UNIVERSITÄT DER BUNDESWEHR MÜNCHEN
Fakultät für Elektrotechnik und Informationstechnik

Large-Scale
Gradual Change Detection

Heiko Hofer

Vorsitzender des Promotionsausschusses:	Prof. Dr.-Ing. G. Bauch
1. Berichterstatter	Priv.-Doz. Dr.-Ing. G. Staude
2. Berichterstatter	Prof. Dr.rer.nat. K. Pilzweger

Tag der Prüfung 8.4.2010

Mit der Promotion erlangter akademischer Grad:
Doktor-Ingenieur
(Dr.-Ing.)

Neubiberg, den 12. April 2010

Abstract

Observation is one of the import tasks of scientific work. A quantified observation over time is termed a signal. A signal is a documentation of changes in the system and can be used for analysing the past states of the system. This empirical method is a central concept in science since the European Renaissance.

In this thesis, a method is developed for the computerised analysis of signals. The problem is the segmentation of a signal in segments with constant system behaviour.

The computer aided analysis has the advantages to decide objectively and to give reproducible results. In addition to that, it allows to process a large amount of data in reasonable time. These two points are especially important in intensive care monitoring, where a reliable segmentation is the precondition for the description of the patients state.

Other applications of segmentation can be found in almost all scientific areas. Among them is the study of global climate changes and the detection of trends in econometrics.

The algorithm presented in this thesis is termed SEMUG and aims specifically on the segmentation problem where the signal is a linear function on the segments. The signal is assumed to be composed of adjacent ramp-steps where a ramp-step is a linear transition between two steady states. It is a generalisation of the often used step and ramp profiles. Since it includes these profiles as special cases it can be applied for many traditional segmentation problems, as well.

Segmentation is a search in the finite set of all possible partitions which becomes challenging for long signals, since the number of partitions grow exponentially making a complete search impossible. SEMUG avoids the exponential grow with a sequential search strategy where the change-points are detected one after another.

The performance of SEMUG has been evaluated on simulated and measured data. The most notable result is the robustness on either stochastic or deterministic errors. This thesis does not only provide an algorithm for the segmentation problem, additionally it provides an easy and transparent system for tuning SEMUG. In the applications of SEMUG there is typically not much information about the system's

statistical properties so that the tuning has to be based on a few known parameters, only. The tuning system developed in this thesis is based on visually apparent properties of the least significant change in the data set and it is shown by example that with this information the segmentation of the data set can be well performed.

This thesis gives a solution for an open problem in change-point analysis. However, new questions arose from the found insights. A central one is whether a more complex model could enhance the performance while preserving the reliability of SEMUG, which will be an interesting question for new research projects in the area of large-scale gradual change detection.

Zusammenfassung

Das Beobachten ist eine wichtige Tätigkeit eines Wissenschaftlers. Quantifiziert man die Beobachtungen und schreibt sie über die Zeit auf, so erhält man Signale, die die zeitliche Veränderung des beobachteten Systems beschreiben und die für Aussagen über das System genutzt werden können. Diese empirische Herangehensweise wird seit der Renaissance in der Wissenschaft genutzt und durchdringt heutzutage alle Realwissenschaften.

In dieser Dissertation wird eine Methode für die computergestützte Analyse von systembeschreibenden Signalen entwickelt. Die Aufgabe der Methode soll das Zerlegen der Signale in Segmente mit konstantem Systemverhalten sein.

Die computergestützte Analyse hat den Vorteil, dass eine objektive und reproduzierbare Entscheidung gefällt wird. Des Weiteren können mit dem Computer Datenmengen verarbeitet werden, die manuell nur mit einem unverhältnismäßigen Personalaufwand möglich wären. Ein Beispiel sind die Analyseverfahren in der Intensivmedizin. Eine objektive und fortlaufende Segmentierung der gemessenen Signale bedeutet hier, den Gesundheitszustand des Menschen zu beschreiben, was Voraussetzung für eine sachgemäße Behandlung ist.

Weitere Anwendungsgebiete der Segmentierung lassen sich in allen Realwissenschaften finden, darunter ist die Ursachenforschung der Veränderungen im globalen Klima, das Erkennen von Trends im Wirtschaftssystem oder die Beurteilung des Zustands einer Industrieanlage, um die Qualität des Endproduktes zu gewährleisten.

In der Dissertation wird ein spezielles, bisher nicht zufriedenstellend gelöstes, Segmentierungs-Problem angegangen. Es handelt sich um die Segmentierung eines zeitdiskreten Signals dessen Verlauf aus einer Verkettung von Rampensprüngen besteht. Ein Rampensprung ist ein linearer Übergang zwischen zwei konstanten Phasen. Der Rampensprung beinhaltet die in der Literatur oft vorkommenden Profile Sprung und Rampe. Deshalb kann der Algorithmus auch auf alle Probleme angewendet werden, bei denen man sich sonst für eines dieser Profile entscheiden musste. Der Algorithmus wird im folgenden als SEMUG bezeichnet welches ein Akronym

für Sequential Detection of Multiple Gradual Changes ist.

Die Segmentierung eines zeitdiskreten Signals ist ein endliches Suchproblem in der Menge aller denkbaren Segmentierungen. Die Mächtigkeit dieser Menge steigt exponentiell mit der Signallänge was eine vollständige Suche unmöglich macht. SEMUG umgeht das Problem durch die Anwendung einer sequentiellen Methode, die Segmente in chronologischer Reihenfolge detektiert.

Die Güte von SEMUG wird in der Dissertation anhand theoretischer Betrachtungen, simulierter Signale und durch die Anwendung auf physiomotorische Signale bewertet. Neben dem Algorithmus und anderen wissenschaftlichen Erkenntnissen, wird eine Methode zur Parametrisierung von SEMUG entwickelt, die ausschließlich auf visuell bestimmbaren Größen beruht. Statistische Kenngrößen der gemessenen Signale, die üblicherweise zur Parametrisierung notwendig sind, müssen nicht bekannt sein. Dadurch wird die Hürde zur Anwendung verringert.

Diese Dissertation bietet eine Lösung für ein offenes Problem aus dem Gebiet der Segmentierung von Signalen. Durch die Arbeit an diesem Problem sind neue offenen Fragen entstanden. Eine der wichtigsten ist, ob ein komplexeres Signalmodell die Güte tatsächlich erhöht und ob durch diese Erhöhung der Komplexität die Robustheit von SEMUG beeinträchtigt wird. Beides werden wichtige Fragen für weiterführende Forschungsprojekte sein.

Acknowledgement

I would like to gratefully acknowledge the enthusiastic supervision of Dr. Gerhard Staude during this work. I thank Prof. Werner Wolf, for his detailed and constructive comments, and for his important support throughout this work and I thank Prof. Ulrich Appel for making this work possible.

Special thanks to my colleagues at the Institute of Mathematics and Computer Science, University of A. F. Munich, for the nice working atmosphere.

I am forever indebted to my parents and all family members for their unconditional love. Finally, I thank my wife Monika for her understanding, endless patience and encouragement when it was most required.

Contents

1. Introduction

This thesis gives a solution to a problem out of the field of change-point analysis. Methods from change-point analysis are used in signal processing when the monitored signal comprises one or more structural changes. An example for such a signal is given in Figure 1.1. It is a quarterly index of import prices of petroleum products for Germany. For this signal, change-point analysis is concerned with two questions. Does the signal comprise significant abrupt shifts and when do they happen? In other applications these questions might be formulated more generally since changes can happen in the internals of the observed system which must not necessarily cause abrupt shifts in the monitored signal. However, the two questions are the core problems of change-point analysis. It is the detection whether a change has happened, termed *change detection*, and the estimate of its location, termed *change-point estimation*.

That the signal in Figure 1.1 comprises significant changes is undoubtedly true. There are three major shifts in the oil price. The first is caused by the Arab oil embargo after the Fourth Arab-Israeli War also known as the Yom Kippur war in 1973. The second price increase in 1979 marks the Iranian revolution followed by the war between Iran and Iraq (1980-1988) which had a negative effect on the oil production. Finally, in the mid-80s several minor effects cause a decrease of the oil price. Among them are an increase of production in Great Britain, Norway and Mexico, and internal quarrels in the OPEC cartel.

An algorithm which might be utilised for the oil price data must be capable to solve the *multiple change-point problem* which is the detection and the localisation of more than one change-point. However, in the early days of change-point

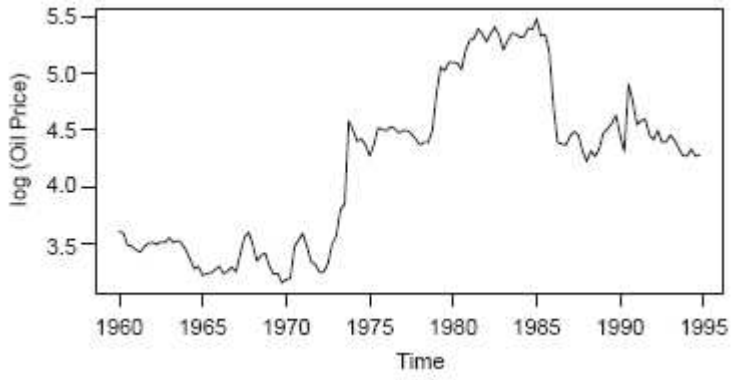

Figure 1.1.: A quarterly index of import prices of petroleum products. The data was obtained from the Statistisches Bundesamt Deutschland (Federal Statistical Office, Germany). (source: Zeileis et al. [2003])

research, the scientists focussed on a simpler model with a single abrupt change, only. Solutions for the single change-point problem were developed in the early 30s of the 20th century in the field of process control [Shewhart, 1931] (see also [Montgomery, 1985, Timmer and Pignatiello Jr., 2003]). Process supervision is typically a safety critical part with the aim to detect if a process, e. g., a chemical plant, gets out of control. Both, the methodologies as well as the variety of applications has evolved since then. Nowadays change-point related problems can be found in almost every scientific area. The applications range from the partitioning of audio signals [Desobry and Davy, 2003] over an improved trend analysis in econometrics [Perron and Zhu, 2005] and ecological science Toms and Lesperance [2003] to life sciences [Hall et al., 2003, Staude, 2001].

Note, that not in every mentioned application the signals are inherently discrete like the quarterly index of import prices depicted in Figure 1.1. In fact, most sensors give an analogue signal which could be analysed by analogue circuits, but the current and future trend is the sampling of analogue signals via an analogue-to-digital converter (ADC) resulting in a discrete-time signal, which is then processed by a digital computer. Thus, this thesis focuses on discrete-time signals and throughout the thesis, the term signal refers to a discrete-time signal.

This thesis addresses the multiple change-point problem, i. e., the partitioning

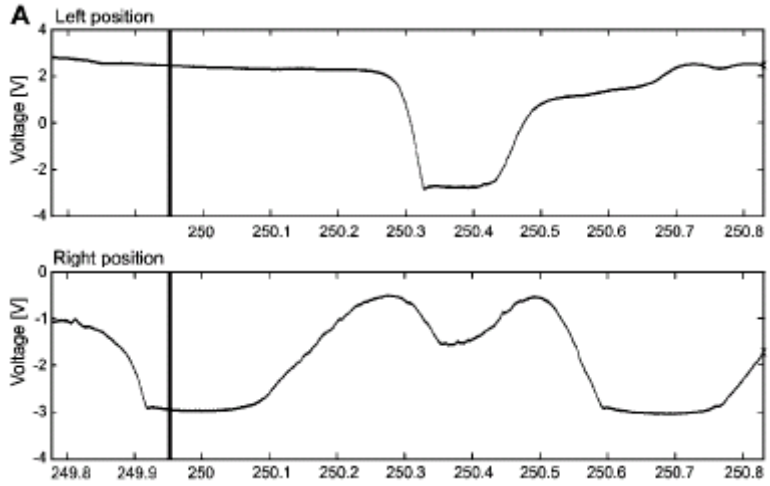

Figure 1.2.: Position of the index fingers (proportional to the depicted voltage value) recorded during a finger tapping experiment. Aim is to identify and locate single finger movements. Note, that the bar at time 249,95 is a marker and not part of the signal. (source: Cong Khac et al. [2007])

of signals in segments with constant system behaviour, which is also termed *segmentation*. Certainly, what is meant by constant system behaviour depends on the modelling. Cong Khac et al. [2007] addressed the segmentation of the finger position recorded during finger tapping experiments (see Figure 1.2). Objective is to separate rest-phases from movements in order to determine the beginning of a movement precisely. Following the physical modelling of a moving mass (the finger), a dynamical system would be an appropriate model for finger tapping. A change in the output value of a dynamical system is either caused by a change in the input signal or a change in the system's internal parameters. However, since neither the system's internals nor its input (the signal from the brain to the motor neurons of the finger) are measurable, the segmentation problem Cong Khac et al. [2007] dealt with, is to split up the monitored biomechanical signal in movements modelled by a piecewise linear function.

Because of its simplicity and flexibility the linear model can be found in many other applications, too. In cognitive science, segmentation is used in studies about the development of strategies. In Luwel et al. [2001] a study is presented where

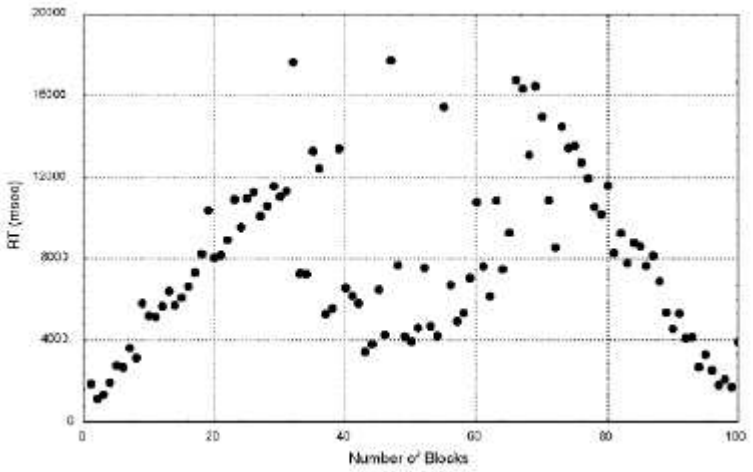

Figure 1.3.: The time needed for counting the number of coloured blocks on a 10 x 10 grid. Objective is to estimate the number of strategy shifts. The depicted recording comprises two strategy shifts at approximately 30 and 70 number of blocks. (source: Luwel et al. [2001])

subjects had to determine the number of coloured square blocks presented in a 10 x 10 grid. The time needed for counting over the number of squared blocks is depicted in Figure 1.3. Objective is to identify different strategies which is, e. g., to count the non-coloured blocks instead of the coloured blocks, typically used if the number of coloured blocks is high, which causes the decrease for a number of blocks greater than 70 in Figure 1.3. The linear increase for a low number of coloured blocks is caused by a second strategy where the number of coloured blocks are counted. Moreover, the response time pattern in Figure 1.3 shows a third, qualitatively different, strategy for quantity judgement, the so-called estimation strategy [Verschaffel et al., 1998].

While the change of a strategy might happen suddenly which is termed an *abrupt change*, a change in climatological time series is rather smooth which is called a *gradual change*. Figure 1.4 displays a temperature time series with annual measurements in the 20th century. A regression model with two change-points at 1935 and 1980 shows a local warming in the last 20 years, while a linear regression (dashed line) would falsely suggest a local cooling.

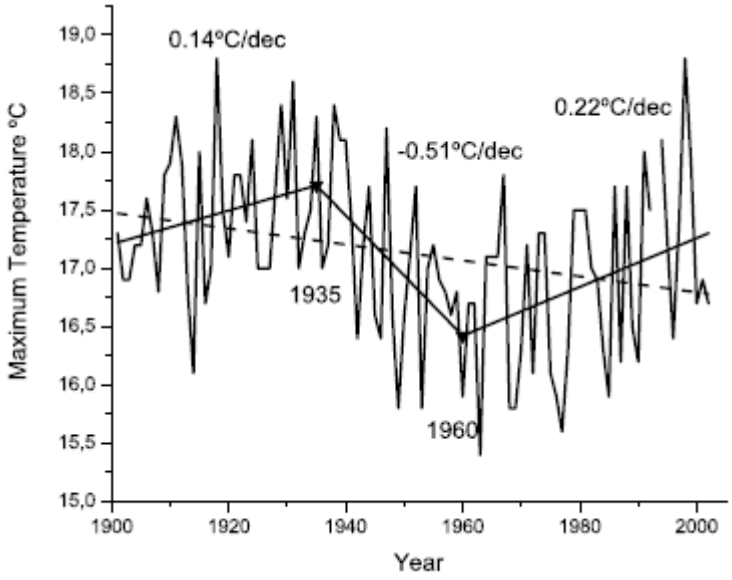

Figure 1.4.: Maximum December temperature at Angra do Heroismo (Azores), change-points (1935 and 1960), partial tendencies, in °C/decade, and the linear trend. Aim is to identify trends correctly which involves the partitioning in segments with constant trend. (source: Tome and Miranda [2004])

Another area where the correct identification of trends is of great importance is intensive care monitoring. Charbonnier et al. [2004] utilises change-point analysis to pre-process data describing the patient's state. The pre-processing splits up the signal in linear segments in order to identify trends and to remove artefacts. Two recordings of the oxygen saturation (SaO2) over time are depicted in Figure 1.5. The recorded signals (thin solid line) as well as the pre-processed signals (thick solid line) are displayed. Objective is to give an alarm when the oxygen saturation decreases significantly, whereas all the decreases not detected correspond to artefact measurements, according to Charbonnier et al. [2004]. Since their algorithm must be aware of fast transitions it identifies many changes in rather smooth phases so that it gives a vast amount of segments for the signal in Figure 1.5.

The four examples displayed in the Figures 1.2-1.5 share the common objective to split up the signal in linear segments. The mathematical problem is to decide which of the possible segmentations is the most likely one. The standard approach

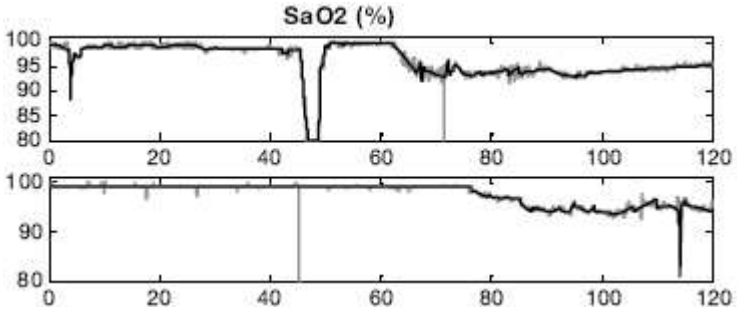

Figure 1.5.: Two recordings of the oxygen saturation (SaO2) over time (thin solid line) and the pre-processed signals (thick solid line). Pre-processing should identify trends (segmentation) and remove artefact measurements. (source: Charbonnier et al. [2004])

to solve such a problem is to define a so-called objective function, which gives for every possible segmentation a real number representing the likelihood that this segmentation is the true one. The objective function is then computed for every possible segmentation, and the segmentation which produces the maximum value of the objective function is called the optimal, most likely, segmentation.

While this approach solves the problem in principle, it is not feasible in every situation. It has especially its drawbacks in the large-scale case. Among the examples, intensive care monitoring and finger tapping are large scale problems. These problems are long signals with many change-points which results in a huge amount of possible segmentations so that enumerating every possible segmentation would not give the result in reasonable time. Note, that the number of segmentations grow exponentially with the size of the signal. Although previous research has been done focussing large-scale change-point problems [Vostrikova, 1981, Bai and Perron, 2003, Charbonnier et al., 2004], a complete solution for the large-scale case has not been found yet.

This thesis makes a contribution for closing this gap. It proposes an algorithm for the large-scale segmentation problem, where the algorithm is particularly suited for signals comprising linear and non-abrupt changes. The approach is based on statistical detection and estimation theory and it was successfully applied to biomechanical signals from psychophysiological experiments (see Chapter 4 for a detailed

description of the problem statement).

The outline of the thesis is organised as follows. Chapter 2 gives a brief overview about estimation and detection theory in statistical signal processing. Readers who are familiar with this topic may directly start with Chapter 3 where recent developments in related scientific areas are presented. This is followed by a detailed problem description in Chapter 4 with relations to the state of the art. In this chapter, the problem is defined that should be solved by the proposed algorithm. A detailed description of the algorithm is provided by Chapter 5. Chapter 6 is devoted to application engineers facing the problem of tuning the parameters of the algorithm. Easy to use rules are given in this chapter. The performance of the algorithm is analysed in Chapter 7 on simulated and measured signals, whereas the algorithm's properties when applied to signals with low signal to noise ratio (SNR) is focussed in Chapter 8, followed by some concluding remarks in Chapter 9.

2. A Brief Review of Statistical Signal Processing

Statistical signal processing is part of many modern signal processing systems. An example are radar systems at the airport which are installed to keep track of the aircrafts close to the airport. The interesting value is the distance to the aircraft. A radar system consists of a transmitter and a receiver. The transmitter sends out short bursts of radio waves. It then shuts off, and the receiver listens for the echoes which are reflections by the aircraft. The elapsed time is proportional to the distance.

Statistical signal processing involved in a radar system is responsible to (i) detect whether the monitored signal comprises a reflection and (ii) a precise estimation of the reflections onset. Both, detection and estimation theory will be treated in this chapter. Followed by a brief review of classification.

2.1. Estimation Theory

2.1.1. Introduction

Change-point estimation belongs to the general mathematical field called *estimation theory*. While change-point estimation deals with the problem of determining a single point in time, namely the change-point, estimation theory does generalise this objective to the problem of determining arbitrary parameters from a monitored signal. These parameters may describe several properties of the system like its dynamical behaviour or its structure.

Mathematically, a finite length discrete-time signal is a N-point *data set* $\{y[1], y[2], \ldots, y[N]\}$ which is for now supposed to depend on a single unknown parameter denoted by θ. Estimation means to determine θ based on the data, or, in other words, to define an *estimator* $g(y[1], y[2], \ldots, y[N])$, which is a mapping from the data space to the parameters space. The parameter estimated by g is denoted by $\hat{\theta}$

$$\hat{\theta} = g(y[1], y[2], \ldots, y[N]) \; . \tag{2.1}$$

The major aims of estimation theory are to find *suitable estimators* for a wide range of applications and to assess the *estimation error* due to noise and modelling errors.

2.1.2. Optimal Estimators

In any application it is desired to have an estimator which gives the best estimation result. To be able to define and verify optimality with mathematical methods, an application must be described by a mathematical model. Since monitored signals are inherently random, due to measurement errors and random effects in the system, the output of the system is described by its *probability density function* (PDF). The PDF is denoted by $p(y[1], y[2], \ldots, y[N]; \theta)$. This notation separates the variables from the parameters of the function p by a semicolon. Note that a parameter is viewed to have a fixed value which distinguishes it to a variable. Remember, that a variable represents *any* element within a specified set.

An easy to understand example from signal processing is the estimation of the direct current (DC) level. One can do a simple experimental setup with a source driving, e. g., a motor and measuring the current to the motor. After waiting some time there are multiple observations modelled by

$$y[t] = \theta + w[t] \quad \text{with} \quad t = 1, 2, \ldots, N \tag{2.2}$$

where $y[t]$ is the observation at *discrete-time* t, θ is the unknown DC level and $w[t]$ is zero mean white Gaussian noise (WGN) with variance σ^2, i. e., $y[t]$ is normally

distributed with mean θ and variance σ^2. The discrete time is a nonnegative integer equal to the time difference to the start of observation divided by the sampling time of the ADC plus one, where the sampling time is considered to be constant.

Let $\boldsymbol{y} = (y[1], y[2], \ldots, y[N])^T$ be a vector collecting the observations, the PDF of $\boldsymbol{y}$ can be expressed by $p(y[t]; \theta)$, the PDFs of the single observation. Since the observations $\{y[1], y[2], \ldots, y[N]\}$ are assumed to be statistically independent, the PDF $p(\boldsymbol{y}; \theta)$ is the product of the single PDFs so that

$$p(\boldsymbol{y}; \theta) = \prod_{t=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}\left(y[t] - \theta\right)^2\right] \tag{2.3}$$

$$= \frac{1}{\left(2\pi\sigma^2\right)^{N/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{t=1}^{N}\left(y[t] - \theta\right)^2\right] . \tag{2.4}$$

The PDF $p(\boldsymbol{y}; \theta)$ is often referred to as the *joint* PDF of the single PDFs $p(y[t]; \theta)$.

After the mathematical model is done, estimators for the DC level must be found. Two estimators are considered in the following. The first is the mean over the observations and the second is equal to the first observation

$$\hat{\theta} = \frac{1}{N}\sum_{t=1}^{N} y[t] \tag{2.5}$$

$$\breve{\theta} = y[1] . \tag{2.6}$$

Two methods to assess the performance of estimators are viewed next. The first is a *Monte Carlo technique* which is a numerical approach and the second method evaluates the PDF of the estimator analytically.

Since the observations are corrupted by noise, the performance of estimators cannot be compared by using just an one sample sequence. Applying the Monte Carlo technique, several sequences that attain the model, i. e., that attain (2.4) with varying value of the parameter θ are generated. For 100 realisations with $N = 20$, $\sigma^2 = 1$, and $\theta = 1$ the histograms are shown in Figure 2.1. It should now be evident that $\hat{\theta}$ is a better estimator than $\breve{\theta}$ because the values obtained are more concentrated around the true value.

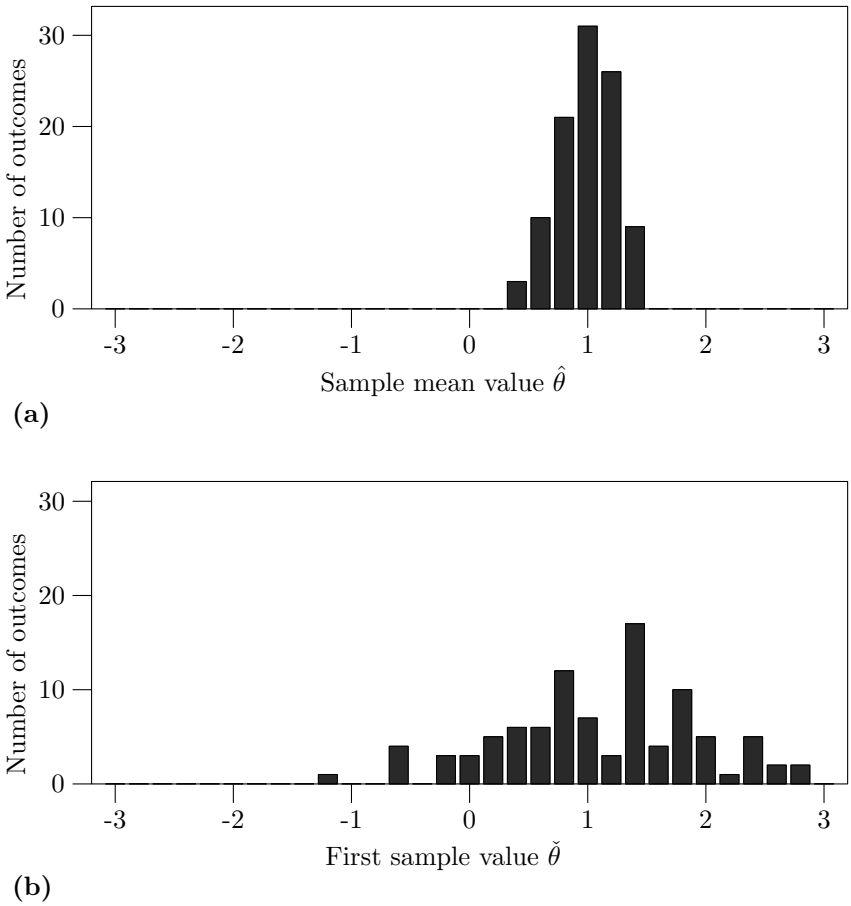

Figure 2.1.: Histograms of a Monte Carlo experiment estimating the DC level with (a) sample mean and (b) first sample estimator. The true value is at $\theta_0 = 1$.

Surely, for a real prove this experiment must be repeated for different θ, since the performance of estimators might depend on the true parameter θ. The computation effort rises exponentially with the number of parameters, which is the main drawback of the Monte Carlo technique. However, the power of this technique is that it can be used for any estimator.

A real *prove* that $\hat{\theta}$ has a better performance than $\check{\theta}$ can be done by evaluating their PDFs. The sum of N normally distributed random variables $\mathcal{N}(\theta, \sigma^2)$ is a random variable which is also normally distributed with mean $N\theta$ and variance $N\sigma^2$. Furthermore, a normally distributed variable $\mathcal{N}(N\theta, N\sigma^2)$ multiplied by $1/N$ results in a random variable that is $\mathcal{N}(\theta, \sigma^2/N)$ distributed. Consequently,

$\hat{\theta}$ according to (2.5) is $\mathcal{N}(\theta, \sigma^2/N)$ distributed in contrast to $\check{\theta}$ according to (2.6) which is $\mathcal{N}(\theta, \sigma^2)$ distributed.

A common performance measure is the bias of an estimator. An estimator is *unbiased* if it yields on average to the true value of the unknown parameter. Mathematically, an estimator is unbiased if its expectation is equal to the true value

$$\mathrm{E}(\hat{\theta}) = \theta \; . \tag{2.7}$$

Even though $\hat{\theta}$ and $\check{\theta}$ are both unbiased, the estimator $\hat{\theta}$ is preferable since its variance $\mathrm{var}(\hat{\theta}) = \sigma^2/N$ is lower than the variance of the second estimator $\mathrm{var}(\check{\theta}) = \sigma^2$.

Two methods, the Monte Carlo technique and the evaluation of the PDF have proven that the estimator $\hat{\theta}$ is better than $\check{\theta}$. However, there is still the possibility that there exist a better estimator than $\hat{\theta}$. In searching for optimal estimators one needs to adopt some optimality criterion. A natural one among many others is the *mean square error* (MSE), defined as

$$\mathrm{mse}(\hat{\theta}) = \mathrm{E}\left[(\hat{\theta} - \theta)^2\right] . \tag{2.8}$$

This measures the average mean squared deviation of the estimator from the true value. The MSE is a trade-off between the variance and the bias of the estimator. (2.8) can be rewritten as

$$\mathrm{mse}(\hat{\theta}) = \mathrm{E}\left\{\left[\Big(\hat{\theta} - \mathrm{E}(\hat{\theta})\Big) + \Big(\mathrm{E}(\hat{\theta}) - \theta\Big)\right]^2\right\} \tag{2.9}$$

$$= \mathrm{var}(\hat{\theta}) + \big[\underbrace{\mathrm{E}(\hat{\theta}) - \theta}_{\text{bias}}\big]^2 \tag{2.10}$$

which shows that the MSE is composed of errors due to the variance of the estimator as well as the bias. Since the bias depends on the true parameter θ, most estimators minimising the MSE depend on θ, too. Because of this dependency on the unknown parameter, those estimators are not realisable. An alternative approach is to constrain the bias to be zero and find the estimator which minimizes

the variance. Such an estimator is termed the minimum variance unbiased (MVU) estimator. In most practical situations engineers try to find the MVU estimator.

2.1.3. Minimum Variance Unbiased Estimators

A special property of the MVU estimators is that there exist a lower bound for their variance, namely the Cramer-Rao Lower Bound (CRLB). The CRLB allows to assert that an estimator is the MVU estimator. This will be the case if the estimator attains the bound for all values of the unknown parameter θ.

It is assumed that the PDF $p(\boldsymbol{y};\theta)$ satisfies the condition

$$\mathrm{E}\left\{\frac{\partial \ln p(\boldsymbol{y};\theta)}{\partial\theta}\right\} = 0 \quad \text{for all } \theta \tag{2.11}$$

often referred as the regularity condition. The MVU estimator $\hat{\theta} = g(\boldsymbol{y})$ attaining the CRLB has a variance of

$$\mathrm{var}(\hat{\theta}) = \frac{1}{I(\theta)} \tag{2.12}$$

which is minimal among the MVU estimators, where $I(\theta)$ is called the Fischer information. The Fischer information $I(\theta)$ as well as the estimator $g(\boldsymbol{y})$ can be determined from the equation

$$\frac{\partial \ln p(\boldsymbol{y};\theta)}{\partial\theta} = I(\theta)(g(\boldsymbol{y}) - \theta) \tag{2.13}$$

which holds for every MVU estimator [Cramér, 1946]. This will be illustrated for the example of an unknown DC level in WGN. The PDF $p(\boldsymbol{y};\theta)$ is given in (2.4). Taking the first derivative

$$\frac{\partial \ln p(\boldsymbol{y};\theta)}{\partial\theta} = \frac{\partial}{\partial\theta}\left[-\ln\left[\left(2\pi\sigma^2\right)^{N/2}\right] - \frac{1}{2\sigma^2}\sum_{t=1}^{N}\left(y[t]-\theta\right)^2\right] \tag{2.14}$$

$$= \frac{1}{\sigma^2}\sum_{t=1}^{N}\left(y[t]-\theta\right) \tag{2.15}$$

$$= \frac{N}{\sigma^2}\left(\left[\frac{1}{N}\sum_{t=1}^{N}y[t]\right] - \theta\right) \tag{2.16}$$

and comparing (2.16) with (2.13) it is obvious that the sample mean estimator attains the bound and must therefore be the MVU estimator. Also, the minimum variance is given by $I(\theta)^{-1} = \sigma^2/N$ which means that there is no unbiased estimator with a lower variance than that.

2.1.4. Maximum Likelihood Estimators

In many applications there are several unknown parameters. Although the CRLB can be extended to a vectorial form, it might be that a solution for (2.13) can not be found for every parameter. In such cases an estimator based on the maximum likelihood (ML) principle can be applied. The ML principle, originally developed by R. A. Fisher in the 1920s (see Aldrich [1997] for a historical overview), states that the desired probability distribution be the one that makes the observed data most likely. In the sense of ML, the PDF is viewed as a function of the unknown parameter θ, with $\boldsymbol{y}$ being a fixed (recorded) signal. From this point of view the PDF migrates to the *likelihood function* denoted by $\mathrm{L}(\theta; \boldsymbol{y})$. The maximum likelihood estimate (MLE) is obtained by seeking the value of the parameter vector that maximises the likelihood function. Although ML has a statistical motivation, it does not generally fulfil an optimality criterion like MSE or MVU estimators, although sometimes the ML principle produces the MVU estimator. But a distinct advantage of the likelihood is that it can always be computed for a given data set numerically (see Section 2.1.6) even though an analytical solution is preferable which will be discussed next.

An extension of the DC level example is a model with a time varying mean signal depending on the unknown parameter θ

$$y[t] = u[t] + w[t] \quad t = 1, 2, \ldots, N \tag{2.17}$$

with $u[t]$ being a function of t and θ. The likelihood for θ when $\boldsymbol{y}$ is observed is equal to

$$\mathrm{L}(\theta; \boldsymbol{y}) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{t=1}^{N}(y[t]-u[t])^2\right] . \tag{2.18}$$

Especially when dealing with WGN the log-likelihood function $\ln \mathrm{L}(\theta; \boldsymbol{y})$ is used instead of the likelihood function, since it is easier to compute. Thus, the estimate of θ, according to the ML principle, is the argument which maximises the log-likelihood function

$$\hat{\theta} = \arg\max_{\theta} \ln \mathrm{L}(\theta; \boldsymbol{y}) . \tag{2.19}$$

Note that, since the logarithm is strictly increasing, it does not change the location of the maximum, which means that the log-likelihood gives the same results as the likelihood. From (2.18) it follows for the log-likelihood

$$\ln \mathrm{L}(\theta; \boldsymbol{y}) = -\frac{N}{2}\ln\left(2\pi\sigma^2\right) + \left[-\frac{1}{2\sigma^2}\sum_{t=1}^{N}(y[t]-u[t])^2\right] \tag{2.20}$$

and since the $\arg\max_\theta$-operator is invariant with respect to summation and multiplication with positive factors not depending on θ, (2.19) can be condensed to

$$\hat{\theta} = \arg\max_{\theta} \Lambda(\theta; \boldsymbol{y}) \tag{2.21}$$

with

$$\Lambda(\theta; \boldsymbol{y}) = -\sum_{t=1}^{N}(y[t]-u[t])^2 . \tag{2.22}$$

The optimisation (2.21) must be performed in order to obtain the MLE. The fact that the objective function (2.22) is relatively simple is an important reason why ML is the method of choice in many applications.

The formulas (2.21) and (2.22) show a strong connection of the MLE to another popular method, namely the *least squares* (LS) approach. An equivalent minimisation problem to (2.21) has the objective function

$$J(\theta; \boldsymbol{y}) = \sum_{t=1}^{N}(y[t]-u[t])^2 \tag{2.23}$$

which is (2.22) multiplied by minus one. $J(\theta; \boldsymbol{y})$ is the sum of the squared difference between the monitored signal $y[t]$ and the model $u[t] = f(t, \theta)$, well known as the objective function of the LS approach. This method dates back to 1795 when Gauss used the method to study planetary motions. In contrast to the ML principle, the LS approach is not build on a statistical model of the system making it a popular method when prior knowledge about the system is limited.

2.1.5. The Basic Change-Point Problem

The ML principle will be used next to solve the basic change-point problem which is a shift in the DC Level at an unknown change-point ν. Note that because of the abrupt change, the observation vector's PDF cannot be differentiated with respect to the change location and hence, the MVU estimator cannot be found using the CLRB. Although there exist recent work to overcome this issue (see Tourneret et al. [2004], Swami and Sadler [1998], Reza and Doroodchie [1996]), it remains unsolved so that the ML method is still state of the art for solving change-point problems.

Consider the model (2.17) with the mean signal

$$u[t] = \begin{cases} 0 & \text{for } t \leq \nu \\ 1 & \text{for } t > \nu \end{cases} \quad \text{with} \quad t = 1, 2, \ldots, N \tag{2.24}$$

From (2.21) if follows that the estimate $\hat{\nu}$ is obtained by maximising $\Lambda(\nu; \boldsymbol{y})$, which is equal to

$$\begin{aligned} \Lambda(\nu; \boldsymbol{y}) &= -\sum_{t=1}^{N} (y[t] - u[t])^2 \\ &= -\sum_{t=1}^{\nu} (y[t])^2 - \sum_{t=\nu+1}^{N} (y[t] - 1)^2 \end{aligned} \tag{2.25}$$

when the definition of $u[t]$ (2.24) is taken into account. The most obvious way to determine the maximum of $\Lambda(\nu; \boldsymbol{y})$ is to compute (2.25) for every possible change-point and take the maximum, which is well known as the *grid search* method. Figure 2.2 (a) displays a signal corrupted by WGN with $\sqrt{\sigma^2} = 0.5$ and a change

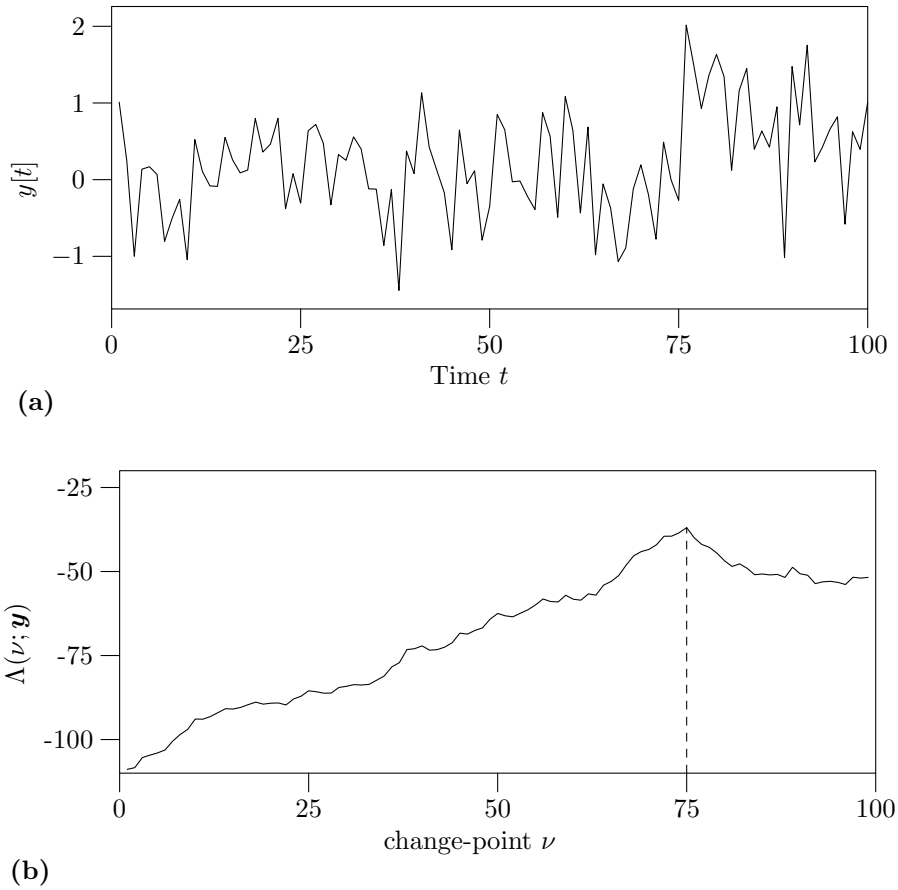

Figure 2.2.: (a) A signal corrupted by WGN with a change in DC level at $\nu_0 = 75$. (b) The log-likelihood function $\Lambda(\nu; \boldsymbol{y})$ which has its maximum at the true change-point.

of the DC level at $\nu_0 = 75$. $\Lambda(\nu; \boldsymbol{y})$ is displayed in Figure 2.2 (b). It has a clear maximum at $\nu = 75$ which is equivalent to the true change-point.

2.1.6. Newton-Raphson Method

A distinct advantage of the MLE is that it can always be found for a given data set *numerically*, although standard calculus is preferred in the case when the likelihood function is differentiable. If, as in the change-point example, the likelihood function is not differentiable but, the unknown parameter is confined to a finite set $\nu \in 1, 2, \ldots, N-1$ then a *grid search* over the set can be performed which is guaranteed

to find the MLE for the given data set, numerically. In cases where θ lives on an infinite interval so that grid search may not be computationally feasible, *iterative maximisation procedures* like the Newton-Raphson method might be applied for a numerical solution. In general, this method will produce the MLE if the initial guess is close to the true maximum. If not, convergence may not be attained, or only convergence to a local maximum. The difficulty with the use of iterative methods is that convergence can not be guaranteed and, even if convergence is attained, it can not be guaranteed that the value produced is the MLE. A special issue with ML is that the function to be maximised is not known a priori, since the likelihood function changes for each data set, requiring the maximisation of a random function. Nevertheless, iterative methods can at times produce good results but must be used with caution.

The Newton-Raphson method dates back to the 17th century where Newton applied an iterative scheme for finding a root of a polynomial [Newton, 1664–1671]. On top of that Raphson built a method which is close to the current notation [Raphson, 1690]. See Kollerstrom [1992] for a historical overview.

The Newton-Raphson method attempts to maximise the log-likelihood function by finding a zero of the derivative function. To do so, the derivative is taken and set equal to zero, yielding

$$\frac{\partial \ln \mathrm{L}(\theta; \boldsymbol{y})}{\partial \theta} = 0 \; . \tag{2.26}$$

Then, the method solves this equation iteratively. Let

$$g(\theta) = \frac{\partial \ln \mathrm{L}(\theta; \boldsymbol{y})}{\partial \theta} \tag{2.27}$$

and assume that θ_0 is a good initial guess for the solution to (2.26). Then, if $g(\theta)$ is approximately linear near θ_0, it can be approximated by

$$g(\theta) \approx g(\theta_0) + \left.\frac{dg(\theta)}{d\theta}\right|_{\theta=\theta_0} (\theta - \theta_0) \tag{2.28}$$

which are the first two elements of the Taylor expansion. Next, (2.28) is used to

solve for the zero θ_1, so that upon setting $g(\theta)$ equal to zero according to (2.26) and solving for θ_1 one gets

$$\theta_1 = \theta_0 - \left[\left. \frac{dg(\theta)}{d\theta} \right|_{\theta=\theta_0} \right]^{-1} g(\theta_0) \; . \tag{2.29}$$

Again the method linearises $g(\theta)$ but now at the new guess, θ_1, and repeats the previous procedure to find the new zero. In general, the Newton-Raphson iteration finds the new guess θ_{i+1} based on the previous one θ_i using

$$\theta_{i+1} = \theta_i - \left[\left. \frac{dg(\theta)}{d\theta} \right|_{\theta=\theta_i} \right]^{-1} g(\theta_i) \; . \tag{2.30}$$

Note that at convergence $\theta_{i+1} = \theta_i$, and from (2.30) $g(\theta_i) = 0$, as desired. Since $g(\theta)$ is the derivative of the log-likelihood function, the MLE is found by

$$\theta_{i+1} = \theta_i - \left[\frac{d^2 \ln \mathrm{L}(\theta; \boldsymbol{y})}{d\theta^2} \right]^{-1} \left. \frac{d \ln \mathrm{L}(\theta; \boldsymbol{y})}{d\theta} \right|_{\theta=\theta_i} \; . \tag{2.31}$$

In most situations there are more than one unknown parameters collected in a parameter vector. The Newton-Raphson method can easily be extended to the vector parameter case where the unknown parameters are optimised simultaneously. The Newton-Raphson iteration becomes

$$\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i - [H(\boldsymbol{\theta}; \boldsymbol{y})]^{-1} \left. \frac{\partial \ln \mathrm{L}(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_i} \; . \tag{2.32}$$

where

$$H(\boldsymbol{\theta}; \boldsymbol{y}) = \frac{\partial^2 \ln \mathrm{L}(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \tag{2.33}$$

is the Hessian of the log-likelihood function and $\frac{\partial \ln \mathrm{L}(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \boldsymbol{\theta}}$ is the gradient. When the number of unknown parameters is ρ, the gradient is a $\rho \times 1$ vector with the elements

$$\left[\frac{\partial \ln \mathrm{L}(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \boldsymbol{\theta}} \right]_i = \frac{\partial \ln \mathrm{L}(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \theta_i} \qquad i = 1, 2, \ldots, \rho \tag{2.34}$$

where θ_i is the i-th element of $\boldsymbol{\theta}$. Similarly, the Hessian is a $\rho \times \rho$ matrix with elements

$$[H(\boldsymbol{\theta};\boldsymbol{y})]_{i,j} = \frac{\partial^2 \ln \mathrm{L}(\boldsymbol{\theta};\boldsymbol{y})}{\partial\theta_i \partial\theta_j} \qquad i = 1, 2, \ldots, \rho\,;\ j = 1, 2, \ldots, \rho\,. \tag{2.35}$$

Note that, when implementing (2.32), inversion of the Hessian is not required. Rewriting (2.32) as

$$H(\boldsymbol{\theta};\boldsymbol{y})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_i}\,\boldsymbol{\theta}_{i+1} = H(\boldsymbol{\theta};\boldsymbol{y})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_i}\,\boldsymbol{\theta}_i - \left.\frac{\partial \ln \mathrm{L}(\boldsymbol{\theta};\boldsymbol{y})}{\partial\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_i}\,. \tag{2.36}$$

the new iterate $\boldsymbol{\theta}_{i+1}$, can be found from the previous iterate $\boldsymbol{\theta}_i$, by solving this set of ρ linear equations.

In general, the Newton-Raphson method will produce the MLE if the initial guess is close to the true maximum. If not, it might converge to a local maximum or convergence may not be attained. The reason for this is that the Newton-Raphson method is designed to give a one-step convergence for quadratic functions. When the function values between the current iterate and the minimum cannot be approximated very well by a quadratic function, the Newton-Raphson method tends to be unstable.

The effect of the Hessian in (2.32) is to control the step size and to modify the step direction. If the Hessian is be replaced by the identity matrix, the method degenerates to a pure gradient method. The influence of the Hessian is especially important close to the true maximum since the gradient becomes then very small. The Newton-Raphson method can become more stable when far from the true maximum the influence of the Hessian is confined. One common way is the procedure proposed by Levenberg [1944]. Then an approximation

$$R_L(\lambda) = H(\boldsymbol{\theta};\boldsymbol{y}) + \lambda\boldsymbol{I} \tag{2.37}$$

is used for the Hessian where $\boldsymbol{I}$ is the $\rho \times \rho$ identity matrix and λ is a positive

scalar, so that the update rule (2.32) becomes

$$\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i - [H(\boldsymbol{\theta};\boldsymbol{y}) + \lambda \boldsymbol{I}]^{-1} \left.\frac{\partial \ln \mathrm{L}(\boldsymbol{\theta};\boldsymbol{y})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_i} . \tag{2.38}$$

With $\lambda = 0$ the Levenberg procedure is equal to the Newton-Raphson method. Increasing λ means that the step size is decreased and the search direction is turned towards the gradient. The value of λ is dynamically changed from iteration to iteration. If the likelihood goes up following an update, it implies that $\boldsymbol{\theta}_{i+1}$ is closer to the maximum so that λ can be reduced (usually by a factor of 10) so that the influence of the gradient is reduced. On the other hand, if the likelihood goes down the step is retracted and λ is increased by the same factor, which means that in the next trial the gradient will have a greater influence.

A popular variation to the Levenberg procedure was introduced by Marquardt [1963] who replaced the identity matrix in (2.37) by the diagonal of the Hessian

$$R_M(\lambda) = H(\boldsymbol{\theta};\boldsymbol{y}) + \lambda \,\mathrm{diag}[H(\boldsymbol{\theta};\boldsymbol{y})] \tag{2.39}$$

resulting in the Levenberg-Marquardt update rule

$$\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i - [H(\boldsymbol{\theta};\boldsymbol{y}) + \lambda \,\mathrm{diag}[H(\boldsymbol{\theta};\boldsymbol{y})]]^{-1} \left.\frac{\partial \ln \mathrm{L}(\boldsymbol{\theta};\boldsymbol{y})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_i} . \tag{2.40}$$

The Levenberg-Marquardt update rule takes advantage of the Hessian even when λ is high. This is especially beneficial when the likelihood function forms a valley. In this case the Levenberg update rule does, when λ is high, a small step along the valley and a big step up the hill according to the gradient. Since the Hessian is proportional to the curvature of the function the Levenberg-Marquardt rule has the superior behaviour doing a big step a long the valley and only a small step up the hill which prevents the algorithm jumping out of the valley.

2.2. Detection Theory

2.2.1. Introduction

Detection theory is important in many modern signal processing systems. An example is the radar system at an airport keeping track of the aircrafts close to the airport. A signal processing problem which has to be solved in this context is to decide whether an aircraft has entered the airspace controlled by the radar system. To accomplish this task, the radar transmits an electromagnetic pulse, which if reflected by a large moving object, will indicate the presence of an aircraft. The received waveform will either consist of the reflected pulse if an aircraft is present or noise only, if there is no aircraft close to the airport. This kind of decision making problem is central in *detection theory*; being able to decide when an event of interest occurs and then to determine more information about that event. Where the latter is typically done with methods from estimation theory described in the previous section.

2.2.2. Neyman-Pearson Detector

The simplest detection problem is to decide whether a known signal is present, which, usually, is embedded in noise, or if only noise is present. Since this is a decision between two possible alternatives, this is termed *binary hypothesis testing problem*. Since the data are inherently random in nature, a statistical approach is necessary with the goal to use the received data as efficiently as possible.

The discussion about signal detection will be centred around the example, where on the basis of N observations it should be distinguished if the DC level is zero or one. The problem is formulated with the two hypotheses

$$\mathcal{H}_0 : y[t] = w[t] \tag{2.41}$$

$$\mathcal{H}_1 : y[t] = 1 + w[t] \quad \text{with} \quad t = 1, 2, \ldots, N \tag{2.42}$$

where $w[t]$ is a realisation of a WGN process with zero mean and variance σ^2. $\mathcal{H}_0$

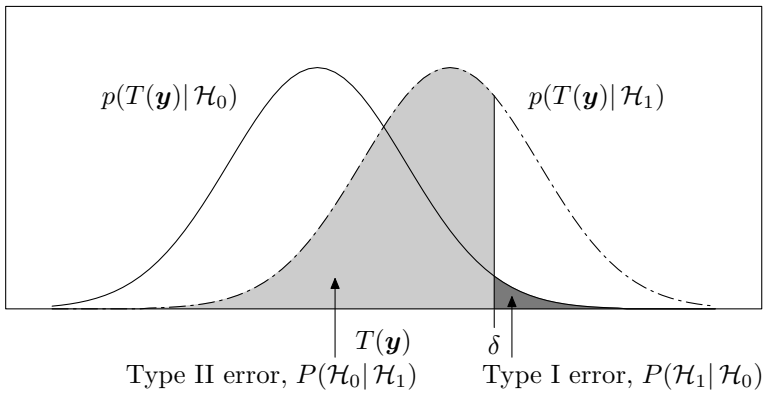

Figure 2.3.: PDFs of the test statistic $T(\boldsymbol{y})$ if $\mathcal{H}_0$ is true and $\mathcal{H}_1$ is true, respectively. Shaded are the two types of error. Type I error: Decision for $\mathcal{H}_1$ when $\mathcal{H}_0$ is true. Typ II error: Decision for $\mathcal{H}_0$ when $\mathcal{H}_1$ is true.

is referred as the *noise only hypothesis* and $\mathcal{H}_1$ as the *signal + noise hypothesis*. In statistics literature they are also termed *null hypothesis* and *alternate hypothesis*, respectively.

The decision is based on a test statistic $T(\boldsymbol{y})$ which is a mapping of the observed signal to a real value. When the test statistic exceeds a real valued threshold, denoted by δ, $\mathcal{H}_0$ is rejected and $\mathcal{H}_1$ is accepted, respectively. Note with this scheme one can make two types of errors. If the test decides $\mathcal{H}_1$ but $\mathcal{H}_0$ is true which is called a *type I error*. On the other hand, if the test decides $\mathcal{H}_0$ but $\mathcal{H}_1$ is true which is called a *type II error*. Figure 2.3 illustrates the PDF of $T(\boldsymbol{y})$ when $\mathcal{H}_0$ is true and the PDF of $T(\boldsymbol{y})$ when $\mathcal{H}_1$ is true. The probability of the two types of errors is shaded. The notation $P(\mathcal{H}_i|\,\mathcal{H}_j)$ indicates the probability of deciding $\mathcal{H}_i$ when $\mathcal{H}_j$ is true. The two errors are unavoidable to some extent but may be traded off against each other. To do so, the threshold δ must be changed in Figure 2.3.

The type I error, $P(\mathcal{H}_1|\,\mathcal{H}_0)$ is referred to as the *probability of false alarm* in engineering literature and is denoted by P_{FA}. The probability of a false alarm is crucial. E. g., in military applications, a falsely detected enemy aircraft may initiate an attack; or in intensive care monitoring, too many false alarms may cause that any further alarm might be ignored by the operator with disastrous effects. In any

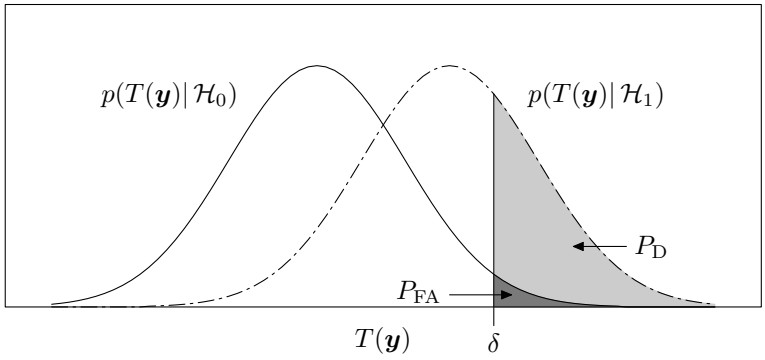

Figure 2.4.: PDFs of the test statistic $T(\boldsymbol{y})$ if $\mathcal{H}_0$ is true and $\mathcal{H}_1$ is true, respectively. Shaded are the probability of false alarm P_{FA} and the probability of detection P_{D}.

case, it is often desired that the P_{FA} is small and known a priori which leads to the *Neyman-Pearson* approach of optimal detection.

The Neyman-Pearson theorem provides the test statistic that maximises for a given $P_{\mathrm{FA}} = \alpha$ the probability $P(\mathcal{H}_1|\,\mathcal{H}_1)$ which is called the *probability of detection*, also denoted by P_{D}. P_{FA} and P_{D} are shaded in Figure 2.4. The test statistic according to Neyman-Pearson is the quotient of the likelihood for $\boldsymbol{y}$ under $\mathcal{H}_1$ and the likelihood for $\boldsymbol{y}$ under $\mathcal{H}_0$. Hence, the Neyman-Pearson detector decides $\mathcal{H}_1$ if

$$\mathrm{T}_{\mathrm{LR}}(\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\,\mathcal{H}_1)}{p(\boldsymbol{y}|\,\mathcal{H}_0)} > \delta \tag{2.43}$$

where the threshold δ is found from

$$P_{\mathrm{FA}} = \int_{\boldsymbol{y}:\mathrm{T}_{\mathrm{LR}}(\boldsymbol{y})>\delta} p(\boldsymbol{y}|\,\mathcal{H}_0)d\boldsymbol{y} = \alpha. \tag{2.44}$$

The function $\mathrm{T}_{\mathrm{LR}}(\boldsymbol{y})$ is termed the *likelihood ratio* since it indicates for each value of $\boldsymbol{y}$ the likelihood of $\mathcal{H}_1$ versus the likelihood of $\mathcal{H}_0$. The entire test is called the *likelihood ratio test* (LRT).

The PDFs under each hypothesis of the DC level example are

$$p(\boldsymbol{y}|\,\mathcal{H}_0) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{t=1}^{N} y[t]^2\right] \tag{2.45}$$

$$p(\boldsymbol{y}|\,\mathcal{H}_1) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{t=1}^{N} (y[t]-1)^2\right] . \tag{2.46}$$

The LRT decides $\mathcal{H}_1$ if

$$\frac{\frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{t=1}^{N} (y[t]-1)^2\right]}{\frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{t=1}^{N} y[t]^2\right]} > \delta\ . \tag{2.47}$$

Taking the logarithm of both sides results in

$$-\frac{1}{2\sigma^2}\left(\sum_{t=1}^{N}(y[t]-1)^2 - \sum_{t=1}^{N} y[t]^2\right) > \ln\delta \tag{2.48}$$

$$\Leftrightarrow \qquad -\frac{1}{2\sigma^2}\left(-2\sum_{t=1}^{N} y[t] + N\right) > \ln\delta \tag{2.49}$$

which simplifies to

$$\frac{N}{\sigma^2}\left(\frac{1}{N}\sum_{t=1}^{N} y[t] - \frac{1}{2}\right) > \ln\delta\ . \tag{2.50}$$

The LRT compares basically the sample mean with a threshold. The relation of the threshold δ and the probability of false alarms P_{FA} is obtained by solving (2.44). The integral is in this case equal to the complementary cumulative distribution function of a normally distributed random vector which can be found in tables or computed numerically.

2.2.3. Generalised Likelihood Ratio Test

Previously, complete knowledge of the PDFs under $\mathcal{H}_0$ and $\mathcal{H}_1$ has been assumed, allowing the design of the optimal Neyman-Pearson detector. In the more realistic problem the PDF is not completely known. For example, the radar return from a target will be delayed depending on the distance from the sender to the target. As

a result, the arrival time is generally unknown. The design of good detectors when the PDFs have unknown parameters is therefore of great practical importance.

The general class of hypothesis tests with unknown parameters are termed *composite hypothesis tests*. The PDF under $\mathcal{H}_0$ or under $\mathcal{H}_1$ or under both hypotheses may not be completely specified. An example is a similar problem as considered in the previous section with the difference that the DC level under hypothesis $\mathcal{H}_1$ is not known a priori. So, consider the detection problem

$$\mathcal{H}_0 : y[t] = w[t] \tag{2.51}$$

$$\mathcal{H}_1 : y[t] = A + w[t] \quad \text{with} \quad t = 1, 2, \ldots, N \tag{2.52}$$

where A is the unknown parameter and $w[t]$ is a realisation of a WGN process. Since the amplitude A is unknown, the PDF under $\mathcal{H}_1$ is not completely specified. The PDF belongs to a family of PDFs, one for each value of A. The PDF is said to be *parametrised* by A.

$$p(\boldsymbol{y}; A|\,\mathcal{H}_1) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{t=1}^{N}(y[t]-A)^2\right] \tag{2.53}$$

As a result of 2.50 the LRT in this scenario is basically the sample mean. The test statistic $\mathrm{T_{LR}}$ changes therefore its sign with A so that a comparison with a single threshold for $A > 0$ and $A < 0$ is not possible. The Neyman-Pearson detector can therefore not be applied.

Intuitively the test should be changed so that the absolute value of $\mathrm{T_{LR}}$ is compared to the threshold. This is exactly the solution that is obtained with the generalised *likelihood ratio test* (GLRT). The GLRT replaces the unknown parameters by their maximum likelihood estimates (MLEs). Unlike the LRT based on the Neyman-Pearson theorem there is no optimality associated with the GLRT, but it is a method which works very well in many applications. In general, the GLRT decides $\mathcal{H}_1$ if

$$\mathrm{T_{GLR}}(\boldsymbol{y}) = \frac{p(\boldsymbol{y}; \hat{\boldsymbol{\theta}}_1|\,\mathcal{H}_1)}{p(\boldsymbol{y}; \hat{\boldsymbol{\theta}}_0|\,\mathcal{H}_0)} > \delta \tag{2.54}$$

where $\hat{\boldsymbol{\theta}}_1$ is the MLE of $\boldsymbol{\theta}_1$ assuming $\mathcal{H}_1$ is true, and $\hat{\boldsymbol{\theta}}_0$ is the MLE of $\boldsymbol{\theta}_0$ assuming $\mathcal{H}_0$ is true. The approach also provides information about the unknown parameters since the first step in determining $\mathrm{T_{GLR}}(\boldsymbol{y})$ is to find the MLEs.

In the DC level example, $\boldsymbol{\theta}_1$ is the unknown post-change level A and there are no unknown parameters under $\mathcal{H}_0$. Thus, the GLRT decides $\mathcal{H}_1$ if

$$\mathrm{T_{GLR}}(\boldsymbol{y}) = \frac{p(\boldsymbol{y}; \hat{A}|\, \mathcal{H}_1)}{p(\boldsymbol{y}; \mathcal{H}_0)} > \delta \; . \tag{2.55}$$

The MLE of A is the sample mean

$$\hat{A} = \frac{1}{N} \sum_{t=1}^{N} y[t] \tag{2.56}$$

as derived in Section 2.1.4. Thus, with (2.45), (2.53), and (2.55)

$$\mathrm{T_{GLR}}(\boldsymbol{y}) = \frac{\frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{t=1}^{N} \left(y[t] - \hat{A}\right)^2\right]}{\frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{t=1}^{N} y[t]^2\right]} > \delta \; . \tag{2.57}$$

Taking the logarithm of both sides results in

$$-\frac{1}{2\sigma^2} \left(\sum_{t=1}^{N} \left(y[t] - \frac{1}{N} \sum_{i=1}^{N} y[i] \right)^2 - \sum_{t=1}^{N} y[t]^2 \right) > \ln \delta \tag{2.58}$$

$$\Leftrightarrow \quad -\frac{1}{2\sigma^2} \left(-2 \frac{1}{N} \sum_{t=1}^{N} y[t] \sum_{t=1}^{N} y[t] + N \left(\frac{1}{N} \sum_{t=1}^{N} y[t] \right)^2 \right) > \ln \delta \tag{2.59}$$

$$\Leftrightarrow \quad \frac{N}{2\sigma^2} \left(\frac{1}{N} \sum_{t=1}^{N} y[t] \right)^2 > \ln \delta \tag{2.60}$$

or the GLRT decides $\mathcal{H}_1$ if

$$\left(\frac{1}{N} \sum_{t=1}^{N} y[t] \right)^2 > \delta' \tag{2.61}$$

where

$$\delta' = \frac{2\sigma^2}{N} \ln \delta \; . \tag{2.62}$$

2.2.4. Change Detection

The methods derived in the previous section will be used next to solve the classical change detection problem. The objective of this problem is to detect a jump at unknown time, when the level before and after the jump is unknown, too.

Consequently, the noise only hypothesis is

$$\mathcal{H}_0 : y[t] = \mu + w[t] \quad \text{with} \quad t = 1, 2, \ldots, N \tag{2.63}$$

and the noise + signal hypothesis is formulated as

$$\mathcal{H}_1 : y[t] = \begin{cases} \mu_1 + w[t] & \text{with} \quad t = 1, 2, \ldots, \nu \\ \mu_2 + w[t] & \text{with} \quad t = \nu + 1, \nu + 2, \ldots, N \end{cases} \tag{2.64}$$

where ν is the unknown date of the single change. It is supposed that $w[t]$ is WGN with a known variance σ^2. Thus the assumption $\sigma^2 = 1$ is done without loss of generality.

The PDFs under the two hypothesis are defined by

$$p(\boldsymbol{y}; \mu | \, \mathcal{H}_0) = \frac{1}{\left(\sqrt{2\pi}\right)^N} \exp\left[-\frac{1}{2}\sum_{t=1}^{N}(y[t] - \mu)^2\right] \tag{2.65}$$

$$p(\boldsymbol{y}; \nu, \mu_1, \mu_2 | \, \mathcal{H}_1) = \frac{1}{(\sqrt{2\pi})^N} \exp\left[-\frac{1}{2}\sum_{t=1}^{\nu}(y[t] - \mu_1)^2 - \frac{1}{2}\sum_{t=\nu+1}^{N}(y[t] - \mu_2)^2\right] \tag{2.66}$$

Using the GLRT, the test statistic is given by

$$\mathrm{T}_{\mathrm{GLR}}(\boldsymbol{y}) = \frac{p(\boldsymbol{y}; \hat{\nu}, \hat{\mu}_1, \hat{\mu}_2 | \, \mathcal{H}_1)}{p(\boldsymbol{y}; \hat{\mu} | \, \mathcal{H}_0)} > \delta \tag{2.67}$$

where $\hat{\mu}$, $\hat{\mu}_1$, and $\hat{\mu}_2$ are the MLEs defined by

$$\hat{\mu} = \frac{1}{N} \sum_{t=1}^{N} y[t] \tag{2.68}$$

$$\hat{\mu}_1 = \frac{1}{\nu} \sum_{t=1}^{\nu} y[t] \tag{2.69}$$

$$\hat{\mu}_2 = \frac{1}{N-\nu} \sum_{t=\nu+1}^{N} y[t] \; . \tag{2.70}$$

The change-point ν is integer valued and lives on the interval $[1, N-1]$, so that the MLE of the change-point is guaranteed to be found by a grid search method, as explained in Section 2.1.5. It follows for the test statistic

$$\mathrm{T_{GLR}}(\boldsymbol{y}) = \frac{\max_\nu p(\boldsymbol{y}; \nu, \hat{\mu}_1, \hat{\mu}_2 | \, \mathcal{H}_1)}{p(\boldsymbol{y}; \hat{\mu} | \, \mathcal{H}_0)} \tag{2.71}$$

and since the denominator is independent of ν

$$\mathrm{T_{GLR}}(\boldsymbol{y}) = \max_\nu \frac{p(\boldsymbol{y}; \nu, \hat{\mu}_1, \hat{\mu}_2 | \, \mathcal{H}_1)}{p(\boldsymbol{y}; \hat{\mu} | \, \mathcal{H}_0)} \tag{2.72}$$

Using the abbreviations

$$S_\nu = \sum_{t=1}^{\nu} (y[t] - \hat{\mu}_1)^2 + \sum_{t=\nu+1}^{N} (y[t] - \hat{\mu}_2)^2$$

$$S = \sum_{t=1}^{N} (y[t] - \hat{\mu})^2$$

and taking the logarithm, the test statistic can be reduced to

$$\begin{aligned}
\ln \mathrm{T}_{\mathrm{GLR}}(\boldsymbol{y}) &= \max_{\nu} \ln \frac{p(\boldsymbol{y}; \nu, \hat{\mu}_1, \hat{\mu}_2 |\, \mathcal{H}_1)}{p(\boldsymbol{y}; \hat{\mu} |\, \mathcal{H}_0)} \\
&= \max_{\nu} \ln \frac{\exp\left[-\frac{1}{2}\sum_{t=1}^{\nu}(y[t]-\hat{\mu}_1)^2 - \frac{1}{2}\sum_{t=\nu+1}^{N}(y[t]-\hat{\mu}_2)^2\right]}{\exp\left[-\frac{1}{2}\sum_{t=1}^{N}(y[t]-\hat{\mu})^2\right]} \\
&= \max_{\nu} \ln \frac{\exp\left[-\frac{1}{2}S_\nu\right]}{\exp\left[-\frac{1}{2}S\right]} \\
&= \max_{\nu} \frac{1}{2}(-S_\nu + S) \\
&= \max_{\nu} \frac{1}{2} V(\nu, N)
\end{aligned}$$

introducing $V(\nu, N) = S - S_\nu$. The following algebra allows a simplification of $V(\nu, N)$.

$$\begin{aligned}
V(\nu, N) &= S - S_\nu \\
&= \sum_{t=1}^{N}(y[t]-\hat{\mu})^2 - \sum_{t=1}^{\nu}(y[t]-\hat{\mu}_1)^2 - \sum_{t=\nu+1}^{N}(y[t]-\hat{\mu}_2)^2 \\
&= \sum_{t=1}^{N} y[t]^2 - 2\hat{\mu}\sum_{t=1}^{N} y[t] + N\hat{\mu}^2 \\
&\quad - \sum_{t=1}^{\nu} y[t]^2 + 2\hat{\mu}_1 \sum_{t=1}^{\nu} y[t] - \nu\hat{\mu}_1^2 \\
&\quad - \sum_{t=\nu+1}^{N} y[t]^2 + 2\hat{\mu}_2 \sum_{t=\nu+1}^{N} y[t] - (N-\nu)\hat{\mu}_2^2
\end{aligned} \tag{2.73}$$

splitting the sum over the samples

$$\begin{aligned}
V(\nu, N) &= -2\hat{\mu}\left(\sum_{t=1}^{\nu} y[t] + \sum_{t=\nu+1}^{N} y[t]\right) + \nu\hat{\mu}^2 + (N-\nu)\hat{\mu}^2 \\
&\quad + 2\hat{\mu}_1 \sum_{t=1}^{\nu} y[t] - \nu\hat{\mu}_1^2 + 2\hat{\mu}_2 \sum_{t=\nu+1}^{N} y[t] - (N-\nu)\hat{\mu}_2^2 \\
&= -2\nu\hat{\mu}_1\hat{\mu} - 2(N-\nu)\hat{\mu}_2\hat{\mu} + \nu\hat{\mu}^2 + (N-\nu)\hat{\mu}^2 \\
&\quad + 2\nu\hat{\mu}_1^2 - \nu\hat{\mu}_1^2 + 2(N-\nu)\hat{\mu}_2^2 - (N-\nu)\hat{\mu}_2^2 \\
&= -2\nu\hat{\mu}_1\hat{\mu} - 2(N-\nu)\hat{\mu}_2\hat{\mu} + \nu\hat{\mu}^2 + (N-\nu)\hat{\mu}^2 + \nu\hat{\mu}_1^2 + (N-\nu)\hat{\mu}_2^2 \\
&= \nu\left(\hat{\mu}_1^2 - 2\hat{\mu}_1\hat{\mu} + \hat{\mu}^2\right) + (N-\nu)\left(\hat{\mu}_2^2 - 2\hat{\mu}_2\hat{\mu} + \hat{\mu}^2\right) \\
&= \nu(\hat{\mu}_1 - \hat{\mu})^2 + (N-\nu)(\hat{\mu}_2 - \hat{\mu})^2
\end{aligned} \tag{2.74}$$

so that finally the change-point hypothesis $\mathcal{H}_1$ is accepted if

$$\max_{\nu} \frac{1}{2} V(\nu, N) > \ln \delta \tag{2.75}$$

or

$$\max_{\nu} V(\nu, N) > \delta' \tag{2.76}$$

where $V(\nu, N)$ is given by (2.74).

2.2.5. On-line Change Detection

In an on-line setup the change in mean should be detected during data acquisition. This is a common task in process control where information produced by the sensors is employed to detect abrupt changes in process variables. The consequence of detecting a change might be different. In case of a severe failure the process will be shut down while a small change will effect that the control law is adapted. The time when the change is detected is commonly referred as the alarm time denoted by t_a. Let n be the current time then a change is detected utilising the GLRT if $\max_\nu V(\nu, n)$ exceeds the threshold δ. The stopping rule is consequently formulated

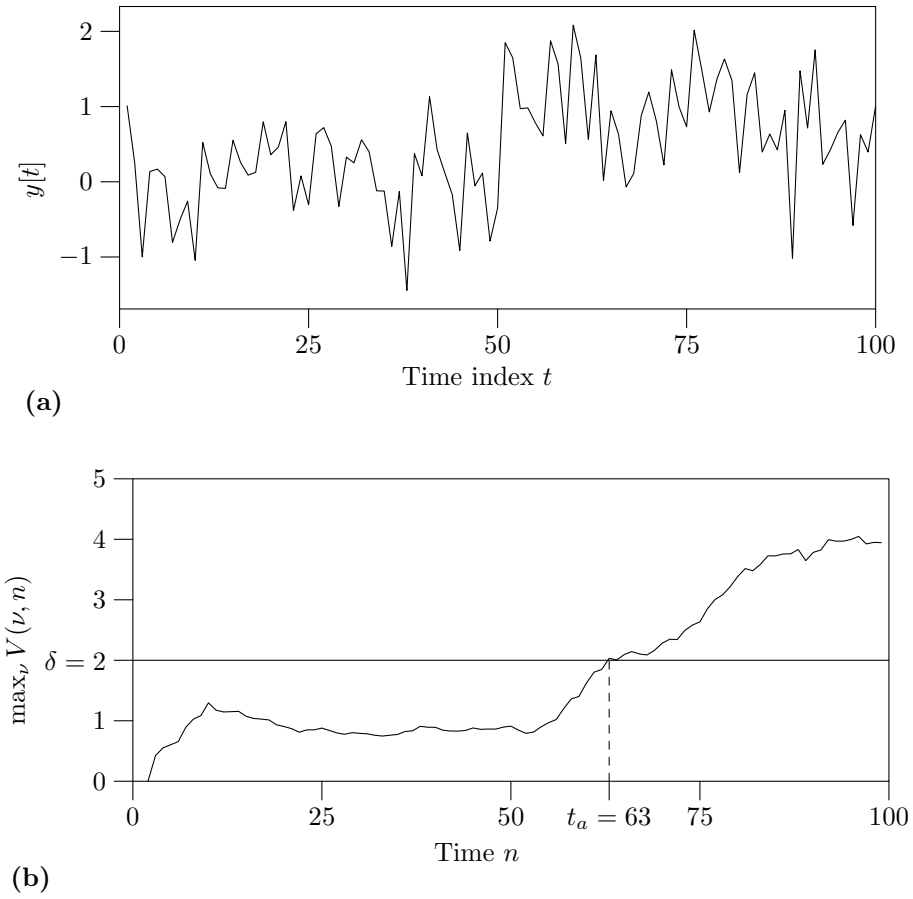

Figure 2.5.: (a) A signal corrupted by WGN with a change in DC level at $\nu_0 = 50$. (b) The test statistic $\max_\nu V(\nu, n)$. For a threshold $\delta = 2$ the GLRT detects the change at $t_a = 63$.

as

$$t_a = \min\{n \,:\, \max_{\nu} V(\nu, n) > \delta\} \,. \tag{2.77}$$

Figure 2.5 displays an example where the GLRT is applied to a simulated signal with a change at $\nu_0 = 50$. It is evident that the test statistic is less noisy than the signal. This is due to the sums in $V(\nu, n)$ reducing stochastic errors. After the change has happened at $\nu_0 = 50$ the test statistic rises so that the GLRT detects the change at $t_a = 63$ when a threshold $\delta = 2$ is used.

2.3. Classification

2.3.1. Introduction

In this section the previous detection scenario is extended to the case where one wishes to distinguish between c hypotheses, with $c > 2$. Such a problem arises quite frequently in communications, in which one of c signals must be detected where the observed signal is classified to one of the c signals. Beside *classification* this problem is sometimes termed *multiple hypotheses testing* or simply *signal detection*.

2.3.2. Minimum Probability of Error

Assume that one wishes to decide among the c possible hypotheses $\{\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_c\}$. A commonly used criterion is the *minimum probability of error*. The *probability of error* P_e is the probability that the decision is wrong and is defined as

$$P_e = \sum_{i=1}^{c} \sum_{\substack{j=1 \\ j \neq i}}^{c} P(\mathcal{H}_i | \, \mathcal{H}_j) P(\mathcal{H}_j) \tag{2.78}$$

where $P(\mathcal{H}_j)$ is the a priori probability for hypothesis j and $P(\mathcal{H}_i | \, \mathcal{H}_j)$ is the conditional probability deciding for hypothesis i when hypothesis j is true. Aim is to find a decision criterion that minimises P_e. Let $R_i = \{\boldsymbol{y} : \textit{decide } \mathcal{H}_i\}$ be a subset of $\mathbb{R}^N$ where each $\boldsymbol{y} \in R_i$ is decided to be a realisation of hypothesis i. R_i is termed the *decision region* of $\mathcal{H}_i$. The partitioning of the space $\mathbb{R}^N$ is non overlapping whereas the union extends to the whole space, so that

$$\bigcup_{i=1}^{c} R_i = \mathbb{R}^N \quad i = 1, 2, \ldots, c \, . \tag{2.79}$$

From this it follows that $P(\mathcal{H}_i | \, \mathcal{H}_j)$ can be rewritten to

$$P(\mathcal{H}_i | \, \mathcal{H}_j) = \int_{R_i} p(\boldsymbol{y} | \, \mathcal{H}_j) d\boldsymbol{y} \tag{2.80}$$

so that

$$P_e = \sum_{i=1}^{c} \sum_{\substack{j=1 \\ j \neq i}}^{c} \int_{R_i} p(\boldsymbol{y}|\, \mathcal{H}_j) d\boldsymbol{y} P(\mathcal{H}_j) \tag{2.81}$$

$$= \sum_{i=1}^{c} \int_{R_i} \sum_{\substack{j=1 \\ j \neq i}}^{c} p(\boldsymbol{y}|\, \mathcal{H}_j) P(\mathcal{H}_j) d\boldsymbol{y} \; . \tag{2.82}$$

Utilising Bayes law

$$P(\mathcal{H}_j|\, \boldsymbol{y}) = \frac{p(\boldsymbol{y}|\, \mathcal{H}_j) P(\mathcal{H}_j)}{p(\boldsymbol{y})} \tag{2.83}$$

(2.82) is equal to

$$P_e = \sum_{i=1}^{c} \int_{R_i} \sum_{\substack{j=1 \\ j \neq i}}^{c} P(\mathcal{H}_j|\, \boldsymbol{y}) p(\boldsymbol{y}) d\boldsymbol{y} \; . \tag{2.84}$$

The term

$$C_i^* = \sum_{\substack{j=1 \\ j \neq i}}^{c} P(\mathcal{H}_j|\, \boldsymbol{y}) p(\boldsymbol{y}) d\boldsymbol{y} \tag{2.85}$$

is the contribution to P_e if the observation $\boldsymbol{y}$ is assigned to R_i. In order to minimise P_e, $\boldsymbol{y}$ should assigned to that decision region so that the contribution C_i^* is minimal. Since $p(\boldsymbol{y})$ in C_i^* is constant, it can be cancelled so that in summary the *minimum probability error detector* decides $\mathcal{H}_i$ for which

$$C_i(\boldsymbol{y}) = \sum_{\substack{j=1 \\ j \neq i}}^{c} P(\mathcal{H}_j|\, \boldsymbol{y}) \tag{2.86}$$

is minimal. C_i is termed the cost of deciding $\mathcal{H}_i$ when $\boldsymbol{y}$ is observed.

There is a strong connection of the minimum P_e detector to a maximum likelihood based detector. To show this, $C_i(\boldsymbol{y})$ can be rewritten to

$$C_i(\boldsymbol{y}) = \sum_{j=1}^{c} P(\mathcal{H}_j|\, \boldsymbol{y}) - P(\mathcal{H}_i|\, \boldsymbol{y}) \tag{2.87}$$

where the sum $\sum_{j=1}^{c} P(\mathcal{H}_j|\, \boldsymbol{y})$ is the probability that $\boldsymbol{y}$ is assigned to any hypothesis, which is one because of (2.79); $C_i(\boldsymbol{y})$ is therefore minimised by maximising $P(\mathcal{H}_i|\, \boldsymbol{y})$.

Thus, the minimum P_e decision rule is equivalent to decide $\mathcal{H}_i$ if

$$P(\mathcal{H}_i|\,\boldsymbol{y}) > P(\mathcal{H}_j|\,\boldsymbol{y}) \quad \text{with} \quad j = 1, 2, \ldots, c\ ;\ i \neq j\ . \tag{2.88}$$

With Bayes law (2.83) it follows that the minimum P_e decides for $\mathcal{H}_i$ if the product of the likelihood $p(\boldsymbol{y}|\,\mathcal{H}_i)$ and the prior probability $P(\mathcal{H}_i)$ is maximal

$$p(\boldsymbol{y}|\,\mathcal{H}_i)P(\mathcal{H}_i) > p(\boldsymbol{y}|\,\mathcal{H}_j)P(\mathcal{H}_j) \quad \text{with} \quad j = 1, 2, \ldots, c\ ;\ i \neq j\ . \tag{2.89}$$

In case that the prior probabilities are equal so that $P(\mathcal{H}_j) = 1/c$ the decision reduces to comparing the likelihoods. This is the maximum likelihood (ML) decision rule which decides for $\mathcal{H}_i$ if

$$p(\boldsymbol{y}|\,\mathcal{H}_i) > p(\boldsymbol{y}|\,\mathcal{H}_j) \quad \text{with} \quad j = 1, 2, \ldots, c\ ;\ i \neq j\ . \tag{2.90}$$

The ML decision rule is often used in applications where the prior probabilities are not known, so that there is no argument against the uniform distribution. There is an interesting analogy to the MLE (see Section 2.1.4). In fact the MLE is equal to the ML decision rule in case that $\boldsymbol{\theta}$ represents a finite set.

The ML decision rule is illustrated on the problem of deciding among three DC levels. Assume that one of the three hypotheses

$$\mathcal{H}_1 : y[t] = 1 + w[t] \tag{2.91}$$

$$\mathcal{H}_2 : y[t] = 2 + w[t] \tag{2.92}$$

$$\mathcal{H}_3 : y[t] = 3 + w[t] \quad \text{with} \quad t = 1, 2, \ldots, N \tag{2.93}$$

are possibly observed, where $w[t]$ is WGN with variance σ^2. Furthermore, if the prior probabilities are equal or $P(\mathcal{H}_1) = P(\mathcal{H}_2) = P(\mathcal{H}_3) = 1/3$, then the ML decision rule applies. Consider first the simple case of $N = 1$ with the PDFs shown in Figure 2.6. By symmetry it is clear from (2.90) that the ML decision rule decides for $\mathcal{H}_2$ if $y[1]$ is in the range $R_2 = [1.5, 2.5]$, for $\mathcal{H}_1$ if $y[1]$ is at the left and $\mathcal{H}_3$ if

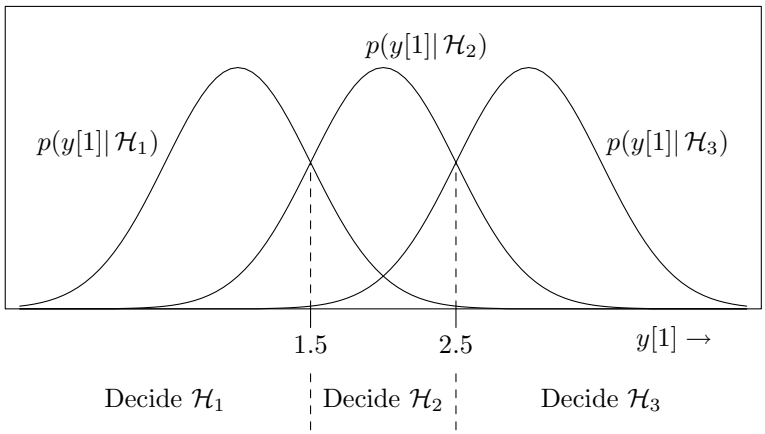

Figure 2.6.: The PDFs and the decision regions for the three DC levels example (N=1).

$y[1]$ is at the right of this range.

In order to get the decision regions for multiple samples ($N > 1$) the conditional PDFs must be evaluated. The PDF $p(\boldsymbol{y}|\,\mathcal{H}_i)$ is analogue to (2.4)

$$p(\boldsymbol{y}|\,\mathcal{H}_i) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{t=1}^{N}(y[t]-i)^2\right] . \tag{2.94}$$

The decision of (2.90) remains unchanged when taking the logarithm on both sides since it is a strictly increasing function, so that with (2.94) the ML rule decides for

$\mathcal{H}_i$ if

$$\ln p(\boldsymbol{y}|\,\mathcal{H}_i) > \ln p(\boldsymbol{y}|\,\mathcal{H}_j) \tag{2.95}$$

$$\ln \frac{1}{(2\pi\sigma^2)^{N/2}} + \left[-\frac{1}{2\sigma^2}\sum_{t=1}^{N}(y[t]-i)^2\right] > \ln \frac{1}{(2\pi\sigma^2)^{N/2}} + \left[-\frac{1}{2\sigma^2}\sum_{t=1}^{N}(y[t]-j)^2\right] \tag{2.96}$$

$$-\frac{1}{2\sigma^2}\sum_{t=1}^{N}(y[t]-i)^2 > -\frac{1}{2\sigma^2}\sum_{t=1}^{N}(y[t]-j)^2 \tag{2.97}$$

$$\sum_{t=1}^{N}(y[t]-i)^2 < \sum_{t=1}^{N}(y[t]-j)^2$$

$$\text{with} \quad j = 1,2,3\ ;\ i \neq j\ . \tag{2.98}$$

The final inequality has a nice geometrical interpretation. The sum of squares is the squared *Euclidean norm* a commonly used distance measure in $\mathbb{R}^N$ which means that $\boldsymbol{y}$ is assigned to the closest hypothesis.

2.3.3. Time Varying Templates

In the example of the previous section there are three simple hypotheses representing different DC levels. This example will now be extended to the case of c hypotheses $\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_c$ representing the signals $\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_c$. This means, when $\mathcal{H}_i$ is true the observation $y[t]$ is equal to $p_i[t]$ corrupted by WGN $w[t]$

$$\mathcal{H}_i : y[t] = p_i[t] + w[t] \quad \text{with} \quad t = 1,2,\ldots,N\ . \tag{2.99}$$

The special case when $p_i[t]$ is constant is treated in the example of the previous section. Proceeding in an analogous way with

$$p(\boldsymbol{y}|\,\mathcal{H}_i) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{t=1}^{N}(y[t]-p_i[t])^2\right] \tag{2.100}$$

the ML rule decides for $\mathcal{H}_i$ if

$$\sum_{t=1}^{N} \left(y[t] - p_i[t]\right)^2 < \sum_{t=1}^{N} \left(y[t] - p_j[t]\right)^2 \quad \text{with} \quad j = 1, 2, \ldots, c \; ; \; i \neq j \tag{2.101}$$

or in a vectorial notation

$$(\boldsymbol{y} - \boldsymbol{p}_i)^T (\boldsymbol{y} - \boldsymbol{p}_i) < (\boldsymbol{y} - \boldsymbol{p}_j)^T (\boldsymbol{y} - \boldsymbol{p}_j) \quad \text{with} \quad j = 1, 2, \ldots, c \; ; \; i \neq j \; . \tag{2.102}$$

After factorisation and cancelling the constant term $\boldsymbol{y}^T\boldsymbol{y}$ it remains

$$-2\boldsymbol{y}^T\boldsymbol{p}_i + \boldsymbol{p}_i{}^T\boldsymbol{p}_i < -2\boldsymbol{y}^T\boldsymbol{p}_j + \boldsymbol{p}_j{}^T\boldsymbol{p}_j \quad \text{with} \quad j = 1, 2, \ldots, c \; ; \; i \neq j \tag{2.103}$$

and, after resorting, one can get the form

$$2(\boldsymbol{p}_j - \boldsymbol{p}_i)^T\boldsymbol{y} < \boldsymbol{p}_j{}^T\boldsymbol{p}_j - \boldsymbol{p}_i{}^T\boldsymbol{p}_i \quad \text{with} \quad j = 1, 2, \ldots, c \; ; \; i \neq j \; . \tag{2.104}$$

Finally these inequalities can be written in a convenient matrix form

$$A_i\boldsymbol{y} < b \tag{2.105}$$

with

$$A_i = \begin{pmatrix} 2(\boldsymbol{p}_1 - \boldsymbol{p}_i)^T \\ \vdots \\ 2(\boldsymbol{p}_{i-1} - \boldsymbol{p}_i)^T \\ 2(\boldsymbol{p}_{i+1} - \boldsymbol{p}_i)^T \\ \vdots \\ 2(\boldsymbol{p}_c - \boldsymbol{p}_i)^T \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} \boldsymbol{p}_1{}^T\boldsymbol{p}_1 - \boldsymbol{p}_i{}^T\boldsymbol{p}_i{}^T \\ \vdots \\ \boldsymbol{p}_{i-1}{}^T\boldsymbol{p}_{i-1} - \boldsymbol{p}_i{}^T\boldsymbol{p}_i{}^T \\ \boldsymbol{p}_{i+1}{}^T\boldsymbol{p}_{i+1} - \boldsymbol{p}_i{}^T\boldsymbol{p}_i{}^T \\ \vdots \\ \boldsymbol{p}_c{}^T\boldsymbol{p}_c - \boldsymbol{p}_i{}^T\boldsymbol{p}_i{}^T \end{pmatrix} . \tag{2.106}$$

The solution set of the inequality $A_i\boldsymbol{y} < b$ is the decision region R_i and every $\boldsymbol{y} \in R_i$ will be assigned to $\mathcal{H}_i$. In linear algebra, the solution space of $A_i\boldsymbol{y} < b$ is called *convex polyhedron*. In case that R_i is bounded, the term *convex polytope* is

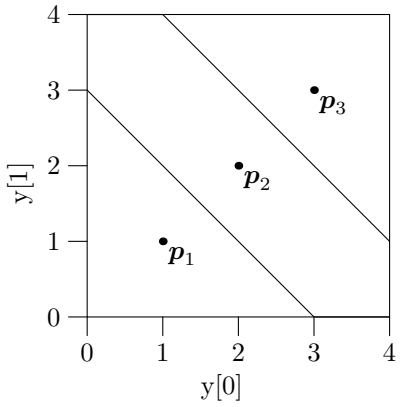

Figure 2.7.: The Voronoi diagram for the three DC levels example (N=2).

commonly used. The faces of the polyhedron R_i are hyperplanes of the form

$$2(\boldsymbol{p}_j - \boldsymbol{p}_i)^T \boldsymbol{y} = \boldsymbol{p}_j{}^T \boldsymbol{p}_j - \boldsymbol{p}_i{}^T \boldsymbol{p}_i \quad \text{with} \quad j = 1, 2, \ldots, c\,;\ i \neq j \tag{2.107}$$

they are also termed *decision boundaries*. A diagram that displays the decision boundaries and the vectors $\boldsymbol{p}_i$ is called a *Voronoi diagram*. The Voronoi diagram for the three DC levels example (N=2) is displayed in Figure 2.7. In this example the three hypotheses are represented by

$$\begin{aligned} \boldsymbol{p}_1 &= (1,1)^T \\ \boldsymbol{p}_2 &= (2,2)^T \\ \boldsymbol{p}_3 &= (3,3)^T \end{aligned} \tag{2.108}$$

whereas the decision boundaries are simply straight lines.

3. State of the Art in Multiple Change Detection

3.1. Introduction

In a survey about the change-point problem Dragalin [1997] pointed out, that the change-point problem can be considered to be one of the central problems of statistical inference. One of the pioneering works is due to Shewhart [1931], who proposed a control chart for detecting a change in the quality characteristics of a manufacturing process. Since that time, the subject has undergone tremendous growth, both in the scope of its applications and in methodological advances.

The currently used, sometimes irritating, terminology is a result of this process. Several terms are used for the change-point, among them are break-point, jump-point and switch-point. Similarly the change-point problem is called fault detection, change detection and segmentation depending on the objective and application of the analysis. The following classification of the change-point problem tries to clarify the differences in terminology.

1. **Objective** The change-point problem can be considered as either the problem of *estimating the location* of the change-point, or the problem of *quickest on-line detection* of a change, which is also termed *fault detection* or *change detection*.

   The former is a typical *a posteriori* or *retrospective* problem, performed when the process of data acquisition is completed, whereas the latter is performed on-line with the process of data acquisition, which is termed *sequential* or

prospective change-point problem.

2. **Level of a priori information** Depending on available a priori information one can distinguish between *parametric* and *non-parametric* methods. Parametric statistical methods are mathematical procedures for statistical hypothesis testing which assume that the distributions of the variables being assessed belong to known parametrised families of probability distributions. In that case we speak of a parametric model. Non-parametric methods were developed to be used in cases when the researcher has limited knowledge about the systems structure.

3. **Characteristics of data** Change-points may occur *temporal*. E. g., machinery that is operating satisfactorily may all of a sudden experience difficulty, which is modelled by a random process where the change-point is considered as a moment in time when some characteristics of the process change. Changes in system parameters can occur *spatially* as well. Examples are the change in the speed of sound as a sound wave traverses a boundary such as an air-water interface, which is modelled by a change in a random field.

   Furthermore the observations can be statistically *independent* or *dependent*.

4. **Type of changes** The majority of publications describe *abrupt changes* in the characteristics of observations. Abrupt is meant in the sense that changes in characteristics occur very fast with respect to the sampling period of the measurements. Smooth change transitions covering more than a few samples are termed *gradual changes*.

   Depending on the number of change-points, one can distinguish between *single change* and *multiple change-point* problems.

The focus of this thesis is on the multiple change-point problem in the temporal characteristics of a random-process with the objective of testing for a change and dating the change-point. This is well known as the *segmentation* problem.

In most physical systems segmentation means to detect and locate gradual changes rather than abrupt jumps. E. g., position signals in classical mechanics cannot comprise jumps because of the ubiquitous masses in these systems. A jump would mean that the kinetic energy tends to infinity at this moment which is inconsistent with the law of conservation of energy.

Since even a break needs time to happen, models with gradual changes should be used for these systems out of the field of classical mechanics. They are commonly modelled by linear, often time invariant, dynamical systems. For a band limited input signal these systems give a band limited output allowing to sample the signal with the possibility of a complete reconstruction from the samples. The well known Shannon sampling theorem states that an exact reconstruction is possible when the sampling frequency is greater than twice of the signal's bandwidth [Shannon, 1948]. When the Shannon sampling theorem is attained, the monitored signal will comprise smooth gradual changes from one state to another.

Detecting changes in dynamical systems is rather complex [Willsky and Jones, 1976] so that often low order models are used. The computational effort is usually high, since for each possible change-point the parameters of the dynamical system model have to be estimated. An objective cost function like the mean squared error is then used to give an estimate for the change-point [Björklund and Ljung, 2003]. Methods commonly used for system identification are the instrumental variable method [Söderström and Stoica, 2002] and the prediction error method [Ljung, 2002].

In change-point analysis this computational effort is often avoided by approximating the signal with piecewise linear functions, so that every segment between two change-points is described by two parameters, only. When the adjoining linear functions meet at the change-point, the model is a continuous function which is termed *continuous change-point model*.

Although continuous change-point models do model the continuous nature of physical systems more precisely, a lot of work has been done for piecewise linear models where the adjoining linear functions do not necessarily meet at the change-

point. These models comprising abrupt changes are termed *discontinuous change-point models*. Their usage is valid in two cases. (i) The discontinuous change-point model is not directly used for the monitored signal. Instead, it is used to model a systems internal parameter. A method based on a model which is a dynamical system with switching parameters was published, e. g., by Timmer and Pignatiello Jr. [2003]. While the output signal of the dynamical system changes smoothly, its internal parameters may change abruptly. (ii) When the bias of the wrong modelling is within the desired accuracy bounds, algorithms based on discontinuous change-point models are especially interesting for large-scale problems. The state of the art in change-point analysis is that there are efficient algorithms for the segmentation of large-scale *discontinuous* problems, but the segmentation of large-scale *continuous* problems in reasonable time is still an open problem, which is not solved satisfactorily. The algorithmic differences will be discussed in the next two sections.

Finally, it should be mentioned that other than physical systems may comprise abrupt changes. E. g., in econometrics when an event happens during the closing hours of the stock exchange. The share prices will then jump suddenly after the opening of the stock exchange, see e. g. Muggeo [2003]. A second example are undocumented relocations of meteorological stations, effecting the measured climatological time series like temperature or pressure, see e. g. Ducré-Robitaille et al. [2003]. Algorithms based on discontinuous change-point models are then used to homogenise these climatological time series.

3.2. Discontinuous Change-Point Models

A *discontinuous change-point model* allows the regression function to be discontinuous in the change point which are sometimes referred as *jump models*. Most prominent is the single step model where the epochs before and after the change are restricted to be constant, which was first investigated by Page [1955]. A discontinuous model with not constant but linear pre- and post-change function was

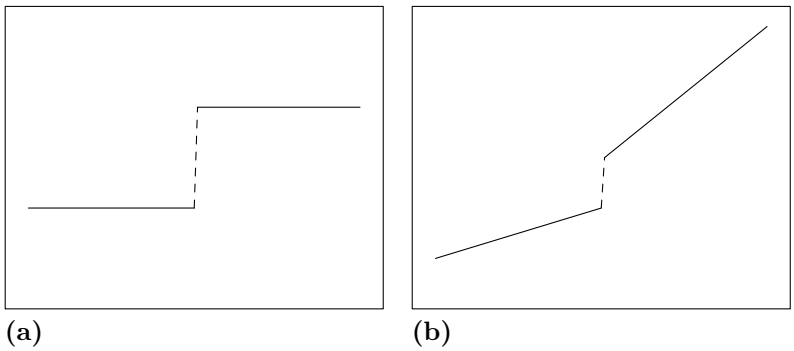

Figure 3.1.: Two prominent discontinuous change-point models. (a) the step model and (b) the two phase jump model.

first considered by Hudson [1966]. In literature, this model is called the *two-phase jump model*. Both models are illustrated in Figure 3.1.

In cognitive science the two phase model was utilised by Beem [1995]. For the same application, namely for modelling strategy shifts, Luwel et al. [2001] had used a three phase model comprising two change-points (see Figure 1.3 in the introduction of the thesis). Similar models are used for homogenising climatological series, too. Inhomogeneities in, e. g., temperature series are often caused by non climate factors such as: changes in measurement practices, station relocations, changes in the surroundings of a station over the years, etc. [Ducré-Robitaille et al., 2003, Lund and Reeves, 2002, Wang, 2003]. Therefore, it becomes essential to take these factors into account in order to retain only the climate signal of interest. To accomplish this, abrupt changes in climatological series are identified and the time series is then homogenised. Ducré-Robitaille et al. [2003] compare methods for homogenising climatological series of artificial data with multiple inhomogeneities, where only methods designed for at most one change-point are incorporated.

Most algorithms that are used in cognitive and geophysical sciences are *grid search* algorithms. They compute for each segmentation a real number which represents the goodness of fit. The computational effort depends then on the number of possible segmentations. Following the notation that $P_{N,K}$ is the set of possible segmentations for a signal of length N with K change-points, the size of this set,

denoted by $\#P_{N,K}$, is equal to the binomial coefficient. This is a well known fact in enumerative combinatorics, since the number of possible segmentations for a signal of size N with K change-points is equal to the number of possible combinations without repetitions where the order does not matter. Thus, the number of possible segmentations is defined by

$$\begin{aligned} \#P_{N,K} &= \binom{N-1}{K} \\ &= \frac{(N-1)!}{K!(N-1-K)!} \,. \end{aligned} \tag{3.1}$$

Since the signals in the previously mentioned applications, namely cognitive and geophysical sciences, are short with just a view change-points, this number is small which allows to solve these problems with grid search in reasonable time.

The number calculated by (3.1) gives the number of segmentations when K is known. The number of segmentations when K is unknown, denoted by $\#P_N$, is the sum over the segmentations for fixed K, which is

$$\#P_N = \sum_{K=1}^{K=N} \#P_{N,K} \tag{3.2}$$

$$= 2^{N-1} \,. \tag{3.3}$$

This exponential growth can be explained easily. When a signal is prolonged by one sample, this sample can be a change-point or not, doubling the possible number of segmentations in comparison to the original signal. With the fact that a signal of length two has two possible segmentation, which are a single change (the values of the two samples are different) and no change, it is evident that the exponent in (3.3) has to be $N-1$.

Because of the exponential growth, the number of possible segmentations rises drastically for large sample sizes, which triggers the desire for more efficient algorithms. An efficient algorithm is the *binary splitting* algorithm introduced by Vostrikova [1981]. The binary splitting algorithm is retrospective and works recursively. First, a statistical test for a single change-point is applied on the complete

series. If the change hypothesis is accepted, the two subsequences before and after the change-point are processed in the same manner, separately. In comparison to grid search, the number of segmentations considered by this algorithm is less than N^2 which is the main reason for its computational efficiency.

Recently another method that scales well with the sample size has gained much attention. It is the *dynamic programming* method successfully applied to change detection by Bai and Perron [2003] and Hawkins [2001] (reviewed in Zeileis et al. [2003]). The dynamic programming algorithm is build on the *Markov property* that is fulfilled for discontinuous change-point models, since the signal properties on the segments are independent from each other. Dynamic programming decreases the computational complexity to a more manageable level, i. e., it is proven that dynamic programming finds the same segmentation that grid search would choose, by considering a subset of the possible segmentations, only. But, in contrast to grid search and binary splitting the computational effort of dynamic programming increases linearly with the sample size, making it the method of choice. However, when the model is continuous at the change-point so that the Markov property is not fulfilled, dynamic programming may give suboptimal results, i. e., it does not necessarily choose the same segmentation as grid search.

3.3. Continuous Change-Point Models

Continuous change-point models comprise smooth, continuous change transitions. They are sometimes referred as *join models*. Early work was contributed by Hudson [1966] and Hinkley [1969] who utilised the two phase join model, depicted in Figure 3.2 (a).

Beside the classical two-phase model, a second continuous change-point model namely the ramp-step model gained much attention (see Figure 3.2 (b)). The ramp-step model is, e. g., used by Friede et al. [2001] to model dose-response curves. A dose-response curve relates the amount of a drug or toxin given, to the response of the organism to that drug. The first point where a response above

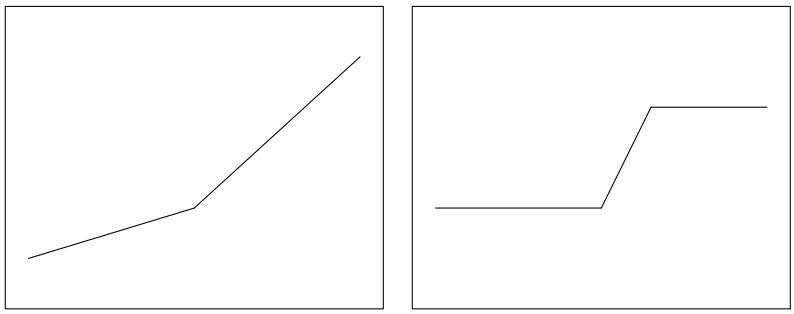

Figure 3.2.: (a) The two phase join model which is the most used continuous change-point model. (b) The ramp-step, a three phase join model is a generalisation of the step model (see Figure 3.1).

zero is reached, is usually referred to as the threshold-dose. Aim is to identify the threshold-dose and the slope of the dose-response curve for doses greater than that [Pastor-Barriuso et al., 2003].

Another application where the ramp-step model is utilised is the modelling of bacterial growth. Aim is to identify the parameters of the bacterial growth curve which is roughly divided into the lag, exponential growth and a stationary growth phases. During the lag phase, the cells are assumed to be non-replicating, as they adapt themselves to their environment. Once adapted, the cells begin to grow at a rate that is maximal for the microorganism in the specific environment. Once the stationary phase has been reached, there is no increase in population and the specific growth rate (the logarithm of the growth rate) returns to zero [Buchanan et al., 1997, Garthright, 1997, Lopez et al., 2004]. Figure 3.3 shows a measurement of the specific growth rate together with fits of the ramp-step model (solid line) and two smooth change models (dashed lines).

In geophysical science, the ramp-step model is not utilised so often, although examples can be found, e. g., Mudelsee [2000], see Figure 3.4. In this scientific area the more general piecewise linear join model is utilised more frequently [Solow, 1987, Tome and Miranda, 2004], see Figure 1.4 in the introduction of the thesis. The piecewise linear join model is an extension to the two-phase join model comprising

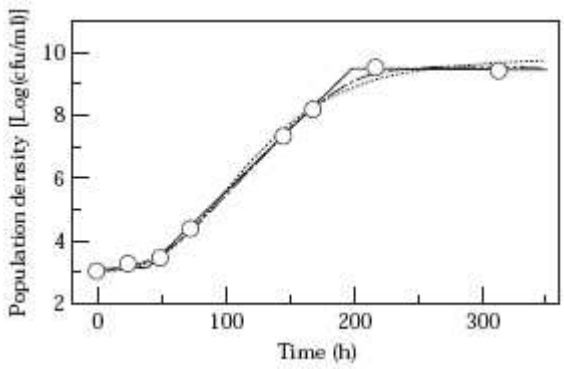

Figure 3.3.: The fits of the ramp-step model (solid line) compared to two smooth models (dashed lines). The circles are measurements of the specific bacterial growth rate (source: Buchanan et al. [1997])

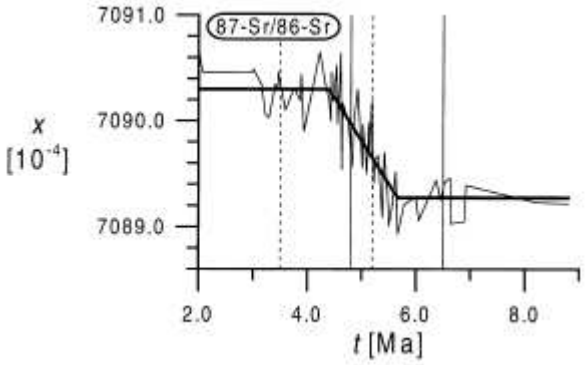

Figure 3.4.: The ratio of strontium isotopes 87-Sr/86-Sr (solid line) which documents the geochemical cycling of strontium used to infer hydrothermal circulation at mid-ocean ridges in the late Neogene ocean around two to eight million years ago. The heavy line is the fit of the ramp-step function. (source: Mudelsee [2000])

more than two linear phases. Recent work is contributed by Tome and Miranda [2004] fitting a piecewise linear join model to climate data. The proposed algorithm is computationally expensive, since a least squares fit must be performed for every possible segmentation. The method of Tome and Miranda [2004] has the benefit that the global optimum is found at the cost of a high computational burden. However, for problems where the number of possible segmentations is small, this is the method of choice. Note, that the methods used to fit the two-phase join model and the ramp-step model can be viewed as variations of the generally applicable approach of Tome and Miranda [2004].

A method that scales well with the signal size was proposed by Charbonnier et al. [2004]. It is a sequential method which finds application in the analysis of physiological parameters of a patients state in an intensive care unit (ICU). Analysing ICU signals is quite challenging. Changes can be very quick, or rather slow. Variations can occur on the monitored signal that do not correspond to a physiological change but that are due to extraneous causes (measurement artefacts, patient turning in bed, coughing, etc.) [Avent and Charlton, 1990]. Therefore, Charbonnier et al. [2004] utilised a rather complex general piecewise linear model where the function might be continuous or discontinuous in the change-points. The algorithm of Charbonnier et al. [2004] is a two stage method using an on-line change detection algorithm at the first stage. Once a change is detected, the change-point is estimated at the second stage. This sequential approach of testing for a change and estimating the change-point is a common strategy in change-point analysis which was utilised before by, e. g., Staude [2001]. This strategy allows to solve large-scale problems in reasonable time, since not every possible segmentation is considered. However, since it does a locally optimal estimation of the change-point, the segmentation is satisfactory in many applications.

The method of Charbonnier et al. [2004] faces the problem that the estimation step involves the decision whether the change is continuous or not, solved in an heuristic way, which is computationally efficient but does not fulfil an optimisation criterion and might therefore give suboptimal results.

4. The Large-Scale Gradual Change Detection Problem

The algorithm presented in this thesis should be able to automatically perform a segmentation of a signal composed of adjacent gradual changes, modelled by linear transitions between two constant phases, namely the ramp-step model (see Figure 3.2 (b)). The algorithm should satisfy two important requests, (i) to process large-scale signals in reasonable time, and (ii) to be reliable and easily tunable.

Whether a signal is large-scale depends on the length of the signal as well as on the number of change-points, since both have an influence on the number of possible segmentations which is an objective measure for the size of the problem. But, whether the runtime needed for solving the problem is reasonable depends on the application. Reasonable time can therefore be months when, e. g., computing climate models, or just a millisecond for, e. g., on-line applications. The algorithm presented in this thesis should be particularly well suited for biomechanical signals from psychophysiological experiments. In this application, signals are traditionally analysed by visual inspection. In order that the algorithm will gain acceptance, reasonable time means to be at least as fast as a well trained operator. So that for a signal with 100 change-points assuming that the operator's decision time is 2s per change-point the desired runtime is approximately 3 min or less.

The second request demands that the influence of the tuning parameters on the achieved segmentation is predictable for the operator. This implies a clear coupling of the segmentation to the tuning parameters as well as a definite and reliable guideline how tuning parameters should be adopted to a particular segmentation task.

Straight forward and transparent tuning is an important contribution to the reliability of an algorithm. Beside that, a reliable algorithm should indicate when its results have to be taken with caution. In signal processing it is crucial to know how random errors obfuscate the results, i. e., to be aware of the behaviour of the algorithm for signals with a low SNR.

The ancestor of the ramp-step model is the step model so that it seems to be natural to review algorithms based on the step model when dealing with the segmentation of ramp-step change profiles. Indeed, algorithms like binary splitting and dynamic programming might be applied with good results when the transition phase covers only a few samples, however, since the problem should not be restricted to fast transitions these algorithms give biased results since they do not have the necessary complexity of the ramp-step model.

An algorithm based on the piecewise linear join model could be used for the segmentation of a signal composed of ramp-steps. Since the ramp-step is a special case of the piecewise linear model, it can be expected that algorithms based on the more general piecewise linear model give unbiased results. The globally optimal solution to the piecewise linear segmentation problem can be obtained by the algorithm published by Tome and Miranda [2004]. This method reviews every possible segmentation and chooses the optimal one utilising the least squares method. As explained in Section 3.3 this complete search strategy does not scale well with the size of the problem. The size of the problem is given by the number of possible segmentations defined by (3.1). Which means that for the climatological time series addressed in [Tome and Miranda, 2004], the number of segmentations to be tested is 4,950 ($N = 100$ and $K = 2$). However, in a large-scale application, e. g. tapping, this number rises significantly. So that for a one minute recording of a typical tapping signal (see Cong Khac et al. [2007]) approximately 7.67×10^{453} different segmentations might be found ($N = 60,000$ and $K \approx 150$). With the current computer architecture and the usually estimated annually doubling of speed, this problem cannot be solved completely in reasonable time in the foreseeable future. Therefore, grid search algorithms cannot be used, instead an algorithm which

does not enumerate every possible segmentation must be applied. A modern one is a sequential algorithm and was published by Charbonnier et al. [2004], which is based on the general piecewise linear model. In order to satisfy the complexity of this model, the method of Charbonnier et. al. has several shortcomings. So is the localisation of a change-point suboptimal since it is based on a heuristic and the tuning process is obfuscated by the necessity to define additional parameters.

The algorithm presented in this thesis overcomes these shortcomings by attaining locally optimal dating of the change-point as well as providing a transparent tuning. The details are described in the chapters 5 and 6 followed by Chapter 7 that analyses the performance of the presented algorithm on simulated and monitored data whereas the performance evaluation is round up in Chapter 8 by considering signals with very low SNR.

5. Sequential Detection of Gradual Changes

5.1. Model

The signal of interest is modelled by a deterministic function over time with an additional error term. When the considered signal is denoted by $y[t]$ it is assumed to have the time varying mean $u[t]$, and the additive error term is zero-mean white Gaussian noise $w[t]$ with constant variance σ^2. This follows the notation introduced in Section 2.1.4.

$$y[t] = u[t] + w[t] \quad t = 1, 2, \ldots, N \tag{5.1}$$

A single change in $u[t]$ is modelled by a three-phase linear function (ramp-step) defined on the interval $[a, b]$, as depicted in Figure 5.1. The change-point k and the rise time τ define the beginning and the duration of the change, respectively. Moreover, the offset d and the magnitude h are introduced since neither the level before nor the level after the change is known a priori.

By principle, the offset d and the magnitude h are real-valued whereas the change-point k and rise time τ are integer-valued, due to the discrete signal representation in the time domain. The change-point is an element of the interval $[a, b-1]$ and the rise time of the interval $[1, b-k]$ so that $k + \tau \leq b$ holds. If $\tau = 1$, the ramp-step model migrates to the simple step model, and for $\tau = b - k$, it migrates to a pure ramp model. The ramp-step model is therefore capable to model abrupt, gradual and ongoing changes.

The regression model of the complete signal (length N) is considered to be com-

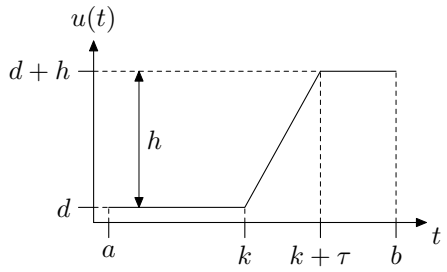

Figure 5.1.: The ramp-step function (defined on $[a, b]$) modelling a single gradual change in mean from an initial level d to a new level $d + h$ with rise time τ. The objective is to precisely estimate the change-point k.

posed of K adjacent ramp-steps, one for each change. The number K and the locations of the ramp-steps are supposed to be unknown.

5.2. Method

The method described next is based on a statistical framework namely the likelihood approach. Evaluating the likelihood is a well established approach in the field of change-point analysis (see Basseville and Nikiforov [1993]). The main concepts and properties of the likelihood principle are described in Chapter 2. Basic methods and formulas for detecting an abrupt change as well as estimating its location are covered by this chapter. In this section these methods are extended to the gradual change detection problem.

The change detection algorithm described below consists of two subsequent units: (i) a GLRT starts at the first sample of $y[t]$ to search for the first change. If indicated, then (ii) a local estimate of the ramp-step parameters is established by using the ML method. Afterwards, step (i) is restarted at the next available sample. This two-step procedure is repeated until the end of the signal is reached. Separating the detection from the localisation of the change is a common paradigm of modern change detection algorithms, which was formerly suggested by Charbonnier et al. [2004] and Staude [2001].

The outcome of this process will be a list of change-points together with their

ramp-step parameters $k^{(i)}$, $\tau^{(i)}$, $h^{(i)}$ and $d^{(i)}$ which are defined on $[a^{(i)}, b^{(i)}]$ with $i = 1, 2, \ldots, K$; for the initial condition, $a^{(1)} = 1$.

5.2.1. Detection of a Change

A change is detected by using a GLRT to choose among two competing hypotheses which are the "no change hypothesis" $\mathcal{H}_0$ against the "change hypothesis" $\mathcal{H}_1$ that a relevant change has happened at time ν with $a^{(i)} \leq \nu < n$. The test statistic of the GLRT for this problem can be formulated as

$$\mathrm{T}_{\mathrm{GLR}}(\boldsymbol{y}) = \frac{p(\boldsymbol{y}; \hat{\nu}, \hat{\mu}_1, \hat{\mu}_2 | \, \mathcal{H}_1)}{p(\boldsymbol{y}; \hat{\mu} | \, \mathcal{H}_0)} \tag{5.2}$$

This is the ratio of the PDFs under $\mathcal{H}_1$ and $\mathcal{H}_0$. The mean values before and after the supposed change-point as well as the mean value in the no change hypothesis are replaced by their MLE, i. e.,

$$\hat{\mu}(n) = \frac{1}{n - a^{(i)} + 1} \sum_{t=a^{(i)}}^{n} y[t] \tag{5.3}$$

$$\hat{\mu}_1(\nu, n) = \frac{1}{\nu - a^{(i)} + 1} \sum_{t=a^{(i)}}^{\nu} y[t] \tag{5.4}$$

$$\hat{\mu}_2(\nu, n) = \frac{1}{n - \nu} \sum_{t=\nu+1}^{n} y[t] \, . \tag{5.5}$$

The standard GLRT computes the test statistic for each possible change-point ν in order to get an estimate for ν which causes a high computational load. In this thesis, a sliding window algorithm is preferred which assumes the change to be happen L samples before the current time n so that the variable ν can be eliminated by replacing ν with $n - L$. This approach is called the approximated GLRT (see [Staude, 2001]). Working on-line, the test statistic of the approximated GLRT denoted by $V(n; L)$ is computed for each time n and the first time instant n where $V(n; L)$ exceeds an appropriately chosen threshold δ is called the alarm time

denoted by t_a. This is consequently formulated as the stopping rule

$$t_a = \min\{n \,:\, V(n;L) > \delta\} \tag{5.6}$$

with

$$V(n;L) = (n-L)(\hat{\mu}_1(n-L,n) - \hat{\mu}(n))^2 + L(\hat{\mu}_2(n-L,n) - \hat{\mu}(n))^2 \,. \tag{5.7}$$

This formula follows directly from the test statistic of the standard GLRT by setting $\nu = n - L$. The detailed calculus for the GLRT can be found in Section 2.2.4.

The sliding window technique is an approximation of the standard GLRT. Whether an approximation makes sense, depends on the errors and uncertainties that it introduces. This problem is discussed in Chapter 6 and it turns out that the sliding window technique has the same reliability as the standard approach in case that the window width L is chosen appropriately.

5.2.2. Estimation of the Ramp-Step Function

Once a change has been indicated, the ML method is used to fit a ramp-step to the signal on the interval $[a^{(i)}, b^{(i)}]$ with $b^{(i)} = t_a$. To enhance the readability, a vectorial notation will be chosen next. The vectors $\boldsymbol{y} = (y[a^{(i)}], y[a^{(i)}+1], \ldots, y[b^{(i)}])^T$, $\boldsymbol{u} = (u[a^{(i)}], u[a^{(i)}+1], \ldots, u[b^{(i)}])^T$, and $\boldsymbol{w} = (w[a^{(i)}], w[a^{(i)}+1], \ldots, w[b^{(i)}])^T$, are introduced so that the signal model (5.1) is rewritten to the vectorial form

$$\boldsymbol{y} = \boldsymbol{u} + \boldsymbol{w} \,. \tag{5.8}$$

The vector $\boldsymbol{u}$ is supposed to be a ramp-step, with parameters (k, τ, h, d), i. e.,

$$\boldsymbol{u}_t = \begin{cases} d & \text{if} \quad a^{(i)} \leq t \leq k \\ \frac{h}{\tau}(t-k) + d & \text{if} \quad k < t \leq k+\tau \\ d+h & \text{if} \quad k+\tau < t \leq b^{(i)} \end{cases} \tag{5.9}$$

where $\boldsymbol{u}_t$ denotes the t-th element of the vector $\boldsymbol{u}$. The parameters (k, τ, h, d) are estimated using the likelihood principle which gives an optimal fit of $\boldsymbol{u}$ to the observed signal $\boldsymbol{y}$.

In order to derive formulas for the MLEs of (k, τ, h, d) the parameters are separated by introducing $\boldsymbol{p}$, which is a ramp-step with unit norm and zero mean, i. e., it depends on the variables (k, τ), only. The elements of $\boldsymbol{p}$ are defined by

$$\boldsymbol{p}_t = \begin{cases} d_p & \text{if} \quad a^{(i)} \leq t \leq k \\ \frac{h_p}{\tau}(t-k) + d_p & \text{if} \quad k < t \leq k+\tau \\ d_p + h_p & \text{if} \quad k+\tau < t \leq b^{(i)} \end{cases} \tag{5.10}$$

with the magnitude

$$h_p = \frac{2\sqrt{3m\tau}}{\sqrt{2m(\tau(6k+2\tau-3)+1) - 3\tau(2k+\tau-1)^2}} \tag{5.11}$$

and the offset

$$d_p = -\frac{\sqrt{3\tau}(2m-2k-\tau+1)}{\sqrt{m}\sqrt{2m(\tau(6k+2\tau-3)+1) - 3\tau(2k+\tau-1)^2}} \tag{5.12}$$

where m is the length of the vector $\boldsymbol{p}$, i. e., the length of the interval $[a^{(i)}, b^{(i)}]$

$$m = b^{(i)} - a^{(i)} + 1 \; . \tag{5.13}$$

The formulas for h_p and d_p are derived by solving the linear equation system

$$\boldsymbol{u}^T\boldsymbol{u} = 1 \tag{5.14}$$

$$\bar{\boldsymbol{u}} = 0 \tag{5.15}$$

where $\bar{\boldsymbol{u}}$ denotes the mean over the elements of $\boldsymbol{u}$.

Now, the scale factor α and the offset β are introduced which allows to formulate $\boldsymbol{u}$ in terms of $\boldsymbol{p}$

$$\boldsymbol{u} = \alpha\boldsymbol{p} + \beta\boldsymbol{1} \tag{5.16}$$

where $\mathbf{1}$ is defined as an vector of length m with elements that are equal to one. The estimation of α and β is the standard linear regression problem (see e. g. Kundu and Ubhaya [2001]) with the solution

$$\hat{\alpha} = \boldsymbol{y}^T\boldsymbol{p} \tag{5.17}$$

$$\hat{\beta} = \bar{\boldsymbol{y}} \tag{5.18}$$

With that and the definitions (5.9) and (5.10) the MLEs of h and d can be inferred as

$$\hat{h} = \hat{\alpha}h_p \tag{5.19}$$

$$= \boldsymbol{y}^T\boldsymbol{p}\, h_p \tag{5.20}$$

and

$$\hat{d} = \hat{\alpha}d_p + \hat{\beta} \tag{5.21}$$

$$= \boldsymbol{y}^T\boldsymbol{p}\, d_p + \bar{\boldsymbol{y}}, \tag{5.22}$$

respectively. The MLEs of the remaining parameters, namely the change-point k and the rise time τ, are finally derived by replacing $\boldsymbol{u}$ in the equation

$$J(k, \tau; \boldsymbol{y}) = \sum_{t=1}^{N} (y_t - u_t)^2 \tag{5.23}$$

$$= \sum_{t=1}^{N} \Big(y_t - (\hat{\alpha}p_t + \hat{\beta})\Big)^2 \tag{5.24}$$

with (5.16) and performing its minimisation. $J(\theta; \boldsymbol{y})$ is the sum of the squared difference between the monitored signal $\boldsymbol{y}$ and the model $\boldsymbol{u}$ which is well known as to be the objective function of the least squares method. Note that in Section 2.1.4 it is shown that least squares produces the MLE.

Equation (5.24) can further be simplified so that the estimates of k and τ are

obtained by

$$(\hat{k}, \hat{\tau}) = \arg \max_{(k,\tau)} \left| \boldsymbol{y}^T \boldsymbol{p} \right| \ . \tag{5.25}$$

The complete derivation is given in the Appendix A. The method described there is called scale and shift invariant MLE which solves the problem in a broader context and is not bounded to the ramp-step template.

Since the number of pairs (k, τ) is finite, a complete grid search algorithm can be used for the maximisation in (5.25). While this guaranties to give the global optimum, other methods with a significant better performance can be utilised as described in Section 5.3.

5.2.3. Recursive Estimates on a Growing Domain

When the signal to noise ratio is high, the algorithm usually detects a change while it is still ongoing, i. e., the interval $[a^{(i)}, b^{(i)} = t_a]$ does not necessarily cover the whole transition. Therefore, the upper bound of the interval $[a^{(i)}, b^{(i)}]$ is recursively increased until it covers a whole transition. The post-change duration $s = b^{(i)} - (k^{(i)} + \tau^{(i)})$, i. e., the duration of the constant-signal epoch after the transition, serves as a decision criterion for this condition. The recursion is performed by increasing $b^{(i)}$ and fitting the ramp-step on $[a^{(i)}, b^{(i)}]$ until s exceeds an adequately chosen threshold s_{min}.

Figure 5.2 illustrates this process. The shaded epoch in Figure 5.2 (a) indicates the interval $[a^{(i)}, t_a]$ where the change is initially detected; the change is still in the transition phase. Subsequently, as shown by Figure 5.2 (b), the estimation of the ramp-step function is continued with a growing window until s fits to the threshold criterion s_{min} resulting in the final interval $[a^{(i)}, b^{(i)}]$.

5.2.4. Sequential Detection of Multiple Changes

For detecting and locating multiple changes, the steps described in the previous sections are repeated with the detection of the next change being initialised by $a^{(i+1)} = \hat{k}^{(i)} + \hat{\tau}^{(i)}$. A side effect of this initialisation is that the adjacent ramp-

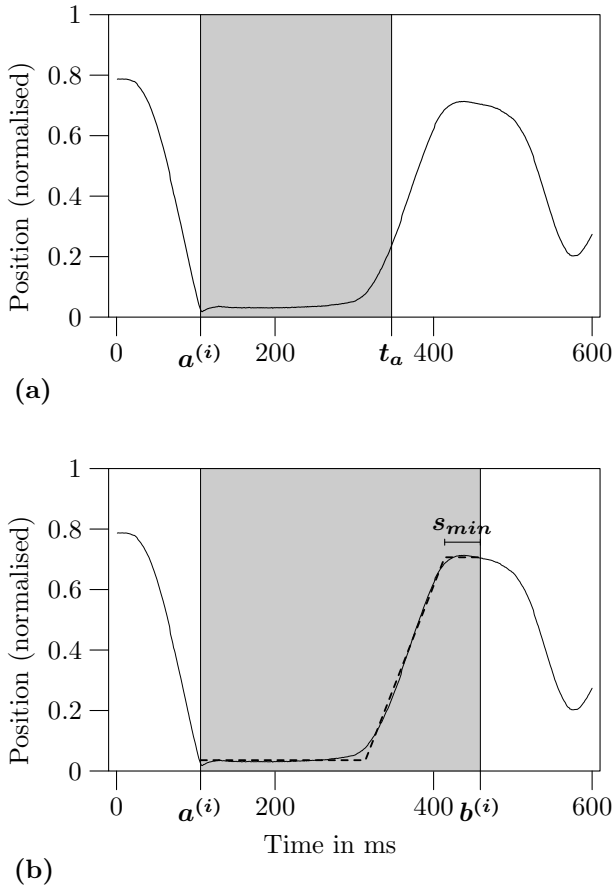

Figure 5.2.: Illustration of the recursive estimation process using a growing window. (a) The end of the shaded epoch is the alarm time t_a at which the change is detected. During the recursive estimation procedure, the window is enlarged until the post-change duration s exceeds a minimal value s_{min}. (b) The final range $[a^{(i)}, b^{(i)}]$ covers the whole transition, so that the ramp-step function (dashed line) is correctly estimated.

$a \leftarrow 1$
while $a < N$ **do**
  $b \leftarrow$ detect change on $[a, N]$
  $(\hat{k}, \hat{\tau}, \hat{h}, \hat{d}) \leftarrow$ estimate ramp-step on $[a, b]$
  while $b - (\hat{k} + \hat{\tau}) < s_{min}$ **do**
    $b \leftarrow b + 1$
    $(\hat{k}, \hat{\tau}, \hat{h}, \hat{d}) \leftarrow$ estimate ramp-step on $[a, b]$
  end while
  $a \leftarrow \hat{k} + \hat{\tau}$
end while

Figure 5.3.: Pseudo code representation of the algorithm for a sequential detection of multiple gradual changes (SEMUG).

steps overlap which improves the accuracy of the estimated mean $\hat{\mu}_0$ before change in contrast to a non-overlapping initialisation

The complete algorithm for a SEquential detection of MUltiple Gradual changes (SEMUG) is summarised in pseudo code in Figure 5.3.

5.3. Computational Aspects

Computational efficiency is very important for the acceptance of an algorithm. This section presents several ideas how to increase the computational efficiency of SEMUG. They range from an intelligent implementation of the formulas, over the use of heuristics, to an iterative optimisation procedure.

5.3.1. Effecient Implementation

When implementing SEMUG, the program listed in Figure 5.3 must be realised in the desired programming language. This involves implementation of (5.6) and (5.7) for the detection of a change and (5.20), (5.22) and (5.25) for the estimation of the ramp-step. They have in common that several sums must be computed for the mean values needed in (5.7) or for the inner products in (5.20), (5.22) and (5.25). This can efficiently be implemented by once calculating the cumulative sum (CS)

over the signal samples

$$\mathrm{CS}[t] = \begin{cases} 0 & \text{if } t = 0; \\ \sum_{j=1}^{t} y[j] & \text{if } t > 0 \end{cases} \tag{5.26}$$

and inferring the desired values from this sum. With that, the mean values of (5.3)-(5.5) are then computed with a few instructions by

$$\hat{\mu}(n) = \frac{1}{n - a^{(i)} + 1}(\mathrm{CS}(n) - \mathrm{CS}(a^{(i)} - 1)) \tag{5.27}$$

$$\hat{\mu}_1(\nu, n) = \frac{1}{\nu - a^{(i)} + 1}(\mathrm{CS}(\nu) - \mathrm{CS}(a^{(i)} - 1)) \tag{5.28}$$

$$\hat{\mu}_2(\nu, n) = \frac{1}{n - \nu}(\mathrm{CS}(n) - \mathrm{CS}(\nu)) \; . \tag{5.29}$$

The inner product $\boldsymbol{y}^T\boldsymbol{p}$ found in (5.20), (5.22) and (5.25) cannot be rewritten in terms of the simple cumulative sum, only. Additionally another cumulative sum denoted by CSI must be used in order to obtained an efficient implementation. CSI is defined as

$$\mathrm{CSI}[t] = \begin{cases} 0 & \text{if } t = 0; \\ \sum_{j=1}^{t} jy[j] & \text{if } t > 0 \; . \end{cases} \tag{5.30}$$

The inner product $\boldsymbol{y}^T\boldsymbol{p}$ can be rewritten as

$$\boldsymbol{y}^T\boldsymbol{p} = \sum_{j=a^{(i)}}^{k} d_p y[j] + \sum_{j=k+1}^{k+\tau} \left[\frac{h_p}{\tau}(j-k) + d_p\right] y[j] + \sum_{j=k+\tau+1}^{b^{(i)}} (h_p + d_p) y[j] \tag{5.31}$$

$$= \sum_{j=a^{(i)}}^{k} d_p y[j] + \sum_{j=k+1}^{k+\tau} \frac{h_p}{\tau} jy[j] - \sum_{j=k+1}^{k+\tau} \frac{h_p}{\tau} ky[j] + \sum_{j=k+1}^{k+\tau} d_p y[j] + \sum_{j=k+\tau+1}^{b^{(i)}} h_p y[j] + \sum_{j=k+\tau+1}^{b^{(i)}} d_p y[j] \tag{5.32}$$

which is in terms of CS and CSI

$$
\begin{aligned}
\boldsymbol{y}^T\boldsymbol{p} &= d_p(\mathrm{CS}(k) - \mathrm{CS}(a^{(i)} - 1)) + \frac{h_p}{\tau}(\mathrm{CSI}(k+\tau) - \mathrm{CSI}(k)) \\
&\quad - \frac{h_p}{\tau}k(\mathrm{CS}(k+\tau) - \mathrm{CS}(k)) + d_p(\mathrm{CS}(k+\tau) - \mathrm{CS}(k)) \\
&\quad + h_p(\mathrm{CS}(b^{(i)}) - \mathrm{CS}(k+\tau)) + d_p(\mathrm{CS}(b^{(i)}) - \mathrm{CS}(k+\tau)) \qquad (5.33) \\
&= d_p\mathrm{CS}(k) - d_p\mathrm{CS}(a^{(i)} - 1) + \frac{h_p}{\tau}\mathrm{CSI}(k+\tau) - \frac{h_p}{\tau}\mathrm{CSI}(k) \\
&\quad - \frac{h_p}{\tau}k\mathrm{CS}(k+\tau) + \frac{h_p}{\tau}k\mathrm{CS}(k) + d_p\mathrm{CS}(k+\tau) - d_p\mathrm{CS}(k) \\
&\quad + h_p\mathrm{CS}(b^{(i)}) - h_p\mathrm{CS}(k+\tau) + d_p\mathrm{CS}(b^{(i)}) - d_p\mathrm{CS}(k+\tau) \qquad (5.34) \\
&= \frac{h_p}{\tau}\mathrm{CSI}(k+\tau) - \frac{h_p}{\tau}\mathrm{CSI}(k) - d_p\mathrm{CS}(a^{(i)} - 1) + \frac{h_p}{\tau}k\mathrm{CS}(k) \\
&\quad - \left(\frac{h_p}{\tau}k + h_p\right)\mathrm{CS}(k+\tau) + (h_p + d_p)\mathrm{CS}(b^{(i)})\,. \qquad (5.35)
\end{aligned}
$$

Using these equations instead of the original ones, the computational costs reduces drastically. However, the problem of estimating the parameters of a ramp-step is still of order $O(n^2)$ due to the grid search of k and τ. This issue will be treated in the next section.

5.3.2. Reducing Computational Costs using Heuristics

The grid search used for estimating the change-point k and the rise time τ demands a high computational effort. This effort is multiplied by the number of recursive estimates which are done until the ramp-step covers the complete change as described in Section 5.2.3. While being computationally expensive, this method guarantees that the global optimum is found, since the complete grid is computed. In the following, three heuristics will be presented which have the effect that parts of the grid, which are unlikely to contain the optimum, will be omitted, which reduces the computational load.

Ramp Pre-Estimate

This heuristic utilises the fact that the estimation of the ramp-step is preceded by a detection with the GLRT. When the mean time between changes is high, the

detected change will be at the end of the interval $[a, b]$ for adequately chosen tuning parameters. In this case, the grid search will waste a lot of time investigating the beginning of $[a, b]$. In order to avoid this, a lower bound for the change-point denoted by a_R is estimated before starting the grid search which then only considers change-points in the range $[a_R, b]$. The lower bound a_R is determined by fitting a ramp on $[a, b]$ which is done in $O(n)$. In Hofer et al. [2007] it is derived that the ramp has the lowest change-point estimate among the ramp-step templates which gives a data driven estimate of a_R.

Adaptively Growing Domain

In SEMUG, it is checked if the estimated ramp-step covers the whole change by comparing the post-change duration s with a minimal threshold denoted by s_{min}. The size of the domain is then increased by one if the estimated duration after change is lower than s_{min}. When s is significantly lower than s_{min} it is unlikely that a ramp-step fitted to a domain which is just one sample greater fulfils the criterion $s > s_{min}$. In order to avoid unnecessary growing steps, this heuristic introduces an adaptively growing domain. The domain is always increased by $s_{min} - s$.

Fixed Change-Point

The idea to this heuristic is closely coupled to the aforementioned. It utilises the fact that on can expect an unbiasede change-point estimate, invariant with respect to the size of the domain as long as the domain covers the true change-point. Thus, after the first complete ramp-step estimate, the change-point is kept fixed, which has the advantage that the following ramp-step fit on the bigger domain has only a computational complexity of $O(n)$ since the only remaining parameter is the rise time. Although this heuristic works perfectly when the true signal shape is an undisturbed signal it is recommended to perform a complete ramp-step estimate after the recursive estimation phase is completed. This will enhance the accuracy in many applications with unexpected disturbances.

5.3.3. Iterative Maximisation Procedure

The basics of iterative maximisation procedures are described in Section 2.1.6. These procedures start with an initial guess, followed by a directed search towards the optimal parameter combination. In change-point analysis many scientists make use of iterative methods, see e. g. Buchanan et al. [1997], Chiu [2002], Muggeo [2003], Pastor-Barriuso et al. [2003]. In this section a prominent one, namely the Levenberg-Marquardt algorithm is utilised for estimating the change-point and the rise time. This problem is solved by maximising the objective function

$$F(\boldsymbol{\theta}; \boldsymbol{y}) = \left| \boldsymbol{y}^T \boldsymbol{p} \right| \tag{5.36}$$

with the parameter vector $\boldsymbol{\theta} = (k, \tau)^T$. Note, that standard techniques from linear programming like the simplex method [Dantzig, 1963, Nash, 2000] cannot be applied, since the objective function does not depend linearly on the parameter vector.

The Levenberg-Marquardt update rule (2.40) for this problem is

$$\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i - \left[H(\boldsymbol{\theta}; \boldsymbol{y}) + \lambda \operatorname{diag}[H(\boldsymbol{\theta}; \boldsymbol{y})] \right]^{-1} \frac{\partial F(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \boldsymbol{\theta}} \bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_i} \tag{5.37}$$

where $\frac{\partial F(\boldsymbol{\theta};\boldsymbol{y})}{\partial \boldsymbol{\theta}}$ is the gradient

$$\frac{\partial F(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \boldsymbol{\theta}} = \begin{pmatrix} \frac{\partial F(\boldsymbol{\theta};\boldsymbol{y})}{\partial k} \\ \frac{\partial F(\boldsymbol{\theta};\boldsymbol{y})}{\partial \tau} \end{pmatrix} . \tag{5.38}$$

and $H(\boldsymbol{\theta}; \boldsymbol{y})$ is the Hessian

$$H(\boldsymbol{\theta}; \boldsymbol{y}) = \begin{pmatrix} \frac{\partial^2 F(\boldsymbol{\theta};\boldsymbol{y})}{\partial k^2} & \frac{\partial^2 F(\boldsymbol{\theta};\boldsymbol{y})}{\partial k \partial \tau} \\ \frac{\partial^2 F(\boldsymbol{\theta};\boldsymbol{y})}{\partial \tau \partial k} & \frac{\partial^2 F(\boldsymbol{\theta};\boldsymbol{y})}{\partial \tau^2} \end{pmatrix} . \tag{5.39}$$

Note that for the given size of the Hessian, which is only 2×2, there exist the

explicit formula for the inverse in (5.37), which is

$$[H(\boldsymbol{\theta};\boldsymbol{y}) + \lambda \operatorname{diag}[H(\boldsymbol{\theta};\boldsymbol{y})]]^{-1} = C \begin{pmatrix} \frac{\partial^2 F(\boldsymbol{\theta};\boldsymbol{y})}{\partial \tau^2} & -\frac{\partial^2 F(\boldsymbol{\theta};\boldsymbol{y})}{\partial k \partial \tau} \\ -\frac{\partial^2 F(\boldsymbol{\theta};\boldsymbol{y})}{\partial \tau \partial k} & \frac{\partial^2 F(\boldsymbol{\theta};\boldsymbol{y})}{\partial k^2} \end{pmatrix} \tag{5.40}$$

with

$$C = \left((\lambda+1)^2 \frac{\partial^2 F(\boldsymbol{\theta};\boldsymbol{y})}{\partial k^2} \frac{\partial^2 F(\boldsymbol{\theta};\boldsymbol{y})}{\partial \tau^2} - \frac{\partial^2 F(\boldsymbol{\theta};\boldsymbol{y})}{\partial k \partial \tau} \frac{\partial^2 F(\boldsymbol{\theta};\boldsymbol{y})}{\partial \tau \partial k} \right)^{-1} . \tag{5.41}$$

This avoids a time consuming computation needed for higher order problems.

A numerical solution for the partial derivatives in (5.38)–(5.41) can be obtained by a Taylor expansion. For the change-point k the Taylor expansion is

$$\begin{aligned} F(k+\varepsilon,\tau;\boldsymbol{y}) = F(k,\tau;\boldsymbol{y}) &+ \varepsilon \frac{\partial F(k,\tau;\boldsymbol{y})}{\partial k} + \frac{\varepsilon^2}{2!} \frac{\partial^2 F(k,\tau;\boldsymbol{y})}{\partial k^2} \\ &+ \frac{\varepsilon^3}{3!} \frac{\partial^3 F(k,\tau;\boldsymbol{y})}{\partial k^3} + \cdots . \end{aligned} \tag{5.42}$$

With that, it is apparent that

$$\begin{aligned} \frac{F(k+\varepsilon,\tau;\boldsymbol{y}) - F(k,\tau;\boldsymbol{y})}{\varepsilon} &= \frac{\partial F(k,\tau;\boldsymbol{y})}{\partial k} + \frac{\varepsilon}{2!} \frac{\partial^2 F(k,\tau;\boldsymbol{y})}{\partial k^2} \\ &\quad + \frac{\varepsilon^2}{3!} \frac{\partial^3 F(k,\tau;\boldsymbol{y})}{\partial k^3} + \cdots \\ &= \frac{\partial F(k,\tau;\boldsymbol{y})}{\partial k} + O(\varepsilon) \end{aligned} \tag{5.43}$$

holds, which gives an approximation for the partial derivative along k with an error term of order ε. This formula is known as the *forward difference derivative*. Advanced and more often applied is the *central difference formula* which uses information from both before and after the point to be evaluated. With the backward Taylor expansion

$$\begin{aligned} F(k-\varepsilon,\tau;\boldsymbol{y}) = F(k,\tau;\boldsymbol{y}) &- \varepsilon \frac{\partial F(k,\tau;\boldsymbol{y})}{\partial k} + \frac{\varepsilon^2}{2!} \frac{\partial^2 F(k,\tau;\boldsymbol{y})}{\partial k^2} \\ &- \frac{\varepsilon^3}{3!} \frac{\partial^3 F(k,\tau;\boldsymbol{y})}{\partial k^3} + \cdots \end{aligned} \tag{5.44}$$

if follows for the central difference formula

$$\begin{aligned}\frac{F(k+\varepsilon,\tau;\boldsymbol{y})-F(k-\varepsilon,\tau;\boldsymbol{y})}{2\varepsilon} &= \frac{\partial F(k,\tau;\boldsymbol{y})}{\partial k} \\ &\quad + \frac{\varepsilon^2}{3!}\frac{\partial^3 F(k,\tau;\boldsymbol{y})}{\partial k^3} + \cdots \\ &= \frac{\partial F(k,\tau;\boldsymbol{y})}{\partial k} + O(\varepsilon^2) \qquad (5.45)\end{aligned}$$

that the error term is of order ε^2 which is superior for small ε. Note, for the endpoints of k ($k = 1$ and $k = m - 1$) the central difference formula can not be applied. For these two points the formula

$$\begin{aligned}\frac{-F(k+2\varepsilon,\tau;\boldsymbol{y})+4F(k+\varepsilon,\tau;\boldsymbol{y})-3F(k,\tau;\boldsymbol{y})}{2\varepsilon} &= \frac{\partial F(k,\tau;\boldsymbol{y})}{\partial k} \\ &\quad - 2\frac{\varepsilon^2}{3!}\frac{\partial^3 F(k,\tau;\boldsymbol{y})}{\partial k^3} + \cdots \\ &= \frac{\partial F(k,\tau;\boldsymbol{y})}{\partial k} + O(\varepsilon^2) \qquad (5.46)\end{aligned}$$

must be used instead.

With (5.45) and (5.46) the gradient can be computed, where the partial derivative along τ is derived analogously and the Hessian is computed by building simply the derivative of the derivative. Since k and τ are integer-valued, the smallest possible ε is one, which is numerically not optimal. This is a major drawback, since a less accurate estimate of the gradient may lead to a divergence of the Levenberg-Marquardt procedure.

The fact that k and τ are constrained to integers raises another issue that an iteration step lower than 0.5 does not change anything due to rounding. The consequence would be that the algorithm gets stuck especially close to the maximum where the step sizes are typically small. To overcome this issue, a local grid search around the current position is performed when the computed step size is lower than 0.5.

The Levenberg-Marquardt procedure was applied to the signal depicted in Figure 5.4 which was monitored during a finger tapping experiment (see Section 7.3

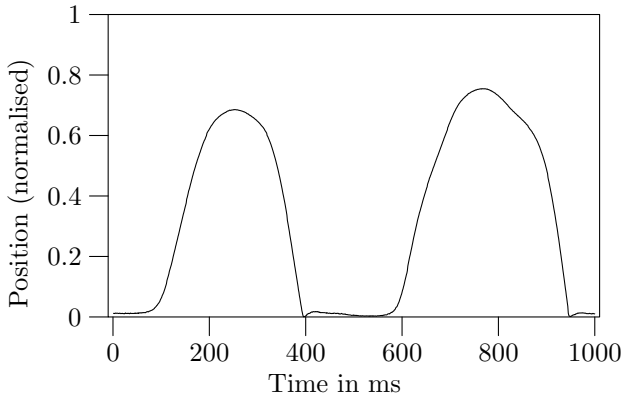

Figure 5.4.: A segment of a typical signal recorded in a tapping experiment. Two subsequent tapping movements are shown. The normalised position indicates the location of the tip of the index finger with respect to ground level [Cong Khac et al., 2007].

for more details on finger tapping). The initial guess θ_0 was obtained by a ramp-estimate and the method was combined with an adaptively growing domain as described in Section 5.3.2. The results were compared with the grid-search method and the optimised grid-search which makes use of the heuristics introduced in Section 5.3.2. The estimated change-points and rise times are summarised in Table 5.1 and Table 5.2. The grid-search serves as a reference since it is guaranteed to find the best fitting segmentation. The optimised grid-search gives almost the same estimates for the k and τ and also for a, which is evident since an iteration is initialised by $a^{(i+1)} = k^{(i)} + \tau^{(i)}$.

The Levenberg-Marquardt algorithm attains convergence and gives results which are close to the grid-search with difference of just a few samples. The only exception is $b^{(1)}$ which deviates 15 samples. However, this difference has had not influence on the estimated change-point and rise time.

For this particular signal, the optimised grid-search is preferable to the Levenberg-Marquardt procedure, because of accuracy and reliability. Comparing the computation time the two algorithms are almost identical, processing the signal in approximately 100 ms on a PC with a 2.8 GHz Intel© Pentium© 4 CPU, but one order lower than the computation time of the complete grid-search algorithm, which was

Table 5.1.: The estimated change-points and rise times when applying SEMUG on the signal depicted in Figure 5.4.

	Grid-Search		Grid-Search optimised (Diff. to Grid-Search)		Levenberg-Marquardt (Diff. to Grid-Search)	
i	$k^{(i)}$	$\tau^{(i)}$	$k^{(i)}$	$\tau^{(i)}$	$k^{(i)}$	$\tau^{(i)}$
1	92	111	0	0	-2	4
2	320	76	0	0	1	-2
3	581	132	2	-3	-1	2
4	866	87	0	0	2	-4

Table 5.2.: The boundaries of the domains $[a^{(i)}; b^{(i)}]$ of the estimated ramp-steps when applying SEMUG on the signal depicted in Figure 5.4.

	Grid-Search		Grid-Search optimised (Diff. to Grid-Search)		Levenberg-Marquardt (Diff. to Grid-Search)	
i	$a^{(i)}$	$b^{(i)}$	$a^{(i)}$	$b^{(i)}$	$a^{(i)}$	$b^{(i)}$
1	1	294	0	0	0	15
2	203	486	0	0	2	-1
3	396	803	0	-1	-1	2
4	713	1000	-1	0	1	0

approximately 6 s.

It is impossible to generalise these results. Which method should be used depends on the application. But, when the computation time of grid search is acceptable it should be preferred over the others.

6. Tuning Parameters

6.1. Introduction

The application of signal processing algorithms is often handicapped by the problem that their parametrisation is difficult due to the lacking knowledge about the properties of the signal. Thus, parameters are usually tuned by a tedious trial and error process, which often ends up with frustration, since the more complex the algorithm, the larger is the available parameter space, and the less probable is the chance to find some optimum.

SEMUG has three tuning parameters, namely the detection threshold δ, the window width L and the minimal post-change duration s_{min}. Especially δ is hard to guess appropriately since it is not directly linked to visually apparent signal properties. To overcome this problem, a method was developed which determines reasonable tuning parameters from the signal's apparent properties like change magnitude and rise time of the individual changes.

This approach is not the standard method in detection theory. In the sense of detection theory the threshold δ is a trade of between the probability of a false alarm and a miss. Though this method is preferable, it needs to know the probability density function of the test statistic. In general, the derivation of the test statistics' PDF is not a trivial for change-point problems (see i. e. Basseville and Nikiforov [1993]) and it has yet not been done for the detection of a gradual, ramp-step like change.

The approach presented next is based on the noise-free deterministic case. Although there is no proof that the found equations can be used for stochastic signals, they have been successfully applied to biophysical signals with moderate noise level

(see Chapter 7).

6.2. Tuning with the Signal's Deterministic Properties

As explained in Section 5.2.1, SEMUG uses a sliding window algorithm which assumes the change to happen L samples before the current time n. In contrast, the standard GLRT considers every sample before n as a potential change-point, which is the only difference between the standard GLRT and the windowed algorithm. Figure 6.1 is a contour plot of the test statistic of the standard GLRT, denoted by $V(\nu, n)$, when the signal is ramp-step with true parameters $d_0 = 1$, $h_0 = 1$, $k_0 = 50$, and $\tau_0 = 50$. Since the ramp-step is a continuous but piecewise defined function, $V(\nu, n)$ is continuous and piecewise defined, too. It comprises the nonoverlapping regions (see also Figure 6.1)

- **Region A:** $1 \leq n \leq k_0$ and $1 \leq \nu \leq k_0$
- **Region B:** $k_0 < n \leq k_0 + \tau_0$ and $1 \leq \nu \leq k_0$
- **Region C:** $k_0 < n \leq k_0 + \tau_0$ and $k_0 < \nu \leq k_0 + \tau_0$
- **Region D:** $k_0 + \tau_0 < n \leq N$ and $1 \leq \nu \leq k_0$
- **Region E:** $k_0 + \tau_0 < n \leq N$ and $k_0 < \nu \leq k_0 + \tau_0$
- **Region F:** $k_0 + \tau_0 < n \leq N$ and $k_0 + \tau_0 < \nu \leq N$

Formulas for $V(\nu, n)$ for every region are derived and listed in Appendix B. Furthermore, it is proven there that $V(\nu, n)$ is strictly increasing with n in all regions. Together with the fact that $V(\nu, n)$ is strictly increasing with ν in region D but strictly decreasing with ν in region F, it can be concluded that the global maximum of $V(\nu, n)$ always occurs in region E.

The sliding window approach does not compute every single value of $V(\nu, n)$ which is the reason for its high computational efficiency. Since $n - \nu = L$ is constant, the first value computed is $V(1, L + 1)$, the second is $V(2, L + 2)$ and so on. This fact is illustrated by the dashed line in Figure 6.1; it has unit slope and intersects

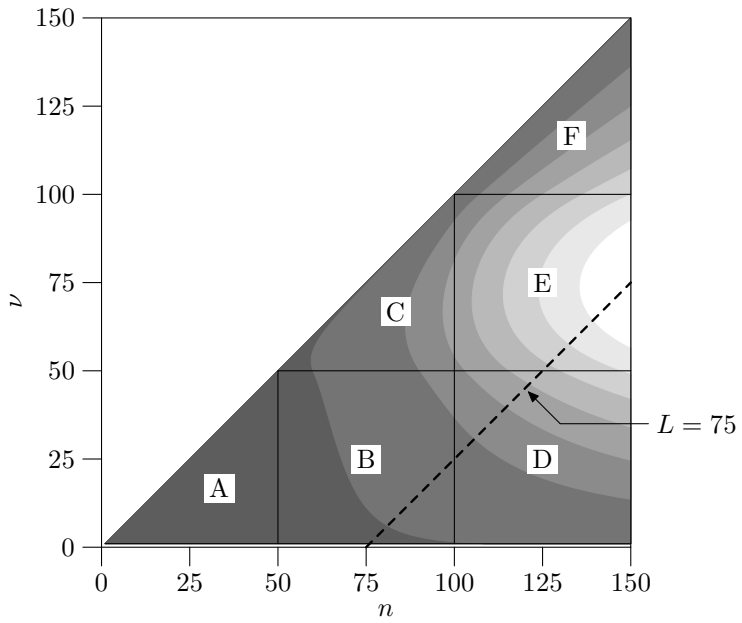

Figure 6.1.: A contour plot of the test statistic $V(\nu, n)$ obtained for a ramp-step with $d_0 = 1$, $h_0 = 1$, $k_0 = 50$ and $\tau_0 = 50$ serving as the input signal. Increasing brightness indicates larger values of $V(\nu, n)$. $V(\nu, n)$ is defined on six regions indicated by capital letters A to F. The maximum of $V(\nu, n)$ always occurs in region E (see text for details). The sliding window technique does not scan the complete area, but it walks on a straight line defined by the sliding window width L. The dashed line is the special case for $L = 75$.

the abscissa at $n = L$. The windowed detection walks on this straight line. For a window width L that is smaller are higher than the displayed one, the maximum value of $V(n;L)$ would be smaller than the maximum of $V(\nu,n)$. This means that in this case the change is detected only for a smaller threshold. But a small threshold causes a rising number of false alarms. Therefore, an appropriately chosen window width L maximises $V(n;L)$ with the consequence that the maximum value of $V(\nu,n)$ (standard GLRT) is equal to the maximum value of $V(n;L)$ (windowed detection).

Equivalent to $V(\nu,n)$, the global maximum of $V(n;L)$ occurs in region E, which means that $k_0+\tau_0 < n \leq N$ and $n-(k_0+\tau_0) < L \leq n-k_0$, holds. For this case the sums in (5.3)-(5.5) can explicitly be solved with

$$\begin{aligned}
\hat{\mu}(n) &= \frac{1}{n}\left(\sum_{j=1}^{k_0} d_0 + \sum_{j=k_0+1}^{k_0+\tau_0}\left(\frac{h_0}{\tau_0}(j-k_0)+d_0\right) + \sum_{j=k_0+\tau_0+1}^{n}(d_0+h_0)\right)\\
\hat{\mu}_1(n) &= \frac{1}{n-L}\left(\sum_{j=1}^{k_0} d_0 + \sum_{j=k_0+1}^{n-L}\left(\frac{h_0}{\tau_0}(j-k_0)+d_0\right)\right)\\
\hat{\mu}_2(n) &= \frac{1}{L}\left(\sum_{j=n-L+1}^{k_0+\tau_0}\left(\frac{h_0}{\tau_0}(j-k_0)+d_0\right) + \sum_{j=k_0+\tau_0+1}^{n}(d_0+h_0)\right)
\end{aligned}$$

so that after further calculations the test statistic of the sliding window algorithm (5.7) can be reformulated to

$$\begin{aligned}
V_E(n;L) = \frac{h_0^2}{4}\frac{1}{L(n-L)n\tau_0^2}\big((-1+k_0+L-n)(k_0+L-n)n\\
+(L-n)(1-2k_0+2n)k_0+(-L+n)\tau_0^2\big)^2 . \qquad (6.1)
\end{aligned}$$

With this formula it is apparent that individual changes with a smaller magnitude h_0 give a smaller test statistic $V_E(n;L)$, i. e., they are less significant. Generally, a change is only detected when the maximum value of the test statistic $V_E(n;L)$ is smaller or equal to the threshold δ. But a small δ causes more false alarms so that it is advisable to choose δ equal to maximum of the test statistic for the least

significant change.

As stated above, the maximum of $V_E(n;L)$ is reached for $n = k_0 + \tau_0 + s_0$. Which means that the least significant change is the one with minimal values for k_0, τ_0 and s_0 and since the post-change duration s_0 of the i-th change is identical to the change-point of the (i+1)-th change, the equation $k_0 = s_0$ holds for the least significant change. So that $V_E(n;L)$ migrates to $V_E^*(L)$ for $n = k_0 + \tau_0 + s_0$ and $k_0 = s_0$ so that

$$\begin{aligned} V_E^*(L) &= V_E(n;L) \\ &= -\frac{h_0^2}{4}\frac{1}{L(L-(2s_0+\tau_0))\tau_0^2(2s_0+\tau_0)}\big((2s_0+\tau_0)L^2 \\ &\quad -(4s_0^2+\tau_0^2+s_0(2+4\tau_0))L+s_0(1+s_0)(2s_0+\tau_0)\big)^2 \end{aligned} \tag{6.2}$$

when

$$n = k_0 + \tau_0 + s_0 \quad \text{and} \quad k_0 = s_0 \tag{6.3}$$

holds. With theses two constraints the definition range of L shrinks to $L \in]s_0,\, \tau_0 + s_0]$. $V_E^*(L)$ is a rational function in L with singularities outside of the definition range. It can be shown that this function has a single maximum in the definition range. This allows to derive the maximum by partial derivation although L is an integer valued variable. Hence the optimal window width L, which maximises $V_E^*(L)$, is the solution of the differential equation

$$\frac{\partial V_E^*(L)}{\partial L} = 0 \quad \text{with} \quad s_0 < L \leq \tau_0 + s_0\,. \tag{6.4}$$

With standard calculus it is found that

$$L = \frac{\tau_0+1}{2} + s_0 \quad \text{for odd}\,\tau_0 \tag{6.5}$$

and

$$L = \frac{\tau_0}{2} + s_0 \quad \text{for even}\,\tau_0\,. \tag{6.6}$$

With (6.5) and (6.6) formulas are found for tuning the window width L. With those a tuning formula for δ can be found by replacing L in (6.2). For even τ_0 the window width is replaced by (6.6) so that (6.2) reduces to

$$V_E^*(L) = \frac{h_0^2(4s_0 + \tau_0)^2}{16(2s_0 + \tau_0)} \; . \tag{6.7}$$

As stated above, the threshold δ should tuned to be equal to $V_E^*(L)$ for minimal values of h_0, τ_0 and s_0. So that finally the tuning equations are

$$\mathrm{L} = \mathrm{round}\left(\frac{\tau_{0\,min}}{2} + s_{0\ \mathrm{min}}\right) \tag{6.8}$$

and

$$\delta = \frac{h_{0\,min}^2(4s_{0\,min} + \tau_{0\,min})^2}{16(2s_{0\,min} + \tau_{0\,min})} \; . \tag{6.9}$$

Beside L and δ, SEMUG has a third tuning parameter s_{min} which is used for terminating the recursive increase of the interval $[a, b]$ (see Section 5.2.3). It stops when the post-change duration will be greater than or equal to s_{min}. Hence, s_{min} will simply be set to the minimal expected post-change duration

$$s_{min} = s_{0\,min} \; . \tag{6.10}$$

The equations (6.8), (6.9), and (6.10) define SEMUG's tuning parameters L, δ and s_{min} using expert knowledge about the least significant change in the signal ($h_{0\,min}$, $\tau_{0\,min}$, and $s_{0\,min}$). Note that this does not mean that every change with $\tau_0 < \tau_{0\,min}$ will be rejected. There is the chance to detect such steep changes if their magnitudes are high enough. The magnitude $h_{0\,min}$ is squared in (6.9) and, therefore, it is the most influential parameter among the three.

7. Performance Evaluation

7.1. Introduction

A performance evaluation does support a theory by testing an algorithm on either simulated or measured data. In this chapter both will be done for evaluating the performance of SEMUG. Simulated data are computer generated signals that attain the signal model of SEMUG. In contrast to that measured signals do not exactly attain the model. There are always aspects in measured data that are not reflected in the model. This usually causes performance degeneration in comparison to the simulated signals. Hence, the performance evaluation on simulated data is often used as a reference.

For performance analysis the measured signals have the disadvantage that the true values for the parameters of interest, e. g., the true change-points, are rarely known. However, to overcome this issue the estimation results can be compared to the decision of an expert who has done a visually inspection of the signal. This is not the same as the comparison of the algorithms' results to the true values like in the performance analysis with simulated signals. This is a comparison of two algorithms where the decision of the expert is viewed as a decision of an algorithm which is based on the experience of the expert. This decision may have a bias and it surely comprises random effects since decisions of humans are not exactly reproducible.

Reproducibility is one advantage of the performance evaluation with simulated data. Another is that it gives the freedom to choose the signal profile so that special properties and effects of the algorithm can be demonstrated. Simulated signals can also be used to give error bounds and confidence intervals. However, these values

cannot directly transferred to measured data because of modelling issues. It is important to keep in mind that a model does describe the reality only to some extent.

7.2. Results on Simulated Data

A set of data was created, consisting of 10,000 simulated signals. Each simulation consists of three ramp-step changes where the first two have positive magnitude modelling an interrupted gradual change and the last has a negative magnitude so that the signal is in the end back to the initial level. This means that the data set is composed of 30,000 gradual changes. This large number does allow statistical evident reasoning.

The simulated signals are corrupted by WGN as well as by a deterministic error in the rest phase before the second gradual change which is modelled by a ramp-step with small negative magnitude. The algorithm should be tuned so that the three major ramp-steps are found but not the one considered as a deterministic error.

Figure 7.1 displays a representative simulated signal in the upper panel and its regression function in the lower panel. For each simulated signal, the parameters of the ramp-steps I-IV are randomly chosen with the restriction that the offset of I is zero and that the magnitude of IV is determined by the fact that the signal should be back to zero in the last steady range. Except that, it holds that for the ramp-steps I, III and IV the change-point is in $[1, 50]$, the rise-time is in $[40, 80]$ and the magnitude is in $[0.5, 1]$ for I, III and $[-2, -1]$ for IV. The ramp-step in II uses the same range for the change-point but the rise-time is in $[1, 40]$ and the magnitude is in $[-0.25, 0]$. The simulated signals were obfuscated by WGN with a standard deviation up to 75% of the magnitude of the ramp-steps in I, III and IV.

SEMUG has analysed the $10,000$ simulated signals with the same tuning parameters which are $h_{0\,min} = 0.4$, $\tau_{0\,min} = 40$, and $s_{0\,min} = 30$. One can expect that with these tuning parameters almost every change will be detected since $h_{0\,min} = 0.4$ is 20 % less than the true minimal value which results in a more sensitive detection.

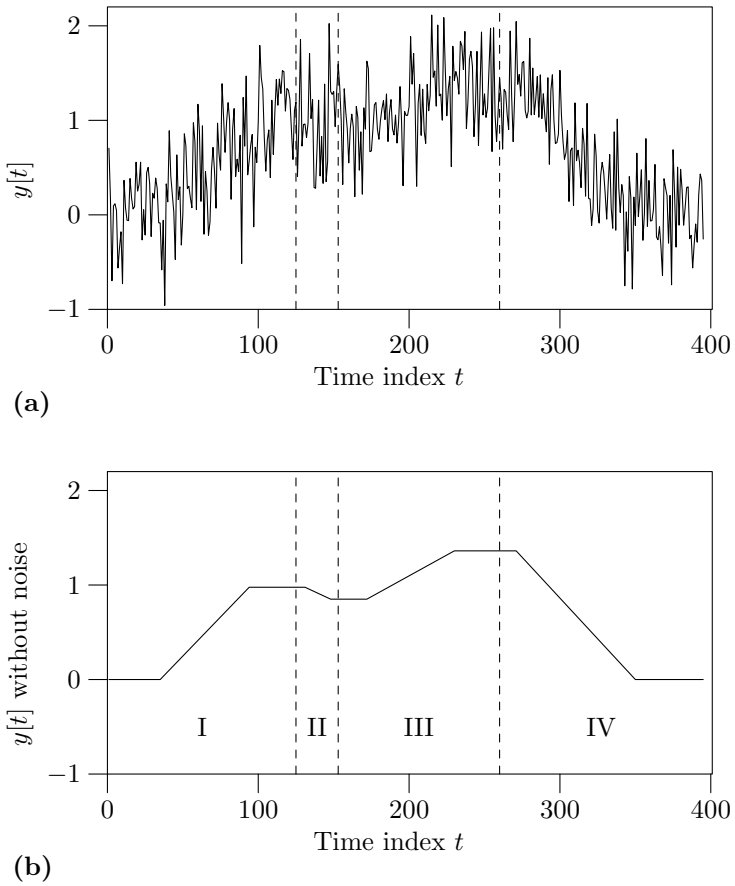

Figure 7.1.: (a) A representative simulated signal and (b) the regression function of the signal. The signal comprises four parts, with three gradual changes (I, III and IV) and one deterministic error (II) which are all modelled by ramp-steps.

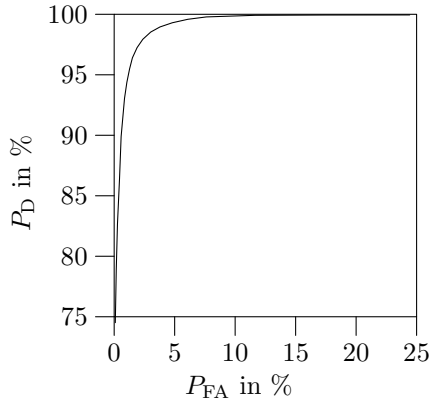

Figure 7.2.: The receiver operating characteristic (ROC) of SEMUG obtained by analysing the simulated data set with different tuning parameters. The ROC curve is a graphical plot of the probability of true positives (probability of detection P_{D}) over the probability of false positive (probability of a false alarm P_{FA}).

In fact, only 63 out of 30,000 changes where missed which is around 0.2 %. In contrast to that 7.6 % of the detected changes are false alarms. This high number has several reasons. One is the sensitive tuning in combination with the high noise level of some signals. Another is the deterministic error in II. It is the reason why the false alarms are not equally distibuted among the four parts (I-IV) of the signal. With 38 % of all false alarms being in II there is a significant increase of the false alarm rate. In general, there is always a trade-off between the probability of a flase alarm denoted by P_{FA} and the probability of a correctly detected change denoted by P_{D}. This trade-off is visualised by the receiver operating characteristic (ROC) displayed in Figure 7.2.

Until know, only the detection performance of SEMUG was regarded. In table 7.1 the median of the estimation errors is listed. Note, that the median is used in favour of the mean since outliers have such a big influence to the mean so that it would not be a good measure for the average result. The outliers are caused by those simulated signals with low SNR which will be discussed in detail in Chapter 8.

Apparent in Table 7.1 is the difference of the figures in column two to the first and the last column. The estimation errors listed in the second column are caused

Table 7.1.: The median of the estimation error of SEMUG when applied to the set of 10,000 simulated signals.

SEMUG	Intervals		
	I	II+III	IV
median$(\hat{k}^{(i)} - k_0^{(i)})$	2.0	7.0	1.0
median$(\hat{\tau}^{(i)} - \tau_0^{(i)})$	−3.0	−10.0	−1.0
median$(\hat{h}^{(i)} - h_0^{(i)})$	−0.0088	−0.0763	0.0055
median$(\hat{d}^{(i)} - d_0^{(i)})$	0.0008	0.0563	−0.0025

Table 7.2.: The median of the estimation error of MLE when applied to the set of 10,000 simulated signals.

MLE	Intervals		
	I	II+III	IV
median$(\hat{k}^{(i)} - k_0^{(i)})$	0.0	5.0	0.0
median$(\hat{\tau}^{(i)} - \tau_0^{(i)})$	−1.0	−5.0	0.0
median$(\hat{h}^{(i)} - h_0^{(i)})$	0.0007	−0.0586	−0.0011
median$(\hat{d}^{(i)} - d_0^{(i)})$	−0.0001	0.0599	0.0004

by both, stochastic and the deterministic error in the interval II resulting in a biased estimation. The negative magnitude of the neglected ramp-step in II causes a negative bias of the offset which is the reason for the positive bias of the magnitude. Note, that the two errors are almost offset against each other so that the bias of the offset in III is one magnitude lower. This means that the deterministic error in II has only a local influence on the performance.

Since SEMUG is a rather complex algorithm to determine the exact cause for the estimation errors in Table 7.1. For that reason the data set was analysed by a second algorithm. It is an algorithm that knows the boundaries of the intervals I-IV and computes the MLE of the ramp-steps. This algorithm gives the best results that can be reached with likelihood based algorithms. The estimation error of the MLE is listed in table 7.2. Apparent is again the bias of the k and τ in the second column which is caused by the deterministic error. The overall performance of SEMUG is worse but close the MLE.

The results obtained on the simulated data show that the segmentation algorithm is able to almost eliminate deterministic and random errors of the signal while still providing the correct segmentation, when the signal is made up of ramp-steps. In the next section these restriction are relaxed. SEMUG will be applied to signals obtained from finger tapping experiments. These signals have less predictable errors and their regression lines are not known a priori.

7.3. Application to Finger Tapping

7.3.1. Experimental Setup

The data was taken from a research project investigating the behaviour of the human motor system in dual-task situations. For a detailed description of material and methods, see [Cong Khac et al., 2007]. In short, the subject was sitting at a table and was required to tap with the left and right index fingers, according to the given instruction. Two laser distance sensors fixed above the finger tips measured their vertical position which was digitised by an ADC at a sampling frequency of 1 kHz. Figure 7.3 displays a representative section of the position signal. The finger of the right hand moves down (flexion movement) in the beginning and hits the table top at approximately $n = 100$; afterwards, it starts to move upwards (extension movement) towards the resting position at approx. $n = 300$. While the subject was instructed to tap rhythmically with the index finger of the right hand, a single tap should be executed with the other index finger in response to a visual stimulus. The unexpected movement of the second finger interferes with the periodic movement; e. g., it sometimes causes an interruption of the flexion movement (see Figure 7.3).

7.3.2. Demonstration on Short-Term Signals

SEMUG is utilised next for the segmentation of two short term signals recorded in a tapping experiment (see Figure 7.3 and Figure 7.5 (a)). First, a signal comprising a flexion and an extension movement as well as an interrupted flexion movement which consists of three changes (see Figure 7.3) were investigated. After a visual inspection

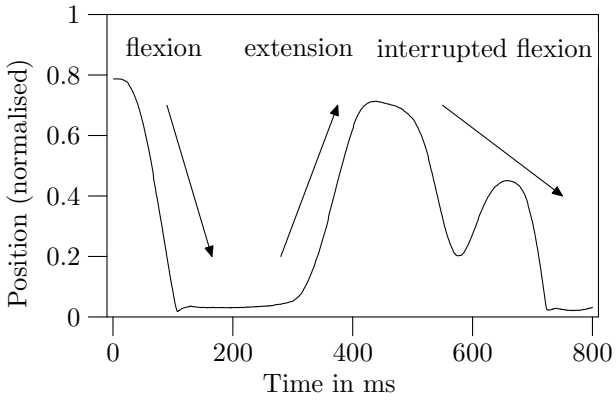

Figure 7.3.: A segment of a typical kinematic signal recorded in a tapping experiment. Two subsequent tapping movements are shown. The normalised position indicates the location of the tip of the index finger with respect to ground level. Note that the second flexion movement is interrupted, which occurs frequently in coordination experiments [Cong Khac et al., 2007].

of the small changes during the interrupted flexion movement, the minimal change magnitude was set to $h_{0\,min} = 0.2$, the minimal rise time was set to $\tau_{0\,min} = 40$, and the minimal duration after change was set to $s_{0\,min} = 30$. Using these parameters, SEMUG detected five changes as depicted in Figure 7.4, shaded are the estimated transitions. The segmentation shows by example that SEMUG can be successfully applied to data that does not strictly fulfil the ramp-step model.

The signal shown in Figure 7.5 (a) comprising an extension and a flexion movement serves as a second example. This example will show how to control the segmentation with properly chosen tuning parameters. The extension movement is prolonged by a short pause splitting the transient into two submovements. The analysis was done twice, (i) with the parameters already used in the previous example ($h_{0\,min} = 0.2$, $\tau_{0\,min} = 40$, and $s_{0\,min} = 30$) and (ii) by using less sensitive parameters ($h_{0\,min} = 0.4$, $\tau_{0\,min} = 70$, and $s_{0\,min} = 90$) in order to detect only the major changes. Figure 7.5 (b) and (c) display the detected changes and the estimated ramp-steps. The results confirm the expected behaviour of the algorithm. SEMUG detected the major changes with the less sensitive tuning parameters, whereas the submovements were detected additionally using the more sensitive setup.

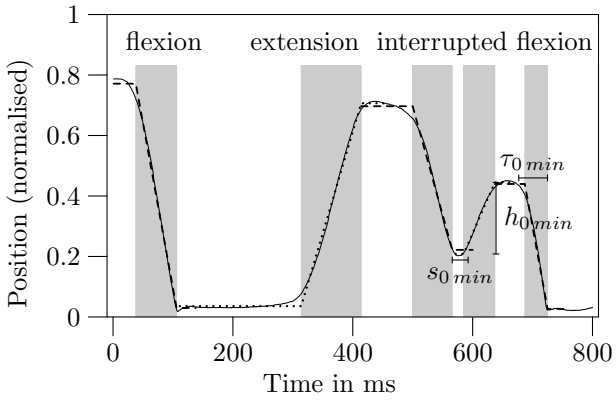

Figure 7.4.: Estimated ramp-step functions for the analysed signal of Figure 7.3. The five detected transitions are shaded. Dotted lines indicate positive magnitudes and dashed lines indicate negative magnitudes, respectively. The characteristics $h_{0\,min}$, $k_{0\,min}$ and $\tau_{0\,min}$ of the small change during the interrupted flexion movement were used for tuning the algorithm.

7.3.3. Analysis of a Long-Term Signal

In order to investigate the sensitivity of the method to varying tuning parameters in a real world setup, SEMUG was applied to a long-term segment recorded during a tapping experiment of $27\,s$ duration (i. e. a signal length of 27,000 samples). A Monte-Carlo-Simulation was performed on this signal with randomly varying window width L and detection threshold δ. Since, as will be shown below, s_{min} does not have direct influence on the test statistic, it was kept fixed at $s_{min} = 40$. The simulation comprised 10.000 runs with the aim to find parameter combinations (L, δ) with which SEMUG detected the changes correctly. The reference was given by an expert, who visually inspected the signal. The expert provided tolerance intervals for every parameter so that an outcome of SEMUG was labelled to be correct, only if every parameter of every detected change was in the respective tolerance interval. Neither a single false alarm (false positive) nor a single miss (false negative) was allowed.

Figure 7.6 shows a segment of the long-term tapping recording comprising a single movement. The tolerance intervals specified by the expert define an upper and a lower bound for correct ramp-step estimates, as indicated by the dashed lines.

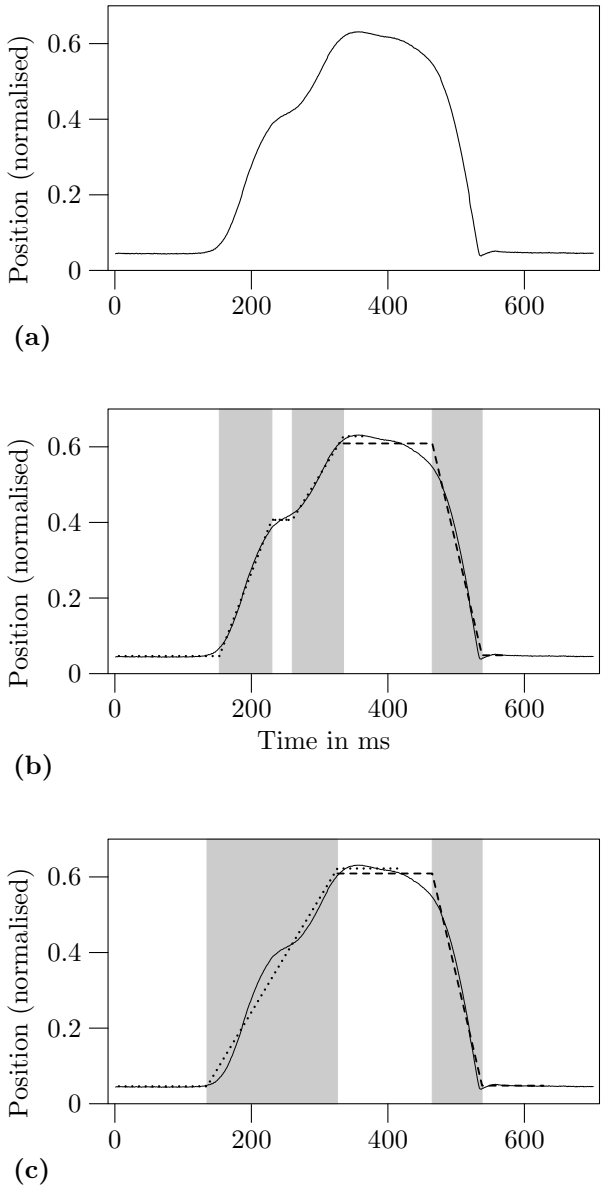

Figure 7.5.: (a) An extension movement comprising two submovements followed by a flexion movement. (b) Using sensitive tuning parameters ($h_{0\,min} = 0.2$, $\tau_{0\,min} = 40$, and $s_{0\,min} = 30$), the extension movement is modeled by two ramp-step functions. (c) With less sensitive tuning parameters ($h_{0\,min} = 0.4$, $\tau_{0\,min} = 70$, and $s_{0\,min} = 90$), the major changes are detected.

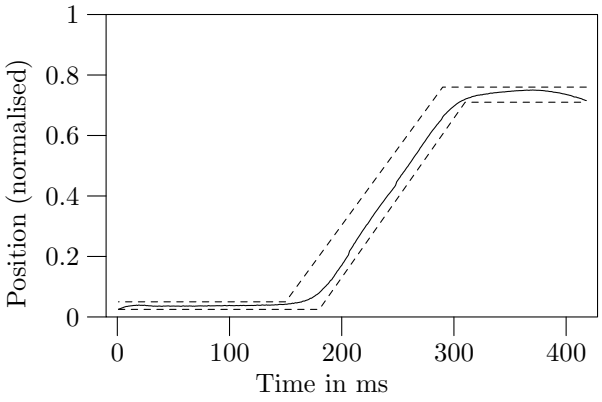

Figure 7.6.: A signal segment out of the long-term tapping recording. The area enclosed by the dashed lines is the tolerance region specified by an expert. When the estimated ramp-step is within this region, it is considered as a correct detection of this movement.

The expert identified 100 major changes with five of them comprising submovements (as in Figure 7.5). If the parameters (L, δ) were located in region A (Figure 7.7), SEMUG detected all the 100 major movements. If they were located in the more narrow region B (Figure 7.7), all submovements were additionally detected.

These results show, that the segmentation is predictable. Tuning parameters resulting in the same segmentation built contiguous intervals so that small changes in the tuning parameters do not change the segmentation significantly. This fact is studied in detail, next.

To round up the performance evaluation on measured data, an analysis of the robustness of change localisation against the tuning parameters was performed. The long-term tapping signal was analysed with (i) constant $\delta = 8$ and $s_{min} = 40$ and varying $L \in \{100, 125, 150, 175, 200\}$, (ii) constant $L = 150$ and $s_{min} = 40$ and varying $\delta \in \{6, 7, 8, 9, 10\}$, and (iii) constant $\delta = 8$ and $L = 150$ and varying $s_{min} \in \{30, 35, 40, 45, 50\}$. The mean values of the alarm time t_a and the change-point k across the 100 changes are depicted in Figure 7.8.

Since t_a and k are strongly dependent on the true change-points (which are not exactly known for measured data by principle), only relative differences are shown with the outcome of the parameter combination ($\delta = 8$, $L = 150$, and $s_{min} = 40$)

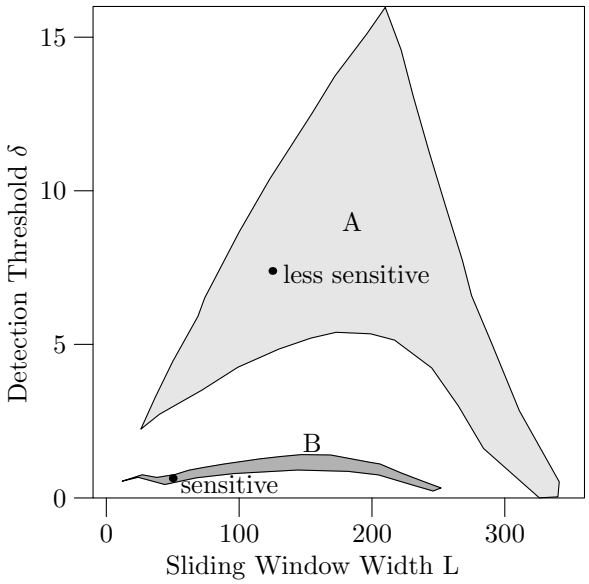

Figure 7.7.: Parameter mapping for the Monte-Carlo-Simulation performed on the long-term tapping signal. Tuning parameters (L, δ) in region A detected the 100 major taps. Parameters in region B are more sensitive and all the five submovements were detected additionally. The parameter combinations labelled by dots were used for the analysis shown in Figure 7.5.

serving as a reference.

Despite the large deviation of alarm times (Figure 7.8 (a), (b)), the change-point estimates in Figure 7.8 (d) to (f) show small fluctuations resulting in an average standard deviation below one. Note, that different scales are used for the panels (a), (b) and (c) to (f), respectively. The minimal post-change duration s_{min} does not have any direct influence on the alarm time (Figure 7.8 (c)), since this tuning parameter is used in the estimation process, only. The small deviation of t_a caused by s_{min} is due the fact that a difference in the estimation result in the i-th iteration causes a variation of $a^{(i+1)}$ which may cause a variation of t_a in the (i+1)-th iteration. But, since the effect of s_{min} on the change-point estimate k is below one sample (Figure 7.8 (f)), the effect on t_a is small, too. In summary, the results depicted in Figure 7.8 show that the change-points estimated with SEMUG are robust against some variation of tuning parameters.

This observation is supported by another study where the long-term tapping signal was analysed with in total $10,000$ randomly chosen parameter combinations (L, δ, s_{min}). The sample standard deviation of the alarm time t_a, the change-point k, the rise time τ, the magnitude h, and the offset d were computed for each change. The average standard deviation of the alarm time was high with $\overline{\sigma}_{ta} = 31.89$ samples in contrast to the standard deviation of the change-point $\overline{\sigma}_k = 1.34$ samples. Thus, it can be concluded that the localisation of the changes is robust against the time instant where the change is detected, which means in addition that it is robust against changes in the tuning parameters L and δ. The average standard deviation of the remaining parameters $\overline{\sigma}_\tau = 4.73$ samples, $\overline{\sigma}_h = 0.013$, and $\overline{\sigma}_d = 0.003$ was small, too. The reason for these good results is that the intervals $[a^{(i)}, b^{(i)}]$ where the ramp-steps are defined on, are fairly constant with $\overline{\sigma}_a = 3.51$ and $\overline{\sigma}_b = 17.40$ samples.

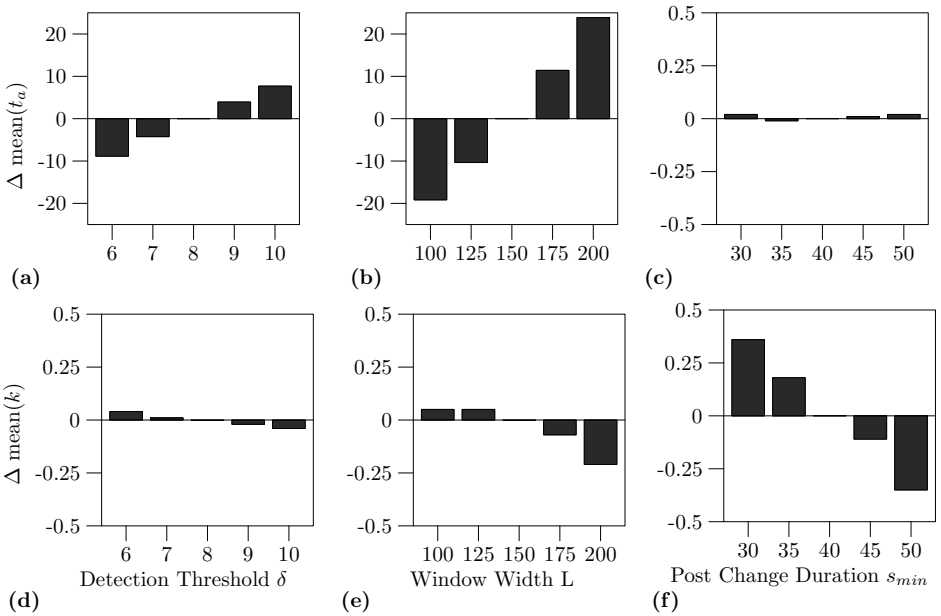

Figure 7.8.: The mean of the alarm time t_a and the change-point k obtained by analysing the long-term tapping signal. Depicted are the results relative to the outcome obtained by the parameter combination $\delta = 8$, $L = 150$, and $s_{min} = 40$. Note, that different scales are used for panels (a), (b) and (c) to (f), respectively. Despite the large variability of alarm times in (a), (b), the estimated change-point in (d) to (f) varies only in the sub-sample range, which demonstrates the robustness of the method.

8. Systems with High Disturbances

8.1. Introduction

In the last sections, it has been shown that SEMUG gives reliable results for signals with moderate and high SNR. Next, the behaviour of the algorithm will be studied on a simulated signal where noise has a significant influence. In addition, the case will be treated that the observed signal does not have a deterministic component at all, i. e., a pure noise signal. The simulated signal was composed of 10,000 adjacent ramp-step templates with randomly chosen parameters. The change magnitude was either positive or negative with an absolute value $|h_0| \in [0.2, 1]$, a rise time $\tau_0 \in [10, 21]$ and a post-change duration of at least 20 samples. SEMUG analysed the signal with four different noise levels. The first without noise as a reference study, the second with significant noise $\sigma = 0.1$, which is equivalent to half of the minimal change magnitude, the third with high noise $\sigma = 1$, and the last is a pure noise signal ($\sigma = 1, h_0 = 0$). In Figure 8.1, change patterns for the four different noise levels are depicted. It is obvious that the higher the noise level the more difficult the change is to detect.

SEMUG was tuned with $h_{0\,min} = 0.2$, $\tau_{0\,min} = 10$, and $s_{0\,min} = 20$. The most interesting outcome of this study, is the estimated rise time $\hat{\tau}$. Figure 8.2 displays histograms of $\hat{\tau}$ for the different noise conditions. The first panel shows the outcome of a perfect detection result. Since SEMUG detects every change correctly, $\hat{\tau}$ is uniformly distributed in the same range as the true rise time τ_0.

Adding significant noise blurs this distribution as depicted in panel (b) of Figure 8.2. Most of the changes were detected correctly and there was almost no false alarms and only a few of the smaller changes were missed.

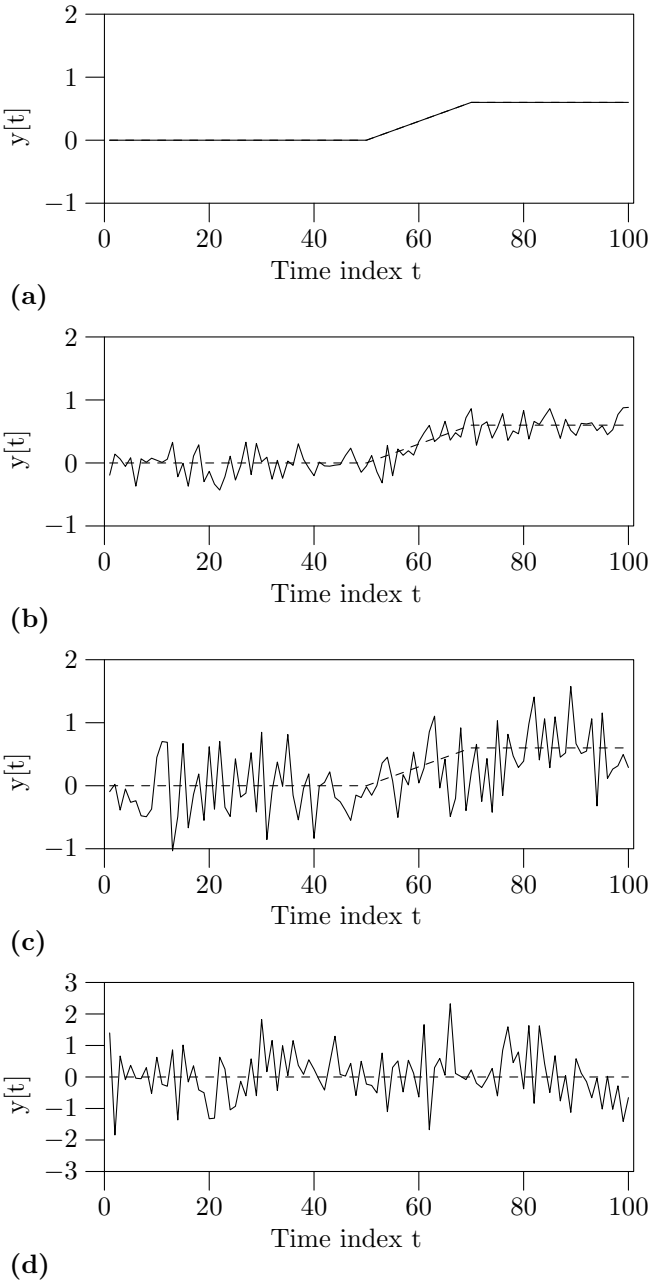

Figure 8.1.: Example change patterns for four different noise levels. The dashed line is the change pattern without noise and the solid line is disturbed by WGN with (a) $\sigma = 0$, (b) $\sigma = 0.1$, (c) $\sigma = 0.5$, and (d) is pure noise with $\sigma = 1$, $h_0 = 0$. In (a), (b) and (c) the change magnitude is kept constant with $h_0 = 0.6$.

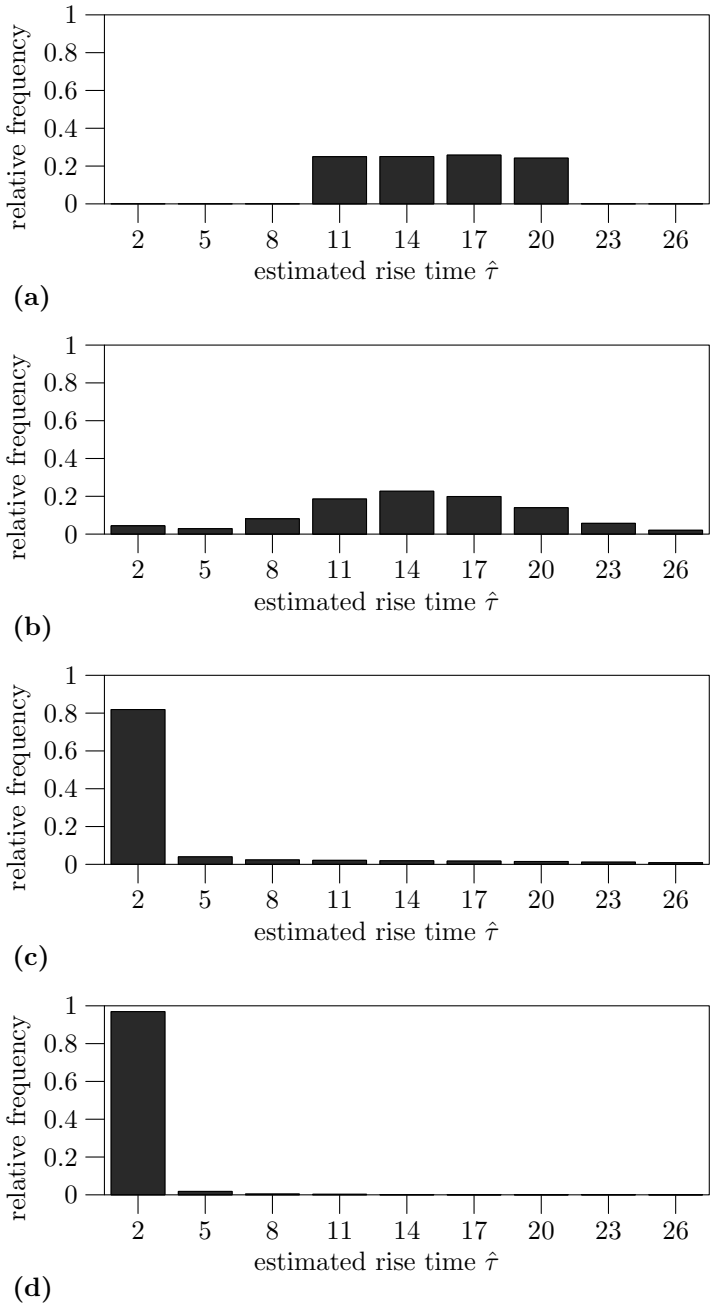

Figure 8.2.: Histograms of $\hat{\tau}$. The true rise time was randomly chosen out of the range $[10, 21]$. The study was repeated with four different noise conditions (a) $\sigma = 0$, (b) $\sigma = 0.1$, (c) $\sigma = 0.5$, and (d) a pure noise signal with $\sigma = 1$, $h_0 = 0$.

This changes for $\sigma = 0.5$. The number of false alarms raises drastically resulting in twice as many detected changes as there are in the signal. The noise is dominant which is clearly shown by the similarity to panel (d) representing the pure noise condition. The typical random effect of blurring the results visualised in panel (b) changed for a lower SNR to a seemingly deterministic result in panel (c) and (d). This chapter will clarify the reason for this.

Analysing systems with random effects can lead to wrong results when the noise component of the signal changes the signal's shape. This effect is illustrated in Figure 8.3. Displayed are two fragments of the simulated signal together with the true regression function (dashed line) and the estimated ramp-step (solid line). The error signal has a standard deviation $\sigma = 0.1$, which is therefore related to Figure 8.2 (b). The time is normalised to the beginning of the estimated ramp-step's domain. In the upper panel in Figure 8.3 the estimation result is good despite the presence of significant noise. In the lower panel random effects are the cause that the rise time is underestimated. The additive error signal is positive in the beginning of the transition phase and negative in the end which changes the signal's shape to a ramp-step with lower rise time. The frequency of these cases is depicted by the first two bars in the histogram of Figure 8.2 (b).

The histograms Figure 8.2 (c) and (d) cannot be reasoned by occasionally occurring noise effects since they are significantly different to the histograms in (a) and (b). The outcome in the pure noise case will next be explained by (i) interpreting the ML estimate in terms of classification and (ii) calculating the probability that a signal is classified to a ramp-step with a specific rise time.

8.2. Scale and Shift Invariant Classification

In this section, the estimation of a ramp-step (see Section 5.2.2) will be treated in terms of classification. Classification is the problem of making a decision among multiple hypotheses.

The part of SEMUG that can be solved by classification is the estimation of the

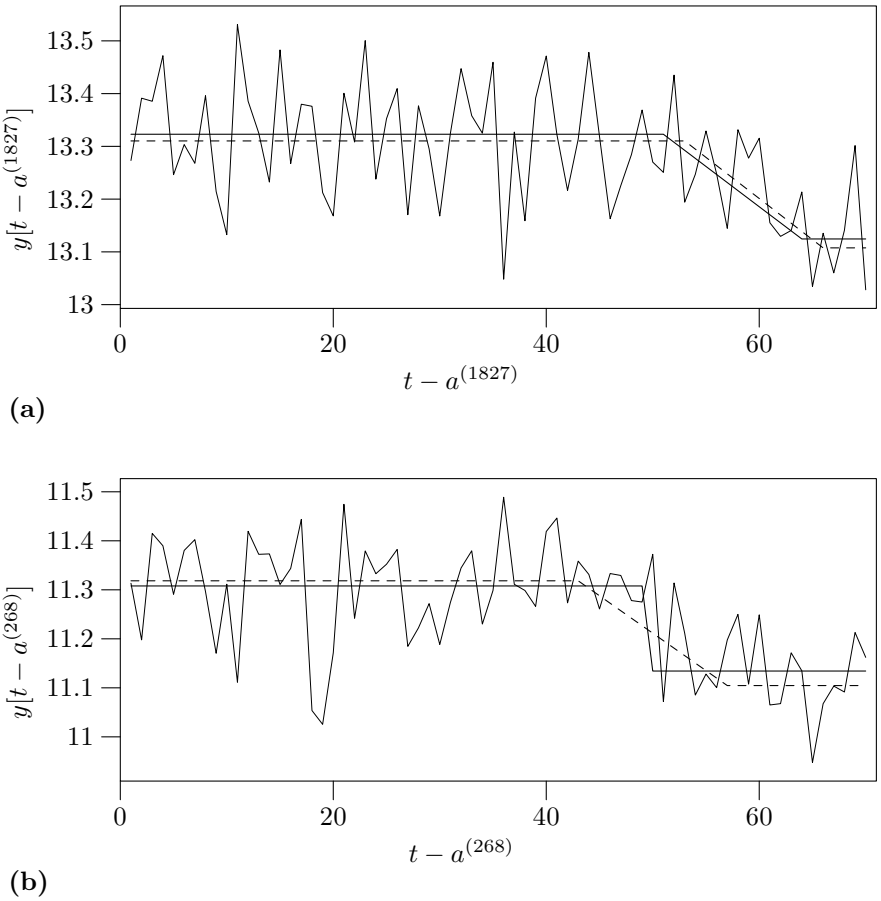

Figure 8.3.: Two fragments of the simulated signal displayed together with the true regression function (dashed line) and the estimated ramp-step (solid line). The standard deviation of the error signal is $\sigma = 0.1$ and the time is normalised to the beginning of the estimated ramp-step's domain. In the upper panel the estimated ramp-step is close the the true function despite the influence of the error signal. In the lower panel the rise time is underestimated due to an unfortunate noise signal during the transition phase. The additive error signal is positive in the beginning of the transition phase and negative in the end which changes the signal's shape to a ramp-step with lower rise time.

ramp-steps' change-point k and rise time τ. For this problem it is supposed that the signal is one of the c hypotheses $\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_c$ where every hypothesis has a unique parameter combination (k, τ). The number of possible combinations of these two parameters is finite which gives a finite number of hypotheses. For, e. g., $m = 3$ there are three combinations $(k = 1, \tau = 1)$, $(k = 1, \tau = 2)$, and $(k = 2, \tau = 1)$. In general, there are $c = m(m-1)/2$ combinations for $m >= 2$.

Every hypothesis $\mathcal{H}_i$ is represented by a unique change pattern. The problem of classification is now to make the decision which hypothesis best fits to the measured signal. A common criterion for this decision is to minimise the probability for a wrong decision, i. e., minimise the probability of an error denoted by P_e. In Section 2.3.2 it is shown that there is a strong connection between classification and maximum likelihood when the minimum P_e criterion is used. In fact both methods give the same results when the hypotheses have equal prior probabilities. For the ramp-step estimation there is no reason to not assign equal prior probabilities to the hypotheses.

The notable difference between maximum likelihood and classification is that classification is rather a selection than an estimation and it is restricted to a finite set of hypotheses. In classification a template is represented by a point in $\mathbb{R}^m$ denoted by $\boldsymbol{p}$ where the t-th component of $\boldsymbol{p}$ is equal to the ramp-step signal at time index t. This definition allows to transfer the formulas from Section 5.2.2 one-to-one so that with (5.25) the minimum P_e method decides for $\mathcal{H}_i$ if

$$\left|\boldsymbol{y}^T\boldsymbol{p}_i\right| \geq \left|\boldsymbol{y}^T\boldsymbol{p}_j\right| \quad \text{with} \quad j = 1, 2, \ldots, c\,;\; i \neq j\,;\; \boldsymbol{p}_i, \boldsymbol{p}_j \in P\,. \tag{8.1}$$

where P denotes the set of possible templates and the templates $\boldsymbol{p}_i$ and $\boldsymbol{p}_j$ have zero mean and unit norm so that they fulfil the regularity conditions

$$\bar{\boldsymbol{p}}_i = 0 \tag{8.2}$$

$$\|\boldsymbol{p}_i\| = 1 \quad \text{with} \quad j = 1, 2, \ldots, c\,;\; \boldsymbol{p}_i \in P \tag{8.3}$$

as defined in Section 5.2.2. With (8.1) the decision is done in favour of $\mathcal{H}_i$ when

the absolute value of the inner product of the signal $\boldsymbol{y}$ and template $\boldsymbol{p}_i$ is maximal. Without loss of generality, it can be assumed, that for each $\boldsymbol{p}_i \in P$ there exist a $\boldsymbol{p}_j \in P$ with $\boldsymbol{p}_i = -\boldsymbol{p}_j$. As a result, if the minimum P_e method decides for $\mathcal{H}_i$, the hypothesis $\mathcal{H}_j$ is true, too and vice versa. As a consequence, the absolute value operation in (8.1) can be cancelled, so that the final decision rule is

$$\boldsymbol{y}^T\boldsymbol{p}_i \geq \boldsymbol{y}^T\boldsymbol{p}_j \quad \text{with} \quad j = 1, 2, \ldots, c\,;\; i \neq j \tag{8.4}$$

These inequalities can be written in a convenient matrix form

$$A_i\boldsymbol{y} \leq 0 \tag{8.5}$$

with

$$A_i = \begin{pmatrix} (\boldsymbol{p}_1 - \boldsymbol{p}_i)^T \\ \vdots \\ (\boldsymbol{p}_{i-1} - \boldsymbol{p}_i)^T \\ (\boldsymbol{p}_{i+1} - \boldsymbol{p}_i)^T \\ \vdots \\ (\boldsymbol{p}_r - \boldsymbol{p}_i)^T \end{pmatrix} . \tag{8.6}$$

A geometrical interpretation of (8.5) is that $A_i\boldsymbol{y} \leq 0$ defines a region in $\mathbb{R}^m$ and if the measured signal $\boldsymbol{y}$ is in this region, the ML rule decides for $\boldsymbol{p}_i$. This region is termed decision region and is denoted by R_i. Its size and shape is defined by the templates in P. As in the general case, treated in Section 2.3.3, the geometrical shape of the decision region is a convex polyhedron.

In 2-dimensions a convex polyhedron is termed convex polygon which is a closed path composed of a finite sequence of straight line segments. The straight line segments are called edges and they are the decision boundaries of R_i. In 3-dimensions the decision boundaries are planes and in higher dimensions they are termed hyperplanes.

The convex polyhedron defined by (8.5) has some special properties. One is that the decision boundaries are always in the middle of two templates. Since they are

defined by

$$\boldsymbol{y}^T\boldsymbol{p}_i = \boldsymbol{y}^T\boldsymbol{p}_j \tag{8.7}$$

$$\Leftrightarrow \quad (\boldsymbol{p}_i - \boldsymbol{p}_j)^T\boldsymbol{y} = 0\ . \tag{8.8}$$

This defines a hyperplane through the origin which is orthogonal to the vector $(\boldsymbol{p}_i - \boldsymbol{p}_j)$ and since $\|\boldsymbol{p}_i\| = \|\boldsymbol{p}_j\| = 1$ the midpoint between $\boldsymbol{p}_i$ and $\boldsymbol{p}_j$ is on this hyperplane, which is simply proven by replacing $\boldsymbol{y}$ with the midpoint $(\boldsymbol{p}_i + \boldsymbol{p}_j)/2$ in (8.8)

$$(\boldsymbol{p}_i - \boldsymbol{p}_j)^T\left(\frac{\boldsymbol{p}_i + \boldsymbol{p}_j}{2}\right) = \frac{1}{2}\left(\boldsymbol{p}_i{}^T\boldsymbol{p}_i - \boldsymbol{p}_j{}^T\boldsymbol{p}_j\right) \tag{8.9}$$

$$= 0\ . \tag{8.10}$$

A second property is that independent from A_i, the origin attains (8.5) which means that the origin is element of every decision region and if an observation $\boldsymbol{y}$ attains (8.5) the scaled observation $\alpha\boldsymbol{y}$ attains (8.5), too when $\alpha \geq 0$, which means that the decision regions are unbounded ($\alpha \to \infty$). A simple proof follows from the commutativity law that $A_i(\alpha\boldsymbol{y}) = \alpha A_i\boldsymbol{y}$ and since $A_i\boldsymbol{y} \leq 0$, multiplication with a positive constant does not change the sign.

A polyhedron with these properties is called a *polyhedral cone*. Note that it has the origin as an extreme point. In 3-dimensional space this extreme point is called apex of the cone. Figure 8.4 displays a polyhedral cone in the 3-dimensional space.

8.3. The Pure Noise Case

In the previous section it has be explained that a scale and shift invariant classification leads to a partitioning of $\mathbb{R}^m$ where every decision region is a polyhedral cone (see Figure 8.4).

Since $\boldsymbol{y}$ is normally distributed with mean zero—in the pure noise case—and since the normal distribution is rotationally symmetric, the probability that a par-

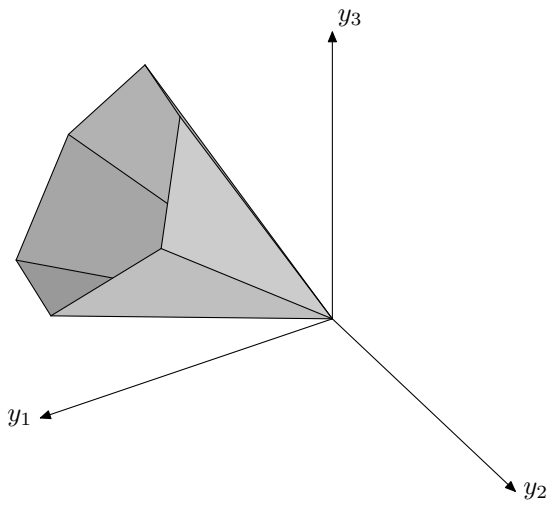

Figure 8.4.: A polyhedral cone in 3-dimensional space.

ticular template $\boldsymbol{p}_i$ is selected in a pure noise signal, denoted by $P(\boldsymbol{y} \in R_i | \, \boldsymbol{u} = 0)$, depends only on the size of R_i and since R_i is solely defined by the templates in P other statistical properties like the noise variance σ have no influence. A rotational symmetric distribution of $\boldsymbol{y}$ means that many realisations of $\boldsymbol{y}$ build up a spherical cloud with a centre in the origin. A higher variance σ does only expand this cloud but since the decision regions are polyhedral cones, the expansion does not change the decision region where the individual $\boldsymbol{y}$ are placed.

This is especially important in practical applications where statistical properties like σ^2 are scarcely known. The invariance from σ^2 in the pure noise case means that, i. e., the histogram in Figure 8.2 (d) is unique, it is obtained when noise is too high for reliable results. Application scientists should be alerted when observing such a distribution when they have doubts that the signal comprises only abrupt changes.

The histogram in Figure 8.2 (d) can be determined by computing the size of the surface area of the m-dimensional unit sphere that intersects the individual decision regions. This surface area divided by the complete surface area of the m-dimensional unit sphere is called solid angle.

When Ω_i is the solid angle of the polyhedral cone R_i the equation

$$P(\boldsymbol{y} \in R_i | \, \boldsymbol{u} = 0) = \Omega_i / S_m \tag{8.11}$$

holds. S_m is the surface area of the m-dimensional unit sphere

$$S_m = 2\frac{\pi^{m/2}}{\Gamma(m/2)} \tag{8.12}$$

where Γ is Euler's Gamma function. An alternative formula is

$$S_m = \begin{cases} \frac{m(2\pi)^{m/2}}{2 \cdot 4 \cdots m} & \text{if } m \text{ is even;} \\ \frac{2m(2\pi)^{(m-1)/2}}{1 \cdot 3 \cdots m} & \text{if } m \text{ is odd.} \end{cases} \tag{8.13}$$

The computation of the solid angle in higher dimensions is an open problem in computational geometry (see Hajja and Walker [2002], Nunemacher [1999]). A prove of concept will be given next for low dimensions.

8.4. Solid angle in the 2-dimensional space

For $m = 3$ there are three ramp-step templates with parameters $(k = 1, \tau = 1)$, $(k = 1, \tau = 2)$, and $(k = 2, \tau = 1)$. The templates are accordingly

$$\boldsymbol{p}_1 = \left(\frac{-2\sqrt{6}}{6}, \frac{\sqrt{6}}{6}, \frac{\sqrt{6}}{6} \right)^T \tag{8.14}$$

$$\boldsymbol{p}_2 = \left(\frac{-\sqrt{2}}{2}, 0, \frac{\sqrt{2}}{2} \right)^T \tag{8.15}$$

$$\boldsymbol{p}_3 = \left(\frac{-\sqrt{6}}{6}, \frac{-\sqrt{6}}{6}, \frac{2\sqrt{6}}{6} \right)^T \quad . \tag{8.16}$$

The set of templates is extended by the negative counterparts of $\boldsymbol{p}_1$, $\boldsymbol{p}_2$, and $\boldsymbol{p}_3$, which allows an easier computation as explained in the previous section (see page 105). Consequently the templates $P = \{\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_6\}$ with $\boldsymbol{p}_4 = -\boldsymbol{p}_1$, $\boldsymbol{p}_5 = -\boldsymbol{p}_2$, and $\boldsymbol{p}_6 = -\boldsymbol{p}_3$ are considered, where each template fulfils two regularity

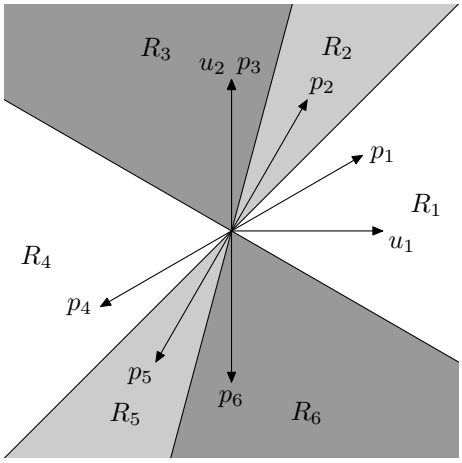

Figure 8.5.: The ramp-step templates for $N = 3$. The templates are all element of the plane $U_m = \{\boldsymbol{y} \in \mathbb{R}^m | \sum_{i=1}^m y[i] = 0\}$ which allows a 2d-plot. The decision regions that belong together are shaded in the same colour.

conditions which are (i) zero mean and (ii) unit norm. From the first regularity condition it follows that each template is element of the plane

$$U_m = \left\{\boldsymbol{y} \in \mathbb{R}^m | \sum_{i=1}^{m} y[i] = 0\right\}, \tag{8.17}$$

which is a plane trough the origin. The intersection $R_i^U = R_i \cap U_m$ is also a polyhedral cone with solid angle Ω_i^U and the ratio of this Ω_i^U to the sphere S_{m-1} is the same as the ratio of Ω_i to S_m so that the equation

$$\frac{\Omega_i^U}{S_{m-1}} = \frac{\Omega_i}{S_m} \tag{8.18}$$

holds. The probability that $\boldsymbol{p}_i$ is selected in a pure noise signal, defined by (8.11), can therefore be solved in the subspace U_m. The projections of the template vectors on U_3 using the orthonormal basis $u_1 = (-1, 1, 0)^T/\sqrt{2}$ and $u_2 = (-1, -1, 2)/\sqrt{6}$ are displayed in Figure 8.5. Apparent is the symmetry of the templates which results in a symmetry of the decision regions $R_1, \ldots, R_6$. The decision boundaries are rays from the origin through the midpoint of the line segment between two templates.

Table 8.1.: The solid angle of the decision regions and the probability that $\boldsymbol{y}$ is in R_i if $\boldsymbol{y}$ is pure noise, denoted by $P(\boldsymbol{y} \in R_i \,|\, \boldsymbol{u} = \boldsymbol{0}) = \Omega_i / S_2$.

	V_1	V_2	V_3	V_4	V_5	V_6
Solid angle Ω_i	75°	30°	75°	75°	30°	75°
$P(\boldsymbol{y} \in R_i \,\|\, \boldsymbol{u} = \boldsymbol{0})$	0.208	0.083	0.208	0.208	0.083	0.208

The solid angle for each decision region can be computed using simple trigonometry. They are listed in Table 8.1 together with the probability that $\boldsymbol{y}$ is in R_i for a pure noise signal. This probability can be computed according to (8.11) by dividing the solid angle by 360°. The distribution over the estimated rise time can be obtained by combining the templates with equal rise time. The templates $\boldsymbol{p}_1$, $\boldsymbol{p}_3$, $\boldsymbol{p}_4$, and $\boldsymbol{p}_6$ have a rise time of one where as the templates $\boldsymbol{p}_2$, and $\boldsymbol{p}_5$ have a rise time of two. Thus the probability that a signal with estimated rise time $\hat{\tau} = 1$ is observed

$$P(\hat{\tau} = 1 |\, \boldsymbol{u} = \boldsymbol{0}) = (\Omega_1 + \Omega_3 + \Omega_4 + \Omega_6)/S_2 \tag{8.19}$$

$$= 300°/360° = 0.833\,. \tag{8.20}$$

Analogous the probability for observing $\hat{\tau} = 2$

$$P(\hat{\tau} = 2 |\, \boldsymbol{u} = \boldsymbol{0}) = (\Omega_2 + \Omega_5)/S_2 \tag{8.21}$$

$$= 60°/360° = 0.167\,. \tag{8.22}$$

The higher probability for the step template is apparent which has also been observed in a more general setup (see Figure 8.2 (d)).

8.5. Solid angle in the 3-dimensional space

For $m = 4$ the number of possible ramp-step templates rises to six. Adding the negative counterpart ($\boldsymbol{p}_{i+6} = -\boldsymbol{p}_i$) results in a set of templates comprising twelve templates $P = \{\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_{12}\}$ summarised in Table 8.2. Objective is to determine

Table 8.2.: The change-point and the rise time of the ramp-step templates for $m = 4$.

		τ		
		1	2	3
k	1	$\boldsymbol{p}_1, \boldsymbol{p}_7$	$\boldsymbol{p}_2, \boldsymbol{p}_8$	$\boldsymbol{p}_3, \boldsymbol{p}_9$
	2	$\boldsymbol{p}_4, \boldsymbol{p}_{10}$	$\boldsymbol{p}_5, \boldsymbol{p}_{11}$	—
	3	$\boldsymbol{p}_6, \boldsymbol{p}_{12}$	—	—

the solid angle of the decision regions R_i. As explained in the previous section, the problem can be transformed on the plane U_4 defined by (8.17), where the orthonormal basis $u_1 = (1, -1, 0, 0)^T/\sqrt{2}$, $u_2 = (0, 0, 1, -1)/\sqrt{2}$, and $u_3 = (1, 1, -1, -1)/2$ can be used.

The first step in order to compute the solid angle of R_i is to find the neighbours of $\boldsymbol{p}_i$. A template $\boldsymbol{p}_j$ is termed a neighbour of $\boldsymbol{p}_i$ if the decision region R_i increases when removing $\boldsymbol{p}_j$ from P. Finding the neighbours is equal to removing redundant inequalities from (8.5). This is a subject of a so-called *vertex enumeration* algorithm like presented in Avis and Fukuda [1992]. A vertex enumeration algorithm determines the vertices of a polyhedron where redundant inequalities have no effect on the result. The algorithm of Avis and Fukuda [1992] implies that the polyhedron is bounded. This can be achieved by adding the constraint $\boldsymbol{p}_i{}^T\boldsymbol{y} \leq \boldsymbol{p}_i{}^T\boldsymbol{p}_i$ which restricts the extend of the polyhedral cone to $\boldsymbol{p}_i{}^T\boldsymbol{p}_i$. Note that this polyhedron would not be bounded in the degenerated case when all templates are on a single plane, but in this case the problem can be transformed in two dimensions by using a orthonormal basis of this plane and can then be treated as shown in the previous section.

In the non degenerated case the polyhedron $R_i' = \{\boldsymbol{y} |\, A_i'\boldsymbol{y} \leq b'\}$ is bounded with

$$A_i' = \begin{pmatrix} A_i \\ \boldsymbol{p}_i{}^T\boldsymbol{y} \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} \mathbf{0} \\ \boldsymbol{p}_i{}^T\boldsymbol{p}_i \end{pmatrix} . \tag{8.23}$$

The result of vertex enumeration will be a minimal set of points $\{o, v_1, v_2, ..., v_{m_i}\}$ that define R_i' where o is the origin and m_i is the number of neighbours of $\boldsymbol{p}_i$.

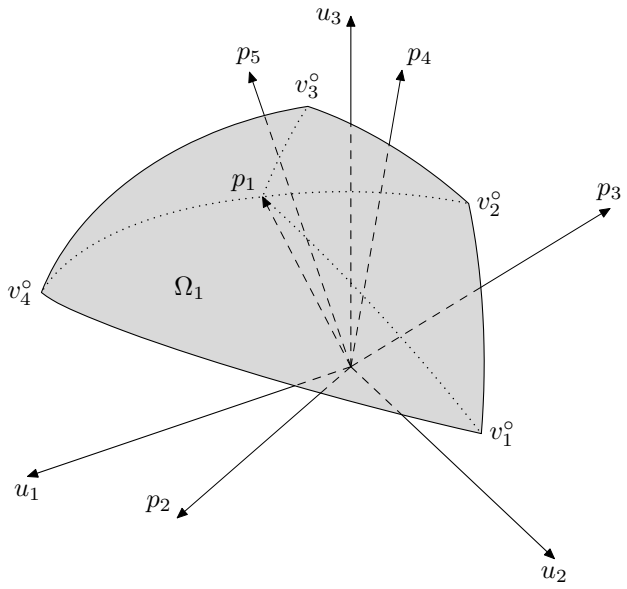

Figure 8.6.: The decision region R_1. Its solid angle is the sum of the four spherical triangles $p_1v_1^\circ v_2^\circ$, $p_1v_2^\circ v_3^\circ$, $p_1v_3^\circ v_4^\circ$, and $p_1v_4^\circ v_1^\circ$.

This allows to describe R_i by the minimal set of rays $\{\overrightarrow{ov}_1, \overrightarrow{ov}_2, \ldots, \overrightarrow{ov}_{m_i}\}$. The points where these rays intersect the unit sphere, define the solid angle Ω_i. They are denoted by $\{v_1^\circ, v_2^\circ, ..., v_{m_i}^\circ\}$. Figure 8.6 illustrates an example where $\boldsymbol{p}_1$ has four neighbours namely $\boldsymbol{p}_2$, $\boldsymbol{p}_3$, $\boldsymbol{p}_4$, and $\boldsymbol{p}_5$. The solid angle Ω_1 is shaded and the vertices $\{v_1^\circ, v_2^\circ, v_3^\circ, v_4^\circ\}$ are depicted. The solid angle Ω_1 is now calculated by summing up the area of the spherical triangles $p_1v_1^\circ v_2^\circ$, $p_1v_2^\circ v_3^\circ$, $p_1v_3^\circ v_4^\circ$, and $p_1v_4^\circ v_1^\circ$. Measuring the area of a spherical triangle dates back to work of Euler [1778] and Lagrange [1798]. Although solved, it lasts until present days until a simple formula was found. Eriksson [1990] proposes that the area Ω_{ijk} of a spherical triangle $p_iv_j^\circ v_k^\circ$ can be expressed in terms of the inner products and the triple product $[p_i, v_j^\circ, v_k^\circ] = p_i^T(v_j^\circ \times v_k^\circ)$ as

$$\tan\frac{\Omega_{ijk}}{2} = \frac{|[p_i, v_j^\circ, v_k^\circ]|}{1 + p_i^T v_j + p_i^T v_k + v_j^T v_k} \,. \tag{8.24}$$

Table 8.3 lists the probabilities of the ramp-step templates in the pure noise condition. The last row sum up the probabilities for the templates with equal rise

Table 8.3.: The probability that a ramp-step template is estimated in a pure noise signal for m=4.

		τ		
		1	2	3
k	1	0.239	0.06	0.087
	2	0.179	0.256	—
	3	0.179	—	—
Σ		0.597	0.316	0.087

time. The preference for low rise times is evident as in higher dimensions (see Figure 8.2 (d)).

9. Conclusions

The problem addressed in this thesis is the computer-aided segmentation of signals comprising gradual changes. The segmentation problem has a long history dating back to the 1930s. In the early days the solutions allowed at most a single change-point which splits up the signal in two segments. In parallel to the progress of the computer systems other segmentation problems with higher computational effort were solved so that algorithms were developed for the segmentation of signals with complex modelling and for the segmentation of large scale signals with many segments. This thesis presents an algorithm termed SEMUG that fits in the latter branch of research.

SEMUG is based on the often made assumption that the signal is a linear function on the segments (see examples in the chapters 1 and 3). This model is not only suitable for systems where the output is a linear function but it may also be used for systems where the output value comprises complex change patterns. Because for some cases, the complex change patterns in the output value can be traced back to linearly changing internal parameters so that the model is used for the internal parameters except of the output value.

As described in Chapter 3 the piecewise linear segmentation problem can be divided in two groups. The first tackles the problem where the signal may jump at the change-points. This problem can be viewed to be solved even in the large-scale case. The second group deals with the problem where the signal is continuous in the change-points. The principle solution to this problem fits the model by considering every possible segmentation. This has the shortcoming that it does not scale well with the number of change-points and the signal size, which also means that it

is hardly on-line applicable since with proceeding time the computational effort rises unrestrictedly. SEMUG fills this gap. It is an algorithm specifically suited for processing large-scale segmentation problems.

SEMUG solves the segmentation problem by separating the detection from the localisation of a change event. It works sequentially where the change-points are detected one after another. This approach has two advantages. (i) The computational complexity does only rise linearly with the signal size so that it is a suitable method to solve large scale problems. (ii) It is on-line applicable since no future samples are used in the computations.

The detection of the next change is done with a windowed GLRT. It is faster than the standard GLRT while having the same reliability with a proper tuning (see Chapter 6). Other authors propose other change detection methods like the CUSUM which is utilised by Charbonnier et al. [2004]. However, Han and Tsung [2005] show that the GLRT outperforms the CUSUM when detecting a dynamic mean change that finally approaches a steady state.

SEMUG uses not only the likelihood principle for detecting a change but also for determining its location. The localisation is based on the ramp-step model. The ramp-step is a change profile with a linear transition between two steady states. It is the generalisation of the often used step and ramp profiles. Since it includes these profiles as special cases it can be applied for many change-point problems with fast or slow transitions.

Despite its wide applicability, the ramp-step model depends on four parameters, only, which define the beginning of the change (the change-point) and the duration of the linear transition as well as the level of the two steady states. For the latter two, explicit formulas are found to compute their MLE which is describe in Section 5.2.2 and in the Appendix A in more detail. These explicit formulas reduce the objective function for the remaining parameters to a single inner product allowing a very easy and efficient implementation.

Note that for more complex transition models the objective function would not reduce so nicely. Imaginable are transition profiles with additional change-points

in order to model the beginning phase and the end phase of a transition more accurately. The ramp-step can therefore be seen as a trade-off between modelling accuracy and computational complexity so that large scale signals can be processed in reasonable time which is the first important request on SEMUG defined in Chapter 4.

In this chapter it is stated that the per change computation time should be below 2 s. SEMUG is far below this threshold on a standard PC with a 2.8 GHz Intel© Pentium© 4 CPU. It process a biomechanical signal with 100 gradual changes in 15 s which is a per change computation time of 0.15 s (see Section 7.3 for a description of the biomechanical signal).

The locally optimal MLE of the ramp-step has a good performance, influences of stochastic and deterministic errors are rather low as shown in the sections 7.2 and 7.3 for simulated and measured data, respectively.

Important for sequential algorithms is that errors do not propagate from one iteration to another. The performance evaluation on simulated data show that a deterministic error does not effect the localisation of the next change. SEMUG is especially robust against variations of the time instant when the change is detected (the alarm time). It has been demonstrated on measured data that a varying alarm time, due to variations in the tuning parameters, has only little effect on the estimated change-point location and change duration and since the detection of a subsequent change is initialised with these estimates, the overall process gives robust and reliable results, too. The reliability is the second important request on the algorithm stated in the problem definition in Chapter 4.

Another request found in the problem definition is that the tuning should be easy and transparent. The tuning system developed in Chapter 6 is based on prototypical change transitions defined by the operator. The signal characteristics used are the minimal values of the change's magnitude and rise time as well as the minimal post-change duration. These are only a few and visually apparent signal characteristics. This is a big advantage in comparison to specifying the detection threshold directly. However, a tuning system that suits every needs can hardly

be found. A good tuning system depends only on available knowledge about the system. It is therefore always a trade-off between the amount of required knowledge against accuracy of the results.

The proposed tuning system does not require to know detailed statistical properties of the system, e. g., the variance of the error signal, but it requires that the operator has some knowledge about the least significant change in the data set. The performance analysis on simulated and measured data has shown that the tuning system gives the expected segmentation.

In summary SEMUG does attain the following requests:

- Automatically perform a segmentation of a signal composed of adjacent ramp-step profiles.
- Process large-scale signals in reasonable time.
- To provide a reliable and transparent tuning system.

With that, the goals specified in Chapter 4 are fulfilled completely.

This thesis provides a solution for a specific change-point problem which can be used as it is or it can serve as a source of ideas for the solution of related problems. E. g., the separation of detection and estimation has many advantages in a large-scale setup. It allows a fine grained and time consuming estimation with locally optimal results while preserving an acceptable overall computational complexity. However, a globally optimal solution for the segmentation of the continuous piecewise linear model does remain as a topic for future research.

The exponential growth of the number of possible segmentations is the key problem that has to be solved for algorithms based on the continuous piecewise linear model. In a large-scale setup it is necessary to ignore segmentations which are unlikely to be the true one.

This idea let to the dynamic programming algorithm. It is a sophisticated search strategy starting from a two segment partition which will be used to reduce the number of considered three segment partitions (see Bai and Perron [2003] for de-

tails). The number of change-points is determined in a post-processing step by using an information criterion.

A different search strategy is utilised by the algorithm presented by Vostrikova [1981]. This algorithm starts with the two segment partition, too, but it then splits the signal at the found change-point and does further processing on the two parts separately.

These two approaches reduce the considered segments by objective criteria. Other approaches use a-priori knowledge about the application. This would be a minimal size of the segments or maximal number of change-points. When it is known that the change-points are uniformly distributed the signal could be split up in sufficient long parts in a pre-processing step, where every part is processed separately. The list of possible constraints is long and a good algorithm is able to profit from them. In Section 5.3.2 examples are given how a priori knowledge can be incorporated into SEMUG.

In fact, the two stage approach of SEMUG does a significant contribution in this sense. Based on the assumption of a reliable detection the signal is split up in segments with exactly one change which is then located on the second stage. It is not an easy task to define the conditions in which this data-driven split gives good results. This will be a topic for future work.

Additionally, future work will ease the restriction of the linear model by using a model which incorporates more information about the change transition profile to increase the accuracy of the method. A prominent one are dynamical system. In the introduction of Chapter 3 it is explained that dynamical systems are common models in classical mechanics and other scientific areas. In comparison to the linear model they give a better approximation close to the true change-point since the transition of one to another segment is smooth in many applications. However, the estimation of the dynamical system's parameters must be repeated for every possible change-point which is time consuming.

In change-point analysis the identification of the system is different to the standard identification problem since the input signal is not necessarily measurable

which is termed blind identification. E. g., for finger tapping the input signal is the signal of the brain to the motor neurons of the finger. As a first approximation this excitation can be modelled by an impulse which is the initial trigger for the movement. With this input signal and the monitored output the parameters of the system can be identified, where it makes sense to use a weighting mechanism so that the samples right after the impulse have a greater impact since the other parts of the output signal might be obscured with subsequent excitation of the motor neurons.

Future work will incorporate this approach in SEMUG seamlessly, where SEMUG with the ramp-step template is responsible for a reliable detection and localisation of the change transitions. Then, the advanced dynamical system model is used to improve the estimate of the change-point.

The performance of SEMUG for systems with high disturbances will be addressed in future work, too. This thesis started by focusing on pure noise signals where it could be shown that the distribution of the rise time can serve as an indicator whether the segmentation is reliable (see Chapter 8). Future work may regard signals with arbitrary SNR, which will require research about the properties of polyhedral cones in high dimensional spaces. Together with the dynamic model of the transition profile it is another item on the task list on the route to the goal to provide a most sophisticated but simple to use segmentation algorithm for signal processing.

A. Scale and Shift Invariant Template Estimation

Estimation theory deals with the problem to determine the parameters of the model which describes the monitored signal $y[t]$. Aim is to gain information describing the state of the observed process. The signal model

$$y[t] = u[t] + w[t] \tag{A.1}$$

is considered which is equal to that used in Section 2.1.4, where $u[t]$ is termed the (noise-fee) regression function and $w[t]$ is a pure noise component. In Section 2.1.4 it is described that the ML method is a widely used and flexible method for estimating the parameters $\boldsymbol{\theta}$ on those $u[t]$ depends. Aim in this appendix is to extend the ML method in order to give a scale and shift invariant estimate. The scale and shift invariant MLE should be preferred when the amplitude and the offset of $u[t]$ holds only minor information.

The derivation of the scale and shift invariant MLE is based on Section 2.1.4 where the MLE for a change in mean is derived for a model with an arbitrary mean signal $u[t]$. $u[t]$ depends an a parameter θ which is estimated by seeking the value of the parameter that maximises the likelihood function

$$\hat{\theta} = \arg \max_{\theta} \ln \mathrm{L}(\theta; \boldsymbol{y})$$

which in the Gaussian case is equal to maximising Λ

$$\hat{\theta} = \arg\max_{\theta} \Lambda(\theta; \boldsymbol{y})$$

with

$$\Lambda(\theta; \boldsymbol{y}) = -\sum_{t=1}^{N} \left(y[t] - u[t]\right)^2 \ .$$

Suppose now that the regression function denoted by $p[t]$ depends on several unknown parameters collected in the parameter vector $\boldsymbol{\theta}$. The MLE of $\boldsymbol{\theta}$ should be invariant with respect to multiplication or summation of $p[t]$ with any constant, denoted by α and β, respectively. The model for the observed signal is consequently

$$y[t] = \alpha p[t] + \beta + w[t] \qquad t = 1, 2, \ldots, N \tag{A.2}$$

or in vectorial notation

$$\boldsymbol{y} = \alpha \boldsymbol{p} + \beta \mathbf{1} + \boldsymbol{w} \tag{A.3}$$

where $\mathbf{1}$ is a vector with every element equal to 1 and $\boldsymbol{w}$ is a Gaussian distributed random vector.

Viewing the regression function $\boldsymbol{u}$ as a scaled and shifted template $\boldsymbol{p}$

$$\boldsymbol{u} = \alpha \boldsymbol{p} + \beta \mathbf{1} \tag{A.4}$$

the results from Section 2.1.4 can directly be transferred so that the MLE of α, β and $\boldsymbol{\theta}$ can be computed by the maximisation of $\Lambda(\alpha, \beta, \boldsymbol{\theta}; \boldsymbol{y})$

$$\hat{\theta} = \arg\max_{\alpha,\beta,\theta} \Lambda(\alpha, \beta, \boldsymbol{\theta}; \boldsymbol{y}) \tag{A.5}$$

with

$$\Lambda(\alpha, \beta, \boldsymbol{\theta}; \boldsymbol{y}) = -(\boldsymbol{y} - \boldsymbol{u})^T (\boldsymbol{y} - \boldsymbol{u}) \ . \tag{A.6}$$

It follows then for $\Lambda(\alpha, \beta, \boldsymbol{\theta}; \boldsymbol{y})$

$$\begin{aligned}\Lambda(\alpha, \beta, \boldsymbol{\theta}; \boldsymbol{y}) &= -(\boldsymbol{y} - (\alpha\boldsymbol{p} + \beta\mathbf{1}))^T(\boldsymbol{y} - (\alpha\boldsymbol{p} + \beta\mathbf{1})) \\ &= -\boldsymbol{y}^T\boldsymbol{y} - \alpha^2\boldsymbol{p}^T\boldsymbol{p} - m\beta^2 + 2\alpha\boldsymbol{y}^T\boldsymbol{p} + 2m\beta\bar{\boldsymbol{y}} - 2m\alpha\beta\bar{\boldsymbol{p}} \qquad \text{(A.7)}\end{aligned}$$

with $\bar{\boldsymbol{y}}$ and $\bar{\boldsymbol{p}}$ denoting the mean over the elements of $\boldsymbol{y}$ and $\boldsymbol{p}$ respectively.

The parameters α and β are real valued. The MLE of α and β can therefore be computed by setting the partial derivatives to zero

$$\frac{\partial\Lambda(\alpha, \beta, \boldsymbol{\theta}; \boldsymbol{y})}{\partial\alpha} = 2\left(-\alpha\boldsymbol{p}^T\boldsymbol{p} + \boldsymbol{y}^T\boldsymbol{p} - m\beta\bar{\boldsymbol{p}}\right) \qquad \text{(A.8)}$$

$$\frac{\partial\Lambda(\alpha, \beta, \boldsymbol{\theta}; \boldsymbol{y})}{\partial\beta} = 2m\left(-\beta + \bar{\boldsymbol{y}} - \alpha\bar{\boldsymbol{p}}\right). \qquad \text{(A.9)}$$

Let $\hat{\alpha}$ and $\hat{\beta}$ be the solutions of the system of equations

$$\frac{\partial\Lambda(\alpha, \beta, \boldsymbol{\theta}; \boldsymbol{y})}{\partial\alpha} = 0 \qquad \text{(A.10)}$$

$$\frac{\partial\Lambda(\alpha, \beta, \boldsymbol{\theta}; \boldsymbol{y})}{\partial\beta} = 0 \qquad \text{(A.11)}$$

so that

$$2\left(-\hat{\alpha}\boldsymbol{p}^T\boldsymbol{p} + \boldsymbol{y}^T\boldsymbol{p} - m\hat{\beta}\bar{\boldsymbol{p}}\right) = 0 \qquad \text{(A.12)}$$

$$2m\left(-\hat{\beta} + \bar{\boldsymbol{y}} - \hat{\alpha}\bar{\boldsymbol{p}}\right) = 0 \qquad \text{(A.13)}$$

which is equal to

$$\hat{\alpha} = \frac{\boldsymbol{y}^T\boldsymbol{p}}{\boldsymbol{p}^T\boldsymbol{p}} - m\hat{\beta}\,\frac{\bar{\boldsymbol{p}}}{\boldsymbol{p}^T\boldsymbol{p}} \qquad \text{(A.14)}$$

$$\hat{\beta} = \bar{\boldsymbol{y}} - \hat{\alpha}\bar{\boldsymbol{p}}\,. \qquad \text{(A.15)}$$

This system of equations can be solved in the following way

$$\xrightarrow{\text{(A.15) in (A.14)}} \quad \hat{\alpha} = \frac{\boldsymbol{y}^T\boldsymbol{p}}{\boldsymbol{p}^T\boldsymbol{p}} - m\left(\bar{\boldsymbol{y}} - \hat{\alpha}\bar{\boldsymbol{p}}\right)\frac{\bar{\boldsymbol{p}}}{\boldsymbol{p}^T\boldsymbol{p}}$$

$$\Leftrightarrow \quad \hat{\alpha}\boldsymbol{p}^T\boldsymbol{p} = \boldsymbol{y}^T\boldsymbol{p} - m\bar{\boldsymbol{y}}\bar{\boldsymbol{p}} + \hat{\alpha}m\bar{\boldsymbol{p}}^2$$

$$\Leftrightarrow \quad \hat{\alpha} = \frac{\boldsymbol{y}^T\boldsymbol{p} - m\bar{\boldsymbol{y}}\bar{\boldsymbol{p}}}{\boldsymbol{p}^T\boldsymbol{p} - m\bar{\boldsymbol{p}}^2} \tag{A.16}$$

$$\xrightarrow{\text{(A.16) in (A.15)}} \quad \hat{\beta} = \bar{\boldsymbol{y}} - \frac{\boldsymbol{y}^T\boldsymbol{p} - m\bar{\boldsymbol{y}}\bar{\boldsymbol{p}}}{\boldsymbol{p}^T\boldsymbol{p} - m\bar{\boldsymbol{p}}^2}\,\bar{\boldsymbol{p}}$$

$$\Leftrightarrow \quad \hat{\beta} = \frac{\bar{\boldsymbol{y}}\boldsymbol{p}^T\boldsymbol{p} - \bar{\boldsymbol{y}}m\bar{\boldsymbol{p}}^2 - \boldsymbol{y}^T\boldsymbol{p}\bar{\boldsymbol{p}} + m\bar{\boldsymbol{y}}\bar{\boldsymbol{p}}^2}{\boldsymbol{p}^T\boldsymbol{p} - m\bar{\boldsymbol{p}}^2}$$

$$\hat{\beta} = \frac{\bar{\boldsymbol{y}}\boldsymbol{p}^T\boldsymbol{p} - \bar{\boldsymbol{p}}\boldsymbol{y}^T\boldsymbol{p}}{\boldsymbol{p}^T\boldsymbol{p} - m\bar{\boldsymbol{p}}^2}\,. \tag{A.17}$$

Note that the same results are obtained when viewing the problem as a linear regression problem of $\boldsymbol{y} = \alpha\boldsymbol{p} + \beta\mathbf{1}$ [Kundu and Ubhaya, 2001].

Now, the maximisation of the likelihood is simplified by substituting α and β by their estimates in (A.4).

$$\boldsymbol{u} = \hat{\alpha}\boldsymbol{p} + \hat{\beta}\mathbf{1} \tag{A.18}$$

$$= \frac{\boldsymbol{y}^T\boldsymbol{p} - m\bar{\boldsymbol{y}}\bar{\boldsymbol{p}}}{\boldsymbol{p}^T\boldsymbol{p} - m\bar{\boldsymbol{p}}^2}\,\boldsymbol{p} + \frac{\bar{\boldsymbol{y}}\boldsymbol{p}^T\boldsymbol{p} - \bar{\boldsymbol{p}}\boldsymbol{y}^T\boldsymbol{p}}{\boldsymbol{p}^T\boldsymbol{p} - m\bar{\boldsymbol{p}}^2}\mathbf{1} \tag{A.19}$$

After factorisation, the terms $-m\bar{\boldsymbol{y}}\bar{\boldsymbol{p}}^2\mathbf{1} + m\bar{\boldsymbol{y}}\bar{\boldsymbol{p}}^2\mathbf{1}$ are added to the numerator. Resorting leads then to the equations

$$\boldsymbol{u} = \frac{\boldsymbol{y}^T\boldsymbol{p}\left(\boldsymbol{p} - \bar{\boldsymbol{p}}\mathbf{1}\right) - m\bar{\boldsymbol{y}}\bar{\boldsymbol{p}}\left(\boldsymbol{p} - \bar{\boldsymbol{p}}\mathbf{1}\right) + \bar{\boldsymbol{y}}\left(\boldsymbol{p}^T\boldsymbol{p} - m\bar{\boldsymbol{p}}^2\right)\mathbf{1}}{\boldsymbol{p}^T\boldsymbol{p} - m\bar{\boldsymbol{p}}^2} \tag{A.20}$$

$$= \frac{\left(\boldsymbol{y}^T\boldsymbol{p} - m\bar{\boldsymbol{y}}\bar{\boldsymbol{p}}\right)\left(\boldsymbol{p} - \bar{\boldsymbol{p}}\mathbf{1}\right) + \bar{\boldsymbol{y}}\left(\boldsymbol{p}^T\boldsymbol{p} - m\bar{\boldsymbol{p}}^2\right)\mathbf{1}}{\boldsymbol{p}^T\boldsymbol{p} - m\bar{\boldsymbol{p}}^2} \tag{A.21}$$

$$= \frac{\left(\boldsymbol{y}^T\boldsymbol{p} - m\bar{\boldsymbol{y}}\bar{\boldsymbol{p}}\right)\left(\boldsymbol{p} - \bar{\boldsymbol{p}}\mathbf{1}\right)}{\boldsymbol{p}^T\boldsymbol{p} - m\bar{\boldsymbol{p}}^2} + \bar{\boldsymbol{y}}\mathbf{1} \tag{A.22}$$

$$= \frac{\left(\boldsymbol{y} - \bar{\boldsymbol{y}}\mathbf{1}\right)^T\boldsymbol{p}\left(\boldsymbol{p} - \bar{\boldsymbol{p}}\mathbf{1}\right)}{\boldsymbol{p}^T\boldsymbol{p} - m\bar{\boldsymbol{p}}^2} + \bar{\boldsymbol{y}}\mathbf{1}\,. \tag{A.23}$$

Without loss of generality one can assume that $\boldsymbol{p}$ has zero mean and unit norm,

i. e. that the conditions

$$\bar{\boldsymbol{p}} = 0 \tag{A.24}$$

$$\|\boldsymbol{p}\| = 1 \tag{A.25}$$

hold, which are termed the *regularity conditions* of the templates. A prove that these conditions do not effect the generality is given next.

An arbitrary template $\boldsymbol{p}'$ can be constructed from a template $\boldsymbol{p}$ that fulfils the regularity conditions by multiplying and adding constants to $\boldsymbol{p}$, i. e.

$$\boldsymbol{p}' = \phi\boldsymbol{p} + \psi\mathbf{1} \; . \tag{A.26}$$

With this definition the mean and the norm of $\boldsymbol{p}'$ is equal to

$$\begin{aligned} \bar{\boldsymbol{p}}' &= \frac{1}{m}\phi\bar{\boldsymbol{p}}^T\mathbf{1} + \frac{1}{m}\psi\mathbf{1}^T\mathbf{1} \\ &= \phi\bar{\boldsymbol{p}} + \psi \end{aligned} \tag{A.27}$$

and

$$\begin{aligned} \boldsymbol{p}'^T\boldsymbol{p}' &= (\phi\boldsymbol{p} + \psi\mathbf{1})^T(\phi\boldsymbol{p} + \psi\mathbf{1}) \\ &= \phi^2\boldsymbol{p}^T\boldsymbol{p} + 2\phi\psi\boldsymbol{p}^T\mathbf{1} + \psi^2\mathbf{1}^T\mathbf{1} \\ &= \phi^2\boldsymbol{p}^T\boldsymbol{p} + 2\phi\psi m\bar{\boldsymbol{p}} + m\psi^2 \; . \end{aligned} \tag{A.28}$$

Replacing $\boldsymbol{p}$ with $\boldsymbol{p}'$ in (A.23) leads to

$$\boldsymbol{u} = \frac{(\boldsymbol{y} - \bar{\boldsymbol{y}}\mathbf{1})^T\boldsymbol{p}'\,(\boldsymbol{p}' - \bar{\boldsymbol{p}}'\mathbf{1})}{\boldsymbol{p}'^T\boldsymbol{p}' - m\bar{\boldsymbol{p}}'^2} + \bar{\boldsymbol{y}}\mathbf{1} \tag{A.29}$$

so that with (A.27) and (A.28)

$$\boldsymbol{u} = \frac{(\boldsymbol{y} - \bar{\boldsymbol{y}}\mathbf{1})^T(\phi\boldsymbol{p} + \psi\mathbf{1})\,(\phi\boldsymbol{p} + \psi\mathbf{1} - (\phi\bar{\boldsymbol{p}} + \psi)\mathbf{1})}{\phi^2\boldsymbol{p}^T\boldsymbol{p} + 2\phi\psi m\bar{\boldsymbol{p}} + m\psi^2 - m(\phi\bar{\boldsymbol{p}} + \psi)^2} + \bar{\boldsymbol{y}}\mathbf{1} \tag{A.30}$$

$$= \frac{(\phi\boldsymbol{y}^T\boldsymbol{p} + \psi\boldsymbol{y}^T\mathbf{1} - \phi\bar{\boldsymbol{y}}\boldsymbol{p}^T\mathbf{1} - \psi\bar{\boldsymbol{y}}\mathbf{1}^T\mathbf{1})\,(\phi\boldsymbol{p} + \psi\mathbf{1} - \phi\bar{\boldsymbol{p}}\mathbf{1} - \psi\mathbf{1})}{\phi^2\boldsymbol{p}^T\boldsymbol{p} + 2\phi\psi m\bar{\boldsymbol{p}} + m\psi^2 - m\phi^2\bar{\boldsymbol{p}}^2 - 2m\phi\psi\bar{\boldsymbol{p}} - m\psi^2} + \bar{\boldsymbol{y}}\mathbf{1} \tag{A.31}$$

$$= \frac{(\phi\boldsymbol{y}^T\boldsymbol{p} + \psi m\bar{\boldsymbol{y}} - \phi\bar{\boldsymbol{y}}\boldsymbol{p}^T\mathbf{1} - \psi m\bar{\boldsymbol{y}})\,(\phi\boldsymbol{p} - \phi\bar{\boldsymbol{p}}\mathbf{1})}{\phi^2\boldsymbol{p}^T\boldsymbol{p} - m\phi^2\bar{\boldsymbol{p}}^2} + \bar{\boldsymbol{y}}\mathbf{1} \tag{A.32}$$

$$= \frac{(\phi\boldsymbol{y}^T\boldsymbol{p} - \phi\bar{\boldsymbol{y}}\boldsymbol{p}^T\mathbf{1})\,(\phi\boldsymbol{p} - \phi\bar{\boldsymbol{p}}\mathbf{1})}{\phi^2\boldsymbol{p}^T\boldsymbol{p} - m\phi^2\bar{\boldsymbol{p}}^2} + \bar{\boldsymbol{y}}\mathbf{1} \tag{A.33}$$

$$= \frac{(\boldsymbol{y} - \bar{\boldsymbol{y}}\mathbf{1})^T\boldsymbol{p}\,(\boldsymbol{p} - \bar{\boldsymbol{p}}\mathbf{1})}{\boldsymbol{p}^T\boldsymbol{p} - m\bar{\boldsymbol{p}}^2} + \bar{\boldsymbol{y}}\mathbf{1} \tag{A.34}$$

which is identical to (A.23). Since the mapping (A.26) does not change $\boldsymbol{u}$ one can assume that the templates fulfil the regularity conditions without loss of generality.

Incorporating the regularity conditions in (A.16) and (A.17) simplifies the estimates of the magnitude and offset

$$\hat{\alpha} = \boldsymbol{y}^T\boldsymbol{p} \tag{A.35}$$

$$\hat{\beta} = \bar{\boldsymbol{y}} \tag{A.36}$$

as well as $\boldsymbol{u}$ when incorporating them in (A.23) which leads to

$$\boldsymbol{u} = (\boldsymbol{y} - \bar{\boldsymbol{y}}\mathbf{1})^T\boldsymbol{p}\,\boldsymbol{p} + \bar{\boldsymbol{y}}\mathbf{1}\ . \tag{A.37}$$

The final step is to replace $\boldsymbol{u}$ by (A.37) in (A.6) which is the definition of Λ .

$$
\begin{aligned}
\Lambda(\hat{\alpha}, \hat{\beta}, \boldsymbol{\theta}; \boldsymbol{y}) &= -(\boldsymbol{y} - \boldsymbol{u})^{T} (\boldsymbol{y} - \boldsymbol{u}) \\
&= -(\boldsymbol{y} - \boldsymbol{u})^{2} \\
&= -\left(\boldsymbol{y} - \left((\boldsymbol{y} - \bar{\boldsymbol{y}}\mathbf{1})^{T}\boldsymbol{p}\,\boldsymbol{p} + \bar{\boldsymbol{y}}\mathbf{1}\right)\right)^{2} && \text{(A.38)} \\
&= -\left(\boldsymbol{y} - (\boldsymbol{y} - \bar{\boldsymbol{y}}\mathbf{1})^{T}\boldsymbol{p}\,\boldsymbol{p} - \bar{\boldsymbol{y}}\mathbf{1}\right)^{2} && \text{(A.39)} \\
&= -\left((\boldsymbol{y} - \bar{\boldsymbol{y}}\mathbf{1}) - (\boldsymbol{y} - \bar{\boldsymbol{y}}\mathbf{1})^{T}\boldsymbol{p}\,\boldsymbol{p}\right)^{2} && \text{(A.40)} \\
&= -(\boldsymbol{y} - \bar{\boldsymbol{y}}\mathbf{1})^{T}(\boldsymbol{y} - \bar{\boldsymbol{y}}\mathbf{1}) + 2(\boldsymbol{y} - \bar{\boldsymbol{y}}\mathbf{1})^{T}\boldsymbol{p}\ (\boldsymbol{y} - \bar{\boldsymbol{y}}\mathbf{1})^{T}\boldsymbol{p} \\
&\quad - \left((\boldsymbol{y} - \bar{\boldsymbol{y}}\mathbf{1})^{T}\boldsymbol{p}\right)^{2}\boldsymbol{p}^{T}\boldsymbol{p} && \text{(A.41)} \\
&= -(\boldsymbol{y} - \bar{\boldsymbol{y}}\mathbf{1})^{T}(\boldsymbol{y} - \bar{\boldsymbol{y}}\mathbf{1}) + 2\left((\boldsymbol{y} - \bar{\boldsymbol{y}}\mathbf{1})^{T}\boldsymbol{p}\right)^{2} \\
&\quad - \left((\boldsymbol{y} - \bar{\boldsymbol{y}}\mathbf{1})^{T}\boldsymbol{p}\right)^{2} && \text{(A.42)} \\
&= -(\boldsymbol{y} - \bar{\boldsymbol{y}}\mathbf{1})^{T}(\boldsymbol{y} - \bar{\boldsymbol{y}}\mathbf{1}) + \left((\boldsymbol{y} - \bar{\boldsymbol{y}}\mathbf{1})^{T}\boldsymbol{p}\right)^{2} && \text{(A.43)} \\
&= -(\boldsymbol{y} - \bar{\boldsymbol{y}}\mathbf{1})^{T}(\boldsymbol{y} - \bar{\boldsymbol{y}}\mathbf{1}) + \left(\boldsymbol{y}^{T}\boldsymbol{p} - m\,\bar{\boldsymbol{y}}\bar{\boldsymbol{p}}\right)^{2} && \text{(A.44)} \\
&= -(\boldsymbol{y} - \bar{\boldsymbol{y}}\mathbf{1})^{T}(\boldsymbol{y} - \bar{\boldsymbol{y}}\mathbf{1}) + \left(\boldsymbol{y}^{T}\boldsymbol{p}\right)^{2} && \text{(A.45)}
\end{aligned}
$$

Since the first term is independent from $\boldsymbol{\theta}$, it can be cancelled when estimating $\boldsymbol{\theta}$ so that the scale and shift invariant ML estimate of $\boldsymbol{\theta}$ is equal to

$$
\begin{aligned}
\hat{\boldsymbol{\theta}} &= \arg\max_{\boldsymbol{\theta}} \Lambda(\hat{\alpha}, \hat{\beta}, \boldsymbol{\theta}; \boldsymbol{y}) \\
&= \arg\max_{\boldsymbol{\theta}} \left(\boldsymbol{y}^{T}\boldsymbol{p}\right)^{2} && \text{(A.46)} \\
&= \arg\max_{\boldsymbol{\theta}} \left|\boldsymbol{y}^{T}\boldsymbol{p}\right| \ . && \text{(A.47)}
\end{aligned}
$$

Note that this equation does only hold when $\boldsymbol{p}$ has unit norm and zero mean which does not effect the generality as stated above.

B. The Test Statistic if the Signal is a Ramp-Step

In this appendix closed form expressions are given for the test statistic $V(\nu, n)$ in the deterministic case where noise is supposed to be zero. As explained in Section 6.2, $V(\nu, n)$ is piecewise defined on six regions termed A-F. The derived formulas prove that $V(\nu, n)$ is strictly increasing with n in all regions, strictly increasing with ν in region D and strictly decreasing with ν in region F; so that the global maximum is in region E.

Region A: $1 \leq n \leq k_0$ **and** $1 \leq \nu \leq k_0$

This case is trivial. Since $y[i] = d$ the mean values $\hat{\mu}$, $\hat{\mu}_1$ and $\hat{\mu}_2$ are identical and therefore $V(\nu, n)$ is equal to zero for each pair (ν, n).

$$V_A(\nu, n) = 0 \tag{B.1}$$

Region B: $k_0 < n \leq k_0 + \tau_0$ **and** $1 \leq \nu \leq k_0$

In this case ν is smaller than k_0 and n is greater than k_0 so that it follows for the mean values

$$\begin{aligned}
\hat{\mu}_1 &= \frac{1}{\nu} \sum_{j=1}^{\nu} d_0 \\
\hat{\mu}_2 &= \frac{1}{n-\nu} \left(\sum_{j=\nu+1}^{k_0} d_0 + \sum_{j=k_0+1}^{n} \left(\frac{h_0}{\tau_0}(j-k_0) + d_0 \right) \right) \\
\hat{\mu} &= \frac{1}{n} \left(\sum_{j=1}^{k_0} d_0 + \sum_{j=k_0+1}^{n} \left(\frac{h_0}{\tau_0}(j-k_0) + d_0 \right) \right) .
\end{aligned}$$

The test statistic $V(\nu, n)$ is then

$$V_B(\nu, n) = \frac{h_0^2}{4} \frac{\nu(n-k_0)^2(n-k_0+1)^2}{n\tau_0^2(n-\nu)} . \tag{B.2}$$

Substituting $\varphi = \nu - k_0$ and $\psi = n - k_0$ the partial derivatives $\frac{\partial V_B(\nu,n)}{\partial n}$ and $\frac{\partial V_B(\nu,n)}{\partial \nu}$ are

$$\begin{aligned}
\frac{\partial V_B(\nu, n)}{\partial n} &= \frac{h_0^2}{4} \frac{\psi^2(1+\psi)^2}{\tau_0^2(\varphi+\psi)^2} \\
\frac{\partial V_B(\nu, n)}{\partial \nu} &= \frac{h_0^2}{4} \frac{1}{\tau_0^2(\varphi+\psi)^2(\varphi+\psi+\nu)^2} \big(\nu\psi(1+\psi)(2(\varphi+\psi) \\
&\qquad (\varphi + 2\varphi\psi + \psi^2) + (2\varphi + \psi + 4\varphi\psi + 3\psi^2)\nu) \big) .
\end{aligned}$$

Since φ and ψ are positive by definition the partial derivatives contain only positive summands and factors and are therefore positive, too. As a consequence, $V_B(\nu, n)$ is a strictly increasing function.

Region C: $k_0 < n \leq k_0 + \tau_0$ **and** $k_0 < \nu \leq k_0 + \tau_0$

The mean values in this case are

$$\begin{aligned}
\hat{\mu}_1 &= \frac{1}{\nu}\left(\sum_{j=1}^{k_0} d_0 + \sum_{j=k_0+1}^{\nu}\left(\frac{h_0}{\tau_0}(j-k_0)+d_0\right)\right) \\
\hat{\mu}_2 &= \frac{1}{n-\nu}\sum_{j=\nu+1}^{n}\left(\frac{h_0}{\tau_0}(j-k_0)+d_0\right) \\
\hat{\mu} &= \frac{1}{n}\left(\sum_{j=1}^{k_0} d_0 + \sum_{j=k_0+1}^{n}\left(\frac{h_0}{\tau_0}(j-k_0)+d_0\right)\right) \ .
\end{aligned}$$

The function $V(\nu, n)$ is then equal to

$$V_C(\nu, n) = \frac{h_0^2}{4}\,\frac{(n-\nu)(\nu n - k_0^2 + k_0)^2}{\nu \tau_0^2 n} \ . \tag{B.3}$$

The partial derivative $\frac{\partial V_C(\nu,n)}{\partial n}$ is equal to

$$\begin{aligned}
\frac{\partial V_C(\nu, n)}{\partial n} = \frac{h_0^2}{4}\,\frac{1}{\tau_0^2(\varphi+\psi+k_0)^2}\big(&\varphi^2 + 3\varphi\psi + 2\psi^2 \\
&+ (1+2\varphi+3\psi)k_0)(\varphi(\varphi+\psi) + (1+2\varphi+\psi)k_0)\big)
\end{aligned} \tag{B.4}$$

where the substitutions $\varphi = \nu - k_0$ and $\psi = n - \nu$ are used. Both are again positive and for the same reason as in case B, $V_C(\nu, n)$ is strictly increasing along n. Along k there is a local maximum.

Region D: $k_0 + \tau_0 < n \leq N$ **and** $1 \leq \nu \leq k_0$

The mean values in this case are

$$\begin{aligned}
\hat{\mu}_1 &= \frac{1}{\nu}\sum_{j=1}^{\nu} d_0 \\
\hat{\mu}_2 &= \frac{1}{n-\nu}\left(\sum_{j=\nu+1}^{k_0} d_0 + \sum_{j=k_0+1}^{k_0+\tau_0}\left(\frac{h_0}{\tau_0}(j-k_0)+d_0\right)\right. \\
&\quad \left. + \sum_{j=k_0+\tau_0+1}^{n}(d_0+h_0)\right) \\
\hat{\mu} &= \frac{1}{n}\left(\sum_{j=1}^{k_0} d_0 + \sum_{j=k_0+1}^{k_0+\tau_0}\left(\frac{h_0}{\tau_0}(j-k_0)+d_0\right)\right. \\
&\quad \left. + \sum_{j=k_0+\tau_0+1}^{n}(d_0+h_0)\right).
\end{aligned}$$

The function $V(\nu, n)$ is then equal to

$$V_D(\nu, n) = \frac{h_0^2}{4}\,\frac{\nu(2n-2k_0-\tau_0+1)^2}{n(n-\nu)}\,. \tag{B.5}$$

The partial derivative

$$\frac{\partial V_D(\nu,n)}{\partial \nu} = \frac{h_0^2}{4}\,\frac{(2n-2k_0-\tau_0+1)^2}{(n-\nu)^2}$$

is positive so that $V_D(\nu, n)$ is strictly increasing with ν. The proof that $V_D(\nu, n)$ is strictly increasing with n is more sophisticated. $V_D(\nu, n)$ is a rational function of the form

$$V_D(\nu, n) = \alpha\frac{(n-\beta)^2}{n(n-\nu)}$$

with α and β

$$\begin{aligned}
\alpha &= \frac{h_0^2\nu}{4} \\
\beta &= k_0 + \frac{\tau_0 - 1}{2}
\end{aligned}$$

not depending on n. $V_D(\nu, n)$ is not defined at $n = 0$ and $n = \nu$ and since ν is a positive integer, $V_D(\nu, n)$ is not negative for $n \geq \nu$. If $\beta \geq \nu$ the function $V_D(\nu, n)$ is strictly increasing for $n > \beta$ with the limit α as n approaches infinity. So if it can be proven, that $\beta \geq \nu$ and $\beta \leq k_0 + \tau_0$ then it is proven that $V_D(\nu, n)$ is strictly increasing on $[k_0 + \tau_0, \infty[$.

Since $\nu \leq k_0$ and $\beta \geq k_0$ (for $\tau_0 = 1$) and $\beta \leq k_0 + \tau_0$ by definition, this proof is trivial

Region E: $k_0 + \tau_0 < n \leq N$ **and** $k_0 < \nu \leq k_0 + \tau_0$

The mean values in this case are

$$\begin{aligned}
\hat{\mu}_1 &= \frac{1}{\nu}\left(\sum_{j=1}^{k_0} d_0 + \sum_{j=k_0+1}^{\nu}\left(\frac{h_0}{\tau_0}(j-k_0)+d_0\right)\right) \\
\hat{\mu}_2 &= \frac{1}{n-\nu}\left(\sum_{j=\nu+1}^{k_0+\tau_0}\left(\frac{h_0}{\tau_0}(j-k_0)+d_0\right) + \sum_{j=k_0+\tau_0+1}^{n}(d_0+h_0)\right) \\
\hat{\mu} &= \frac{1}{n}\left(\sum_{j=1}^{k_0} d_0 + \sum_{j=k_0+1}^{k_0+\tau_0}\left(\frac{h_0}{\tau_0}(j-k_0)+d_0\right) + \sum_{j=k_0+\tau_0+1}^{n}(d_0+h_0)\right) .
\end{aligned}$$

The function $V(\nu, n)$ is then equal to

$$V_E(\nu, n) = \frac{h_0^2}{4}\,\frac{(n(\nu-k_0+1)(\nu-k_0)-(2n-2k_0+1)\nu\tau_0+\nu\tau_0^2)^2}{n(n-\nu)\nu\tau_0^2} . \qquad \text{(B.6)}$$

The proof that $V_E(\nu, n)$ is strictly increasing with n will be done in the same manner as in Case D. $V_E(\nu, n)$ is a rational function of the form as $V_D(\nu, n)$

$$V_E(\nu, n) = \alpha\frac{(n-\beta)^2}{n(n-\nu)}$$

with

$$\begin{aligned}
\alpha &= \frac{h_0^2}{4\nu\tau_0^2} \\
\beta &= \frac{-\tau_0(2k_0+\tau_0-1)\nu}{\nu^2-(2(k_0+\tau_0)-1)\nu+k_0(k_0-1)} .
\end{aligned}$$

B. The Test Statistic if the Signal is a Ramp-Step

Since α is not negative the proof the $V_E(\nu, n)$ is strictly increasing with n follows directly, as in Case D, from the proof that $\beta \in [\nu, k_0 + \tau_0]$.

This time β depends on ν and it is a rational function in ν of the form

$$\beta = \frac{-\tau_0(2k_0 + \tau_0 - 1)\nu}{(\nu - \nu_1)(\nu - \nu_2)}$$

which is not defined for

$$\nu_{1,2} = k_0 + \tau_0 - 0.5 \pm \sqrt{4\tau_0^2 + 8k_0\tau_0 + \tau_0 - 4\tau_0 + 1}$$

Since k_0 and τ_0 are positive integers, the value of the square root is greater than $2\tau_0$ with the consequence that ν_1 is at the right and ν_2 at the left side of the domain $\nu \in [k_0, k_0 + \tau_0]$. So, if the range $]\nu_1, \nu[$ is denoted by D_ν and the domain is denoted by D, this means that D_ν is a subset of D.

From simple calculus it follows that β has one local minimum on D at $\nu_{min} = \sqrt{k_0^2 - k_0}$ which is at the left of D_ν so that β is increasing on D_ν with the consequence that the maximum value of β is at $\nu = k_0 + \tau_0$

$$\max_\nu \beta = k_0 + \tau_0 \quad \text{for} \quad \nu \in D_\nu$$

which proves the first condition that $\beta \leq k_0 + \tau_0$. The second condition that $\beta \geq \nu$ is proven by evaluating the function

$$\begin{aligned} \beta^* &= \beta - \nu \\ &= \frac{-\tau_0(2k_0 + \tau_0 - 1)\nu - (\nu - \nu_1)(\nu - \nu_2)\nu}{(\nu - \nu_1)(\nu - \nu_2)} \\ &= \frac{-\nu(\nu - (k_0 + \tau_0 - 1))(\nu - (k_0 + \tau_0))}{(\nu - \nu_1)(\nu - \nu_2)} \,. \end{aligned}$$

It has the same domain as β and is not negative on D_ν so that it can be concluded that $\beta \geq \nu$ with the consequence that and $V_E(\nu, n)$ is strictly increasing with n.

Region F: $k_0 + \tau_0 < n \leq N$ **and** $k_0 + \tau_0 < \nu \leq N$

The mean values in this case are

$$\begin{aligned}
\hat{\mu}_1 &= \frac{1}{\nu}\left(\sum_{j=1}^{k_0} d_0 + \sum_{j=k_0+1}^{k_0+\tau_0}\left(\frac{h_0}{\tau_0}(j-k_0)+d_0\right) + \sum_{j=k_0+\tau_0+1}^{\nu}(d_0+h_0)\right) \\
\hat{\mu}_2 &= \frac{1}{n-\nu}\sum_{j=\nu+1}^{n}(d_0+h_0) \\
\hat{\mu} &= \frac{1}{n}\left(\sum_{j=1}^{k_0} d_0 + \sum_{j=k_0+1}^{k_0+\tau_0}\left(\frac{h_0}{\tau_0}(j-k_0)+d_0\right) + \sum_{j=k_0+\tau_0+1}^{n}(d_0+h_0)\right) .
\end{aligned}$$

The function $V(\nu, n)$ is then equal to

$$V_F(\nu, n) = \frac{h_0^2}{4}\,\frac{(n-\nu)(2k_0+\tau_0-1)^2}{n\nu} . \tag{B.7}$$

Since the partial derivative along n

$$\frac{\partial V_F(\nu,n)}{\partial n} = \frac{h_0^2}{4}\,\frac{(n-\nu)(2k_0+\tau_0-1)^2}{n^2}$$

is positive, $V_F(\nu, n)$ is strictly increasing with n. In contrast, the partial derivative

$$\frac{\partial V_F(\nu,n)}{\partial \nu} = -\frac{h_0^2}{4}\,\frac{(n-\nu)(2k_0+\tau_0-1)^2}{\nu^2}$$

is negative so that $V_F(\nu, n)$ is strictly decreasing with ν.

Glossary of Symbols and Abbreviations

Symbols

(Boldface characters denote vectors. All others are scalars.)

$\hat{}$	denotes an estimate
$\bar{}$	denotes the sample mean
$\lvert \cdot \rvert$	denotes the absolute value of a scalar
$\lVert \cdot \rVert$	denotes the norm of a vector
$_0$	(0 subscript) denotes the true value
^{T}	(T superscript) denotes the transpose of a vector
a	a ramp-step is defined on $[a, b]$
b	a ramp-step is defined on $[a, b]$
c	number of classes, i. e., hypotheses in a multiple hypotheses test
d	offset of the ramp-step
δ	threshold
k	change-point of the ramp-step
K	the number of change-points in a signal
E	expectation
h	change magnitude of the ramp-step
$H(\boldsymbol{\theta}; \boldsymbol{y})$	the Hessian of the log-likelihood function

$\mathcal{H}_0$	noise only hypothesis or null hypothesis
$\mathcal{H}_1$	noise + signal hypothesis or alternate hypothesis
$\mathcal{H}_i$	i-th hypothesis in multiple hypotheses tests
I	Fisher information
L	likelihood function
$\ln \mathrm{L}$	log-likelihood function
m	number of observations for the local ramp-step fit
	$(m = b - a + 1)$
μ	mean
n	current time in an on-line setup
ν	time of an abrupt change when testing for a change in mean
N	number of observations
$\mathcal{N}(\mu, \sigma^2)$	normal distribution with mean μ and variance σ^2
Ω_i	solid angle of R_i
P	set of templates in a multiple hypotheses test
$p(\boldsymbol{y}; \theta)$	probability density function of $\boldsymbol{y}$ with θ as parameter
$P(\mathcal{H}_i \vert \mathcal{H}_j)$	probability of deciding $\mathcal{H}_i$ when $\mathcal{H}_j$ is true
P_D	probability of detection $(P(\mathcal{H}_1 \vert \mathcal{H}_1))$
P_e	probability of error in a multiple hypotheses test
P_FA	probability of false alarm $(P(\mathcal{H}_1 \vert \mathcal{H}_0))$
$\#P_N$	number of segmentations of a signal with size N
$\#P_{N,K}$	number of segmentations of a signal with size N and K change-points
R_i	the decision region for hypothesis $\mathcal{H}_i$
t	sequence index

τ	rise time of the ramp-step
S_m	surface area of the m-dimensional unit sphere
σ^2	variance
$\mathrm{T_{LR}}$	likelihood ratio function
$\mathrm{T_{GLR}}$	generalised likelihood ratio function
$\theta(\boldsymbol{\theta})$	unknown parameter (vector)
$u[t]$	time varying mean at time t
$w[t]$	observation noise at time t
$y[t]$	observation at time t
$\boldsymbol{y}$	vector of observations $(y[1], y[2], ..., y[N])^T$

Abbreviations

ADC	analogue-to-digital converter
CUSUM	cumulative sum chart
DC	direct current
GLRT	generalised likelihood ratio test
ICU	intensive care unit
LS	least squares
LRT	likelihood ratio test
ML	maximum likelihood
MLE	maximum likelihood estimate
MSE	mean squared error
MVU	minimum variance unbiased
OPEC	organisation of the petroleum exporting countries
PDF	probability density function

ROC	receiver operating characteristic
SEMUG	sequential detection of multiple gradual changes
WGN	white Gaussian noise

Bibliography

J. Aldrich. R. A. Fischer and the making of maximum likelihood 1912–1922. *Statistical Science*, 12(3):162–176, 1997.

R. Avent and J. Charlton. A critical review of trend-detection methodologies for biomedical systems. *Critical Reviews in Biomedical Engineering*, 17:621–659, 1990.

D. Avis and K. Fukuda. A pivoting algorithm for convex hull and vertex enumeration of arrangements and polyhedra. *Discrete & Computational Geometry*, 8(3): 295–313, 1992.

J. Bai and P. Perron. Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, 18:1–22, Oct 2003.

M. Basseville and I. Nikiforov. *Detection of abrupt changes: Theory and application.* Prentice-Hall, Inc., 1993.

A.-L. Beem. A program for fitting two-phase segmented-curve models with an unknown change point, with an application to the analysis of strategy shifts in a cognitive task. *Behavior Research Methods, Instruments, & Computers*, 27(3): 392–399, 1995.

S. Björklund and L. Ljung. A review of time-delay estimation techniques. In *Proceedings - IEEE Conference on Decision and Control*, pages 2502–2507, Maui, Hawaii USA, December 2003.

R. L. Buchanan, R. C. Whiting, and W. C. Damert. When is simple good enough:

a comparison of the Gompertz, Baranyi, and three-phase linear models for fitting bacterial growth curves,. *Food Microbiology*, 14(4):313–326, Aug. 1997.

S. Charbonnier, G. Becq, and L. Biot. On-line segmentation algorithm for continuously monitored data in intensive care units. *IEEE Transactions on Biomedical Engineering*, 51(3):484–492, 2004.

J. Chen and A. Gupta. On change point detection and estimation. *Communications in Statistics - Simulation and Computation*, 30(3):665–697, 2001.

G. Chiu. *Bent-cable regression for assessing abruptness of change*. PhD thesis, Simon Fraser University, 2002.

D. Cong Khac, G. Staude, and W. Wolf. Motor timing and more – additional options using advanced registration and evaluation of tapping data. *Biomedizinische Technik / Biomedical Engineering*, 52(1):156–163, 2007.

H. Cramér. *Mathematical Methods of Statistics*. Princeton University Press, 1946.

G. B. Dantzig. *Linear Programming and Extensions*. Princeton University Press, 1963.

F. Desobry and M. Davy. Support vector-based online detection of abrupt changes. *Proceedings - ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*, 4:872–875, 2003.

V. Dragalin. The sequential change point problem. *Economic Quality Control*, 12 (2):95–122, 1997.

J.-F. Ducré-Robitaille, L. A. Vincent, and G. Boulet. Comparison of techniques for detection of discontinuities in temperature series. *International Journal of Climatology*, 23(9):1087–1101, 2003.

F. Eriksson. On the measure of solid angles. *Mathematics Magazine*, 63(3):184–187, 1990.

L. Euler. De mensura angulorum solidorum. *Opera omnia*, 26:204–223. (Orig. in *Acta acad. sc. Petrop.* 1778).

T. Friede, F. Miller, W. Bischoff, and M. Kieser. A note on change point estimation in dose-response trials. *Computational Statistics & Data Analysis*, 37(2):219–232, Aug. 2001.

W. E. Garthright. The three-phase linear model of bacterial growth: a response. *Food Microbiology*, 14(4):395–397, Aug. 1997.

M. Hajja and P. Walker. The measure of solid angles in n-dimensional Euclidean space. *International Journal of Mathematical Education in Science & Technology*, 33(5):725–800, 2002.

C. B. Hall, J. Ying, L. Kuo, and R. B. Lipton. Bayesian and profile likelihood change point methods for modeling cognitive function over time. *Computational Statistics & Data Analysis*, 42:91–109, 2003.

D. Han and F. Tsung. Comparison of the CUSCORE, GLRT and CUSUM control charts for detectin a dynamic mean change. *Annals of the Insitute of Statistical Mathematics*, 57(3):531–552, 2005.

D. Hawkins. Fitting multiple change-point models to data. *Computational Statistics & Data Analysis*, 37:323–341, 2001.

D. V. Hinkley. Inference about the intersection in two-phase regression. *Biometrika*, 56:495–504, 1969.

H. Hofer, G. Staude, and W. Wolf. Change-point estimation by template matching – some basic aspects. *Informatik, Biometrie and Epidemiologie in Med. u. Biol.*, 35, 2004a.

H. Hofer, G. Staude, T. Zaiser, and W. Wolf. Change point estimation with fuzzy pattern. In *Analysis of biomedical signals and images*, pages 78–80, Brno, Czech Republic, 2004b.

H. Hofer, G. Staude, and W. Wolf. Change detection in continuous tapping data. In *IFMBE Proceedings 2005*, volume 11, 2005. ISSN 1727-1983.

H. Hofer, G. Staude, and W. Wolf. Sequential segmentation of time series containing gradual changes in mean: Fitting multiple ramp-step templates. In *ESGCO Proceedings*, pages 92–95, 2006a.

H. Hofer, G. Staude, and W. Wolf. A method for locating gradual changes in time series. In *WCB*, 2006b.

H. Hofer, G. Staude, and W. Wolf. A method for locating gradual changes in time series. *Biomedizinische Technik / Biomedical Engineering*, 52(1):137–142, 2007.

D. J. Hudson. Fitting segmented curves whose join points have to be estimated. *Journal of the American Statistical Association*, pages 1097–1129, 1966.

S. M. Kay. *Fundamentals of statistical signal processing, Volume I: Estimation theory*. Prentice Hall PTR, 1993.

S. M. Kay. *Fundamentals of statistical signal processing, Volume II: Detection theory*. Prentice Hall PTR, 1998.

N. Kollerstrom. Thomas Simpson and "Newton's method of approximation": an enduring myth. *British Journal for the History of Science*, 25:347–354, 1992.

S. Kundu and V. A. Ubhaya. Fitting a least squares piecewise linear continuous curve in two dimensions. *Computers & Mathematics with Applications*, 41(7-8): 1033–1041, Apr. 2001.

J. L. Lagrange. Solutions de quelques problèmes relatifs au triangles sphériques. *Oeuvres*, 7:331–359. (Orig. in *J. de l'Ecole Polyt.* 1798).

K. Levenberg. A method for the solution of certain nonlinear problems in least squares. *Quarterly of Applied Mathematics*, 2:164–168, 1944.

L. Ljung. Prediction error estimation methods. *Circuits, Systems, and Signal Processing*, 21(1):11–21, 2002.

S. Lopez, M. Prieto, J. Dijkstra, M. S. Dhanoa, and J. France. Statistical evaluation of mathematical models for microbial growth. *International Journal of Food Microbiology*, 96(3):289–300, Nov. 2004.

R. Lund and J. Reeves. Detection of undocumented changepoints: a revision of the two-phase regression model. *Journal of Climate*, 15(17):2547–2554, 2002.

K. Luwel, A. Beem, P. Onghena, and L. Verschaffel. Using segmented linear regression models with unknown change points to analyze strategy shifts in cognitive tasks. *Behavior Research Methods, Instruments, & Computers*, 33(4):470–478, 2001.

D. W. Marquardt. An algorithm for least squares estimation of nonlinear parameters. *SIAM Journal on Applied Mathematics*, 11:431–441, 1963.

D. C. Montgomery. *Introduction to Statistical Quality Control*. Wiley, New York, 1985.

M. Mudelsee. Ramp function regression: a tool for quantifying climate transitions. *Computers & Geosciences*, 26(3):293–307, 2000.

V. M. R. Muggeo. Estimating regression models with unknown break-points. *Statistics in Medicine*, 22(19):3055–3071, 2003.

J. C. Nash. The (Dantzig) simplex method for linear programming. *Computing in Science and Engineering*, 2(1):29–31, 2000.

I. Newton. *De Methodus Fluxionum et Serierum infinitorum*. 1664–1671.

J. Nunemacher. On solid angles and the volumes of regular polyhedra. *Mathematics Magazine*, 72(1):56–58, 1999.

E. S. Page. A test for a change in a parameter occurring at an unknown point. *Biometrika*, 42:523–527, 1955.

R. Pastor-Barriuso, E. Guallar, and J. Coresh. Transition models for change-point estimation in logistic regression. *Statistics in Medicine*, 22:1141–1162, 2003.

P. Perron and X. Zhu. Structural breaks with deterministic and stochastic trends. *Journal of Econometrics*, 129:65–129, 2005.

J. Raphson. *Analysis aequationum universalis*. 1690.

A. M. Reza and M. Doroodchie. Cramer-Rao lower bound on locations of sudden changes in a steplike signal. *IEEE Transactions on Signal Processing*, 44(10): 2551–2556, oct 1996.

T. Söderström and Stoica. Instrumental variable methods for system identification. *Circuits, Systems, and Signal Processing*, 21(1):1–9, 2002.

C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 and 623–656, 1948.

W. A. Shewhart. *Economic Control of Manufactured Products*. Van Nostrand Reinhold, 1931.

A.-R. Solow. Testing for climate change: an application of the two-phase regression model. *Journal of Applied Meteorology*, 26(10):1401–1405, 1987.

G. Staude. Precise onset detection of human motor responses using a whitening filter and the log-likelihood ratio test. *IEEE Transactions on Biomedical Engineering*, 48(11):1292–1305, 2001.

A. Swami and B. Sadler. Cramer-Rao bounds for step-change localization in additive and multiplicative noise. In *Ninth IEEE Signal Processing Workshop on Statistical Signal and Array Processing*, pages 403–406, 1998.

D. Timmer and J. Pignatiello Jr. Change point estimates for the parameters of an AR(1) process. *Quality and Reliability Engineering International*, 19:355–369, 2003.

A. R. Tome and P. M. A. Miranda. Piecewise linear fitting and trend changing points of climate parameters. *Geophysical Research Letters*, 31:1–4, 2004.

J. D. Toms and M. L. Lesperance. Piecewise regression: A tool for identifying ecological thresholds. *Ecology*, 84(8):2034–2041, 2003.

J. Y. Tourneret, A. Ferrari, and S. A. Cramer-Rao lower bounds for change points in additive and multiplicative noise. *Signal Processing*, 84(7):1071–1088, July 2004.

L. Verschaffel, E. De Corte, C. Lamote, and N. Dhert. The acquisition and use of an adaptive strategy for estimating numerosity. *European Journal of Psychology of Education*, 13:347–370, 1998.

L. J. Vostrikova. Detecting 'disorder' in multidimensional random processes. *Soviet Mathematics Doklady*, 24:55–59, 1981.

X. L. Wang. Comments on "Detection of undocumented changepoints: a revision of the two-phase regression model". *Journal of Climate*, 16(20):3383–3388, 2003.

A. Willsky and H. Jones. A generalized likelihood ratio approach to the detection and estimation of jumps in linear systems. *IEEE Transactions on Automatic Control*, AC-21(February):108–112, 1976.

A. Zeileis, C. Kleiber, W. Kramer, and K. Hornik. Testing and dating of structural changes in practice. *Computational Statistics & Data Analysis*, 44(1-2):109–123, Oct. 2003.