Robust Statistics

KJM

November 17, 2024

Contents

1 Motivation

In statistics and mathematical modeling in general, in addition to the sample that we observe, any inference or conclusion one may reach is facilitated by a host of assumptions. These may include the distributional form from which we sample, independence, randomness, and so on. This is a necessary step, not designed to produce a model absolutely resembling reality, but rather to produce something that is reasonable with only small error, as well as mathematically convenient. Any conclusions we draw, therefore, should consequently be reasonable and small in error.

However, countless examples reveal that common statistical procedures are remarkably sensitive to small departures from some set of common assumptions. Most notably, the sample mean can be highly variable under slight departures from normality. In reality, distributions can be skewed, they can have heavy tails, and random samples often have outliers, leading to a plethora of practical problems even if these deviations are minor in scale. For example, the standard error of the sample mean is inflated substantially by heavy tails, resulting in the probability coverage of standard methods for computing confidence intervals differing greatly from the nominal value, and the usual sample variance can give a distorted view of the amount of dispersion. Thus, these small deviations from normality give a false sense of the amount of spread of the data, and the central or "average" value around which most the population are distributed.

Consequently, it is desirable to find estimators that do not vary so greatly due to small deviations from our assumptions, since this can lead to a false sense of the distribution we are working with, as well as any inferences we make from it. That is, we want to use techniques that are robust. Robustness of a given statistical procedure signifies insensitivity to small deviations from the assumptions upon which the procedure was constructed.

The scenario below is a classic example, due to Tukey, demonstrating how a small deviation from normality can greatly affect the well-known sample standard deviation. It turns out that the mean absolute deviation, to be defined below, is a more robust measure of scale (an indication of how spread out the values in a distribution are), and how it can greatly outperform the standard deviation in the face of outliers.

Imagine some proportion of observations, $1 - \epsilon$, is generated by a normal model and the remaining proportion by some unknown mechanism, in our case another normal with greater variance. For example, we are taking repeated measurements of something: 95% of the time, we take measurements correctly but the remaining 5% there is some error like an apparatus failure. Suppose the distribution for the correctly measured data has distribution $G = \mathcal{N}(\mu, \sigma^2)$ and erroneous data $H = \mathcal{N}(\mu, 9\sigma^2)$. Our data will thus follow the distribution F, where:

$$F = (1 - \epsilon)G + \epsilon H$$

So, in our experiment the data we collect follows a distribution F, but in our model we assume that we are sampling from a distribution resembling normality. Now consider two measures of scale, mean absolution deviation (MAD), d_n, and root mean square deviation (RMSD), s_n:

$$d_n = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|, \qquad s_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

The asymptotic relative efficiency (ARE) is a statistical measure used to compare the efficiency of two estimators when drawing from a particular distribution as the sample size n grows large. Here, the ARE is given when sampling from the mixed normal, dependent on ϵ:

$$\text{ARE}(\epsilon) = \lim_{n\to\infty} \frac{Var(s_n)/(\mathbb{E}(s_n)^2)}{Var(d_n)/(\mathbb{E}(d_n)^2)} = \frac{\frac{1}{4}[\frac{3(1+80\epsilon)}{(1+8\epsilon)^2}] - 1}{\frac{\pi(1+8\epsilon)}{2(1+2\epsilon)^2} - 1}$$

Intuitively, we take the ratio of the variances of the two estimators as n grows large. But since the MAD and RMSD are estimates of different measures of scale, their raw variances alone wouldn't provide a fair comparison. By using the squared expectation in the denominator for each, the variance is scaled relative to the central tendency of each estimator. In the case above, if $\text{ARE}(\epsilon) > 1$ then we conclude the RMSD is less efficient than MAD.

Assuming absolute normality ($\epsilon = 0$), we get an ARE of roughly 0.88. That is, RMSD is 12% more efficient than MAD. Different values of ϵ yield the following:

ϵ	ARE(ϵ)
0	0.876
0.001	0.948
0.01	1.439
0.05	2.035
0.10	1.903
0.5	1.017
1.0	0.876

Table 1: Table of ϵ and its corresponding ARE(ϵ)

Therefore, if ϵ is only equal to 0.01 (there is a 1% chance of taking erroneous measurements), the 12% advantage of the RMSD has vanished and been replaced by a 44% advantage for MAD. Namely, an experiment with a very small error rate disproportionately increases the variance in RMSD compared to MAD, thus rendering it suspect as an ideal measure of scale.

This egregious inflation of the variance of the RMSD indeed has dire consequences upon inferences one may make from the data when assuming normality. Set μ, σ, and ϵ to 0, 1, and 0.1 respectively. The probability distribution functions of G and F are plotted below. The standard normal has variance 1 but the mixed normal has a variance of (...). As a result, if sampling is from the mixed normal, the length of the standard confidence interval for the population mean, μ, will be over (...) times longer than it would be when sampling from the standard normal distribution instead.

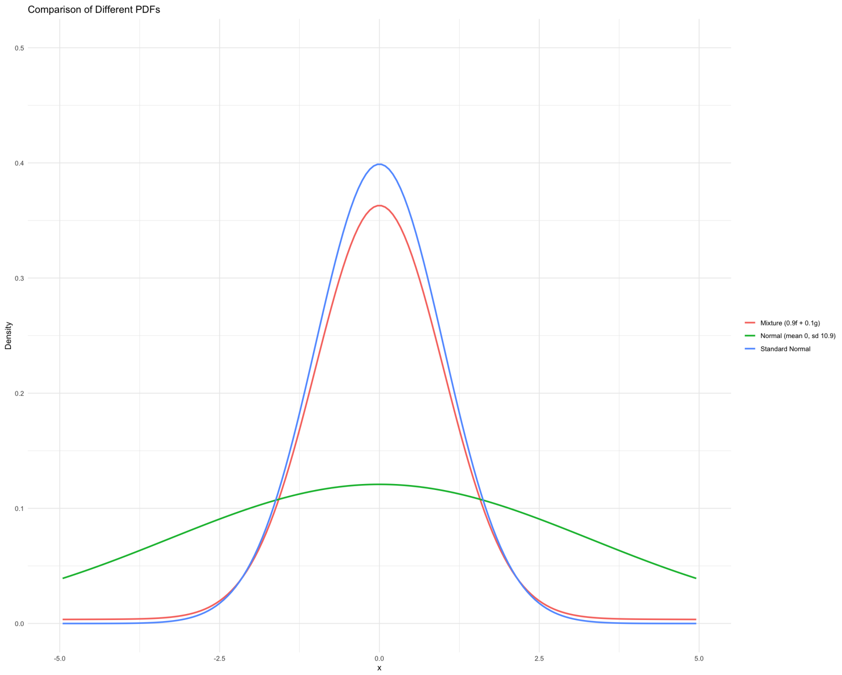

Figure 1: Tukey's Mixed Normal. The diagram below is a plot of the probability distribution function (pdf) of X, in blue, and the mixed normal, in red.

From the image, it seems that the normal distribution provides a good approximation of the contaminated normal. Only the tails of the mixed normal are slightly above the tails of the normal. But this small proportion of the population has had an inordinately large effect on the variance of the distribution and the RMSD, thus underscoring the need to find alternative measures of a distribution that do not result in wildly different outcomes as a result of a slight shift from our assumptions.

One might question, however, the necessity of robust methods. If we were in the scenario above, could we not apply some rule for outlier rejection and then proceed to use standard techniques? Only one reason which is in the negative is due to masking. This is when the presence of an outlier prevents an outlier rejection test from identifying another outlier. For instance, the presence of an outlier can inflate the standard error of the sample mean, increasing the width of a given confidence interval that could envelop other outliers.

Robust methods, therefore, are a collection of new and improved methods that provide a more accurate and more nuanced understanding of data. under general conditions they can have substantially higher power compared to more traditional techniques. Robust methods should strike a balance between the following three traits:

i **Efficiency**: Given the assumed model we are sampling from, the statistical procedure should have low variance

ii **Stability**: Small deviations from the model assumptions should only impair model performance to a small extent

iii **Breakdown**: Large outliers should not cause a drastic change in the performance of the model.

2 Fundamentals

A measure of distribution is a statistical value or function that helps summarize and describe the characteristics of a probability distribution. Measures of location and scale are two examples of such measures that characterize a given distribution. In what follows, we shall largely be discussing measures of location, defined below.

Let X be a random variable with distribution F, and $\theta(X)$ some measure of F. $\theta(X)$ is said to be a measure of location if it satisfies the following conditions. For any given constants a and b:

i $\theta(X+b)=\theta(X)+b$ (location equivariance)

ii $\theta(-X)=-\theta(X)$

iii $X\geq 0 \Rightarrow \theta(X)\geq 0$

iv $\theta(aX)=a\theta(X)$ (scale equivariance)

Also referred to as a measure of central tendency, these provide a central or "average" value around which the data points are distributed. Examples include the sample mean. This can be shown using standard properties of expectation.

2.1 Measuring Robustness

A measure is called robust if slight changes in the distribution from which it is sampled results in a relatively small effect on its value. As demonstrated earlier, the sample mean and standard deviation are not robust. There are three different tools that are utilized to asses whether a given measure has the desired robustness properties.

2.1.1 Qualitative Robustness

Often, our measure of location can be viewed as functionals of their empirical distributions or distribution functions. For example, the sample and population means from a continuous distribution can be found with respect to the functional T:

$$T(F)=\mathbb{E}_f(X)=\int_{-\infty}^{\infty} xf(x)dx, \qquad T(\hat{F})=\mathbb{E}_{\hat{f}}(X)=\sum_{i=1}^{n} x_i\hat{f}(x_i)$$

Say we are sampling from some true underlying distribution H, which we have modeled as G. Say that G is "close" to H - in the sense of some metric. Then, it should be the case that our measure of location when sampling from G, $T(G)$ is also "close" to that same measure had we sampled from the true distribution, $T(H)$.

Another intuitive idea behind robustness is that if we have some sample $(x_1,\dots,x_n)$ and another sample $(y_1,\dots,y_n)$, where the difference between them is either small differences between observation i in each, that is $x_i\approx y_i$, or large changes between x_i and y_i for only some i due to gross errors or blunders but otherwise $x_i=y_i$. Then we should have that the empirical distributions of each are "close" and therefore so are the estimates we obtain.

More formally, a measure, T, is said to be qualitatively robust at H if T is continuous at H according to the metric d. That is,

$$\forall\,\epsilon>0\;\exists\,\delta>0\;:\;d(H,G)<\delta\;\Rightarrow\;|T(H)-T(G)|<\epsilon$$

2.1.2 Infinitesimal Robustness

Given our functional T, we want to be able to measure the sensitivity and amount of change of an estimator due to outliers. More specifically, we want to measure the rate of change of the estimator T when a small amount of contamination is added. This is done using the influence function (IF), defined below:

$$IC(x,F,T)=\lim_{\epsilon\to 0^+}\frac{T((1-\epsilon)F+\epsilon\delta_x)-T(F)}{\epsilon}$$

Intuitively, we are measuring the rate of change of T when an infinitesimal amount of contamination has been added at point x. As the value of the IF varies with x, we say that F is infinitesimally robust if the IF is bounded.

2.1.3 Quantitative Robustness

Let Θ be the set of values that a parameter θ can take. For our measure of location or scale, this is the interval $[-\infty, \infty]$ or $[0, \infty]$ respectively. Additionally, consider $F_{x,\epsilon} = (1-\epsilon)F + \epsilon\delta_x$.

The breakdown point of the estimator T at F, denoted by $\epsilon^*(T, F)$ is the largest $\epsilon^* \in (0, 1)$ such that:

$$\forall\epsilon < \epsilon^* \ \forall G \ \exists K \subset \Theta \ : \ T((1-\epsilon)F + \epsilon G) \in K$$

Where K is closed and bounded. The breakdown point is the smallest fraction of bad observations that may cause an estimator to take on arbitrarily large values.

2.2 M-Estimates of Location

2.2.1 Generalities

An M-estimate, $\hat{\theta}$, of some population parameter, θ, is one that is found by minimizing a chosen loss function $\rho(x;\theta)$ applied to our observed data. Namely, given some sample $(x_1, \dots, x_n)$ and loss function $\rho(x;\theta)$ then the M-estimator is defined:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} \rho(x_i;\theta)$$

or defined by the implicit equation:

$$\sum_{i=1}^{n} \psi(x_i;\hat{\theta}) = 0,$$

Where $\partial\rho(x;\theta)/\partial\theta = \psi(x;\theta)$. In our case, we are mainly concerned with measures of location. When searching for a measure of location, we are looking for some value that is in some sense close to all of our data points. We aim to find the estimate $\hat{\theta}$ that minimizes the total error across all data points by summing up the individual errors, measured by the function ρ. So in this setup, we want to find $\hat{\theta}$ such that:

$$\sum_{i=1}^{n} \rho(x_i - \hat{\theta}) = \min!, \qquad \sum_{i=1}^{n} \psi(x_i - \hat{\theta}) = 0.$$

It can be shown that if ψ is monotone nondecreasing, and there exists $r \in \mathbb{R}_{>0}$ such that $\psi(-r) < 0 < \psi(r)$, then there exists a solution to the above optimization problem. If ψ is strictly monotone increasing then this solution is unique. Note then that all symmetric unimodal distributions have this property.

To evaluate the performance of M-estimators, it is necessary to calculate their distributions. However, for finite samples there is no close-form or analytical formula that describes the exact probability distribution of the M-estimator, and so must be approximated. It can be shown that M-estimators are asymptotically normally distributed:

$$\hat{\theta} \rightsquigarrow \mathcal{N}(\theta_0, \frac{\nu}{n})$$

Where:

- θ_0 is the asymptotic value of $\hat{\theta}$, where:

$$\mathbb{E}[\psi(X - \theta_0)] = 0$$

- ν is the asymptotic variance of $\hat{\theta}$, given by:

$$\nu = \frac{\mathbb{E}[\psi(X)^2]}{(\mathbb{E}[\psi'(X)])^2}$$

It should be noted that maximum likelihood estimation (MLE) is a special case of M-estimation. Consider the log-likelihood below.

$$\mathcal{L}(\theta; x_1, \dots, x_n) = \sum_{i=1}^{n} \log\ f_0(x_i - \theta)$$

The maximum likelihood estimate ($\hat{\theta}_{\text{MLE}}$) is then given by the value of θ that maximises the log-likelihood. If we set $\rho = -\log\ f_0$, then this is equivalent to:

$$\hat{\theta}_{\text{MLE}} = \operatorname*{argmin}_{\theta} \sum_{i=1}^{n} \rho(x_i - \theta)$$

Which is the same form of an M-estimator. It can, in fact, be shown that $\hat{\theta}_{\text{MLE}}$ is the "optimal" estimate, in that it attains the lowest possible asymptotic variance among a "reasonable" class of estimators, given F. But maximum likelihood estimation is not in general robust, and in reality we are not truly sampling from distribution F but rather something that is "close" to it.

We can further see that the sample mean and median are special cases of M-estimation. Consider the following case where $\rho(x) = x^2$ and thus $\psi(x) = 2x$, with observed sample x_i for $i = 1, 2, \dots, n$. We want to find $\hat{\theta}$ minimising (...):

$$\sum_{i=1}^{n} \psi(x_i - \hat{\theta}) = \sum_{i=1}^{n} (x_i - \hat{\theta}) = 0$$

$$\Rightarrow \hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{\mathbf{x}}$$

Thus we obtain the sample mean. Now consider setting $\rho(x) = |x|$ and $\psi(x) = \text{sgn}(x)$. We then find $\hat{\theta}$ by the same procedure:

$$\sum_{i=1}^{n} \psi(x_i - \hat{\theta}) = \sum_{i=1}^{n} \text{sgn}(x_i - \hat{\theta}) = 0$$

In this situation, we want to pick $\hat{\theta}$ such that the number of x_i where $x_i - \hat{\theta}$ is negative is equal to the number of x_i so that $x_i - \hat{\theta}$ is positive. So $\hat{\theta}$ is greater than half the values and less than the other half, namely a sample median. Thus it has been shown that the sample mean and median are indeed M-estimates.

An intuitive view

In most cases that will be discussed, it will be the case that $\psi(0) = 0$ and $\psi'(0)$ exists. Define

$$W(x) = \begin{cases} \frac{\psi(x)}{x} & \text{if } x \neq 0, \\ \psi'(0) & \text{if } x = 0. \end{cases}$$

So we can re-write:

$$\sum_{i=1}^{n} \psi(x_i - \hat{\theta}) = \sum_{i=1}^{n} W(x_i - \hat{\theta})(x_i - \hat{\theta}) = 0.$$

$$\Rightarrow \hat{\theta} = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i}, \quad \text{where } w_i = W(x_i - \hat{\theta})$$

So we can re-write our M-estimate as a weighted average of our observations. These weights depend on $\hat{\theta}$ itself, but gives intuition on how we calculate the M-estimate.

In the cases we consider, ρ will be of an order less than quadratic as $|x|$ grows large, and therefore ψ will be less than linear this region. Therefore, $W(x)$ is decreasing for large $|x|$. So any observations far from our estimate will have very small weight.

2.2.2 Picking the ρ-function

We want an estimator that is “nearly optimal” when sampling from exactly F and also from distributions “close” to F. The function ρ, or its derivative, ψ, can be chosen in such a way to provide the estimator with such desirable properties.

Huber functions

Bisquare functions

3 Linear Regression

Before delving into the intricacies of robust regression, it is essential to first address the principles underpinning standard linear regression. With this foundation, we can then examine how robust regression adapts the approach of linear regression to accommodate potential deviations in our assumptions, such as the presence of outliers. In what follows, the key assumptions of linear regression will be introduced, as well as the method of least squares for estimating model parameters, and the approaches used for statistical inference and prediction, which will serve as a baseline for developing robust methods.

3.1 Core Principles

Assume that p unknown parameters $\beta_1, \dots, \beta_p$ are to be estimated from n observations $y_1, \dots, y_n$ to which they are linearly related by:

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i$$

Which can be represented by matrix notation:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

The notation is defined as follows:

- $\mathbf{Y^T} = (y_1, \dots, y_n)$ is the vector of responses, each y_i called the response variable,
- $\mathbf{X}$ is the design matrix, each x_{ij} called a predictor variable,
- $\boldsymbol{\beta^T} = (\beta_1, \dots, \beta_p)$ is the p-dimensional parameter vector, each β_i fixed and unknown coefficients, called parameters,
- $\boldsymbol{\epsilon^T} = (\epsilon_1, \dots, \epsilon_n$ is the vector of errors, each ϵ_i called an error variable.

We also make the following assumptions:

i **Linearity**: $\mathbb{E}(\epsilon_i) = 0$

ii **Homoscedasticity**: $\text{Var}(\epsilon_i) = \sigma^2$

iii **Independence**: $\text{Cov}(\epsilon_i, \epsilon_j) = 0, \forall i \neq j$

iv **Normality**: $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$

Consequently, $\mathbf{y}$ is an n-dimensional normally distributed random vector. Specifically, $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, where $\mathbf{I}$ is the n-dimensional identity matrix.

Commonly, one estimates the parameters of the linear regression model using the method of least squares. In our framework, it can be demonstrated that least-squares estimation is equivalent to maximum likelihood estimation. Looking ahead to the next chapter, where we will explore robust regression through the framework of M-estimation, it is instructive to re-frame standard linear regression through the lens of maximum likelihood, thus providing a natural bridge to understanding how robust methods extend and generalize these ideas to handle violations of key assumptions.

We obtain our estimates by finding $\beta_1, \dots, \beta_p$ that minimizes the set of equations:

$$\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 = \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^T\boldsymbol{\beta} \right)^2$$

Equivalently, we can find our estimate $\hat{\boldsymbol{\beta}}^T = (\hat{\beta}_1, \dots, \hat{\beta}_p)$ by differentiating with respect to β and setting to zero:

$$\sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^T\boldsymbol{\beta} \right)\mathbf{x}_i = 0$$

Which is equivalent to the linear equations:

$$\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y}$$

This set of linear equations is called the “normal equations”.

3.2 Inference and Prediction

3.3 Diagnostics

3.4 Limitations

discussion of assumptions that could be violated

4 Robust Regression

- How it adapts concepts from linear regression:

  i Replacing least squares with alternative criteria (e.g., Huber, Tukey loss).

  ii Weighted least squares or iterative methods.

- Practical applications and algorithms (e.g., IRLS, MM-estimation).

- Examples or case studies comparing linear and robust regression outcomes (highlight trade-offs in robustness, efficiency, and computational complexity).

1. **Core Adjustments** Explain how robust regression modifies least squares to handle outliers and assumption violations. Link this back to M-estimation and the ψ-functions from the Fundamentals chapter.

2. **Methods** Iteratively reweighted least squares (IRLS) as an intuitive algorithm. Examples of robust methods (Huber regression, Tukey's bisquare).

3. **Comparison** Use small simulations or examples (e.g., synthetic datasets with and without outliers) to show the advantages of robust regression over OLS.

Chemometrics Article

- The robust estimators aim to describe well the data majority regardless the data contamination. The robustness of an estimator can be described by its breakdown point, a concept introduced by Hampel, however also other robustness criteria exist.

- In general, have qualitative and quantitative robustness. Qualitative robustness aims to express the differences between two studied distributions by means of the **Prohorov distance**. When this distance is small then the difference between the distributions of estimations is also small. There are two concepts that target quantitative robustness issue (i) the breakdown point of an estimator and (ii) its **influence function**, expressing global and local sensitivity, respectively.

- The **breakdown point** of an estimator is the proportion of incorrect data points that can be handled before the estimator produces an incorrect result. Mean is 0% and median is 50%. The influence function of an estimator aims to describe the influence of objects upon the estimator with respect to infinitesimal perturbations. Namely, sensitivity to changes in sample.

- **Efficiency** of an estimator expresses how good estimates the estimator yields for non-contaminated data compared to a classical estimator

- In general, a good robust estimator should have a high efficiency, high breakdown point and smoothed influence function.

- Types of robust estimators:

  i **Parametric**: Parametric estimators assume a certain data distribution, for instance that the data majority follows a normal distribution and thus such estimators simply eliminate outliers. Such parametric estimators work well for contaminated data but also offer better estimates compared to classical ones obtained for other non-normal models of the data distribution (Cauchy, t, Laplace, etc.).

  ii **Non-parametric**: Non-parametric estimators are robust in their nature because they do not require knowledge about the data distribution at hand.

  iii **Semi-non-parametric**: Semi-non-parametric approach, where any type of the non-normality can be handled. Contrary to parametric estimators, semi-non-parametric estimators do not reject outliers but transform the data.

Mean and Median

The lack of robustness of the mean estimator, can be explained by its least squares nature. The mean of a random variable is a point minimizing the Euclidean distances to all data objects. This condition is expressed as

$$\min_{\mu} \sum_{i=1}^{n} \|x_i - \mu(\mathbf{x})\|$$

where $\|...\|$ is the L2-Euclidean norm.

The median is a robust alternative, being the middle value element for the sorted elements. Has a breakdown point of 50%, which is the highest possible.

For multidimensional data, can compute column means or medians.

Variance and Standard Deviation

The standard deviation, σ, and the variance, σ^2, of the random variable x are used to describe the data spread (scale). They are defined as:

$$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \mu(\mathbf{x}))^2$$

Similarly to the mean, it has a breakdown point of 0%. There are three different well-known robust estimates of the standard deviation:

i **Median of Absolute Deviation (MAD)**, σ_{MAD}: This estimator is defined

$$\sigma_{\text{MAD}} = c \cdot \text{median}\left(|x_i - \text{median}(\mathbf{x})|\right)$$

where $c = 1.4826$. The σ_{MAD} estimator has a 50% breakdown point. Although it is easy to compute, it suffers from a low efficiency

ii **Sn scale estimator**, σ_{Sn}: This is a more efficient estimatre, along with a breakdown point of 50%. Defined: TBD

iii **Qn scale estimator**, σ_{Qn}: TBD

Classical and robust methods of data transformation

In data analysis often the studied data require preprocessing (a transformation). Data preprocessing aims to correct undesired effects, for instance, (to remove offset from the data?), to give equal impact of every variable in the analysis, to enhance the quality and interpretability of the obtained results.

Affine translations a combination of linear transformation and translation. Maps a straight line to a straight line for example, and given

$$\mathbf{AX} - \mathbf{1b}^T$$

For example a popular transformation is the z-transformation to center the data around 0 and make all variances the same. However, in the presence of outliers must use robust centering and auto-scaling.

References

[1] Huber, P. J., and Ronchetti, E. M. (2009). *Robust Statistics.* 2nd ed., Wiley.

[2] Wilcox, R. R. (2021). *Introduction to Robust Estimation and Hypothesis Testing.* 5th ed., Academic Press.

[3] Maronna, R. A., Martin, D. R., and Yohai, V. J. (2006). *Robust Statistics: Theory and Methods.* Wiley.