ESTIMATOR

There are two kinds of robust estimators – monotone and descending – from shape of ψ function. Huber's function is monotone and all others (Hampel's, Tukey's, Andrew's) are descending.

The impotance of the the shape of the function can be show on the Newton's method. The convergence criterii of Newton's method is known (Ralston and Rabinowitz (2012))

$$\epsilon^{(i+1)} = \frac{|f''(x^{(i)})|}{2|f'(x^{(i)})|}(\epsilon^{(i+1)})^2, \tag{13.65}$$

where the condition must be satisfied

$$|f''(x)| < 2|f'(x)|. \tag{13.66}$$

The region of convergence can be easy discovered from(13.65) and see Figure 13.2.

Dale ukazat ze treba pro tukeye nebo andrewse to neni splneno na intevralu $-a \ldots a$. Pekne grafy to ukazou.

13.10 QUESTION OF OUTLIERS

Robust estimators excellently works with "small changes from presasuptions", but the large deviations like data from another statistical sample, errorneous data commonly known as outliers, can destroy robust estimate.

There are two alternatives which minimizes fluence of outliers: a choice of robust function which vanishes in infinity and removing or replacing of outliers.

Robust functions which vanishes in infinity are Hammpel (11.3), Andrews (11.4) or Tukey (11.5). The condition of vanishing requires non-monotone function. The condition implicates that the Newton's method

(13.45) may diverge. The convergence of the method requires that ratio $\sum \psi / \sum \psi' < 1$ is small. Unfortunatelly, in certain range of parameter the condition is not meet which leads to divergence. Therefore the use of the functions cannot be reccomended for Newtwon method. The use of a gradient free method can be satisfactory when estimation of uncertainities is not required. Huber (1981) has his "A Word of Caution" in section 4.8 Descending M-estimates.

WINSORIZATION

As more reliable way for handling with outliers is a technique which replaces outliers. The replacement by the formula (for definition of sign function see (11.2))

$$x_i^* = \begin{cases} x_i, & |x_i - \mu| \le cs, \\ \mu + cs \ \operatorname{sign} x_i, & |x_i - \mu| > cs, \end{cases} \tag{13.67}$$

is known as data winsorization (winsorizing – we are Winsorize the data. Procedue is named aster inventor Charles P. Winsor (1895 – 1951)). The estimation of the parameter must be executed by a robust method – by median μ as I give a hint. The parameter setting limit c should by set to an appropriate value $1 < c < 2$. See Huber (1981), Section 1.7. My experiences with using on data with outliers shows that better interval is $1 < c < a$ (the upper constant is given by Hubber $a = 1.349$), say $0.9a$.

An alterative for winsorization is well known "clipping" in which the outliers are removed from a sample. Usually on base of similar criterii as in winsorisation. The result will generally similar except for short samples, when clipping can remove significant amount of data.

13.11 AN ALGORITHM

There is a summary of the development of this section in the form of an algorithm for computation of robust mean. The algorithm has been heavy tested as the part of Munipack code. One can be considered as a prototype of a robust algorithm.

Prerequisite. Lets $\{x_1, x_2, \ldots, x_N\}$ is a set of N single non-identical numbers from $\mathbb{R}$. The data should be represented in computer by an array of real (floating point) numbers.

Robust Mean Algorithm.

i) The initial estimation of central moment μ is given by median (13.52)

$$\mu = \text{median}\{x_1, x_2, \ldots, x_N\}. \tag{13.68}$$

ii) The initial estimation standard deviation s. Lets absolute deviations are

$$d_i = |x_i - \mu|, \quad \text{for } i = 1, \ldots, N \tag{13.69}$$

and median of the absolute deviations (mad) is

$$d_{\text{mad}} = \text{median}\{d_1, d_2, \ldots, d_N\}. \tag{13.70}$$

The estimation of standard deviation will be finally by (13.53)

$$s = \frac{d_{\text{mad}}}{0.6745} \tag{13.71}$$

It is strongly recommended to check the condition $s > \epsilon$ (ϵ is non-zero positive constant – larger than machine precision).

iii) Winsorisation according to (13.67) with substitution $\chi = 1.2\, s$

$$x_i^* = \begin{cases} x_i, & |x_i - \mu| \le \chi, \\ \mu + \chi \operatorname{sign} x_i, & |x_i - \mu| > \chi, \end{cases} \tag{13.72}$$

iv) Location of minimum of robust function. By defining of residuals

$$r_i = \frac{x_i^* - \mu}{s}, \quad \text{for } i = 1, \ldots, N \tag{13.73}$$

we use function

$$\ln L(x_i|\tilde{x}, s) = \sum_{i=1}^{N} \varrho(r_i) + N \ln s \tag{13.74}$$

where the integral of robust function $\varrho(x)$ is given by (**??**).

This steps locates of extreme $\ln L$ with certain precision. Recommended method for minimisation is the simplex method (Nelder and Mead (1965)) or any method using no derivations.

v) Robust estimation by minimising of set of equations against to parameters $\tilde{x}, s$ (via r_i)

$$\sum_{i=1}^{N} \psi(r_i) = 0, \tag{13.75a}$$

$$\sum_{i=1}^{N} \psi(r_i) \cdot r_i = N. \tag{13.75b}$$

Recommended method for minimising is Levendberg-Marquart (Marquardt (1963)) with analytic Jacobian given by (13.32). The method is regularised (insensitive for errors), fast and provides the most precise solution.

vi) The uncertainties of the robust mean σ can be estimated in the minimum as (13.34) using of results of previous step

$$\sigma^2 = s^2 \frac{N}{N-1} \frac{\sum_{i=1}^{N} \psi^2(r_i)}{[\sum_{i=1}^{N} \psi'(r_i)]^2} \tag{13.76}$$

Result. The result of the algorithm is estimation of robust mean with uncertainty

$$\tilde{x} \pm \sigma \tag{13.77}$$

and the standard deviation s.

Recommendations. A reliable implementation should check than $N > 0, s > 0$ (s during all interactions).

13.12 AN SIMPLIFIED ALGORITHM

The general algorithm in Section 13.11 simultaneously minimizes both the paramteres $\tilde{x}, s$. This simplified version estimates the scale parameter (standard deviation) s by median od absolute deviation. This reduces space of parameters in one dimension. Newton's method of root finding can be used and it importantly speed-up iteraction due to quadratic convergence. There is a small loss of precision (up to $10 - 20$ %).

The simplified algorithm has assumptions and notation the same as a general algorithm of Section 13.11. Initial steps i) – iii) (winsorisation) are the same and the alternative way starts the steps iv) and v) are replaced by the single step

v) The next step are by (13.45) where $\tilde{x}^{(0)} = \mu$, $r_i^{(k)} = (x_i^* - \tilde{x}^{(k)})/s$:

$$\tilde{x}^{(k+1)} = \tilde{x}^{(k)} + s \frac{\sum_{i=1}^{N} \psi(r_i^{(k)})}{\sum_{i=1}^{N} \psi'(r_i^{(k)})} \quad \text{for } k = 0, \ldots \tag{13.78}$$

The iterations can stop when $|\tilde{x}^{(k+1)} - \tilde{x}^{(k)}| < \epsilon$ where ϵ is a required precision (machine precision).

Estimation of uncertainties is again by step vi).

An reliable implementation should check $N > 0, s > 0$ when initialisation is finished. The interaction can converge only when the second correction element is $|s \sum \psi(r_i)| < |\sum \psi'(r_i)| < 1$ and also $\sum \psi'(r_i) \neq 0$. When no convergence occurs, the k should by limited on an appropriate amount of interactions (42).

13.13 AN EXAMPLE

There are an illustration of the methods on numerical example. The example can be also used for testing purposes. All results are rounded on 4 digits although ones has been computed on at least 16 digits.

As a test set a sequence of 15 elements has been generated[5] both from Normal distrobution with the same dispersion but centre at point 1 (good) and 0 (bad):

$$\{x_i \in \mathcal{N}(1, 0.1), i = 1, 13\} + \{x_i \in \mathcal{N}(0, 0.1), i = 14, 15\} \tag{13.80}$$

with the result

$$\begin{aligned}\{&0.719, 0.983, 0.818, 0.933, 1.034, 1.005, 1.145, 1.255,\\ &1.039, 1.041, 1.078, 1.111, 0.872, 0.288, 0.137\}.\end{aligned} \tag{13.81}$$

Amount of data is small and it is instructive only. The data of bad distribution represents 13% of all points.

ARITHMETIC MEAN

Results of deriving of arithmetic mean as has been introduced in Section 13.2 are

$$\bar{x} = 0.8972, \tag{13.82a}$$
$$s = 0.3088, \tag{13.82b}$$
$$\sigma = 0.0797. \tag{13.82c}$$

At minimum, the hessian is

$$\hat{H} = \begin{pmatrix} -157.2703, & 0.0000, \\ 0.0000, & -314.5407 \end{pmatrix}. \tag{13.83}$$

[5] A generator of random numbers from a standard library of Fortran compiler has been used. The elements are selection from Uniform distribution $u \in \mathcal{U}(0, 1)$ (in interval $0 \le u < 1$) and u represents probability. Normal distribution has been established from inverse to cumulative distribution fuinction of $x \in \mathcal{N}(\mu, \sigma)$ as $x = \mu - \sqrt{2}\sigma\,\mathrm{erf}^{-1}(2u - 1)$ where the inverse erf function has an approximation with its precision better than $3.5 \cdot 10^{-4}$:

$$\mathrm{erf}^{-1}(x) \approx \mathrm{sign}(x)\sqrt{\sqrt{\left(\frac{2}{\pi a} + \frac{\ln(1 - x^2)}{2}\right)^2 - \frac{\ln(1 - x^2)}{a}} - \left(\frac{2}{\pi a} + \frac{\ln(1 - x^2)}{2}\right)}, \tag{13.79}$$

where

$$a = \frac{8(\pi - 3)}{3\pi(4 - \pi)}.$$

This approximation has been published only at Wikipedia (2016).

The results demonstrates strong bias toward the bad data. Results can not be accepted.

Figure 13.3 shows likelihood function of Normal distribution for the data. Maximum of the fuinction is visible shifted. The confidence interval does not includes expected center location.

ROBUST MEAN

Application of single steps of algorithm from Section 13.11:

i) Initial estimation of robust mean by median is

$$\mu = 0.9940, \tag{13.84}$$

ii) and it scatter

$$s = 0.1490. \tag{13.85}$$

iii) Limits for winsorisation is ± 1.21 so values are in range $0.872 \ldots 1.116$. The original set is transformed to (changed values are denoted)

$$\begin{aligned}\{&0.872^*, 0.983, 0.818, 0.933, 1.034, 1.005, 1.145, 1.116^*,\\ &1.039, 1.041, 1.078, 1.111, 0.872, 0.872^*, 0.872^*\}.\end{aligned} \tag{13.86}$$

Total number of changed values is 4 (27%). Because limits was under 1.21 of $\mathcal{N}(0,1)$, the expected number of changes was 11% ($1-2$ elements) and we have only two outliers, which is exactly what we expected.

iv) Robust estimation by minimizing of integral of Hubber function with Simplex algorithm gives

$$\tilde{x} = 0.9753, \tag{13.87}$$
$$s = 0.1205. \tag{13.88}$$

v) The same result (due rounding) gives Marquart-Levenberg minimisation of Hubber's function

$$\tilde{x} = 0.9753, \tag{13.89}$$
$$s = 0.1386, \tag{13.90}$$
$$\sigma = 0.0547 \tag{13.91}$$

Hessian at minimum is

$$\hat{H} = \begin{pmatrix} -894.9690, & 185.7405, \\ 185.7405, & -1780.6926 \end{pmatrix}. \tag{13.92}$$

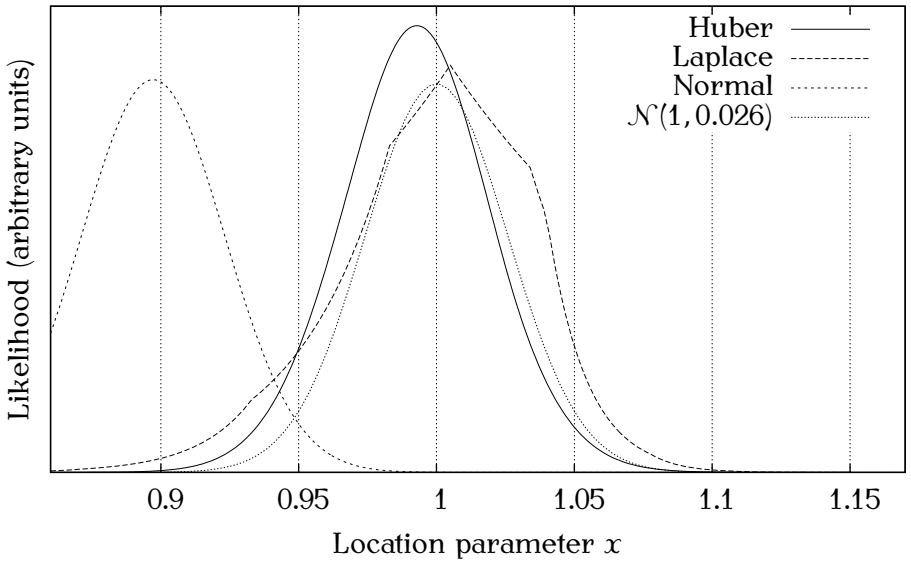

Figure 13.3: Likelihood function for various distributions

Another point of view shows graph of cummulative distribution function on Fig. 13.4. The maximum difference between theoretical and empiriocal is at point 0.288 and is 0.13 and it is abowe critical value of Colmogorov-Smirnoff test (xxx) which confirms hypothesis that the point violates the Normal distribution.

The teoretical distribution function on Fig. 13.4 is cummulative of $\mathcal{N}(0.9753, 0.14)$, eg. Normal with parameters by robust estimation. The graph confims that the fit is appropriate, the original data set has biased mean due to limited number of points (confidence looks better asympoticaly). The digram is relative steping due to small amount of data.

Fig. 13.3 shows likelihood functions for both initial estimation and robust function. The initial estimation as "Laplace" shows piecewise profile. In maximum, the point is equivalent to median. Robust likelihood "Huber" is shifted from expected value which must be considered as a random coincidence by generated data. The appropriate amount of data confirms this hypothesis.

Hessian shows weak depencnce of both parameters. It reveales effects of winsorising.

SIMPLIFIED ROBUST MEAN

Result of siplified algorithm by Section 13.12 are

$$\tilde{x} = 0.9688, \tag{13.93}$$

$$\sigma = 0.0499, \tag{13.94}$$

$$s = 0.1654. \tag{13.95}$$

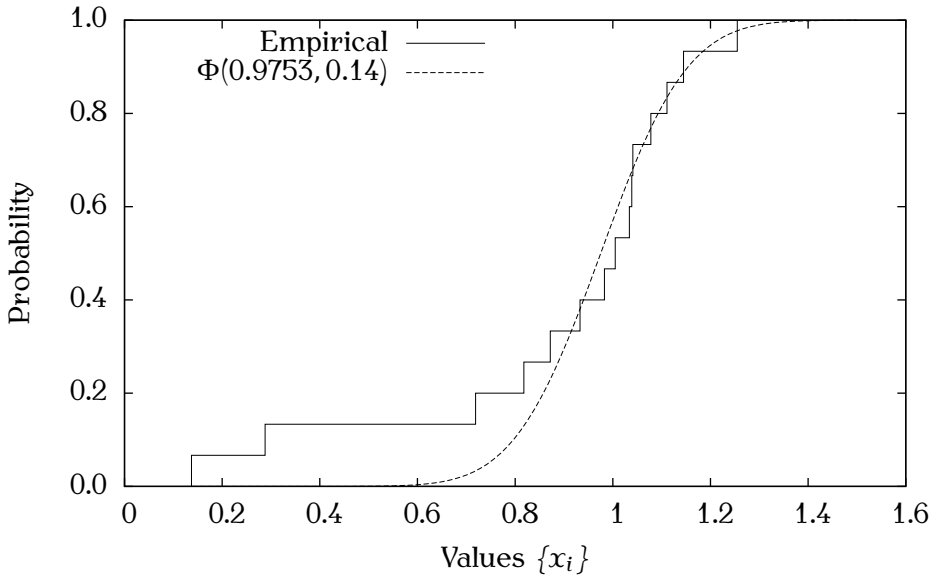

Figure 13.4: Distribution functions of the example data

While the estimation is a little bit worse. On the other side, the algorithm was significantly simpler and faster. The proper rounding will give for both the algorithms the same value 0.97 ± 0.05 so there is no important difference.