# Robust Speech Filtering
# In Impulsive Noise Environments

Christelle Ledoux

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
in
Electrical Engineering

Lamine Mili, chair
Jeffrey Reed
Louis Beex

December 13, 1999
Blacksburg, Virginia

Keywords: robust statistics, impulsive noise, linear prediction coding,
M-estimators, GM-estimators, projection statistics, Kalman filter.

# Robust Speech Filtering in Impulsive Noise Environments

Christelle Ledoux

## Abstract

This thesis presents a new robust filtering technique that suppresses impulsive noise in speech signals. The method makes use of Projection Statistics based on medians to detect segments of speech with impulses. The autoregressive model employed to smooth out the speech signal is identified by means of a robust nonlinear estimator known as the Schweppe-type Huber GM-estimator. Simulation results are presented that demonstrate the effectiveness of the filter. Another contribution of the work is the development of a robust version of the Kalman filter based on the Huber M-estimator. The performances of this filter are evaluated for a simple autoregressive process.

# Chapter 1

# Introduction

*Problem statement*

The purpose of this research is to develop robust filtering methods to suppress impulses in speech signals. This research was motivated by the growing importance of mobile communication systems, and therefore the need to find new methods to process speech signals [15, 39]. In impulsive noise environments, the use of robust statistics is very appropriate, which will be shown all along this thesis.

The evolution of mobile communications is probably one of the most important technological changes in modern time. In fact, it is much more than that, it is a real revolution from an economical and social viewpoint. In a very near future, we can expect that the number of wireless phone customers in the world will be equal to that of conventional lines.

The presence of fading channels and interference in cellular phone communication are among the trickiest problems [32]. They are also the main sources of impulsive noise in a speech signal. For this reason, suppression of this type of noise is of great importance. Impulsive noise is characterized by a relatively large amplitude and a short duration [37]. It leads to the complete destruction of the information of the signal on the segment where it is present. From a statistical viewpoint, the corruption of a signal by impulsive noise can be seen as the introduction of outliers among the data. Using robust statistics is a direct way to suppress them.

The theories of robust statistics are a generalization of the parametric estimation theory. While parametric estimation relies on the assumption that the probability distribution of the noise is known, robust estimation deals with the situation were some

unknown perturbations are introduced. It leads to the construction of estimators that provide good estimates even in the presence of outliers [12, 14].

*State of the art*

If speech enhancement has been the subject of a great deal of research in the past decades, only little attention was paid to the problem of suppressing impulsive noise [20, 27]. This is a very difficult problem, because the approximation of Gaussianity that is used in most of the filtering techniques can not be assumed. However, different methods using robust statistics have already been tested.

The median filter, which can be seen as the simplest robust filter, has been advocated as a mean to suppress pulses. However, the window of the filter has to be twice as large as the pulse, which is often too much to keep a signal of good quality [7, 37]. Other alternatives are based on the Linear Prediction Coding (LPC) technique, which is one of the most widely used methods to model and filter a speech signal. The principle of the method consists in identifying an autoregressive model of the signal and comparing the estimated values to the associated observations. If the residual has a large amplitude, the associated observation is considered as corrupted by impulsive noise and replaced by an estimate [38]. But this method requires an estimation process that is not influenced by the impulsive noise. Some research has been reported using the Huber M-estimator [18]. But because the LPC method can be seen as a linear regression, robust statistics tell us that only an estimation using GM-estimators enables us to detect efficiently both leverage points and vertical outliers.

*Contribution of this research*

In this study, a LPC method using the Schweppe-type Huber GM-estimator [13, 14, 26] is developed and tested through simulations. It enables us to estimate the parameters of a model in a robust way. The identification of the impulsive noise is realized by means of a new technique too. It makes use of the so-called Projection Statistics, which is a very robust technique to detect outliers among a point cloud in high dimension [4, 9, 25, 34, 35]. Specifically, the signal is divided into small segments, and an observation matrix containing consecutive values of the signal is built. Each row of

this matrix is associated with its Projection Statistic, which is then compared to a threshold value. This method is all the more interesting than the simplicity of its algorithm makes its utilization in real time applications possible. Moreover, it presents the advantage that the probability distribution of the Projection Statistics is approximately known. Therefore, the threshold value under which the amplitude of the signal is considered as too high can be determined.

Part of the work is also devoted to the study of the Kalman filter [1, 3, 8, 10, 11, 21, 24, 28] and to the development of a robust version of this filter based on the Huber M-estimator [14, 17, 23]. This research was motivated by the good behavior of the classical Kalman filter for speech enhancement [8, 11, 28]. Besides, a robust filter is able to filter the noise with longer tails than the Gaussian distribution. It is particularly useful in speech processing, knowing that the probability distribution of speech is close to a Laplacian distribution. Successful simulations of this robust filter are carried out in a simple case, but the application to speech filtering would require more research and development.

*Overview of the research*

The thesis is organized as follows. Chapter 2 gives an overview of speech processing as applied to the impulsive noise problem. The main properties of speech signals are presented and the LPC method is explained. The sources and characteristics of impulsive noise are exposed and the current existing techniques reviewed.

Chapter 3 presents the fundamentals of parametric and robust estimation. The maximum likelihood approach is presented, associating each probability distribution with an optimal estimator. Then, the class of the robust M-estimators is introduced, with emphasis on the Huber M-estimator.

Chapter 4 is devoted to the linear regression problem. Both classical and robust methods are explained. It leads to the definition of Projection Statistics and of the Schweppe-type Huber GM-estimator, which enables us to decrease the influence of both vertical outliers and leverage points.

Chapter 5 presents a robust version of the Kalman filter based on the Huber M-estimator. Simulations are made in a simple case to verify its ability to attenuate impulsive noise.

Chapter 6 makes use of all the preceding Chapters to develop a robust method to suppress impulsive noise in speech signals. Projection Statistics are used to detect the impulses and corrupted values are replaced by estimates calculated by means of the Schweppe-type Huber GM-estimator. The improvement realized is very encouraging and appears clearly in both time domain and frequency domain representations. This successful result has been confirmed by listening tests.

# Chapter 2

# Speech processing

# in presence of impulsive noise

## 2.1. Introduction

This Chapter gives an overview of the impulsive noise problem along with some basic speech processing techniques. In a first step, the sources of impulsive noise are presented, together with its modeling. The representation in the time and frequency domains is given to illustrate the problem. Then, the speech mechanism and the Linear Prediction Coding (LPC) method are described. This method allows us to model a speech signal with few coefficients, which significantly decreases the memory necessary to store it. Finally, the existing techniques to remove impulsive noise in speech signals are exposed and different ways to evaluate the quality and intelligibility of speech are described.

## 2.2. Impulsive noise

### 2.2.1. Sources of impulsive noise

Impulsive noise is characterized by a relatively large amplitude, which highly damages the information contained in the signal. It may be found in several situations.

One such a situation is provided by systems that are submitted to a discontinuity in the process of transmission or of measurement. For example, this happens when there is a discontinuity in the channel due to the closing and opening of switches.

Impulsive noise may also be generated by the process of transmission itself, if there is a discontinuity of the phase in a FM modulated signal while a differentiator is used in the demodulation [2]. Such phase discontinuities appear very often in wireless communication. For example, this is the case when a channel is submitted to fading or to interference [30].

Fading is the result of multipath propagation. In a wireless system, the wave can use several paths to travel from the transmitter to the receiver. Therefore, the received signal consists of several replications of the transmitted signal arriving at slightly different times. The resulting signal is characterized by very rapid variations in amplitude and in phase.

Impulsive noise is also a consequence of adjacent or co-channel interference. The presence of interference is inherent to the principle of frequency reuse, which is a way to highly increase the number of users in a defined area. Each area is divided into several zones, in which the signals are transmitted with a particular frequency. Different frequencies are assigned to neighbor cells, but the same frequency is reused in cells that are further away. If the frequencies assigned to two adjacent cells are too close, the spectrum of the signals of these cells is susceptible to overlap, which creates adjacent channel interference. If the distance between two cells that use the same frequency is too short, a second type of interference called co-channel interference occurs. In both cases, it results in the introduction of impulsive noise.

### 2.2.2. Modeling of impulsive noise

Impulsive noise can be modeled as a stochastic process. Different representations can be used, the Poisson-Gaussian model being one of the most common example [5, 37]. Impulsive noise will be seen as the response of a filter h to a train of impulses x with amplitude following a Gaussian distribution and interarrival time following a Poisson distribution.

The filter, which models the shape of the pulse, can be seen as an exponential function. Its impulse response is then given by

$$h(t) = e^{-t/T_p} u(t), \qquad (1)$$

where $T_p$ is the effective duration of the impulsive noise sequence and u is the unit step function.

Usually, impulsive noise consists of two consecutive impulses of opposite signs. It may be modeled as

$$h(t) = \begin{cases} e^{(t-T_p/2)/T_p} & t < \dfrac{T_p}{2} \\ -e^{(-t+T_p/2)/T_p} & \dfrac{T_p}{2} \le t \end{cases}. \qquad (2)$$

The input signal x is given by

$$x(t) = \sum_{i=1}^{N} a(t_i)\,\delta(t - t_i), \qquad (3)$$

where the sequence $\{a(t_i), i = 1,\ldots,N\}$ is drawn from a Gaussian distribution and where the intervals $\Delta_i = t_{i+1} - t_i$ follow the Poisson distribution defined as

$$P(k,T) = \frac{(\lambda T)^k}{k!} e^{-\lambda T}. \qquad (4)$$

Here T is the duration of the signal and $\lambda$ is the average number of pulses desired on this duration.

### 2.2.3. Illustration in time and frequency domains

Figure 1 and Figure 2 represent the shape of an impulsive noise in the time and frequency domains. Here we took $T_p$ = 3 seconds, which creates a pulse with a duration of about 3 milliseconds. We sampled it with a sample frequency of 8 kHz, and calculated the Fast Fourier Transform (FFT). The spectrum we obtain is almost flat, like that of a Dirac impulse. Therefore, all the frequencies from 0 to 4 kHz (half the sample frequency) are present with equal energy.

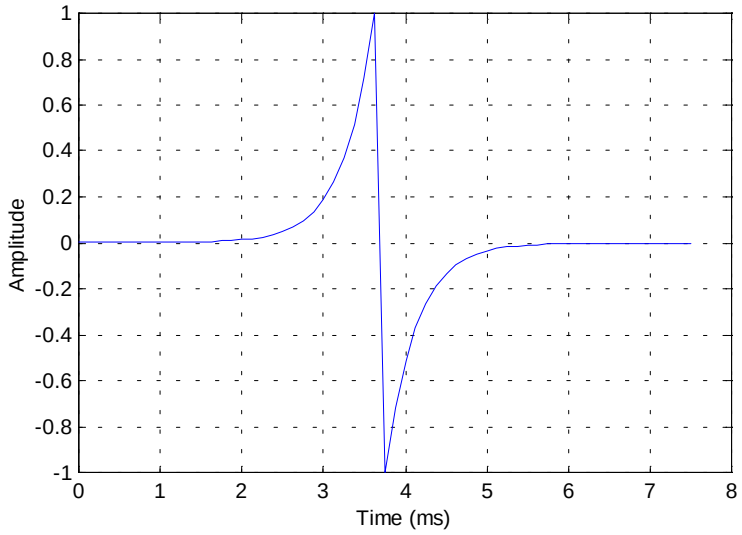



Figure 1: Impulsive noise in time domain

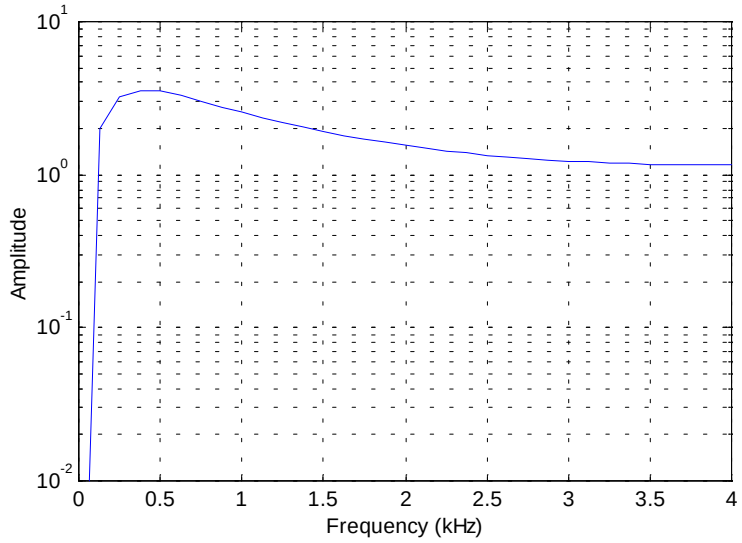



Figure 2: Impulsive noise in frequency domain

Figure 3 displays the time domain representation of the sentence “Back one message”. This message was professionally recorded and can be considered as noiseless. Figure 4 represents the same signal after addition of impulsive noise.

Figure 5 and Figure 6 are the corresponding spectrograms. The more a frequency is present in the signal, the darker it is represented. The impulsive noise appears very

clearly under the form of dark stripes. The information contained in the signal in a window where there is an impulse is completely unusable.

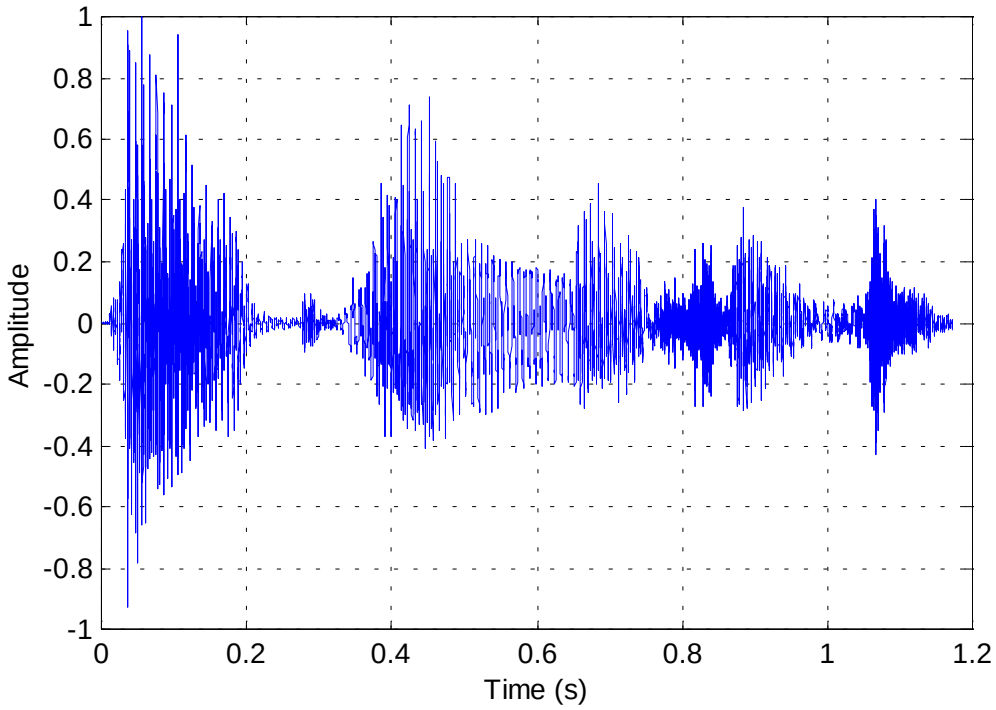



Figure 3: Clean signal in time domain

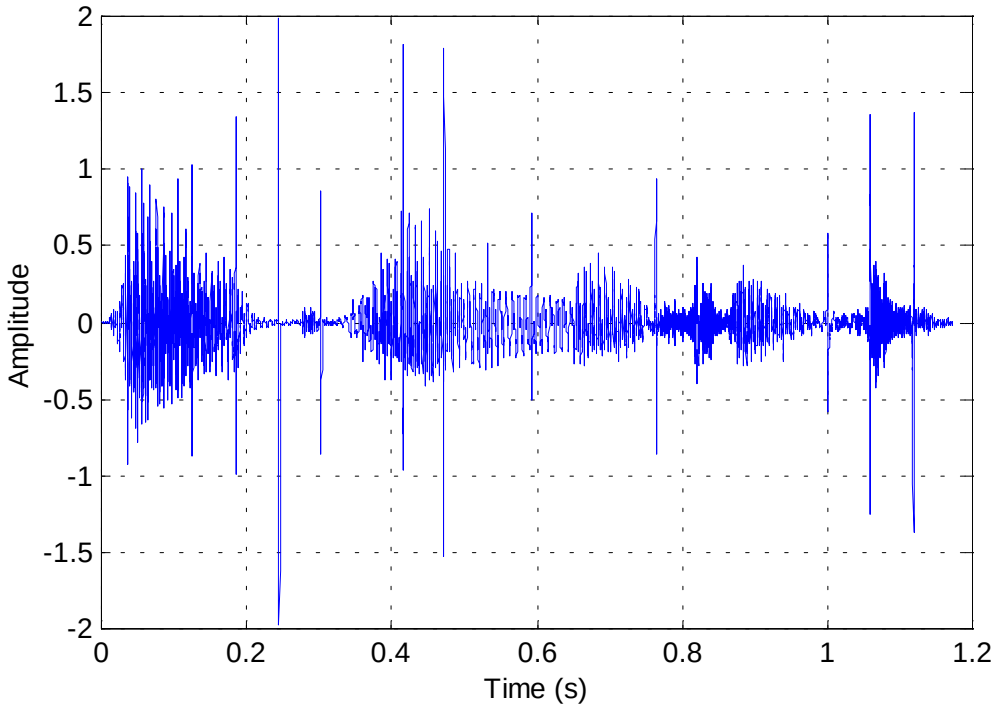



Figure 4: Noisy signal in time domain

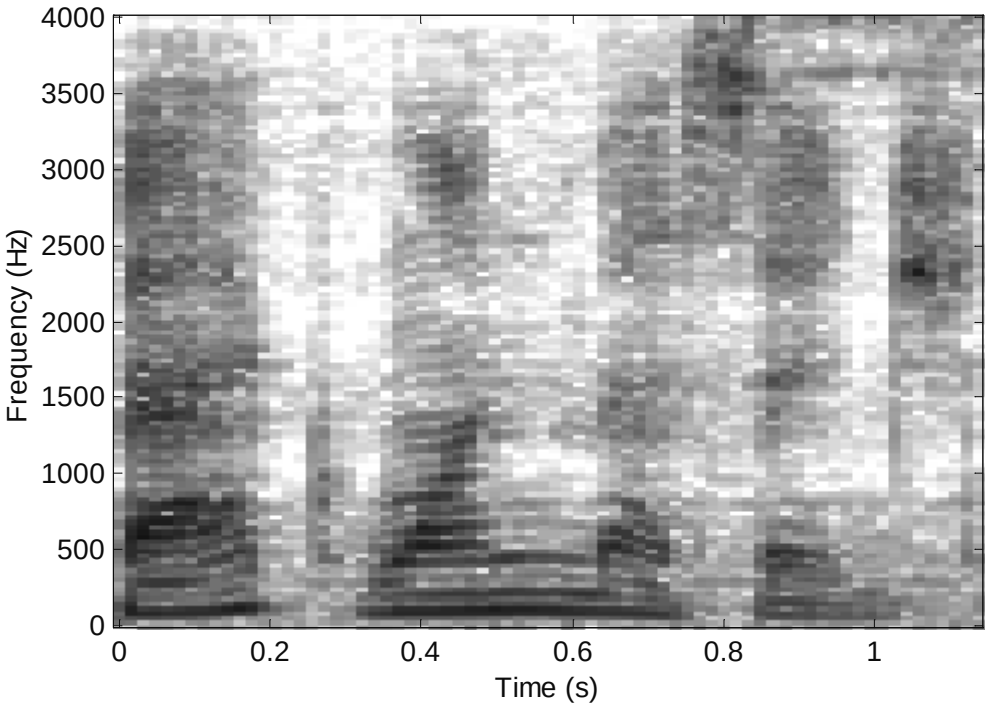



Figure 5: Clean signal in frequency domain

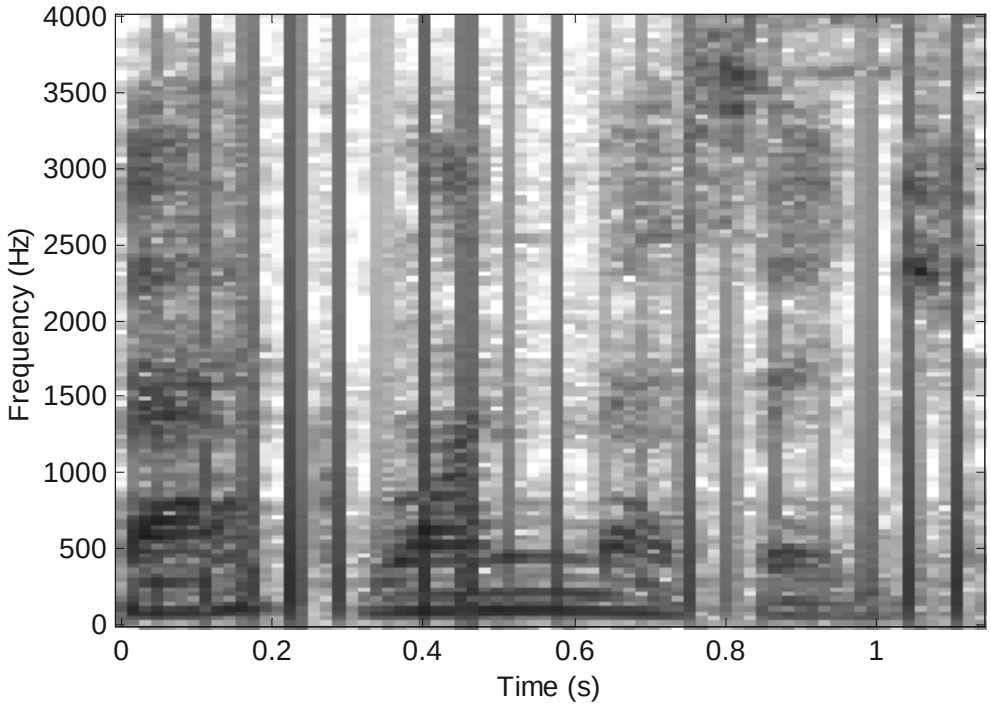



Figure 6: Noisy signal in frequency domain

## 2.3. Modeling of speech signals

### 2.3.1. Introduction

To be able to use digital signal processing techniques on speech signals, we need to convert the acoustic wave into a discrete signal. Therefore, the first step in the modeling is to sample and quantize the signal. This process is called a waveform coding and results in a waveform representation of the signal.

Several waveform coding techniques have been developed to decrease the number of bits necessary to code the signal while introducing only little distortion. They can be classified into 2 categories: time domain and frequency domain coders. In the time domain, the simplest technique is the well-known Pulse Code Modulation (PCM). Some other methods using differential modulation like Delta Modulation (DM) or Adaptive Differential Pulse Code Modulation (ADPCM) are more elaborate and give usually better results. In the frequency domain, we can basically distinguish Sub-Band Coding (SBC) and Adaptive Transform Coding (ATC).

Nevertheless, waveform representation does not decrease the number of bits required to store the signal in a very efficient way. To have better results, it is necessary to use specific properties of speech signals. Such coders are known as source coders. The latter source coders are based on a study of the mechanism by which speech is produced. As this very elaborate work requires a high knowledge of physics laws, we will only summarize the results that have been described in the literature. It appears that the modeling of a speech signal can be divided into 2 parts: the vocal tract model and the excitation model. For more information on this issue, the reader is referred to [30].

### 2.3.2. Vocal tract model

Speech production can be seen as a linear filtering. The filter consists of the vocal tract, composed by the pharynx, and the oral cavity. The input signal, or excitation, is created by the arrival of air coming from the lungs, which will be described in the next paragraph.

The vocal tract is modeled with the following transfer function:

$$V(z) = \frac{G}{1 - \sum_{k=1}^{m} a_k z^{-k}}, \tag{5}$$

where the parameters $a_k$ and the gain G are estimated by means of appropriate system identification methods. The order m of the model determines the precision of the results: a higher value will give us better results but require a higher memory storage [30].

To model the radiation produced by the lips, we need to pass the signal coming from V into another filter R given by

$$R(z) = R_0\left(1 - z^{-1}\right). \tag{6}$$

Here, the value of $R_0$ is the only unknown parameter to be determined.

### 2.3.3. Excitation model

We can distinguish 2 types of excitation, which results in 2 types of sounds: voiced and unvoiced sounds. Voiced sounds are the result of quasi-periodic pulses of air

while unvoiced sounds are produced by a turbulence that can be assimilated to a white noise excitation.

*Voiced sounds*

The excitation related to voiced sounds is basically a train of impulses spaced according to the pitch period of the signal. The shape of these pulses has been modeled as

$$g(n)=\begin{cases}\frac{1}{2}\left[1-\cos(\pi n/N_1)\right] & 0\le n\le N_1\\ \cos(\pi(n-N_1)/2N_2) & N_1\le n\le N_1+N_2\\ 0 & \textit{otherwise}\end{cases}. \qquad (7)$$

A representation of this function is given in Figure 7. The multiplication by a gain $A_v$ models the amplitude of this pulse.

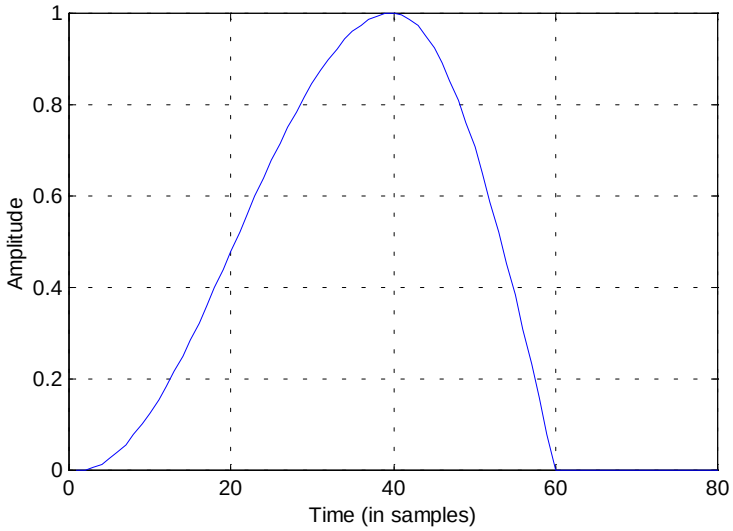



Figure 7: Glottal pulse shape

*Unvoiced sounds*

The excitation for unvoiced sounds may be modeled as a random noise generator multiplied by a gain $A_N$. The distribution of the noise does not matter very much as long as the spectrum is flat. In fact, a white Gaussian noise with zero mean and unit variance appears to be the most reasonable way to model the excitation of unvoiced sounds.

### 2.3.4. The final model

The excitation and vocal tract models are assembled to constitute the complete model of speech generation. The change from voiced to unvoiced excitation is realized with a switch as shown in Figure 8.

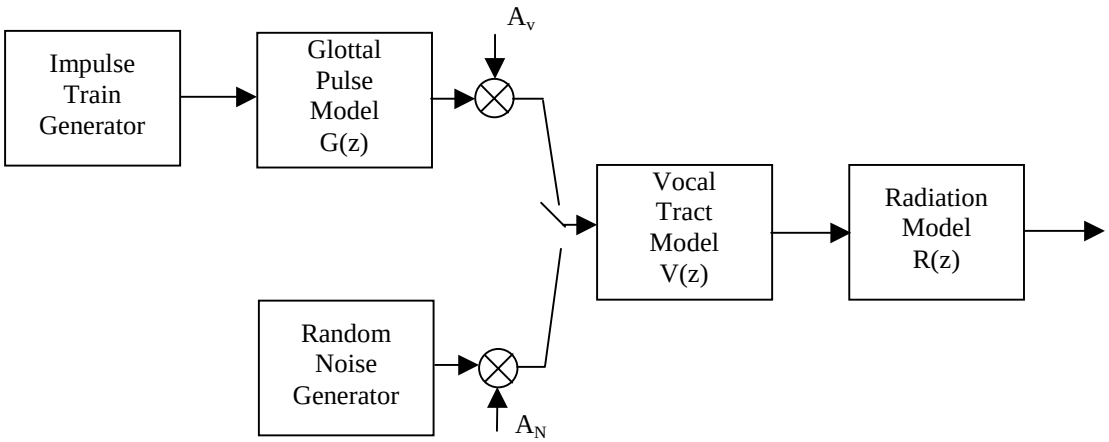



Figure 8: Complete model for speech production

## 2.4. Linear Predictive Coding (LPC)

### 2.4.1. Principle

The Linear Predictive Coding technique is probably the most common way to analyze speech signals because it can provide very good results with relative easy computations. It relies on the model that we derived in the preceding section and on parametric estimation theory to identify models in an optimal way. Different techniques have been developed that have various computational efficiencies. The most popular ones are the covariance and autocorrelation methods. Because we are not primarily interested in numerical robustness and efficiency, we will not present them in this report, but they can be found in [30]. For our simulations, we chose to use the so-called covariance method because being a Least Squares estimation, it will allow us to draw conclusions from a statistical viewpoint.

LPC is based on a simplified model of the vocal tract in that it neglects the influence of radiation, yielding a transfer function of the form

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{1 - \sum_{k=1}^{m} a_k z^{-k}}. \quad (8)$$

Here S and U are the z-transforms of the signal and of the excitation, respectively. The LPC method consists in estimating the coefficients $a_k$, the gain G, and the excitation U.

### 2.4.2. Estimation of the parameters of the model

In the time domain, the transfer function reduces to

$$s(n) = \sum_{k=1}^{m} a_k s(n-k) + G u(n). \quad (9)$$

To estimate the parameters $a_k$, the excitation will be neglected, which results in the following equation:

$$\tilde{s}(n) = \sum_{k=1}^{m} a_k s(n-k). \quad (10)$$

A justification of this approximation is that it accounts for all the correlation in the signal. Therefore, we do not need to consider the influence of the excitation to estimate the parameters of the model.

Because speech is a non-stationary process, the coefficients ought to be evaluated over small periods of the signal, called windows. If N denotes the length of a window, the system that has to be solved can be written in a matrix form given by

$$\begin{bmatrix} s(m+1) \\ s(m+2) \\ \cdots \\ s(N-1) \\ s(N) \end{bmatrix} = \begin{bmatrix} s(m) & s(m-1) & \cdots & s(2) & s(1) \\ s(m+1) & s(m) & \cdots & s(3) & s(2) \\ \cdots & & & & \cdots \\ s(N-2) & s(N-3) & \cdots & s(N-m) & s(N-m-1) \\ s(N-1) & s(N-2) & \cdots & s(N-m+1) & s(N-m) \end{bmatrix} \cdot \begin{bmatrix} a_m \\ a_{m-1} \\ \cdots \\ a_2 \\ a_1 \end{bmatrix} + \begin{bmatrix} e(m+1) \\ e(m+2) \\ \cdots \\ e(N-1) \\ e(N) \end{bmatrix}. \quad (11)$$

where e(n) is the error between the real and the estimated values, that is

$$e(n) = s(n) - \tilde{s}(n). \quad (12)$$

The solution of these equations is found by minimizing the sum $E_k$ of the square of the errors on each window, which is written as

$$E_k = \sum_{n=m+1}^{N} e^2(n). \qquad (13)$$

This provides a Least Squares estimation of the parameters. We will see in Chapter 3 that it is an optimal method in the case of errors following a Gaussian distribution.

The length N of the window is an important parameter since it greatly affects the accuracy of the estimates. It should correspond to a time interval over which the signal can be considered as stationary. Because of the great variety of speech signals, it is not easy to determine the optimal value for N. Nevertheless, it has been shown that a duration of 20 to 40 ms was appropriate, 30 ms being a typical value. If the signal is sampled at a rate of 8 kHz, N will be equal to 240 samples.

Another issue is related to the type of weighting function that needs to be used to isolate a segment of the signal. Applying a basic rectangular window seems not to be appropriate because the sudden change from zeros to signal values introduces distortions in the frequency domain. Therefore, it has been recommended to use a Hamming window, which consists in multiplying the signal by the following function:

$$w(n) = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{N}\right) & 0 \le n \le N-1 \\ 0 & \text{otherwise} \end{cases}, \qquad (14)$$

whose graph is depicted in Figure 9. This segmentation requires windows to overlap, so that the influence of the extremities is decreased. A typical value for this overlap is 10 ms.

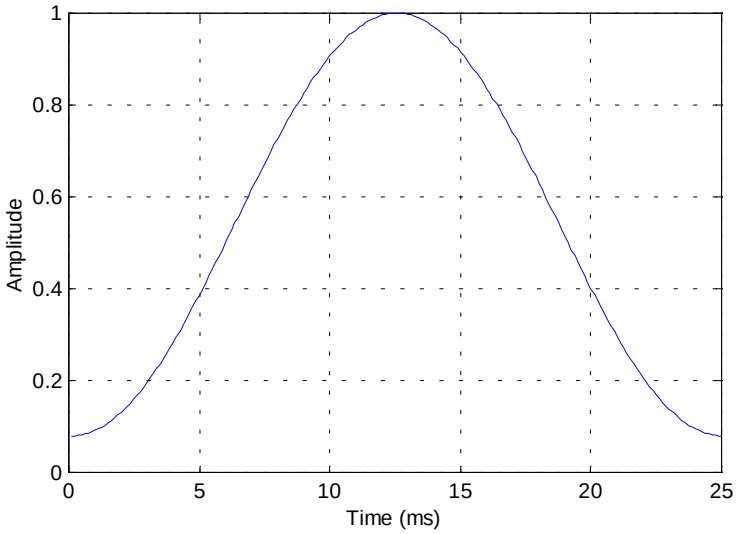



Figure 9: Shape of a Hamming window

### 2.4.3. Estimation of the excitation

The method we used to filter impulsive noise does not require an estimation of the excitation. Nevertheless, part of this process, which includes pitch estimation, allows us to study properties of a speech signal. Therefore, some basic techniques held our attention, which will be presented next.

The estimation of the excitation is made of several steps. First, because there are two types of excitations, we need to distinguish voiced from unvoiced parts of the signal. One possibility would be to study the short-time autocorrelation function of the signal over successive windows. Indeed, this is a way to underline the differences between these two types of speech signals according to frequency and periodicity properties.

For a voiced signal, the autocorrelation function clearly reveals the presence of two main frequencies as shown in Figure 10. The highest periodicity of the signal corresponds to the pitch frequency, or fundamental frequency of the signal, and is in the order of 100 Hz. In other words, the pitch period is equal to the time delay between the higher peaks of the function. The other frequency is in the range of 1 kHz and corresponds to the voiced feature of the signal.

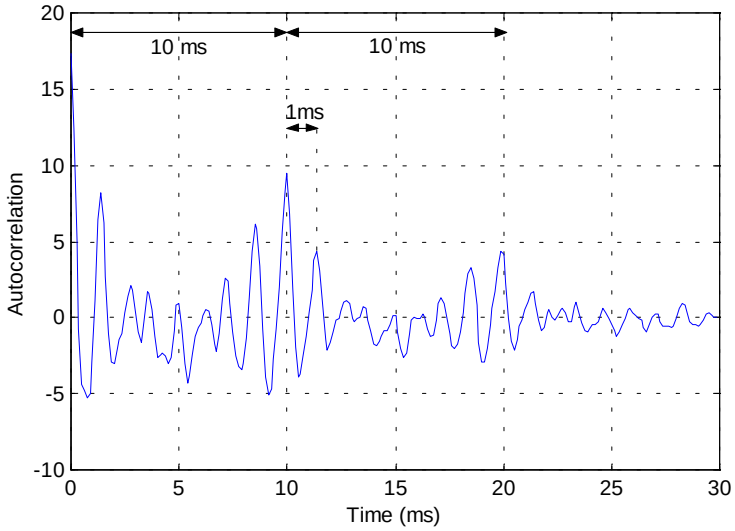



Figure 10: Autocorrelation for a voiced speech

Figure 11 depicts the shape of the autocorrelation function calculated over an unvoiced segment of a speech signal. As seen in that figure, an unvoiced signal is characterized by a main frequency in the order of 3 kHz. The variations of the autocorrelation function are therefore much larger than those of a voiced speech. Moreover, unvoiced signals have no periodic feature, so that no peak is distinguishable.

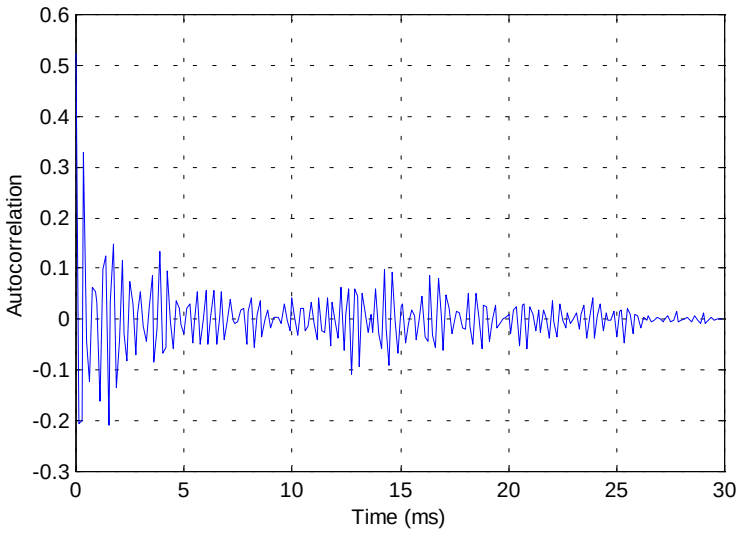



Figure 11: Autocorrelation for an unvoiced speech

Once the distinction between voiced and unvoiced sounds has been made, the gain and the pitch period have to be found. We implicitly saw a way to determine the pitch period using the autocorrelation function. An accurate method requires the application of some preprocessing techniques that make peaks easier to detect. Even with such precautions, the evaluation is not always very good. Therefore, when values of the pitch period have been found for each window, a smoothing of the obtained curve has to be carried out.

#### 2.4.4. Estimation of the gain

The computation of the gain is a much easier problem than that of the excitation. It is based on the assumption that for every window, the energy in the error signal e(n) is equal to the energy of the excitation, which can be written as

$$E_k = \sum_{n=m+1}^{N} e^2(n) = G_k{}^2 \sum_{n=m+1}^{N} u^2(n). \qquad (15)$$

The function u(n) denotes a Dirac impulse for voiced speech and a white noise with zero mean and unity variance for unvoiced speech. Then, it can be shown in both cases that the gain $G_k$ is equal to the square root of $E_k$, that is,

$$G_k = \sqrt{E_k}\,. \qquad (16)$$

### 2.5. Speech enhancement techniques in presence of impulsive noise

#### 2.5.1. Introduction

In communication systems, the presence of noise is unavoidable. This is why the problem of speech enhancement has been studied for many years. Different approaches have been proposed dealing with various properties of speech signals and with various types of noise [20].

The problem of removing impulsive noise stems from the fact that most of the filtering theories are based on the hypothesis that noise is mostly Gaussian. A literature review will be given in this section. Also, the issue of quantification of speech

enhancement will be considered and the most common measures of speech quality and speech intelligibility will be reviewed.

### 2.5.2. Existing techniques to remove impulsive noise

*Median filters*

Median filters are the most widely used methods to filter impulsive noise [7, 37]. A median filter consists in passing a window along the signal and replacing the point located in the center by the median of all the points in the window. Because we know that the sample median is a robust estimator, we understand easily that it is an efficient way to reject outliers induced by impulsive noise. However, replacing every point by the median of its neighbors tends to decrease the quality of a speech signal. Another problem is that impulsive noise can corrupt 20 or even 40 consecutive samples, which means that the size of the window has to reach values up to 80. This is a too large window to keep a speech signal of good quality.

*Detection and interpolation*

Let us describe the method proposed in [37, 38] for detection and interpolation. The first step consists in detecting the samples of the signal corrupted by impulsive noise. It uses the LPC method to model the signal, and uses the fact that this modeling will be good only if there is no impulsive noise. After having defined a threshold, the decision will be taken to keep the value of the signal or replace it. Values that have to be changed can be interpolated. This step can be realized thanks to several methods. One of them consists in a two-sided predictor, which uses both past and future predictions to give accurate results.

The method we developed uses the same idea: we will first detect impulsive noise and then estimate values to be replaced. However, we will use robust statistics in both detection and interpolation steps. Indeed, impulsive noise can be seen as leverage points in the observation matrix used by the LPC method, and robust estimators will enable us to detect them and estimate new values without being influenced by them.

*Robust LPC method*

We saw that LPC is nothing else than a Least Squares estimation of the parameters modeling a speech signal. Instead of solving the regression using Least Squares, one can use a robust M-estimator. This would give us a robust estimation of the parameters, which can then be used to build a Kalman or Wiener filter. This method has been tried and leads to satisfactory results. In [16], Least Squares are replaced by the Lp-norm. In [18], the Huber M-estimator and then a robust version of Kalman filter is used. However, because the observation matrix considered in this regression contains values of the signal, it is obviously corrupted by the impulsive noise. We will see in Chapter 4 that M-estimators are not robust against leverage points. Therefore this solution will not be advocated. Instead, we make use of the Schweppe-type Huber GM-estimator, which can cope with bad leverage points.

### 2.5.3. Speech quality and speech intelligibility evaluation

Several objective measures of speech quality have been defined [6], [29]. Some of them quantify the waveform distortion while others are a measure of spectral distortion. The judicious choice of one or the other measure depends on the type of the communication system that is studied. Time domain measures are used more often because they are easier to compute. Nevertheless, it seems that spectral measures are generally more conform to the result given by subjective listening tests. In the time domain, a common measure of quality is the Signal to Noise Ratio, and an improved version of it is called Segmental Signal to Noise Ratio. In the frequency domain, we can use Spectral Distances or Likelihood Ratio. Some subjective tests can also be applied, the most popular one being the Mean Opinion Score. As far intelligibility is concerned, subjective tests like the Diagnostic Rhyme Test are the only resource.

The case of impulsive noise is quite particular. Indeed, it is more a destruction of some parts of the signal than an addition of noise. Therefore, objective measures of the quality are not very well suited. For example, if we calculate the Signal to Noise Ratio after the removal of impulsive noise, the improvement is always significant because of the large difference between the estimates, which are close to the normal signal amplitude, and the impulses. For this reason, we will base our analysis on two other tests,

namely, subjective listening tests and a graphical comparison between temporary and frequency representations of the clean, the noisy and the filtered signal.

### 2.6. Conclusions

This Chapter presented basic methods of speech processing that will be useful in the development of our impulsive noise filtering method. First, we exposed the parametric representation of speech and the principle of Linear Prediction Coding. Then, we focused on the specific situation where we have additive impulsive noise. Then, we explained what are the sources of such a noise and we saw how it can be modeled as a stochastic process. Finally, we reviewed various methods that can be used to suppress impulses and to quantify speech quality and intelligibility. We also indicated which ones will be used in our study.

# Chapter 3

# Parametric and robust estimation

## 3.1. Introduction

The purpose of our study is to investigate the use of robust statistics in detecting and removing impulsive noise in speech signals. Therefore, a brief review of this approach is provided in this Chapter. First, the main concepts related to classical parametric statistics are presented and their main properties described. Then, the class of Maximum-Likelihood estimators is introduced. The interesting feature of these estimators is that each of them is optimal at a given probability density function of the noise. Then, the robustness approach is summarized [12, 14]. Its purpose is to develop estimators that are reliable even if the probability distribution of the noise is not exactly known and may have long tails. Finally, the class of robust estimators called M-estimators is defined. The emphasis is placed on the Huber estimator, which plays a key role in our study.

## 3.2. What is a “good” estimator?

At a given probability distribution, the parametric estimation theory defines some concepts that give an answer to the question: what is a good estimator at that distribution?

Firstly, a good estimator should converge towards the true value of the parameter to be estimated when the number of measurements increases to infinity. This property is called Fisher consistency. Secondly, an estimator should be unbiased, that is, its mean value should be equal to the true value for any sample size. Thirdly, a good estimator should have a fast rate of convergence towards the true value. Fourthly, the variance of the estimates should be in the vicinity of the lower bound. When the lower bound is attained, the estimator is said to be efficient.

### 3.2.1. Fisher consistency

Given a sample $\underline{z} = \{z_1, z_2, \ldots, z_m\}$ of m observations following a distribution with a probability density function $f$ and a probability cumulative function F, we can define the empirical probability density and cumulative functions $f_m$ and $F_m$ as

$$f_m(z) = \frac{1}{m}\sum_{i=1}^{m}\delta(z - z_i) \text{ and } F_m(z) = \frac{1}{m}\sum_{i=1}^{m}\Delta(z - z_i), \qquad (17)$$

where $\delta$ is the Dirac function and $\Delta$ the unit step function. These 2 functions are illustrated in Figure 12 and Figure 13, respectively, where a sample of size 10 drawn from a Gaussian distribution with zero mean and variance equal to 4 has been considered. By virtue of the Glivenko-Cantelli theorem [31], we have

$$f_m \to f \quad and \quad F_m \to F \quad as \quad m \to \infty .$$

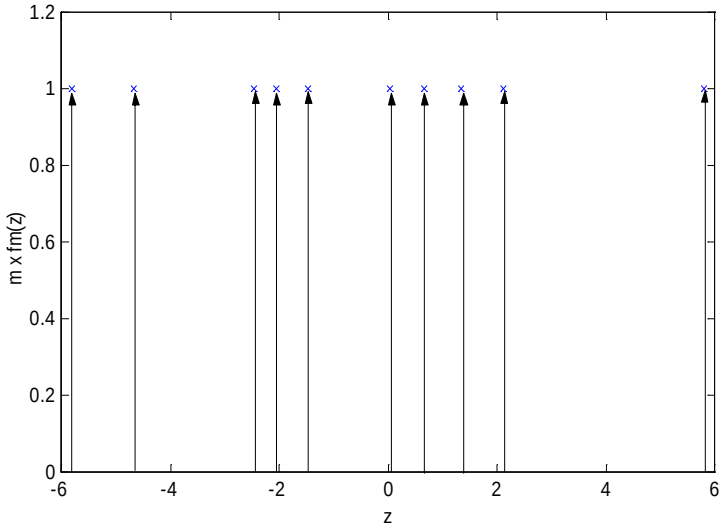



Figure 12: Empirical probability density function

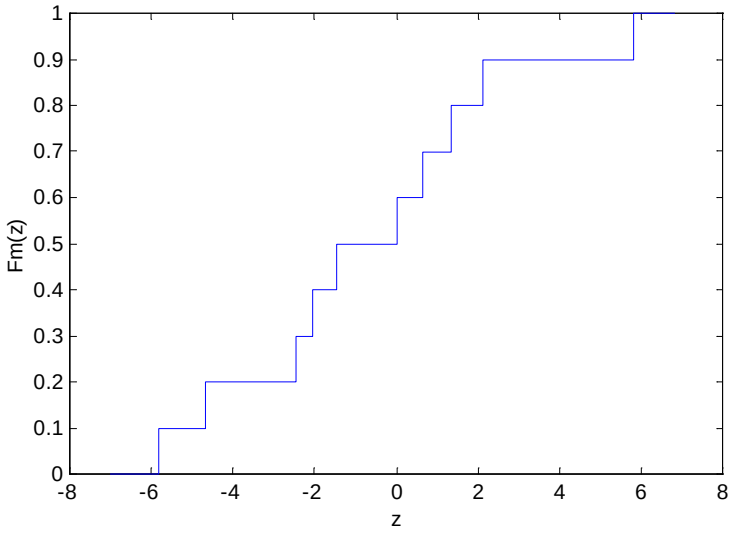



Figure 13: Empirical probability cumulative function

An estimator is said to be Fisher consistent if the estimator viewed as a function of $F_m$ converges to the true value, $\theta_t$, as m tends to infinity. Formally,

$$\lim_{m\to\infty} \hat{\theta}(F_m) = \hat{\theta}(F) = \theta_t . \qquad (18)$$

It can be shown that the sample mean and the sample median are both Fisher consistent if the true value is defined respectively as the mean and the median of the distribution.

### 3.2.2. Unbiasedness

Given a sample of m observations, an estimator $\hat{\theta}_m$ is said to be unbiased if

$$\forall\, m, \quad E\left[\hat{\theta}_m\right] = \theta_t . \qquad (19)$$

When an infinite number of samples of size m is drawn and an estimate is calculated for each of them, the mean of all these estimates is equal to the true value for an unbiased estimator. The sample mean and the sample median satisfy this property.

### 3.2.3. Rate of convergence

The rate of convergence of an estimator is the rate at which its variance tends to zero as the sample size m tends to infinity. For the sample mean and the sample median, it is equal to $1/\sqrt{\mathrm{m}}$. This is the so-called normal rate of convergence. We define the normalized estimator as $\sqrt{m}\hat{\theta}_m$ and its variance as the normalized variance.

### 3.2.4. Efficiency

The efficiency is based on the concept of Fisher information, which is defined as

$$I_f = E\left[\left(\frac{\partial \ln f(z,\theta)}{\partial \theta}\right)^2\right], \tag{20}$$

where *f* is the probability density function of the distribution.

Let us consider an unbiased estimator $\hat{\theta}_m$. Cramer and Rao [19] established the following inequality:

$$\mathrm{var}\left(\sqrt{m}\hat{\theta}_m\right) \geq \frac{1}{I_f}. \tag{21}$$

Based on this inequality, the efficiency of an estimator is defined as

$$e_m = \frac{1/I_f}{\mathrm{var}\left(\sqrt{m}\hat{\theta}_m\right)}. \tag{22}$$

An estimator is said to be efficient if $\mathrm{e_m}$ is equal to 1, that is to say, if its normalized variance reaches the lower Cramer-Rao bound. For example, it can be shown that the sample mean is efficient at the Gaussian distribution for all m.

We can also define the asymptotic relative efficiency of an estimator $\hat{\theta}'_{\mathrm{m}}$ with respect to another estimator $\hat{\theta}''_{\mathrm{m}}$. It is given by

$$\mathrm{e}_{\hat{\theta}'_{\mathrm{m}},\hat{\theta}''_{\mathrm{m}}} = \lim_{\mathrm{m}\to\infty} \frac{var\left(\sqrt{\mathrm{m}}\hat{\theta}''_{\mathrm{m}}\right)}{var\left(\sqrt{\mathrm{m}}\hat{\theta}'_{\mathrm{m}}\right)}. \tag{23}$$

For example, at the Gaussian distribution, the sample median is 36.3% less efficient than the sample mean while at the Laplacian distribution, the sample mean is 50% less efficient than the sample median.

We will see in the next section that for a Gaussian mixture, the sample median becomes very quickly more efficient than the sample mean when the tails of the distribution becomes longer.

### 3.3. Maximum Likelihood Estimators

#### 3.3.1. Definition

Consider a sample $\underline{z} = \{z_1, z_2, \ldots, z_m\}$ of m observations that are independent and identically distributed according to a probability distribution F with probability density $f(z;\theta)$. The likelihood function $L(\theta;\underline{z})$ is defined as

$$L(\theta;\underline{z}) = c \cdot f(\underline{z};\theta), \tag{24}$$

where c is a constant and $f(\underline{z};\theta)$ is viewed as a function of $\theta$ given $\underline{z}$. Because the value of c does not affect the value of $\theta$ at which L is maximum, it is usually put equal to one. For convenience, we transform the problem of maximizing $L(\theta;\underline{z}) = f(\underline{z};\theta)$ into the problem of minimizing $-\ln(f(\underline{z};\theta))$. This transformation is targeted toward the very prevalent exponential family, which includes the Gaussian and the Laplacian distributions. If we introduce the residuals $r_i = z_i - \theta$, our problem can finally be stated as follows:

$$\min_\theta J(\theta) = \sum_{i=1}^{m} -\ln f(r_i). \tag{25}$$

Called an objective function, the function J defines a large class of estimators known as Maximum Likelihood (ML) estimators [19].

Let us emphasize again that there is no ideal estimator that can be used in every case and still give the best possible result. An ML-estimator is optimal at the associated probability distribution and several ML-estimators have been derived for the most common distributions. For example, the Least Squares estimator is the ML-estimator at the Gaussian distribution while the L1-norm is the ML-estimator at the Laplacian distribution. In the one dimensional case of estimating a parameter of location, the Least Squares reduces to the sample mean and the L1-norm to the sample median. Therefore, the sample mean and the sample median are also ML-estimators.

The rate of convergence of the Least Squares and of the L1-norm is in the order of $1/\sqrt{m}$, where m is the sample size. In addition, the Least Squares estimator is efficient at the Gaussian distribution while the L1-norm is asymptotically efficient at the Laplacian distribution.

### 3.3.2. An example of a Gaussian mixture

Let us consider a Gaussian mixture given by $G=(1-\varepsilon)N(0,1)+\varepsilon\, N(0,25)$. When $\varepsilon$ is equal to 0, the most efficient estimator is the sample mean. But as $\varepsilon$ becomes larger, the tails of the distribution increase and the sample median becomes a better estimator. In this example, it can be shown that the asymptotic efficiency of the sample mean is given by

$$\mathrm{var}(\bar{z},G)=1+24\varepsilon\,, \tag{26}$$

while the asymptotic variance of the sample median is written as

$$\mathrm{var}(z_{med},G)=\frac{25\pi}{2(5-4\varepsilon)^2}\,. \tag{27}$$

Therefore, the relative efficiency e of the sample median with respect to the sample mean is equal to

$$e=\frac{(1+24\varepsilon)(5-4\varepsilon)^2}{25\pi}\,, \tag{28}$$

whose graph is plotted in Figure 14.

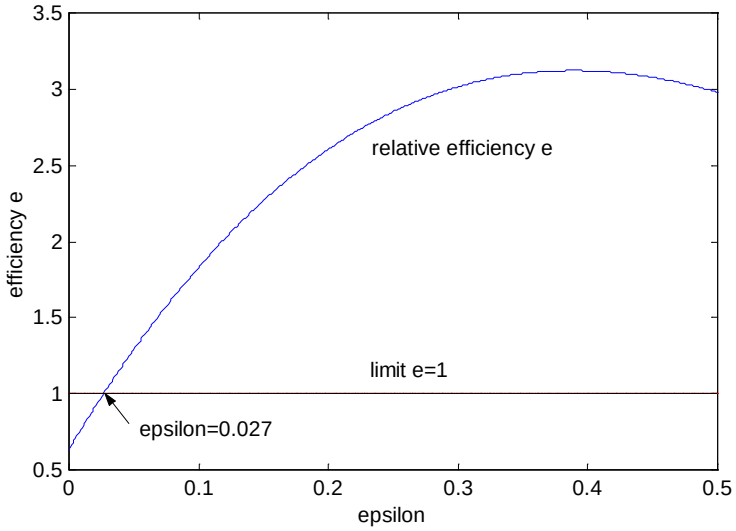



Figure 14: Relative efficiency of the sample median
with respect to the sample median

It is found that e becomes greater than 1 when $\varepsilon$ is larger than 2.7%, which indicates that a tiny increase of the tails leads to a higher efficiency of the sample median with respect to the sample mean. This result is all the more interesting that many real data are closer to a Gaussian mixture than to a perfect Gaussian distribution. Besides, the difference between the Gaussian distribution N(0,1) and the Gaussian mixture G with $\varepsilon = 2.7\%$ is almost unnoticeable. In Figure 15, a plot of the probability density functions of these two distributions illustrates this assertion.

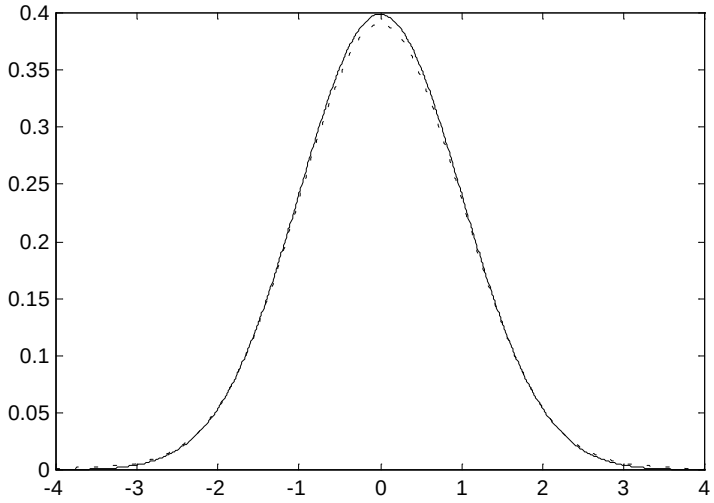



Figure 15: Shape of a Gaussian probability density function (solid line)
and of a Gaussian mixture with $\varepsilon = 0.027$ (dotted line)

### 3.4. Robustness concept

#### 3.4.1. Introduction

The parametric estimation theory defines concepts that allow the statistician to compare the performance of various estimators for a broad range of distributions. But these concepts are based on the assumption that the distribution of the data is known and that no observation deviates from this model. On the other hand, robust statistics aim to extend this theory to cases where the assumptions are slightly violated. We can distinguish three aspects of robustness related to the type of perturbation that is considered, namely, qualitative, global and local robustness.

#### 3.4.2. Qualitative robustness

Qualitative robustness means that small perturbations in the assumptions have small effects on the estimate. Specifically, let us consider a sample that follows a distribution G. Let us suppose that this distribution is not exactly known and is therefore approximated by F. Let us call $L_G$ the distribution of the estimator when the sample follows exactly G, and $L_F$ the distribution of the estimator when the sample follows F.

The estimator is said to be qualitatively robust at F if

$$\forall \delta > 0, \quad \exists \varepsilon > 0 \quad \textit{so that} \quad \forall\, m, \quad d(F,G) < \varepsilon \Rightarrow d(L_F, L_G) < \delta\,, \qquad (29)$$

where $\mathrm{d}(\cdot)$ is a measure of the distance between two distribution functions. One of the most commonly used metric $\mathrm{d}(\cdot)$ is the Kolgomorov distance defined as

$$d(F,G) = \sup\left(\left|F(r) - G(r)\right|\right). \qquad (30)$$

From a practical viewpoint, we will be more interested in global and local robustness.

### 3.4.3. Global robustness

Global robustness aims to evaluate the largest perturbation that an estimator can handle. It is related to the concepts of maximum bias and breakdown point.

*Maximum bias curve*

Let us suppose that the true distribution G followed by the observations is a mixture of the assumed distribution F and an unknown distribution H, yielding

$$G = (1-\varepsilon)F + \varepsilon H\,. \qquad (31)$$

This $\varepsilon$-contaminated distribution defines a neighborhood $N_\varepsilon(F)$ of F. Let $\hat{\theta}(F)$ be an estimator based on a sample of infinite size drawn from the distribution F and let $\hat{\theta}(G)$ be that estimator based on the distribution G. The asymptotic maximum bias is then given by

$$b_{\max}(\varepsilon) = \sup_H \left|\hat{\theta}(F) - \hat{\theta}(G)\right|. \qquad (32)$$

According to this definition, we have to consider, for a given $\varepsilon$, the distribution H that will have the worst effect on the estimator. The sample mean is not robust since a single outlier can carry its bias to infinity. On the other hand, Huber showed that the sample median is a minimax estimator from a bias viewpoint in that at any distribution, it minimizes the maximum bias curve over all location equivariant estimators.

*Breakdown point*

In the location case, the breakdown point corresponds to the maximum value of $\varepsilon$ for which the estimator has a finite maximum bias, that is,

$$\varepsilon^* = \max\{\varepsilon;\ b_{\max}(\varepsilon)\ finite\}. \tag{33}$$

The highest possible value for $\varepsilon^*$ is $\frac{1}{2}$, and is reached by the sample median. In the finite case, the breakdown point is defined as

$$\varepsilon^* = \max\left\{\varepsilon = \frac{v}{m};\ b_{\max}(\varepsilon)\ finite\right\} \tag{34}$$

where $v$ is the number of outliers among the m observations.

For estimators of location that are location equivariant, $\varepsilon^*$ is no larger than $\left[\frac{m-1}{2}\right]\Big/m$, that is,

$$\varepsilon^* \le \left[\frac{m-1}{2}\right]\Big/m, \tag{35}$$

where [ x ] denotes the integer part of the real number x.

There are two ways that an estimator of scale may break down, namely, by explosion or by implosion. A scale estimator explodes when it becomes infinitely large and implodes when it reduces to zero. Hence, we may define the breakdown point of a scale estimator as the maximum fraction of outliers that can bring the estimator neither to infinity nor to zero. An estimator is said to be non-robust if $\varepsilon^* = 0$. Then, the higher the value of $\varepsilon^*$, the more robust the estimator.

### 3.4.4. Local robustness

A third aspect of robustness consists in assessing the effects of infinitesimal perturbations on the bias and the variance of an estimator, which are measured by means of the influence function and the change-of-variance function, respectively.

*Influence function*

Let us consider a sample $\{z_1, z_2, \ldots, z_{m-1}, z\}$ of m observations, where $\{z_1, z_2, \ldots, z_{m-1}\}$ follow a distribution F and z takes any value on the real line. In the finite sample case, the influence function is defined as

$$IF_m(z, F) = m \cdot \left(\hat{\theta}_m(z_1, \ldots, z_{m-1}, z) - \hat{\theta}_{m-1}(z_1, \ldots, z_{m-1})\right). \quad (36)$$

It is therefore a measure of the effect of a single outlier on the estimate. In the asymptotic case, it is expressed as

$$IF(z, F) = \lim_{\varepsilon \to 0} IF_m(z, F) = \left.\frac{\partial \hat{\theta}(G)}{\partial \varepsilon}\right|_{\varepsilon = 0}, \quad (37)$$

where

$$G = (1 - \varepsilon)F + \varepsilon \Delta z, \quad and\ \varepsilon = \frac{1}{m}. \quad (38)$$

In the next paragraph, we will derive the influence function of the M-estimators of location.

*Gross-error sensitivity*

The gross-error sensitivity of an estimator at F is defined as

$$\gamma^* = \sup_z \left|IF(z, F)\right| \quad (39)$$

If $\gamma^*$ is finite, that is, if the influence function is bounded, then the estimator is said to be B-robust, where "B" stands for bias. Interestingly, $\gamma^*$ is the slope of the maximum bias curve at $\varepsilon = 0$. Hampel et al. [12] showed that the sample median is the most B-robust estimator of location in the sense that it has the smallest possible $\gamma^*$.

### 3.5. M-estimators

The class of M-estimators is an extension of that of ML-estimators. The "M" stands for "Maximum Likelihood type". An M-estimator is defined as a minimizer of an objective function J expressed as

$$J(\theta) = \sum_{i=1}^{m} \rho\left(\frac{r_i}{\sigma}\right), \quad (40)$$

where $\rho$ is a function of the residuals, $r_i$, and $\sigma$ is a robust residual scale. Typically, $\sigma$ is chosen equal to the so-called Median Absolute Deviation from the median (MAD) defined as

$$MAD = 1.4826\ median_{\,i}\ \left|\ r_i - median_{\,j}\left(r_j\right)\right|. \tag{41}$$

Without loss of generality, let us assume that $\sigma$ =1 in the sequel. From (40), it appears that an M-estimator is an ML-estimator if and only if there exists a probability density function *f* such that

$$\rho(r) = -\ln\left(f(r)\right). \tag{42}$$

An interesting subclass of M-estimators is the one defined by convex $\rho$ functions. It includes the Least Squares and the L1-norm estimators with $\rho(r) = \frac{r^2}{2}$ and $\rho(r) = |r|$, respectively.

Huber proposed an estimator that is equal to the Least Squares when the residuals are small and to the L1-norm when they are large. Formally, we have

$$\rho(r) = \begin{cases} \frac{r^2}{2} & for\ |r| \le b \\ b|r| - \frac{b^2}{2} & for\ |r| > b \end{cases}. \tag{43}$$

The intent here is to combine the efficiency of the Least Squares at the Gaussian distribution with the robustness of the L1-norm in a single estimation method.

An M-estimator is robust if it has a bounded influence function. In the location case, it is given by

$$\mathrm{IF(r, F)} = \frac{\psi(\mathrm{r})}{\mathrm{E}[\psi'(\mathrm{r})]} \tag{44}$$

where

$$\psi(\mathrm{r}) = \frac{\mathrm{d}\rho(\mathrm{r})}{\mathrm{dr}} \tag{45}$$

and

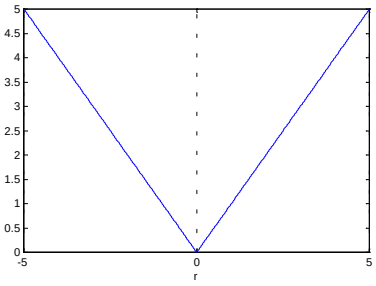
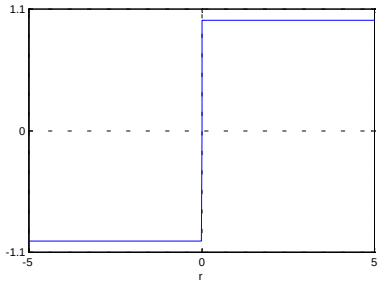
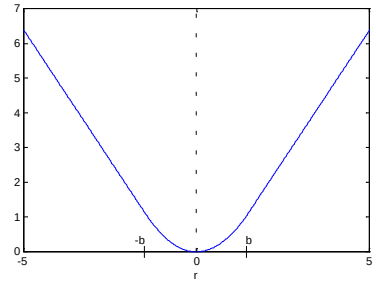
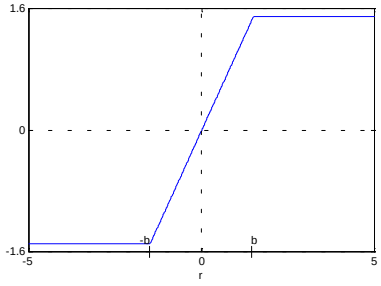
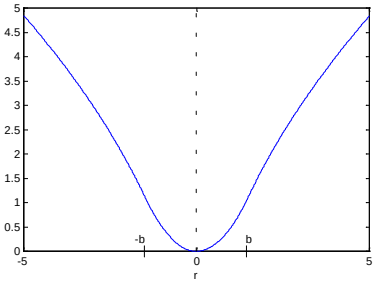
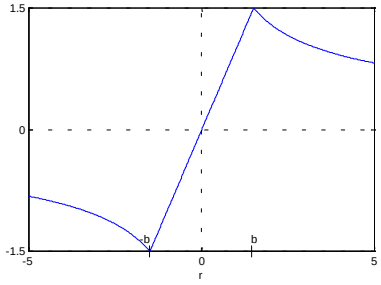
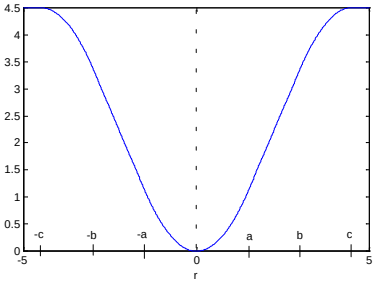
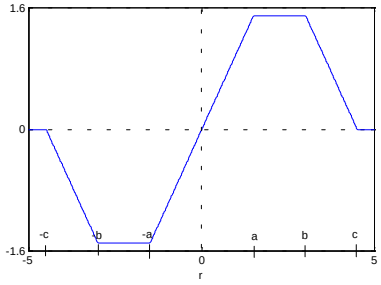
$$\psi'(r) = \frac{d\psi(r)}{dr}. \tag{46}$$

Table 1 and Table 2 provide the equation and the graph of $\rho$ and $\psi$ for the most widely used M-estimators, respectively.

Table 1: Mathematical form of $\rho$ and $\psi$ for the most widely used M-estimators

| Estimator | Domain | ρ (r) | ψ (r) |
|---|---|---|---|
| L1-norm | $\Re$ | $\lvert r\rvert$ | sign (r) |
| Huber | $\lvert r\rvert \le b$ | $\frac{1}{2}r^2$ | r |
| | $\lvert r\rvert > b$ | $b\lvert r\rvert - \frac{b^2}{2}$ | b sign (r) |
| Merrill-Schweppe | $\lvert r\rvert \le b$ | $\frac{1}{2}r^2$ | r |
| | $\lvert r\rvert > b$ | $2b^{3/2}\sqrt{\lvert r\rvert} - \frac{3}{2}b^2$ | $b^{3/2}\frac{\text{sign}(r)}{\sqrt{\lvert r\rvert}}$ |
| Hampel | $\lvert r\rvert \le a$ | $\frac{1}{2}r^2$ | r |
| | $a < \lvert r\rvert \le b$ | $a\lvert r\rvert - \frac{a^2}{2}$ | a sign (r) |
| | $b < \lvert r\rvert \le c$ | $a\frac{c\lvert r\rvert - \frac{1}{2}r^2}{c-b} - \frac{a^2}{2} - \frac{ab^2}{2(c-b)}$ | $a\frac{c-\lvert r\rvert}{c-b}\text{sign}(r)$ |
| | $\lvert r\rvert > c$ | $\frac{a}{2}(c+b-a)$ | 0 |

Table 2: Graph of $\rho$ and $\psi$ for the most widely used M-estimators

| Estimator | ρ (r) | ψ (r) |
|---|---|---|
| L1-norm |  |  |
| Huber |  |  |
| Merrill-Schweppe |  |  |
| Hampel |  |  |

### 3.6. Conclusions

This Chapter gave an overview of classical and robust parametric estimation theories. From the classical viewpoint, we came to the notion of optimality of an estimator: the optimal estimator at a given distribution is the one that has the highest efficiency. But it was shown that optimal estimators like the Least Squares can have a very unreliable behavior if the distribution of the observations slightly deviates from the ideal model. To deal with this problem, robust statistics were proposed. The Huber M-estimator can be seen as a good compromise in case of a distribution that is mostly Gaussian: it combines the efficiency of the Least Squares with the robustness of the L1-norm.

# Chapter 4

# Robust estimation in linear regression

## 4.1. Introduction

Parameter estimation is usually formulated as a linear regression problem, which is based on a model given by

$$z = Hx + e, \tag{47}$$

where x is a $(n \times 1)$-dimensional vector containing the parameters to be estimated;

z is a $(m \times 1)$-dimensional vector containing the measurements;

H is a $(m \times n)$-dimensional matrix called the observation matrix, also known as the design matrix;

e is a $(m \times 1)$-dimensional vector containing the measurement errors.

When m>n, we are dealing with an overdetermined system of m equations and n unknowns. The matrix H can be written as

$$H = \begin{bmatrix} h_{1,1} & h_{1,2} & \cdots & h_{1,n} \\ \vdots & \vdots & & \vdots \\ h_{m,1} & h_{m,2} & \cdots & h_{m,n} \end{bmatrix}. \tag{48}$$

In the case where there is an intercept, the associated column entries are equal to ones, say the last column, which yields

$$H = \begin{bmatrix} h_{1,1} & \cdots & h_{1,n-1} & 1 \\ \vdots & & \vdots & \vdots \\ h_{m,1} & \cdots & h_{m,n-1} & 1 \end{bmatrix}. \tag{49}$$

The problem is hence to estimate the slopes and the intercept of a hyperplane in a n-dimensional space passing through m data points. If the entries of H are uniformly distributed, the design of the space is said to be balanced. The case where they are

normally distributed is also very common. As seen in Chapter 3, the probability distribution of the error vector e is a very important factor in the accuracy of the estimation.

### 4.2. Weighted Least Squares estimator

Consider that each entry $e_i$ of the error vector follows a probability distribution characterized by a zero mean and a variance $\sigma_i^2$. Assume that they are uncorrelated so that the covariance matrix R of the vector e can be written as

$$R = \begin{bmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_m^2 \end{bmatrix}. \tag{50}$$

The Weighted Least Squares (WLS) estimator minimizes an objective function given by

$$J(x) = \frac{1}{2}\sum_{i=1}^{m}\left(\frac{r_i}{\sigma_i}\right)^2, \tag{51}$$

where the residuals $r_i$ are defined as

$$r_i = z_i - h_i^T \hat{x}. \tag{52}$$

The result of this minimization gives an estimate vector of the form

$$\hat{x} = \left(H^T R^{-1} H\right)^{-1} H^T R^{-1} z. \tag{53}$$

As we know, this is the best estimate in the particular case where the random vector e is Gaussian. This can be shown by checking that the covariance matrix $\Sigma_x$ of the estimate is equal to the inverse of the Fisher information matrix $I_f$ given by

$$\Sigma_x = I_f^{-1} = \left(H^T R^{-1} H\right)^{-1}. \tag{54}$$

From (53), we see that each of the $\hat{x}_i$ is written as a linear combination of the observations $\{z_j,\ j = 1,\ldots m\}$. Because the weights are constant, a single $z_j$ moving to infinity will take the estimator to infinity. Hence, the WLS estimator is not robust. This main disadvantage can be overcome by using robust M-estimators, what will be shown in the next section.

### 4.3. M-estimators

#### 4.3.1. Principle

A linear regression problem can be solved by means of an estimator defined as a minimizer of an objective function given by

$$J(x) = \frac{1}{2}\sum_{i=1}^{m} \rho\left(\frac{r_i}{\sigma_i}\right). \tag{55}$$

In this case, we may use the Iterative Reweighted Least Squares (IRLS) algorithm until convergence, that is,

$$\hat{\mathrm{x}}^{(\upsilon+1)} = \left(\mathrm{H}^{\mathrm{T}}\mathrm{R}^{-1}\mathrm{Q}^{(\upsilon)}\mathrm{H}\right)^{-1}\mathrm{H}^{\mathrm{T}}\mathrm{R}^{-1}\mathrm{Q}^{(\upsilon)}\mathrm{z}\,, \tag{56}$$

where $\upsilon$ is the iterative index and where Q is a diagonal matrix given by

$$\mathrm{Q} = \mathrm{diag}\left(\mathrm{q}\left(\frac{\mathrm{r_i}}{\sigma_{\mathrm{i}}}\right)\right). \tag{57}$$

The diagonal elements of Q are defined as

$$\mathrm{q}\left(\frac{\mathrm{r_i}}{\sigma_{\mathrm{i}}}\right) = \frac{\psi\left(\mathrm{r_i}/\sigma_{\mathrm{i}}\right)}{\mathrm{r_i}/\sigma_{\mathrm{i}}}. \tag{58}$$

The matrix Q will adjust the weight associated to each measurement: a smaller weight will be given to the observations with higher residuals, which will decrease their influence in the final result. By this way, the influence of the outliers is bounded. This is the reason why Q is called the weight matrix.

A simple formula of the covariance matrix $\Sigma_{\mathrm{x}}$ of the estimate can only be derived in the asymptotic case. If $r = z - H\,\hat{x}$ denotes the vector of the residuals and if $\sigma$ represents its standard deviation, it can be shown that [14]

$$\Sigma_{\mathrm{x}} = \frac{\mathrm{E}\left[\psi^2\left(\frac{\mathrm{r}}{\sigma}\right)\right]}{\left\{\mathrm{E}\left[\psi'\left(\frac{\mathrm{r}}{\sigma}\right)\right]\right\}^2}\left(\mathrm{H}^{\mathrm{T}}\mathrm{R}^{-1}\mathrm{H}\right)^{-1}, \tag{59}$$

where

$$\psi'\left(\frac{r}{\sigma}\right) = \frac{d\psi\left(r/\sigma\right)}{d\left(\mathrm{r}/\sigma\right)}. \tag{60}$$

In the finite case, some approximations have been derived. Tuckey proposes to approximate $\Sigma_x$ by

$$\hat{\Sigma}_x = \frac{m \cdot \sum_{i=1}^{m} \psi^2\left(\frac{r_i}{\sigma_i}\right)}{\left[\sum_{i=1}^{m} \psi'\left(\frac{r_i}{\sigma_i}\right)\right]\left[\sum_{i=1}^{m} \psi'\left(\frac{r_i}{\sigma_i}\right) - 1\right]} \left(H^T R^{-1} H\right)^{-1}. \qquad (61)$$

In the case where the M-estimator is a WLS estimator, we have $\psi(r/\sigma) = r/\sigma$ and $\psi'(r/\sigma) = 1$.

It turns out that, from a robustness viewpoint, the distribution of the points in the space associated with the matrix H matters. This space is called the design space. An observation $(z_i, h_i)$, where $h_i^T$ is the associated row of H, is said to be a leverage point if the point identified by $h_i$ is an outlier in the design space of H. Otherwise, it is called a vertical outlier [34].

It can be shown that while an M-estimator with a bounded $\psi$ function is robust against vertical outliers, it breaks down in presence of bad leverage points, that is, leverage points that have either wrong z values or wrong $h_I$ values.

### 4.3.2. Example in the 2 dimensional case

Suppose we have 10 noisy observations that are generated by adding a Gaussian error, $N(0,0.01)$ to true values that exactly satisfy a straight line with a slope equal to 1 and an intercept equal to 0. This yields

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{10} \end{bmatrix} = \begin{bmatrix} h_{1,1} & 1 \\ h_{2,1} & 1 \\ \vdots & \vdots \\ h_{10,1} & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_{10} \end{bmatrix}. \qquad (62)$$

These 10 observations lie in a 3 dimensional space: 2 dimensions for H and 1 for y. However, as the second column of H is equal to a vector of ones, we may delete that dimension. The first column h1 of H was drawn from a Normal distribution N(0,1) and the error e follows a Gaussian distribution with variance 0.01 and mean 0. To test the robustness of the Least Squares method and that of the Huber M-estimator, we replaced

the value of y associated with the largest value of h1 by 15. The result of the estimations is plotted in Figure 16. We see clearly that the solution of the Least Squares (dotted line) was highly influenced by the outlier while that of the Huber estimator (solid line) is not.

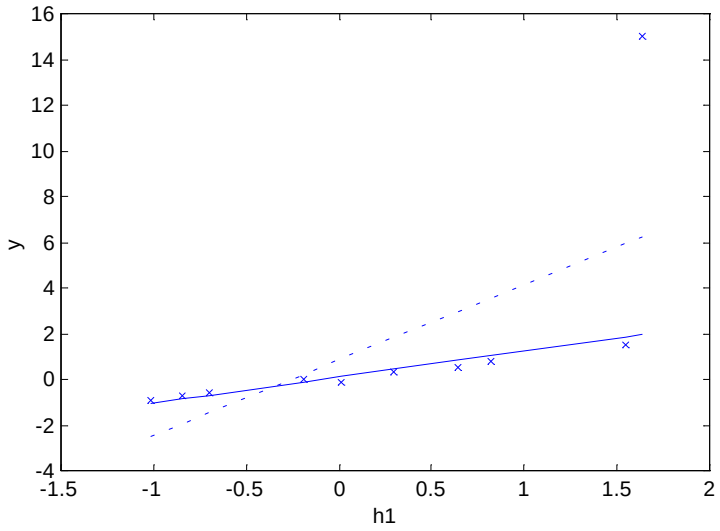



Figure 16: Estimation of the slope and intercept of a straight line with the Least Squares (dotted line) and with Huber M-estimator (solid line)

### 4.4. Leverage point identification

#### 4.4.1. Influence function

In a linear regression, the influence function given by (36) in the second Chapter has to be modified. While in the location case, the estimator is based on one-dimensional observations, in multiple dimensions the distribution of the points in the design space is important. This results in an influence function of an M-estimator that is expressed as the product of the influence of residuals IR and the influence of position IP. Assuming that the residuals r and the explanatory variable h follow respectively the distributions F and $\Lambda$, Hampel [12] showed that the influence function of an M-estimator is expressed as

$$\mathrm{IF}(\mathrm{r},\mathrm{h},\mathrm{F},\Lambda) = \mathrm{IR}(\mathrm{r},\mathrm{F})\cdot\mathrm{IP}(\mathrm{h},\Lambda). \qquad (63)$$

The IR is equal to the IF that we defined for the location case, that is,

$$\mathrm{IF}(\mathrm{r},\mathrm{F}) = \frac{\psi(\mathrm{r})}{\mathrm{E}[\psi'(\mathrm{r})]}, \qquad (64)$$

while the influence of position is given by

$$IP(h,\Lambda) = \left\{E\left[hh^T\right]\right\}^{-1} h\,. \tag{65}$$

As IR is bounded for robust M-estimators, we have to bound IP if we want a bounded influence function, IF. To do so, a weight function w is introduced in the implicit equation (55) that defines an M-estimator. This weight function decreases with an increase of the standardized distance of a point $h_i$ with respect to the point cloud in the design space of H.

In this study, we are mainly interested in distances that are robust to outliers, called Robust Distances (RD). Two methods to calculate RD are described in the next two paragraphs: the first method is a robust version of the Mahalanobis Distances while the second one makes uses of one-dimensional projections, yielding the so-called Projection Statistics.

### 4.4.2. Mahalanobis Distances

To identify leverage points, we need to define in a robust way the center of the design space and the standardized distances of each point with respect to this center. We will first derive the classical Mahalanobis Distances and then the Robust Distances.

The Mahalanobis Distances are defined as

$$\mathrm{MD}_i = \sqrt{\left(\mathrm{h}_i - \bar{\mathrm{h}}\right)^{\mathrm{T}} \mathrm{C}^{-1} \left(\mathrm{h}_i - \bar{\mathrm{h}}\right)}, \tag{66}$$

where

$$\bar{\mathrm{h}} = \frac{1}{\mathrm{m}} \sum_{i=1}^{\mathrm{m}} \mathrm{h}_i \tag{67}$$

and

$$\mathrm{C} = \frac{1}{\mathrm{m}-1} \sum_{i=1}^{\mathrm{m}} \left(\mathrm{h}_i - \bar{\mathrm{h}}\right)\left(\mathrm{h}_i - \bar{\mathrm{h}}\right)^{\mathrm{T}}. \tag{68}$$

We see that if one of the $h_i$ becomes infinitely large, the mean vector $\bar{\mathrm{h}}$ will become infinite too and the covariance matrix C will explode, which means that these distances are no longer reliable.

The Robust Distances proposed by Huber [14] include weight functions to bound the influence of bad data. They are calculated iteratively until convergence by means of the following algorithm:

1. Calculate the Huber weight functions:

$$u_i = \begin{cases} \dfrac{a^2}{RD_i^{\,2}} & \text{for} \quad |RD_i| \le a \\ 1 \quad \text{for} \quad a \le |RD_i| \le b \\ \dfrac{b^2}{RD_i^{\,2}} & \text{for} \quad |RD_i| \ge b \end{cases} \tag{69}$$

$$v_i = \begin{cases} 1 & \text{for} \quad |RD_i| \le c \\ \dfrac{c}{RD_i} & \text{for} \quad |RD_i| \ge c \end{cases} \tag{70}$$

2. Calculate $\begin{cases} y_i = L^{-1} h_i \\ \bar{y} = \dfrac{1}{m} \sum_{i=1}^{m} v_i y_i \end{cases}$ (71)

3. Calculate $C = \dfrac{1}{d \cdot m} \sum_{i=1}^{m} u_i (y_i - \bar{y})(y_i - \bar{y})^T = LL^T$ (72)

4. Calculate $RD_i = \sqrt{(y_i - \bar{y})^T (y_i - \bar{y})}$. (73)

The initial condition is usually taken as $RD_1$=MD.

It can be shown that these Robust Distances follow a $\chi^2_n$ distribution if h is normally distributed [14]. Therefore, the parameter b and c are typically set to $b = c = \chi^2_{n,0.975}$. This means that asymptotically, we will only assign weights greater or equal to 1 to 97.5% of the points in the design space, should the latter be normally distributed. The influence of the 2.5% furthest points will progressively diminish, while points close to the center of the point cloud will be given more weight. The parameters a and d are tuned depending on the problem under study. Typically, they are put equal to $a = 0.2\, b$ and d = 1.

### 4.4.3. Projection Statistics

Carrying out one-dimensional projections provides another way to robustly estimate distances, yielding the so-called Projection Statistics [9]. As the algorithm to calculate them does not require any iterative calculation, it is much faster than the one we saw above. Moreover, they are calculated through a non-iterative algorithm that consists in the following steps:

1. Calculate $m = \text{median}_{i,j}\left(h_{i,j}\right)$ (74)
2. Calculate $v_i = h_i - m,\ i = 1,2,\ldots,m$ (75)
3. For every direction $v_i$, calculate $z_{i,j} = \dfrac{\left|y_{i,j} - L_i\right|}{S_i},\ j = 1,2,\ldots,m$ (76)

where $y_{i,j} = h_j v_i^T$, $L_i = \text{median}_j\left(y_{i,j}\right)$, $S_i = \text{length of the shortest half of } y_{i,j}$ .
The latter is defined as the smallest differences

$$\{y_{(i,\nu)} - y_{(i,1)}\ ,\ y_{(i,\nu+1)} - y_{(i,2)}\ ,\ \ldots,\ y_{(i,m)} - y_{(i,m-\nu+1)}\},$$

where $\nu = [m/2] + 1$ and $y_{(i,\nu)}$ is the $\nu$th ordered observation.

4. Calculate the projection statistic of the $i^{th}$ data point, which is given by

$$PS_i = \max_j\left(z_{i,j}\right). \quad (77)$$

The Projection Statistics follow roughly a $\chi^2_n$ distribution with n degrees of freedom, as shown by Rousseuw and Van Zomeren [34]. Therefore, by comparing them to $b = \chi^2_{n,0.975}$, we are able to identify outliers. Let $\upsilon$ be their number. Assume that these outliers are associated with the last rows of H, namely $h_j$, $j = m - \upsilon,\ldots,m$. Then, the Robust Distances are defined as

$$RD_i = \sqrt{\left(h_i - \bar{h}\right)^T C^{-1}\left(h_i - \bar{h}\right)}, \quad (78)$$

where

$$\bar{h} = \frac{1}{m-\upsilon}\sum_{i=1}^{m-\upsilon} h_i \quad (79)$$

$$C = \frac{1}{m-\upsilon-1}\sum_{i=1}^{m-\upsilon}\left(h_i - \bar{h}\right)\left(h_i - \bar{h}\right)^T. \quad (80)$$

From the foregoing equations, we see that the mean value $\bar{h}$ as well as the covariance matrix are calculated without the $\upsilon$ outliers. Note that these Robust Distances are not affine equivariant, which may be a limiting factor should a transformation be applied to the data points.

4.4.4. A 2-dimensional example

Let us consider a matrix H with 1000 rows and 2 columns, whose values follow a Gaussian distribution with zero mean and unit variance. The squared Robust Distances should roughly follow a $\chi^2_n$ distribution with 2 degrees of freedom.

Figure 17 displays the scaled relative frequency histogram of $RD_i^2$ together with the $\chi^2_2$ density function (solid line). As expected, we see that the hights of the rectangles are approximately $\chi^2_2$ distributed. In the case where H does not contain any outlier, the $RD_i$ given by the two foregoing methods are very close.

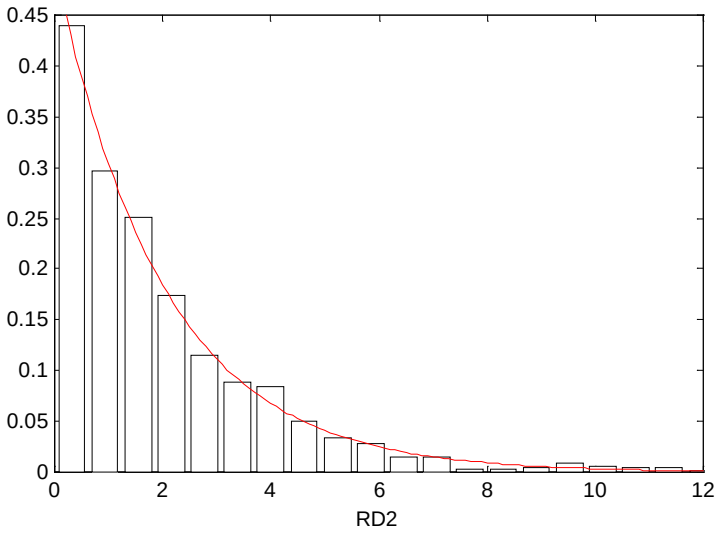



Figure 17: Scaled relative frequency histogram of $RD_i^2$ and the $\chi^2_2$ density function

Figure 18 depicts the point cloud in the design space of H together with the 97.5% confidence ellipse given by

$$h^T C^{-1} h = \sqrt{\chi^2_{2,0.975}} = 2.71\,. \qquad (81)$$

We observe that the ellipse is very close to a circle centered at the origin and containing roughly 97.5% of the points.

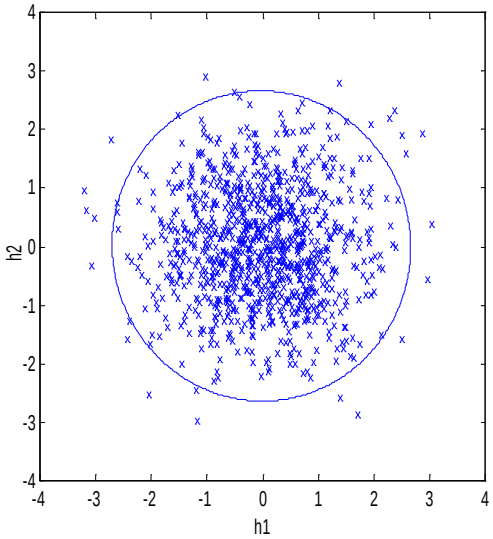



Figure 18: 97.5% confidence ellipse of the points in the design space

We also carried out some simulations to verify the robustness of our algorithms. The simulations consists of introducing more and more outliers in H and checking if the center of H is still near the origin and if the 97.5% confidence ellipse is still a circle of radius 2.71 and hence, contains only good data points.

Figure 19 illustrates the robustness of RD. Out of 1000 points of H, we set 20 points equal to (20,20), which amounts to introducing 2% of outliers. We see that if we estimate the center of the point cloud and the covariance matrix by means of the classical methods, the ellipse is attracted by the outliers (dotted ellipse). On the other hand, both estimations based on RD efficiently reject the bad values.

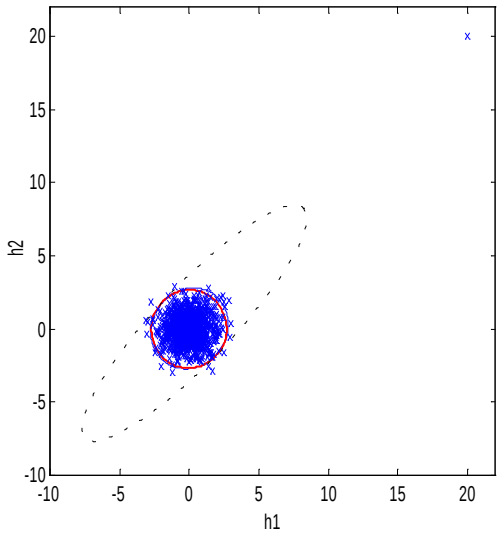



Figure 19: 97.5% confidence ellipses based on the classical Mahalanobis Distances (dotted ellipse), the Huber Robust Distances (thin solid ellipse), and the Projection Statistics (thick solid ellipse) when 2% of the 1000 points are outliers.

Figure 20 illustrates the fact that the method based on Projection Statistics is the most robust one. We set 200 of the 1000 points of H to the value (20,20), which means that 20% of the points are outliers. It turns out that only Projection Statistics are able to reject these outliers and to yield a confidence ellipse that encloses only good data points. Figure 21 shows the histogram of the Huber RD. While the outliers distort the distribution, they are at a value inferior to $\chi^2_{2,0.975} \approx 7.38$, which means that the method fails to identify them.

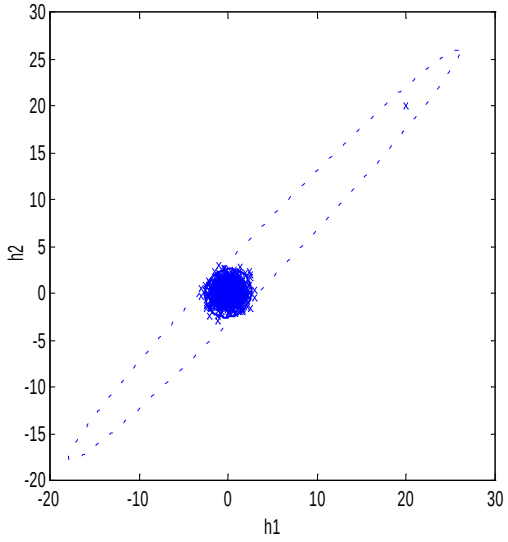



Figure 20: 97.5% confidence ellipses based on Mahalanobis Distances (dotted ellipse) and on Projection Statistics (solid ellipse) when 20% of the points are outliers.

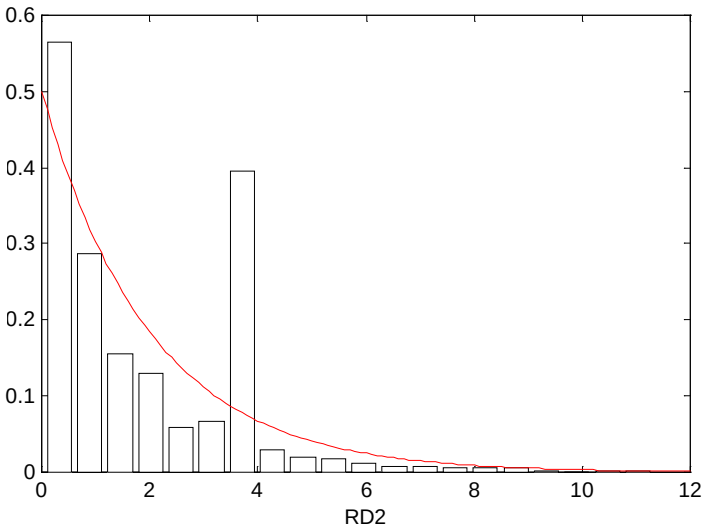



Figure 21: Scaled relative frequency histogram of the Huber $RD_i^2$.

## 4.5. GM-estimators

### 4.5.1. Principle

Robust generalized M-estimators make use of Robust Distances and hence, are able to disregard bad leverage points. Let us show how they are derived from the M-estimators. An M-estimator minimizes an objective function written as

$$J(x)=\frac{1}{2}\sum_{i=1}^{m}\rho\left(\frac{r_i}{\sigma_i}\right), \tag{82}$$

which implies that it is solution to the necessary condition of optimality given by

$$\frac{\partial J(x)}{\partial x}=\sum_{i=1}^{m}\frac{h_i}{\sigma_i}\cdot\psi\left(\frac{r_i}{\sigma_i}\right)=0\,. \tag{83}$$

In order to bound the influence of outlying $h_i$ in (82), Schweppe [13] proposed to multiply $h_i$ by a weight function $w(h_i)$ and to divide $r_i$ by $w(h_i)$, yielding

$$\sum_{i=1}^{m}w(h_i)\cdot\frac{h_i}{\sigma_i}\cdot\psi\left(\frac{r_i}{\sigma_i w(h_i)}\right)=0\,, \tag{84}$$

where $w(h_i)$ are based on the Mahalanobis distances. A robust version of these weights is written as

$$w(h_i)=\begin{cases}1 & \text{for} \quad |RD_i|\le b\\ \dfrac{b^2}{RD_i^{\,2}} & \text{for} \quad |RD_i|\ge b\end{cases}. \tag{85}$$

The solution of this equation, calculated thanks to the IRLS algorithm, is then as follows:

$$\hat{x}^{(\upsilon+1)}=\left(H^T R^{-1}Q^{(\upsilon)}H\right)^{-1}H^T R^{-1}Q^{(\upsilon)}z\,, \tag{86}$$

where the matrix Q is defined as

$$Q=\mathrm{diag}\left(q\left(\frac{r_i}{\sigma_i w(h_i)}\right)\right). \tag{87}$$

### 4.5.2. Example of a simple regression

Let us consider again the example given in Section 4.3.2. Here, we will make the $10^{th}$ observation $(y_{10},h_{10,1})$ not a vertical outlier, but a bad leverage point. To this end, we

put $h_{10,1}$ equal to 6 and $y_{10}$ equal to 0. Figure 22 shows that the Schweppe-type Huber GM-estimator (solid line) is able to reject the bad leverage point while the Huber M-estimator (dotted line) is not.

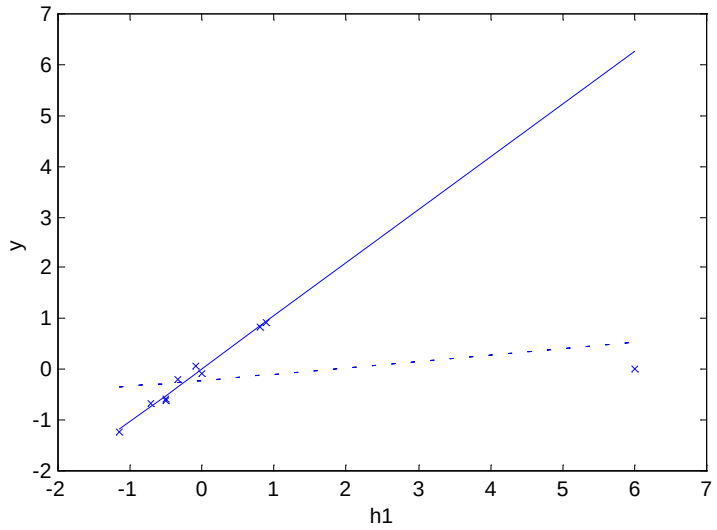



Figure 22: Estimation of the slope and intercept of a straight line with the Huber M-estimator (dotted line) and with the Schweppe-type Huber GM-estimator (solid line)

### 4.6. Conclusions

This Chapter described a robust estimation method in linear regression. First, M-estimators were derived and their robustness analyzed in presence of outliers. Then, the unbounded influence of bad leverage points on M-estimators was observed and the class of robust GM-estimators defined. The emphasis was on the Schweppe-type Huber GM-estimator, which is an improved version of the Huber M-estimator that is able to withstand bad leverage points. We also compared two methods to calculate Robust Distances, and demonstrated through simulations that the method based on Projection Statistics were the most robust.

# Chapter 5

# Robust Kalman filter

## 5.1. Introduction

Because speech signals are non-stationary, it is not simple to find filters that reduce noise efficiently. Different methods of speech enhancement have been proposed in the literature [20, 27], among which the Wiener filter is one of the most widely used methods. Paliwal and Basu [28] compared the performance of this filter with the Kalman filter in a white noise environment. It turned out that the Kalman filter was able to provide better results. Then, Gibson et al. [11] extended this study to colored noise and showed that in this second case, the enhancement was also better. Another method of speech enhancement based on the Kalman filter can be found in [8]. These results encourage us to opt for Kalman filtering. Moreover, Kalman filter has the advantage to use a time representation of the signal, which seems more convenient in the case were impulsive noise is present. Also, we intend to develop a robust version of the Kalman filter using the Huber M-estimator. This robust filter is presented in this Chapter.

Kalman filtering is a process that estimates the state variable of a linear dynamic system. It is based on 2 different pieces of information: a dynamic model of the system and some measurements of the output. By putting this information in an appropriate form, we can reformulate the problem as a generalized linear regression estimation. It turns out that the Kalman filter is a Weighted Least Squares state estimation. By introducing the weight function of the Huber M-estimator, we add some robustness to the filter, what will be demonstrated through some simulations.

## 5.2. Classical Kalman filter

### 5.2.1. Discrete linear dynamic systems

Usually, a linear dynamic system can be modeled by means of a dynamic state equation and an observation equation together with some conditions on noise and initial values [1, 3, 10, 24]. Here, we will restrict attention to a Bayesian model given by

$$x_{k+1} = F_k x_k + G_k w_k + C_k u_k \tag{88}$$

$$z_k = H_k x_k + e_k , \tag{89}$$

where

$x_k$ is the state vector at time $t_k$;

$w_k$ is the system error vector at time $t_k$;

$u_k$ is the input vector at time $t_k$;

$z_k$ is the measurement vector at time $t_k$;

$e_k$ is the measurement error vector at time $t_k$;

$F_k$ is the transition matrix at time $t_k$;

$H_k$ is the observation matrix at time $t_k$;

$C_k$ and $G_k$ are 2 known matrices at time $t_k$.

The dynamic state equation given by (88) indicates that the system is completely defined at a time $t_k$ by the sate variable $x_k$. To obtain the state at the next instant, we just need to know the new input vector of the system, $u_{k+1}$. The noise $w_k$ accounts for the uncertainties in the model, possibly in the linear approximation that is used and in the entries of the transition matrix $F_k$. We can notice that all the matrices are time-variant, which allows us to model a system whose behavior is not constant in time.

The observation equation given by (89) relates the observation vector $z_k$ to the state vector $x_k$ at time $t_k$. It can be obtained by sensing several variables of the system, referred to as measurement redundancy, or by placing several systems in parallel, referred to as structural redundancy, or by repeating the experiment several times. If there is only

one sensor, one system, and one experiment, the vector $z_k$ reduces to a scalar. The matrix $H_k$ relates the state vector to the measurement vector through an additive error vector $e_k$. The latter contains all the errors that can affect the measurement as well as the uncertainty in the model. Since the error $e_k$ is usually assumed to have zero mean, systematic errors are not accounted for in the model.

The Bayesian model is based on a certain number of assumptions. First, the state variable $x_0$ at time $t_0$ is supposed to be a random variable with zero mean and known covariance matrix given by

$$E[x_0] = 0 \tag{90}$$

$$E[x_0 x_0'] = \psi . \tag{91}$$

Then, the error vectors $w_k$ and $e_k$ are assumed to be independent random variables with zero mean and known covariance matrices $W_k$ and $R_k$, yielding

$$E[w_k] = 0 \quad , \quad E[e_k] = 0$$

$$W_k = \begin{bmatrix} \alpha_{1,k}^{\,2} & & 0 \\ & \ddots & \\ 0 & & \alpha_{m,k}^{\,2} \end{bmatrix} \quad , \quad R_k = \begin{bmatrix} \beta_{1,k}^{\,2} & & 0 \\ & \ddots & \\ 0 & & \beta_{m,k}^{\,2} \end{bmatrix} \tag{92}$$

$$E[w_k e_l] = 0 \quad \forall\, k, l .$$

5.2.2. Kalman filter seen as a linear regression

The Kalman filter, which is based on the foregoing dynamic state equations, can be seen as a two-steps process: the prediction step and the filtering step.

*The prediction step*

In the dynamic state equation, the state error vector, $w_k$, represents the error between the real system and its assumed model. Because this random vector is supposed to have zero mean, we can predict the value of $x_k$ thanks to the following equation:

$$\hat{x}_{k|k-1} = F_{k-1}\hat{x}_{k-1|k-1} + C_{k-1}u_{k-1} \tag{93}$$

where $\hat{x}_{k|k-1}$ is the predicted value of the state vector $x_k$ at time $t_k$, given the state estimate at time $t_{k-1}$, $\hat{x}_{k-1|k-1}$. Let $\delta_{k|k-1}$ be the prediction error defined as

$$\delta_{k|k-1} = x_k - \hat{x}_{k|k-1}. \tag{94}$$

Then, we have

$$\hat{x}_{k|k-1} = x_k - \delta_{k|k-1}. \tag{95}$$

*The filtering step*

This step makes use of the observation equation given by

$$z_k = H_k x_k + e_k, \tag{96}$$

together with the state prediction $\hat{x}_{k|k-1}$.

At this point, we combine equations (95) and (96) in a single equation to build a new linear regression model given by

$$\begin{bmatrix} z_k \\ \hat{x}_{k|k-1} \end{bmatrix} = \begin{bmatrix} H_k \\ I \end{bmatrix} x_k + \begin{bmatrix} e_k \\ \delta_{k|k-1} \end{bmatrix}. \tag{97}$$

The foregoing equation can be written in compact form as

$$\tilde{z}_k = \tilde{H}_k x_k + \tilde{e}_k, \tag{98}$$

where

$$\operatorname{cov}(\tilde{e}_k) = \tilde{R}_k = \begin{bmatrix} R_k & 0 \\ 0 & \Sigma_{k|k-1} \end{bmatrix}. \tag{99}$$

Based on this equation, the Kalman filter estimates $x_k$ using the Weighted Least Squares estimator, which minimizes an objective function defined as

$$J(x_k) = (\tilde{z}_k - H_k x_k)^T \tilde{R}_k^{-1} (\tilde{z}_k - \tilde{H}_k x_k). \tag{100}$$

The solution is then written as

$$\hat{x}_{k|k} = \left(\tilde{H}_k^T \tilde{R}_k^{-1} \tilde{H}_k\right)^{-1} \tilde{H}_k^T \tilde{R}_k^{-1} \tilde{z}_k. \tag{101}$$

Since the Weighted Least Squares estimator is the maximum likelihood estimator at the Gaussian distribution, it will give us the best estimate of $x_k$ if the errors of both the

dynamic model and the observation model are Gaussian variables with zero mean. If the error vector is non Gaussian with long tails, then the performance of the estimator will degrade. Also, in presence of a single outlier, the bias of the estimator can be arbitrarily large. The Kalman filter is not robust in this sense.

### 5.2.3. The recursive Kalman filter

The Kalman filter equations can be written in a more useful form, which is known as the recursive Kalman filter. This form is derived by making use of the inversion lemma given by

$$\left(\mathrm{A}-\mathrm{CB}^{-1}\mathrm{D}\right)^{-1}=\mathrm{A}^{-1}+\mathrm{A}^{-1}\mathrm{C}\left(\mathrm{B}-\mathrm{DA}^{-1}\mathrm{C}\right)^{-1}\mathrm{DA}^{-1}, \qquad (102)$$

where A and B are inversible matrices, and C and D are matrices with adequate dimensions.

The predicted state vector $\hat{x}_{k|k-1}$ is related to the state estimate $\hat{x}_{k-1|k-1}$ through

$$\hat{\mathrm{x}}_{\mathrm{k|k-1}}=\mathrm{F}_{\mathrm{k-1}}\hat{\mathrm{x}}_{\mathrm{k-1|k-1}}. \qquad (103)$$

Let $\Sigma_{k-1|k-1}$ denote the covariance matrix of $\hat{x}_{k-1|k-1}$ and $\Sigma_{\mathrm{k|k-1}}$ denote the covariance matrix of $\hat{x}_{k|k-1}$. The latter is expressed as

$$\Sigma_{\mathrm{k|k-1}}=\mathrm{F}_{\mathrm{k-1}}\Sigma_{\mathrm{k-1|k-1}}\mathrm{F}_{\mathrm{k-1}}{}^{\mathrm{T}}+\mathrm{G}_{\mathrm{k-1}}\mathrm{W}_{\mathrm{k-1}}\mathrm{G}_{\mathrm{k-1}}{}^{\mathrm{T}}. \qquad (104)$$

Since we have

$$\tilde{H}_k{}^{T}\tilde{R}_k{}^{-1}\tilde{H}_k=H_k R_k{}^{-1}H_k+\Sigma_{k|k-1}, \qquad (105)$$

it follows that

$$\left[\tilde{H}_k{}^{T}\tilde{R}_k{}^{-1}\tilde{H}_k\right]^{-1}=\Sigma_{k|k-1}-\Sigma_{k|k-1}H_k{}^{T}\left[H_k\Sigma_{k|k-1}H_k{}^{T}+R_k\right]^{-1}H_k\Sigma_{k|k-1}, \qquad (106)$$

which can be derived by using the inversion matrix lemma.

By defining the gain matrix $\mathrm{K_k}$ as

$$K_k=\Sigma_{k|k-1}H_k{}^{T}\left[H_k\Sigma_{k|k-1}H_k{}^{T}+R_k\right]^{-1}, \qquad (107)$$

and substituting it in (101), we get

$$\hat{\mathrm{x}}_{\mathrm{k|k}}=\left(\Sigma_{\mathrm{k|k-1}}-\mathrm{K}_{\mathrm{k}}\mathrm{H}_{\mathrm{k}}\Sigma_{\mathrm{k|k-1}}\right)\left[\mathrm{H}_{\mathrm{k}}{}^{\mathrm{T}}\quad\mathrm{I}\right]\begin{bmatrix}\mathrm{R}_{\mathrm{k}}{}^{-1} & 0\\ 0 & \Sigma_{\mathrm{k|k-1}}{}^{-1}\end{bmatrix}\begin{bmatrix}\mathrm{z}_{\mathrm{k}}\\ \hat{\mathrm{x}}_{\mathrm{k|k-1}}\end{bmatrix}. \qquad (108)$$

The latter equation reduces to

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + K_k\left(z_k - H_k\hat{x}_{k|k-1}\right). \quad (109)$$

It follows that the covariance matrix $\Sigma_{k|k}$ of $\hat{x}_{k|k}$ is expressed as

$$\Sigma_{k|k} = \left(I - K_k H_k\right)\Sigma_{k|k-1}. \quad (110)$$

In summary, the recursive Kalman filter is defined by the following set of equations:

$$\hat{x}_{k|k-1} = F_{k-1}\hat{x}_{k-1|k-1} \quad (111)$$

$$\Sigma_{k|k-1} = F_{k-1}\Sigma_{k-1|k-1}F_{k-1}^{T} + G_{k-1}W_{k-1}G_{k-1}^{T} \quad (112)$$

$$K_k = \Sigma_{k|k-1}H_k^{T}\left[H_k\Sigma_{k|k-1}H_k^{T} + R_k\right]^{-1} \quad (113)$$

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + K_k\left(z_k - H_k\hat{x}_{k|k-1}\right) \quad (114)$$

$$\Sigma_{k|k} = \left(I - K_k H_k\right)\Sigma_{k|k-1}. \quad (115)$$

The Kalman gain $K_k$ is called the optimal gain because it minimizes the Weighted Least Squares objective function. In other words, it will give the best correction between $\hat{x}_{k|k-1}$ and $\hat{x}_{k|k}$ at the Gaussian distribution. An interesting advantage of this form is that only the very last state is needed to compute the new one.

### 5.2.4. Stability of the Kalman filter

One of the main issues that needs to be addressed is the stability of the Kalman filter. In the case of a time-invariant system model characterized by $F_k$, $C_k$, $G_k$, $H_k$, $R_k$, $W_k$ invariant with time, a necessary condition of stability of the filter is that $F_k$ is controllable and observable, $R_k>0$ and $W_k>0$. It can be shown that these conditions are equivalent to a condition on the covariance matrix relative to the noise $w_k$, which states that $\Sigma_{k|k}$ converges towards a matrix $\Sigma_\infty$ and there exists $\alpha>0$ and $\beta<\infty$ so that $\alpha I < \Sigma_\infty < \beta I$ [1]. Therefore, the stability of the filter resides in the good behavior of the matrix $\Sigma_{k|k}$.

### 5.3. Robust Kalman filter

Several methods are presented in the literature to robustify the Kalman filter, based on a robust AR model [17, 23, 36]. Here, instead of solving the regression problem using the Weighted Least Squares estimator, we build a robust filter using the Huber M-estimator. To insert the weight function of the Huber M-estimator in the equations of the Kalman filter, we first transform our model so that the covariance of the noise vector is equal to the identity matrix. By using Cholesky's technique, we can factorize it as

$$\tilde{R}_k = \left(\sqrt{\tilde{R}_k}\right)\left(\sqrt{\tilde{R}_k}\right)^T. \tag{116}$$

Then, we multiply the model equation given by (98) from the left by $\sqrt{\tilde{R}_k^{-1}}$ to get

$$\left(\sqrt{\tilde{\mathrm{R}}_\mathrm{k}}\right)^{-1}\tilde{\mathrm{z}}_\mathrm{k} = \left(\sqrt{\tilde{\mathrm{R}}_\mathrm{k}}\right)^{-1}\tilde{\mathrm{H}}_\mathrm{k}\mathrm{x}_\mathrm{k} + \left(\sqrt{\tilde{\mathrm{R}}_\mathrm{k}}\right)^{-1}\tilde{\mathrm{e}}_\mathrm{k}. \tag{117}$$

Therefore, the model becomes

$$\mathrm{y}_\mathrm{k} = \mathrm{A}_\mathrm{k}\mathrm{x}_\mathrm{k} + \eta_\mathrm{k} \quad \text{with} \quad \mathrm{cov}(\eta_\mathrm{k}) = \mathrm{I}. \tag{118}$$

The Iterative Reweighted Least Squares algorithm, that minimizes the objective function associated with the Huber M-estimator, is written as

$$x_{k|k}^{(\upsilon+1)} = (A_k^T Q^{(\upsilon)} A_k)^{-1} A_k^T Q^{(\upsilon)} y_k, \tag{119}$$

where

$$Q^{(\upsilon)} = diag\left(q^{(\upsilon)}(r_i/\sigma_i)\right). \tag{120}$$

In practice, we will not use this formulation to calculate our solution because the matrix $\left(A_k^T Q^{(\upsilon)} A_k\right)$ can have a large size and be difficult to invert. Similarly to the classical filter, we use the inversion matrix lemma given by (102) to derive the following algorithm:

Step 1: Use Cholesky method to factorize $\Sigma_{\mathrm{k|k-1}}$ into

$$\Sigma_{\mathrm{k|k-1}} = \left(\sqrt{\Sigma_{\mathrm{k|k-1}}}\right)\left(\sqrt{\Sigma_{\mathrm{k|k-1}}}\right)^{\mathrm{T}} \tag{121}$$

Step 2: Iterate until convergence the following equations:

$$\Sigma_{k|k-1}^{*(\upsilon)} = \left(\sqrt{\Sigma_{k|k-1}}\right) Q_x^{-1(\upsilon)} \left(\sqrt{\Sigma_{k|k-1}}\right)^T \tag{122}$$

$$R_k^{*(\upsilon)} = R_k Q_z^{-1(\upsilon)} \tag{123}$$

$$K_k^{*(\upsilon)} = \Sigma_{k|k-1}^{*(\upsilon)} H_k^T \left(H_k \Sigma_{k|k-1}^{*(\upsilon)} H_k^T + R_k^{*(\upsilon)}\right) \tag{124}$$

$$x_{k|k}^{(\upsilon+1)} = \hat{x}_{k|k-1} + K_k^{*(\upsilon)} \left(z_k - H_k \hat{x}_{k|k-1}\right) \tag{125}$$

$$rz_k^{(\upsilon+1)} = z_k - H_k x_{k|k}^{(\upsilon+1)} \tag{126}$$

$$rx_k^{(\upsilon+1)} = \hat{x}_{k|k-1} - x_{k|k}^{(\upsilon+1)} \tag{127}$$

$$Q_z^{(\upsilon+1)} = \mathrm{diag}\left(q\left(\frac{rz_i^{(\upsilon+1)}}{\sigma_i}\right)\right) \tag{128}$$

$$Q_x^{(\upsilon+1)} = \mathrm{diag}\left(q\left(\frac{rx_i^{(\upsilon+1)}}{\sigma_i}\right)\right) \tag{129}$$

Step 3: Calculate the state covariance matrices,

$$\Sigma_{k|k} = \left(I - K_k^* H_k\right) \Sigma_{k|k-1}^* \tag{130}$$

$$\Sigma_{k+1|k} = F_k \Sigma_{k|k} F_k^T + G_k W_k G_k^T \tag{131}$$

and the state vector,

$$\hat{x}_{k+1|k} = F_k \hat{x}_{k|k} . \tag{132}$$

We notice that the only difference with the standard Kalman filter is the introduction of the matrices $Q_x$ and $Q_z$ in the covariance of the noise. As these weight matrices are linked with the residuals and with the value of the state variables, we need to iterate until convergence.

## 5.4. A simple case of AR(3)

### 5.4.1. The model

To simulate the performances of our robust Kalman filter, we used an autoregressive model of order 3 defined as

$$y_{k+1} = 0.5y_k + 0.3y_{k-1} + 0.1y_{k-2} + v_{k+1}.$$

We can identify this equation with the model given by (88) and (89), where

$$x_k = \begin{bmatrix} y_{k-2} \\ y_{k-1} \\ y_k \end{bmatrix} \quad , \quad F = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0.1 & 0.3 & 0.5 \end{bmatrix} \quad , \quad G = I \quad , \quad w_k = \begin{bmatrix} 0 \\ 0 \\ v_k \end{bmatrix} \quad , \quad H = [0\ 0\ 1] \qquad (133)$$

As we are more interested in the steady state of the system than in its response to a particular input, we will consider that there is no input signal. We can also notice that the matrices are independent with time, which simplifies the matter.

### 5.4.2. Simulation results

In these simulations, we consider the error on the model and the error on the observations as Gaussian variables with mean zero and a variance of 0.04. Then, we introduce an outlier at the unit time t=15 by altering the value of $e_{15}$ to a higher value equal to 3. We filter the obtained signal using the classic and robust versions of the Kalman filter. For the robust filter, we choose the threshold value b of the Huber M-estimator equal to 3.5, which means that every standardized residual higher than 3.5 will receive a weight smaller than 1.

We see clearly from Figure 23 that in the portions of the signal where there are no impulse, the two filters give the same output value. In the presence of spikes, the robust filter is closer to the real values. The SNR confirms the good behavior of the robust filter. The reason why the robust Kalman filter does not completely remove the peak introduced by the outlier is that there is not enough redundancy in the filtering process. Indeed, the linear regression made at each unit time is based on 1 observation and 3 predicted values. A way to give more influence to the estimated value is to decrease the parameter of the

Huber M-estimator used to calculate the matrix Qz. However, this may damage the signal in regions where there is no outlier.

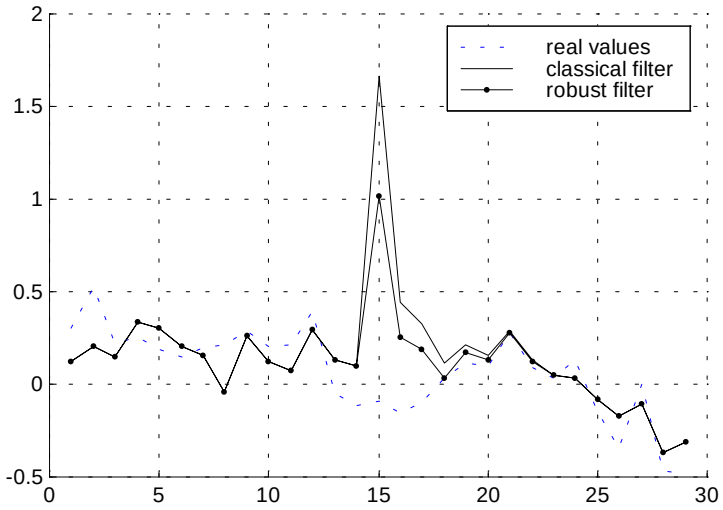



Figure 23: Outputs of the classical and robust Kalman filters (b=3.5) in presence of an outlier.

The introduction of weight matrices in the equations of the filter may be a factor of risk for the stability of the filter.

Figure 24 shows the evolution of the first diagonal element of the covariance matrix $\Sigma_{k|k}$ when k increases. A similar behavior can be observed for other variances. The initial value is 1, and we see that for k=4, $\Sigma_{4|4}$ has already converged to a steady state value equal to 0.1. However, the correction made by the robust filter at time t=15 due to the outlier introduces a perturbation at time t=16. This situation is only temporary, and no case was found where $\Sigma_{k|k}$ does not return back to the steady state value.

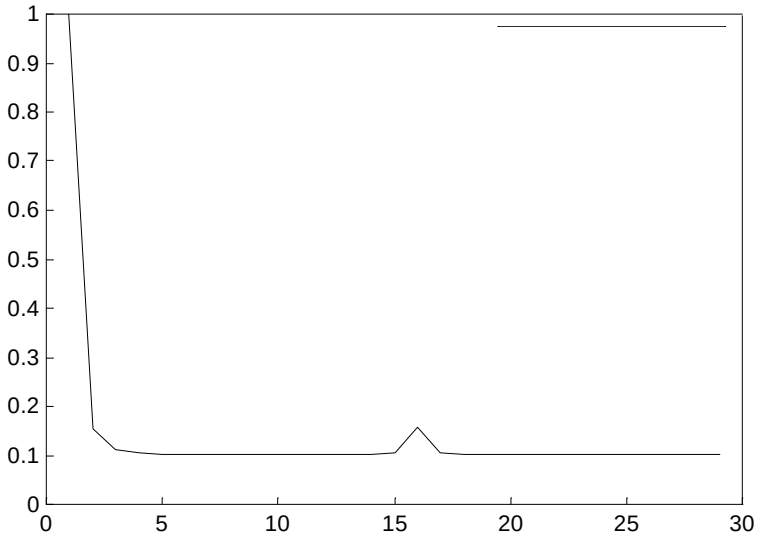



Figure 24: Covariance matrix $\Sigma_{k|k}$ in presence of 1outlier with b=3.5.

Figure 25 and Figure 26 show the result of a similar experiment if the cutoff value of the Huber M-estimator is put equal to 1. With this lower cutoff value, the filter exhibits a better outlier suppression capability. The covariance matrix is obviously more perturbed, but there is still no problem of stability since it converges to a steady-state value.

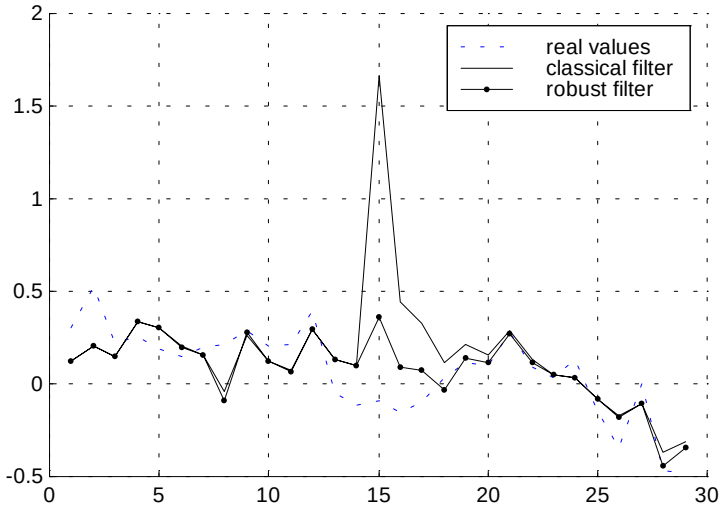



Figure 25: Output of the classical and robust Kalman filters (b=1) in presence of an outlier.

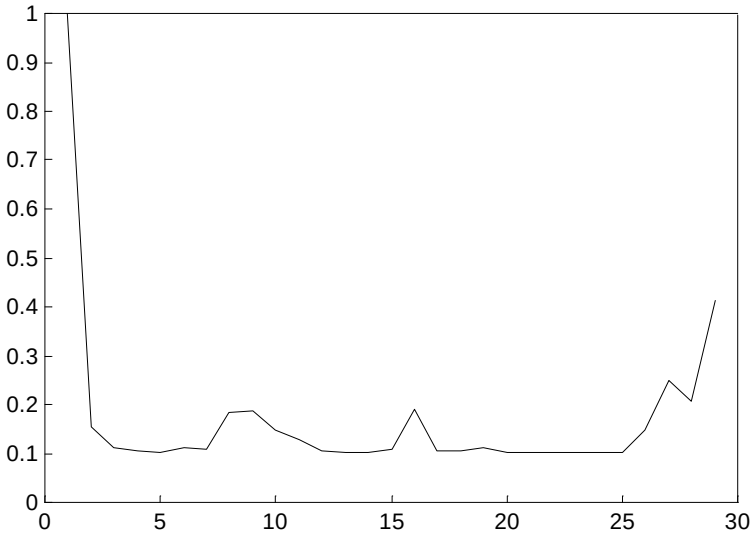



Figure 26: Covariance matrix $\Sigma_{k|k}$ in presence of 1 outlier with b=1.

Figure 27 and Figure 28 display the output of the classical and robust Kalman filters when 3 consecutive outliers have been introduced, specifically, $e_{15}=e_{16}=e_{17}=3$. The cutoff value of the Huber M-estimator is maintained at 1. We see that the robust filter still

succeeds in attenuating the pulse. However, the covariance matrix becomes more and more unstable because of the corrections that have to be brought.

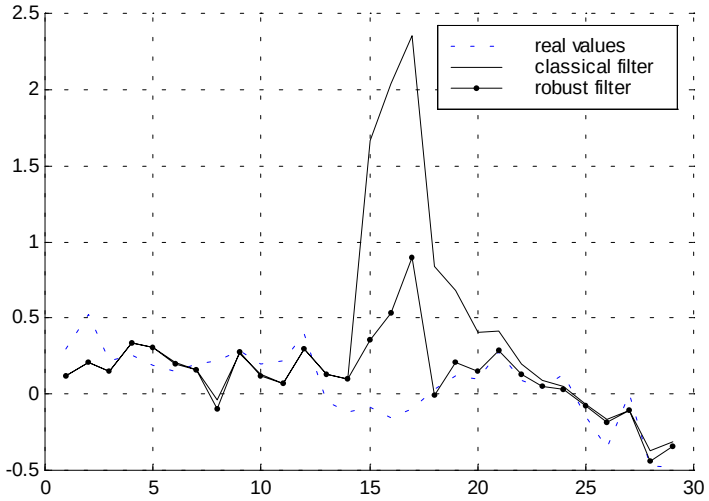



Figure 27: Output of the classical and robust Kalman filters (b=1) in presence of 3 outliers.

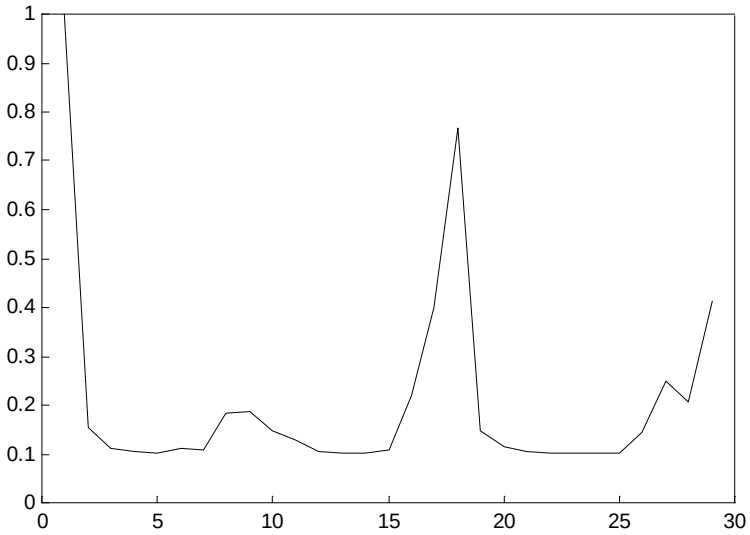



Figure 28: Covariance matrix $\Sigma_{k|k}$ in presence of 3 outliers with b=1.

Some other simulations have been made to test the robust filter in the presence of Laplacian noise. Because the best estimator at the Laplacian noise is the L1-norm, which

is equivalent to a Huber M-estimator with a parameter b equal to 0, we tried to decrease the parameter to see how the filter behaves.

Figure 29 and Figure 30 show the result of such a simulation when b=2. Here, it appears that the robust filter is able to smooth out some segments of the signal better than the classical filter. Besides, the covariance matrix keeps value very close to the final state value.

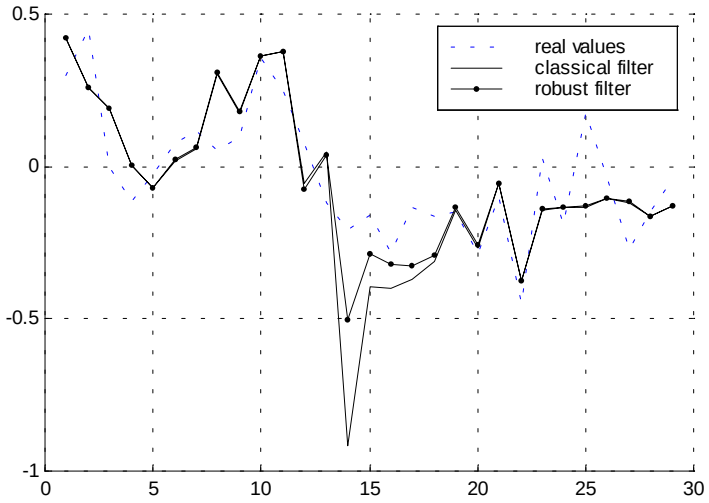



Figure 29: Output of the classical and robust Kalman filters (b=2) in presence of Laplacian noise.

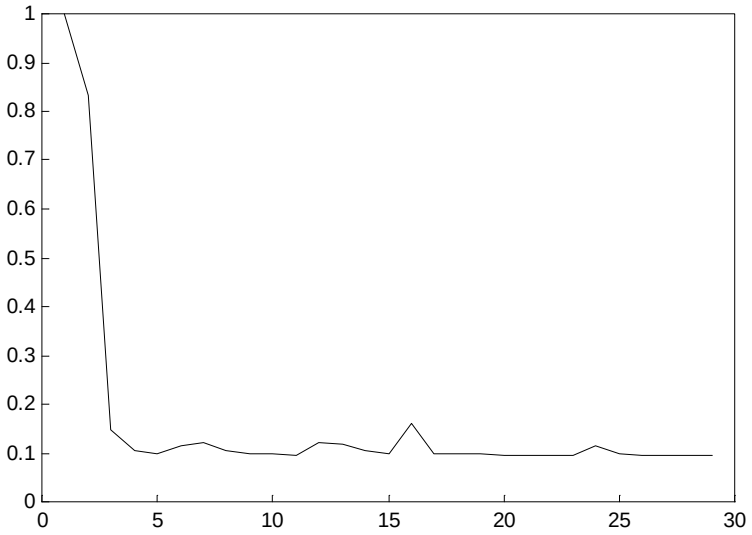



Figure 30: Covariance matrix $\Sigma_{k|k}$ in presence of Laplacian noise with b=2.

The simulations were pursued with b=0.1. The results are displayed in Figure 31 and Figure 32. The robust filter treats almost every value of the noisy signal as an outlier, and allocates very small weights to the observations, which increases the values of the covariance matrix. The robust filter is no longer an improvement of the classical filter. This is due to a poor estimation of the covariance matrix , $\Sigma_{k|k}$ .

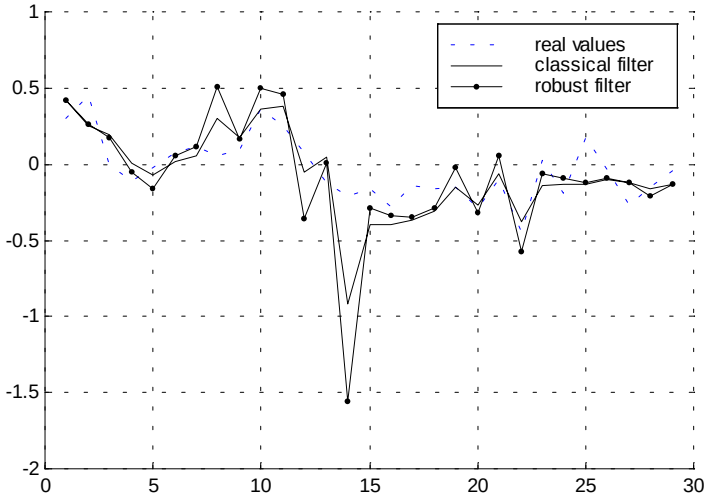



Figure 31: Output of the classical and robust
Kalman filters (b=0.1) in presence of Laplacian noise.

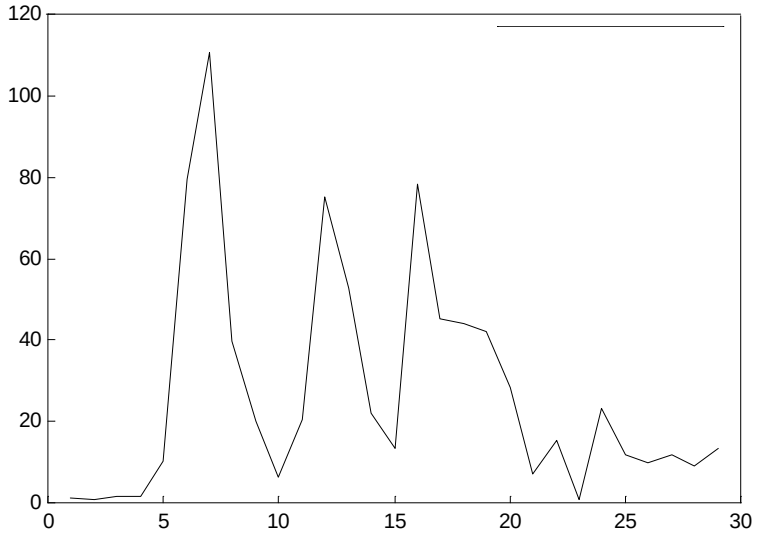



Figure 32: Covariance matrix $\Sigma_{k|k}$ in presence of Laplacian noise with b=0.1.

### 5.5. Conclusions

A robust version of the Kalman filter was presented. It may be seen as a Huber M-estimator in a generalized linear regression. Some simulation results have been applied to a autoregressive model of order 3 to compare the conventional Kalman filter and the robust filter. The result given by the robust filter appeared to be better in terms of SNR when some outliers were introduced among Gaussian noise. The study also showed that the cutoff value of the Huber M-estimator has to be chosen not too small to maintain a good accuracy of the covariance matrix estimate and thereby, to have good convergence properties of the filter.

# Chapter 6

# Simulation results

## 6.1. Introduction

This Chapter presents a robust filtering method based on both speech processing and robust statistics. In a first step, a robust LPC method is developed. The segments of the signal corrupted by impulsive noise are detected, the parameters modeling the signal are evaluated, and bad data are replaced with estimated values. In a second step, the robust version of the Kalman filter derived in Chapter 5 is tested. Unfortunately, results were not successful because of the inaccuracy and the lack of redundancy of the model. However, the filtering using the robust LPC method is able to detect most of the pulses, which has been clearly confirmed by listening tests.

## 6.2. Detection of impulsive noise

### 6.2.1. Segmentation of the signal

The classical method for signal segmentation was explained in Chapter 2. In this chapter, some modifications of this method are introduced to detect impulsive noise, which requires redundancy.

The detection of impulsive noise is done using Projection Statistics, as described in [4, 9, 25, 34, 35]. An observation matrix H containing values of the signal is built and each row of this matrix is associated with its projection statistic. The most important point in this process is to have enough redundancy: we need a sample of the signal in which there are enough good values compared to bad values. The strength of Projection Statistics is precisely that the breakdown point is very high. According to our experience,

only twice as much good values as bad values is enough to get good results. Generally, the length of a pulse does not exceed 40 consecutive samples. Therefore, a window of 120 samples seems long enough under the assumption that only one pulse will occur in such a small period of time.

Usually, the matrix H is built so that there is only one unit time difference between 2 consecutive rows: if the first row contains consecutive samples from 1 to m, the second row will contain samples from 2 to m+1. Therefore, if the data point s(m) of the signal is wrong, the m first rows of the matrix H will be corrupted. To avoid this problem, we chose to build H so that each value of the signal only appears once:

$$H = \begin{bmatrix} s(m) & \cdots & s(2) & s(1) \\ s(2m) & \cdots & s(m+2) & s(m+1) \\ \cdots & \cdots & \cdots & \cdots \\ s(N-1) & \cdots & s(N-m+1) & s(N-m) \end{bmatrix}. \qquad (134)$$

The percentage of Projection Statistics susceptible of being too high for this matrix will be equivalent to the one we would have had if the classical method were to be used. However, the situation where we have only one bad data in a row, which is more difficult to detect because the associated Projection Statistic will not be very large, will occur less often. Another advantage of this method is that we will have to calculate m times less Projection Statistics, which will decrease computing time.

The classical method recommends using a Hamming window. But here, the amplitude of the signal is the key of the detection of pulses. Therefore, a rectangular window was used. While estimates of the coefficients may be less accurate, the probability that some pulses stay undetected will be smaller.

After testing different values for m and N, we finally opted for m=4 and N=128, which is equivalent to a window of 16 ms.

The value of m is smaller than what is usually advocated in the literature. But here again, we are interested in a method that is robust rather than accurate. To take a small value of m presents some advantages in this respect. First, in the case where only one value in a row is wrong, it increases the fraction of bad data in this row, equal to 1/m. Therefore the wrong data will have more influence and will be easier to detect. Then, for a fixed value of N, decreasing m will increase the redundancy. In our case, we pick a

redundancy of 8, specifically, we consider a matrix H with 32 rows, which is 8 times the number of parameters we are estimating.

The value of N is also quite small. We came to this solution after doing some simulations, but it is motivated by the fact that we chose a small value for m. Indeed, a small value for m will decrease the accuracy of the coefficients while a small value of N will increase it. Therefore, the small value for N can be seen as a way to compensate the small value chosen for m.

### 6.2.2. Calculation of Projections Statistics

Once the matrix H has been built, the Projection Statistics algorithm is used to detect outliers among the row vectors of H.

Table 3 gives the range of the Projection Statistics calculated for the noisy signal already taken as an example in Chapter 2. Figure 33 is the corresponding histogram. We already mentioned that Projection Statistics roughly follow a $\chi_n^2$ distribution, where n is here equal to 4. We chose a threshold value equal to $\chi^2_{\mathrm{n},0.95} = 9.49$. From Table 3, we see that the presence of impulsive noise turns into values of Projection Statistics much higher than this threshold.

It can be observed from Figures 34-36 that the developed robust filter does a good job in suppressing impulses. Compare the clean signal displayed in Figure 34 with the impulsive signal in Figure 35 and the filtered signal depicted in Figure 36.

Table 3: Frequencies of the Projection Statistics of the speech signal shown in Figure 35.

| Range | Frequency | Range | Frequency | Range | Frequency | Range | Frequency |
|---|---|---|---|---|---|---|---|
| 0-2 | 538 | 20-22 | 6 | 40-50 | 7 | 140-150 | 1 |
| 2-4 | 1046 | 22-24 | 5 | 50-60 | 10 | 150-160 | 0 |
| 4-6 | 436 | 24-26 | 5 | 60-70 | 5 | 160-170 | 0 |
| 6-8 | 135 | 26-28 | 2 | 70-80 | 2 | 170-180 | 1 |
| 8-10 | 44 | 28-30 | 6 | 80-90 | 4 | | |
| 10-12 | 24 | 30-32 | 2 | 90-100 | 0 | | |
| 12-14 | 15 | 32-34 | 3 | 100-110 | 1 | | |
| 14-16 | 9 | 34-36 | 3 | 110-120 | 1 | | |
| 16-18 | 7 | 36-38 | 4 | 120-130 | 3 | | |
| 18-20 | 7 | 38-40 | 2 | 130-140 | 2 | | |

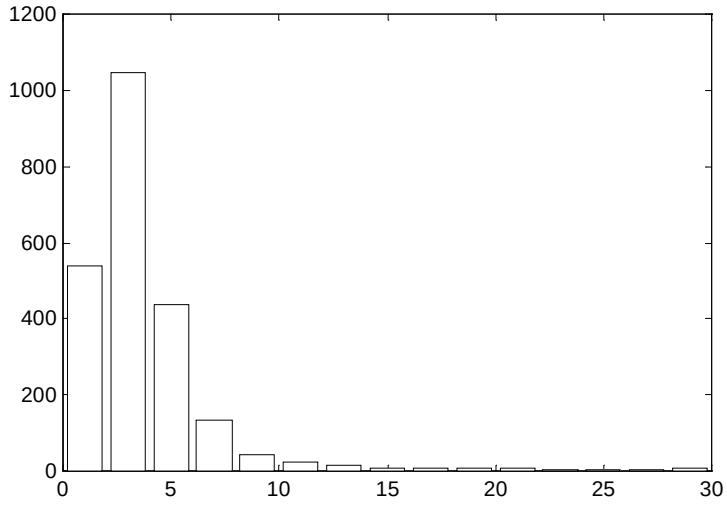



Figure 33: Histogram of the Projection Statistics for the noisy signal

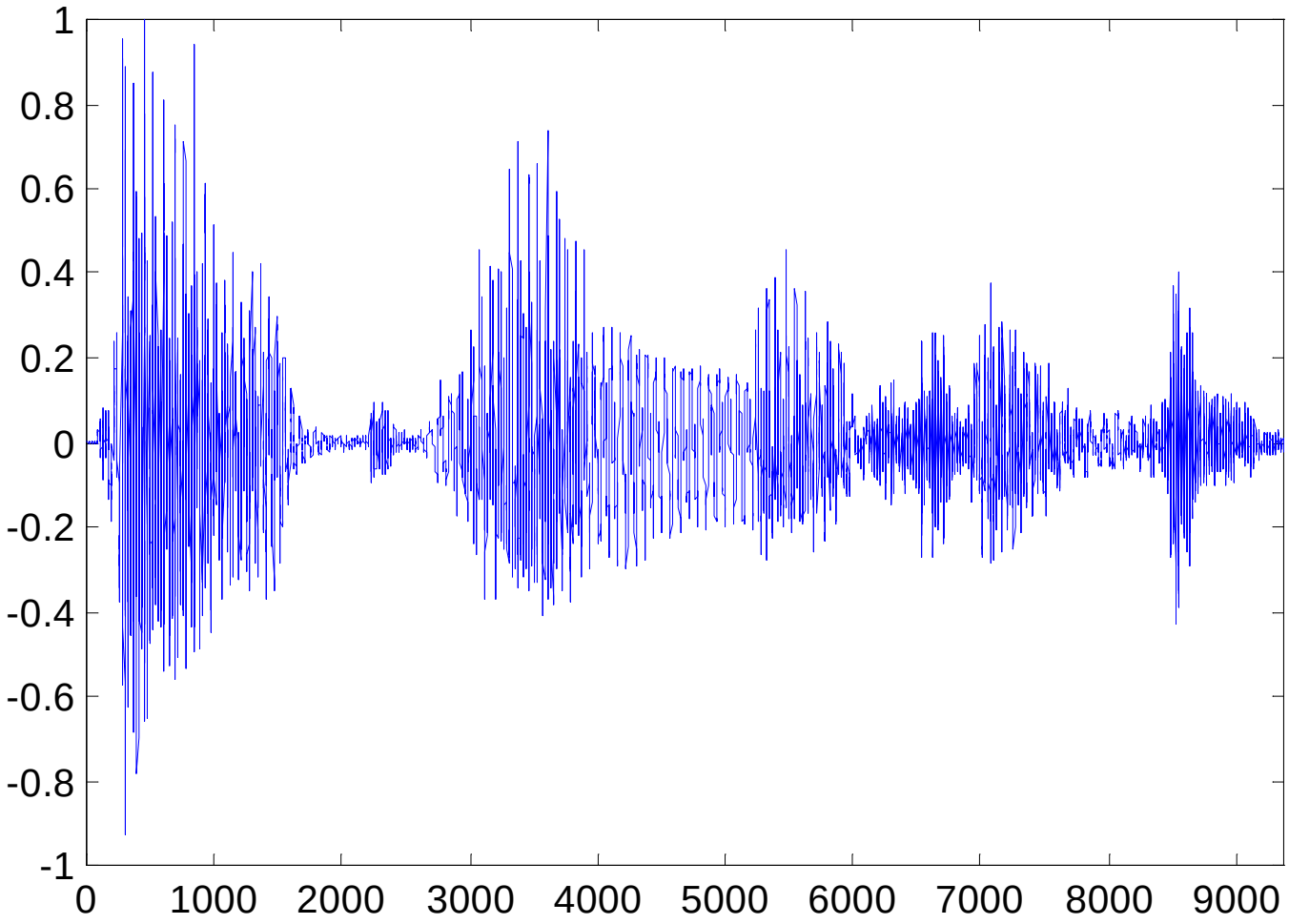



Figure 34: Clean speech signal

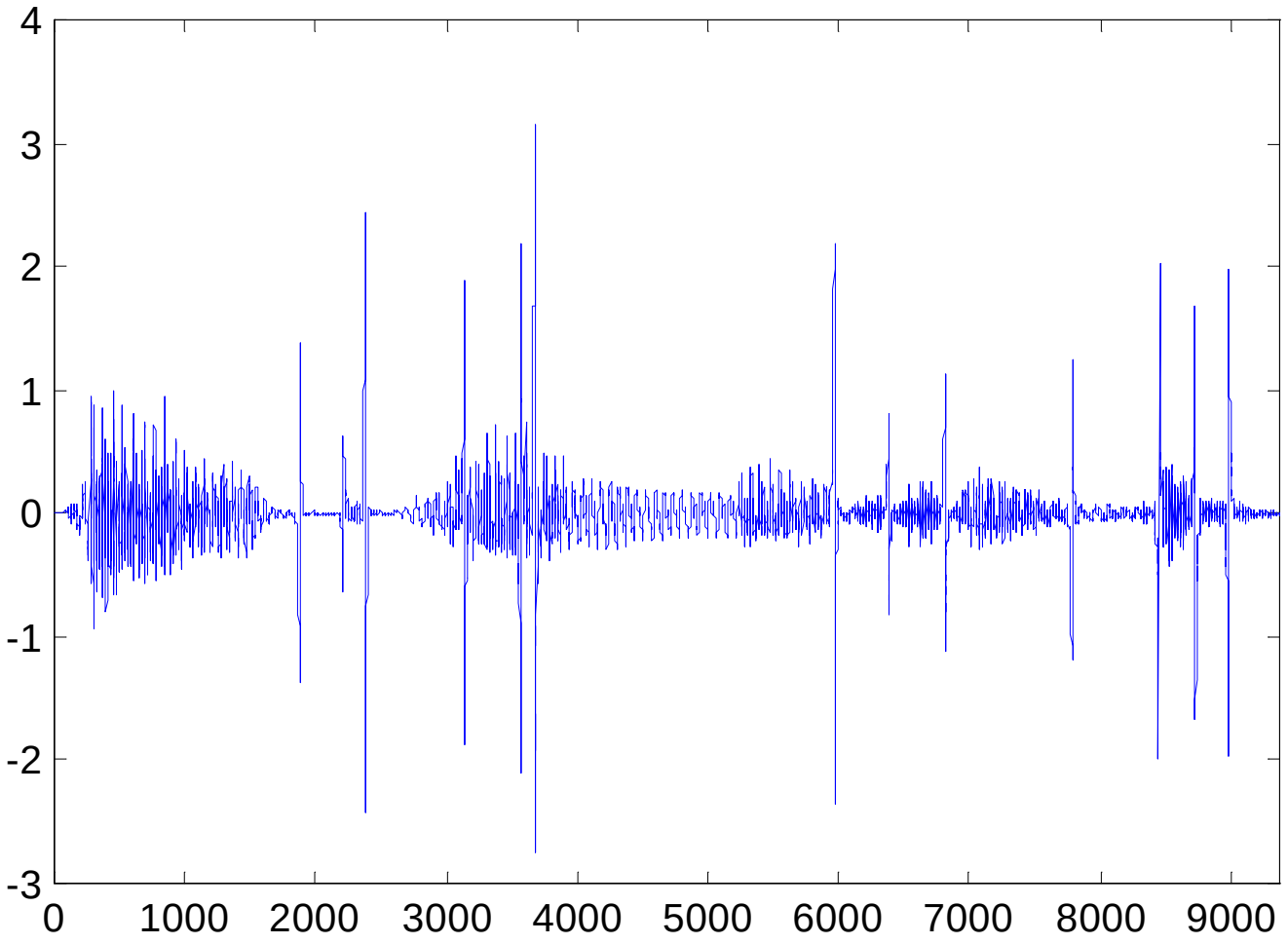



Figure 35: Noisy speech signal

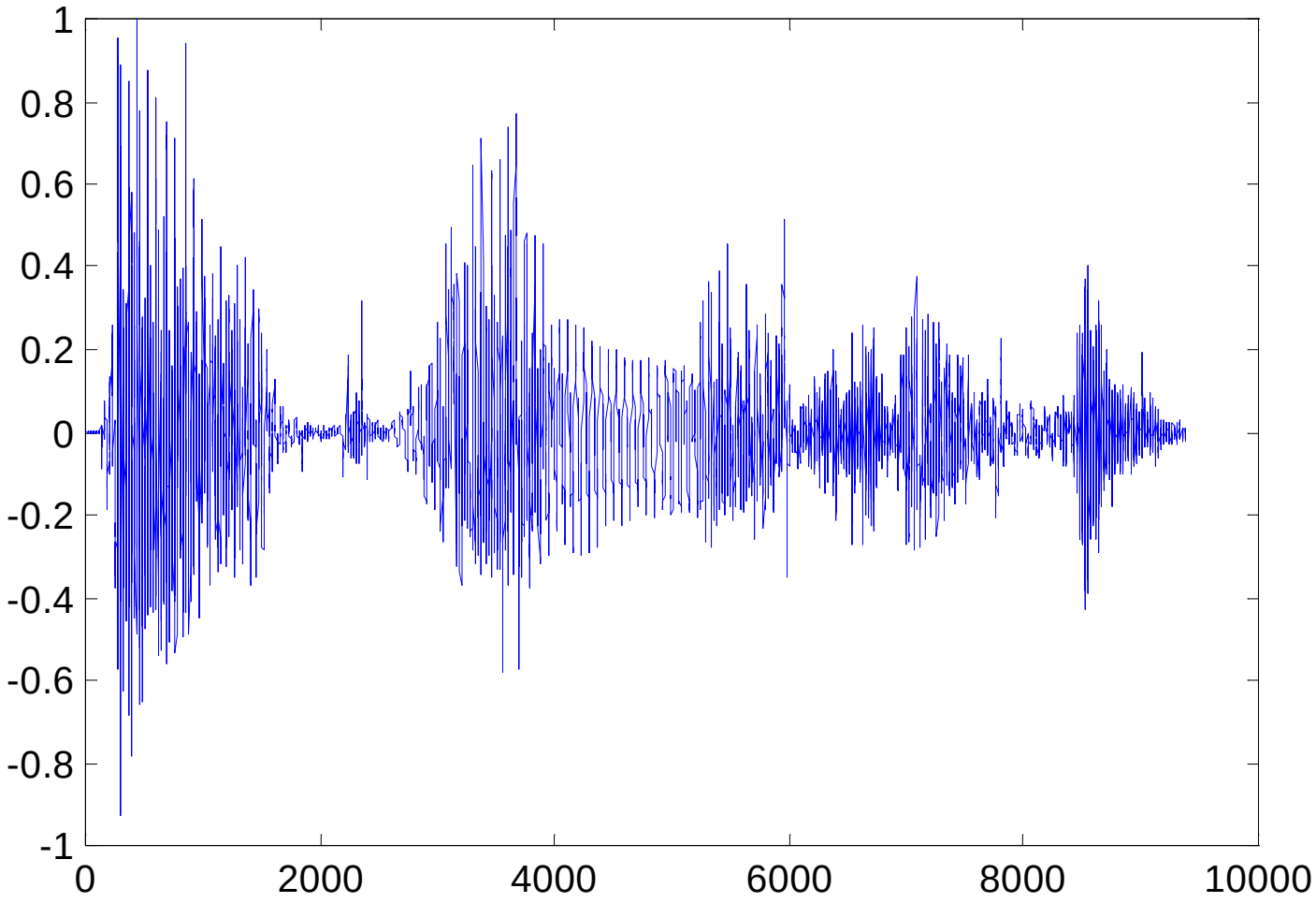



Figure 36: Filtered speech signal

Figures 37 and 38 allow us to better evaluate the improvement realized after filtering. Figure 37 is an histogram of the absolute values of the errors between the clean signal and the noisy signal. Figure 38 is the same histogram for the errors between the clean signal and the filtered signal. It appears very clearly that the errors have been decreased. While the maximum error is almost equal to 3 in the first case, it is below 1 in the second case.

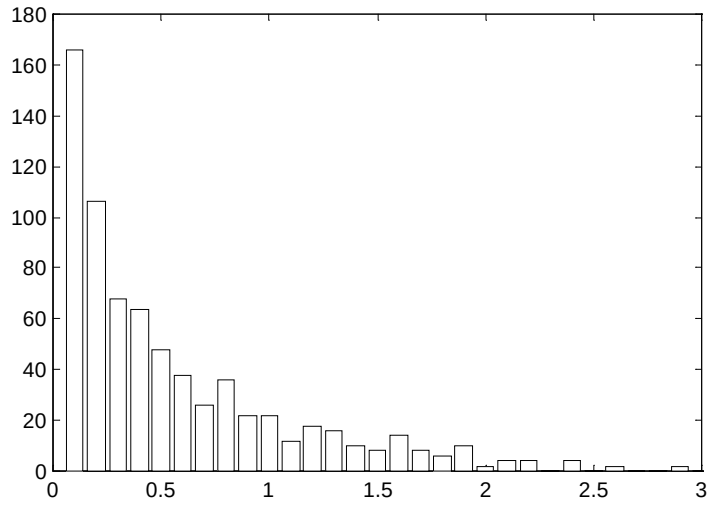



Figure 37: Histogram of the absolute values of the errors between the clean and the noisy signals

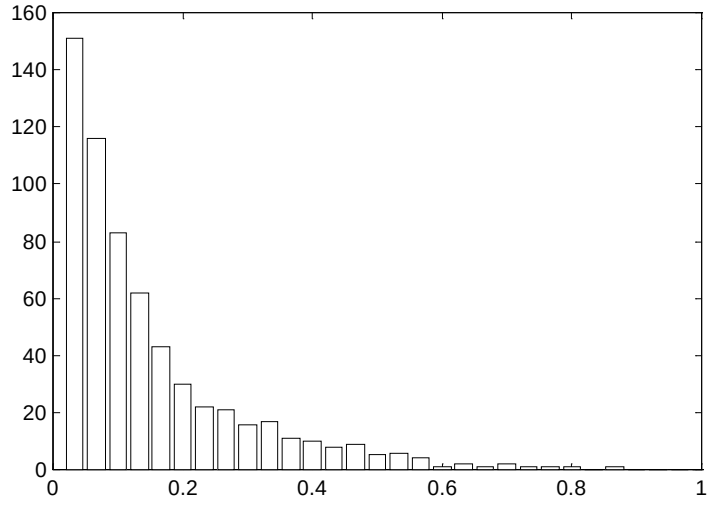



Figure 38: Histogram of the absolute values of the errors between the clean and the filtered signals

To better understand the detection process, we focused on a segment of the signal where there is a pulse. Figure 39 depicts the clean and the impulsive signals between unit times 3541 and 3601. The associated Projection Statistics are plotted on Figure 40.

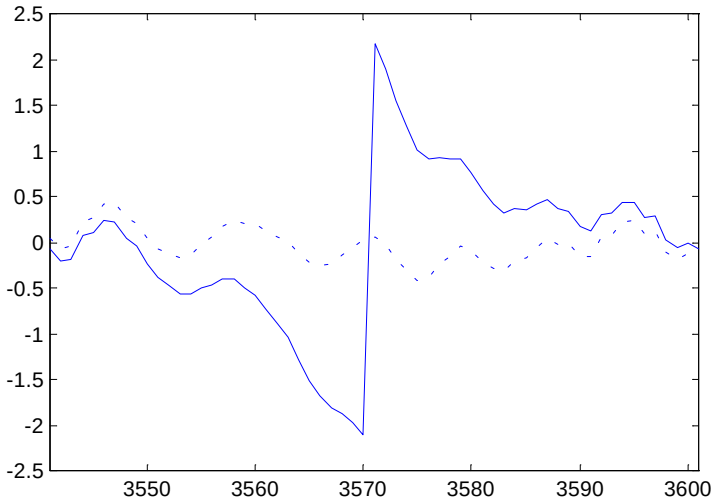



Figure 39: Zooming on the clean signal (dotted line) and noisy signal (solid line)

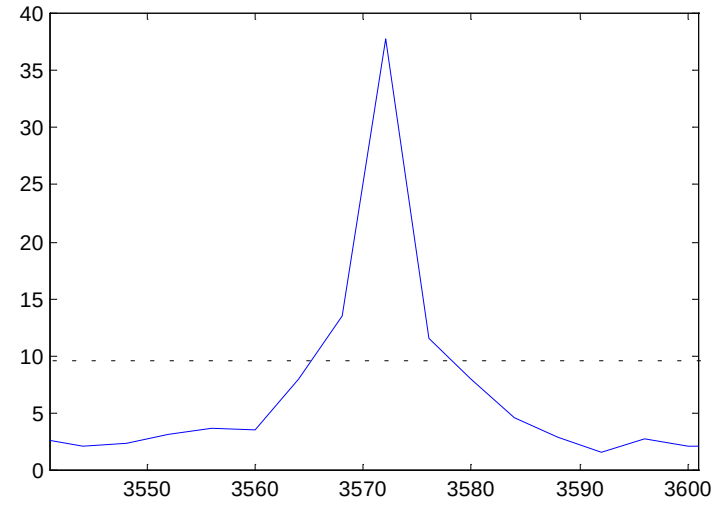



Figure 40: Projection Statistics associated with the pulse of Figure 39, where the cutoff value is shown as a dotted line.

It can be noticed that the beginning and the end of the pulse are not actually detected. Therefore, the decision was made to estimate the values of the signal for the

rows preceding and following a region where impulsive noise was detected. Besides, we tried to avoid that some values of the signal were considered as impulsive noise when it was not the case. To this end, the algorithm does not replace a portion of the signal with estimated values whose Projection Statistic is larger than the threshold but where the precedent and subsequent data points have Projection Statistics smaller than the threshold.

### 6.3. Reconstruction of the signal

In segments of the signal where impulsive noise has been detected, we need to estimate new values. The parameters of the autoregressive model representing the signal can be found by means of a linear regression estimator. Instead of using the Least Squares estimator as it is usually done, we resort to the Schweppe-type Huber GM-estimator. This estimator is applied over a window of length N, yielding

$$\begin{bmatrix} s(n+m+1) \\ s(n+m+2) \\ \cdots \\ s(n+N) \end{bmatrix} = \begin{bmatrix} s(n+m) & s(n+m-1) & \cdots & s(n+1) \\ s(n+2m) & s(n+2m-1) & \cdots & s(n+m+1) \\ \cdots & \cdots & \cdots & \cdots \\ s(n+N-1) & s(n+N-2) & \cdots & s(n+N-m) \end{bmatrix} \cdot \begin{bmatrix} a_1 \\ a_2 \\ \cdots \\ a_m \end{bmatrix} + \begin{bmatrix} e(n+m+1) \\ e(n+m+2) \\ \cdots \\ e(n+N) \end{bmatrix} \qquad (135)$$

Using the Schweppe-type Huber GM-estimator bounds the influence of outliers among the measurements, be they leverage points or not. Figure 41 represents the weights $w(h_i)$ calculated for the rows of the H matrix corresponding to the pulse shown in Figure 39. We see that the amplitude of the peak is so high that the corresponding weight is put to zero.

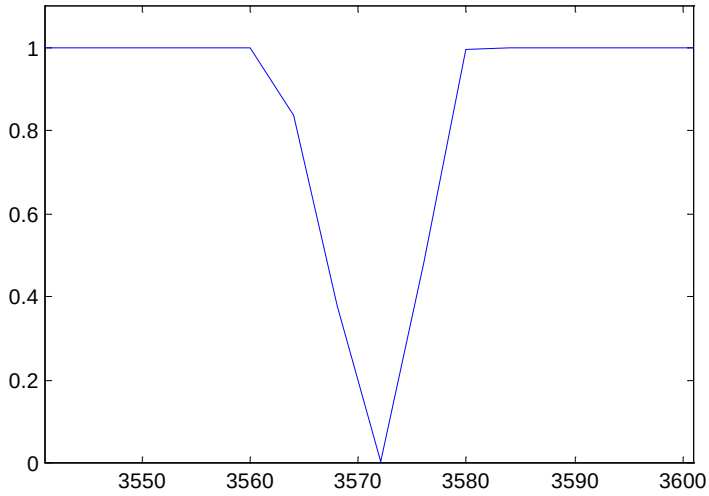



Figure 41: Weights $w(h_i)$ of an impulsive portion of the speech signal.

The estimation of the parameters of an AR(4) is not very accurate because the order of the model is not large enough. Applying a second higher order AR model would improve the estimates once the pulses have been removed; however, it would double the computing times. Moreover, the signal is to be reconstructed over portions that last around 4 ms, which will yield small improvements that may not be audible.

Once the AR(4) is identified, it is used to reconstruct the impulsive portion of the speech signal. To this end, we excite the AR(4) by a Gaussian noise with a standard deviation equal to a robust scale estimate of the signal over the appropriate window. The latter has been calculated by means of the median-absolute-deviation-from-the median (MAD) after having deleted the impulses.

Figure 42 shows the reconstructed signal on the segment already studied. We see that the impulses have been removed in a quite efficient way and that the trend of the estimated values is good. Figure 43 presents a different view, which allows us to better appreciate the quality of the smoothed signal. The amplitude of the clean signal on this portion reaches values around 0.6. It is probably the reason why the signal was not changed until a sampling time of 3560. At least the reconstruction allows us to smoothly replace the contaminated portion of the signal with a signal that has a similar frequency. Figure 44 provides a comparison of the spectra. The improvement is obvious since the spectrum envelope of the filtered signal fits quite well that of the clean signal.

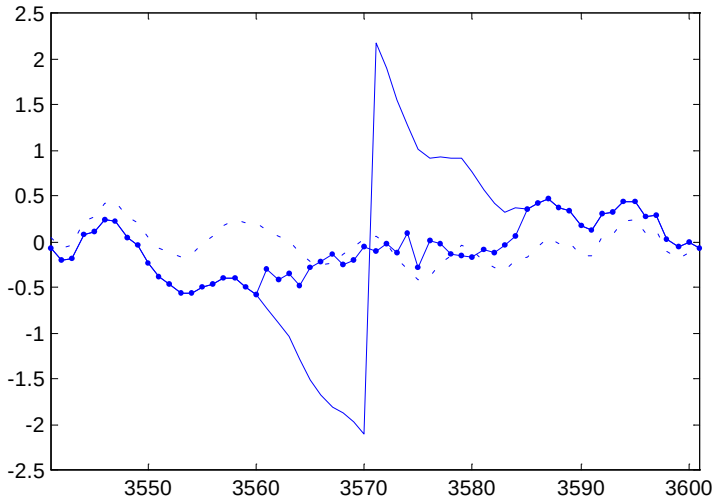



Figure 42: The filtered signal (solid line with point markers) is much closer to the clean signal (dotted line) than the noisy signal (solid line).

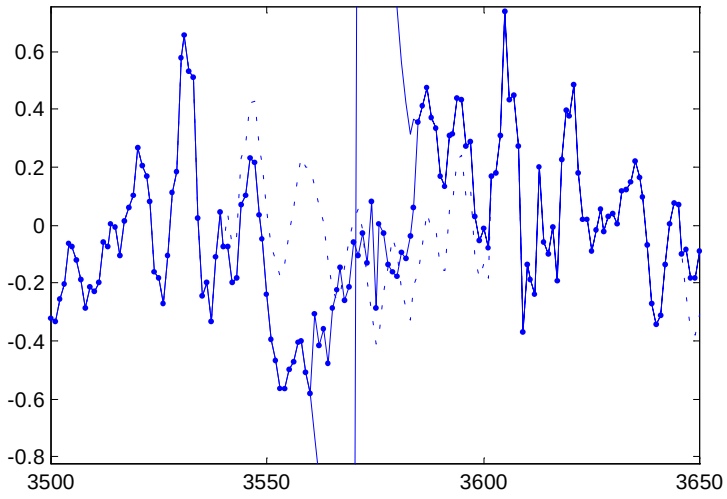



Figure 43: A zoom of the portion of the signal shown in Figure 42.

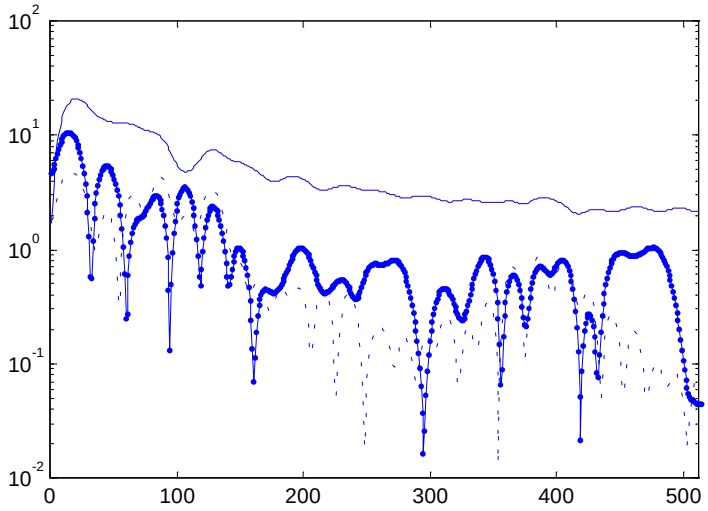



Figure 44: Spectrum of the clean signal (dotted line), noisy signal (solid line), and filtered signal (solid line with point markers) over the portion shown in Figure 42.

Similar observations can be made all along the signal. Three examples are given in Figures 45-50. They focus on the portions of the signal around the sample time 6380 (Figure 45 and Figure 46), 5970 (Figure 47 and Figure 48), and 8730 (Figure 49 and Figure 50). A listening test confirms that a real improvement has been realized: the clicks due to the presence of the impulsive noise are much less audible.

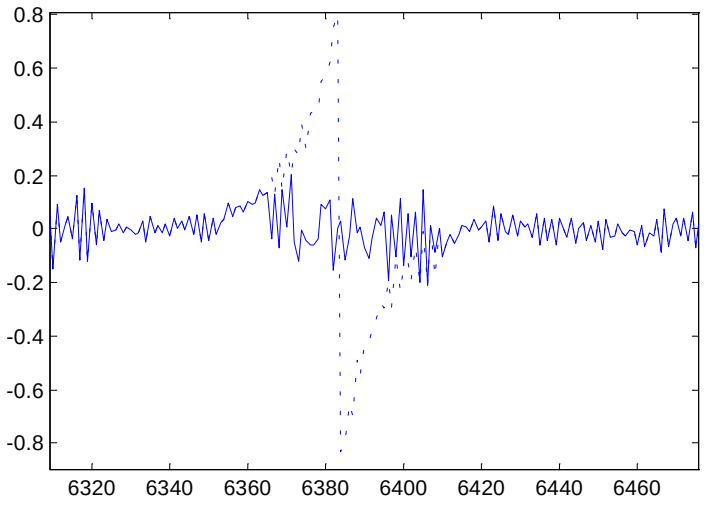



Figure 45: Example 1 in the time domain
(noisy signal: dotted line, filtered signal: solid line)

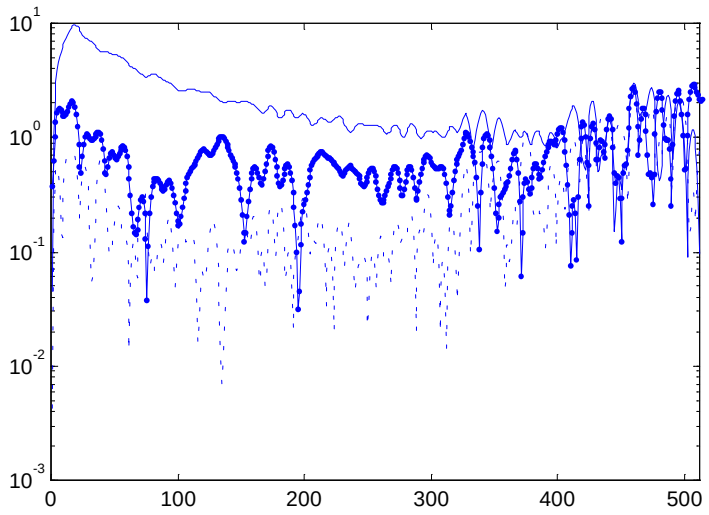



Figure 46: Example 1 in the frequency domain (clean signal: dotted line, noisy signal: solid line, filtered signal: solid line with point markers)

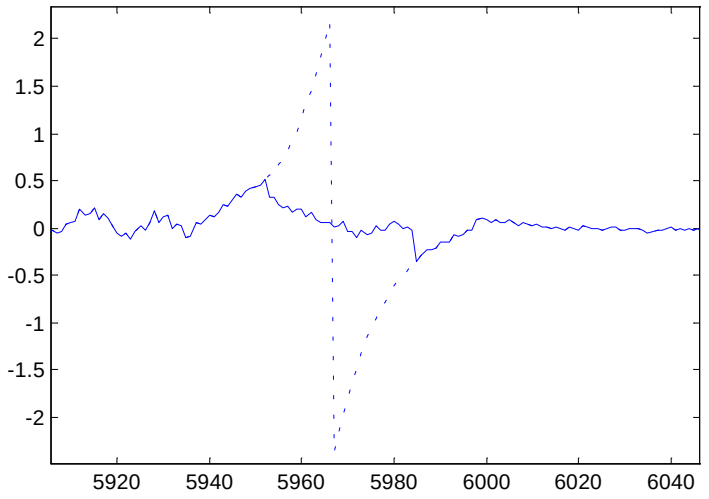



Figure 47: Example 2 in the time domain
(noisy signal: dotted line, filtered signal: solid line)

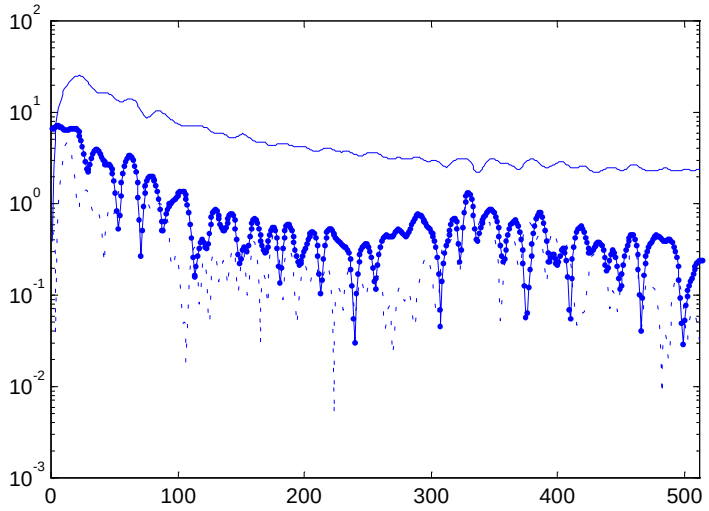



Figure 48: Example 2 in the frequency domain (clean signal: dotted line, noisy signal: solid line, filtered signal: solid line with point markers)

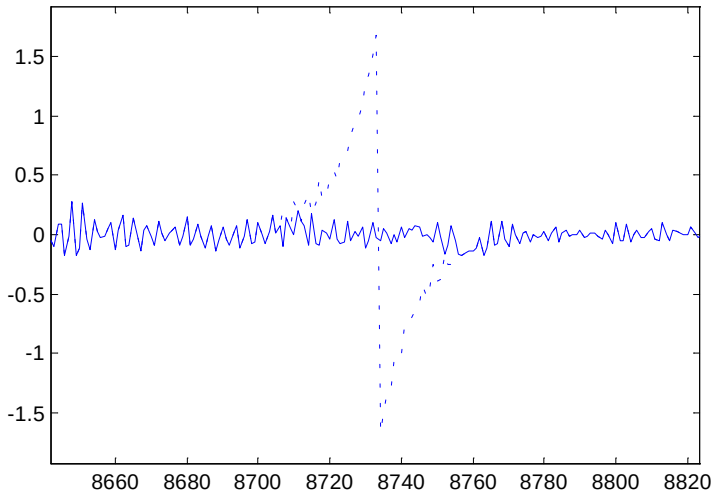



Figure 49: Example 3 in the time domain
(noisy signal: dotted line, filtered signal: solid line)

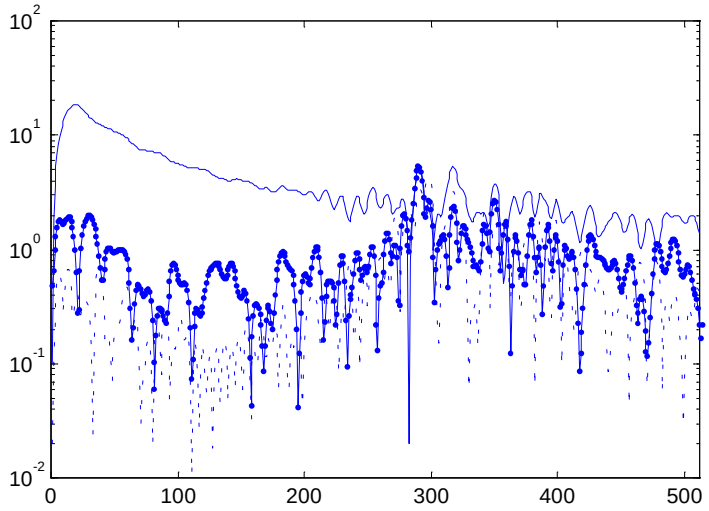



Figure 50: Example 3 in the frequency domain(clean signal: dotted line, noisy signal: solid line, filtered signal: solid line with point markers)

### 6.4. Robust Kalman filter

The robust Kalman filter developed in Chapter 5, and based on the Huber M-estimator, can present a double advantage in the filtering of impulsive noise in speech signals. First, it should be able to remove the impulsive noise that may remain after the filtering using the robust LPC method. Second, it should be more efficient than the classical Kalman filter because of the nature of speech. Indeed, speech signals have a probability distribution that is very close to a Laplacian distribution, as is illustrated in [30]. Because the L1-norm estimator is the best estimator at this distribution and that the Huber M-estimator behaves as the L1-norm for high residuals, the efficiency of the robust Kalman filter should be better than that of the conventional one.

The Kalman filter relies on the model of the dynamic system that is studied. The principle is to predict values based on this model and to consider them as supplementary observations. If the uncertainty on the model is higher than that on the observations, the filtering is obviously not possible. Therefore, we need an accurate model. This can be done by proceeding to an LPC filtering, with a number of parameters between 10 to 14. Let us assume that over each window, 12 parameters are to be estimated from 12 predicted values and 1 observation. Hence, we have to find 12 new estimates from a set

of 13 data, which provides a very weak redundancy of 13/12. In the example of Chapter 5, it was shown that a redundancy of 4/3 enabled us to significantly attenuate pulses. Here, the redundancy appears to be too small, and the robust filter does not behave better than the classical filter.

To increase the redundancy, it would be necessary to decrease the number of parameters to model the system. However, the complexity of a speech signal does not allow us to choose an AR model with an order smaller than 10. If the order is smaller, the model becomes so inaccurate that noisy observations are more reliable than predicted values. Some more work would have to be done to solve the tradeoff between accuracy and robustness of the filter.

### 6.5. Conclusions

This Chapter presented a method to filter impulsive noise in a speech signal using robust statistics. Simulation results revealed that the method is very effective in detecting pulses. The smoothing of the corrupted segments was made by means of the Schweppe-type Huber GM-estimator. Listening tests confirmed that a significant improvement was achieved.

# Chapter 7

# Conclusions

This thesis presented a new robust method to suppress impulsive noise in speech signals. In a first part, an introduction to speech processing was given. The principle of speech production and the Linear Prediction Coding method were explained. Particular attention was paid to the problem of impulsive noise, exposing its sources, the way it can be modeled, and the existing techniques to suppress it. Then, a review of parametric and robust estimation methods was given. The main properties of estimators, like efficiency, were presented, and the class of the ML-estimators defined. Robust concepts were explained, which led to the class of M-estimators. The interest of the Huber M-estimator was particularly underlined. The problem of linear regression was exposed, including detection of leverage points and vertical outliers. A calculation of Robust Distances by means of Projection Statistics was seen and some simulations were made to illustrate the theoretical results. Finally, the Schweppe-type Huber GM-estimator was presented, which appears as the appropriate method to solve a linear regression in an impulsive noise environment. Another Chapter was devoted to the derivation of a robust Kalman filter using the Huber M-estimator. It was tested on a dynamic system modeled by an autoregressive model of order 3. A significant attenuation of the pulses was noticed.

All the preceding techniques were presented and tested on simple examples in the perspective of combining them and deriving a method to filter impulsive noise in speech signals. The procedure that was developed can be seen as a robust LPC method. It is based on the classical LPC, but modifications were made to adapt it to the removal of impulsive noise.

The method can be divided into two steps, namely, impulsive noise detection and signal reconstruction. For the detection step, the Projection Statistics were used. After

having segmented the signal into portions of 16 ms, Projection Statistics were calculated for groups of 4 consecutive data. Because the probability distribution of these Projection Statistics is roughly a $\chi^2_{n=4}$ distribution, a threshold value equal to $\chi^2_{n=4,0.95}$ was chosen. Portions of the signal that had a Projection Statistic value higher than this threshold were considered as susceptible of containing impulsive noise.

Because the beginning and the end of pulses are difficult to detect accurately, the algorithm considered that a portion of the signal preceding or following a region with high Projection Statistic had to be reestimated too. On the other hand, some sudden variations in a speech signal can be identified as impulsive noise by mistake. If the Projection Statistic of a group of 4 data was above the threshold and those of the preceding and subsequent groups were not, the signal was not smoothed out.

To reconstruct the signal, the parameters were estimated by means of the Schweppe-type Huber GM-estimator. The priority was given to the robustness of the method and not to its accuracy. Therefore, the order of the AR model was kept equal to 4. A simple model of the excitation was proposed using a Gaussian process. Its variance was set equal to the variance of the signal, estimated in a robust way for each segment of 16 ms. Estimated values of the signal were calculated for each portion of the signal found to be corrupted by impulses. Each time, the standardized residuals were calculated and the value of the signal was replaced by its estimate if the associated standardized residual was higher than 3.

The results of this method showed that most of the impulsive noise has been removed, and that only few non-noisy values were smoothed out by mistake. The reconstruction of the signal provided a good approximation, especially in the frequency domain. The listening comparison of the noisy signal and the filtered signal confirmed these observations.

A more extensive research could be done to improve the algorithm, especially in the reconstruction step. An AR model of higher order would probably lead to a better accuracy. Besides, some more work could be done on the robust Kalman filter. Indeed, the filter that was developed in this thesis did not provide interesting results when used to suppress impulses. The redundancy available in the system was not large enough to notice an improvement between the classical Kalman filter and the robust one.

This study is also a proof of the interest that relies in combining engineering and statistical methods. The solution that was given is really based on both speech processing techniques and robust estimation. A large number of problems can be approached this way and give birth to simple and efficient methods.

# References


[1] Brian D.O. Anderson and John B. Moore, *Optimal Filtering*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1979.

[2] Nelson M. Blachman, *Noise and its Effect on Communication*, McGraw-Hill, New York, 1966.

[3] George E. P. Box, and Gwilym M. Jenkins, *Time Series Analysis; Forecasting and Control*, Holden-Day, San Francisco, 1970.

[4] D.L. Donoho, "Breakdown Properties of Multivariate Location Estimators", qualifying paper, Harvard University, Boston, Maryland, 1982.

[5] Garry A. Einicke and Langford B. White, "Robust Model Based Methods for Speech Enhancement", *Proc. 1997 IEEE, Trends in Electronics Conference*, Vol. 2, pp. 471-474, 1997.

[6] Sadaoki Furui and M. Mohan Sondhi, *Advances in Speech Signal Processing*, Marcel Dekker, Inc., New York, 1992.

[7] N.C. Gallagher and G.L. Wise, "A Theoretical Analysis of the Properties of Median Filters", *IEEE Trans. On Acoustics, Speech and Signal Processing*, Vol. 29, pp. 1136-1141, 1981.

[8] Sharon Gannot, David Burshtein and Ehud Weinstein, "Iterative and Sequential Kalman Filter-Based Speech Enhancement Algorithms", *IEEE Trans. Speech and Audio Proc*., Vol. 6, No. 4, pp. 373-385, July 1998.

[9 ] M. Gascot, and D. Donoho, "Influential Observation in Data Analysis", *American Statistical Association, Proceedings of the Business and Economic Statistics Section*, pp. 104-110, 1982.

[10] Arthur Gelb and others, *Applied Optimal Estimation*, M.I.T. Press, Cambridge, Massachusetts, 1974.

[11] Jerry D. Gibson, Boneung Koo and Steven D. Gray, "Filtering of Colored Noise for Speech Enhancement and Coding", *IEEE Trans. Signal Processing*, Vol. 39, No. 8, pp. 1732-1741, Aug. 1991.

[12] F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw and W.A. Stahel, *Robust Statistics: The Approach Based on Influence Functions*, John Wiley and Sons, Inc., New York, 1986.

[13] E. Handschin, F.C. Schweppe, J. Kohlas and A. Fiechter, "Bad Data Analysis for Power System State Estimation", *IEEE Transactions on Power Apparatus and Systems*, Vol. 94, No. 2, pp. 329-337, March/April 1975.

[14] Peter J. Huber, *Robust Statistics*, John Wiley and Sons, Inc., New York, 1981.

[15] B.H. Juang and others, "The Past, Present, and Future of Speech Processing", *IEEE Signal Processing Magazine*, Vol. 1053, pp. 24-28, May 1998.

[16] James L. Lansford, "Lp Models in Speech Coding and Markov Chains in Speech Recognition", Ph.D. dissertation, Oklahoma State University, 1988.

[17] Nhu D. Le, Adrian E. Raftery and R. Douglas Martin, "Robust Bayesian Model Selection for Autoregressive Processes With Additive outliers", *Journal of the American Statistical Association*, Vol. 91, No. 433, March 1996.

[18] Ki Y. Lee and others, "Robust Estimation of AR Parameters and its Application for Speech Enhancement", *Proc. 1992 IEEE, Int. Conf. Acoustics, Speech, and Signal Proc.*, Vol. 1, pp. 309-312, 1992.

[19] E.L. Lehmann, *Theory of Point Estimation*, John Wiley and Sons, 1983.

[20] Jae S. Lim, *Speech Enhancement*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1983.

[21] Lennart Ljung, *System Identification: Theory for the User*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1987.

[22] John D. Markel and A. H. Jr. Gray, *Linear Prediction of Speech*, Springer-Verlag, Berlin, New York, 1976.

[23] C. Johan Masreliez and R. Douglas Martin, "Robust Bayesian Estimation for the Linear Model and Robustifying the Kalman Filter", *IEEE Trans. on Automatic Control*, Vol. 22, pp. 361-371, June 1977.

[24] Jerry M. Mendel, *Lessons in Digital Estimation Theory*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1987.

[25] L. Mili and al., "Robust Estimation Based on Projection Statistics", *IEEE Trans. on Power Systems*, Vol. 11, No. 2, pp. 1118-1127, May 1996.

[26] L. Mili, V. Phaniraj and P.J. Rousseeuw, "Robust Estimation Theory for Bad Data Diagnostics in Electric Power Systems", in *Control and Dynamic Systems: Advances in Industrial Systems*, C.T. Leondes, Academic Press, 1990.

[27] Russell J. Niederjohn and James A. Heinen, "Understanding Speech Corrupted by Noise", *Proc. 1996 IEEE, Int. Conf. Industrial Technology*, pp P1-P5, 1996.

[28] K. K. Paliwal, and A. Basu, "A speech Enhancement Method Based on Kalman Filtering", *Proc. 1987 IEEE, Int. Conf. Acoustics, Speech, and Signal Processing*, Vol. 1, pp. 177-180, 1987.

[29] Panos E. Papamichalis, *Practical Approaches to Speech Coding*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1987.

[30] Lawrence R. Rabiner and Ronald W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1978.

[31] C.R. Rao, *Linear Statistical Inference and Its Applications*, 2nd edition, John Wiley and Sons, 1973.

[32] Theodore S. Rappaport, *Wireless Communications: Principles and Practice*, Prentice-Hall PTR, Upper Saddle River, New Jersey, 1996.

[33] Brian D. Ripley, *Stochastic Simulations*, John Wiley and Sons, Inc., New York, 1987.

[34] P.J. Rousseeuw and B.C. Van Zomeren, "Unmasking Multivariate Outliers and Leverage Points", *Journal of the American Statistical Association*, Vol. 85, No. 411, pp. 633-651, Sept. 1990.

[35 ] W.A. Stahel, "Robuste Schätzungen: Infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen", Ph.D. Dissertation, E.T.H. Zürich, Swiss, 1981.

[36] G.C. Tiao, "Autoregressive Moving Average Models, Intervention Problems and Outlier Detection in Time Series", in E.J. Hannan, P.R. Krishnaiah and M.M. Rao, eds., *Handbook of Statistics*, Vol. 5, pp 85-118, 1985.

[37] Saeed V. Vaseghi, *Advanced Signal Processing and Digital Noise Reduction*, John Wiley and Sons Ltd, Chichester, England, and B.G. Teubner, Stuttgart, Germany, 1996.

[38] Saeed V. Vaseghi and P.J.W. Rayner, “Detection and Suppression of Impulsive Noise in Speech Communication Systems”, *IEE Proc-I Communications Speech and Vision*, pp. 38-46, Feb. 1990.

[39] F.A. Westall, “Review of Speech Technologies for Telecommunications”, *Electronics and Communication Engineering Journal*, Vol. 95, pp. 197-207, Oct. 1997.

# Vita

Christelle Ledoux was born in Chambray-lès-Tours, France on January 22, 1977. She received a French Diplôme d'Ingénieur from the Ecole Nationale d'Ingénieurs Electriciens de Grenoble (ENSIEG), part of the Institut National Polytechniques de Grenoble (INPG) in September 1999. She is currently in the process of completing the requirements for the M.S. degree in Electrical Engineering at the Virginia Polytechnic Institute and State University. Her main interests are communications and robust statistics.