DISSERTATION

On Robust Spectral Density Estimation

ausgeführt zum Zwecke der Erlangung des akademischen Grades eines
Doktors der technischen Wissenschaften unter der Leitung von

O.Univ.-Prof. Dipl.-Ing. Dr.techn. Rudolf Dutter
E 107 – Institut für Statistik und Wahrscheinlichkeitstheorie

eingereicht an der Technischen Universität Wien,
Fakultät für Mathematik und Geoinformation

von

Dipl.-Ing. Bernhard Spangl
Matr.-Nr.: 9426095
1100 Wien, Ranzonigasse 1

Wien, am 10. September 2008   Bernhard Spangl

To my wife, Barbara, and my parents, Brigitte and Ferdinand,
and in memory of Harald Strelec

Kurzfassung

Im folgenden befassen wir uns mit der robusten Spektraldichteschätzung und ihrer Anwendung in der Herzratenvariabilitätsanalyse.

Da die klassischen Spektraldichteschätzer empfindlich auf Ausreißer reagieren, sind robuste Verfahren von Bedeutung. Daher konzentrieren wir uns auf die robuste Schätzung der Spektraldichtefunktion und stellen verschiedene, bereits existierende, aber auch neue Methoden vor, die robust gegenüber Ausreißern sind.

Wir betrachten Spektraldichteschätzer, die auf einer Robustifizierung der Fourier-Transformation und auf der robusten Schätzung der Autokovarianzfunktion basieren.

Um allerdings verlässliche Schätzwerte der Spektraldichtefunktion zu erhalten, die unempfindlich gegenüber Ausreißern sind, ist es besser, zuerst die Ausreißer mittels eines geeigneten robusten Verfahrens herauszufiltern und anschließend die Spektraldichtefunktion der bereinigten Zeitreihe zu berechnen.

Diese Bereinigung der Daten leistet ein robustifizierter Kalman-Filter. Wir schlagen dafür einen neuen multivariaten ACM-Typ Filter für Zustandsraummodelle vor. Dieser neuen Filter verallgemeinert den ursprünglichen ACM-Typ Filter, der auf eindimensionale Beobachtungen beschränkt ist. Unser neuer Filter wird mit einem weiteren Ansatz, dem gegenwärtig verwendeten rLS-Filter, verglichen.

Alle beschriebenden Methoden sind in der Programmiersprache R implementiert und werden mit Hilfe umfangreicher Simulationsstudien miteinander verglichen. Das beste Verfahren wird zur Analyse der Herzratenvariabilität bei Diabetikern mit unterschiedlich schwerer kardiovaskulärer autonomer Neuropathie herangezogen.

Abstract

In the following we deal with robust spectral density estimation and its application to the analysis of heart rate variability.

As classical spectral density estimators are sensitive to outlying observations, robustness is an issue. Hence, we focus on the problem of estimating the spectral density function robustly and present different methods, existing and new ones, that are resistant to outliers.

We consider spectral density estimators based on robustifying the Fourier transformation and on robust autocovariance estimation.

However, in order to get a reliable estimate of the spectral density function, that is insensitive to outlying observations, it turned out that cleaning the time series in a robust way first and calculating the spectral density function afterwards leads to encouraging results.

The data-cleaning operation wherein the robustness is introduced, is accomplished by a robustified version of the Kalman filter. We propose a new multivariate approximate conditional-mean (ACM) type filter for state-space models. This new filtering method generalizes the original ACM-type filter which is bound to the univariate setting. We compare our new filtering method to a second approach, the currently used rLS filter, which is also described.

All presented methods are implemented in the open source language R and compared by extensive simulation studies. The most competitive method is also applied to actual heart rate variability data of diabetic patients with different degrees of cardiovascular autonomic neuropathy.

Preface

The present research deals with robust spectral density estimation and its application to the analysis of heart rate variability (HRV). The analysis of heart rate variability as non-invasive method is increasingly used in medicine, especially when investigating cardiovascular autonomic neuropathy in diabetic patients. This work was motivated by the following remark on the frequency domain analysis of HRV in Pumprla et al. (2002):

> "Before this type of analysis can be performed, extensive editing and review of the electrocardiogram by an experienced operator is required to remove/edit non-sinus ectopic beats, pauses, tape artefact and non-periodic R-R interval changes."

Classical spectral density estimators are sensitive to outlying observations, such as ectopic beats and other artifacts, yielding poor results. The advantage of estimates obtained by robust methods is that they are not much affected by outliers. Neither removing nor editing of the data needs to be done.

All presented methods for estimating the spectral density function robustly are described and explained in detail. They are implemented in the open source language R (R Development Core Team, 2005) and are available on request from the author in order to enable others to understand and reproduce our work.

In a discussion with Dave Thomson, currently at Queen's University, Kingston, he mentioned that there exist many different robust methods for the same research question and it is hard to decide which one should be used in a certain practical application. As this happens in many areas of robust statistics, we therefore put special emphasis on the comparison of methods which is accomplished by extensive simulation studies.

Contents

v

vi

List of Figures

Chapter 1

Introduction

The spectral density function is a commonly used tool when analyzing time series in the frequency domain. Areas of applications are signal processing (cf. Thomson, 1994), geophysics (cf. Chave et al., 1987; Jones and Hollinger, 1997) and medicine (cf. Hartikainen et al., 1998).

However, it is well known that classical estimators of the spectral density function are sensitive to outlying observations. In the presence of outliers spectral density estimates may therefore be markedly affected in shape and power.

Furthermore, in many applications the observed signal will be instationary or its frequency content may change over time. In such situations the standard Fourier analysis is quite inadequate because it summarizes the information in the series as a function of frequency and does not preserve any information in time. Moreover, we have to assume stationarity of the data. To overcome these limitations of standard Fourier analysis we need to represent the signal in time and frequency.

An example that this is worth considering is given in Figure 1.1. Three synthetic signals are plotted in the first row of Figure 1.1. In the second row the estimated spectral density functions of the signals are shown whereas in the third row the time-frequency decompositions of the signals are displayed in greyscales according to the magnitude of the corresponding decomposition value (small values close to white, large values almost black).

The first two signals are composed of two sine waves with fixed frequencies. On the first signal both frequencies occur during the whole observation interval, i.e., the first signal was obtained by simply adding the two sine
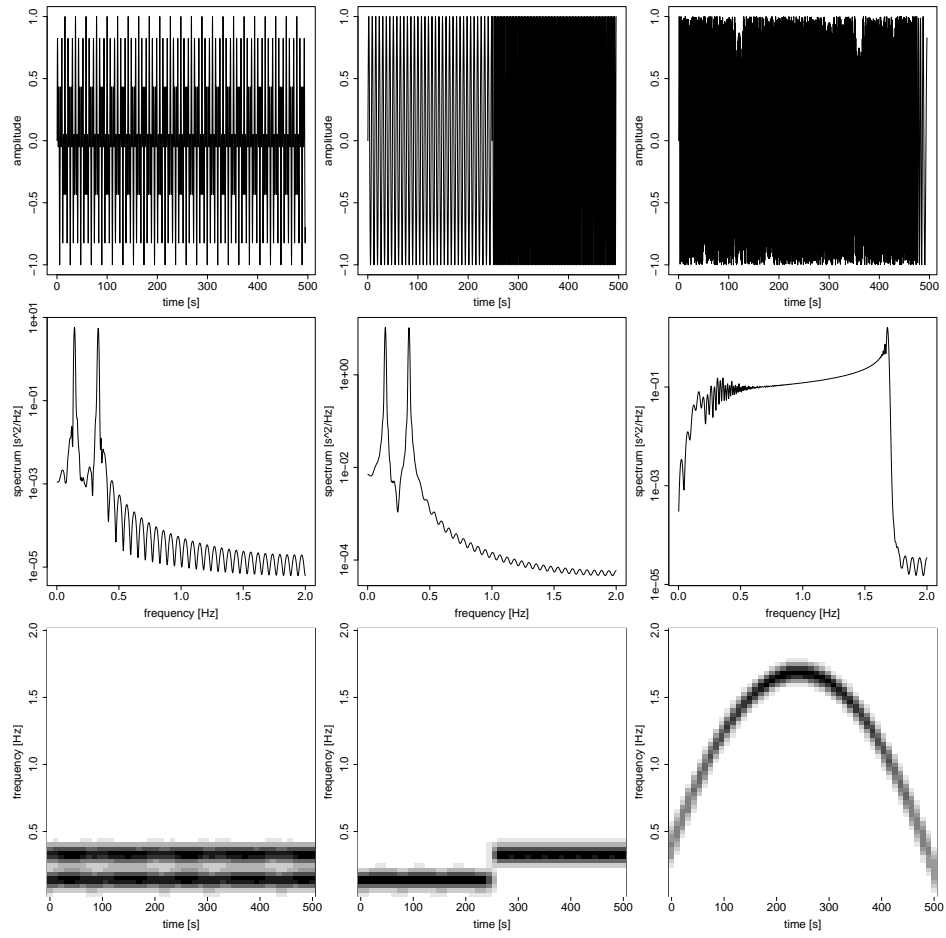
Figure 1.1: Standard and dynamic Fourier analysis of synthetic signals

waves, whereas on the second they are temporally localized on the first and second halves of the time interval. Although the two signals can hardly be distinguished by only looking at their estimated spectral density functions, the time-frequency decompositions clearly reveal the temporal discrepancy of the two components. Two bands (dark color) are visible on the whole interval for the first signal, whereas for the second signal we can clearly identify two temporally localized bands. However, the varying of the frequency content of the third signal, a so called quadratic chirp, is only revealed by its time-frequency decomposition.

In literature there exist several time-frequency representations. We will only mention a few but this listing is by no means exhaustive. One is the dynamic Fourier analysis, also known as windowed Fourier transform, short-time Fourier transform, or Gabor transform (cf. Shumway and Stoffer, 2000; Gençay et al., 2002). It is obtained by sliding a window of fixed width across the time series and estimating the spectral density function in each window. Another possibility of a more flexible time-frequency representation (or, to be exact, time-scale decomposition) are wavelets. Details about wavelet analysis may be found in Ogden (1997) or Percival and Walden (2000). A third is time-frequency analysis using Wigner distributions (cf. Matz and Hlawatsch, 2003). Although these methods are not robust in the presence of outliers, some attempts have been made to robustify wavelet analysis using median-interpolating refinement (cf. Donoho and Yu, 2000) and time-frequency analysis using minimax robust time-varying Wiener filters (cf. Matz and Hlawatsch, 2000).

We will however restrict our further research to dynamic Fourier analysis which is sensitive to outlying observations unless we estimate the spectral density function in each window robustly. Hence, we will focus in the following on the problem of robust spectral density estimation and present different methods, existing and new ones, that are resistant to outliers.

Each chapter is dedicated to a specific topic. To improve readability and increase understanding the related simulation studies and results together with a discussion are enclosed at the end of each chapter.

Part I containing Chapters 2 to 9 considers the theory of robust spectral density estimation whereas in Part II, as a special practical application, we focus on the frequency-domain analysis of actual heart rate variability measurements.

Chapter 2 sums up the existing classical spectral density estimators, in-

troduces different kinds of time series outliers and points out the sensitivity of classical methods to outlying observations. In Chapter 3 we consider spectral density estimators based on robustifying the Fourier transformation. Chapter 4 presents several methods based on robust autocovariance estimation. These include a highly robust autocovariance function and robust autocovariance estimation based on rank correlation coefficients and partial autocorrelation coefficients. Here we also cope with the problem of ensuring positive semidefiniteness.

Chapters 5 to 7 finally deal with robust filtering algorithms. Chapter 5 considers two different ways of estimating the hyper parameters relevant for the robust filters in the subsequent chapters, namely bounded-influence autoregression and robust autoregressive parameter estimation via the MCD estimator. In Chapter 6 we present different approaches for robustifying the Kalman filter. We propose a new multivariate approximate conditional-mean (ACM) type filter for state-space models. This new filtering method generalizes the original ACM-type filter which is bound to the univariate setting. We compare our new filtering method to a second approach, the currently used rLS filter, which is also described. In Chapter 7 we consider prewhitened spectral density estimation based on the robust filtering techniques described in the previous chapter.

Chapter 8 gives an overview on further related methods that have also been used in the context of spectral density estimation. In Chapter 9 conclusions and remarks to the methods described in Part I are given.

Part II discusses the analysis of heart rate variability data in the frequency domain. The most competitive method is applied to short-term heart rate variability measurements of diabetic patients with different degrees of cardiovascular autonomic neuropathy and the corresponding results are presented.

4

Part I

Robust Spectral Density Estimation

Chapter 2

Classical Spectral Density Estimation

2.1 Nonparametric Estimation

The nonparametric estimation of the spectral density function is based on smoothing the periodogram.

Let $\{x_t, t = 1, \ldots, n\}$ denote the observed process which is assumed to be second-order stationary and to have zero mean. We further suppose that the time intervals between two consecutive observations are equally spaced with duration Δt. Then the *periodogram* is defined as follows:

$$\widehat{S}^{(p)}(f) = \frac{\Delta t}{n} \left| \sum_{t=1}^{n} x_t e^{-i2\pi f t \Delta t} \right|^2 = \Delta t \sum_{h=-(n-1)}^{(n-1)} \widehat{\gamma}_x(h) e^{-i2\pi f h \Delta t} , \qquad (2.1)$$

where $\widehat{\gamma}_x(h)$ denotes the sample autocovariance function of the time series x_t. $\widehat{S}^{(p)}(f)$ is defined over the interval $[-f_{(n)}, f_{(n)}]$, where $f_{(n)}$ is called the *Nyquist frequency* and is given by

$$f_{(n)} = \frac{1}{2\Delta t} . \qquad (2.2)$$

If we restrict our attention to just the frequency $f = f_k = k/(n\Delta t)$, where k is an integer such that $|k| \leq \lfloor n/2 \rfloor$, we see that the periodogram at that frequency f_k is—except for a scaling factor—the squared modulus of $X(f_k)$, the k-th component of the discrete Fourier transform of the sequence

$x_1, \ldots, x_n$, which is given by

$$X(f_k) = \Delta t \sum_{t=1}^{n} x_t e^{-i2\pi f_k t \Delta t} \ . \tag{2.3}$$

Although $\widehat{S}^{(p)}(f)$ is an asymptotically unbiased estimator of the true spectral density function $S(f)$, it is well known that $\widehat{S}^{(p)}(f)$ may be badly biased in some cases. There are two common techniques for reducing the bias in the periodogram: tapering and prewhitening. The latter will be described in Section 2.3.

Data tapering leads to the *direct spectral estimator*, which is defined by

$$\widehat{S}^{(d)}(f) = \Delta t \left| \sum_{t=1}^{n} h_t x_t e^{-i2\pi f t \Delta t} \right|^2 , \tag{2.4}$$

where $\{h_t, t = 1, \ldots, n\}$ is called the data taper sequence with $\sum_{t=1}^{n} h_t^2 = 1$.

The main idea behind tapering is to choose the h_t's so that the beginning and the end of the observed series x_t is downweighted toward 0 in a smooth way. A particularly interesting class of data tapers among others (e.g. the cosine or Hanning tapers) are those formed from 0th-order discrete prolate spheroidal sequences. Because of their remarkable properties we will use these data tapers. Details may be found in a series of papers by D. Slepian, of which the most recent is Slepian (1978), and in Thomson (1977, 1982).

Because $\widehat{S}^{(d)}(f)$ is defined for all $f \in [-f_{(n)}, f_{(n)}]$, we can smooth it using a continuous convolution over a continuous set of frequencies. Thus, we consider an estimator of the form

$$\begin{aligned}
\widehat{S}^{(lw)}(f) &= \int_{-f_{(n)}}^{f_{(n)}} W_m(f - \nu) \widehat{S}^{(d)}(\nu) d\nu \\
&= \Delta t \sum_{h=-(n-1)}^{n-1} w_{h,m} \widehat{s}_h^{(d)} e^{-i2\pi f h \Delta t} , \tag{2.5}
\end{aligned}$$

where $w_{h,m}$ and $W_m(.)$ are a Fourier transform pair and $w_{h,m} = 0$ for $|h| \geq n$. $\widehat{s}_h^{(d)}$ is the estimator of the autocovariance sequence corresponding to $\widehat{S}^{(d)}(f)$, i.e., its inverse Fourier transform.

The function $W_m(.)$ is a symmetric real-valued $2f_{(n)}$ periodic function for all choices of m, which is square integrable over $[-f_{(n)}, f_{(n)}]$, and m is a

7

smoothing parameter that controls the degree of smoothing. $W_m(.)$ is called a *smoothing window*, its inverse Fourier transform $w_{h,m}$ is called a *lag window*. Hence, we call $\widehat{S}^{(lw)}(f)$ a *lag window spectral estimator* of $S(f)$.

On the other hand by smoothing the direct spectral estimator $\widehat{S}^{(d)}(f)$ with a discrete convolution over a discrete set of frequencies we obtain a so called *discretely smoothed direct spectral estimator* of $S(f)$. However, any lag window spectral estimator $\widehat{S}^{(lw)}(.)$ can be expressed at the Fourier frequencies as a discretely smoothed direct spectral estimator.

We note that the definition of the lag window spectral estimator $\widehat{S}^{(lw)}(f)$ in (2.5) is essentially a weighting applied to the autocovariance sequence estimator $\widehat{s}_h^{(d)}$ corresponding to the direct spectral estimator $\widehat{S}^{(d)}(f)$.

Many different lag and smoothing windows, e.g., the Bartlett, the Daniell, the Parzen or the Papoulis window, just to name a few of them, have been proposed and discussed in literature (see, for example, Priestley, 1981). Mainly two criteria should be taken into account for evaluating different lag window spectral estimators: The smoothing window leakage should be small and the smoothing window should have good smoothing properties, i.e., its corresponding lag window should decay monotonically to 0.

In all simulation studies as well as for our real data examples we use the Parzen window because it seems to accomplish the trade-off between small smoothing window leakage and good smoothing properties best. The Parzen lag window, Parzen (1961), is defined as follows:

$$
w_{h,m} = \begin{cases} 1 - 6(h/m)^2 + 6(|h|/m)^3 & \text{if} \quad |h| \leq m/2 \\ 2(1 - |h|/m)^3 & \text{if} \quad m/2 < |h| \leq m \\ 0 & \text{if} \quad m < |h| \,. \end{cases} \tag{2.6}
$$

However, more important than the choice of the smoothing window is the choice of the smoothing parameter m. We will use an approach known as *window closing*. The idea is to compute a sequence of different lag window spectral estimates for the same set of data using different smoothing parameters which range from small to large. In the case of the Parzen window, the estimates will look smooth for small values of m, but, as m increases, the estimates will reveal more and more details. Based upon the examination of all different estimates, we choose a value of m that is appropriate in the sense that the resulting estimate is neither too smooth nor too erratic. Although $\widehat{S}^{(lw)}(.)$ is a smoothed version of $\widehat{S}^{(d)}(.)$, the former should still capture the important features of the latter. According to our experience, a value of m between $n/4$ and $n/2$ seems to be appropriate.

Further details about the above described nonparametric spectral density estimators, their statistical properties and the different smoothing windows may be found in Priestley (1981) or Percival and Walden (1993).

2.2 Parametric Estimation

The most widely used form of parametric spectral density estimation uses an autoregressive model of order p as the underlying functional form for $S(f)$. A stationary $\text{AR}(p)$ process $\{x_t, t \in \mathbb{Z}\}$ with zero mean satisfies the equation

$$x_t - \sum_{j=1}^{p} \phi_j x_{t-j} = \varepsilon_t \; , \tag{2.7}$$

where ε_t is a white noise process with zero mean and variance σ_ε^2. Thus, the spectral density function satisfies the following equation

$$\left| 1 - \sum_{j=1}^{p} \phi_j e^{-i2\pi f j \Delta t} \right|^2 S(f) = \Delta t \sigma_\varepsilon^2 \; . \tag{2.8}$$

Substituting the maximum likelihood or least squares estimators of the model parameters, denoted by $\widehat{\phi}_1, \ldots, \widehat{\phi}_p$ and $\widehat{\sigma}_\varepsilon^2$, we obtain a parametric spectral density estimator

$$\widehat{S}^{(AR)}(f) = \frac{\Delta t \widehat{\sigma}_\varepsilon^2}{\left| 1 - \sum_{j=1}^{p} \widehat{\phi}_j e^{-i2\pi f j \Delta t} \right|^2} \; , \quad |f| \leq f_{(n)}. \tag{2.9}$$

2.3 Semi-parametric Estimation

The lag window spectral density estimator and the parametric spectral density estimator are the limiting versions of prewhitening the process before estimating the spectral density function. The first approach leads to no prewhitening and the latter corresponds to total prewhitening if we assume that the process is a finite-order autoregressive process with known order p.

Let $\{y_t, t = 1, \ldots, n\}$ denote the observed values of a second-order stationary process with zero mean. Then the prewhitened spectral density estimate

originally suggested by Blackman and Tukey (1958) is defined as

$$\widehat{S}(f) = \frac{\widehat{S}_r^{(lw)}(f)}{|\widehat{H}_p(f)|^2} \ , \quad |f| \leq f_{(n)} \ , \tag{2.10}$$

where $\widehat{S}_r^{(lw)}(f)$ is a lag window spectral density estimate of the prediction residuals $r_t = y_t - \sum_{j=1}^{p} \widehat{\phi}_j y_{t-j}$, $t = p+1, \ldots, n$ and $\widehat{H}_p(f)$ is the transfer function given as

$$\widehat{H}_p(f) = 1 - \sum_{j=1}^{p} \widehat{\phi}_j e^{-i2\pi f j \Delta t} \ . \tag{2.11}$$

2.4 Time Series Outliers

Generally, three types of outliers are considered (cf. Denby and Martin, 1979): innovation outliers (IO), which affect all subsequent observations, additive outliers (AO) and substitutive outliers (SO). The latter two have no effect on subsequent observations. Innovation outliers and additive outliers were first introduced by Fox (1972).

The second-order stationary ARMA(p, q) process $\{x_t, t \in \mathbb{Z}\}$ is said to have *innovation outliers* if the innovations ε_t have a heavy-tailed distribution, for instance a γ-contaminated normal distribution

$$\mathcal{CN}(\gamma, \sigma_0, \sigma) = (1 - \gamma)\mathcal{N}(0, \sigma_0^2) + \gamma \mathcal{N}(0, \sigma^2) \tag{2.12}$$

where $\sigma_0^2 \ll \sigma^2$ and γ is small. $\mathcal{N}(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2.

Another commonly used model for outliers in time series is the additive outlier model (AO model). The AO model consists of a stationary core process, x_t, to which occasional outliers have been added. The observed process $\{y_t, t \in \mathbb{Z}\}$ is said to have *additive outliers* if it is defined by

$$y_t = x_t + v_t \tag{2.13}$$

where the contaminations v_t are independent and identically distributed with $\mathcal{F}_{v_t}$. $\mathcal{F}_{v_t}$ is equal to a contaminated normal distribution with degenerate central component, i.e.,

$$\mathcal{F}_{v_t} = \mathcal{CN}(\gamma, 0, \sigma^2) = (1 - \gamma)\mathcal{N}(0, 0) + \gamma \mathcal{N}(0, \sigma^2) \ . \tag{2.14}$$

Hence, the core process x_t is observed with probability $1 - \gamma$ whereas the core process plus a disturbance v_t is observed with probability γ. We shall also assume that x_t and v_t are independent.

Of course, there is no reason that the mixing component should be the normal distribution $\mathcal{N}(0, \sigma^2)$. We could use any other suitably heavy-tailed distribution $\mathcal{H}$ instead. However, for most theoretical results mentioned in this work we will need the assumptions given above.

We remark that the AO model can be generalized in the following way. Let B_t be a Bernoulli process with $P(B_t = 1) = \gamma$ and $P(B_t = 0) = 1 - \gamma$, and let the observed process be

$$y_t = (1 - B_t)x_t + B_t\tilde{v}_t \ . \tag{2.15}$$

This means that the core process x_t is observed with probability $1 - \gamma$ and replaced by a contaminating process $\tilde{v}_t$ with probability γ. Thus, the process y_t is said to have *substitutive outliers*. We note that when $\tilde{v}_t = x_t + v'_t$ with $v'_t \sim \mathcal{N}(0, \sigma^2)$ this is equivalent to model (2.13) with $v_t \sim \mathcal{CN}(\gamma, 0, \sigma^2)$.

2.5 Non-robustness of Classical Spectral Density Estimation

It is well known that usual variance estimators are not robust in the presence of outliers. Since classical spectral density estimation procedures can be interpreted as estimators of variance on a frequency basis, they share this lack of robustness.

Innovation outliers do not pose serious problems for spectral density estimation if we are primarily interested in the shape of the spectral density function. We only have to estimate the scaling factor robustly.

The real problem for spectral density estimation are additive outliers. Let us consider the special case of a single fixed additive outlier, i.e., $v_j = A, v_t = 0, t = 1, \ldots, n, t \neq j$. Since $|z_1 + z_2|^2 = |z_1|^2 + |z_2|^2 + 2\mathfrak{Re}\{z_1\overline{z_2}\}, z_1, z_2 \in \mathbb{C}$, the periodogram $\widehat{S}_{y_t}^{(p)}(f_k)$ of y_t,

$$\widehat{S}_{y_t}^{(p)}(f_k) = \widehat{S}_{x_t}^{(p)}(f_k) + \Delta t \frac{A^2}{n} + \Delta t \frac{A}{n} 2\mathfrak{Re}\left\{X(f_k)e^{i2\pi f_k j \Delta t}\right\} \ , \tag{2.16}$$

is raised by $\Delta t(A^2/n)$ and retains its shape only if the oscillatory term $\Delta t(A/n)2\mathfrak{Re}\{X(f_k)e^{i2\pi f_k j \Delta t}\}$ is considerably smaller than $\widehat{S}_{x_t}^{(p)}(f_k)$ for all frequencies f_k.

Since autoregressive parameter estimates obtained by classical estimation procedures are not robust in situations where the time series is contaminated by additive outliers, there is also a corresponding effect on parametric spectral density estimates.

For further details see Kleiner et al. (1979) or Martin and Thomson (1982).

Chapter 3

Simple Robust Approaches

3.1 Robust Fourier Transformation

Let $\{x_t, t = 1, \ldots, n\}$ denote the observed process which is assumed to be second-order stationary and to have zero mean. Then, the *periodogram* is defined as follows:

$$\widehat{S}^{(p)}(f) = \frac{\Delta t}{n} \left| \sum_{t=1}^{n} x_t e^{-i2\pi f t \Delta t} \right|^2 . \tag{3.1}$$

If we restrict our attention to just the frequency $f = k/(n\Delta t)$, where k is an integer such that $|k| \leq \lfloor n/2 \rfloor$, we see that the periodogram at that frequency is the squared modulus of the k-th component of the discrete Fourier transform of the sequence $x_1, \ldots, x_n$.

Transforming the above equation (3.1) we get

$$\widehat{S}^{(p)}(f) = n\Delta t \left| \frac{1}{n} \sum_{t=1}^{n} x_t e^{-i2\pi f t \Delta t} \right|^2 . \tag{3.2}$$

Hence, if we keep the frequency f fixed, $n^{-1} \sum_{t=1}^{n} x_t e^{-i2\pi f t \Delta t}$ is just the arithmetic mean of a complex variable, or, in other words, the multivariate location estimate of 2-dimensional observations.

In order to obtain a robust estimate of the spectral density function we simply replace the arithmetic mean by a robust measure of location. A straightforward way would be to take the trimmed mean or the median, which will be applied coordinate-wise. A more sophisticated alternative would be

to replace the mean by a robust multivariate location estimator, e.g., the L_1-median, also called spatial median, or the center of the ellipsoid that corresponds to the robustly estimated covariance matrix using Rousseeuw's MCD estimator.

3.2   Simulation Study

The outline of our simulation study is as follows: First we simulate a core process x_t of length $n = 100$. x_t is chosen to be an autoregressive process of order 2 given by

$$x_t = x_{t-1} - 0.9 x_{t-2} + \varepsilon_t \ , \tag{3.3}$$

with $\varepsilon_t \sim \mathcal{N}(0,1)$. Additionally, the additive outliers v_t are simulated from a contaminated normal distribution with degenerate central component, i.e.,

$$\mathcal{CN}(\gamma, 0, \sigma^2) = (1-\gamma)\mathcal{N}(0,0) + \gamma \mathcal{N}(0, \sigma^2) \ , \tag{3.4}$$

with $\sigma^2 = 10^2$. The contamination γ is varied from 0% to 20% by steps of 5%. That means that with probability γ, v_t is an additive outlier with $v_t \neq 0$. To obtain the contaminated process y_t, the v_t's are added to the core process x_t. For each level of contamination this is done 400 times.

For each contaminated series we robustly estimate the spectral density function using the methods based on the trimmed mean and Rousseeuw's MCD estimator.

Then, the deviation of each estimated spectral density function from the true spectral density function is measured in the sense of the squared L_2-norm, i.e.,

$$err^2_{\widehat{S}(f)} := \|\widehat{S}(f) - S(f)\|^2 = \int (\widehat{S}(f) - S(f))^2 df \ , \tag{3.5}$$

where $\widehat{S}(f)$ and $S(f)$ denote the estimated and true spectral density functions.

3.3   Results

Regarding the computation time the robust spectral density estimation based on the trimmed mean is much faster than the method using Rousseeuw's

MCD estimator. This is due to the fact that the coordinate-wise application of the trimmed mean can be calculated straightforward whereas for the MCD estimator several iterations need to be computed.

Figure 3.1 tries to visualize the results of our simulation study. For both methods and each level of contamination seven curves are plotted on a logarithmic scale. The bold black line represents the true spectral density function, whereas the thin black line is the spectral density estimate of one realization out of 400. Moreover, we calculate frequency by frequency the minimum and maximum, the first and third quartile and median value of all spectral density estimates. Connecting all median values we obtain the grey line, to which we will refer hereafter as median spectral density function. In the same sense we refer to all minimum values as minimum spectral density function, and so on. Hence, the lower and upper dotted lines are the minimum and maximum spectral density functions, whereas the lower and upper dashed lines represent the first and third quartile spectral density functions. This plot may be seen as generalization of the well-known boxplot. The results obtained by using the MCD estimator are plotted in the left column, whereas the results based on the trimmed mean are displayed in the right column.

Surprisingly, for both methods the dispersion of the spectral density estimates stays almost the same for all different levels of contamination. However, for both methods we obtain bad spectral density estimates even in the case of no contamination, whereas the method based on the trimmed mean performs better than the one using the MCD estimator, especially if no outliers are present. At its peak, i.e., at a frequency of about 0.16, the true spectral density function is always underestimated but at all other frequencies, especially at higher ones, we see that the spectral density function is all the more overestimated the higher the contamination.

Next, we try to visualize the squared errors of the estimated spectral density functions. First, the logarithm of the squared errors is taken. For both methods Figure 3.2 shows boxplots of the squared errors in eight equally-sized frequency bands as well as the total squared errors (bottom right) for all different levels of contamination. As expected by the results plotted in Figure 3.1 the errors do not differ very much for both methods and, looking at the different frequency bands separately, stay almost the same independently of the amount of contamination. However, looking at the total squared errors, we see that the method based on the trimmed mean performs better than the one using the MCD estimator. The largest contribution to the total

15
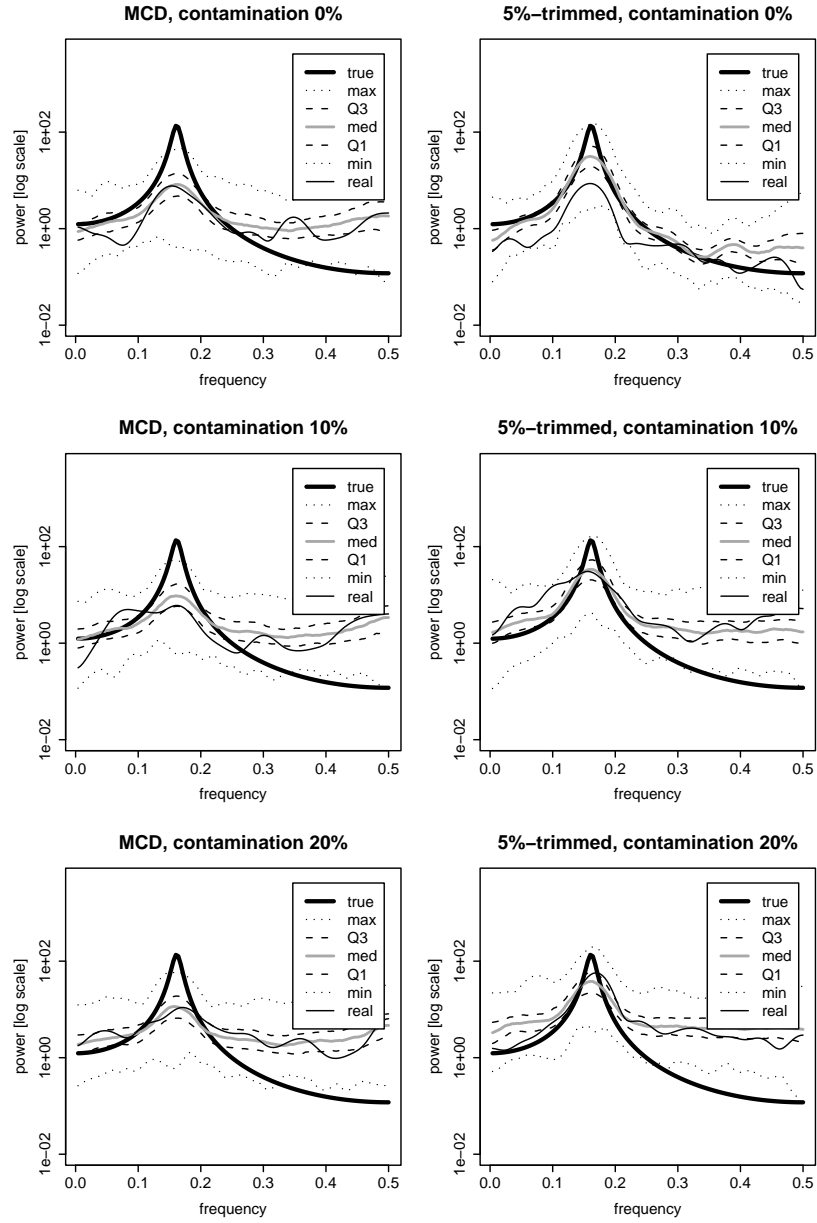
Figure 3.1: Robust spectral density estimates of the simulated data, left column 'MCD', right '5%-trimmed mean'
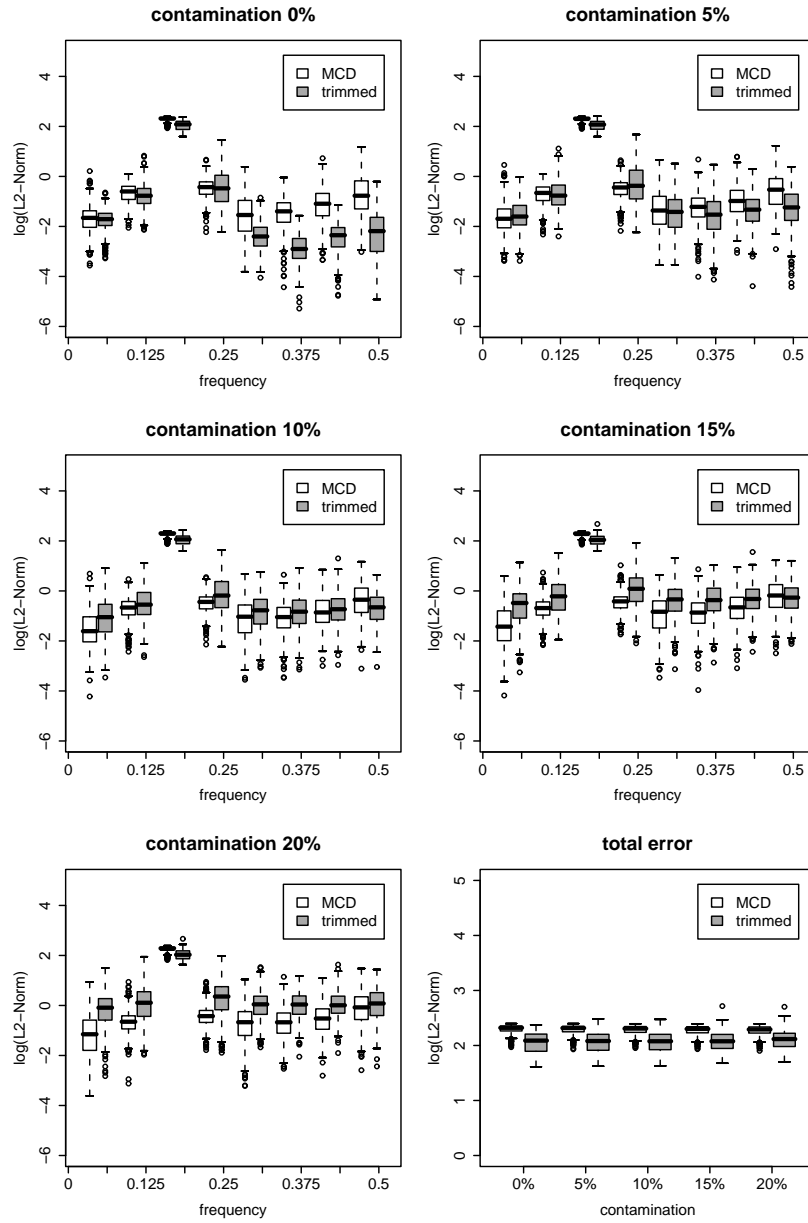
Figure 3.2: Boxplots of the errors

squared error is the amount of the frequency band where the spectral density function has its peak. There the squared errors using the MCD estimator are larger than the ones based on the trimmed mean. Moreover, we see that all squared errors are in the same range for all contamination levels.

3.4 Discussion

In this chapter we tried to get robust spectral density estimates by directly using a robustified version of the Fourier transformation. The classical Fourier transformation can be interpreted as taking the mean of 2-dimensional observations. In order to robustly estimate the spectral density function, we simply replaced the mean by a robust multivariate measure of location. We compared estimates obtained by two methods, one based on the trimmed mean, applied coordinate-wise, and the other using Rousseeuw's MCD estimator.

The results of the simulation study suggest that the method based on the trimmed mean performs better than the one using the MCD estimator for all different levels of contamination, which is rather surprising. We would have expected that the method based on the trimmed mean only performs best in cases where the amount of trimming coincide with the amount of outliers.

We also used other robust multivariate location estimators, namely the median, which was again applied coordinate-wise in the same way as the trimmed mean, and the L_1-median (or spatial median). The L_1-median has already been implemented in R based on the algorithm described in Hössjer and Croux (1995) and is available in the R-package `pcaPP`. However, both methods underestimate the true spectral density function even in the case of no contamination and the results of the corresponding simulation studies are therefore not published here.

Chapter 4

Robust Autocovariance Functions

4.1 Spectral Density Estimation via the Autocovariance Function

Since the spectral density function is the Fourier transform of the autocovariance function, Blackman and Tukey (1958) suggest an estimator of the form

$$\widehat{S}^{(p)}(f) = \Delta t \sum_{h=-(n-1)}^{(n-1)} \widehat{\gamma}_x(h) e^{-i2\pi f h \Delta t} \; , \tag{4.1}$$

where $\widehat{\gamma}_x(h)$ is just the sample autocovariance function of an observed process $\{x_t, t = 1, \ldots, n\}$.

In order to get a robust version of the spectral density function we simply replace $\widehat{\gamma}_x(h)$ by a robust autocovariance function. We will compare estimates obtained by a highly robust autocovariance function (Ma and Genton, 2000) to others based on Spearman's rank correlation coefficient (Ahdesmäki et al., 2005) or on partial autocorrelation (Möttönen et al., 1999). Unfortunately, most of these robust alternatives only yield robust autocorrelation coefficients. Moreover, the maximum lag for which the autocovariance or autocorrelation coefficients can be calculated using these methods is $n - 2$.

Anyway, setting $C = \gamma_x(0) = \mathrm{var}(x_t)$ and estimating it by using a robust

measure of scale, we propose the following estimate

$$\widehat{S}^{(rob)}(f) = \Delta t \widehat{C} \sum_{h=-(n-2)}^{(n-2)} \widehat{\rho}_x^{(rob)}(h) e^{-i2\pi f h \Delta t} \ , \tag{4.2}$$

where $\widehat{\rho}_x^{(rob)}(h)$ denotes one of the above mentioned robust autocorrelation functions. These methods will be described in detail in the next sections.

Replacing the autocovariance function by a lag-weighted version obtained by multiplying $\widehat{\gamma}_x(h)$ in (4.1) point by point by some weight function $w(h)$ can be used to improve the properties of the spectral estimator.

4.2 A Highly Robust Autocovariance Function

A highly robust autocovariance estimator was suggested by Ma and Genton (2000). Their idea was that covariance estimation can be based on a scale approach using the following identity (Huber, 1981):

$$\mathrm{cov}(x, y) = \frac{1}{4\alpha\beta} \left(\mathrm{var}(\alpha x + \beta y) - \mathrm{var}(\alpha x - \beta y) \right) \ , \tag{4.3}$$

with $\alpha, \beta \in \mathbb{R}$. In general, the variables x and y may be measured in different units. However, in time series analysis, x and y will represent the same variable and we set $\alpha = \beta = 1$.

Let $\{x_t, t = 1, \ldots, n\}$ be a stationary time series. Then the highly robust autocovariance function estimator is defined as follows. Extract the first $n-h$ observations of $\{x_t, t = 1, \ldots, n\}$ to produce a vector $\boldsymbol{u}$ of length $n - h$ and the last $n - h$ observations of $\{x_t, t = 1, \ldots, n\}$ to produce a vector $\boldsymbol{v}$ of length $n - h$. Then:

$$\widehat{\gamma}_x^{(Q)}(h) = \frac{1}{4} \left(Q_{n-h}^2(\boldsymbol{u} + \boldsymbol{v}) - Q_{n-h}^2(\boldsymbol{u} - \boldsymbol{v}) \right) \ . \tag{4.4}$$

The function $Q_m(\boldsymbol{z})$ is a highly robust estimator of scale proposed by Rousseeuw and Croux (1992, 1993) and is defined by

$$Q_m(\boldsymbol{z}) = c \left\{ |z_i - z_j| : i < j \right\}_{(k)} \ , \tag{4.5}$$

20

where $\boldsymbol{z} = (z_1, \ldots, z_m)^\top$ is the sample and

$$k = \binom{l}{2} \ , \ \text{with } l = \left\lfloor \frac{m}{2} \right\rfloor + 1 \ ,$$

where $\lfloor . \rfloor$ denotes the integer part. The factor $c = 2.2219$ is for consistency in case of normally distributed data. We note that Q_m does not rely on any location estimate and is therefore said to be location-free.

Moreover, Croux and Rousseeuw (1992) determined an appropriate correction factor for finite samples. Hence, they redefine Q_m as

$$Q_m(\boldsymbol{z}) = d_m c \left\{ |z_i - z_j| : i < j \right\}_{(k)} \ , \tag{4.6}$$

where the correction factor d_m is given by

m	2	3	4	5	6	7	8	9
d_m	0.399	0.994	0.512	0.844	0.611	0.857	0.669	0.872

for $m \leq 9$, and for $m > 9$ it is defined as

$$d_m = \frac{m}{m + 1.4} \quad \text{if } m \text{ odd}$$
$$d_m = \frac{m}{m + 3.8} \quad \text{if } m \text{ even.}$$

Finally, to obtain a highly robust estimator of the autocorrelation function, Ma and Genton (2000) propose the following estimator:

$$\widehat{\rho}_x^{(Q)}(h) = \frac{Q_{n-h}^2(\boldsymbol{u} + \boldsymbol{v}) - Q_{n-h}^2(\boldsymbol{u} - \boldsymbol{v})}{Q_{n-h}^2(\boldsymbol{u} + \boldsymbol{v}) + Q_{n-h}^2(\boldsymbol{u} - \boldsymbol{v})} \ . \tag{4.7}$$

The above definition ensures that $\widehat{\rho}_x^{(Q)}$ is bounded between -1 and 1 as we would expect from an autocorrelation function. We note that $\widehat{\gamma}_x^{(Q)}$ depends on the choices of the consistency constant c and the finite sample correction factor d_m in (4.6), whereas $\widehat{\rho}_x^{(Q)}$ is independent of both.

Ma and Genton (2000) note that their proposed estimator of the autocovariance function is a robust alternative to the classical, not necessarily positive semidefinite one, i.e., the one that is divided by $n - h$ instead of n (see also Appendix A.1). Hence, in order to be comparable with the classical positive semidefinite sample autocovariance function, we suggest to modify

the estimates in (4.4) and (4.7) by multiplying them with the factor $\frac{n-h}{n}$ at lag h and, additionally, we also have to ensure positive semidefiniteness.

The highly robust estimator Q_m has already been implemented in R based on the fast algorithm described in Croux and Rousseeuw (1992) and is available in the R-package `robustbase`. In R the definition of Q_m given in (4.6) is used by default.

4.3 Robust Autocovariance Estimation Based on Rank Correlation Coefficients

Ahdesmäki et al. (2005) consider a rank-based autocorrelation estimator originally introduced by Pearson et al. (2003) for the problem of robust, non-parametric spectral density estimation.

Let $\{x_t, t = 1, \ldots, n\}$ be a stationary time series. Then the proposed estimator is an extension of Spearman's rank correlation coefficient, quantifying the association between x_{t+h} and x_t. More specifically, we consider the correlation coefficient between the data ranks $r_x(i)$ and $r'_x(i)$, defined by

$$
\begin{aligned}
\rho_x^{(S)}(h) &= \\
&= \frac{1}{K} \frac{12}{(n-h)^2 - 1} \sum_{i=1}^{n-h} \left(r_x(i) - \frac{n-h+1}{2} \right) \left(r'_x(i) - \frac{n-h+1}{2} \right) \;,
\end{aligned} \quad (4.8)
$$

where K is a normalisation factor, $r_x(i)$ denotes the rank of x_i in the set $\{x_t, t = 1, \ldots, n-h\}$ and $r'_x(i)$ denotes the rank of x_{i+h} in the set $\{x_{t+h}, t = 1, \ldots, n-h\}$.

By choosing either $K = n - h$ or $K = n$ in (4.8) we obtain the unbiased estimate of the rank correlation coefficient or the biased one, respectively. We shall in future use the biased estimator, i.e., $K = n$, in order to be comparable with the classical sample autocovariance function. Moreover, we note that $n - 2$ is the largest lag for which formula (4.8) can be computed.

Although this procedure has already been implemented in R and is available in the R-package `GeneTS` (cf. also Wichert et al., 2004) we slightly have to modify it. This is because Ahdesmäki et al. (2005) originally intended to detect periodicities in time series and not to obtain estimates of the spectral density function. Therefore they only took the Fourier transform of the autocorrelation function whereas for spectral density estimation rescaling with a robust measure of scale is necessary.

4.4 Robust Autocovariance Estimation Based on Partial Autocorrelation Coefficients

Möttönen et al. (1999) proposed a robust autocorrelation estimation method which is based on estimating the partial autocorrelation coefficients first. (For details about partial autocorrelation coefficients the reader is referred to Appendix A.2.) They suggested to robustly estimate the partial autocorrelation coefficients by using transformations of Spearman's rho.

Let $\{x_t, t = 1, 2, \dots\}$ be a stationary time series with zero mean and $\phi_{11} = \rho_x(1) = \mathrm{corr}(x_t, x_{t-1})$, where ρ_x denotes the autocorrelation function of x_t. The partial autocorrelation coefficients ϕ_{hh}, $h = 2, 3, \dots$, can be calculated recursively as follows. We first write

$$\nu_{ht} = x_t - \sum_{i=1}^{h-1} \phi_{h-1,i} x_{t-i} \tag{4.9}$$

and

$$\omega_{h,t-h} = x_{t-h} - \sum_{i=1}^{h-1} \phi_{h-1,i} x_{t-(h-i)} \tag{4.10}$$

for the prediction errors of x_t and x_{t-h}, respectively, with the linear effect of $\{x_{t-(h-1)}, \dots, x_{t-1}\}$ on each removed. Then,

$$\phi_{hh} = \mathrm{corr}(\nu_{ht}, \omega_{h,t-h}) \tag{4.11}$$

is the correlation between the prediction errors. The recursions for the next coefficients are

$$\phi_{hi} = \phi_{h-1,i} - \phi_{hh}\phi_{h-1,h-i} , \quad i = 1, \dots, h-1 . \tag{4.12}$$

The autocorrelations $\rho_x(h)$, $h = 2, 3, \dots$, are then obtained by

$$\rho_x(h) = \sum_{i=1}^{h-1} \phi_{h-1,i}\rho_x(h-i) + \phi_{hh}\left(1 - \sum_{i=1}^{h-1} \phi_{h-1,i}\rho_x(i)\right) . \tag{4.13}$$

Estimates obtained by using formula (4.13) will be denoted by $\widehat{\rho}_x^{(pacf)}(h)$ where we use the biased version of Spearman's rank correlation coefficient (4.8) to obtain robust partial autocorrelation coefficients in (4.11).

23

We note that formula (4.12) is just taken from the Durbin-Levinson algorithm whereas (4.13) can be easily obtained from it by rearranging

$$\phi_{hh} = \frac{\rho_x(h) - \sum_{i=1}^{h-1} \phi_{h-1,i}\rho_x(h-i)}{1 - \sum_{i=1}^{h-1} \phi_{h-1,i}\rho_x(i)} \ .\qquad (4.14)$$

(Formulas (4.14) and (4.12) are referred to as the well-known Durbin-Levinson algorithm.)

This approach was originally suggested by Masarotto (1987), who estimated the partial autocorrelation coefficients using an M-estimator of the type considered for independent and identically distributed multivariate data by Maronna (1976).

In both articles, in Möttönen et al. (1999) as well as in Masarotto (1987), it is stated that if $|\phi_{hh}| < 1$, the sequence $\rho_x(1), \rho_x(2), \dots,$ obtained by (4.13) is positive definite. Although the ϕ_{hh}'s satisfy $|\phi_{hh}| < 1$ except when the process x_t is deterministic (cf. Maronna et al., 2006), because the ϕ_{hh}'s are correlation coefficients, this is only true in theory, but definitely not for a realisation of x_t. We are able to give at least one counter-example that yields a non-positive definite sequence although $|\phi_{hh}| < 1$.

4.5 Ensuring Positive Semidefiniteness

We note that the classical estimators of the autocovariance function as well as of the autocorrelation function usually used in time series analysis are positive semidefinite. Unfortunately, all estimators described in the previous sections are not necessarily positive semidefinite. Hence, using those estimators, we should ensure non-negative definiteness of the sample autocovariance function and the sample autocorrelation function, respectively.

In their article Ma and Genton (2000) state that this can be achieved by shrinking (linear or nonlinear), the eigenvalue method or by the scaling method which were proposed by Rousseeuw and Molenberghs (1993). However, of the three methods only shrinking is applicable in the time series context. This is due to the fact that shrinking will always by definition transform equal correlations to the same value. In order to obtain a positive semidefinite autocorrelation function the proposed methods are applied to the corresponding autocorrelation matrix which is a Toeplitz matrix. If we now use the eigenvalue method or the scaling method we will obtain a positive

semidefinite matrix but it will not necessarily be a Toeplitz matrix anymore because equal correlations may be transformed to different values. If we further estimate the sample autocorrelation function by, e.g., taking the first row of this positive semidefinite matrix it may not be positive semidefinite itself!

In the following we will briefly describe the *nonlinear shrinking method* which was originally proposed by Devlin et al. (1975). If the corresponding autocorrelation matrix of the robustly estimated sample autocorrelation function is not positive semidefinite, then each off-diagonal element r_{ij}, $i \neq j$, is replaced by r_{ij}^* via the following transformation procedure

$$r_{ij}^* = \begin{cases} f^{-1}(f(r_{ij}) + \delta) & \text{if} \quad r_{ij} < -f^{-1}(\delta) \\ 0 & \text{if} \quad |r_{ij}| \leq f^{-1}(\delta) \\ f^{-1}(f(r_{ij}) - \delta) & \text{if} \quad r_{ij} > f^{-1}(\delta) \end{cases} \quad . \tag{4.15}$$

Here, δ is a small positive constant, e.g., $\delta = 0.05$, and f is a monotone increasing function with $f : [-\infty, \infty] \rightarrow [-1, 1]$. This procedure is repeated until the resulting matrix is positive semidefinite. Devlin et al. (1975) suggest to define f as follows:

$$\begin{aligned} f(x) &= \frac{e^x - e^{-x}}{e^x + e^{-x}} = \tanh(x) \\ f^{-1}(x) &= \frac{1}{2}\ln(\frac{1+x}{1-x}) = \tanh^{-1}(x) \ . \end{aligned} \tag{4.16}$$

The authors remark that f^{-1} in (4.16) is Fisher's variance-stabilizing transform for correlation coefficients.

The shrinking method has the advantage of being easy to implement, because (4.15) operate on each correlation coefficient separately.

4.6 Simulation Study

The outline of our simulation study is as follows: In both cases we simulate a core process x_t of length $n = 100$. For the first example x_t is chosen to be a moving average process of order 1 given by

$$x_t = \varepsilon_t - 0.5\varepsilon_{t-1} \ , \tag{4.17}$$

with $\varepsilon_t \sim \mathcal{N}(0,1)$. We note that for moving average processes the true autocorrelation function can be easily calculated. Here, it is given by

$$\rho_x(h) = \begin{cases} 1 & \text{if} \quad h = 0 \\ \frac{\theta_1}{1+\theta_1^2} = -0.4 & \text{if} \quad |h| = 1 \\ 0 & \text{if} \quad |h| > 1 \end{cases}. \tag{4.18}$$

For the second example x_t is chosen to be an autoregressive process of order 2 given by

$$x_t = x_{t-1} - 0.9x_{t-2} + \varepsilon_t , \tag{4.19}$$

again with $\varepsilon_t \sim \mathcal{N}(0,1)$. Additionally, for both models the additive outliers v_t are simulated from a contaminated normal distribution with degenerate central component, i.e.,

$$\mathcal{CN}(\gamma, 0, \sigma^2) = (1 - \gamma)\mathcal{N}(0,0) + \gamma\mathcal{N}(0, \sigma^2) , \tag{4.20}$$

with $\sigma^2 = 10^2$. The contamination γ is varied from 0% to 20% by steps of 5%. That means that with probability γ, v_t is an additive outlier with $v_t \neq 0$. To obtain the contaminated process y_t, the v_t's are added to the core process x_t. For both models as well as for each level of contamination 400 series are simulated.

For each contaminated series we compute the sample autocorrelation function using the three methods described above, i.e., the highly robust autocorrelation function (Ma and Genton, 2000) and the ones based on Spearman's rank correlation coefficient (Ahdesmäki et al., 2005) and on partial autocorrelation (Möttönen et al., 1999). Additionally, the mean-squared error, denoted by *MSE*, between the true autocorrelation function and the different estimates averaged over the lags h is computed, i.e.,

$$MSE_{\widehat{\rho}_y(h)} := \underset{h}{\text{mean}}((\rho_y(h) - \widehat{\rho}_y^{(rob)}(h))^2) \tag{4.21}$$

where $\rho_y(h)$ and $\widehat{\rho}_y^{(rob)}(h)$ denote the true autocorrelation function and the estimate, respectively, and $rob \in \{Q, S, pacf\}$.

Furthermore, we apply nonlinear shrinking to obtain positive semidefinite sample autocorrelation functions. Again, the mean-squared error between the true autocorrelation function and the different positive semidefinite estimates averaged over the lags h is computed.

These positive semidefinite sample autocorrelation functions of the two most competitive methods are used to calculate spectral density estimates of each process. The factor $\widehat{C}$ in (4.2) is estimated by $\widehat{\gamma}_y^{(Q)}(0)$. Last, the deviation of each estimated spectral density function from the true spectral density function is measured in the sense of the squared L_2-norm, i.e.,

$$err^2_{\widehat{S}(f)} := \|\widehat{S}(f) - S(f)\|^2 = \int (\widehat{S}(f) - S(f))^2 df \ , \qquad (4.22)$$

where $\widehat{S}(f)$ and $S(f)$ denote the estimated spectral density function and true one.

4.7 Results

In Figure 4.1 the true autocorrelation functions together with typical robust estimates of the above described methods for different levels of contamination are presented. The autocorrelation functions of the moving average process are plotted in the left column, whereas the ones of the autoregressive process are displayed in the right column. Each plot contains four graphs. The thick black line is the true autocorrelation function, whereas the light grey line represents the estimate obtained by the highly robust autocorrelation function denoted by 'Q'. The dashed grey line is the estimate using Spearman's rank correlation coefficient and the estimate based on partial autocorrelation is plotted as a dotted dark grey line. The latter estimates are denoted by 'S' and 'pacf', respectively.

In Figure 4.1 it is visible that the deviation between the true autocorrelation function and the different estimates becomes larger the higher the amount of contamination. Especially, this can be seen for the autoregressive process at higher lags. The difference is largest for the estimates based on partial autocorrelation.

This observation is also confirmed by Figure 4.2 which displays the mean-squared errors between the true autocorrelation function and each estimate for all levels of contamination. The estimator based on partial autocorrelation performs worst yielding the largest errors. Depending on the simulated time series model the estimates based on Spearman's rank correlation coefficient are slightly better or equal to the estimates obtained by the highly robust autocorrelation function.
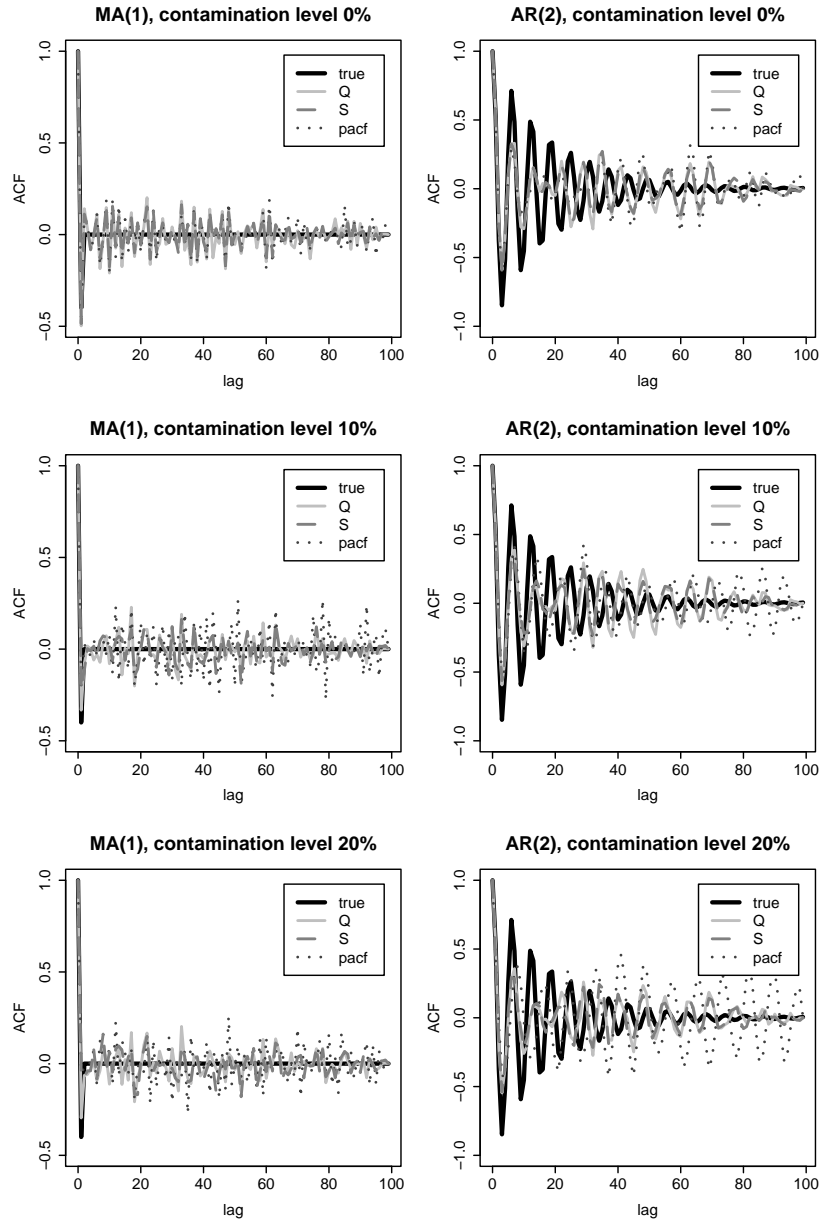
Figure 4.1: Typical robust estimates of the autocorrelation function, left column MA(1) process, right AR(2) process
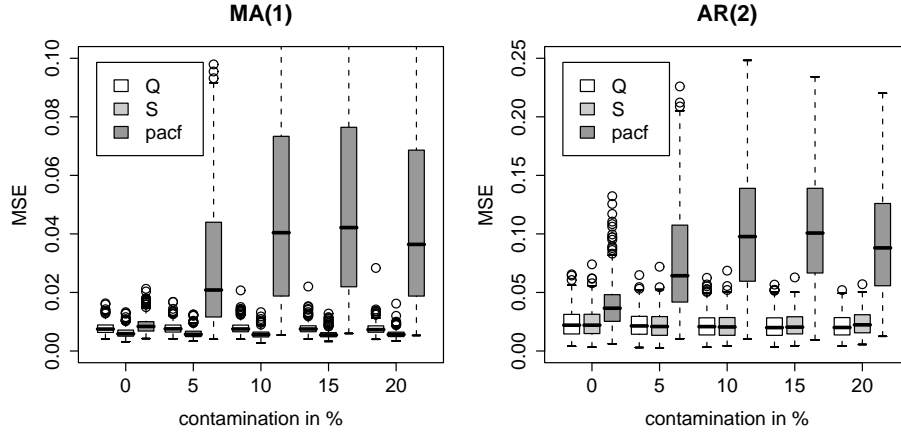
Figure 4.2: Boxplots of the errors, left MA(1) process, right AR(2) process

As mentioned before the sample autocorrelation functions are not necessarily positive semidefinite. Hence, we apply nonlinear shrinking to get positive semidefinite estimates. In Figure 4.3 typical results are plotted in the same way as in Figure 4.1. Again, the true autocorrelation function and the positive semidefinite estimates of the moving average process are plotted in the left column, whereas the ones of the autoregressive process are displayed in the right column.

In Figure 4.3 we see that due to nonlinear shrinking the estimated autocorrelations are shrunken to zero at many lags. This is especially true for estimates obtained by the highly robust autocorrelation function.

Again, in Figure 4.4 the mean-squared errors between the true autocorrelation function and each positive semidefinite estimate for all levels of contamination are displayed. We get the same result as before except for the moving average process where the method based on Spearman's rank correlation coefficient now yields slightly larger errors than the one using the highly robust autocorrelation function. We note that this is solely due to the effect of nonlinear shrinking because, as already mentioned before, especially the estimated autocorrelations obtained by the highly robust autocorrelation function are shrunken to zero at most lags and due to the fact that the true autocorrelation function of a moving average process of order q is zero at lags h with $h > q$.

Based on the previous results we use the estimates obtained by the highly

29
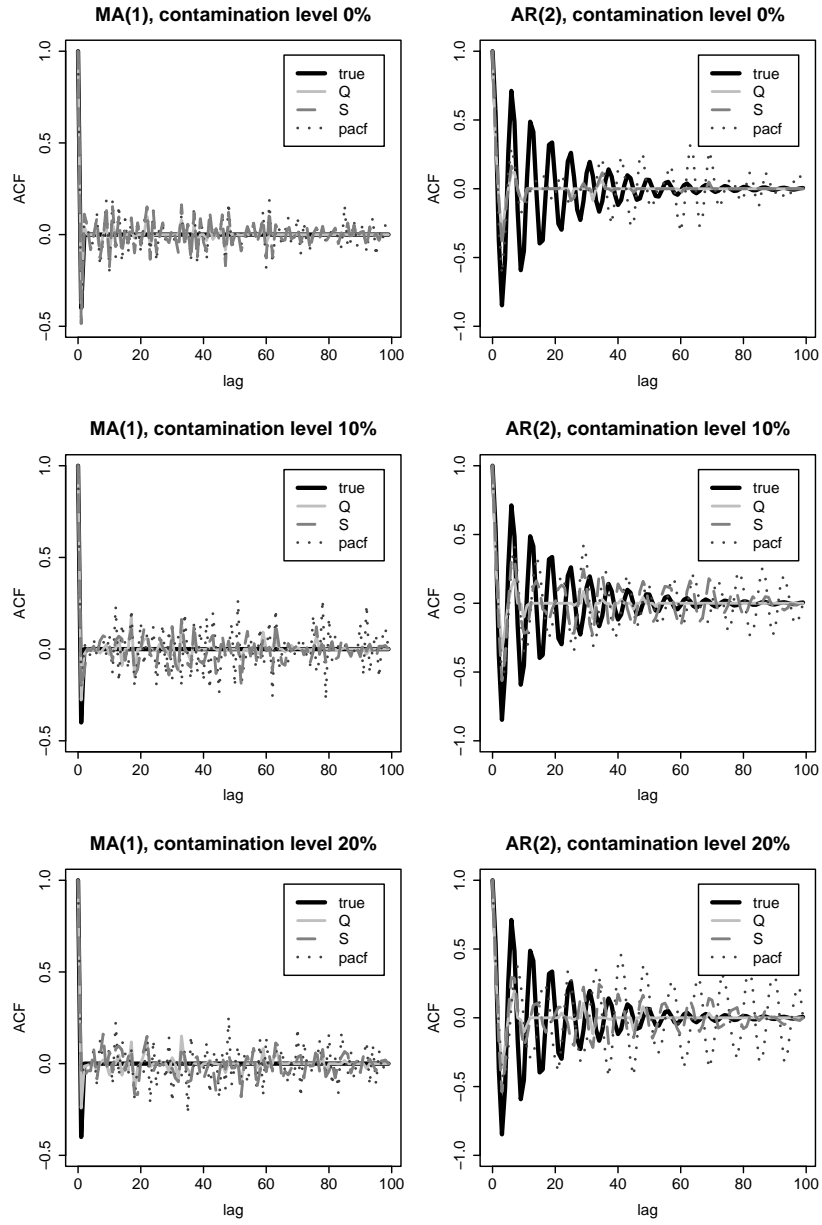
Figure 4.3: Typical robust positive semidefinite estimates of the autocorrelation function, left column MA(1) process, right AR(2) process
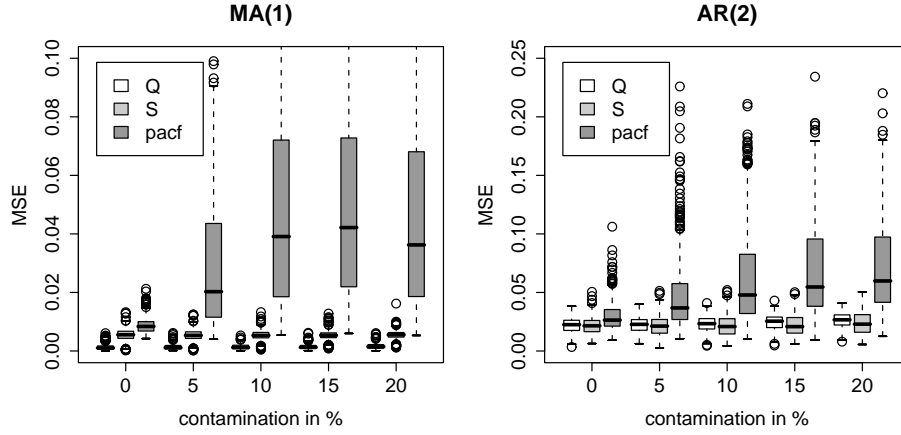
Figure 4.4: Boxplots of the errors in the positive semidefinite case, left MA(1) process, right AR(2) process

robust autocorrelation function and the ones based on Spearman's rank correlation coefficient to estimate the spectral density functions of both time series models.

Figure 4.5 and 4.7 try to visualize the results of our simulation study. For both methods and each level of contamination seven curves are plotted on a logarithmic scale. The bold black line represents the true spectral density function, whereas the thin black line is the spectral density estimate of one realization out of 400. Moreover, we calculate frequency by frequency the minimum and maximum, the first and third quartile and median value of all spectral density estimates. Connecting all median values we obtain the grey line, to which we will refer hereafter as median spectral density function. In the same sense we refer to all minimum values as minimum spectral density function, and so on. Hence, the lower and upper dotted lines are the minimum and maximum spectral density functions, whereas the lower and upper dashed lines represent the first and third quartile spectral density functions. The results obtained by using the highly robust autocorrelation function are plotted in the left column, whereas the results based on Spearman's rank correlation coefficient are displayed in the right column.

In Figure 4.5 the results for the moving average process are displayed. As expected, for both methods the dispersion of the spectral density estimates becomes larger the higher the amount of contamination. Additionally we see
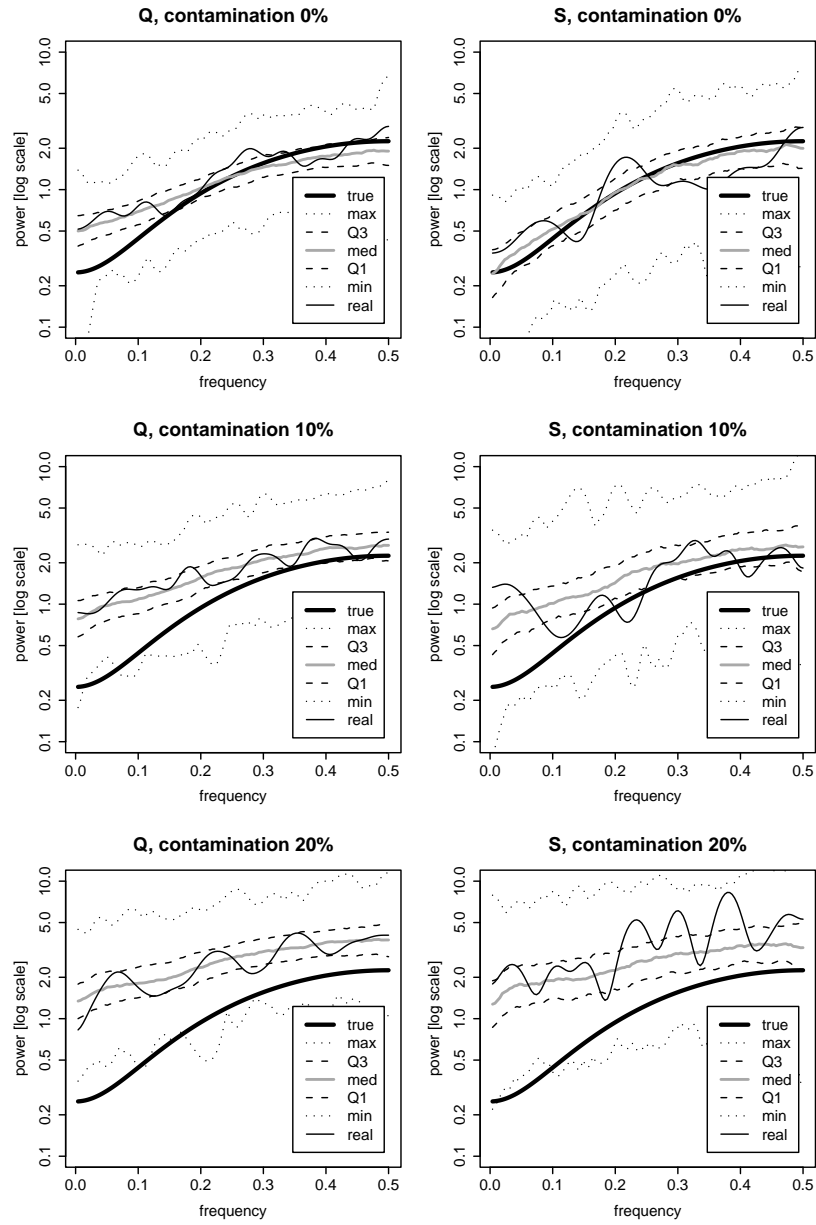
Figure 4.5: Robust spectral density estimates of the MA(1) process, left column 'Q', right 'S'
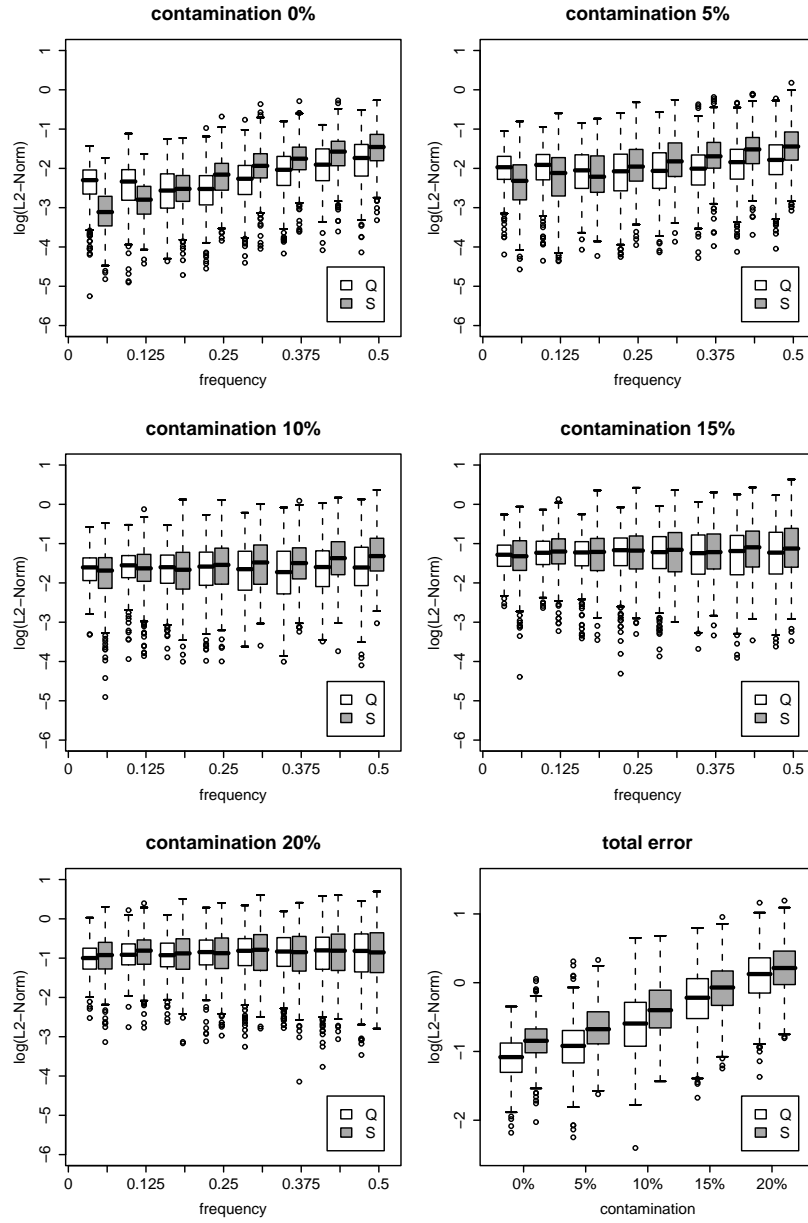
Figure 4.6: Boxplots of the errors, MA(1) process

that the spectral density function is all the more overestimated the higher the contamination.

However, this effect is not visible for the autoregressive model in Figure 4.7. For both methods we get bad spectral density estimates even in the case of no contamination, whereas the dispersion of the estimates stays almost the same for all different levels of contamination.

Next, we try to visualize the squared errors of the estimated spectral density functions. First, the logarithm of the squared errors is taken. For both models Figure 4.6 and 4.8 show boxplots of the squared errors in eight equally-sized frequency bands as well as the total squared errors (bottom right) for all different levels of contamination.

As we can see for the moving average process in Figure 4.6, the squared errors become larger the higher the contamination, especially at higher frequencies. Moreover, the method based on Spearman's rank correlation coefficient yields slightly larger errors than the one using the highly robust autocorrelation function in almost all frequency bands for all levels of contamination. We recall that the true autocorrelation function of an MA(q) process is zero at lags h with $h > q$. This and the fact, that by nonlinear shrinking especially the estimated autocorrelations obtained by the highly robust autocorrelation function are shrunken to zero at most lags, are responsible for the better performance.

For the autoregressive process the squared errors are displayed in Figure 4.8. As expected by the results plotted in Figure 4.7 the errors do not differ very much for both methods and, looking at the different frequency bands separately, stay almost the same independently of the amount of contamination. However, looking at the total squared errors, we see that the method based on Spearman's rank correlation coefficient performs slightly better than the one using the highly robust autocorrelation function. The largest contribution to the total squared error is the amount of the frequency band where the spectral density function has its peak. There the squared errors using the highly robust autocorrelation function are larger than the ones based on Spearman's rank correlation coefficient. Moreover, we see that the total squared errors are in the same range for all contamination levels.
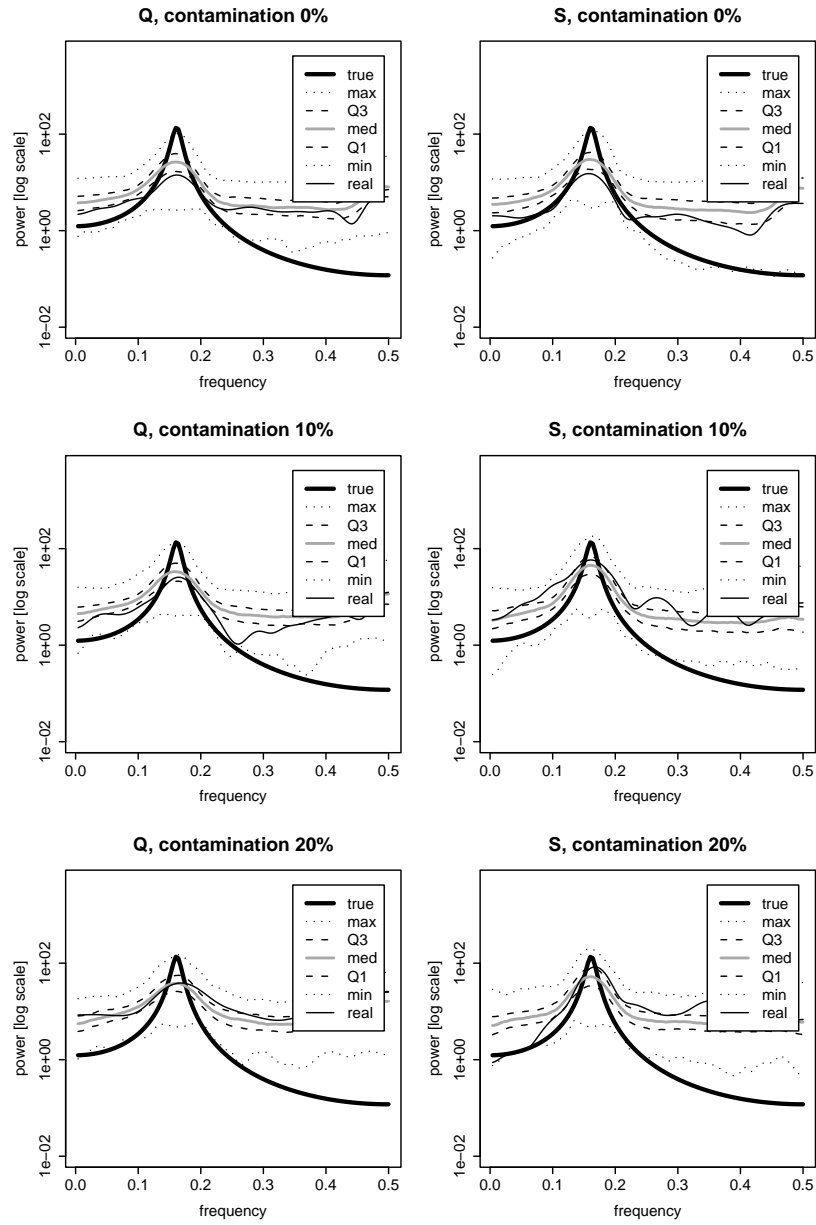
Figure 4.7: Robust spectral density estimates of the AR(2) process, left column 'Q', right 'S'
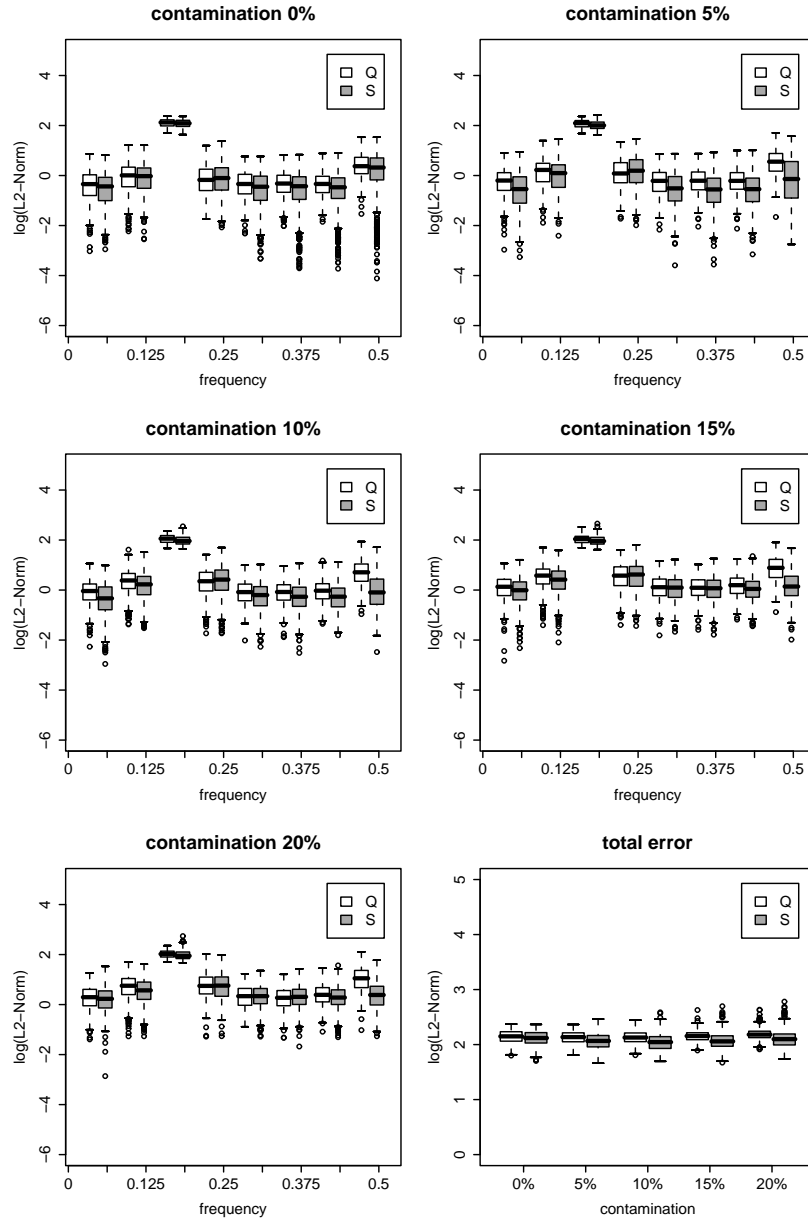
Figure 4.8: Boxplots of the errors, AR(2) process

36

4.8 Discussion

In this chapter we used the spectral density estimator (4.1) originally proposed by Blackman and Tukey (1958) which is based on the Fourier transform of the autocovariance function. In order to robustly estimate the spectral density function, we simply replaced the autocovariance function by a robust estimate. We compared estimates obtained by a highly robust autocovariance function (Ma and Genton, 2000) to others based on Spearman's rank correlation coefficient (Ahdesmäki et al., 2005) or on partial autocorrelation (Möttönen et al., 1999).

Unfortunately, all these robust alternatives yield estimates that may not necessarily be positive semidefinite. Hence, we applied nonlinear shrinking to ensure non-negative definiteness.

The results of the simulation study suggest that the method based on Spearman's rank correlation coefficient performs better than the one using the highly autocorrelation function and the method based on partial autocorrelation performs worst. As already mentioned before, the slightly better performance of the method using the highly robust autocorrelation function in case of the moving average process is on one hand due to the effect of nonlinear shrinking because especially the estimated autocorrelations obtained by this method are shrunken to zero at most lags and on the other hand due to the fact that the true autocorrelation function of an $MA(q)$ process is zero at lags h with $h > q$.

We note that there exist many other alternatives to estimate correlation robustly. However, Croux and Filzmoser (2003) already studied several bivariate measures of association and found out that Spearman's rank correlation coefficient has good robustness properties. Hence, we limit our survey to the above described methods.

Chapter 5

Estimation of Hyper Parameters

5.1 Generalized Maximum Likelihood Type Estimation of Autoregressive Models

As autoregressive parameter estimates obtained by maximum likelihood type estimation (M-estimation) are not robust toward situations where the given time series is contaminated by additive outliers—the special case of first-order autoregressive models was treated in Denby and Martin (1979) and Martin and Jong (1976)—one might use other procedures, such as the generalized M-estimation procedure (GM-estimation) described in this section, to compute robust estimates.

5.1.1 Estimating the Location Parameter

We now concentrate on estimating an autoregressive model of order p (AR(p) model). Let $\{y_t, t = 1, \ldots, n\}$ denote the observed process. First, we center the data robustly by using an ordinary location M-estimate $\widehat{\mu}$ (Huber, 1964). This is analogous to the usual approach for estimating autoregression parameters via least squares estimation where the sample mean is used to center the data.

If the robustly centered observation is—for notational convenience—again denoted by y_t, then the AR(p) model can be written in the linear model form

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\phi} + \boldsymbol{\varepsilon} \ , \tag{5.1}$$

where $\boldsymbol{y} = (y_{p+1}, \ldots, y_n)^\top$, $\boldsymbol{Z} = [\boldsymbol{z}_{p+1}, \ldots, \boldsymbol{z}_n]^\top$ with $\boldsymbol{z}_t = (y_{t-1}, \ldots, y_{t-p})^\top$, $t = p+1, \ldots, n$, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)^\top$ and $\boldsymbol{\varepsilon} = (\varepsilon_{p+1}, \ldots, \varepsilon_n)^\top$.

5.1.2 Stating the Problem

As the influence curve of the M-estimator of $\boldsymbol{\phi}$ is bounded in y but unbounded in $\boldsymbol{z}$, which is an undesirable property if additive outliers occur, the basic idea of GM-estimation is to modify the M-estimation problem so that the resulting influence curve is a bounded and continuous function of the data. The GM-estimators $\widehat{\boldsymbol{\phi}}$ and $\widehat{\sigma}_\varepsilon$ are analogues of bounded-influence regression estimators and in context of autoregression, are also called bounded-influence autoregression (BIAR) estimators. They are given as an extension of Huber's proposal (Huber, 1973) for robust regression by the general minimum problem

$$h(\boldsymbol{\phi}', \sigma'_\varepsilon) = \sum_{t=p+1}^n u_t v_t \rho_1 \left(\frac{y_t - \boldsymbol{z}_t^\top \boldsymbol{\phi}'}{u_t \sigma'_\varepsilon} \right) \sigma'_\varepsilon + c\sigma'_\varepsilon = \min! \ , \tag{5.2}$$

where the minimum in $\boldsymbol{\phi}'$ and σ'_ε has to be achieved and $\rho_1(.)$ is a symmetric robustifying loss function. The constant c is chosen so that $\widehat{\sigma}_\varepsilon$ is a consistent estimate of σ_ε when the series is free of outliers with $\mathcal{N}(0, \sigma_\varepsilon^2)$-distributed innovations. The role of the u_t's and v_t's is to downweight those summands in (5.2) for which $\boldsymbol{z}_t$ is a poor predictor because it is "large" due to an outlier in one or more of its components.

Differentiating $h(\boldsymbol{\phi}', \sigma'_\varepsilon)$ with respect to $\boldsymbol{\phi}'$ and σ'_ε and equating the resulting expressions to zero yields a system of equations defining $\widehat{\boldsymbol{\phi}}$ and $\widehat{\sigma}_\varepsilon$

$$\sum_{t=p+1}^n v_t \psi_1 \left(\frac{y_t - \boldsymbol{z}_t^\top \widehat{\boldsymbol{\phi}}}{u_t \widehat{\sigma}_\varepsilon} \right) \boldsymbol{z}_t = \boldsymbol{0} \ , \tag{5.3}$$

where $\psi_1(s) = \mathrm{d}\rho_1(s)/\mathrm{d}s$ is odd and should be bounded and continuous, and

$$\sum_{t=p+1}^n u_t v_t \chi_1 \left(\frac{y_t - \boldsymbol{z}_t^\top \widehat{\boldsymbol{\phi}}}{u_t \widehat{\sigma}_\varepsilon} \right) = c \ , \tag{5.4}$$

with $\chi_1(s) = s\psi_1(s) - \rho_1(s)$. If *Huber's monotone psi-function* (Huber, 1964), which is defined by

$$\psi_H(s) = \begin{cases} s & \text{if} \quad |s| \leq c_H \\ c_H \operatorname{sgn}(s) & \text{if} \quad |s| > c_H \ , \end{cases} \tag{5.5}$$

is used, i.e., $\psi_1(s) = \psi_H(s)$, then $\chi_1(s) = \frac{1}{2}\psi_1^2(s)$ and (5.4) becomes

$$\sum_{t=p+1}^{n} u_t v_t \psi_1^2 \left(\frac{y_t - \boldsymbol{z}_t^\top \widehat{\boldsymbol{\phi}}}{u_t \widehat{\sigma}_\varepsilon} \right) = 2c \; . \tag{5.6}$$

Further, for computational convenience, we set $B = 2c / \sum_{t=p+1}^{n} u_t v_t$, and get

$$\sum_{t=p+1}^{n} u_t v_t \left(\psi_1^2 \left(\frac{y_t - \boldsymbol{z}_t^\top \widehat{\boldsymbol{\phi}}}{u_t \widehat{\sigma}_\varepsilon} \right) - B \right) = 0 \; . \tag{5.7}$$

For a Mallows type GM-estimator (Mallows, 1976) every u_t is equal to 1 and $v_t = \psi_2(d_t)/d_t$, where d_t denotes the "largeness" of $\boldsymbol{z}_t$ (distance in the $\boldsymbol{z}_t$-space). A Schweppe type GM-estimator (Schweppe et al., 1975) uses $u_t = v_t = \psi_2(d_t)/d_t$. ψ_2 again is odd, bounded and continuous, for example one of the well-known psi-functions. We will use *Tukey's redescending bisquare psi-function* (Beaton and Tukey, 1974), which is given by

$$\psi_B(s) = \begin{cases} s(1 - (s/c_B)^2)^2 & \text{if} \quad |s| \le c_B \\ 0 & \text{if} \quad |s| > c_B \; . \end{cases} \tag{5.8}$$

5.1.3 Computational Details

The "largeness" d_t of $\boldsymbol{z}_t$ can be assessed by

$$d_t = d_t(\boldsymbol{z}_t) = \left(p^{-1} \boldsymbol{z}_t^\top \widehat{\boldsymbol{C}}_p^{-1} \boldsymbol{z}_t \right)^{\frac{1}{2}} \; , \tag{5.9}$$

where $\widehat{\boldsymbol{C}}_p$ is a robust estimate of the unknown $p \times p$ covariance matrix $\boldsymbol{C}_p$ of the outlier-free AR(p) process. Martin (1980) estimates $\boldsymbol{C}_p^{-1}$ in the following way: Suppose that $\{x_t\}$ is a zero mean Gaussian process (not necessarily an autoregression process of order p) with $p \times p$ covariance matrix $\boldsymbol{C}_p$ and let $\phi_{k1}, \ldots, \phi_{kk}$, $k = 1, \ldots, p-1$, be the coefficients of the minimum mean squared error predictor of x_t based on $x_{t-1}, \ldots, x_{t-k}$. The corresponding prediction error variance is denoted by $\sigma_{\varepsilon,k}^2$. Then $\boldsymbol{C}_p^{-1}$ has the factorization (Akaike, 1969)

$$\boldsymbol{C}_p^{-1} = \boldsymbol{A}_p^\top \boldsymbol{A}_p \; , \tag{5.10}$$

where

$$(\boldsymbol{A}_p)_{ij} = \begin{cases} \dfrac{-\phi_{p-i,j-i}}{\sigma_{\varepsilon,p-i}} & \text{if} \quad j > i \\ \dfrac{1}{\sigma_{\varepsilon,p-i}} & \text{if} \quad j = i \\ 0 & \text{if} \quad j < i \;, \end{cases}$$

with $\sigma_{\varepsilon,0}$ denoting the scale of the x_t's. Thus $\boldsymbol{C}_p^{-1}$ has a representation in terms of prediction coefficients and prediction error variances. The zero length predictor is just the mean value.

When, as in the present research, a location M-estimate $\widehat{\mu}$ is used to estimate the location parameter, a natural estimate of $\sigma_{\varepsilon,0}$ is the scale estimate of Huber's Proposal 2 (Huber, 1964) obtained during the computation of $\widehat{\mu}$.

When $p > 1$ we also need estimates of $\phi_{p-i,1}, \ldots, \phi_{p-i,p-i}$ and $\sigma_{\varepsilon,p-i}$, $i < p$. For $\phi_{p-i,j}$ we use the j-th element of the GM-estimate $\widehat{\boldsymbol{\phi}}$ from an AR$(p-i)$ fit. For $\sigma_{\varepsilon,p-i}$ we have the robust residual scale estimate $\widehat{\sigma}_\varepsilon$ from that fit. Thus we successively fit autoregressive models of orders $1, \ldots, p$ so that for each autoregression $\widehat{\boldsymbol{C}}_p^{-1}$ can be obtained via (5.10) from the lower order results.

Before estimating $\boldsymbol{\phi}$ and σ_ε the weights u_t and v_t, $t = p+1, \ldots, n$, which are constant for a fixed p and time series $y_1, \ldots, y_n$, must be determined.

We note that (5.3) can be written as follows:

$$\sum_{t=p+1}^{n} \frac{r_t}{\widehat{\sigma}_\varepsilon} \frac{v_t}{u_t} \frac{\psi_1(r_t/u_t\widehat{\sigma}_\varepsilon)}{r_t/u_t\widehat{\sigma}_\varepsilon} \boldsymbol{z}_t = \boldsymbol{0} \;, \tag{5.11}$$

where r_t denotes the residual $r_t = y_t - \boldsymbol{z}_t^\top \widehat{\boldsymbol{\phi}}$.

This reveals that GM-estimators can be regarded as weighted least squares estimators with weights $w_t = v_t\psi_1(r_t/u_t\widehat{\sigma}_\varepsilon)/(r_t/\widehat{\sigma}_\varepsilon)$ whose weights depend on the residuals and therefore on $\widehat{\boldsymbol{\phi}}$. Hence, (5.11) is only an implicit equation. However, the following _iteratively reweighted least squares_ (IWLS) _algorithm_ could be used to estimate $\boldsymbol{\phi}$ and σ_ε simultaneously. Dutter (1983) uses a similar algorithm to compute bounded-influence estimators for linear regression.

5.1.4 IWLS Algorithm

Let starting values $\widehat{\boldsymbol{\phi}}^{(0)}$ and $\widehat{\sigma}_\varepsilon^{(0)}$, and a tolerance level κ be given.

(i) Set the iteration counter $m = 0$.

(ii) Denote $r_t^{(m)} = y_t - \boldsymbol{z}_t^\top \widehat{\boldsymbol{\phi}}^{(m)}$, $t = p+1, \ldots, n$.

(iii) Compute a new estimate for σ_ε using (5.7):

$$(\widehat{\sigma}_\varepsilon^{(m+1)})^2 = \left(B \sum_{t=p+1}^n u_t v_t \right)^{-1} \left(\sum_{t=p+1}^n u_t v_t \psi_1^2 \left(\frac{r_t^{(m)}}{u_t \widehat{\sigma}_\varepsilon^{(m)}} \right) \right) (\widehat{\sigma}_\varepsilon^{(m)})^2 \ .$$

(iv) Calculate weights $w_t^{(m)}$ considering that $u_t = 1$ for a Mallows type estimator and $u_t = v_t$ for the Schweppe type estimator:

$$w_t^{(m)} = \begin{cases} v_t \psi_1 \left(\frac{r_t^{(m)}}{u_t \widehat{\sigma}_\varepsilon^{(m+1)}} \right) \Big/ \left(\frac{r_t^{(m)}}{\widehat{\sigma}_\varepsilon^{(m+1)}} \right) & \text{if } r_t^{(m)} \neq 0, \ u_t \neq 0 \\ v_t / u_t & \text{if } r_t^{(m)} = 0, \ u_t \neq 0 \\ 1 & \text{if } r_t^{(m)} = u_t = v_t = 0 \\ 0 & \text{if } r_t^{(m)} \neq 0, \ u_t = v_t = 0 \ , \end{cases}$$

where $t = p+1, \ldots, n$. Define a diagonal matrix $\boldsymbol{W}^{(m)}$ with $w_t^{(m)}$ as its $(t-p)$-th diagonal element.

(v) Solve

$$\sum_{t=p+1}^n (r_t^{(m)} - \boldsymbol{z}_t^\top \widehat{\boldsymbol{\tau}}^{(m)})^2 w_t^{(m)} = \min! \ ,$$

for $\widehat{\boldsymbol{\tau}}^{(m)}$, which can be computed by

$$\widehat{\boldsymbol{\tau}}^{(m)} = (\boldsymbol{Z}^\top \boldsymbol{W}^{(m)} \boldsymbol{Z})^{-1} \boldsymbol{Z}^\top \boldsymbol{W}^{(m)} \boldsymbol{y} - \widehat{\boldsymbol{\phi}}^{(m)} \ ,$$

where $\boldsymbol{Z}$ and $\boldsymbol{y}$ are defined by (5.1).

(vi) Compute new estimates for $\boldsymbol{\phi}$ by

$$\widehat{\boldsymbol{\phi}}^{(m+1)} = \widehat{\boldsymbol{\phi}}^{(m)} + \omega \widehat{\boldsymbol{\tau}}^{(m)} \ ,$$

where $0 < \omega < 2$ is an arbitrary relaxation factor.

(vii) Stop, if

$$|\widehat{\sigma}_\varepsilon^{(m)} - \widehat{\sigma}_\varepsilon^{(m+1)}| < \kappa \widehat{\sigma}_\varepsilon^{(m+1)}$$

and if the difference of all parameters from the m-th step to the $(m+1)$-th is less than κ times their approximate standard deviation, i.e.,

$$|\omega \widehat{\tau}_k^{(m)}| < \kappa \widehat{\sigma}_\varepsilon^{(m+1)} \sqrt{z^{kk}} \ , \quad k = 1, \ldots, p \ ,$$

where z^{kk} is the k-th diagonal element in $(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}$.

(viii) Augment $m = m + 1$ and go to Step (ii).

As starting values $\widehat{\phi}^{(0)}$ and $\widehat{\sigma}_\varepsilon^{(0)}$ one can use the least squares estimates of ϕ and σ_ε (cf. Martin, 1980; Martin and Thomson, 1982; Stockinger and Dutter, 1987). In our case we use the least squares estimate $\widehat{\phi}_{LS}$ of ϕ and the median absolute deviation (MAD) of the residuals given by

$$\widehat{\sigma}_\varepsilon^{(0)} = \underset{t}{\mathrm{med}}\, |r_t^{(0)} - \underset{s}{\mathrm{med}}\, r_s^{(0)}|/.6745 \ ,$$

with $r_t^{(0)} = y_t - \mathbf{z}_t^\top \widehat{\phi}_{LS}$. This was proposed by Martin and Zeh (1978).

It can be shown that the estimating equations (5.3) and (5.7) have a unique solution if ψ_1 is non decreasing, e.g. $\psi_1 = \psi_H$. However, there may be many solutions if ψ_1 is redescending.

Nevertheless, a redescending psi-function, e.g. Tukey's ψ_B, shows better performances at extremely heavy-tailed outlier distributions. Thus the following overall computational strategy has been adopted.

First, to compute GM-estimates $\widehat{\phi}$ and $\widehat{\sigma}_\varepsilon$, several iterations are made using a non decreasing ψ_1 such as Huber's ψ_H until there is relatively little change in the estimates. The resulting GM-estimates are then used as starting values for one or two iterations using a redescending ψ_1 such as Tukey's ψ_B in place of ψ_H. Martin and Thomson (1982) report that in case of higher order autoregression further iterations using ψ_B may lead to bad estimates because of multiple roots.

The motivation for the above strategy is rather obvious. It is hoped that the GM-estimates based on a non decreasing ψ_1 are close to the "appropriate" solution of the estimating equations based on a redescending ψ_1.

Although the transformation from equation (5.4) to (5.7) only holds for Huber's monotone psi-function and Stockinger and Dutter (1987) proposed to modify the IWLS algorithm by omitting Step (iii) when using a redescending ψ_1, we always keep this step and use equation (5.7) to update $\widehat{\sigma}_\varepsilon$ regardless of the used psi-function as proposed by Martin and Zeh (1978). This method works well in practice and simulation studies indicate that it is sometimes crucial to obtain estimates based on ψ_B.

The equation for B which insures that the innovations' scale estimate $\widehat{\sigma}_\varepsilon$ of (5.3) and (5.7) is a consistent estimate of σ_ε when the series is free of outliers with $\mathcal{N}(0, \sigma_\varepsilon^2)$-distributed innovations is

$$B = E(\psi_1^2(\zeta)) \ , \ \text{with } \zeta \sim \mathcal{N}(0, 1) \ . \tag{5.12}$$

43

For example, the use of $c_H = 1.5$ for $\psi_1 = \psi_H$ and $c_B = 5.0$ for $\psi_1 = \psi_B$ yields $B_H = 0.7785$ and $B_B = 0.6395$ respectively. Note, if $\psi_1(s) = s$ which is equivalent to $c_H = c_B = \infty$ then $B = E(\zeta^2) = \text{Var}(\zeta) = 1$ with $\zeta \sim \mathcal{N}(0, 1)$.

5.1.5 Selection of Order p

Martin and Thomson (1982) propose the following procedure to select the order p of the autoregressive approximation. For increasing orders p BIAR estimates are computed and the estimated innovation scale estimates $\widehat{\sigma}_\varepsilon(p)$ are examined for each order. The final order is selected as that value of p for which $\widehat{\sigma}_\varepsilon(p + 1)$ is not much smaller than $\widehat{\sigma}_\varepsilon(p)$, e.g., less than a 10-percent decrement as suggested by Martin and Thomson (1982).

Another robust order-selection rule based on BIAR estimates and motivated by Akaike's minimization criterion Akaike (1974) was proposed by Martin (1980).

5.2 Robust Autoregressive Parameter Estimation via the MCD Estimator

Instead of applying GM-estimation to obtain robust autoregressive parameter estimates another way would be to use the *method of moments*. Liu et al. (2004) propose to estimate the autocorrelation coefficients by a robust multivariate location and scatter estimator separately. They suggest to use the minimum covariance determinant (MCD) estimator developed by Rousseeuw (1984) and Rousseeuw and Van Driessen (1999). Once all the $p + 1$ autocorrelation coefficients up to lag p are estimated the Yule-Walker equations are solved to obtain the AR(p) model parameters.

Assuming that the observed process is denoted by $\{y_t, t = 1, \ldots, n\}$, the proposed algorithm consists of the following steps:

(i) Estimate the mean μ and variance $\gamma_y(0)$ of $\{y_t, t = 1, \ldots, n\}$ using a univariate robust estimator of location and scale, e.g., the M-estimator of Huber's Proposal 2 (Huber, 1964).

(ii) Form new bivariate datasets $\boldsymbol{Y}_t^{(h)} = \{(y_t, y_{t-h})^\top, t = h + 1, \ldots, n\}$, $h = 1, \ldots, p$. For each bivariate dataset $\boldsymbol{Y}_t^{(h)}$ the covariance matrix,

denoted by

$$\mathbf{\Gamma}^{(h)} = \begin{pmatrix} \gamma_{11}^{(h)} & \gamma_{12}^{(h)} \\ \gamma_{21}^{(h)} & \gamma_{22}^{(h)} \end{pmatrix} \, ,$$

can be estimated by any robust multivariate location and scatter estimator, e.g., Rousseeuw's MCD estimator. Then, the estimated autocovariance coefficient at lag h is given by $\widehat{\gamma}_y(h) = \gamma_{12}^{(h)}$, whereas the estimated autocorrelation coefficient at lag h can be calculated by $\widehat{\rho}_y(h) = \gamma_{12}^{(h)}/\sqrt{\gamma_{11}^{(h)}\gamma_{22}^{(h)}}$.

(iii) Solve the Yule-Walker equations to obtain AR(p) model parameters $\widehat{\phi}_1, \ldots, \widehat{\phi}_p$ and $\widehat{\sigma}_\varepsilon$.

In order to avoid matrix inversion we suggest to use the Durbin-Levinson algorithm to solve the Yule-Walker equations.

5.3 Simulation Study

The outline of our simulation study is as follows: First we simulate a core process x_t of length $n = 100$. x_t is chosen to be an autoregressive process of order 2 given by

$$x_t = x_{t-1} - 0.9x_{t-2} + \varepsilon_t \, , \tag{5.13}$$

with $\varepsilon_t \sim \mathcal{N}(0,1)$. Additionally, the additive outliers v_t are simulated from a contaminated normal distribution with degenerate central component, i.e.,

$$\mathcal{CN}(\gamma, 0, \sigma^2) = (1 - \gamma)\mathcal{N}(0,0) + \gamma\mathcal{N}(0, \sigma^2) \, , \tag{5.14}$$

with $\sigma^2 = 10^2$. The contamination γ is varied from 0% to 20% by steps of 5%. That means that with probability γ, v_t is an additive outlier with $v_t \neq 0$. To obtain the contaminated process y_t, the v_t's are added to the core process x_t. For each level of contamination this is done 400 times.

For each contaminated series we compute estimates of the innovations scale $\widehat{\sigma}_\varepsilon$ and the autoregressive parameters $\widehat{\phi}_1, \ldots, \widehat{\phi}_p$ using bounded-influence autoregression as well as the method proposed by Liu et al. (2004). We then compare them to the non-robust parameter estimates obtained by solving the Yule-Walker equations based on the classically estimated autocorrelation coefficients.

The order p of the autoregressive model is chosen according to the order-selection criterion proposed by Martin and Thomson (1982), which yields values of p from 2 to 3 subject to the contamination level. In order to be able to compare the results we choose an equal order p for all levels of contamination and fix it equal to 3.

5.4 Results

In Figure 5.1 the results of the simulation study are presented. The parameter estimates of the three compared methods and all contamination levels are displayed in Figure 5.1 using grouped boxplots.

The difference between the classically estimated autoregressive parameters and those obtained by the method proposed by Liu et al. (2004) labeled 'ACF' and 'MCD', respectively, is that for the first method the autocorrelation coefficients are estimated using the classical autocorrelation function whereas for the second the autocorrelation coefficients are computed by Rousseeuw's MCD estimator. Once all autocorrelation coefficients up to lag p are calculated the Yule-Walker equations are solved to obtain the AR(p) model parameters. However, the method proposed by Liu et al. (2004) unfortunately yields negative estimates of the innovations variance in 5 to 10% of the cases depending on the amount of contamination. The parameter estimates obtained by bounded-influence autoregression are labeled by 'arGM'.

The dotted horizontal lines, one in each plot, represent the true values of the model parameters.

It is clearly visible in Figure 5.1 that the deviation between all estimated model parameters and their true values becomes larger the higher the amount of contamination. As expected, the difference is largest for the classical estimates. However, it is also larger for the estimates obtained by the method proposed by Liu et al. (2004) than for those of the bounded-influence autoregression. Moreover, the method of Liu et al. (2004) that uses the MCD estimator to obtain robust autocorrelation coefficients yields parameter estimates with a larger variation than the other two methods.

Hence, for all contamination levels the bounded-influence autoregression yields parameter estimates that are closest to their true values.
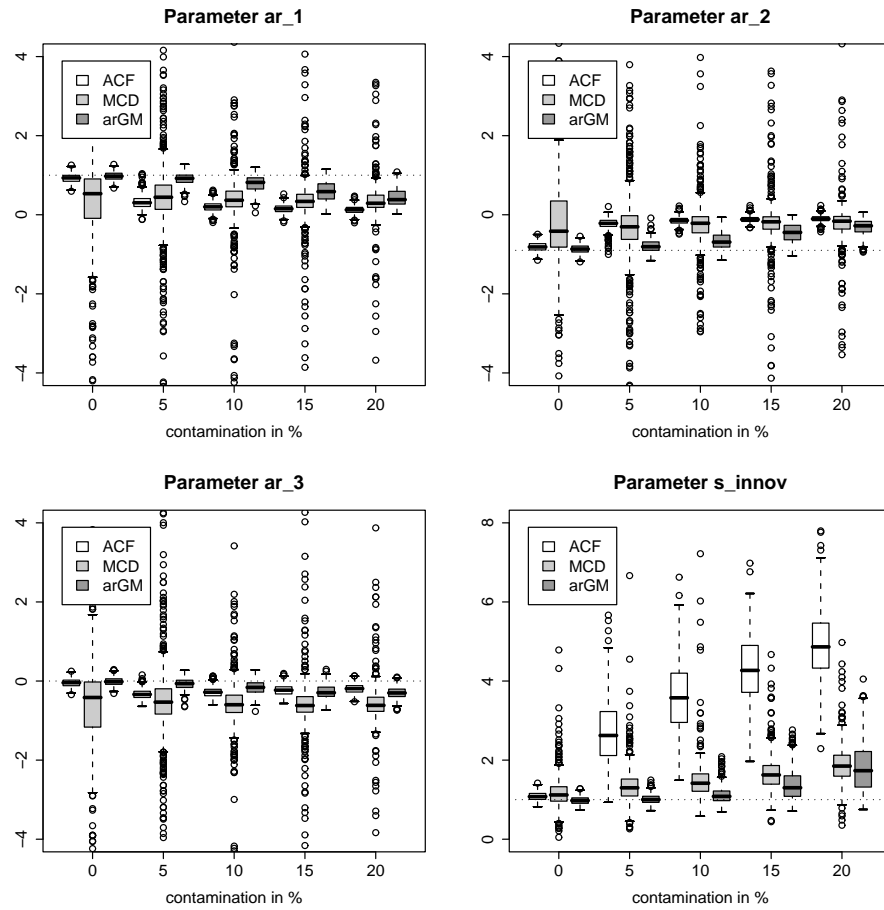
Figure 5.1: Boxplots of the autoregressive parameters and the innovations scale

5.5 Discussion

The results of our simulation experiments show that the bounded-influence autoregression performs better compared to the method proposed by Liu et al. (2004).

Further, there are two main disadvantages using the method of Liu et al. (2004): First, the variation of the parameter estimates obtained by this method is larger compared to the other two tested methods. Second, in some cases the method yields negative estimates of the innovations variance which cannot be used further on.

Moreover, the method proposed by Liu et al. (2004) yields estimates that are worse compared to those using bounded-influence autoregression. They suggest to estimate the autocorrelation coefficients separately, forming new bivariate datasets first, and then computing all corresponding 2×2 covariance matrices using the MCD estimator. Once all autocorrelation coefficients up to lag p are calculated the Yule-Walker equations are solved to obtain the AR(p) model parameters.

The proposed algorithm can be further improved by forming a new $p+1$-dimensional dataset $\{(y_t, \ldots, y_{t-p})^\top, t = p+1, \ldots, n\}$ and computing the $(p+1) \times (p+1)$ covariance matrix using the MCD estimator. The autocorrelation coefficients up to lag p can be calculated by averaging the diagonals of the upper (or lower) triangular matrix of the previously obtained covariance matrix using the trimmed mean or the median. Then again the Yule-Walker equations are solved to obtain the AR(p) model parameters. This modification yields better parameter estimates. However, the two above mentioned disadvantages still remain: The variation of the obtained parameters are large and there are still cases where we get negative estimates of the innovations variance.

Additional simulation studies trying other autoregressive models with different parameters have also been done yielding similar results and are therefore not published here.

Chapter 6

Robustifying the Kalman Filter

6.1   State-space Models

Let us assume we observe a q-dimensional, vector-valued process $\{\boldsymbol{y}_t,\ t = 1, \ldots, n\}$ which is only a linear transformation of an unobserved p-dimensional signal $\boldsymbol{x}_t$ with some noise added. Then the *state-space model* can be defined as follows:

$$
\begin{aligned}
\boldsymbol{x}_t &= \boldsymbol{\Phi}\boldsymbol{x}_{t-1} + \boldsymbol{\varepsilon}_t \\
\boldsymbol{y}_t &= \boldsymbol{H}\boldsymbol{x}_t + \boldsymbol{v}_t \ ,
\end{aligned}
\tag{6.1}
$$

where $\boldsymbol{x}_t$ is the unobserved p-dimensional vector called *state vector*. The first equation in (6.1) is named *state equation* and the second *observation equation*. It is assumed that $\boldsymbol{\varepsilon}_t$ has dimension p, $\boldsymbol{\Phi}$ is a $p \times p$ matrix and $\boldsymbol{H}$ is a $q \times p$ matrix. We further assume that $\boldsymbol{x}_t$ is independent of future values of $\boldsymbol{\varepsilon}_t$, and that $\boldsymbol{\varepsilon}_t$ and $\boldsymbol{v}_t$ are zero mean independent and identically distributed (iid) sequences which also are independent of each other but could be non-Gaussian. A more general definition of state-space models considering correlated errors as well as more complex models including exogenous variables or selection matrices can be found in Shumway and Stoffer (2000) or Durbin and Koopman (2001).

6.2   The Classical Kalman Filter

The primary aim of any analysis using state-space models as defined by (6.1) is to find estimators of the underlying unobserved signal $\boldsymbol{x}_t$, given the data

$\boldsymbol{Y}_s = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_s\}$, up to time s. If $s < t$, $s = t$ or $s > t$, the problem is called *prediction*, *filtering* or *smoothing*, respectively.

In addition, we want to get estimators T_t of $\boldsymbol{x}_t$ given $\boldsymbol{Y}_s$ which are best in the sense of the minimum mean-squared error, i.e.,

$$E(\|\boldsymbol{x}_t - T_t(\boldsymbol{Y}_s)\|^2) = \min_{T_t}! \ . \tag{6.2}$$

The solution is the conditional mean of $\boldsymbol{x}_t$ given $\boldsymbol{Y}_s$, i.e.,

$$T_t(\boldsymbol{Y}_s) = E(\boldsymbol{x}_t | \boldsymbol{Y}_s) \ , \tag{6.3}$$

and will further on be denoted by $\boldsymbol{x}_{t|s}$.

However, in general the conditional mean is hard to calculate and therefore we restrict ourselves to the class of linear estimators. Then the solution to these problems is accomplished via the *Kalman filter* and *smoother* (cf. Kalman, 1960; Kalman and Bucy, 1961). The estimators we obtain are the minimum mean-squared error estimators within the class of linear estimators.

In the following we will just focus on the Kalman filter. Its advantage is that it specifies how to update the filter values from $\boldsymbol{x}_{t-1|t-1}$ to $\boldsymbol{x}_{t|t}$ once a new observation $\boldsymbol{y}_t$ is obtained, without having to reprocess the entire data set $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_t$. The Kalman filter recursions can be split into three steps:

(i) Initialization ($t = 0$):

$$\boldsymbol{x}_{0|0} = \boldsymbol{\mu}_0 \ , \quad \boldsymbol{P}_0 = \boldsymbol{\Sigma}_0 \tag{6.4}$$

where $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ are the unconditional mean and $p \times p$ covariance matrix of $\boldsymbol{x}_0$.

(ii) Prediction ($t \geq 1$):

$$\begin{aligned} \boldsymbol{x}_{t|t-1} &= \boldsymbol{\Phi}\boldsymbol{x}_{t-1|t-1} \\ \boldsymbol{M}_t &= \boldsymbol{\Phi}\boldsymbol{P}_{t-1}\boldsymbol{\Phi}^\top + \boldsymbol{Q} \end{aligned} \tag{6.5}$$

(iii) Correction ($t \geq 1$):

$$\begin{aligned} \boldsymbol{x}_{t|t} &= \boldsymbol{x}_{t|t-1} + \boldsymbol{K}_t(\boldsymbol{y}_t - \boldsymbol{H}\boldsymbol{x}_{t|t-1}) \\ \boldsymbol{P}_t &= \boldsymbol{M}_t - \boldsymbol{K}_t\boldsymbol{H}\boldsymbol{M}_t \\ \text{with} \quad \boldsymbol{K}_t &= \boldsymbol{M}_t\boldsymbol{H}^\top(\boldsymbol{H}\boldsymbol{M}_t\boldsymbol{H}^\top + \boldsymbol{R})^{-1} \end{aligned} \tag{6.6}$$

The $p \times q$ matrix $\boldsymbol{K}_t$ is called the *Kalman gain*. The $p \times p$ matrix $\boldsymbol{M}_t$ is the conditional prediction error covariance matrix,

$$\boldsymbol{M}_t = E\big((\boldsymbol{x}_t - \boldsymbol{x}_{t|t-1})(\boldsymbol{x}_t - \boldsymbol{x}_{t|t-1})^\top | \boldsymbol{Y}_{t-1}\big) \ , \tag{6.7}$$

and the conditional filtering error covariance matrix $\boldsymbol{P}_t$ is given by

$$\boldsymbol{P}_t = E\big((\boldsymbol{x}_t - \boldsymbol{x}_{t|t})(\boldsymbol{x}_t - \boldsymbol{x}_{t|t})^\top | \boldsymbol{Y}_t\big) \ . \tag{6.8}$$

The $p \times p$ matrix $\boldsymbol{Q}$ and the $q \times q$ matrix $\boldsymbol{R}$ denote the covariance matrices of $\boldsymbol{\varepsilon}_t$ and $\boldsymbol{v}_t$, respectively.

6.3 Approximate Conditional-mean Type Filtering

The robust filter described in this section is an approximate conditional-mean (ACM) type filter motivated by Masreliez's result (Masreliez, 1975).

6.3.1 Masreliez's Theorem

We will use the following notation: $\boldsymbol{Y}_t = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_t\}$ denote the first t observations. The conditional mean of $\boldsymbol{x}_t$ given $\boldsymbol{Y}_t$ is written as $\boldsymbol{x}_{t|t} = E(\boldsymbol{x}_t|\boldsymbol{Y}_t)$ whereas the conditional mean of $\boldsymbol{x}_t$ given $\boldsymbol{Y}_{t-1}$ is denoted by $\boldsymbol{x}_{t|t-1} = E(\boldsymbol{x}_t|\boldsymbol{Y}_{t-1})$. The estimates $\widehat{\boldsymbol{x}}_{t|t}$ and $\widehat{\boldsymbol{x}}_{t|t-1}$ are called *filter estimate* and *one-step-ahead prediction*, respectively. It is easily verified that $\boldsymbol{x}_{t|t-1} = \boldsymbol{\Phi}\boldsymbol{x}_{t-1|t-1}$ under the model assumptions of (6.1). The state prediction density, i.e., the density of $\boldsymbol{x}_t$ conditioned on prior observations $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_{t-1}$, is denoted by $f_{\boldsymbol{x}_t}(.|\boldsymbol{Y}_{t-1})$. Similarly, $f_{\boldsymbol{y}_t}(.|\boldsymbol{Y}_{t-1})$ is the observation prediction density conditioned on the past observations. In the following we will use the symbol "$\simeq$" to associate a density function with its related distribution.

If $\boldsymbol{\varepsilon}_t$ and $\boldsymbol{v}_t$ are Gaussian a straightforward calculation of $\boldsymbol{x}_{t|t} = E(\boldsymbol{x}_t|\boldsymbol{Y}_t)$ yields the Kalman filter recursion equations (see, for example, Jazwinski, 1970). For non-Gaussian $\boldsymbol{v}_t$ the calculation of the exact $\boldsymbol{x}_{t|t}$ is difficult. However, Masreliez (1975) made the simplifying assumption that the state prediction density $f_{\boldsymbol{x}_t}(.|\boldsymbol{Y}_{t-1})$ is Gaussian, i.e.,

$$\boldsymbol{x}_t | \boldsymbol{Y}_{t-1} \sim \mathcal{N}_p(\boldsymbol{x}_{t|t-1}, \boldsymbol{M}_t) \ , \tag{6.9}$$

where $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The covariance matrix $\boldsymbol{M}_t$ in (6.9) is the conditional error covariance matrix for predicting $\boldsymbol{x}_t$

$$\boldsymbol{M}_t = E\big((\boldsymbol{x}_t - \boldsymbol{x}_{t|t-1})(\boldsymbol{x}_t - \boldsymbol{x}_{t|t-1})^\top | \boldsymbol{Y}_{t-1}\big) \ . \tag{6.10}$$

For the definition of the ACM filter we also need the conditional filtering error covariance matrix

$$\boldsymbol{P}_t = E\big((\boldsymbol{x}_t - \boldsymbol{x}_{t|t})(\boldsymbol{x}_t - \boldsymbol{x}_{t|t})^\top | \boldsymbol{Y}_t\big) \ . \tag{6.11}$$

We note that in the pure Gaussian situation $\boldsymbol{x}_{t|t}$, $\boldsymbol{M}_t$ and $\boldsymbol{P}_t$, $t = 1, \ldots, n$, are obtained by the Kalman filter recursions and $\boldsymbol{M}_t$ and $\boldsymbol{P}_t$ do not depend upon $\boldsymbol{Y}_{t-1}$ and $\boldsymbol{Y}_t$ respectively, which is a rather special feature of the Gaussian case. One should not expect $\boldsymbol{M}_t$ and $\boldsymbol{P}_t$ to be independent of the data in general, and in fact it turns out that both, $\boldsymbol{M}_t$ and $\boldsymbol{P}_t$, depend upon the data in an intuitively appealing manner in Masreliez's result, which we now state.

For the following ACM filter theorem we assume that the observations are generated by (6.1) and that $\boldsymbol{\Phi}$, the covariance matrix $\boldsymbol{Q}$ of the $\boldsymbol{\varepsilon}_t$ and the density $f_{\boldsymbol{v}_t}$ of the $\boldsymbol{v}_t$ are known in advance.

Theorem (Masreliez, 1975). If $\boldsymbol{x}_t | \boldsymbol{Y}_{t-1} \sim \mathcal{N}_p(\boldsymbol{x}_{t|t-1}, \boldsymbol{M}_t)$, $t \geq 1$, then $\boldsymbol{x}_{t|t} = E(\boldsymbol{x}_t | \boldsymbol{Y}_t)$, $t \geq 1$, is generated by the recursions

$$\begin{align} \boldsymbol{x}_{t|t} &= \boldsymbol{x}_{t|t-1} + \boldsymbol{M}_t \boldsymbol{H}^\top \Psi_t(\boldsymbol{y}_t) \tag{6.12}\\ \boldsymbol{P}_t &= \boldsymbol{M}_t - \boldsymbol{M}_t \boldsymbol{H}^\top \Psi_t'(\boldsymbol{y}_t) \boldsymbol{H} \boldsymbol{M}_t \tag{6.13}\\ \boldsymbol{M}_{t+1} &= \boldsymbol{\Phi} \boldsymbol{P}_t \boldsymbol{\Phi}^\top + \boldsymbol{Q} \ , \tag{6.14} \end{align}$$

where $\Psi_t(\boldsymbol{y}_t)$ is a q-dimensional vector with components

$$(\Psi_t(\boldsymbol{y}))_i = -(\partial/\partial y_i) \log f_{\boldsymbol{y}_t}(\boldsymbol{y} | \boldsymbol{Y}_{t-1}) \tag{6.15}$$

and is usually called the *score function* for the observation prediction density $f_{\boldsymbol{y}_t}(. | \boldsymbol{Y}_{t-1})$, and $\Psi_t'(\boldsymbol{y}_t)$ is a $q \times q$ matrix with elements

$$(\Psi_t'(\boldsymbol{y}))_{ij} = (\partial/\partial y_j)(\Psi_t(\boldsymbol{y}))_i \ . \tag{6.16}$$

If $f_{\boldsymbol{y}_t}(. | \boldsymbol{Y}_{t-1})$ is Gaussian, it is easy to verify that Masreliez's filter reduces to the Kalman filter. In this case we have

$$\begin{align} \Psi_t(.) &= (\boldsymbol{H} \boldsymbol{M}_t \boldsymbol{H}^\top + \boldsymbol{R})^{-1}(. - \boldsymbol{H} \boldsymbol{x}_{t|t-1}) \tag{6.17}\\ \Psi_t'(.) &= (\boldsymbol{H} \boldsymbol{M}_t \boldsymbol{H}^\top + \boldsymbol{R})^{-1} \ . \tag{6.18} \end{align}$$

Although Masreliez (1975) did not specify initial conditions for the above recursions, appropriate $\boldsymbol{x}_{0|0}$ and $\boldsymbol{M}_1$ may be set to $\boldsymbol{x}_{0|0} = E(\boldsymbol{x}_0) = \boldsymbol{\mu}_0$ and $\boldsymbol{M}_1 = E(\boldsymbol{x}_1 \boldsymbol{x}_1^\top) = \boldsymbol{\Sigma}_0$, i.e., the unconditional mean and covariance of $\boldsymbol{x}_1$ (cf. Martin, 1981). However, in order to agree with the definition of the classical Kalman filter recursions we will specify the initial conditions for the above recursions by setting $\boldsymbol{x}_{0|0} = E(\boldsymbol{x}_0) = \boldsymbol{\mu}_0$ and $\boldsymbol{P}_0 = E(\boldsymbol{x}_0 \boldsymbol{x}_0^\top) = \boldsymbol{\Sigma}_0$ (see, for example, Shumway and Stoffer, 2000; Durbin and Koopman, 2001). This will lead to slightly different results only for the first few $\boldsymbol{M}_t$'s and $\boldsymbol{P}_t$'s.

6.3.2 Masreliez's Filter for Autoregressive Models

Any zero mean p-th order autoregressive process

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t \tag{6.19}$$

which is free of additive outliers can be written in state-space form (6.1) by setting

$$\boldsymbol{\Phi} = \begin{pmatrix} \phi_1 & \cdots & \phi_{p-1} & \phi_p \\ 1 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 \end{pmatrix}, \tag{6.20}$$

$$\boldsymbol{\varepsilon}_t = (\varepsilon_t, 0, \ldots, 0)^\top \text{ and} \tag{6.21}$$

$$\boldsymbol{H} = (1, 0, \ldots, 0). \tag{6.22}$$

We note that the process y_t of the observation equation is not vector-valued any more but univariate.

For the state prediction density $f_{\boldsymbol{x}_t}(.|\boldsymbol{Y}_{t-1})$ we still assume that $\boldsymbol{x}_t|\boldsymbol{Y}_{t-1} \sim \mathcal{N}_p(\boldsymbol{x}_{t|t-1}, \boldsymbol{M}_t)$. Then the observation prediction density $f_{y_t}(.|\boldsymbol{Y}_{t-1})$ in Masreliez's theorem is obtained by convolving the prediction density $f_{z_t}(.|\boldsymbol{Y}_{t-1})$ of $z_t|\boldsymbol{Y}_{t-1} \sim \mathcal{N}(\boldsymbol{H}\boldsymbol{x}_{t|t-1}, \boldsymbol{H}\boldsymbol{M}_t\boldsymbol{H}^\top)$ for $z_t = \boldsymbol{H}\boldsymbol{x}_t$ with the observation noise density f_{v_t}. For autoregressive models $\boldsymbol{H} = (1, 0, \ldots, 0)$, and therefore $f_{z_t}(.|\boldsymbol{Y}_{t-1})$ is just the marginal state prediction density $f_{\boldsymbol{x}_t}(.|\boldsymbol{Y}_{t-1})$ for the first component $x_t = (\boldsymbol{x}_t)_1$ of $\boldsymbol{x}_t$, which is

$$f_{x_t}(.|\boldsymbol{Y}_{t-1}) \simeq \mathcal{N}(\xi_t, m_{11,t}), \tag{6.23}$$

where

$$\xi_t = (\boldsymbol{x}_{t|t-1})_1 = (\boldsymbol{\Phi}\boldsymbol{x}_{t-1|t-1})_1 = \sum_{j=1}^p \phi_j (\boldsymbol{x}_{t-1|t-1})_j \tag{6.24}$$

53

is the first component of $\boldsymbol{x}_{t|t-1}$, $(\boldsymbol{x}_{t-1|t-1})_j$ is the j-th component of $\boldsymbol{x}_{t-1|t-1}$, $1 \le j \le p$, and $m_{11,t} = (\boldsymbol{M}_t)_{11}$ is the 1-1 element of $\boldsymbol{M}_t$.

We now suppose that the v_t's have a contaminated normal distribution

$$\mathcal{CN}(\gamma, \sigma_0^2, \sigma^2) = (1 - \gamma)\mathcal{N}(0, \sigma_0^2) + \gamma\mathcal{N}(0, \sigma^2) , \tag{6.25}$$

with γ, the amount of contamination, not too large. Then, since $\eta_t = E(y_t|\boldsymbol{Y}_{t-1}) = E(x_t|\boldsymbol{Y}_{t-1}) = \xi_t$, convolution of f_{x_t} with $f_{v_t} \simeq \mathcal{CN}(\gamma, \sigma_0^2, \sigma^2)$ gives

$$f_{y_t}(.|\boldsymbol{Y}_{t-1}) \simeq (1 - \gamma)\mathcal{N}(\eta_t, \sigma_{0t}^2) + \gamma\mathcal{N}(\eta_t, \sigma_t^2) , \tag{6.26}$$

with $\sigma_{0t}^2 = m_{11,t} + \sigma_0^2$ and $\sigma_t^2 = m_{11,t} + \sigma^2$. The observation prediction density f_{y_t} can also be written as

$$f_{y_t}(.|\boldsymbol{Y}_{t-1}) = g_t(. - \eta_t) , \tag{6.27}$$

where the density g_t is obtained by convolution:

$$g_t \simeq \mathcal{N}(0, m_{11,t}) * \mathcal{F}_{v_t} , \tag{6.28}$$

where $\mathcal{F}_{v_t}$ denotes the distribution of the observation noise v_t. To ease notation we will further on set $\Delta y_t = y_t - \eta_t$.

Thus, noting that $\boldsymbol{H} = (1, 0, \ldots, 0)$ and denoting the first column of $\boldsymbol{M}_t$ by $\boldsymbol{m}_{.1,t} = (\boldsymbol{M}_t)_{.1}$, the versions of (6.12)–(6.14) for autoregressive models of order p are

$$\boldsymbol{x}_{t|t} = \boldsymbol{\Phi}\boldsymbol{x}_{t-1|t-1} + \boldsymbol{m}_{.1,t}\Psi_t(\Delta y_t) \tag{6.29}$$

$$\boldsymbol{P}_t = \boldsymbol{M}_t - \Psi_t'(\Delta y_t)\boldsymbol{m}_{.1,t}\boldsymbol{m}_{.1,t}^\top \tag{6.30}$$

$$\boldsymbol{M}_{t+1} = \boldsymbol{\Phi}\boldsymbol{P}_t\boldsymbol{\Phi}^\top + \boldsymbol{Q} , \tag{6.31}$$

where the $p \times p$ covariance matrix

$$\boldsymbol{Q} = \begin{pmatrix} \sigma_\varepsilon^2 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix} \tag{6.32}$$

of $\boldsymbol{\varepsilon}_t$ is singular and where $\Psi_t(\Delta y_t) = -(\partial/\partial\Delta y_t)\log g_t(\Delta y_t)$ with its first derivative Ψ_t' .
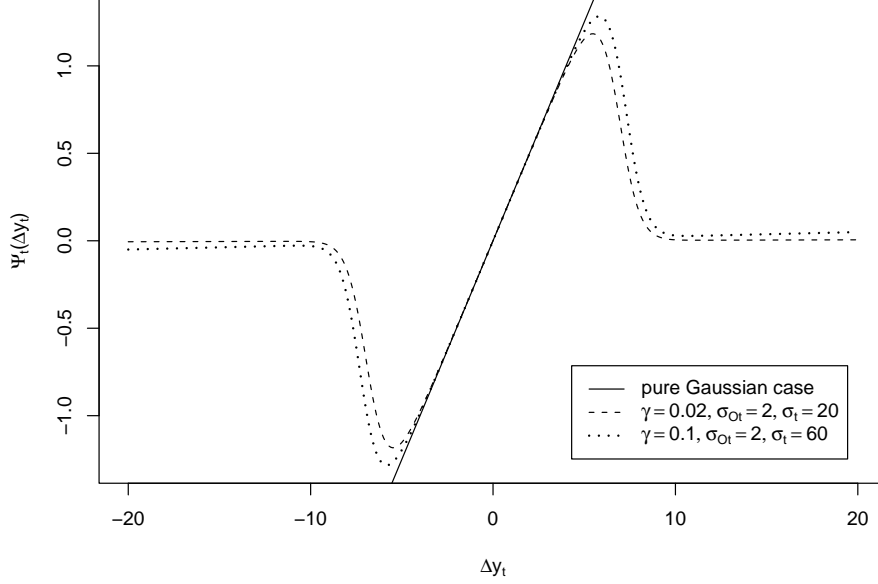
Figure 6.1: Shapes of scalar-valued score functions

It may be noted that both $\Psi_t(\Delta y_t)$ and $\Psi_t'(\Delta y_t)$ are scalar-valued. The shapes of $\Psi_t(\Delta y_t)$ in the Gaussian situation and for two combinations of values of γ, σ_{0t} and σ_t are shown in Figure 6.1. If $\mathcal{F}_{v_t}$ is Gaussian, i.e., $\gamma = 0$, the score function is linear with slope $1/\sigma_{0t}^2$. For heavy-tailed g_t the shape of Ψ_t is such that large prediction residuals are downweighted and $\boldsymbol{P}_t \cong \boldsymbol{M}_t$. More details may be found in Martin (1979).

We can proceed one step further and represent g_t in the form

$$g_t(\Delta y_t) = \frac{1}{s_t} g(\frac{\Delta y_t}{s_t}) \; , \tag{6.33}$$

where

$$g \simeq \mathcal{N}(0, a) * \mathcal{F}_{v_t, b} \; , \tag{6.34}$$

with

$$\mathcal{F}_{v_t, b}(r) = \mathcal{F}_{v_t}(\frac{r}{b}) \; , \tag{6.35}$$

55

and s_t, a and b are appropriately specified. Equation (6.33) is valid if $\mathcal{F}_{v_t}$ is Gaussian or a mixture of Gaussian distributions with zero mean and arbitrary variances. Unfortunately, this is generally not true if $\mathcal{F}_{v_t}$ is non-Gaussian. However, in the case of $\mathcal{F}_{v_t}$ being a contaminated normal distribution, it is correct to set

$$s_t^2 = \sigma_{0t}^2 = m_{11,t} + \sigma_0^2 \ , \tag{6.36}$$

$$a = m_{11,t}/s_t^2 \ , \quad \text{and} \quad b = 1/s_t \ , \tag{6.37}$$

and use formula (6.33). If $\gamma = 0$, i.e., in the pure Gaussian situation, the idea is simply to standardize Δy_t by s_t. In the case of γ not too large we may still use the same Gaussian distribution as an approximation. Actually the above approximation should work reasonably well for any heavy-tailed distribution $\mathcal{F}_{v_t}$ which is nearly Gaussian in the middle.

The same considerations used above now give some kind of time-invariant version of the score function Ψ_t:

$$\Psi_t(\Delta y_t) \cong \frac{1}{s_t}\Psi\left(\frac{\Delta y_t}{s_t}\right) \ , \tag{6.38}$$

and

$$\Psi_t'(\Delta y_t) \cong \frac{1}{s_t^2}\Psi'\left(\frac{\Delta y_t}{s_t}\right) \ , \tag{6.39}$$

where

$$\Psi(r) = -(\partial/\partial r)\log g(r) \ . \tag{6.40}$$

This yields simplified versions of (6.29)–(6.31):

$$\boldsymbol{x}_{t|t} = \boldsymbol{\Phi}\boldsymbol{x}_{t-1|t-1} + \boldsymbol{m}_{.1,t}\frac{1}{s_t}\Psi\left(\frac{\Delta y_t}{s_t}\right) \tag{6.41}$$

$$\boldsymbol{P}_t = \boldsymbol{M}_t - \frac{1}{s_t^2}\Psi'\left(\frac{\Delta y_t}{s_t}\right)\boldsymbol{m}_{.1,t}\boldsymbol{m}_{.1,t}^\top \tag{6.42}$$

$$\boldsymbol{M}_{t+1} = \boldsymbol{\Phi}\boldsymbol{P}_t\boldsymbol{\Phi}^\top + \boldsymbol{Q} \ , \tag{6.43}$$

where the definition of $\boldsymbol{Q}$ remains the same.

6.3.3 ACM-type Filters for Autoregressive Models

Since $\mathcal{F}_{v_t}$ will rarely be known in practice, we follow the usual M-estimate route by replacing the score function Ψ by a good robustifying psi-function ψ. Thus, the preceding observations suggest how to define an approximate conditional-mean (ACM) type filter for autoregressive models of order p. Let ψ be any bounded and continuous function chosen with robust considerations in mind, and let

$$s_t^2 = m_{11,t} + \sigma_0^2 \ , \tag{6.44}$$

where σ_0^2 is the variance of the uncontaminated Gaussian distribution of the observation noise v_t. Then, noting that $\Delta y_t = y_t - \sum_{j=1}^p \phi_j(\boldsymbol{x}_{t-1|t-1})_j$, the recursions (6.41)–(6.43) are

$$\boldsymbol{x}_{t|t} = \boldsymbol{\Phi}\boldsymbol{x}_{t-1|t-1} + \frac{\boldsymbol{m}_{.1,t}}{s_t^2}s_t\psi\left(\frac{y_t - \sum_{j=1}^p \phi_j(\boldsymbol{x}_{t-1|t-1})_j}{s_t}\right) \tag{6.45}$$

$$\boldsymbol{P}_t = \boldsymbol{M}_t - \psi'\left(\frac{y_t - \sum_{j=1}^p \phi_j(\boldsymbol{x}_{t-1|t-1})_j}{s_t}\right)\frac{\boldsymbol{m}_{.1,t}\boldsymbol{m}_{.1,t}^\top}{s_t^2} \tag{6.46}$$

$$\boldsymbol{M}_{t+1} = \boldsymbol{\Phi}\boldsymbol{P}_t\boldsymbol{\Phi}^\top + \boldsymbol{Q} \ . \tag{6.47}$$

The definition of $\boldsymbol{Q}$ remains the same and ψ' again denotes the first derivative of ψ.

Martin (1979) proposed to use a special form of *Hampel's three-part re-descending psi-function* (Hampel, 1968),

$$\psi_{HA}(s) = \begin{cases} s & \text{if} \quad |s| \le a \\ a\operatorname{sgn}(s) & \text{if} \quad a < |s| \le b \\ \frac{a}{b-c}(s - c\operatorname{sgn}(s)) & \text{if} \quad b < |s| \le c \\ 0 & \text{if} \quad c < |s| \ , \end{cases} \tag{6.48}$$

for the above ψ, namely, *Hampel's two-part redescending psi-function*, where $b = a$.

If we do not omit the constant part that is unequal to zero in Hampel's psi-function or use Huber's monotone psi-function for ψ instead, it will result in $\psi(s) \ne 0$ but $\psi'(s) = 0$, if $a < |s| < b$, so that $\boldsymbol{P}_t = \boldsymbol{M}_t$, while at the same time $\boldsymbol{x}_{t|t}$ is not equal to $\boldsymbol{\Phi}\boldsymbol{x}_{t-1|t-1}$ as one might expect if $\boldsymbol{P}_t = \boldsymbol{M}_t$. On the other hand the ACM filter based on Hampel's two-part redescending psi-function has the appealing feature that $\boldsymbol{x}_{t|t} = \boldsymbol{\Phi}\boldsymbol{x}_{t-1|t-1}$ and $\boldsymbol{P}_t = \boldsymbol{M}_t$ by virtue of $\psi(s) = \psi'(s) = 0$ if $|s| > c$. This appears to be the natural embodiment of an outlier-rejection rule in the filtering context.

6.3.4 One-sided Outlier Interpolation

An important use of ACM filters is for time series situations where we assume that the uncontaminated Gaussian distribution of v_t has variance $\sigma_0^2 = 0$, i.e., in the case of additive outliers. Now $\widehat{x}_{t|t} = (\widehat{\boldsymbol{x}}_{t|t})_1 = (\widehat{E}(\boldsymbol{x}_t|\boldsymbol{Y}_t))_1 = \widehat{E}(x_t|\boldsymbol{Y}_t)$ is the filtered value at time t based on the vector estimate $\widehat{\boldsymbol{x}}_{t|t}$, and $\widehat{y}_{t|t-1} = \sum_{j=1}^{p} \phi_j (\widehat{\boldsymbol{x}}_{t-1|t-1})_j$ is the prediction of y_t based on $\boldsymbol{Y}_{t-1}$. Suppose ψ is Hampel's two-part redescending psi-function. If the prediction residual $\widehat{\Delta y_t} = y_t - \widehat{y}_{t|t-1}$ satisfies $|\widehat{\Delta y_t}|/s_t > c$ then $\widehat{x}_{t|t}$ is just the predicted value $\widehat{x}_{t|t} = \widehat{y}_{t|t-1}$, and if $|\widehat{\Delta y_t}|/s_t < a$ then $\widehat{x}_{t|t} = \widehat{y}_{t|t-1} + (y_t - \widehat{y}_{t|t-1}) = y_t$. When a time series contains only a rather small fraction of outliers, and the constants a and c are appropriately adjusted, we find that $\widehat{x}_{t|t} = y_t$ a large fraction of time and $\widehat{x}_{t|t} \cong \widehat{y}_{t|t-1}$ a small fraction of time. Therefore Martin (1979) refers to an ACM-type filter operating under such conditions as a *one-sided interpolator*.

When considering autoregressive models of order $p \geq 2$ there is the question concerning which coordinate of $\widehat{\boldsymbol{x}}_{t|t}$ should be used as filter output. For the general ACM filter context as well as for the one-sided interpolator the first coordinate $\widehat{x}_{t|t} = (\widehat{\boldsymbol{x}}_{t|t})_1$ of $\widehat{\boldsymbol{x}}_{t|t}$ gives an estimate of the conditional mean $x_{t|t} = E(x_t|\boldsymbol{Y}_t)$ for filtering. However we could also make the choice $\widehat{x}_{t-p+1|t} = (\widehat{\boldsymbol{x}}_{t|t})_p$, the last coordinate of $\widehat{\boldsymbol{x}}_{t|t}$. Since $(\widehat{\boldsymbol{x}}_{t|t})_p = \widehat{E}(x_{t-p+1}|\boldsymbol{Y}_t)$ this results in a "fixed-lag" smoother of lag $p-1$.

6.4 Approximate Conditional-mean Type Filtering for Vector-valued Observations

6.4.1 Masreliez's Filter for State-space Models

Proceeding from Masreliez's Theorem (cf. Section 6.3.1) we still assume for the state prediction density $f_{\boldsymbol{x}_t}(.|\boldsymbol{Y}_{t-1})$ that $\boldsymbol{x}_t|\boldsymbol{Y}_{t-1} \sim \mathcal{N}_p(\boldsymbol{x}_{t|t-1}, \boldsymbol{M}_t)$. Then the observation prediction density $f_{\boldsymbol{y}_t}(.|\boldsymbol{Y}_{t-1})$ in Masreliez's theorem is obtained by convolving the prediction density $f_{\boldsymbol{z}_t}(.|\boldsymbol{Y}_{t-1})$ of $\boldsymbol{z}_t|\boldsymbol{Y}_{t-1} \sim \mathcal{N}_q(\boldsymbol{H}\boldsymbol{x}_{t|t-1}, \boldsymbol{H}\boldsymbol{M}_t\boldsymbol{H}^\top)$ for $\boldsymbol{z}_t = \boldsymbol{H}\boldsymbol{x}_t$ with the observation noise density $f_{\boldsymbol{v}_t}$.

Now, if we further suppose that the $\boldsymbol{v}_t$'s have a contaminated multivariate normal distribution

$$\mathcal{CN}_q(\gamma, \boldsymbol{R}, \boldsymbol{R}_c) = (1-\gamma)\mathcal{N}_q(\boldsymbol{0}, \boldsymbol{R}) + \gamma\mathcal{N}_q(\boldsymbol{0}, \boldsymbol{R}_c) , \tag{6.49}$$

with γ, the amount of contamination, again not too large, convolution of $f_{\boldsymbol{z}_t}$ with $f_{\boldsymbol{v}_t} \simeq \mathcal{CN}_q(\gamma, \boldsymbol{R}, \boldsymbol{R}_c)$ gives

$$f_{\boldsymbol{y}_t}(.|\boldsymbol{Y}_{t-1}) \simeq (1-\gamma)\mathcal{N}_q(\boldsymbol{H}\boldsymbol{x}_{t|t-1}, \boldsymbol{R}_t) + \gamma\mathcal{N}_q(\boldsymbol{H}\boldsymbol{x}_{t|t-1}, \boldsymbol{R}_{c,t}) \ , \qquad (6.50)$$

with $\boldsymbol{R}_t = \boldsymbol{H}\boldsymbol{M}_t\boldsymbol{H}^\top + \boldsymbol{R}$ and $\boldsymbol{R}_{c,t} = \boldsymbol{H}\boldsymbol{M}_t\boldsymbol{H}^\top + \boldsymbol{R}_c$. Then we can also write the observation prediction density $f_{\boldsymbol{y}_t}$ as

$$f_{\boldsymbol{y}_t}(.|\boldsymbol{Y}_{t-1}) = g_t(. - \boldsymbol{H}\boldsymbol{x}_{t|t-1}) \ , \qquad (6.51)$$

where the density g_t is obtained by convolution, i.e.,

$$g_t \simeq \mathcal{N}_q(\boldsymbol{0}, \boldsymbol{H}\boldsymbol{M}_t\boldsymbol{H}^\top) * \mathcal{F}_{\boldsymbol{v}_t} \ , \qquad (6.52)$$

and $\mathcal{F}_{\boldsymbol{v}_t}$ denotes the distribution of the observation noise $\boldsymbol{v}_t$. To ease notation we will further on set $\Delta\boldsymbol{y}_t = \boldsymbol{y}_t - \boldsymbol{H}\boldsymbol{x}_{t|t-1}$.

Thus, Masreliez's filter recursions (6.12)–(6.14) become

$$\boldsymbol{x}_{t|t} = \boldsymbol{x}_{t|t-1} + \boldsymbol{M}_t\boldsymbol{H}^\top\Psi_t(\Delta\boldsymbol{y}_t) \qquad (6.53)$$
$$\boldsymbol{P}_t = \boldsymbol{M}_t - \boldsymbol{M}_t\boldsymbol{H}^\top\Psi'_t(\Delta\boldsymbol{y}_t)\boldsymbol{H}\boldsymbol{M}_t \qquad (6.54)$$
$$\boldsymbol{M}_{t+1} = \boldsymbol{\Phi}\boldsymbol{P}_t\boldsymbol{\Phi}^\top + \boldsymbol{Q} \ , \qquad (6.55)$$

where $\Psi_t(\Delta\boldsymbol{y}_t)$ is a q-dimensional vector with components

$$(\Psi_t(\boldsymbol{y}))_i = -(\partial/\partial y_i)\log g_t(\boldsymbol{y}) \qquad (6.56)$$

and $\Psi'_t(\Delta\boldsymbol{y}_t)$ is a $q \times q$ matrix with elements

$$(\Psi'_t(\boldsymbol{y}))_{ij} = (\partial/\partial y_j)(\Psi_t(\boldsymbol{y}))_i \ . \qquad (6.57)$$

In Figures 6.2 and 6.3 the situations for two combinations of values of γ, $\boldsymbol{R}_t$ and $\boldsymbol{R}_{c,t}$ are shown in the two-dimensional case, i.e., for $q = 2$. We only vary $\boldsymbol{R}_t$ in Figures 6.2 and 6.3 and set

$$\gamma = 0.1 \ , \quad \boldsymbol{R}_{c,t} = \begin{pmatrix} 100 & 0 \\ 0 & 100 \end{pmatrix} \ , \text{ and}$$

$$\boldsymbol{R}_t = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \text{ or } \boldsymbol{R}_t = \begin{pmatrix} 2 & -0.2 \\ -0.2 & 2 \end{pmatrix} \ . \qquad (6.58)$$

In plot (a) the graph of the contaminated bivariate normal density function is displayed. Plot (b) shows the logarithm of the contaminated bivariate
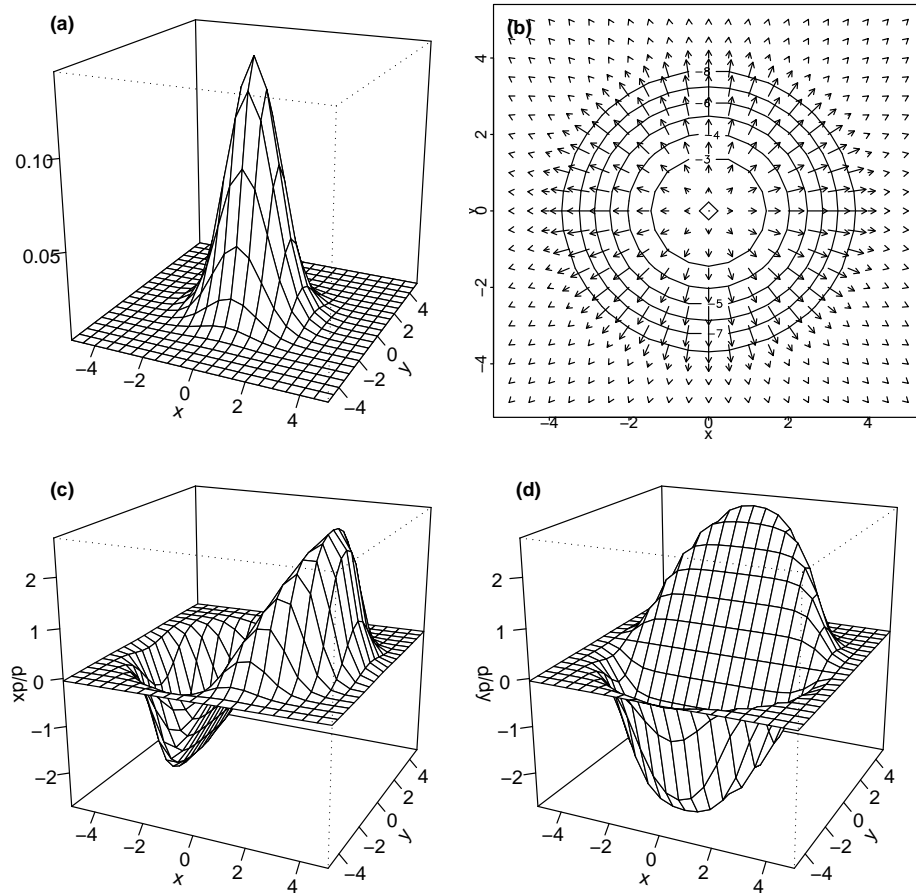
Figure 6.2: Score function of contaminated bivariate normal distribution (I)
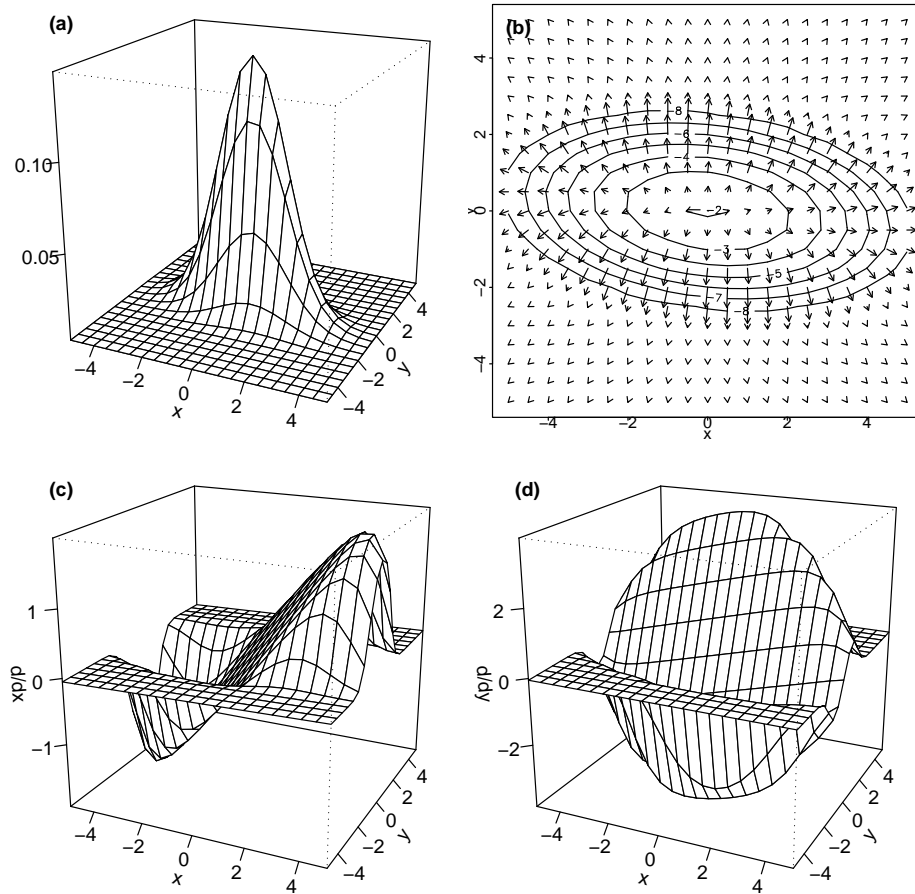
Figure 6.3: Score function of contaminated bivariate normal distribution (II)

normal density function as contour plot, where the ellipses connect points of the same value, together with the corresponding score function, which is in the case of $q = 2$ a two-dimensional vector field of gradients. The graphs of the first and second coordinate of the corresponding score function are shown in plots (c) and (d), respectively.

As in the univariate case we can proceed one step further and represent g_t in the form

$$g_t(\Delta \boldsymbol{y}_t) = |\boldsymbol{S}_t| g(\boldsymbol{S}_t \Delta \boldsymbol{y}_t) \ , \tag{6.59}$$

where

$$g \simeq \mathcal{N}_q(\boldsymbol{0}, \boldsymbol{A}) * \mathcal{F}_{\boldsymbol{v}_t, \boldsymbol{B}} \ , \tag{6.60}$$

with

$$\mathcal{F}_{\boldsymbol{v}_t, \boldsymbol{B}}(\boldsymbol{r}) = \mathcal{F}_{\boldsymbol{v}_t}(\boldsymbol{B}\boldsymbol{r}) \ , \tag{6.61}$$

and $\boldsymbol{S}_t$, $\boldsymbol{A}$ and $\boldsymbol{B}$ are appropriately specified. It is easily proven that Equation (6.59) is valid if $\mathcal{F}_{\boldsymbol{v}_t}$ is Gaussian or a mixture of Gaussian distributions with mean vector $\boldsymbol{0}$ and arbitrary covariance matrices. As in the univariate case, this is generally not true if $\mathcal{F}_{\boldsymbol{v}_t}$ is non-Gaussian. However, in the case of $\mathcal{F}_{\boldsymbol{v}_t}$ being a contaminated multivariate normal distribution, it is correct to set

$$\boldsymbol{S}_t = \boldsymbol{V} \boldsymbol{D}^{-1/2} \boldsymbol{V}^\top \tag{6.62}$$

with $\boldsymbol{R}_t = \boldsymbol{V} \boldsymbol{D} \boldsymbol{V}^\top = \boldsymbol{H} \boldsymbol{M}_t \boldsymbol{H}^\top + \boldsymbol{R}$, and further,

$$\boldsymbol{A} = \boldsymbol{S}_t \boldsymbol{H} \boldsymbol{M}_t \boldsymbol{H}^\top \boldsymbol{S}_t \ , \quad \text{and} \quad \boldsymbol{B} = \boldsymbol{S}_t^{-1} \ , \tag{6.63}$$

and use formula (6.59). The matrix $\boldsymbol{D}$ denotes the diagonal matrix containing the eigenvalues of $\boldsymbol{R}_t$ and $\boldsymbol{V}$ is the matrix of the corresponding eigenvectors.

If $\gamma = 0$, i.e., in the pure Gaussian situation, the idea is simply to standardize $\Delta \boldsymbol{y}_t$ by $\boldsymbol{S}_t$. In the case of γ not too large we may still approximate g_t by a multivariate Gaussian distribution with mean vector $\boldsymbol{0}$ and covariance matrix $\boldsymbol{R}_t = \boldsymbol{H} \boldsymbol{M}_t \boldsymbol{H}^\top + \boldsymbol{R}$ as before. Hence, the above approximation should work reasonably well for any heavy-tailed distribution $\mathcal{F}_{\boldsymbol{v}_t}$ which is nearly Gaussian in the middle.

We note that $\boldsymbol{S}_t$ is simply the square root of the inverse of $\boldsymbol{R}_t$, i.e., $\boldsymbol{S}_t\boldsymbol{S}_t = \boldsymbol{R}_t^{-1}$. We further note that $\boldsymbol{S}_t = \boldsymbol{V}\boldsymbol{D}^{-1/2}\boldsymbol{V}^\top$ is symmetric and is just a distortion of the q-dimensional space without any rotation.

Based on the same considerations as before we obtain some kind of time-invariant version of the score function Ψ_t:

$$\Psi_t(\Delta\boldsymbol{y}_t) \cong \boldsymbol{S}_t\Psi(\boldsymbol{S}_t\Delta\boldsymbol{y}_t) \ , \tag{6.64}$$

and

$$\Psi'_t(\Delta\boldsymbol{y}_t) \cong \boldsymbol{S}_t\Psi'(\boldsymbol{S}_t\Delta\boldsymbol{y}_t)\boldsymbol{S}_t \ , \tag{6.65}$$

where $\Psi(\boldsymbol{S}_t\Delta\boldsymbol{y}_t)$ is a q-dimensional vector with components

$$(\Psi(\boldsymbol{y}))_i = -(\partial/\partial y_i)\log g(\boldsymbol{y}) \tag{6.66}$$

and $\Psi'(\boldsymbol{S}_t\Delta\boldsymbol{y}_t)$ is a $q \times q$ matrix with elements

$$(\Psi'(\boldsymbol{y}))_{ij} = (\partial/\partial y_j)(\Psi(\boldsymbol{y}))_i \ . \tag{6.67}$$

This yields simplified versions of (6.53)–(6.55):

$$\begin{aligned}
\boldsymbol{x}_{t|t} &= \boldsymbol{x}_{t|t-1} + \boldsymbol{M}_t\boldsymbol{H}^\top\boldsymbol{S}_t\Psi(\boldsymbol{S}_t\Delta\boldsymbol{y}_t) & (6.68) \\
\boldsymbol{P}_t &= \boldsymbol{M}_t - \boldsymbol{M}_t\boldsymbol{H}^\top\boldsymbol{S}_t\Psi'(\boldsymbol{S}_t\Delta\boldsymbol{y}_t)\boldsymbol{S}_t\boldsymbol{H}\boldsymbol{M}_t & (6.69) \\
\boldsymbol{M}_{t+1} &= \boldsymbol{\Phi}\boldsymbol{P}_t\boldsymbol{\Phi}^\top + \boldsymbol{Q} \ . & (6.70)
\end{aligned}$$

6.4.2 ACM-type Filters for State Space Models

Since $\mathcal{F}_{\boldsymbol{v}_t}$ will rarely be known in practice, we follow on the same lines as proposed by Martin (1979). Therein he suggests how to define an approximate conditional-mean (ACM) type filter for autoregressive models of order p. We now generalize his results and define a multivariate approximate conditional-mean (ACM) type filter for state-space models with vector-valued observations.

Following the usual M-estimate route we replace the score function Ψ by a good robustifying psi-function ψ, which is bounded and continuous and leaves "small" vectors unchanged. Again, let

$$(\boldsymbol{S}_t\boldsymbol{S}_t)^{-1} = \boldsymbol{R}_t = \boldsymbol{H}\boldsymbol{M}_t\boldsymbol{H}^\top + \boldsymbol{R} \ , \tag{6.71}$$

where $\boldsymbol{R}$ is the covariance matrix of the uncontaminated multivariate Gaussian distribution of the observation noise $\boldsymbol{v}_t$. Then, noting that $\Delta \boldsymbol{y}_t = \boldsymbol{y}_t - \boldsymbol{H}\boldsymbol{x}_{t|t-1}$, the recursions (6.68)–(6.70) are

$$
\begin{aligned}
\boldsymbol{x}_{t|t} &= \boldsymbol{x}_{t|t-1} + \boldsymbol{M}_t \boldsymbol{H}^\top \boldsymbol{S}_t \psi(\boldsymbol{S}_t(\boldsymbol{y}_t - \boldsymbol{H}\boldsymbol{x}_{t|t-1})) & (6.72) \\
\boldsymbol{P}_t &= \boldsymbol{M}_t - \boldsymbol{M}_t \boldsymbol{H}^\top \boldsymbol{S}_t \psi'(\boldsymbol{S}_t(\boldsymbol{y}_t - \boldsymbol{H}\boldsymbol{x}_{t|t-1}))\boldsymbol{S}_t \boldsymbol{H} \boldsymbol{M}_t & (6.73) \\
\boldsymbol{M}_{t+1} &= \boldsymbol{\Phi}\boldsymbol{P}_t\boldsymbol{\Phi}^\top + \boldsymbol{Q} , & (6.74)
\end{aligned}
$$

with $\boldsymbol{x}_{t|t-1} = \boldsymbol{\Phi}\boldsymbol{x}_{t-1|t-1}$. Further, $\psi(\boldsymbol{S}_t(\boldsymbol{y}_t - \boldsymbol{H}\boldsymbol{x}_{t|t-1}))$ is a q-dimensional vector and $\psi'(\boldsymbol{S}_t(\boldsymbol{y}_t - \boldsymbol{H}\boldsymbol{x}_{t|t-1}))$ denotes a $q \times q$ matrix with elements

$$
(\psi'(\boldsymbol{y}))_{ij} = (\partial/\partial y_j)(\psi(\boldsymbol{y}))_i . \tag{6.75}
$$

At first glance it is not clear how to define a vector-valued psi-function in the multivariate case. Ruckdeschel (2001) proposed a multivariate analogue of Huber's psi-function (cf. also Section 6.5). On the same lines we suggest to define the multivariate analogue of Hampel's three-part redescending psi-function in the following way:

$$
\psi_{HA}(\boldsymbol{s}) = \begin{cases} \boldsymbol{s} & \text{if} \quad \|\boldsymbol{s}\| \le a \\ \frac{a}{\|\boldsymbol{s}\|}\boldsymbol{s} & \text{if} \quad a < \|\boldsymbol{s}\| \le b \\ \frac{a}{c-b}(c - \|\boldsymbol{s}\|)\frac{\boldsymbol{s}}{\|\boldsymbol{s}\|} & \text{if} \quad b < \|\boldsymbol{s}\| \le c \\ \boldsymbol{0} & \text{if} \quad c < \|\boldsymbol{s}\| . \end{cases} \tag{6.76}
$$

Furthermore, as in the univariate case, we propose to use the multivariate analogue of Hampel's two-part redescending psi-function where we set $a = b$ for the robustifying psi-function ψ of the above filter recursions. The Figures 6.4 and 6.5 show the approximated score functions of Figures 6.2 and 6.3 using the approximation (6.64) and the multivariate analogue of Hampel's two-part redescending psi-function with parameters $a = b = 2.5$ and $c = 5.0$. Plot (a) shows the graph of the first coordinate of the approximation of the score function originally displayed in plot (c), whereas plot (b) corresponds to the original plot (d).

Hence, the ACM filter for state-space models based on the multivariate analogue of Hampel's two-part redescending psi-function has the appealing feature that $\boldsymbol{x}_{t|t} = \boldsymbol{\Phi}\boldsymbol{x}_{t-1|t-1}$ and $\boldsymbol{P}_t = \boldsymbol{M}_t$ by virtue of $\psi(\boldsymbol{s}) = \psi'(\boldsymbol{s}) = \boldsymbol{0}$ if $\|\boldsymbol{s}\| > c$. This characteristic is exactly the one we would expect from an outlier-rejection rule in the filtering context.
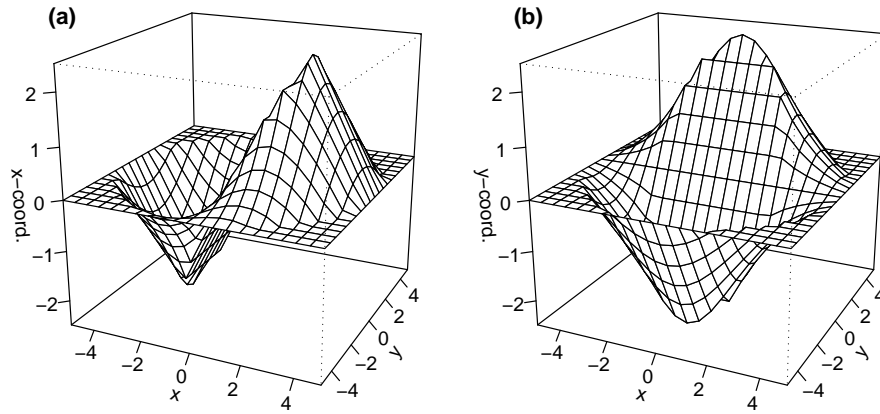
Figure 6.4: Approximated score function of contaminated bivariate normal distribution (I)
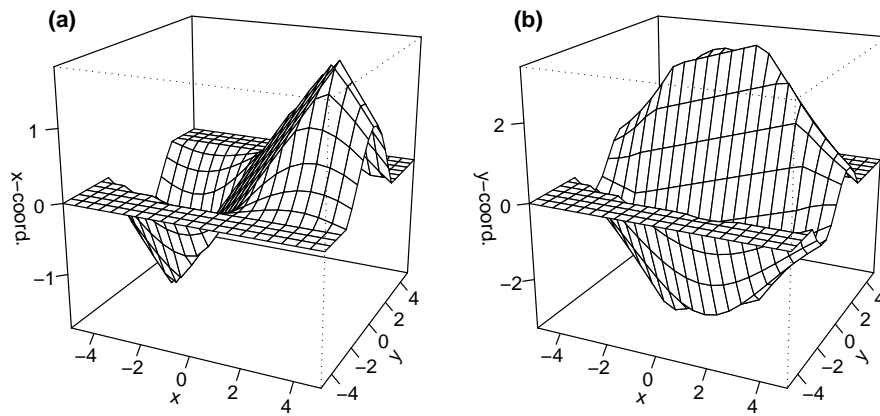

Figure 6.5: Approximated score function of contaminated bivariate normal distribution (II)

6.5 The Robust Least Squares (rLS) Filter Algorithm

In the following we describe a robustified version of the Kalman filter which was proposed by Ruckdeschel (2001).

6.5.1 Robustified Optimization Problem

The idea is to reduce the influence of an observation $\boldsymbol{y}_t$, that is affected by an additive outlier, in the correction step (6.6) of the classical Kalman filter. Instead of $\boldsymbol{K}_t \Delta \boldsymbol{y}_t$ with $\Delta \boldsymbol{y}_t = \boldsymbol{y}_t - \boldsymbol{H} \boldsymbol{x}_{t|t-1}$ we use a huberized version of it, i.e.,

$$H_{b_t}(\boldsymbol{K}_t \Delta \boldsymbol{y}_t) = \boldsymbol{K}_t \Delta \boldsymbol{y}_t \min\{1, \frac{b_t}{\|\boldsymbol{K}_t \Delta \boldsymbol{y}_t\|}\} \ , \tag{6.77}$$

so that the obtained result will be equal to the one of the classical Kalman filter, if $\|\boldsymbol{K}_t \Delta \boldsymbol{y}_t\|$ is not too large. If on the other hand $\|\boldsymbol{K}_t \Delta \boldsymbol{y}_t\|$ is too large, the direction will remain unchanged and it will be projected on the q-dimensional ball with radius b_t.

Additionally we can replace the Euclidean norm in (6.2) by the following loss function ρ_c defined as

$$\rho_c(\boldsymbol{x}) = \left\{ \begin{array}{ll} \|\boldsymbol{x}\|^2 & \text{if} \quad \|\boldsymbol{x}\| \leq c \\ 2c\|\boldsymbol{x}\| - c^2 & \text{if} \quad \|\boldsymbol{x}\| > c \ . \end{array} \right. \tag{6.78}$$

We note that setting $c = \infty$ again yields the Euclidean norm. This leads to a robustified optimization problem given by

$$E(\rho_c(\Delta \boldsymbol{x}_t - H_{b_t}(\boldsymbol{K}_t \Delta \boldsymbol{y}_t))) = \min_{\boldsymbol{K}_t}! \ , \tag{6.79}$$

where $\Delta \boldsymbol{x}_t = \boldsymbol{x}_t - \boldsymbol{x}_{t|t-1}$ denotes the prediction error. The above optimization problem is equivalent to the optimization problem (6.2) of the classical Kalman filter and its solution is named $\boldsymbol{K}_t^{\text{rLS}}$.

6.5.2 The rLS Filter

Hence, this gives us the following filter recursions, named *robust least squares* (rLS) *filter*:

(i) Initialization ($t = 0$):

$$\boldsymbol{x}_{0|0}^{\text{rLS}} = \boldsymbol{\mu}_0 \tag{6.80}$$

(ii) Prediction ($t \geq 1$):

$$\boldsymbol{x}_{t|t-1}^{\text{rLS}} \quad = \quad \boldsymbol{\Phi} \boldsymbol{x}_{t-1|t-1}^{\text{rLS}} \tag{6.81}$$

(iii) Correction ($t \geq 1$):

$$\boldsymbol{x}_{t|t}^{\text{rLS}} \quad = \quad \boldsymbol{x}_{t|t-1}^{\text{rLS}} + H_{b_t}(\boldsymbol{K}_t^{\text{rLS}}(\boldsymbol{y}_t - \boldsymbol{H}\boldsymbol{x}_{t|t-1}^{\text{rLS}})) \tag{6.82}$$

Because the calculation of $\boldsymbol{K}_t^{\text{rLS}}$ is computationally extensive, Ruckdeschel (2001) proposed to use $\boldsymbol{K}_t^{\text{KK}}$ instead, where $\boldsymbol{K}_t^{\text{KK}}$ denotes the Kalman gain obtained by the classical Kalman filter recursions. Simulation studies therein have shown that the worsening, in the sense of a larger mean-squared error, is only small if using $\boldsymbol{K}_t^{\text{KK}}$ instead of $\boldsymbol{K}_t^{\text{rLS}}$. Moreover, it does not make any numerical difference using the Euclidean norm in the robustified optimization problem (6.79), i.e., setting $c = \infty$, or using the modified loss function. Hence, these simplifying modifications almost yield the classical Kalman filter recursions with the only exception of replacing the first line of the correction step in (6.6) by

$$\boldsymbol{x}_{t|t} \quad = \quad \boldsymbol{x}_{t|t-1} + H_{b_t}(\boldsymbol{K}_t^{\text{KK}}(\boldsymbol{y}_t - \boldsymbol{H}\boldsymbol{x}_{t|t-1})) \ . \tag{6.83}$$

From now on, if speaking of the rLS filter, we will only consider this modified version.

Moreover, Ruckdeschel (2001) proved that the rLS filter is SO-optimal under certain side conditions. SO stands for substitutive outlier and means that, instead of disturbing $\boldsymbol{v}_t$, contamination affects $\boldsymbol{y}_t$ directly, replacing it by an arbitrarily distributed variable $\boldsymbol{y}_t'$ with some low probability. For further details we refer the reader to Ruckdeschel (2001).

Still, there remains the open problem of fixing the clipping value b_t.

6.5.3 Fixing the Clipping Value b_t

In order to properly choose b_t, Ruckdeschel (2001) proposed an assurance criterion: How much efficiency in the ideal model relative to the optimal procedure, i.e., the Kalman filter, am I ready to pay in order to get robustness

under deviations from the ideal model? This loss of efficiency, which we will obtain if we use a robust version instead of the classical Kalman filter, is quantified as the relative worsening of the mean-squared error in the ideal model. Hence, for a given relative worsening $\delta > 0$ we solve

$$E(\|\Delta \boldsymbol{x}_t - H_{b_t}(\boldsymbol{K}_t^{\text{rLS}} \Delta \boldsymbol{y}_t)\|^2) \overset{!}{=} (1+\delta) E(\|\Delta \boldsymbol{x}_t - \boldsymbol{K}_t^{\text{KK}} \Delta \boldsymbol{y}_t\|^2) . \quad (6.84)$$

(The symbol $\overset{!}{=}$ means that b_t is chosen in a way to achieve equality.)

Again, we use the simplifying modifications just mentioned and replace $\boldsymbol{K}_t^{\text{rLS}}$ by $\boldsymbol{K}_t^{\text{KK}}$. Moreover, in most time-invariant situations, the sequence of $\boldsymbol{M}_t$ (and hence also of $\boldsymbol{P}_t$ and $\boldsymbol{K}_t^{\text{KK}}$) stabilizes due to asymptotic stationarity. Thus, once $\boldsymbol{M}_t$ does not change for more than a given tolerance level, we can stop calibration and use the finally calculated b_t for all subsequent times $s, s > t$. The Kalman gain and filtering error covariance matrix used in this previous calibration step will be denoted by $\boldsymbol{K}_\infty^{\text{KK}}$ and $\boldsymbol{P}_\infty$, respectively. For details we refer to Anderson and Moore (1979) and Moore and Anderson (1980). Further we make another simplifying modification and assume that for all t

$$\Delta \boldsymbol{x}_t \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{M}_t) \quad \text{and} \quad \boldsymbol{v}_t \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{R}) . \quad (6.85)$$

Thus, we may solve

$$\begin{aligned} E(\|\Delta \boldsymbol{x} - H_b(\boldsymbol{K}_\infty^{\text{KK}} \Delta \boldsymbol{y})\|^2) &\overset{!}{=} (1+\delta) E(\|\Delta \boldsymbol{x} - \boldsymbol{K}_\infty^{\text{KK}} \Delta \boldsymbol{y}\|^2) \\ &= (1+\delta) \operatorname{tr} \boldsymbol{P}_\infty , \end{aligned} \quad (6.86)$$

in b, uniquely for a given loss of efficiency δ, where $\operatorname{tr} \boldsymbol{P}_\infty$ denotes the trace of the conditional filtering error covariance matrix. We note that the relatively time-consuming calibration, i.e., finding b to a given δ, can be done beforehand. Additional details may be found in Ruckdeschel (2001) and Ruckdeschel (2000).

6.6 Simulation Study

To test how well the multivariate ACM-type filter works we simulate two different state-space models with different additive outlier situations and different amounts of contamination and compare the results to those of the rLS filter.

For the first state-space model we use the following hyper parameters:

$$\boldsymbol{\mu_0} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad \boldsymbol{\Sigma_0} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix},$$
$$\boldsymbol{\Phi} = \begin{pmatrix} 0.5 & 0.3 \\ 0.6 & 0.5 \end{pmatrix}, \quad \boldsymbol{H} = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}, \qquad \qquad (6.87)$$
$$\boldsymbol{Q} = \begin{pmatrix} 3 & 2 \\ 2 & 3 \end{pmatrix}, \qquad \boldsymbol{R} = \begin{pmatrix} 2 & -0.2 \\ -0.2 & 0.5 \end{pmatrix}.$$

For the second state-space model we further use the following hyper parameters:

$$\boldsymbol{\mu_0} = \begin{pmatrix} 20 \\ 0 \end{pmatrix}, \quad \boldsymbol{\Sigma_0} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix},$$
$$\boldsymbol{\Phi} = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}, \quad \boldsymbol{H} = \begin{pmatrix} 0.3 & 1 \\ -0.3 & 1 \end{pmatrix}, \qquad \qquad (6.88)$$
$$\boldsymbol{Q} = \begin{pmatrix} 0 & 0 \\ 0 & 9 \end{pmatrix}, \quad \boldsymbol{R} = \begin{pmatrix} 9 & 0 \\ 0 & 9 \end{pmatrix}.$$

We note that the first coordinate of the above state process is a random walk and therefore non-stationary, whereas the second coordinate is just white noise.

Additionally, for the first state-space model we simulate the $\boldsymbol{v}_t$'s of the observation process from a contaminated bivariate normal distribution (6.49) with

$$\boldsymbol{R} = \begin{pmatrix} 2 & -0.2 \\ -0.2 & 0.5 \end{pmatrix} \quad \text{and} \quad \boldsymbol{R}_c = \begin{pmatrix} 100 & 0 \\ 0 & 100 \end{pmatrix}, \qquad (6.89)$$

whereas for the second state-space model the $\boldsymbol{v}_t$'s are simulated from a contaminated bivariate normal distribution with a contaminating distribution given by

$$\mathcal{N}_2\left(\begin{pmatrix} 25 \\ 30 \end{pmatrix}, \begin{pmatrix} 0.9 & 0 \\ 0 & 0.9 \end{pmatrix} \right). \qquad (6.90)$$

We note that the mean of the contaminating normal distribution is now unequal to zero. This case was not considered in Section 6.4 and therefore is a slightly more generalized definition of the contaminated multivariate normal distribution than the one originally given in (6.49).

The contamination γ is varied from 0% to 20% by steps of 5%. For our first and second state-space model that means that with probability γ, $\boldsymbol{v}_t$ is an additive outlier originating either from

$$\mathcal{N}_2\left(\left(\begin{array}{c} 0 \\ 0 \end{array}\right), \left(\begin{array}{cc} 100 & 0 \\ 0 & 100 \end{array}\right)\right) \quad \text{or} \quad \mathcal{N}_2\left(\left(\begin{array}{c} 25 \\ 30 \end{array}\right), \left(\begin{array}{cc} 0.9 & 0 \\ 0 & 0.9 \end{array}\right)\right). \quad (6.91)$$

For each level of contamination we simulate 400 realisations.

Then the bivariate state processes of each simulated state-space model are computed using the multivariate ACM-type filter and the rLS filter proposed in Section 6.4 and by Ruckdeschel (2001), respectively. For the ACM-type filter we used the multivariate analogue of Hampel's three-part redescending psi-function with $a = b = 2.5$ and $c = 5.0$ for both state-space models. The rLS filter was calibrated to a given loss of efficiency $\delta = 0.1$ for both models and the clipping value b was set to $b = 3.7$ and $b = 2.4$ for the first and second state-space model, respectively. Moreover, also the classical Kalman filter estimates were calculated.

Finally, the mean-squared error, denoted by MSE, between the true state process and the different filter estimates averaged over time is computed, i.e.,

$$MSE_{\widehat{\boldsymbol{x}}_{t|t}} := \operatorname*{mean}_{t}(\|\boldsymbol{x}_t - \widehat{\boldsymbol{x}}_{t|t}\|^2) \quad (6.92)$$

where $\boldsymbol{x}_t$ and $\widehat{\boldsymbol{x}}_{t|t}$ denote the true state vector and the filter estimate, respectively.

6.7   Results

Regarding the computation time the rLS filter performs slightly better than the ACM-type filter once the clipping value b is fixed. This is due to the fact that additional computations have to be done within the correction step of the ACM-type filter.

In Figure 6.6 the results of the first simulation study are presented. Typical realisations of the state process together with their filter estimates are displayed for different levels of contamination. The first coordinate of the state vector is plotted in the left column, whereas the second coordinate is displayed in the right column. Each plot contains four graphs. The thick black line is the true state process, whereas the light grey line represents the filter estimate of the classical Kalman filter. The dotted dark grey line is
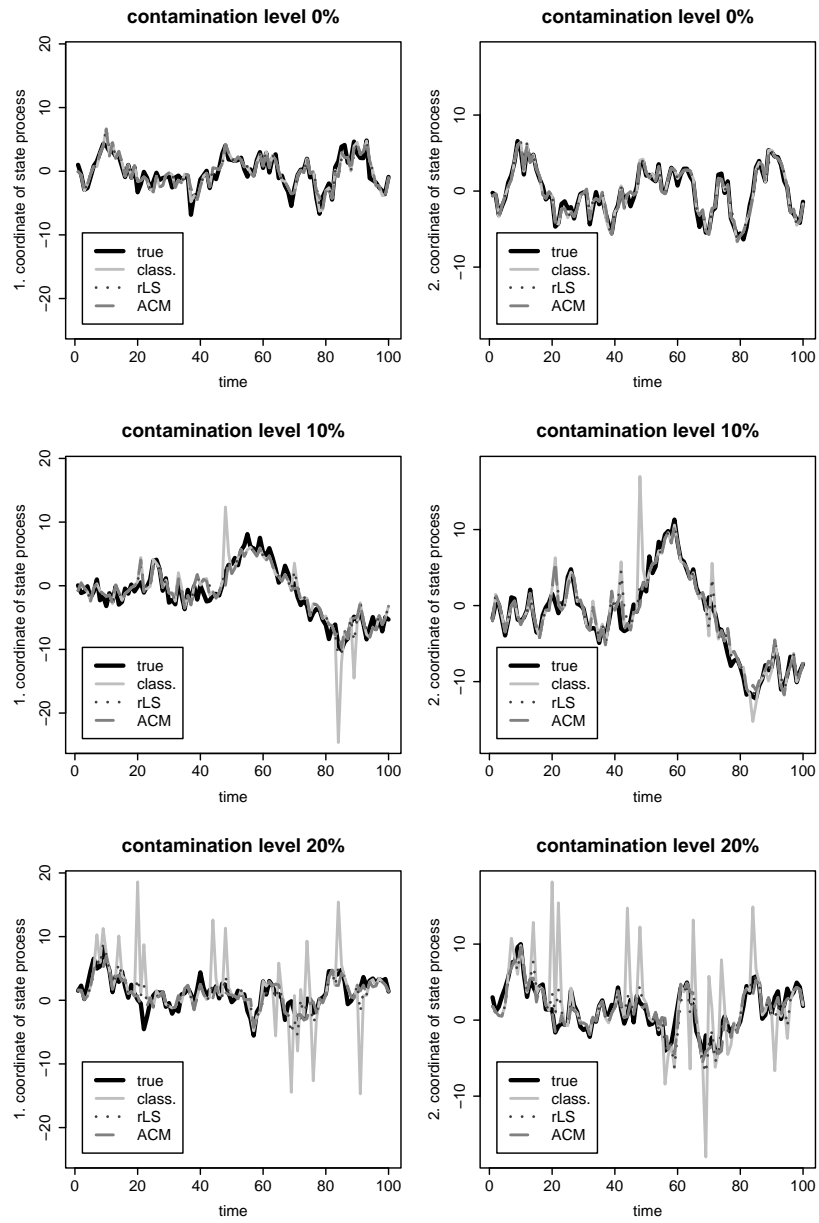
Figure 6.6: Typical results of the first state-space model at different levels of contamination, left column first coordinate of state process, right second coordinate of state process

the filter estimate obtained by the rLS filter and the filter estimate using the ACM-type filter is plotted as a dashed grey line.

We note that the contamination by additive outliers cannot be seen directly in Figure 6.6 because only the observation equation is affected. However, the effect of additive outliers is indirectly observable: The spikes, especially visible for the filter estimate of the classical Kalman filter, are caused by additive outliers.

In Figure 6.7 the same graphs as in Figure 6.6 are plotted for the second state-space model. Again, the first coordinate of the state vector is displayed in the left column and the second coordinate is plotted in the right column.

It is even more visible in Figure 6.7 than in Figure 6.6 that the deviation between the true state process and the different filter estimates becomes larger the higher the amount of contamination. As expected, this difference is largest for the classical Kalman filter estimates. However, it is also larger for the filter estimates obtained by the rLS filter than for those of the ACM-type filter.

This observation is also confirmed by Figure 6.8. For our first simulation experiment plot (a) displays the mean-squared errors between the true state vector and each filter estimate for all levels of contamination. Similarly, the mean squared errors for the second state-space model are seen in plot (b).

Moreover, because of the fact that the rLS filter was calibrated to a given loss of efficiency $\delta = 0.1$ it yields larger errors in the case of no contamination. This has already been noted by Ruckdeschel (2001) and is especially visible in plot (b) of Figure 6.8. Furthermore, we see that according to the mean-squared errors the ACM-type filter performs best for all contamination levels.

6.8 Discussion

Based on the work of Masreliez (1975) and Martin (1979) we developed a new multivariate approximate conditional-mean (ACM) type filter for state-space models with vector-valued observations which generalizes Martin's results (Martin, 1979).

The results of our simulation experiments show that the multivariate ACM-type filter performs very well compared to the rLS filter proposed by Ruckdeschel (2001). Moreover, the ACM-type filter yields remarkably good results not only in situations where the $\boldsymbol{v}_t$'s have a contaminated multivariate normal distribution but also in outlier situations where the mean of the
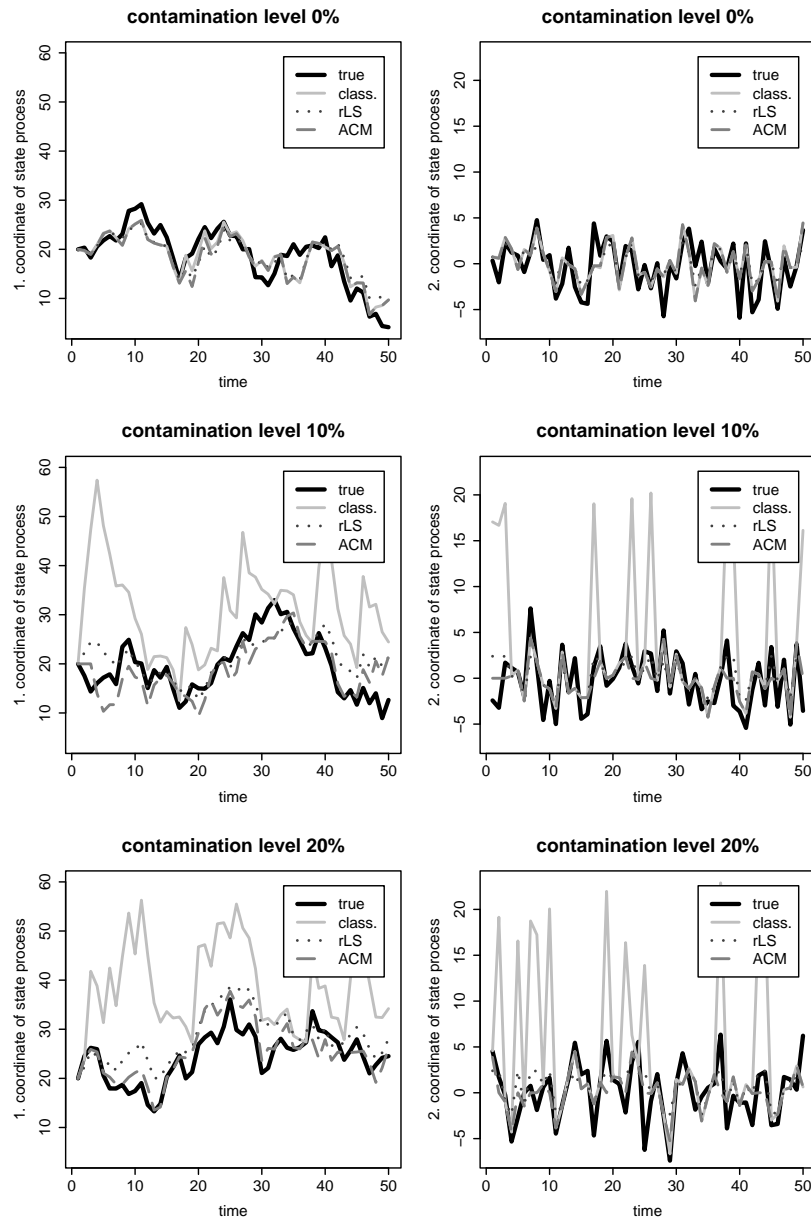
Figure 6.7: Typical results of the second state-space model at different levels of contamination, left column first coordinate of state process, right second coordinate of state process

73
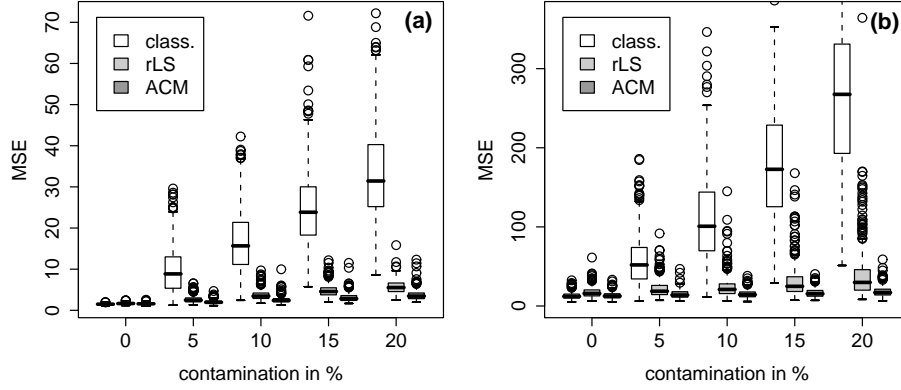
Figure 6.8: Boxplots of the errors, first (a) and second (b) state-space model

contaminating distribution is additionally unequal to zero.

We note that the weighting in the correction step of the univariate ACM-type filter as well as of the multivariate one is a discontinuous function if using Hampel's two-part redescending psi-function or its multivariate analogue. However, in the univariate case we have explicitly implemented the first derivative of Hampel's psi-function, whereas in the multivariate case the calculation of the Jacobian matrix is accomplished via numerical differentiation. Moreover, in the univariate case the first derivative of the psi-function can be replaced by a continuous weight function as already proposed by Martin and Thomson (1982).

Additional simulation studies using different state-space models have already been done yielding similar results and are therefore not published here.

Chapter 7

Robust Prewhitened Spectral Density Estimation

7.1 The Robust Filter-cleaner Algorithm

The procedure to obtain a robust spectral density estimate proposed by Martin and Thomson (1982) incorporates an important data-cleaning operation wherein the robustness is introduced.

7.1.1 Robust Prewhitening

Let $\{y_t, t = 1, \ldots, n\}$ again denote the observed process which is assumed to be second-order stationary and to have zero mean. The cleaning operator C maps the original data y_t into the cleaned data Cy_t. In the context of the AO model (2.13), we want the Cy_t to reconstruct the core process x_t, and so we will use the labeling $Cy_t = \widehat{x}_{t|t}$, where $\widehat{x}_{t|t}$ denotes an estimate of x_t at time t. The second index of $\widehat{x}_{t|t}$ should indicate that the kind of data cleaning procedure we have in mind here is a robust filtering procedure which uses the past and present data values $y_1, \ldots, y_t$ to produce a cleaned filter estimate $\widehat{x}_{t|t}$ of x_t, $t = 1, \ldots, n$. For AO models with a fraction of contamination γ not too large, it turns out that the data cleaner has the property that $Cy_t = y_t$ most of the time, that is about $(1 - \gamma) \times 100$ percent of the time.

The filter-cleaner procedure involves a robust estimation of an autoregressive approximation to the core process x_t of order p, with estimated

coefficients $\widehat{\phi}_1, \ldots, \widehat{\phi}_p$. Now, the residual process

$$r_t = Cy_t - \sum_{j=1}^{p} \widehat{\phi}_j Cy_{t-j} , \quad t = p+1, \ldots, n , \tag{7.1}$$

can easily be formed. Since cleaned data are used to obtain these residuals, and the $\widehat{\phi}_i$ are robust estimates, the transformation (7.1) is called a robust prewhitening operation. The benefit in the use of prewhitening in the context of spectral density estimation is to reduce the bias, i.e., the transfer of power from one frequency region of the spectral density function to another, known as leakage (cf. Blackman and Tukey, 1958).

The robust estimate of the spectral density function is based on the above robust prewhitening as follows. Let

$$\widehat{H}_p(f) = 1 - \sum_{j=1}^{p} \widehat{\phi}_j e^{-i2\pi f j \Delta t} \tag{7.2}$$

be the transfer function of the prewhitening operator (7.1) at frequency f, and let $\widehat{S}_r^{(lw)}(f)$ denote a lag window spectral density estimate based on the residual process r_t. Then the prewhitened spectral density estimate is defined as

$$\widehat{S}(f) = \frac{\widehat{S}_r^{(lw)}(f)}{|\widehat{H}_p(f)|^2} , \tag{7.3}$$

where $\widehat{S}(f)$ is evaluated at the Fourier frequencies $f_k = k/(n\Delta t)$, $k = 1, \ldots, \lfloor n/2 \rfloor$.

7.1.2 The Robust Filter-cleaner

The filter-cleaner algorithm as presented in the paper of Martin and Thomson (1982) relies on the p-th order autoregressive approximation of the underlying process x_t, which can be represented in state-space form (6.1) as follows. Assuming that x_t satisfies

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + \varepsilon_t$$

the state space model can be written as

$$\begin{aligned} \boldsymbol{x}_t &= \boldsymbol{\Phi} \boldsymbol{x}_{t-1} + \boldsymbol{\varepsilon}_t \\ y_t &= x_t + v_t , \end{aligned} \tag{7.4}$$

with

$$\boldsymbol{x}_t = (x_t, x_{t-1}, \dots, x_{t-p+1})^\top , \qquad (7.5)$$

$$\boldsymbol{\varepsilon}_t = (\varepsilon_t, 0, \dots, 0)^\top \text{ and} \qquad (7.6)$$

$$\boldsymbol{\Phi} = \begin{pmatrix} \phi_1 & \cdots & \phi_{p-1} & \phi_p \\ 1 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 \end{pmatrix} . \qquad (7.7)$$

Additionally, we set

$$\text{cov}(\boldsymbol{\varepsilon}_t) = \boldsymbol{Q} = \begin{pmatrix} \sigma_\varepsilon^2 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix} \text{ and} \qquad (7.8)$$

$$\text{var}(v_t) = \boldsymbol{R} = \sigma_0^2 . \qquad (7.9)$$

The algorithm computes robust estimates $\widehat{\boldsymbol{x}}_{t|t}$ of the unobservable $\boldsymbol{x}_t$ according to the following recursion:

$$\widehat{\boldsymbol{x}}_{t|t} = \boldsymbol{\Phi}\widehat{\boldsymbol{x}}_{t-1|t-1} + \frac{\boldsymbol{m}_{.1,t}}{s_t^2} s_t \, \psi\left(\frac{y_t - \widehat{y}_{t|t-1}}{s_t}\right) \qquad (7.10)$$

with $\boldsymbol{m}_{.1,t}$ being the first column of $\boldsymbol{M}_t$ which is computed recursively as

$$\boldsymbol{M}_{t+1} = \boldsymbol{\Phi}\boldsymbol{P}_t\boldsymbol{\Phi}^\top + \boldsymbol{Q} \qquad (7.11)$$

$$\boldsymbol{P}_t = \boldsymbol{M}_t - w\left(\frac{y_t - \widehat{y}_{t|t-1}}{s_t}\right) \frac{\boldsymbol{m}_{.1,t}\boldsymbol{m}_{.1,t}^\top}{s_t^2} . \qquad (7.12)$$

The weight function w is defined by

$$w(r) = \frac{\psi(r)}{r} , \qquad (7.13)$$

where ψ stands for some robustifying psi-function. The scale s_t is set to

$$s_t^2 = m_{11,t} \qquad (7.14)$$

and $\widehat{y}_{t|t-1}$ denotes a robust one-step-ahead prediction of y_t based on $\boldsymbol{Y}_{t-1} = \{y_1, \dots, y_{t-1}\}$, and is given by

$$\widehat{y}_{t|t-1} = (\boldsymbol{\Phi}\widehat{\boldsymbol{x}}_{t-1|t-1})_1 . \qquad (7.15)$$

77

Finally, the cleaned process at time t results in

$$\widehat{x}_{t|t} = (\widehat{\boldsymbol{x}}_{t|t})_1 \ . \tag{7.16}$$

It should be noted that if ψ is the identity function and $w \equiv 1$, and (7.14) is replaced by $s_t^2 = m_{11,t} + \sigma_0^2$ with $\sigma_0^2 = \mathrm{var}(v_t)$ in the AO model, the above recursions are those of the Kalman filter. The use of $\sigma_0^2 = 0$ in (7.14) corresponds to the assumptions that $v_t = 0$ a large fraction of time and that a contaminated normal distribution with degenerate central component, i.e.,

$$\mathcal{CN}(\gamma, 0, \sigma^2) = (1 - \gamma)\mathcal{N}(0,0) + \gamma \mathcal{N}(0, \sigma^2) \ , \tag{7.17}$$

provides a reasonable model. Correspondingly, $\boldsymbol{M}_t$ and $\boldsymbol{P}_t$ are the prediction and filtering error-covariance matrices as described in the previous section (Section 6.3). Again, in order to agree with the definition of the classical Kalman filter recursions, we specify the initial conditions for the above recursions by setting $\widehat{\boldsymbol{x}}_{0|0} = \boldsymbol{0}$ and $\boldsymbol{P}_0 = \widehat{\boldsymbol{\Sigma}}_0$ where $\widehat{\boldsymbol{\Sigma}}_0$ is an estimate of the $p \times p$ covariance matrix of the state process. As mentioned before, there also exists another way to specify those initial conditions (see Martin, 1981).

The psi-function ψ and the weight function w which are essential to obtain robustness should be bounded and continuous. Additionally, it is highly desirable that both have zero values outside a bounded, symmetric interval around the origin. Furthermore, $\psi(s)$ is odd and should look like the identity function for small values of s (see Martin, 1979). Boundedness assures that no single observation has an arbitrarily large effect on the filter-cleaner. Continuity assures that small variations, e.g., due to rounding, will not have a major effect. Compact support results in the following behavior which is desirable for a filter-cleaner: If an observation y_t deviates from its prediction $\widehat{y}_{t|t-1}$ by a sufficiently large amount, then $\widehat{\boldsymbol{x}}_{t|t}$ will be the pure prediction $\widehat{\boldsymbol{x}}_{t|t} = \boldsymbol{\Phi}\widehat{\boldsymbol{x}}_{t-1|t-1}$ and the filtering error covariance $\boldsymbol{P}_t$ is set equal to the prediction error covariance $\boldsymbol{M}_t$. Martin and Thomson (1982) proposed to use for ψ a special form of Hampel's three-part redescending psi-function (6.48), namely, Hampel's two-part redescending psi-function, with $b = a$, which has all the desirable properties.

Moreover, the robust filter-cleaner just described differs from the simple robust filter of Kleiner et al. (1979) in two important aspects. First (7.10)–(7.15) are vector and matrix recursions, and the structure of the factor $\boldsymbol{m}_{.1,t}/s_t^2$ is such that good data points following an outlier can be used to improve the estimate of x_t at the outlier position. This can yield improved

estimates of x_u at times $u > t$, as well. Secondly, the current filter incorporates the data-dependent scale s_t whose values satisfy $s_{t+1} > s_t$ if a gross outlier occurs at time t. Particularly, this is an important feature when using a redescending psi-function. If a fixed scale s were used, the filter could lose track of the data and never regain them for the duration of a fixed length data set $y_1, \ldots, y_n$. Although this may also be possible if using s_t, the properties of s_t suggest that the filter will tend to regain track more quickly than if a fixed scale is used.

7.1.3   An Approximate Optimality Result

There is an approximate optimality result for the filter described above if we replace (7.14) by

$$s_t^2 = m_{11,t} + \sigma_0^2 \;, \tag{7.18}$$

and w in (7.13) by

$$w(r) = \psi'(r) = \frac{\partial}{\partial r}\psi(r) \;. \tag{7.19}$$

Namely, under the assumption that the state prediction density $f_{\boldsymbol{x}_t}(.|\boldsymbol{Y}_{t-1})$ is Gaussian and that $\psi(r) = -(\partial/\partial r)\log g(r)$, where g is an approximation of the observation prediction density $f_{y_t}(.|\boldsymbol{Y}_{t-1})$, the filter is the conditional-mean filter proposed by Masreliez (1975). Further details may be found in Section 6.3. The preceding assumption will never hold exactly under an AO model where v_t is non-Gaussian (see Martin, 1979, Sec. 5). However, there is some evidence that $f_{\boldsymbol{x}_t}(.|\boldsymbol{Y}_{t-1})$ is nearly Gaussian and that the filter is a good approximation to the exact conditional-mean filter. Therefore the filter is referred to as an approximate conditional-mean (ACM) filter. More details can be found in Martin (1979). The results therein suggest that the use of Hampel's two-part redescending psi-function is reasonable when the observation noise v_t has a contaminated normal distribution. However, the weight function w given by (7.19) is discontinuous if using Hampel's two-part redescending psi-function, and therefore Martin and Thomson (1982) prefer to specify w by (7.13).

7.1.4   Fixed-lag Smoother-cleaners

As already mentioned in Section 6.3, if one uses the last coordinate of the filter estimate $\widehat{\boldsymbol{x}}_{t|t}$ to produce cleaned data, then one has that $\widehat{x}_{t-p+1} = (\widehat{\boldsymbol{x}}_{t|t})_p$

is an estimate of x_{t-p+1} based on the observations $\boldsymbol{Y}_t$ up to time t. Such an estimate is usually called a fixed-lag smoother, with lag $p-1$ in this case.

7.1.5 Estimation of Hyper Parameters and Order Selection

To use the filter-cleaner algorithm we need robust estimates $\widehat{\boldsymbol{\phi}}$, $\widehat{\sigma}_\varepsilon$ and $\widehat{\boldsymbol{\Sigma}}_0$ of the AR(p) parameter vector $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)^\top$, the innovations scale σ_ε and the $p \times p$ covariance matrix of the state process, respectively. Martin and Thomson (1982) proposed to get initial estimates using bounded-influence autoregression (BIAR) via the *iteratively reweighted least squares* (IWLS) *algorithm*. Details about BIAR may be found in Section 5.1.

The order p of the autoregressive model is chosen according to Section 5.1.5.

7.2 Simulation Study

We use the same simulation setup as in Section 5.3: First we simulate a core process x_t of length $n = 100$. x_t is chosen to be an autoregressive process of order 2 given by

$$x_t = x_{t-1} - 0.9 x_{t-2} + \varepsilon_t , \tag{7.20}$$

with $\varepsilon_t \sim \mathcal{N}(0, 1)$. The variance of the core process x_t, i.e., the value of the autocovariance function at lag zero, can be calculated by numerical integration and is given approximately by $\mathrm{var}(x_t) \approx 7.27$. Additionally, the additive outliers are simulated from a contaminated normal distribution with degenerate central component (7.17) with $\sigma^2 = 10^2$. We again vary the contamination γ from 0% to 20% by steps of 5%. That means that with probability γ, v_t is an additive outlier with $v_t \neq 0$. To obtain the contaminated process y_t, the v_t's are added to the core process x_t. For each level of contamination this is done 400 times.

For each contaminated series, estimates of the hyper parameter, i.e., the innovations scale $\widehat{\sigma}_\varepsilon$, the autoregressive parameters $\widehat{\phi}_1, \ldots, \widehat{\phi}_p$ and the $p \times p$ covariance matrix $\widehat{\boldsymbol{\Sigma}}_0$ of the state process $\boldsymbol{x}_t$, are computed via bounded-influence autoregression. The order p of the autoregressive approximation is chosen according to the order-selection criterion proposed by Martin and

Thomson (1982), which yields values of p from 2 to 3 subject to the contamination level. In order to be able to compare the results we choose an equal order p for all levels of contamination and fix it equal to 3. Using an order $p = 2$ in cases of lower contamination levels, where this is appropriate, we obtain almost perfect fits for both filtering algorithms. But, although the simulated core process is of order 2, the estimated BIAR parameters we obtain setting p equal to 3 are similar to those of the original core process, i.e., the first two AR parameters are close to the original ones and the third AR parameter is almost zero, as one would expect.

Then each process is cleaned using the ACM-type filter and the rLS filter proposed by Martin and Thomson (1982) and Ruckdeschel (2001), respectively. Afterwards, the hyper parameters of the filtered series are estimated again.

Those re-estimated hyper parameters are used to calculate a prewhitened spectral density estimate for each process. Last, the deviation of each estimated spectral density function from the true spectral density function is measured in the sense of the squared L_2-norm, i.e.,

$$err^2_{\widehat{S}(f)} := \|\widehat{S}(f) - S(f)\|^2 = \int (\widehat{S}(f) - S(f))^2 df \ , \tag{7.21}$$

where $\widehat{S}(f)$ and $S(f)$ denote the estimated and true spectral density functions.

7.3   Results

Regarding the computation time the rLS filter performs better than the ACM-type filter as we expected. This is due to the fact that additional weights have to be computed within the correction step of the ACM-type filter.

Figure 7.1 tries to visualize the results of our simulation study. For both methods and each level of contamination seven curves are plotted on a logarithmic scale. The bold black line represents the true spectral density function, whereas the thin black line is the spectral density estimate of one realization out of 400. Moreover, we calculate frequency by frequency the minimum and maximum, the first and third quartile and median value of all spectral density estimates. Connecting all median values we obtain the grey line, to which we will refer hereafter as median spectral density function. In the same

sense we refer to all minimum values as minimum spectral density function, and so on. Hence, the lower and upper dotted lines are the minimum and maximum spectral density functions, whereas the lower and upper dashed lines represent the first and third quartile spectral density functions. The results obtained by using the ACM-type filter are plotted in the left column, whereas the results of the rLS filter are displayed in the right column.

As expected, for both methods the dispersion of the spectral density estimates becomes larger the higher the contamination. However, this effect is more visible, especially at higher frequencies, when using the ACM-type filter.

Next, we try to visualize the squared errors of the estimated spectral density functions. First, the logarithm of the squared errors is taken. For both methods Figure 7.2 shows boxplots of the squared errors in eight equally-sized frequency bands as well as the total squared errors (bottom right) for all different levels of contamination. Again, the squared errors become larger the higher the contamination, especially at higher frequencies. And, this effect again is larger, when using the ACM-type filter. However, these errors are very small and, looking at the total squared errors, we see that the ACM-type filter performs better than the rLS filter. The largest contribution to the total squared error is the amount of the frequency band where the spectral density function has its peak. There the squared errors using the rLS filter are larger than the ones using the ACM-type filter. Moreover, we see that all squared errors are in the same range for all contamination levels.

7.4   Discussion

In order to get a robust estimate of the spectral density function, it turns out that cleaning the series in a robust way first and calculating a prewhitened spectral density estimate afterwards leads to encouraging results. This data-cleaning operation wherein the robustness is introduced is solved by two different robustified versions of the Kalman filter. Although, as far as we know, there exist no theoretical results on the statistical properties of both proposed multi-step procedures, the empirical results based on simulations and real data sets promise those procedures to be of high quality. The results of the simulation study suggest that the ACM-type filter algorithm performs slightly better than the rLS filter algorithm.

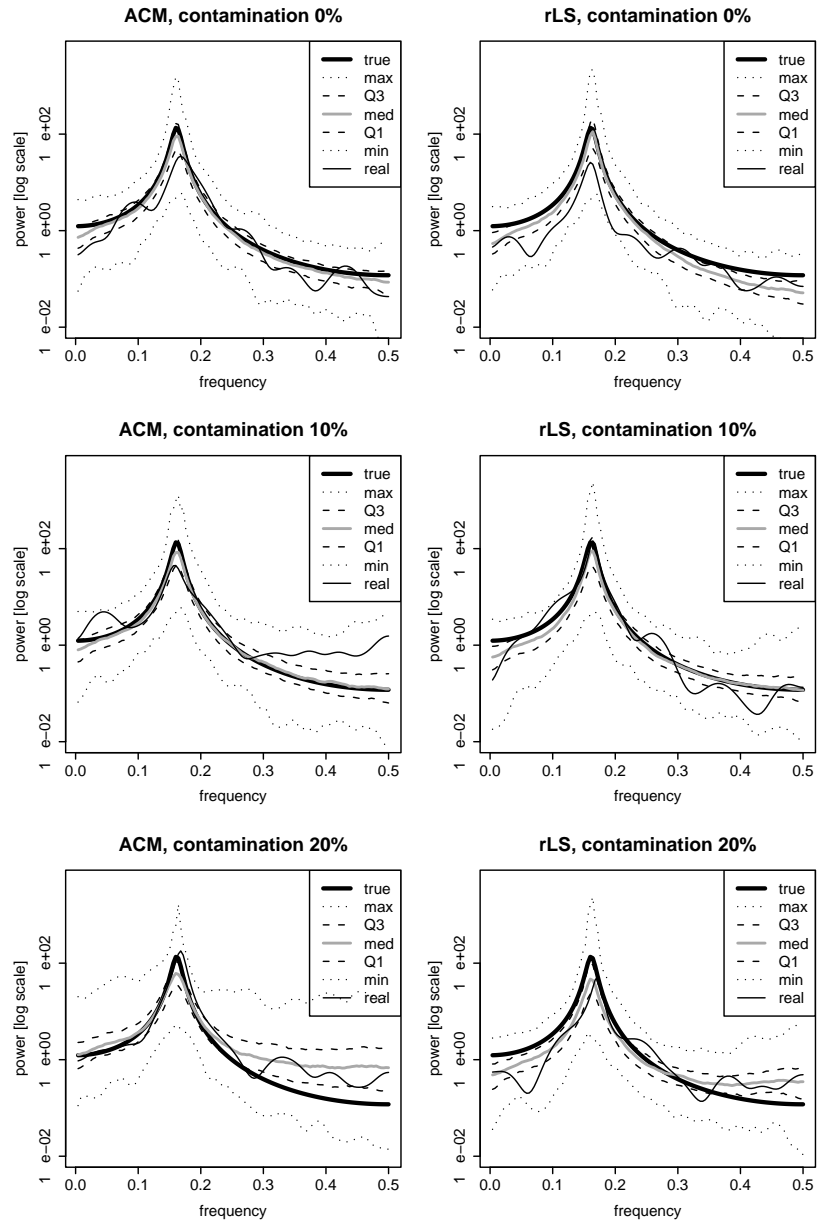In Spangl and Dutter (2005) we compare the ACM-type filter approach

Figure 7.1: Robust spectral density estimates of the simulated data, left column 'ACM', right 'rLS'
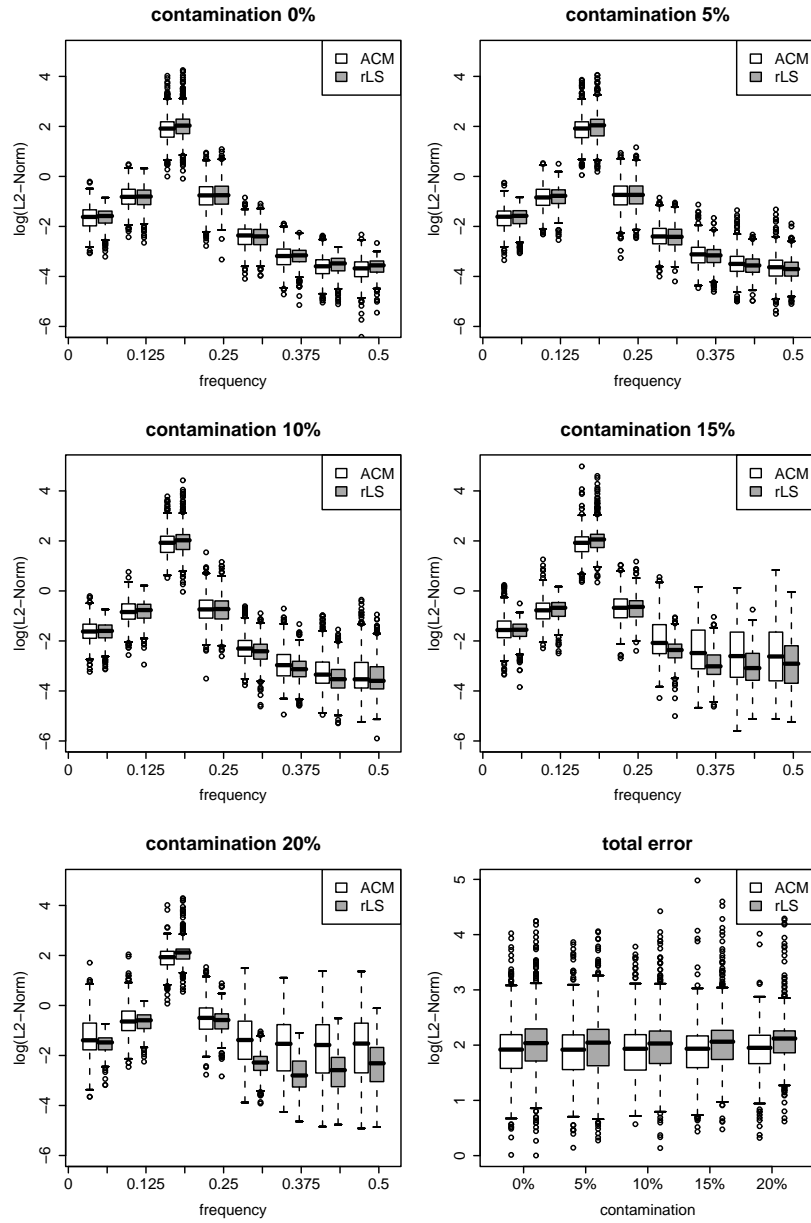
Figure 7.2: Boxplots of the errors

with another approach proposed by Tatum and Hurvich (1993a). This procedure, called biweight filter-cleaner (cf. Section 8.2), also yields good results, but tends to underestimate the core process slightly. Moreover it is computationally intensive.

The problem of estimating the hyper parameters was accomplished by bounded-influence autoregression. An alternative way would be to use a highly robust autocovariance function estimator (cf. Ma and Genton, 2000) and calculate estimates of the hyper parameters via the Yule-Walker equations. Hyper parameters may also be obtained by computing a robust covariance matrix via the MCD algorithm (cf. Rousseeuw and Van Driessen, 1999) and estimate the parameters again using the Yule-Walker equations (see Chapter 5). Recently, Maronna et al. (2006) propose to use τ-estimates.

Chapter 8

Further Related Methods

8.1 Robustifying Welch's Overlapped Segment Averaging

Using conventional nonparametric methods we obtain a direct spectral estimator by using an appropriate data taper (in order to reduce the bias of the periodogram estimator) and by then calculating the periodogram of the tapered data series.

In order to get a smoother spectral estimator with better variance properties Welch (1967) suggested to split a time series into a number of overlapping blocks, compute a direct spectral estimate based on the data in each block and average the individual spectral estimates to form an overall one. This method is known as Welch's Overlapped Segment Averaging (WOSA).

Based on this approach Chave et al. (1987) proposed a method that is both, robust and data-adaptive. The spectral density estimates are not just averaged together but are combined in such a way that individual estimates corresponding to blocks contaminated by outliers are downweighted.

In detail, let $\{y_t,\, t = 1, \ldots, n\}$ be the observed process again.

The process is split into n_B overlapping blocks of length n_S.

Additionally, direct spectral estimates are calculated for different blocks of n_S contiguous data values

$$\widehat{S}_j^{(d)}(f) = \Delta t \left| \sum_{t=1}^{n_S} h_t y_{t+(j-1)N} e^{-i2\pi f t \Delta t} \right|^2 , \quad j = 1, \ldots, n_B , \qquad (8.1)$$

where $h_1, \ldots, h_{n_S}$ is a data taper. $N = (1 - \lambda)n_S \leq n_S$ and $\lambda < 1$ is the given percentage of overlapping.

Then the WOSA spectral estimator is defined by

$$\widehat{S}^{(WOSA)}(f) = \frac{1}{n_B} \sum_{j=1}^{n_B} \widehat{S}_j^{(d)}(f) \ . \tag{8.2}$$

A robust spectral estimator can be obtained by replacing the sample mean in (8.2) by an M-estimator, i.e.,

$$\min_{\theta} \sum_j \rho \left(\frac{x_j - \theta}{s} \right) \ , \text{ or, equivalently, } \quad \sum_j \psi \left(\frac{x_j - \theta}{s} \right) = 0 \ , \tag{8.3}$$

where x_j corresponds to $\widehat{S}_j^{(d)}(f)$ and $\psi(r) = \rho'(r)$. The solution $\widehat{\theta}$ is called M-estimate.

Because outlier contamination can only result in a spectrum that is biased upwards (cf. Chave et al., 1987), a special asymmetric psi-function is used. It is defined by

$$\psi(r) = r \exp(- \exp(\beta(r - \beta))) \ . \tag{8.4}$$

The solution is then calculated using iteratively reweighted least squares (IWLS) with proper initial values, e.g., the sample median and a corrected version of the median absolute deviation (MAD).

8.2 The Biweight Filter-cleaner Algorithm

Tatum and Hurvich (1993a) propose a frequency domain approach to the problem of cleaning outliers in time series. Their high breakdown method assumes only that the core process is Gaussian and has a continuous spectrum. It is nonparametric in the sense that it does not assume a finite parameter model, e.g. an ARMA model, for the core process. The approach uses robust trigonometric regression to fit a sine and cosine coefficient at each Fourier frequency and, hence, to obtain a robustified discrete Fourier transform. These coefficients are then inverse Fourier transformed to get a filtered version of the data. This procedure is termed biweight filter (BF). On this basis, a cleaned version of the data is constructed in which observations that appear to be outliers are replaced and most of the original series remain

unchanged. The replacement values are found by a linear interpolation of all "non-outlying" data points. The interpolation is based on an estimate of the autocovariance function of the filtered series. This additional step is called biweight filter-cleaner (BFC).

8.2.1   The Biweight Filter

Any time series $\{y_t, t = 0, \ldots, n-1\}$ (note that we have changed the indexing in this section for ease of notation) can be represented as the sum of n cosines and sines at the Fourier frequencies $\omega_k = 2\pi k/n$, $k = 0, 1, \ldots, \lfloor n/2 \rfloor$. The representation is

$$y_t = A_0 + \sum_{0<k<n/2} (A_k cos(\omega_k t) + B_k sin(\omega_k t)) + (-1)^t A_{n/2} \ , \qquad (8.5)$$

where the last term is only included if n is even.

The Fourier coefficients, $A_k, k = 0, \ldots, \lfloor n/2 \rfloor$, and $B_k, k = 1, \ldots, \lfloor n/2 \rfloor - 1$, are identical to those that would be found by least squares regression of y_t on the Fourier sinusoids, $cos(\omega_k t)$ and $sin(\omega_k t)$. If a core process, x_t, is subjected to contamination, then the Fourier transform of the resulting series y_t will also reflect the contamination. Hence, inversion of the Fourier transform will simply return the contaminated data.

Tatum and Hurvich (1993a) propose a robust Fourier transform whose aim is to obtain estimates of the sine and cosine coefficients of the core process x_t that are insensitive to the contaminating series v_t.

The robust regression is based on reducing the influence of large residuals by using *Tukey's biweight psi-function*,

$$\psi(u) = \left\{ \begin{array}{ll} u(1 - u^2)^2 & |u| \leq 1 \\ 0 & |u| > 1 \ . \end{array} \right. \qquad (8.6)$$

The cosine and sine coefficients at frequency ω_k, A'_k and B'_k, are minimizing the function

$$\sum_{t=0}^{n-1} \rho(u_{k,t}) \qquad (8.7)$$

where $\rho'(r) = \psi(r)$ and $u_{k,t} = (y_t - A'_k cos(\omega_k t) - B'_k sin(\omega_k t))/(Cs_k)$. s_k is a scale parameter of $u_{k,t}$, e.g. the median absolute deviation, and C is a tuning constant.

The regression can resist a high proportion of outliers as long as it is supplied with good starting values. Tatum and Hurvich (1993a) used Siegel's repeated median (cf. Siegel, 1982) to supply high breakdown initial values.

When the series is of prime length, the repeated median has an asymptotic breakdown point of 50%, as shown in Tatum and Hurvich (1993b). We will only consider this case in the following. In the case n is not prime, details can be found in Tatum and Hurvich (1993a).

The series is centered by removing a robust location estimate $\tilde{y}$. At each Fourier frequency, a sine and cosine coefficient is separately estimated by robust regression. The sum of the squared coefficients at each frequency gives a robust periodogram. Since the core process is assumed to have a continuous spectrum, the robust periodogram is smoothed using an appropriate lag window to obtain a lag window spectral estimate. Tatum and Hurvich (1993a) suggest to determine the amount of smoothing by the frequency domain version of the corrected AIC proposed by Hurvich and Beltrão (1990).

In the biweight filter the order in which the frequencies will be used is determined by this smoothed periodogram, largest periodogram values first. Using that order, a sinusoid is fitted at each Fourier frequency by robust regression and then swept out of the residuals from the previous step.

The cosine and sine coefficients found by the biweight filter algorithm are then inverse Fourier transformed to give a filtered series, y_t^F, $t = 0, \ldots, n-1$, where

$$y_t^F = \tilde{y} + \sum_{0 < k < n/2} \left(A_k' cos(\omega_k t) + B_k' sin(\omega_k t) \right) \ . \tag{8.8}$$

8.2.2 The Biweight Filter-cleaner

On the basis of the biweight filter, Tatum and Hurvich (1993a) develop a biweight filter-cleaner which has output exactly equal to the input for most values and interpolates the remaining values.

The biweight filter-cleaner initially compares the filtered series to the original and flags discrepant points using the residuals relative to a robust scale estimate of the residuals. The flagged values are replaced by linear interpolation. Finally, the estimated interpolation variance is used to compare the distance between the interpolated points and their original values. When this distance is "relatively small", the original observation is reintroduced. "Relatively small" here means in respect to the estimation (or interpolation)

error of the *filter-cleaned* value. This last step helps to counterbalance the tendency of the first step to flag too many uncontaminated observations.

A data point y_t is flagged, if $|u_t| > 1$. The data $u_t, t = 0, \ldots, n-1$ denote the scaled residuals, i.e.,

$$u_t = (y_t - y_t^F)/(Ks) \ , \tag{8.9}$$

where K is a tuning constant and s is a robust scale estimate of the residuals, e.g., the median absolute deviation.

The linear interpolator is based on finding a vector of weights on the basis of an estimate of the autocovariance function. A preliminary estimate is directly obtained from the filtered data,

$$\widehat{\gamma}_{y_t^F}(h) = \frac{1}{n} \sum_{t=0}^{n-h-1} (y_t^F - \bar{y}_t^F)(y_{t+h}^F - \bar{y}_t^F) \ . \tag{8.10}$$

This autocovariance function is Fourier transformed and the resulting periodogram is again smoothed using an appropriate lag window. The corresponding autocovariance function, $\tilde{\gamma}_{y_t^F}(h)$, is a lag-weighted version of $\widehat{\gamma}_{y_t^F}(h)$.

Weights $a_{t,u}$ are calculated by minimizing the interpolation error at time t, i.e.,

$$\min E \left(\sum_{u \neq t} (a_{t,u} z_u) - z_t \right)^2 \ , \tag{8.11}$$

subject to the constraints $a_{t,u} = 0$ if observation y_u has been flagged. z_t is assumed to be a zero mean stationary process with autocovariance function $\tilde{\gamma}_{y_t^F}(h)$. The constraints ensure that flagged observations do not enter the calculation. Adding the additional constraint $a_{t,t} = -1$ allows (8.11) to be written as

$$\min E \left(\sum_{u=0}^{n-1} a_{t,u} z_u \right)^2 \ , \tag{8.12}$$

or, in vector notation,

$$\min E(\boldsymbol{a}_t^\top \boldsymbol{z} \boldsymbol{z}^\top \boldsymbol{a}_t) = \min (\boldsymbol{a}_t^\top \boldsymbol{C} \boldsymbol{a}_t) \ , \tag{8.13}$$

where $\boldsymbol{C} = E(\boldsymbol{z}\boldsymbol{z}^\top)$, $\boldsymbol{z} = (z_0, \ldots, z_{n-1})^\top$, and $\boldsymbol{a}_t = (a_{t,0}, \ldots, a_{t,n-1})^\top$.

If a total of m observations, say, $\{v_1, \ldots, v_m\}$, have been flagged, the constraints can be written as $\boldsymbol{M}\boldsymbol{a}_{v_j} = \boldsymbol{b}_{v_j}$, where $\boldsymbol{M}$ is an $m \times n$ matrix and $\boldsymbol{b}_{v_j} = (b_{v_j,1}, \ldots, b_{v_j,m})^{\top}$. If y_{v_j} has been flagged, then the j-th row of $\boldsymbol{M}$ will contain 1 in the v_j-th entry and zero in the remaining entries. The vector $\boldsymbol{b}_{v_j}$ is defined $b_{v_j,j} = -1$ and $b_{v_j,i} = 0$ for $i \neq j$.

Minimizing $\boldsymbol{a}_t^{\top}\boldsymbol{C}\boldsymbol{a}_t$ subject to $\boldsymbol{M}\boldsymbol{a}_{v_j} = \boldsymbol{b}_{v_j}$ has an explicit solution as long as $\boldsymbol{C}$ is positive definite and $\boldsymbol{M}$ is of full rank. Using an appropriate lag window guarantees that $\boldsymbol{C}$ is positive definite. The matrix $\boldsymbol{M}$ is clearly of full rank. Hence, the solution is given by

$$\boldsymbol{a}_t = \boldsymbol{C}^{-1}\boldsymbol{M}^{\top}(\boldsymbol{M}\boldsymbol{C}^{-1}\boldsymbol{M}^{\top})^{-1}\boldsymbol{b}_t \ . \tag{8.14}$$

The interpolated value at time t is

$$\widehat{y}_t = \sum_{u \neq t} a_{t,u}(y_u - \bar{y}^F) + \bar{y}^F \tag{8.15}$$

with an estimated interpolation variance given by

$$\widehat{\sigma}_t^2 = \sum_{u,v} a_{t,u} a_{t,v} \tilde{\gamma}_{y_t^F}(|u - v|) = \boldsymbol{a}_t^{\top}\boldsymbol{C}\boldsymbol{a}_t \ . \tag{8.16}$$

The studentized distance, d_t, between the original and the interpolated data point at time t, is $d_t = (y_t - \widehat{y}_t)/\widehat{\sigma}_t$. Now a cleaned data set, y_t^C, is constructed as follows,

$$y_t^C = \begin{cases} y_t & : \ |d_t| < a \\ \alpha_t y_t + (1 - \alpha_t)\widehat{y}_t & : \ a \leq |d_t| < b \\ \widehat{y}_t & : \ |d_t| \geq b \ , \end{cases} \tag{8.17}$$

where $\alpha_t = (b - |d_t|)/(b - a)$ with appropriate values a and b.

Chapter 9

Conclusions

We presented several approaches to robustly estimate the spectral density function.

It is well known that the periodogram is a naive estimator of the spectral density function. Moreover, it is the basis of nonparametric spectral density estimation. Namely, at a fixed frequency the periodogram is the squared modulus of the discrete Fourier transform of the observed process.

Hence, we first tried to get robust spectral density estimates by directly using a robustified version of the Fourier transformation. The classical Fourier transformation can be interpreted as taking the mean of 2-dimensional observations. In order to obtain a robust estimate of the spectral density function, we simply replaced the mean by a robust multivariate measure of location. We compared estimates obtained by two methods, one based on the trimmed mean, applied coordinate-wise, and the other using Rousseeuw's MCD estimator.

It is also well known that the spectral density function is the Fourier transform of the autocovariance function.

Therefore we next used a spectral density estimator originally proposed by Blackman and Tukey (1958) which is exactly based on the Fourier transform of the autocovariance function. In order to robustly estimate the spectral density function, we replaced the autocovariance function by a robust estimator. We compared estimates obtained by a highly robust autocovariance function (Ma and Genton, 2000) to others based on Spearman's rank correlation coefficient (Ahdesmäki et al., 2005) or on partial autocorrelation (Möttönen et al., 1999).

The problem that all these robust alternatives yield estimates that may not necessarily be positive semidefinite was solved by applying nonlinear shrinking to ensure non-negative definiteness.

Unfortunately, all above mentioned procedures to robustly estimate the spectral density function have not turned out to be satisfactory.

However, in order to get a robust estimate of the spectral density function, that is insensitive to outlying observations, it turned out that cleaning the time series in a robust way first and calculating the spectral density function afterwards leads to encouraging results.

This procedure was proposed by Martin and Thomson (1982). The data-cleaning operation wherein the robustness is introduced is accomplished by a robustified version of the Kalman filter. We compared two different versions, namely the ACM-type filter (Martin, 1979), which was originally used by Martin and Thomson (1982), and a newer alternative, the rLS filter proposed by Ruckdeschel (2001). Although both procedures yield good spectral density estimates, the results of the simulation study suggest that the ACM-type filter algorithm performs slightly better than the rLS filter algorithm.

The additional problem of estimating the hyper parameters is best solved by bounded-influence autoregression.

Moreover, while the original ACM-type filter is bound to the univariate setting, we proposed a generalized ACM-type filter for multivariate observations (cf. Spangl and Dutter, 2008).

However, there exist further robust filter and data-cleaning procedures. In Spangl and Dutter (2005) we compared the ACM-type filter approach with another approach proposed by Tatum and Hurvich (1993a). This procedure, called biweight filter-cleaner also yields good results, but tends to underestimate the core process slightly. Moreover it is computationally intensive.

Furthermore, a whole bundle of robust time series filters have already been implemented in R and are available in the R-package `robfilter` (Fried and Schettlinger, 2008). However, these filters are specialized to reveal trends, trend changes or shifts of an underlying, possibly nonstationary signal in the presence of outliers and, according to our experience, extremely smooth the underlying core process. Hence, these filters are not applicable if one wants to estimate the spectral density function.

A completely different approach is a robustified version of Welch's Over-lapped Segment Averaging (WOSA) proposed by Chave et al. (1987). Although widely used in geophysical applications, it will only yield a good

spectral density estimate if a small fraction of the data segments are contaminated by outliers. Even for a small portion of outlying observations, if we assume that the outliers are randomly distributed within the series and are not clustered, most of the segments will contain at least one contaminated data point, and therefore this method will yield an inflated and biased estimate of the spectral density function. The same concerns have already been made by Martin and Thomson (1982) discussing a similar procedure.

We note that cleaning the time series in a robust way first and calculating the spectral density function afterwards has the additional advantage that outliers can be easily spotted. Moreover, the resulting spectral density estimate based on the ACM-type filter, which is according to our simulation studies the most competitive method, can even be improved by using an ACM-type smoother instead (cf. Martin and Thomson, 1982). Once all filter estimates have been calculated, computation of the smoother estimates is easily done by some kind of backward prediction. For details see Martin (1979) and Martin and Thomson (1982).

Although our application, the analysis of heart rate variability data, described in the next chapter is solely retrospective, one may be interested in online (bio)signal processing in future research projects. In Martin and Thomson (1982) the authors made a proposal to compute robust spectral density estimates in real time, which is based on filtering. To be able to cover this important topic we therefore focused our research mainly on filtering.

Part II

Applications

Chapter 10

Analysis of Heart Rate Variability Data

10.1 Introduction

Our present work has been motivated by the frequency-domain analysis of short-term heart rate variability (HRV) recordings. This is a non-invasive method which has been increasingly used in medicine (cf. Task Force of The European Society of Cardiology and The North American Society of Pacing and Electrophysiology, 1996; Howorka et al., 1997, 1998; Hartikainen et al., 1998; Pumprla et al., 2002).

To access the variability of heart rate in the frequency domain the spectral density function of the tachogram is estimated. The tachogram is the series of time intervals between consecutive heart beats. These time intervals are also called R-R-intervals, i.e., the periods between an R-peak and the next R-peak in an electrocardiogram (cf. Figure 10.1). The intervals normally have a duration of about $750\,\mathrm{ms}$ corresponding to a heart rate of 80 beats per minute.

In the tachogram (an example is displayed in Figure 10.2), outlying observations can be caused by ventricular ectopic beats and other artifacts (cf. Hartikainen et al., 1998). Ectopic beats are usually premature and produce a very short R-R-interval followed by a compensatory delay and therefore a prolonged R-R-interval. Typical tachogram patterns caused by ectopic beats can be seen in Figure 10.2 around heart beat number 90 and 1090. Correspondingly, missed beats result in erroneously prolonged R-R-intervals (sum
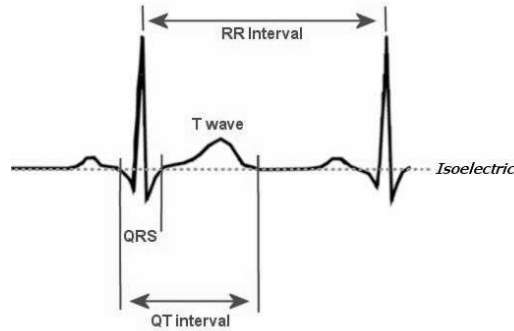
Figure 10.1: Ideal electrocardiogram signal

of two consecutive R-R-intervals). Typical patterns caused by missed beats are visible in Figure 10.2 around beat number 730.

These outlying tachogram measurements affect the spectral analysis of heart rate variability if we use classical spectral density estimators which are sensitive to outliers (cf. also Section 2.5).

We therefore aim to access the heart rate variability by estimating the spectral density function of the tachogram series using robust methods that are insensitive to outlying tachogram values caused by ectopic beats or other artifacts.

Furthermore, as ectopic or missing beats do not affect successive heart beats, the additive outlier model (cf. Section 2.4) seems to be an appropriate model when analyzing heart rate variability data.

We do not compute the spectral density function of the entire tachogram, but calculate several estimates within overlapping windows (cf. Pumprla et al., 2002). This is to ensure stationarity in each window and to deal with signals whose frequency content changes over time. The result of the so called dynamic Fourier analysis applied to the tachogram series plotted in Figure 10.2 is displayed in Figure 10.3. Each slice parallel to the frequency-spectrum plane in Figure 10.3 represents the spectral density estimate of the corresponding time window.

A high variability in heart rate indicates good adaptability, implying a healthy person with well functioning autonomic control mechanisms. Conversely, lower variability is often an indicator of abnormal and insufficient adaptability of the autonomic nervous system.

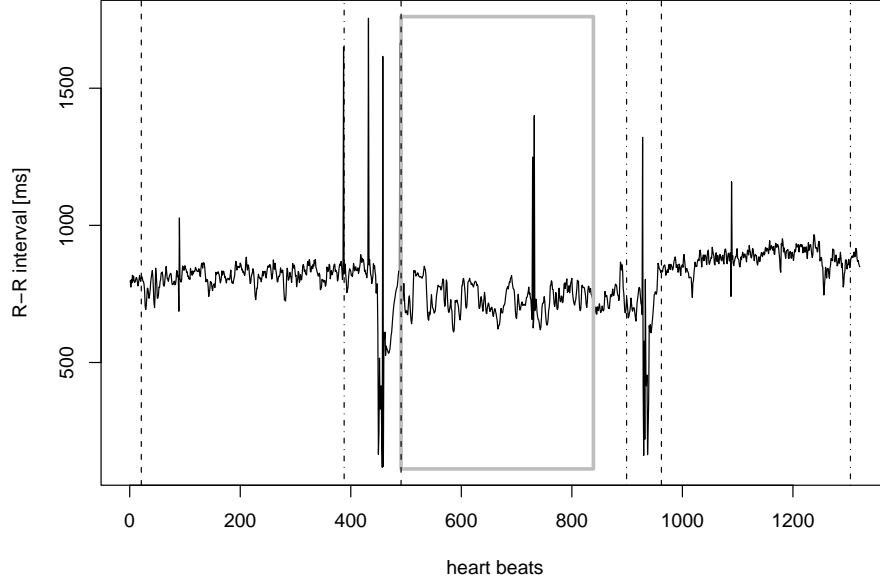Hence, we do not use the entire tachogram series but several overlapping

Figure 10.2: Tachogram of 1321 consecutive heart beats

windows to access the heart rate variability and only focus on an analysis in the frequency domain. We are neither interested in modeling the heart rate in the time domain nor in forecasting as this is often the aim in the context of online-monitoring.

In the following we consider the problem of robust spectral analysis of short-term HRV data. To obtain a robust estimate of the spectral density function we suggest to use the multi-step procedure described in Chapter 7. This procedure was proposed by Martin and Thomson (1982) and incorporates an important robust filtering operation which is accomplished by an approximate conditional-mean (ACM) type filter. According to our research, this is the most competitive method for estimating the spectral density function robustly.
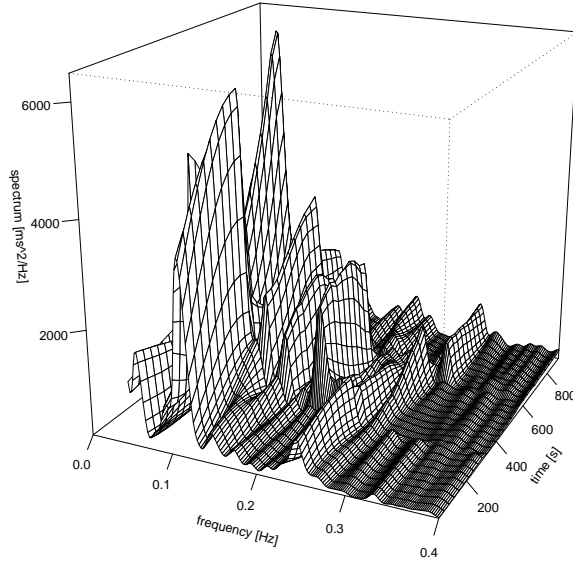
Figure 10.3: Robust dynamic Fourier analysis of the original short-term HRV data displayed in Figure 10.2

10.2   Methods

All analyzed tachogram series come from diabetic patients with different degrees of cardiovascular autonomic neuropathy (CAN) and were provided by J. Pumprla and K. Howorka, Department of Biomedical Engineering and Physics, Medical University of Vienna.

HRV is composed of certain well-defined rhythms which contain information about the contribution of different regulatory mechanisms of cardiovascular control.

In short-term HRV recordings three spectral components can be normally distinguished: high frequency (HF, 0.15-0.4 Hz), low frequency (LF, 0.04-0.15 Hz), and very low frequency (VLF, 0-0.04 Hz) components. The HF component represents parasympathetic activity whereas the sympathetic nervous system is the main contributor of the LF component (cf. Hartikainen et al., 1998). The physiological explanation of the VLF component has not

been well defined. Additionally, it represents very low frequency oscillations that need a relatively long recording (about 1 hour) to be assessed reliably. Thus, the VLF component is recommended to be discarded when interpreting short-term HRV measurements (cf. also Task Force of The European Society of Cardiology and The North American Society of Pacing and Electrophysiology, 1996).

The analysis of heart rate variability as proposed in the review article by Pumprla et al. (2002) is in fact a combination of three short-term HRV recordings. The duration of each recording lasts 5 min as recommended by the Task Force of The European Society of Cardiology and The North American Society of Pacing and Electrophysiology (1996). Moreover, the proposed modified orthostatic test where the individual lies supine for 5 min, stands for 5 min and lies supine again for another 5 min is also recommended when investigating patients with cardiovascular autonomic neuropathy in order to separate sympathetic from parasympathetic abnormalities.

An advantage of short-term HRV recordings is that due to the restricted duration they can easily be performed in strictly standardized conditions to ensure stationarity of the tachogram signal. It is therefore also essential, that after a stimulus (e.g. standing) the HRV analysis should not be performed from a period that includes the early change in heart rate in response to the stimulus, but from the moment after the heart rate has stabilised to a new level. In addition, when the patient assumes the supine position again the heart rate starts to decline and results in a positive trend and nonstationarity of the tachogram recording. This period should also be omitted and the analysis should be started again as soon as the heart rate has stabilised (cf. Hartikainen et al., 1998).

As the tachogram recording is a discrete event series it is an irregularly time-sampled signal (cf. also Drews, 1983). To obtain a regularly sampled series we interpolate the original tachogram recording using cubic splines and re-sample at equidistant points in time. The re-sampling frequency has to be sufficiently high so that the Nyquist frequency of the spectral density function is not within the frequency range of interest. We therefore choose a re-sampling period of 0.25 seconds.

As mentioned before, to ensure stationarity of the signal, the duration of the recording should not be extended too much. On the other hand, for frequency-domain analysis of HRV it is recommended that the signal is sufficiently long. Hence, to assess the HF component a recording of approximately 1-minute duration is required. Correspondingly, a 4-minute recording

is needed for the assessment of the LF component (cf. Hartikainen et al., 1998).

The outline of our dynamic Fourier analysis is as follows: We calculate a robust prewhitened spectral density estimate of the interpolated tachogram recording every 5 seconds for a time window with a duration of 256 seconds using the algorithm proposed by Martin and Thomson (1982) and described in detail in Chapter 7. For each window the tachogram series is cleaned in a robust way by using an approximate conditional-mean type filter first. The hyper parameters of the approximating autoregressive process of order 5 are estimated robustly by bounded-influence autoregression (cf. Section 5.1). According to the order-selection rule described in Section 5.1.5 an order of 5 seems to be sufficient.

Then, we calculate the spectral density function using a prewhitened spectral density estimator. The lag window spectral estimate of the prediction residuals therein is obtained by using a 0th-order discrete prolate spheroidal sequence as data taper and a Parzen window for smoothing.

Finally we display the results three dimensionally where we only plot the frequency range of interest, i.e., the LF and HF components. Moreover, we note that it is recommended in medicine to display the spectral density estimates on a metric and not, as usually, on a logarithmic scale.

10.3   Results

In Figure 10.4 we present intermediate results of our robust dynamic Fourier analysis described in the previous section. To show how the suggested robust multi-step procedure works applied to the HRV data, we take one single time window of the data displayed in Figure 10.2. The chosen example is the first 256-second window of the 5-minute standing period indicated by the grey frame. Each of the three 5-minute periods, that correspond to the supine position, standing, and the supine position again, is indicated in Figure 10.2 by a vertical dashed line at the beginning and a dot-dashed one at the end.

Plot (a) of Figure 10.4 shows the original tachogram recording (black line) along with the cubic spline interpolation (light grey line). As obviously seen the two are almost identical.

In Plot (b) the interpolation result is displayed along with the robust filter estimate obtained by the approximate conditional-mean type filter. The filter estimate (dashed grey line) is equivalent to the interpolated tachogram series
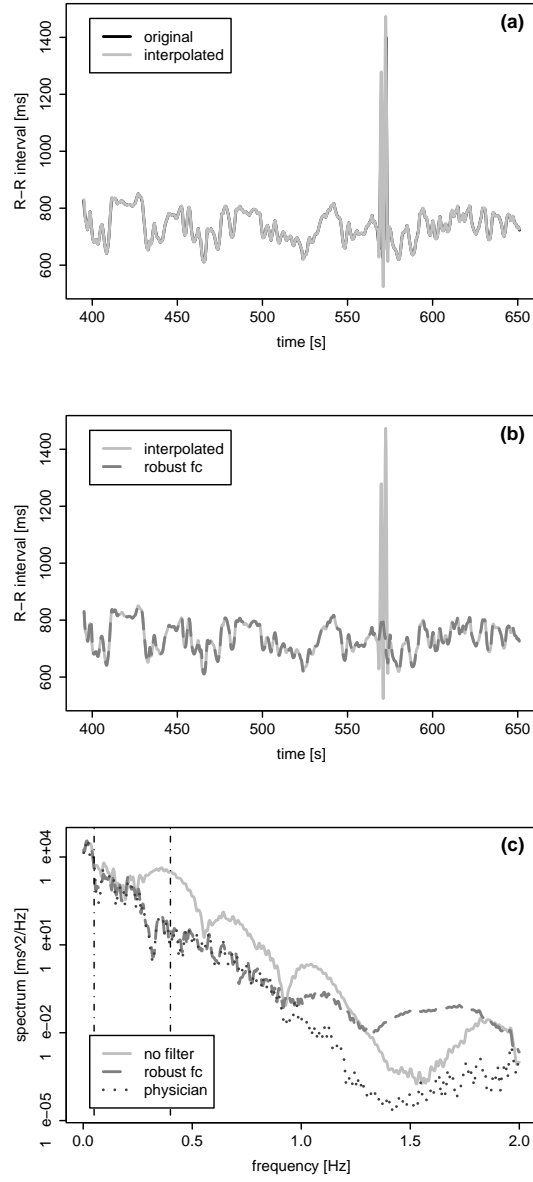
Figure 10.4: Intermediate results of the suggested robust spectral analysis applied to the short-term HRV data displayed in Figure 10.2
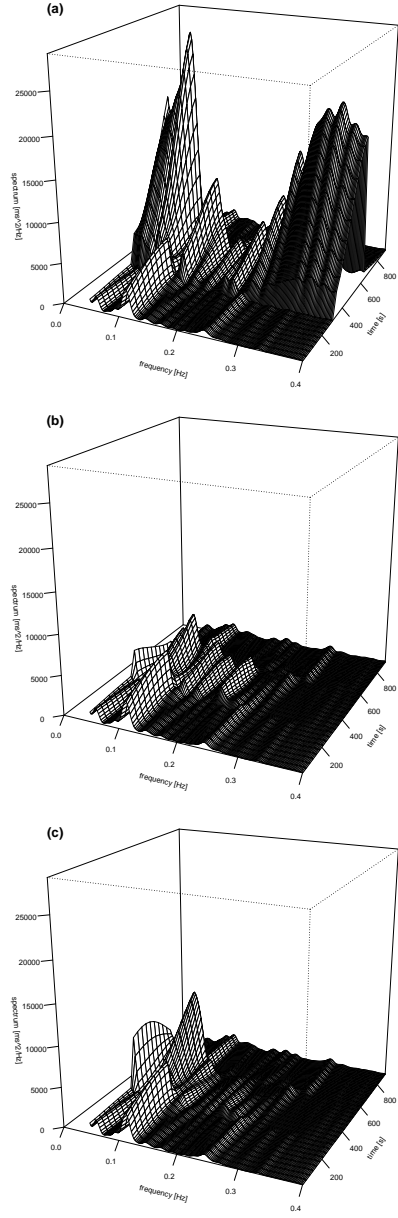
Figure 10.5: Dynamic Fourier analysis of the original short-term HRV data displayed in Figure 10.2

103

(light grey line) in cases where no outliers are present. Additionally, it is not affected by outlying observations around 570 seconds that are caused by missed heart beats.

Plot (c) shows several spectral density estimates of the HRV data. The prewhitened spectral density estimate of the robustly filtered tachogram series (dashed grey line) is similar in shape and power to the lag window spectral estimate of the tachogram series that was manually cleaned by the physician and will be considered as benchmark in the following (dotted dark grey line). This correspondence is extremely well within the frequency range of interest, i.e., between 0.04 and 0.4 Hz, indicated by the two vertical dot-dashed lines. Moreover, the lag window spectral estimate of the original tachogram series (light grey line) is markedly affected by outlying observations. We further note that, as we have chosen a re-sampling period of 0.25 seconds, the Nyquist frequency in this case is equal to 2 Hz.

The final result of the dynamic Fourier analysis is displayed in Figure 10.5. Plot (a) shows the classical non-robust lag window spectral estimates of the original tachogram series. In Plot (b) the result of robust dynamic Fourier analysis is displayed whereas the result in Plot (c) is obtained by using the same estimator as in Plot (a) but now applied to the manually filter tachogram series. As before (cf. Figure 10.4, Plot (c)), the results in Plot (b) and (c) of Figure 10.5 are very similar. However, we see that the spectral density estimates in Plot (a) are markedly affected in shape and power by outlying observations. Moreover, we note that Figure 10.3 is equal to Plot (b), but for the latter we use the same scaling on the vertical axis as in Plot (a) to be able to compare them.

We now present the results of the spectral analysis of short-term HRV data of diabetic patients with different degrees of cardiovascular autonomic neuropathy (CAN).

In Figure 10.6 the original tachogram recordings of three diabetic patients with (a) no CAN, (b) moderate CAN, and (c) severe CAN are displayed. In each plot the three 5-minute periods of the modified orthostatic test are again indicated by the dashed and dot-dashed vertical lines.

In Figure 10.7 the final results of the robust dynamic Fourier analysis are plotted on the same vertical scale. Plot (a) (no CAN) shows a clear predominance of parasympathetic activity associated with higher values of the spectral density function in the HF component (around 0.25 Hz) during both supine positions. During standing mostly sympathetic autonomic activity
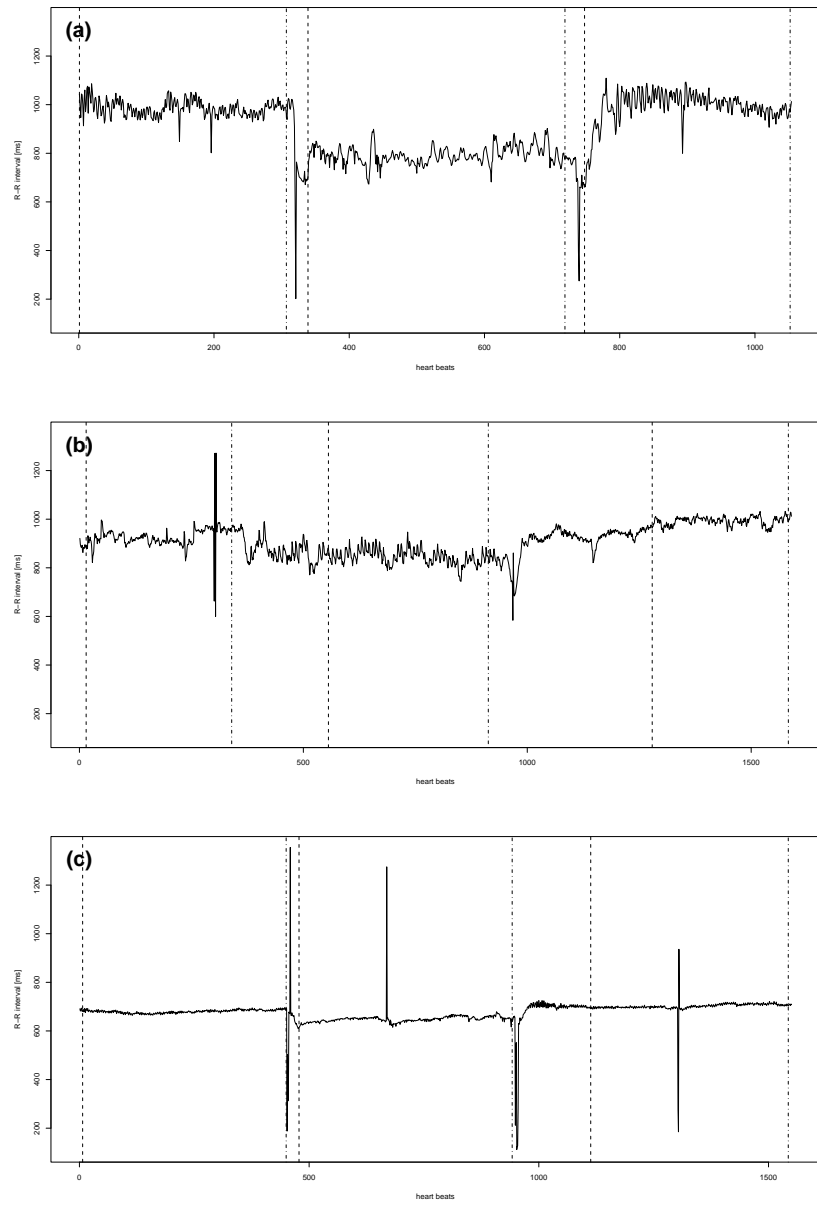
Figure 10.6: Tachogram of diabetic patients with different degrees of CAN, (a) no CAN, (b) moderate CAN, and (c) severe CAN
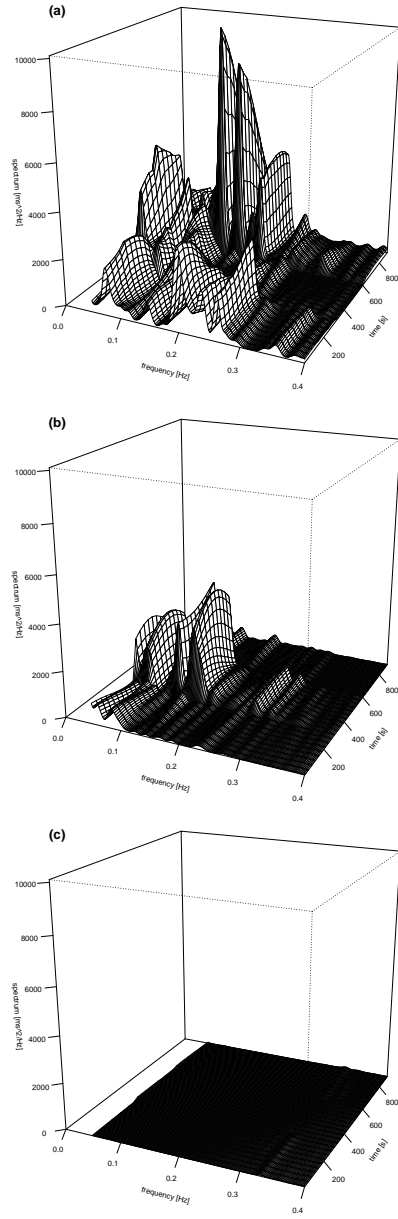
Figure 10.7: Dynamic Fourier analysis of the short-term HRV data displayed in Figure 10.6, (a) no CAN, (b) moderate CAN, and (c) severe CAN

106

predominates associated with a decrease of power in the HF component and an increase in the LF component (around 0.1 Hz).

In early stages of cardiovascular autonomic neuropathy (Figure 10.7, Plot (b)) parasympathetic activity is reduced, whereas sympathetic autonomic activity is preserved in the LF component.

In patients with severe cardiovascular autonomic neuropathy (Figure 10.7, Plot (c)) activity is reduced throughout the whole frequency range (HF and LF components) with only minimal reaction in heart rate variability to the interventions of the orthostatic test.

In addition to the visual representation, the amount of heart rate variability of each frequency component may also be numerically expressed as power, i.e., the integral of the spectral density function within the frequency bounds of the corresponding component (cf. Hartikainen et al., 1998; Pumprla et al., 2002). To capture the amount of variablity in case of the above used spectral analysis of short-term HRV measurements, additional averaging of the power of consecutive spectral density functions is needed. In Howorka et al. (1998) ranges for the spectral power of each frequency component have been established to allow a rapid classification of the severity of cardiovasular autonomic neuropathy.

We further note that similar results have already been published in Spangl and Dutter (2005, 2007).

10.4 Discussion

The measurement of heart rate variability provides a useful tool for assessing the status of the cardiac autonomic regulation. Usually, HRV analysis is based on measuring the duration of R-R-intervals of consecutive heart beats.

In this chapter we focused on the spectral analysis of short-term HRV data. As mentioned, to assess heart rate variability in the frequency domain, the spectral density function of the corresponding tachogram series has to be estimated. This generates a visual representation of the heart rate variability that is easily understood. However, in literature, it is recommended that the electrocardiogram is reviewed by a physician or experienced operator to remove or edit outlying observations, such as ectopic beats or other artifacts, in order not to introduce an error into the HRV analysis (cf., for example, Hartikainen et al., 1998; Pumprla et al., 2002). This is due to the fact that classical spectral density estimators are sensitive to outliers.

Hence, to obtain a robust estimate of the spectral density function we suggest to use a multi-step procedure based on robust filtering instead. This procedure is insensitive to outlying observations, and therefore provides fully automated signal processing which will facilitate reliable and reproducible HRV analysis with minimal operator input. Neither removing nor editing of outliers needs to be done, as they are appropriately downweighted by the proposed robust method. Furthermore, the suggested procedure, if slightly adapted, may also be used for online data processing.

However, we want to remark that our proposed method should not replace manual reviewing by an expert but may help to increase clinical accessibility. Moreover, it can actually be used to identify and mark outlying observations.

Additionally, according to our previous research, the described multi-step procedure seems to be the best method for estimating the spectral density function robustly.

Part III

Appendix

Appendix A

Time Series Analysis

A.1 Autocorrelation and Cross-correlation Function

The *mean value function* of a univariate time series $\{x_t,\ t = 1, \ldots, n\}$ is defined as

$$\mu_t = E(x_t) \ . \tag{A.1}$$

The lack of independence between two adjacent values x_s and x_t of the same time series can be assessed numerically, as in classic statistics, using the notion of covariance. Hence, the *autocovariance function* is defined as the second moment product

$$\gamma_x(s, t) = E((x_s - \mu_s)(x_t - \mu_t)) \ , \ \text{for all } s \text{ and } t. \tag{A.2}$$

The autocovariance function measures the linear dependence between two points of the same series observed at different times.

The preceding definitions of the mean and autocovariance functions are completely general. However, in order to do some statistical inference, we have to assume that a sort of regularity may exist over time in the behavior of a time series. Therefore we introduce the concept of stationarity. A *weakly stationary time series* has to fulfill the following properties: (i) the mean value function μ_t in (A.1) is constant, i.e., it does not depend on time t, and (ii) the autocovariance function $\gamma_x(s, t)$ in (A.2) depends on s and t only by their difference $h = s - t$, where h is called the *lag*. Now we are able to define the mean value and autocovariance function of a weakly stationary

time series as

$$E(x_t) = \mu_x \tag{A.3}$$

and

$$\gamma_x(h) = E((x_{t+h} - \mu_x)(x_t - \mu_x)) \ , \tag{A.4}$$

where, for convenience, we write $\gamma_x(h)$ instead of $\gamma_x(t+h,t)$. We note that the autocovariance function satisfies $\gamma_x(h) = \gamma_x(-h)$. The *autocorrelation function (ACF)* of a stationary time series can be written as

$$\rho_x(h) = \frac{\gamma_x(t+h,t)}{\sqrt{\gamma_x(t+h,t+h)\gamma_x(t,t)}} = \frac{\gamma_x(h)}{\gamma_x(0)} \ . \tag{A.5}$$

When several stationary time series are available, say, x_t and y_t, we often would like to measure the predictability of the series y_t from the series x_t, leading to the notion of the *cross-covariance function* of stationary time series,

$$\gamma_{xy}(h) = E((x_{t+h} - \mu_x)(y_t - \mu_y)) \ . \tag{A.6}$$

We note that the cross-covariance function satisfies $\gamma_{xy}(h) = \gamma_{yx}(-h)$. The scaled version of the above, called *cross-correlation function (CCF)* of stationary time series, is defined as

$$\rho_{xy}(h) = \frac{\gamma_{xy}(h)}{\sqrt{\gamma_x(0)\gamma_y(0)}} \ . \tag{A.7}$$

Estimation

Assuming stationarity we are able to estimate the mean value function (A.3) if it is constant by replacing the average over the population, denoted by E, with an average over the sample, say,

$$\bar{x} = \frac{1}{n} \sum_{t=1}^{n} x_t \tag{A.8}$$

and the theoretical autocovariance (A.4) by the *sample autocovariance function*

$$\widehat{\gamma}_x(h) = \frac{1}{n} \sum_{t=1}^{n-h} (x_{t+h} - \bar{x})(x_t - \bar{x}) \ , \quad h = 0, \ldots, n-1 \ , \tag{A.9}$$

with $\widehat{\gamma}_x(-h) = \widehat{\gamma}_x(h)$. The estimator in (A.9) is generally preferred to the one that would be obtained by dividing by $n - h$ because (A.9) is a non-negative definite function. We note that neither dividing by n nor $n - h$ in (A.9) yields an unbiased estimator of $\gamma_x(h)$ (cf. Shumway and Stoffer, 2000; Deistler and Scherrer, 1994).

The *sample autocorrelation function* of a stationary time series is defined, analogously to (A.5), as

$$\widehat{\rho}_x(h) = \frac{\widehat{\gamma}_x(h)}{\widehat{\gamma}_x(0)} \ . \tag{A.10}$$

The estimators for the cross-covariance function $\gamma_{xy}(h)$ as given in (A.6) and the cross-correlation function $\rho_{xy}(h)$ in (A.7) are given by the *sample cross-covariance function* of stationary time series

$$\widehat{\gamma}_{xy}(h) = \frac{1}{n} \sum_{t=1}^{n-h} (x_{t+h} - \bar{x})(y_t - \bar{y}) \ , \quad h = 0, \dots, n - 1 \ , \tag{A.11}$$

where $\widehat{\gamma}_{xy}(-h) = \widehat{\gamma}_{yx}(h)$ determines the function for negative lags, and the *sample cross-correlation function*

$$\widehat{\rho}_{xy}(h) = \frac{\widehat{\gamma}_{xy}(h)}{\sqrt{\widehat{\gamma}_x(0)\widehat{\gamma}_y(0)}} \ . \tag{A.12}$$

A.2 Partial Autocorrelation Function

Formally, for a stationary time series $\{x_t, t = 1, 2, \dots\}$ we define the *partial autocorrelation function (PACF)* ϕ_{hh}, $h = 1, 2, \dots$, by

$$
\begin{aligned}
\phi_{00} &= 1 = \rho_x(0) \\
\phi_{11} &= \text{corr}(x_t, x_{t-1}) = \rho_x(1) \\
\phi_{hh} &= \text{corr}(x_t - x_{t|t-(h-1):t-1}, x_{t-h} - x_{t-h|t-(h-1):t-1}) \ , \quad h \geq 2,
\end{aligned} \tag{A.13}
$$

where $x_{s|m:n}$ denotes the linear minimum mean squared error predictor of x_s based on $\{x_m, \dots, x_n\}$. To ease notation, we will just write $x_{s|n}$ if $m = 1$, e.g., $x_{t|t-1}$ in the case of the one-step-ahead predictor of x_t based on all previous observations. Hence, $x_{t|t-(h-1):t-1}$ can be thought of as the linear regression of x_t on the past, $x_{t-1}, \dots, x_{t-(h-1)}$, and $x_{t-h|t-(h-1):t-1}$ can be thought of as the linear regression of x_{t-h} on the future, $x_{t-(h-1)}, \dots, x_{t-1}$. We will

go into detail later, for now we only note that both $(x_t - x_{t|t-(h-1):t-1})$ and $(x_{t-h} - x_{t-h|t-(h-1):t-1})$ are uncorrelated with $\{x_{t-(h-1)}, \ldots, x_{t-1}\}$. Because of stationarity, the PACF ϕ_{hh} is the correlation between x_t and x_{t-h} with the linear effect of $\{x_{t-(h-1)}, \ldots, x_{t-1}\}$ on each removed.

Moreover we can interpret the PACF in the following way. To ease notation we assume that x_t is stationary with zero mean. Let

$$\nu_{ht} = x_t - \sum_{j=1}^{h-1} a_j x_{t-j} \tag{A.14}$$

and

$$\omega_{h,t-h} = x_{t-h} - \sum_{k=1}^{h-1} b_k x_{t-(h-k)} \tag{A.15}$$

be the two residuals where $\{a_1, \ldots, a_{h-1}\}$ and $\{b_1, \ldots, b_{h-1}\}$ are chosen so that they minimize the mean squared errors

$$E(\nu_{ht}^2) \text{ and } E(\omega_{h,t-h}^2) \ .$$

Then the PACF at lag h can be defined as the cross-correlation between ν_{ht} and $\omega_{h,t-h}$, i.e.,

$$\phi_{hh} = \frac{E(\nu_{ht}\omega_{h,t-h})}{\sqrt{E(\nu_{ht}^2)E(\omega_{h,t-h}^2)}} \ . \tag{A.16}$$

We note that the coefficients in (A.14) and (A.15) are the same, i.e., $a_i = b_i$, which means that, for stationary processes, linear prediction forward in time is equivalent to linear prediction backward in time.

Partial Autocorrelation Function and Best Linear Predictor

Let us suppose that x_t is a stationary time series and we want to predict future values x_{n+m}, $m = 1, 2, \ldots$, based on given data $\{x_{n-(h-1)}, \ldots, x_n\}$. Generally the *minimum mean squared error predictor* of x_{n+m} is $x_{n+m|n-(h-1):n} = E(x_{n+m}|x_n, \ldots, x_{n-(h-1)})$ because the conditional expectation minimizes the mean squared error

$$E(x_{n+m} - g(x_{n-(h-1)}, \ldots, x_n))^2 \ , \tag{A.17}$$

where $g(x_{n-(h-1)}, \ldots, x_n)$ denotes a measurable function of the observations $x_{n-(h-1)}, \ldots, x_n$. Linear predictors of the form

$$x_{n+m|n-(h-1):n} = a_0 + \sum_{k=1}^{h} a_k x_{n-(k-1)} , \qquad (A.18)$$

that minimize the mean squared error (A.17) are called *best linear predictors (BLP)*.

For a stationary process we have the following property. Given data $x_{n-(h-1)}, \ldots, x_n$, the BLP, $x_{n+m|n-(h-1):n} = a_0 + \sum_{k=1}^{h} a_k x_{n-(k-1)}$, of x_{n+m}, for $m \geq 1$, is found by solving

$$\begin{aligned} E(x_{n+m} - x_{n+m|n-(h-1):n}) &= 0 \\ E((x_{n+m} - x_{n+m|n-(h-1):n})x_{n-(k-1)}) &= 0 , \quad k = 1, \ldots, h . \end{aligned} \qquad (A.19)$$

The equations specified in (A.19) are called the *prediction equations*.

Now we consider *one-step-ahead prediction* and assume, to ease notation, that $E(x_t) = 0$, which means $a_0 = 0$. The BLP of x_{n+1}, given $x_{n-(h-1)}, \ldots, x_n$, is

$$x_{n+1|n-(h-1):n} = \phi_{h1}x_n + \phi_{h2}x_{n-1} + \cdots + \phi_{hh}x_{n-(h-1)} , \qquad (A.20)$$

where we have written a_k in (A.18) as ϕ_{hk} in (A.20), for $k = 1, \ldots, h$. Using the prediction equations, the coefficients $\phi_{h1}, \ldots, \phi_{hh}$ satisfy

$$E((x_{n+1} - \sum_{j=1}^{h} \phi_{hj}x_{n+1-j})x_{n+1-k}) = 0 , \quad k = 1, \ldots, h ,$$

or

$$\sum_{j=1}^{h} \phi_{hj}\gamma_x(k-j) = \gamma_x(k) , \quad k = 1, \ldots, h . \qquad (A.21)$$

The prediction equation (A.21) can be written in matrix notation as

$$\mathbf{\Gamma}_h \boldsymbol{\phi}_h = \boldsymbol{\gamma}_h , \qquad (A.22)$$

where $\mathbf{\Gamma}_h = (\gamma_x(k-j))_{j,k=1}^{h}$ is an $h \times h$ matrix, $\boldsymbol{\phi}_h = (\phi_{h1}, \ldots, \phi_{hh})^\top$ is an $h \times 1$ vector, and $\boldsymbol{\gamma}_h = (\gamma_x(1), \ldots, \gamma_x(h))^\top$ is an $h \times 1$ vector.

Furthermore, it can be proved that the following equation holds:

$$\phi_{hh} = \frac{\rho_x(h) - \widetilde{\boldsymbol{\rho}}_{h-1}^{\top} \boldsymbol{R}_{h-1}^{-1} \boldsymbol{\rho}_{h-1}}{1 - \widetilde{\boldsymbol{\rho}}_{h-1}^{\top} \boldsymbol{R}_{h-1}^{-1} \widetilde{\boldsymbol{\rho}}_{h-1}} = a_h \ , \tag{A.23}$$

where $\boldsymbol{\rho}_{h-1} = (\rho_x(1), \ldots, \rho_x(h-1))^{\top}$, $\widetilde{\boldsymbol{\rho}}_{h-1} = (\rho_x(h-1), \ldots, \rho_x(1))^{\top}$ and $\boldsymbol{R}_h$ is the $h \times h$ matrix with elements $\rho_x(i-j)$, $i, j = 1, \ldots, h$ (cf. also Shumway and Stoffer, 2000).

A.3   Autoregressive Moving Average Models

Now we proceed with the general definition of *autoregressive moving average (ARMA) models* for stationary time series. As before, to ease notation, we assume that the time series x_t has zero mean. A univariate time series $\{x_t, t = 0, \pm 1, \pm 2, \ldots\}$ is said to be ARMA(p, q) if x_t is stationary and

$$x_t = \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} \ , \tag{A.24}$$

with $\phi_p \neq 0$ and $\theta_q \neq 0$. In (A.24) ε_t denotes a white noise process with $\sigma_w^2 > 0$. The parameters p and q are called the autoregressive and moving average orders, respectively. If $q = 0$ the model is called an autoregressive model of order p, AR(p), and if $p = 0$ the model is called a moving average model of order q, MA(q).

Additionally we also require in (A.24) that $\phi(z)$ and $\theta(z)$ have no common factors, where the *AR* and *MA polynomials*, $\phi(z)$ and $\theta(z)$, are defined as

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p \ , \quad \phi_p \neq 0 \ , \tag{A.25}$$

and

$$\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q \ , \quad \theta_q \neq 0 \ , \tag{A.26}$$

respectively, where z is a complex number.

Using the *backshift operator* B, defined as

$$B x_t = x_{t-1} \ , \tag{A.27}$$

we can write the ARMA(p, q) model in (A.24) as

$$\phi(B) x_t = \theta(B) \varepsilon_t \ . \tag{A.28}$$

115

Further in order to obtain models that do not depend on the future and are unique, we will require some additional restrictions on the model parameters in (A.24). An ARMA(p, q) model, $\phi(B)x_t = \theta(B)\varepsilon_t$, is said to be *causal*, if the time series $\{x_t, t = 0, \pm 1, \pm 2, \dots\}$ can be written as a one-sided linear process:

$$x_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j} = \psi(B)\varepsilon_t \ , \tag{A.29}$$

where $\psi(B) = \sum_{j=0}^{\infty} \psi_j B^j$ and $\sum_{j=0}^{\infty} |\psi_j| < \infty$; we set $\psi_0 = 1$.

Moreover an ARMA(p, q) model, $\phi(B)x_t = \theta(B)\varepsilon_t$, is said to be *invertible*, if the time series $\{x_t, t = 0, \pm 1, \pm 2, \dots\}$ can be written as

$$\pi(B)x_t = \sum_{j=0}^{\infty} \pi_j x_{t-j} = \varepsilon_t \ , \tag{A.30}$$

where $\pi(B) = \sum_{j=0}^{\infty} \pi_j B^j$ and $\sum_{j=0}^{\infty} |\pi_j| < \infty$; π_0 is set equal to one.

In general we have the following property. An ARMA(p, q) model is causal only if the roots of $\phi(z)$ lie outside the unit circle, i.e., $\phi(z) = 0$ only if $|z| > 1$. Analogously an ARMA(p, q) model is invertible only if the roots of $\theta(z)$ lie outside the unit circle.

Appendix B

The Spectral Representation Theorem

Every stationary process $\{x_t\}$ can be approximated with arbitrary accuracy by an harmonic process, i.e., there is a sequence $\{x_{t,n} : t \in \mathbb{Z}\}_{n \in \mathbb{N}}$, of harmonic processes $\{x_{t,n} : t \in \mathbb{Z}\}$ such that

$$\underset{n \to \infty}{\text{l.i.m}}\, x_{t,n} = x_t \tag{B.1}$$

holds for every t, where the notation $\text{l.i.m}_{n \to \infty} x_{t,n} = x_t$ means that the limit is understood in the mean squares sense.

More precisely, a sequence of random variables $\{x_k\}_{k \in \mathbb{N}}$ is said to converge to x_0 in *mean squares sense* if

$$E x_0^* x_0 < \infty \tag{B.2}$$

and

$$\lim_{k \to \infty} E(x_k - x_0)^*(x_k - x_0) = 0 \tag{B.3}$$

holds.

A stochastic process $\{z(f) : f \in [-1/2, 1/2]\}$ with random variables $z(f) : \Omega \to \mathbb{C}^p$ is called a *process of orthogonal increments* if the following conditions are satisfied:

(i) $z(-1/2) = 0$ a.e. and $z(1/2) = x_0$ a.e.

(ii) $\text{l.i.m}_{\varepsilon \downarrow 0}\, z(f + \varepsilon) = z(f)$ for $f \in [-1/2, 1/2)$ (right continuity)

(iii) $Ez(f)^* z(f) < \infty$ for all $f \in [-1/2, 1/2]$

(iv) $E(z(f_4) - z(f_3))(z(f_2) - z(f_1))^* = 0$ for all $f_1 < f_2 \leq f_3 < f_4$.

We note that $\{z(f) : f \in [-1/2, 1/2]\}$ is a stochastic process with a continuous index set $[-1/2, 1/2]$ and we will interpret these indices f not as time points but as frequencies.

Suppose we have given a deterministic, scalar function $g : [-1/2, 1/2] \to \mathbb{C}$, $f \mapsto g(f)$ and a partition $-1/2 = f_0^n < f_1^n < \cdots < f_n^n = 1/2$ of the interval $[-1/2, 1/2]$. We then define a finite sum

$$\mathrm{I}_n(g) = \sum_{i=0}^{n-1} g(f_i^n)(z(f_{i+1}^n) - z(f_i^n)) . \tag{B.4}$$

If for all sequences of partitions with $\max_i(f_{i+1}^n - f_i^n) \to 0$ for $n \to \infty$ the limit in mean squares sense of $\mathrm{I}_n(g)$ exists and is the same, then we define

$$\mathrm{I}(g) = \int_{-1/2}^{1/2} g(f) dz(f) = \underset{n \to \infty}{\mathrm{l.i.m}} \, \mathrm{I}_n(g) . \tag{B.5}$$

$\mathrm{I}(g)$ is called the *stochastic integral* of $g(.)$ with respect to the process $z(f)$.

We now state the *Spectral Representation Theorem* (cf. also Hannan, 1970; Brockwell and Davis, 1991, for details). For every second-order stationary process $\{x_t\}$ there exists a process $\{z(f) : f \in [-1/2, 1/2]\}$ with orthogonal increments such that

$$x_t = \int_{-1/2}^{1/2} e^{i2\pi ft} \, dz(f) \tag{B.6}$$

holds. The process $\{z(f)\}$ is a.e. uniquely determined by $\{x_t\}$.

The process $\{z(f)\}$ defines a function $F : [-1/2, 1/2] \to \mathbb{C}^{n \times n}$ by $F(f) = Ez(f)z(f)^*$ where the following relations hold:

$$
\begin{aligned}
F(-1/2) &= 0 \\
F(1/2) &\geq 0 \\
F(f_2) - F(f_1) &= E(z(f_2) - z(f_1))(z(f_2) - z(f_1))^* \text{ for } f_1 \leq f_2 .
\end{aligned}
\tag{B.7}
$$

Thus $F(.)$ is a non-decreasing right continuous function, where non-decreasing means that the difference $F(f_2) - F(f_1)$ is a non-negative definite matrix for all $f_1 \leq f_2$.

If $\{z(f)\}$ is the orthogonal increment process corresponding to $\{x_t\}$ the function $F(.)$ is called the *spectral distribution function* of $\{x_t\}$. If there exists a function $S : [-1/2, 1/2] \to \mathbb{C}^{n \times n}$ such that

$$F(f) = \int_{-1/2}^{f} S(\nu) \, d\nu \, , \qquad (B.8)$$

where ν denotes the Lebesgue measure, then $S(.)$ is called the *spectral density function* of $\{x_t\}$. Other commonly used terms for $S(.)$ are *spectral density*, *spectrum* or *power spectrum*.

If we assume (to ease notation) that $Ex_t = 0$ we have

$$
\begin{aligned}
\gamma(h) &= Ex_h x_0^* \\
&= E \int_{-1/2}^{1/2} e^{i2\pi fh} \, dz(f) \left(\int_{-1/2}^{1/2} e^{i2\pi f0} \, dz(f) \right)^* = \int_{-1/2}^{1/2} e^{i2\pi fh} \, dF(f) \, ,
\end{aligned} \qquad (B.9)
$$

which is the *spectral representation of the autocovariance function*. If $S(.)$ exists, we further get

$$\gamma(h) = \int_{-1/2}^{1/2} e^{i2\pi fh} S(f) \, df \, . \qquad (B.10)$$

One condition to ensure the existence of the spectral density function is if the autocovariance function is absolutely summable, i.e.,

$$\sum_{h=-\infty}^{\infty} ||\gamma(h)|| < \infty \, . \qquad (B.11)$$

Then the spectral distribution function is absolutely continuous with $dF(f) = S(f) \, df$.

Under this condition the autocovariance at lag h, $\gamma(h)$, $h \in \mathbb{Z}$, are the Fourier coefficients of $S(.)$ and thus we can represent $S(f)$ as

$$S(f) = \sum_{h=-\infty}^{\infty} \gamma(h) e^{-i2\pi fh} \, . \qquad (B.12)$$

We note that

$$\gamma(0) = \text{var}(x_t) = \int_{-1/2}^{1/2} S(f) \, df \qquad (B.13)$$

which follows from (B.10) and expresses the total variance as the integrated spectral density over all the frequencies.

119

Bibliography

M. Ahdesmäki, H. Lähdesmäki, R. Pearson, H. Huttunen, and O. Yli-Harja. Robust detection of periodic time series measured from biological systems. *BMC Bioinformatics*, 6, 2005.

H. Akaike. Power spectrum estimation through autoregressive model fitting. *Ann. Inst. Statist. Math.*, 21:407–419, 1969.

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, AC-19:716–722, 1974.

B.D.O. Anderson and J.B. Moore. *Optimal filtering*. Prentice Hall, Englewood Cliffs, NJ, 1979.

A.E. Beaton and J.W. Tukey. The fitting of power series, meaning polynomials, illustrated on bandspectroscopic data. *Technometrics*, 16(2):147–192, 1974.

R.B. Blackman and J.W. Tukey. *The Measurement of Power Spectra*. Dover, New York, 1958.

P.J. Brockwell and R.A. Davis. *Time Series: Theory and Methods*. Springer, New York, 2nd edition, 1991.

A.D. Chave, D.J. Thomson, and M.E. Ander. On the robust estimation of power spectra, coherences, and transfer functions. *J. Geophys. Res.*, 92 (B1):633–648, 1987.

C. Croux and P. Filzmoser. Projection pursuit based measures of association. Research report 0341, K.U.Leuven, Dept. of Applied Econometrics, 2003.

C. Croux and P.J. Rousseeuw. Time-efficient algorithms for two highly robust estimators of scale. *Computational Statistics*, 2:411–428, 1992.

M. Deistler and W. Scherrer. The Prague Lectures Econometrics II. Dept. of Econometrics, Operations Research and System Theory, Vienna University of Technology, Vienna, 1994.

L. Denby and R.D. Martin. Robust estimation of the first-order autoregressive parameter. *J. Am. Statist. Assoc.*, 74(365):140–146, 1979.

S.J. Devlin, R. Gnanadesikan, and J.R. Kettenring. Robust estimation and outlier detection with correlation coefficients. *Biometrika*, 62(3):531–545, 1975.

D.L. Donoho and T.P.-T. Yu. Nonlinear pyramid transforms based on median-interpolation. *SIAM Journal of Math. Anal.*, 31(5):1030–1061, 2000.

I. Drews. Zur statistischen Auswertung der Herzperiodendauer. Berlin, 1983.

J. Durbin and S.J. Koopman. *Time Series Analysis by State Space Methods.* Oxford University Press, New York, 2001.

R. Dutter. Computer program BLINWDR for robust and bounded influence regression. Technical Report 8, Inst. Statist., Techn. Univ., Graz, 1983.

A.J. Fox. Outliers in time series. *J. Royal Statist. Soc.*, 34(3):350–363, 1972.

R. Fried and K. Schettlinger. `robfilter`: *Robust Time Series Filters.* Web: `http://cran.r-project.org/web/packages/robfilter/`, 2008.

R. Gençay, F. Selçuk, and B. Whitcher. *An Introduction to Wavelets and Other Filtering Methods in Finance and Economics.* Academic Press, San Diego, 2002.

F.R. Hampel. *Contributions to the Theory of Robust Estimation.* PhD thesis, University of California, Berkeley, 1968.

E.J. Hannan. *Multiple Time Series.* John Wiley & Sons, New York, 1970.

J.E.K. Hartikainen, K.U.O. Tahvanainen, and T.A. Kuusela. Short-term measurement of heart rate variability. In Malik, editor, *Clinical Guide to Cardiac Autonomic Tests*, pages 149–176. Kluwer, Dordrecht, 1998.

O. Hössjer and C. Croux. Generalizing univariate signed rank statistics for testing and estimating a multivariate location parameter. *Nonparametric Statistics*, 4:293–308, 1995.

K. Howorka, J. Pumprla, P. Haber, J. Koller-Strametz, J. Mondrzyk, and A. Schabmann. Effects of physical training on heart rate variability in diabetic patients with various degrees of cardiovascular autonomic neuropathy. *Cardiovasc. Res.*, 34(1):206–214, 1997.

K. Howorka, J. Pumprla, and A. Schabmann. Optimal parameters of short-term heart rate spectrogram for routine evaluation of diabetic cardiovascular autonomic neuropathy. *J. Auton. Nerv. Syst.*, 69(2-3):164–172, 1998.

P.J. Huber. Robust estimation of a location parameter. *Ann. Math. Statist.*, 35(1):73–101, 1964.

P.J. Huber. Robust regression: Asymptotics, conjectures and Monte Carlo. *Ann. Statist.*, 1(5):799–821, 1973.

P.J. Huber. *Robust Statistics*. John Wiley & Sons, New York, 1981.

C.M. Hurvich and Beltrão. Cross-validatory choice of a spectrum estimate and its connections with AIC. *J. Time Series Analysis*, 11(2):121–137, 1990.

A. Jazwinski. *Stochastic Processes and Filtering Theory*. Academic Press, New York, 1970.

A.G. Jones and K. Hollinger. Spectral analysis of the KTB sonic and density logs using robust nonparametic methods. *J. Geophys. Res.*, 102(B8): 18391–18403, 1997.

R.E. Kalman. A new approach to linear filtering and prediction problems. *J. Basic Eng.—Trans. ASME*, 82:35–45, 1960.

R.E. Kalman and R.S. Bucy. New results in filtering and prediction theory. *J. Basic Eng.—Trans. ASME*, 83:95–108, 1961.

R. Kleiner, R.D. Martin, and D.J. Thomson. Robust estimation of power spectra. *J. Royal Statist. Soc. B*, 41(3):313–351, 1979.

H. Liu, S. Shah, and W. Jiang. On-line outlier detection and data cleaning. *Computers and Chemical Engineering*, 28:1635–1647, 2004.

Y. Ma and M.G. Genton. Highly robust estimation of the autocovariance function. *J. Time Series Analysis*, 21(6):663–684, 2000.

C.L. Mallows. On some topics in robustness. Technical report, Bell Laboratories, Murray Hill, New Jersey, 1976.

R. Maronna. Robust M-estimation of multivariate location. *Ann. Statist.*, 4 (1):51–67, 1976.

R.A. Maronna, R.D. Martin, and V.J. Yohai. *Robust Statistics: Theory and Methods.* John Wiley & Sons, New York, 2006.

R.D. Martin. Approximate conditional-mean type smoothers and interpolators. In Gasser and Rosenblatt, editors, *Smoothing Techniques for Curve Estimation.* Springer-Verlag, New York, 1979.

R.D. Martin. Robust estimation of autoregressive models. In D.R. Brillinger and G.C. Tiao, editors, *Directions in Time Series*, pages 228–254. Inst. Math. Statist. Publications, Haywood, CA, 1980.

R.D. Martin. Robust methods for time series. In Findley, editor, *Applied Time Series II.* Academic Press, New York, 1981.

R.D. Martin and J.M. Jong. Asymptotic properties of robust generalized M-estimates for the first-order autoregressive parameter. Technical report, Bell Laboratories, Murray Hill, New Jersey, 1976.

R.D. Martin and D.J. Thomson. Robust-resistant sprectrum estimation. In *Proceedings of the IEEE*, volume 70, pages 1097–1115. IEEE, 1982.

R.D. Martin and J.E. Zeh. Generalized M-estimates for autoregressions, including small-sample efficiency robustness. Technical Report 214, Dept. of Electrical Engineering, Univ. Washington, Seattle, 1978.

G. Masarotto. Robust identification of autoregressive moving average models. *Appl. Statist.*, 36(2):214–220, 1987.

C.J. Masreliez. Approximate non-Gaussian filtering with linear state and observation relations. In *IEEE Transactions on Automatic Control*, pages 107–110, AC-20, 1975.

G. Matz and F. Hlawatsch. Minimax robust nonstationary signal estimation based on a p-point uncertainty model. *J. Franklin Inst.*, 337(4):403–419, 2000.

G. Matz and F. Hlawatsch. Wigner distributions (nearly) everywhere: time-frequency analysis of signals, systems, random processes, signal spaces, and frames. *Signal Processing*, 83(7):1355–1378, 2003.

J.B. Moore and B.D.O. Anderson. Coping with singular transition matrices in estimation and control stability theory. *Int. J. Control*, 31:571–586, 1980.

J. Möttönen, V. Koivunen, and H. Oja. Robust autocovariance estimation based on sign and rank correlation coefficients. In *Proc. of the IEEE Signal Processing Workshop on Higher-Order Statistics*, pages 187–190, 1999.

R.T. Ogden. *Essential Wavelets for Statistical Applications and Data Analysis*. Birkhäuser, Boston, 1997.

E. Parzen. Mathematical considerations in the estimation of spectra. *Technometrics*, 3:167–190, 1961.

R. Pearson, H. Lähdesmäki, H. Huttunen, and O. Yli-Harja. Detecting periodicity in nonideal datasets. In *Proc. of the SIAM International Conference on Data Mining*, 2003.

D.B. Percival and A.T. Walden. *Spectral Analysis for Physical Applications: Multitaper and Conventional Univariate Techniques*. Cambridge University Press, Cambridge, 1993.

D.B. Percival and A.T. Walden. *Wavelet Methods for Time Series Analysis*. Cambridge University Press, Cambridge, 2000.

M.B. Priestley. *Spectral Analysis and Time Series*, volume 1 of *Probability and Mathematical Statistics*. Academic Press, London, 1981.

J. Pumprla, K. Howorka, D. Groves, M. Chester, and J. Nolan. Functional assessment of heart rate variability: physiological basis and practical applications. *Int. J. Cardiology*, 84:1–14, 2002.

R Development Core Team. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, 2005. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

P.J. Rousseeuw. Least median of squares regression. *J. Am. Statist. Assoc.*, 79:871–880, 1984.

P.J. Rousseeuw and C. Croux. Explicit scale estimators with high breakdown point. In Y. Dodge, editor, L_1 *Statistical Analyses and Related Methods*, pages 77–92. North-Holland, Amsterdam, 1992.

P.J. Rousseeuw and C. Croux. Alternatives to the median absolute deviation. *J. Am. Statist. Assoc.*, 88:1273–1283, 1993.

P.J. Rousseeuw and G. Molenberghs. Transformation of non positive semidefinite correlation matrices. *Comm. Statist., Theory Meth.*, 22(4):965–984, 1993.

P.J. Rousseeuw and K. Van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41:212–223, 1999.

P. Ruckdeschel. Robust Kalman filtering. In Härdle, Hlávka, and Klinke, editors, *XploRe. Application Guide*, chapter 18, pages 483–516. Springer, 2000.

P. Ruckdeschel. *Ansätze zur Robustifizierung des Kalman-Filters*, volume 64 of *Bayreuther Mathematische Schriften*. Mathematisches Institut, Universität Bayreuth, Bayreuth, 2001. PhD thesis.

F. Schweppe, E. Handschin, J. Kohlas, and A. Friechter. Bad data analysis for power system state estimation. In *IEEE Transactions on Power Apparatur and Systems*. 94, 2, 329-337, 1975.

R.H. Shumway and D.S. Stoffer. *Time Series Analysis and Its Applications*. Springer, New York, 2000.

A.F. Siegel. Robust regression using repeated medians. *Biometrika*, 69:242–244, 1982.

D. Slepian. Prolate spheriodal wave functions, Fourier analysis, and uncertainty-V: The discrete case. *Bell System Techn. J.*, 57:1371–1430, 1978.

B. Spangl and R. Dutter. On robust estimation of power spectra. *Austrian Journal of Statistics*, 34(2):199–210, 2005.

B. Spangl and R. Dutter. Estimating spectral density functions robustly. *REVSTAT - Statistical Journal*, 5(1):41–61, 2007.

B. Spangl and R. Dutter. Approximate conditional-mean type filtering for vector-valued observations. Technical Report TR-AS-08-1, Institut für Angewandte Statistik und EDV, Universität für Bodenkultur, Wien, 2008.

N. Stockinger and R. Dutter. Robust time series analysis: A survey. *Kybernetika*, Supplement 23:1–90, 1987.

Task Force of The European Society of Cardiology and The North American Society of Pacing and Electrophysiology. Heart rate variability. *Circulation*, 93:1043–1065, 1996.

L.G. Tatum and C.M. Hurvich. A frequency domain approach to robust time series analysis. In Morgenthaler, Ronchetti, and Stahel, editors, *New Directions in Statistical Data Analysis and Robustness*. Birkhäuser-Verlag, Basel, 1993a.

L.G. Tatum and C.M. Hurvich. High breakdown methods of time series analysis. *J. Royal Statist. Soc. B*, 55(4):881–896, 1993b.

D.J. Thomson. Spectrum estimation techniques for characterization and development of WT 4 Waveguide-I. *Bell System Techn. J.*, 56(4):1769–1815, 1977.

D.J. Thomson. Spectrum estimation and harmonic analysis. In *Proceedings of the IEEE*, volume 70, pages 1055–1096. IEEE, 1982.

D.J. Thomson. An overview of multiple-window and quadratic-inverse spectrum estimation methods. In *Proceedings of the IEEE ICASSP*, volume 6, pages 185–194. IEEE, 1994.

P.D. Welch. The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics*, 15:70–73, 1967.

S. Wichert, K. Fokianos, and K. Strimmer. Identifying periodically expressed transcripts in microarray time series data. *Bioinformatics*, 20(1):5–20, 2004.

Curriculum Vitae

June 22, 1975	born in Vienna, Austria
1981–1985	elementary school, Vienna
1985–1993	Bundesrealgymnasium XVI, Vienna
1993–1994	military service
1994–2003	Student at the Vienna University of Technology
2000–2003	Tutor at the Dept. of Statistics & Probability Theory, Vienna University of Technolgy
2001–2003	working for the Österreichische Nationalbank, Abt. für Zahlungsbilanz & Devisenstatistik
2003	MSc from the Vienna University of Technology
since 2003	University Assistent at the Inst. of Applied Statistics & Computing, University of Natural Resources & Applied Life Sciences, Vienna (BOKU)