# 1 Introduction and Foundations

In this chapter, the basic concepts of robust estimation in a signal processing framework are introduced. After a brief historical recount, we discuss outlier and heavy-tailed distribution models. These models are common in engineering practice as is evident from numerous measurements made in different fields, for example, in digital communication. Among other heavy-tailed noise models, we introduce in this chapter the epsilon mixture model that is used extensively in subsequent chapters. We then consider the estimation of location and scale parameters in the real data case. The principles underpinning this case are demonstrated by considering the simple problem of estimating the direct current (DC) value when measurements are subject to random fluctuations that are independently and identically distributed (i.i.d.) from sample to sample. In this problem, the $M$-estimator is introduced and this highlights the intuitive link to maximum likelihood estimation for different noise models. Important measures of robustness then follow and these include the influence function (IF) and the breakdown point (BP). The introduction of robustness in estimation comes at the price of decreased statistical efficiency of the estimator and the trade-off between robustness and efficiency is discussed. This trade-off is demonstrated by considering location estimation based on the sample median. An understanding of this trade-off is likely to facilitate the signal processing practitioner to design robust estimators for location and scale. Because this chapter is intended to serve as an easy-to-read introduction to robustness concepts, only the real-valued case is discussed. The complex-valued case is treated in Chapter 2, where the linear regression model is introduced, which contains as a special case the location (or location-scale) model.

Several examples, along with the associated MATLAB© code that allows users to reproduce the results, are included in the downloadable RobustSP toolbox.

## 1.1 History of Robust Statistics

Statistical signal processing is an important area of research that has been successfully applied to generations of engineering problems where the extraction of useful information from empirical data is required. An effective way to incorporate knowledge from empirical data is to use parametric stochastic models. Important foundations were established in the 1920s by R. A. Fisher (Fisher, 1925) who derived many useful statistical models and methods. When applying parametric methods to real-world problems, the

situation often arises that the observations do not exactly follow the assumptions made to model the problem. In these cases, the nominally excellent performance can drastically degrade.

From a practitioner's viewpoint, however, it is essential that the results associated with the parametric method used be acceptable in situations where the distributional assumptions underpinning the assumed model do not hold. One approach is to make as few assumptions as possible about the data and resort to a nonparametric statistical model. In some signal processing applications, for example in spectrum estimation, nonparametric approaches, based on the periodogram, have become widely popular. However, in today's engineering practice, parametric models continue to play an important role. This is especially the case in complex applications, where, to retrieve meaningful information, it is necessary to incorporate some knowledge about the system under consideration. So which strategy should one follow? Everyone who deals with real-world problems can relate to the famous remark by G. E. P. Box on robustness in statistical model building, "All models are wrong, but some are useful" (Box, 1979). If we acknowledge that the data model we use is at best a close approximation to the true model from which real measurements have been obtained, it is then only a small step to robust statistics.

Robustness, as treated in this book, deals with deviations from the distributional assumptions, and we mainly consider deviations from a Gaussian (normal) probability model. The word *robust* was introduced into the statistics literature by G. E. P. Box in 1953 (Box, 1953). The study of robustness, however, predates even this pioneering work. According to D. Bernoulli (Bernoulli, 1777), outlier rejection was already common practice in 1777. Mixture models and estimators that down-weight outliers were known in the 1800s and S. Newcomb even "preinvented" a kind of one-step Huber-estimator (Stigler, 1973). The question of how best to characterize uncertainties in observations has been an ongoing discussion since the early days of statistics. The first scientist to note in print that measurement errors deserve a systematic and scientific treatment was G. Galileo in 1632 (Galilei, 1632).

Since its discovery in 1733 by A. de Moivre (de Moivre, 1733), the normal distribution has played a central role in statistical modeling. It was named after C. F. Gauss, who derived it to justify his use of the least squares criterion in astronomy to locate an orbit that best fitted known observations (Gauss, 1809). He developed a theory of errors that is based on the following assumptions: (i) small errors are more likely than large errors; (ii) the likelihood of the errors being positive or negative is equal; and (iii) in the presence of several measurements of the same quantity, the most likely value of the quantity being measured is their average. On this basis, Gauss derived the formula for the normal probability density of the errors (Stahl, 2006), and this formula has since been justified in many different ways and shown to be applicable in many different contexts such that it is the default model that is used is many applications.

As H. Poincaré pointed out in 1904 (Poincaré, 1904), "Physicists believe that the Gaussian law has been proved in mathematics while mathematicians think that it was experimentally established in physics." Even today, many methods encountered in engineering practice rely on the Gaussian distribution of the data, which in many situations

is well justified and motivated (Kim and Shevlyakov, 2008) by the central limit theorem. Assuming Gaussianity can be practical in many situations, for example, using the Gaussian error model can be based on the argument that it minimizes the Fisher information over the class of distributions with a bounded variance, and the Fourier transform of a Gaussian function is another Gaussian function. Assuming Gaussianity also enables a simple derivation of likelihood functions. In summary, the main justification for assuming a normal distribution is twofold. On the one hand, it provides an approximate representation for many real-world data sets. On the other hand, it is convenient from a theoretical viewpoint as it facilitates the derivation of closed-form expressions for optimal detectors or estimators. Optimality is clearly a desirable property for a detector or an estimator. Optimality, only under the assumed (nominal) distribution, however, is useless if the estimator is applied to data that does not follow this distribution. As highlighted by Tukey in 1960, even slight deviations from the assumed distribution may cause the estimator's performance to drastically degrade or to completely break down (Zoubir et al., 2012).

Robust statistics formalizes the theory of approximate parametric models. On the one hand, robust methods are able to leverage a parametric model, but on the other hand, such methods do not depend critically on the exact fulfillment of the model assumptions. In this sense, robust statistics are consistent with engineering intuition and signal processing demands. Robust methods are designed in such a way that they behave nearly optimally, if the assumed model is correct, while small deviations from the model assumptions degrade performance only slightly and larger deviations do not cause a catastrophe (Huber and Ronchetti, 2009). The theory of robust statistics was established in the middle of the twentieth century by the pioneering work of J. P. Tukey, P. J. Huber, and F. R. Hampel, who are often called the "founding fathers" of robust statistics. In 1960, J. W. Tukey (Tukey, 1960) summarized his work in the 1940s and 1950s on the effect of a small amount of contaminating data (outliers) on the sample mean and standard deviation. He introduced a contamination model and proposed some estimators that are robust against such contamination.

The first attempt toward a unified framework for robust statistics was undertaken in the seminal paper of P. J. Huber on robust location estimation in 1964 (Huber, 1964). After defining neighborhoods around a true distribution that generates the data, he proposed an estimator that yields minimax optimal performance over the entire neighborhood. This means that the estimator is optimal for the worst-case distribution within the neighborhood. For details on Huber's approach, the reader is referred to the book by P. J. Huber and E. M. Ronchetti (Huber and Ronchetti, 2009).

Further fundamental concepts of robust statistics were introduced by F. R. Hampel in 1968 (Hampel, 1968). His so-called infinitesimal approach is based on three central concepts: qualitative robustness, the IF and the BP. Intuitively, they correspond to the continuity and first derivative of a function and the distance to its nearest singularity. Interested readers are referred to the book by F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel (Hampel et al., 2011).

In engineering, robust estimators and detectors have been of interest since the early days of digital signal processing; see the review paper on robust methods published

by Kassam and Poor in 1985 (Kassam and Poor, 1985) and references therein. Since then, many aspects of robustness have been utilized in signal processing associated with communications systems, radar and sonar, pattern recognition, biomedicine, speech, and image processing, amongst others. The increasing complexity of the data (and models) that are analyzed today have triggered new areas of research in robust statistics. Today, for example, robust methods have to deal with high-dimensional, sparse, multivariate, and/or complex-valued data. Nevertheless, much of today's work is based on the ideas that were formalized in the middle of the last century.

In many areas of engineering today, the distribution of the measurement data is far from Gaussian as it contains outliers, which cause the distribution to be heavy-tailed. In particular, measurement data from a diversity of areas (Blankenship et al., 1997; Abramovich and Turcaj, 1999; Middleton, 1999; Etter, 2003) have confirmed the presence of impulsive (heavy-tailed) noise, which can cause optimal signal processing techniques, especially the ones derived under the Gaussian probability model, to be biased or to even break down.

The occurrence of impulsive noise has been reported, for example, in outdoor mobile communication channels due to switching transients in power lines or automobile ignition (Middleton, 1999), in radar and sonar systems as a result of natural or man-made electromagnetic and acoustic interference (Abramovich and Turcaj, 1999; Etter, 2003), and in indoor wireless communication channels, owing, for example, to microwave ovens and devices with electromechanical switches, such as electric motors in elevators, printers, and copying machines (Blankenship et al., 1997). Moreover, biomedical sensor array measurements of brain activity, such as in magnetic resonance imaging (MRI) and associated with regions of the human brain where complex tissue structures are known to exist, have been found to be subject to non-Gaussian noise and interference (Alexander et al., 2002).

In geolocation position estimation and tracking, non-line-of-sight (NLOS) signal propagation, caused by obstacles such as buildings or trees, results in outliers in measurements, to which conventional position estimation methods are very sensitive (Hammes et al., 2009). In classical short-term load forecasting, the prediction accuracy is adversely influenced by outliers, which are caused by nonworking days or exceptional events such as strikes, soccer's World Cup, or natural disasters (Chakhchoukh et al., 2010). Moreover, on a computer platform, various components, such as the liquid crystal display (LCD) pixel clock and the peripheral component interconnect (PCI) express bus, cause impulsive interference that degrades the performance of the embedded wireless devices (Nassar et al., 2008). These studies show that in real-world applications, robustness against departures from Gaussianity is important. It is therefore not surprising that robustness is becoming an important area of engineering practice, and more emphasis has been given in recent years to the design and development of robust systems. The complexity of new engineering systems and the high robustness requirements in many applications suggest the urgent need to further revisit robust estimation techniques and present them in an accessible manner.

## 1.2 Robust *M*-estimators for Single-Channel Data

In this section, robust *M*-estimators for single-channel data are introduced. *M*-estimation is easily accessible in the single-channel context; in later chapters, we will show how this concept can be applied to other areas such as multichannel data and linear regression.

### 1.2.1 Location and Scale Estimation

The robust estimation of the location and scale parameters of a univariate random variable is considered to be the origin of what we know today as robust statistics. In the 1940s and 1950s, J. W. Tukey, one of the pioneers of statistics of the twentieth century, investigated the effect of small amounts of contaminating data (outliers) on the sample mean and standard deviation. Tukey also proposed robust estimators that are not severely affected by outliers (Tukey, 1960). In 1964, P. J. Huber formalized robust statistical theory and introduced *M*-estimation in his seminal paper on robust location estimation (Huber, 1964).

Consider the Thevenin equivalent model of a DC electrical system, as illustrated in Figure 1.1, with a Thevenin equivalent voltage of $\mu$ and a Thevenin equivalent resistance of $R$ Ohm. Thermal noise in the resistance leads to a time-varying noise signal, denoted $v(t)$, in series with the DC voltage source. The voltage at the system output is denoted $y(t)$.

The noise signal arising from a resistor has a uniform power spectral density over a band that usually well exceeds the bandwidth of a measurement system, and time samples from such a signal are consistent with samples from a Gaussian probability density function (pdf). From a random process perspective, the noise arising from a resistor has a white power spectral density and a Gaussian amplitude pdf, that is, the noise is additive white Gaussian noise (AWGN). Consistent with this, and in an electrical context, the measurement of a DC level is modeled according to

$$y_i = \mu + v_i, \quad i = 1, \ldots, N, \tag{1.1}$$

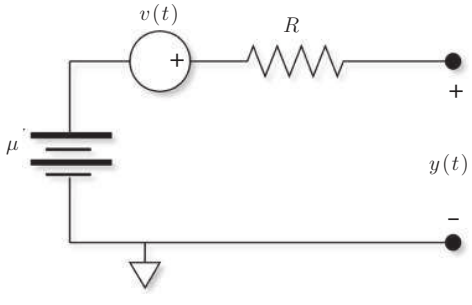


**Figure 1.1** Thevenin equivalent model for a DC electrical system.

where $y_i$ are random variables[1] that model measurements taken at time instances $t_i, i \in \{1,\dots,N\}$ and $v_i$ are identically and independently distributed (i.i.d.) variables for $i = 1,\dots,N$. Because the DC voltage is constant over time, it is represented by the deterministic scalar quantity $\mu$. A common assumption is that the random variable $v_i$ follows the zero-mean Gaussian distribution, whose pdf is given by

$$f\left(v_i\middle|\mu_v,\sigma_v^2\right) = \frac{1}{\sqrt{2\pi\sigma_v^2}} e^{-\frac{(v_i-\mu_v)^2}{2\sigma_v^2}}, \tag{1.2}$$

with $\mu_v = 0$. Under these assumptions, the pdf of the measurements is

$$f\left(y_i\middle|\mu,\sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\mu)^2}{2\sigma^2}}, \tag{1.3}$$

where $\sigma^2 = \sigma_v^2$.

For many of today's engineering applications, however, the AWGN model is not an adequate representation. When considering the measurement of a DC voltage, impulsive non-Gaussian noise can be injected, for example, by DC–DC converters, switching mode power supplies found in light dimmers, or switching thermostats in fridges or cookers.

### Maximum Likelihood Estimation of Location and Scale

The goals of location and scale estimation are to determine the values $\mu$ and $\sigma$, which best model the observations/measurements $\mathbf{y} = (y_1,\dots,y_N)^\top$ from (1.1). Provided that the Gaussian noise assumption is fulfilled, that is, that the pdf of $v_i$ is given by (1.2), the maximum likelihood estimators (MLEs) are the sample mean and the sample standard deviation.

Assuming statistical independence of $y_i$, $i = 1,\dots,N$, this directly follows from (1.1) and (1.3) by taking the partial derivatives of the Gaussian negative log-likelihood function

$$L_{\text{ML}}(\mu,\sigma|\mathbf{y}) = \frac{N}{2}\ln(2\pi\sigma^2) + \frac{\sum_{i=1}^N (y_i-\mu)^2}{2\sigma^2} \tag{1.4}$$

with respect to the unknown parameters $\mu,\sigma$

$$\frac{\partial}{\partial\mu} L_{\text{ML}}(\mu,\sigma|\mathbf{y}) = -\frac{2\sum_{i=1}^N (y_i-\mu)}{2\sigma^2}$$

$$\frac{\partial}{\partial\sigma} L_{\text{ML}}(\mu,\sigma|\mathbf{y}) = \frac{N}{\sigma} - \frac{\sum_{i=1}^N (y_i-\mu)^2}{\sigma^3}$$

and setting them equal to zero. Thus, the sample mean is such that

$$\sum_{i=1}^N (y_i - \hat{\mu}) = 0$$

[1] Throughout the book, we will not explicitly differentiate between a random variable $X$ and its realization $x$. This should be understood from the context.

and the sample standard deviation is such that

$$N - \frac{\sum_{i=1}^{N}(y_i - \hat{\mu})^2}{\hat{\sigma}^2} = 0$$
$$\Leftrightarrow \frac{1}{N}\frac{\sum_{i=1}^{N}(y_i - \hat{\mu})^2}{\hat{\sigma}^2} = 1.$$

Solving yields the well-known estimators of location

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{1.5}$$

and scale

$$\hat{\sigma} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{\mu})^2}. \tag{1.6}$$

Because the objective function that is defined in (1.4) is a jointly convex function in $(\mu, 1/\sigma^2)$, a global minimizer can be found. Provided that the Gaussian assumption is fulfilled, the sample mean and sample standard deviation that are defined, respectively, in (1.5) and (1.6) are optimal in the sense that they attain the Cramér–Rao lower bound (CRLB). This means that the distributions of $\sqrt{N}(\hat{\mu} - \mu)$ and $\sqrt{N}(\hat{\sigma} - \sigma)$ for $N \to \infty$ tend to Gaussian distributions whose mean values are the true values (consistency) and whose covariance is equal to the inverse of the Fisher information matrix (efficiency).

For the Gaussian distribution, and consistent with (1.5) and (1.6), it is optimal to give all observations equal importance in the objective function. However, if the noise pdf $f\left(v_i|\mu_v, \sigma_v^2\right)$ is that of a non-Gaussian random variable, and the measured data contains outliers, our intuition dictates that we weight the observations $y_i, i = 1, \ldots, N$, in a manner to give more importance to observations that are close to the measurement model as compared to the ones that are unlikely to occur. This is precisely what robust location and scale estimation is about.

A frequently used approach to derive robust estimators is to compute the MLE for a heavy-tailed noise model, for example, the Laplace or the Cauchy noise distribution (see Figure 1.2). The Laplace distribution has the pdf

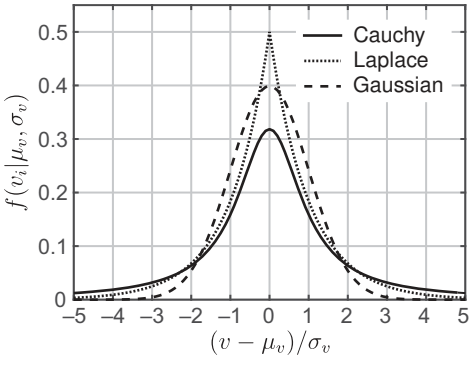


**Figure 1.2** The Gaussian, Laplace, and Cauchy pdfs.

**Table 1.1** The probability that a random variable takes on a value that is more than a few $\sigma$ away from $\mu$.

| | Gaussian | Laplace | Cauchy |
|---|---|---|---|
| $\text{Prob}(\lvert y-\mu\rvert > 3\sigma)$ | 0.0027 | 0.0498 | 0.2048 |
| $\text{Prob}(\lvert y-\mu\rvert > 4\sigma)$ | $6.3342 \cdot 10^{-5}$ | 0.0183 | 0.1560 |
| $\text{Prob}(\lvert y-\mu\rvert > 5\sigma)$ | $5.7330 \cdot 10^{-7}$ | 0.0067 | 0.1257 |
| $\text{Prob}(\lvert y-\mu\rvert > 10\sigma)$ | 0 | $4.5400 \cdot 10^{-5}$ | 0.0635 |

$$f(v_i|\mu_v,\sigma_v) = \frac{1}{\sqrt{2}\sigma_v} e^{-\frac{\sqrt{2}|v_i-\mu_v|}{\sigma_v}}, \tag{1.7}$$

where $\mu_v$ and $\sigma_v$ are the location and scale parameters, respectively. The Cauchy distribution has the pdf

$$f(v_i|\mu_v,\sigma_v) = \frac{1}{\pi\sigma_v} \cdot \frac{\sigma_v^2}{(v_i-\mu_v)^2+\sigma_v^2}. \tag{1.8}$$

As shown in Table 1.1, the probability that a Laplace, or Cauchy, distributed random variable takes on a value that is more than three standard deviations away from $\mu$ is significantly different from zero. For the Cauchy-distributed random variable, even the probability of taking on a value that is more than ten $\sigma$ away from $\mu$ is 0.0635.

If the model for the noise is the Laplace distribution, that is, (1.7) holds with $\mu_v = 0$, the pdf of $y_i$ becomes

$$f(y_i|\mu,\sigma) = \frac{1}{\sqrt{2}\sigma} e^{-\frac{\sqrt{2}|y_i-\mu|}{\sigma}},$$

where $\sigma = \sigma_v$. For $N$ observations, the negative log-likelihood function is then given by

$$L_{\text{ML}}(\mu,\sigma|\mathbf{y}) = N\ln(\sqrt{2}\sigma) + \frac{\sqrt{2}}{\sigma}\sum_{i=1}^{N}|y_i-\mu|. \tag{1.9}$$

Taking the partial derivative with respect to $\mu$ yields

$$\begin{aligned}\frac{\partial}{\partial\mu}L_{\text{ML}}(\mu,\sigma|\mathbf{y}) &= \frac{\sqrt{2}}{\sigma}\sum_{i=1}^{N}\frac{\partial|y_i-\mu|}{\partial\mu}\\ &= -\frac{\sqrt{2}}{\sigma}\sum_{i=1}^{N}\text{sign}(y_i-\mu)\end{aligned} \tag{1.10}$$

where the identity

$$\frac{\partial|x|}{\partial x} = \frac{x}{|x|} = \text{sign}(x).$$

has been used and the sign function is defined as

$$\mathsf{sign}(x) = \begin{cases} +1, & \text{if } x > 0, \\ 0, & \text{if } x = 0, \\ -1, & \text{if } x < 0. \end{cases} \tag{1.11}$$

The MLE of the location parameter $\mu$, is the solution of

$$\frac{\sqrt{2}}{\sigma} \sum_{i=1}^{N} \mathsf{sign}(y_i - \mu) = 0$$

and the sample median, that is,

$$\mathsf{med}(\mathbf{y}) = \begin{cases} y_{\left(\frac{N+1}{2}\right)}, & \text{if } N \text{ is odd}, \\ \frac{1}{2}\left(y_{\left(\frac{N}{2}\right)} + y_{\left(\frac{N}{2}+1\right)}\right), & \text{if } N \text{ is even}, \end{cases} \tag{1.12}$$

given ordered samples $\{y_{(1)} \leq \ldots \leq y_{(N-1)} \leq y_{(N)}\}$. From

$$\frac{\partial}{\partial \sigma} L_{\mathrm{ML}}(\mu, \sigma | \mathbf{y}) = 0$$

the MLE of the scale parameter turns out to be the mean of the absolute deviations from the median, that is,

$$\hat{\sigma} = \frac{1}{N} \sum_{i=1}^{N} |y_i - \mathsf{med}(\mathbf{y})|. \tag{1.13}$$

Weighted medians are addressed in Section 2.4.1 in the context of linear regression. Median and weighted median filters are discussed in Section 7.2.

The negative log-likelihood function for the Cauchy distribution given a sample size $N$ is

$$L_{\mathrm{ML}}(\mu, \sigma | \mathbf{y}) = N \ln(\sigma \pi) + \sum_{i=1}^{N} \ln\left(1 + \left(\frac{y_i - \mu}{\sigma}\right)^2\right). \tag{1.14}$$

Taking the partial derivatives of (1.14) with respect to the unknown parameters $\mu$ and $\sigma$ yields

$$\frac{\partial}{\partial \mu} L_{\mathrm{ML}}(\mu, \sigma | \mathbf{y}) = -2\sigma \sum_{i=1}^{N} \frac{y_i - \mu}{\sigma^2 + (y_i - \mu)^2} \tag{1.15}$$

and

$$\frac{\partial}{\partial \sigma} L_{\mathrm{ML}}(\mu, \sigma | \mathbf{y}) = \frac{N}{\sigma} - \frac{2}{\sigma} \sum_{i=1}^{N} \frac{1}{\sigma^2 + (y_i - \mu)^2}. \tag{1.16}$$

To find the Cauchy location and scale estimates, a numerical solution to

$$\sum_{i=1}^{N} \frac{y_i - \mu}{\sigma^2 + (y_i - \mu)^2} = 0$$
$$\sum_{i=1}^{N} \frac{1}{\sigma^2 + (y_i - \mu)^2} = \frac{N}{2}$$

is required. Kalluri and Arce (2000), for example, provide an algorithm that employs a fixed-point (FP) search that is guaranteed to converge to a local minimum. However, one needs to be careful with the choice of the local minimum because the Cauchy likelihood can have multiple spurious roots (Reeds, 1985). Finding a global optimum for the location and scale of non-Gaussian ML functions is still an open problem.

### *M*-estimation of Location and Scale

An important class of robust estimators are *M*-estimators (Huber and Ronchetti, 2009), which are a generalization of MLEs. Because this chapter is intended to serve as an easy-to-read introduction to robustness concepts, only the real-valued case is discussed. *M*-estimation of location and scale is extended to the complex-valued case in Section 2.5 of the next chapter, where linear regression is discussed.[2]

*M*-estimators replace the negative log-likelihood function $L_{\text{ML}}(\mu, \sigma|\mathbf{y})$ with a different objective function $L_{\text{M}}(\mu, \sigma|\mathbf{y}) = \rho(\mu, \sigma|\mathbf{y})$. If $\rho(\cdot)$ is differentiable, with

$$\psi(x) = \frac{d\rho(x)}{dx}, \tag{1.17}$$

then the *M*-estimating equations follow:

$$\sum_{i=1}^{N} \psi\left(\frac{y_i - \hat{\mu}}{\hat{\sigma}}\right) = 0 \tag{1.18}$$

and

$$\frac{1}{N}\sum_{i=1}^{N} \psi\left(\frac{y_i - \hat{\mu}}{\hat{\sigma}}\right) \cdot \left(\frac{y_i - \hat{\mu}}{\hat{\sigma}}\right) = b. \tag{1.19}$$

Here, $b$ is a positive constant that must satisfy $0 < b < \rho(\infty)$. If $f(y_i|\mu, \sigma)$ is symmetric, then $\rho(\mu, \sigma|\mathbf{y})$ is even and, hence, $\psi(\mu, \sigma|\mathbf{y})$ is odd. MLEs are included within the class of *M*-estimators by setting $\rho(\mu, \sigma|\mathbf{y}) = L_{\text{ML}}(\mu, \sigma|\mathbf{y})$. For example, in the Gaussian noise case, the MLE is obtained by letting $\psi(x) = x$ and $b = 1$ in (1.18) and (1.19).

*M*-estimators are classified into two categories depending on the shape of $\psi(x)$, namely the *monotone* and the *redescending* *M*-estimators. Within the redescending class, the *M*-estimators for which $\psi(x)$ returns exactly to zero, that is, $\psi(x_0) = 0$ for some value $x_0$, are called *strongly redescending*. For a detailed discussion of different $\psi$ functions, the interested reader is referred, for example, to Huber and Ronchetti (2009, chapter 4).

[2] The linear regression model contains as a special case the location (or location-scale) model when the design matrix is a column vector of ones.