Detection in Random Fields

E. P. Hupkens
Department of Applied Mathematics
University of Twente
P. O. Box 217
7500 AE Enschede
The Netherlands

Druk: Universiteits drukkerij Universiteit Twente
ISBN 90-36510058

DETECTION IN RANDOM FIELDS

PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit Twente,
op gezag van de rector magnificus,
prof. dr. F. A. van Vught,
volgens besluit van het College voor Promoties
in het openbaar te verdedigen
op vrijdag 10 oktober 1997 te 15.00 uur.

door

Erik Peter Hupkens
geboren op 9 januari 1970
te Coevorden

Dit proefschrift is goedgekeurd
door de promotor
prof. dr. A. Bagchi

Preface

Finally, after four years of staring at the wall in my office, and occasionally writing down some ideas, the time has come to apply the last few changes to my thesis. In a way, it feels rather strange to me to write a preface as the concluding part; of course I know that for the reader these are indeed the first words to read. Also, I realize that this is the only time for me to be able to make some meaningful comments about the time spent in the process of writing this thesis. If I would have written this preface four years ago, there would not have been much to say. Now that I've almost finished my thesis, some words of thanks are in order.

First of all, I would like to thank Arun Bagchi for being my supervisor. Already as a student, I enjoyed the freedom given by Arun, knowing that he would still be there if something went wrong. During the last four years, we have spent many hours discussing the problems that I stumbled upon, and Arun always managed to direct me in the right way. The people that also had some direct influence on the development of my thesis are the members of my *begeleidingscommissie*. Bart de Jong, Rob Luesink, Ferdi van der Heijden and Wim Albers always carefully watched over the progress I made. In particular, I am indebted to Rob Luesink and Ferdi van der Heijden, for carefully reading my thesis and commenting on the contents and the presentation, thus allowing me to make considerable improvements. The members of my *promotiecommissie*, Ravi Mazumdar, Wim Albers, Huibert Kwakernaak en Paul Regtien, and in particular Henk Blom, also provided me with several useful remarks that improved the quality of the thesis.

On a less professional level, I would like to thank all the people that have made my stay in Enschede a pleasant one. Al the (ex-)colleagues from the System- and Control Group, and in particular Robert van der Geest, who had to endure my presence in the same room for the last part of my stay, provided me with a pleasant working atmosphere. Also I would like to mention Pancho, Thomas, Krzysztof, Antonio and all the guests that took part in our daily lunch festivities in the Bastille. Unfortunately, that tradition has left with you all.

All friends, relatives and teammates offered me the opportunity to relax from time to time and play an occasional football match, thus keeping me both physically and mentally healthy.

Erik Hupkens Enschede, September 5 1997

Contents

Chapter 1

Introduction

During the past centuries mankind has been trying to gain control over the environment. With the increasing demand for products and the decreasing supply of human labour, the need for machines to produce more and with less effort has been growing rapidly. Since these machines are required to operate autonomously, the occurrence of errors have to be detected automatically. In general, this is realized by monitoring a certain signal obtained from a sensor that is measuring an output of the machine. As soon as this signal deviates from its nominal value, an alarm is given. In many situations, we may have the possibility to adjust the inputs of the machine. This way we may avoid the occurrence of errors to some extent without having to shut down the machine. This type of control is sufficient whenever the errors that occur are within the specifications of the model. However, if some part of the machine becomes damaged, it may become uncontrollable. In that case, the machine has to be turned off in order to be repaired.

Classical fault detection theory basically deals with problems of this type. A certain process is monitored, for which the stochastic characteristics are known. At some point in time, a change in the process may occur, thus disrupting the stochastic behaviour of the process. The objective of the classical fault detection problem is to detect such changes as soon as possible after they occur. This implies that the detection problem basically is a statistical hypothesis testing problem, where the hypotheses represent the nominal state of the process and the changed state of the process. Due to the unknown change time, the tests that are used generally are sequential tests.

The detection problem that is dealt with in this thesis is essentially different from the classical fault detection problems. Instead of monitoring a certain one-parameter process, i.e., a process with one independent variable, for instance time, we are dealing with two-parameter processes. Two-parameter processes appear frequently in image processing, remote sensing, surface inspection, agriculture and other geographical applications. For example, an image may be considered as a two-parameter process; instead of having time as the independent variable, the position on a two-dimensional plane is the independent variable. The changes that may appear in these two-parameter processes vary from objects to be recognized in images to scratches on surfaces and ecological disasters on planet surfaces. Although this problem differs substantially from the classical fault detection problem, we may use several results from classical detection theory. In Chapter 2 we give an outline of the results from detection theory that will be used in this thesis.

Instead of having time as a parameter, we use the position on a two-dimensional plane as our parameters. The main difference with one-parameter processes is the lack of ordering in two-parameter processes. In general, there does not exist a natural unique ordering on a two-dimensional plane. Due to this lack of ordering, the stochastic nature of the two-parameter process may be more complex than that of a one-parameter process. A one-parameter process is said to be non-causal if the present state depends on the future states. In general, when time is our only parameter, the present only depends on the past so that any physically meaningful one-parameter process is causal. However, if we impose some ordering on a two-dimensional plane, we create a one-parameter process that may very well be non-causal.

The two-parameter processes that will be used in this thesis are also known as *random fields* in the literature [49, 24]. One of the first contributors to the theory on two-parameter processes is Whittle [45], who used these processes for agricultural applications. Throughout this thesis, we only use random fields that are defined on a discrete plane. This *grid* or *index set* is denoted by $\mathcal{G}$. The elements of $\mathcal{G}$ are termed *sites* and are generally denoted by the letter t or s in this thesis. The exact structure of $\mathcal{G}$ depends on the situation; in Chapters 3 and 4, where $\mathcal{G}$ is simply a grid on a certain two-dimensional plane, it may be modelled as a subset of $\mathbb{Z}^2$. In Chapter 5 the structure of $\mathcal{G}$ is more abstract; hence we use the term index set instead of grid.

Chapter 3 deals with the detection of global changes in autoregressive random fields. In the literature, autoregressive fields are mostly used to model textured images. Since these autoregressive fields are strongly non-causal, the only sequential detection algorithms that are considered in this case are sequential in time. The non-sequential detection algorithm may be considered as a classification algorithm. This approach is particularly useful when all data is easily obtainable. For example, in image processing the entire image is immediately available; there is no need to make our observations on the plane in a sequential way.

If the changes have a local nature, i.e., they appear only in some part of the two-dimensional plane, a non-sequential test may not be accurate enough. For example, if we want to detect a small scratch on a large surface, the deviation in the test statistic resulting from this scratch may very well be absorbed by the noise over the entire surface. Hence, a non-sequential test, which may also be regarded as a global image operator in this context, may not detect such local changes. Therefore, a local image operator that only performs tests on small parts of the plane may be superior to a global image operator. However, to be able to use such a sequential test with some theoretical foundation we need some causal ordering of the process on the two-dimensional plane. Hence, in Chapter 4 we consider local changes in causal and semi-causal random fields. The structural properties of the process determine the ordering that is used for our sequential detection algorithms.

Finally, in Chapter 5 we try to detect both global and local changes with the least possible effort. The effort mentioned here results from the measurements of the individual data points. In other words, we desire to minimize the number of measurements that are needed to detect a change. The one-parameter process resulting from any ordering on a random field is assumed to be causal in this chapter. Although this restricts the class of random fields considerably, it allows us to extend some of the results from the quickest detection theory for one-parameter processes. Two different approaches are used to define a quickest detection problem for random fields; the Neyman-Pearson approach and the Bayesian approach. Using the Neyman-Pearson approach, we simply try to minimize the number of observations that are needed to make a decision, when the probabilities of making an incorrect decision are bounded by given constants. In the Bayesian approach, a cost function that contains both the number of observations and the error probabilities is minimized. These two approaches will be introduced in Chapter 2.

Chapter 2

Detection Theory

In this chapter we focus on the tools from detection theory that will be used in the following chapters. Most of the theory given here is taken from the books by Basseville and Nikiforov [2], Poor [36], Lehmann [25] and Wald [43, 44]. Section 2.1 describes some of the statistical concepts on which detection theory is constructed. In Section 2.2 we focus on sequential detection algorithms that are commonly used in the literature.

2.1 Statistical tests

In general, a detection problem may be described as a comparison between a nominal model and a model under faulty operation. Based on some measurements from the process, one of those models has to be selected as the correct one. Clearly, in case the models are deterministic this is a rather straightforward task. However, in general we may not be able to model the process exactly. To deal with this lack of certainty, a stochastic component is added to the models. The assignment of a measurement to one of two stochastic models may be done with the help of a statistical test.

First we formulate two hypotheses; the *null hypothesis* that assumes the nominal model to hold, and the *alternative hypothesis* that assumes some other model to hold. We will use $\mathcal{H}_0$ and $\mathcal{H}_1$ to denote the null- and the alternative hypothesis, respectively. The model under the alternative hypothesis may be completely known, or may be contained in a certain class of models. For the largest part of this thesis we assume the class of models under the alternative hypothesis to be countable, i.e., we may assign a parameter θ to each model, and the class of models may be parameterized by

$$\Theta = \{\theta_1, \theta_2, \ldots\}$$

Only in Chapters 3 and 4 we give some results for a parameter set that is not countable. In those cases, the parameters may assume any real value in a certain interval. Note that θ may be vector-valued.

If Θ consists of one element only, the alternative hypothesis is said to be *simple*; otherwise it is said to be *composite*. The nominal model is said to have parameter $\theta = \theta_0$. We emphasize here that Θ only contains the parameters of the models under the alternative hypothesis. We denote

$$\Theta^* = \Theta \cup \{\theta_0\}$$

as the parameter set including the models under both hypotheses. The true parameter will be denoted by $\boldsymbol{\theta}$. We assume that $\boldsymbol{\theta}$ is a random vector variable taking values in the parameter set Θ^*.

Let us assume that we have a process Y_k, $k = 1, \ldots, n$, where each Y_k is a random scalar variable. For the moment we assume n to be finite, so that the process may be represented by a random vector variable $\mathbf{Y} = (Y_1, \ldots, Y_n)$. We assume the outcome of the process to be given by $\mathbf{y} \in \mathbb{R}^n$. This vector is also termed the *data* of the process. These data may be measured, giving us the *measurement* of the process. For now we assume the measurement to be equal to the data; all available data is measured simultaneously and the measurement is perfect. In Section 2.2 we allow the data to enter sequentially and n may be infinite, and in Chapters 3 and 4 the measurements may be noisy.

Based on the measurement $\mathbf{y}$, we want to accept either the null hypothesis or the alternative hypothesis.

Definition 2.1 (Decision function) *A statistical test, or decision function, is defined as a function δ from $\mathbb{R}^n$ to $[0, 1]$. For a given measurement $\mathbf{y}$, the null hypothesis then is rejected with probability $\delta(\mathbf{y})$. A decision function that only takes the values 0 or 1 is said to be non-randomized. The class of decision functions is denoted by $\mathcal{D}$.*

In practice, this implies that to each measurement $\mathbf{y}$ a number $\delta(\mathbf{y})$ between 0 and 1 is assigned. The decision then depends on some arbitrary test (for example, throwing a coin) that decides in favour of the null hypothesis with probability $1-\delta(\mathbf{y})$ and in favour of the alternative hypothesis with probability $\delta(\mathbf{y})$.

It follows that a non-randomized test divides the measurement space $\mathbb{R}^n$ in two regions; the *acceptance region* $S_0(\delta)$ and the *rejection region* $S_1(\delta)$. In this thesis, randomization is only used in degenerate cases.

Given a model, we may derive a probability distribution function for the process $\mathbf{Y}$. We denote the distribution of $\mathbf{Y}$ corresponding to the model with parameter θ as

$$F_\theta(\mathbf{y}) = \mathbf{Pr}[\mathbf{Y}_1 \leq \mathbf{y}_1, \ldots, \mathbf{Y}_n \leq \mathbf{y}_n | \boldsymbol{\theta} = \theta]$$

We assume this distribution function to be absolutely continuous throughout this thesis. Hence, the probability density function of $\mathbf{Y}$ exists; it will be denoted as $f_\theta(\mathbf{y})$. Although formally we should write f_{θ_0} to denote the density function under the null hypothesis, we use f_0 instead to simplify the notations. Moreover, in case both hypotheses are simple, we may use 0 and 1 instead of θ_0 and θ_1 as the indices.

Clearly, to be able to detect a change, there has to be some difference in the stochastic nature of the process under both hypotheses. A measure that is widely used to indicate this difference is the *Kullback-Leibler information* [22].

Definition 2.2 (Kullback-Leibler information) *The Kullback-Leibler information for a test between parameters θ_1 and θ_2 is defined as*

$$I(\theta_1, \theta_2) = \mathbf{E}_{\theta_1}[\log \frac{f_{\theta_1}(\mathbf{Y})}{f_{\theta_2}(\mathbf{Y})}]$$

where $\mathbf{E}_\theta$ denotes the expectation under the assumption that $\boldsymbol{\theta} = \theta$, i.e.,

$$\mathbf{E}_\theta[h(\mathbf{Y})] = \int_{\mathbf{y} \in \mathbb{R}^n} h(\mathbf{y}) f_\theta(\mathbf{y}) d\mathbf{y}$$

Since in our situation $\theta_2 = \theta_0$ is the nominal parameter, we generally use the notation

$$I(\theta) = I(\theta, \theta_0)$$

The Kullback-Leibler information may be used as an indicator of the *detectability* of a change [2].

Definition 2.3 (Detectability) *A change θ is said to be detectable if the Kullback-Leibler information is strictly positive.*

In fact, according to this definition, a change is not detectable if and only if $f_\theta \equiv f_0$ almost everywhere, i.e., the functions may only differ on sets of zero probability.

2.1.1 Neyman-Pearson approach

One way of defining the quality of a test is to use the error probabilities; the probabilities of accepting an incorrect hypothesis. A *false alarm* is defined as the incorrect rejection of the null hypothesis. A *miss* is defined as the incorrect acceptance of the null hypothesis. The probability of false alarm may now be written as

$$\begin{aligned} \alpha(\delta) &= \mathbf{E}_{\theta_0}[\delta(\mathbf{Y})] \\ &= \int \delta(\mathbf{y}) f_0(\mathbf{y}) d\mathbf{y} \end{aligned}$$

A test that has a false alarm probability equal to α is said to have *size* α.

The miss probability is not that easily defined, since the alternative hypothesis is not necessarily simple. The opposite of missing a change obviously is to detect a change. A change is detected if the null hypothesis is correctly rejected. Let us define the *power function*, or *detection probability*, as

$$\begin{aligned} \beta^{\theta}(\delta) &= \mathbf{E}_{\theta}[\delta(\mathbf{Y})] \\ &= \int \delta(\mathbf{y}) f_{\theta}(\mathbf{y}) d\mathbf{y} \end{aligned}$$

The miss probability may now be given as a function of the parameter θ,

$$\gamma^{\theta}(\delta) = 1 - \beta^{\theta}(\delta)$$

Hence, if $\boldsymbol{\theta} = \theta$ then the probability that the change will not be detected when using the decision function δ is given by $\gamma^{\theta}(\delta)$.

In case the probability distribution function of $\boldsymbol{\theta}$ is known, we may actually calculate the overall detection- and miss probabilities. Let us define ξ as the probability distribution of $\boldsymbol{\theta}$ on Θ, i.e.,

$$\xi(\theta) = \mathbf{Pr}[\boldsymbol{\theta} = \theta | \boldsymbol{\theta} \in \Theta]$$

denotes the probability that $\boldsymbol{\theta}$ equals θ, given that a change is present. Then

$$\begin{aligned} \beta^{\xi}(\delta) &= \sum_{\theta \in \Theta} \xi(\theta) \beta^{\theta}(\delta) \\ \gamma^{\xi}(\delta) &= \sum_{\theta \in \Theta} \xi(\theta) \gamma^{\theta}(\delta) \end{aligned}$$

In fact, the presence of the distribution ξ implies that the alternative hypothesis may be treated as a simple hypothesis. The distribution function of $\mathbf{y}$ is under the alternative hypothesis then given by

$$F_1(\mathbf{y}) = \sum_{\theta \in \Theta} \xi(\theta) F_{\theta}(\mathbf{y})$$

Definition 2.4 (Most powerful test) *The most powerful (MP) test is defined as the test between two simple hypotheses that maximizes the probability of detection, under the constraint that the size of the test is bounded by a certain constant, i.e.,*

$$\delta^{MP} = \{\delta \in \mathcal{D}_{\alpha} | \forall \delta' \in \mathcal{D}_{\alpha}, \beta(\delta) \geq \beta(\delta')\}$$

where $\mathcal{D}_{\alpha}$ is defined as the class of decision functions with size smaller than or equal to α.

From the well-known Neyman-Pearson lemma we know that this test is in fact a likelihood ratio test.

Definition 2.5 (Likelihood ratio test) *A test between two simple hypotheses is said to be a likelihood ratio test if there exists a constant λ such that*

$$\delta(\mathbf{y}) = \begin{cases} 0 & \textit{if } \ell(\mathbf{y}) < \lambda \\ q(\mathbf{y}) & \textit{if } \ell(\mathbf{y}) = \lambda \\ 1 & \textit{if } \ell(\mathbf{y}) > \lambda \end{cases}$$

where

$$\ell(\mathbf{y}) = \frac{f_1(\mathbf{y})}{f_0(\mathbf{y})}$$

is defined as the likelihood ratio and $q(\mathbf{y})$ takes values between 0 and 1.

Without a proof, we now state the Neyman-Pearson lemma. For a proof, see Lehmann [25, Section 3.2] or Poor [36, p. 23].

Lemma 2.1 (Neyman-Pearson) *For a pair of simple hypotheses the following holds:*

- **Optimality.** *Any likelihood ratio test with size α is a most powerful test.*

- **Existence.** *For every $\alpha \in (0, 1)$ such a likelihood ratio test exists with $q(\mathbf{y}) = q_0$.*

- **Uniqueness.** *If δ' is a most powerful test of size α, then it is a likelihood ratio test. Clearly, on the regions of zero probability under both hypotheses, the uniqueness is not guaranteed.*

In case the alternative hypothesis is composite, and we do not have a prior distribution function on the parameter set, some other approaches for measuring the quality of a test exist. If we generalize the principle of the most powerful test, we arrive at the *uniformly most powerful* (UMP) tests. A test $\delta \in \mathcal{D}_\alpha$ is said to be UMP if it maximizes the power function for all $\theta \in \Theta$. However, such tests only exist in a limited number of situations.

An alternative is to use the so-called *generalized likelihood ratio*, or *maximum-likelihood ratio*, to construct our tests. The generalized likelihood ratio (GLR) is defined as

$$T(\mathbf{y}) = 2 \sup_{\theta \in \Theta} \log \frac{f_\theta(\mathbf{y})}{f_0(\mathbf{y})}$$

Note that this test is based on the log-likelihood ratio. A generalized likelihood ratio test is then constructed in a similar way as the likelihood ratio test. As is well known in the literature, these tests will generally not be uniformly most powerful [25].

An interesting approximation for the generalized likelihood ratio is known in case the parameter set is given by $\Theta = \mathbb{R}^p$. So, each parameter θ is a vector from $\mathbb{R}^p$. For simplicity, we assume $\theta_0 = 0$, so that the null hypothesis is given by

$$\mathcal{H}_0 : \boldsymbol{\theta} = 0$$

If the changes are small, that is $\|\theta\|$ is close to zero, and the density functions are differentiable with respect to θ, the GLR may be approximated by the *score test* statistic (Basseville and Nikiforov [2])

$$S(\mathbf{y}) = \mathcal{Z}_0(\mathbf{y})'\Gamma(0)^{-1}\mathcal{Z}_0(\mathbf{y})$$

Here

$$\mathcal{Z}_\theta(\mathbf{y}) = \frac{\partial}{\partial\theta}\log f_\theta(\mathbf{y})$$

is the *efficient score* and

$$\Gamma(\theta) = \mathbf{E}_\theta[\mathcal{Z}_\theta(\mathbf{Y})\mathcal{Z}_\theta(\mathbf{Y})']$$

is the *Fisher information matrix.*

Theorem 2.1 *If $\|\boldsymbol{\theta}\| \to 0$ and $n \to \infty$ then $T(\mathbf{y}) \to S(\mathbf{y})$ for all $\mathbf{y}$.*

Proof. Writing down the second order Taylor expansion of the log-likelihood ratio around 0 (with respect to the parameter θ), we obtain

$$\ell_\theta(\mathbf{y}) = \ell_0(\mathbf{y}) + \theta'\mathcal{Z}_0(\mathbf{y}) + \frac{1}{2}\theta'\left(\frac{\partial^2\ell_\theta(\mathbf{y})}{\partial\theta^2}\right)_{\theta=0}\theta + o(\|\theta\|^2)$$

where $\ell_\theta(\mathbf{y}) = \log\frac{f_\theta(\mathbf{y})}{f_0(\mathbf{y})}$. The parameter $\bar{\theta}$ that maximizes $\ell_\theta(\mathbf{y})$ satisfies

$$\mathcal{Z}_{\bar{\theta}}(\mathbf{y}) = 0$$

The expansion of the efficient score may be written as

$$\mathcal{Z}_\theta(\mathbf{y}) = \mathcal{Z}_0(\mathbf{y}) + \left(\frac{\partial^2\ell_\theta(\mathbf{y})}{\partial\theta^2}\right)_{\theta=0}\theta + o(\|\theta\|)$$

so that

$$\begin{aligned}\mathcal{Z}_0(\mathbf{y}) &= \mathcal{Z}_{\bar{\theta}}(\mathbf{y}) - \left(\frac{\partial^2\ell_\theta(\mathbf{y})}{\partial\theta^2}\right)_{\theta=0}\bar{\theta} + o(\|\theta\|) \\ &= -\left(\frac{\partial^2\ell_\theta(\mathbf{y})}{\partial\theta^2}\right)_{\theta=0}\bar{\theta} + o(\|\theta\|)\end{aligned}$$

which gives us the following expression for the maximum of the log-likelihood ratio

$$\ell_{\bar{\theta}}(\mathbf{y}) = -\frac{1}{2}\mathcal{Z}_0(\mathbf{y})'\left(\frac{\partial^2 \ell_\theta(\mathbf{y})}{\partial \theta^2}\right)^{-1}_{\theta=0} \mathcal{Z}_0(\mathbf{y}) + o(\|\theta\|^2)$$

From Serfling [39] we know that

$$-\left(\frac{\partial^2 \ell_\theta(\mathbf{y})}{\partial \theta^2}\right)_{\theta=0} \to \Gamma(0)$$

with probability 1 as n goes to infinity. As $\boldsymbol{\theta}$ converges to zero the order term vanishes, so that we obtain

$$\ell_{\bar{\theta}}(\mathbf{y}) \to \frac{1}{2}\mathcal{Z}_0(\mathbf{y})'\Gamma(0)^{-1}\mathcal{Z}_0(\mathbf{y}) \text{ w.p. } 1$$

and finally,

$$T(\mathbf{y}) = 2\ell_{\bar{\theta}}(\mathbf{y}) \to \mathcal{Z}_0(\mathbf{y})'\Gamma(0)^{-1}\mathcal{Z}_0(\mathbf{y}) = S(\mathbf{y}) \text{ w.p. } 1$$

□

Since the efficient score $\mathcal{Z}_0(\mathbf{y})$ is under the null hypothesis asymptotically distributed as a Gaussian variable with zero mean and covariance matrix $\Gamma(0)$, we now find that under the null hypothesis S is χ^2-distributed with p degrees of freedom. As a result, the GLR is asymptotically χ^2_p-distributed under the null hypothesis.

2.1.2 Bayesian approach

Let us assume that the prior distribution of $\boldsymbol{\theta}$ on the parameter set Θ^* is known, say

$$\xi(\theta) = \mathbf{Pr}[\boldsymbol{\theta} = \theta]$$

Note that the nominal model is included here. Suppose that we may assign a certain cost to making incorrect decisions, that is, we may define a cost function

$$K(\theta, d) = \begin{cases} c_0 & \text{if } \theta = \theta_0 \text{ and } d = 1 \\ c_1 & \text{if } \theta \neq \theta_0 \text{ and } d = 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.1}$$

Basically, $K(\theta, d)$ is the cost of making decision d when θ is the correct parameter. Note that this cost function has a rather simple form. All parameters in Θ give the same cost if not detected. Although the theory exists to deal with more general cost functions, we only use them in this form.

The expected cost resulting from the use of the decision function δ, given that θ is the correct parameter, may be calculated as

$$\begin{aligned} K(\theta,\delta) &= \mathbf{E}_\theta[\mathbf{E}_{\delta(\mathbf{Y})}[K(\boldsymbol{\theta},d)]] \\ &= \int E_{\delta(\mathbf{y})}[K(\theta,d)]f_\theta(\mathbf{y})d\mathbf{y} \qquad (2.2) \\ &= \int\{\delta(\mathbf{y})K(\theta,1)+(1-\delta(\mathbf{y}))K(\theta,0)\}f_\theta(\mathbf{y})d\mathbf{y} \end{aligned}$$

The unconditional expected cost resulting from a decision function δ may then be written as

$$\begin{aligned} K(\xi,\delta) &= \mathbf{E}_\xi[K(\boldsymbol{\theta},\delta)] \\ &= \sum_{\theta\in\Theta^*}\xi(\theta)K(\theta,\delta) \end{aligned}$$

Definition 2.6 (Bayes decision function) *A Bayes decision function is defined as a test that minimizes the expected cost, i.e.,*

$$\delta_\xi^B=\{\delta\in\mathcal{D}|\forall\delta'\in\mathcal{D},K(\xi,\delta)\le K(\xi,\delta')\}$$

The solution to this problem may easily be obtained by rewriting the expected cost as

$$K(\xi,\delta)=\int\delta(\mathbf{y})[c_0\xi(\theta_0)f_0(\mathbf{y})-c_1\sum_{\theta\in\Theta}\xi(\theta)f_\theta(\mathbf{y})]d\mathbf{y}+c_1(1-\xi(\theta_0))$$

It then follows that the Bayes decision function may be written as

$$\delta_\xi^B(\mathbf{y})=\begin{cases}0 & \text{if } c_0\xi(\theta_0)f_0(\mathbf{y})>c_1\sum_{\theta\in\Theta}\xi(\theta)f_\theta(\mathbf{y})\\ q(\mathbf{y}) & \text{if } c_0\xi(\theta_0)f_0(\mathbf{y})=c_1\sum_{\theta\in\Theta}\xi(\theta)f_\theta(\mathbf{y})\\ 1 & \text{if } c_0\xi(\theta_0)f_0(\mathbf{y})<c_1\sum_{\theta\in\Theta}\xi(\theta)f_\theta(\mathbf{y})\end{cases} \qquad (2.3)$$

where $q(\mathbf{y})$ may be chosen arbitrarily between 0 and 1.

As we may see, the Bayes decision function is also a likelihood ratio test. Indeed, the inequalities defining the decision function may be rewritten as

$$\sum_{\theta\in\Theta}\xi(\theta)\ell_\theta(\mathbf{y})\gtrless\frac{c_0\xi(\theta_0)}{c_1}$$

where

$$\ell_\theta(\mathbf{y})=\frac{f_\theta(\mathbf{y})}{f_0(\mathbf{y})}$$

Note, however, that the probabilities on the left hand side do not add up to 1. If we insist on having this property, we may replace ξ by

$$\xi'=\frac{\xi}{1-\xi(\theta_0)}$$

so that the Bayes decision function may be rewritten as

$$\sum_{\theta \in \Theta} \xi'(\theta) \ell_\theta(\mathbf{y}) \gtrless \frac{c_0 \xi'(\theta_0)}{c_1}$$

2.1.3 Comparison of approaches

Although both the Neyman-Pearson approach and the Bayesian approach lead to likelihood ratio tests, there remains a difference in the interpretation. This difference may perhaps best be explained by a simple example.

A farmer lives, together with his family, on a piece of land where he earns his livings from the sale of milk from his cows. One day he is informed by one of his neighbours that a certain disease has been spreading among the cows of some of his colleagues. If his cows would also have this disease, they could be cured by some treatment. If they did not receive this treatment, they might die. On the other hand, if they did not have this disease, but did receive the treatment, the milk production would be halved. Fortunately, there exists a test method to determine whether or not the cows are infected. Unfortunately, this test method is not perfectly reliable. If the cows are healthy, the outcome of the test may be described by a Gaussian variable with mean 0 and variance 1. If the cows are infected by the disease, the outcome may be described by a Gaussian variable with mean 1 and variance 1.

Suppose our farmer uses a most powerful test to make a decision on the state of health of his cows. Halving the milk production unnecessarily is not an attractive option for him, so that he is inclined to use a small size of the test, say $\alpha = 0.01$. Using the previous theory, we may now find the most powerful test resulting from these settings, which declares the cows ill if the outcome of the test is larger than 2.33. The resulting detection probability then is 0.09, which is clearly far too small.

Now let us assume the farmer uses the Bayesian approach. The cost of incorrectly declaring the cows ill is halving of the milk production, and the cost of incorrectly declaring the cows healthy is the entire milk production. So, let us say $c_0 = 1/2$ and $c_1 = 1$. According to some epidemiological model, the probability that his cows are ill is calculated as 0.2. The Bayes decision function then gives us a threshold 1.19, which results in a probability of false alarm equal to 0.12 and a probability of detection equal to 0.42.

As we may see, both tests have their weak points. For the Neyman-Pearson approach, we have to make a suitable choice for the size of the test. For the Bayesian approach, we have to make a choice for the prior probability of a change. Although the Bayes decision function seems to be better in this case, this is merely the effect of the choices made by the farmer; if he would have chosen a size equal to 0.12, he would have obtained the same test for both situations.

Obviously, the main conclusion that should be drawn from this example is that the test method should be drastically improved. At the end of the following section we get back to this example, and show that we may obtain far better results.

2.2 Sequential detection

So far we assumed all data to be measured simultaneously. In practice, this may not be very efficient as the number of data points may be very large, or the cost of taking these measurements may be too high to include all of them. Moreover, the data may not be available at all time. For example, for a process that evolves in time, the future data points are not known yet. Furthermore, in many situations only a few data points may contain sufficient information to reach a qualitatively good decision. Hence, instead of including all data points, we may select a subset of data points, a *sample*, that should be measured. Based on this measurement we may now choose between one of the hypotheses. If the information obtained from the sample is not conclusive, we may decide to add another sample and continue this procedure until we have selected one of the hypotheses or there are no more data points remaining.

At this point we assume the data to be ordered in some sense. That is, we may consider the data to be a function of $\mathbb{N}$ to $\mathbb{R}$, where the kth data point may only be accessed if all $k-1$ previous data points have been processed. In general, a sequence of data points will simply be called a process. If these data points may be described by a random variable, we call the sequence a *stochastic process*. A stochastic process is said to be independent and identically distributed (i.i.d.) if the random variables are all independent and have the same distribution.

Before we proceed, let us first introduce some basic concepts.

Definition 2.7 (Stopping time) *A random variable τ is said to be a stopping time with respect to a certain process if it assumes only values in $\mathbb{N}$ and if the event $\{\tau = k\}$ is determined by the outcome of the process up to the kth data point. The class of stopping times is denoted by $\mathcal{N}$.*

In other words, a stopping time may be used to determine when we should stop sampling. Note that τ now is equal to the number of data points that are processed before we make a decision. The expectation of this stopping time is said to be the *average sample number.*

After we stop sampling, we have to choose between the two hypotheses, for which we may use a decision function as defined in Section 2.1.

Definition 2.8 (Stopping rule) *A sequential decision rule, or simply stopping rule, is defined as a pair* $\mathbf{d} = (\tau, \delta)$*, where* τ *denotes a stopping time and* δ *denotes the decision function that is used to make a decision after the procedure is stopped. The class of stopping rules is denoted by* $\mathcal{R} = \mathcal{N} \times \mathcal{D}$*.*

A stopping rule is said to be *truncated* at k if $\tau \leq k$ with probability 1 under both hypotheses. Furthermore, a stopping rule is said to be *concentrated* at k if $\tau = k$ with probability 1 under both hypotheses. Clearly, any non-sequential test is a concentrated stopping rule.

We now wish to find a stopping rule that optimizes the decision quality in some sense. As in the non-sequential case, the decision quality is measured by the error probabilities. However, in the sequential case we also desire to minimize the number of measurements that are needed to reach a decision. Since these two goals are conflicting, we have to find some balance between the two. In the next sections we describe some of the approaches to this problem.

2.2.1 Quickest detection

Let us once more return to our farmer and his cows. After testing his cows the farmer decided that they were not infected yet. However, as the disease was still spreading, he decided to keep testing his cows once every day. Naturally, our farmer wishes to detect a possible contamination as soon as possible after it has occurred.

The classical quickest detection problems, as defined by Shiryaev [40], Lorden [27], Pollak [35], Moustakides [30], and Basseville and Nikiforov [2] are all based on a problem of this type. Basically, it consists of detecting a change in a certain process that appears at some unknown time. For our farmer, the change is the infection of the cows. We only consider this problem in its simplest form, where the samples of the process are independent and identically distributed. Furthermore, the number of samples is countably infinite. Each sample may contain more than one data point, so that we use a vector $\mathbf{y}_k$ to denote the kth sample.

The hypotheses may now be written as

- $\mathcal{H}_0$: The samples $\mathbf{Y}_k$ are i.i.d. with density function $f_0(\mathbf{y}_k)$ for all $k \in \mathbb{N}$.
- $\mathcal{H}_1$: There exists a $t_c \in \mathbb{N}$ such that the samples $\mathbf{Y}_k$ are i.i.d. with density function $f_0(\mathbf{y}_k)$ for all $k < t_c$, and the samples are i.i.d. with density function $f_1(\mathbf{y}_k)$ for all $k \geq t_c$.

Our objective is to minimize the *delay of detection*, which is defined as the difference between the stopping time τ and the *change time* t_c. Clearly, in this context the acceptance of the null hypothesis will never be a logical decision, because at each stage we do not know anything about the future outcomes of the process. At any future stage, a change may appear. Hence, stopping at a certain time may be regarded to be equivalent to rejecting the null hypothesis. As a result, the decision function does not play any role here. Hence, instead of using the class of stopping rules, we may as well use the class of stopping times.

Note that, although the distribution of the process is known after a change has occurred, the alternative hypothesis is composite due to the unknown nature of the change time. The parameter set may be written as

$$\Theta = \{1, 2, 3, \ldots\}$$

where each element denotes a certain change time.

As in the non-sequential case, we have to impose some constraints on the error probabilities. Since in this case we never accept the null hypothesis, the only error that may occur is a false alarm. Basically two ways of imposing a constraint on the class of stopping times exist. The first one assumes the probability of false alarm to be not larger than some constant α. The second one assumes the *mean time between false alarms*, the average sample number under the null hypothesis, to be larger than some number. Here we only consider the first one, simply because of the similarity with the non-sequential problem.

The stopping time that is most commonly used for the quickest detection problem is Page's CUSUM algorithm. This algorithm actually is a sequential version of the maximum likelihood ratio test. In its most transparent form it is given by

$$\tau^P = \min\{k \in \mathbb{N} | \max_{1 \leq j \leq k} S_j^k(\mathbf{y}_1, \ldots, \mathbf{y}_k) > \lambda\}$$

where

$$S_j^k(\mathbf{y}_1, \ldots, \mathbf{y}_k) = \sum_{i=j}^{k} \log \ell(\mathbf{y}_i)$$

Hence, the procedure stops as soon as a sequence of samples has been encountered that leads to a sufficiently large increase of the likelihood ratio. In its final version, which is more easily computable, the CUSUM stopping time is written as

$$\begin{aligned}
\tau^P &= \min\{k \in \mathbb{N} | g_k(\mathbf{y}_1, \ldots, \mathbf{y}_k) > \lambda\} \\
g_k(\mathbf{y}_1, \ldots, \mathbf{y}_k) &= \max\{0, g_{k-1}(\mathbf{y}_1, \ldots, \mathbf{y}_{k-1}) + \log \ell(\mathbf{y}_k)\} \\
g_0 &= 0
\end{aligned}$$

Due to the composite nature of the alternative hypothesis, the optimality of this test can not be guaranteed.

2.2.2 Neyman-Pearson approach

Although there exist several ways of approaching the sequential detection problem from the Neyman-Pearson point of view [40], we only consider one way of dealing with this problem. This approach is based on the minimization of the average sample number, given that both error probabilities are bounded by some given constants. It is frequently used in the classical literature on sequential analysis [43, 44, 2]. The stopping rule that achieves this goal for each parameter θ, is termed the *uniformly most efficient* stopping rule. If a stopping rule only has this property for a parameter θ, then this stopping rule is said to be most efficient at θ.

Definition 2.9 (Uniformly most efficient (UME) stopping rule) *The uniformly most efficient stopping rule is defined as the stopping rule that minimizes the average sample number for each $\theta \in \Theta^*$, given that the error probabilities are bounded by given constants, i.e.,*

$$\mathbf{d}^{UME}(\alpha,\gamma) = \{\mathbf{d} \in \mathcal{R}_{\alpha,\gamma} | (\forall \tau' \in \mathcal{R}_{\alpha,\gamma})(\forall \theta \in \Theta^*), \mathbf{E}_\theta[\tau] \leq \mathbf{E}_\theta[\tau']\}$$

where $\mathcal{R}_{\alpha,\gamma} = \{\mathbf{d} \in \mathcal{R} | \alpha(\delta) \leq \alpha, \gamma(\delta) \leq \gamma\}$.

Note that this approach does include the average sample number under both hypotheses. Furthermore, it is also defined for a composite alternative hypothesis. However, as in the non-sequential case, the existence of a UME stopping rule is not likely for a composite alternative hypothesis.

In the remainder of this section, we assume the alternative hypothesis to be simple. The composite alternative hypothesis will not be treated from the Neyman-Pearson approach, since there generally does not exist a solution. In Section 2.2.3 both simple and composite hypotheses will be used.

As in the non-sequential case, the UME stopping rule may be shown to be based on the likelihood ratio. First let us define the sequential equivalent of the likelihood ratio test.

Definition 2.10 (Sequential probability ratio test) *A sequential probability ratio test (SPRT) is a stopping rule defined by*

$$\begin{aligned} \tau &= \min\{k | \ell(\mathbf{y}_k) \notin (a,b)\} \\ \delta &= \begin{cases} 0 & \text{if } \ell(\mathbf{y}_\tau) \leq a \\ 1 & \text{if } \ell(\mathbf{y}_\tau) \geq b \end{cases} \end{aligned}$$

where (a,b) are the thresholds, and

$$\ell(\mathbf{y}_k) = \frac{f_1(y_1,\ldots,y_k)}{f_0(y_1,\ldots,y_k)}$$

is the likelihood ratio up to the kth data point.

A more general definition of the SPRT includes randomization at the boundaries of the stopping regions, i.e., when the likelihood ratio equals a or b. However, in general it is sufficient to use the non-randomized form. Only in degenerate cases randomization may be required to obtain certain error probabilities.

In some situations we may use a generalized version of a sequential probability ratio test, where the thresholds vary with k. Such a test is termed a *generalized sequential probability ratio test*, or GSPRT. For example, we may want to use a truncated test to guarantee the average sample numbers to be bounded by some constant. Clearly, truncation of a GSPRT at the kth data point is obtained by choosing $a_k = b_k$. If the GSPRT is also concentrated at k, the thresholds at the previous stages are $(-\infty, \infty)$.

Without proof, we now state the following well-known Wald-Wolfowitz lemma. A proof may be found in most standard books on statistics, for example Ghosh [11, Section 3.1.4].

Lemma 2.2 (SPRT is UME) *For a test between two simple hypotheses, where the process is i.i.d. under both hypotheses, any uniformly most efficient stopping rule is a SPRT.*

Although this lemma restricts the class of stopping rules considerably, the problem remains to find the thresholds a and b that define this UME stopping rule. Wald [43] came up with some approximations for these thresholds. If the *overshoot* over the boundaries is negligible, the thresholds are approximately equal to

$$(a, b) \approx (\frac{\gamma}{1-\alpha}, \frac{1-\gamma}{\alpha})$$

The exact values of the thresholds are implicitly given by

$$\begin{aligned} P_0(0) &= 1-\alpha \\ P_1(0) &= \gamma \end{aligned}$$

where $P_i(x)$ is defined as the probability that the null hypothesis will be accepted, given that at this point the log-likelihood ratio already is equal to x, and that the ith hypothesis is correct. Hence,

$$P_i(x) = \mathbf{Pr}[\delta(\mathbf{Y}) = 0 | \boldsymbol{\theta} = \theta_i]$$

If the distribution of the increment of the log-likelihood ratio is known, we may find an expression for these probabilities. In fact, from Basseville and Nikiforov [2, p. 160-161], we may find

$$P_i(z) = G_i(A-z) + \int_A^B P_i(x) dG_i(x-z)$$

where G_i denotes the probability distribution function of the increment of the log-likelihood ratio,

$$G_i(x) = \mathbf{Pr}[\log \ell(Y) < x | \boldsymbol{\theta} = \theta_i]$$

where $i = 0$ or $i = 1$, and the thresholds A and B are equal to $\log a$ and $\log b$, respectively. It is not difficult to see that these equations uniquely define the UME test in case both hypotheses are simple.

The average sample number may be calculated similarly. Indeed,

$$\mathbf{E}_i[\tau] = N_i(0)$$

where

$$N_i(z) = 1 + \int_A^B N_i(x) dG_i(x - z)$$

Wald also gives an approximation for these numbers, which is fairly accurate if the overshoot is negligible. It is given by

$$\begin{aligned} \mathbf{E}_0[\tau] &\approx \frac{A}{I_0} \approx I_0^{-1} \log \frac{\gamma}{1-\alpha} \\ \mathbf{E}_1[\tau] &\approx \frac{B}{I_1} \approx I_1^{-1} \log \frac{1-\gamma}{\alpha} \end{aligned}$$

where I_0 and I_1 are the Kullback-Leibler information numbers, given by

$$\begin{aligned} I_1 &= I(\theta_1) \\ I_0 &= I(\theta_0, \theta_1) \end{aligned}$$

and using the notation of Definition 2.2.

2.2.3 Bayesian approach

From the Bayesian point of view, the sequential detection problem is very similar to the non-sequential detection problem (Berger [3]). However, the sequential problem is a lot more difficult to solve than the non-sequential problem.

Let us assume that the prior probability distribution of the changes is given by ξ, and the miss cost and the false alarm cost are given by c_1 and c_0, respectively. Apart from the cost resulting from the errors we make, we now also have a sampling cost. Throughout this thesis we assume the sampling cost to be equal to the number of data points used before stopping. So, the cost corresponding to a decision d after k data points have been processed, and given that θ is the true parameter, is given by

$$L(\theta, (k, d)) = k + K(\theta, d)$$

where $K(\theta, d)$ is defined by equation (2.1). Using the prior distribution ξ, we may calculate the expected cost for any stopping rule $\mathbf{d} = (\tau, \delta)$ according to

$$L(\xi, \mathbf{d}) = \mathbf{E}_\xi[\tau + K(\boldsymbol{\theta}, \delta)]$$

where $K(\theta, \delta)$ is defined by equation (2.2) and $\mathbf{E}_\xi$ is the expectation given the distribution ξ of $\boldsymbol{\theta}$, as defined in Section 2.1.2. After each measurement, we may update the prior distribution ξ to obtain the posterior distribution using Bayes' rule

$$T^y\xi(\theta) = \frac{\xi(\theta) f_\theta(y)}{\sum_{\theta' \in \Theta^*} \xi(\theta') f_{\theta'}(y)}$$

Hence, T denotes the operator that gives the posterior estimate of ξ. This way we may interpret the sequence $\{\xi, T\xi, TT\xi, \ldots\}$ as a stochastic process as well. We define $\xi_{k+1} = T\xi_k$, where $\xi_0 = \xi$.

Definition 2.11 (Bayes stopping rule) *The Bayes stopping rule is defined as the stopping rule that minimizes the expected cost, i.e.,*

$$\mathbf{d}_\xi^B = \{\mathbf{d} \in \mathcal{R} | \forall \mathbf{d}' \in \mathcal{R}, L(\xi, \mathbf{d}) \leq L(\xi, \mathbf{d}')\}$$

The expected cost that corresponds with the Bayes stopping rule is termed the *Bayes cost*, and is denoted as

$$\rho(\xi) = L(\xi, \mathbf{d}_\xi^B) = \inf_{\mathbf{d} \in \mathcal{R}} L(\xi, \mathbf{d})$$

Clearly, the optimal decision function is again defined by the Bayes decision function δ^B, defined by (2.3). The optimal stopping time may not be found that easily. Logically, we continue processing as long as the cost of continuation is smaller than the cost of stopping. This cost of continuation, however, is rather difficult to compute. The theory on optimal stopping is extensively treated by Shiryayev [41]. Here we focus on the tools that are required for solving our specific problem.

Suppose that we have measured k data points, say $y_1, \ldots, y_k$. The posterior distribution of the parameters is now given by ξ_k. If we stop sampling at this point, and apply the Bayes decision function to make a choice between the hypotheses, then the *additional* cost may easily be shown to be

$$\rho_0(\xi_k) = \min\{\xi_k(\theta_0)c_0, (1 - \xi_k(\theta_0))c_1\}$$

The alternative is to take another data point, so that we arrive at a similar situation for $k + 1$. The additional Bayes cost then is equal to the minimum of the cost of stopping and the cost of continuation, i.e.,

$$\rho(\xi_k) = \min\{\rho_0(\xi_k), 1 + \mathbf{E}_{\xi_k}[\rho(\xi_{k+1})]\}$$

The Bayes stopping time may now be written as

$$\tau_\xi^B = \min\{k|\rho(\xi_k) = \rho_0(\xi_k)\}$$

Note that we use the *additional* cost here. The advantage of using this approach is that after each measurement, the problem is automatically rewritten in its original form. Since the stopping rule is entirely determined by ξ_k, this implies that the same test may be applied at each stage of the procedure. A well-known result from decision theory (Wald [44]) is that the Bayes cost $\rho(\xi)$ is a concave function of ξ. Consequently, the stopping region is of the form

$$S^a \cup S^r$$

where S^a and S^r are closed and convex regions, containing those ξ for which the null hypothesis is accepted and rejected, respectively. In case the alternative hypothesis is simple, this implies that the Bayes stopping time may be written as

$$\tau_\xi^B = \min\{k|\xi_k(\theta_0) \notin (\eta^b, \eta^a)\}$$

The interval $[0, \eta^b]$ then is the rejection region, and the interval $[\eta^a, 1]$ is the acceptance region. Again, the resulting test may be rewritten in the form of a likelihood ratio test. In fact, for testing between two simple hypotheses, the Bayes stopping rule is a sequential probability ratio test, with thresholds

$$\begin{aligned} a &= \frac{\xi(\theta_0)}{1-\xi(\theta_0)}\frac{1-\eta^a}{\eta^a} \\ b &= \frac{\xi(\theta_0)}{1-\xi(\theta_0)}\frac{1-\eta^b}{\eta^b} \end{aligned}$$

To calculate these thresholds, we may use the results from the Neyman-Pearson approach. Since the expected cost corresponding with a stopping rule $\mathbf{d}$ may be written as

$$L(\xi, \mathbf{d}) = \xi(\theta_0)(\mathbf{E}_0[\tau] + c_0\alpha(\delta)) + (1-\xi(\theta_0))(\mathbf{E}_1[\tau] + c_1\gamma(\delta))$$

and we know that the Bayes stopping rule is a SPRT, we may find the optimal values of the thresholds by minimizing this cost function over a and b. All quantities from the cost function may be calculated using the integral equations from Section 2.2.2. However, these equations require the knowledge of the probability distribution of the increments of the log-likelihood ratio.

The approach mentioned above only applies if the hypotheses are simple, and the distribution of the likelihood ratio is known. In case these conditions are not satisfied, we may still find some solutions. A constructive way of approximating the process $\rho(\xi_k)$ is the following. Define

$$\rho_n(\xi) = \min\{\rho_0(\xi), 1 + \mathbf{E}_\xi[\rho_{n-1}(T\xi)]\}$$

where $T\xi$ denotes the posterior distribution given ξ, as defined before. Then

$$\rho(\xi) = \lim_{n\to\infty} \rho_n(\xi)$$

The function $\rho_n(\xi)$ may now be calculated iteratively, although this will be of great numerical complexity.

The quantity $\rho_n(\xi)$ may be seen as the minimal expected cost, when no more than n additional data points may be processed. So, if the class of stopping rules consists of all stopping rules that are truncated at or before some point N, this procedure gives us the exact optimal stopping rule. In that case, the Bayes cost is given by

$$\rho(\xi) = \rho_N(\xi)$$

Note that after each measurement the amount of remaining data points changes, so that the stopping problem also depends on k explicitly. Consequently, the stopping regions corresponding with the Bayes stopping rule vary with k.

2.2.4 Comparison of approaches

Let us return to our farmer and his cows. After a few days he hears that the disease is no longer spreading. However, he is not yet entirely sure that his cows are not contaminated. Apart from the extra work load resulting from the tests, there is also some cost involved with each test. Therefore, he now wishes to stop testing as soon as possible. The problem therefore may be written as a sequential test between two simple hypotheses. This implies that for both the Neyman-Pearson and the Bayesian approach a sequential probability ratio test is optimal.

Let us first consider the UME stopping rule. The farmer insists on having both the false alarm probability and the miss probability not larger than 0.01. Using Wald's approximations, this gives us thresholds $(a, b) \approx (0.01, 99)$, or, using the log-likelihood ratio, $(A, B) = (-4.6, 4.6)$.

Since the distribution of the increments of the log-likelihood ratio is known, the exact values of the thresholds may be obtained from the equations given in Section 2.2.2. Using these equations, the actual thresholds may be calculated as $(A, B) = (-4.0, 4.0)$. The error probabilities that follow from using the estimates are both equal to 0.0056, so that indeed the estimates are too conservative. The average sample number is, under both hypotheses, equal to 10.5 for the estimated thresholds, and equal to 9.2 for the calculated thresholds. Using Wald's estimates for these quantities, we obtain $(-I_0 = I_1 = \frac{1}{2})$ a value of 9.2.

If our farmer instead uses the Bayesian approach, some more information is required. Again, the prior probability of contamination is supposed to be

0.2. The cost function is scaled on the sampling cost, so that each test is supposed to have a cost equal to 1. The miss cost equals the (scaled) value of his cows, which is 1000. The false alarm cost is supposed to be half of the miss cost, so that $c_1 = 500$. The optimal thresholds are computed as $(A, B) = (-4.3, 5.9)$, with a minimal cost equal to 12.8. The resulting error probabilities are $(\alpha, \gamma) = (0.0016, 0.0074)$, and the average sample numbers are given by $(\mathbf{E}_0\tau, \mathbf{E}_1\tau) = (10.1, 13.1)$.

The difference in the miss cost and the false alarm cost, together with the prior probability of the null hypothesis result in a loss of symmetry in the thresholds. Note that, from the thresholds obtained here we may calculate the thresholds ξ^a and ξ^b that are independent of the prior distribution. In this case, we have $(\xi^b, \xi^a) = (0.0108, 0.9966)$. With these thresholds, we may now calculate the optimal test for each prior distribution. For example, if $\xi(\theta_0) = 0.5$, we obtain $(A, B) = (-5.7, 4.5)$, which illustrates the greater importance that is given to misses in comparison with false alarms.

Let us compare these results with the results from the Neyman-Pearson approach. The continuation region is larger in the Bayesian case, resulting in a larger average sample number and smaller error probabilities. Apparently, the value of the cows is so large that the conservative error bounds that were made in the Neyman-Pearson approach are not conservative enough. A rough conclusion may be that, for testing between two simple hypotheses, the Bayesian and the Neyman-Pearson approach are very similar. However, the use of some additional information in the Bayesian approach allows us to give theoretically justified error bounds. Clearly, the question remains whether this additional information is accurate enough.

Chapter 3

Global Changes

During the last few decades considerable research has been done in the area of multi-parameter stochastic processes, better known as random fields in the two-parameter case. This started with Whittle [45] who used autoregressive fields to model some agricultural phenomena. He showed that the extension from one-parameter processes to multi-parameter ones is not all that trivial. The lack of ordering in a multidimensional space results in a lack of causality in the random field models.

Several researchers have further developed the theory of autoregressive models (Woods [49], Besag [5], Larimore [24]). A lot of research has been done towards the estimation and identification of autoregressive models (Balram and Moura [1], Isaksson [14], Kashyap [20], Demoment [9], Kulkarni [21], Kartikeyan and Sarkar [18], Haralick [13]).

The real application of these autoregressive models is mainly found in image processing. The models are used for coding, filtering, and restoration of images. Review papers on these subjects have been written by Jain [16], Kashyap [19] and Cohen [7].

Once a decent model has been found that describes the field with sufficient accuracy, new applications are possible. For example, automated inspection of surfaces and texture classification are some of the possibilities. Cohen, Fan and Attali [8] already used these models for the inspection of textile fabrics.

In this chapter we examine the detectability of global changes in autoregressive models, using the existing theory of change detection (Basseville and Nikiforov [2]). These changes are characterized by a change in the *structure* of the model. This implies that the parameters of the model are changing in such a way that structural properties like the correlation between two variables are changing.

In Section 3.1 we describe the two models that are used in this chapter. The tests that are used for the detection of parametric changes are given in Section 3.2. Finally, in Section 3.3 the theory is illustrated with some simulations.

3.1 Autoregressive random fields

In Chapter 1 we introduced the index set $\mathcal{G}$ on which a random field is defined. In this chapter, we assume $\mathcal{G}$ to be constructed on a two-dimensional grid, i.e., $\mathcal{G} \subset \mathbb{Z}^2$. Hence, each site t may be identified by its location on the grid. We denote $t = (k, l)$, where k and l are elements of $\mathbb{Z}$.

Throughout this chapter we assume all random fields to have zero mean and to be stationary, so that the covariance function may be written as

$$R_X(t-s) = \mathbf{E}[X_t X_s']$$

We may only have access to the field through some measurements of the form

$$Y = X + V$$

where V is a white noise with covariance $I\delta_{t,s}$. Here I is the identity matrix and $\delta_{t,s}$ is the Kronecker delta function, defined by

$$\delta_{t,s} = \begin{cases} 1 & \text{if } t = s \\ 0 & \text{otherwise} \end{cases}$$

Furthermore, all random variables are assumed to be Gaussian. Finally, we assume that all fields are scalar; X_t may only take values in $\mathbb{R}$. This last assumption is made only to simplify notations; it does not lead to a loss of generality.

Due to the discrete and finite nature of the fields, we may easily rewrite them in a vector notation. Let us denote the number of sites on the field by N. By stacking all elements on top of each other, we obtain one large column vector of length N. This operation is denoted as

$$\mathcal{X} = \operatorname{vec} X$$

The corresponding covariance function may now be rewritten as a covariance matrix $R_{\mathcal{X}}$. The probability density function for a Gaussian field with zero mean and covariance $R_{\mathcal{X}}$ is given by

$$f(\mathcal{X}) = \frac{1}{(2\pi)^{N/2}|R_{\mathcal{X}}|^{1/2}} \exp(-\frac{1}{2}\mathcal{X}' R_{\mathcal{X}}^{-1}\mathcal{X})$$

Since we are mainly interested in the global characteristics of the field, this notation will be sufficient for our purposes.

Throughout this chapter we use the model that was introduced by Whittle [45], which is basically an extension of the one-parameter autoregressive model. Therefore, we name the fields that may be described by this model *autoregressive fields.* The value of the field at a certain site may be expressed as a linear combination of the field values at its surrounding sites, together with some field noise.

$$X_t = \sum_{s \in \mathcal{G}_0} A_s X_{t+s} + W_t \tag{3.1}$$

where $\mathcal{G}_0$ is the set of indices defining the neighbourhood of a site. For example, $\mathcal{G}_0 = \{(1,0),(0,1),(-1,0),(0,-1)\}$ defines a neighbourhood as shown in Figure 3.1. We assume all values outside the grid to be zero; a field of this type is also known in the literature as a Dirichlet field.

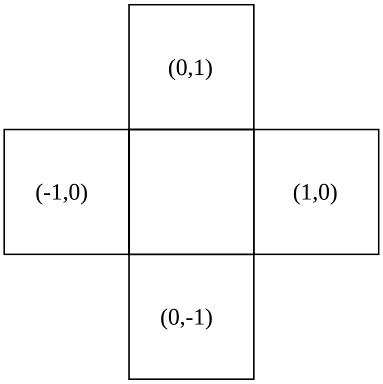

Figure 3.1: Neighbourhood $\mathcal{G}_0 = \{(1,0),(0,1),(-1,0),(0,-1)\}$

In vector form, equation (3.1) reduces to the very simple equation

$$\mathbf{L}\mathcal{X} = \mathcal{W}$$

where the large matrix $\mathbf{L}$ consists of the A_ss. Note that this matrix is square, since we assume the vectors $\mathcal{X}$ and $\mathcal{W}$ to have identical dimensions. This matrix is termed the *field matrix.* If the grid $\mathcal{G}$ is rectangular, then $\mathbf{L}$ is block Toeplitz, which may be helpful for the calculation of the inverse of this matrix.

Assumption 3.1 (Invertibility of field matrix) *The field matrix* $\mathbf{L}$ *is assumed to be invertible.*

Due to this assumption, a restriction is imposed on the parameters of the field.

The second order characteristics of the field are determined by the stochastic characteristics of the field noise. In the literature two different autoregressive fields frequently appear; the non-causal autoregressive fields, and the Gauss Markov random fields.

The *non-causal autoregressive field* (NCAR), also known as the *simultaneous autoregressive field*, is driven by a white noise. A white noise

$$R_{\mathcal{W}} = \mathbf{I}$$

results in this case in a field covariance matrix

$$\begin{aligned} R_{\mathcal{X}} &= \mathbf{E}[\mathcal{X}\mathcal{X}'] \\ &= \mathbf{L}^{-1} R_{\mathcal{W}} (\mathbf{L}')^{-1} \\ &= (\mathbf{L}'\mathbf{L})^{-1} \end{aligned}$$

and a cross-covariance between the field and the field noise

$$\begin{aligned} R_{\mathcal{X}\mathcal{W}} &= \mathbf{E}[\mathcal{X}\mathcal{W}'] \\ &= R_{\mathcal{X}}\mathbf{L}' \\ &= \mathbf{L}^{-1} \end{aligned}$$

Since this cross-covariance matrix is non-diagonal, the field $\mathcal{X}$ and the noise $\mathcal{W}$ are not independent for NCAR fields.

The *Gauss Markov random field* (GMRF) is based on the Markov property for uni-directional processes. In plain words, the Markov property states that the future is independent of the past, given the present state. An extension of this Markov property to two-parameter processes requires a new definition of past, present, and future. From equation (3.1) we may find a straightforward definition of future as the left hand term of this equation, and the present as all field values in the right hand term that have corresponding nonzero A_s. The remaining sites on the field may then be regarded as the past. The Markov property now implies that the field noise W_t should be independent of all X_s, where $s \neq t$. In vector form, this may be written as

$$R_{\mathcal{X}\mathcal{W}} = \mathbf{I}$$

This results in a covariance matrix of the field equal to

$$\begin{aligned} R_{\mathcal{X}} &= \mathbf{E}[\mathcal{X}\mathcal{X}'] \\ &= R_{\mathcal{X}\mathcal{W}}(\mathbf{L}')^{-1} \\ &= \mathbf{L}^{-1} \end{aligned}$$

and a noise covariance

$$\begin{aligned} R_{\mathcal{W}} &= \mathbf{E}[\mathcal{W}\mathcal{W}'] \\ &= \mathbf{L}R_{\mathcal{X}\mathcal{W}} \\ &= \mathbf{L} \end{aligned}$$

Note that, being an invertible covariance matrix, $\mathbf{L}$ has to be symmetric and positive definite.

3.1.1 Relationship between NCAR and GMRF models

Although there seems to be quite a difference between the NCAR models and the GMRF models, we may show that it is only a difference in representation. This is stated in the following lemma.

Lemma 3.1 *The random field X may be described by a NCAR model if and only if there exists a GMRF model that describes X.*

Proof. Suppose the field X may be described by the NCAR model

$$\mathbf{L}\mathcal{X} = \mathcal{W}$$

with $R_{\mathcal{W}} = \mathbf{I}$, $R_{\mathcal{X}} = (\mathbf{L}'\mathbf{L})^{-1}$ and $R_{\mathcal{X}\mathcal{W}} = \mathbf{L}^{-1}$. It is not difficult to see that by premultiplying this equation with $\mathbf{L}'$, the transpose of $\mathbf{L}$, we obtain a GMRF representation of the field.

$$\mathbf{K}\mathcal{X} = \mathcal{V}$$

where $\mathbf{K} = \mathbf{L}'\mathbf{L}$ and $\mathcal{V} = \mathbf{L}'\mathcal{W}$.

Alternatively, suppose X may be described by a GMRF model

$$\mathbf{L}\mathcal{X} = \mathcal{W}$$

with $R_{\mathcal{W}} = \mathbf{L}$, $R_{\mathcal{X}} = \mathbf{L}^{-1}$ and $R_{\mathcal{X}\mathcal{W}} = \mathbf{I}$. Since $\mathbf{L}$ is positive definite, we may find an invertible matrix $\mathbf{K}$ such that $\mathbf{L} = \mathbf{K}'\mathbf{K}$. The field X may now be described by the NCAR model

$$\mathbf{K}\mathcal{X} = \mathcal{V}$$

where $\mathcal{V} = (\mathbf{K}')^{-1}\mathcal{W}$. Indeed, the noise covariance

$$\begin{aligned} R_{\mathcal{V}} &= (\mathbf{K}')^{-1} R_{\mathcal{W}} \mathbf{K}^{-1} \\ &= (\mathbf{K}')^{-1} \mathbf{K}'\mathbf{K}\mathbf{K}^{-1} \\ &= \mathbf{I} \end{aligned}$$

equals the identity, which implies that the field noise is white. □

In this paper we only compare NCAR models and GMRF models that are described by the same matrix $\mathbf{L}$. This implies that the actual fields we are looking at are structurally different.

3.1.2 Field sequences

In some applications we may have to deal with a series of samples taken from a certain field. For example, a satellite circling around the earth may take a picture of a certain region once per cycle. Clearly, instead of considering only the present measurement it may be better to use the outcomes of the previous measurements as well. Therefore, we introduce the notion of field sequences.

Definition 3.1 (Field sequence) *A field sequence is defined as a mapping from $\mathcal{T} \subset \mathbb{Z}$ to $\ell_2(\mathcal{G})$.*

At each moment in time, which is assumed to be discrete, a measurement is taken from the entire field. We indicate the presence of a field sequence by using extra parentheses as in $X(\cdot)$, and we use $X(k)$ to denote the field at time k. We assume that all fields are independent in time. If this assumption does not hold, we may use some transformation (like a Kalman filter) to obtain independence.

Through the introduction of field sequences, the resulting process has the same characteristics as a standard stochastic process in discrete time. In the following sections we will see how this similarity may be used for the detection of changes in a field sequence.

3.2 Parametric changes

Our farmer has been growing pumpkins on part of his fields for many years now. Throughout the years, he has been collecting data on his yields. This way, he discovered that the weight of the pumpkins may be described by an autoregressive field. In fact, the weight of a pumpkin is negatively influenced by the weights of the surrounding pumpkins. This year, the farmer is trying to alter this relation by adding manure to the field. At harvesting time, he wants to find out if his actions were successful. In other words, he wants to test for a global parametric change in his field.

Suppose that under normal circumstances the field may be described by

$$\mathbf{L}\mathcal{X} = \mathcal{W}$$

Then, after the occurrence of a change in the parameters, the new model describing the field is

$$(\mathbf{L} + \boldsymbol{\Delta})\mathcal{X} = \mathcal{W}$$

where the form of $\boldsymbol{\Delta}$ follows from the basic autoregressive equation (3.1). The exact value of $\boldsymbol{\Delta}$ is unknown. The p-dimensional vector containing the changes in the original A_s parameters is denoted by $\theta = (\theta_1, \ldots, \theta_p)$. Note that θ_i now denotes the ith component of the vector θ.

Because of the introduction of a time-axis $\mathcal{T}$, it may be possible to identify a change time $k_c \in \mathcal{T}$, as in Chapter 2. The detection problem may then be written as the classical test between the hypotheses

$$\begin{aligned}
\mathcal{H}_0 &: \mathcal{Y}(k) \sim N(0, R_0) \text{ for all } k \in \mathcal{T} \\
\mathcal{H}_1 &: \exists k_c \in \mathcal{T} \to \begin{cases} \mathcal{Y}(k) \sim N(0, R_0) \text{ for } k < k_c \\ \mathcal{Y}(k) \sim N(0, R_1) \text{ for } k \geq k_c \end{cases}
\end{aligned}$$

where $\mathcal{Y}$ is the (possibly exact) measurement of the field $\mathcal{X}$. The covariance matrices R_0 and R_1 depend on the characteristics of the field and the measurement noise. If $\mathcal{T}$ only consists of one point in time, the detection problem reduces to a classification problem.

Because of the unknown change θ and the unknown change time k_c, there is no known UME test. Therefore, the generalized likelihood ratio (GLR) seems to be the most appropriate test statistic. The GLR may in this case be written as

$$\begin{aligned}
T_k(\mathcal{Y}(1), \ldots, \mathcal{Y}(k)) &= 2 \sup_{k_c, \theta} \sum_{i=k_c}^{k} \log \frac{f_\theta(\mathcal{Y}(i))}{f_0(\mathcal{Y}(i))} \\
&= \sup_{k_c, \theta} \sum_{i=k_c}^{k} \log \frac{|R_0|}{|R_1|} - \mathcal{Y}(i)'(R_1^{-1} - R_0^{-1})\mathcal{Y}(i)
\end{aligned}$$

Unfortunately, it is not possible to use the CUSUM test here. This is caused by the fact that the estimate of $\boldsymbol{\theta}$ varies with k. Hence, we may not use a recursive algorithm to calculate the generalized likelihood ratio. Obviously, if Θ only contains one element, we do not have this problem, and the GLR is equivalent to the CUSUM test.

To find the maximum likelihood estimate, we have to calculate the most likely parameter θ for all possible values of k_c, and choose the combination that maximizes the GLR. For a given k_c, we may use some numerical method to find the optimal value of θ. In general, this will be computationally expensive.

Let us first focus on the classification problem. From Theorem 2.1 we know that for small $\|\theta\|$, the GLR may be approximated by the score test

$$S(\mathcal{Y}) = \mathcal{Z}_0(\mathcal{Y})'\mathbf{\Gamma}(0)^{-1}\mathcal{Z}_0(\mathcal{Y})$$

The computation of the decision function S does not require any maximization. The ith component of the efficient score may be calculated according to

$$\begin{aligned}
\{\mathcal{Z}_0(\mathcal{Y})\}_i &= \{\frac{\partial}{\partial \theta_i} \log f_\theta(\mathcal{Y})\}_{\theta=0} \\
&= -\frac{1}{2} \frac{\partial}{\partial \theta_i} [\log |R(\theta)| + \mathcal{Y}' R(\theta)^{-1} \mathcal{Y}]_{\theta=0}
\end{aligned}$$

$$\begin{aligned}
&= -\frac{1}{2}[\mathrm{tr}\, R(0)^{-1}\left(\frac{\partial R(\theta)}{\partial\theta_i}\right)_{\theta=0} + \mathcal{Y}'\left(\frac{\partial R(\theta)^{-1}}{\partial\theta_i}\right)_{\theta=0}\mathcal{Y}] \\
&= \frac{1}{2}\mathrm{tr}\,(R_0 - \mathcal{Y}\mathcal{Y}')D_i
\end{aligned}$$

where

$$D_i = \left(\frac{\partial R(\theta)^{-1}}{\partial\theta_i}\right)_{\theta=0} \tag{3.2}$$

The derivative appearing in the expression for D_i given above depends on the model we are using. The calculation of D_i is straightforward for both of the NCAR and the GMRF models. Note that the inverse of the covariance matrix is quadratic in θ_i for the NCAR model, and linear in θ_i for the GMRF model.

Similarly, the Fisher information may be computed as

$$\begin{aligned}
\Gamma_{ij}(0) &= \mathbf{E}_0[\mathcal{Z}_0(\mathcal{Y})\mathcal{Z}_0(\mathcal{Y})']_{ij} \\
&= \mathbf{E}_0[-\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f_\theta(\mathcal{Y})]_{\theta=0} \\
&= \mathbf{E}_0[-\frac{1}{2}\mathrm{tr}\,\{\left(\frac{\partial^2 R(\theta)^{-1}}{\partial\theta_i\partial\theta_j}\right)_{\theta=0}[R_0 - \mathcal{Y}\mathcal{Y}'] \\
&\quad + \left(\frac{\partial R(\theta)}{\partial\theta_j}\right)_{\theta=0}\left(\frac{\partial R(\theta)^{-1}}{\partial\theta_i}\right)_{\theta=0}\}] \\
&= -\frac{1}{2}\mathrm{tr}\,[\left(\frac{\partial R(\theta)}{\partial\theta_j}\right)_{\theta=0}\left(\frac{\partial R(\theta)^{-1}}{\partial\theta_i}\right)_{\theta=0}] \\
&= \frac{1}{2}\mathrm{tr}\,[R_0 D_j R_0 D_i]
\end{aligned}$$

where D_i and D_j are defined by (3.2).

In case $\mathcal{T}$ consists of more than one element, time gets involved in the problem. Hence, not only do we have to detect the change θ, also the change time k_c plays an important role. Due to the presence of a change time we may split the time axis $\mathcal{T}$ into two parts. During the first part, when $k < k_c$, the nominal model holds and during the second part a change has appeared. This implies that we have to split the data in our test statistic as well.

The score test statistic may be generalized in several ways. One option is to try to find an equivalent of the CUSUM test. Since the score test does not require an estimation of the parameter $\boldsymbol{\theta}$, it may very well be possible to find a recursive generalization. The most suitable choice then would be to sum the quantities $S(\mathcal{Y}(i)) - 1$ over $i = 1, \ldots, k$. This gives us the possibility to calculate the optimal change time recursively as in the CUSUM test. However, the variance of this statistic also grows with k. For that reason, we choose two other alternatives.

In analogy with Basseville and Nikiforov [2, p.146], we define the efficient score over the sequence $\mathcal{Y}(l..k) \triangleq \{\mathcal{Y}(l), \ldots, \mathcal{Y}(k)\}$ as

$$\mathcal{Z}_\theta(\mathcal{Y}(l..k)) = \frac{1}{\sqrt{k-l+1}} \sum_{i=l}^{k} \mathcal{Z}_\theta(\mathcal{Y}(i))$$

where $\mathcal{Z}_\theta(\mathcal{Y}(i))$ denotes the efficient score of the field at time i. Under the condition that θ is the correct parameter, it has zero mean and covariance

$$\begin{aligned}
\mathbf{E}_\theta[\mathcal{Z}_\theta(\mathcal{Y}(l..k))\mathcal{Z}_\theta(\mathcal{Y}(l..k))'] &= \frac{1}{k-l+1} \sum_{i=l}^{k} \sum_{j=l}^{k} \mathbf{E}_\theta[\mathcal{Z}_\theta(\mathcal{Y}(i))\mathcal{Z}_\theta(\mathcal{Y}(j))'] \\
&= \frac{1}{k-l+1} \sum_{i=l}^{k} \Gamma(\theta) \\
&= \Gamma(\theta)
\end{aligned}$$

Now, remembering the fact that the sequence is split into two parts, we come up with the following possibilities for the sequential score test:

$$\begin{aligned}
\tilde{S}_k(\mathcal{Y}(1..k)) &= \sup_{k_c} \mathcal{Z}_0(\mathcal{Y}(k_c..k))'\Gamma(0)^{-1}\mathcal{Z}_0(\mathcal{Y}(k_c..k)) \\
\hat{S}_k(\mathcal{Y}(1..k)) &= \mathcal{Z}_0(\mathcal{Y}(k-l..k))'\Gamma(0)^{-1}\mathcal{Z}_0(\mathcal{Y}(k-l..k))
\end{aligned}$$

The first statistic searches for the optimal change time, whereas the second statistic simply uses the $l+1$ latest measurements, i.e., it uses a *window* of constant length $l+1$. Clearly, the first statistic is computationally more expensive. The problem with the second statistic is to find a reasonable window length. Here we have to deal with a tradeoff between the delay of detection and the error probabilities. If we increase the window length, the error probabilities will decrease; however, the delay of detection will increase.

For the same reason that the window length should not be chosen too small, we should also impose a minimum delay of detection in the first statistic. It is not at all unlikely that one particular sample, say the kth one, may be regarded as an out-lier, resulting in

$$\mathcal{Z}_0(\mathcal{Y}(k))'\Gamma(0)^{-1}\mathcal{Z}_0(\mathcal{Y}(k)) > \lambda$$

Then obviously, $\tilde{S}_k$ will be larger than the threshold λ as well, leading to a rejection of the null hypothesis. By imposing a minimum delay of detection, this will be avoided.

3.2.1 Detectability

To indicate the quality of detection that may be expected, the notion of detectability was introduced in Chapter 2. A change is said to be detectable if and only if the Kullback-Leibler information is strictly positive. For the situation described in the previous sections, the Kullback-Leibler information may be calculated rather easily. For the classification problem where $\mathcal{T}$ consists of one element, the Kullback-Leibler information may be calculated as

$$I(\theta) = \frac{1}{2}(\log \frac{|R_0|}{|R_1|} + \operatorname{tr}(R_1 R_0^{-1} - \mathbf{I}))$$

It follows that a change θ is detectable if and only if $R_1 = R(\theta)$ is not equal to $R_0 = R(0)$.

In case $\mathcal{T}$ consists of more than one element, a change θ at time k_c has a Kullback-Leibler information equal to $(k - k_c + 1)I(\theta)$ at time k, which follows easily from the definition of the Kullback-Leibler information. This implies that it grows linearly in time.

Since the GLR is defined as a function of the log-likelihood ratio, there exists a natural connection between the GLR and the Kullback-Leibler information. Indeed, given that $\boldsymbol{\theta} = \theta$, we may find the following lower bound for the expectation of the GLR for an individual sample of the field.

$$\begin{aligned} \mathbf{E}_\theta[T(\mathcal{Y})] &= \mathbf{E}_\theta[2 \sup_\vartheta \log \frac{f_\vartheta(\mathcal{Y})}{f_0(\mathcal{Y})}] \\ &\geq 2 \sup_\vartheta \mathbf{E}_\theta[\log \frac{f_\vartheta(\mathcal{Y})}{f_0(\mathcal{Y})}] \\ &= 2I(\theta) \end{aligned}$$

where the last equality follows from the fact that (Kullback [22])

$$\mathbf{E}_1[\log \frac{f_1(\mathcal{Y})}{f_0(\mathcal{Y})}] \geq \mathbf{E}_1[\log \frac{f_2(\mathcal{Y})}{f_0(\mathcal{Y})}]$$

so that $\vartheta = \theta$ maximizes the expectation. If $\boldsymbol{\theta} = 0$, i.e., if no change is present in the sample, this lower bound equals 0.

A lower bound for the expectation of the GLR for a field sequence, given that $\boldsymbol{\theta} = \theta$ and k_c is the change time, may similarly be found as

$$\mathbf{E}_{k_c,\theta}[T_k(\mathcal{Y}(1..k))] \geq 2(k - k_c + 1)I(\theta) \tag{3.3}$$

Suppose now that we are using the threshold λ in our GLR test. We now introduce the *expected detection time* as the first time for which the lower bound (3.3) crosses the rejection boundary. By solving the equation $2(k - k_c + 1)I(\theta) = \lambda$, we obtain the expected detection time

$$E_T[k_d(\theta, k_c)] = k_c - 1 + \frac{\lambda}{2I(\theta)}$$

Note that this quantity is not the same as the expectation of the detection time. In fact, since it uses a lower bound of the expectation of the generalized likelihood ratio, this expected detection time will be larger than the expectation of the detection time.

The *mean delay of detection*, defined as the difference between the expected detection time and the change time, may now easily be found as

$$\bar{\tau}_T(\theta) = \frac{\lambda}{2I(\theta)} - 1 \tag{3.4}$$

It follows that the mean delay of detection is inversely proportional to the Kullback-Leibler information.

Similar expressions may be found for our score test statistics. To do this, we need to know the expectation and the covariance matrix of the efficient score under the alternative hypothesis.

Theorem 3.1 *Given that $\boldsymbol{\theta} = \theta$, the mean $\mu(\theta)$ and covariance matrix $\bar{\Gamma}(\theta)$ of the efficient score are given by*

$$\begin{aligned} \mu_i(\theta) &= \mathbf{E}_\theta[\{\mathcal{Z}_0(\mathcal{Y})\}_i] \\ &= \frac{1}{2} tr\,(R_0 - R_1)D_i \end{aligned}$$

and

$$\begin{aligned} \bar{\Gamma}_{ij}(\theta) &= \mathbf{E}_\theta[(\{\mathcal{Z}_0(\mathcal{Y})\}_i - \mu_i(\theta))(\{\mathcal{Z}_0(\mathcal{Y})\}_j - \mu_j(\theta))] \\ &= \frac{1}{2} tr\,R_1 D_i R_1 D_j \end{aligned}$$

Proof. First we note that

$$\begin{aligned} \{\mathcal{Z}_0(\mathcal{Y})\}_i &= \frac{1}{2}\mathrm{tr}\,(R_0 - \mathcal{Y}\mathcal{Y}')D_i \\ &= \frac{1}{2}\mathrm{tr}\,(R_1 - \mathcal{Y}\mathcal{Y}')D_i + \frac{1}{2}\mathrm{tr}\,(R_0 - R_1)D_i \\ &= \bar{\mathcal{Z}}_i(\mathcal{Y}) + \mu_i(\theta) \end{aligned}$$

where

$$\bar{\mathcal{Z}}_i(\mathcal{Y}) = \frac{1}{2}\mathrm{tr}\,(R_1 - \mathcal{Y}\mathcal{Y}')D_i$$

and

$$\begin{aligned} \mu_i(\theta) &= \mathbf{E}_\theta[\{\mathcal{Z}_0(\mathcal{Y})\}_i] \\ &= \frac{1}{2}\mathrm{tr}\,(R_0 - R_1)D_i \end{aligned}$$

This implies that

$$\begin{aligned}\mathbf{E}_\theta[\bar{\mathcal{Z}}_i(\mathcal{Y})] &= 0\\ \bar{\Gamma}_{ij}(\theta) &= \mathbf{E}_\theta[\bar{\mathcal{Z}}_i(\mathcal{Y})\bar{\mathcal{Z}}_j'(\mathcal{Y})]\\ &= \frac{1}{4}\mathbf{E}_\theta[\mathrm{tr}\,(R_1-\mathcal{Y}\mathcal{Y}')D_i\mathrm{tr}\,(R_1-\mathcal{Y}\mathcal{Y}')D_j]\\ &= \frac{1}{4}\mathbf{E}_\theta[\{\mathcal{Y}'D_i\mathcal{Y}-\mathbf{E}_\theta[\mathcal{Y}'D_i\mathcal{Y}]\}\{\mathcal{Y}'D_j\mathcal{Y}-\mathbf{E}_\theta[\mathcal{Y}'D_j\mathcal{Y}]\}]\end{aligned}$$

The quadratic forms in this expression may be simplified by using the transformations

$$x = T'R^{-1/2}\mathcal{Y}$$

and

$$y = S'R^{-1/2}\mathcal{Y}$$

where the orthogonal matrices S and T are such that

$$T'R^{1/2}D_iR^{1/2}T = \Lambda_i$$

and

$$S'R^{1/2}D_jR^{1/2}S = \Lambda_j$$

where Λ_i and Λ_j are diagonal matrices. This is always possible since D_i and D_j are symmetric, so that their eigenvalues are real. The quadratic forms now reduce to

$$\mathcal{Y}'D_i\mathcal{Y} = x'\Lambda_i x = \sum_{k=1}^{n}\lambda_{ik}x_k^2$$

and

$$\mathcal{Y}'D_j\mathcal{Y} = y'\Lambda_j y = \sum_{l=1}^{n}\lambda_{jl}y_l^2$$

This implies that

$$\begin{aligned}\bar{\Gamma}_{ij}(\theta) &= \frac{1}{4}\mathbf{E}_\theta[\sum_{k=1}^{n}\lambda_{ik}(x_k^2-\mathbf{E}_\theta[x_k^2])\sum_{l=1}^{n}\lambda_{jl}(y_l^2-\mathbf{E}_\theta[y_l^2])]]\\ &= \frac{1}{4}\sum_{k=1}^{n}\sum_{l=1}^{n}\lambda_{ik}\lambda_{jl}\mathbf{E}_\theta[(x_k^2-1)(y_l^2-1)]\\ &= \frac{1}{4}\sum_{k=1}^{n}\sum_{l=1}^{n}\lambda_{ik}\lambda_{jl}\{\mathbf{E}_\theta[x_k^2y_l^2]-1\}\end{aligned}$$

Denote the columns of the matrices T and S as T_k and S_l, respectively. Then we may write

$$\begin{aligned}x_k &= T_k'R^{-1/2}\mathcal{Y}\\ y_l &= S_l'R^{-1/2}\mathcal{Y}\end{aligned}$$

These variables are jointly Gaussian distributed with zero mean and covariance matrix

$$\mathbf{E}\begin{pmatrix} x_k \\ y_l \end{pmatrix} \begin{pmatrix} x_k & y_l \end{pmatrix} = \begin{pmatrix} 1 & T_k' S_l \\ T_k' S_l & 1 \end{pmatrix}$$

For zero mean Gaussian variables $w_i, i = 1, \ldots, 4$, it is well known that

$$\mathbf{E}[w_1 w_2 w_3 w_4] = \sigma_{12}\sigma_{34} + \sigma_{13}\sigma_{24} + \sigma_{14}\sigma_{23}$$

where $\sigma_{ij} = \mathbf{E}[w_i w_j]$. This implies that the remaining expectation may be evaluated according to

$$\begin{aligned} \mathbf{E}_\theta[x_k^2 y_l^2] &= \mathbf{E}_\theta[x_k^2]\mathbf{E}_\theta[y_l^2] + 2\mathbf{E}_\theta[x_k y_l]\mathbf{E}_\theta[x_k y_l] \\ &= 1 + 2(T_k' S_l)^2 \end{aligned}$$

Finally, this gives us

$$\begin{aligned} \bar{\Gamma}_{ij}(\theta) &= \frac{1}{4}\sum_{k=1}^{n}\sum_{l=1}^{n} \lambda_{ik}\lambda_{jl} 2 T_k' S_l S_l' T_k \\ &= \frac{1}{2}\sum_{k=1}^{n}\sum_{l=1}^{n} \lambda_{ik}(T'S)_{kl}\lambda_{jl}(S'T)_{lk} \\ &= \frac{1}{2}\mathrm{tr}\, T\Lambda_i T' S \Lambda_j S' \\ &= \frac{1}{2}\mathrm{tr}\, R_1 D_i R_1 D_j \end{aligned}$$

□

From this theorem, we may find

$$\begin{aligned} \mathbf{E}_\theta[\mathcal{Z}_0(\mathcal{Y})\mathcal{Z}_0(\mathcal{Y})'] &= \mathbf{E}_\theta[(\bar{\mathcal{Z}} + \mu(\theta))(\bar{\mathcal{Z}} + \mu(\theta))'] \\ &= \bar{\Gamma}(\theta) + \mu(\theta)\mu(\theta)' \end{aligned}$$

We now define an equivalent of the Kullback-Leibler information for the score test statistic $\tilde{S}$ by

$$\begin{aligned} I_k^{\tilde{S}}(\theta, t_c) &= \mathbf{E}_{\theta,k_c}[\mathcal{Z}_0(\mathcal{Y}(k_c..k))'\Gamma(0)^{-1}\mathcal{Z}_0(\mathcal{Y}(k_c..k))] \\ &= \frac{1}{k-k_c+1}\mathbf{E}_\theta[\sum_{i=k_c}^{k} \mathcal{Z}_0(\mathcal{Y}(i))'\Gamma(0)^{-1} \sum_{i=k_c}^{k} \mathcal{Z}_0(\mathcal{Y}(i))] \\ &= \frac{1}{k-k_c+1}\{\sum_{i=k_c}^{k} \mathbf{E}_\theta[\mathcal{Z}_0(\mathcal{Y}(i))\Gamma(0)^{-1}\mathcal{Z}_0(\mathcal{Y}(i))] + \\ &\quad \sum_{i=k_c}^{k}\sum_{j\neq i} \mathbf{E}_\theta[\mathcal{Z}_0(\mathcal{Y}(i))'\Gamma(0)^{-1}\mathcal{Z}_0(\mathcal{Y}(j))]\} \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{k-k_c+1}\{(k-k_c+1)\mathrm{tr}\,\mathbf{E}_\theta[\mathcal{Z}_0(\mathcal{Y})\mathcal{Z}_0(\mathcal{Y})']\Gamma(0)^{-1} + \\
&\quad (k-k_c+1)(k-k_c)\mu'\Gamma(0)^{-1}\mu \\
&= \mathrm{tr}\,\bar{\Gamma}(\theta)\Gamma(0)^{-1} + (k-k_c+1)\mu'\Gamma(0)^{-1}\mu
\end{aligned}$$

Note that this expectation is defined conditional on the assumption that the change time is estimated correctly.

Similarly, we may find an equivalent of the Kullback-Leibler information for the statistic $\hat{S}$. However, in this case we have to distinguish two situations. If the window only contains changed samples ($k_c \leq k-l$) we may use the previous result to get

$$\begin{aligned}
I_k^{\hat{S}}(\theta, k_c) &= \mathbf{E}_{\theta,k_c}[\hat{S}_k(\mathcal{Y}(k-l..k))] \\
&= \mathrm{tr}\,\bar{\Gamma}(\theta)\Gamma(0)^{-1} + (l+1)\mu'\Gamma(0)^{-1}\mu \qquad (3.5)
\end{aligned}$$

If the change occurs inside the window ($k-l < k_c \leq k$), we get

$$\begin{aligned}
I_k^{\hat{S}}(\theta, k_c) &= \mathbf{E}_{\theta,k_c}[\hat{S}_k(\mathcal{Y}(k-l..k))] \\
&= \frac{1}{l+1}\mathbf{E}_{\theta,k_c}[\sum_{i=k-l}^{k}\mathcal{Z}_0(\mathcal{Y}(i))'\Gamma(0)^{-1}\sum_{i=k-l}^{k}\mathcal{Z}_0(\mathcal{Y}(i))] \\
&= \frac{1}{l+1}\{\mathbf{E}_0[\sum_{i=k-l}^{k_c-1}\mathcal{Z}_0(\mathcal{Y}(i))'\Gamma(0)^{-1}\sum_{i=k-l}^{k_c-1}\mathcal{Z}_0(\mathcal{Y}(i))] \\
&\quad +\mathbf{E}_\theta[\sum_{i=k_c}^{k}\mathcal{Z}_0(\mathcal{Y}(i))'\Gamma(0)^{-1}\sum_{i=k_c}^{k}\mathcal{Z}_0(\mathcal{Y}(i))]\} \\
&= \frac{1}{l+1}\{(k_c-k+l)p + (k-k_c+1)\mathrm{tr}\,\bar{\Gamma}(\theta)\Gamma(0)^{-1} \\
&\quad +(k-k_c+1)^2\mu'\Gamma(0)^{-1}\mu\} \\
&= p + \frac{1}{l+1}[(k-k_c+1)(\mathrm{tr}\,\bar{\Gamma}(\theta)\Gamma(0)^{-1} - p) \\
&\quad +(k-k_c+1)^2\mu'\Gamma(0)^{-1}\mu] \qquad (3.6)
\end{aligned}$$

The mean delay of detection for the first statistic, again ignoring the maximization, may now be computed as

$$\bar{\tau}_{\tilde{S}}(\theta) = \frac{\lambda - \mathrm{tr}\,\bar{\Gamma}(\theta)\Gamma(0)^{-1}}{\mu'\Gamma(0)^{-1}\mu} - 1 \qquad (3.7)$$

For the second statistic, we may now find a lower bound on the window length:

$$l+1 \geq \frac{\lambda - \mathrm{tr}\,\bar{\Gamma}(\theta)\Gamma(0)^{-1}}{\mu'\Gamma(0)^{-1}\mu} \qquad (3.8)$$

Note that this bound does not guarantee that the change θ will be detected; it merely says that on average it will be detected. As is well known for χ^2-distributed variables, the variance of the statistics we use is large. This implies that these results may not be that useful in practice.

The mean delay of detection given in (3.4) and (3.7) may be used as upper bounds to the mathematical expectation of the delay of detection. This follows from the fact that the Kullback-Leibler information numbers that are used underestimate the true expectation of the test statistics, since the maximizations are ignored. For the same reason, the lower bound (3.8) is rather conservative. Even for smaller values of the window length, the detection quality may be sufficient. Initially, the number of samples that are processed will be smaller than the window length. This implies that a decision at such an early stage may not be accurate. However, since the bound is rather conservative, it may not be necessary to wait until $l+1$ samples have been processed before rejecting the null hypothesis.

3.3 Nearest neighbour model

To illustrate some of the previous results we restrict our attention to one of the most basic autoregressive fields. The simultaneous nearest-neighbour model ([28],[1]) is defined by

$$X_t + h(X_{t+t_h} + X_{t-t_h}) + v(X_{t+t_v} + X_{t-t_v}) = W_t$$

where $t_h = (1,0)$ is the horizontal shift and $t_v = (0,1)$ is the vertical shift. The grid $\mathcal{G}$ is assumed to be square, of size $M \times M$. The driving noise W_t is assumed to be zero-mean Gaussian with a variance equal to one. Depending on the cross-covariance between W_t and X_s this equation defines a NCAR or GMRF model. The measurements are given by the vector $\mathbf{y} \in \mathbb{R}^{M^2}$. These measurements may contain additive Gaussian white noise. In case we are dealing with a field sequence, the measurements are denoted by $\mathbf{y}(k)$, where $k = 1, 2, \ldots$ may denote time.

In the following simulations we examine the detectability of a change in the parameters h and v. Such a change represents a change in the structure of the field; the relation between neighbouring sites on the field becomes stronger or weaker. The changes in the parameter are denoted by

$$\theta = \begin{pmatrix} \theta_h & \theta_v \end{pmatrix}$$

The classification problem may be defined as the test between hypotheses

$$\begin{aligned} \mathcal{H}_0 &: \theta = (0,0) \\ \mathcal{H}_1 &: \theta \neq (0,0) \end{aligned}$$

In case we have access to a field sequence, the change may appear at a certain point in time. For example, if the field sequence describes a series of satellite images of a forest, a change in the structure may be the result of a large fire. In this case, the hypotheses may be written as

$$\begin{aligned}\mathcal{H}_0 &: \text{for all } k, \text{ we have } \theta = (0,0)\\ \mathcal{H}_1 &: \text{there exists a } k_c \geq 1,\\ & \quad \text{such that } \theta = (0,0) \text{ for } k < k_c \text{ and } \theta \neq (0,0) \text{ for } k \geq k_c\end{aligned}$$

The field matrix $\mathbf{L}$ corresponding to the nearest neighbour model has the form

$$\mathbf{L} = \begin{pmatrix} L_0 & L_1 & 0 & \cdots & \cdots & \cdots & 0 \\ L_1 & L_0 & L_1 & 0 & \cdots & \cdots & \vdots \\ 0 & L_1 & L_0 & L_1 & 0 & \cdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \cdots & 0 & L_1 & L_0 & L_1 & 0 \\ \vdots & \cdots & \cdots & 0 & L_1 & L_0 & L_1 \\ 0 & \cdots & \cdots & \cdots & 0 & L_1 & L_0 \end{pmatrix}$$

where L_0 is a tridiagonal matrix with 1 on the diagonal and v on the first upper and lower diagonals, and L_1 is a diagonal matrix with h as diagonal elements. The matrix $\mathbf{\Delta}$ may be obtained from $\mathbf{L}$ by replacing every h, v and 1 by θ_h, θ_v and 0, respectively. The assumption that $\mathbf{L}$ has to be invertible results in the following conditions for the parameters

$$\begin{aligned} v^2 &\neq 1 \\ h^2 &\neq 1 \\ h^2 + v^2 &\neq 1 \end{aligned}$$

For the GMRF model we need the additional assumption that $\mathbf{L}$ is symmetric and positive definite. This gives us the additional constraint (Balram and Moura [1])

$$|v| + |h| < \frac{1}{2\cos\frac{\pi}{M+1}}$$

Obviously, the changes θ_h and θ_v should be such that the new parameters $h + \theta_h$ and $v + \theta_v$ still satisfy these constraints.

Simulation 3.1 *We choose the parameters $h = -0.1$ and $v = -0.2$, satisfying all constraints. The grid $\mathcal{G}$ is chosen to have a size 16×16. Using a 95 percent confidence interval, we obtain a rejection boundary of 5.99 (χ^2-test with two degrees of freedom).*

θ_h	θ_v	S	T
0	0	25	25
0.01	0.01	0	0
0.05	0.05	10	15
0.10	0.10	25	25

Table 3.1: Number of correct decisions (out of 25). NCAR without measurement noise.

θ_h	θ_v	S	T
0	0	22	23
0.01	0.01	3	2
0.05	0.05	3	7
0.10	0.10	13	16

Table 3.2: Number of correct decisions (out of 25). NCAR with measurement noise.

θ_h	θ_v	S	T
0	0	21	22
0.01	0.01	0	0
0.05	0.05	5	5
0.10	0.10	16	19

Table 3.3: Number of correct decisions (out of 25). GMRF without measurement noise.

θ_h	θ_v	S	T
0	0	25	25
0.01	0.01	0	1
0.05	0.05	0	0
0.10	0.10	3	5

Table 3.4: Number of correct decisions (out of 25). GMRF with measurement noise.

The classification problem is considered first. For each field, 25 samples without a change, 25 samples with a small change, 25 samples with an intermediate change and 25 samples with a large change are generated. Using both the score test (S) and the GLR test (T), the number of correct classifications is shown in Tables 3.1–3.4. From these tables we may see that the GLR test performs slightly better than the score test; however both tests are incapable of detecting small changes.

From the tables it is clear that adding measurement noise causes the number of correct decisions to decrease rapidly. Note that the intensity of the measurement noise is relatively large, since the driving noise and the measurement noise have the same variance. If we compare the results for the NCAR models with the results from the GMRF models, it seems that the NCAR models give better results. However, a comparison between these models is not really fair. From Lemma 3.1 we know that any NCAR model has an equivalent GMRF model. In this case, we use the same parameterization for both models, so that these models are not equivalent; they describe different random fields. Hence, we may not expect the classification results to be comparable.

To see if these small changes may be detected if the amount of data increases, we generate field sequences. In this case we use the three statistics $\hat{S}_k$, $\tilde{S}_k$ *and* T_k*. Figure 3.2 shows the value of the test statistics as a function of the sample number, for the first 25 samples. In all situations,* $k_c = 5$ *and*

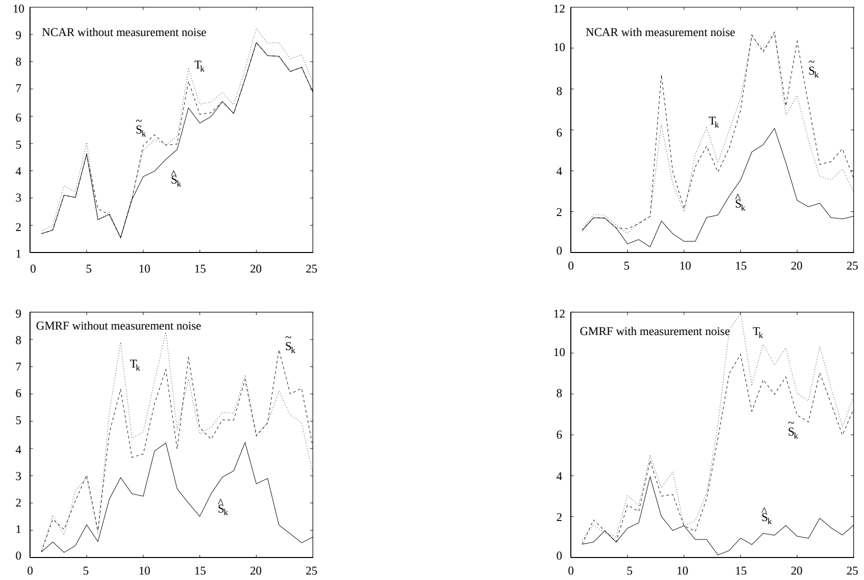

Figure 3.2: Sequential test results. Values of the test statistics as a function of k. Change $\theta = (0.01, 0.01)$ appears at $k = 5$.

$(\theta_h, \theta_v) = (0.01, 0.01)$. *The Kullback-Leibler information for these changes, listed in Table 3.5, is quite small. Note the difference between the Kullback-Leibler information for the NCAR and the GMRF models. The expected detection times are given in Table 3.5. The mean delay of detection follows by subtracting 5 from each of these numbers.*

	NCAR-	*NCAR+*	*GMRF-*	*GMRF+*
$I(\theta)$	*0.1726*	*0.0912*	*0.0438*	*0.0171*
$E_{\hat{S}} k_d$	*20*	*32*	*56*	*133*
$E_{\tilde{S}} k_d$	*17*	*29*	*53*	*130*
$E_T k_d$	*22*	*37*	*73*	*180*
$l + 1$	*13*	*25*	*49*	*126*

Table 3.5: Kullback-Leibler information, expected detection times, and lower bound for window length for small change ($\theta = (0.01, 0.01)$).

Note that we use a window of infinite length for the score test $\hat{S}_k$*, i.e., all samples are included. This does not cause any major problems due to the small change time. This implies that the change time will always be contained in the window, so that the Kullback-Leibler equivalent (3.6) may be used to*

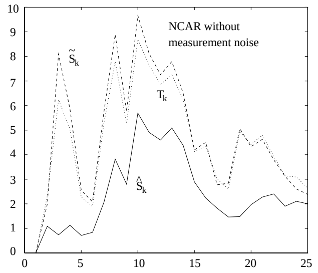

Figure 3.3: Sequential testing without the presence of a change.

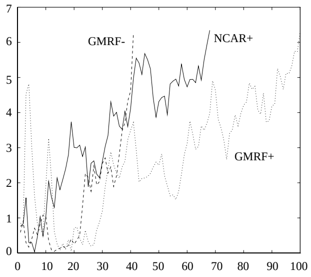

Figure 3.4: $\hat{S}_k$ with the presence of a change.

calculate an expected detection time for this particular case.

As we may see from Figure 3.2, both T_k and $\tilde{S}_k$ reject the null hypothesis much earlier than the expected detection times. To check the quality of these decisions, we simulate a field sequence where no changes occur. Figure 3.3 shows that both statistics give rise to several false alarms, whereas $\hat{S}_k$ remains inside the acceptance region. This may be explained by the fact that no minimum bound on the delay of detection was used. If the kth sample deviates only slightly from the original model, the estimated change time will be k, and the statistic will attain a large value, possibly resulting in a false alarm. This implies that to obtain better results with these statistics, a minimum bound on the delay of detection should be used. This implies that we use averaging, so that possible outliers do not automatically lead to false alarms.

As we may see from Figure 3.4, the null hypothesis eventually is rejected for all models if a change is present, using the statistic $\hat{S}_k$. Comparing the detection time with the expected detection times, we may see that the expected detection times indeed are larger than the actual detection times, with one exception for the NCAR model with measurement noise. The lower bounds for the window length defined by (3.8) are given in Table 3.5. We may see that these window lengths are sufficient to detect the changes in the simulations, again with the exception of the NCAR model with measurement noise.

Finally, from Figure 3.5 we may see that if no change is present, the fluctuations in the statistic remain reasonably small so that the probability of false alarm is not very large.

From the simulations we may have seen that the score test and the generalized likelihood ratio test are both incapable of detecting small changes in the classification problem. However, the grid that is used in the simulation is also relatively small. In image processing applications, the images generally have far more pixels than 16×16. As the size of the grid grows, the detection qual-

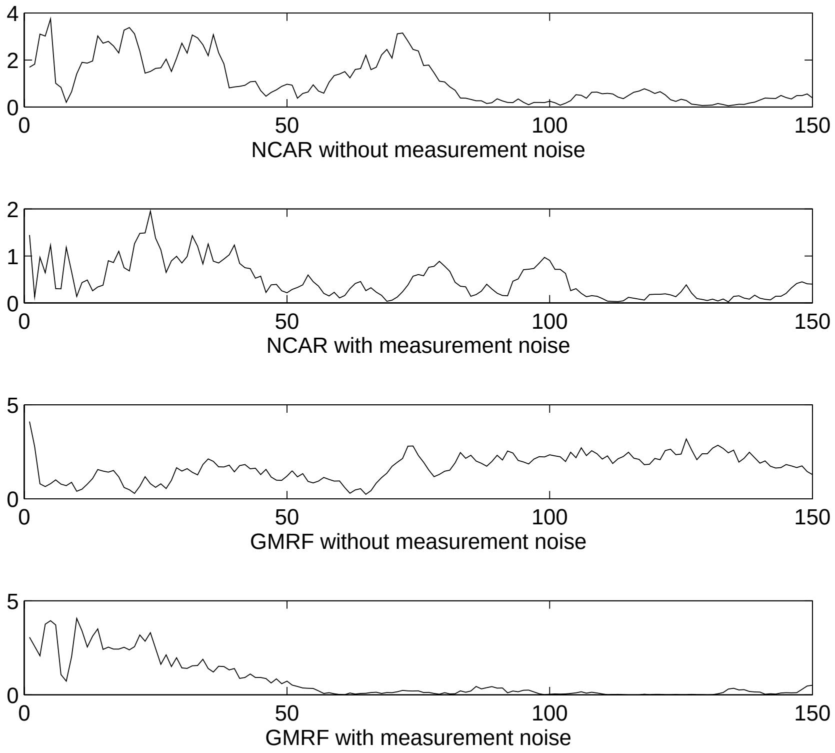

Figure 3.5: $\hat{S}_k$ without the presence of a change.

ity will improve. Since the computational load of the generalized likelihood ratio test, for which a maximization is required, increases rapidly when using larger grids, this test may not be suitable for dealing with real applications in image processing. The increase in computational load for the score test is only marginal if compared with the generalized likelihood ratio.

In case a sequence of images is available, and the change appears at a certain point in time, the simulations indicate that the adapted score test using $\hat{S}_k(\mathcal{Y}(k-l..k))$ may give some good results. In the simulations, the window length was assumed to be infinite. In practical situations, including the entire history of the process may give some problems. Since the statistic averages over the entire window length, the delay of detection will be proportional with the number of samples that are processed before a change has appeared. In the simulations, this did not create much problems, since only four samples were processed before the change appeared. In practice, a suitable window length has to be chosen. An indication for a suitable choice is given by the lower bound (3.8). For this particular example, this lower bound may become rather large; however, the simulations show that this may indeed be required.

Chapter 4

Local Changes

The detection of changes in dynamic systems received an impulse in the 60's when the Kalman filter was successfully applied for this purpose. Newbold and Ho [31], Mehra and Peschon [29] and Willsky [46] developed the theory of fault detection using the Kalman filter.

Since then the use of Kalman filtering techniques has only increased in applications. For example, in the introduction of [33], Clark, Frank and Patton mention an impressive number of detection problems that have been dealt with using the Kalman filter.

In the previous chapter all changes were present on the entire field. This allowed us to process all data on the grid simultaneously. In this chapter we study the detection of local changes in random fields. Hence, a change may only be present at certain sites on the grid. Throughout this chapter, we only consider random fields that may be described in the form of a stochastic dynamic system. For example, the field may be scanned in a column-by-column fashion, where every column may be written as a function of the previous columns. Although not all fields may be described in such a form, it still covers a large class of random fields [17].

Example 4.1 *Let us once more return to our farmer. This time he wants to harvest his cornfields. Putting several harvesters next to each other, he can cover an entire side of the field. This way, he may harvest the entire field in one sweep. For some reason, he has the feeling that part of the field has been damaged by an unfriendly neighbour. To separate the resulting damaged corn from the good corn, he has built a small device that can be installed on each of the harvesters. This device roughly measures the quality of the corn that is harvested. Hence, the farmer may detect a damaged part of his field while he is harvesting by monitoring the outcome of the device.*

As mentioned already, we use Kalman filtering techniques for the detection of local changes in random fields. Applications may be found, for example, in the detection of flaws on surfaces or object recognition. In Chapter 3 one single test is performed on the entire field to determine whether or not the structure of the field has changed. The reason for using a local approach here is twofold. Firstly, the changes are local, that is, the part of the field that is affected by the change is relatively small. Therefore, a global statistic may not be sensitive enough to result in a detection of the change. Secondly, we desire to use as little data as needed to detect a change. If the amount of data on the field is very large, the computational cost for processing all data may be enormous. Using a local approach, these problems may be reduced.

The main difference between the problem presented here and the classical detection problems concerns the form of the changes. Where changes in standard uni-directional processes generally appear at a certain time, and remain present for some period (which often lasts forever), in random fields the changes may take rather unstructured forms. Depending on the method of processing the field, the changes may appear and disappear at arbitrary instants. Therefore, the parameterization of the changes is an important aspect of the detection problem.

The general model that will be used in this chapter is given in Section 4.1. This section also contains the main results of this chapter. In Section 4.2 we restrict our attention to a simplified version of the general model. A closer look is taken at the structure of the field and the parameterization of the changes. Finally, Section 4.3 contains some simulations to illustrate the theory.

4.1 The general model

Several models have been developed for the description of discrete random fields, mainly for the use in image processing. An overview is given by Jain [16] and Dubes and Jain [10]. One particular model is the one-dimensional state space representation. In this model, the field is divided into sets of equal sizes, and the sites in each set are gathered in a vector X. The sets are ordered in such a way that the process at sites in set $k+1$ only depend on the process at sites in set k. Obviously, this limits our choice of the sets considerably.

Under normal circumstances, the field may be described by some known state space model. Whenever a change in the field occurs, the model is no longer accurate. Two possible strategies exist to detect the occurrence of a change. The first one consists of a goodness-of-fit test that simply tries to detect any deviation from the nominal model (Mehra and Peschon [29]). However, since we assume the changes to be local, the deviations from the nominal model are expected to be rather small. Therefore, this approach may not be very effective in this case.

The second approach consists of parameterizing the changes. In case the possible changes are known, it may be possible to find a model for the changed field. Although the assumption of completely known changes is not very realistic, this approach may be more useful than the goodness-of-fit approach. This is because of the fact that although we do not know the changes exactly, we may still estimate them using the partial knowledge we have on the changes. For example, if we know that the change causes a jump in the mean, but we do not know where the change appears, we may easily find a parameterization of the random field.

In Chapter 3 we used non-causal autoregressive random fields to describe the data. This lack of causality makes it impossible to use sequential tests on the grid. Therefore, we assume the random fields to be causal. We assume that the random field may be described by

$$\begin{aligned} X_{k+1} &= A_k(\theta)X_k + F_k(\theta)W_k + E_k^X(\theta) && (4.1)\\ Y_k &= C_k(\theta)X_k + V_k + E_k^Y(\theta) && (4.2) \end{aligned}$$

The disturbances W_k and V_k are assumed to be zero mean i.i.d. Gaussian variables with covariances $R_k^W(\theta)$ and $R_k^V(\theta)$, respectively. We assume these matrices to have full rank for all values of θ. The state vector X_k is initialized as $X_0 = 0$. The vector Y_k denotes the noisy measurement of the field on the kth set. The vectors E_k^X and E_k^Y are deterministic. Note that E_k^X starts at $k = 0$ and E_k^Y starts at $k = 1$. For notational simplicity, we may sometimes write A_k when we mean $A_k(0)$, and the same holds for other system matrices. Also, the additive terms E_k^X and E_k^Y are assumed to vanish for $\theta = 0$. Clearly, in this form the random field is nothing else than a stochastic system. The detection theory for stochastic systems has been developed by Willsky [46, 47]. The only difference with our model is that Willsky only allowed changes to appear in the additive terms E_k^X and E_k^Y.

The variable $\boldsymbol{\theta}$ represents the change in the field. It may take values in a certain parameter set Θ. The precise structure of this set depends on the changes that may appear. Under the null hypothesis, no change is present in the field. The nominal model of the field coincides with the parameter value $\boldsymbol{\theta} = 0$.

We now want to test whether a change is present or not. This implies that we test whether the null hypothesis

$$\mathcal{H}_0 : \boldsymbol{\theta} = 0$$

holds.

Depending on the parameter set Θ we may choose a strategy with the purpose of detecting all possible changes. According to Chapter 2, the generalized likelihood ratio is a suitable statistic for detecting an unknown change.

In case Θ consists of a finite number of parameters, the GLR may be obtained by explicitly calculating the likelihood ratio for each parameter (Willsky [46, 47]). For each $\theta \in \Theta$ we may use a Kalman filter to find the log-likelihood ratio

$$\begin{aligned} \ell_n(Y,\theta) &= \log \frac{f_\theta(Y)}{f_0(Y)} \\ &= \log \prod_{k=1}^{n} \frac{f_\theta(Y_k|Y_{k-1},\dots,Y_1)}{f_0(Y_k|Y_{k-1},\dots,Y_1)} \\ &= \sum_{k=1}^{n} \log \frac{f_\theta(\nu_k(\theta))}{f_0(\nu_k(0))} \end{aligned} \tag{4.3}$$

Here $\nu_k(\theta)$ is the innovation of the process, which may be obtained from the Kalman filter equations

$$\hat{X}_{k+1|k}(\theta) = A_k(\theta)\hat{X}_{k|k}(\theta) + E_k^X(\theta) \tag{4.4}$$

$$\hat{X}_{k|k}(\theta) = \hat{X}_{k|k-1}(\theta) + K_k(\theta)\nu_k(\theta) \tag{4.5}$$

$$\nu_k(\theta) = Y_k - C_k(\theta)\hat{X}_{k|k-1}(\theta) - E_k^Y(\theta) \tag{4.6}$$

$$P_{k+1|k}(\theta) = A_k(\theta)P_{k|k}(\theta)A_k(\theta)' + F_k(\theta)R_k^W(\theta)F_k(\theta)' \tag{4.7}$$

$$R_k(\theta) = C_k(\theta)P_{k|k-1}(\theta)C_k(\theta)' + R_k^V(\theta) \tag{4.8}$$

$$K_k(\theta) = P_{k|k-1}(\theta)C_k(\theta)'(R_k(\theta))^{-1} \tag{4.9}$$

$$P_{k|k}(\theta) = [I - K_k(\theta)C_k(\theta)]P_{k|k-1}(\theta) \tag{4.10}$$

Since $X_0 = 0$, the initial estimate $\hat{X}_{0|0}(\theta)$ also equals zero. Hence, the initial value of the covariance matrix $P_{0|0}$ is the zero matrix.

In our Gaussian setting, we may evaluate the LR as

$$\ell_n(Y,\theta) = \frac{1}{2}\sum_{k=1}^{n} \log \frac{|R_k(0)|}{|R_k(\theta)|} + \nu_k(0)'R_k(0)^{-1}\nu_k(0) - \nu_k(\theta)'R_k(\theta)^{-1}\nu_k(\theta) \tag{4.11}$$

where R_k denotes the covariance matrix of the innovation.

The detectability of a change θ is determined by the Kullback-Leibler information, which may in this case be written as

$$\begin{aligned} I_n(\theta) &= \mathbf{E}_\theta \ell_n(Y,\theta) \\ &= \frac{1}{2}\sum_{k=1}^{n} \{\log \frac{|R_k(0)|}{|R_k(\theta)|} - N + \operatorname{tr} R_k(0;\theta)R_k(0)^{-1} \\ &\quad + \mu_k(0;\theta)'R_k(0)^{-1}\mu_k(0;\theta)\} \end{aligned} \tag{4.12}$$

where

$$\begin{aligned}
\mu_k(0;\theta) &= \mathbf{E}_\theta \nu_k(0) \\
&= C_k(\theta)\mu_{k|k-1}(\theta;\theta) - C_k\mu_{k|k-1}(0;\theta) + E_k^Y(\theta) \qquad (4.13) \\
\mu_{k+1|k}(\theta;\theta) &= \mathbf{E}_\theta \hat{X}_{k+1|k}(\theta) \\
&= A_k(\theta)\mu_{k|k-1}(\theta;\theta) + E_k^X(\theta) \qquad (4.14) \\
\mu_{k+1|k}(0;\theta) &= \mathbf{E}_\theta \hat{X}_{k+1|k}(0) \\
&= A_k[\mu_{k|k-1}(0;\theta) + K_k\mu_k(0;\theta)] \qquad (4.15)
\end{aligned}$$

and

$$\begin{aligned}
R_k(0;\theta) &= \mathbf{E}_\theta(\nu_k(0) - \mu_k(0;\theta))(\nu_k(0) - \mu_k(0;\theta))' \\
&= R_k(\theta) + C_k(\theta)Q_{k|k-1}(\theta,\theta;\theta)C_k(\theta)' - C_k(\theta)Q'_{k|k-1}(0,\theta;\theta)C'_k \\
&\quad -C_kQ_{k|k-1}(0,\theta;\theta)C_k(\theta)' + C_kQ_{k|k-1}(0,0;\theta)C'_k \qquad (4.16)
\end{aligned}$$

where

$$\begin{aligned}
Q_{k+1|k}(\theta,\theta;\theta) &= A_k(\theta)[Q_{k|k-1}(\theta,\theta;\theta) + K_k(\theta)R_k(\theta)K_k(\theta)']A_k(\theta)' \\
Q_{k+1|k}(0,\theta;\theta) &= A_k[(I - K_k(0)C_k)Q_{k|k-1}(0,\theta;\theta) \\
&\quad +K_k(0)C_k(\theta)Q_{k|k-1}(\theta,\theta;\theta) \\
&\quad +K_k(0)R_k(\theta)K_k(\theta)']A_k(\theta)' \\
Q_{k+1|k}(0,0;\theta) &= A_k[Q_{k|k-1}(0,0;\theta) + Q_{k|k-1}(0,\theta;\theta)C_k(\theta)'K_k(0)' \\
&\quad -Q_{k|k-1}(0,0;\theta)C'_kK_k(0)' + K_k(0)C_k(\theta)Q'_{k|k-1}(0,\theta;\theta) \\
&\quad -K_k(0)C_kQ_{k|k-1}(0,0;\theta) + K_k(0)R_k(0;\theta)K_k(0)']A'_k
\end{aligned}$$

The initial values of all μ- and Q-variables are zero.

Lemma 4.1 *The change θ is not detectable if and only if $R_k(\theta) = R_k(0;\theta) = R_k(0)$ and $\mu_k(0;\theta) = 0$ for all k.*

Proof. Let us take one element of the sum in $I_n(\theta)$ as given by (4.12). This element consists of two parts; a part that represents the structural change and a part that represents the additional change. The structural change part is

$$\log\frac{|R_k(0)|}{|R_k(\theta)|} - N + \operatorname{tr} R_k(0;\theta)R_k(0)^{-1}$$

Since

$$R_k(0;\theta) = R_k(\theta) + \text{cov.matrix}$$

we know that $R_k(0;\theta) \geq R_k(\theta)$. Hence,

$$\begin{aligned}
\log\frac{|R_k(0)|}{|R_k(\theta)|} - N + \operatorname{tr} R_k(0;\theta)R_k(0)^{-1} &\geq \\
\log|R_k(0)R_k(\theta)^{-1}| - N + \operatorname{tr} R_k(\theta)R_k(0)^{-1} &= \\
-\log|R_k(\theta)R_k(0)^{-1}| - N + \operatorname{tr} R_k(\theta)R_k(0)^{-1} &= \\
\sum_{i=1}^{N} -\log\lambda_i(\theta) - 1 + \lambda_i(\theta) &\geq 0
\end{aligned}$$

where $\lambda_i(\theta), i = 1,\ldots,N$ are the eigenvalues of the matrix $R_k(\theta)R_k(0)^{-1}$. The last inequality follows from the fundamental inequality $x-1 \geq \log x$. Since the equality only holds if $x = 1$, the last inequality may be replaced by an equality if and only if $\lambda_i(\theta) = 1$ for all i. This implies that $R_k(\theta) = R_k(0)$. Clearly, the first inequality reduces to an equality if and only if $R_k(0;\theta) = R_k(\theta)$.

The additional change part is given by

$$\mu_k(0;\theta)R_k(0)^{-1}\mu_k(0;\theta)$$

We may easily see that this expression is zero if and only if $\mu_k(0;\theta) = 0$. □

The conditions given in the lemma are not that transparent. We would like to translate these conditions in terms of the system matrices.

Lemma 4.2 *The additional changes are not detectable if and only if*

$$E_k^Y(\theta) = -C_k(\theta)\mu_{k+1|k}(\theta;\theta)$$

Proof. Suppose that $\mu_k(0;\theta) = 0$ for all k. This implies that

$$E_k^Y(\theta) = -C_k(\theta)\mu_{k|k-1}(\theta;\theta) + C_k\mu_{k|k-1}(0;\theta) \tag{4.17}$$

for all k. The second term in the right hand side of (4.17) is always zero, since it may only be triggered by $\mu_k(0;\theta)$, which was assumed to be zero. The statement of the lemma follows easily. □

The conditions for a change in the structure may not be translated into conditions on the system parameters that easily.

The statistic on which we base our decision is the generalized likelihood ratio (GLR)

$$T_n(Y) = 2\sup_{\theta\in\Theta}\ell_n(Y,\theta)$$

If $T_n(Y)$ is larger than a certain threshold λ, then the null hypothesis is rejected. Because of the complex nature of the statistic, the choice of the threshold is not that straightforward. Although we know that the innovation is a

Gaussian process, the distribution of the maximum of a quadratic form is not that easily found. An experimental way of determining a threshold, is to process a large number of samples without a change, and use the results to determine a threshold value for which the relative number of false alarms is not too large.

If the number of parameters in Θ is not finite, this approach obviously fails. A possible solution is to discretize the parameter set and apply the same procedure for the resulting problem. However, this approach introduces some new robustness problems. Kumamaru, Sagara and Söderström [23] used a similar approach. They fixed the parameter vector such that only one of its components, the so-called monitoring parameter, had to be estimated. For this one-dimensional parameter a discretization of the parameter space may be quite accurate for a relatively small number of parameters. The drawback of this approach is that any change should affect the monitoring parameter, which may not always be the case.

Instead of discretizing the parameter set, we may also try to find the supremum of the likelihood ratio with the help of some numerical algorithms. In some situations, the supremum may actually be found analytically. Assuming Θ to be equal to $\mathbb{R}^r$, we know from Chapter 2 that the GLR may be approximated for small values of θ by the score test statistic

$$S_n(Y) = \mathcal{Z}_n(Y,0)'\Gamma_n(0)^{-1}\mathcal{Z}_n(Y,0)$$

where

$$\mathcal{Z}_n(Y,\theta) = \frac{\partial}{\partial\theta}\ell_n(Y,\theta)$$

is the efficient score, and

$$\begin{aligned}\Gamma_n(\theta) &= \mathbf{E}_\theta \mathcal{Z}_n(Y,\theta)\mathcal{Z}_n(Y,\theta)' \\ &= -\mathbf{E}_\theta \frac{\partial^2}{\partial\theta^2}\ell_n(Y,\theta)\end{aligned}$$

is the Fisher information matrix.

For the computation of this new statistic we do not need to know the parameter θ; it may completely be calculated under the null hypothesis. If the true parameter $\boldsymbol{\theta}$ is not close to zero, the score test is no longer a sufficient approximation of the GLR test. To find the value of the GLR test in this case, we may use something like a gradient search algorithm. However, this has the disadvantage that the Kalman filter equations have to be calculated for every iteration in this algorithm. Moreover, the convergence of these algorithms may be very slow so that a large number of iterations is needed to find the maximum likelihood estimate. Algorithms for the computation of the efficient score are given by Wilson and Kumar [48], Segal and Weinstein [38] and Leland [26].

4.1.1 A special case

Let us now focus on the special case where the changes in the system matrices A, F and C are known, and the only unknown is a scaling parameter in the additional term. Basically, this implies that we are testing for the presence of a given change with an unknown mean.

Suppose that, under the alternative hypothesis, the field may be described by

$$\begin{aligned} X_{k+1} &= A_k(\theta_0)X_k + F_k(\theta_0)W_{k+1} + c^X E_k^X(\theta_0) && (4.18)\\ Y_k &= C_k(\theta_0)X_k + V_k + c^Y E_k^Y(\theta_0) && (4.19) \end{aligned}$$

where θ_0 is known. The only unknowns therefore are the real numbers c^X and c^Y. Since only the additive terms depend on these unknowns, the covariance $R_k(\theta_0)$ is independent of these constants. Only the mean of the process is affected by these parameters. In fact, we obtain a similar situation as the one dealt with by Willsky [47]. The optimal value of the vector $c = (c^X, c^Y)$ may be calculated exactly, so that the generalized likelihood ratio also may be found without using any maximization.

Theorem 4.1 *The GLR for the test between hypotheses*

$$\begin{aligned} \mathcal{H}_0 &: \theta = 0, c = (0,0)\\ \mathcal{H}_1 &: \theta = \theta_0, c = (c^X, c^Y) \end{aligned}$$

where c^X and c^Y are unknown is given by

$$\begin{aligned} T_n(Y) &= \sum_{k=1}^{n}(\log\frac{|R_k(0)|}{|R_k(\theta_0)|} + \nu_k(0)'R_k(0)^{-1}\nu_k(0)\\ &\quad -G_k(\theta_0)'R_k(\theta_0)^{-1}G_k(\theta_0)) + U_n'(\theta_0)Q_n(\theta_0)^{-1}U_n(\theta_0) && (4.20) \end{aligned}$$

where

$$\begin{aligned} G_k(\theta_0) &= Y_k - C_k(\theta_0)\eta_k(\theta_0) && (4.21)\\ \eta_{k+1}(\theta_0) &= A_k(\theta_0)[I - K_k(\theta_0)C_k(\theta_0)]\eta_k(\theta_0) + A_k(\theta_0)K_k(\theta_0)Y_k && (4.22) \end{aligned}$$

and

$$\begin{aligned} Q_n(\theta_0) &= \sum_{k=1}^{n} H_k(\theta_0)'R_k(\theta_0)^{-1}H_k(\theta_0) && (4.23)\\ U_n(\theta_0) &= \sum_{k=1}^{n} H_k(\theta_0)'R_k(\theta_0)^{-1}G_k(\theta_0) && (4.24)\\ H_k(\theta_0) &= (\, C_k(\theta_0)\zeta_k^X(\theta_0) \quad C_k(\theta_0)\zeta_k^Y(\theta_0) + E_k^Y(\theta_0)\,) && (4.25)\\ \zeta_{k+1}^X(\theta_0) &= A_k(\theta_0)[I - K_k(\theta_0)C_k(\theta_0)]\zeta_k^X(\theta_0) + E_k^X(\theta_0) && (4.26)\\ \zeta_{k+1}^Y(\theta_0) &= A_k(\theta_0)[I - K_k(\theta_0)C_k(\theta_0)]\zeta_k^Y(\theta_0)\\ &\quad -A_k(\theta_0)K_k(\theta_0)E_k^Y(\theta_0) && (4.27) \end{aligned}$$

and the processes η_k, ζ_k^X and ζ_k^Y have zero initial values.

Proof. The derivative of the log-likelihood ratio with respect to $c = (c^X, c^Y)'$ is given by

$$\begin{aligned}\mathcal{Z}_n(Y, c) &= \frac{\partial \ell_n(Y, c)}{\partial c}\\ &= \frac{1}{2}\sum_{k=1}^{n} \nu_k(c)' R_k(\theta_0)^{-1} \frac{\partial \nu_k(c)}{\partial c}\end{aligned}$$

Let us write

$$\hat{X}_{k|k-1}(c) = (\, \zeta_k^X(\theta_0) \quad \zeta_k^Y(\theta_0)\,)\, c + \eta_k(\theta_0)$$

Then

$$\begin{aligned}\hat{X}_{k+1|k}(c) &= c^X \zeta_{k+1}^X(\theta_0) + c^Y \zeta_{k+1}^Y(\theta_0) + \eta_{k+1}(\theta_0)\\ &= A_k(\theta_0)\hat{X}_{k|k}(c) + c^X E_k^X(\theta_0)\\ &= A_k(\theta_0)(\hat{X}_{k|k-1}(c) + K_k(\theta_0)\nu_k(c)) + c^X E_k^X(\theta_0)\\ &= A_k(\theta_0)[I - K_k(\theta_0)C_k(\theta_0)]\hat{X}_{k|k-1}(c) + A_k(\theta_0)K_k(\theta_0)Y_k\\ &\quad -c^Y A_k(\theta_0)K_k(\theta_0)E_k^Y(\theta_0) + c^X E_k^X(\theta_0)\\ &= A_k(\theta_0)[I - K_k(\theta_0)C_k(\theta_0)](c^X\zeta_k^X(\theta_0) + c^Y\zeta_k^Y(\theta_0) + \eta_k(\theta_0))\\ &\quad +A_k(\theta_0)K_k(\theta_0)Y_k - c^Y A_k(\theta_0)K_k(\theta_0)E_k^Y(\theta_0) + c^X E_k^X(\theta_0)\end{aligned}$$

so that we may write

$$\begin{aligned}\zeta_{k+1}^X(\theta_0) &= A_k(\theta_0)[I - K_k(\theta_0)C_k(\theta_0)]\zeta_k^X(\theta_0) + E_k^X(\theta_0)\\ \zeta_{k+1}^Y(\theta_0) &= A_k(\theta_0)[I - K_k(\theta_0)C_k(\theta_0)]\zeta_k^Y(\theta_0) - A_k(\theta_0)K_k(\theta_0)E_k^Y(\theta_0)\\ \eta_{k+1}(\theta_0) &= A_k(\theta_0)[I - K_k(\theta_0)C_k(\theta_0)]\eta_k(\theta_0) + A_k(\theta_0)K_k(\theta_0)Y_k\end{aligned}$$

where $\zeta_0^X(\theta_0) = \zeta_0^Y(\theta_0) = \eta_0(\theta_0) = 0$.

As a result, we may write the innovation as

$$\begin{aligned}\nu_k(c) &= Y_k - C_k(\theta_0)\eta_k(\theta_0) - c^X C_k(\theta_0)\zeta_k^X(\theta_0) - c^Y[C_k(\theta_0)\zeta_k^Y(\theta_0) + E_k^Y(\theta_0)]\\ &= G_k(\theta_0) - H_k(\theta_0)c\end{aligned}$$

where

$$\begin{aligned}G_k(\theta_0) &= Y_k - C_k(\theta_0)\eta_k(\theta_0)\\ H_k(\theta_0) &= (\, C_k(\theta_0)\zeta_k^X(\theta_0) \quad C_k(\theta_0)\zeta_k^Y(\theta_0) + E_k^Y(\theta_0)\,)\end{aligned}$$

Note that $H_k(\theta_0)$ is a deterministic variable.

Substituting the innovation in the derivative of the log-likelihood ratio, and putting the result equal to zero, we obtain

$$\sum_{k=1}^{n}(G_k(\theta_0) - H_k(\theta_0)\hat{c}(n))' R_k(\theta_0)^{-1} H_k(\theta_0) = 0$$

which has a solution

$$\begin{aligned}\hat{c}(n) &= (\sum_{k=1}^{n} H_k(\theta_0)' R_k(\theta_0)^{-1} H_k(\theta_0))^{-1} \sum_{k=1}^{n} H_k(\theta_0)' R_k(\theta_0)^{-1} G_k(\theta_0) \\ &= [Q_n(\theta_0)]^{-1} U_n(\theta_0)\end{aligned}$$

where $Q_n(\theta_0)$ and $U_n(\theta_0)$ are defined as in the statement of the theorem.

The maximum value of twice the LR is now given by

$$\begin{aligned}T_n(Y) &= \sum_{k=1}^{n}(\log \frac{|R_k(0)|}{|R_k(\theta_0)|} + \nu_k(0)' R_k(0)^{-1} \nu_k(0) \\ &\quad - G_k(\theta_0)' R_k(\theta_0)^{-1} G_k(\theta_0)) + U_n'(\theta_0)[Q_n(\theta_0)]^{-1} U_n(\theta_0)\end{aligned}$$

□

From the proof of this theorem we may deduce that we should monitor the three new variables $\zeta_k^X(\theta_0)$, $\zeta_k^Y(\theta_0)$ and $\eta_k(\theta_0)$. From these variables the optimal value of c and the corresponding maximum value of the LR may be calculated.

Note that we only require the actual change to be detectable, i.e., the Kullback-Leibler information has to be strictly positive. Lemma's 4.1 and 4.2 specify the conditions that guarantee the detectability of the changes. Clearly, if θ_0 is such that the structural part of the change already is detectable, the value of c does not influence the detectability of the change anymore. Suppose now that θ_0 is such that $R(\theta_0) = R(0; \theta_0) = R(0)$. Hence, the structural part of the change is not detectable. From Lemma 4.2 we may then see that the change is not detectable if and only if

$$c^Y E_k^Y(\theta_0) = -C_k(\theta_0)\mu_{k|k-1}(\theta_0; \theta_0) \tag{4.28}$$

where

$$\mu_{k+1|k}(\theta_0; \theta_0) = A_k(\theta_0)\mu_{k|k-1}(\theta_0; \theta_0) + c^X E_k^X(\theta_0) \tag{4.29}$$

and $\mu_{0|-1}(\theta_0) = 0$. Equations (4.28) and (4.29) may be rewritten as

$$c^Y E_k^Y(\theta_0) = -c^x C_k(\theta_0) \sum_{i=1}^{k} \prod_{j=1}^{i-1} A_{k-j}(\theta_0) E_{k-i}^X(\theta_0) \tag{4.30}$$

Clearly, if $c = 0$ equation (4.30) holds so that the change is not detectable. Suppose now that there exists a $c_1 \neq 0$ for which the change is not detectable either. We may then easily see that equation (4.30) also holds for rc_1, for each $r \in \mathbb{R}$. If there exists another $c_2 \neq rc_1$, for any r, for which the equation also holds, it follows that the change is not detectable for any $c \in \mathbb{R}^2$. However, both $E_k^Y(\theta_0)$ and $\mu_{k|k-1}(\theta_0;\theta_0)$ have to be zero for all k for this to hold. These observations may be gathered in the following corollary.

Corollary 4.1 *The additional part of the change in the system (4.18)-(4.19) is not detectable if and only if either* $c = (0,0)$ *or*

$$E_k^Y(\theta_0) = -\kappa C_k(\theta_0) \sum_{i=1}^{k} \prod_{j=1}^{i-1} A_{k-j}(\theta_0) E_{k-i}^X(\theta_0)$$

for some $\kappa \in \mathbb{R}$ *and* c *is such that* $c^X = \kappa c^Y$.

Similarly, we may also apply the same procedure in case the parameter θ_0 is unknown, but contained in a finite parameter set Θ. In this situation, the maximum likelihood estimate of c has to be calculated at each time for all possible changes $\theta \in \Theta$. This has the disadvantage that for each parameter a relatively large number of variables has to be stored in memory. In fact, from the proof of the theorem it follows that we need approximately three times as much memory space for this problem compared to the problem with known intensity. However, if we use a gradient search algorithm to find the optimal value of the vector c, the computational load is bound to be considerably higher.

4.2 Detecting local changes in random fields

From now on we use a simplified model for the field. The matrices in the defining equations are assumed to be constant over the field. This implies that, under the null hypothesis, the field may be written as

$$X_{k+1} = AX_k + FW_{k+1} \tag{4.31}$$

$$Y_k = CX_k + V_k \tag{4.32}$$

where $X_k \in \mathbb{R}^m$. This state vector may represent a column of the field, part of a column, multiple columns, or just some sites from the field. At the moment we do not specify the exact nature of the state. The vectors W_k, V_k and Y_k are denoted as the field noise, measurement noise, and measurements of the field. The covariance matrices are given by R^W and R^V.

Several types of changes may occur in a field.

- A change in the matrix A may be the result of a change in the structure of the field; the relation between the field values at different sites has changed.

- A so-called drift or trend in the field may be modelled by an additive change in the state equation.

- A change in the matrix C may be the result of several events. In most cases, such a change will be accompanied by an extra additive change in the measurement equation.

  Firstly, the measurement device may be defective, thus producing incorrect measurements.

  Secondly, some object may be blocking the view of the measurement device. These objects may vary from clouds when taking satellite pictures, to pieces of dirt when scanning a metal surface with a sonar device.

  Finally, an abrupt change in the field may in some situations be described more efficiently by a change in the C-matrix, possibly together with an additive change in the measurement equation, than by a change in the state equation. This may be explained by the lack of memory in the measurement equation, as opposed to the Markovian nature of the state equation. For a fault that is described by a change in the A-matrix, the model is bound to give changed values at the sites in the neighbourhood of the fault. However, if the change is abrupt, the actual field has not changed at these sites, so that the model becomes less accurate.

In the sequel we only consider changes of the third type; a change in the matrix C, accompanied with an additive change in the measurement equation.

The location of the changes clearly has to be parameterized. This implies that the position of the changed parameters in the system matrices also have to be regarded as (discrete) parameters.

Assume we have a model describing a square grid (say N by N) under normal circumstances. The state X_k is the kth column of the field. Suppose we want to detect an object of known size and form, say a square of size M by M, where M is smaller than N. Suppose that k is the smallest number for which the kth column of the grid is covered by the object. The rows of C_k corresponding to the elements of the column that are covered now change to zero. This implies that the ith row up to the $(i + M - 1)$th row of C_k are zero. The parameters k and i may in this case be used to denote the position of the change on the grid, as illustrated in Figure 4.1. Additionally, an extra vector, say E_k, appears in the measurement equation. It is zero everywhere, apart from the elements that correspond to the position of the object. Then

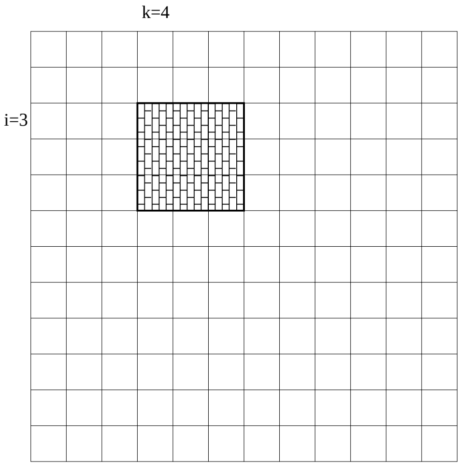

Figure 4.1: Position of a change of size 3 by 3 on a 12 by 12 grid.

we may describe the new measurement as

$$Y_k = C_k X_k + E_k + V_k$$

The vectors E_k together form a matrix E that is exactly equal to the zero field plus the object. For this new model we may now use the detection method as described in the previous sections. Since we do not know the position of the object, we have to repeat this procedure for all possible positions. Hence, we have to estimate the parameters k and i.

For each possible position, an evaluation of the Kalman filter equations is required for both hypotheses. Although the change itself does not contain any texture, the fact that it does not cover the entire field implies that we still have to use a Kalman filter to extract the information we need. To avoid this, we may also use a local model of the change. This implies that we simply use a window of size M by M and perform a test for each position of this window. However, the model under the null hypothesis may give some problems in this case, so that it may not be as accurate as the previous approach. Furthermore, since this advantage is only present in this particular situation, we prefer to use the other approach.

4.3 Simulations

Here we demonstrate the theoretical results we obtained in the previous sections. To illustrate the quality of the tests we plot the probability of detection as a function of the probability of false alarm. These plots are also known as the *receiver operating characteristics* (ROC) [36].

The changes we consider are all of the type as described in the example of the previous section. This implies that some of the rows of the matrix C change to zero, while simultaneously an additive term in the measurement equation appears. If we construct a matrix by placing all vectors E_k^Y next to each other, we obtain a blueprint of the scratch or damage that we are looking for, as we may see in Figure 4.12. The intensity of the scratch is defined as the (assumed constant) value of the elements in E_k^Y.

We use 100 samples of the field without change to obtain the thresholds corresponding to a certain probability of false alarm. An additional 100 samples with change are used to obtain the probabilities of detection. A change is said to be detected if the GLR ($T_k(Y)$) crosses the threshold while the change is ϵ-detectable, that is, the expectation of the log-likelihood ratio has to be larger than ϵ. This extra condition is imposed to make the statistic more robust. For example, we do not always desire to reject the null hypothesis if only one pixel has a value that deviates from its nominal value. We choose $\epsilon = 0.1$, which implies that almost all changes are detectable.

The original model for the (causal) field is given by

$$\begin{aligned} X_t &= 0.2X_{t-t_h-t_v} + 0.4X_{t-t_h} + 0.3X_{t-t_h+t_v} + W_t \\ Y_t &= X_t + V_t \end{aligned}$$

where W_t and V_t are independent Gaussian noise processes with zero mean and a variance equal to 1. The constants t_h and t_v are the horizontal and vertical shift as defined in Chapter 3. We assume the boundary values to be zero. The corresponding system matrices are given by

$$A = \begin{pmatrix} 0.4 & 0.3 & 0 & \cdots & 0 \\ 0.2 & 0.4 & 0.3 & \ddots & \vdots \\ 0 & 0.2 & 0.4 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & 0.3 \\ 0 & \cdots & 0 & 0.2 & 0.4 \end{pmatrix}$$

and $F = C = I_{10}$. The covariance matrices R^W and R^V are also identical to the identity matrix.

Note that this field is actually causal in the k direction, so that the state space description becomes very natural. In general, causal fields may be more complex, as was shown by Jain [15]. However, all causal and semi-causal fields with zero-boundary conditions may be described in a state-space form.

4.3.1 Known change

First we assume the change to be known. The parameter space Θ contains one element only,

$$\Theta = \{\theta\}$$

The alternative hypothesis now is simple, so that from Chapter 2 we know that the sequential probability ratio test actually is optimal in the Neyman-Pearson sense. In this case Wald's inequality for the probability of false alarm also holds;

$$\alpha = \mathbf{Pr}[(\exists n)(T_n(Y,\theta) > \lambda)|\boldsymbol{\theta} = 0] \leq e^{-\lambda/2}$$

In this particular case, the overshoot over the threshold λ may be rather large. Therefore, this inequality is rather rough, so that the thresholds that may be obtained from this inequality are too large to be of practical use.

We consider three types of changes;

- **Parallel scratches.** In case we know that the only possible change is a certain vertical scratch on the field, the model for this change becomes quite simple. In fact,

  $$E_k^Y = cE\delta_{k-k_c}$$

  where c is the known intensity of the scratch, and E is a known vector consisting of zeroes and ones only. The known number k_c denotes the column where the scratch appears.

  Four different intensities are used, varying from zero to 2. The power of the test is plotted against the size of the test in Figure 4.2.

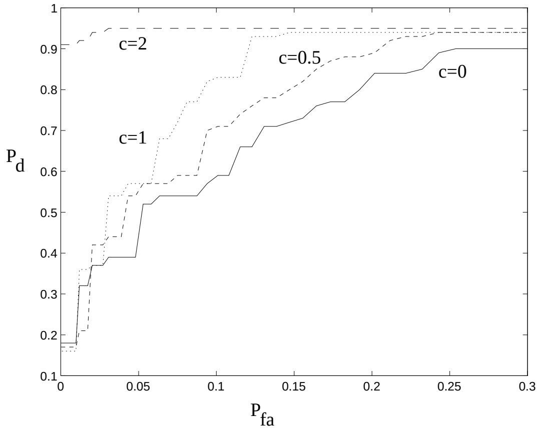

Figure 4.2: Receiver operating characteristic for parallel scratches.

- **Orthogonal scratches.** In case the only possible change is a horizontal scratch, the model remains quite simple. Denoting k_0 and k_1 as the number of the first and last column in which the scratch appears, we find

  $$E_k^Y = \sum_{i=k_0}^{k_1} cE\delta_{k-i}$$

where E now contains one 1 only, at a fixed position.

The same four intensities were used as for the parallel scratches. The ROC is given in Figure 4.3.

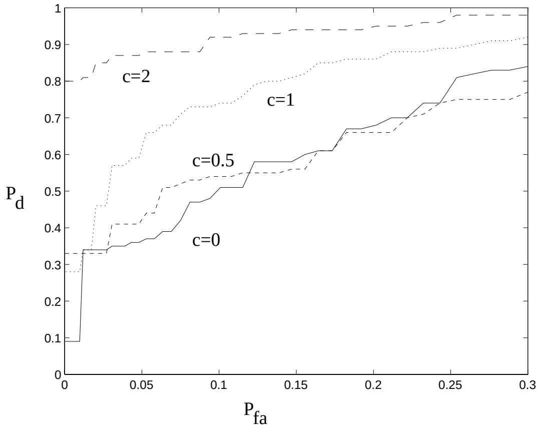

Figure 4.3: Receiver operating characteristic for orthogonal scratches.

- **General damages.** Finally, a combination of the previous changes gives us more general changes. For simplicity, we only consider changes that are constructed of parallel scratches that are connected with each other. The corresponding ROC is given in Figure 4.4.

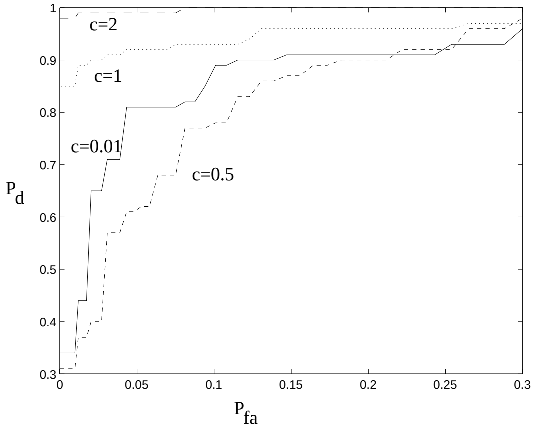

Figure 4.4: Receiver operating characteristic for general damages.

From the figures we may observe that a higher value of the intensity improves the detection quality of the test. If the intensity is smaller than 1, this difference is not that clear anymore. Furthermore, the general damages are detected most accurately, which may be explained by the fact that the general damages contain on average more sites than the other scratches. Finally, the parallel scratches are more easily detected than the orthogonal scratches. All sites in a parallel scratch are contained in one column, whereas the sites in the orthogonal scratches are spreaded over a large number of columns. This implies that the relative number of affected sites is larger for parallel scratches, which may explain the higher quality of detection.

4.3.2 Unknown intensity

In case the position and the form of the scratch are known, but the intensity is unknown, we may find some direct results. If we write θ_0 as the change with intensity 1, then the system under the alternative hypothesis may be written as

$$\begin{aligned} X_{k+1} &= AX_k + FW_{k+1} \\ Y_k &= C_k(\theta_0)X_k + V_k + cE_k^Y(\theta_0) \end{aligned}$$

As we may see, this system is of the form as treated in Section 4.1.1. The only difference is that c now is an element of $\mathbb{R}$ in stead of $\mathbb{R}^2$, so that the resulting expectation of the test statistic changes to

$$\begin{aligned} \mathbf{E}_c T_n(Y) &= 1 + \sum_{k=1}^{n} \log \frac{|R_k(0)|}{|R_k(\theta_0)|} - N + \operatorname{tr} R_k(0;\theta_0)R_k(0)^{-1} \\ &\quad + c^2 \rho_k(0;\theta_0)' R_k(0)^{-1} \rho_k(0;\theta_0) \end{aligned}$$

Applying the theory on a set of samples with intensities varying between -2 and 2, we obtain a ROC as given in Figure 4.5. As before, this figure shows that the best results are obtained for general damages, followed by parallel scratches and the orthogonal scratches again come in last.

As we have seen before, the detection quality improves if the true value of the parameter c is large. Therefore, for a consistent evaluation of these results we have to interpret them in an appropriate way. A first attempt may be to consider the detection rate as a function of the intensity. However, since the number of sites in a scratch also varies, this may not be sufficient. Therefore, we use the Kullback-Leibler information to indicate the detectability of a scratch. Figure 4.6 shows the relative number of detections as a function of the Kullback-Leibler information. The points in the graph indicate the relative number of detections for changes with a Kullback-Leibler information in an interval of length one half.

As expected, Figure 4.6 clearly illustrates a correlation between the Kullback-Leibler information in a change and the quality of detection.

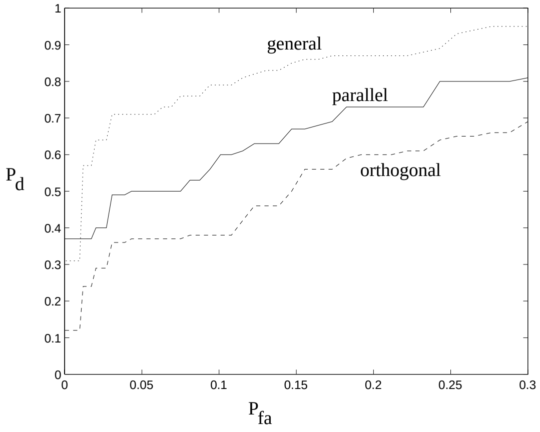

Figure 4.5: Receiver operating characteristic for changes of unknown intensity and known position and form.

4.3.3 Unknown change of given intensity

If the position and the size of the change are unknown, we may use a generalized likelihood ratio test, for which the ROCs are shown in Figures 4.7, 4.8, and 4.9. The intensity of the scratch is supposed to be known.

To save computation time we consider only a limited number of parameters for the general scratch case. For the first column of the field we evaluate all possible (parallel) scratches, and store only the 100 changes with the largest likelihood ratios. At every following column, we consider only the 100 previous changes, all possible extensions of these changes, and all possible new (parallel) scratches. Again the 100 most likely changes are stored.

From the figures we may see that the detection quality for general damages is reasonable, whereas the detection quality for both parallel and orthogonal scratches may only be considered reasonable for an intensity of 2. Note that the difference between parallel and orthogonal scratches is smaller than in the previous situations.

4.3.4 Unknown change of unknown intensity

Finally, if apart from the position and the size of the scratch, the intensity is also unknown, we may use a combination of the previous methods to detect these changes. For all possible changes, after each iteration, we calculate the most likely value of the intensity of the scratch and its corresponding likelihood ratio. The maximum value of these ratios is compared with a threshold to decide whether or not a change is present.

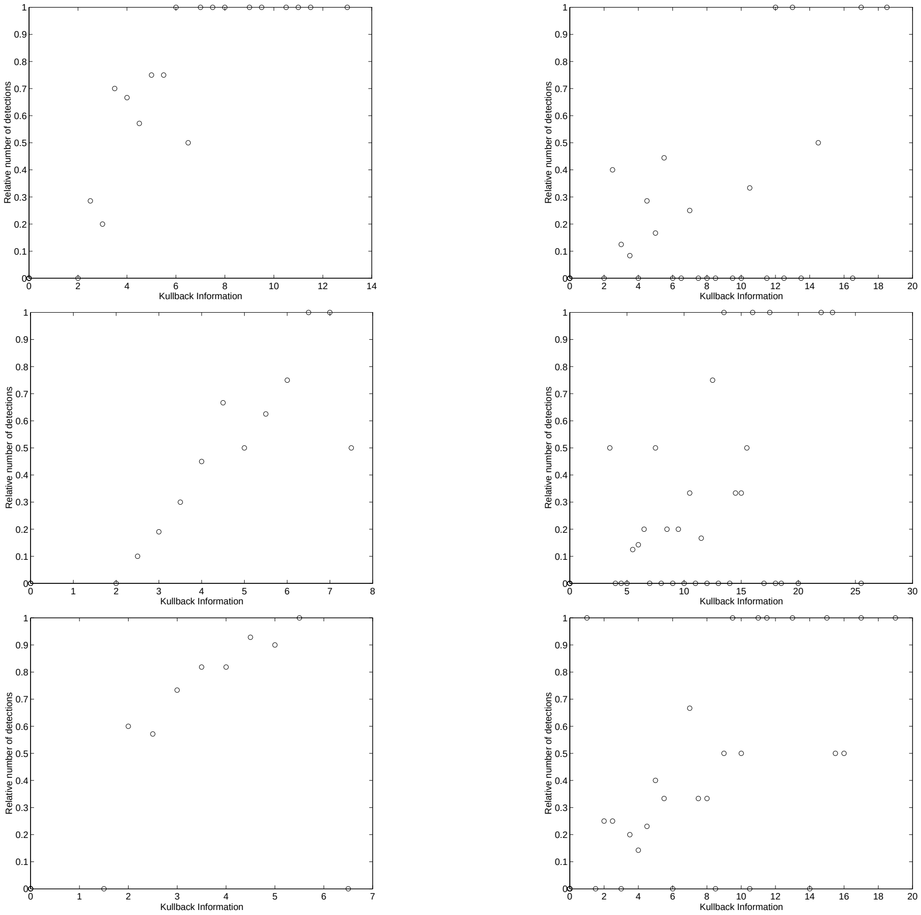

Figure 4.6: Detection rates as function of Kullback-Leibler information. From top to bottom: Parallel scratches, orthogonal scratches and general damages. From left to right: Known form of change and unknown form of change.

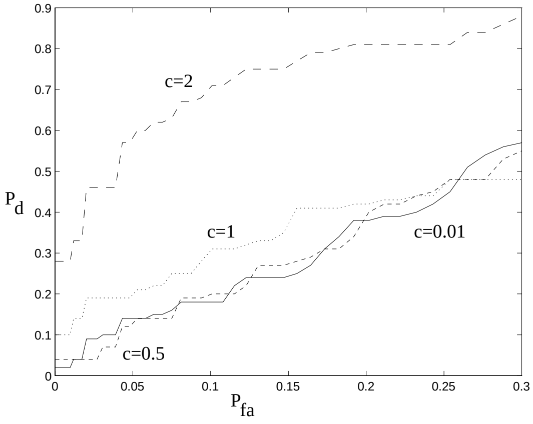

Figure 4.7: Receiver operating characteristic for parallel scratches of unknown length and position.

As before, the value of the threshold is determined empirically. One hundred samples of the field without and one hundred samples of the field with a change are used to obtain a ROC. The resulting plot for these experiments is shown in Figure 4.10. We may see that the detection quality has dropped considerably compared with the previous results. The detection quality for general damages is still superior to the other two classes of scratches. The difference between the orthogonal and the parallel scratches has vanished completely.

As for the scratches of known form, these results are also presented using the Kullback-Leibler information. Figure 4.6 illustrates the detection rates as a function of the Kullback-Leibler information. As might be expected, a correlation between the Kullback-Leibler information and the detection rate still exists. However, if we compare the figures on the right with the figures on the left, we may see that the correlation is not that clear anymore.

As a final illustration, one particular realization of the field has been highlighted. In Figure 4.11 the measurements of a field containing a change are shown as a function of the position on the grid. The change has an intensity of 1.64 and is positioned as illustrated in Figure 4.12. The generalized likelihood ratio for this field is plotted in Figure 4.13, together with its expected values, calculated under the assumption that the estimated changes are correct.

From the 100 samples without a change we may find that to obtain a false alarm rate of 0.1 the threshold should be 44. To obtain a false alarm rate of 0.3, the threshold may be chosen as 38. From the figure we may see that this second threshold is reached after the 9th column, and the first threshold is reached after the 10th column. However, if we examine the estimates at the 9th and the 10th column we find that only the latter coincides with a correct

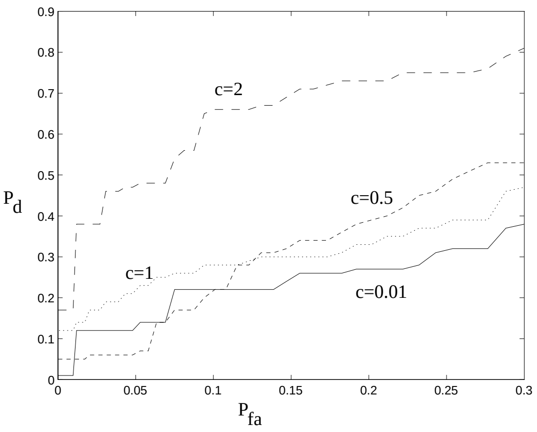

Figure 4.8: Receiver operating characteristic for orthogonal scratches of unknown length and position.

detection. In Figure 4.14 the positions of these estimated changes are shown. Their estimated intensities are -1.45 (9th column) and 1.61 (10th column).

4.3.5 Discussion

The use of the Kalman filter for detection problems has been proven valuable already by Willsky [46, 47]. In this chapter we applied this approach for the detection of changes in random fields. Willsky used the Kalman filter to detect additive changes only. The local changes that may appear in random fields however may not always be modeled by additive terms in the system equations (4.18)-(4.19). The main problem that arises here is the parameterization of the changes, due to the use of a sequential test on a two-dimensional grid.

From the different receiver operating characteristics, we may see that the detection quality of the generalized likelihood ratio test varies with the complexity of the problem. The more we know about the change, the better the detection quality. Furthermore, the size of the region of change and the intensity of the change positively influence the detection quality. These two characteristics of a change are measured through the Kullback-Leibler information. Indeed, Figure 4.6 shows a positive relation between the Kullback-Leibler information and the relative number of detections.

The noise that has been used in the simulations is rather strong; both the field noise W_t and the measurement noise V_t have a variance equal to one. This implies that the realizations of the field have a rather random-looking appearance, as we may see from Figure 4.11. In practice, the noise may be a lot smaller, thus resulting in more structured realizations. This will also

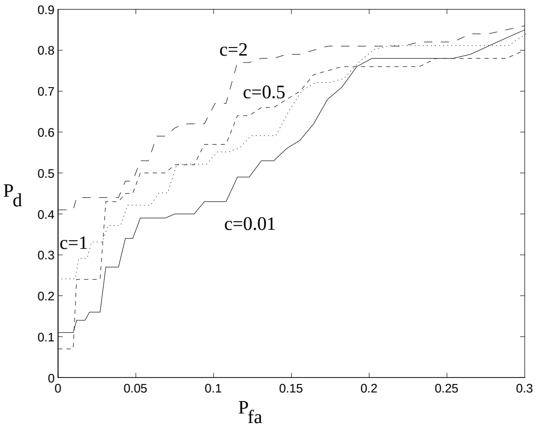

Figure 4.9: Receiver operating characteristic for general damages of unknown shape and position.

improve the detection quality.

The main drawback of the use of the generalized likelihood ratio is the maximization that is required whenever the position of the change is unknown. In these simulations we used a grid that is quite small; it only contains 10 by 10 grid-points. In practice, the grids will be much larger. Even for a grid this small the number of computations required is very large. Hence, it seems necessary that to obtain an algorithm that is practically useful for the detection of changes at unknown positions, we need to simplify the search procedure. In the last simulation, we already used a minor simplification by only consider the 100 most likely changes. Experiments have shown that the detection quality does not deteriorate seriously if this number is lowered to around 20.

Parallel scratches appear to be easier to detect than orthogonal scratches. Note, however, that this also strongly depends on the random field. It may happen that a random field has the shape of wave in the column direction; the field is more or less constant in each column, and varies between the columns. In that case it may be clear that a parallel scratch may be very difficult to detect, whereas orthogonal scratches seem to be easy to detect.

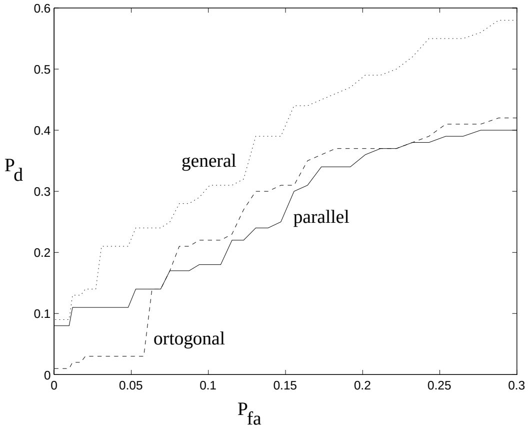

Figure 4.10: Receiver operating characteristic for changes of unknown intensity, unknown position and unknown shape.

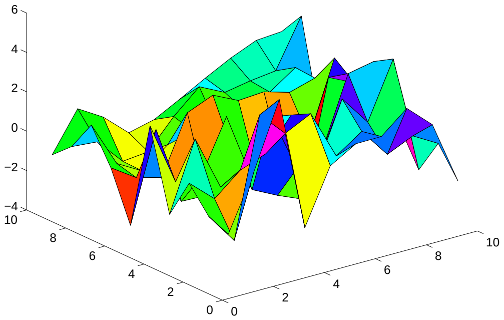

Figure 4.11: Measurements of a field with a change as a function of the position on the grid.

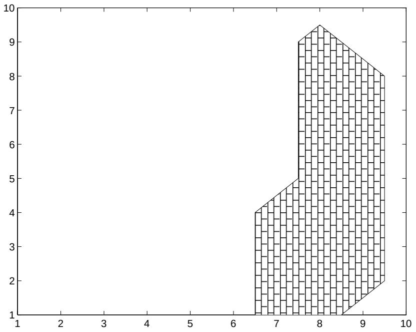

Figure 4.12: Position of the region of change on the grid.

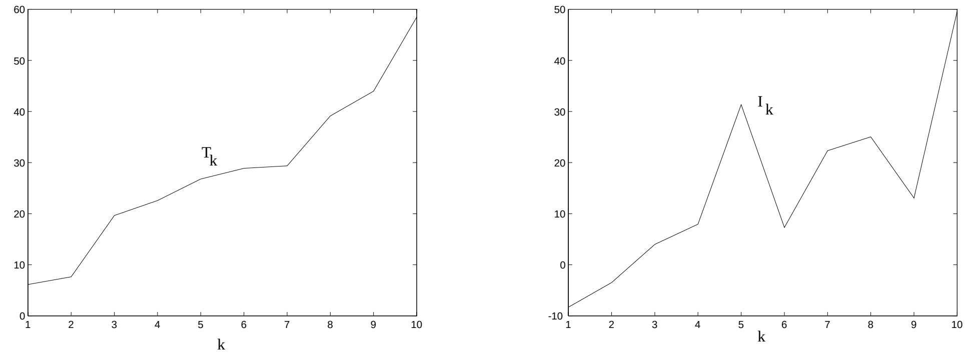

Figure 4.13: Generalized likelihood ratio (left) and Kullback-Leibler information (right) as function of k for the realization shown in Figure 4.11.

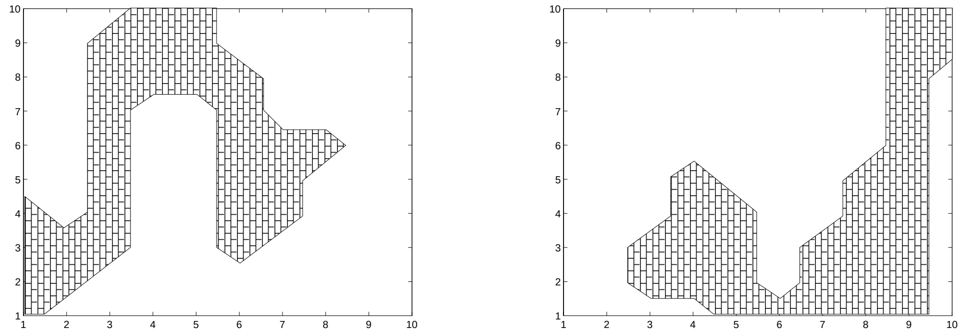

Figure 4.14: Position of estimated changes after 9th column (left) and 10th column (right).

Chapter 5

Quickest Detection

One of the main obstacles in the detection of global changes, as treated in Chapter 3, is formed by the size of the grid. In applications from image processing, an image may easily contain thousands of pixels, so that the statistical tests from Chapter 3 may lead to serious computational problems.

Since it is in general not necessary to use all available data to reach a decision of sufficient quality, a *sequential* approach seems to be more appropriate for the detection of global changes on a large grid. This implies that a sequence of statistical tests is performed on an increasing set of sites from the grid. At each stage of this procedure, a decision is made whether to continue with the next stage or to stop and to choose between one of the hypotheses. In fact, already in Chapter 4 a sequential approach has been used to detect local changes. However, in that case the logic for using a sequential approach was the structural properties of the process that was defined on the grid. Viewed from this perspective we may expect some problems in the implementation of a sequential approach in case the random field is non-causal. Note the difference between time-sequential tests and the sequential tests used here.

Due to the lack of ordering on a two-dimensional plane, the sequence of measurements that may be used to detect a change is generally not unique. In fact, we may use this freedom in the choice of sequences to minimize the number of measurements needed to detect a change. For example, if a local change is likely to be present in a certain region of the plane, it will probably be most efficient to start searching in that area. It follows that the a-priori information on the changes that is available is rather important for the strategy that should be used.

In this chapter we define the quickest detection problem for random fields. We first approach this problem from the classical Neyman-Pearson point of view, where the class of feasible tests is restricted by constraints on the error probabilities. In Section 5.3 we use the Bayesian approach. In this case, the prior distribution of the changes are assumed to be known, as well as the cost of making an error, relative to the sampling cost. The theory provided by Cairoli and Dalang [6] formed the major inspiration for this chapter, in particular the Bayesian approach.

5.1 Definitions and tools

In contrast to the previous chapters, the random fields in this chapter do not have to be defined on grids. Instead, we assume that the two-dimensional plane may be divided in several regions that have similar characteristics. For example, an image may be segmented based on texture. Each of these regions is represented by a site. The set of sites is in this case termed an *index set* and is denoted by $\mathcal{G}$.

In order to obtain measurements from the random field, we introduce an *observation mechanism.* At each site, we may make one or more observations. These observations are described by a stochastic process $Y_t(k)$, denoting the kth measurement at site t. The random field is now defined as the set of processes $Y_t(k)$ for all $t \in \mathcal{G}$. We assume $Y_t(k)$ to be statistically independent of $Y_s(l)$ for all $s \neq t$ and for all k and l. Furthermore, $Y_t(k)$ is also assumed to be independent of $Y_t(l)$ for all $k \neq l$. Clearly, these restrictions are rather severe. Although it may be possible to define the same problem for a more general random field, in this chapter we only focus on these so-called independent random fields.

The number of measurements that may be taken from each site is termed the *resources* of the observation mechanism. We denote M_t as the resources at site t, and M as the vector containing all M_t. If we may take infinitely many measurements from each site, the observation mechanism is said to be *inexhaustive.* Clearly, after taking a measurement from a site t of a random field with an exhaustive observation mechanism, the resources at that site reduce with 1.

On a random field the sequence of measurements is not unique. In particular, an independent field allows any ordering of the sites. Hence, we need to define a *sample path* that tells us where to make our next observations.

Definition 5.1 (Sample path) *A sample path is defined as a sequence of samples,* $\mathbf{t} = \{\mathbf{t}_1, \mathbf{t}_2, \ldots\}$. *A sample* $\mathbf{t}_k$ *is defined as a set of sites from the field.*

A sample path is said to be deterministic if all samples are fixed with probability 1.

If each sample only contains one site, the sample path is said to be a *site-by-site* sample path. One possibility for a sample to contain more than one site is that there is more than one measurement device available. Alternatively, we may also make a series of measurements before evaluating the results.

If the sample path that is used is deterministic, the measurements taken at the described sites may be characterized as an ordinary one-parameter stochastic process. Hence, stopping rules may be defined without any problem for these sample paths, as in Chapter 2.

Example 5.1 *One day, our farmer hears some rumors about a jar containing gold coins being buried on his land. He divides his land in squares and picks one of these squares to start digging. After an hour he has not found anything yet, and continues digging at one of the other squares.*

We may consider the set of squares as an index set with an exhaustive observation mechanism. Each observation coincides with the result of one hour of digging. The sample path is the sequence of squares that the farmer uses.

For stochastic sample paths, we may also find a class of stopping rules. As for deterministic sample paths, we may define a stopping time τ and a decision function δ, which together form a stopping rule. However, the stochastic nature of the sample path increases the complexity of the calculations involved with these sample paths considerably, since each next sample depends on the outcome of the previous measurements.

Example 5.2 *After the farmer has searched several of the squares, he finds one gold coin in one of the squares. He concludes that the jar must have been broken so that its contents have spread over a certain region. Hence, he continues his search in the squares next to the square where he found this gold coin.*

The resulting sample path is stochastic; depending on the outcome of the measurements, the next square is chosen.

It is clear now that not only do we need a stopping rule but also a sample path to solve a quickest detection problem on a random field.

Definition 5.2 (Stopping strategy) *A stopping strategy is defined as the combination of a sample path and a stopping rule, and is denoted as* $\mathbf{d} = (\mathbf{t}, \tau, \delta)$. *The class of stopping strategies is denoted as* $\mathcal{R}$.

The quickest detection problem may now be defined as the minimization of some cost function over the class of stopping strategies. This cost function clearly depends on the approach we are using. Generally, it will contain the number of measurements needed to reach a decision.

Throughout this chapter we assume the set of states the random field may take to be finite. We introduce a set of parameters, $\Theta^* = \Theta \cup \{0\}$, to denote these states. The nominal state, the random field without a change, is characterized by the parameter 0. Each of the parameters in Θ correspond with one particular change on the random field. Since not all of the sites in the index set have to be influenced by the change, we also introduce the *region of change* as the index set $\mathcal{G}_c(\theta) \subset \mathcal{G}$ that contains all sites that are influenced by the change θ. The null hypothesis now corresponds with the random field being in the nominal state. Clearly, the alternative hypothesis then corresponds with the random field being in one of the changed states.

5.2 Neyman-Pearson approach

Using the Neyman-Pearson approach, we want to minimize the number of measurements needed to make a decision, where the error probabilities are bounded by some given constants. This implies that we constrain the class of stopping strategies to the class

$$\mathcal{R}_{\alpha,\gamma} = \{\mathbf{d} \in \mathcal{R} | \alpha(\mathbf{d}) \le \alpha, \gamma_\theta(\mathbf{d}) \le \gamma \text{ for all } \theta \in \Theta\}$$

where $\alpha(\mathbf{d})$ and $\gamma_\theta(\mathbf{d})$ are the probability of false alarm and the probability of missing the change θ, respectively, when using the stopping strategy $\mathbf{d}$.

Let us define the *average site number* as

$$N_\theta(\mathbf{d}) = \mathbf{E}_\theta[\sum_{i=1}^{\tau} |\mathbf{t}_i|]$$

for all $\theta \in \Theta^*$. Here $|\mathbf{t}_i|$ denotes the number of elements in the sample $\mathbf{t}_i$. The average site number simply denotes the expected number of measurements that has to be taken from the field in order to reach a decision. In analogy with the UME stopping rules from Chapter 2, we may now define the UME stopping strategies.

Definition 5.3 (UME stopping strategy) *The uniformly most efficient (UME) stopping strategy is defined as the stopping strategy that minimizes all average site numbers,*

$$\mathbf{d}^{NP} = \{\mathbf{d} \in \mathcal{R}_{\alpha,\gamma} | N_\theta(\mathbf{d}) \le N_\theta(\mathbf{d}'), \textit{ for all } \mathbf{d}' \in \mathcal{R}_{\alpha,\gamma}, \textit{ for all } \theta \in \Theta^*\}$$

As in the one-parameter case, a UME stopping strategy may not exist.

Pelkowitz and Schwartz [34] and Nikiforov [32] considered similar problems in a one-parameter setting, where an unknown change time was involved. Both papers are concerned with the minimization of the mean delay of detection for a given mean time between false alarms. Pelkowitz and Schwartz use a test that is based only on the latest sample, thus neglecting all previous information. Their goal is to find the optimal size of the samples that should be used to obtain the best results. Nikiforov gives some asymptotic results for the worst case mean delay of detection.

In the following sections we examine the quickest detection problem for some specific cases. In Section 5.2.1 we assume the alternative hypothesis to be simple. Then, in Section 5.2.2 the case where the alternative hypothesis is composite is dealt with for some special cases.

5.2.1 Simple hypotheses

Let us assume the alternative hypothesis to be simple. Under the null hypothesis, the density function of the random field is given by f_0 at each site, and under the alternative hypothesis it is given by f_t at site t. We assume $f_t \not\equiv f_s$ for $s \neq t$. This is no restriction because when both density functions are equal we may join both sites into one site without loss of generality. In case the observation mechanism is exhaustive, the resources of both sites then have to be added.

Example 5.3 *Our farmer decides to start growing lettuce on his land. Not willing to take any further risk, he first wants to find out if his land is fertile enough for growing lettuce. He knows that the presence of a particular organic substance is a perfect indicator for the fertility of his land. The problem is that this substance is hard to detect. However, he does know that the presence of the substance tends to darken the soil. Therefore, he divides his land in several regions, based on the darkness of the soil. Then, by taking samples from the darkest regions and having them examined in a laboratory he may conclude whether or not the substance is present, and from that whether or not the land is fertile.*

For simplicity, let us consider a field whose index set only contains one site, and that has an inexhaustive observation mechanism. The class of sample paths is now entirely determined by the sample sizes.

Theorem 5.1 *For a test between two simple hypotheses on a field with index set $\mathcal{G} = \{t\}$ and with an inexhaustive observation mechanism, the UME stopping strategy in $\mathcal{R}_{\alpha,\gamma}$ consists of the sample path with sample size 1 and the SPRT that gives error probabilities equal to α and γ.*

Proof. We may see that any stopping strategy in $\mathcal{R}_{\alpha,\gamma}$, may be rewritten as a stopping strategy with sample size 1, by defining the stopping rule appropriately. According to Lemma 2.2, this stopping rule may never be more efficient than the SPRT for sample size 1. As a result, the SPRT for sample size 1 will be most efficient. □

If the distribution function of the increment of the log-likelihood ratio is known, we may calculate the error probabilities and the average site number for these tests, using the integral equations from Chapter 2.

Example 5.4 *Let us consider a random field with an index set that consists of one site, and whose observation mechanism is inexhaustive. The process is Gaussian with variance 1. Under the null hypothesis the mean is equal to zero, and under the alternative hypothesis it is equal to μ. According to Theorem 5.1, the UME stopping strategy in the class $\mathcal{R}_{0.05,0.05}$ consists of the sample path with fixed sample size* 1 *and a SPRT as stopping rule. Here we investigate the difference in the average site numbers for different values of the fixed sample size. For each sample size, we may now calculate the optimal SPRT using the integral equations from Chapter 2.*

Let us consider a sample path with sample size n. The increment of the log-likelihood ratio has a Gaussian distribution with variance $n\mu^2$. Under the null hypothesis, the mean is equal to $-n\mu^2/2$ and under the alternative hypothesis the mean is equal to $n\mu^2/2$. The integral equations from which the error probabilities may be obtained are

$$P_i(z) = G_i(A - z) + \int_A^B P_i(x) dG_i(x - z)$$

where $G_i(x)$ is the probability distribution function of the log-likelihood ratio, given that f_i is the correct density function. Since in this particular case we have the relation

$$G_1(x) = 1 - G_0(-x)$$

we may show that

$$P_1(z) = 1 - P_0(-z)$$

if $A = -B$. This implies that to find the optimal SPRT when $\alpha = \gamma$ we only have to consider the symmetric SPRT's. The resulting average site numbers are given by $N_0(0)$ and $N_1(0)$, where

$$N_i(z) = n + \int_A^B N_i(x) dG_i(x - z)$$

Due to the symmetry of the tests we know that $N_0(z) = N_1(-z)$, so that the average site numbers are the same for both hypotheses. In Figure 5.1 the average site number is shown as a function of the sample size, for $\mu = 1$. Indeed, in accordance with Theorem 5.1, the average site number is minimal for $n = 1$. In fact, the average site number is increasing with n and converges to n rather rapidly. Figure 5.2 illustrates the effect of the mean μ on the average site number. Clearly, a sample size equal to one is used to calculate the average site number for each value of the mean. We may see that the average site number is decreasing with μ, which is not very surprising.

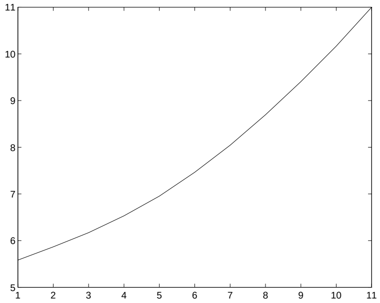

Figure 5.1: Average site number as function of sample size for $\mu = 1$

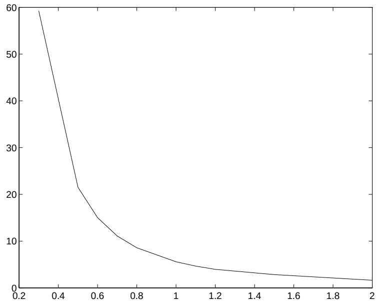

Figure 5.2: Average site number as function of mean

In case the index set consists of more than one site, i.e., if there exist regions on the plane where the change has a different effect, Theorem 5.1 no longer holds. However, in many situations we may find a site for which the detection quality is better than for other sites. Using the terminology of Example 5.3, the site with the darkest soil may have this property. In such a case, the UME stopping strategy uses a sample path that is concentrated at one particular site. Hence, the actual problem is equivalent to a detection problem with a single site.

Definition 5.4 (Most informative) *A site t is said to be most informative in $\mathcal{S} \subset \mathcal{R}_{\alpha,\gamma}$ if*

$$N_i(\mathbf{d}_t) \leq N_i(\mathbf{d})$$

for all $\mathbf{d} \in \mathcal{S}$ and for $i = 0, 1$, where $\mathbf{d}_t$ denotes the UME stopping strategy concentrated at site t.

Clearly, if t is most informative in $\mathcal{R}_{\alpha,\gamma}$, then the UME stopping strategy is given as in Theorem 5.1.

Example 5.5 *For a random field whose index set consists of an arbitrary number of sites, where the process has a Gaussian distribution at each site,* *there always exists a site that is most informative in $\mathcal{R}_{\alpha,\gamma}$. In fact, the site with the largest Kullback-Leibler divergence may be shown to be most informative.*

$$\begin{aligned} J_t &= \mathbf{E}_1 \log \frac{f_t}{f_0} - \mathbf{E}_0 \log \frac{f_t}{f_0} \\ &= \mu_t^2/\sigma_t^2 \end{aligned}$$

From Example 5.4 we know that the average site number decreases with μ if $\sigma = 1$. For other values of the variance, the quantity μ/σ plays the role of the mean, hence explaining the importance of the Kullback-Leibler divergence in this case.

If more than one site have the same Kullback-Leibler divergence, each one of these sites is most informative. In fact, these sites are statistically equivalent so that they may be joined in one site.

We may ask ourselves whether a simple hypothesis problem exists for which the random field does not have a most informative site. Indeed, it is not easy to find a counterexample. Suppose that a certain site is optimal at a certain stage. After taking a measurement at this site, something has to change such that this site no longer is optimal. We have to be able to find a certain variable that indicates which site should be chosen next.

For example, consider the following hypothetical situation. Under the null hypothesis, a certain site t is most informative. Under the alternative hypothesis, another site s is most informative. Since we initially do not know which hypothesis is true, we may randomly select the first site. Depending on the outcome of the first measurement, the null hypothesis becomes more or less likely than the alternative hypothesis. If the null hypothesis becomes more likely, the next measurement should then be taken at site t. Otherwise, site s should be used to take the next measurement. The variable we were looking for appears to be given by the values of the likelihood ratios ℓ_t and ℓ_s. If they both are smaller than 1, site t should be chosen next, if they both are larger than 1, site s should be chosen next. However, in the remaining part of the state space, where one of the likelihood ratios is larger than 1 and one of them is smaller than 1, it is not clear what to do.

Exhaustive observation mechanisms

So far we only considered inexhaustive observation mechanisms. If the observation mechanism is exhaustive, the problem changes rather drastically. Clearly, since the resources of the random field change with each measurement, the optimal stopping strategy has to depend on the resources. This implies that, even for random fields whose index set consists of one site, Theorem 5.1 does not apply anymore. Due to the finiteness of the resources, we may use an iterative approach to solve the quickest detection problem.

Suppose we have a random field with resources M. Let us first consider the stopping strategy that exhausts the resources using only one sample. That is, all possible measurements are made, and based on these measurements a decision is made. From Chapter 2 we know that the optimal decision in this case has to be based on a likelihood ratio test. Let us denote $\mathbf{d}_\lambda$ as this stopping strategy, where λ is the threshold of the likelihood ratio test. The following lemma tells us that this stopping strategy is the one with the smallest possible error probabilities.

Lemma 5.1 *For every stopping strategy* $\mathbf{d}$ *there exists a stopping strategy* $\mathbf{d}_\lambda$ *based on a likelihood ratio test that exhausts the resources of the random field, such that either*

$$\begin{aligned}\alpha(\mathbf{d}) &= \alpha(\mathbf{d}_\lambda)\\ \gamma(\mathbf{d}) &\geq \gamma(\mathbf{d}_\lambda)\end{aligned}$$

or

$$\begin{aligned}\alpha(\mathbf{d}) &\geq \alpha(\mathbf{d}_\lambda)\\ \gamma(\mathbf{d}) &= \gamma(\mathbf{d}_\lambda)\end{aligned}$$

Proof. Consider an arbitrary stopping strategy $\mathbf{d} = (\mathbf{t}, \tau, \delta)$. According to this stopping strategy, each point $\mathbf{y}$ in the measurement space is located either in the acceptance region $\Omega_a(\mathbf{d})$ or the rejection region $\Omega_r(\mathbf{d})$. It follows that

$$\begin{aligned}\alpha(\mathbf{d}) &= F_0(\Omega_r(\mathbf{d}))\\ \gamma(\mathbf{d}) &= F_1(\Omega_a(\mathbf{d}))\end{aligned}$$

The rejection region for the stopping strategy $\mathbf{d}_\lambda$ is given by

$$\Omega_r(\mathbf{d}_\lambda) = \{\mathbf{y}|f_1(\mathbf{y}) > \lambda f_0(\mathbf{y})\}$$

and the acceptance region is its complement. Suppose that $\alpha(\mathbf{d}) > 0$. (If this does not hold, then certainly $\gamma(\mathbf{d}) > 0$, and the following also holds with the error probabilities interchanged.) From the absolute continuity of the density functions it follows that there exists a λ for which

$$\alpha(\mathbf{d}_\lambda) = \alpha(\mathbf{d})$$

The difference in the miss probabilities for these two tests is now given by

$$\begin{aligned}\gamma(\mathbf{d}) - \gamma(\mathbf{d}_\lambda) &= F_1(\Omega_a(\mathbf{d})) - F_1(\Omega_a(\mathbf{d}_\lambda))\\ &= \int_{S(\mathbf{d},\mathbf{d}_\lambda)} f_1(\mathbf{y})d\mathbf{y} - \int_{S(\mathbf{d}_\lambda,\mathbf{d})} f_1(\mathbf{y})d\mathbf{y}\end{aligned}$$

$$\begin{aligned}
&> \int_{S(\mathbf{d},\mathbf{d}_\lambda)} \lambda f_0(\mathbf{y})d\mathbf{y} - \int_{S(\mathbf{d}_\lambda,\mathbf{d})} \lambda f_0(\mathbf{y})d\mathbf{y} \\
&= \lambda\{\int_{\Omega_a(\mathbf{d})} f_0(\mathbf{y})d\mathbf{y} - \int_{\Omega_a(\mathbf{d}_\lambda)} f_0(\mathbf{y})d\mathbf{y}\} \\
&= \lambda(1-\alpha(\mathbf{d}) - 1 + \alpha(\mathbf{d}_\lambda)) \\
&= 0
\end{aligned}$$

where $S(\mathbf{d}_1, \mathbf{d}_2)$ denotes the part of $\Omega_a(\mathbf{d}_1)$ that is not contained in $\Omega_a(\mathbf{d}_2)$. The inequality follows from the fact that $S(\mathbf{d}, \mathbf{d}_\lambda)$ is contained in $\Omega_r(\mathbf{d}_\lambda)$. It follows that the error probabilities are equal if and only if the stopping regions are exactly the same. □

Clearly, this lemma is only of minor practical use, since the average site numbers for this stopping strategy are maximal. However, it does give an interesting result for the class $\mathcal{R}_{\alpha,\gamma}$. Since, if there does not exist a threshold λ for which the error probabilities are smaller than α and γ, there does not exist any other strategy for which these inequalities are satisfied. Hence, the class $\mathcal{R}_{\alpha,\gamma}$ is nonempty if and only if there exists a λ such that $\mathbf{d}_\lambda$ is contained in this class.

Example 5.6 *For the detection of a change in the mean of a Gaussian random field with one site only, Lemma 5.1 gives us a lower bound on the size of the resources. If the resources are smaller than this lower bound, the class $\mathcal{R}_{\alpha,\gamma}$ is empty.*

The error probabilities when the resources are given by M are given by

$$\begin{aligned}
\alpha(\mathbf{d}_\lambda) &= 1 - \Phi(\frac{\lambda + M\mu^2/2}{\sqrt{M}\mu}) \\
\gamma(\mathbf{d}_\lambda) &= \Phi(\frac{\lambda - M\mu^2/2}{\sqrt{M}\mu})
\end{aligned}$$

From these equations we may find the minimal values the resources should have to guarantee the existence of a stopping rule in the class $\mathcal{R}_{\alpha,\gamma}$. In Figure 5.3 the minimal values of the resources are plotted as a function of μ, for $\alpha = \gamma = 0.05$.

Suppose now that $\mathbf{d}_\lambda \in \mathcal{R}_{\alpha,\gamma}$ for some λ. This stopping strategy is UME if and only if it is the only stopping strategy in $\mathcal{R}_{\alpha,\gamma}$. Therefore, we would like to know what other strategies are contained in this class.

Considering the complexity of this problem, we restrict our index set to a one-site index set with resources M. Clearly, if M goes to infinity the observation mechanism becomes inexhaustive and the optimal stopping rule is a SPRT. For large values of M, the stopping rule for an exhaustive observation

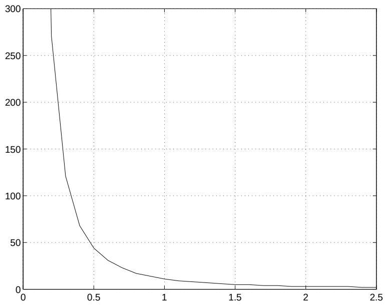

Figure 5.3: Minimal resources as function of mean ($\alpha = \gamma = 0.05$)

mechanism may have similar characteristics. If we use a truncated SPRT as our stopping rule, we may obtain some reasonable results. However, it is more likely that the optimal stopping rule uses a GSPRT with decreasing upper bound and increasing lower bound. Since, when the observation mechanism is almost exhausted the probability that the likelihood ratio changes considerably decreases rapidly. As a result, if this likelihood ratio is close to one of the boundaries, the final decision already is known with large probability. Hence, this decision may as well be made at this stage, thus decreasing the continuation region.

This problem has also been considered by Ghosh [11, Section 5.3], among others. Unfortunately, no solution has been found yet. Ghosh discusses some approaches to this problem that have been used in the literature. Basically, these approaches are based on GSPRT's with thresholds that are defined by some special functions.

For an index set consisting of more than one site, a similar approach seems to be the best we can do. Generally, the random field will have a most informative site, in which case the problem becomes exactly the same. Otherwise, we may still use a similar GSPRT with a sample path that is not concentrated at one site.

5.2.2 Composite hypothesis

Let us assume the alternative hypothesis to be composite, i.e., the parameter set Θ consists of more than one element. We know from Chapter 2 that generally there does not exist a UME stopping rule in case the sample path is deterministic. Although the class of sample paths is larger in this case, we may generally not be able to find a UME stopping strategy in this case either. For example, for two changes that appear at different sites in the index set, the individual most efficient stopping strategies are concentrated at these sites. Any combination of the two stopping strategies leads to an increase in at least one of the average site numbers, so that we may never find a UME stopping strategy for the composite problem.

Throughout the remainder of this section we limit our attention to the so-called *myopic* sample paths. These sample paths originate from the theory of search (Stone [42]) and bandit problems (Berry and Fristedt [4], Gittins [12]). In the context of bandit problems, where the objective is to maximize the outcome of a sequence of experiments, the myopic sample path is based on a *stay-with-the-winner* principle. The theory of search is closely connected to our detection problem. A classical example is the drawer problem, where a certain object that is hidden in one of the drawers is to be found as soon as possible. The difference with our problem lies in the distributions; in search theory, the probability of detection generally is assumed to be known exactly. In the discrete case, the probability of detection on the kth look in the jth drawer is given. Myopic sample paths are based on an optimization of the immediate result. That is, we imagine that the next measurement will be the last one. The next site is then chosen as the site that minimizes the error probabilities.

In our case only the stochastic nature of the random field under both hypotheses is known, and the probability of detection is only implicitly defined. Nevertheless, we may define a myopic sample path in a similar way. The formal definition depends on the specific situation, and is therefore not given here. Although these sample paths may be optimal in many situations, and have an intuitively appealing character, their optimality does not hold in general. A stopping strategy for which the sample path is myopic is also said to be a myopic stopping strategy.

To approach the problem from another point of view, we assume a prior distribution on the parameter set Θ to be known. This distribution is defined conditional on the alternative hypothesis, i.e.,

$$\xi(\theta) = \mathbf{Pr}[\boldsymbol{\theta} = \theta | \boldsymbol{\theta} \in \Theta]$$

for all $\theta \in \Theta$. In a way, this transforms the alternative hypothesis into a simple hypothesis. The resulting average site number under the alternative hypothesis is given by

$$N_1(\mathbf{d}) = \sum_{\theta \in \Theta} \xi(\theta) N_\theta(\mathbf{d})$$

The error constraints are now given by

$$\begin{aligned} \mathbf{Pr}[\delta = 1 | \boldsymbol{\theta} = 0] &\leq \alpha \\ \sum_{\theta \in \Theta} \xi(\theta) \mathbf{Pr}[\delta = 0 | \boldsymbol{\theta} = \theta] &\leq \gamma \end{aligned}$$

In the following sections, we examine these approaches for some specific situations.

Disjoint regions of change

Suppose a change may be present at one site only. This implies that each parameter θ represents the presence of a change at one of the sites. In other words, the parameter set Θ may be chosen equal to the set of sites in the index set. The parameter $\theta = t$ represents the presence of a change at site t only. We assume that under the null hypothesis the process has density function f_0 at each site. Under the alternative hypothesis, the density functions are given by

$$f_{\theta,t}(y) = \begin{cases} f_1(y) & \text{if } \theta = t \\ f_0(y) & \text{otherwise} \end{cases}$$

Intuitively, we may be inclined to use a myopic strategy; at every stage choose the most likely site. Since in this case each change may be linked to a particular site, the likelihood ratio for each change indicates the likelihood of a site. Hence, for each change (or site) we have to store the (log-)likelihood ratio.

Since a change may only be present at one site, as soon as one of the likelihood ratios becomes large we may accept the presence of a change at that site. On the other hand, if one of the likelihood ratios becomes very small, we may reject the presence of a change at that site. However, in that case there may still be a change at one of the other sites. Hence, we reject the null hypothesis as soon as one of the likelihood ratios grows beyond a certain upper threshold, and accept the null hypothesis as soon as all likelihood ratios are smaller than a certain lower threshold. This implies that the stopping rule is based on a generalized likelihood ratio test, as defined in Chapter 2. Note that this is not a UME stopping strategy, but merely an intuitively appealing stopping strategy that is not too complex.

Example 5.7 *The gold coin problem from Example 5.1 clearly is of this type. Each square represents a different site, and the jar is the change which may only be present at one site. Clearly, the null hypothesis is that the rumors are nothing but rumors, so that there is no jar at all.*

For the myopic strategy described here, the probability of rejection may be computed in a similar way as the acceptance probability for the simple hypothesis problem. We define $P_\theta(\mathbf{x})$ to be the probability of rejection given that θ is the true parameter. Here $\mathbf{x}$ denotes the current values of the log-likelihood ratios, i.e., $\mathbf{x}_t = \log \ell_t$. This probability may be written as the sum of the probability of rejection at the next stage and the probability of rejection after the next stage. Let us assume that $\mathbf{x}_t$ is the maximal element of $\mathbf{x}$. Then we may find

$$P_\theta(\mathbf{x}) \quad = \quad 1 - G_{\theta,t}(B - \mathbf{x}_t) + \int_{-\infty}^{B} P_\theta(\mathbf{y}(t, \mathbf{x}, u)) dG_{\theta,t}(u - \mathbf{x}_t)$$

$$\begin{aligned} &= 1 - G_{\theta,t}(B - \mathbf{x}_t) + G_{\theta,t}(A - \mathbf{x}_t)P_\theta(\mathbf{x}^*) \\ &\quad + \int_A^B P_\theta(\mathbf{y}(t,\mathbf{x},u))dG_{\theta,t}(u - \mathbf{x}_t) \end{aligned}$$

where

$$\mathbf{y}_s(t,\mathbf{x},u) = \begin{cases} \mathbf{x}_s & \text{if } s \neq t \\ u & \text{if } s = t \end{cases}$$

Furthermore, $G_{\theta,t}$ equals G_1 if $\theta = t$ and G_0 otherwise. Finally, $\mathbf{x}^*$ is the vector $\mathbf{x}$ where the maximal value is removed. Hence, $\mathbf{x}^*$ contains one element less than $\mathbf{x}$. The probability $P_\theta(\mathbf{x}^*)$ is needed when the site t may be rejected, i.e., if $\mathbf{x}(t)$ becomes smaller than A. Since at least one of the other values of $\mathbf{x}$ is larger than A, this site will never be selected again. If more than one site has maximal likelihood, we use randomization in order to maintain symmetry. So, if there are two sites that both have maximal likelihood ratio, each of these sites is chosen with probability 1/2.

It is not difficult to see that in this specific situation, the problem is symmetric in the sense that

$$P_t(\mathbf{0}) = P_s(\mathbf{0})$$

for all t and s. Therefore, we also use $P_1(\mathbf{0})$ to denote this probability.

The error constraints are now given by

$$\begin{aligned} 1 - P_0(\mathbf{0}) &\leq \alpha \\ P_1(\mathbf{0}) &\leq \gamma \end{aligned}$$

The average site numbers may now be calculated in a similar fashion. This gives us

$$N_\theta(\mathbf{x}) = 1 + G_{\theta,t}(A - \mathbf{x}_t)N_\theta(\mathbf{x}^*) + \int_A^B N_\theta(\mathbf{y}(t,\mathbf{x},u))dG_{\theta,t}(u - \mathbf{x}_t)$$

As for the probabilities P_t, the average site numbers also are symmetric in the sense that

$$N_t(\mathbf{0}) = N_s(\mathbf{0})$$

for all t and s. This implies that the only quantities of interest are $N_0(\mathbf{0})$ and $N_1(\mathbf{0})$, the average site numbers under the null and the alternative hypothesis, respectively.

Example 5.8 *Let us consider a random field whose index set consists of two sites s and t. Using the myopic stopping strategy with stopping boundaries A and B, we obtain the following expression for the probability of rejecting the null hypothesis. If $x < y$,*

$$P_\theta(x,y) = 1 - G_{\theta,t}(B - y) + G_{\theta,t}(A - y)P_\theta^s(x) + \int_A^B P_\theta(x,u)dG_{\theta,t}(u - y)$$

and if $x > y$,

$$P_\theta(x,y) = 1 - G_{\theta,s}(B-x) + G_{\theta,s}(A-x)P_\theta^t(y) + \int_A^B P_\theta(u,y)dG_{\theta,s}(u-x)$$

If $x = y$, we choose each of the sites with the same probability, which gives us

$$\begin{aligned} P_\theta(x,x) &= \frac{1}{2}[2 - G_{\theta,t}(B-x) - G_{\theta,s}(B-x) \\ &\quad + G_{\theta,t}(A-x)P_\theta^s(x) + G_{\theta,s}(A-x)P_\theta^t(x) \\ &\quad + \int_A^B P_\theta(x,u)dG_{\theta,t}(u-x) + \int_A^B P_\theta(u,x)dG_{\theta,s}(u-x)] \end{aligned}$$

Here x and y denote the present value of the log-likelihood ratios $\log \ell_s$ and $\log \ell_t$, respectively. Furthermore, $G_{\theta,t}(x)$ denotes the distribution function of the increment of the log-likelihood ratio, given θ and observing at t. If $\theta = t$, this distribution is $G_1(x)$, otherwise it is $G_0(x)$. Finally, we have

$$P_\theta^t(x) = 1 - G_{\theta,t}(B-x) + \int_A^B P_\theta^t(u)dG_{\theta,t}(u-x)$$

as the probability of rejection when only site t is being used.

Then, for the stopping strategy to be an element of $\mathcal{R}_{\alpha,\gamma}$, we require

$$\begin{aligned} P_0(0,0) &\leq \alpha \\ P_s(0,0) &\leq 1-\gamma \\ P_t(0,0) &\leq 1-\gamma \end{aligned}$$

Since we know that $P_s(0,0) = P_t(0,0)$, the last two constraints actually are the same. From these inequalities we may find all eligible thresholds A and B.

Similarly, we may find

$$N_\theta(x,y) = \begin{cases} 1 + G_{\theta,t}(A-y)N_\theta^s(x) + \int_A^B N_\theta(x,u)dG_{\theta,t}(u-y) & \textit{if } x < y \\ 1 + G_{\theta,s}(A-x)N_\theta^t(y) + \int_A^B N_\theta(u,y)dG_{\theta,s}(u-x) & \textit{if } x > y \end{cases}$$

and, if $x = y$,

$$\begin{aligned} N_\theta(x,x) &= 1 + \frac{1}{2}[G_{\theta,t}(A-x)N_\theta^s(x) + G_{\theta,s}(A-x)N_\theta^t(x) \\ &\quad + \int_A^B N_\theta(x,u)dG_{\theta,t}(u-x) + \int_A^B N_\theta(u,x)dG_{\theta,s}(u-x)] \end{aligned}$$

where

$$N_\theta^t(x) = 1 + \int_A^B N_\theta^t(u)dG_{\theta,t}(u-x)$$

For a Gaussian random field with variance equal to one, and mean equal to zero under the null hypothesis and equal to μ under the alternative hypothesis, the resulting average site numbers and thresholds are shown in Figure 5.4 as functions of μ. For small values of μ, the convergence of the numerical algorithms that were used is so slow that these values are not included in the figures.

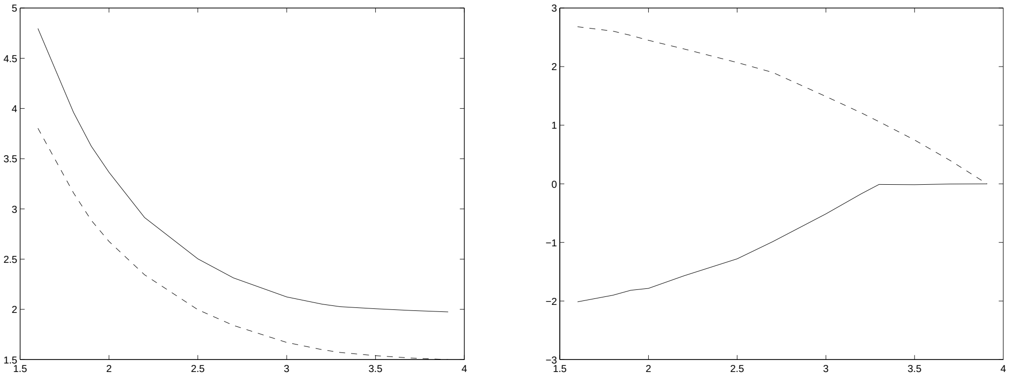

Figure 5.4: Average site numbers (left) and thresholds (right) as function of mean for myopic strategy

As the mean becomes larger, the thresholds A and B approach each other, until at some point they both become equal to zero. For this particular case, the smallest value of μ for which the thresholds are zero may be calculated as 3.909*. We may also see that the lower threshold reaches zero far earlier than the upper threshold. Clearly, this is a consequence of the use of the generalized likelihood ratio. As soon as this threshold reaches zero, the miss probability of the test becomes smaller than γ. Similarly, for values of μ larger than* 3.909*, the false alarm probability also becomes smaller than α.*

In the figure on the left, the average site number under the null hypothesis is shown by the solid line and the average site number under the alternative hypothesis is shown by the dashed line. We may see that the N_0 is always larger than N_1. This is a result of the use of the generalized likelihood ratio test; only one measurement may be sufficient to reject the null hypothesis, while each site has to be visited at least once before the null hypothesis may be accepted.

Instead of using the generalized likelihood ratio, another approach to the problem is to use the averaged likelihood ratio. This approach assumes the knowledge of a prior distribution function on Θ. However, assuming the presence of a distribution on the parameter set has some more consequences.

Let us assume the conditional probability distribution of the parameters to be known;

$$\xi(t) = \mathbf{Pr}[\boldsymbol{\theta} = t | \boldsymbol{\theta} \in \Theta]$$

A myopic sample path may now be interpreted as the sample path that chooses the next site as the one with the largest probability $\xi(t)$. Since these probabilities are updated using the likelihood ratios ℓ_t, the myopic sample path is essentially the same as the myopic sample path in the previous case. The only difference is that the initial values of the likelihood ratio now are determined by the prior distribution ξ.

The assumption of a prior distribution on Θ allows us to use the averaged likelihood ratio instead of the generalized likelihood ratio. The averaged likelihood ratio test may be written as

$$\sum_t \ell_t(\mathbf{y}) \in (a, b)$$

where $\mathbf{y}$ is the vector containing all measurements made so far, and ℓ_t is the likelihood ratio at site t,

$$\ell_t(\mathbf{y}) = \xi(t) \prod_i \ell(y_i)$$

Here $\ell(y) = f_1(y)/f_0(y)$ and i is the index of those elements of $\mathbf{y}$ that were measured at site t.

The use of the averaged likelihood ratio instead of the generalized likelihood ratio results in a difference in the shape of the stopping regions. Typical stopping regions for these two tests are shown in Figure 5.5. In this figure, the stopping regions for a problem with two changes are shown in the likelihood ratio space. That is, ℓ_1 is shown on the horizontal axis and ℓ_2 is shown

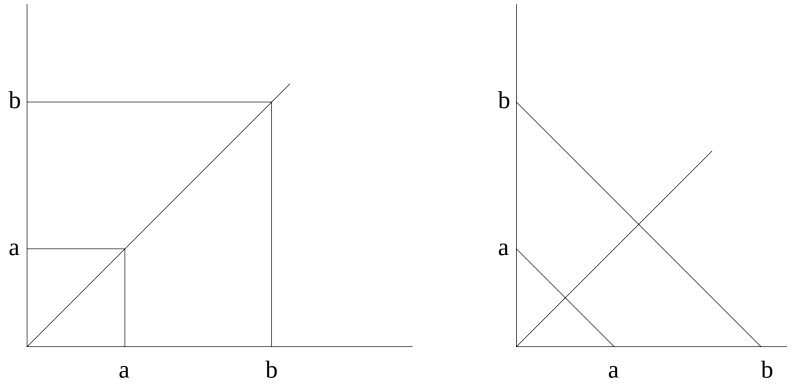

Figure 5.5: Stopping regions for generalized likelihood ratio test (left) and averaged likelihood ratio test (right)

on the vertical axis. For the figure on the left, the procedure starts at the point $(1, 1)$, and for the figure on the right the procedure starts at the point $(\xi_1, \xi_2) = (\xi_1, 1-\xi_1)$. The sampling process may now be regarded as a random walk on this plane that stops on first entrance of one of the stopping regions.

The error probabilities may be calculated using the same approach as before. Denote $P_\theta(\mathbf{x})$ as the probability of rejecting the null hypothesis given that $\boldsymbol{\theta} = \theta$ and when the current values of the likelihood ratios are given by $\mathbf{x}$, i.e., $\ell_t = \mathbf{x}_t$. Suppose now that $\mathbf{x}_t$ is maximal, so that the next measurement is taken at site t. To reject the null hypothesis immediately after this next measurement, the updated averaged likelihood ratio should be at least equal to b. That is,

$$\sum_s \mathbf{x}_s + \mathbf{x}_t \ell(y) \geq b$$

If $\ell(y)$ is such that the averaged likelihood ratio is contained in (a, b), we may start the same procedure again. If the distribution of the log-likelihood ratio is known, this gives us the following expression for the probability of rejection.

$$\begin{aligned} P_\theta(\mathbf{x}) &= 1 - G_{\theta,t}(\log \frac{b - x(-t)}{\mathbf{x}_t}) \\ &\quad + \int_{\max\{a, x(-t)\}}^{b} P_\theta(\mathbf{y}(t, \mathbf{x}, u)) dG_{\theta,t}(\log \frac{u - x(-t)}{\mathbf{x}_t}) \end{aligned}$$

where $x(-t) = \sum_{s \neq t} \mathbf{x}_s$ and

$$\mathbf{y}_s(t, \mathbf{x}, u) = \begin{cases} \mathbf{x}_s & \text{if } s \neq t \\ u - x(-t) & \text{if } s = t \end{cases}$$

and $G_{\theta,t}$ denotes the distribution function of the log-likelihood ratio for one single measurement at site t, given that $\boldsymbol{\theta} = \theta$. The lower bound in the integral follows from the fact that $\ell(y)$ may not be negative. If the maximum of $\mathbf{x}$ is not unique, we use randomization.

The error constraints may now be written as

$$\begin{aligned} P_0(\xi) &\leq \alpha \\ 1 - \sum_\theta \xi(\theta) P_\theta(\xi) &\leq \gamma \end{aligned}$$

For the average site number we obtain similar equations. We may write

$$N_\theta(\mathbf{x}) = 1 + \int_{\max\{a, x(-t)\}}^{b} N_\theta(\mathbf{y}(t, \mathbf{x}, u)) dG_{\theta,t}(\log \frac{u - x(-t)}{\mathbf{x}_t})$$

given that $\mathbf{x}_t$ is the maximum of $\mathbf{x}$.

Example 5.9 *Let us consider a random field whose index set consists of two sites t and s. The process is Gaussian under both hypotheses, with variance equal to one. Under the null hypothesis, the mean is equal to zero, and under the alternative hypothesis one of the sites has mean μ. Under the alternative hypothesis, the change is present at site t with probability η and at site s with probability $1-\eta$. Without loss of generality, we assume $\eta \geq 1/2$.*

Although the class of myopic sequential probability ratio tests not necessarily has to contain the optimal stopping strategy, we restrict our attention to this class. The rejection probability may in this case be written as

$$P_\theta(x_t, x_s) = 1 - G_{\theta,t}(\log\frac{b-x_s}{x_t}) + \int_{\max\{a,x_s\}}^{b} P_\theta(u-x_s, x_s) dG_{\theta,t}(\log\frac{u-x_s}{x_t})$$

if $x_t > x_s$,

$$P_\theta(x_t, x_s) = 1 - G_{\theta,s}(\log\frac{b-x_t}{x_s}) + \int_{\max\{a,x_t\}}^{b} P_\theta(x_t, u-x_t) dG_{\theta,s}(\log\frac{u-x_t}{x_s})$$

if $x_t < x_s$, and

$$\begin{aligned} P_\theta(x_t, x_s) \quad = \quad & (2 - G_{\theta,t}(\log\frac{b-x_s}{x_t}) - G_{\theta,s}(\log\frac{b-x_t}{x_s}) \\ & + \int_{\max\{a,x_s\}}^{b} P_\theta(u-x_s, x_s) dG_{\theta,t}(\log\frac{u-x_s}{x_t}) \\ & + \int_{\max\{a,x_t\}}^{b} P_\theta(x_t, u-x_t) dG_{\theta,s}(\log\frac{u-x_t}{x_s}))/2 \end{aligned}$$

if $x_t = x_s$.

We may see that, as μ goes to infinity, the thresholds do not converge to 1. Since if they did, the miss probability would converge to $1-\eta$ instead of 0. This difference with the previous case is caused by the use of the averaged likelihood ratio instead of the generalized likelihood ratio; only one measurement may already result in an acceptance of the null hypothesis, while the other site is not yet examined.

Overlapping regions of change

In general we may not be able to distinguish the regions of change that easily. It may very well be the case that a certain site may be contained in different regions of change. Here we assume the regions of change to have a constant shape. For this particular problem, let us introduce the variable T_c as the site that indicates the position of the region of change. For example, it may be the center site, or one of the corner sites of the region of change. The set of changes may now be parameterized by the sites from the index set. Although not every site may host T_c, we include all sites in Θ for symmetry reasons. The region of change is now denoted as $G_c(T_c)$.

Example 5.10 *The extended gold coins problem of Example 5.2, where the jar is assumed to be broken may be interpreted as a problem of this type. Let us assume that the gold coins are distributed over nine squares that are positioned in a three by three grid. The variable T_c may for example be chosen as the middle square of the nine that contain the gold coins. Obviously, in that case T_c may never be equal to any of the squares on the boundary of the land.*

Under the null hypothesis the field is assumed to be independent and identically distributed with density function f_0 at each site. Under the alternative hypothesis, the sites that are contained in the region of change are also identically distributed. That is,

$$f_{\theta,t}(y) = \begin{cases} f_1(y) & \text{if } t \in \mathcal{G}_c(\theta) \\ f_0(y) & \text{otherwise} \end{cases}$$

In case the regions of change are disjoint, the first measurement may be taken at any site. However, in this case several sites are contained in more than one region of change. Therefore, it seems more efficient to take the first measurement at one of the sites that is contained in as many as possible regions of change. For each change, we may calculate the likelihood ratio based on a measurement y at site t,

$$\ell_\theta(y) = \frac{f_{\theta,t}(y)}{f_0(y)}$$

Then, using initial values equal to 1, we may update the likelihood ratio for each change after each measurement according to

$$\ell_\theta(\mathbf{y}_n) = \ell_\theta(\mathbf{y}_{n-1})\ell_\theta(y)$$

A myopic sample path may now be defined by taking the next measurement at that site for which the sum of the likelihood ratios

$$l(t) = \sum_{\{\theta | t \in \mathcal{G}_c(\theta)\}} \ell_\theta(\mathbf{y}_n)$$

is maximal. This implies that using this approach is equivalent to the assumption that the prior distribution of the changes is uniform.

A logical choice for the stopping rule is the one based on the generalized likelihood ratio. As soon as one of the changes becomes quite likely (the likelihood ratio ℓ_θ becomes larger than a certain threshold for some θ), we stop and reject the null hypothesis. If at a certain stage all of the changes are not likely to be present (ℓ_θ is smaller than a certain threshold for all θ), we stop and accept the null hypothesis. Using this stopping strategy, the probability of rejecting the null hypothesis when θ is the true parameter, and given that $l(t)$ is maximal, is given by

$$\begin{aligned} P_\theta(\mathbf{x}) &= 1 - G_{\theta,t}(B - \max(\mathbf{x})) + G_{\theta,t}(A - \max(\mathbf{x}))P_\theta(\mathbf{x}^*) \\ &\quad + \int_A^B P_\theta(\mathbf{y}(t,\mathbf{x},u))dG_{\theta,t}(u - \max(\mathbf{x})) \end{aligned}$$

where $\mathbf{x}$ is the vector containing the log-likelihood ratios, and

$$\mathbf{y}_\theta(t,\mathbf{x},u) = \begin{cases} \mathbf{x}_\theta & \text{if } t \notin \mathcal{G}_c(\theta) \\ \mathbf{x}_\theta + u - \max(\mathbf{x}) & \text{if } t \in \mathcal{G}_c(\theta) \end{cases}$$

The vector $\mathbf{x}^*$ only consists of the elements $\mathbf{x}_\theta$ for which $t \notin \mathcal{G}_c(\theta)$. If more than one site exists for which $l(t)$ is maximal, we use randomization to proceed.

Alternatively, we may also assume the conditional distribution of the changes to be known.

$$\xi(t) = \mathbf{Pr}[T_c = t|\boldsymbol{\theta} \in \Theta]$$

For those sites that are physically incapable of hosting T_c, this probability will be zero. For example, if T_c denotes the middle site of nine sites ordered in a square, T_c may never be equal to one of the sites on the boundary of the grid.

Furthermore, let us define

$$P(t) = \mathbf{Pr}[t \in \mathcal{G}_c(T_c)|\boldsymbol{\theta} \in \Theta]$$

as the probability that site t is contained in the region of change. Clearly, $P(t)$ plays the role of $l(t)$ in the myopic sample path as described before. In analogy with the disjoint regions of change, we may find similar expressions for the probability of rejection. As before, let us assume that $P(t)$ is maximal, so that the next measurement is taken at site t. Then, with $\mathbf{x}$ denoting the vector of likelihood ratios ℓ_θ,

$$\begin{aligned} P_\theta(\mathbf{x}) &= 1 - G_{\theta,t}(log\frac{b - x(-t)}{x(+t)}) \\ &\quad + \int_{\max\{a,x(-t)\}}^b P_\theta(\mathbf{y}(t,\mathbf{x},u))dG_{\theta,t}(\log\frac{u - x(-t)}{x(+t)}) \end{aligned}$$

where

$$\begin{aligned} x(-t) &= \sum_{\{\theta|t\notin\mathcal{G}_c(\theta)\}} \mathbf{x}_\theta \\ x(+t) &= \sum_{\{\theta|t\in\mathcal{G}_c(\theta)\}} \mathbf{x}_\theta \end{aligned}$$

and

$$\mathbf{y}_\theta(t,\mathbf{x},u) = \begin{cases} \mathbf{x}_\theta & \text{if } t \notin \mathcal{G}_c(\theta) \\ \frac{u-x(-t)}{x(+t)}\mathbf{x}_\theta & \text{if } t \in \mathcal{G}_c(\theta) \end{cases}$$

The error constraints may now be written as

$$\begin{aligned} P_0(\xi) &\leq \alpha \\ 1 - \sum_{\theta\in\Theta} \xi(\theta) P_\theta(\xi) &\leq \gamma \end{aligned}$$

For the average site numbers we find

$$N_\theta(\mathbf{x}) = 1 + \int_{\max\{a,x(-t)\}}^{b} N_\theta(\mathbf{y}(t,\mathbf{x},u)) dG_{\theta,t}(\log \frac{u-x(-t)}{x(+t)})$$

As for the disjoint regions of change, these equations may be solved numerically in order to find the optimal thresholds and the resulting average site numbers.

5.3 Bayesian approach

In case we have some additional prior information about the distribution of the parameters and the costs of making incorrect decisions, the Bayesian approach may be used. Cairoli and Dalang [6] already created a theory to deal with the optimization of multi-parameter processes in a Bayesian setting. Their approach basically is focused towards the multiple-armed bandit problems, where a certain gain has to be optimized by sequentially pulling one of many arms. Translated to our problem, the arms may be regarded as the sites, and the gain is something like a detection quality. In Chapter 6 of Cairoli and Dalang [6], a problem is formulated that gives a better match with our quickest detection problem. The objective of Cairoli and Dalang is to estimate a parameter θ, whereas our objective is to detect a change in the parameter θ. The results given in this section are derived from the results given by Cairoli and Dalang.

As in Chapter 2, a change may be seen as a stochastic variable $\boldsymbol{\theta}$. The prior distribution of the parameters is given by ξ, defined by

$$\xi(\theta) = \mathbf{Pr}[\boldsymbol{\theta} = \theta]$$

for all $\theta \in \Theta^*$. After taking a measurement at a site t, the posterior distribution is denoted by $T_t\xi$. The cost of making an incorrect decision is assumed to be given by

$$K(\theta,d) = \begin{cases} c_0 & \text{if } \theta = 0, d = 1 \\ c_1 & \text{if } \theta \neq 0, d = 0 \\ 0 & \text{otherwise} \end{cases}$$

where c_0 and c_1 are the cost of false alarm and the miss cost, respectively. The sampling cost is equal to the number of measurements that are taken.

Using a stopping strategy $\mathbf{d} = (\mathbf{t}, \tau, \delta)$, in analogy with Section 2.2.3 this gives us an expected cost

$$L(\xi, \mathbf{d}) = \mathbf{E}_\xi[\tau + K(\boldsymbol{\theta}, \delta)]$$

The main objective throughout this section is to minimize this quantity.

Definition 5.5 (Bayes cost) *The Bayes cost is defined as*

$$\rho(\xi) = \inf_{\mathbf{d}} L(\xi, \mathbf{d})$$

In other words, the Bayes cost is the minimal expected cost when ξ is the prior distribution of the parameters.

Definition 5.6 (Bayes stopping strategy) *The stopping strategy that minimizes the expected cost is said to be the Bayes stopping strategy, and is denoted by $\mathbf{d}^B(\xi)$.*

As in the one-parameter case, the following lemma holds.

Lemma 5.2 (Concavity of the Bayes cost) *The Bayes cost $\rho(\xi)$ is concave in ξ.*

Proof. We may easily see that $L(\xi, \mathbf{d})$ is linear in ξ. Then,

$$\begin{aligned}
\rho(\lambda\xi_1 + (1-\lambda)\xi_2) &= \inf_{\mathbf{d}} L(\lambda\xi_1 + (1-\lambda)\xi_2, \mathbf{d}) \\
&= \inf_{\mathbf{d}} \{\lambda L(\xi_1, \mathbf{d}) + (1-\lambda) L(\xi_2, \mathbf{d})\} \\
&\geq \lambda \inf_{\mathbf{d}} L(\xi_1, \mathbf{d}) + (1-\lambda) \inf_{\mathbf{d}} L(\xi_2, \mathbf{d}) \\
&= \lambda\rho(\xi_1) + (1-\lambda)\rho(\xi_2)
\end{aligned}$$

which completes the proof. □

The optimization problem in this context is more complicated than the optimization problem for one-parameter processes, due to the large number of possible sample paths that may be used. In Section 5.1 we defined a site-by-site sample path as a sample path that takes one measurement only in each sample. This implies that each sample is taken at one particular site, thus explaining the name site-by-site sample path. The following theorem shows that we may restrict our attention to the class of site-by-site sample paths.

Theorem 5.2 (Completeness of site-by-site class) *The class of site-by-site sample paths is complete in the sense that any sample path may be written as a site-by-site sample path.*

Proof. We show that for each sample path there exists a site-by-site sample path for which a stopping rule exists that results in the same cost. Let us start at the first sample. Divide this sample of the sample path into its individual sites, where these sites are arbitrarily ordered. For all sites but the last one, we continue with probability 1 and select the next site according to the ordering we chose. At the last site, the same stopping rule is applied as for the original sample path. The next sample is then treated just like the previous one.

Clearly, this gives us the same cost as for the original sample path. □

Note that this theorem only involves the class of sample paths and not the class of stopping strategies. It only says that we do not have to consider sample paths that contain more than one site in a sample, since there always will exist a site-by-site sample path with the same (or even smaller) cost. Hence, from now on we only use the class of site-by-site sample paths.

As before we may distinguish two different situations. The first one deals with random fields with inexhaustive observation mechanisms, that is, from each site an unlimited number of measurements may be taken without losing the independence property. The second situation is based on fields with exhaustive observation mechanisms, where only a limited number of measurements may be taken from each site. The vector M contains the resources, i.e. the number of measurements that may be taken from each site. So, from site t we may take M_t measurements. We denote $M(t)$ as the vector M, where M_t is decreased by 1, i.e.,

$$M_s(t) = \begin{cases} M_t - 1 & \text{if } s = t \\ M_s & \text{otherwise} \end{cases}$$

Furthermore, t is said to be an element of M ($t \in M$), if $M_t > 0$. Hence, if the site t is an element of M, there are some resources available at t. Since the resources of a random field constrain the sample paths that may be used, we introduce the class $\mathcal{R}_M$ as those stopping strategies for which the sample paths are constrained by M.

The Bayes cost and the Bayes stopping strategies may be obtained from the following theorem.

Theorem 5.3 (Bayes cost) *For the random field with inexhaustive observation mechanism, the Bayes cost is given by*

$$\rho(\xi) = \min\{\rho_0(\xi), 1 + \min_t \mathbf{E}_\xi[\rho(T_t\xi)]\} \tag{5.1}$$

For the random field with exhaustive observation mechanism, the Bayes cost is given by

$$\rho^M(\xi) = \min\{\rho_0(\xi), 1 + \min_{t\in M} \mathbf{E}_\xi[\rho^{M(t)}(T_t\xi)]\} \tag{5.2}$$

Proof. Let us start with the exhaustive observation mechanism. The proof follows from induction on the sum of the elements of M, denoted by $|M|$. Clearly, if $|M| = 0$ the Bayes cost equals $\rho_0(\xi)$, since there are no sites to process. Now suppose that for all M for which $|M| \leq k$ the Bayes cost is given by equation (5.2). Let us construct M' by increasing one of the elements of M by 1, that is $M'(t) = M$ for some t. Clearly, we have to compare the cost of stopping with the cost of continuation. If we continue processing, we select that site for which the expected Bayes cost will be minimal. Since this expected Bayes cost is calculated with $|M| = k$, equation (5.2) may be used by induction. As a result, the Bayes cost may be written as

$$\rho^{M'}(\xi) = \min\{\rho_0(\xi), 1 + \min_{t \in M'} \mathbf{E}_\xi \rho^{M'(t)}(T_t\xi)\}$$

Since this holds for all M' for which $|M'| = k + 1$, the proof is complete.

The result for the inexhaustive observation mechanism follows from the exhaustive one by letting M go to infinity, in the sense that $M_t \to \infty$ for all t. To prove that $\rho^M(\xi)$ converges to $\rho(\xi)$ let us consider the class of stopping strategies $\mathcal{R}_M$. The function $\rho^M(\xi)$ is decreasing with M in the sense that if the resources are growing, then the Bayes cost decreases. If M' is larger than M, i.e. $M'_t \geq M_t$ for all t, then clearly $\mathcal{R}_M \subset \mathcal{R}_{M'}$. It follows that

$$\begin{aligned} \rho^M(\xi) &= \inf_{\mathbf{d} \in \mathcal{R}_M} L(\xi, \mathbf{d}) \\ &\geq \inf_{\mathbf{d} \in \mathcal{R}_{M'}} L(\xi, \mathbf{d}) \\ &= \rho^{M'}(\xi) \end{aligned}$$

Since $\rho^M(\xi)$ is also bounded from below by zero, it has to converge. If M_t goes to infinity for each t, the class $\mathcal{R}_M$ clearly converges to the class $\mathcal{R}$, so that indeed $\rho^M(\xi)$ converges to $\rho(\xi)$. □

The Bayes stopping strategies may now be described as follows.

- The sample path $\mathbf{t}$ follows from the minimization over the future Bayes cost.
- The stopping time τ is given by the first time for which $\rho(\xi) = \rho_0(\xi)$.
- The decision function δ is the Bayes decision function as defined in Section 2.1.2.

Finding the Bayes sample path and the Bayes stopping time are rather complex tasks. For exhaustive observation mechanisms, they may be calculated through extensive numerical computation. For inexhaustive observation mechanisms, this calculation is less straightforward. However, the infiniteness of the resources also has some advantages. In the remainder of this section we give some theoretical results for random fields with inexhaustive observation mechanisms that will be used in later sections.

From Theorem 5.3 we may see that the vector ξ gives sufficient information for our stopping strategy. Hence, the space of all possible values of ξ, which we denote as S,

$$S = \{\xi | \xi(\theta) \geq 0, \sum_{\theta \in \Theta^*} \xi(\theta) = 1\}$$

may be divided in several regions.

The acceptance region is given by

$$S^a = \{\xi | \rho(\xi) = (1 - \xi(0))c_1\}$$

and the rejection region is given by

$$S^r = \{\xi | \rho(\xi) = \xi(0)c_0\}$$

Then, the continuation region C is the complement of the union of these two regions.

Theorem 5.4 (Convexity of stopping regions) *Both the acceptance region S^a and the rejection region S^r are convex.*

Proof. Suppose that ξ_1 and ξ_2 are both contained in S^r. We now show that $\lambda\xi_1 + (1-\lambda)\xi_2$ is also contained in S^r. From the concavity of $\rho(\xi)$, we know that

$$\rho(\lambda\xi_1 + (1-\lambda)\xi_2) \geq \lambda\rho(\xi_1) + (1-\lambda)\rho(\xi_2)$$

On the other hand, by definition we have

$$\rho(\lambda\xi_1 + (1-\lambda)\xi_2) \leq (\lambda\xi_1(0) + (1-\lambda)\xi_2(0))c_0$$

The right hand side of the last inequality may be written as

$$\begin{aligned}(\lambda\xi_1(0) + (1-\lambda)\xi_2(0))c_0 &= \lambda\xi_1(0)c_0 + (1-\lambda)\xi_2(0)c_0 \\ &= \lambda\rho(\xi_1) + (1-\lambda)\rho(\xi_2)\end{aligned}$$

where the last equality follows from the fact that ξ_1 and ξ_2 are contained in S^r. Combining both inequalities, we may see that

$$\rho(\lambda\xi_1 + (1-\lambda)\xi_2) = (\lambda\xi_1(0) + (1-\lambda)\xi_2(0))c_0$$

which implies that $\lambda\xi_1 + (1-\lambda)\xi_2$ is contained in S^r. Hence, S^r is convex.

The proof for S^a is similar, and is omitted here. □

Note that this theorem also holds for random fields with exhaustive observation mechanisms.

The two hyperplanes that form the boundaries between the stopping regions and the continuation region are defined by

$$\begin{aligned} B^r &= \{\xi | \xi(0)c_0 = 1 + \min_t \mathbf{E}_\xi \rho(T_t \xi)\} \\ B^a &= \{\xi | (1-\xi(0))c_1 = 1 + \min_t \mathbf{E}_\xi \rho(T_t \xi)\} \end{aligned}$$

Finding these boundaries is equivalent with solving the entire problem.

Inside the continuation region, the choice for the next site to be processed is based on the expectation of the future Bayes cost. Let us define

$$\varrho(\xi, t) = \mathbf{E}_\xi \rho(T_t \xi)$$

Clearly, the Bayes sample path may be obtained from minimizing this function over t for all ξ.

Definition 5.7 (Most informative) *A site t is said to be most informative at ξ if*

$$\varrho(\xi, t) = \min_s \varrho(\xi, s)$$

It is denoted as $t_\infty(\xi)$.

In case there exists more than one site for which the future cost is minimal, the most informative site $t_\infty(\xi)$ is no longer uniquely defined. Since at this stage each of these sites give the same expected future cost, we may select any one of these sites to take our next measurement. Hence, we should use randomization to select the next site. In that case we may assume $t_\infty(\xi)$ to be a random variable with uniform distribution on the set of sites for which the expected future cost is minimal.

Furthermore, let $\varrho(\xi) = \varrho(\xi, t_\infty(\xi))$ denote the future Bayes cost. The continuation region may now be divided in a number of sub-regions, according to which site is most informative. Let us define

$$C(t) = \{\xi \in C | t_\infty(\xi) = t\}$$

Clearly, $C(t)$ denotes the part of the continuation region where t is most informative. Hence, if $\xi \in C(t)$ at some stage, the next measurement will be taken at site t. On the boundaries between the different continuation regions randomization is used to select the next site.

5.3.1 Approximate solution

In many situations, the number of sites that will be processed is limited. For example, if the cost of error is relatively small, it is not likely that an extremely large number of sites is required to make a decision. Therefore, a good approximation of the Bayes cost may be found by using a truncated test.

Let us define

$$\rho_k(\xi) = \inf_{\mathbf{d}\in\mathcal{R}_k} L(\xi,\mathbf{d})$$

as the k-truncated Bayes cost, where $\mathcal{R}_k$ is the class of stopping strategies that are truncated at or before the kth measurement.

We may define the following truncated equivalents of $\varrho(\xi,t)$ and $t_\infty(\xi)$,

$$\begin{aligned}
\varrho_k(\xi,t) &= \mathbf{E}_\xi \rho_{k-1}(T_t\xi) \\
t_k(\xi) &= \{t|\varrho_k(\xi,t) = \min_s \varrho_k(\xi,s)\} \\
\varrho_k(\xi) &= \varrho_k(\xi,t_k(\xi))
\end{aligned}$$

where the same randomization is used in case the minimum is not unique.

Theorem 5.5 *The k-truncated Bayes cost is concave in ξ, decreasing with k and converges with k to the Bayes cost. It is given by*

$$\rho_k(\xi) = \min\{\rho_0(\xi), 1 + \min_t \mathbf{E}_\xi \rho_{k-1}(T_t\xi)\}$$

Furthermore, the k-truncated Bayes stopping strategy converges to the Bayes stopping strategy as k goes to infinity.

Proof. The concavity of the truncated Bayes cost immediately follows from the linearity of L, as for the original Bayes cost.

We may easily see that the class $\mathcal{R}_k$ is contained in the class $\mathcal{R}_{k+1}$. Hence,

$$\begin{aligned}
\rho_k(\xi) &= \inf_{\mathbf{d}\in\mathcal{R}_k} L(\xi,\mathbf{d}) \\
&\geq \inf_{\mathbf{d}\in\mathcal{R}_{k+1}} L(\xi,\mathbf{d}) \\
&= \rho_{k+1}(\xi)
\end{aligned}$$

which proves that the truncated Bayes cost is decreasing with k. Since the cost is bounded from below (by zero), this implies that the Bayes cost has to converge for each ξ. Since the class $\mathcal{R}_k$ converges to $\mathcal{R}$, the truncated Bayes cost has to converge to the Bayes cost. From the same argument we may conclude that the truncated Bayes stopping strategy converges to the Bayes stopping strategy.

The iterative solution of the truncated Bayes cost follows as for the original Bayes cost, with the exception that, after taking a next measurement, the remaining number of possible measurements is reduced by one. This gives us

$$\rho_k(\xi) = \min\{\rho_0(\xi), 1 + \min_t \mathbf{E}_\xi \rho_{k-1}(T_t\xi)\}$$

□

We may have noted the similarity between the use of truncation and the use of finite resources. However, for random fields whose index set contains more than one site, there is a clear difference between both situations. In fact, it is not difficult to see that both situations are equal if and only if either $|M| = 0$ and $k = 0$ or $M_t = \infty$ for all t and $k = \infty$. For any other situation, we may easily see that the classes $\mathcal{R}_M$ and $\mathcal{R}_k$ may never be the same.

The stopping regions corresponding with the k-truncated Bayes cost may now be written as

$$\begin{aligned} S_k^a &= \{\xi | (1-\xi(0))c_1 = \rho_k(\xi)\} \\ S_k^r &= \{\xi | \xi(0)c_0 = \rho_k(\xi)\} \end{aligned}$$

and, similarly, the corresponding continuation region is denoted as C_k, which again may be divided in sub-regions $C_k(t)$.

The nature of these approximations to the original stopping regions follows from the following lemma.

Lemma 5.3 *The convex regions S_k^a and S_k^r are decreasing with k and converge to S^a and S^r, respectively.*

Proof. First note that $\rho_k(\xi)$ denotes the Bayes cost for a k-truncated sampling strategy. Since any k-truncated sampling strategy is part of the class of $(k+1)$-truncated sampling strategies, it follows that $\rho_k(\xi) \geq \rho_{k+1}(\xi)$. It then easily follows that $S_{k+1}^a \subset S_k^a$ and $S_{k+1}^r \subset S_k^r$. The convergence to S^a and S^r follows from the fact that $\rho_k(\xi)$ converges to $\rho(\xi)$.

The convexity of the stopping regions may be proven similarly as in the proof of Theorem 5.4. □

For the hyperplanes B^r and B^a we may find similar approximations. Define

$$\begin{aligned} B_k^r &= \{\xi | \xi(0)c_0 = 1 + \varrho_k(\xi)\} \\ B_k^a &= \{\xi | (1-\xi(0))c_1 = 1 + \varrho_k(\xi)\} \end{aligned}$$

Clearly, these hyperplanes also converge to B^r and B^a, respectively.

Let us now define the n-step look-ahead Bayes stopping strategy as the stopping strategy that follows from the minimization of $\rho_n(\xi)$ after each measurement. So, after each measurement, we consider at most n future measurements in our minimization. Clearly, if we let n go to infinity this strategy converges to the Bayes stopping strategy. Since we apply the same test at each stage, the stopping regions S_n^r and S_n^a are constant. Moreover, from Lemma 5.3 we know that the optimal stopping regions are contained in these stopping regions.

Let us consider the 1-step look-ahead procedure. The stopping boundaries B_1^a and B_1^r may be obtained rather easily. They are implicitly given by

$$\begin{aligned} \xi(0)c_0 &= 1+\varrho_1(\xi) && (5.3)\\ (1-\xi(0))c_1 &= 1+\varrho_1(\xi) && (5.4) \end{aligned}$$

Let us define $S_1^a(t)$ and $S_1^r(t)$ as the acceptance and rejection region corresponding to a 1-step look-ahead procedure that is concentrated at site t. A stopping strategy is said to be *concentrated* at t if it only uses measurements from the site t, i.e., its sample path is given by $\mathbf{t} = \{t, t, t, \dots\}$.

Lemma 5.4 *The stopping regions for the 1-step look-ahead procedure are given by the intersection of the stopping regions per site, i.e.,*

$$\begin{aligned} S_1^a &= \cap_t S_1^a(t)\\ S_1^r &= \cap_t S_1^r(t) \end{aligned}$$

Proof. Follows easily from the fact that the class $\mathcal{R}_1$ consists of concentrated stopping strategies only. □

If we consider more than one step in our look-ahead procedure, the equality no longer holds. Since, for $k > 1$ the class $\mathcal{R}_k$ also contains stopping strategies that are not concentrated. However, we may still find

$$\begin{aligned} S_n^a &\subset \cap_t S_n^a(t)\\ S_n^r &\subset \cap_t S_n^a(t) \end{aligned}$$

for $n > 1$.

Finally, let us state the following lemma's that will be useful later on.

Lemma 5.5 *For each $t \in \mathcal{G}$,*

$$\mathbf{E}_\xi T_t \xi = \xi$$

Proof. Calculating the expectation of $T_t\xi(\theta)$ given ξ, for arbitrary θ and t, gives us

$$\begin{aligned}
\mathbf{E}_\xi[T_t\xi(\theta)] &= \sum_{\vartheta\in\Theta^*} \xi(\vartheta)\mathbf{E}_\vartheta[T_t\xi(\theta)] \\
&= \int \sum_{\vartheta\in\Theta^*} \xi(\vartheta) f_\vartheta^t(y) \frac{\xi(\theta)f_\theta^t(y)}{\sum_{\vartheta\in\Theta^*}\xi(\vartheta)f_\vartheta^t(y)} dy \\
&= \int \xi(\theta) f_\theta^t(y) dy \\
&= \xi(\theta)
\end{aligned}$$

Since this holds for all t and θ, the proof is complete. □

We may find the following bounds on the minimal cost of continuation.

Lemma 5.6 *For each $k > 0$ (including infinity), the expected k-truncated Bayes cost satisfies*

$$0 \leq \varrho_k(\xi) \leq \rho_{k-1}(\xi)$$

or, even stronger,

$$0 \leq \varrho_k(\xi, t) \leq \rho_{k-1}(\xi)$$

for each t.

Proof. We only prove the second statement, since the first follows from the second. The statement of the lemma follows from the concavity of the Bayes cost, together with Jensen's inequality, and Lemma 5.5:

$$\begin{aligned}
\varrho_k(\xi, t) &= \mathbf{E}\rho_{k-1}(T_t\xi) \\
&\leq \rho_{k-1}(\mathbf{E}_\xi T_t\xi) \\
&= \rho_{k-1}(\xi)
\end{aligned}$$

The lower bound is rather obvious, since the Bayes cost is always non-negative. □

In the following sections we examine the optimal cost function for some special types of changes. In Section 5.3.2 we assume the alternative hypothesis to be simple, so that the problem may be interpreted as the detection of a known global change.

In case the alternative hypothesis is composite, we may still distinguish several situations. Here we focus on local changes of known form and size, but of unknown position. In Section 5.3.3 we assume the possible regions of change to be disjoint. Finally, in Section 5.3.4 the regions of change may be overlapping.

5.3.2 Simple hypothesis

In case the alternative hypothesis is simple, the change will be present all over the plane. Hence, it seems that it does not matter where we make our observations. However, although the change is present everywhere, it may not be equally easy to detect at every site.

Example 5.11 *Once more, let us return to our farmer and his lands. He is growing lettuce on one part of his land, and wants to know if a certain insect has spreaded its larvae over his land. Since these larvae are hard to detect, he has to inspect the heads of lettuce one by one with a magnifying glass. Unfortunately, he can only do this by destroying the lettuce. Even then he may still miss a lot of larvae, as the insects tend to place their offspring deep inside the heads, and the larvae may easily be mistaken for the natural pigment of the lettuce. The farmer assumes that either all of the heads of lettuce are invaded by the insects, or none of them are. If the heads contain the larvae, he has to use some chemicals to exterminate them before they transform into caterpillars and start eating his lettuce. Clearly, this problem is of the same type as the problem our farmer had with his cows in Chapter 2. The fact that the heads of lettuce are positioned on a two-dimensional piece of land is of no importance here, due to the independence and the identical distribution under both hypotheses. The order in which the farmer tests his heads of lettuce does not have any influence on the expected cost.*

If under both hypotheses the random field is independent and identically distributed, we may use the exact same approach as in Section 2.2.3. The order of the measurements may be chosen as any arbitrary ordering on the field. A more interesting case arises when we assume the process not to be identically distributed on the field. To return to the farmer and his lettuce, there may be some regional difference in the distribution of the larvae over the heads of lettuce. For example, in the more sunny part of the field, there may be more larvae than in the other part of the field. In this particular case it may be obvious that the heads of lettuce from the sunny part of the field should be examined first. However, in general this may not be so clear.

We assume that the field is independent and identically distributed under the null hypothesis, say with density function f_0. Under the alternative hypothesis, we only assume the field to be independent, with a density function at site t that is given by f_t. The case that the field is identically distributed under the alternative hypothesis as well is obtained by choosing $f_t \equiv f_1$ for all t. The prior distribution of the parameters is given by

$$\begin{aligned}\xi(0) &= \mathbf{Pr}[\boldsymbol{\theta} = 0]\\ \xi(1) &= \mathbf{Pr}[\boldsymbol{\theta} \neq 0] = 1 - \xi(0)\end{aligned}$$

Since this distribution function only depends on the scalar $\xi(0)$, we denote $\eta = \xi(0)$.

The calculation of the posterior probability of the null hypothesis, after measuring at site t, is rather easy:

$$T_t\eta = \frac{\eta}{\eta + (1-\eta)\ell_t}$$

where $\ell_t = f_t/f_0$ is the likelihood ratio at site t.

Let us first assume the random field to have an inexhaustive observation mechanism. So, we may take unlimited measurements from each site, without losing independence. In that case, after we take a measurement, the only thing that changes is the distribution of the parameters.

Corollary 5.1 (Bayes stopping time for inexhaustive fields) *For a test between two simple hypotheses on a random field whose observation mechanism is inexhaustive, there exist constants η^a and η^r such that the Bayes stopping time may be written as*

$$\tau = \min\{k | \eta_k \notin (\eta^r, \eta^a)\}$$

Here η_k is defined by

$$\eta_k = T_{\mathbf{t}_k}\eta = T_{t_k}\eta_{k-1}$$

Proof. Follows directly from the convexity of the stopping regions (Theorem 5.4). □

The problem that remains is twofold. Firstly, we do not know the constants η^a and η^r. Unfortunately, as in the one-parameter case, these may not be found that easily. One possible way is to use the integral equations from Section 2.2.2. Since these thresholds also depend on the minimization over the sites, the problem is more complicated in this case. If we only use one particular site t for our measurements, we may calculate the optimal thresholds $\eta^r(t)$ and $\eta^a(t)$, using the integral equations from Chapter 2. If we repeat this procedure for each site, we obtain a sequence of stopping boundaries. From the discussion following Lemma 5.4 we know that

$$\begin{aligned}\eta^r &\leq \min_t \eta^r(t)\\ \eta^a &\geq \max_t \eta^a(t)\end{aligned}$$

The second problem is the optimal choice of the sites. Using the terminology of Definition 5.7 we have to find the most informative sites. However, it is not obvious what makes one site more informative than another one.

Moreover, if a certain site is most informative for some η, this does not automatically imply that this site is most informative for all other η. If the site t is most informative for all η between $\eta^r(t)$ and $\eta^a(t)$, then the Bayes sample path is given by $\mathbf{t} = \{t, t, t, \ldots\}$. Clearly, in this case the thresholds are given by $(\eta^r, \eta^a) = (\eta^r(t), \eta^a(t))$. Here we may establish a link with the Neyman-Pearson approach through the following theorem.

Theorem 5.6 *If the site t is most informative for all η, then it is also most informative in the class $\mathcal{R}_{\alpha,\gamma}$.*

Proof. First note that the Bayes stopping strategy in this case equals the stopping strategy $\mathbf{d}_t$ from Section 5.2.1. To show that the site t is most informative in the Neyman-Pearson sense, we have to prove that $N_i(\mathbf{d}_t) \leq N_i(\mathbf{d})$ for all possible stopping strategies $\mathbf{d} \in \mathcal{R}_{\alpha,\gamma}$. We know that

$$L(\xi, \mathbf{d}_t) \leq L(\xi, \mathbf{d})$$

for all ξ and all $\mathbf{d}$, where we used $\xi(0) = \eta$ and $\xi(1) = 1 - \eta$. Note that we may write

$$L(\xi, \mathbf{d}) = \eta(\mathbf{E}_0\tau + c_0\alpha(\mathbf{d})) + (1-\eta)(\mathbf{E}_1\tau + c_1\gamma(\mathbf{d}))$$

We may then easily see that if $\alpha(\mathbf{d}) \leq \alpha(\mathbf{d}_t)$ and $\gamma(\mathbf{d}) \leq \gamma(\mathbf{d}_t)$, the required inequalities hold. Since this is exactly the case in the class $\mathcal{R}_{\alpha,\gamma}$, the result follows. □

Let us examine the 1-step look-ahead procedure for the simple hypotheses problem. From Lemma 5.4 we know that the stopping boundaries may be obtained by considering each site separately. Therefore, let us now focus on a single site t. The function $\varrho_1(\eta, t)$ may be calculated as

$$\varrho_1(\eta, t) = \eta c_0 \mathbf{Pr}[\ell_t > \frac{\eta c_0}{(1-\eta)c_1} | \boldsymbol{\theta} = 0] + (1-\eta)c_1 \mathbf{Pr}[\ell_t < \frac{\eta c_0}{(1-\eta)c_1} | \boldsymbol{\theta} \neq 0]$$

Although this may already be rather difficult to solve, in general it is possible to find a numerical solution. From the equations

$$\begin{aligned} \eta c_0 &= 1 + \varrho_1(\eta, t) \\ (1-\eta)c_1 &= 1 + \varrho_1(\eta, t) \end{aligned}$$

we may find the thresholds $(\eta_1^r(t), \eta_1^a(t))$. Repeating this procedure for each site on the field, we obtain a sequence of thresholds, from which we may select

$$\begin{aligned} \eta_1^r &= \min_t \eta_1^r(t) \\ \eta_1^a &= \max_t \eta_1^a(t) \end{aligned}$$

according to Lemma 5.4. From Lemma 5.3 we know that $\eta_k^r \geq \eta_{k+1}^r$ and $\eta_k^a \leq \eta_{k+1}^a$. Hence, η_1^r and η_1^a are upper and lower bounds to η^r and η^a, respectively. In the following example we examine the quality of these approximations.

Example 5.12 *Consider a random field whose index set consists of one site only. The process at this site is Gaussian, with a variance equal to* 1*. Under the null hypothesis the mean is equal to* 0*, and under the alternative hypothesis the mean is equal to* μ*.*

Then, the log-likelihood ratio is distributed as a Gaussian random variable with a mean that is equal to $-\mu^2/2$ *under the null hypothesis and* $\mu^2/2$ *under the alternative hypothesis. The variance is given by* μ^2*.*

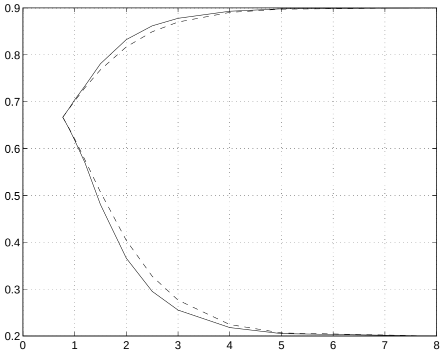

Figure 5.6: The stopping boundaries as a function of the mean (solid line), and their approximation using 1-step look-ahead Bayes cost (dashed), with error costs $(c_0, c_1) = (5, 10)$

Using an arbitrary prior $\xi(0)$*, we may now calculate the optimal thresholds* η^a *and* η^r *for any value of* μ*. The Bayes cost may be calculated according to*

$$\rho(\eta) = \inf_{a,b}[\eta\{N_0(0) + c_0(1 - P_0(0))\} + (1 - \eta)\{N_1(0) + c_1 P_1(0)\}]$$

where

$$\begin{aligned} N_i(z) &= 1 + \int_a^b N_i(x) g_i(x - z) dx \\ P_i(z) &= \int_{-\infty}^a g_i(x - z) dx + \int_a^b P_i(x) g_i(x - z) dx \end{aligned}$$

Here $g_i(x)$ *denotes the density function of the log-likelihood ratio, given that* f_i *is the correct density function. Since these density functions are known, these integral equations may be solved numerically for each pair* (a, b)*. Then,* *the optimal thresholds may be obtained by minimizing the cost function over (a, b) and rewriting*

$$\begin{aligned}\eta^a &= \frac{\xi(0)}{\xi(0) + (1-\xi(0))\exp(a)} \\ \eta^r &= \frac{\xi(0)}{\xi(0) + (1-\xi(0))\exp(b)}\end{aligned}$$

Clearly, these thresholds depend on the distribution functions G_0 and G_1. The special case where $\eta^a = \eta^r = \frac{c_1}{c_0+c_1}$ appears when

$$G_0(0) - G_1(0) = \frac{1}{c_0} + \frac{1}{c_1}$$

This equation may be regarded as a lower bound to the amount of information the hypotheses should hold in order to guarantee the need for measurements. If $G_0(0) - G_1(0)$ is smaller than the right hand side of this equation, it is more optimal to make a decision right now than to take a measurement.

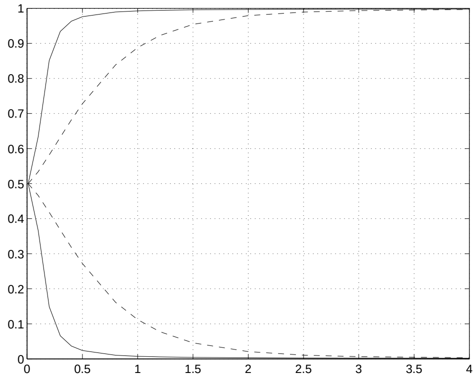

Figure 5.7: The stopping boundaries as a function of the mean (solid line), and their approximation using 1-step look-ahead Bayes cost (dashed), with error costs $(c_0, c_1) = (500, 500)$

In Figure 5.6-5.7 the optimal thresholds are shown as a function of μ, together with their approximations from the 1-step look-ahead strategies. Here we only considered the cases where the error costs are given by $(c_0, c_1) = (5, 10)$ and $(c_0, c_1) = (500, 500)$. We may see that in the first case the 1-step look-ahead approximations are already quite good. Due to the small values of the error cost, the number of measurements needed to make a decision will not be very large. Therefore, the n-step look-ahead strategies are quite accurate for small values of n.

For the larger error costs, the approximations are less accurate. This gives us reason to believe that n should be quite large to find good approximations to the true stopping bounds.

In the previous example we may notice that the thresholds η^r and η^a converge as the value of μ goes to infinity. This convergence holds in general as well, and the limiting values of the thresholds depend only on the error costs. Instead of letting μ go to infinity, we may as well let the variance σ^2 go to zero. So, the hypotheses become deterministic, which implies that they are *perfectly distinguishable* by one measurement only.

Definition 5.8 (Perfectly distinguishable) *Two hypotheses are said to be perfectly distinguishable if there exists a decision function, based on one measurement, for which both error probabilities are zero.*

Obviously, whether or not two hypotheses are perfectly distinguishable depends on the statistical parameters under both hypotheses. In Example 5.12 the mean under the alternative hypothesis determines the distinguishability of the hypotheses.

The following lemma now is straightforward.

Lemma 5.7 *Let G_0 and G_1 denote the probability distribution functions of the (log-)likelihood ratio for a test between two hypotheses, under the null and the alternative hypothesis, respectively. The hypotheses are perfectly distinguishable if and only if there exists a λ for which*

$$G_0(\lambda) - G_1(\lambda) = 1$$

Proof. If the hypotheses are perfectly distinguishable, and there exist distribution functions for the (log-)likelihood ratio, then there exists a λ such that

$$\begin{aligned}\alpha &= \mathbf{Pr}[T > \lambda | \boldsymbol{\theta} = 0] = 1 - G_0(\lambda) = 0\\ \gamma &= \mathbf{Pr}[T < \lambda | \boldsymbol{\theta} \neq 0] = G_1(\lambda) = 0\end{aligned}$$

from which it follows that $G_0(\lambda) - G_1(\lambda) = 1$.

Alternatively, if $G_0(\lambda) - G_1(\lambda) = 1$ for some λ, then necessarily $G_0(\lambda) = 1$ and $G_1(\lambda) = 0$, from which perfect distinguishability follows. □

In general two hypotheses will not be perfectly distinguishable. In the following theorem we mention hypotheses that "become" perfectly distinguishable. Here we mean to say that the parameters that define the statistical characteristics of the model under both hypotheses are converging in such a way that the hypotheses become perfectly distinguishable. Now let us state the more general theorem that defines the limiting thresholds of a simple test.

Theorem 5.7 *If there exists a site t for which the hypotheses become perfectly distinguishable, the thresholds converge to*

$$\eta^r \rightarrow \frac{1}{c_0}$$

and

$$\eta^a \rightarrow 1 - \frac{1}{c_1}$$

Proof. If the hypotheses are perfectly distinguishable, we only need to take one measurement to make a perfect decision. So,

$$\begin{aligned} \mathbf{E}_0\tau &= \mathbf{E}_1\tau = 1 \\ \alpha(\delta) &= \gamma(\delta) = 0 \end{aligned}$$

so that $\rho(\eta) \leq 1$. The limiting thresholds may now be obtained from

$$\begin{aligned} 1 &= \eta^r c_0 \\ 1 &= (1 - \eta^a)c_1 \end{aligned}$$

The result of the theorem follows easily. □

At the other extreme, we may find a sufficient condition for the necessity of measurements.

Theorem 5.8 *Let G_0 and G_1 denote the probability distribution functions of the (log-)likelihood ratio at site t for a test between two hypotheses, under the null and the alternative hypothesis, respectively. If there exists a λ such that*

$$G_0(\lambda) - G_1(\lambda) > \frac{1}{c_0} + \frac{1}{c_1}$$

then the region of continuation is non-empty, i.e., there exists a ξ for which $\rho(\xi) < \rho_0(\xi)$.

Proof. From the concavity of the Bayes cost, and the fact that ρ_0 has its maximum at $c_1/(c_0 + c_1)$, we may see that if the region of continuation is non-empty, the point $c_1/(c_0 + c_1)$ will be part of it. If we assume a λ to exist for which the inequality in the statement of the theorem holds, we may write

$$\begin{aligned} \rho(\frac{c_0}{c_0 + c_1}) &= \inf_{\mathbf{d}} L(\frac{c_1}{c_0 + c_1}, \mathbf{d}) \\ &\leq \frac{c_1}{c_0 + c_1}(1 + c_0(1 - G_0(\lambda))) + \frac{c_0}{c_0 + c_1}(1 + c_1 G_1(\lambda)) \\ &= \frac{c_0 c_1}{c_0 + c_1} + 1 - \frac{c_0 c_1}{c_0 + c_1}(G_0(\lambda) - G_1(\lambda)) \\ &< \frac{c_0 c_1}{c_0 + c_1} \\ &= \rho_0(\frac{c_1}{c_0 + c_1}) \end{aligned}$$

Here we simply used the stopping rule $\mathbf{d} = (\{t\}, 1, \delta)$, where δ accepts the null hypothesis if the (log-)likelihood ratio is smaller than λ, and rejects it otherwise. □

Let us now focus on the continuation regions. Clearly, since the parameter η is scalar, the continuation regions may be characterized by intervals on the real line. Due to the concavity of the functions $\varrho_k(\eta, t)$, we also know that each site t may only be most informative on at most two separated intervals.

Example 5.13 *Suppose we have a random field whose index set consists of two sites. The process is Gaussian at both sites, with variance $\sigma(t)^2$ depending on the site. Under the null hypothesis, the mean at both sites is zero, whereas under the alternative hypothesis, the mean is equal to μ for both sites. As in the previous case, we use the log-likelihood ratio as our sufficient statistic. Under the null hypothesis, it has a Gaussian distribution with mean $-\mu^2/(2\sigma(t)^2)$. Under the alternative hypothesis, the mean only reverses sign. The variance equals $\mu^2/\sigma(t)^2$ under both hypotheses. Clearly, the effect of having variance σ^2 instead of 1 is that the mean μ may be replaced by μ/σ. Hence, we may expect the site with the smallest variance to be most informative for all η.*

First let us consider a 1-step look-ahead approach. We may calculate $\varrho_1(\eta, t)$ as

$$\varrho_1(\eta, t) = \eta c_0 \alpha(t) + (1-\eta) c_1 \gamma(t)$$

where

$$\begin{aligned} \alpha(t) &= 1 - \Phi(\frac{\sigma(t)}{\mu}\lambda + \frac{\mu}{2\sigma(t)}) \\ \gamma(t) &= \Phi(\frac{\sigma(t)}{\mu}\lambda - \frac{\mu}{2\sigma(t)}) \end{aligned}$$

and

$$\lambda = \log \frac{\eta c_0}{(1-\eta)c_1}$$

The thresholds $(\eta_1^r(t), \eta_1^a(t))$ may easily be found for both sites, thus providing us with (η_1^r, η_1^a).

Suppose that $\sigma(t) < \sigma(s)$. To find the continuation region $C_1(t)$, we have to find those η for which $\varrho_1(\eta, t) < \varrho_1(\eta, s)$. We may see that this inequality certainly holds as long as $|\lambda| < \mu/2\sigma$, since in that case both error probabilities are increasing with σ. This condition may be translated as

$$\frac{c_1}{c_1 + c_0 \exp(\frac{1}{2}(\frac{\mu}{\sigma})^2)} < \eta < \frac{c_1 \exp(\frac{1}{2}(\frac{\mu}{\sigma})^2)}{c_0 + c_1 \exp(\frac{1}{2}(\frac{\mu}{\sigma})^2)}$$

If these inequalities do not hold, one of the error probabilities decreases with σ. However, since the other error probability still increases with σ, it is not *clear if the site with larger variance ever becomes more informative. From numerical experiments - compare with Examples 5.4 and 5.5 - we expect this never to happen.*

5.3.3 Disjoint regions of change

In case the alternative hypothesis is composite, the problem is considerably more complex. In this section we examine one of the simplest composite hypothesis problems that may occur.

Example 5.14 *After some more thinking, our farmer comes to the conclusion that only part of the field may be covered with larvae. He divides the field into several regions, and assumes that if the insects have left their offspring in his lettuce field, it must be in only one of those regions. He now wants to know what strategy he should use to make a decision about the presence of the larvae with the lowest possible cost.*

Under the null hypothesis, the process has density function f_0 at each site. Under the alternative hypothesis, the density function at only one of the sites changes to f_1. This implies that all regions of change are assumed to be disjoint. Moreover, each region of change only consists of one site. Each θ from Θ corresponds with a change at one of the sites. The parameter set may then as well be given by the set of sites, i.e., $\theta = t$ denotes the presence of a change at site t.

The posterior probabilities are calculated from

$$T_t\xi(\theta) = \begin{cases} \frac{\xi(\theta)\ell}{1+\xi(t)(\ell-1)} & \text{if } \theta = t \\ \frac{\xi(\theta)}{1+\xi(t)(\ell-1)} & \text{otherwise} \end{cases}$$

The first guess for the Bayes sample path would be to choose those sites that have the largest probability of being contained in the region of change. Indeed, for the 1-step look-ahead Bayes stopping strategy, the sample path has this form.

Theorem 5.9 (Continuation regions) *For the 1-step look-ahead procedure the continuation regions for the disjoint change problem are given by*

$$C_1(t) = \{\xi \in C_1 | \xi(t) > \xi(s) \textit{ for all } s \neq t\}$$

Proof. We have to prove that $\varrho_1(\xi, t) < \varrho_1(\xi, s)$ if $\xi(t) > \xi(s)$.

We may easily see that $\varrho_1(\xi, t) = \varrho_1(\xi, s)$ if $\xi(s) = \xi(t)$. Suppose now that $\xi(s) = 0$. Then $T_s\xi = \xi$, so that

$$\begin{aligned} \varrho_1(\xi, s) &= \mathbf{E}_\xi \rho_0(\xi) \\ &= \rho_0(\xi) \end{aligned}$$

Then, using Lemma 5.6 it follows that $\varrho_1(\xi, t) \leq \varrho_1(\xi, s)$ for all t, so that the site s may never be more informative than any other site.

Let us consider the hyperplane parameterized by

$$\begin{aligned} \xi(s) &= \lambda \\ \xi(t) &= a - \lambda \end{aligned}$$

where λ varies from 0 to a, and the remaining parameters are fixed. From the previous, we know that $\varrho_1(\xi, t) < \varrho_1(\xi, s)$ for $\lambda = 0$, $\varrho_1(\xi, t) = \varrho_1(\xi, s)$ for $\lambda = a/2$, and $\varrho_1(\xi, t) > \varrho_1(\xi, s)$ for $\lambda = a$.

From the concavity and the symmetry of the Bayes cost, we may see that we now have two possibilities for the continuation regions. Either $\varrho_1(\xi, s)$ reaches its maximal value for $\lambda > a/2$, or for $\lambda \leq a/2$. Clearly, if the second case holds, the statement of the theorem follows easily. We show that the maximum is positioned at $\lambda = 0$, i.e., $\varrho_1(\xi, s)$ is decreasing with λ.

For $\lambda = 0$ we already have

$$\varrho_1(\xi, s) = \rho_0(\xi)$$

Since $\rho_0(\xi)$ is independent of λ, and $\varrho_1(\xi, s) \leq \rho_0(\xi)$ for all λ, it follows that $\varrho_1(\xi, s)$ has its maximum in $\lambda = 0$. Hence, $\varrho_1(\xi, t) < \varrho_1(\xi, s)$ for $\lambda < a/2$, or, equivalently, for $\xi(t) > \xi(s)$. Since the remaining part of ξ was chosen arbitrarily, and the previous holds for each a, it follows that the continuation regions of the 1-step look-ahead procedure indeed are as given in the statement of the theorem. □

Although we expect this theorem to hold as well for n-step look-ahead strategies in general, we have not been able to prove this.

Comparing this result with the Neyman-Pearson approach, we may find some similarities. Indeed, the 1-step look-ahead Bayes sample path is a myopic sample path. This is a rather obvious result, since the basis for the myopic sample path also is to look only one step ahead.

Using Lemma 5.4 we may find the stopping boundaries for the 1-step look-ahead procedure by considering each site separately. We may calculate

$$\varrho_1(\xi, t) = (1 - \xi(0) - \xi(t))c_1 + (\xi(0)c_0 - (1 - \xi(0) - \xi(t))c_1)\alpha(t) + \xi(t)c_1\gamma(t)$$

where

$$\begin{aligned} \alpha(t) &= \mathbf{Pr}[\ell > \lambda(t) | \boldsymbol{\theta} = 0] \\ \gamma(t) &= \mathbf{Pr}[\ell < \lambda(t) | \boldsymbol{\theta} \in \Theta] \end{aligned}$$

and

$$\lambda(t) = \frac{\xi(0)c_0 - (1 - \xi(0) - \xi(t))c_1}{\xi(t)c_1}$$

The stopping boundaries may now be calculated numerically. We may see that the stopping boundary only depends on $\xi(t)$ and $\xi(0)$. Hence, for each pair $(\xi(0), \xi(t))$ that solve the equality, the local part of the stopping boundary is given by the hyperplane

$$\{\xi' \in S | \xi'(0) = \xi(0), \xi'(t) = \xi(t), \xi'(s) < \xi(t) \text{ for all } s \neq t\}$$

The following example illustrates this procedure.

Example 5.15 *Let us consider a Gaussian random field whose index set consists of two sites. The parameter set also contains two elements, each of which represents the presence of a change at one of the sites. Under the null hypothesis, the process has zero mean on both sites. Under the alternative hypothesis, one of the sites has mean μ, and the other still has mean zero. The variance is equal to* 1 *under all circumstances. The error costs are given by $c_0 = 50$ and $c_1 = 100$.*

Using a 1-step look-ahead procedure, the stopping regions may be calculated using the previous results. The stopping boundaries are calculated for both sites separately, and are shown in Figure 5.8. From Lemma 5.4, the actual stopping regions according to the 1-step look-ahead procedure are then given by the intersection of the separate stopping regions.

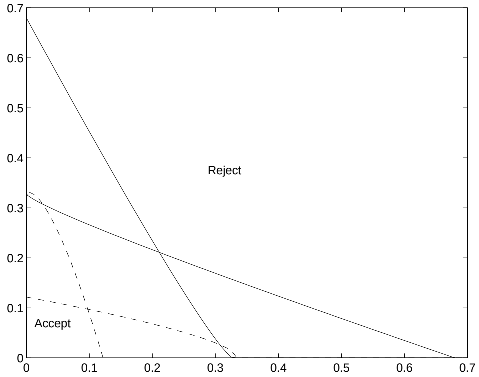

Figure 5.8: Stopping regions according to 1-step look-ahead procedure.

The boundary between the regions $C_1(s)$ and $C_1(t)$ is defined by the straight line $\xi(s) = \xi(t)$. If $\xi(s) > \xi(t)$, then the site s is more informative than the site t.

5.3.4 Overlapping regions of change

A generalization of the previous results is to allow the regions of change to be overlapping.

Example 5.16 *Our farmer still is not very happy with his approach. He knows that the insects typically spread their larvae over a region of a specific form, but the possible regions do not have to be disjoint at all. So, he wants to include the possibility that the regions of change are overlapping.*

We assume the shape of the region of change to be known; hence, the index set has to carry some information about the original position on the plane. In fact, in this case the use of a grid is intuitively more clear. The only unknown aspect of the change is its position on the grid. The position of a change is supposed to be characterized by the random variable T_c, which may take any $t \in \mathcal{G}$ as its values. The value of T_c corresponds with a certain pre-defined characteristic point of the region of change. For example, this may be the center, or one of the corners of the region of change. The region of change is now denoted as $\mathcal{G}_c(T_c)$. Each parameter from the parameter set Θ may then be identified with a certain site of the grid, i.e., the site that coincides with T_c. The prior distribution of T_c is given by ξ, that is,

$$\xi(\theta) = \mathbf{Pr}[T_c = \theta]$$

and the probability of no change follows from

$$\xi(0) = 1 - \sum_{\theta \in \Theta} \xi(\theta)$$

For a given site t, the probability of being contained in the region of change may be calculated as

$$P(t) = \sum_{\{s | t \in \mathcal{G}_c(s)\}} \xi(s)$$

The posterior distribution, based on a measurement at site t, may then be derived as

$$T_t\xi(\theta) = \begin{cases} \frac{\xi(\theta)\ell}{1+P(t)(\ell-1)} & \text{if } t \in \mathcal{G}_c(\theta) \\ \frac{\xi(\theta)}{1+P(t)(\ell-1)} & \text{otherwise} \end{cases}$$

Let us now focus on the optimal sample path. As for the disjoint regions of change, we may be inclined to choose the next site as the one that has the largest probability of being contained in the region of change. Indeed, as in the case where the regions of change are disjoint, this procedure is optimal for the 1-step look-ahead procedure.

Theorem 5.10 (Continuation regions) *For the 1-step look-ahead procedure, the continuation regions for the overlapping regions of change are given by*

$$C_1(t) = \{\xi \in C_1 | P(t) > P(s) \textit{ for all } s \neq t\}$$

Proof. Let us first calculate

$$\begin{aligned} \varrho_1(\xi, t) &= \mathbf{E}_\xi \rho_0(T_t \xi) \\ &= \mathbf{E}_\xi \min\{T_t\xi(0)c_0, (1 - T_t\xi(0))c_1\} \end{aligned}$$

Since $T_t\xi(0)$ only depends on $P(t)$ and $\xi(0)$, it follows that $\varrho_1(\xi, t) = \varrho_1(\xi, s)$ if $P(t) = P(s)$.

Now let us suppose that $P(t) = 0$, which implies that $\xi(s) = 0$ for all s such that $t \in \mathcal{G}_c(s)$. It follows that $T_t\xi = \xi$. Hence,

$$\varrho_1(\xi, t) = \rho_0(\xi) = c(\xi(0))$$

where we use the function c to indicate that this expression only depends on $\xi(0)$.

So far the proof has been similar to the proof of Theorem 5.9. However, instead of comparing two sites, in this case we have to approach the problem from a different perspective. Let us consider the hyperplane where $\xi(0)$ equals a certain constant. We may easily see that this hyperplane is convex in the ξ-space. Hence, the function $\varrho_1(\xi, t)$ is concave on this hyperplane. Furthermore, since we know that $\varrho_1(\xi, t) < \rho_0(\xi) = c(\xi(0))$ on this hyperplane, it follows that $\varrho_1(\xi, t)$ is maximal if $P(t) = 0$.

From the concavity and the fact that $\varrho_1(\xi, t) = \varrho_1(\xi, s)$ if $P(t) = P(s)$ we may conclude that

$$C_1(t) = \{\xi \in C_1 | P(t) > P(s) \text{ for all } s \neq t\}$$

which was to be proved. □

At the boundaries between the different continuation regions, we have to use randomization to choose the next site.

As for the disjoint regions of change, the 1-step look-ahead Bayes sample path is myopic.

If there exists a site t for which $P(t) = 1$, this site is most informative for all ξ. We may easily see that the resulting test is simple.

Example 5.17 *Consider a grid consisting of three sites, that are positioned next to each other. The region of change covers two neighbouring sites on this grid. So, there are two possible positions for a change; the left side or the right side of the grid. Since $P(t) = 1$ for the middle site, this site is most informative for all ξ. Hence, the actual stopping problem becomes simple. This implies that the problem becomes equivalent with Example 5.12.*

Using a 1-step look-ahead Bayes strategy, the expected Bayes cost may be written as

$$\begin{aligned} \varrho_1(\xi,t) &= \mathbf{E}_\xi \rho_0(T_t\xi) \\ &= (1-\xi(0)-P(t))c_1 + (\xi(0)c_0 - (1-\xi(0)-P(t))c_1)\alpha(t) \\ &\quad +P(t)c_1\gamma(t) \end{aligned}$$

Here,

$$\begin{aligned} \alpha(t) &= \mathbf{Pr}[\ell > \lambda(t) | \boldsymbol{\theta} = 0] \\ \gamma(t) &= \mathbf{Pr}[\ell < \lambda(t) | \boldsymbol{\theta} \in \Theta] \end{aligned}$$

where

$$\lambda(t) = 1 + \frac{\xi(0)c_0 - (1-\xi(0))c_1}{P(t)c_1}$$

From Lemma 5.4 we know that the exact boundaries for the 1-step look-ahead procedure may be found by considering each site separately. Suppose that we use the site t in our next step. The stopping boundaries are defined by the equalities

$$\xi(0)c_0 = 1 + \varrho_1(\xi,t) \tag{5.5}$$

$$(1-\xi(0))c_1 = 1 + \varrho_1(\xi,t) \tag{5.6}$$

Since these equations are similar to the ones in the previous section, the stopping boundaries will also be similar. However, the role of $\xi(t)$ is now replaced by $P(t)$. If the stopping boundaries for the disjoint problem are found, we immediately have obtained the stopping boundaries for the overlapping problem as well. For each pair $(\xi(0), P(t))$ that solves the equality, the local part of the stopping boundary is given by

$$\{\xi' \in S | \xi'(0) = \xi(0), P'(t) = P(t), P'(s) < P(t) \text{ for all } s \neq t\}$$

In practice, we may construct the stopping boundaries by starting with $\xi(t) = P(t)$, and $\xi(s) = 0$ for all $s \neq t$ for which $t \in \mathcal{G}_c(s)$. From the skeleton that is formed this way we may obtain the final stopping boundaries by linear interpolation between the legs of the skeleton.

Example 5.18 *Consider a random field defined on a grid consisting of four sites, positioned on one row. As in the previous example, the regions of change cover two sites that are positioned next to each other. So, in this case there are three possible changes on the field.*

Using a 1-step look-ahead strategy, we know from Theorem 5.10 that the continuation regions are given by

$$C_1(t) = \{\xi \in C_1 | P(t) > P(s) \textit{ for all } s \neq t\}$$

However, before we can use this result we have to know C_1. To find this region we have to calculate the stopping boundaries. These may be found numerically from equations (5.5) and (5.6).

5.4 Summary

In this chapter, the quickest detection problem for random fields has been introduced. Stopping strategies are introduced, consisting of a sample path, a stopping time and a decision function. These stopping strategies are to our problem what the statistical tests are to the classical quickest detection problem. It is shown that the generally very large class of sample paths may be reduced to the much smaller class of site-by-site sample paths. The quickest detection problem is approached from two different points of view; a Neyman-Pearson approach and a Bayesian approach.

Using the Neyman-Pearson approach, the objective of the quickest detection problem is to minimize the expected number of observations that is needed to reach a decision, given that the error probabilities are bounded by given constants. Since this problem generally does not have a solution, we restricted ourselves to an intuitively appealing stopping strategy that may be used for this problem. This so-called myopic stopping strategy is based on an optimization of the immediate result; in other words, it only looks one step ahead.

If we use a Bayesian approach to the quickest detection problem, the objective is to minimize a certain cost function over the class of all stopping strategies. In this case, there is no constraint on the error probabilities. The cost function is defined as a linear combination of the error cost and the sampling cost. The quickest detection problem is shown to have a solution in the Bayesian setting. However, this solution is implicitly defined and may not be found that easily. Approximations are given by the so-called n-step look-ahead procedures.

Chapter 6

Conclusions

In this final chapter we look back at the results obtained in this thesis. We point out what significance these results have, and what consequences they may have. Furthermore, we discuss what further extensions of the theory may be possible and useful.

The complexity of the random field more or less defines the scale on which we may work. To start with the most complex random fields, the non-causal random fields from Chapter 3, we have shown that in a non-sequential setting the changes in these fields may be dealt with as in the classical detection theory. Spatially sequential tests have not been considered for these fields, because of the complications these tests would give for the calculation of probabilities and average site numbers. The detection problem that was dealt with may be interpreted as a classification problem. Both hypotheses are represented by a class of models that may describe the random field. The class that contains the model that describes the measured data best is selected.

Alternatively, if the observation mechanism allows us to make a sequence of observations of the entire field in time, the detection problem may be rewritten as a classical quickest detection problem. Apart from the classical generalized likelihood ratio test, two adaptations of the score test are used to detect small changes in random fields. Upper bounds are given for the expectation of the delay of detection for all statistics.

In Chapter 3 we only considered global changes, although local changes may be dealt with similarly. However, to detect local changes we need either a parameterization of the region of change, or a window that only allows certain regions of the field to enter in the test statistic. This matter is discussed in Chapter 4, where we are looking for local changes in (semi-)causal fields. The complexity of the random fields is reduced in order to be able to use spatially sequential tests. In this chapter, we used the Kalman filter to extract all information on the possible changes that is contained in the measurements.

Due to the dynamic systems describing the random field in this case, the detectability of the changes is not that straightforward. Indeed, nonzero changes may not be detectable. Necessary and sufficient conditions for a change to be detectable are given in this chapter.

The thresholds used in the statistical tests of Chapter 4 are all based on data obtained from simulations. Hence, an improvement would be to determine these thresholds on a theoretical basis. Since the test statistic that is used is based on a maximization of quadratic forms in Gaussian variables, this may not be easy. However, it may be possible to find better approximations.

Finally, in Chapter 5 we introduced the theory of quickest detection for random fields. To allow total freedom in the observation mechanism, the random fields in this chapter were assumed to be independent in the sense that the process is independent between the sites. An interesting next step may be to drop this assumption of independence. In that case, we may approach the problem from two different points of view. Firstly, we may restrict the class of sample paths in such a way that all resulting processes are causal. Note that in some cases this restricted class of sample paths may be degenerate; for example, for the random fields of Chapter 3 the only sample path that satisfies this restriction is the sample path that includes the entire field in its first sample. Alternatively, if we do not impose this constraint on the class of sample paths, the statistics of the resulting non-causal processes may get very complicated.

For both cases, due to the dependence thus obtained, the given formulae for the calculations of probabilities and average site numbers are no longer valid. Nevertheless, it may still be possible to approximate the optimal strategy for this problem in a similar way as was done in our independent case. However, it is unclear how well this approach will work.

Two classical approaches were used to define the quickest detection problem. Using the Neyman-Pearson approach, the objective is to minimize the average site numbers under the constraint that the error probabilities are bounded by given constants. Only for some elemental cases a uniformly most efficient strategy may be found. In case the problem becomes only slightly more general, such strategies may not exist. An intuitively appealing approach is to use myopic stopping strategies; strategies that give the best immediate result, but may be less efficient on the long term. Here it may be interesting to find out under what circumstances these myopic strategies actually are optimal.

In order to be able to use the Bayesian approach, some more information has to be available. The cost of making errors has to be known, relative to the cost of making observations. Furthermore, a prior distribution has to be known on the parameter set. Using this information, a cost function may be defined which may be minimized over the class of stopping strategies. Although this problem is shown to have a solution, this solution may not always be computed that easily. The approximations given by the n-step look-ahead procedure have some nice properties like the monotone convergence of the resulting cost to the Bayes cost. However, the rate of convergence is not known.

The stopping strategy resulting from the 1-step look-ahead procedure is myopic. As for the Neyman-Pearson approach, it may be interesting to determine the circumstances that are required for the Bayes stopping strategies to be myopic in general. For example, the specific problems treated in Sections 5.3.3 and 5.3.4 may very well have this property.

Comparing the Bayesian approach with the Neyman-Pearson approach, we may draw similar conclusions as for the classical detection problems of Chapter 2. To be able to use the Bayesian approach, some more information is required, but a solution always exists. Using the Neyman-Pearson approach, a solution generally does not exist. However, the definition of the quickest detection problem in the Neyman-Pearson setting requires far more from the solution than in the Bayesian setting. Hence, it may be better to loosen the requirements in the Neyman-Pearson approach of the quickest detection problem, so that this problem may also have a solution in general. For example, we may replace the condition that the stopping strategy has to be uniformly most efficient by the condition that the stopping strategy only has to be most efficient for some particular parameters.

A comparison with existing theory may be made from two perspectives. Firstly, the detection of objects and the classification of textures has received considerable attention in the world of image processing. The observation mechanism does not play an actual role here, since all data from an image is immediately available. Hence, for most image processing applications non-sequential tests are used. This implies that a comparison with Chapter 3 is most appropriate. For texture classification problems there generally does not exist a nominal model. The parameters of the model are simply estimated by maximizing a likelihood function, or some simplification of the likelihood function. In our case, we do have a nominal model, so that we may formulate a statistical test. Although the use of the score test appears to be promising for the detection of small changes, in practice a change in texture may give rather large changes in the parameters of the model. Hence, the work presented in Chapter 3 should merely be considered of theoretical importance.

Also, for the detection of line objects as in Chapter 4, several techniques have been developed in image processing theory. Most of the classical approaches, like the Hough transform as described in Pratt [37], have serious problems with textured images. Of course, for images without any texture, the stochastic model used in Chapter 4 would assume an almost trivial form. Hence, the applications that our approach is better suited for are those concerning textured images. A practical problem like the detection of roads on satelite images is of this type. The methods used so far to deal with these problems generally are nonparametric. In our parametric setting, the problem is that the nominal model should be known. Nevertheless, in many situations it may be possible to obtain quite accurate estimates of the nominal model.

The second perspective is from the existing detection theory. The detection theory that is used for Chapters 3 and 4 is not new. In Chapter 3 the problem has been rewritten as a standard detection problem, and the generalized likelihood ratio and the score test are used to detect changes. In Chapter 4 Willsky's approach to the detection problem using a Kalman filter has been used. Willsky only considered changes that were either additive or parametric. In our setting the changes are have an additive and parametric part, so that both methods given by Willsky are combined to find a detection algorithm.

The formulation of the quickest detection problem in Chapter 5 is original. The sample paths, which are borrowed from Cairoli and Dalang [6], are used to construct a multi-parameter detection problem in a similar fashion as the one-parameter detection problem. Although Cairoli and Dalang already defined a similar estimation problem, they did not address the multi-parameter detection problem. This implies that the basics of the problem definition could be obtained from the work of Cairoli and Dalang, and the theory has been further developed using the results from one-parameter detection theory.

Bibliography

[1] Nikhil Balram and José M. F. Moura. Noncausal Gauss Markov random fields: Parameter structure and estimation. *IEEE Transactions on Information Theory*, 39:1333–1355, 1993.

[2] Michèle Basseville and Igor V. Nikiforov. *Detection of Abrupt Changes: Theory and Applications*. Information and System Sciences Series. Prentice Hall, 1993.

[3] James O. Berger. *Statistical Decision Theory - Foundations, Concepts, and Methods*. Springer Series in Statistics. Springer-Verlag, 1980.

[4] Donald A. Berry and Bert Fristedt. *Bandit Problems: Sequential Allocation of Experiments*. Monographs on Statistics and Applied Probability. Chapman and Hall, 1985.

[5] Julian E. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society B*, 36:192–236, 1974.

[6] R. Cairoli and Robert C. Dalang. *Sequential Stochastic Optimization*. Wiley Series in Probability and Statistics. John Wiley & Sons, 1996.

[7] Fernand S. Cohen. Markov random fields for image modelling & analysis. In Uday B. Desai, editor, *Modelling and Applications of Stochastic Processes*, chapter 10, pages 243–272. Kluwer Academic Publishers, 1986.

[8] Fernand S. Cohen, Zhigang Fan, and Stephane Attali. Automated inspection of textile fabrics using textural models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13:803–808, 1991.

[9] Guy Demoment. Image reconstruction and restoration: Overview of common estimation structures and problems. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37:2024–2036, 1989.

[10] Richard C. Dubes and Anil K. Jain. Random field models in image analysis. *Journal of Applied Statistics*, 16:131–164, 1989.

[11] B. K. Ghosh. *Sequential Tests of Statistical Hypotheses.* Addison-Wesley Publishing Company, 1970.

[12] J. C. Gittins. *Multi-armed Bandit Allocation Indices.* Interscience Series in Systems and Optimization. John Wiley & Sons, 1989.

[13] Robert M. Haralick. Statistical and structural approaches to texture. *Proceedings of the IEEE*, 67:786–804, 1979.

[14] Alf J. Isaksson. Analysis of identified 2-D noncausal models. *IEEE Transactions on Information Theory*, 39:525–534, 1993.

[15] Anil K. Jain. Partial differential equations and finite-difference methods in image processing, part 1: Image representation. *Journal of Optimization Theory and Applications*, 23:65–91, 1977.

[16] Anil K. Jain. Advances in mathematical models for image processing. *Proceedings of the IEEE*, 69:502–528, 1981.

[17] Anil K. Jain and Jaswant R. Jain. Partial differential equations and finite difference methods in image processing, part II: Image restoration. *IEEE Transactions on Automatic Control*, 23:817–834, 1978.

[18] B. Kartikeyan and A. Sarkar. An identification approach for 2-D autoregressive models in describing textures. *CVGIP: Graphical Models and Image Processing*, 53:121–131, 1991.

[19] Rangasami L. Kashyap. Analysis and synthesis of image patterns by spatial interaction models. In Laveen N. Kanal and Azriel Rosenfeld, editors, *Progress in Pattern Recognition*, volume 1, pages 149–186. North-Holland Pub. Co, 1981.

[20] Rangasami L. Kashyap. Characterization and estimation of two-dimensional ARMA models. *IEEE Transactions on Information Theory*, 30:736–745, 1984.

[21] P. M. Kulkarni. Estimation of parameters of a two-dimensional spatial autoregressive model with regression. *Statistics & Probability Letters*, 15:157–162, 1992.

[22] Solomon Kullback. *Information Theory and Statistics.* Wiley Publications in Statistics. John Wiley & Sons, 1959.

[23] K. Kumamaru, S. Sagara, and T. Söderström. Some statistical methods for fault diagnosis for dynamical systems. In Patton et al. [33], chapter 13, pages 439–475.

[24] Wallace E. Larimore. Statistical inference on stationary random fields. *Proceedings of the IEEE*, 65:961–970, 1977.

[25] E. L. Lehmann. *Testing Statistical Hypotheses*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, second edition, 1986.

[26] Robert P. Leland. A new formula for the log-likelihood gradient for continuous-time stochastic systems. *IEEE Transactions on Automatic Control*, 40:1295–1300, 1995.

[27] G. Lorden. Procedures for reacting to a change in distribution. *The Annals of Mathematical Statistics*, 42:1897–1908, 1971.

[28] Rob Luesink. On the likelihood ratio for two-parameter discrete space stochastic processes. *Journal of Multivariate Analysis*, 48:275–296, 1994.

[29] R. K. Mehra and J. Peschon. An innovations approach to fault detection and diagnosis in dynamic systems. *Automatica*, 7:637–640, 1971.

[30] George V. Moustakides. Optimal stopping times for detecting changes in distributions. *The Annals of Statistics*, 14:1379–1387, 1986.

[31] P. M. Newbold and Yu-Chi Ho. Detection of changes in the characteristics of a Gauss-Markov process. *IEEE Transactions on Aerospace and Electronic Systems*, 4:707–718, 1968.

[32] Igor V. Nikiforov. Two strategies in the problem of change detection and isolation. *IEEE Transactions on Information Theory*, 43:770–776, 1997.

[33] Ron Patton, Paul Frank, and Robert Clark, editors. *Fault Diagnosis in Dynamic Systems: Theory and Applications*. Prentice Hall International Series in Systems and Control Engineering. Prentice Hall, 1989.

[34] L. Pelkowitz and S. C. Schwartz. Asymptotically optimum sample size for quickest detection. *IEEE Transactions on Aerospace and Electronic Systems*, 23:263–272, 1987.

[35] Moshe Pollak. Optimal detection of a change in distribution. *The Annals of Statistics*, 13:206–227, 1985.

[36] H. Vincent Poor. *An Introduction to Signal Detection and Estimation*. Springer-Verlag, second edition, 1994.

[37] William K. Pratt. *Digital Image Processing*. John Wiley & Sons, 1978.

[38] Mordechai Segal and Ehud Weinstein. A new method for evaluating the log-likelihood gradient (score) of linear dynamic systems. *IEEE Transactions on Automatic Control*, 33:763–766, 1988.

[39] Robert J. Serfling. *Approximation Theorems of Mathematical Statistics.* Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, 1980.

[40] A. N. Shiryaev. On optimum methods in quickest detection problems. *Theory of Probability and Its Applications*, 8:22–46, 1963.

[41] A. N. Shiryayev. *Optimal Stopping Rules*, volume 8 of *Applications of Mathematics.* Springer-Verlag, 1978.

[42] Lawrence D. Stone. *Theory of Optimal Search*, volume 118 of *Mathematics in Science and Engineering.* Academic Press, 1975.

[43] Abraham Wald. *Sequential Analysis.* Wiley Series in Probablity and Mathematical Statistics. John Wiley & Sons, 1947.

[44] Abraham Wald. *Statistical Decision Functions.* Chelsea Publishing Company, second edition, 1971.

[45] P. Whittle. On stationary processes in the plane. *Biometrika*, 41:434–449, 1954.

[46] Alan S. Willsky. A survey of design methods for failure detection in dynamic systems. *Automatica*, 12:601–611, 1976.

[47] Alan S. Willsky. Detection of abrupt changes in dynamical systems. In Michèle Basseville and Albert Benveniste, editors, *Detection of Abrupt Changes in Signals and Dynamical Systems*, volume 77 of *Lecture Notes in Control and Information Sciences*, pages 27–49. Springer, 1986.

[48] David A. Wilson and Alok Kumar. Derivative computations for the log likelihood function. *IEEE Transactions on Automatic Control*, 27:230–232, 1982.

[49] John W. Woods. Two-dimensional discrete Markovian fields. *IEEE Transactions on Information Theory*, 18:232–240, 1972.

Index

Summary

In general, a detection problem arises when changes may appear in a certain process. These changes may be errors in a dynamic process, objects in images, scratches on surfaces, or anything that is not supposed to be present.

This thesis is concerned with a generalization from one-parameter detection theory to multi-parameter detection theory. The difference between these problems is not just an increase in dimension of the independent parameter. Consider for example a process that is defined as a function of time. At a certain moment in time, a change in the process may appear, and the objective of the problem is to detect this change as soon as possible after it has occurred. On the other hand, if the parameter is given by the position on a surface, so that it is two-dimensional, we cannot speak of a change time. The way the change appears in the measurements depends on the observation mechanism that is used to obtain these measurements. So, instead of a one-dimensional time-scale that is perfectly ordered, a two-dimensional grid, or mult-dimensional index set without any natural ordering has to be used.

It is assumed that a model exists that describes the process under normal circumstances, i.e., when no change is present. Furthermore, models are assumed to exist that describe the process when a change is present. Based on these models, a statistical test may be created in order to detect these changes.

Two classes of statistical tests exist; non-sequential tests and sequential tests. A non-sequential test uses all available data to make a decision about the presence of a change. The performance of such a test is measured by the error probabilities. The disadvantage of these tests is that all data has to be used, which may be expensive and not always necessary.

A sequential test performs a sequence of tests on subsets of the available data. At each stage, a decision is made whether to continue with the next subset or to stop and make a final decision on the presence of a change. An additional performance measure for these tests is given by the number of data points that is required to make a decision.

Chapter 3 deals with the detection of global changes in autoregressive fields. An autoregressive field is characterized by a statistical dependence between neighbouring sites on the grid. Because these random fields are strongly

non-causal, sequential tests may not be evaluated that easily. Hence, only non-sequential tests are examined in this case. The resulting detection problem is rewritten as a standard one-parameter detection problem. In case the parameter becomes three-dimensional, where the third parameter may represent something like time or depth, and there is independence in this third dimension, it is possible to use a sequential test. Again, the resulting detection problem is rewritten as a sequential one-parameter detection problem.

The changes that may appear in the random fields of Chapter 3 are global in the sense that they cover the entire grid. In practice, this will only rarely be the case. Therefore, in Chapter 4 the emphasis lies on the detection of local changes; only small parts of the grid are covered by the change. The random fields in this chapter are assumed to be semi-causal; there exists a dependence between the sites on the grid, but it is possible to define a causal ordering on the grid. The random fields are described by a stochastic dynamic system, and the changes are detected by using a bank of Kalman filters. Although the tests that are used are sequential, the emphasis does not lie on quickest detection.

Finally, in Chapter 5 the quickest detection problem is addressed. So far the observation mechanism has been fixed. In practice it may be more optimal to allow some freedom in the observation mechanism. For example, if previous observations indicate the presence of a change in a certain region, it may be best to continue measuring in that region. The concept of a stopping strategy is introduced to denote the combination of an observation mechanism or sample path and a statistical test.

All random fields are assumed to be independent. The problem is approached from both the Neyman-Pearson point of view and the Bayesian point of view. From the Neyman-Pearson point of view, the quickest detection problem is defined as the minimization of the expected number of observations that are needed to make a decision, when the error probabilities are bounded by given constants. The stopping strategy that minimizes this average site number for all possible changes is said to be uniformly most efficient. Unfortunately, such a stopping strategy generally does not exist. A simplification is given by the use of so-called myopic strategies that only optimize the immediate result.

Using the Bayesian approach, a cost function, defined by a linear combination of the error probabilities and the average site numbers, has to be minimized. The weights of the different terms are determined by the prior probabilities of the changes and the relative cost of the errors with respect to the cost of making an observation. This problem does have a solution, but the optimal stopping strategy is only implicitly defined. Approximations to the optimal strategy are given by n-step look-ahead procedures, that only consider the n next observations in the cost function.

Samenvatting

In het algemeen ontstaat een detectie-probleem als veranderingen in een zeker proces op kunnen treden. Denk hierbij aan fouten in een dynamisch proces, beschadigingen op oppervlakken, voorwerpen op een foto, of willekeurig iets wat niet aanwezig hoort te zijn.

In dit proefschrift wordt het detectie-probleem gegeneraliseerd van één onafhankelijke parameter naar meerdere onafhankelijke parameters. Het verschil tussen deze problemen ligt niet alleen in de verhoogde dimensie van de onafhankelijke parameter. Beschouw bijvoorbeeld een proces dat gedefinieerd is als functie van de tijd. Op een bepaald moment kan een verandering optreden, en het doel van het detectie-probleem is deze verandering zo snel mogelijk te detecteren. Bekijk nu een proces dat gedefinieerd is op een oppervlak, zo dat de onafhankelijke parameter twee-dimensionaal is. In dat geval bestaat er niet zoiets als het moment waarop een verandering optreedt. De manier waarop de verandering in de metingen binnenkomt is afhankelijk van het gekozen observatie-mechanisme. In plaats van een geordende tijds-as zal een ongeordend rooster gebruikt moeten worden als onafhankelijke parameter.

Zowel het proces zonder verandering als het proces met verandering wordt verondersteld volledig gemodelleerd te zijn. Met behulp van deze modellen kunnen statistische toetsen opgesteld worden waarmee een verandering gedetecteerd kan worden.

Er bestaan twee klasses van statistische toetsen. Een niet-sequentiële toets gebruikt alle aanwezige data om in een enkele toets een beslissing te nemen over de aanwezigheid van een verandering. De kwaliteit van zo'n toets wordt gemeten met behulp van de foutkansen. Het gebruiken van alle data kan kostbaar en niet geheel noodzakelijk zijn.

Een sequentiële toets voert een serie van toetsen uit op een groeiende data verzameling. Tijdens elke stap wordt beslist of er gestopt kan worden, of dat er nog meer data nodig is om een uiteindelijke beslissing te kunnen nemen. Een extra kwaliteitsmaat voor deze toetsen is gegeven door de hoeveelheid data die nodig is om een beslissing te nemen.

In Hoofdstuk 3 wordt de detectie van globale veranderingen in autoregressieve velden behandeld. Een autoregressief veld wordt gekarakteriseerd door een statistische afhankelijkheid tussen verschillende punten op het rooster. Omdat deze velden niet causaal zijn is het moeilijk om met sequentiële toetsen te werken. Daarom wordt alleen naar niet-sequentiële toetsen gekeken. Het detectie-probleem is herschreven als een standaard één-parameter detectie-probleem. Als de parameter drie-dimensionaal wordt, waarbij de derde parameter iets als de tijd of een diepte voor zou kunnen stellen, en er onafhankelijkheid is in deze derde dimensie, dan kan er alsnog gebruik gemaakt worden van een sequentiële toets. Ook hier is het probleem herschreven als een standaard één-parameter detectie-probleem.

De veranderingen die voor kunnen komen in de stochastische velden van Hoofdstuk 3 beslaan het complete rooster waarop het veld gedefinieerd is. In de praktijk zal dit niet vaak voorkomen. Daarom wordt in Hoofdstuk 4 de nadruk gelegd op de detectie van lokale veranderingen; slechts een gedeelte van het rooster wordt beïnvloed door de verandering. De stochastische velden in dit hoofdstuk zijn semi-causaal; er bestaat een afhankelijkheid tussen de punten op het rooster, maar het is mogelijk om een volgorde op het veld te definiëren. De velden worden beschreven door stochastische dynamische systemen en de veranderingen worden gedetecteerd met behulp van een verzameling Kalman filters. Hoewel de toetsen sequentieel zijn, ligt de nadruk niet op het zo snel mogelijk detecteren van de veranderingen.

Tenslotte wordt in Hoofdstuk 5 het snelste detectie-probleem beschouwd. Het observatie-mechanisme had tot nu toe altijd vastgelegen. In de praktijk lijkt het echter beter om hier ook enige vrijheid in te laten. Bijvoorbeeld, als op basis van eerdere metingen het vermoeden bestaat dat een verandering in een bepaald gebied zal liggen, zal de volgende meting logischerwijs in dat gebied gedaan worden. De strategie die gebruikt wordt om een verandering te detecteren bestaat dus niet alleen uit een statistische toets, maar ook uit een observatie-mechanisme.

Alle stochastische velden worden verondersteld onafhankelijk te zijn. Het probleem wordt op twee manieren benaderd. Vanuit het perspectief van Neyman-Pearson wordt het snelste detectie-probleem gedefinieerd als het minimaliseren van het verwachte aantal observaties dat nodig zal zijn om een verandering te detecteren, gegeven dat de foutkansen begrensd zijn door bepaalde constantes. Helaas heeft dit probleem niet altijd een oplossing. Ook is gekeken naar een vereenvoudigde versie van dit probleem, waarbij alleen naar het directe resultaat gekeken wordt.

Vanuit het perspectief van Bayes wordt een kostenfunctie geminimaliseerd. De gewichten van de termen in deze kostenfunctie worden bepaald door een a-priori kansverdeling van de veranderingen en de relatieve kosten van een fout met respect tot de meetkosten. Dit probleem heeft wel een oplossing, maar de optimale strategie is impliciet gedefinieerd. Benaderingen van de oplossing worden gegeven door slechts een beperkt aantal stappen vooruit te kijken.

About the author

Erik Peter Hupkens was born on January the 9th of the year 1970 in Coevorden, a small city in the northern part of the Netherlands. After finishing highschool in Hardenberg, he studied Applied Mathematics at the University of Twente in Enschede, where he specialized in System- and Control theory. His final year project was concerned with the use of neural networks for pattern recognition and system identification. Part of this project was performed at the Jet Propulsion Laboratory in Pasadena, in the USA, where he stayed for six months in the year 1992. After finishing his studies in 1993, he continued his stay at the University of Twente to obtain a PhD in Applied Mathematics, leading to the present thesis. His main interests are in stochastic system theory and detection theory.

In his spare time, he enjoys playing football (soccer). Furthermore, listening to music, reading literature and watching cycling competitions on television are some of the more passive passings of time that he likes.