IDEA League

Master of Science in Applied Geophysics

Research Thesis

Test

Subtitle

Fabian Antonio Stamm

March 24, 2017

Test
Subtitle

Master of Science Thesis

for the degree of Master of Science in Applied Geophysics at
Delft University of Technology
ETH Zürich
RWTH Aachen University
by

Fabian Antonio Stamm

March 24, 2017

Department of Geoscience & Engineering · Delft University of Technology
Department of Earth Sciences · ETH Zürich
Faculty of Georesources and Material Engineering · RWTH Aachen University

Delft University of Technology

Printed in The Netherlands, Switzerland, Germany

IDEA LEAGUE
JOINT MASTER'S IN APPLIED GEOPHYSICS

Delft University of Technology, The Netherlands
ETH Zürich, Switzerland
RWTH Aachen, Germany

Dated: *March 24, 2017*

Supervisor(s):

Prof. Florian Wellmann, Ph.D.

Prof. Dr. Janos Urai

Committee Members:

Prof. Florian Wellmann, Ph.D.

Prof. Dr. Janos Urai

Miguel de la Varga, M.Sc.

Abstract

Please pay particular attention to the preparation of your abstract; use this text as a guide. Every master thesis report must be accompanied by an informative abstract of no more than one paragraph (max 300 words). The abstract should be self-contained. No references, figures, tables, or equations are allowed in an abstract. Do not use new terminology in an abstract unless it is defined or is well-known from the literature. The abstract must not simply list the topics covered in the paper but should (1) state the scope and principal objectives of the research, (2) describe the methods used, (3) summarize the results, and (4) state the principal conclusions. Do not refer to the master thesis report itself in the abstract. For example, do not say, "In this thesis we will discuss". Furthermore the abstract must stand alone as a very short version of the master thesis report rather than as a description of the contents. Remember that the abstract will be the first and most widely read portion of the master thesis report. Readers will be influenced by the abstract to the point that they decide to read the master thesis report or not.

Acknowledgements

First of all I want to thank all the people who have participated in this project .. Remember, often more people have (in some way) contributed to your final thesis than you would initially think of....

RWTH Aachen University | Fabian Antonio Stamm
March 24, 2017

Table of Contents

List of Figures

List of Tables

Acronyms

RWTH Aachen University

March 24, 2017

Chapter 1

Introduction

Welcome to the standard layout for your IDEA LEAGUE MSc thesis written in LaTeX. LaTeX has a variety of advantages over conventional/ standard text editing programs, which you will soon enough discover yourself. LaTeX almost forms a standard in the Scientific Community, especially due to its effective and straightforward mathematical capabilities.
This is Chapter 1. If you want to know more about LaTeX you better read [Knuth, 1984] or use the extensive help available on the internet. . This 'hidden' index command helps you making an index at the end of your thesis. You can add this flag anywhere you want to make an index hit. You can see here also how to use acronyms, like **DUT!** (**DUT!**). The acronyms are automatically listed in the corresponding section. Also, hyperlinks are created automatically with the developed class file, such that your digital PDF version of your thesis can be read dynamically. Have fun with LaTeX and your M.Sc. research project and good luck!

The purpose of the introduction is to tell readers why they should want to read what follows the introduction. This chapter should provide sufficient background information to allow readers to understand the context and significance of the problem. This does not mean, however, that authors should use the introduction to rederive established results or to indulge in other needless repetition. The introduction should (1) present the nature and scope of the problem; (2) review the pertinent literature, within reason; (3) state the objectives; (4) describe the method of investigation; and (5) describe the principal results of the investigation.

Part I

First Part

Chapter 2

Methods

The methods utilized in this work are presented in the following chapter. This includes introductions into Bayesian analysis and decision theory. In particular Bayesian inference, estimation of uncertain values and how loss function can be used in this context. These methods are incorporated in the context of numerical modeling of structural geological settings. This is done by programming in a python environment. Central tools for model construction and conduction of the statistical evaluation are Geomodeler3D, GemPy and PyMC in particular and are also presented in this chapter.

2-1 Bayesian analysis and decision theory

Problems and reasoning behind decision making are examined in the field of decision theory, as implied by the name [?]. Such decision problems are commonly influenced by parameters that are uncertain. In statistical decision theory, the presence of statistical knowledge is used to gain information on the nature of these uncertainties. (Such uncertain parameters can be considered as numerical quantities.) In order to find the best decision to a problem, it is possible to combine sample information with other aspects such as the possible consequences of decision making and the availability of prior information on our uncertainties. Decision consequences are expressed as gains in economic decision theory and as losses, which equal negative gains, in statistics. Prior information might be given for example due to experience from previous similar problems or from expert knowledge (see Batvold and Begg). The approach of utilizing such as priors is known as Bayesian analysis and is explained in the following [?]. (It goes well with decision theory.)

2-2 Basic elements

Some basic elements are to be defined. The unknown (uncertain?) quantity influencing decision making is usually named denominated as θ, the state of nature. Given statistical information on θ in the form of probability distributions, θ is called the parameter. Decisions are also referred to as actions a. The outcome of statistical tests in form of information or statistical evidence is denoted as X. Loss is defined as $L(\theta, a)$, so $L(\theta_1, a-1)$ is the actual loss incurred when action a_1 is taken while the true state of nature is θ_1 [?]. Loss, expected loss and loss functions are further explained below.

2-2-1 Bayesian inference

Bayesian inference is most importantly characterized by its preservation of uncertainty, in contrast to standard statistical inference [?]. Probability is seen as a measure of belief for an event to occur. It has been argued by [?] that this Bayesian approach is intuitive and inherent in the natural human perspective. These beliefs can furthermore be assigned to individuals [?]. Thus, different and even contradicting beliefs about the probability of an event might be held by different individuals, based on variations and disparities in the information available to each one individual [?]. The initial belief or guess about an event A can be denoted as P(A). This is used as the so-called prior probability on which Bayesian updating is based. The beliefs about the occurrence of an event are revalued in the presence of additional information, i.e. the observation of new evidence which can be denominated as X. These observations are included as likelihoods P(X|A). This process of updating results in a posterior probability P(A|X) [?]. It is important to note that the prior is not simply discarded but re-weighted by Bayesian updating. It was also pointed out by [?] that by utilizing an uncertain prior, the potential for wrongfulness of the initial guess is already included. This means that Bayesian updating is about reducing uncertainty in a belief and reaching a guess that is less wrong [?]. Bayesian updating is defined by and conducted via the following equation, called the Bayes' Theorem:

$$P(A|X) = \frac{P(X|A)P(A)}{P(X)} \propto P(X|A)P(A) \tag{2-1}$$

-Law of Large Numbers!

2-2-2 Estimation

The resulting posterior distribution can be used to acquire point estimates for the true state of nature θ. Common and simple examples for such estimators are the mode (i.e. the generalized maximum likelihood estimate), the mean and the median of a distribution [?]. The presentation of a point estimate should usually come with a measure for its estimation error. According to [?], the posterior variance is most commonly used as an indication for estimate accuracy. However, it is argued by [?] that by using pure accuracy metrics, while this technique is objective, it ignores the original intention of conducting the statistical inference in cases, in which payoffs of decisions are valued more than their accuracies. A more appropriate approach can be seen in the introduction of loss and the use of loss functions [?].

2-2-3 Expected loss and loss functions

Loss is a statistical measure of how bad an estimate is, i.e. how much is lost by making a certain decision. Gains are considered by statisticians as negative losses [?]. The magnitude of an estimate's loss is defined by a loss function, which is a function of the estimate of the parameter and the true value of the parameter [?]:

$$L(\theta, \hat{\theta}) = f(\theta, \hat{\theta}) \tag{2-2}$$

So, "how bad" a current estimate is, depends on the way a loss function weights accuracy errors and returns respective losses. Two standard loss functions are the absolute-error and the squared-error loss function. Both are simple to understand and commonly used [?].

As implied by its name, the absolute-error loss function returns loss as the absolute error, i.e. difference between the estimate and the true parameter [?]:

$$L(\theta, \hat{\theta}) = |\theta - \hat{\theta}| \tag{2-3}$$

Accordingly, losses increasing linearly with the distance to the true value are returned for respective estimates.

Using the squared-error loss function returns losses that increase quadratically with distance of the estimator to the true parameter value [?]:

$$L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|^2 \tag{2-4}$$

This exponential growth of loss also means that large errors are weighted much stronger than small errors. This might come with overvaluation of distant outliers and misrepresentation of magnitudes in distance. Regarding this, the absolute-error loss function can be seen as more robust [?].

- each give mean or median –¿ estimator with accuracy metric? objective and symmetric! aim at high precision.

-shift in prespective. look at outcome instead of estimation precision!

[?] and (LF DESIGN PAPER?) propose that it might be useful to move away from these type of objective loss functions to the design of new loss functions that are specifically customized to an reflect an individual's, i.e. the decision maker's objectives, preferences and outcomes.

The standard loss function defined above are symmetric, but can easily be adapted to be asymmetric, for example by weighting on the negative side stronger than those on the positive side. Preference over estimates larger than the true value is thus incorporated in an uncomplicated way [?]. Much more complicated designs of loss functions are possible, depending on purpose, objective and application [?]. Some case-specific loss functions are designed in the following chapters of this work.

- Cite from paper about the design of loss functions! Subjectivity of customized loss functions!

The presence of uncertainty during decision making implies that the true parameter is unknown and thus the truly incurred loss $L(\theta, a)$ cannot be known at the time of making the decision [?, ?]. The Bayesian perspective considers unknown parameters as random variables and samples that are drawn from the posterior distribution as possible realizations of the unknown parameter, i.e. all possible true values are represented by this distribution [?]. A suitable alternative to the actual loss is to consider a decision's expected loss and to make a decision that is optimal in relation to this expected loss [?].

Given a posterior distribution $P(\theta|X)$, the expected loss of choosing an estimate $\hat{\theta}$ over the true parameter θ (after evidence X has been observed) is defined by the function below [?]:

$$l(\hat{\theta}) = E_\theta[L(\theta, \hat{\theta})] \tag{2-5}$$

The expectation symbol E is subscripted with θ, by which it is indicated that θ is the respective unknown variable. This expected loss as defined above, is also referred to the risk of estimate $\hat{\theta}$ [?].

By the Law of Large Numbers, the expected loss of $\hat{\theta}$ can be approximated drawing N samples from the posterior distribution, respectively applying a loss function L and averaging of the number of samples [?]:

$$\frac{1}{N}\sum_{i=1}^{N} L(\theta_i, \hat{\theta}) \approx E_\theta[L(\theta, \hat{\theta})] = l(\hat{\theta}) \tag{2-6}$$

Minimization of 2-5 or 2-6 returns a Bayesian point estimate known as Bayes action $\delta^P(X)$ (???), which is the estimate, action or decision with the least expected loss according to the loss function [?]. For a unimodal and symmetric absolute-error loss function, the Bayes action is simply the median of the posterior distribution, while squared-error loss it is the mean [?, ?]. The MAP estimate is returned when using zero-one loss [?] (WRITE THIS?). The possibility of more than on minimum also implies that several Bayes actions can exist for one problem [?].

- MAP?

[?] implemented different risk affinities by simply introducing a risk parameter into the loss function. By using different values for this parameter, it can be represented how comfortable an individual is with being wrong and furthermore which "side of wrong" is preferred by this decision maker [?]. In the chapter below, different risk parameters are introduced to alter the weighting in the loss functions (???).

2-2-4 Value of information

- considering the change in gain (or loss) after regarding new information or evidence, the value of this information can be calculated - in the presence of uncertainty we look at expected value of information - this can be used as a measure, to assess if the effort to attain new evidence or information is worth it - also as a measure to see how much additional information is worth to different actors with different risk affinities - here we use the comparison between Bayes actions before and after Bayesian updating

2-3 Application in structural geological modeling

- applying these statistical theories in the context of geological modeling - as done by De la Varga (2015)

- equivalence of basic elements in geologic models, some examples

- sampling methods

- numerical implementation in python using pyMC

- construction of 1D model in python and pyMC

- construction of 3D model using GeoModeller3D and GeMPy additionally

- what about theory of petroleum reservoir and volumetric calculation? explain this in later chapter of 3D model???

2-4 Example section

This is the section. Referring to equations, figures and tables can easily be done by the commands `\eqnref{}`, `\figref{}` and `\tabref{}`.

$$H(s) = \frac{1}{s+2} \tag{2-7}$$

You see? Refer to equations like this Eq. (2-7), i.e. the name of the label you have given the specific equation, figure or table.

2-4-1 The first subsection

Now I demonstrate, numbering equations, using subequations:

$$\nabla \times \mathbf{L} = \frac{\partial \mathbf{G}}{\partial t} \tag{2-8a}$$

$$\nabla \times \mathbf{G} = \frac{\partial \mathbf{L}}{\partial t} + \mathbf{J} \tag{2-8b}$$

$$\mathbf{G} = \sigma \mathbf{J} \tag{2-8c}$$

Or we can make matrices:

$$\mathbf{Q}_{12} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

This can also be done using the `\align{}` command. Equation arrays are also possible:

$$\nabla \times \mathbf{L} = \frac{\partial \mathbf{G}}{\partial t} \tag{2-9}$$

$$\nabla \times \mathbf{G} = \frac{\partial \mathbf{L}}{\partial t} + \mathbf{J} \tag{2-10}$$

$$\mathbf{G} = \sigma \mathbf{J} \tag{2-11}$$

The first sub-subsection with a very very very long title, but in the table of contents one can only see the short title in square brackets

Impressed by the capabilities? If you want to know more about the capabilities of LaTeX, take a look at the "**The Not So Short Introduction to LaTeX 2$_\varepsilon$**", which can be found on the internet.

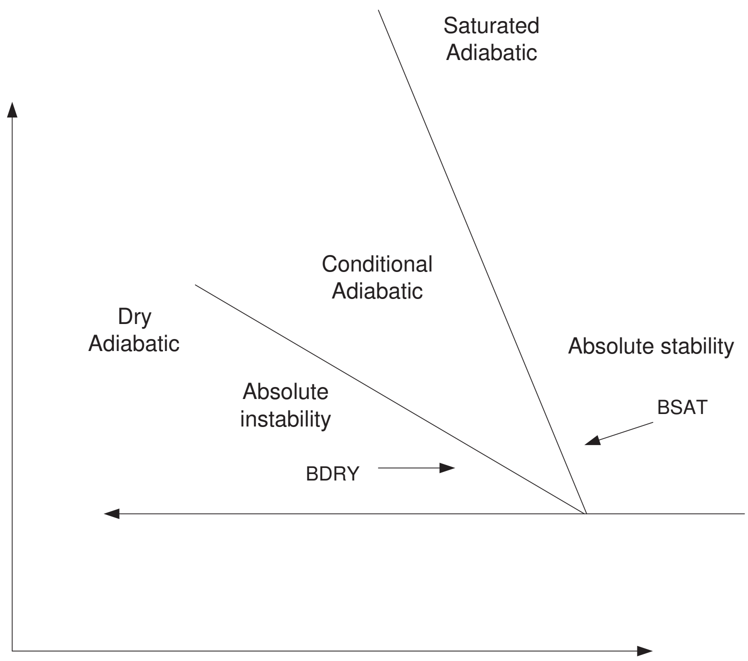

Figure 2-1: Stability conditions for the vertical stability of saturated and unsaturated air.

Next paragraph. .
And finally I end this example file with a table which will be centered in the middle of the following page

Data **files listed in a table**	
First part	
Fabracadabra.m	- Saturation computation
Fobracadabra.m	- Pressure computation
Fibricadibri.m	- Permeability computation
Structural rock model	
struct.m	- Rock structural data using symmetric boundary condition
bstruct.m	- Rock structural data using anti-symmetric boundary condition

Table 2-1: Good-looking program data deck files.

Bibliography

[Knuth, 1984] Knuth, D. E. (1984). *The TEXbook*. Addison-Wesley.

Appendix A

The back of the thesis

A-1 An appendix section

A-1-1 An appendix subsection with C++ Lisitng

```cpp
//
// C++ Listing Test
//

#include <stdio.h>

for(int i=0;i<10;i++)
{
    cout << "Ok\n";
}
```

A-1-2 A Matlab Listing

```matlab
%
% Comment
%
n=10;
for i=1:n
    disp('Ok');
end
```

Appendix B

Yet another appendix

B-1 Another test section

Ok, all is well.

Index