Risk Assessments of Minefields in Humanitarian Mine Action – A Bayesian Approach

Jan Bastholm Vistisen

Kgs. Lyngby 2006
IMM-PhD-2006-161

Technical University of Denmark
Informatics and Mathematical Modelling
Building 321, DK-2800 Lyngby, Denmark
Phone +45 45253351, Fax +45 45882673
reception@imm.dtu.dk
www.imm.dtu.dk

IMM-PhD: ISBN 87-643-0037-4
ISSN 0909-3192

Foreword

The management of the large number of areas found in many post-conflict countries, suspected or verified of being contaminated by mines, poses a major challenge to decision makers involved in the administration of national mine action programmes. Analytical tools are therefore needed which can facilitate the identification of the most important minefields with respect to mine clearance. In February 2002, the Danish Defence Research Establishment initiated in collaboration with the Technical University of Denmark a Ph.D.-project to investigate whether the application of operations research or statistics can support the Humanitarian Mine Action sector to make the prioritization of mine clearance operations more effective. The present Ph.D.-thesis summarizes the results from the completed research project.

English Summary

During the last 10-15 years, the international community has become aware of the devastating mine contamination problems experienced in many post-conflict countries. As a consequence, a considerable amount of money and time is spent on research and development in new ways of locating buried mines and unexploded ordnance in a fast and secure way. A major breakthrough is however still waiting, and a large fraction of the mine clearance, which still remains to be done, will therefore hinge on slow and dangerous procedures based on prodders and metal detectors.

Realizing that landmine contamination is a phenomenon which cannot be eliminated overnight but is a problem which has to managed in several years to come, it is essential that the resources a national government in a mine affected country spends on mine clearance are used on the right projects. However, the identification of the mine clearance projects with the greatest impact is a delicate task. More systematic approaches to the *ranking* of minefields with respect to mine clearance can be found in the literature, but these methods are either founded on simple scoring rules or are of a more qualitative nature. Thus nobody seems yet to have examined the usefulness of the analytical tools which might be provided by operations research and statistics in order to support decision makers involved in national mine clearance programmes.

In February 2002, the Danish Defence Research Establishment initiated in collaboration with the Technical University of Denmark a Ph.D.-project to investigate whether the application of operations research and statistics can support decision makers in Humanitarian Mine Action to make the prioritization of mine clearance operations more effective. The main part of that project, which is presented in the enclosed thesis, has concentrated on the development of a risk model quantifying to what extent a minefield poses a risk to a society.

The risk model is derived in two steps: First, a general model, which requires detailed information about the mined area in question, is derived. Secondly, by the introduction of

two additional assumptions, the general model is turned into a simple binomial model depending on two parameters m and θ. In this context the integer m denotes the number of so-called *functional mines* in the minefield under consideration, and the parameter θ denotes the probability of a randomly selected mine being encountered by a person, a vehicle, etc... during a predefined observation period.

The true values of the binomial parameters, which jointly characterize the state of the mined area, will rarely be known in advance, but beliefs about these based on whatever information is available can conveniently be expressed in terms of probability distributions $p(m)$ and $p(\theta)$. This prepares the way for the introduction of Bayesian data analysis by which updates of the probability distributions can be generated from incoming accident statistics.

The major obstacle to a real-life application of the derived risk model seems to be the lack of actual information about the binomial parameter θ. A considerable part of the enclosed thesis focuses therefore on ways to provide information about θ through statistical modelling. Depending on the level of *historical information* available to a hypothetical decision maker, two different proposed models are examined as ways of extracting information about θ : 1) A simple *hierarchical model* which as input requires accident statistics and clearance reports from already cleared minefields; 2) A *finite mixture model* where only accident statistics and the specification of certain prior distributions are needed as input data. Common to both models is the generation of posterior distributions of the parameter θ. To extract information about θ from these distributions various simulation techniques are applied including importance sampling and Markov Chain simulation.

The possibility of making updates of the entering probability distributions $p(m)$ and $p(\theta)$ through incoming accident statistics by the use of *Bayes' rule* makes the suggested risk model dynamic. Moreover, the application of Bayesian data analysis gives the derived risk model a very flexible structure which allows an accommodation to the varied circumstances found in Humanitarian Mine Action with respect to the amount of accessible information. The present thesis closes with an overall prescription for the synthesis of different pieces of information based on the concept of *reference priors*.

Danish Summary / Dansk Resumé

Indenfor de seneste 10-15 år er det internationale samfund i stigende grad blevet opmærksom på de ødelæggende mineforureningsproblemer, som eksisterer i mange post-konflikt lande. Som en konsekvens heraf investeres i dag en betragtelig mængde af penge og tid på forskning og udvikling af hurtigere og pålideligere metoder til lokalisering af nedgravede miner og ueksploderet ammunition. Et større teknisk gennembrud lader imidlertid vente på sig. Det må derfor forventes, at velprøvede men langsommelige minerydningsteknikker baseret på minesonder og metaldetektorer også i fremtiden vil spille en betydelig rolle – og mineforureningen vil derfor være et fænomen i de berørte lande, som skal håndteres i mange år fremover.

I denne situation er det afgørende, at de begrænsede økonomiske ressourcer, som et land afsætter til minerydning, udnyttes optimalt. Udpegningen af de rydningsprojekter, hvis gennemførelse vil have den største samfundsmæssige effekt – herunder reducere risikoen for fremtidige mineulykker - er imidlertid en vanskelig opgave. Mere systematiske tilgange til prioriteringen af minefelter med henblik på senere minerydning kan findes i litteraturen, men disse metoder er enten simple kvantitative metoder eller er af en mere kvalitativ karakter. De muligheder, som eksempelvis inddragelsen af analytiske redskaber hentet fra operationsanalyse eller statistik kunne tilvejebringe, er derimod mangelfuldt beskrevet.

I februar 2002 igangsatte Forsvarets Forskningstjeneste i et samarbejde med Danmarks Tekniske Universitet et PhD-projekt med det formål at undersøge, hvorvidt inddragelsen af operationsanalyse eller statistik kan støtte beslutningstagere indenfor humanitær minerydning med henblik på at opnå en optimal ressourceudnyttelse. Hovedparten af dette projekt, der præsenteres i vedlagte PhD-afhandling, har koncentreret sig om udviklingen af en risikomodel, som kvantificerer den trussel et minefelt udgør for det omkringliggende samfund.

Ovenstående risikomodel udledes i to trin: Indledningsvis udledes en overordnet model, som kræver detaljeret information om minefeltet, der ønskes risikovurderet. Ved

anvendelsen af to forsimplende antagelser transformeres den overordnede model til en simpel binomial model, der afhænger af parametrene m og θ. Heltalsparameteren m angiver antallet af såkaldte *funktionelle miner* i minefeltet under vurdering, mens parameteren θ angiver sandsynligheden for, at en tilfældigt udvalgt mine i minefeltet bliver antruffet af en person, et køretøj, etc... indenfor en nærmere angivet observationsperiode.

De sande værdier af ovenstående binomialparametre, som tilsammen karakteriserer det pågældende minefelts tilstand, vil sjældent være kendte på forhånd, men vurderinger af disse baseret på den tilgængelige information kan passende udtrykkes i form af sandsynlighedsfordelinger $p(m)$ og $p(\theta)$. Dette baner vejen for introduktionen af Bayesiansk dataanalyse, som muliggør opdateringer af de opstillede sandsynlighedsfordelinger via *Bayes' regel*.

En betydelig del af PhD-afhandlingen fokuserer på metoder til tilvejebringelse af information om parameteren θ gennem statistisk modellering. Afhængig af mængden af *historisk information*, som er tilgængelig for en hypotetisk beslutningstager, undersøges to forskellige metoder til ekstraktion af information om θ: 1) En simpel *hierarkisk model* hvor ulykkesstatistikker og rydningsrapporter fra allerede ryddede minefelter udgør inddata; 2) En *finite mixture model* hvor kun ulykkesstatistikker samt specifikationen af visse *a priori* fordelinger indgår som inddata. Fælles for begge modeller er frembringelsen af *posteriori* fordelinger for parameteren θ. For at udtrække information om θ fra disse fordelinger anvendes forskellige simulationsteknikker, eksempelvis *importance sampling* og *Markov Chain simulation*.

Opdateringen af de indgående sandsynlighedsfordelinger $p(m)$ og $p(\theta)$ via indkommende ulykkesstatistikker gør den udledte risikomodel dynamisk. Anvendelsen af Bayesiansk dataanalyse giver derudover risikomodellen en fleksibel struktur, hvilket muliggør en tillempning af modellen til de meget varierende forhold som forefindes indenfor humanitær minerydning. Den vedlagte PhD-afhandling afslutter med en overordnet forskrift på syntesen af forskellige fragmenter af relevant information og dets overførsel til risikomodellen baseret på konceptet *reference priors*.

Acknowledgements

The work leading to the enclosed PhD-thesis was along the way supported by various people. In this connection I would like to thank my supervisors *Torben Christensen* from the Danish Defence Research Establishment and *Jens Clausen* from the Technical University of Denmark.

For their continued moral support and encouragement during the ongoing project I would like to thank *Ole Nymann* from Nordic Demining Research Forum, *Jan Larsen* from the Technical University of Denmark (IMM), *Bjarne Haugstad* from the Norwegean Defence Research Establishment, and *Svend Clausen* from the Danish Defence Research Establishment.

Thanks also to *Bo Bischoff*, former Head of Danish Demining Group, who introduced me to many of the practical aspects and problems which are encountered in Humanitarian Mine Action, and who took the time to read and comment my initial writings.

Most of all, however, I should like to thank my wife, *Lone*.

Jan Vistisen, Copenhagen, 15.1.2006.

Contents

Chapter 1
The Land Mine Problem

Globally, land mines claim an estimated 15,000-20,000 civilian victims per year in 90 countries, and about 40-50 million mines remain to be cleared [MacDonald et al., 2003]. Besides the suffering and death caused by mines, the sheer presence of mines or the mere suspicion of their presence has far reaching consequences in terms of blockage of reconstruction and economic growth in many mine affected countries. The recognition of the size of the global land mine problem made in 1994 the United Nations (UN) to declare that "*land mines may be one of the most widespread, lethal and long-lasting forms of pollution we have yet encountered*" [United Nations, 1994].

One manifestation of the growing international understanding of the land mine problem is the emergence of the civilian discipline *Humanitarian Mine Action* (HMA) whose core activities include mine clearance operations in post-conflict countries. Since its advent in the late eighties the HMA sector has undergone a tremendous development. Another manifestation is the intensification in research aiming at improving the mine detection technology. Unfortunately the search for a replacement of the simple metal detector used in manual demining has turned out to be a much larger technological challenge than anticipated at first. As a consequence, the predominant part of mine clearance operations in the foreseeable future will still hinge on manual demining. Mine clearance remains thus to be a very slow, troublesome and dangerous business. At the current rate, the clearing of all existing minefields will approximately require 450-500 years [MacDonald et al., 2003].

Realizing that landmine contamination is a phenomenon which cannot be eliminated overnight but is a problem which has to managed in several years to come, it is essential that the resources a national government in a mine affected country spends on mine clearance are used on the right projects. However, the identification of the mine clearance projects with the greatest impact is a delicate task. More systematic approaches to the *ranking* of minefields with respect to mine clearance can be found in the literature but these methods are either founded on simple scoring rules [GICHD, 2001] or are of a more

qualitative nature [Millard, 2000, 2001]. Thus nobody seems yet to have examined the potential usefulness of the strong analytical tools provided by operations research and statistics to support the decision makers involved in HMA.

By the present thesis the first step in the above direction has been taken. Thus in the chapters which follow a general framework based on Bayesian data analysis is introduced which can support decision makers in their efforts to identify the most important minefields with respect to mine clearance. It is not claimed that the suggested mathematical models provide the full picture of all facets of the landmine problem in a given country. Alternative methods taking a more qualitative approach are therefore still needed to complement the analysis. The outlined framework nevertheless represents a very structured way of collecting and synthesizing information which can minimize the risk of future minefield accidents.

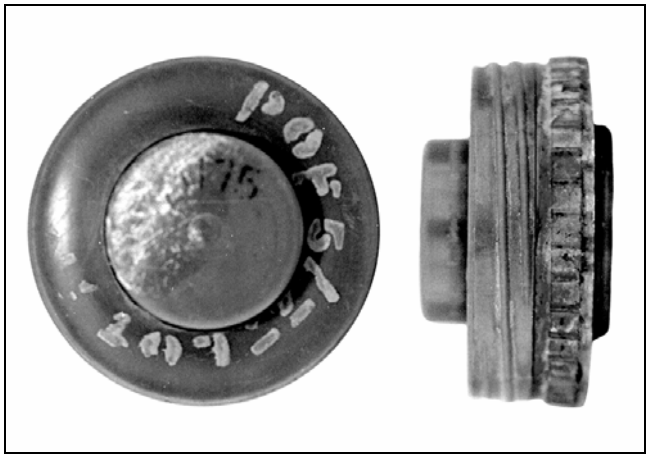
The coming chapters 2-15 are of a quite technical character and to appreciate their contents, the present chapter provides a background to certain aspects of the global landmine problem. Thus in paragraph 1.1, the meaning of the word "mine" is defined, and a brief historical account of the origin and development of HMA is given. The main contents of paragraph 1.1 are based on the publication "A guide to Mine Action" by the Geneva International Centre for Humanitarian Demining [GICHD, 2004]. Paragraph 1.2 summarizes the current state of HMA. In paragraph 1.3 the discussion about impact and prioritizations in HMA is introduced, and the merits and shortcomings of the so-called *mine impact score model* are mentioned. In paragraph 1.4 the research objectives of the present thesis are defined, and possible techniques from operations research or statistics which might be brought into play to reach the defined objectives are discussed. Finally paragraph 1. 5 outlines the contents of the chapters 2-15.

1.1 Introduction to Humanitarian Mine Action

According to the *Anti-Personnel Mine Ban Convention* [for a thorough introduction, see GICHD, 2004] a mine is defined as "*a munition designed to be placed under, on or near the ground or other surface area and to be exploded by the presence, proximity or contact of a person or a vehicle*". As illustrated in fig. 1.1, a landmine is in principle a very simple piece

of device. It consists of a casing made by metal, plastic or wood containing a piece of explosive material. The casing contains furthermore a fuzing mechanism to initiate the detonation of the explosive which is typically activated by a vertical pressure on the casing or by the extension of a connected tripwire. Certain types of mines may also be activated from distance by remote control.

Fig. 1.1. Anti-personnel mine (AP). Photo: Danish Demining Group.



Landmines are manufactured in a variety of different sizes and shapes but may generally be classified as either *anti-tank mines* (*AT-mines*) or *anti-personnel mines* (*AP-mines*) depending on whether the intended victim is a vehicle or a person. Where the threshold "pressure" to activate an AP-mine is typically of the order of 10 kg or less, an AT-mine usually demands a vertical pressure equivalent to several hundreds of kg. Depending on how the mine injuries its victim, AP-mines may be classified further as blast-, fragmentation-, bounding-, or directional fragmentation-mines. There are today approximately 700 types of manufactured AP-mines excluding the improvised (home made) mines [Handicap International, 2000] .

Even though landmines have been used excessively in international or local conflicts at several occasions during the 20th century, the emergence of *Humanitarian Mine Action* (HMA) as a discipline is of relatively recent date. Its origin can thus be traced back to October 1988, where the United Nations for the first time appealed for funds for *humanitarian demining* in Afghanistan [GICHD, 2004]. At that time, the Soviet troops were about to leave Afghanistan, and the Afghan society was left with a severe mine

contamination problem but without a functioning national army to address the clearance of the minefields.

As a result of the UN initiative, more than 10,000 Afghan refugees received basic mine clearance training by military contingents from donor countries. The UN furthermore supported the creation of a number of NGO's (Non Governmental Organizations) to survey, map, mark and clear minefields and support the civilian population through mine awareness campaigns.

The initiatives seen in 1988 in Afghanistan were notable for various reasons: Firstly, the term *humanitarian demining* implied demining activities for humanitarian purposes, and the phrase was thus deliberately used to distinguish it from military demining (so-called *breaching*). Secondly, where mine clearance previously had been entrusted to military units, mine clearance and related activities became now a possible civilian occupation.

The end of the Gulf War in 1991 marked the second major event in mine action. During the subsequent mine clearance programme in Kuwait which lasted from 1991-1993, mechanical mine clearance with flails and tillers was introduced, and several commercial companies entered the field of mine action.

In the following years from 1992-1994, UN-assisted mine action programmes were planned and initiated in Cambodia, Mozambique and Angola with varying degrees of success. An important event was the establishment of the Cambodian Mine Action Centre (CMAC), which was set up in 1992 and was intended a leading and coordinating role of the Cambodian mine action programme. This programme has since then turned into one the largest mine action programmes worldwide. Similar mine action centres have been established in a variety of mine affected countries during the nineties.

Important lessons were learned during the first half of the nineties. Firstly, the presence of national authorities capable of regulating, coordinating and sustaining programme objectives were prerequisites for successful completions of national mine action programmes. Secondly, with an increasing number of actors with various backgrounds involved in mine action, there was a need to standardize the different components of mine

action. Consequently, a conference on international standards for humanitarian mine clearance programmes was launched in Denmark in 1996, and proposals from the conference were subsequently by a UN-led working group developed into the standards *International Standards for Humanitarian Mine Clearance Operations*, released in 1997 (these standards have since 2001 been superseded by the *International Mine Action Standards*, *IMAS*).

Besides the increasing number of mine action programmes which were set up during the last half of the nineties, e.g., Albania, Bosnia and Herzegovinia, Northern Iraq, etc., the launch of the *Convention on the Prohibition of the Use, Stockpiling, Production and Transfer of Anti-Personnel Mines and on Their Destruction* (in short, the *Anti-Personnel Mine Ban Convention*) in 1997 contributed to an enhanced public awareness of the impact of the global mine contamination problem. Signatory States of this convention undertake never under any circumstances to use, produce, develop, stockpile, or transfer anti-personnel mines, or to assist, encourage, or induce anyone to commit such acts. Signatories are furthermore obliged to clear all anti-personnel mines in mined areas under their jurisdiction not later than 10 years after they become Parties to the Convention. When the *Anti-Personnel Mine Ban Convention* entered into force in 1999, 133 States had signed the Convention. Today, i.e., 2005, more than two thirds of the States in the world have signed the Convention.

1.2 Humanitarian Mine Action Today

The main objective of humanitarian demining is to clear all mines and other explosive remnants of war from a given area such that the area is safe to the civilian population. Unfortunately, no existing mine clearance method applied in HMA can guarantee a 100% clearance.

In manual demining, which is the most frequently used method under mine clearance operations, a metal detector is used for the location of buried metal containing mines, and an excavator or prodder is subsequently used to uncover the mine. The repeated process of detection and uncovering is dangerous and time consuming due to the high false alarm rate by the metal detector.

Fig. 1.3 (right): Deminer working with a prodder.
(Photo: Danish Demining Group).
Fig. 1.2 (below): Manual demining.
(Photo: Danish Demining Group).

The search for a replacement of the simple metal detector used in manual demining has turned out to be a much larger technological challenge than anticipated at first. This is revealed by the spectrum of technologies which have been put on test including ground penetrating radar (GPR), nuclear quadrupole resonance, infrared imaging (IR), ion mobility spectrometry, photoacoustic spectroscopy, thermal neutron analysis, reversal electron attachment detector, antibodies, artificial noses (Bio-mimics), and various methods based on chemical detection. The list of animals trained to detect mines includes dogs, rats and various insects, and the development of plants genetically modified to change colour by the induction of TNT or some of its degradation products has reach a stage where actual plants are being tested in controlled minefields. However, in spite of the efforts made by the research community, a technological breakthrough seems not to be impending, and the major part of mine clearance operations in the foreseeable future will therefore still hinge on manual demining.

Besides manual demining, two supplementary methods of increasing importance are *mine dog detection* and *mechanical mine clearance*. In mine dog detection, the detection tool *is* the dog due the dogs outstanding capacity to detect odours including explosives as TNT in very small concentrations. In contrast with metal detectors, dogs can detect mines with a

low metal content buried in soil characterized by a high metal content. Mine dogs function optimally in areas with a low mine density and are therefore typically used in the process of *area reduction*, i.e., the process through which an area initially suspected of being contaminated with mines is reduced to a smaller area. In areas characterized by a high mine density, mine dogs can get confused, and other factors such as fatigue or climatic conditions might affect the reliability of mine dogs.

Fig. 1.4. A deminer handling his dog in the Tete province, Mozambique (photo: GICHD))

Fig. 1.5. A Hydrema flail system used for mechanical mine clearance.

In mechanical mine clearance, machines like flails and tillers are used to detonate or destroy mines, typically tripwire-operated mines, or as vegetation cutters prior to manual mine clearance. The major advantage of mechanical mine clearance is obviously speed, but its usefulness as a clearance method depends on the terrain of the mine affected area. The quality of the clearance achieved by mechanical mine clearance has been questioned, and mechanical mine clearance is therefore rarely used alone but typically as an assisting tool to manual clearance.

It has been one of the essential lessons learned from a decade of ongoing mine action that collection of accurate and timely information about the scale, form and impact of a mine contamination problem is a prerequisite for a successful national mine action programme. Standardized *Landmine Impact Surveys* have been completed in a number of severely mine affected countries since 1999. The essential information provided by these surveys is the geographical distribution of *mine affected* communities. In this context a community is being referred to as mine affected if it contains one or several areas which are believed or

verified to contain mines. Also included in the surveys are accident statistics from the mine affected areas. Fig. 1.6 below illustrates the distribution of mine affected communities according to the landmine impact survey undertaken in Mozambique in the period 1999-2001. Table 1.1 contains the corresponding accident statistics where *recent victims* refers to the number casualties recorded two years prior to the survey.

Fig. 1.6. Mine affected communities in Mozambique. Reprinted from Canadian International Demining Corps et al., 2001.

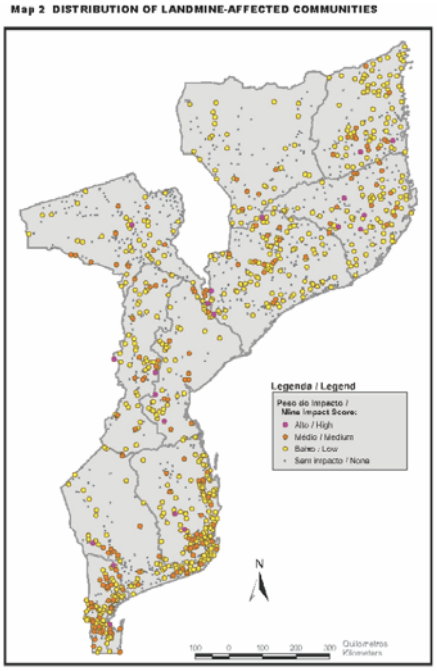


Table 1.1. Mine accident statistics from Mozambique. Source: Canadian International Demining Corps et al., 2001.

# of recent victims	# of communities
0	**710**
1	**45**
2	**11**
3	**13**
4	**2**
5	**3**
8	**1**
10	**1**
25	**1**
unknown	**4**
TOTAL	**791**

1.3 Impact and Prioritizations in Humanitarian Mine Action

The present lack of a fast and reliable mine detection technology means that the mine contamination problem found in many post-conflict countries cannot be eliminated overnight but has to be managed in several years to come. This entails that only a subset of the mine affected areas in a given country can be subject to mine clearance in the foreseeable future. To contain the mine contamination problem as effective as possible it is therefore essential that the national authorities are able to rank or prioritize the minefields according to the expected gain from a potential clearance operation.

Ignoring the emergency phase which may follow immediately after the ending of a war, the prioritization issue outlined above is in general a complicated matter. A contributory factor to this complexity is the multiple set of objectives which may influence the final prioritizations in a national coordinated mine action programme. For example, to reduce the direct dangers of explosive accidents will in most cases be a prominent objective in a mine clearance programme, but there are situations in which the relief of the indirect effects of mine contamination, i.e. the blockage of reconstruction and economic growth, are just as significant.

A second factor which complicates the prioritization process is the inability to measure the impact of mine clearance operations. In the early days of HMA the impact was simply considered to be proportional to the number of eliminated mines - or the size of the area cleared. Nowadays the situation is realized to be more complex. As a matter of fact, in the *GICHD* publication "*A Study of Socio-Economic Approaches to Mine Action*", the situation in HMA is summarized as follows: "*We remain unable to determine the impact of mine action in total, let alone estimate the decline in accidents due to the various components of mine action such as mine awareness or clearance*" [GICHD, 2001].

It goes without saying that the inability to measure the impact of HMA is a serious problem for at least two reasons: Firstly, it makes it difficult for decision makers to allocate resources into HMA optimally as the impact of a potential clearance task is unknown. Secondly, the lack of documentation could in the long run result in a reduced interest in HMA from national or international financial donors.

In spite of GICHD's rather pessimistic statement made above, certain attempts have been made to quantify the impact of mine contamination. The most prominent among the more quantitatively orientated models is the *mine impact score* model which has been implemented into the so-called IMSMA database [see GICHD, 2004, chapter 12]. The mine impact score is a weighted linear combination of 13 variables which includes the number of recent victims, certain livelihood and institutional blockage variables characterizing the mine affected community under study, and binary variables indicating whether mines or UXO (i.e., unexploded ordnance) have been present. Some weights are fixed, for example the weight associated with the number of recent victims, while others can be adjusted

within certain limits. The working hypothesis is that communities scoring high most likely are the ones in which mine action has the greatest potential for reducing future suffering [GICHD, 2001].

Possibly due to its implementation into the IMSMA data base, the mine impact score model has been used as a prioritization tool in the published landmine impact surveys which were mentioned in the previous paragraph. Figure 1.7 below illustrates the variables and the used weights in the report from the survey conducted in the Republic of Mozambique. According to the authors of the report [Canadian International Demining Corps et al., 2001] the used weights were chosen on the basis of "*the CIDC's experience, discussions with knowledgeable persons, and a review of the relevant literature*".

Figure 1.7. Reprinted from Canadian International Demining Corps et al. , 2001.

IMPACT SCORE VARIABLES

Variable	Weight
Group 1	
There were landmines	2[1]
There was UXO	1[2]
Group 2	
Access to some rainfed cropland was blocked	2
Access to some irrigated cropland was blocked	0
Access to some fixed pasture was blocked	2
Access to some migratory pasture was blocked	0
Access to some non-agricultural land was blocked	1[3]
Access to drinking water was blocked	2
Access to water for other uses was blocked	1[4]
Access to a housing area was blocked	0
One or more roads were blocked	1[5]
Access to some other infrastructure was blocked	1[6]
Group 3	
There were landmine victims in the last 24 months	2[7]

The mine impact score system permits a classification of the mine affected communities into three classes: "Low", "Medium" and "High". As an example, fig. 1.8 and fig. 1.9 on the following page show the distribution of mine impact scores and the final impact classification based on the landmine impact survey conducted in Yemen 1999-2000.

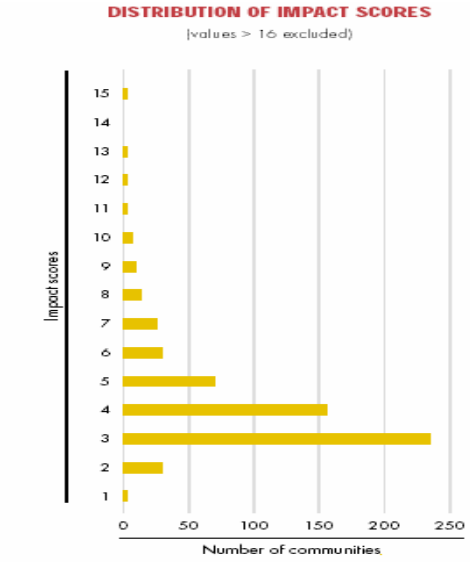
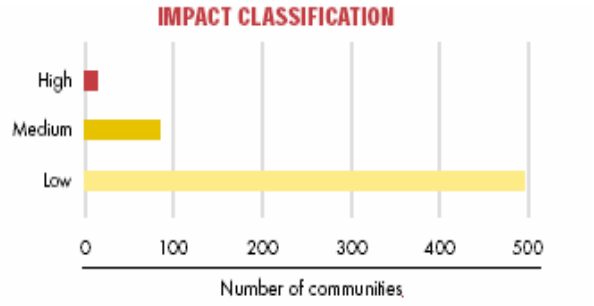


Figure 1.8 (left): Impact scores in Yemen. Reprinted from Survey Action Centre et al. , 2000.
Figure 1.9 (below): Impact classification in Yemen. Reprinted from Survey Action Centre et al. , 2000.



As confirmed by simulation runs performed by the *Survey Action Centre* who developed the model, the mine impact score is drawn to communities with comparatively many recent victims. On the contrary, communities with no record of recent mine victims will never be classified as "High" no matter how the weights of the blockage indicators are varied [Canadian International Demining Corps et al., 2001]. The number of recent victims is thus a variable attached central importance.

The mine impact score model is easy to comprehend and calculate, and it keeps information costs down. Through its blend of entering variables it takes into consideration the risk aspect of mine contamination (i.e., *Group 1* and *Group 3* variables in fig. 1.7) as well as its socio-economic impact (*Group 2* variables in fig. 1.7) even though the relative magnitudes of the attached weights appear arbitrary. The mine impact score model suffers however from a number of shortcomings which will be commented here. First of all, it is questionable, whether the number of recorded casualties is a reliable measure of the threat a given minefield poses to the surrounding society. That is, due to the stochastic nature of mine accidents, two identical minefields may display very different accident patterns even if the local population's degree of exposure to the minefields are identical.

Secondly, the high emphasis on the number of recent victims in the mine impact score model causes a problem as the majority of the mine affected communities show a record of very few or none reported victims (see for example table 1.1). Consequently, most of the

communities are classified as “Low” which makes the mine impact score less suited for long-term planning purposes.

Thirdly, the binary nature of the variable indicating whether mines have been present excludes the possibility of a more graduated estimate of the mine contamination.

Fourthly, the mine impact score model does not prescribe how to make a balanced updated risk assessment of the minefield if new information arrives.

Finally, the mine impact score model does not quantify the risk associated with a given minefield in such a way that comparisons to other sources of risk in the society can be made.

1.4 Research Objectives of Thesis

As remarked in the introduction to the present chapter, nobody seems yet to have examined the potential usefulness of the strong analytical tools provided by operations research and statistics to support the decision makers involved in HMA. Taking the observations made in connection with the impact score model into account, the aim of the present thesis is to analyze and give suggestions to how the situation in HMA, as to making qualified ranking of minefields, can be improved through the involvement of operations research or statistics.

In the previous paragraph it was noted that the mine impact score model considers the risk aspect and to a certain extent also the socio-economic impact of mine contamination. To simplify matters we will deal exclusively with the risk aspect of mine contamination. This limitation does not intend to downplay the importance of socio-economic considerations in relation to HMA. In other words, it is to be understood that any systematic risk assessment based on the approach outlined in the following chapters should be properly counterbalanced by some kind of socio-economic analysis before a final ranking of minefields can be made.

The word *risk* is used in many different contexts. Most expressions of risk are compound measures describing both the probabilities and severities of a set of damaging events.

Lowrance [Lowrance, 1976], for example, defines risk as a measure of the probability and severity of the consequences of undesirable events. Some risk measures attempt to describe the vulnerability of the society as a hole to a certain hazard, while other measures pay attention to particular groups or individuals. In the present context the most flexible measure of risk seems to be obtained if we define the risk associated with a given minefield as the *probability of mine accidents in the minefield within an observation period of predefined length.* Consequently, our primary objective is to derive a mathematical model from which the probability of mine accidents within an observation period can be calculated.

A mathematical model of the above kind should permit a ranking of an arbitrary number of minefields according to risk. However, to be useful within the framework of HMA it should additionally be flexible enough to accommodate the varied circumstances found in HMA with respect to accessible data. A second objective of the present work is therefore to provide methods which enable a decision maker to extract and transfer essential information from a variety of different sources into the mathematical model.

Finally, the shortcomings identified in the mine impact score model should be overcome by the introduction of the mathematical model.

As to the possible techniques from operations research or statistics which might be brought into play, the stated primary objective points in the direction of a *descriptive stochastic* mathematical model. That is, mine accidents are by nature stochastic events, and the frequency by which they happen might be envisaged as a function of some underlying variables describing the state of the minefield under study in a given observation period. As the state of the minefield may change over time, we are also looking for a *dynamical* model. Types of models which fit the above specifications include stochastic variables characterized by parametric probability distributions with time dependent parameters, and Markov processes.

What complicates HMA in particular is the lack of solid information. Most mine affected areas do in fact show a record of zero accidents. Whatever the choice of a descriptive stochastic dynamical model, the parameters which enter into such a model will be very

hard to estimate from the recorded accident statistics alone. Consequently, complementary information has to be added. In the case of HMA complementary information of potential relevance might be very diversified, and different levels of credibility might be attached to different pieces of information. A type of stochastic model which allows such diversified information to be added is a *Bayesian probability model* where previous information enters as *a priori* information.

Finally, one of the shortcomings identified in the case of the mine impact score model was its inability to make a balanced updated risk assessment of the minefield if new information arrives. A Bayesian type of model might show its relevance here too due to its ability to generate updated *posterior* distributions based on incoming observations.

1.5 "Road Map" to Thesis

To provide the reader with an overview of the contents of the present thesis, fig. 1.10 on page 16 includes a "road map" showing the interrelationships between the last 14 chapters of the thesis (excluding various appendices).

The key chapter in the thesis is chapter 2 where it is shown that a minefield accident under fairly general conditions can be considered to be the outcome of a binomial process. Consequently, the state of a minefield in a given observation period can be described by just two binomial parameters, i.e. the integer m and the probability parameter θ. The two binomial parameters will rarely be known in advance but have to be estimated.

Chapter 3 describes carefully the generation and the features of the simulated data to be used in the following chapters.

Depending on the character of the available information, the present report suggests two different ways of obtaining information about the probability parameter θ through the application of Bayesian data analysis. Thus given that accident statistics *and* mine clearance data are available, an estimate of θ in terms of a probability distribution can be generated by the use of a simple hierarchical Bayesian model as derived in chapter 4.

If only accidents statistics are available, the extraction of information about θ is made difficult. However, given that an estimate of the degree of mine contamination in an "average" minefield can be provided in terms of an informed prior distribution, it is possible to estimate θ through the application of so-called finite mixture models. This approach is discussed in the chapters 5-13. The applied techniques include Markov Chain Monte Carlo sampling and finite mixture models with a varying number of components.

A unified strategy for the synthesis of the various pieces of information is suggested in chapter 14 through the application of the reference prior approach.

Chapter 15 closes with a summary, conclusions, and suggestions for further work.

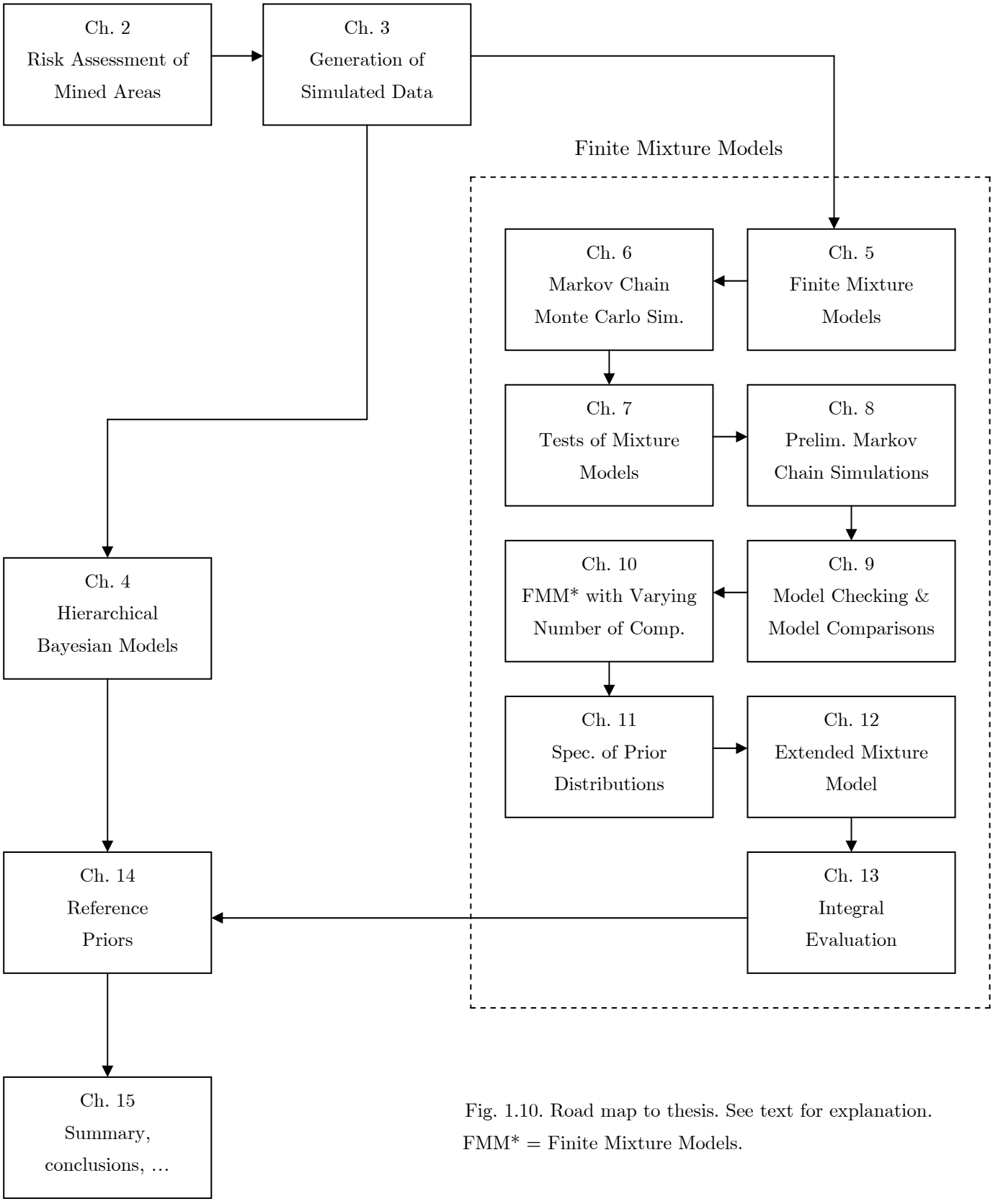


Fig. 1.10. Road map to thesis. See text for explanation. FMM* = Finite Mixture Models.

Chapter 2

Risk Assessments of Mined Areas – a Bayesian Approach in Mine Action

2.1 Introduction

To keep the discussion at a general level we will as our point of departure consider a hypothetical post-conflict region or country containing a large number of mine affected communities as sketched in fig. 2.1 and 2.2 below. In the present context a community is being referred to as mine affected if it contains one or several areas within the community border which are believed or verified to contain mines. Similarly, an area which is believed or verified to contain mines will be termed "a mine affected area". In what follows the word "minefield" and the concept "mine affected area" will be used interchangeably.

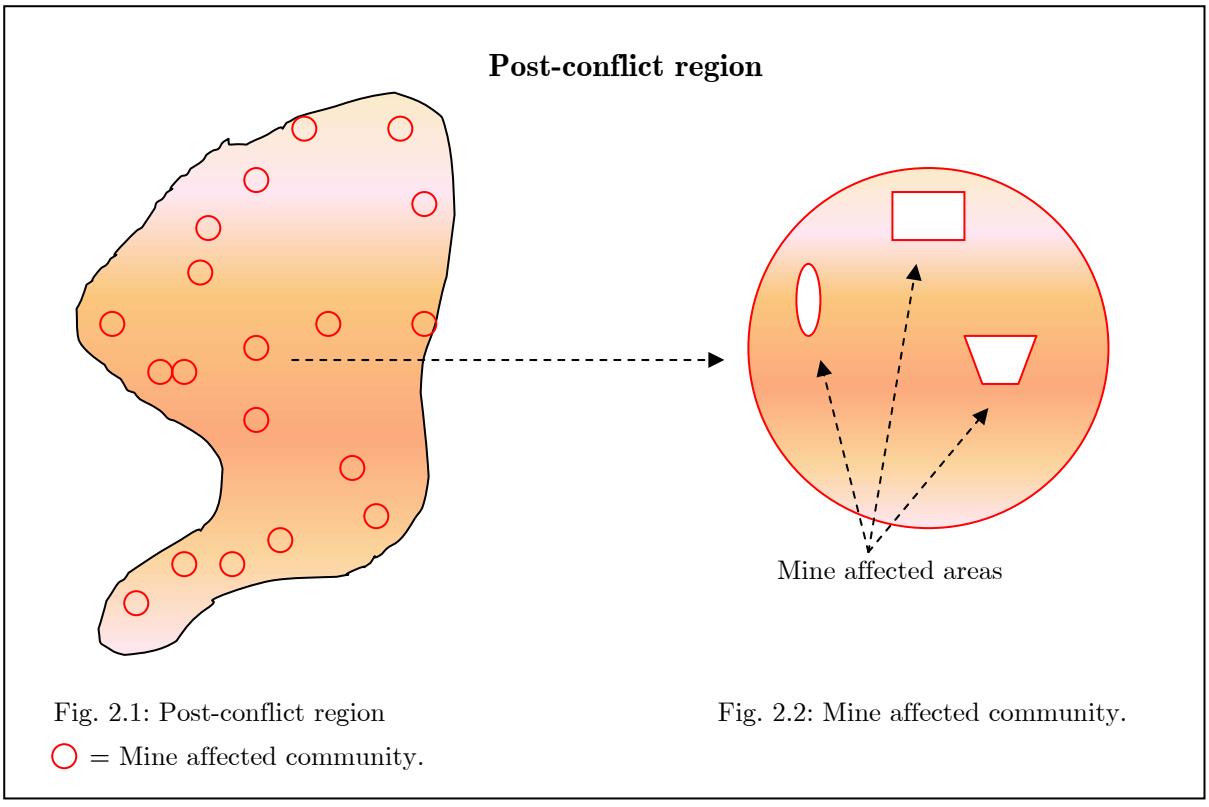


Fig. 2.1: Post-conflict region
○ = Mine affected community.

Fig. 2.2: Mine affected community.

Concerning the mine affected areas , we will make the following few assumptions:

- A mine affected area can contain an arbitrary number of mines (including zero) of various types and in various conditions.
- The mines present in a given area can be distributed in a random or non-random pattern, each mine being positioned either at the ground of the surface or buried to a certain depth.
- Information available to a decision maker about types and numbers of mines in a mine affected area may include detailed mine maps, assessments from regional or local experts, or no information at all.

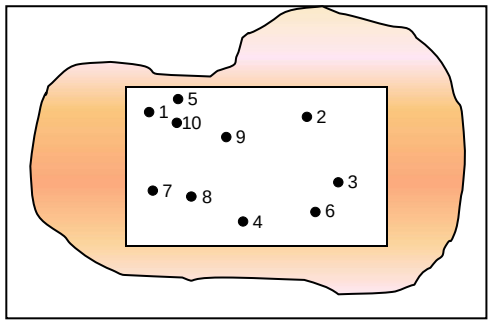
In the present chapter we will present a simple stochastic risk model designed for risk assessments of mine affected areas. The risk model will be derived in two steps: First, a general model which requires detailed information about the mined area in question will be derived. Secondly, by the introduction of two additional assumptions the general model turns into a simple 2-parameter binomial model. The true values of the binomial parameters which jointly characterize the state of the mined area will rarely be known in advance, but beliefs about these based on whatever information is available can conveniently be expressed in terms of probability distributions. This prepares the way for the introduction of Bayesian data analysis by which updates of the probability distributions can be generated from incoming accident statistics.

After having derived the risk model, illustrative examples showing how the ranking of mine affected areas can be accomplished through Bayesian data analysis will be given.

2.2 Derivation of General Risk Model

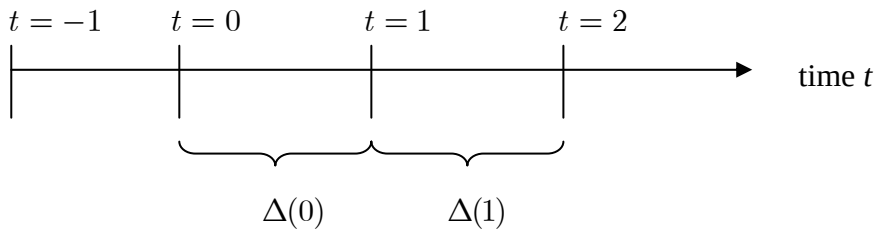
Consider some minefield which at time $t = 0$ contains m mines as sketched in fig. 2.3, where each mine has been assigned a number $k \in \{1,2,...,m\}$

Fig. 2.3. Minefield containing $m = 10$ mines.



As minefield accidents by nature are random events, the central quantity in a risk assessment of the above minefield is the probability distribution $p(z)$, where z denotes the number of accidents in the minefield during a future observation period of a certain length. In what follows, an observation which starts at time t and ends at time $t+1$ will be denoted $\Delta(t)$ as indicated in fig. 2.4. The time unit in fig. 2.4 is arbitrary, but as accident statistics in so-called Landmine Impact Surveys typically report the number of casualties observed during a two-year period, we will assume that $|\Delta(t)| = 2$ years for all t.

Fig. 2.4. Time axis



Now, let $Z_0 \in \{1,2,...,m\}$ denote the number of minefield accidents which might occur during $\Delta(0)$ in the minefield from fig. 2.3. To calculate $p(z_0)$ we will by way of introduction look at mine no. 1 from fig. 2.3. During $\Delta(0)$ mine no. 1 will either detonate or not. To record this event, let Z_0^1 denote the binary random variable which takes the value 1 if mine no. 1 is set off and 0 otherwise.

To calculate $p(z_0^1)$, that is, the probability of mine no. 1 being set off during $\Delta(0)$, it is valuable to consider the sequence of events which is a prerequisite for a detonation: Firstly, during $\Delta(0)$ there has to be a "contact" between mine no. 1 and a person, a vehicle, etc. Secondly, to detonate during the "contact", mine no. 1 has to be exposed to a pressure which is equal to or exceeds a certain threshold value.

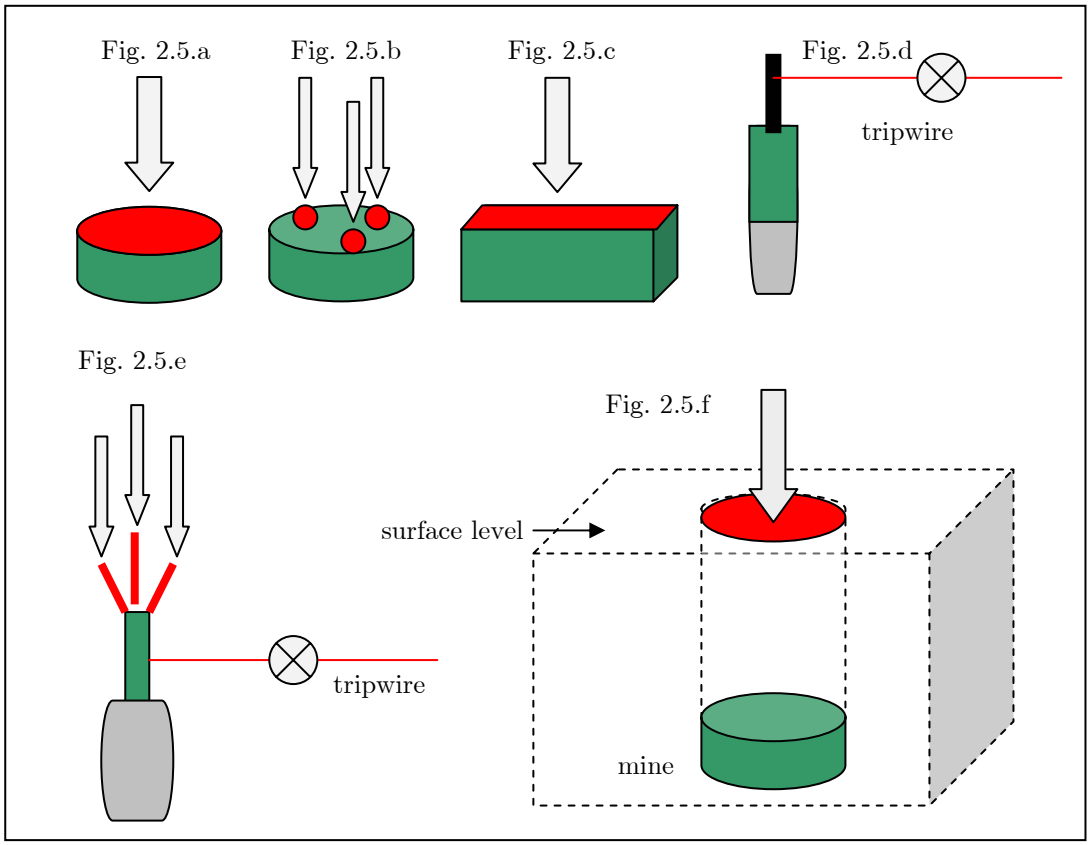
The very simplified account given above covers up certain difficulties. First of all, the notion a "contact" is ill-defined, as the triggering of a mine not necessarily implies a physical contact between the mine and say a person. Secondly, to set off a mine the triggering pressure has to be exerted at the right part of the mine or at the right part of the ground above a buried mine.

To overcome the above difficulties and to keep our model considerations simple, we will assume that every mine can be characterized by an individual *contact zone*, that is, a surface in 3D-space with the following properties:

1) To set off the mine, a pressure equal to or exceeding a certain *threshold pressure* (*TP*) has to be exerted within the boundary of the contact zone.
2) The threshold pressure is constant over the contact zone.

Examples of contact zones for different types of mines are sketched in fig. 2.5 below. Depending on whether the mine is located on the surface of the ground or buried, the contact zone may or may not coincide with parts of the casing of the mine.

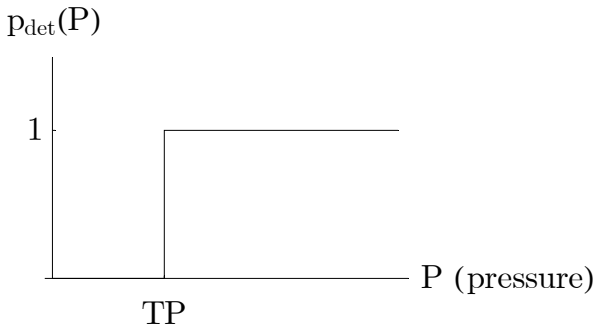
Fig. 2.5 Contact zones of mines. The red coloured areas denote the contact zones of mines of various designs.



The introduction of contact zones allows us to clarify the "contact" concept: Whenever a person, a vehicle, etc., touches the contact zone of a mine, we will refer to the event as a "contact".

The idealized model of a uniform threshold pressure can be sketched as in fig. 2.6 below.

Fig. 2.6. Probability of detonation. $p_{\text{det}}(\text{P})$ denotes the probability of detonation given a pressure P is exerted on the contact zone of a mine. The value "TP" denotes the threshold pressure of the mine.



It should be noted that not all mines fit into the idealized model sketched in fig. 2.6. We will however ignore cases such as the PFM-1 anti-personnel mine which can be triggered by the accumulated effect of successive contacts due to its pressure fuzed liquid explosive.

The magnitude of the threshold pressure of a mine will in general depend on factors such as

- type of mine (AP-mine, AT-mine)
- fuzing mechanism
- condition of mine (ageing, corrosion)
- vertical position of mine.

Whether the threshold pressure of a mine is reached during a random contact will in general depend on the kind of activity during the contact (walking, driving, ...). In addition, for a given type of activity the pressure exerted on a mine will presumably vary from contact to contact due to its stochastic nature. To incorporate this variability into our model we will assign the minefield from fig. 2.3 a probability distribution $p(CP)$ which denotes the probability of observing a *contact pressure* of magnitude CP during a contact with a randomly selected mine. The contact pressure is here defined as the *maximum* pressure exerted on a randomly selected mine during a contact.

It follows from the considerations above that mine no. 1 subsequent a random contact only will detonate with a certain probability ϕ_1 which can be calculated as

$$\phi_1 = \int_{TP_1}^{\infty} p(CP)\, dCP, \tag{2.01}$$

where TP_1 in equation (2.01) denotes the threshold pressure of mine no. 1. The parameter ϕ_1 will be denoted the *conditioned probability of detonation* of mine no. 1.

After having introduced these facilitating concepts, a closed expression for $p(z_0^1)$ can be obtained in the following way: Let X_1 denote the random variable which counts the number of times the contact zone of mine no. 1 is struck during the period $\Delta(0)$. The probability of mine no. 1 not being set off can be written as

$$\begin{aligned} p(Z_0^1 = 0) = {} & p(X_1 = 0) + \\ & p(X_1 = 1)(1-\phi_1) + \\ & p(X_1 = 2)(1-\phi_1)^2 + \ldots \\ = {} & \sum_{i=0}^{\infty} p(X_1 = i)(1-\phi_1).^i \end{aligned} \tag{2.02}$$

If X_1 follows a Poisson distribution with intensity λ_1, that is

$$p(x_1) = e^{-\lambda_1} \frac{{\lambda_1}^{x_1}}{x_1\,!}, \tag{2.03}$$

where $E[X_1] = \lambda_1$, it follows that

$$\begin{aligned} p(Z_0^1 = 0) &= \sum_{i=0}^{\infty} e^{-\lambda_1} \frac{{\lambda_1}^i}{i!} (1-\phi_1)^i \\ &= e^{-\lambda_1 \phi_1}. \end{aligned} \tag{2.04}$$

Consequently $p(z_0^1)$ takes the form

$$p(z_0^1) = \begin{cases} 1 - e^{-\lambda_1 \phi_1} \text{ if } z_0^1 = 1 \\ e^{-\lambda_1 \phi_1} \text{ if } z_0^1 = 0. \end{cases} \tag{2.05}$$

If the stochastic variables $Z_0^1, Z_0^2, .., Z_0^m$ furthermore are independent, it follows that the distribution of $Z_0 = \sum_{k=1}^{m} Z_0^k$ can be calculated as

$$p(z_0) = \sum \coprod_{k=1}^{m} p\,(z_0^k), \tag{2.06}$$

where $p(z_0^k)$ is given as

$$p(z_0^k) = \begin{cases} 1 - e^{-\lambda_k \phi_k} \text{ if } z_0^k = 1 \\ e^{-\lambda_k \phi_k} \text{ if } z_0^k = 0, \end{cases} \quad \textbf{(2.07)}$$

and the sum denoted by Σ in equation (2.06) includes all vectors $(z_0^1, z_0^2, ..., z_0^m)$ for which $z_0^1 + z_0^2 + ...z_0^m = z_0$. In spite of the simple structure of equation (2.07) the model embedded in this equation reflects the combined action of several factors, that is,

- the types, conditions and vertical locations of the mines present (reflected by TP_k)
- the activities taking place in the mined area (reflected through $p(CP)$)
- the intensities of the activities taken place in the mined area (reflected by λ_k).

The utility of the model may be questioned as neither m nor the true values of the parameters $\{\{\phi_k, \lambda_k\}\}$ will be known in the general case. We might however have some, albeit incomplete information at hand which makes it possible to make a qualified guess at their true values by means of probability distributions $p(m)$, $p(\phi)$ and $p(\lambda)$. From these distributions $p(z_0)$ can be calculated numerically.

In the present chapter we will follow a slightly different course. That is, by introducing two additional assumptions the stochastic variable Z_0 from (2.06) can be turned into a binomially distributed variable. Apart from its simple analytical structure the binomial model demands as input only two parameters to calculate $p(z_0)$.

Table 2.1. Applied notation in minefield model.

Factor	**Represents**	**Factor**	**Represents**
t	time	TP_k	Threshold pressure of mine no. k.
$\Delta(t)$	Observation period [t ; t+1]	CP	Contact Pressure
m	Number of mines	$p(CP)$	Probability of CP during contact.
Z_t	Number of accidents in $\Delta(t)$	ϕ_k	The probability of detonation of mine no. k given a random contact.
$p(z_t)$	Probability of observing z_t accidents in $\Delta(t)$.	X_k	Number of random contacts with mine no. k during $\Delta(t)$
Z_t^k	0-1 variable. Indicates whether mine no. k has been set off in $\Delta(t)$.	λ_k	The expected value of X_k

2.3 Derivation of a Binomial Model

2.3.1 Homogeneous minefields

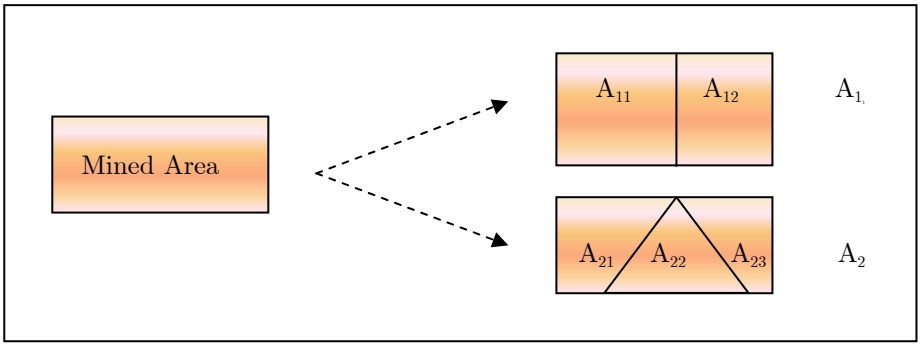
The presence of mines is obviously a prerequisite for mine accidents, but the intensity of the activity taking place in a mined area may have a profound effect on the probability of mine accidents as well. If several activities of different intensities are going on in a given area, the making of a risk assessment becomes complex.

To sketch how a thorough risk assessment may be structured in a complex environment, assume that a number of activities A_1, A_2 …, A_N which might cause the triggering of a mine takes place in a mined area. With respect to activity A_i, we will assume that the mined area in question can be split up into homogeneous sub-areas A_{i1}, A_{i2},…,$A_{iK(i)}$ within which the intensity of activity A_i may be taken as uniform. A mined area characterized by an activity of uniform intensity will be termed a *homogeneous minefield*, and we will assign all contact zones within the borders of a homogenous minefield the same Poisson parameter whatever the number of mines present. For a homogeneous minefield A_{ij} we thus have that

$$p_{ij}(z_0^k) = \begin{cases} 1 - e^{-\lambda_{ij}\phi_{ik}} & \text{if } z_0^k = 1 \\ e^{-\lambda_{ij}\phi_{ik}} & \text{if } z_0^k = 0, \end{cases} \qquad \textbf{(2.08)}$$

where $k \in \{1,2,...,m\}$, and m is the number of mines present in sub-area A_{ij}. The probability distribution of the contact pressure CP in minefield A_{ij} may similarly be denoted $p_{ij}(CP)$. Fig. 2.7 illustrates the partitioning of a mined area into homogeneous minefields for two activities A_1 and A_2. As sketched in fig. 2.7, the partitioning may depend on the activity considered.

Fig. 2.7: Partitioning of mined area into homogeneous minefields for two different activities A_1 and A_2.



From the considerations above it follows that a homogeneous minefield plays a pivotal role. In other words, if $p_{ij}(z_t)$ can be calculated for an arbitrary homogeneous minefield, the probability of accidents in any minefield can be determined by combining the probability distributions $p_{ij}(z_t)$ from the underlying homogeneous minefields. In the remaining paragraphs we will focus exclusively on the determination of $p(z_t)$ for a homogeneous minefield characterized by a single activity through $p(CP)$.

2.3.2 Functional Mines

To simplify equation (2.08) we will look into the variation among the values taken by the parameters $\{\phi_1, \phi_2, ..., \phi_m\}$ in a homogeneous minefield which according to equation (2.01) will be a function of the frequency of threshold pressures and the probability distribution $p(CP)$. It turns out that ϕ_k in many cases will take a value of either zero or one.

Let us, to keep things simple, assume that $p(CP)$ can be represented by a normal distribution in a homogeneous minefield characterized by a single activity. If $CP \sim N(\mu, \sigma)$ it follows from equation (2.01) that

$$\phi_k = \int_{TP_k}^{\infty} N(CP \mid \mu, \sigma)\, dCP, \qquad \textbf{(2.09)}$$

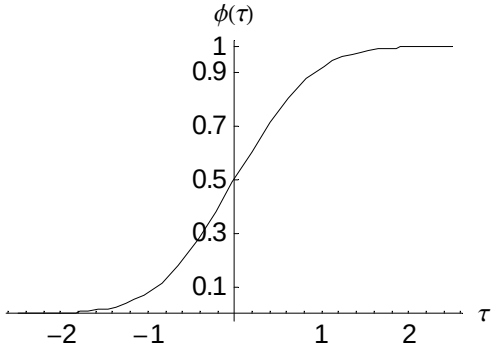
which can be expressed as

$$\phi_k = \phi(\tau_k) = 1 - \tfrac{1}{2} erfc(\tau_k), \qquad \textbf{(2.10)}$$

where *erfc* denotes the complementary error function, and

$$\tau_k = \frac{\mu - TP_k}{\sqrt{2}\sigma}. \qquad \textbf{(2.11)}$$

Fig. 2.8 $\phi(\tau) = 1 - \tfrac{1}{2} erfc(\tau)$.



It is evident from equation (2.11) and fig. 2.8 that if $|\tau_k| \gg 1$ for all k in a minefield, the individual ϕ_k takes either a value of zero or one. This simplifies equation (2.08) considerably and turns eventually Z_0 from equation (2.06) into a binomially distributed variable.

To prove the above assertion, consider an arbitrary minefield which at t=0 contains m mines characterized by the threshold pressures $TP_1, TP_2, ..., TP_m$.

From the minefield above we construct a sequence of related minefields labelled $n = 1,2,...$, all characterized by the same set of threshold pressures as above and with $p_n(CP) \sim N(\mu, \frac{\sigma}{n})$. It follows that $\tau_{k,n}$ for minefield n is given by

$$\tau_{k,n} = \frac{\mu - TP_k}{\sqrt{2}(\frac{\sigma}{n})}, \tag{2.12}$$

where $k \in \{1,2,...,m\}$.

If $Z^k_{0,n}$ denotes the 0-1 variable which records whether mine no. k in minefield n is set off during $\Delta(0)$, we have that $p(z^k_{0,n})$ for $k \in \{1,2,...,m\}$ is given by

$$p\,(z^k_{0,n}) = \begin{cases} 1 - e^{-\lambda\phi(\tau_{k,n})} \text{ if } z^k_0 = 1 \\ e^{-\lambda\phi(\tau_{k,n})} \text{ if } z^k_0 = 0 \end{cases}, \tag{2.13}$$

and it follows that the *generating function* $P_n(s)$ of $Z_{0,n}$ for minefield n can be written

$$\begin{aligned} P_n(s) &= \prod_{k=1}^{m} (e^{-\lambda\phi(\tau_{k,n})} + s(1 - e^{-\lambda\phi(\tau_{k,n})})) \\ &= \prod_{k=1}^{\tilde{m}} (e^{-\lambda\phi(\tau_{k,n})} + s(1 - e^{-\lambda\phi(\tau_{k,n})})) \times \\ &\quad \prod_{k=\tilde{m}+1}^{m} (e^{-\lambda\phi(\tau_{k,n})} + s(1 - e^{-\lambda\phi(\tau_{k,n})})), \end{aligned} \tag{2.14}$$

where $\tilde{m}$ in (2.14) denotes the number of mines satisfying $TP_k \leq \mu$. For later convenience we will refer to $\tilde{m}$ as the number of *functional mines*. For $n \to \infty$ we have that $\phi(\tau_{k,n}) \to 1$ if $TP_k < \mu$, and $\phi(\tau_{k,n}) \to 0$ if $TP_k > \mu$. If we exclude the possibility that any of the TP_k's are identical to μ, it follows that

$$\lim_{n\to\infty} P_n(s) \tag{2.15}$$
$$= \prod_{k=1}^{\tilde{m}} (e^{-\lambda} + s(1-e^{-\lambda}))$$
$$= (1 + (s-1)(1-e^{-\lambda}))^{\tilde{m}},$$

where the last line in equation (2.15) is recognized as the generating function for a binomially distributed variable with parameters $\tilde{m}$ and $(1-e^{-\lambda})$. If Z^{bin} is defined as

$$Z^{bin} \sim Bi(\tilde{m}, 1-e^{-\lambda}), \tag{2.16}$$

it follows that $Z_{0,n}$ for $n \to \infty$ converges *in law* to Z^{bin}.

To illustrate the practical significance of the above result, consider table 2.2 below which tabulates the distribution of threshold pressures for a hypothetical minefield containing 10 mines including 6 anti-personnel mines and 4 anti-tank mines. Due to the large difference between the *TP* of a typical anti-personnel mine and an anti-tank mine the *TP*'s in table 2.2 fall into two well separated groups.

Table 2.2. Distribution of threshold pressures (*TP*) for hypothetical minefield containing 10 mines.

mine #	*TP* (kPa)	
1	6	AP-mines
2	8	
3	8	
4	10	
5	11	
6	13	
7	120	AT-mines
8	250	
9	260	
10	280	

In fig. 2.9 below, four different normal distributions each representing a possible choice of $p(CP)$ have been superimposed on a histogram showing the distribution of threshold pressures from table 2.2. For each normal distribution the corresponding values of ϕ_k for the 10 mines are tabulated in table 2.3.

Fig. 2.9: Frequency distribution of threshold pressures and normal distributed *CP*'s. Vertical bars represent threshold pressures from table 2. Black solid curve: *CP* ~ N(60,10); black dashed curve: *CP* ~ N(60,25); blue solid curve: *CP* ~ N(150,10); blue dashed curve: *CP* ~ N(150,25). P(kPa) denotes threshold pressure.

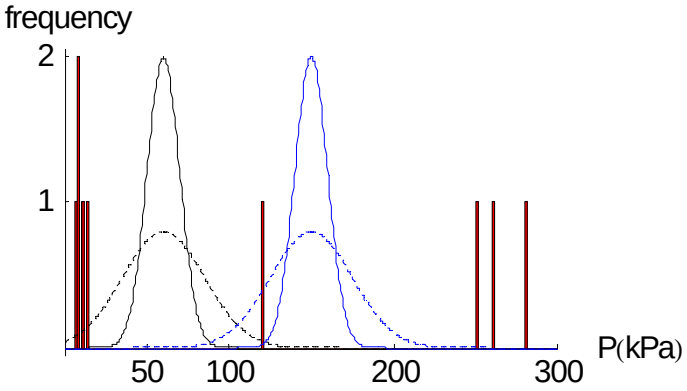


Table 2.3. Calculated values of ϕ for the minefield tabulated in table 2.2 when $CP \sim N(\mu,\sigma)$. $\lambda = 0.1$.

μ	**60**		**150**	
σ	10	25	10	25
ϕ_1	1.000	0.985	1.000	1.000
ϕ_2	1.000	0.981	1.000	1.000
ϕ_3	1.000	0.981	1.000	1.000
ϕ_4	1.000	0.977	1.000	1.000
ϕ_5	1.000	0.975	1.000	1.000
ϕ_6	1.000	0.970	1.000	1.000
ϕ_7	0.000	0.008	0.999	0.885
ϕ_8	0.000	0.000	0.000	0.000
ϕ_9	0.000	0.000	0.000	0.000
ϕ_{10}	0.000	0.000	0.000	0.000

As it emerges from table 2.3, the parameters $\phi_1 \rightarrow \phi_6$ take in the case μ=60 a value of approximately one which corresponds to all mines with a $TP_k \leq \mu$ (se table 2.2). The remaining mines take a value of approximately zero. Similarly, when μ=150, $\phi_1 \rightarrow \phi_7$ take a value of approximately one, and the remaining mines a value of approximately zero. For the four cases tabulated in table 2.3 we may thus infer that $p(z_t)$ approximately follows a binomial distribution, that is, $Z_t \sim Bi(6,1-e^{-\lambda})$ when $\mu = 60$, and $Z_t \sim Bi(7,1-e^{-\lambda})$ when $\mu = 150$. Table 2.4 shows the expected value of Z_t for the four cases above calculated from the general expression given by equation (2.08), and from the expression $\tilde{m}(1-e^{-\lambda})$, respectively. The deviation between the two models is marginal.

Table 2.4: The expected number of accidents. $E[Z_t]$ is calculated from (2.08), $\lambda = 0.1$. Percentage deviation refers to deviation between $E[Z_t]$ and $\tilde{m}(1-e^{-\lambda})$.

μ	**60**		**150**	
σ	10	25	10	25
$E[Z_t]$	0.571	0.560	0.666	0.656
$\tilde{m}(1-e^{-\lambda})$	0.571	0.571	0.667	0.667
Percentage deviation	0.00	1.97	0.02	1.60

The error induced by the use of the binomial model will in the general case depend on the detailed distribution of threshold pressures and the location and spread of $p(CP)$. To provide an upper bound to this error, consider a homogenous minefield which at time t contains m mines characterized by the set $\{\tau_1,\tau_2,...,\tau_m\}$. Let $E[Z_t]$ denote the expected value of Z_t (calculated from (2.08)), and let $E[Z^{bin}] = \tilde{m}(1-e^{-\lambda})$, where Z^{bin} is given by (2.16). It can be shown that

$$| E[Z_t] - E[Z^{bin}] | \leq \tilde{m}\left(e^{-\lambda\left(1-\frac{erfc(\tau^+_{\min})}{2}\right)} - e^{-\lambda}\right) + (m-\tilde{m})\left(1-e^{-\lambda\left(1-\frac{erfc(\tau_{\max})}{2}\right)}\right) \quad \textbf{(2.17)}$$

where

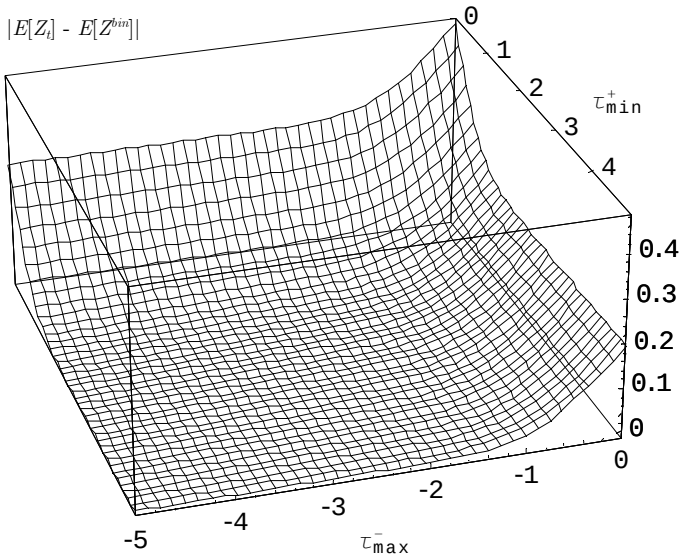
$$\tau^+_{\min} = \min\{\tau_1,\tau_2,...,\tau_m \mid \tau_k \geq 0\}, \quad \textbf{(2.18)}$$

and

$$\tau^{-}_{\max} = \max\{\tau_1, \tau_2, ..., \tau_m \mid \tau_k < 0\}. \qquad \textbf{(2.19)}$$

The deviation of $E[Z^{bin}]$ from $E[Z_t]$ will according to equation (2.17) depend on both $\tilde{m}$ and m, but as $(\tau^{-}_{\max}, \tau^{+}_{\min}) \to (-\infty, \infty)$, the deviation goes inevitably to zero. This is illustrated in fig. 2.10 where the deviation $|\, E[Z_t] - E[Z^{bin}]\,|$ is shown as a function of $\tau^{+}_{\min}$ and $\tau^{-}_{\max}$ for $m = 10$ and $\tilde{m} = 6$.

Fig.2.10. $|\, E[Z_t] - E[Z^{bin}]\,|$ as a function of $\tau^{+}_{\min}$ and $\tau^{-}_{\max}$. $m = 10, \tilde{m} = 6, \lambda = 0.1$.



2.4 Bayesian Data Analysis

The calculation of $p(z_t)$ for a homogeneous minefield demands in general a detailed knowledge of the distribution of threshold pressures and a knowledge of the location and spread of $p(CP)$. However, if information is at hand which allows us to conclude that $|\, \tau_k \,| \gg 1$ for all mines present, it follows from the previous paragraph that

$$Z_t \sim Bi(\tilde{m}, 1 - e^{-\lambda}), \qquad \textbf{(2.20)}$$

or simply

$$Z_t \sim Bi(\tilde{m}, \theta), \qquad \textbf{(2.21)}$$

where $\theta = 1 - e^{-\lambda} \in]0;1[$. Consequently, a binomial distribution will under these circumstances give a satisfactory description of the probability of minefield accidents.

Unfortunately, we do not in general know the true values of either $\tilde{m}$ or θ. We might however have some information at hand which makes it possible to make a qualified guess at their true values. A convenient way to quantify our belief about $\tilde{m}$ or θ is in terms of a probability distribution. Such a probability distribution will necessarily be time-dependent and should be regularly updated by taking the number of accidents observed during future observation periods into consideration.

Updating of probability distributions can be carried out in a convenient way by *Bayes' theorem*. To recast our risk assessment problem into a form which makes it suitable to Bayesian data analysis, let $\pi_t(\tilde{m})$ denote our prior distribution as to the number of functional mines present at time t in the minefield under study. The probability distribution $p(z_t)$ can be written as

$$p(z_t) = \sum_{\tilde{m}=0} p(z_t \mid \tilde{m})\, \pi_t(\tilde{m}) \tag{2.22}$$

where

$$p(z_t \mid \tilde{m} = 0) = \begin{cases} 1 \text{ if } z_t = 0 \\ 0 \text{ else} \end{cases} \tag{2.23}$$

and

$$p(z_t \mid \tilde{m} \geq 1) = \begin{cases} \int p(z_t \mid \tilde{m}, \theta) \pi_t(\theta \mid \tilde{m}) d\theta \text{ if } \tilde{m} \geq \max(1, z_t) \\ 0 \text{ else.} \end{cases} \tag{2.24}$$

The term $\pi_t(\theta \mid \tilde{m})$ in (2.24) denotes our prior distribution of θ conditioned on $\tilde{m}$ covering the period $\Delta(t)$. The inclusion of the term $\tilde{m} = 0$ in the summation in (2.22) simply means that we do not exclude the possibility that the minefield under study actually contains zero functional mines.

What is needed to calculate $p(z_t)$ is consequently the prior distributions

$$\pi_t(\tilde{m}) = \{\pi_t(0), \pi_t(1), \ldots\} \tag{2.25}$$

and

$$\pi_t(\theta \mid \tilde{m}) \text{ for } \tilde{m} \geq 1. \tag{2.26}$$

For $\tilde{m} \geq 1$ we may write $\pi_t(\tilde{m})$ and $\pi_t(\theta \mid \tilde{m})$ collectively as the prior joint distribution

$$\pi_t(\tilde{m}, \theta) = \pi_t(\theta \mid \tilde{m})\, \pi_t(\tilde{m}) \tag{2.27}$$

From $p(z_t)$ in (2.22) we may calculate whatever property of interest and subsequently make a risk assessment of the minefield covering the period $\Delta(t)$.

Assume now that the minefield under study is not selected for mine clearance, and a period $\Delta(t)$ passes away during which z_t minefield accidents are observed. According to Bayes' theorem, the posterior distribution $\pi_t(\tilde{m} \mid z_t)$ for $\tilde{m} = 0$ is given as

$$\pi_t(\tilde{m} = 0 \mid z_t) \propto p(z_t \mid \tilde{m} = 0)\pi_t(\tilde{m} = 0). \tag{2.28}$$

In the case $\tilde{m} \geq 1$ the posterior distribution $\pi_t(\tilde{m}, \theta \mid z_t)$ can be calculated as

$$\pi_t(\tilde{m}, \theta \mid z_t) \propto p(z_t \mid \tilde{m}, \theta)\pi_t(\theta \mid \tilde{m})\pi_t(\tilde{m}). \tag{2.29}$$

From $\pi_t(\tilde{m}, \theta \mid z_t)$ in (2.29) the posterior marginal distribution

$$\pi_t(\tilde{m} \mid z_t) = \{\pi_t(0 \mid z_t), \pi_t(1 \mid z_t), \pi_t(2 \mid z_t)...\}, \tag{2.30}$$

and the posterior conditional distribution

$$\pi_t(\theta \mid \tilde{m}, z_t) \text{ for } \tilde{m} \geq 1 \tag{2.31}$$

can be derived. The link between (2.30) and (2.31) and the corresponding distributions valid at t=1 is given by the relations

$$\pi_{t+\Delta(t)}(\tilde{m}) = \pi_t(\tilde{m} + z_t \mid z_t) \tag{2.32}$$

and

$$\pi_{t+\Delta(t)}(\theta \mid \tilde{m}, z_t) = \pi_t(\theta \mid \tilde{m} + z_t, z_t) \tag{2.33}$$

By use of the updates (2.32) and (2.33) we can make an updated risk assessment covering the period $[t+\Delta(t);t+2\Delta(t)]$ by the calculation of

$$p(z_{t+\Delta(t)}) = \sum_{\tilde{m}=0} p(z_{t+\Delta(t)} \mid \tilde{m})\pi_{t+\Delta(t)}(\tilde{m} \mid z_t) \qquad \textbf{(2.34)}$$

The method outlined above is of course only valid if the conditions determining θ are identical in two successive observation periods. If essential conditions have changed (except the number of mines present), new conditional distributions of θ based on the available information have to be set up.

In the following paragraphs illustrative examples of the application of (2.34) will be given.

2.5. Application of Bayesian Data Analysis: Example 1

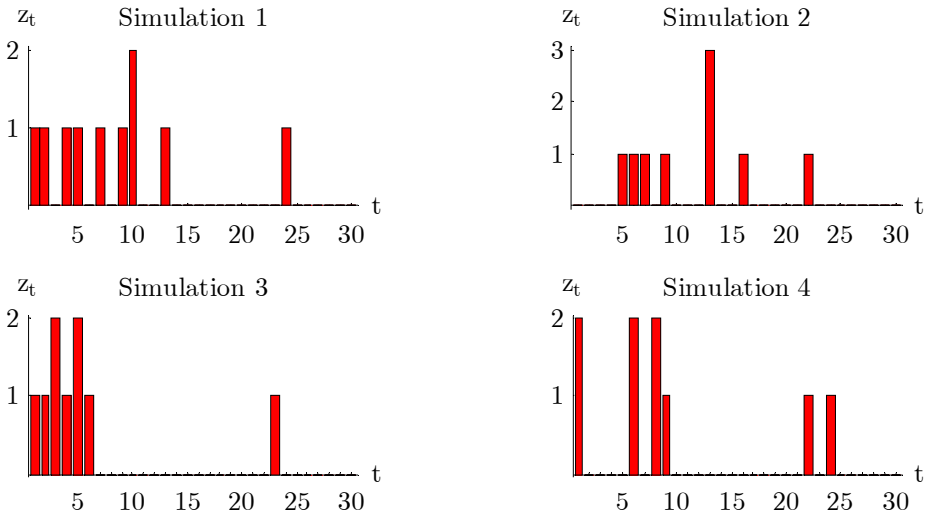
To test the utility of the Bayesian approach outlined above we will illustrate the mode of operation of (2.34) by a hypothetical example covering several observation periods. The example may serve two purposes: 1) support the view that reliable risk assessments of minefields in general have to be based on careful probability calculations; 2) illustrate that (2.34) offers an approach to risk assessment which has the potential of generating reliable estimates.

Now, consider a hypothetical minefield containing 10 functional mines at $t=0$ and characterized by $\theta = 0.1$. Consequently, $Z_0 \sim Bi(10, 0.1)$. More generally we have that $Z_t \sim Bi(\tilde{m}_t, 0.1)$ for all $t \geq 0$ where $\tilde{m}_t$ denotes the number of functional mines left at time t. Due to the stochastic nature of Z_t the accident pattern observed during the coming observations periods might show very different forms. This is illustrated in fig. 2.11 (on the following page) which displays the accident pattern obtained from four simulations covering 30 successive observation periods starting at $t=0$. In each observation period z_t was determined by sampling from a binomial distribution $Bi(\tilde{m}_t, 0.1)$.

A hypothetical observer who has access to the recorded number of casualties within the first few observation periods from one of the simulations in fig. 2.11, and who is ignorant about the true content of mines in the minefield under study, will have great difficulties in making any kind of reliable risk assessment of the minefield. That is, simply counting the

number of minefield accidents within say the first four observation periods does not reveal much about what to be expected in the future. To interpret the recorded observations in a balanced way the observer needs complementary information.

Fig. 2.11. Simulation of accident pattern from hypothetical minefield during 30 successive observation periods. The minefield contains 10 functional mines at $t = 0$, and $\theta = 0.1$. The number of accidents recorded within the first four observations periods goes from zero accidents (simulation 2) to 5 (simulation 3).

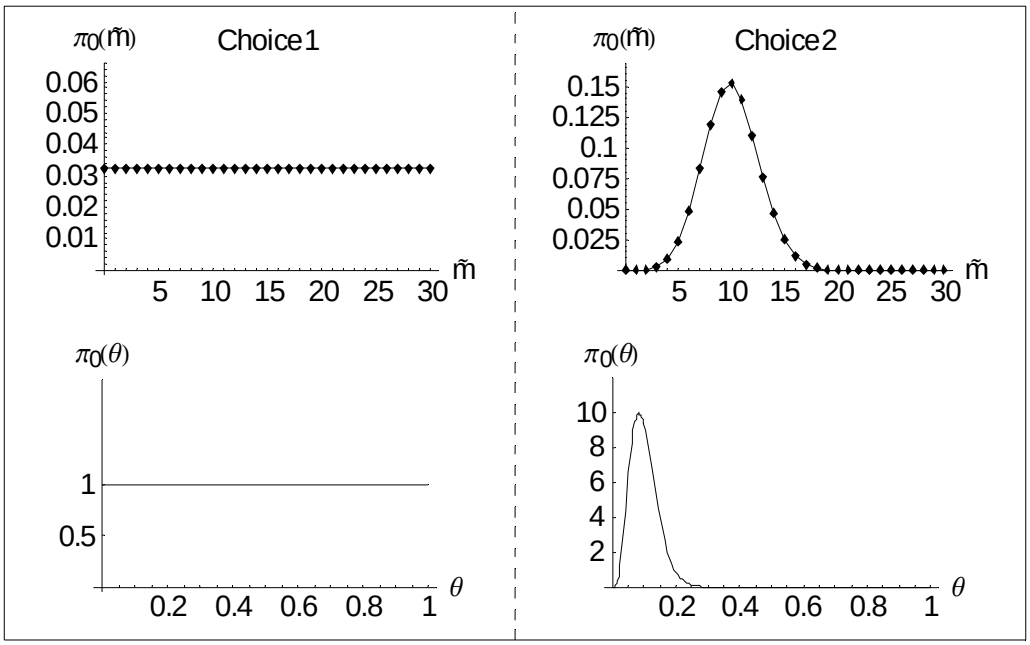


Assume now that our hypothetical observer wishes to interpret the accident pattern from simulation 1 through Bayesian data analysis as outlined in the previous paragraph. More specifically, he wants to make statistical inferences about the true values of $\tilde{m}$ and θ at time t by means of the accident pattern $\{z_0, z_1, ..., z_{t-1}\}$ and Bayesian updating. As to the observer's choice of prior distributions $\pi_0(\tilde{m})$ and $\pi_0(\theta \mid \tilde{m})$, let us consider the two options tabulated in table 2.5 below (and illustrated in fig. 2.12 on the following page). In both cases the observer assumes that $\tilde{m} \leq 30$ at $t = 0$, and $\pi_0(\theta \mid \tilde{m})$ is assumed independent of $\tilde{m}$, i.e., $\pi_0(\theta \mid \tilde{m}) = \pi_0(\theta)$.

Table 2.5. The observer's two sets of prior distributions.

Prior distributions	Choice 1	Choice 2
$\pi_0(\tilde{m})$	$\tilde{m} \sim UD(30)$	$\tilde{m} \sim Bi(30, \frac{1}{3})$
$E[\tilde{m}]$	15	10
$\pi_0(\theta)$	$\theta \sim U(0,1)$	$\theta \sim Be(5,45)$
$E[\theta]$	0.5	0.1

Fig. 2.12. The observer's two sets of prior distributions. See table 2.5 for technical details.



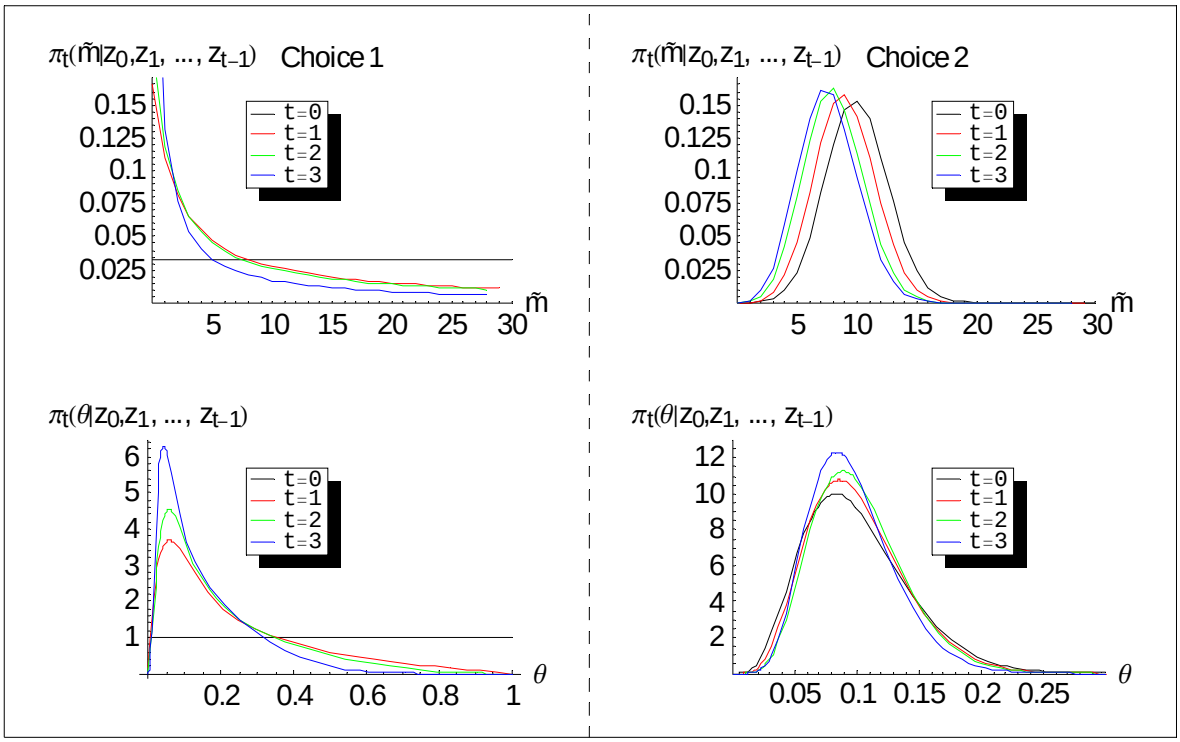
Choice 1 makes up what might be termed a non-informative set of priors. That is, apart from the restriction $\tilde{m} \leq 30$ the prior $\pi_0(\tilde{m})$ assigns equal possibility to all values of $\tilde{m}$. A similar observation goes with $\pi_0(\theta)$. In the case of *Choice 2*, the expected values of $\tilde{m}$ and θ do in fact coincide with the true values of $\tilde{m}$ and θ in the minefield at $t = 0$, but a degree of uncertainty is reflected through the depicted variances of $\tilde{m}$ and θ.

Fig. 2.13 on the following page shows the marginal posteriors $\pi_t(\tilde{m} \mid z_0, z_1, ..., z_{t-1})$ and $\pi_t(\theta \mid z_0, z_1, ..., z_{t-1})$ obtained for successive values of t when the prior distributions are as given in table 2.5. The marginal distribution $\pi_t(\theta \mid z_0, z_1, ..., z_{t-1})$ was for $t > 0$ generated from the conditioned distribution $\pi_t(\theta \mid \tilde{m}, z_0, z_1, ..., z_{t-1})$ by the relation

$$\pi_t(\theta \mid z_0, z_1, ..., z_{t-1}) \propto \sum_{\tilde{m}=1} \pi_t(\theta \mid \tilde{m}, z_0, z_1, ..., z_{t-1})\, \pi_t(\tilde{m} \mid z_0, z_1, ..., z_{t-1}). \qquad \textbf{(2.35)}$$

The impact of the sequence of accidents $\{z_0, z_1, ..., z_{t-1}\}$ on the shape and location of the generated posterior distributions is clearly illustrated in fig. 2.13. Thus if very dispersed distributions are applied at $t = 0$ (*Choice 1*), the generated posterior distributions are highly displaced and reshaped relative to the distributions valid at $t = 0$. On the other hand, if very localized distributions are applied at $t = 0$ (*Choice 2*), the generated posteriors more or less maintain the shapes of the priors applied at $t = 0$.

Fig. 2.13. Marginal posterior distribution of $\tilde{m}$ and θ for successive values of t. The posteriors are based on the priors specified in table 2.5 and the accident pattern from *simulation 1* in fig.2.11.

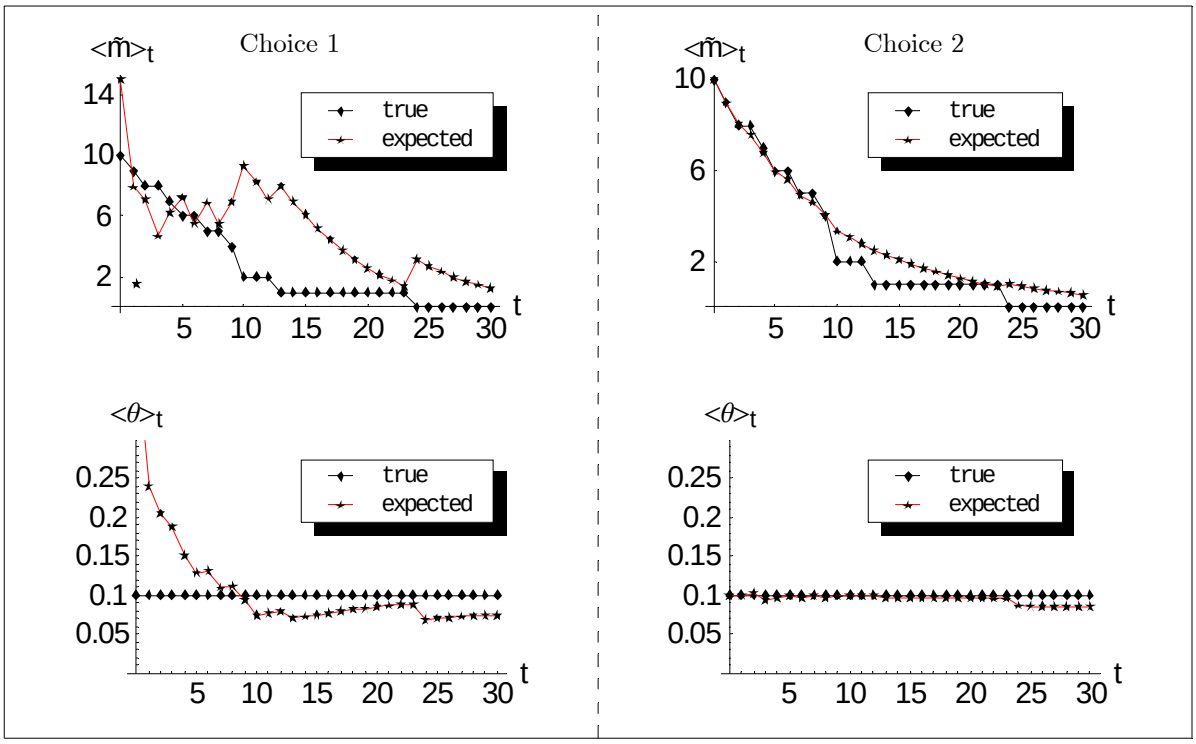


The observations made above seem to agree with common sense. That is, if the observer of the minefield under study has no or very little information at hand about the true values of $\tilde{m}$ and θ, the observer should apply very disperse prior distributions at $t = 0$ reflecting his lack of knowledge. As a consequence, high importance will be attached to the observed number of accidents when the dispersed prior distributions are updated through Bayes' theorem. This seems reasonable as the accident statistics are the only information available. On the contrary, if the observer has very detailed information at hand which allows him to set up very localized priors at $t = 0$, these prior distributions will only be slightly affected by the observed accident pattern. That is, a very extreme accident pattern has to be observed if the observer is to change his initial beliefs about the true values of $\tilde{m}$ and θ.

The true number of functional mines left in the hypothetical minefield at time t can easily be inferred from fig. 2.11. Similarly, from the marginal distributions $\pi_t(\tilde{m} \mid z_0, z_1, \ldots, z_{t-1})$ and $\pi_t(\theta \mid z_0, z_1, \ldots, z_{t-1})$ the expected value of $\tilde{m}$ and θ can be calculated for increasing

values of t. In what follows these quantities will be denoted $< \tilde{m} >_t$ and $< \theta >_t$, respectively. Fig. 2.14 below illustrates to what extent $< \tilde{m} >_t$ and $< \theta >_t$ deviate from their true values for increasing values of t. It emerges clearly from the depicted graphs that the deviations between true and expected values are sensitive to the choice of prior distributions. In the limit $t \to \infty$ it is observed that $< \tilde{m} >_t \to 0$, as expected. As long as there are mines left, $< \theta >_t$ converges to its true value for increasing values of t.

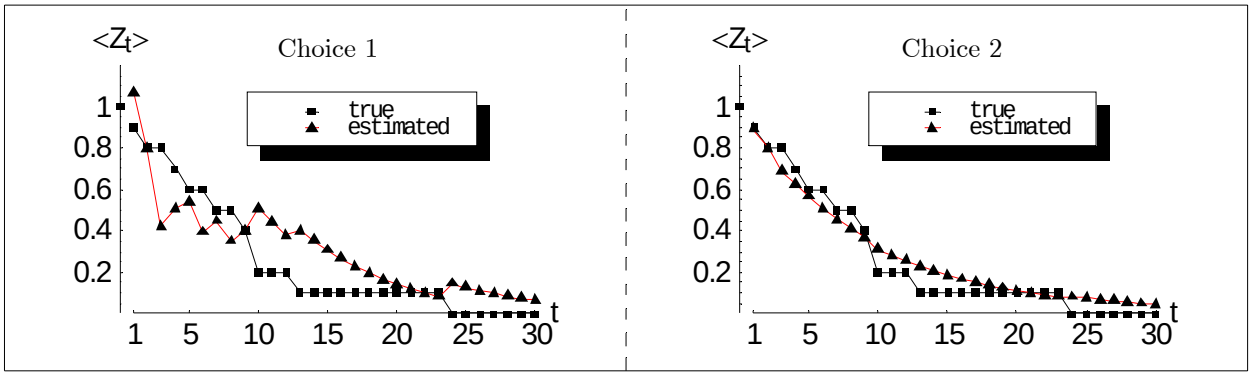
Fig. 2.14. Deviation between true and expected value of $\tilde{m}$ and θ for increasing values of t. *Choice 1* and *Choice 2* refer to the prior distributions defined in table 2.5.



Of major importance in the present context is how the combined action of the marginal and conditioned distributions of $\tilde{m}$ and θ, respectively, determines the distribution $p(z_t)$ (defined in (2.22)). In fig. 2.15 on the following page the expected number of accidents looking one observation period ahead is shown for increasing values of t. Included in the same plot is the true average $< Z_t > = \tilde{m}_t \cdot \theta$, where $\tilde{m}_t$ can be inferred from fig. 2.11. Not surprisingly, the deviations between the true and estimated value of $< Z_t >$ is sensitive to the choice of prior distributions. While the true average inevitable

decreases for every detonated mine, this is not necessarily the case if $< Z_t >$ is calculated by Bayesian updating. However, in the limit $t \to \infty$, $< Z_t > \to 0$ as expected.

Fig. 2.15. The expected number of accidents in the coming observation period as a function of t. The black curve is calculated as $< Z_t > = \tilde{m}_t \cdot \theta$. The red curve is calculated by Bayesian updating.



The theoretical case examined over the last few pages indicates that risk assessments of a minefield based on Bayesian data analysis is a feasible and sound approach as it gives a balanced weighing of prior knowledge and later obtained accident statistics. In a real-life application only reliable accident statistics from a single or a few observation periods will be available. It is therefore essential to provide informative prior distributions. A thorough discussion about how prior distributions based on various types of information can be set up is covered by chapter 4 - 14. Until then an additional example will be given to illustrate how the Bayesian approach can be of support when different minefields are to be ranked according to risk.

2.6 Application of Bayesian Data Analysis: Example 2

In the introduction to the present chapter we considered as our point of departure a hypothetical post-conflict region containing a large number of mine affected areas. Recall that the chief aim of the derivations made so far is to develop a mathematical model which will enable us to rank those mine affected areas in proportion to the risk they pose to the surrounding society.

To illustrate by a simple example how minefields can be ranked according to risk, consider two minefields, i.e., *minefield 1* and *minefield 2*, which at time $t = 0$ have been assigned

the prior distributions $\pi_0(\tilde{m})$ and $\pi_0(\theta \mid \tilde{m})$ listed in table 2.6. Thus the minefields are identical with respect to $\pi_0(\tilde{m})$ but different with respect to $\pi_0(\theta \mid \tilde{m})$. The applied priors are illustrated in fig. 2.16 and fig. 2.17, respectively.

Table 2.6. Features of prior distributions for *minefield 1* and *minefield 2*.

Minefield	$\pi_0(\tilde{m})$	$E[\tilde{m}]$	$V[\tilde{m}]$	$\pi_0(\theta)$	$E[\theta]$	$V[\theta]$	$E[Z_0]$	**Rank**
1	$Bi(30,\frac{1}{3})$	10	6.67	$Be(5,50)$	0.091	0.0015	0.91	1
2	$Bi(30,\frac{1}{3})$	10	6.67	$Be(1,10)$	0.091	0.0069	0.91	1

Fig. 2.16. $\pi_0(\tilde{m})$ for minefield 1 and 2.

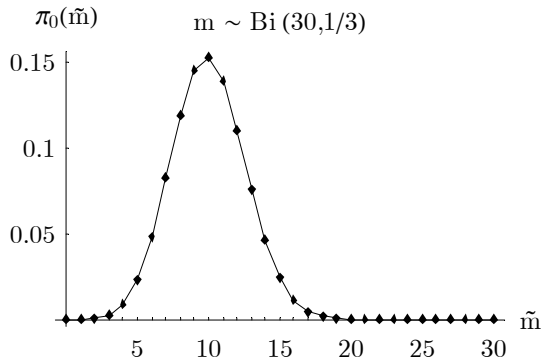
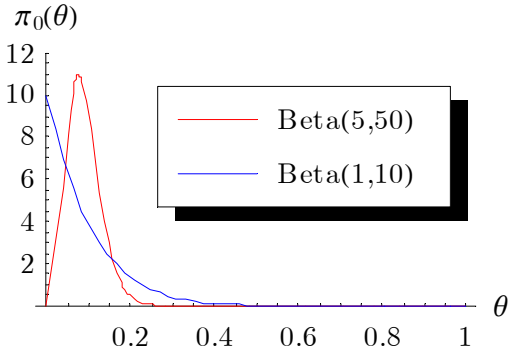


Fig. 2.17. Red curve: $\pi_0(\theta)$ for *minefield 1*; blue curve: $\pi_0(\theta)$ for *minefield 2*.



Based on the given prior distributions the second last column in table 2.6 shows the expected number of casualties in the two minefields for the coming period $\Delta(0)$ where $E[Z_0]$ has been calculated from (2.22). As $E[Z_0] = 0.91$ in both minefields, *minefield 1* and *minefield 2* are ascribed the same rank as shown in the last column of table 2.6. That is, the *risk* of a minefield is equated with the expected number of casualties in the coming observation period. In general, we will ascribe the minefield with the largest value of $E[Z_t]$ a rank of 1 and give it priority with respect to mine clearance.

Assume now that none of the minefields from above are cleared during $\Delta(0)$, and let z_0 denote the number of accidents which are actually observed in *both* minefields during $\Delta(0)$. By means of the posterior distributions $\pi_0(\tilde{m} \mid z_0)$ and $\pi_0(\theta \mid \tilde{m}, z_0)$ and the relations (2.32) and (2.33) updated rankings of the minefields valid at $t = 1$ can be obtained. Table 2.7 below shows the calculated rankings based on two scenarios: $z_0 = 0$ and $z_0 = 1$.

Table 2.7: Updated rankings of *minefield 1* and *minefield 2* at $t = 1$.

Scenario	$z_0 = 0$		$z_0 = 1$	
Minefield	$E[Z_1]$	Rank	$E[Z_1]$	Rank
1	0.73	1	0.83	2
2	0.45	2	0.84	1

In the case $z_0 = 0$, the expected value of accidents for the coming observation period $\Delta(1)$ is readjusted downwards, i.e., $E[Z_1] < 0.91$ for both minefields, see table 2.7. This is explained by the fact that the outcome $z_0 = 0$ is below the expected value of 0.91. However, the adjustment downwards is relatively stronger for *minefield 2*. Consequently, at $t = 1$ *minefield 1* is ranked 1.

The greater sensitivity of *minefield 2* to the observation $z_0 = 0$ is due to the dispersed prior distribution of θ. This is illustrated in fig. 2.18.a where the posterior distribution $\pi_0(\theta \mid 10, z_0)$ for *minefield 2* is displayed together with the prior distribution Beta(1,10). In the case $z_0 = 0$ the posterior distribution of θ is clearly displaced to smaller values of θ which leads to a downwards adjusted value of $E[Z_1]$. The corresponding posterior distribution of θ for *minefield 1* is displaced only slightly relative to its localized prior distribution, as it is seen from fig. 2.18.b.

Fig. 2.18.a.
Conditioned posterior distribution for *minefield 2*.

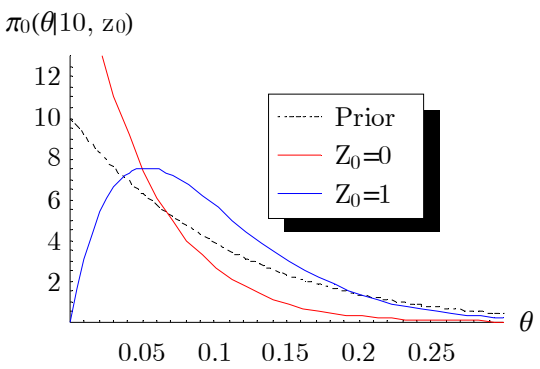
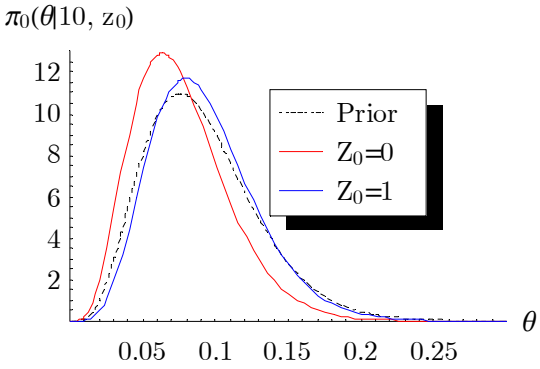


Fig. 2.18.b.
Conditioned posterior distribution for *minefield 1*.



In the case $z_0 = 1$, which is above the expected number of accidents in $\Delta(0)$, the posterior distribution $\pi_0(\theta \mid 10,1)$ for both minefields is displaced to larger values of θ relative to the

corresponding prior distribution. The effect is however largest in the case of *minefield 2* due to its dispersed prior distribution of θ. Consequently, *minefield 2* is ranked 1. Note that the expected number of accidents in $\Delta(1)$ is adjusted downwards as both minefields at $t = 1$ contain one mine less relative to $t = 0$.

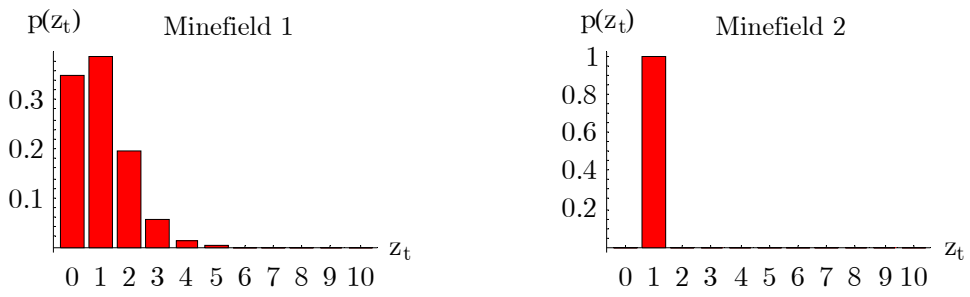
From the observations made above we can once again conclude that the Bayesian approach induces a balanced weighing of prior information and later incoming accident statistics which makes it particularly useful in relation to mine action.

2.7 Further Notes on Ranking of Minefields

In the example studied in paragraph 2.6 just two minefields were ranked, and the ranking was founded on the number of accidents to be expected in the two minefields in a coming observation period. More generally we may consider the ranking of K homogenous minefields, each minefield being characterized by a probability distribution $p(z_t^{(k)})$, and we may consider alternative ways to summarize the contents of $p(z_t^{(k)})$ than simply stating its expected value.

To elaborate on this subject, note that the expected value of $p(z_t^{(k)})$ does make up an important piece of information, but a ranking based on expected values alone may not exploit the full content of information inherent in the collective set of distributions $\{p(z_t^{(k)})\}$. To give an easy comprehensible (but rather artificial) example, consider two hypothetical minefields, *minefield 1* and *minefield 2*, where $Z_t^{(1)} \sim Bi(10, 0.1)$, and $Z_t^{(2)} = 1$ with a probability of 1 as illustrated in fig. 2.19 below.

Fig. 2.19. $p(z_t)$ for two hypothetical minefields. $Z_t^{(1)} \sim Bi(10, 0.1)$, and $Z_t^{(2)} = 1$ with a probability of 1.



As to the expected number of accidents we have that $< Z_t^{(1)} > = < Z_t^{(2)} > = 1$, but the variances are different as $\sigma_1^{\,2} = 0.9$ and $\sigma_2^{\,2} = 0$. Consequently, a ranking based on expected values will assign identical ranks to the two minefields. A simple calculation shows that $p(Z_t^{(1)} < Z_t^{(2)}) = 0.349$, $p(Z_t^{(1)} = Z_t^{(2)}) = 0.387$, and $p(Z_t^{(1)} > Z_t^{(2)}) = 0.264$. In other words, the most probable event is that the same number of accidents are observed in both minefields. However, if the observed number of accidents are different, the number of accidents in *minefield 2* will in the majority of cases be larger than the number of accidents in *minefield 1*. Thus *minefield 2* poses a larger risk than *minefield 1* and *minefield 2* should therefore be assigned a rank of 1.

One way to differentiate between minefields characterized by distributions $p(z_t^{(k)})$ with identical expected values but different variances is to treat the ranks of the minefields as stochastic variables and subsequently rank the minefields according to their expected ranks. To illustrate this approach, let the stochastic rank R_k be defined as

$$R_k = \sum_{j=1}^{K} I(Z_t^{(k)} \leq Z_t^{(j)}) = \sum_{j=1}^{K} I_{jk}, \qquad \textbf{(2.36)}$$

where $I(\cdot)$ is the indicator function, and $I_{jk} = I(Z_t^{(k)} \leq Z_t^{(j)})$ [Laird et al., 1989]. In (2.36) the minefield with the largest outcome of Z_t is assigned a rank of 1. From (2.36) the expected rank $\hat{R}_k$ can be calculated as

$$\hat{R}_k = < R_k > = \sum_{j=1}^{K} P_{jk}, \qquad \textbf{(2.37)}$$

where $P_{kk} = 1$, and $P_{jk} = p(Z_t^{(k)} \leq Z_t^{(j)})$ for all $j \neq k$. Returning to the example from fig. 2.19, the expected rank of *minefield 1* and *2* according to (2.37) turns out to be $\hat{R}_1 = 1.74$ and $\hat{R}_2 = 1.65$, respectively. To obtain integer ranks we simply arrange the expected ranks in increasing order. As $\hat{R}_2 < \hat{R}_1$ we obtain as wished that *minefield 2* is ranked as 1.

An additional merit of the method outlined above is that the variances and covariances of the stochastic ranks can be calculated [for further details, see Laird et al., 1989].

2.8 Conclusions

The conceptual framework build up in the present paper is simple but important as it clarifies the interplay between the key factors behind minefield accidents. It is evident from the preceding discussions that reliable risk assessments entail a balanced weighing of the various pieces of information which may be available to a decision maker. A risk assessment methodology which simply equates the risk of a minefield with the recorded number of accidents, or alternatively the believed number of mines present, is clearly too simplistic an approach.

The introduced risk model appears as a useful decision support tool to decision makers involved in mine action. As the application of the model is founded on Bayesian data analysis, risk assessments based on the model will reflect a balanced weighing of prior information and accident statistics from the minefield. The sensitivity of the risk model to the choice of prior distributions calls however for further analysis, and the development of refined methods for providing prior distributions are needed. Strategies for the provision of prior distributions from historical data and Bayesian modelling will be the main theme in the following chapters.

Chapter 3
Generation of Minefield Data

To carry on the analysis initiated in chapter 2, realistic data sets including accident statistics, mine clearance data, minefield area types, etc. are needed. Unfortunately, the available information about these issues is very sparse. For example, while the previously mentioned landmine impact survey reports contain accident statistics from several mine affected communities covering an observation period of 2 years, the same reports lack detailed information about the nature of the corresponding minefields which limits the statistical utility of the data. Through the included accident statistics the landmine impact surveys do however give an impression of the magnitude of the mine contamination problem and its impact. For comparison, table 3.1 below illustrates the distribution of minefield/UXO accidents in two surveyed countries. It appears from table 3.1 that for both countries, the majority of the mine affected communities has not recorded any accidents due to the presence of mines or UXO within two years prior to the survey.

Surveyed Country	**Yemen**	**Mozambique**
No. of recent victims	**No. of communities**	**No. of communities**
0	514	710
1	39	45
2	23	11
3	5	13
4	4	2
5	1	3
6	1	0
7	3	0
8	1	1
10	0	1
>10	1	1
unknown	0	4

Table 3.1. Source: Canadian International Demining Corps et al., 2001, Survey Action Centre et al., 2000.

As the landmine impact survey reports do not contain any information about the likely number of mines in the minefields under study, nothing can be concluded from the

accident statistics in table 3.1 about the probability of encountering a mine. Concerning information about the observed density distribution of landmines, only a few references in the literature are available including Bajic (Bajic et al., 2003) and Trevelyan (Trevelyan, 1997). While Bajic et al. apply clearance data collected in Croatia to derive empirical statistical models of minefield areas and spatial densities of AP- and AT mines (see fig. 3.1 - 3.2 below), Trevelyan uses clearance reports from mine clearance operations undertaken in Afghanistan to estimate clearance rates (see fig. 3.3 and 3.4 for observed mine densities). Neither the study by Trevelyan nor the study by Bajic et al. include any kind of accident statistics covering the studied minefields.

Fig. 3.1 (left): Lognormal model of minefield areas based on observations made in Croatia 1998-2001. Data source: (Bajic et al., 2003). Fig. 3.2 (right): Lognormal models of mine densities based on Croatia data. Solid line: AT mines, dashed line: AP mines. Data source: (Bajic et al., 2003).

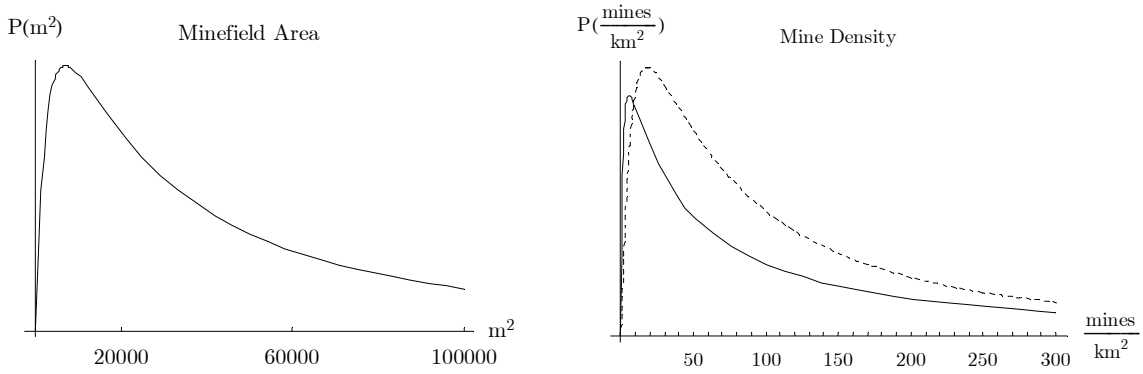
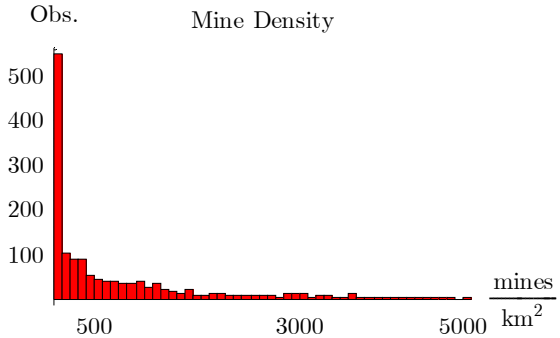


Fig. 3.3 (left): Observed AP mine densities based on approximately 1700 cleared minefields in Afghanistan until mid-May 1997. Data source: (Trevelyan, 1997). Fig. 3.4 (right): Statistical features of frequency distribution shown in fig. 3.3. Data source: (Trevelyan, 1997).



Mine Density Afghanistan (mines/km^2)	
Fraction of minefields containing zero mines	0.23
25% quantile	23
50% quantile	435
75% quantile	1987

From the figures included above it appears that the same asymmetric pattern is observed in both countries as to the mine density, that is, most minefields in a given country display a relatively small mine density, while a few number of minefields have a relatively high mine density. The median mine density is however considerably higher in Afghanistan than in Croatia. Note from fig. 3.4 that around 23% of the areas in Afghanistan originally classified as minefields turned out to be mine free.

In the chapters which follow, various methods which may prepare the way for real-life applications of the binomial model derived in chapter 2 will be introduced. To substantiate the utility of the proposed methods it would be preferable to test each suggested method on one or several relevant data sets picked out from ongoing or completed mine clearance programmes. However, the fragmentary nature of the data available at present in Humanitarian Mine Action excludes the possibility of performing such tests. Examination of the various methods on *simulated* but *realistic* data sets is therefore the only option left.

To generate a simulated data set covering 1000 *virtual* minefields, which suffices in the present context, the following procedure was followed: Firstly, 1000 sets of binomial parameters $(\tilde{m}_j, \theta_j)$ were sampled (for details, see below) where $\tilde{m}_j$ denotes the number of functional mines present in minefield j at time t = -1, and θ_j denotes the probability of a randomly selected mine being triggered by a person during the following observation period. Secondly, based on the 1000 pairs of binomial parameters, accident statistics were simulated by making 1 draw y_j from each of the 1000 binomial distributions, that is, $y_j \sim Bi(\tilde{m}_j, \theta_j)$. Each minefield in the simulated data set is thus characterized by three records as shown in table 3.2.

Table 3.2. Records in simulated data set.

Minefield	$\tilde{m}_j$	θ_j	y_j
1	$\tilde{m}_1$	θ_1	y_1
2	$\tilde{m}_2$	θ_2	y_2
---	---	---	
1000	$\tilde{m}_{1000}$	θ_{1000}	y_{1000}

Fig. 3.5 below illustrates the frequency of virtual minefields containing a given number of functional mines. The outcome depicted in fig. 3.5 was generated by sampling $\tilde{m}_j$ 1000 times from a Log-Series distribution. Table 3.3 tabulates selected quantiles.

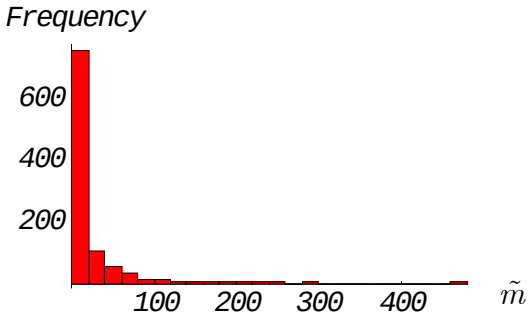


Fig. 3.5 Frequency of minefields containing $\tilde{m}$ functional mines

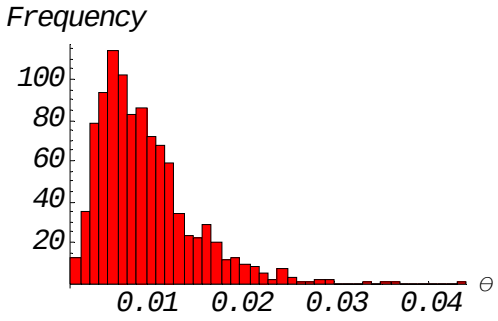


Fig. 3.6. Frequency of θ for virtual minefields.

Table 3.3. Quantiles corresponding to distribution of $\tilde{m}$ in 1000 virtual minefields.

X%	**Number of mines in X% quantile**
10%	0
20%	0
30%	1
40%	2
50%	4
60%	7
70%	14
80%	25
90%	51
100%	475
Median	17.74

The distribution of the sampled values of the probability parameter θ_j corresponding to $\tilde{m}_j$ is depicted in fig. 3.6. This distribution was generated in the following way: Initially, a parameter α_j was drawn 1000 times from a normal distribution $N(\alpha_j \mid \mu, \tau)$. For every drawn α_j, the corresponding θ_j was calculated through the transformation $\theta_j = e^{\alpha_j}(1+e^{\alpha_j})^{-1}$. The specific choice of parameters $(\mu, \tau) = (-4.7, 0.5)$ corresponds to $E[\theta] = 0.010$ which leads to a realistic pattern of accident statistics, see table 3.4 on the following page.

Note that the typical virtual minefield contains a small number of mines or no mines, while a few number of minefields contain a very large number of mines, as it emerges from

fig. 3.5 and table 3.3. The vast majority of the virtual minefields exhibits furthermore no or very few recorded accidents as it emerges from table 3.3.

Table 3.4 Simulated accident statistics from 1000 virtual minefields.

Number of observed casualties	**Number of minefields**
0	887
1	81
2	19
3	7
4	2
5	2
6	2
≥7	0

In the following chapters the simulated data set will be used in two different settings. In chapter 4 it is assumed that a hypothetical decision maker has access to a small sample $\{(\tilde{m}_{k_1}, y_{k_1}),(\tilde{m}_{k_2}, y_{k_2}),...,(\tilde{m}_{k_M}, y_{K_M})\}$ picked at random from the simulated data set. From this sample it is possible to estimate the distribution of the binomial parameters $\{\theta_1, \theta_2, ..., \theta_{1000}\}$ through *Bayesian hierarchical modelling*.

In the chapters 5 – 13 the hypothetical decision maker has access to the complete accident statistics $\{y_1, y_2, ..., y_{1000}\}$ from the simulated data set but does not have information about the mine content in any individual minefield under study. In this case an estimate of the distribution of $\{\theta_1, \theta_2, ..., \theta_{1000}\}$ can be provided through the application of *finite mixture models*.

Chapter 4
Hierarchical Bayesian Models

4.1. Introduction

In chapter 2 it was concluded that the number of casualties in a mine affected area under fairly general assumptions can be considered to be the outcome of a binomial process. A given minefield can therefore be characterized by its current set of binomial parameters $(\tilde{m},\theta)$, and the expected number of casualties in the future can be estimated via estimates of $\tilde{m}$ and θ.

In the Bayesian risk model suggested in chapter 2, estimates of $\tilde{m}$ and θ are requested in terms of probability distributions. More specifically, the following set of prior distributions have to be provided:

$$\pi_t(\tilde{m}) = \{\pi_t(0), \pi_t(1), ...\}, \tag{4.01}$$

$$\pi_t(\theta \mid \tilde{m}) \text{ for } \tilde{m} \geq 1. \tag{4.02}$$

Of the two parameters $\tilde{m}$ and θ, information about $\tilde{m}$ appears at first to be the more accessible. That is, several sources can provide information to a decision maker concerning the possible degree of mine contamination in a mine affected area. These sources include military mine maps and related archives, military staff and other ex-combatants with local knowledge, and local or regional authorities. In the future it may furthermore become technically possible to complement these sources of information by actual geophysical measurements or other kinds of measurements in the minefield. Therefore, through a proper synthesis of the different pieces of information it should be possible to construct a prior $\pi_t(\tilde{m})$.

The major obstacle to a real-life application of the risk model derived in chapter 2 seems therefore to be the lack of actual information about the binomial parameter θ. In the present report we have therefore decided to focus exclusively on ways to extract information about θ through statistical modelling.

With the above aim in mind, two different types of models will be examined. In the present chapter it will be demonstrated through the application of a hierarchical Bayesian model how a probability distribution $p(\theta)$ can be generated by combining accident statistics and clearance data from mine clearance operations. In the chapters 5-13 a so-called *finite mixture model* will be studied which only requires the availability of accident statistics.

The contents of the present chapter will be as follows: In paragraph 4.2, the hierarchical Bayesian model, which is based on the Beta-distribution, is introduced. Paragraph 4.3 follows with a short discussion about the choice of a prior distribution for the parameters specifying the Beta-distribution. In paragraph 4.4, the concept of Monte Carlo importance sampling is introduced. This involves in particular the selection of a usable importance sampling density. A numerical study follows in paragraph 4.5 based on the minefield data introduced in chapter 3. A summary and final conclusions are given in paragraph 4.6.

4.2. A Hierarchical Bayesian Model

To generate a qualified estimate of the binomial parameter θ related to some minefield, consider a group of *previously* mine affected areas, say $\{\text{area}_1, \text{area}_2, ..., \text{area}_J\}$. These areas might for example be geographically located in a different but comparable region or country where a larger mine clearance programme has been completed. For each area from the above list we will assume that the following two observations are available: 1) The number of casualties recorded two years preceding the mine clearance operation; 2) The number of functional mines located during the mine clearance operation. Based on such information the following table can be set up:

Table 4.01. Data from J previously mine affected areas. $\tilde{m}^i_{t-1} = \tilde{m}^i_t + z^i_{t-1}$ for all areas.

Area	Number of mines located at time t	Casualties during $\Delta(t-1)$	Mines at time $t-1$	Unknown
1	$\tilde{m}^1_t$	z^1_{t-1}	$\tilde{m}^1_{t-1}$	θ_1
2	$\tilde{m}^2_t$	z^2_{t-1}	$\tilde{m}^2_{t-1}$	θ_2
...	...	...	...	...
J	$\tilde{m}^J_t$	z^J_{t-1}	$\tilde{m}^J_{t-1}$	θ_J

Table 4.01 makes up what we might term historical data. The main observation to be made is that by adding the numbers $\tilde{m}_t^i$ and z_{t-1}^i, the set $\{z_{t-1}^i, \tilde{m}_{t-1}^i\}$ is accessible for each minefield. We will assume that $\tilde{m}_{t-1}^i > 0$ for all areas included in table 4.01.

The column located at the extreme right in table 4.01 contains the unknown binomial parameters $\{\theta_1, \theta_2, ..., \theta_J\}$ covering the J minefields. To set up a probability distribution $p(\theta)$ we will assume that the members of the set $\{\theta_1, \theta_2, ..., \theta_J\}$ make up random samples from the same probability distribution. This is a valid assumption if no complementary information is available about the individual minefields. As $\theta_j \in]0;1[$, a convenient choice of a common probability distribution is the Beta-distribution. That is, we assume

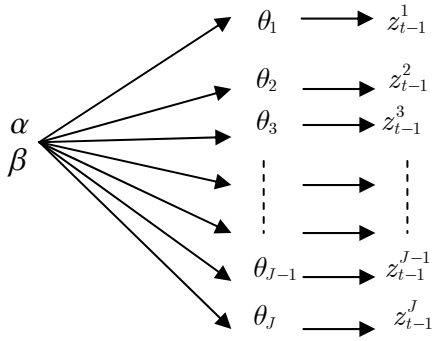
$$\theta_j \sim Beta(\alpha, \beta) \; \forall \; j \qquad \textbf{(4.03)}$$

which implies that

$$p(\theta_j \mid \alpha, \beta) = \frac{1}{B(\alpha, \beta)} \theta_j^{\alpha-1} (1-\theta_j)^{\beta-1}. \qquad \textbf{(4.04)}$$

The hierarchical relationship between the observations z_{t-1}^j, the parameters θ_j and the hyperparameters α and β can be sketched as illustrated in fig. 4.01 below.

Figure 4.01 Hierarchical structure between hyperparameters, binomial parameters, and observations.



To avoid a cluttered notation we will in the equations which follow simply write $\{z_{t-1}^j, \tilde{m}_{t-1}^j\}$ as $\{z_j, m_j\}$. Through the information contained in table 4.01 we can update our knowledge about the hyperparameters (α, β) by the application of Bayes' theorem:

$$p(\alpha, \beta \mid z_1, z_2, ..., z_J, m_1, m_2, ..., m_J) \propto p(z_1, z_2, ..., z_J \mid m_1, m_2, ..., m_J, \alpha, \beta) p(\alpha, \beta) \qquad \textbf{(4.05)}$$

where $p(\alpha,\beta)$ in (4.05) denotes our prior distribution of α and β. We will return to this issue later. The likelihood function appearing in (4.05) can be written as

$$\begin{aligned} p(z_1,z_2,...,z_J \mid m_1,m_2,...,m_J,\alpha,\beta) &= \prod_{j=1}^{J} p(z_j \mid m_1,m_2,...,m_J,\alpha,\beta) \\ &= \prod_{j=1}^{J}\Big[\int p(z_j \mid m_j,\theta_j)p(\theta_j \mid \alpha,\beta)\,d\theta_j\Big] \\ &= \prod_{j=1}^{J}\int \binom{m_j}{z_j}\frac{\theta_j^{\alpha+z_j-1}(1-\theta_j)^{\beta+m_j-z_j-1}}{B(\alpha,\beta)}\,d\theta_j \\ &= \prod_{j=1}^{J}\binom{m_j}{z_j}\frac{B(\alpha+z_j,\beta+m_j-z_j)}{B(\alpha,\beta)}. \end{aligned} \qquad \textbf{(4.06)}$$

Note that the distribution of θ_j is independent of the unit index j. This implies in particular that m_j by assumption conveys no information about θ_j.

Inserting the likelihood function (4.06) into (4.05) we get the following expression for the posterior distribution of α and β:

$$p(\alpha,\beta \mid z_1,z_2,...,z_J,m_1,m_2,...,m_J) \propto \prod_{j=1}^{J}\left[\binom{m_j}{z_j}\frac{B(\alpha+z_j,\beta+m_j-z_j)}{B(\alpha,\beta)}\right]p(\alpha,\beta). \qquad \textbf{(4.07)}$$

Consider now an existing minefield characterized by the binomial parameter θ. If the parameter θ originates from the same "superpopulation" as the parameters $\{\theta_1,\theta_2,...,\theta_J\}$, we can exploit the posterior $p(\alpha,\beta \mid z_1,z_2,...,z_J,m_1,m_2,...,m_J)$ to generate an estimate of the distribution of θ. That is, we will write $p(\theta)$ as

$$\boxed{\begin{aligned} p(\theta) &= \iint p(\theta \mid \alpha,\beta)\,p(\alpha,\beta \mid z_1,z_2,...,z_J,m_1,m_2,...,m_J)\,d\alpha d\beta \\ &\propto \iint \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}\prod_{j=1}^{J}\left[\frac{B(\alpha+z_j,\beta+m_j-z_j)}{B(\alpha,\beta)}\right]p(\alpha,\beta)\,d\alpha d\beta. \end{aligned}} \qquad \textbf{(4.08)}$$

The estimate $p(\theta)$ can be used in different ways. One possible choice is to insert $p(\theta)$ as our prior distribution of θ in the Bayesian risk model derived in chapter 2. Alternatively, we can estimate $E[\theta]$ and possibly $Var[\theta]$ from $p(\theta)$ and use these estimates as *partial*

information. In chapter 14 we will give a detailed account of the construction of priors based on partial information.

We have thus arrived at a method which in a simple way extracts information about the binomial parameter θ. However, the limitations of the method should be fully realized. That is, the use of $p(\theta)$ as a prior distribution for θ for some minefield can only be justified if the minefield under study in all essential features is similar to the minefields which make up the *historical data*.

An alternative to (4.08) can be set up if the data $\{z_j, m_j\}$ are supplemented by explanatory variables $(x_1^j, x_2^j, ..., x_k^j) = x^j$ for all j. One possible choice is to express the relation between θ_j and the explanatory variables x^j as

$$\log \frac{\theta_j}{1-\theta_j} = \alpha + \beta x^j, \tag{4.09}$$

from which it follows that

$$\theta_j = \frac{\exp(\alpha + \beta x^j)}{1 + \exp(\alpha + \beta x^j)}. \tag{4.10}$$

As $z_j \sim Bi(m_j, \theta_j)$ it follows from (4.10) that the corresponding likelihood function takes the form

$$\begin{aligned} &p(z_1, z_2, ..., z_J \mid \alpha, \beta, m_1, m_2, ..., m_J, x^1, x^2, ..., x^J) \\ &= \prod_{j=1}^{J} \binom{m_j}{z_j} \left[\frac{\exp(\alpha + \beta x^j)}{1 + \exp(\alpha + \beta x^j)}\right]^{z_j} \left[1 - \frac{\exp(\alpha + \beta x^j)}{1 + \exp(\alpha + \beta x^j)}\right]^{m_j - z_j}. \end{aligned} \tag{4.11}$$

According to Bayes' theorem the conditioned posterior distribution of α, β can now be obtained as

$$\begin{aligned} &p(\alpha, \beta \mid z_1, z_2, ..., z_J, m_1, m_2, ..., m_J, x^1, x^2, ..., x^J) \\ &\propto p(z_1, z_2, ..., z_J \mid \alpha, \beta, m_1, m_2, ..., m_J, x^1, x^2, ..., x^J)\, p(\alpha, \beta \mid x^1, x^2, ..., x^J), \end{aligned} \tag{4.12}$$

where $p(\alpha, \beta \mid x^1, x^2, ..., x^J)$ denotes the prior distribution of α, β conditioned on the explanatory variables. From the posterior distribution in (4.12), an estimate $p(\theta)$ conditioned on the explanatory variables can subsequently be generated.

In a real-life application, an estimate of $p(\theta)$ based on (4.12) is preferable to (4.08) as the inclusion of explanatory variables improves a decision makers ability to discriminate between various types of minefields. However, to keep the discussions at a general level in the following paragraph we will focus exclusively on model (4.08).

4.3. Specification of Prior Distribution

To apply model (4.08), the prior $p(\alpha,\beta)$ has to be specified. On the assumption that no information about α and β is available, we are looking for a probability distribution whose influence on the posterior $p(\alpha,\beta \mid z_1, z_2, ..., z_J, m_1, m_2, ..., m_J)$ is marginal, that is, the posterior distribution should be dominated by the likelihood function (4.06). Priors carrying this property are generally termed *noninformative priors*. Different principles may be used when constructing noninformative priors, but two well established methods are *Jeffreys' prior* [Jeffreys, 1946] and the *reference prior* approach as defined by Bernardo [Bernardo, 1979]. In the case of the Beta-distribution parameters α and β, both of the above methods identify the noninformative prior as the square root of the Fisher information matrix, that is,

$$p(\alpha,\beta) \propto \left|I(\alpha,\beta)\right|^{1/2}, \tag{4.13}$$

where the Fisher information matrix is given as

$$I(\alpha,\beta) = \begin{bmatrix} \psi'(\alpha) - \psi'(\alpha+\beta) & -\psi'(\alpha+\beta) \\ -\psi'(\alpha+\beta) & \psi'(\beta) - \psi'(\alpha+\beta) \end{bmatrix}, \tag{4.14}$$

$\psi'(x)$ being the trigamma function. Apart from having an intractable analytical expression due to the presence of the trigamma function, the square root of the Fisher information is improper, that is,

$$\int_0^\infty \int_0^\infty \left|I(\alpha,\beta)\right|^{1/2} d\alpha d\beta = \infty \tag{4.15}$$

which is due to the fact that $\left|I(\alpha,\beta)\right|^{1/2} \to \infty$ when either α or β goes to zero. Consequently, the square root of the Fisher information cannot be applied as a prior distribution unless the corresponding posterior $p(\alpha,\beta \mid z_1, z_2, ..., z_J)$ can be proved to be

proper. To avoid such complications we will replace (4.13) by a function "similar" in shape but with a simpler analytical expression. A convenient choice turns out to be

.

$$p(\alpha,\beta) \propto \frac{1}{(\alpha+\beta)^{5/2}}. \qquad \textbf{(4.16)}$$

(4.16) is in fact improper, but it can be shown that the corresponding posterior $p(\alpha,\beta \mid z_1, z_2, ..., z_J)$ is proper if there exists at least one observation z_i where $0 < z_i < m_i$. For a thorough discussion of (4.16), see [Gelman et al., 2003, p. 128]. Fig. 4.02 and 4.03 below illustrate for comparison 3D-plots of $\left|I(\alpha,\beta)\right|^{1/2}$ and $(\alpha+\beta)^{-5/2}$, respectively.

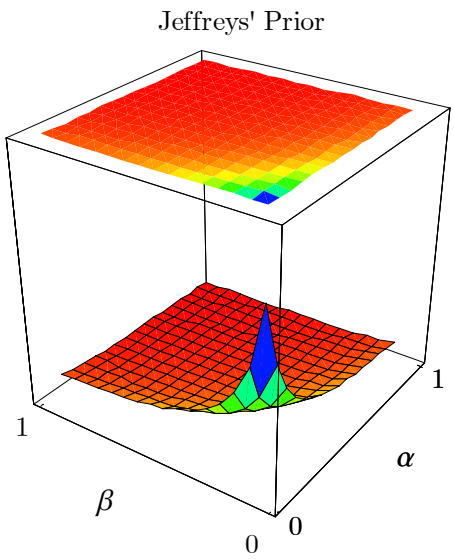


Fig. 4.02. Plot of Jeffreys' prior for the Beta distribution parameters, $p(\alpha,\beta) \propto \left|I(\alpha,\beta)\right|^{1/2}$.

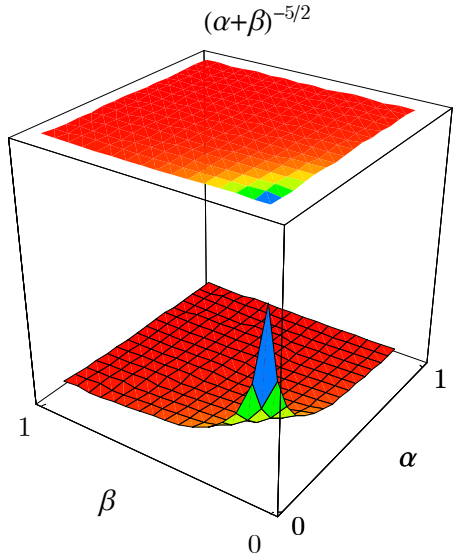


Fig. 4.03: Plot of $p(\alpha,\beta) \propto (\alpha+\beta)^{-5/2}$

4.4. Monte Carlo Integration with Importance Sampling

The integral

$$p(\theta) = \iint p(\theta \mid \alpha,\beta)\, p(\alpha,\beta \mid z_1, z_2, ..., z_J, m_1, m_2, ..., m_J)\, d\alpha d\beta \qquad \textbf{(4.17)}$$

cannot be undertaken analytically, and it is difficult to evaluate (4.17) by some quadrature method for large values of J. An approximation to $p(\theta)$ can however be generated by the method of importance sampling. The main idea in importance sampling is simple. Let

$I(\alpha,\beta)$ denote a distribution which is easy to sample from, and whose support includes the support of $p(\alpha,\beta \mid z_1,z_2,...,z_J,m_1,m_2,...,m_J)$. Writing $p(\theta)$ as

$$\begin{aligned} p(\theta) &= \iint p(\theta \mid \alpha,\beta)\, p(\alpha,\beta \mid z_1,z_2,...,z_J,m_1,m_2,...,m_J)\, d\alpha d\beta \\ &= \iint Be(\theta \mid \alpha,\beta) \frac{p(\alpha,\beta \mid z_1,z_2,...,z_J,m_1,m_2,...,m_J)}{I(\alpha,\beta)} I(\alpha,\beta)\, d\alpha d\beta \\ &= \iint Be(\theta \mid \alpha,\beta)\, w(\alpha,\beta)\, I(\alpha,\beta)\, d\alpha d\beta, \end{aligned} \tag{4.18}$$

$p(\theta)$ can be approximated by $p_m(\theta)$ defined as

$$p_m(\theta) \equiv \sum_{i=1}^{m} Be(\theta \mid \alpha_i,\beta_i)\ \frac{w(\alpha_i,\beta_i)}{\sum_{k=1}^{m} w(\alpha_k,\beta_k)}, \tag{4.19}$$

where

$$w(\alpha_i,\beta_i) = \frac{p(\alpha_i,\beta_i \mid z_1,z_2,...,z_J,m_1,m_2,...,m_J)}{I(\alpha_i,\beta_i)}, \tag{4.20}$$

and $\{(\alpha_1,\beta_1),(\alpha_2,\beta_2),...,(\alpha_m,\beta_m)\}$ denotes points sampled from $I(\alpha,\beta)$. It can be shown that

$$p_m(\theta) \to p(\theta) \tag{4.21}$$

for $m \to \infty$ given that $p(\theta)$ exists and is finite [Geweke, 1989, Tanner 1993]. The density $I(\alpha,\beta)$ is denoted the *importance sampling density,* and $w(\alpha_i,\beta_i)$ in (4.20) is denoted an *importance weight*. The value of $p_m(\theta)$ is invariant with respect to an arbitrary scaling of $I(\alpha,\beta)$ or $p(\alpha,\beta \mid z_1,z_2,...,z_J,m_1,m_2,...,m_J)$. Consequently, the normalization constant of the posterior $p(\alpha,\beta \mid z_1,z_2,...,z_J,m_1,m_2,...,m_J)$ is not needed.

The main result from (4.19) is that $p(\theta)$, through an appropriate choice of $I(\alpha,\beta)$, can be approximated by a linear combination of Beta-distributions. What is uncertain, however, is how to choose $I(\alpha,\beta)$ in the first place. That is, it is unclear how a given choice of $I(\alpha,\beta)$ affects the numerical accuracy of $p_m(\theta)$ and the overall efficiency of the algorithm. To elaborate on that, consider a sampling distribution $I(\alpha,\beta)$ satisfying

$$w(\alpha,\beta) < \overline{w} < \infty \ \forall\ (\alpha,\beta) \in]0;\infty[\times]0;\infty[, \tag{4.22}$$

and

$$\iint Be(\theta \mid \alpha,\beta)^2\, w(\alpha,\beta)\, p(\alpha,\beta \mid z_1, z_2, ..., z_J, m_1, m_2, ..., m_J)\, d\alpha\, d\beta < \infty. \qquad \textbf{(4.23)}$$

It can then be shown [Geweke, 1989] that

$$\sqrt{m}(p_m(\theta) - p(\theta)) \Rightarrow N(0, \sigma^2), \qquad \textbf{(4.24)}$$

where

$$\sigma^2 \propto \iint \left[Be(\theta \mid \alpha,\beta) - p(\theta)\right]^2 w(\alpha,\beta)\, p(\alpha,\beta \mid z_1, z_2, ..., z_J, m_1, m_2, ..., m_J)\, d\alpha\, d\beta. \qquad \textbf{(4.25)}$$

From (4.24) it follows that the numerical accuracy of the estimator $p_m(\theta)$ in general is improved if σ^2 is diminished. It appears from (4.25) and (4.20) that σ^2 is kept small if $I(\alpha,\beta)$ is similar in shape to the posterior $p(\alpha,\beta \mid z_1, z_2, ..., z_J, m_1, m_2, ..., m_J)$. Problems might arise, however, if the tail of $I(\alpha,\beta)$ goes to zero at a higher rate than the posterior itself. In that case, very large weights will show up occasionally which will induce the value of $p_m(\theta)$ to fluctuate severely even after several iterations.

Proof of (4.24) demands in general a detailed mathematical analysis of the involved posterior and importance sampling distribution, in particular investigations of the tail behaviour of both distributions. In the present context we will not spent time on mathematical proofs but instead provide an illustrative example of the potential problems involved in setting up a sampling distribution. Consider therefore the data set in table 4.02 representing historical data from 50 virtual minefields.

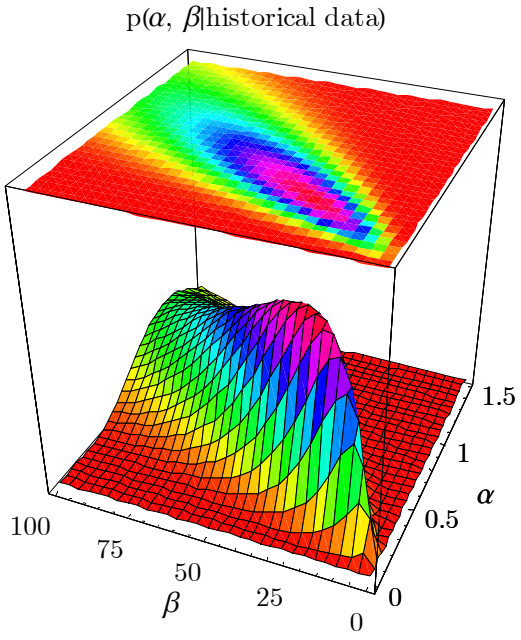
Table 4.02. Historical data. Number of functional mines present in 50 virtual minefields at time $t-1$ and the associated number of casualties observed during the observation period $\Delta(t-1)$.

$\tilde{m}_{t-1}$	z_{t-1}
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 5, 5, 7, 8, 8, 8, 8, 8, 10, 10, 14, 16, 17, 18, 25, 25, 36, 44, 45, 131, 133	0
40, 43, 52, 55, 75	1
295	2
75	3
219	6

The historical data in table 4.02 were constructed from the large data set introduced in chapter 3 (containing 1000 minefields) by random sampling from the subset containing at least one mine. Table 4.02 shows the number of functional mines at time $t-1$ in the 50 sampled minefields grouped according to the associated number of casualties observed during $\Delta(t-1)$.

Fig. 4.04 below shows the posterior $p(\alpha, \beta \,|\, \textit{historical data})$ obtained from the accident statistics and mine data in table 4.02. The posterior distribution was calculated in accordance with (4.05).

Fig. 4.04. Posterior distribution of the Beta-distribution parameters α, β based on accident statistics and clearance data from 50 minefields. "Historical data" refers to table 4.02.



To recast the posterior distribution into a form which resembles a multivariate normal distribution, we will introduce the following reparameterization:

$$r = \log\frac{\alpha}{\beta}, \tag{4.26}$$

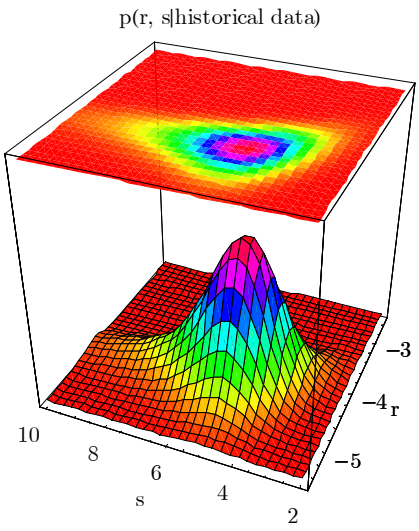
$$s = \log(\alpha + \beta). \tag{4.27}$$

The posterior distribution in terms of the new coordinates r, s can be written as

$$
\begin{aligned}
& p(r,s \mid z_1, z_2, \ldots, z_J, m_1, m_2, \ldots, m_J) \\
& = p(\alpha,\beta \mid z_1, z_2, \ldots, z_J, m_1, m_2, \ldots, m_J) \begin{vmatrix} \frac{d\alpha}{dr} & \frac{d\alpha}{ds} \\ \frac{d\beta}{dr} & \frac{d\beta}{ds} \end{vmatrix} \\
& = p(\alpha,\beta \mid z_1, z_2, \ldots, z_J, m_1, m_2, \ldots, m_J)\, \alpha\beta.
\end{aligned}
\tag{4.28}
$$

Fig. 4.05 below shows the posterior $p(r,s \mid \textit{historical data})$ which resembles a bivariate normal distribution. However, the posterior in fig. 4.05 is characterized by a strong shoulder extending into the positive direction of the s-axis. A bivariate normal approximation $N((r,s) \mid \mu, \Sigma)$ is shown in fig. 4.06 where μ is equal to the mode of $p(r,s \mid \textit{historical data})$, and Σ^{-1} is the negative of the hessian of $p(r,s \mid \textit{historical data})$ evaluated at the mode.

Fig. 4.05 (left figure below). Posterior distribution of r and s based on historical data from table 4.02.



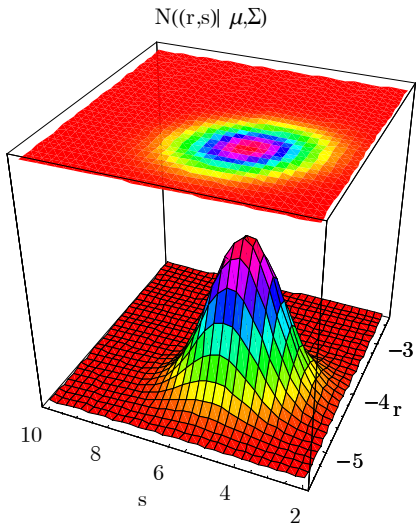
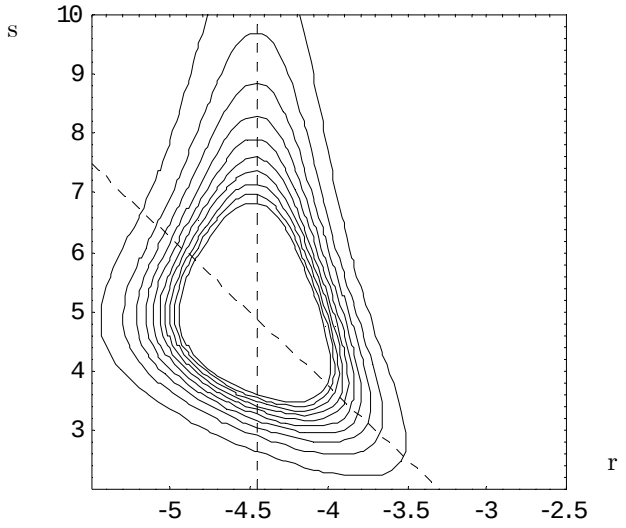


Fig. 4.06 (right figure above). Multivariate normal approximation to $p(r,s \mid \textit{historical data})$ from fig. 4.05. $\mu = (-4.46, 4.89)$, $\Sigma_{11} = 0.100$, $\Sigma_{12} = \Sigma_{21} = -0.0799$, $\Sigma_{22} = 1.057$.

To investigate the usefulness of the normal approximation as an importance sampling density, fig. 4.07 on the following page illustrates the location of the contour lines of the posterior $p(r,s \mid \textit{historical data})$ around its mode $\mu = (-4.46, 4.89)$. Fig. 4.07 clearly shows that the rate of decrease of the posterior is smallest in the positive directions of the superimposed dashed lines.

Fig. 4.07 (left figure below). Contourplot of the posterior $p(r,s \mid \textit{historical data})$. In the positive directions of the dashed lines the rate of decrease of the posterior is particular small.



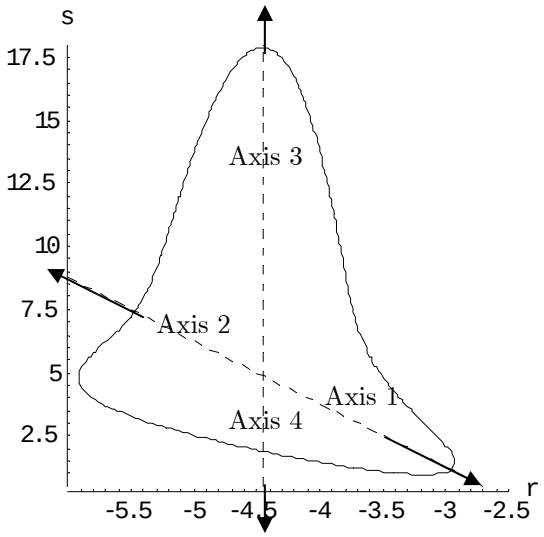


Fig. 4.08 (right figure above). Single contour line of posterior $p(r,s \mid \textit{historical data})$. The probability along the contour line is 0.001 of the probability at the mode $\mu = (-4.46, 4.89)$. Each of the four axis originating from the mode are parameterized $x_j(\delta_i) = \mu + \delta_i \, Te^{(j)}$, $i, j = 1,2$.

In fig. 4.08 the superimposed lines from fig. 4.07 are shown together with the contour line along which the posterior $p(r,s \mid \textit{historical data})$ has decreased to 0.001 of its value at the mode. Based on the superimposed lines we can define four axes as shown in fig. 4.08, every axis originating from the mode μ and parameterized as

$$x_j(\delta_i) = \mu + \delta_i \, Te^{(j)}, \; i, j = 1,2; \qquad \textbf{(4.29)}$$

where $e^{(1)} = (1,0)$, $e^{(2)} = (0,1)$, $\delta_1 \geq 0$, $\delta_2 \leq 0$ and $T = \begin{vmatrix} 0.37 & 0 \\ -0.93 & 1 \end{vmatrix}$.

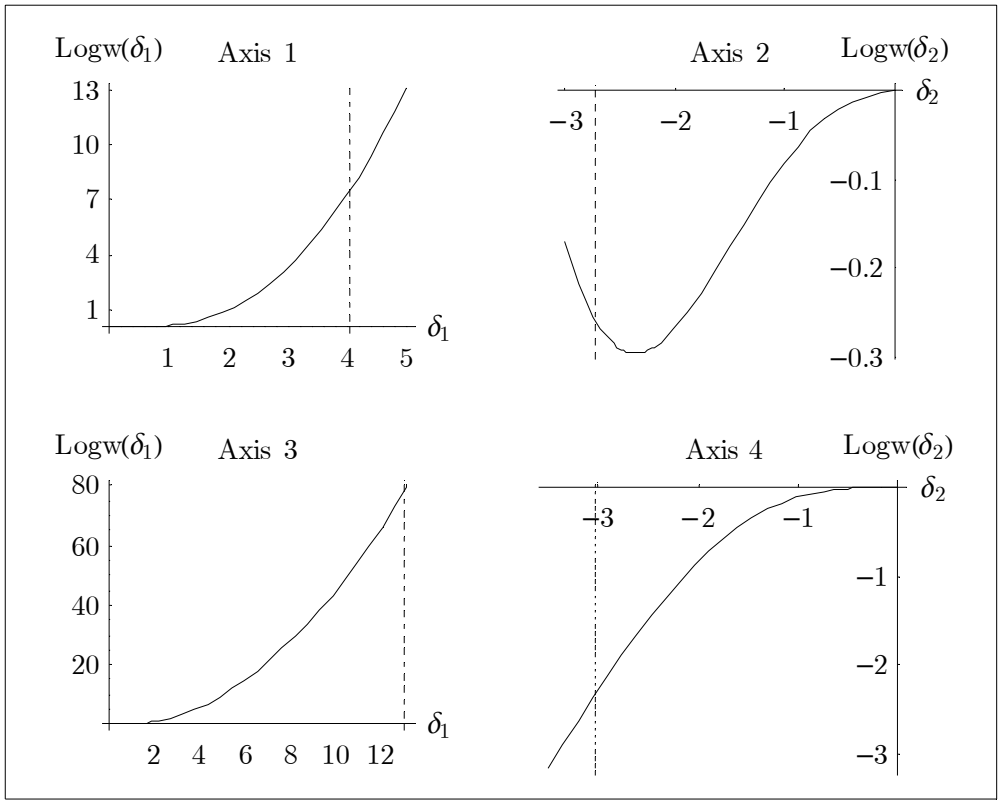
The four axis are labelled according to the scheme given in table 4.03.

Table 4.03. Axis labels.

Axis	1	2	3	4
i	1	2	1	2
j	1	1	2	2

Fig. 4.09 illustrates the sizes of the weights $w(r,s)$ along each of the above axes if the bivariate normal distribution from fig. 4.07 is used as the importance sampling density. Not surprisingly, very large weights show up along *Axis 1* and *Axis 3* as the normal approximation $N((r,s) \mid \mu, \Sigma)$ goes to zero at a faster rate than $p(r,s \mid \textit{historical data})$ in these particular directions. Consequently, one should expect large fluctuations in the value of $p_m(\theta)$ when a point located along or in the vicinity of these axes is sampled by the importance sampling density.

Fig. 4.09. The logarithm to the weight $w(r,s)$ along each of the axis defined in (4.29). The vertical dashed line in each plot indicates the value of δ_i which corresponds to the 0.001-contour line shown in fig. 4.08. Points sampled along or in the vicinity of *Axis 1* and *Axis 3* will in particular give rise to very large weights.



To remedy the above defects, an alternative to the normal approximation is needed. One possibility is the so-called *k-variate split normal density* [Geweke, 1989] which allows one to adjust the spread in each of the directions defined in (4.29). The k-variate split normal density $N^*(\mu, T, q, r)$ is specified by four sets of parameters. In the case $k = 2$, the parameters μ, q and r are two-dimensional vectors with positive components, and T is a

two-by-two matrix. In what follows $\text{sgn}^+(n)$ denotes the indicator function for nonnegative real numbers and $\text{sgn}^-(n) = 1 - \text{sgn}^+(n)$.

A member $x \in \mathbb{R}^2$ of the population $N^*(\mu, T, q, r)$ is constructed in the following way:

$$1) \quad \varepsilon \sim N(0, I_2) \text{ where } I_2 \text{ denotes the identity matrix.} \tag{4.30}$$

$$2) \quad \eta_i = [q_i \, \text{sgn}^+(\varepsilon_i) + r_i \, \text{sgn}^-(\varepsilon_i)]\,\varepsilon_i\,, \quad i = 1, 2. \tag{4.31}$$

$$3) \quad x = \mu + T\eta\,. \tag{4.32}$$

$Log(N^*(x \mid \mu, T, q, r))$ is consequently given as (up to an additive constant)

$$\log N^*(x \mid \mu, T, q, r) \propto -\sum_{i=1}^{2} [\log(q_i)\,\text{sgn}^+(\varepsilon_i) + \log(r_i)\,\text{sgn}^-(\varepsilon_i)] - \frac{\varepsilon'\varepsilon}{2}. \tag{4.33}$$

From (4.31) it follows that the spread of x around the mode μ can be adjusted by changing the parameters q_1, q_2, r_1 and r_2. Assume now that the following inequality holds:

$$\frac{p(x_j(\delta_i) \mid \textit{historical data})}{p(x_j(0) \mid \textit{historical data})} \leq \frac{N^*(x_j(\delta_i) \mid \mu, T, q, r)}{N^*(x_j(0) \mid \mu, T, q, r)} \quad \text{for } 0 < |\delta_i| \leq \Delta_{ij}, \tag{4.34}$$

that is, the rate of decline of $p(x_j(\delta_i) \mid \textit{historical data})$ is larger than the rate of decline of $N^*(x_j(\delta_i) \mid \mu, T, q, r)$ along the parameterized lines $x_j(\delta_i)$ for $0 < |\delta_i| \leq \Delta_{ij}$. It follows that

$$\begin{gathered} \frac{p(x_j(\delta_i) \mid \textit{historical data})}{N^*(x_j(\delta_i) \mid \mu, T, q, r)} \leq \frac{p(x_j(0) \mid \textit{historical data})}{N^*(x_j(0) \mid \mu, T, q, r)} \\ \Downarrow \\ w(x_j(\delta_i)) \leq w(x_j(0)) \quad \text{for } 0 < |\delta_i| \leq \Delta_{ij}. \end{gathered} \tag{4.35}$$

Consequently, the magnitudes of the weights $w(x_j(\delta_i))$ are bounded from above by $w(x_j(0))$ along the parameterized line. A convenient feature of the k-variate split normal distribution is that (4.35) can be satisfied on a parameterized line of finite length by a simple adjustment of the parameters q_1, q_2, r_1 and r_2. By appropriate choices of these

parameters we can thus avoid that points with low probabilities sampled along the parameterized lines are assigned extremely large weights.

To illustrate how suitable values of q_1, q_2, r_1 and r_2 can be determined in the present case, note from (4.30) - (4.33) that the ratios

$$\frac{N^*(x_j(\delta_1) \mid \mu, T, q, r)}{N^*(x_j(0) \mid \mu, T, q, r)} = e^{-\frac{1}{2}\left(\frac{\delta_1}{q_j}\right)^2}, \; j = 1,2; \tag{4.36}$$

and

$$\frac{N^*(x_j(\delta_2) \mid \mu, T, q, r)}{N^*(x_j(0) \mid \mu, T, q, r)} = e^{-\frac{1}{2}\left(\frac{\delta_2}{r_j}\right)^2}, \; j = 1,2; \tag{4.37}$$

where $x_j(\delta_i)$ is given by (4.29). If inequality (4.34) is satisfied for $0 < |\delta_i| \leq \Delta_{ij}$, it follows from (4.36) and (4.37) that

$$\frac{p(x_j(\delta_1) \mid historical\ data)}{p(x_j(0) \mid historical\ data)} \leq e^{-\frac{1}{2}\left(\frac{\delta_1}{q_j}\right)^2} \text{ for } j = 1,2$$

$$\Downarrow \tag{4.38}$$

$$q_j \geq \frac{|\delta_1|}{\sqrt{2 \log \dfrac{p(x_j(0) \mid historical\ data)}{p(x_j(\delta_1) \mid historical\ data)}}} \text{ for } 0 < \delta_1 \leq \Delta_{1j},$$

and

$$\frac{p(x_j(\delta_2) \mid historical\ data)}{p(x_j(0) \mid historical\ data)} \leq e^{-\frac{1}{2}\left(\frac{\delta_2}{r_j}\right)^2} \text{ for } j = 1,2$$

$$\Downarrow \tag{4.39}$$

$$r_j \geq \frac{|\delta_2|}{\sqrt{2 \log \dfrac{p(x_j(0) \mid historical\ data)}{p(x_j(\delta_2) \mid historical\ data)}}} \text{ for } -\Delta_{2j} \leq \delta_2 < 0.$$

If we define $f_j(\delta_i)$ as

$$f_j(\delta_i) = \frac{|\delta_i|}{\sqrt{2\log\dfrac{p(x_j(0)\mid historical\ data)}{p(x_j(\delta_i)\mid historical\ data)}}}, \qquad \textbf{(4.40)}$$

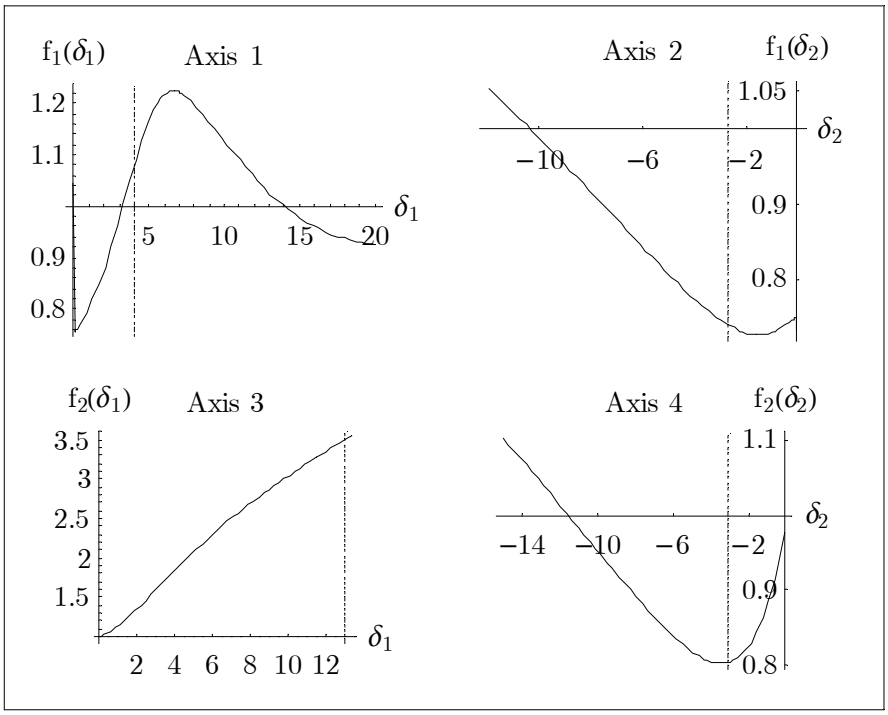
the constraints on q_j and r_j from (4.38) and (4.39) can be rephrased as

$$q_j \geq f_j(\delta_1) \quad \text{for } 0 < \delta_1 \leq \Delta_{1j} \text{ and } j = 1,2; \qquad \textbf{(4.41)}$$

$$r_j \geq f_j(\delta_2) \quad \text{for } -\Delta_{2j} \leq \delta_2 < 0 \text{ and } j = 1,2. \qquad \textbf{(4.42)}$$

Fig. 4.10 below illustrates the behaviour of $f_j(\delta_i)$ along the four axis. Regarding $f_1(\delta_1)$, for example, fig. 4.10 shows that if $q_1 = 1.22$, the corresponding weights $w(x_j(\delta_i)) \leq w(x_j(0))$ for $0 < \delta_1 \leq 20$.

Fig. 4.10. $f_j(\delta_i)$ for $i, j \in \{1,2\}$. Given that q_j or (r_j) is larger or equal to the maximum value of $f_j(\delta_i)$ within a given interval, the corresponding weights $w(x_j(\delta_i)) \leq w(x_j(0))$ within the same interval. The vertical dashed line in each plot indicates the size of δ_i at which the value of the posterior $p(x_j(\delta_i)\mid historical\ data)$ is 0.001 of its value at the mode .



Based on fig. 4.10 we will make the following assignments with respect to q_1, q_2, r_1 and r_2:

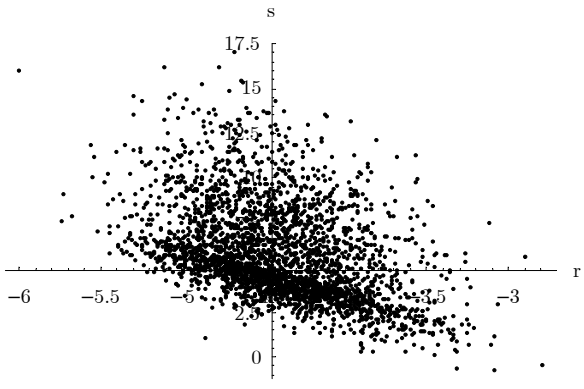
$$\begin{vmatrix} q_1 & r_1 \\ q_2 & r_2 \end{vmatrix} = \begin{vmatrix} 1.22 & 1.00 \\ 3.50 & 1.00 \end{vmatrix} \tag{4.43}$$

With the choice $r_1 = 1$, for example, it follows from fig. 4.10 that inequality (4.34) is violated if $\delta_2 < -10.39$. However, the value of the split normal density evaluated in a point belonging to this "area" is less than 10^{-23} relative to its value at the mode. In other words, the Monte Carlo importance sampling has to include several points if large weights are to show up due to sampling in this area. The most critical assignment is $q_2 = 3.50$. In this case inequality (4.34) is violated if $\delta_1 > 13.0$. At this point, the value of the split normal density is as large as 0.001 relative to its value at the mode.

Fig. 4.11 shows the distribution of 3000 points sampled from the 2-variate normal split density with q_1, q_2, r_1 and r_2 specified in (4.41) , $\mu = (-4.46, 4.89)$ and with T given as

$$T = \begin{vmatrix} 0.37 & 0 \\ -0.93 & 1 \end{vmatrix}. \tag{4.44}$$

Fig. 4.11. 3000 sampled points from $N^*(x \mid \mu, T, q, r)$, the 2-variate split normal density. $\mu = (-4.46, 4.89)$, the matrix T is defined in (4.42) and the vectors q and r are specified in (4.41). The axes intersect the mode of the posterior $p(r, s \mid \textit{historical data})$.

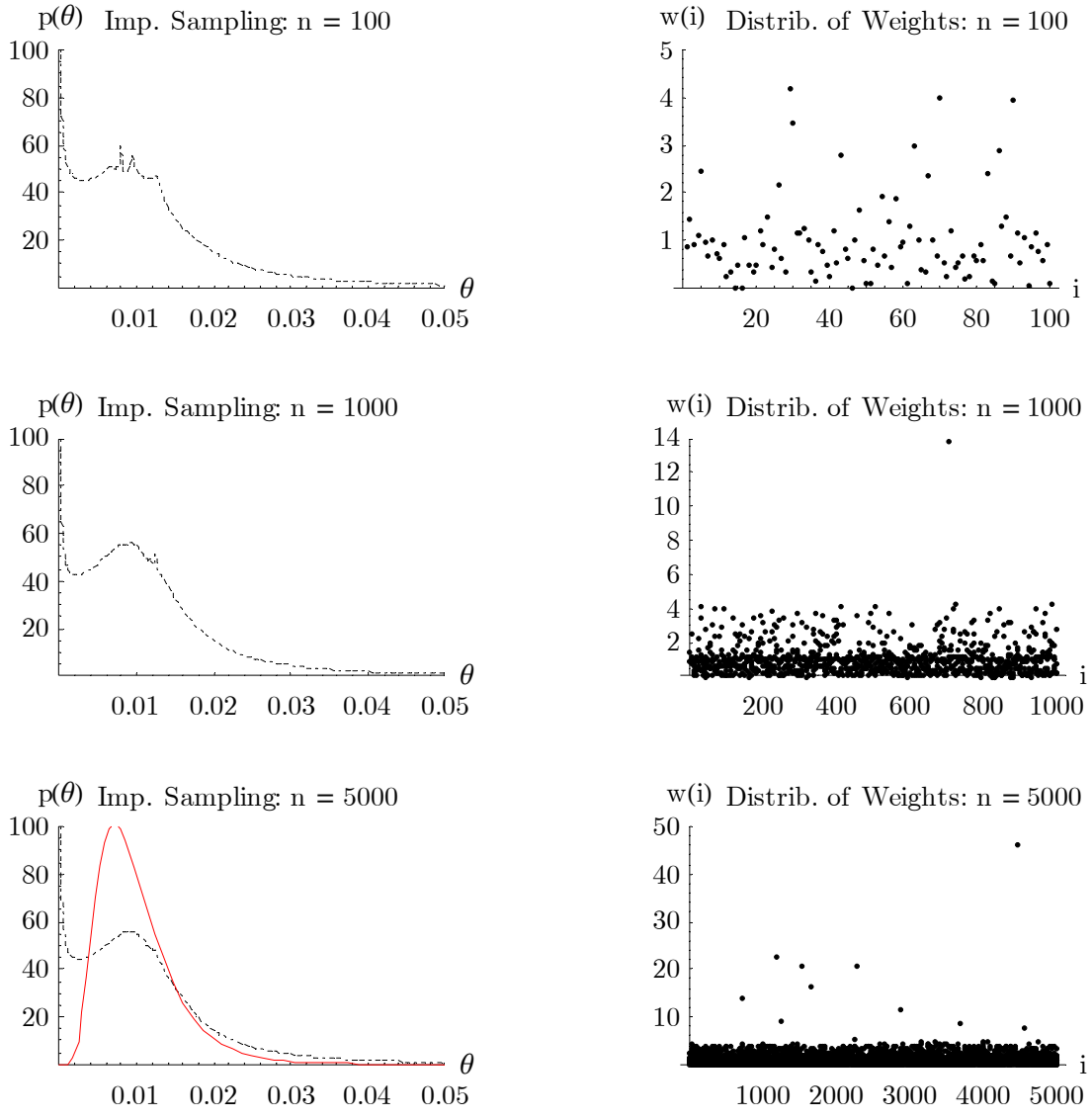


Using the split normal density as the importance sampling density, the following paragraph illustrates how the distribution of the binomial parameter θ can be estimated through Monte Carlo importance sampling.

4.5. Estimation of the Distribution of θ through Monte Carlo Importance Sampling

Having constructed a suitable importance sampling density, the estimation of the distribution of θ can be accomplished through Monte Carlo importance sampling as prescribed by (4.19). The three graphs positioned in the left column of fig. 4.12 show the estimate of $p(\theta)$ (based on the posterior $p(r,s \mid \textit{historical data})$) after 100, 1000 and 5000 sampled points, respectively. The plots positioned in the right column of fig. 4.12 illustrate the corresponding distribution of importance weights $w(r,s)$.

Fig. 4.12. Estimation of distribution of θ through Monte Carlo importance sampling. Sample points are obtained by sampling from the split normal density introduced in paragraph 4.4. n = number of sample points; red solid curve represents the exact distribution of θ according to chapter 3.



A summary of the properties of $p(\theta)$ can be found in table 4.04 below.

Table 4.04. Expected value and spread of θ according to the three estimates of $p(\theta)$ from fig 4.12. The last row "DATA" refers to the expected value and spread of the true distribution of θ, i.e., the distribution generated in chapter 3. *Max w* denotes the largest weight assigned to a sampled point during the Monte Carlo integration.

Sample Points	$E[\theta]$	$\sigma[\theta]$	Max w
100	0.0129	0.0139	4.2
1000	0.0121	0.0122	14.0
5000	0.0117	0.0118	46.4
DATA	0.0102	0.0053	-

It appears from fig. 4.12 that the Monte Carlo importance sampling method works well in the present case. After just 100 sampled points the broad features of $p(\theta)$ are established. More sample points are however needed to smooth out the crisps which appear in $p(\theta)$. The distribution of the importance weights displays a modest spread.

The essential point to be observed in fig. 4.12 is the approximate agreement between the estimate $p(\theta)$ obtained through Bayesian data analysis and the true distribution of θ (as defined in chapter 3). As expected, the spread of θ is overestimated considerably. Notable too is the limiting property that $p(\theta) \to \infty$ when $\theta \to 0$. This phenomenon can be traced back to components $Be(\theta \mid \alpha, \beta)$ (entering into the expression for $p(\theta)$) which have α-values being less than 1.

To investigate the sensitivity of $p(\theta)$ to the number of minefields included as *historical data*, two further studies were made including 25 and 100 minefields, respectively, following the approach outlined above. The data set consisting of 25 minefields was derived from the data set appearing in table 4.02 by discarding 25 minefields selected by random. The data set including 100 minefields was constructed by adding 50 new minefields (selected by random from the data set in chapter 3) to the 50 virtual minefields from table 4.02. The estimates of $p(\theta)$ are shown in fig. 4.13 and summarized in table 4.05. Based on a data set including just 25 minefields, the estimate of $E[\theta]$ is off by approximately 150% from the true value, as shown in table 4.05. The estimate of $E[\theta]$ based on information from 100 minefields is in the examined case essentially equal to the true average of θ.

Fig. 4.13. Left figure: $p(\theta)$ obtained from Monte Carlo importance sampling based on data set including 25 minefields. Right figure: Sampling based on data set including 100 minefields.

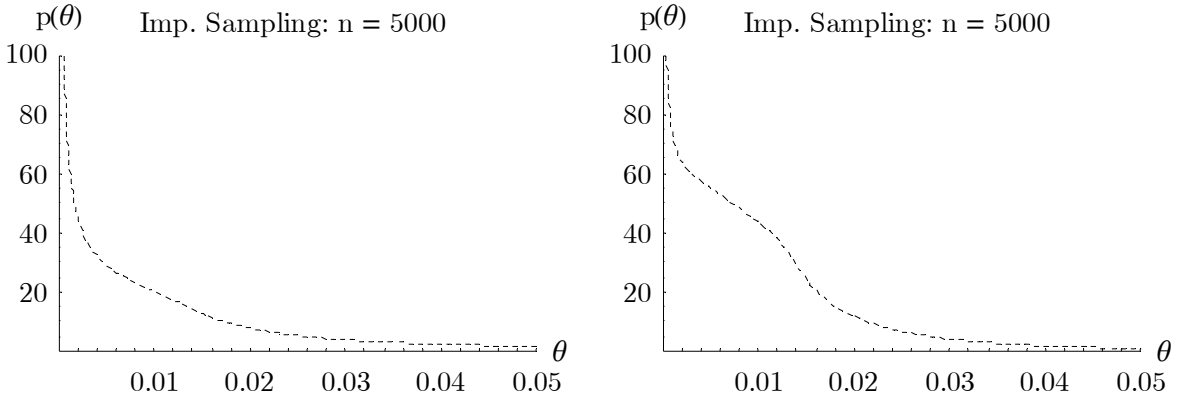


Fig.

Table 4.05. The sensitivity of the average and spread of θ calculated from the estimate $p(\theta)$ to different number of minefields included in the data set. The last row "DATA" refers to the expected value and spread of the true distribution of θ, i.e., the distribution generated in chapter 3.

Number of minefields	$E[\theta]$	$\sigma[\theta]$
25	0.0251	0.0878
50	0.0121	0.0122
100	0.0103	0.0115
DATA	0.0102	0.0053

4.6 Summary and Conclusion

Assessing the risk of a minefield through the risk model derived in chapter 2 presupposes estimates of the binomial parameters $\tilde{m}$ and θ characterizing the state of the minefield. In the present chapter we have shown how an estimate of the probability distribution of θ can be generated through Bayesian data analysis given that clearance data and accident statistics from former minefields similar to the minefield under study are available.

The main assumption underlying the model calculations in the present chapter is the hypothesis that the set of binomial parameters $\{\theta_1, \theta_2, ..., \theta_J\}$ associated with the former minefields are sampled from the same superpopulation, that is, a Beta-distribution characterized by the hyperparameters α and β. Taking the accident statistics from the

former minefields into consideration, a posterior distribution of α and β can subsequently be generated by use of Bayes' theorem. From this posterior distribution, an estimate of $p(\theta)$ can finally be provided through Monte Carlo importance sampling. The use of importance sampling as an integration technique entails the construction of a suitable importance sampling density. In the present context it has been found that the 2-variate split normal density makes up a flexible choice which seems to overcome some of the shortcomings displayed by simple multivariate normal distributions.

To keep the discussions simple, the data set applied in the present chapter has not involved explanatory variables. The inclusion of explanatory variables will however improve a decision makers ability to discriminate between various types of minefields, and explanatory variables should therefore be included if possible in any real-life application.

Chapter 5

Finite Mixture Models

5.1 Introduction

In chapter 4 it was shown how a probability distribution $p(\theta)$ could be generated through Bayesian data analysis by combining accident statistics and clearance data from mine clearance operations. The main theme below is how to estimate $p(\theta)$ if only accident statistics are available. That is, we will assume that the number of casualties caused by mines in minefield j within the last 2 years has been recorded for a total of M minefields, i.e. $j \in \{1,2,...,M\}$. The minefields under study have not been cleared yet, however, and detailed knowledge about the content of mines in the individual minefields is therefore lacking.

To make the above estimation problem computational approachable we will, like it was done in chapter 4, assume that the binomial parameters $\{\theta_1,\theta_2,...,\theta_M\}$ covering the M minefields are sampled from the same probability distribution. As already discussed in chapter 4 this is a perfectly valid assumption if we have no complementary information about the individual minefields. Mathematically, the above assumption can be expressed in various ways. In the present context it is convenient to introduce the auxiliary variable α_j defined as

$$\alpha_j = Log(\frac{\theta_j}{1-\theta_j}), \tag{5.1}$$

and let α_j follow a normal distribution as $\alpha_j \in \mathrm{R}$, i.e.

$$p(\alpha_j) = N(\alpha_j \mid \mu,\tau) \ \forall \ j. \tag{5.2}$$

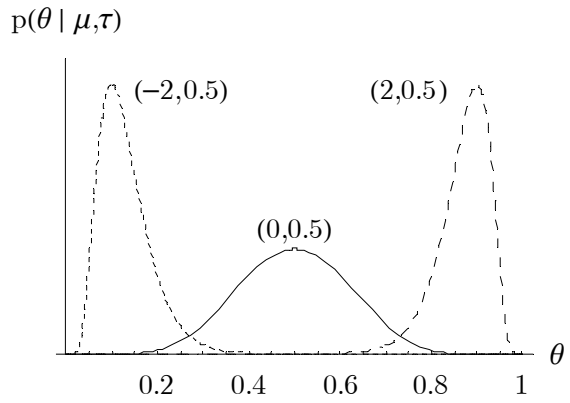
It follows that

$$p(\theta_j \mid \mu,\tau) = p(\alpha_j \mid \mu,\tau)\left|\frac{d\alpha_j}{d\theta_j}\right| = N(\alpha_j \mid \mu,\tau)\,4\,Cosh^2(\frac{\alpha_j}{2}). \tag{5.3}$$

From here on the index in θ_j will be omitted as θ_j follows the same distribution for all j.

Fig. 5.1 below illustrates $p(\theta \mid \mu, \tau)$ for different sets of (μ, τ). A more detailed account of (5.3) will be given in chapter 7.

Fig. 5.1 $p(\theta \mid \mu, \tau)$ calculated according to (5.3) for three different sets of (μ, τ).



Expression (5.2) can easily be modified if a set of explanatory variables $(x_1{}^j, x_2^j, ..., x_n^j)$ is attached each observation y_j. In this case (5.2) can be replaced by the expression

$$p(\alpha_j) = N(\alpha_j \mid \mu_0 + \beta_1 x_1^j + ...\beta_n x_n^j, \tau) \; \forall \; j. \qquad \textbf{(5.4)}$$

In other words, if two observations are ascribed the explanatory variables $(x_1, x_2, ..., x_n)$, the corresponding α's are by assumption sampled from the same normal distribution with average value $\mu_0 + \beta_1 x_1 + ...\beta_n x_n$ and variance τ.

To simplify the following discussions we will in the present report exclusively work with model (5.2). This implies that our initial estimation problem is reduced to the estimation of the normal distribution parameters (μ, τ). We are nevertheless still left with the problem that detailed knowledge about the degree of mine contamination in the individual minefield is lacking. A flexible type of statistical model which allows us to incorporate this uncertainty is the so-called *finite mixture model*.

5.2 Finite Mixture Models

According to the risk model derived in chapter 2 we may consider the observation y_j as the outcome of a stochastic process, where the random variable $Y_j \sim Bi(\tilde{m}_j, \theta_j)$ given that

$\tilde{m}_j > 0$. If $\tilde{m}_j = 0$ we obviously have that $p(Y_j = 0) = 1$. In the present context the parameter $\tilde{m}_j$ refers to the number of functional mines present in minefield j at time t = -1 (that is, 2 years ago). By use of (5.2) and (5.3) this can altogether be written as

$$p(y_j \mid \tilde{m}_j = m) = \begin{cases} I_0(y_j) \text{ if } m = 0 \\ \binom{m}{y_j} \frac{1}{\sqrt{2\pi}\tau} \int_{-\infty}^{\infty} \frac{\exp(\alpha y_j)}{(1+\exp(\alpha))^m} \exp\left(\frac{-(\alpha-\mu)^2}{2\tau^2}\right) d\alpha \text{ if } m > 0 \end{cases} \tag{5.5}$$

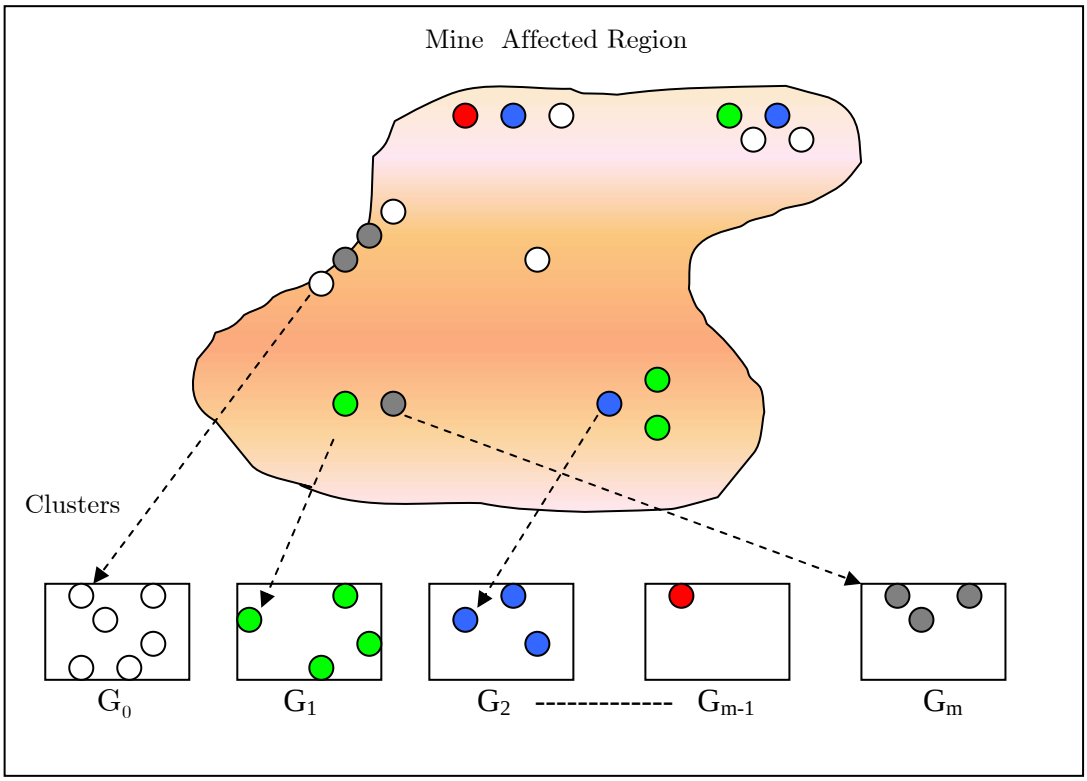
$$\equiv f(y_j \mid m, \mu, \tau),$$

where $I_0(y_j)$ denotes the indicator function defined as

$$I_0(y_j) = \begin{cases} 1 \text{ if } y_j = 0 \\ 0 \text{ else.} \end{cases} \tag{5.6}$$

Due to their different content of mines at t = -1 we may distribute the M minefields into say g clusters. This partitioning is sketched in fig. 5.2 where the cluster denoted G_m contains all minefields which contained m mines at t = -1.

Fig. 5.2. Partitioning of minefields into clusters conditioned on their content of mines at t = -1.



In the general case, the number of different clusters and the number of minefields belonging to each cluster will be unknown to a decision maker. One can, however, make a qualified guess. In a Bayesian framework such a guess can be made by the specification of a vector $\lambda = (\lambda_{m_1}, \lambda_{m_2}, ..., \lambda_{m_g})$ where $0 \leq \lambda_{m_i} \leq 1$ and

$$\sum_{i=1}^{g} \lambda_{m_i} = 1. \tag{5.7}$$

That is, the number of components in the vector λ denotes the believed number of clusters, and the magnitude of λ_{m_i} denotes the probability that a randomly selected minefield contains m_i functional mines at t = -1. Therefore $M \cdot \lambda_{m_i}$ denotes the expected number of minefields belonging to cluster G_{m_i}. Strictly speaking, the probabilities specified in the vector λ are *prior* probabilities in the sense that λ is stated prior to the realization of the outcomes (y_1, y_2, …, y_M).

As $p(\tilde{m}_j = m_i) = \lambda_{m_i}$ for all j it follows from (5.5) that $p(y_j \mid \mu, \tau, \lambda)$ can be written as

$$p(y_j \mid \mu, \tau, \lambda) = \sum_{i=1}^{g} \lambda_{m_i} f(y_j \mid m_i, \mu, \tau). \tag{5.8}$$

The likelihood function given by (5.8) makes up a special case of what might be termed a *finite mixture model*. The quantity λ_{m_i} in (5.8) is termed a *mixture parameter* or simply a *weight*, whereas the distribution $f(y \mid m, \mu, \tau)$ is termed a *mixture component*.

After the realization of the outcome y_j, the *posterior* probability $p(\tilde{m}_j = m_i \mid y_j, \mu, \tau, \lambda)$ is according to Bayes' rule given as

$$p(\tilde{m}_j = m_i \mid y_j, \mu, \tau, \lambda) = \frac{\lambda_{m_i} f(y_j \mid m_i, \mu, \tau)}{p(y_j \mid \mu, \tau, \lambda)}. \tag{5.9}$$

If we finally assume that the M random variables (Y_1, Y_2, …, Y_M) are *independent*, it follows from (5.8) that

$$p(y_1, y_2, ..., y_M \mid \mu, \tau, \lambda) = \prod_{j=1}^{M} p(y_j \mid \mu, \tau, \lambda). \tag{5.10}$$

The extension of (5.8) to the more general case where the individual observations are assigned explanatory variables $(x_1^j, x_2^j, ..., x_k^j) = x^j$ is straightforward. Following the notation from (5.4) we can thus replace (5.8) by the expression

$$p(y_j \mid x^j, \mu_0, \beta, \tau, \lambda) = \sum_{i=1}^{g} \lambda_{m_i} f(y_j \mid m_i, \mu_0 + \beta x^j, \tau), \qquad \textbf{(5.11)}$$

where $\beta = (\beta_1, \beta_2, ..., \beta_k)$. Assuming that the observations (y_1, y_2, ..., y_M) are conditionally independent, it follows from (5.11) that

$$p(y_1, y_2, ..., y_M \mid x^1, x^2, ..., x^M, \mu_0, \beta, \tau, \lambda) = \prod_{j=1}^{M} p(y_j \mid x^j, \mu_0, \beta, \tau, \lambda). \qquad \textbf{(5.12)}$$

In (5.11) the explanatory variables enter exclusively into the expression of the mixture components, that is, they are not informative about the mixture parameters λ_{m_i}. A final generalizing step is to make λ_{m_i} dependent on the explanatory variables as expressed in (5.13):

$$p(y_j \mid x^j, \mu_0, \beta, \tau, \lambda) = \sum_{i=1}^{g} \lambda_{m_i}(x^j, \lambda_{m_i}^0) f(y_j \mid m_i, \mu_0 + \beta x^j, \tau). \qquad \textbf{(5.13)}$$

In (5.13) the variable $\lambda_{m_i}^0$ is just a constant.

Equation (5.13) is a very flexible expression, and a posterior distribution $p(\theta \mid y, x)$ based on (5.13) can be determined through Bayesian data analysis if (5.13) is supplemented by prior distributions for the entering variables. In the present context, however, we will focus on the simple mixture model given by (5.8) to keep discussions simple. The generalizations of (5.8) shown above might therefore seem of minor relevance. They are, however, included to illustrate the large potential of finite mixture models in relation to mine action, and the utility of (5.11) and (5.13) should be tested in the future on real data sets to exploit this potential.

So focusing on the simple mixture model given by (5.8), let us recall that the quantity of primary interest in the present context is the posterior $p(\theta \mid y)$ which can be extracted

from (5.8) in the following way: First, $p(\mu,\tau,\lambda \mid y)$ is calculated by means of Bayes' rule, i.e.

$$\begin{aligned} p(\mu,\tau,\lambda \mid y) &\propto p(y \mid \mu,\tau,\lambda)p(\mu,\tau)p(\lambda) \\ &= \prod_{j=1}^{M} p(y_j \mid \mu,\tau,\lambda)p(\mu,\tau)p(\lambda), \end{aligned} \tag{5.14}$$

where $p(y_j \mid \mu,\tau,\lambda)$ is given by (5.8), and $p(\mu,\tau)$ and $p(\lambda)$ denote the prior distributions of (μ,τ) and λ, respectively. Thereafter $p(\theta \mid y)$ can be extracted from $p(\mu,\tau,\lambda \mid y)$ through the integrations

$$\boxed{\begin{aligned} p(\theta \mid y) &= \int_{-\infty}^{\infty}\int_{0}^{\infty} p(\theta \mid \mu,\tau)p(\mu,\tau \mid y)\, d\tau\, d\mu \\ &= \int_{-\infty}^{\infty}\int_{0}^{\infty} p(\theta \mid \mu,\tau)\int p(\mu,\tau,\lambda \mid y)\, d\lambda\, d\tau\, d\mu, \end{aligned}} \tag{5.15}$$

where $p(\theta \mid \mu,\tau)$ is given by (5.3), and

$$\int p(\mu,\tau,\lambda \mid y)\, d\lambda = \int\int \ldots \int p(\mu,\tau,\lambda_{m_1},\lambda_{m_2},\ldots,\lambda_{m_g} \mid y)\, d\lambda_{m_1} d\lambda_{m_2} \ldots d\lambda_{m_g}. \tag{5.16}$$

As it emerges from (5.15), it should be a simple matter to obtain $p(\theta \mid y)$ through a double integration if the marginal posterior density $p(\mu,\tau \mid y)$ can be provided. Unfortunately, there are many unclarified matters connected with the provision of $p(\mu,\tau \mid y)$. Each of these matters will be thoroughly discussed in the coming chapters, but to give a preliminary impression we will here touch on the major challenges.

First of all, to provide $p(\mu,\tau \mid y)$, prior distributions $p(\mu,\tau)$ and $p(\lambda)$ are needed as input in (5.14). Concerning the vector λ this includes a decision on the dimension of λ which reflects, as may be recalled, the number of minefield clusters underlying the accident statistics (y_1, y_2, …, y_M). Given that the dimension of λ has (somehow) been determined, one has next to decide on the set of integers $\{m_1, m_2, \ldots\}$ to be associated with the components of λ, where m_k signifies the mine content in a minefield which belongs to cluster G_k.

Having determined the dimension of λ and the associated integers $\{m_1, m_2, ...\}$, the next problem is to provide analytical expressions for the prior distributions $p(\mu, \tau)$ and $p(\lambda)$. Concerning $p(\lambda)$ it has become standard in mixture model calculations to assume that $\lambda \sim Dirichlet(\alpha_1, \alpha_2, ..., \alpha_g)$, i.e.

$$p(\lambda) \propto \prod_{i=1}^{g} \lambda_{m_i}^{\alpha_i - 1}. \qquad \textbf{(5.17)}$$

Concerning the normal distribution parameters μ and τ it seems unlikely that information will be available which allows the specification of a very informative prior. It is essential, however, to know the sensitivity of $p(\theta \mid y)$ to various choices of $p(\mu, \tau)$.

Apart from the challenges listed above which are all associated with the specification of prior distributions, the integral $f(y \mid m, \mu, \tau)$ from (5.5) poses in itself a problem in two ways: Firstly, $f(y \mid m, \mu, \tau)$ cannot be evaluated analytically which precludes the possibility of getting an analytical expression for $p(\mu, \tau \mid y)$. Alternatively one might sample from $p(\mu, \tau \mid y)$ through Markov Chain Monte Carlo simulation which is the choice made in the present work. Secondly, the outcome of the Markov Chain simulation process turns out to be very sensitive to the numerical accuracy of the evaluation of $f(y \mid m, \mu, \tau)$. A classical numerical integration formula such as a 20-point Gaussian quadrature formula cannot in general provide the demanded accuracy, and an improved numerical integration algorithm has therefore to be provided.

Each of the problems listed above will be discussed in the coming chapters, and various solutions will be suggested.

Before the closing of this introduction a brief comment will be given on the concept of *indicator variables* which is a computational convenient concept in relation to finite mixture models. An indicator variable is a label vector ζ_j associated each random variable Y_j indicating the component of origin of y_j. Put in another way, if we have g mixture components in (5.8), the associated label vector ζ_j contains g components for all j and

$$\zeta_{jk} = \begin{cases} 1, \textit{ if } y_j \text{ origins from the k'th mixture component} \\ 0, \text{ else.} \end{cases} \qquad \textbf{(5.18)}$$

The true values of the indicator variables $\zeta_1, \zeta_2, ..., \zeta_M$ are by assumption unknown and they are therefore treated as random variables. To clarify this, let $\varepsilon_j = (\varepsilon_{1j}, \varepsilon_{2j}, ..., \varepsilon_{gj})$ denote an outcome of the indicator ζ_j, i.e., only one of the components from ε_j are different from zero. As $p(\tilde{m}_j = m_i) = \lambda_{m_i}$ (the key assumption underlying model (5.8)), we have that

$$p(\zeta_j = \varepsilon_j) = \lambda_{m_1}^{\varepsilon_{1j}} \lambda_{m_2}^{\varepsilon_{2j}} ... \lambda_{m_g}^{\varepsilon_{gj}} \qquad \textbf{(5.19)}$$

from which it follows that $\zeta_j \sim Multinomial(1; \lambda_{m_1}, \lambda_{m_2}, ..., \lambda_{mg})$ for all j. As the Dirichlet distribution (se equation (5.17)) is the conjugate distribution to the Multinomial distribution, the introduction of indicator variables turns the conditioned posterior density of λ into a very simple form, which is very convenient in relation to Markov Chain Monte Carlo simulation (see chapter 6).

The concept of indicator variables will be used throughout the following chapters, and it implies technically that the posterior $p(\mu, \tau, \lambda \mid y)$ is replaced by the enlarged posterior $p(\mu, \tau, \lambda, \zeta \mid y)$. Further details will be given where it is found relevant in the following chapters.

From chapter 8 and further on several implementations of mixture model (5.08) including certain extended versions will be given. Before so it seems appropriate to discuss how sampling from the posterior $p(\mu, \tau, \lambda, \zeta \mid y)$ can be performed through Markov Chain Monte Carlo simulation.

Chapter 6

Markov Chain Monte Carlo Simulation

In a Bayesian context, the aim of doing a Markov Chain Monte Carlo simulation (MCMC) is to make samples from some posterior distribution $p(\phi \mid y)$, often referred to as the *target* distribution, in the correct *proportions*. There are different ways to construct a Markov Chain whose stationary distribution is equal to $p(\phi \mid y)$. In the *Metropolis-Hastings algorithm* [Hastings, 1970], which is a special kind of a Markov Chain simulation method, a sequence of draws $\{\phi^0, \phi^1, \phi^2, ...\}$ is generated in the following way:

Based on some initial value ϕ^0 which satisfies $p(\phi^0 \mid y) > 0$, a candidate point ϕ^* is drawn from a *proposal* distribution $J(\phi^* \mid \phi^0)$. The quotient r defined as

$$r = \frac{p(\phi^* \mid y) / J(\phi^* \mid \phi^0)}{p(\phi^0 \mid y) / J(\phi^0 \mid \phi^*)}, \qquad \textbf{(6.01)}$$

is subsequently calculated. Finally ϕ^1 is determined by the rule

$$\phi^1 = \begin{cases} \phi^* \text{ with probability min(r,1)} \\ \phi^0 \text{ else.} \end{cases} \qquad \textbf{(6.02)}$$

Under quite general conditions, which includes almost any choice of proposal distribution, it can be shown that a sequence of points $\{\phi^0, \phi^1, \phi^2, ...\}$ sampled as prescribed above in their distribution converges to the exact distribution $p(\phi \mid y)$. Further details about regularity conditions, choices of $J(\phi^* \mid \phi^0)$ and related technical matters, see [Gilks et al. 1996].

Typically, the parameter ϕ from the target distribution is a vector $\phi = (\phi_1, \phi_2, ..., \phi_g)$. Instead of updating the complete vector ϕ in a single step as sketched in (6.01) and (6.02) above, it is often more convenient to update the individual components of ϕ successively in g separate steps. More generally, ϕ can be partitioned into blocks of components of various dimension which are then updated one at a time. The above strategy might be

termed *single-component Metropolis-Hastings*. The single-component Metropolis-Hastings algorithm can be sketched as follows [Gilks et al., 1996, page 10]:

Let $\phi_{-i}^{t} = \{\phi_1^{t+1}, \phi_2^{t+1}, ..., \phi_{i-1}^{t+1}, \phi_{i+1}^{t}, ..., \phi_g^{t}\}$ denote the component vector $\phi \setminus \{\phi_i^t\}$ at iteration t+1 after i - 1 completed updating steps. A candidate point ϕ_i^* is sampled from a proposal distribution $J_i(\phi_i^* \mid \phi_i^t, \phi_{-i}^t)$, and the quotient r

$$r = \frac{p(\phi_i^* \mid y, \phi_{-i}^t) \,/\, J_i(\phi_i^* \mid \phi_i^t, \phi_{-i}^t)}{p(\phi_i^t \mid y, \phi_{-i}^t) \,/\, J_i(\phi_i^t \mid \phi_i^*, \phi_{-i}^t)}, \tag{6.03}$$

is subsequently calculated. Finally, ϕ_i^{t+1} is determined by the rule

$$\phi_i^{t+1} = \begin{cases} \phi_i^* \text{ with probability min(r,1)} \\ \phi_i^t \text{ else.} \end{cases} \tag{6.04}$$

Note that $p(\phi_i^* \mid y, \phi_{-i}^t)$ in (6.03) denotes the full conditional distribution of ϕ_i^*, i.e.

$$p(\phi_i^* \mid y, \phi_{-i}^t) = \frac{p(y, \phi_i^*, \phi_{-i}^t)}{\int p(y, \phi_i^*, \phi_{-i}^t)\, d\phi_i^*}. \tag{6.05}$$

Note furthermore that every component ϕ_i is assigned an individual proposal distribution $J_i(\phi_i^* \mid \phi_i^t, \phi_{-i}^t)$. If we, concerning component ϕ_i, make the particular choice

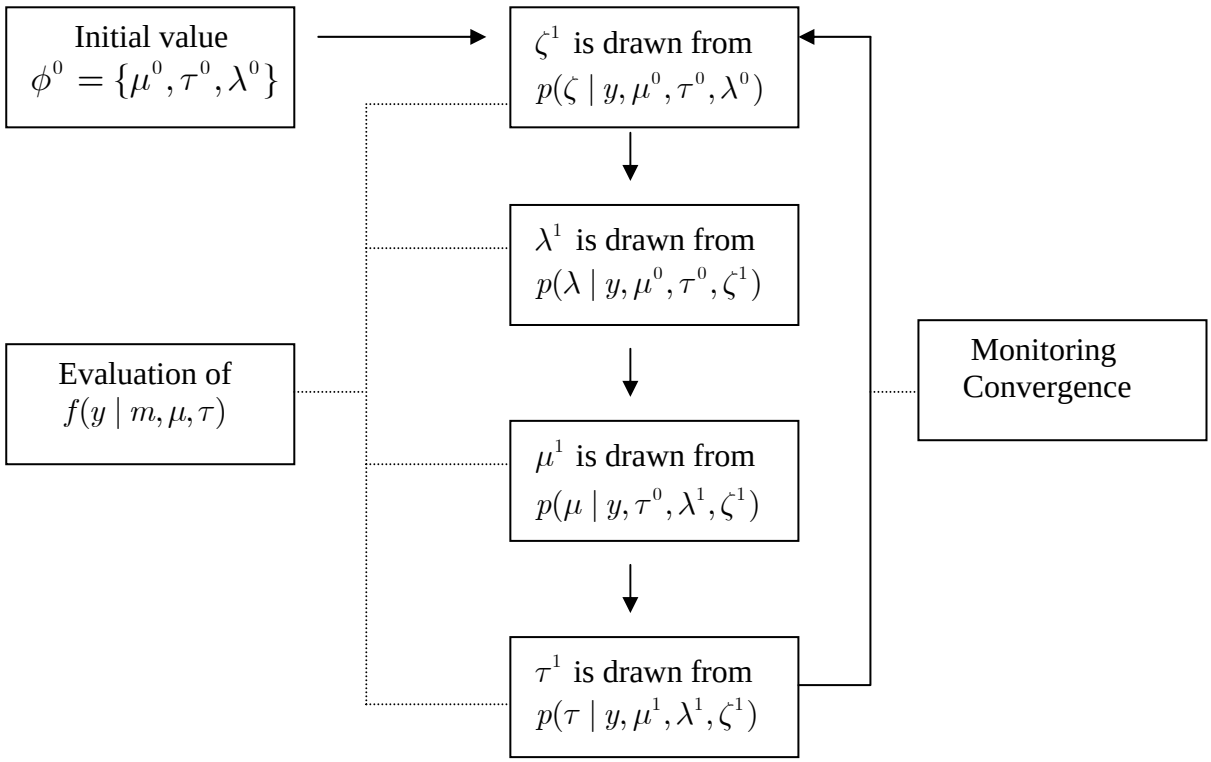
$$J_i(\phi_i^* \mid \phi_i^t, \phi_{-i}^t) = p(\phi_i^* \mid y, \phi_{-i}^t), \tag{6.06}$$

it follows from (6.03) and (6.04) that the candidate point ϕ_i^* is accepted with a probability of 1. The proposal distribution given by (6.06) is termed a *Gibbs sampler*. A particular simple situation arises if the Gibbs samplers for all i take the forms of simple standard distributions which are easy to sample from. In that case every iteration of the single-component Metropolis-Hastings algorithm can be carried out as a sequence of draws from standard distributions. If (6.06) is applied at one or more steps in the single-component Metropolis-Hastings algorithm, this is referred to as *Gibbs sampling*.

In many applications of the single-component Metropolis-Hastings algorithm some of the conditioned distributions derived from a given target distribution take the form of simple standard distributions whereas others do not. This turns out to be the case too if we look at the target distribution $p(\mu,\tau,\lambda,\zeta \mid y)$ introduced in chapter 5. Starting from $p(\mu,\tau,\lambda,\zeta \mid y)$ we can derive the four conditioned distributions $p(\zeta \mid y,\mu,\tau,\lambda)$, $p(\lambda \mid y,\zeta,\mu,\tau)$, $p(\mu \mid y,\zeta,\tau,\lambda)$ and $p(\tau \mid y,\zeta,\mu,\lambda)$. As shown in appendix A, the conditioned distributions $p(\zeta \mid y,\mu,\tau,\lambda)$ and $p(\lambda \mid y,\zeta,\mu,\tau)$ have analytical expressions which allow Gibbs sampling. This is however not the case concerning $p(\mu \mid y,\zeta,\tau,\lambda)$ and $p(\tau \mid y,\zeta,\mu,\lambda)$ which is due to the integral $f(y \mid m,\mu,\tau)$. Analytical expressions for all conditioned distributions can be found in appendix A.

Fig. 6.1 below sketches the sampling algorithm which has been used in the present work to sample from $p(\mu,\tau,\lambda,\zeta \mid y)$. Samples from the conditioned distributions $p(\zeta \mid y,\mu,\tau,\lambda)$ and $p(\lambda \mid y,\zeta,\mu,\tau)$ are obtained directly by Gibbs sampling whereas sampling from $p(\mu \mid y,\zeta,\tau,\lambda)$ and $p(\tau \mid y,\zeta,\mu,\lambda)$ are obtained using a normal distribution and a scaled inverse χ^2–distribution, respectively, as a proposal distribution. Further documentation can be found in appendix A.

Fig 6.1. Markov-chain simulation by single-component Metropolis-Hastings.



The successive samplings from the conditioned distributions constitute the core activity in the single-component Metropolis-Hastings algorithm, but as indicated in fig. 6.1 two additional components are required to initiate and terminate the Markov chain properly. Finally, a numerical integration formula is needed for the evaluation of $f(y \mid m, \mu, \tau)$.

To start the sampling algorithm, an initial vector $\phi^0 = \{\mu^0, \tau^0, \lambda^0\}$ is needed (it is not necessary to provide an initial indicator vector ζ^0). In principle, any vector will do, but to obtain a faster convergence of the Markov Chain we use as ϕ^0 a local maximum of $p(\mu, \tau, \lambda \mid y)$ added some noise. The so-called *EM-algorithm* (*Expected Maximization*) is used to generate a local maximum. A thorough introduction to the EM-algorithm can be found elsewhere [see for example Gelman et al., 2003, ch. 12].

Concerning the termination of the single-component Metropolis-Hastings algorithm, the algorithm can be stopped when a "sufficient" number of draws has been sampled from the converged Markov Chain. What turns out to be a sufficient number will depend on the quantities of interest to be summarized, e.g. modes, quantiles, test statistics or posterior probabilities; and the demanded accuracy of the quantities of interest.

A point of some controversy is the discussion about how to monitor the approximate convergence of the Markov Chain. In the present implementation we have chosen to use the *potential scale reduction factor* $\hat{R}$ as suggested by Gelman [Gelman et al., 1992]. In their approach m Markov Chain simulations are initiated from m overdispersed distributions. After the completion of $2n$ iterations in each chain, the first half of the sampled points from each chain is discarded, and for each scalar quantity ψ of interest the variances B and W defined as

$$B = \frac{n}{m-1} \sum_{j=1}^{m} (\overline{\psi}_{.j} - \overline{\psi}_{..})^2 \qquad \textbf{(6.07)}$$

$$W = \frac{1}{m} \sum_{j=1}^{m} s_j^2 \qquad \textbf{(6.08)}$$

are subsequently calculated, where $\overline{\psi}_{.j,}$ $\overline{\psi}_{..}$ and s_j^2 are defined as

$$\bar{\psi}_{\cdot j} = \frac{1}{n}\sum_{i=1}^{n}\psi_{ij}, \quad \bar{\psi}_{\cdot\cdot} = \frac{1}{m}\sum_{j=1}^{m}\bar{\psi}_{\cdot j}, \quad s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\psi_{ij} - \bar{\psi}_{\cdot j})^2. \tag{6.09}$$

From the expressions above it emerges that the factors B and W represent the *between-sequence* variance and the *within-sequence* variance, respectively.

Based on the factors B and W, the potential scale reduction factor $\hat{R}$ is defined as

$$\hat{R} = \sqrt{\frac{\widehat{\mathrm{var}}(\psi \mid y)}{W}}, \tag{6.10}$$

where

$$\widehat{\mathrm{var}}(\psi \mid y) = \frac{n-1}{n}W + \frac{1}{n}B \tag{6.11}$$

is an estimate of the marginal posterior variance $\mathrm{var}(\psi \mid y)$.

According to Gelman et al., the estimate $\widehat{\mathrm{var}}(\psi)$ represents an overestimate due to the overdispersed starting points, whereas W underestimates $\mathrm{var}(\psi)$ due to the finite length of the individual Markov Chain. As $n \to \infty$, both $\widehat{\mathrm{var}}(\psi)$ and W will approach $\mathrm{var}(\psi)$, and $\hat{R} \to 1$ according to (6.10). If the calculated value of $\hat{R}$ is high after the completion of $2n$ iterations, this seems to indicate that the sampling is far from convergence and improved inferences about ψ can be obtained by continued sampling until $\hat{R} \approx 1$.

In the Markov Chain simulations which are to be presented in the following chapters, each simulation starts with a prescribed number of iterations. The potential scale reduction factor $\hat{R}$ is subsequently calculated for each scalar of interest. If $\hat{R} \leq 1.1$ for all scalars, the sampling algorithm is closed down. Otherwise the sampling continues until $\hat{R} \leq 1.1$ for all scalars of interest.

Concerning the integral which enters into the expression for $f(y \mid m, \mu, \tau)$, the integral has to be evaluated several times during a Markov Chain simulation and in consequence it has to be evaluated fast. Unfortunately, the integration cannot be carried out analytically, and we therefore have to rely on numerical integration.

As the integral appearing in $f(y \mid m, \mu, \tau)$ can be rewritten as $\int g(t)\, e^{-t^2} dt$, it seems natural to use a quadrature formula such as a 20-point Gauss-Hermite quadrature to implement the numerical integration. However, preliminary tests have revealed that the evaluation of $f(y \mid m, \mu, \tau)$ by a 20-point Gaussian quadrature formula is subject to large errors for certain combinations of the parameters (m, y, μ, τ).

Various solutions to the above problem have been examined during the present project. Simply increasing the number of used interpolation points reduces the accuracy problem but does not eliminate it, and the speed of the integration algorithm is furthermore slowed down if every integral is to be evaluated by the summation of a very large but fixed number of terms. An adaptive integration algorithm where the number of included interpolation points varies with (m, y, μ, τ) appears as the required alternative.

In the Markov Chain simulations which are to be presented, the integral appearing in $f(y \mid m, \mu, \tau)$ has been evaluated by an adaptive integration algorithm founded on certain error bound analyses derived by Crouch et al. [Crouch et al., 1990]. The technical details behind the adaptive algorithm are not essential in the coming chapters, and the complete documentation is therefore deferred to chapter 13.

Chapter 7

Tests of Mixture Models

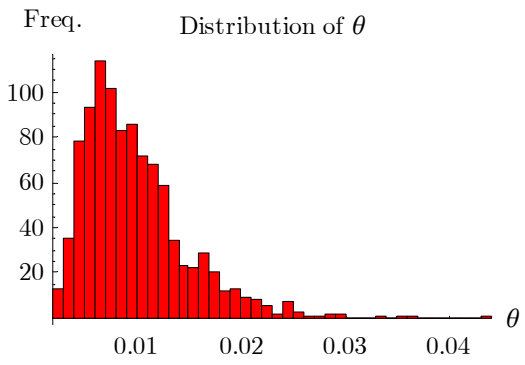
After having introduced the concept of finite mixture models; explained the basic assumptions underlying mixture models in the present context, and discussed various implementation issues, the following chapters will focus on various tests of the mixture model given by (5.8) and certain extensions of (5.8).

In what follows, we will envisage a hypothetical decision maker confronted with the accident statistics from table 7.1 (which were originally introduced in chapter 3). Thus table 7.1 covers accident statistics from 1000 virtual minefields. For completeness, the corresponding frequency distribution of θ for the 1000 minefields is shown in fig. 7.1.

Table 7.1. Simulated accident statistics from 1000 virtual minefields.

Number of observed casualties	*Number of minefields*
0	**887**
1	**81**
2	**19**
3	**7**
4	**2**
5	**2**
6	**2**
≥7	**0**

Fig 7.1. Frequency of θ for 1000 virtual virtual minefields.



Our hypothetical decision maker is assumed to be ignorant of the true underlying frequency distribution of θ depicted in fig. 7.1, but he wants to make statistical inferences about the distribution of θ through the application of the mixture model given by (5.8). To use model (5.8) within a Bayesian framework, the decision maker has to decide on four issues:

- The dimension of λ.
- The set of integers $\{m_1, m_2, ..., m_g\}$ to be associated with the components $(\lambda_{m_1}, \lambda_{m_2}, ..., \lambda_{m_g})$.
- The prior distribution $p(\lambda)$, where $\lambda \sim Dirichlet(\alpha_1, \alpha_2, ..., \alpha_g)$.
- The prior distribution $p(\mu, \tau)$.

A given specification of the above quantities makes up in combination with the mixture model (5.8) what might be termed a *discrete* model. There exists obviously an infinite number of different discrete models to choose from, and the decision maker's particular choice will reflect his level of knowledge about the minefields under study.

From the population of possible discrete models we will assume that the decision maker has selected a subset of k models indexed as say $\{H_1, H_2, ..., H_k\}$ to test on the accident statistics from table 7.1. Based on model H_i the posterior distribution $p(\mu, \tau, \lambda, \zeta \mid y, H_i)$ can be simulated through Markov Chain simulation, and it is now an issue of major importance to investigate the sensitivity of the posterior $p(\theta \mid y, H_i)$ derived from $p(\mu, \tau, \lambda, \zeta \mid y, H_i)$ to the particular model choice H_i. If $p(\theta \mid y, H_i)$ turns out to be sensitive to the choice of model, the decision maker needs analytical tools which enable him to evaluate and compare the predictive quality of the tested models. Based on such evaluations it may be possible to select a single best mode, or alternatively to combine the models into a supermodel $H = \omega_1 H_1 + \omega_2 H_2 + ... \omega_k H_k$, where the weights w_i are somehow derived from the model evaluations. Obviously, the above strategy is only profitable if the analytical tools chosen are able to differentiate among the tested models in terms of predictive quality.

The statistical literature on model checking and model comparisons is vast. One aspect of model checking is so-called posterior predictive checking [see for example Gelman et al., 2003], where a set of *replicated* data y^{rep} conditioned on model H_i is generated from the posterior distribution $p(\phi \mid y, H_i)$, ϕ denoting the vector of model parameters. The replicated data set y^{rep} can be sampled from the distribution

$$p(y^{rep} \mid y, H_i) = \int p(y^{rep} \mid \phi)\, p(\phi \mid y, H_i)\, d\phi. \qquad \textbf{(7.01)}$$

If model H_i fits, it is expected that y^{rep} should look similar to the original data y. To quantify the degree of similarity, *test quantities* $T(y, \phi)$ can be introduced which are scalar summaries of parameters and data. Given a test quantity $T(y, \phi)$ has been defined, a corresponding Bayesian p-value (equivalent to p-values in classical statistics) can be calculated as

$$p_B = p(T(y^{rep}, \phi) \geq T(y, \phi)). \tag{7.02}$$

That is, p_B is the probability that the test statistic based on the replicated data is more extreme than the corresponding test statistic based on the observed data. Formally, p_B under model H_i is calculated as

$$p_B = \int I_{T(y^{rep}, \phi) \geq T(y, \phi)} \; p(y^{rep} \mid y, H_i) \, dy^{rep}, \tag{7.03}$$

but in practice p_B is easily obtained as a by-product from the Markov Chain simulation.

Posterior predictive checking is primarily applied to check the fit of a single model. When comparing several models, a convenient measure termed the *deviance* [Nelder et al., 1972] is defined as minus two times the log-likelihood, i.e.

$$D(y, \phi) = -2 \log p(y \mid \phi), \tag{7.04}$$

and due to its connection with the *Kullback-Leibler information measure* it can be argued that the expected deviance $\hat{D}_{avg}(y)$ under model H_i defined as

$$\hat{D}_{avg}(y) = \frac{1}{L} \sum_{t=1}^{L} D(y, \phi_{H_i}^t), \tag{7.05}$$

is a reasonable measure of the overall fit of model H_i. In (7.05) the variable $\phi_{H_i}^t$ denotes a sample point from a Markov Chain simulation under model H_i.

A somewhat related measure of overall model fit is the *deviance information criterion* (*DIC*) defined as [Spiegelhalter et al. , 2002]

$$DIC = 2\hat{D}_{avg}(y) - D_{\hat{\phi}}(y), \tag{7.06}$$

where

$$D_{\hat{\phi}}(y) = D(y, \hat{\phi}(y)). \qquad \textbf{(7.07)}$$

In (7.07) $\hat{\phi}(y)$ denotes a point estimate of ϕ, for example the mean value of ϕ obtained through a Markov Chain simulation under model H_i.

The following chapters will give several examples of Bayesian p_B-values and deviances obtained under different mixture models. The purpose is twofold: Through the calculation of p_B-values it is revealed whether some or all of the proposed mixture models fail to reproduce certain aspects of the data set from table 7.1. More fundamentally p_B-values may reveal errors in the underlying programming code. Regarding the calculated deviances, it is essential to know whether deviance calculations can support a decision maker when the available information about the minefields under study does not clearly indicate a single best model.

Chapter 8

Preliminary Markov Chain Simulations

Listed in table 8.1 are four discrete models which, do to their very simple structure, will be referred to as "naïve" models during the following discussions. The four models may, if desired, be considered as a small set of competing models picked out by a decision maker for further investigation in relation to the accident statistics from table 7.1.

Table 8.1 Four naïve discrete models.

Model	Dimension of λ	Integers $\{m_1, m_2, ..., m_g\}$	$(\alpha_1, \alpha_2, ..., \alpha_g)$
H_1	11	{0,1,…,10}	(1,1,…,1)
H_2	21	{0,1,…,20}	(1,1,…,1)
H_3	31	{0,1,…,30}	(1,1,…,1)
H_4	41	{0,1,…,40}	(1,1,…,1)

Each model in table 8.1 is specified with respect to the dimension of λ, the integers $\{m_1, m_2, ..., m_g\}$, and the Dirichlet parameters $(\alpha_1, \alpha_2, ..., \alpha_g)$ which define $p(\lambda)$. Common to all models is the prior distribution $p(\mu, \tau) = p(\mu)\, p(\tau)$ specified in (8.01) and (8.02) below.

$$p(\mu) = \begin{cases} \text{constant, if } -10^{k_1} \le \mu \le 10^{k_1},\ k_1 \text{ being a large number} \\ 0 \text{ else,} \end{cases} \qquad \textbf{(8.01)}$$

$$p(\tau) = \begin{cases} \text{constant, if } 0 \le \tau \le 10^{k_2},\ k_2 \text{ being a large number} \\ 0 \text{ else.} \end{cases} \qquad \textbf{(8.02)}$$

In (8.01) and (8.02) the priors $p(\mu)$ and $p(\tau)$ are specified in terms of two for the time being large but undefined constants k_1 and k_2 which cut off the priors at faraway distances and therefore guarantee a proper posterior distribution for the mixture model (5.8). The prior $p(\theta)$ which results from (8.01) and (8.02) is given by

$$
\begin{aligned}
p(\theta) &= \int_{-\infty}^{\infty}\int_{0}^{\infty} p(\theta \mid \mu,\tau)p(\mu,\tau)\,d\mu\,d\tau \\
&\propto Cosh^2(\frac{\alpha}{2})\int_{-10^{k_1}}^{10^{k_1}}\int_{0}^{10^{k_2}} N(\alpha \mid \mu,\tau)\,d\mu\,d\tau \qquad \textbf{(8.03)} \\
&\approx Cosh^2(\frac{\alpha}{2}) = \theta^{-1}(1-\theta)^{-1},
\end{aligned}
$$

where the approximation in the last line of (8.03) can be justified on any closed interval $I \subset]0;1[$ through appropriate choices of k_1 and k_2. The $Beta(0,0)$-distribution is often referred to as a *non-informative* prior distribution.

Similarly, the assignment $\alpha_i = 1$ made in table 8.1 results in what might be termed a non-informative prior distribution for λ as equal density is assigned every λ satisfying the constraint $\sum_{i=1}^{g}\lambda_{m_i} = 1$. The only apparent difference between the four models in table 8.1 is thus the dimension of λ.

Fig 8.1. Marginal posteriors $p(\mu,\tau \mid y,H_i)$ for model H_1,H_2,H_3,H_4 from table 8.1 obtained from Markov chain simulation. Each cluster of points makes up the second half of 2000-2500 sampled points.

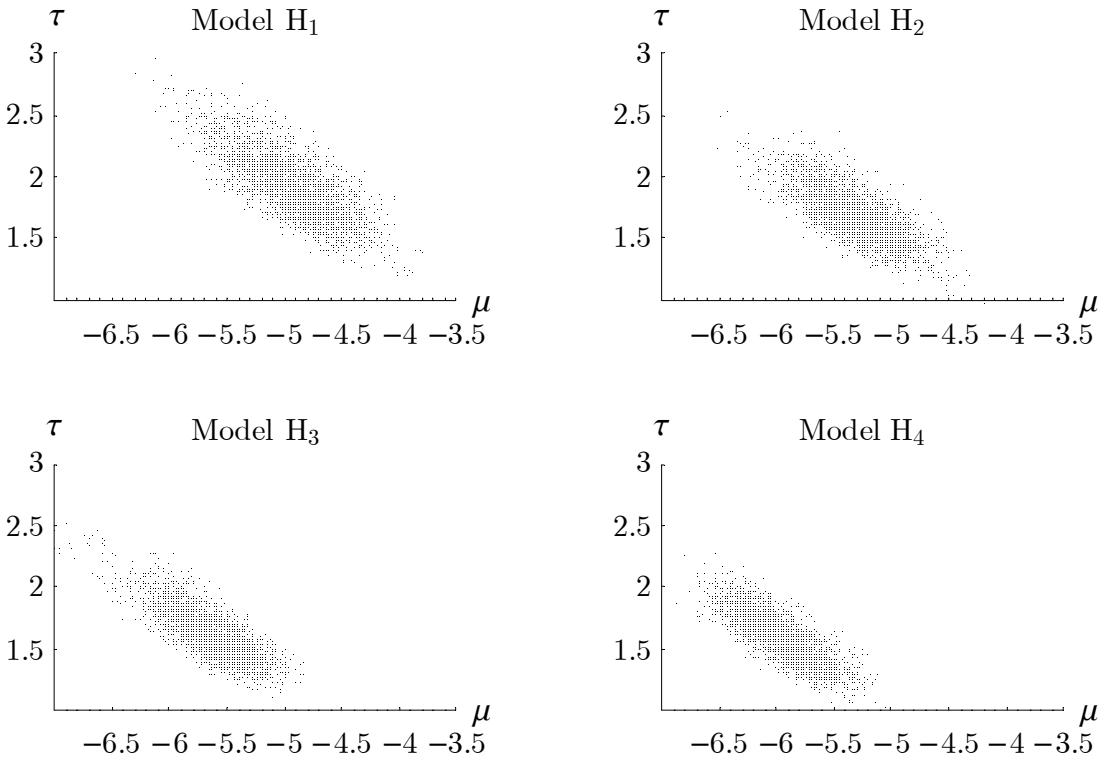
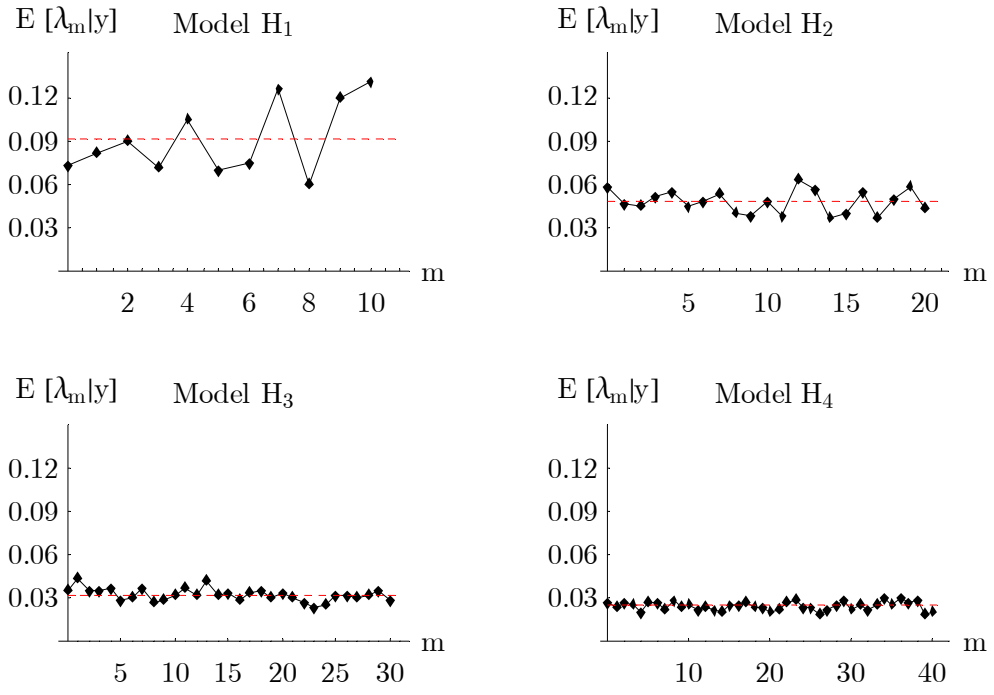


Fig. 8.1 (on the previous page) and fig. 8.2 and 8.3 below illustrate various features of the posteriors $p(\mu,\tau,\lambda,\zeta \mid y,H_i)$ for model H_1,H_2,H_3 and H_4 generated under mixture model (5.8) through Markov Chain simulation. Each sampling which includes 2000-2500 sampled points is based on the accident statistics from table 7.1 and the relevant prior distributions given in table 8.1, where $k_1 = 20$ and $k_2 = 50$. Depicted in fig. 8.1 are the marginal posterior distributions $p(\mu,\tau \mid y,H_i)$. Fig. 8.2 shows the posterior average value of the individual components of λ, and fig. 8.3 shows the posterior variance $Var[\lambda_m \mid y,H_i]$.

Fig. 8.2. Posterior average value of λ_m calculated from $p(\mu,\tau,\lambda,\zeta \mid y,H_i)$ for model H_1,H_2,H_3,H_4. As the number of components in the mixture model increases, the posterior average value of λ_m becomes confined to the vicinity of its prior expected value.

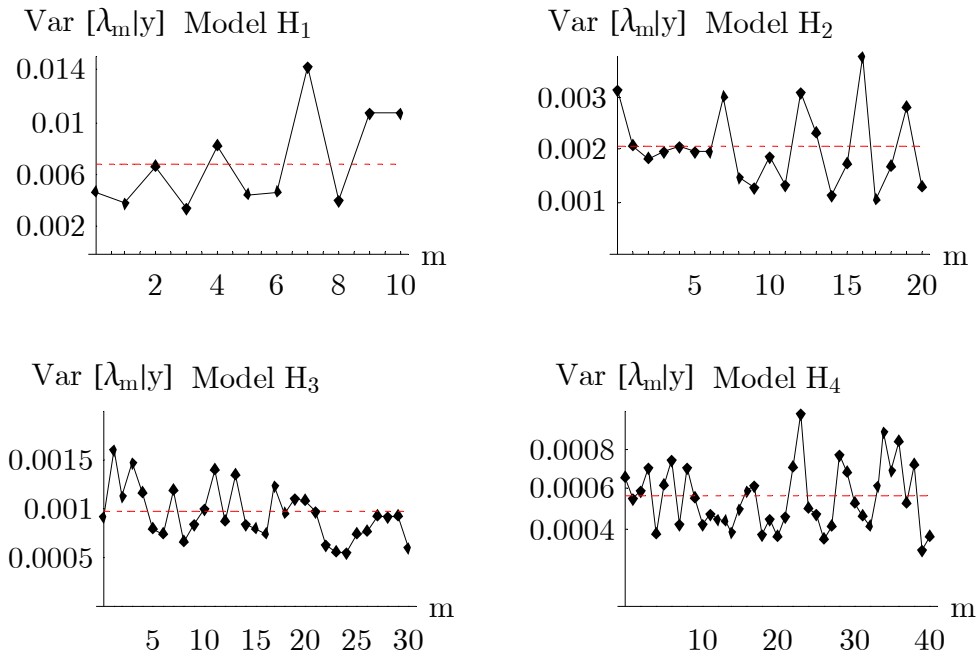


In fig. 8.2 the red dashed lines show the expected value $E[\lambda_m]$ according to the prior distribution $p(\lambda)$ which can be calculated as $E[\lambda_m] = \alpha_m / \alpha_{total}$, where $\alpha_{total} = \sum_{i=1}^{g} \alpha_i$.

Similarly, in fig. 8.3 the red dashed lines show the prior expected value $V[\lambda_m]$ which can be calculated as

$$Var(\lambda_m) = \frac{\alpha_m(\alpha_{total} - \alpha_m)}{\alpha_{total}^2(1+\alpha_{total})}. \qquad \textbf{(8.04)}$$

Fig. 8.3. Posterior variance of λ_m calculated from $p(\mu,\tau,\lambda,\zeta \mid y,H_i)$ for model H_1,H_2,H_3,H_4.



Two observations can be made from the above figures: Firstly, it is evident from fig. 8.2 that $E[\lambda_m \mid y] \approx E[\lambda_m]$ when the dimension of λ is large.

Secondly, from fig. 8.3 it is evident that the variance of the individual components of λ is not, on average, diminished when going from the prior distribution to the posterior distribution. Thus the marginal posterior distribution $p(\lambda \mid y)$ is mainly determined by the prior distribution $p(\lambda)$, i.e., the data from table 7.1 provide negligible information to the determination of λ.

The reason behind the above observations can be explained as follows: When the number of components in λ increases, the prior variance of λ_m decreases for all m according to (8.04), i.e., the prior distribution of λ_m becomes more localized. Consequently, more extreme observations are needed if the posterior distribution of λ_m is to be displaced substantially from its prior distribution. The observations from table 7.1 do not represent extreme observations, that is, most of the observations could origin from any of the used mixture components, and the outcome observed in fig. 8.2 and 8.3 follows.

The observations made above may also be used to explain the location and strong correlation between the sampled values of μ and τ which are observed in fig. 8.1. Due to the fact that $Var[\lambda_m] \to 0$ when the dimension of λ increases, it follows that λ essentially behaves as a deterministic vector when the dimension of λ is large. In other words, the finite mixture model given by (5.8) can be approximated by the simpler model

$$p(y_j \mid \mu, \tau, \lambda^*) = \sum_m \lambda_m^* f(y_j \mid m, \mu, \tau), \qquad \textbf{(8.05)}$$

where the parameter $\lambda_m^* \approx E[\lambda_m]$ is a *constant* fraction (as opposed to a stochastic variable). Even simpler, as no explanatory variables are attached to the individual observation y_j, we may tentatively consider the 1000 virtual minefields as *one* superminefield characterized by a total number of accidents $y_{total} = \sum_{j=1}^{1000} y_j$, and with a total mine content $m_{total} \approx 1000 \cdot \sum_m \lambda_m^* \cdot m$. If we consider $y_{total} \sim Bi(m_{total}, \theta)$ with $\theta \sim Beta(0,0)$ in accordance with (8.03), it follows from Bayes' rule that

$$\theta \mid y_{total} \sim Beta(y_{total}, m_{total} - y_{total}), \qquad \textbf{(8.06)}$$

and $E[\theta \mid y_{total}] = {}^{y_{total}}\!/_{m_{total}}$. From table 7.1 we have that $y_{total} = 170$. Table 8.2 below lists the calculated values of $E[\theta \mid y_{total}]$ based on the information contained in table 8.1.

.

Table 8.2. Expected value of θ calculated from model (8.06). m_{total} denotes the expected number of mines covering all minefields estimated from the prior distribution. y_{total} denotes the total number of accidents recorded from all minefields.

Model	m_{total}	$E[\theta \mid y_{total}]$
H_1	5000	0.0340
H_2	10000	0.0170
H_3	15000	0.0113
H_4	20000	0.0085

To establish the relationship between the results from table 8.2 and the locations of the sampled values of μ and τ as observed in fig. 8.1, recall that the probability parameter θ is connected to the binomial parameters μ, τ through the transformation

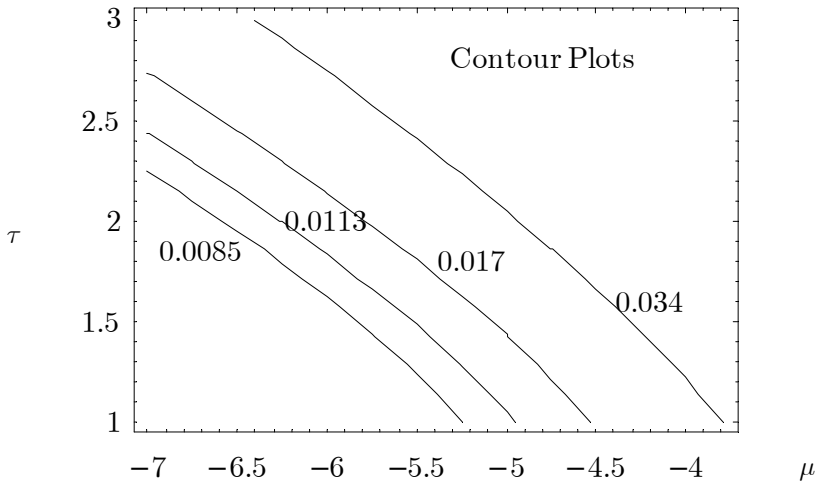
$$p(\theta \mid \mu,\tau) = p(\alpha \mid \mu,\tau)\left|\frac{d\alpha}{d\theta}\right| = N(\alpha \mid \mu,\tau)\,4\,Cosh^2(\frac{\alpha}{2}), \qquad \textbf{(8.07)}$$

where $\alpha = \log\frac{\theta}{1-\theta}$. The average value of $\theta \mid \mu,\tau$ is by definition calculated as

$$E[\theta \mid \mu,\tau] = \int_0^1 p(\theta \mid \mu,\tau)\,\theta\,d\theta. \qquad \textbf{(8.08)}$$

An infinite number of pairs (μ,τ) give rise to distributions $p(\theta \mid \mu,\tau)$ whose average value of θ are identical. This is illustrated in fig. 8.4, where all points (μ,τ) belonging to the same contour line give rise to the same expected value of θ. The four contour lines included in fig. 8.4 correspond to the four values of $E[\theta \mid y_{total}]$ from table 8.2.

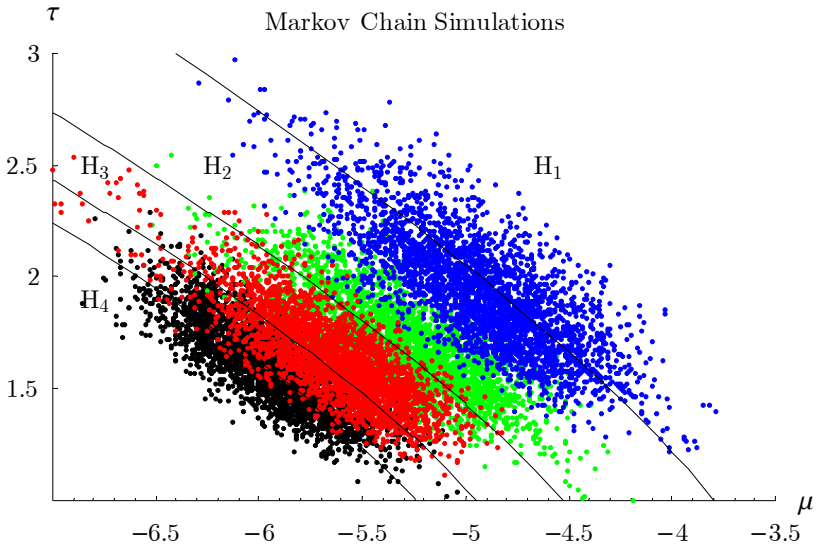
Fig. 8.4 Contour lines. Points (μ,τ) located on the same contour line represent distributions $p(\theta \mid \mu,\tau)$ whose average value of θ are identical. The real numbers indicate the magnitude of $E[\theta]$ for each contour.



In fig. 8.5 on the following page the four Markov Chain simulation plots from fig. 8.1 have been assembled into a single plot upon which the above contour lines are superimposed. The four contour lines nicely traverse the centres of the clusters.

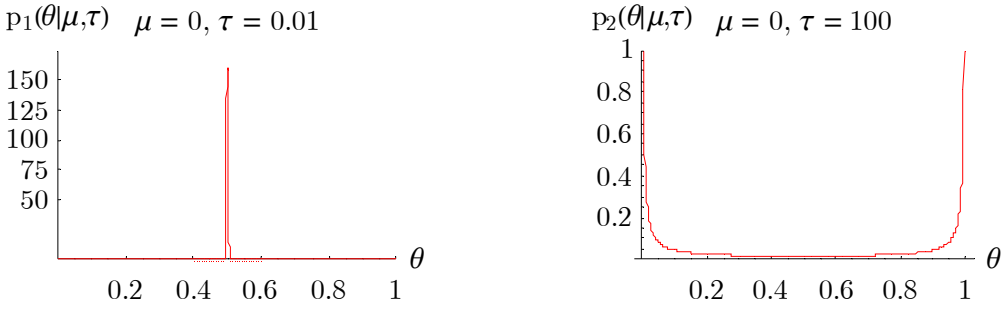
Fig. 8.5 illustrates how to explain the locations of the sampled points. Taking the blue cluster originating from model H_1 as an example, the spread of the sampled points perpendicular to the belonging contour line reflects the uncertainty about $E[\theta]$. Similarly, the spread of the sampled points along a given contour line reflects the uncertainty about the variance of θ for fixed $E[\theta]$.

Fig. 8.5. Contour lines corresponding to $E[\theta] = 0.0085,\ 0.0113,\ 0.017$ and 0.034 obtained from table 8.2 superimposed on the four marginal posteriors $p(\mu,\tau \mid y, H_i)$ from fig. 8.1.



That two different points located on the same contour line are different with respect to the variance of θ is illustrated in fig. 8.6 by a rather extreme example. Both distributions included in fig. 8.6 belong to the contour line characterized by $E[\theta] = 0.5$, but they obviously deviate from each other with respect to the variance of θ.

Fig.8.6. Two distributions characterized by $E[\theta] = 0.5$ but with different variances.



The quantities $E[\theta \mid \mu,\tau]$ and $Var[\theta \mid \mu,\tau]$ can easily be calculated. Simple manipulations show that

$$E[\theta \mid \mu,\tau] = \int_0^1 p(\theta \mid \mu,\tau)\,\theta\, d\theta. \qquad \textbf{(8.09)}$$
$$= \int_{-\infty}^{\infty} N(\alpha \mid \mu,\tau)\, 4\, Cosh^2\left(\frac{\alpha}{2}\right) \frac{e^{2\alpha}}{(1+e^{\alpha})^3}\, d\alpha = f(1 \mid 1,\mu,\tau).$$

More generally it can be shown that

$$E[\theta^n \mid \mu,\tau] = f(n \mid n,\mu,\tau), \tag{8.10}$$

from which it follows that $Var[\theta \mid \mu,\tau] = f(2 \mid 2,\mu,\tau) - f(1 \mid 1,\mu,\tau)^2$.

Due to (8.10), the distributions of $E[\theta]$ and $Var[\theta]$ can easily be extracted from the points (μ,τ) sampled under the four naïve models. The frequency distributions of $E[\theta]$ and $Var[\theta]$ are shown in fig. 8.7 and fig. 8.8. A summary of the findings are given in table 8.3.

Table 8.3. Summary of Markov chain simulations. $E[\theta]^*$ denotes the averages from table 8.2.

Model	$E[E[\theta]]$	$E[\theta]^*$	$\sigma[E[\theta]]$	$\frac{\sigma[E[\theta]]}{E[E[\theta]]}$	$E[Var[\theta]]$
H_1	0.033	0.034	0.0063	0.18	0.0058
H_2	0.018	0.017	0.0035	0.10	0.0019
H_3	0.012	0.0113	0.0021	0.079	0.00097
H_4	0.0089	0.0085	0.0015	0.060	0.00053

Fig. 8.7. The frequency distribution of $E[\theta]$ obtained from the collection of points (μ,τ) sampled under model H_1, H_2, H_3 and H_4. Each point (μ,τ) represents a distribution $p(\theta \mid \mu,\tau)$ where $E[\theta] = f(1 \mid 1,\mu,\tau)$.

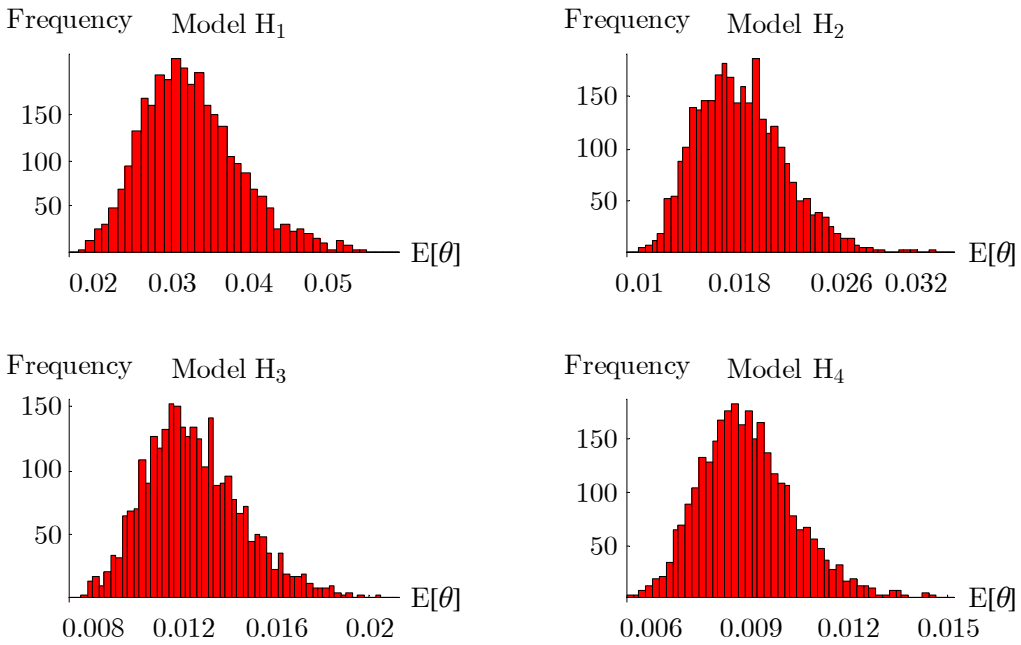


Fig. 8.8. The frequency distribution of $Var[\theta]$ obtained from the collection of points (μ,τ) sampled under model H_1, H_2, H_3 and H_4. Each sampled point (μ,τ) represents a distribution $p(\theta \mid \mu,\tau)$ where $Var[\theta] = f(2 \mid 2,\mu,\tau) - f(1 \mid 1,\mu,\tau)^2$.

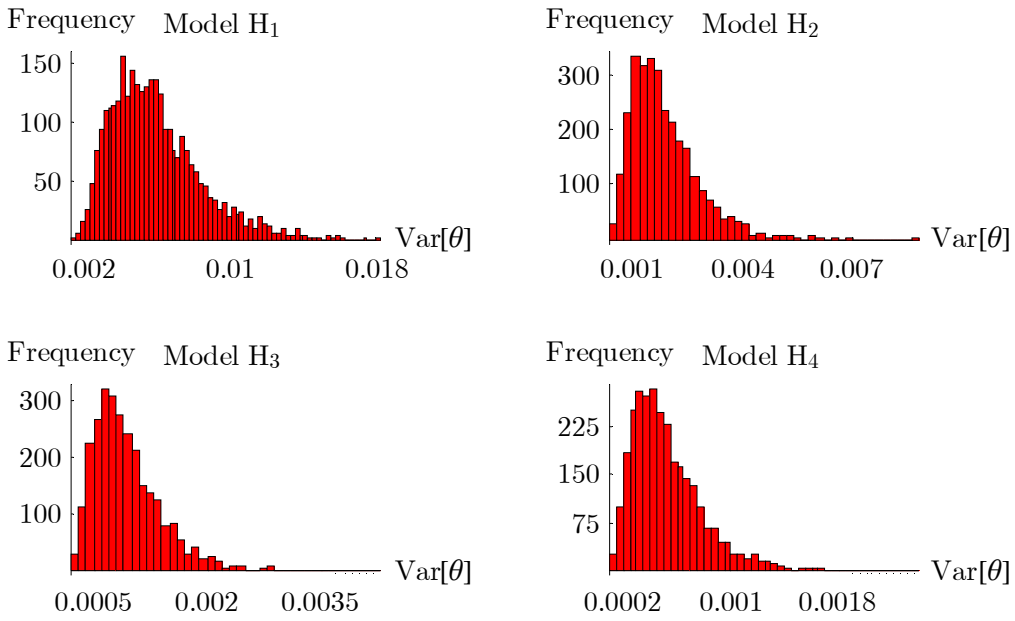
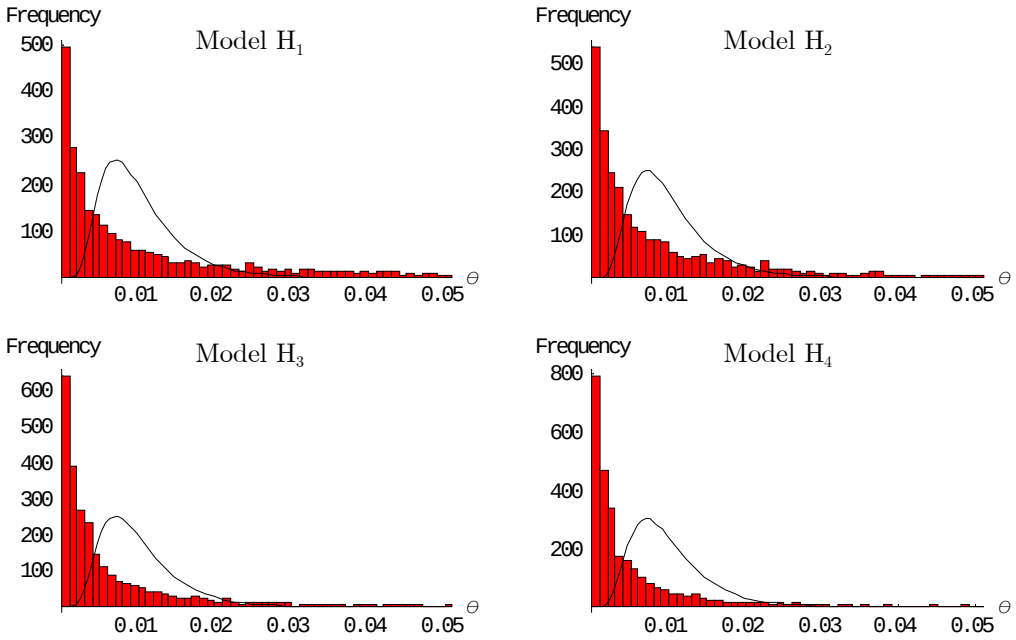


Fig. 8.9. Simulation of $p(\theta \mid y)$. Each histogram is generated from the corresponding Markov-chain simulation depicted in fig. 8.1. Superimposed each histogram is the distribution $p(\theta \mid \mu,\tau) = p(\theta \mid -4.7, 0.5)$ which generated the true frequency distribution of θ shown in fig. 7.1.



The complete posterior distribution $p(\theta \mid y)$ can be obtained through simulation as shown in fig. 8.9 above. The frequency distributions were generated as follows: For every sampled point (μ,τ) from $p(\mu,\tau,\lambda,\zeta \mid y,H_i)$, α was subsequently drawn from the normal distribution $N(\alpha \mid \mu,\tau)$, and θ was calculated as $\theta = e^{\alpha}(1+e^{\alpha})^{-1}$. Note that the θ-axis in fig. 8.9 is cut off at $\theta = 0.05$ as the frequency is practically zero for larger values of θ. Table 8.4 sums up essential descriptors for the generated distributions from fig. 8.9. For clarity the posterior intervals included in table 8.4 are sketched in fig. 8.10.

Table 8.4. Numerical summaries of the posterior distribution $p(\theta \mid y)$. Listed values are based on the histogrammes in fig. 8.9. Values in the last row "DATA" are derived from the distribution $p(\theta \mid \mu,\tau) = p(\theta \mid -4.7, 0.5)$ which generated the true frequency distribution of θ shown in fig. 7.1.

Model	$E[\theta]$	$Var[\theta]$	50% posterior interval for θ	95% posterior interval for θ
H_1	0.032	0.0059	[0.0019,0.026]	[0.00012,0.24]
H_2	0.017	0.0015	[0.0015,0.015]	[0.00012,0.12]
H_3	0.012	0.00078	[0.0011,0.011]	[0.00010,0.080]
H_4	0.0090	0.00053	[0.00092,0.0079]	[0.000095,0.054]
DATA	0.010	0.000028	[0.0064,0.012]	[0.0034,0.023]

Fig. 8.10. Location of 50% and 95% of the posterior density of $p(\theta \mid y)$. In fig. 8.10.a, 25% of the posterior density is located to the left and to the right, respectively, of the horizontal line under a given model. In fig. 8.10.b, 2.5% of the posterior density is located to the left and to the right, respectively, of the horizontal line under a given model. The posterior interval "DATA" is derived from the distribution $p(\theta \mid \mu,\tau) = p(\theta \mid -4.7, 0.5)$ which generated the true frequency distribution of θ shown in fig. 7.1.

Fig. 8.10.a

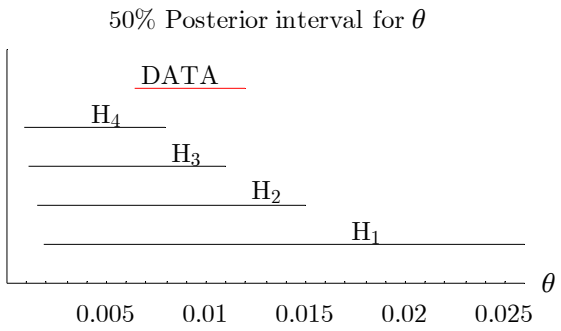
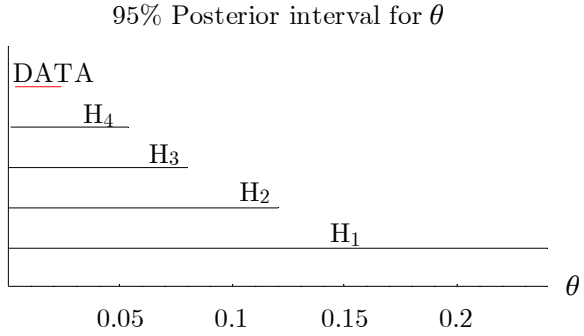
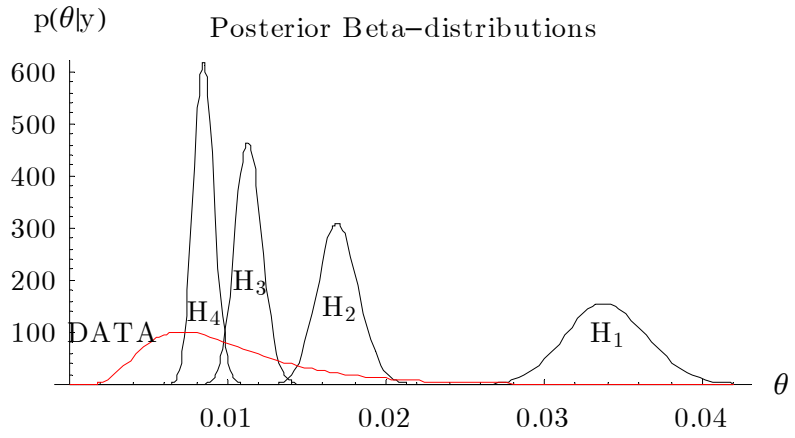


Fig. 8.10.b



In table 8.4 it is found that $E[\theta]$ calculated from the Markov chain simulations and the corresponding $E[\theta]^*$ estimated form the simpler superminefield model are by and large identical. This does not imply that the superminefield model can replace the finite mixture model as the superminefield model underestimates the true variance of θ considerably. This is illustrated in fig. 8.11 where the posterior distribution $Beta(\theta \mid y_{total}, m_{total} - y_{total})$ calculated under each of the naïve models is shown together with the true probability distribution of θ.

Fig. 8.11. The posterior distribution of θ calculated from $Beta(\theta \mid y_{total}, m_{total} - y_{total})$ under the superminefield model. m_{total} denotes the expected number of mines covering all minefields. y_{total} denotes the total number of accidents recorded from all minefields. Red curve: The true probability distribution of θ. The Beta distribution underestimate the true variance of θ considerably.



Another reason for not replacing the finite mixture model with the superminefield model is that to fully exploit the observations y_j one should in a real-life application supplement these by explanatory variables $(x_1^j, x_2^j, ..., x_k^j) = x^j$. As previously discussed, the availability of explanatory variables paves the way for the exploitation of the more advanced mixture models given by (5.11) and (5.13). However, the inclusion of explanatory variables rules out the possibility of merging of the considered minefields into one big minefield, which is the prerequisite for the use of superminefield model.

In all models examined so far, non-informative prior distributions have been used for $p(\mu, \tau)$ and $p(\lambda)$, the only information contained in $p(\lambda)$ being the dimension of the vector λ. By changing the dimension of λ a decision maker obviously changes the number of mixture components to be included in the mixture model, but he also changes the prior

variance of the individual components of λ if $\lambda \sim Dirichlet(1,1,...,1)$. This effect is seen, for example, in the fifth column of table 8.3 where the quantity $\sigma[E[\theta]]/E[E[\theta]]$ decreases for increasing values of the dimension of λ.

The essence of the above observation is that despite of having used non-informative priors, a lot of information may unintentionally be conveyed to the mixture model through the fixing of the dimension of λ. For example, by setting the dimension of λ to a large value the individual components of λ get essentially locked to their prior expected values $E[\lambda_m]$, and one ends up with a mixture model having only two free parameters, i.e. μ, τ. This property of the mixture model is unfortunate and calls for modifications. Various suggestions which may reduce the above defects will be discussed and tested in the coming chapters.

Chapter 9

Model Checking, Model Comparisons and Evaluation of Naïve Models

9.1 Model Checking and Model Comparisons

In a real-life situation, the true distribution of θ will be unknown to a decision maker, and the question is whether the techniques of model checking and model comparisons as discussed in chapter 7 are able to reveal which one among a group of competing models approximates the true distribution of θ best. Alternatively, the decision maker may through model comparisons be able to derive model weights with the purpose of expressing an estimate of $p(\theta)$ as a weighted average of the posteriors generated from the competing set of models.

In chapter 7 two approaches were discussed in relation to model checking and model comparisons. The first approach involved the calculation of posterior predictive distributions and related test statistics, whereas the second approach dealt with the concept of deviance as a measure of predictive accuracy. In what follows, the results from the calculations of various test statistics, deviances and related measures of predictive accuracy will be presented.

Table 9.1 below show the results from the calculation of Bayesian p_B-values based on five different test statistics. Fig. 9.1-9.5 on the following pages show the corresponding frequency distributions of the five test statistics.

Table 9.1. Test statistics and Bayesian p_B-values.

Test Statistic	$T(y_{rep})$	$T(y)$	Bayesian p_B-value			
			H_1	H_2	H_3	H_4
T_1	$\sum_j y_{rep,j}$	170	0.53	0.56	0.55	0.55
T_2	$Var(y_{rep})$	0.345	0.52	0.60	0.62	0.61
T_3	$Max(y_{rep})$	6	0.75	0.84	0.86	0.85
T_4	$\frac{\#\, y_j \in y : y_j = 0}{1000}$	0.887	0.54	0.49	0.51	0.49
T_5	99% quantile	3	0.73	0.69	0.66	0.65

Fig. 9.1. The frequency distribution of the test statistics $T_1 = \sum_{j=1}^{j=1000} y_{rep,j}$ under the naïve models H_1, H_2, H_3 and H_4. $y_{total} = 170$ is marked by the vertical black tab.

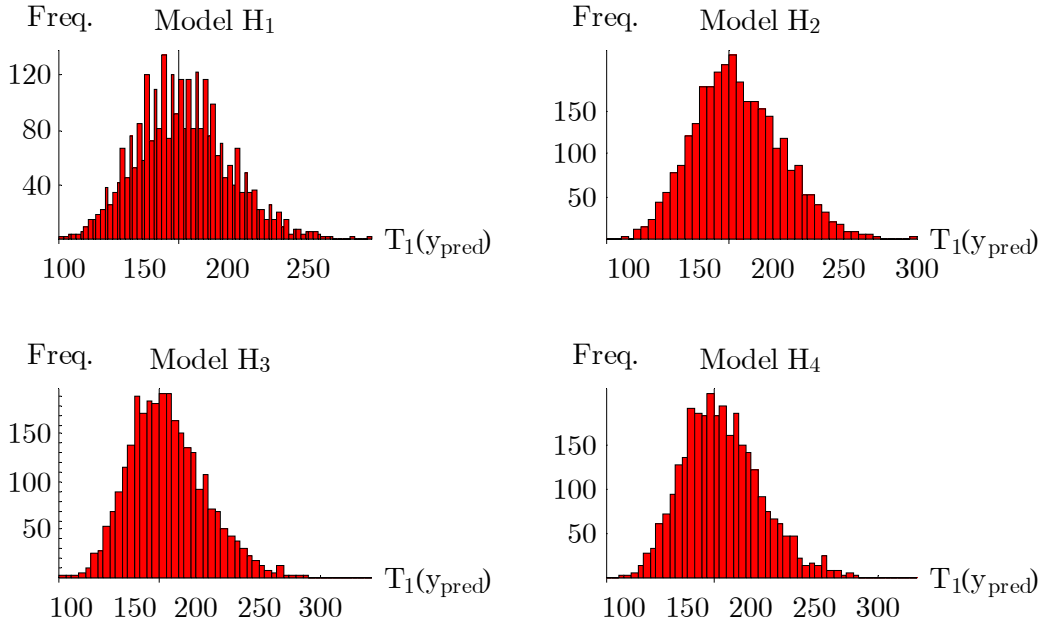


Fig. 9.2. The frequency distribution of the test statistic $T_2 = Var(y_{rep})$ under the naïve models H_1, H_2, H_3 and H_4. $Var(y) = 0.345$ is marked by the vertical black tab.

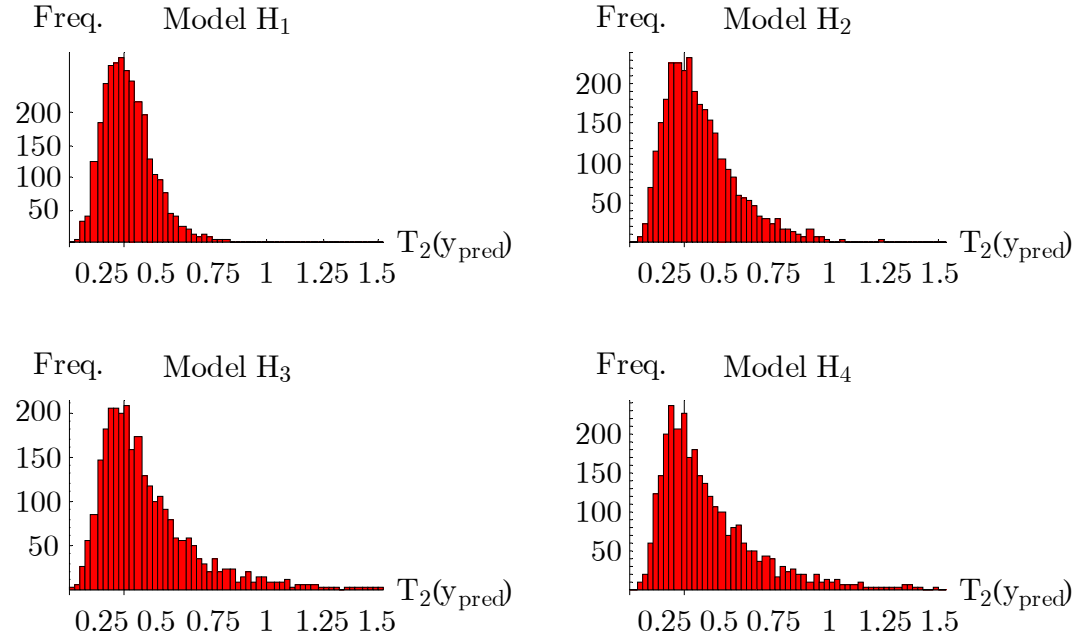


Fig. 9.3. The frequency distribution of the test statistic $T_3 = Max(y_{rep})$ under the naïve models H_1, H_2, H_3 and H_4. $y_{\max} = 6$ is marked by the vertical black tab.

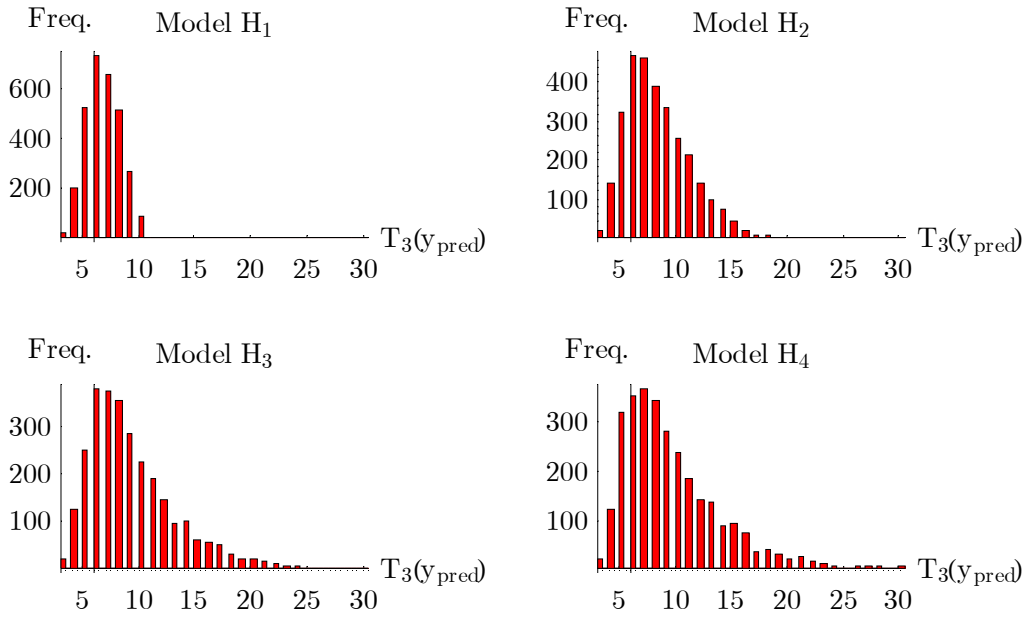


Fig. 9.4. The frequency distribution of the test statistic $T_4 = \dfrac{\#\, y_{rep,j} \in y_{rep} : y_{rep,j} = 0}{1000}$ under the naïve models H_1, H_2, H_3 and H_4. $T_4(y) = 0.887$ is marked by the vertical black tab.

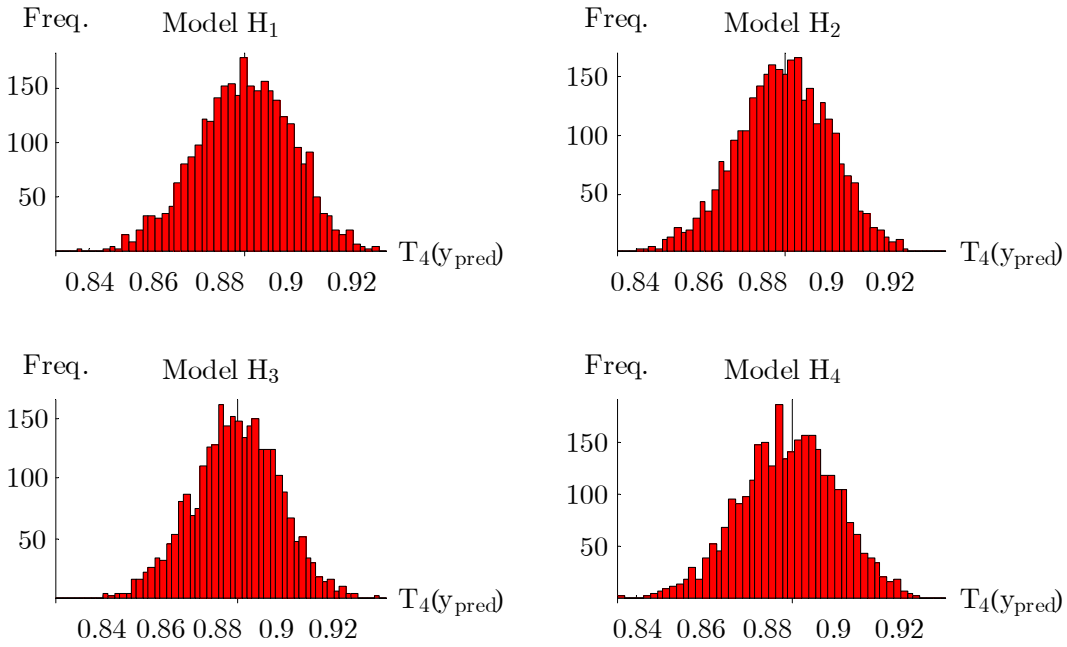
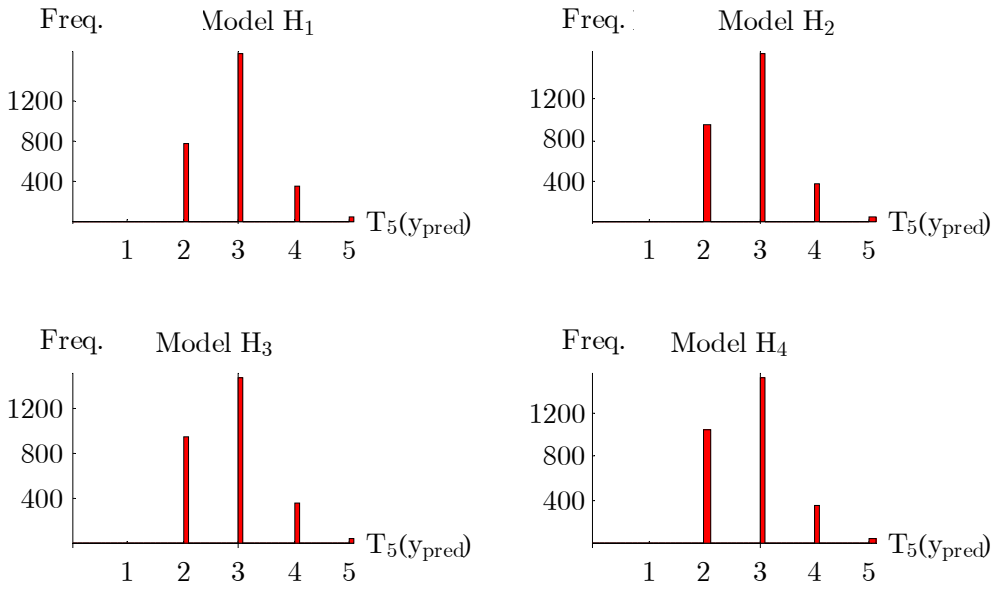


Fig. 9.5. The frequency distribution of the test statistic T_5 (the 99% quantile of a replicated data set) under the naïve models H_1, H_2, H_3 and H_4. $T_5(y) = 3$.



To calculate the Bayesian p_B corresponding to the test statistic $T_1 = \sum y_{rep,j}$ under model H_1, for example, the following procedure is followed: For each point (μ, τ, λ) sampled from $p(\mu, \tau, \lambda, \zeta \mid y, H_1)$ a replicated data set $y_{rep} = \{y_{rep,1}, y_{rep,2}, ..., y_{rep,1000}\}$ is generated by sampling from $p(y_{rep} \mid \mu, \tau, \lambda)$ given by (5.08). The Bayesian p_B is subsequently calculated as the fraction of the $T_1(y_{rep})$'s which equals or exceeds $T_1(y) = 170$.

As it emerges from table 9.1, the calculated p_B' s are not extreme under any of the naïve models. Consequently, the completed model checking does not provide a decision maker with firm arguments for the dismissal of any of the models.

The results from a different line of approach, i.e., the calculation of the expected deviance and the deviance information criterion (*DIC*), are summarized in table 9.2 on the following page. In general, the model with the lowest expected deviance will have the highest posterior probability. As it is seen from table 9.2, the expected deviance varies only slightly under the four naïve models. However, model H_1 does exhibit the smallest expected deviance which is contrary to what might be anticipated. A similar pattern is observed in the *DIC*-column of table 9.2.

Table 9.2. Comparison of models by means of sampled deviances. $\hat{D}_{avg}(y) =$ the average sampled deviance; $DIC =$ deviance information criterion; $p_D^{(2)} =$ model complexity parameter, see definition in the text.

Model	$\hat{D}_{avg}(y)$	DIC	$p_D{}^{(2)}$
H_1	920.257	922.568	2.66302
H_2	921.064	923.317	2.54707
H_3	921.649	924.083	3.41124
H_4	921.742	924.073	2.40732

Included in table 9.2 is the *model complexity* parameter $\mathrm{P_D}^{(2)}$ which is calculated as

$$\begin{aligned} p_D^{(2)} &= \frac{1}{2}\,\widehat{\mathrm{var}}(D(y,\phi)\mid y) \\ &= \frac{1}{2}\frac{1}{L-1}\sum_{l=1}^{L}(D(y,\phi^l)-\hat{D}_{avg}(y))^2. \end{aligned} \qquad \textbf{(9.01)}$$

Thus $\mathrm{P_D}^{(2)}$ is an estimate of half times the posterior variance of the deviance and can be interpreted as the number of *unconstrained* parameters in a Bayesian model. In this context, a parameter is considered as *unconstrained* if it is estimated with no prior information, and it is considered as *constrained* if all information about the parameter comes from the prior distribution. The tabulated values of $\mathrm{P_D}^{(2)}$ in table 9.2 confirms what has previously been discussed: Due to the large number of components in the vector λ, the individual components of λ get essentially locked to their prior expected values $E[\lambda_m]$, and one ends up with a constrained mixture model having approximately two free parameters (the $\mathrm{P_D}^{(2)}$ under model $\mathrm{H_3}$ appears, however, a bit out of line).

9.2 Evaluation of Naïve Models

The mixture models which have been examined so far were in chapter 7 introduced as "naive" discrete models due to the particular simple choice of integers $\{m_1, m_2, ..., m_g\}$, Dirichlet parameters $(\alpha_1, \alpha_2, ..., \alpha_g)$, and prior distribution $p(\mu, \tau)$. The subsequent Markov chain simulations and model checking have revealed several important properties of the finite mixture model (5.8). However, to make (5.8) adaptable to real-life applications and

to remedy some of the shortcomings inherent in (5.8), a couple of modifications have to be implemented.

Firstly, the simple choice of integers and Dirichlet parameters made in connection with the naïve models leads to an unrealistic probability distribution as to the mine content in a randomly selected minefield. This will of course affect the reliability of the generated posterior $p(\theta \mid y)$. Secondly, when the dimension of λ is increased, the mixture model becomes computationally intractable, and the individual components of λ get locked to their prior expected values $E[\lambda_m]$.

An attractive alternative which seems to remedy the above shortcomings is to work with sets of integers sampled from some probability distribution $p(m_1, m_2, ..., m_g)$. The increased flexibility gained hereby gives rise to new choices: 1) the selection of an appropriate sampling distribution; 2) the determination of g.

Another problem revealed by the completed Markov chain simulations is the high sensitivity of the posterior $p(\theta \mid y)$ to the dimension of λ when the prior $p(\mu, \tau)$ is non-informative. The only way to remedy this problem is to replace non-informative priors with vaguely or moderately informative priors. It is uncertain, however, on what criteria such partly informative priors should be derived.

All issues outlined above will be addressed in the following chapters. We begin in chapter 10 by demonstrating how sampling from a distribution $p(m_1, m_2, ..., m_g)$ can be built into the structure of (5.8). There are various ways to accomplish this, but in the present report the choice has fallen on an elegant method developed by Stephens [Stephens, 2000] in which the integer m_i and the associated probability λ_{m_i} is considered as a point in a continuous birth-death point process. Due to the method by Stephens, a Markov chain can be generated which alternately samples from the mixture model (5.8) and from a birth-death point process. As a result, a stationary distribution "averaged" over mixture models of varying dimension (i.e., varying sizes of g) is generated. The outlined extended mixture model can be considered as an alternative to the so-called *reversible jump* methodology [see Richardson and Green, 1997].

After the introduction of the above method, chapter 11 deals with the specification of the various priors which enter into the extended mixture model. Finally, in chapter 12 the results from a variety of Markov chain simulations based on the extended mixture method are presented.

Chapter 10

Finite Mixture Models with Varying Number of Components

Due to the assignment

$$\{m_1, m_2, ..., m_g\} \rightarrow \{0, 1, 2, ..., g-1\} \tag{10.01}$$

which was made in connection with the naïve models considered in chapter 7-9, the mixture model (5.8) took on the simple form

$$\begin{aligned} p(y_j \mid \mu, \tau, \lambda) &= \sum_{i=1}^{g} \lambda_{m_i} \, f(y_j \mid m_i, \mu, \tau) \\ &= \sum_{i=1}^{g} \lambda_{i-1} \, f(y_j \mid i-1, \mu, \tau) \end{aligned} \tag{10.02}$$

from which the posterior distribution could be calculated as

$$\begin{aligned} p(\mu, \tau, \lambda \mid y) &\propto p(y \mid \mu, \tau, \lambda) p(\mu, \tau) p(\lambda) \\ &= \prod_{j=1}^{M} p(y_j \mid \mu, \tau, \lambda) p(\mu, \tau) p(\lambda). \end{aligned} \tag{10.03}$$

We will now abandon the assignment (10.01) and instead consider the set of integers $\{m_1, m_2, ..., m_g\}$ as a stochastic vector distributed according to some probability distribution $p(m_1, m_2, ..., m_g)$. Considering $p(m_1, m_2, ..., m_g)$ as independent of μ, τ and λ, this results in the modified posterior distribution

$$\begin{aligned} p(\mu, \tau, \lambda, m \mid y) &\propto p(y \mid \mu, \tau, \lambda, m) p(\mu, \tau) p(\lambda) p(m) \\ &= \prod_{j=1}^{M} p(y_j \mid \mu, \tau, \lambda, m, g) p(\mu, \tau) p(\lambda) p(m), \end{aligned} \tag{10.04}$$

where $m = (m_1, m_2, ..., m_g)$ and

$$p(y_j \mid \mu, \tau, \lambda, m, g) = \sum_{i=1}^{g} \lambda_{m_i} \, f(y_j \mid m_i, \mu, \tau) \, . \tag{10.05}$$

When $\{m_1, m_2, ..., m_g\}$ is regarded as a stochastic vector, the significance of the number of included components is unclear. Any application of (10.04) should therefore include runs over different values of g to examine the sensitivity of $p(\mu, \tau, \lambda, m \mid y)$ to g. As an alternative to making separate runs for different values of g, one could consider g as a stochastic variable distributed according to a probability distribution $p(g)$, in which case (10.04) is expanded to the expression

$$\begin{aligned} p(\mu, \tau, \lambda, m, g \mid y) &\propto p(y \mid \mu, \tau, \lambda, m) p(\mu, \tau) p(\lambda) p(m) p(g) \\ &= \prod_{j=1}^{M} p(y_j \mid \mu, \tau, \lambda, m, g) p(\mu, \tau) p(\lambda) p(m) p(g). \end{aligned} \tag{10.06}$$

From (10.06) the marginal distribution $p(\mu, \tau \mid y)$ can be obtained, and $p(\theta \mid y)$ is extracted from $p(\mu, \tau \mid y)$ as before.

In what follows we will implement (10.06) by closely following an algorithm derived by Stephens [Stephens, 2000]. To simplify sampling from (10.06) it is advantageous to update $p(\mu, \tau, \lambda, m, g \mid y)$ in two successive steps by sampling from the full conditioned posterior distribution $p(\lambda, m, g \mid y, \mu, \tau)$ and $p(\mu, \tau \mid y, \lambda, m, g)$, respectively. Sampling from $p(\mu, \tau \mid y, \lambda, m, g)$ means sampling from a finite mixture model with a fixed number of components g, and the sampling procedure outlined in chapter 6 can thus be applied if the data y are augmented by indicator variables ζ. What is left is therefore the construction of a Markov chain with stationary distribution $p(\lambda, m, g \mid y, \mu, \tau)$.

The conditioned posterior $p(\lambda, m, g \mid y, \mu, \tau)$ can according to Bayes' rule be written as

$$p(\lambda, m, g \mid y, \mu, \tau) \propto p(y \mid \lambda, m, g, \mu, \tau) p(\lambda, m, g \mid \mu, \tau), \tag{10.07}$$

where

$$p(y \mid \lambda, m, g, \mu, \tau) = \prod_{j=1} [\sum_{i=1}^{g} \lambda_{m_i} f(y_j \mid m_i, \mu, \tau)]. \tag{10.08}$$

It is worthy of note that (10.08) is invariant to permutations of the component labels, i.e.

$$\begin{aligned} &p(y \mid (\lambda_{m_1}, \lambda_{m_2}, ..., \lambda_{m_g}), (m_1, m_2, ..., m_g), g, \mu, \tau) = \\ &p(y \mid (\lambda_{\varepsilon(m_1)}, \lambda_{\varepsilon(m_2)}, ..., \lambda_{\varepsilon(m_g)}), (\varepsilon(m_1), \varepsilon(m_2), ..., \varepsilon(m_g)), g, \mu, \tau) \end{aligned} \tag{10.09}$$

for all permutations ε of $m_1, m_2, ..., m_g$. If we express

$$p(\lambda, m, g \mid \mu, \tau) = p(\lambda \mid \mu, \tau) p(m \mid \mu, \tau) p(g \mid \mu, \tau), \tag{10.10}$$

and assume that $\lambda \mid \mu, \tau \sim Dirichlet(1,1,...,1)$ and the m_i's are independent and identically distributed, it follows that

$$p(\lambda, m, g \mid \mu, \tau) \propto [\prod_{i=1}^{g} p(m_i \mid \mu, \tau)]\, p(g \mid \mu, \tau). \tag{10.11}$$

As $p(\lambda, m, g \mid \mu, \tau)$ in (10.11) is invariant to permutations of the component labels just as (10.08), it follows that the conditioned posterior

$$p(\lambda, m, g \mid y, \mu, \tau) \propto p(y \mid \lambda, m, g, \mu, \tau) p(\lambda, m, g \mid \mu, \tau) \tag{10.12}$$

is also invariant to permutations of the component labels. This property allows us to ignore the component labels and simply consider any set $\{(\lambda_{m_1}, m_1), (\lambda_{m_2}, m_2), ..., (\lambda_{m_g}, m_g)\}$ as g points in $[0,1] \times \mathbb{N}_0$. More specifically, we might view $p(\lambda, m, g \mid y, \mu, \tau)$ as a distribution of points or a *point process* on $[0,1] \times \mathbb{N}_0$ [Stephens, 2000, p. 45].

Looking at $p(\lambda, m, g \mid y, \mu, \tau)$ as a point process provides the way for the introduction of a Markov birth-death process in continuous time with stationary distribution $p(\lambda, m, g \mid y, \mu, \tau)$. In this simulated process, the birth and death of points (λ_{m_i}, m_i) occur as independent Poisson processes, and the dimension of the finite mixture model consequently varies during the simulation process.

To introduce the Markov birth-death process thoroughly, several terms have to be defined. In what follows we will simply write λ_{m_i} as λ_i to avoid cluttered expressions. Firstly, if the process at time t is characterized by the state vector z written as

$$z = \{(m_1, \lambda_1), (m_2, \lambda_2), ..., (m_g, \lambda_g)\}, \tag{10.13}$$

and a birth occurs at the point $(\lambda^*, m^*) \in [0,1] \times \mathbb{N}_0$, the process jumps to the state vector

$$z \cup (\lambda^*, m^*) = \{(\lambda_1(1-\lambda^*), m_1), ..., (\lambda_g(1-\lambda^*), m_g), (\lambda^*, m^*)\}. \quad \textbf{(10.14)}$$

Similarly, if a death occurs at the point (λ_i, m_i), the process jumps to the state vector

$$\begin{aligned} z \setminus (\lambda_i, m_i) = \{(\frac{\lambda_1}{(1-\lambda_i)}, m_1), , ..., (\frac{\lambda_{i-1}}{(1-\lambda_i)}, m_{i-1}), \\ ...(\frac{\lambda_{i+1}}{(1-\lambda_i)}, m_{i+1}), ..., (\frac{\lambda_g}{(1-\lambda_i)}), m_g\}. \end{aligned} \quad \textbf{(10.15)}$$

Whatever the number of points included in the state vector, the conventions made above ensure that $\sum_i \lambda_i = 1$.

To guarantee that the Markov birth-death process has the posterior $p(\lambda, m, g \mid y, \mu, \tau)$ as its stationary distribution, the birth- and death rates have to obey a certain balance equation. Given a birth occurs when the process is at z, this implies that $(\lambda^*, m^*) \in [0,1] \times \mathbb{N}_0$ is chosen according to the density

$$b((\lambda^*, m^*) \mid z) = g(1-\lambda^*)^{g-1} \cdot p(m^* \mid \mu, \tau) \quad \textbf{(10.16)}$$

with the restriction that $b(\lambda^*, m^* \mid z) = 0$ if either the conditioned prior given by (10.10) or the likelihood given (10.08) is equal to zero at the point $z \cup (\lambda^*, m^*)$. The parameter g in (10.16) denotes the number of components in z. From (10.16) it is evident that $\lambda^* \mid z \sim Beta(1, g)$. The density $p(m^* \mid \mu, \tau)$ has yet to be defined. The overall birth rate is set to γ_b.

Regarding the death process, let each point (λ_j, m_j), $j = 1, 2, ..., g$, die independently of the others in a Poisson process with rate $\delta_j(z)$ when the process is at z. The overall death rate amounts then to $\delta(z) = \sum_j \delta_j(z)$. The balance equation implies that if $\delta_j(z)$ is set to

$$\delta_j(z) = \gamma_b \frac{p(y \mid z \setminus (m_j, \lambda_j), g-1, \mu, \tau)}{p(y \mid z, g, \mu, \tau)} \frac{p(g-1 \mid \mu, \tau)}{g\, p(g \mid \mu, \tau)} \text{ for } \forall\ j \quad \textbf{(10.17)}$$

the birth-death process defined above has the stationary distribution $p(\lambda, m, g \mid y, \mu, \tau)$. The density $p(g \mid \mu, \tau)$ appearing in (10.17) has yet to be defined. Similarly to the birth process, (10.17) is restricted by the condition that $\delta_j(z) = 0$ if the conditioned prior (10.10) is equal to zero at the point $z \setminus \{\lambda_j, m_j\}$.

From the summary above it appears that to simulate the Markov birth-death process, three quantities have to be specified: The birth rate γ_b, the density $p(m \mid \mu, \tau)$, and the density $p(g \mid \mu, \tau)$. Given that these quantities have been specified, the simulation of the birth-death process is straightforward. The following sketch of the simulation algorithm follows closely the algorithm suggested by Stephens [Stephens, 2000, page 48]:

1) The birth-death process is run for a virtual time t_0. A convenient choice is $t_0 = 1$.
2) Let the state vector $z = \{(m_1, \lambda_1), (m_2, \lambda_2), ..., (m_g, \lambda_g)\}$ make up the initial model.
3) Calculate the death rate $\delta_j(z)$ for each component j in accordance with (10.17).
4) Calculate the total death rate $\delta(z) = \sum_j \delta_j(z)$.
5) The time t' for the next jump (i.e., birth or death) is simulated by sampling from an exponential distribution with mean $(\gamma_b + \delta(z))^{-1}$.
6) The type of jump at time t' is determined by sampling a real number $r \sim Uniform(0,1)$. The jump is classified as a birth if $r \leq \dfrac{\gamma_b}{\gamma_b + \delta(z)}$.

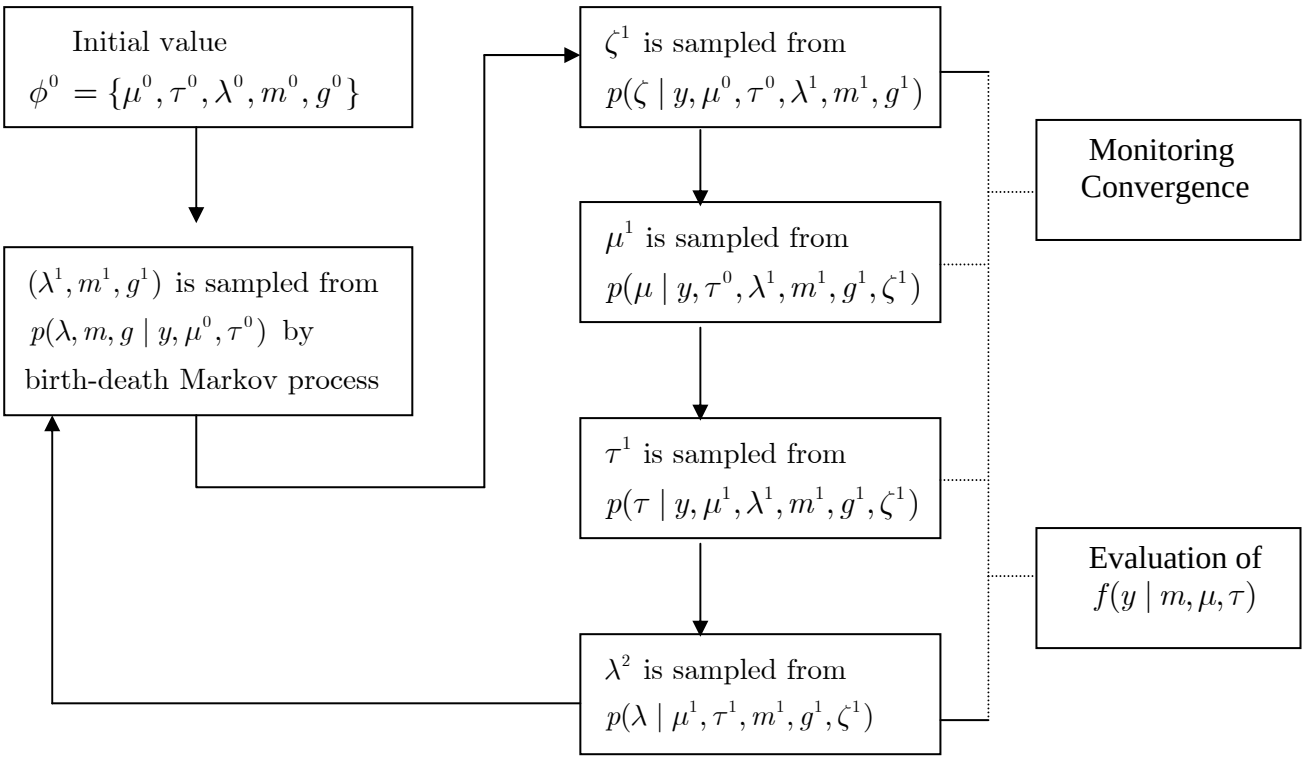
Depending on the type of jump determined at step 6, the state vector z is adjusted in the following way:

7a) Birth: A new point (λ^*, m^*) is determined by sampling independently $\lambda^* \mid z \sim Beta(1, g)$ and m^* from the density $p(m^* \mid \mu, \tau)$.
7b) Death: The component to die is selected with probability $\dfrac{\delta_j(z)}{\delta(z)}$.

The birth-death algorithm returns hereafter to step 3 and continues until the accumulated jump times exceed t_0.

By combining the above birth-death process with the Markov process outlined in chapter 6, sampling from the target distribution $p(\mu, \tau, \lambda, m, g \mid y)$ can be achieved. Fig. 10.1 below gives an overview of the various components in the joined sampling algorithm. Note that the indicators ζ are once again used as auxiliary variables under the updates of $p(\mu, \tau \mid y, \lambda, m, g)$. The indicator variables do not interfere with the birth-death process.

Fig 10.1. Markov-chain simulation including birth-death point process.



It appears from fig. 10.1 that the vector λ is updated twice during every iteration. The second update (i.e., λ^2) is not a prerequisite for convergence of the Markov chain but has simply been included to improve mixing.

To ensure that any Markov chain simulation following the above sampling scheme reaches all important parts of the target distribution, every simulation is initiated from m different starting points $\{\phi_1^0, ..., \phi_m^0\}$ in accordance with the procedure outlined in chapter 6. This time, however, the starting points also differ with respect to g, i.e., the starting points are picked from mixture models of different dimension.

Chapter 11

Specification of Prior Distributions

To apply the extended mixture model introduced in chapter 10, four priors have to be specified:

1) $p(\lambda \mid \mu, \tau)$
2) $p(g \mid \mu, \tau)$
3) $p(m \mid \mu, \tau)$
4) $p(\mu, \tau)$

Concerning $p(\lambda \mid \mu, \tau)$, we will continue with the assignment $\lambda \mid \mu, \tau \sim Dirichlet(1,1,...,1)$ which was made in relation to the naïve models in chapter 8. As to the remaining priors it will in the present chapter be exemplified how informative priors can be set up which only require a modest amount of input from the decision maker. The suggested priors will be thoroughly tested in chapter 12.

11.1 Specification of $p(g \mid \mu, \tau)$

As a matter of simple convenience we will ascribe g a Poisson distribution, i.e.,

$$p(g \mid \mu, \tau) = p(g) \propto \frac{\Lambda^g}{g!}, \qquad \textbf{(11.01)}$$

where Λ in (11.01) is independent of (μ, τ). In chapter 12, Markov chain simulations will be carried out for different values of Λ.

11.2 Specification of $p(m \mid \mu, \tau)$

Unlike the parameter g which does not have a clear physical interpretation, the parameter m from $p(m \mid \mu, \tau)$ is directly linked to the degree of mine contamination in the minefields under study. Consequently, if historical information is available from mine clearance operations completed elsewhere which has revealed the typical content of mines in

minefields of a similar nature, such information should be incorporated into the prior distribution $p(m \mid \mu, \tau)$.

A simple structure to impose on $p(m \mid \mu, \tau)$ is to write $p(m \mid \mu, \tau)$ as

$$p(m \mid \mu, \tau) = p(m) = \begin{cases} p_0 \text{ if } m = 0 \\ \pi(m) \text{ if } m > 0. \end{cases} \qquad \textbf{(11.02)}$$

That is, $p(m)$ is assumed independent of μ, τ, and $m \in \mathbb{N}_0$. The rationale behind (11.02) is the general experience that a considerable fraction of the areas originally classified as minefields during subsequent mine clearance operations turns out to be mine free. Being "mine free" may actually be the most frequent observation made during larger mine clearance programmes. It seems therefore appropriate to ask a decision maker for the probability that a randomly selected minefield actually contains zero mines. The decision maker may give his answer through the point estimate p_0 in (11.02).

Concerning the conditioned distribution $\pi(m)$, a convenient measure of the decision maker's uncertainty (or lack of information) is the *entropy* $H(\pi)$ which for $\pi(m)$ defined in (11.02) takes the form

$$H(\pi) = -\sum_{m=1}^{\infty} \pi(m) \log \pi(m), \qquad \textbf{(11.03)}$$

where $\sum_{m=1}^{\infty} \pi(m) = 1$. Thus $H(\pi)$ is a functional, and it can be shown that $H(\pi) \geq 0$ for any choice of $\pi(m)$. As a matter of fact $H(\pi) = 0$ if and only if $\pi(m)$ takes the form

$$\pi(m) = \begin{cases} 1 \text{ if } m = j \\ 0 \text{ if } m \neq j, \end{cases} \qquad \textbf{(11.04)}$$

where $m, j \in \mathbb{N}$. Given two distributions $\pi_1(m)$ and $\pi_2(m)$ we will say that $\pi_1(m)$ relative to $\pi_2(m)$ reflects a larger uncertainty (i.e., less information) about m if $H(\pi_1) > H(\pi_2)$.

Assume now that the decision maker based on the available information can impose k restrictions on $\pi(m)$ being expressed as

$$\sum_{m=1}^{\infty} g_j(m)\pi(m) = \mathrm{M}_j, \quad j = 1,2,...,k. \tag{11.05}$$

It can then be shown [see for example Berger, 1980] that the distribution which satisfies (11.05) and *maximizes* the entropy $H(\pi)$ is given as

$$\pi(m) = \frac{e^{\sum_{j=1}^{k} g_j(m)\lambda_j}}{\sum_{m'=1}^{\infty} e^{\sum_{j=1}^{k} g_j(m')\lambda_j}}, \tag{11.06}$$

where the coefficients λ_j are to be determined from the k conditions in (11.05). The prior given by (11.06) seems to be a fair choice as no information has been imparted to $\pi(m)$ except from what has been deliberately expressed by (11.05).

In what follows we will assume that information is at hand which allows the decision maker to specify the <u>expected value of m</u> (given $m > 0$). The specification can be written as

$$\sum_{m=1}^{\infty} m\,\pi(m) = \mathrm{M}, \tag{11.07}$$

from which it follows that $\pi(m)$ is given as

$$\pi(m) = \frac{e^{m\lambda}}{\sum_{m'=1}^{\infty} e^{m'\lambda}}. \tag{11.08}$$

By use of the identity $\sum_{m=1}^{\infty} e^{\lambda m} = \frac{e^{\lambda}}{1-e^{\lambda}}$ it turns out that $\lambda = \log(1-\frac{1}{\mathrm{M}})$ from which it follows that

$$\pi(m) = (\frac{1}{\mathrm{M}}) \cdot (1-\frac{1}{\mathrm{M}})^{m-1}, \tag{11.09}$$

i.e., $m \mid m > 0 \sim Ge(\frac{1}{\mathrm{M}})$.

So to conclude,

$$p(m) = \begin{cases} p_0 \text{ if } m = 0 \\ (\frac{1}{\mathrm{M}}) \cdot (1-\frac{1}{\mathrm{M}})^{m-1} \text{ if } m > 0 \end{cases} \tag{11.10}$$

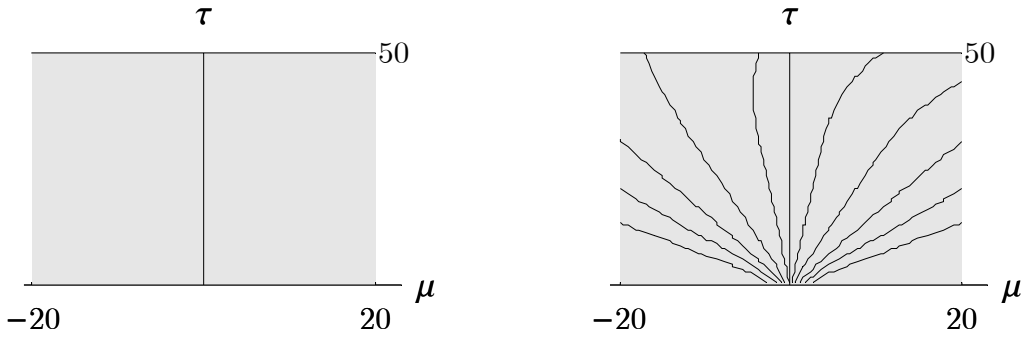
The sensitivity of the posterior $p(\mu,\tau,\lambda,m,g \mid y)$ to various combinations of (p_0, M) will be examined in chapter 12.

11.3 Specification of $p(\mu,\tau)$

The prior $p(m)$ given by (11.10) can be considered as a moderately informative prior as it is based on just two specifications, i.e., the point estimate p_0 and the average value M, whose meanings are intuitively clear. It seems harder, however, to make similar specifications concerning the prior $p(\mu,\tau)$. That is, if historical information is at hand about the likelihood of encountering a mine, such information may presumably be rephrased in quantitative terms by estimates of certain properties of $p(\theta)$, say $E[\theta]$ or the 50% quantile of the distribution of θ. In other words, the estimate refers directly to properties of the binomial parameter θ and not to properties of either μ or τ.

To set up a prior $p(\mu,\tau)$ which is based on prior knowledge about θ, recall that in chapter 8 a non-informative prior distribution was set up in terms of two uniform priors $p(\mu)$ and $p(\tau)$ both being cut off at a faraway distance (specified by the constants k_1 and k_2) to ensure a proper posterior distribution of $p(\mu,\tau,\lambda,\zeta \mid y)$. In fig 11.1.a below, the square at which $p(\mu,\tau) = \text{constant} \neq 0$ is shown for the particular choice $k_1 = 20$ and $k_2 = 50$.

Fig. 11.1. Specification of prior for μ,τ under simple mixture model. Fig. 11.1.a (left figure): Prior distribution used under the naïve models H_1, H_2, H_3, H_4 in chapter 8: $p(\mu,\tau) = \text{constant} \neq 0$ within square. Fig. 11.1.b (right figure): Subset of contour lines traversing the square from fig. 11.1.a .



Recall also that every point (μ,τ) located within the square in fig. 11.1.a corresponds through relation (5.3) to a probability distribution $p(\theta \mid \mu,\tau)$ characterized by an expected

value $E[\theta \mid \mu,\tau] = f(1,1,\mu,\tau)$. In fig. 11.1.b all points (μ,τ) characterized by the same expected value of θ are linked through a contour line.

To set up a moderately informative prior $p(\mu,\tau)$ we will tentatively write $p(\mu,\tau)$ as

$$p(\mu,\tau) \propto \begin{cases} \beta \text{ if } 0 \leq f(1,1,\mu,\tau) \leq E \\ 1 \text{ if } E < f(1,1,\mu,\tau) \leq 1. \end{cases} \tag{11.11}$$

That is, $p(\mu,\tau)$ requires only a specification of the two parameters β and E. To apply expression (11.11), a hypothetical decision maker has to proceed as follows: Firstly, the decision maker divides the square from fig. 11.1 up into two compartments A and B separated by a contour line E of his own choice as sketched in fig. 11.2. Secondly, the decision maker specifies through his choice of β the *prior odds*

$$\beta = \frac{p(\mu_A,\tau_A)}{p(\mu_B,\tau_B)}, \tag{11.12}$$

where (μ_A,τ_A) and (μ_B,τ_B) are arbitrary points belonging to compartment A and B, respectively. It follows that all points belonging to a given compartment are assigned the same probability, a priori.

Fig. 11.2. Compartmentalization by contour line. All points (μ,τ) located on the red solid contour line are characterized by the expected value $E[\theta \mid \mu,\tau] = E$. All points belonging to a given compartment are assigned the same a priori probability.

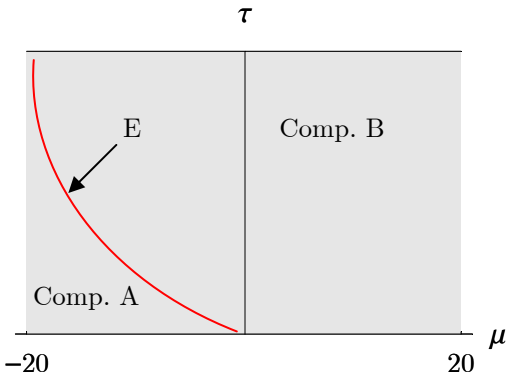


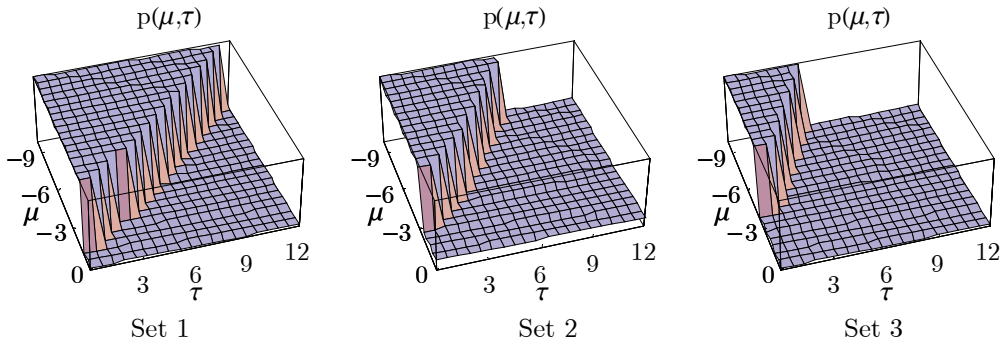
Table 11.1 below tabulates three different combinations of the parameters (E,β), denoted *Set 1*, *Set 2* and *Set 3*. The corresponding priors $p(\mu,\tau)$ are sketched in fig. 11.3. As the

probability mass is more localized in *Set 3* relative to the distributions found in *Set 1* and *Set 2*, we will consider the distribution corresponding to *Set 3* as the most informative among the three prior distributions. The three priors will be applied in chapter 12.

Table 11.1. Three parameter combinations determining three priors $p(\mu,\tau)$.

Parameter combination	E	β
Set 1	0.2	4
Set 2	0.1	9
Set 3	0.02	49

Fig. 11.3 Priors $p(\mu,\tau)$ corresponding to the three parameter combinations from table 11.1.



In table 11.1 we have summarized the parameters to be specified if the priors suggested in the present chapter are to be applied.

Table 11.1. Parameters to be specified in informative prior distributions .

Parameter	**Function**
Λ	The average number of components in finite mixture model
p_0	The probability that a minefield contains zero mines.
M	The expected numbers of mines in a minefield, given that the minefield contains mines.
E	Associated value of contour line dividing the parameter space of μ and τ into compartments A and B.
β	Prior odds for point belonging to compartment A relative to point belonging to compartment B.

Chapter 12

Markov Chain Simulations with Extended Mixture Model

12.1 Introduction

The aim of the following Markov chain simulations is twofold: Firstly, to investigate the utility of the extended mixture model as a way of generating the posterior $p(\theta \mid y)$. Secondly, to examine the sensitivity of the posterior $p(\theta \mid y)$ to different choices of informative priors. In the present chapter the results from 36 Markov chain simulations, all carried out according to the sampling scheme from fig. 10.1, are presented. The complete set of tested models are shown in table 12.1 below.

Table 12.1. Thirty six finite mixture models specified by their prior distribution parameters.

Model	Λ	(E, β)	p_0	M	Model	Λ	(E, β)	p_0	M
1	3	(0.02, 49)	0.1	10	19	10	(0.02, 49)	0.1	10
2	3	(0.02, 49)	0.1	20	20	10	(0.02, 49)	0.1	20
3	3	(0.02, 49)	0.1	30	21	10	(0.02, 49)	0.1	30
4	3	(0.02, 49)	0.2	10	22	10	(0.02, 49)	0.2	10
5	3	(0.02, 49)	0.2	20	23	10	(0.02, 49)	0.2	20
6	3	(0.02, 49)	0.2	30	24	10	(0.02, 49)	0.2	30
7	3	(0.1, 9)	0.1	10	25	10	(0.1, 9)	0.1	10
8	3	(0.1, 9)	0.1	20	26	10	(0.1, 9)	0.1	20
9	3	(0.1, 9)	0.1	30	27	10	(0.1, 9)	0.1	30
10	3	(0.1, 9)	0.2	10	28	10	(0.1, 9)	0.2	10
11	3	(0.1, 9)	0.2	20	29	10	(0.1, 9)	0.2	20
12	3	(0.1, 9)	0.2	30	30	10	(0.1, 9)	0.2	30
13	3	(0.2, 4)	0.1	10	31	10	(0.2, 4)	0.1	10
14	3	(0.2, 4)	0.1	20	32	10	(0.2, 4)	0.1	20
15	3	(0.2, 4)	0.1	30	33	10	(0.2, 4)	0.1	30
16	3	(0.2, 4)	0.2	10	34	10	(0.2, 4)	0.2	10
17	3	(0.2, 4)	0.2	20	35	10	(0.2, 4)	0.2	20
18	3	(0.2, 4)	0.2	30	36	10	(0.2, 4)	0.2	30

A few comments on the applied prior distributions: To investigate the sensitivity of the Markov chain simulations to the number of components included in the extended mixture model, two values of Λ were tested, that is, $\Lambda = 3$ in model $1 \rightarrow 18$, and $\Lambda = 10$ in model $19 \rightarrow 36$.

The various combinations of p_0 and M set up in table 12.1 give rise to different *a priori* estimates of the expected number of mines in a randomly selected minefield. In what follows we will denote such an a priori estimate $< m >$, where $< m >= (1 - p_0)\,\mathrm{M}$. The combinations of p_0 and M in table 12.1 include the estimates $< m > = 8$, 9, 16, 18, 24, and 27. The three applied combinations of (E, β) were mentioned at the end of chapter 11.

12.2 Results from Markov Chain Simulations

The following presentation will be split up into two parts: In the first part, certain features of the Markov chain simulations which arise due to the introduction of the point process will be illustrated. In the second part, various properties of the posteriors $p(\theta \mid y)$ derived from the completed simulations will be calculated and compared.

Features of Point Process

To ensure that any Markov chain simulation reaches all important parts of the target distribution $p(\mu, \tau, \lambda, m, g \mid y)$, every simulation under a given model is initiated from four different starting points. The starting points are different with respect to g, i.e. the number of included mixture components, which for practical reasons has been set to 10, 20, 30 and 40, respectively. Due to the introduction of the point process, g varies in time (i.e. iteration time), and after a few iterations g fluctuates about its average value, which is largely determined by the parameter Λ. This is illustrated for model 21 ($\Lambda = 10$) in fig. 12.1 (on the following page) where g as a function of iteration time is shown for each of the four Markov chains run under model 21.

Due to the point process, the set of integers $\{m_1, m_2, ..., m_g\}$ entering into the extended mixture model varies with time as new integers are constantly born and old integers are eliminated. Fig. 12.2 (on the following page) shows the list of integers which have survived at iteration time $t =$ 336, 337, 338 and 339.

Fig. 12.1. The variation of g with iteration time. The four graphs below illustrate the number of included mixture components as a function of time (i.e. time t = number of iterations). g(0) indicates the number of included mixture components at the start of the simulation. The displayed values of $g(t)$ were recorded during the Markov chain simulations under model 21.

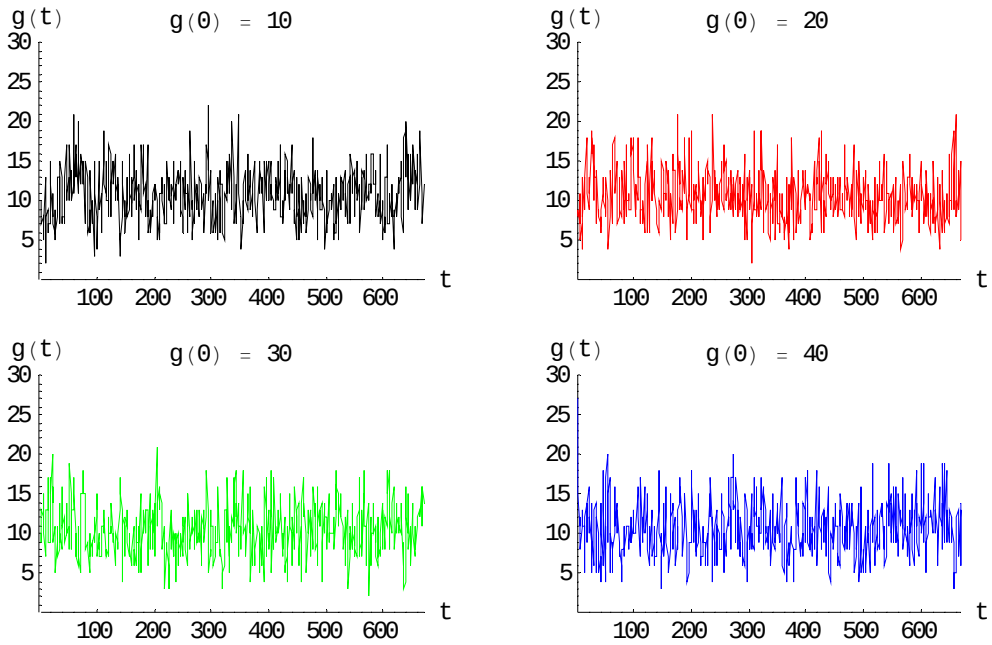
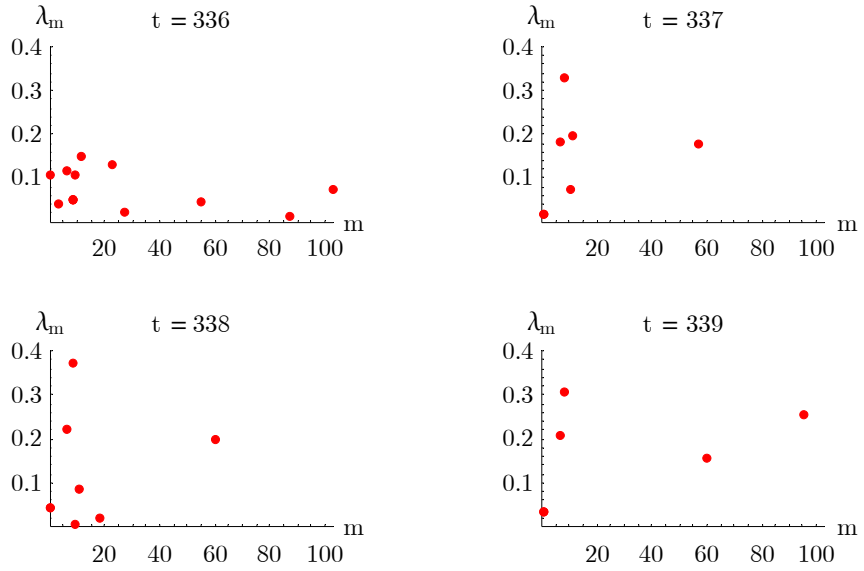


Fig. 12.2. List of integers $\{m_1, m_2, ..., m_g\}$ included in the extended mixture model at four successive iterations during the Markov chain simulation under model 21, $g(0) = 40$.

t = 336: {0,3,6,8,8,8,9,11,22,27,55,87,103}
t = 337: {0,0,6,8,10,11,57}
t = 338: {0,0,6,8,9,10,18,60}
t = 339: {0,0,6,8,60,95}

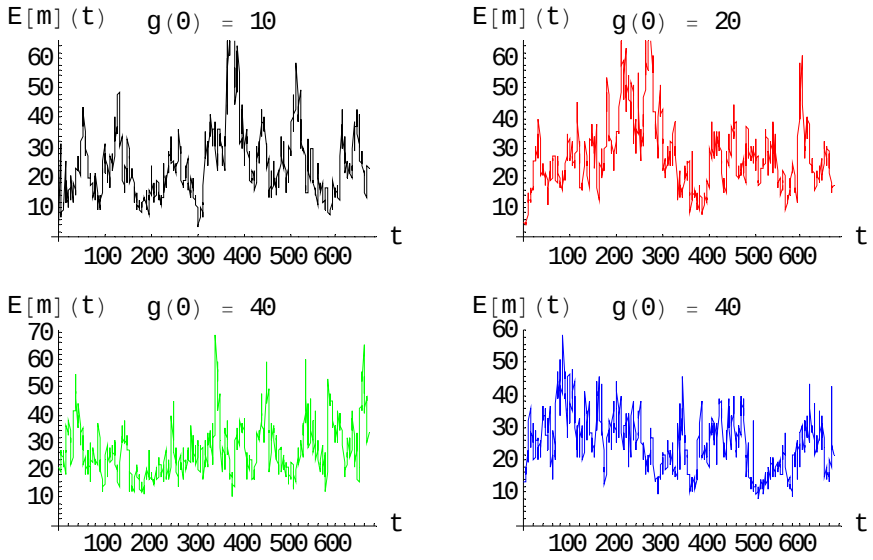
Every integer m is associated with a probability λ_m which varies with iteration time. Fig. 12.3 (on the following page) shows the distribution of the integers and their associated probabilities as points in $[0;1] \times \mathbb{N}_0$. Note that in fig 12.2 a given integer may appear more than once, whereas fig. 12.3 shows the distribution of *different* integers. Consequently, the probabilities in fig. 12.3 do not necessarily sum to 1.

Fig. 12.3. Plots of integers included in the extended mixture model and their associated probabilities at four successive iterations during the Markov chain simulation under model 21, $g(0) = 40$. Note that the multiplicity of an integer in a given list $\{m_1, m_2, ..., m_g\}$ cannot be determined from the plots below.



From the list of integers $\{m_1, m_2, ..., m_g\}$ and the associated set of probabilities $\{\lambda_{m_1}, \lambda_{m_2}, ..., \lambda_{m_g}\}$ the average number of mines, *a posteriori*, can be calculated. This number will be denoted $E[m]$ to distinguish it from the prior estimate $< m >$. Fig. 12.4 shows $E[m]$ as a function of iteration time.

Fig. 12.4. $E[m]$ as a function of iteration time under model 21. The integers and associated probabilities sampled at time t were recorded and the average number of mines was calculated as $E[m] = \sum_{j=1}^{g} \lambda_{m_j} m_j$.



The Distribution of θ: Statistical Inferences from Markov Chain Simulations

In the present context, the parameter of primary interest is the binomial parameter θ whose distribution can be estimated from the sampled values of the normal distribution parameters (μ, τ). To ensure that all important values of (μ, τ) are sampled properly during a Markov chain simulation, four Markov chains initiated from different starting points were run in parallel. At regular intervals the first halves of points (μ, τ) sampled from each chain were temporarily discarded, and the remaining halves were merged into one big chain from which the potential scale reduction factor $\hat{R}$ could be calculated. When $\hat{R}$ was found to be less or equal to 1.1 for both μ and τ, the sampling was stopped.

Fig. 12.5 below shows the distribution of (μ, τ) for each of the Markov chains from fig. 12.1. Each plot includes 670 sampled points. After having discarded half of the sampled points from each chain, the remaining points were merged as shown in fig. 12.6, and the sampling was stopped as $\hat{R}$ was found to be less than 1.1 for both μ and τ.

Fig. 12.5. Simulation of marginal posterior $p(\mu, \tau \mid y)$ under model 21. The four plots illustrate the distribution of the second half of the points (μ, τ) sampled during four Markov chain simulations run in parallel. The four chains are characterized by different starting points. The plot labels $g(0)$ indicate the number of components included in the mixture model at $t = 0$.

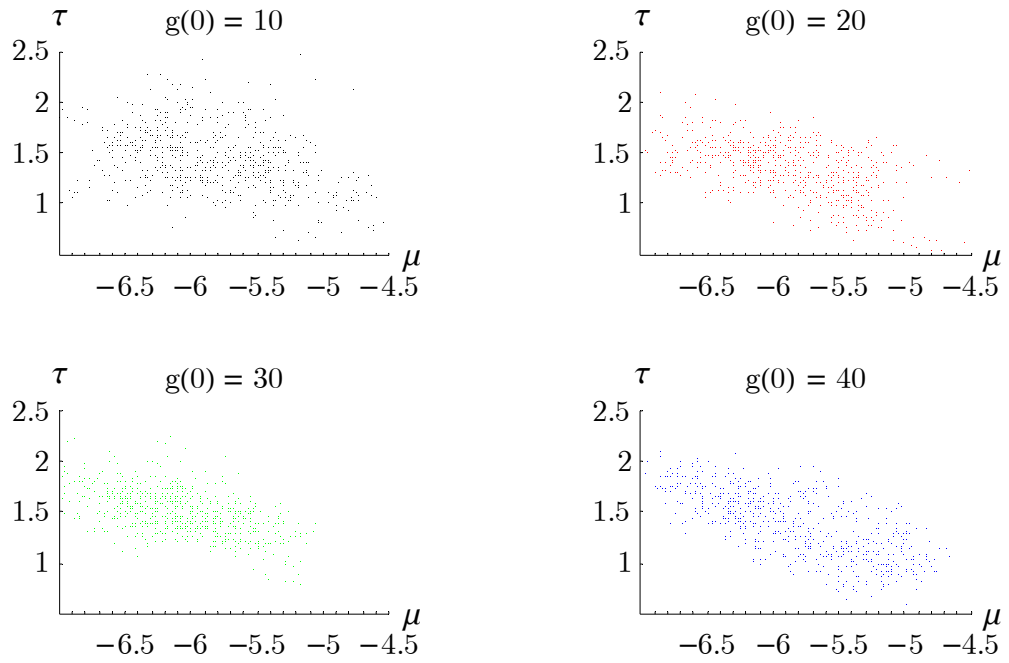
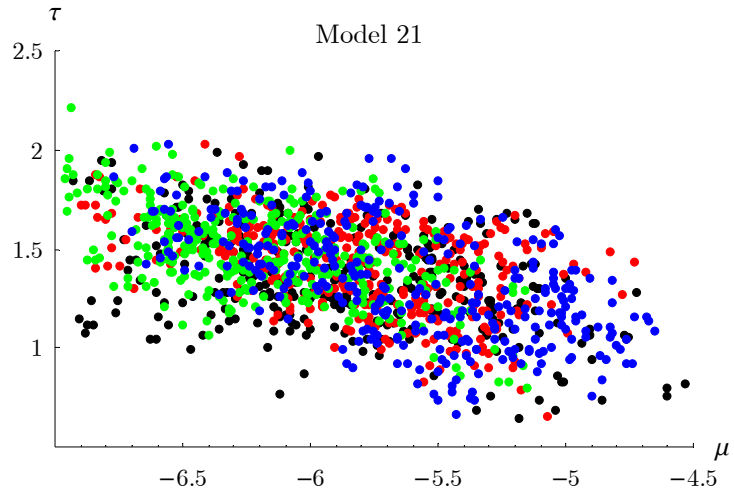


Fig. 12.6. The four marginal posteriors $p(\mu,\tau \mid y)$ from fig. 12.5 merged into one plot. All statistical inferences concerning the distribution of the binomial parameter θ under model 21 are based on the merged plot.



Based on the sample in fig. 12.6 various properties of $p(\theta \mid y)$ can be calculated like it was done in chapter 7 under the naïve models. The sensitivity of $p(\theta \mid y)$ to different choices of priors is the main theme in the remaining part of the present chapter.

Fig. 12.7 below provides an overview of the distribution of $E[\theta \mid y]$ derived from $p(\theta \mid y)$ based on the 36 completed Markov chain simulations. Similarly, fig. 12.8 shows the distribution of $Var[\theta \mid y]$.

Fig. 12.7. The posterior $E[\theta \mid y]$ obtained under the mixture models from table 12.1. The dashed red line indicates the true average value of θ (obtained from the distribution $p(\theta \mid \mu,\tau) = p(\theta \mid -4.7, 0.5)$). The integer located to the right of each point refers to model number according to table 12.1.

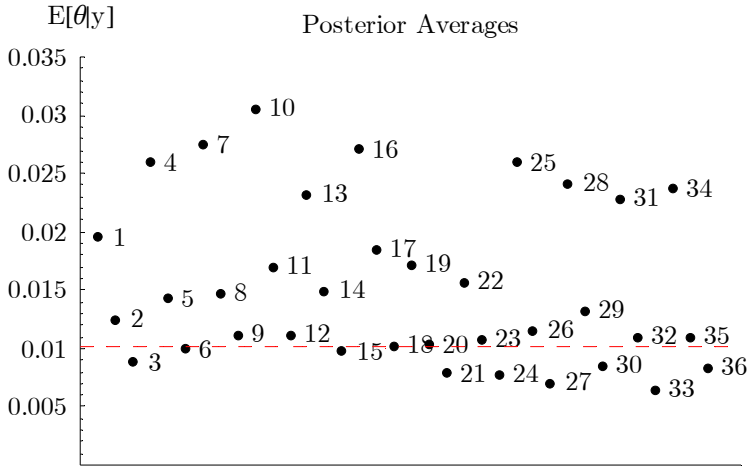


Fig. 12.8 The posterior $Var[\theta \mid y]$ obtained under the mixture models from table 12.1. The dashed red line located at the bottom of the plot indicates the true variance of θ (obtained from the distribution $p(\theta \mid \mu, \tau) = p(\theta \mid -4.7, 0.5)$).

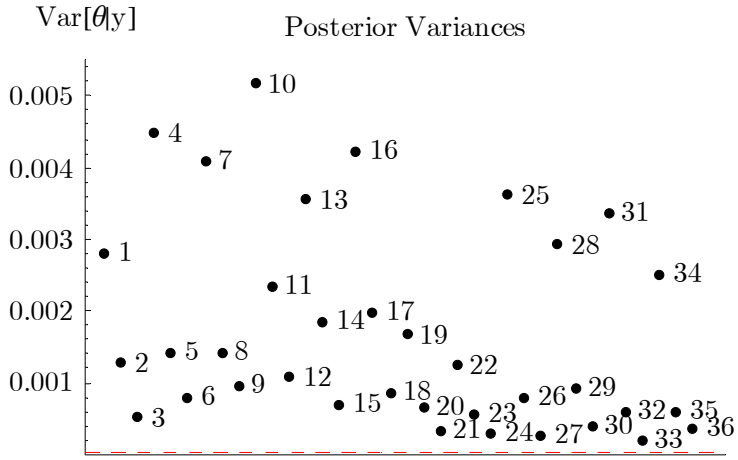


Fig. 12.9 and 12.10 below illustrate the location of the 95% and 50% posterior intervals for θ calculated under the mixture models from table 12.1.

Fig. 12.9 Location of 95% posterior interval for θ under the mixture models from table 12.1. 2.5% of the posterior density of $p(\theta \mid y)$ is located to the left and to the right, respectively, of the horizontal line under a given model. The posterior interval termed "DATA" is derived from the distribution $p(\theta \mid \mu, \tau) = p(\theta \mid -4.7, 0.5)$.

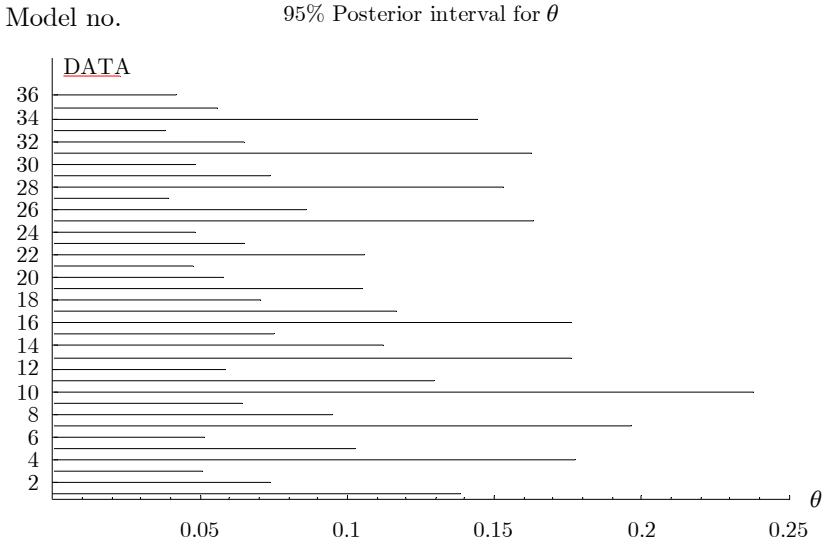


Fig. 12.10. Location of 50% posterior interval for θ under the mixture models from table 12.1. 25% of the posterior density of $p(\theta \mid y)$ is located to the left and to the right, respectively, of the horizontal line under a given model. The posterior interval termed "DATA" is derived from the distribution $p(\theta \mid \mu, \tau) = p(\theta \mid -4.7, 0.5)$.

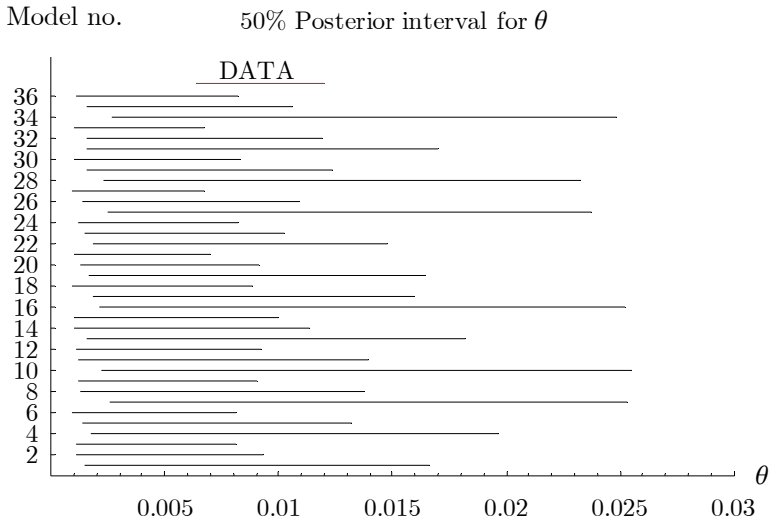


Figure 12.8, 12.9 and 12.10 illustrate as expected that the posterior variance or posterior interval of θ in all examined cases is estimated to be larger than the true variance or the true 50% (or 95%) interval of θ. This overdispersion has two sources: the uncertainty about the actual content of mines in the individual minefields (which is reflected through the use of a finite mixture model), and the overdispersion implied by the introduction of the scale parameter τ which was made in (5.1) and (5.2) .

To analyse the above results thoroughly we will by way of introduction examine whether the average number of components included in the extended mixture model affects essential properties of the posterior $p(\theta \mid y)$. According to table 12.1, model j and model j+18 are identical except from the ascribed value of Λ, where $\Lambda = 3$ for model $1 \rightarrow 18$, and $\Lambda = 10$ for model $19 \rightarrow 36$. Recall that Λ signifies the expected number of components in the mixture model, a priori.

Fig. 12.11 below illustrates the average number of components actually included during the Markov chain simulations for the 36 examined models. Each point in fig. 12.11 is found by averaging over the length of chains similar to the chains depicted in fig. 12.1. These posterior averages are seen to be displaced slightly upwards relative to their respective

prior values (i.e. 3 and 10, respectively), but within each group the posterior averages seem to be unaffected by the values of E, β, p_0, and M.

Fig. 12.11. The average number of included mixture components under the mixture models from table 12.1. $<$ comp. $>$ denotes the average number of included mixture components during the second half of the Markov chain simulation under each model. The red dashed lines indicate the prior average number of components as specified by the parameter Λ.

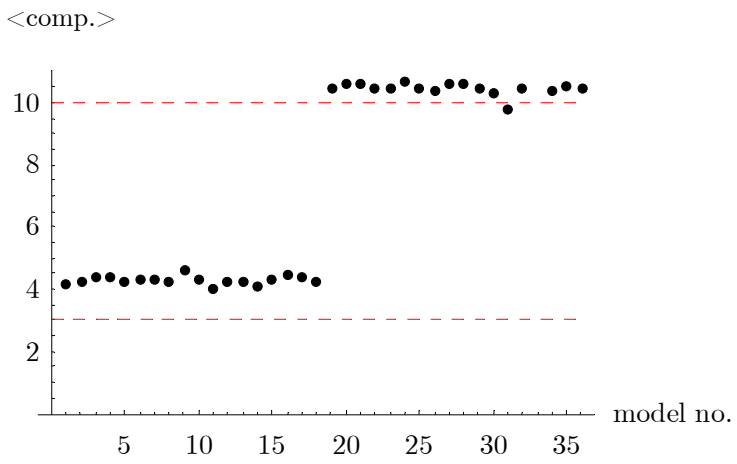
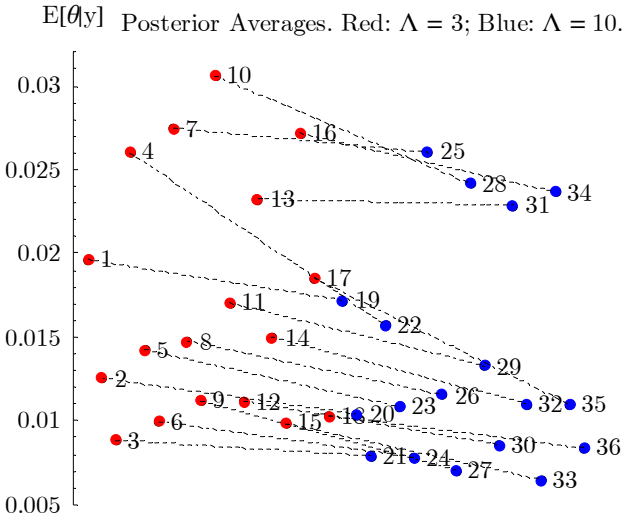


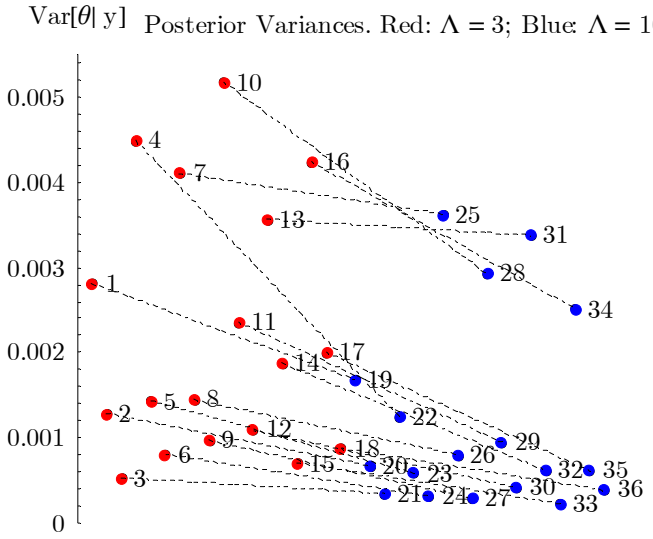
Fig. 12.12. The effect of Λ on the posterior $E[\theta \mid y]$. Two circles linked by a line indicate two models which only differ with respect to Λ. The magnitude of Λ clearly affects the value of $E[\theta \mid y]$. The integer located to the right of each point refers to model number according to table 12.1.



In fig. 12.12 above, the averages $E[\theta \mid y]$ from fig. 12.7 have been reproduced, but this time two models which only differ with respect to the ascribed value of Λ are linked by a

dashed line. Fig. 12.12 shows a clear dependence between the value of Λ and $E[\theta \mid y]$. In general, the expected value of θ is diminished when the value of Λ is increased from 3 to 10. A similar trend is seen with respect to the posterior variance of θ as illustrated in fig. 12.13.

Fig. 12.13. The effect of Λ on $Var[\theta \mid y]$. Two circles linked by a line indicate two models which only differ with respect to the value of Λ. The integer located to the right of each point refers to model number according to table 12.1.



Numerical inaccuracies due to the application of Markov chains of finite length might play a part in the observed differences between $E[\theta \mid y]$ calculated at $\Lambda = 3$ and $\Lambda = 10$. However, as the trends in fig. 12.12 and 12.13 are quite consistent, numerical inaccuracies cannot account fully for the observed patterns. We will return to this problem later.

From table 12.1 it can be seen that models indexed as i, i+6 and i+12 constitute a group of models which are identical with respect to $<m>$ but different with respect to the parameter E and β. Recall that $<m>$ denotes the estimated number of mines, *a priori*, in a randomly selected minefield. In fig. 12.14 and 12.15 the averages $E[\theta \mid y]$ from fig. 12.7 are once again reproduced, but this time the averages are shown as a function of $<m>$. Fig. 12.14 includes all model characterized by $\Lambda = 3$, whereas fig. 12.15 includes all models characterized by $\Lambda = 10$.

Fig. 12.14. $E[\theta \mid y]$ as a function of $< m >$. The mixture models from table 12.1 characterized by $\Lambda = 3$ are distributed into groups of three according to their ascribed value of $< m >$. Solid black line is calculated according to equation (12.01). Red dashed line indicates the true average value of θ.

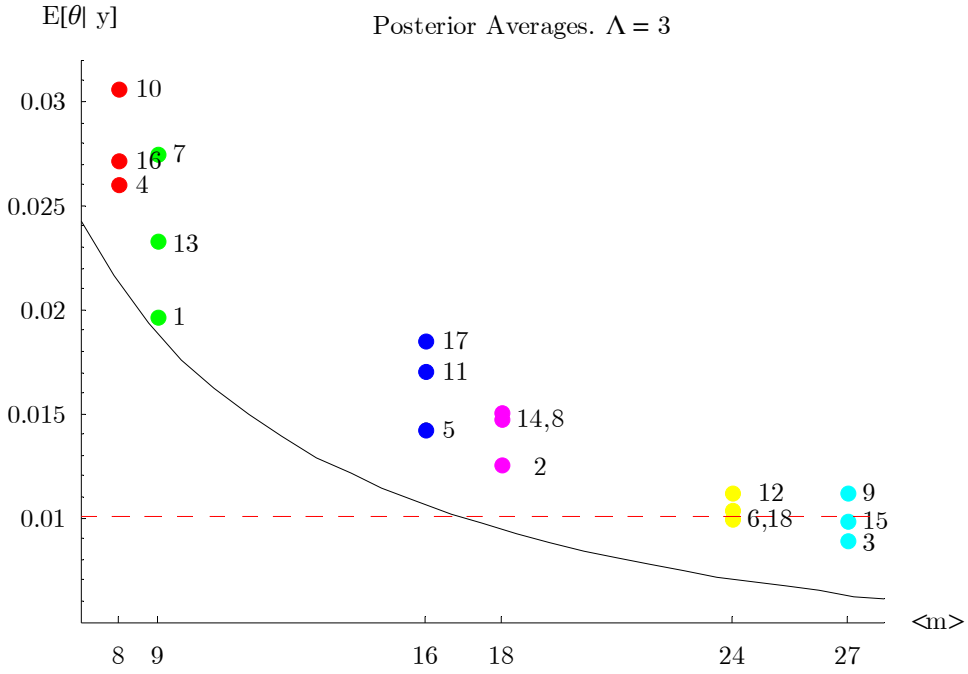
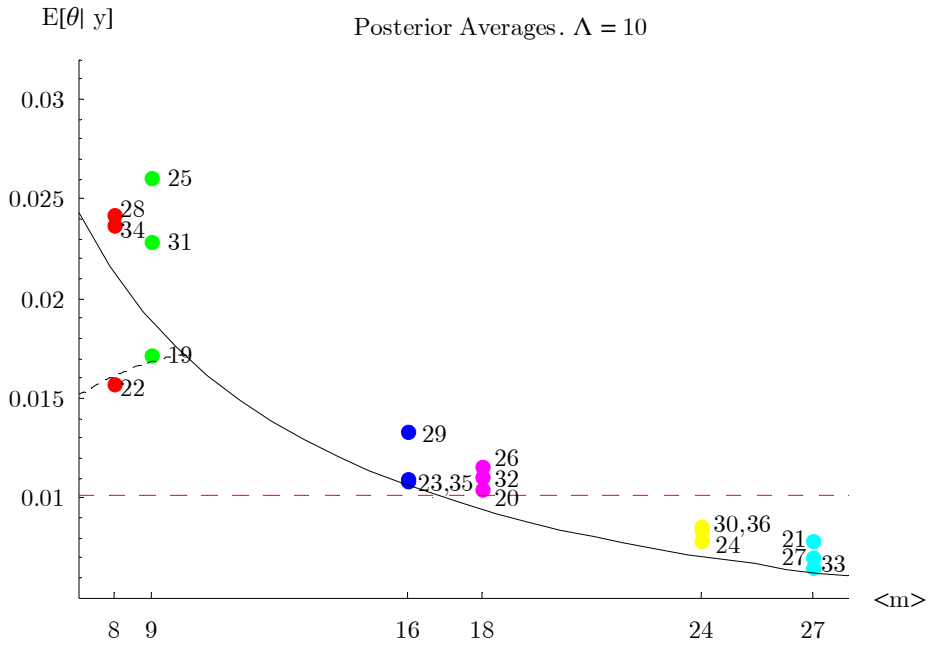


Fig. 12.15. $E[\theta \mid y]$ as a function of $< m >$ for mixture models from table 12.1 characterized by $\Lambda = 10$. Solid black line and dashed red line, see fig. 12.16. Dashed black line indicates the modifying effect of the informative prior $p(\mu, \tau)$ on model 22 and model 19.



Included in fig. 12.14 and 12.15 is also a plot of $E[\theta \mid y_{total}]$ as a function of $<m>$ where $E[\theta \mid y_{total}]$ denotes the estimate of $E[\theta]$ according to the superminefield model from chapter 8, i.e.

$$E[\theta \mid y_{total}] = \frac{y_{total}}{m_{total}} \approx \frac{170}{1000 \cdot < m >}. \qquad \textbf{(12.01)}$$

From fig. 12.14 and 12.15 the following observations can be made:

1) $E[\theta \mid y]$ is highly sensitive to $<m>$ and is roughly inverse proportional to $<m>$. Consequently, the parameters p_0 and M have a substantial effect on $p(\theta \mid y)$.
2) In general $E[\theta \mid y]$ is displaced upwards relative to the estimate given by (12.01). However, the displacement is much stronger when $\Lambda = 3$. The exceptions are model 19 and 22 whose estimates are displaced downwards relative to (12.01).
3) Due to the inverse proportionality between $E[\theta \mid y]$ and $< m >$, the sensitivity of $E[\theta \mid y]$ to variations in $< m >$ diminishes as $< m >$ increases.
4) The effect of $p(\mu, \tau)$ appears to diminish when $< m >$ increases.

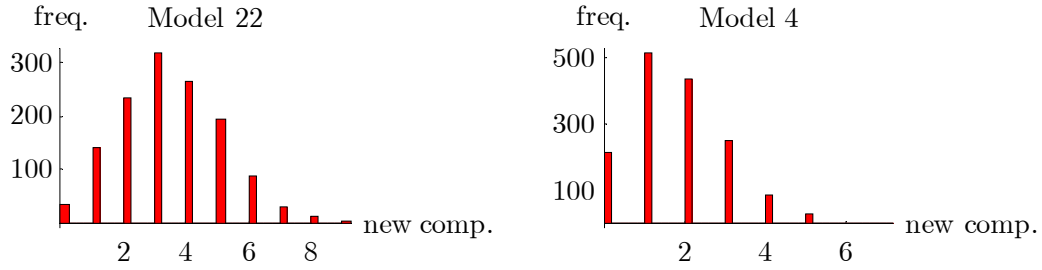
Regarding point 1, the strong dependence of $p(\theta \mid y)$ on $< m >$ means inevitably that very misleading results can be generated if the prior belief about $< m >$ is considerably out of line with the true average value.

Regarding point 2, further tests have to be carried out to detect the reason behind the different results obtained from models specified by $\Lambda = 3$ and $\Lambda = 10$, respectively, and the general displacement from (12.01). However, certain observations point in the direction that the discrepancies are due to premature termination of the Markov chains when $\Lambda = 3$. Firstly, the estimates of $E[\theta]$ provided by model 20, 26 and 32 in fig. 12.15 are very close to the true average value as indicated by the dashed red line in fig. 12.15. This agreement is to be expected as $< m > = 18$ under these models (which is very close to the true average mine content of 17.8). The estimates of $E[\theta]$ from the corresponding models in fig. 17.14 (i.e. model 2, 8 and 14) do not exhibit a similar agreement.

Secondly, a characteristic difference between the point process run at $\Lambda = 3$ and $\Lambda = 10$ is the rate at which new components (i.e., new integers) are introduced into the mixture

model. Fig. 12.16.a and 12.16.b below show for model 22 ($\Lambda = 10$) and model 4 ($\Lambda = 3$) the distribution of the number of new mixture components introduced per iteration based on iterations from the second half of the completed Markov chain simulations. From fig. 12.16 it follows that the average number of new components introduced per iteration is $\simeq 1.7$ ($\Lambda = 3$) and 3.4 ($\Lambda = 10$).

Fig. 12.16. The frequency distribution of the number of new components introduced per iteration. $\Lambda = 10$ for Model 22, and $\Lambda = 3$ for model 4.



A possible consequence of having a large number of new components introduced per iteration is that all parts of the *birth distribution* $p(m)$ are sampled properly including its right-end tail during the finite number of steps in the Markov chain simulation. On the other hand, if the right-end tail is not sampled properly, which may be the case when $\Lambda = 3$, the posterior values of m will, on average, be too low which will increase $E[\theta \mid y]$ as observed in fig. 12.14.

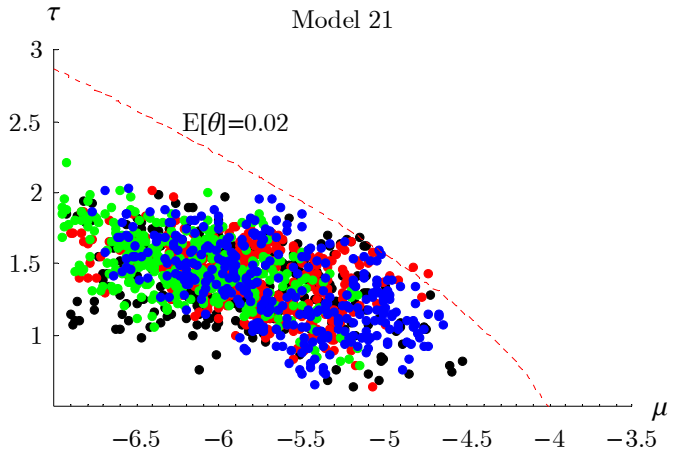
The considerations made above are admittedly speculations, and only through more elaborate tests can the importance of Λ be properly clarified.

The strong sensitivity of $p(\theta \mid y)$ to the prior $p(m)$ is problematic, and for that reason various informative priors $p(\mu, \tau)$ have been used in the present chapter to investigate their possible counter-balancing effect. In connection with point 4 on the previous page it was commented that the effect of $p(\mu, \tau)$ appears to diminish as $< m >$ increases. To elaborate on this observation, note that in fig. 12.14 and 12.15 three models located along the same vertical line are identical with respect to $<m>$ but different with respect to E and β. For convenience we may term such a group of models a triad. The trend observed in fig. 12.14 and 12.15 is that in each triad, the model with $E = 0.02$ generates in general the lowest estimate with respect to $E[\theta]$ (the only exception being model no. 21 in fig.

12.15). As to the order of the estimates of $E[\theta]$ from the remaining two models in each triad, nothing can be concluded in general. However, the variation among the estimates of $E[\theta]$ in a triad is largest when $<m>$ is small and seems to diminish for increasing values of $<m>$.

The above observations can be explained by noting the inverse relationship between $E[\theta \mid y]$ and $<m>$. From (12.01) it is seen that if $< m > \gg 8.5$ it follows that $E[\theta \mid y] \ll 0.02$. Consequently, for the group of mixture models satisfying $< m > \gg 8.5$, the predominant part of the posterior $p(\mu,\tau \mid y)$ will be located along contours in the μ,τ-plane which are well below the 0.02-contour line (i.e. $E[\theta] = 0.02$). This is illustrated in fig. 12.17 in the case of model 21 where the 0.02-contour line has been superimposed on the plot of $p(\mu,\tau \mid y)$ from fig. 12.6. Consequently, the prior $p(\mu,\tau)$ characterized by $E = 0.02$ does not influence the location of $p(\mu,\tau \mid y)$ substantially. The same observation obviously goes with the priors characterized by larger values of E.

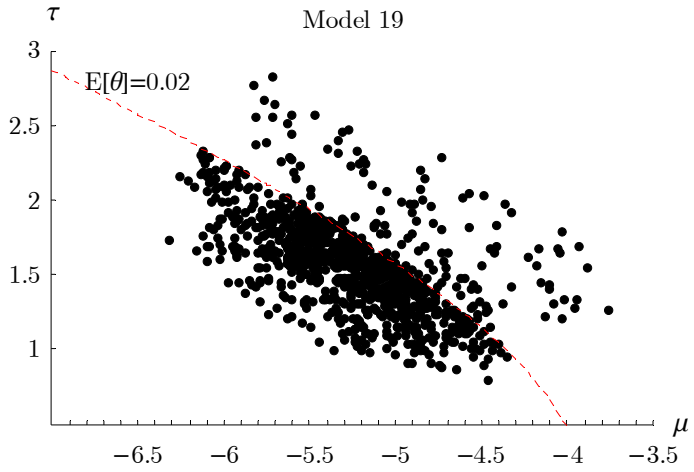
Fig. 12.17. Sampled values from $p(\mu,\tau \mid y)$ obtained under model 21. Due to the large value of $< m >$ in model 21, the values of (μ,τ) sampled during the Markov chain simulation are located well below the contour line $E[\theta] = 0.02$. Consequently, the informative prior $p(\mu,\tau)$ used in model 21 does not influence the location of $p(\mu,\tau \mid y)$.



Now when $<m>$ approaches 8.5 from above, the birth-mechanism from the point process will to an increasing extent give birth to points (λ, m) where m is small. This will be balanced by a posterior $p(\mu,\tau \mid y)$ whose predominant part approaches the contour-line 0.02 from below. If $E = 0.02$ for the prior $p(\mu,\tau)$, areas located in the μ,τ-plane above the 0.02 contour are assigned a lower prior probability relative to areas located below the

contour line. This in turn induces a higher death rate of points (λ, m) characterized by low values of m. The informative prior $p(\mu,\tau)$ thus counter-balances the birth-mechanism, and due to its modifying effect the value of $E[\theta \mid y]$ is displaced downwards relative to what would have been obtained if $p(\mu,\tau)$ was non-informative. Fig. 12.18 below clearly illustrates the imprint of the informative prior in the case of model 19.

Fig. 12.18. Sampled values from $p(\mu,\tau \mid y)$ obtained under model 19. $E = 0.02$ for $p(\mu,\tau)$ and $< m >= 9$. The prior $p(\mu,\tau)$ clearly influences the location of $p(\mu,\tau \mid y)$ which is due to the low value of $< m >$.



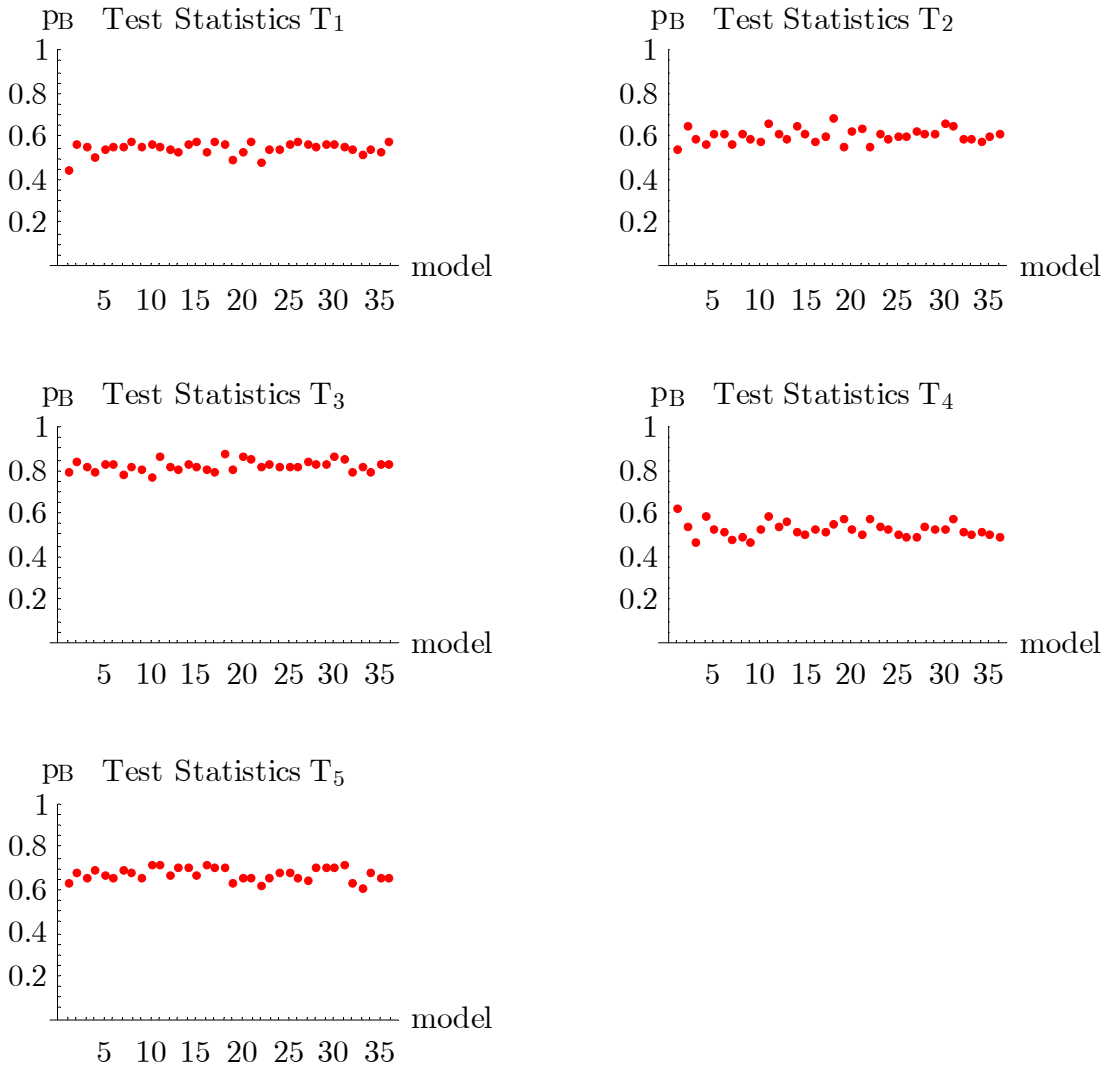
As to the informative priors $p(\mu,\tau)$ tested in the present chapter the following conclusions can be made: If $< m >$ is large, the impact of $p(\mu,\tau)$ on the posterior $p(\theta \mid y)$ is marginal. On the other hand, if the prior $< m >$ is small, the posterior $p(\theta \mid y)$ is very sensitive to changes in $< m >$, and moderate informative priors $p(\mu,\tau)$ do have a counter-balancing impact on the posterior $p(\theta \mid y)$. However, to bring the above conclusions on a firm ground more test calculations should be done. This includes a clarification of the role of the parameter Λ, in particular whether the observed differences between $E[\theta \mid y]$ and relation (12.01) is related to the size of Λ.

12.3 Model Checking and Model Comparisons

The present chapter will close with a few observations dealing with model fit diagnostics in line with the approach taken in chapter 7.

Firstly, fig. 12.19 below illustrates for all models defined in table 12.1 the distribution of the Bayesian p_B-values corresponding to the five test statistics $T_1 \rightarrow T_5$ defined in chapter 9. Like what was seen in the case of the naïve models from chapter 8, none of the calculated p_B-values appearing in fig. 12.19 are extreme, i.e. none of the examined models disqualify themselves due to model misfit under the applied test statistics.

Fig. 12.19. Bayesian p_B-values calculated for the mixture models defined in table 12.1. The five test statistics are identical to the test statistics defined chapter 7.



Secondly, in fig. 12.20 and 12.21 the expected deviance $\hat{D}_{avg}(y)$ and the model complexity parameter $p_D^{(2)}$ are shown for model $1 \rightarrow 36$. Regarding the expected deviance, there is a distinct difference in $\hat{D}_{avg}(y)$ between models characterized by $\Lambda = 3$ (i.e., model $1 \rightarrow 18$) and $\Lambda = 10$ (model $19 \rightarrow 36$) where members of the last group have considerably lower

deviances than members of the first group. Thus the predictive power of the finite mixture model (as measured by the expected deviance) is considerably improved when going from $\Lambda = 3$ to $\Lambda = 10$. The variation of $\hat{D}_{avg}(y)$ among the members of the second group is modest. That is, no single member stands out as superior in terms of predictive power.

Fig. 12.20. Expected deviances calculated for the mixture models defined in table 12.1. Notable is the relatively large expected deviance in model $1 \rightarrow 18$ ($\Lambda = 3$) in comparison with model $19 \rightarrow 36$ ($\Lambda = 10$).

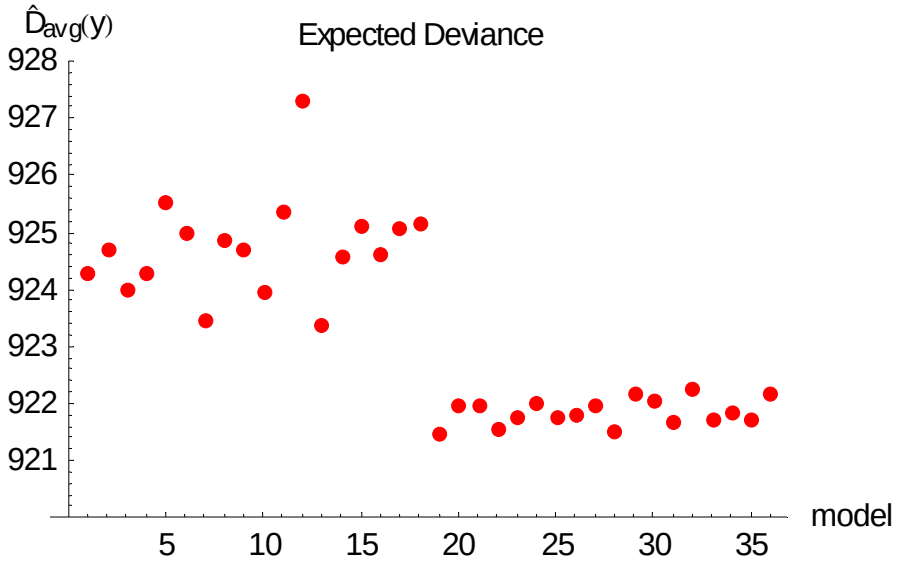
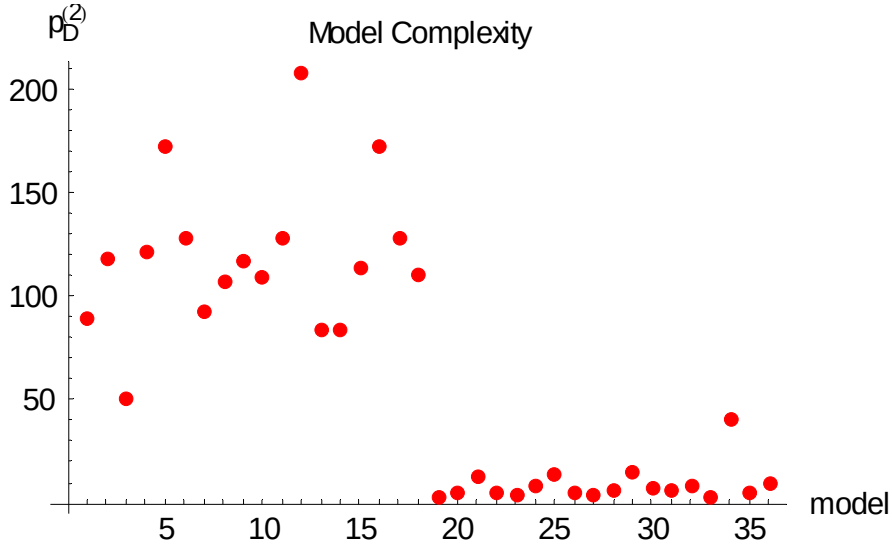


Fig. 12.21. Model complexity parameter $p_D^{(2)}$ calculated for the mixture models defined in table 12.1. The value of $p_D^{(2)}$ is observed to be strongly dependent on the value of the parameter Λ.



The model complexity parameter $p_D^{(2)}$ was in chapter 9 defined as half the posterior variance of the deviance, and $p_D^{(2)}$ was interpreted as the number of unconstrained parameters in the Bayesian model. This interpretation of $p_D^{(2)}$ has lost its meaning in the case of model $1 \rightarrow 18$ where $p_D^{(2)}$ is typically above 100, as seen from fig. 12.21. The value of $p_D^{(2)}$ is strongly dependent on Λ.

Once again it must be concluded that further studies are needed to clarify how and why the choice of Λ affects the progression of the Markov chain simulation. Of particular interest is to determine the value of Λ which minimizes the expected deviance.

12.4 Summary and Conclusions of Finite Mixture Calculations

The overall objective of the chapters 5-12 has been to discuss how a probability distribution $p(\theta)$ can be set up which is solely based on accident statistics from a group of minefields. The major difference between the point of departure taken in chapter 5 and the situation outlined in chapter 4 is therefore the decision maker's uncertainty concerning the degree of mine contamination in the minefields under study. To make inferences about $p(\theta)$ a statistical model explicitly incorporating this additional uncertainty is therefore needed. In this context, the finite mixture model appears to be an obvious choice of model.

The finite mixture model has been introduced in two steps: First, a particular naïve version of the mixture model containing a fixed set of integers $\{m_1, m_2, ..., m_g\}$ was introduced, and Markov chain simulations based on the naïve model were carried out for four different choices of integers. Based on the shortcomings found a more advanced mixture model was subsequently introduced, the major difference from the naïve model being the treatment of the integers $\{m_1, m_2, ..., m_g\}$ as stochastic variables and the averaging over mixture models of varying dimension.

To make the advanced mixture model operational, a hypothetical user has to express his prior belief about the degree of mine contamination in terms of a probability distribution $p(m)$. This information may eventually be complemented by the user's belief about the approximate location of $p(\theta)$ through the specification of the prior $p(\mu, \tau)$. The completed Markov chain simulations based on the advanced mixture model show that the derived

posterior distribution $p(\theta \mid y)$ is highly sensitive to the prior $p(m)$, in particular when the expected number of mines in a randomly selected minefield is small. However, for the range of informative priors $p(\mu,\tau)$ tested, the sensitivity of $p(\theta \mid y)$ to $p(\mu,\tau)$ appears as marginal unless the expected number of mines is small. A third adjustable factor Λ, which largely determines the average number of components included in the mixture model, shows a substantial effect on the location and spread of $p(\theta \mid y)$. It remains to analyze the reason behind the sensitivity of $p(\theta \mid y)$ to Λ and to find an optimal value of Λ.

The observed sensitivity of $p(\theta \mid y)$ to the prior $p(m)$ might from a non-Bayesian point of view appear as open to criticism as the estimate of $p(\theta)$ is certainly biased. It is unclear, however, how an estimate at all can be provided about $p(\theta)$ if previous knowledge about the degree of mine contamination in the minefields under study is completely ignored.

The observed sensitivity of $p(\theta \mid y)$ to a particular choice of $p(m)$ may possibly be reduced if the posteriors $p(\theta \mid y)$ obtained for different choices of $p(m)$ are averaged. It is uncertain, however, how to assign weights to the different posteriors under such an averaging process. An attractive option, in theory at least, is to assign weights according to the relative predictive power of the competing models. Unfortunately, Bayesian p_B-values do not seem to be of any help in this context. Alternatively, the predictive power can be quantified in terms of the expected deviance calculated under each model. As illustrated by the deviance calculations completed in chapter 12, the variation with respect to the expected deviance among the models characterized by $\Lambda = 10$ is unfortunately modest, i.e., the accident statistics do not clearly through the expected deviances rank the competing models. This observation might be due to inadequate sampling during the Markov chain simulations or might simply reflect that the accident statistics do not provide sufficient information for such a ranking. In any case: Further research on the model averaging aspect is needed.

Chapter 13

Integral Evaluation under Markov Chain Simulations

13.1 Introduction

As already noted in chapter 6, the integral which appears in the expression for $f(y \mid m, \mu, \tau)$ defined under the finite mixture model cannot be carried out analytically, and we therefore have to rely on numerical integration. Unfortunately, the use of an approximate summation formula such as a 20-point Gauss-Hermite quadrature generates inaccurate results for certain combinations of the entering variables (y, m, μ, τ). Simply increasing the number of interpolation points reduces the accuracy problem but does not eliminate it, and the integration algorithm is furthermore slowed down. What seems preferable in the present context is therefore an adaptive integration algorithm where the number of included interpolation points varies with (y, m, μ, τ).

In the present chapter we will give a somewhat detailed account of an adaptive numerical integration algorithm (based on work by Crouch et al., 1990) which has been implemented to provide reliable Markov chain simulations. The chapter serves mainly as technical documentation and may without loss of context be skipped on a first reading.

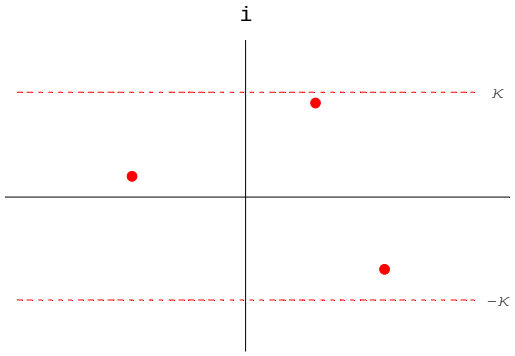
13.2 Numerical Integration Formula

In what follows we will generally factorize $f(y \mid m, \mu, \tau)$ as

$$\begin{aligned} f(y \mid m, \mu, \tau) &= A \cdot \int_{-\infty}^{\infty} g(t)\, e^{-t^2}\, dt \\ &= A \cdot I_g, \end{aligned} \qquad \textbf{(13.01)}$$

where $m \geq 1$, A is a constant independent of t, and I_g is just a shorthand notation for the integral. To set up an adaptive numerical integration algorithm, assume that $g(t)$ from (13.01) is analytical except for isolated poles (not located on the real axis) within some strip $-\kappa < \mathcal{F}(t) < +\kappa$ in the complex plane as sketched in fig. 13.1 below.

Fig. 13.1. The location of the strip $-\kappa < \mathcal{F}(t) < +\kappa$ in the complex plane. Red circles denote isolated poles of the analytical function $g(t)$.



Now, if κ has been chosen such that

$$\int_{-\infty+i\kappa}^{-\infty-i\kappa} g(t)\,e^{-t^2}dt + \int_{\infty-i\kappa}^{\infty+i\kappa} g(t)\,e^{-t^2}dt = 0, \qquad \textbf{(13.02)}$$

it follows from error bound derivations done by Crouch et al. that $\int_{-\infty}^{\infty} g(t)\,e^{-t^2}dt$ can be written as

$$\begin{aligned}
&\int_{-\infty}^{\infty} g(t)\,e^{-t^2}dt = \varepsilon_1(h,\kappa) \\
&+ h\sum_{n=-\infty}^{n=\infty} g(t_0+nh)\,e^{-(t_0+nh)^2} \\
&+ 2\pi i\sum_{below} res\left\{\frac{g(t_j)\,e^{-t_j^2}e^{(-2\pi i(t-t_0)/h)}}{1-e^{(-2\pi i(t-t_0)/h)}}\right\} \qquad \textbf{(13.03)} \\
&+ 2\pi i\sum_{above} res\left\{\frac{g(t_j)\,e^{-t_j^2}e^{(-2\pi i(t-t_0)/h)}}{1-e^{(-2\pi i(t-t_0)/h)}}\right\},
\end{aligned}$$

where the absolute value of $\varepsilon_1(h,\kappa)$ in (13.03) satisfies

$$|\,\varepsilon_1(h,\kappa)\,| \le e^{(\kappa^2-2\pi\kappa/h)} \times \int_{-\infty}^{\infty} \{|\,g(t+i\kappa)\,| + |\,g(t-i\kappa)\,|\}\,e^{-t^2}dt. \qquad \textbf{(13.04)}$$

In (13.03), $\sum_{below} res$ and $\sum_{above} res$ denote the summation over the residues of the poles t_j located respectively below and above the real axis in the strip $-\kappa < \mathcal{F}(t) < +\kappa$. The

parameter h can be considered as a step size of the summation formula, and t_0 is an arbitrary parameter which may be chosen for convenience.

To implement (13.03) on a computer system we necessarily have to truncate the summation which sums over an infinite number of terms. Our final approximation of $\int_{-\infty}^{\infty} g(t)\, e^{-t^2} dt$ can thus be written as

$$\begin{aligned}
&\int_{-\infty}^{\infty} g(t)\, e^{-t^2} dt \approx \\
&h[g(t_0)\, e^{-t_0^2} + \sum_{n=1}^{k_1} g(t_0 + nh)\, e^{-(t_0+nh)^2} + \sum_{n=1}^{k_2} g(t_0 - nh)\, e^{-(t_0-nh)^2}] \\
&+ 2\pi i \sum_{below} res \left\{ \frac{g(t_j)\, e^{-t_j^2} e^{(-2\pi i(t-t_0)/h)}}{1 - e^{(-2\pi i(t-t_0)/h)}} \right\} \\
&+ 2\pi i \sum_{above} res \left\{ \frac{g(t_j)\, e^{-t_j^2} e^{(-2\pi i(t-t_0)/h)}}{1 - e^{(-2\pi i(t-t_0)/h)}} \right\}.
\end{aligned} \tag{13.05}$$

The use of the summation formula in (13.05) in place of $\int_{-\infty}^{\infty} g(t)\exp(-t^2)\, dt$ implies three sources of error: 1) the error term $\varepsilon_1(h,\kappa)$ given by (13.04) whose magnitude can be controlled by appropriate choices of κ and h; 2) the truncation error $\varepsilon_2 = \varepsilon_2(k_1, k_2, h)$ due to the truncation of the summation which sums over an infinite number of terms; 3) the rounding error ε_3 which is due to the finite precision of the applied computer system. The acceptable bounds on the three sources of error will obviously depend on our tolerance concerning the absolute error on the evaluation of $f(y \mid m, \mu, \tau)$. Thus before we proceed with the determination of κ, h, k_1 and k_2 it is essential to clarify the relationship between the absolute error on $f(y \mid m, \mu, \tau)$ and the absolute error on $\int_{-\infty}^{\infty} g(t)\exp(-t^2)\, dt$.

13.3 Error Analysis

In what follows $f^*(y \mid m, \mu, \tau)$ denotes the output from a numerical computation of $f(y \mid m, \mu, \tau)$. In general $f^*(y \mid m, \mu, \tau)$ will be different from $f(y \mid m, \mu, \tau)$ due to various numerical errors. If $f(y \mid m, \mu, \tau) = A\int_{-\infty}^{\infty} g(t)\exp(-t^2)\, dt \equiv A \cdot I_g$, the absolute error amounts

to $| f(y \mid m,\mu,\tau) - f^*(y \mid m,\mu,\tau) | \approx A^*\Delta I_g + I_g^*\Delta A$, where A^* and I_g^* denotes the numerical approximation to A and I_g, respectively, and ΔA and ΔI_g denotes the absolute error on A and I_g, respectively.

Let us as our point of departure aim at an integration algorithm which can provide $f(y \mid m,\mu,\tau)$ with an absolute error not exceeding say 10^{-14}. Consequently,

$$\begin{aligned} & A^*\Delta I_g + I_g^*\Delta A \le 10^{-14} \\ & \Updownarrow \\ & \Delta I_g \le \frac{10^{-14}}{A^*} - \frac{I_g^*}{A^*}\Delta A. \end{aligned} \qquad \textbf{(13.06)}$$

To simplify the following discussion we will ignore the term $\frac{I_g^*}{A^*}\Delta A$ from (13.06). To justify this, let $A^* = a \cdot 10^m$, i.e. A^* is a floating-point real number with a precision of say 16 digits. If by assumption the product $A^* \cdot I_g^* \approx A \cdot I_g \in [0;1]$, we can write $A^* I_g^* = b \cdot 10^{-x}$, where $x \in \mathbb{N}_0$. It follows that (13.06) can be rewritten as

$$\Delta I_g \le \frac{1}{a}10^{-14-m} - \frac{b}{a^2}10^{-x-2m}\Delta A. \qquad \textbf{(13.07)}$$

Now, if the only numerical error associated with A^* is due to the rounding error in relation to the storage of A, this implies that $\Delta A \le 0.5 \cdot 10^{m-16}$ given a precision of 16 digits. Consequently,

$$\frac{b}{a^2}10^{-x-2m}\Delta A \le 0.5\frac{b}{a^2}10^{-x-m-16} \le \frac{1}{a}10^{-14-m}(0.05 \cdot 10^{-x}). \qquad \textbf{(13.08)}$$

From (13.08) it follows that the second term in (13.07) is much smaller than the first term given that $A^* I_g^* \le 1$, and the second term in (13.07) will consequently be ignored. Our final demand on ΔI_g is therefore

$$\Delta I_g \le \frac{1}{a}10^{-14-m}, \qquad \textbf{(13.09)}$$

or simply

$$\Delta I_g \le 10^{-14-m}. \qquad \textbf{(13.10)}$$

In other words, to guarantee that $f(y \mid m, \mu, \tau)$ is calculated with an absolute error being less or equal to 10^{-14}, the parameters κ, h, k_1 and k_2 should be chosen such that

$$\mid \varepsilon_1(\kappa, h) + \varepsilon_2(k_1, k_2, h) + \varepsilon_3 \mid \leq 10^{-14-m}. \tag{13.11}$$

It is instructive to derive an upper bound on the rounding error due to the finite precision of the applied computer system. With that in mind, the process of replacing the integral I_g by an approximating summation followed by the storage of the individual terms on the computer will be sketched as

$$I_g \approx \sum_{i=1}^{k} g_i \approx \sum_{i=1}^{k} g_i^*, \tag{13.12}$$

where $\sum_{i=1}^{k} g_i$ is a shorthand notation for the various summations in (13.05), and g_i^* denotes the stored value of g_i. Whereas the difference between I_g and $\sum_{i=1}^{k} g_i$ is accounted for by $\varepsilon_1(\kappa, h)$ and $\varepsilon_2(k_1, k_2, h)$, the difference between $\sum_{i=1}^{k} g_i$ and $\sum_{i=1}^{k} g_i^*$ is accounted for by ε_3. Thus the only difference between g_i and g_i^* is by assumption due to the finite precision of the applied computer system.

To get a bound on ε_s, note that

$$\mid \varepsilon_s \mid = \mid \sum_{i=1}^{k} g_i - \sum_{i=1}^{k} g_i^* \mid \leq \sum_{i=1}^{k} \mid g_i - g_i^* \mid \leq k \mid \varepsilon^{\max} \mid, \tag{13.13}$$

where $\varepsilon^{\max}$ denotes the largest rounding error. If the parameters κ, h, k_1 and k_2 have been chosen such that the inequality $\mid \varepsilon_1(\kappa, h) + \varepsilon_2(k_1, k_2, h) \mid \leq 10^{-14-m}$ is satisfied, it follows that $\mid I_g - \sum_{i=1}^{k} g_i \mid \leq 10^{-14-m}$ from which it follows that

$$\begin{aligned} \sum_{i=1}^{k} g_i &\leq I_g + 10^{-14-m} \\ &= \frac{b}{a} 10^{-x-m} + 10^{-14-m} \\ &\Downarrow \\ g_{i,\max} &\leq \frac{b}{a} 10^{-x-m} + 10^{-14-m} \leq 10^{-m+1} + 10^{-14-m} \approx 10^{-m+1}, \end{aligned} \tag{13.14}$$

where $g_{i,\max}$ denotes the largest component from the summation $\sum_{i=1}^{k} g_i$. In (13.14) we have written $I_g = \frac{b}{a}10^{-x-m}$ similar to what was done in relation to (13.07).

From (13.14) it follows that

$$| \varepsilon^{\max} | = | g_{i,\max} - g^{*}_{i,\max} | \leq 0.5 \cdot 10^{-m-15} \qquad \textbf{(13.15)}$$

and

$$| \varepsilon_s | \leq 0.5 \cdot k \cdot 10^{-m-15}. \qquad \textbf{(13.16)}$$

Thus the rounding error ε_s is (not surprisingly) proportional to k, i.e. the number of terms included in the approximating summation. Furthermore, if $k \geq 20$ the rounding error may exceed the acceptable tolerance on 10^{-14}. It is therefore essential to choose the step size h as large as possible to ensure that the number of terms to be included in the quadrature formula becomes as small as possible.

With the above aim in mind we now turn to the determination of the parameters κ, h, k_1 and k_2. A wide range of combinations of (κ, h, k_1, k_2) might possibly satisfy the inequality $| \varepsilon_1(\kappa, h) + \varepsilon_2(k_1, k_2, h) + \varepsilon_3 | \leq 10^{-14-m}$, but the fastest integration algorithm is obviously obtained if $k_1 + k_2$ is as small as possible. Thus the optimal combination of parameters might be obtained as a solution to a (restricted) mixed integer minimization problem if explicit bounds for $\varepsilon_1(\kappa, h)$ and $\varepsilon_2(k_1, k_2, h)$ are available. However, the search for the optimal solution is time-consuming, and in the present context we have as presented below applied a less sophisticated approach which seems to give acceptable solutions in relation to the completed Markov chain simulations.

In the approach followed in the present work it is simply demanded that $| \varepsilon_1(\kappa, h) | \leq 10^{-16-m}$ and $| \varepsilon_2(k_1, k_2, h) | \leq 10^{-15-m}$. Consequently, it is guaranteed that the error inherent in the applied quadrature formula is within the fixed limit of 10^{-14-m}. The error bound on ε_s might however exceed the limit 10^{-14-m} when a large number of terms are required to satisfy the inequality $| \varepsilon_2(k_1, k_2, h) | \leq 10^{-15-m}$. This occasional violation appears not to have affected the progression of the completed Markov chain simulations.

The problem might furthermore be eliminated if the integral evaluations are done on a machine with a larger precision than 16 digits.

From the observations made above it is evident that the step size h is a parameter of central importance. To identify the largest possible value of h an explicit expression for $\varepsilon_1(\kappa,h)$ is needed. However, it is intuitively clear that the allowable step size must increase when the bound $|\,\varepsilon_1(\kappa,h)\,| \le 10^{-16-m}$ is relaxed. Through the choice of $A = a\cdot 10^m$ (recall that $f(y \mid m,\mu,\tau) = A\cdot I_g$) it is therefore possible to affect the allowable step size. From the error bound expression it is clear that one should search for factorizations where $A \le 1$, i.e. $m \le 0$.

In what follows, three different factorizations will be derived which ensure that for any $f(y \mid m,\mu,\tau)$ there exists a factorization such that $A \le 1$.

13.4 Factorization of $f(y \mid m,\mu,\tau)$

By definition $f(y \mid m,\mu,\tau)$ is equal to

$$f(y \mid m,\mu,\tau) = \begin{cases} I_0(y) \text{ if } m = 0 \\ \binom{m}{y}\dfrac{1}{\sqrt{2\pi}\tau}\displaystyle\int_{-\infty}^{\infty}\frac{\exp(\alpha y)}{(1+\exp(\alpha))^m}\exp\left(\frac{-(\alpha-\mu)^2}{2\tau^2}\right)d\alpha \text{ if } m > 0. \end{cases} \qquad \textbf{(13.17)}$$

Only the integral expression in (13.17) is of interest in the following derivations. The integral will as a matter of convenience be rewritten as

$$g(y \mid m,\mu,\tau) = \frac{1}{\sqrt{2\pi}\tau}\int_{-\infty}^{\infty}\frac{\exp(xy)}{(1+\exp(x))^m}\exp\left(\frac{-(x-\mu)^2}{2\tau^2}\right)dx. \qquad \textbf{(13.18)}$$

The subject in the present paragraph is therefore through some transformation to accomplish the factorization $g(y \mid m,\mu,\tau) = A\cdot\int_{-\infty}^{\infty} g(t)\,e^{-t^2}dt$. It turns out that three different factorizations suffice to accomplish the goal stated in the previous paragraph. The three factorizations will be referred to as *fac 1*, *fac 2* and *fac 3*.

Concerning *fac 1*: by use of the two transformations

$$1)\ z = x - \mu - y\tau^2$$
$$2)\ t = \frac{z}{\sqrt{2}\tau}$$

in (13.18) it can easily be shown that (13.18) is factorized as

$$g(y \mid m, \mu, \tau) = \frac{\exp(\mu y + \frac{(y\tau)^2}{2})}{\sqrt{\pi}} \times \int_{-\infty}^{\infty} \frac{1}{(1+\exp(\alpha+\beta t))^m} \exp(-t^2)\, dt, \qquad \textbf{(13.19)}$$

where $\alpha = \mu + y\tau^2$ and $\beta = \sqrt{2}\tau$. We have thus obtained a factorization where

$$A = \frac{\exp(\mu y + \frac{(y\tau)^2}{2})}{\sqrt{\pi}}, \qquad \textbf{(13.20)}$$

and

$$g(t) = \frac{1}{(1+\exp(\alpha+\beta t))^m}. \qquad \textbf{(13.21)}$$

Similarly, if only the transformation $t = \frac{x}{\sqrt{2}\tau}$ is used, (13.18) is factorized as

$$g(y \mid m, \mu, \tau) = \frac{\exp(\frac{-\mu^2}{2\tau^2})}{\sqrt{\pi}} \times \int_{-\infty}^{\infty} \frac{\exp(\alpha t)}{(1+\exp(\beta t))^m} \exp(-t^2)\, dt, \qquad \textbf{(13.22)}$$

where $\alpha = \sqrt{2}\tau(y + \frac{\mu}{\tau^2})$ and $\beta = \sqrt{2}\tau$. (13.22) will be termed *fac 2*.

Finally, by use of the identity

$$\int_{-\infty}^{\infty} \frac{1}{(1+\exp(\alpha+\beta t))^m} \exp(-t^2)\, dt = \exp(-m\alpha + \frac{m^2\beta^2}{4}) \times \int_{-\infty}^{\infty} \frac{1}{(1+\exp(-\alpha+\frac{1}{2}m\beta^2+\beta t))^m} \exp(-t^2)\, dt, \qquad \textbf{(13.23)}$$

in (13.19), a third factorization can be derived which takes the form

$$g(y \mid m,\mu,\tau) = \frac{\exp((y-m)[\frac{(y-m)\tau^2}{2}+\mu])}{\sqrt{\pi}} \times \int_{-\infty}^{\infty} \frac{1}{(1+\exp(\gamma+\beta t))^m} \exp(-t^2)\, dt, \tag{13.24}$$

where $\gamma = \tau^2(m-y)-\mu$ and $\beta = \sqrt{2}\tau$. (13.24) is termed *fac 3*. The three factorizations are summarized in table 13.1. In paragraph 13.6 it will be shown that the three tabulated factorizations actually ensure that for any $f(y \mid m,\mu,\tau)$ there exists an $A \leq 1$.

Table 13.1. Derived factorizations.

Fac.	A	$g(t)$	α,β,γ
fac 1	$\frac{\exp(\mu y + \frac{(y\tau)^2}{2})}{\sqrt{\pi}}$	$\frac{1}{(1+\exp(\alpha+\beta t))^m}$	$\alpha = \mu + y\tau^2$ $\beta = \sqrt{2}\tau$
fac 2	$\frac{\exp(\frac{-\mu^2}{2\tau^2})}{\sqrt{\pi}}$	$\frac{\exp(\alpha t)}{(1+\exp(\beta t))^m}$	$\alpha = \sqrt{2}\tau(y+\frac{\mu}{\tau^2})$ $\beta = \sqrt{2}\tau$
fac 3	$\frac{\exp((y-m)[\frac{(y-m)\tau^2}{2}+\mu])}{\sqrt{\pi}}$	$\frac{1}{(1+\exp(\gamma+\beta t))^m}$	$\gamma = \tau^2(m-y)-\mu$ $\beta = \sqrt{2}\tau$

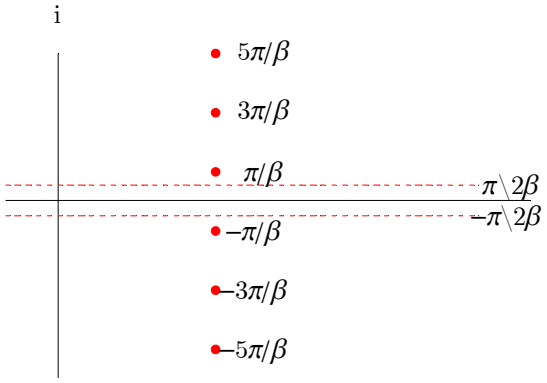
13.5 Adaptive Numerical Integration Algorithm

After having derived three different factorizations what remains is to choose h,κ,k_1 and k_2 such that $\mid \varepsilon_1(\kappa,h) \mid \leq 10^{-16-m}$ and $\mid \varepsilon_2(k_1,k_2,h) \mid \leq 10^{-15-m}$. The specific choice of parameters will in general depend on the used factorization.

Factorization 1

To determine κ and h such that $\mid \varepsilon_1(h,\kappa) \mid \leq 10^{-16-m}$, note that $g(t)$ is defined as $(1+\exp(\alpha+\beta t))^{-m}$ whose poles are the points $t_j = {(i\pi(2j-1)-\alpha)}/{\beta}$ where $j = \pm 0,1,2,...$ The location of a subset of the poles in the complex plane is sketched in fig. 13.2 for the case $\beta > 0$.

Fig. 13.2. The distribution of isolated poles in the complex plane in the case $g(t) = (1 + \exp(\alpha + \beta t))^{-m}$. The confinement of κ to the strip bounded by the red dashed lines eliminates the summation over the poles in (13.05).



The poles of $g(t)$ are not simple poles but poles of order m which complicates the computation of the residues in (13.05). It is therefore convenient to restrict the parameter κ to the interval $0 < \kappa \leq \pi/2\beta$, as indicated in fig. 13.2. This restriction implies two things: Firstly, as no poles are located within the strip $-\kappa < \mathcal{F}(t) < \kappa$ if $0 < \kappa \leq \pi/2\beta$, the last two summations in (13.05) can be ignored. Secondly, it can easily be verified that $| g_1(t \pm i\kappa) | \leq 1$ if $0 < \kappa \leq \pi/2\beta$. Consequently, (13.02) is satisfied. Concerning the error term $\varepsilon_1(h,\kappa)$ we furthermore have that [Crouch et al., 1990, p. 465]

$$\begin{aligned} | \varepsilon_1(h,\kappa) | &\leq e^{(\kappa^2 - 2\pi\kappa/h)} \times \int_{-\infty}^{\infty} \{ | g_1(t + i\kappa) | + | g_1(t - i\kappa) | \} e^{-t^2} dt \\ &\leq e^{(\kappa^2 - 2\pi\kappa/h)} 2\sqrt{\pi}. \end{aligned} \qquad \textbf{(13.25)}$$

Note that (13.25) is minimized if $\kappa = \pi/h$. An efficient choice of the parameters κ and h can according to Crouch et al. be made by use of the following algorithm: We tentatively set $\kappa = \pi/h$ from which it follows that

$$| \varepsilon_1(h,\kappa) | \leq e^{(-\frac{\pi^2}{h^2})} 2\sqrt{\pi}. \qquad \textbf{(13.26)}$$

To determine the step size h, we simply demand that $e^{(-\frac{\pi^2}{h^2})} 2\sqrt{\pi} \leq 10^{-16-m} \equiv \eta$ from which it follows that

$$h \leq \frac{\pi}{\sqrt{\log(\frac{2\sqrt{\pi}}{\eta})}}. \tag{13.27}$$

Consequently, h is set to the upper bound $\frac{\pi}{\sqrt{\log(\frac{2\sqrt{\pi}}{\eta})}}$ as h is wanted as large as possible.

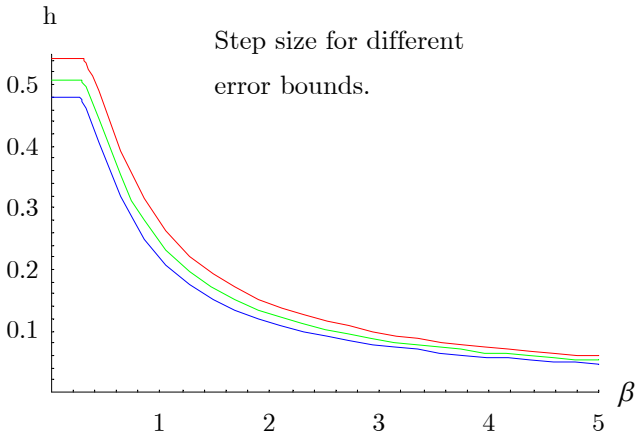
Now, it might turn out that the above choice of κ and h causes the condition $\kappa = \pi/h \leq \pi/2\beta$ to be violated. In that case we make the alternative choice $\kappa = \pi/2\beta$ which by the same set of arguments as before leads to the requirement

$$h \leq \frac{4\beta\pi^2}{(\pi^2 + 4\beta^2 \log(\frac{2\sqrt{\pi}}{\eta})}, \tag{13.28}$$

from which h is determined.

Fig. 13.3 illustrates the step size h as a function of β for three different bounds $\eta = 10^{-16-m}$, i.e. $m = 2, 0, -2$. The step size clearly increases when m decreases. The assignment $\kappa = \pi/h$ is only active when $\beta \leq \frac{\pi}{2}(\log(\frac{2\sqrt{\pi}}{\eta}))^{-1/2}$.

Fig. 13.3. The step size h as a function of β under the bound $\eta = 10^{-16-m}$. Red curve: m = -2; green curve: $m = 0$; blue curve: $m = 2$.



After having determined the step size of the quadrature formula, we next move to the determination of the parameters k_1 and k_2 such that $|\,\varepsilon_2(k_1, k_2, h)\,| \leq 10^{-15-m}$. To derive an error bound of $\varepsilon_2(k_1, k_2, h)$ it is convenient to write out the original summation as

$$h\sum_{n=-\infty}^{n=\infty} g_1(t_0+nh)\,e^{-(t_0+nh)^2} = \\ h\cdot g_1(t_0)\,e^{-t_0^2} + h\cdot\sum_{n=1}^{\infty} g_1(t_0+nh)\,e^{-(t_0+nh)^2} + h\cdot\sum_{n=1}^{\infty} g_1(t_0-nh)\,e^{-(t_0-nh)^2} \tag{13.29}$$

Firstly, to make sure that the terms from each of the two summations in (13.29) diminish in magnitude for increasing values of n, the parameter t_0 is chosen such that $g_1(t)e^{-t^2}$ attains its maximum at $t=t_0$. The determination of t_0 is accomplished by means of a numerical optimization algorithm.

Secondly, let the truncation of (13.29) be written as

$$h\cdot g_1(t_0)\,e^{-t_0^2} + h\cdot\sum_{n=1}^{k_1} g_1(t_0+nh)\,e^{-(t_0+nh)^2} + h\cdot\sum_{n=1}^{k_2} g_1(t_0-nh)\,e^{-(t_0-nh)^2}. \tag{13.30}$$

The error caused by the truncation is a function of the summation limits k_1 and k_2. The error due to the first truncated summation from (13.30) is given by the sum

$$\begin{aligned} & h\cdot\sum_{n=k_{1+1}}^{\infty} g_1(t_0+nh)\,e^{-(t_0+nh)^2} \\ &= h\cdot\sum_{n=k_{1+1}}^{\infty} \frac{1}{\left(1+e^{\alpha+\beta(t_0+nh)}\right)^m} e^{-(t_0+nh)^2} \\ &\le h\cdot\frac{1}{\left(1+e^{\alpha+\beta(t_0+(k_1+1)h)}\right)^m}\sum_{n=k_{1+1}}^{\infty} e^{-(t_0+nh)^2}. \end{aligned} \tag{13.31}$$

Introducing the variable $j=n-k_1-1$, the last term from (13.31) can be rewritten as

$$\begin{aligned} & h\frac{1}{\left(1+e^{\alpha+\beta(t_0+(k_1+1)h)}\right)^m}\sum_{n=k_{1+1}}^{\infty} e^{-(t_0+nh)^2} \\ &= h\frac{1}{\left(1+e^{\alpha+\beta(t_0+(k_1+1)h)}\right)^m}\sum_{j=0}^{\infty} e^{-(t_0+(j+k_1+1)h)^2} \\ &\le h\frac{e^{-(t_0+(k_1+1)h)^2}}{\left(1+e^{\alpha+\beta(t_0+(k_1+1)h)}\right)^m}\sum_{j=0}^{\infty} e^{-2h(t_0+(k_1+1)h)\,j} \\ &= h\frac{e^{-(t_0+(k_1+1)h)^2}}{\left(1+e^{\alpha+\beta(t_0+(k_1+1)h)}\right)^m}\sum_{j=0}^{\infty} x^j, \end{aligned} \tag{13.32}$$

where $x=e^{-2h(t_0+(k_1+1)h)}$.

Assuming that $x<1$ which is equivalent to $t_0+(k_1+1)h>0$ we finally get that

$$\begin{aligned} & h\frac{e^{-(t_0+(k_1+1)h)^2}}{(1+e^{\alpha+\beta(t_0+(k_1+1)h)})^m}\sum_{j=0}^{\infty}x^j \\ &= h\frac{e^{-(t_0+(k_1+1)h)^2}}{(1+e^{\alpha+\beta(t_0+(k_1+1)h)})^m}\frac{1}{1-x} \\ &= h\frac{e^{-\gamma_1{}^2}}{(1+e^{\alpha+\beta\gamma_1})^m(1-e^{-2h\gamma_1})}, \end{aligned} \qquad \textbf{(13.33)}$$

where $\gamma_1=t_0+(k_1+1)h>0$ in (13.33).

Concerning the second truncated summation it can be shown with a similar set of arguments that

$$\begin{aligned} & h\cdot\sum_{n=k_2+1}^{\infty}g_1(t_0-nh)e^{-(t_0-nh)^2} \\ &\le h\frac{e^{-\gamma_2{}^2}}{(1-e^{-2h\gamma_2})}, \end{aligned} \qquad \textbf{(13.34)}$$

where $\gamma_2=t_0-(k_2+1)h<0$ in (13.34).

So to conclude: If the summation $h\sum_{n=-\infty}^{n=\infty}g_1(t_0+nh)e^{-(t_0+nh)^2}$ is replaced by two truncated summations including k_1 and k_2 terms, respectively, the induced truncation error $\varepsilon_2(k_1,k_2,h)$ is bounded according to the inequality

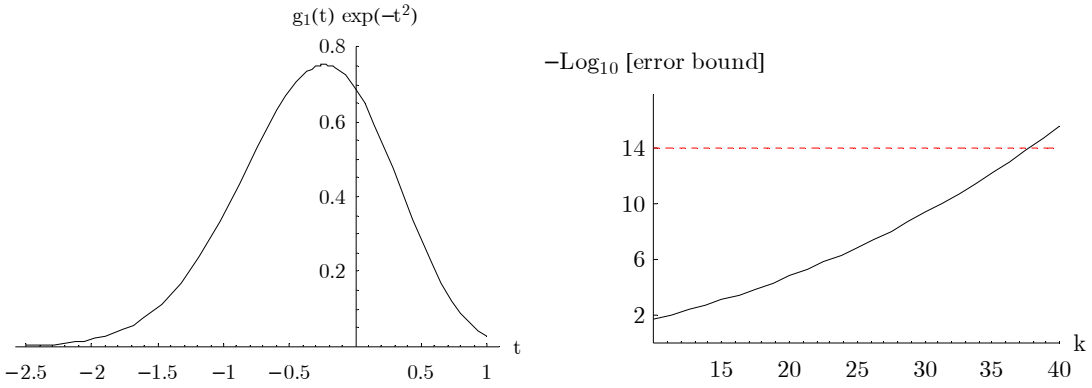
$$|\,\varepsilon_2(k_1,k_2,h)\,|\le h[\frac{e^{-\gamma_1{}^2}}{(1+e^{\alpha+\beta\gamma_1})^m(1-e^{-2h\gamma_1})}+\frac{e^{-\gamma_2{}^2}}{(1-e^{-2h\gamma_2})}] \qquad \textbf{(13.35)}$$

given that $\gamma_1=t_0+(k_1+1)h>0$ and $\gamma_2=t_0-(k_2+1)h<0$.

If we simply set $k_1=k_2=k$, the two restrictions on k are simultaneously satisfied if $k>\frac{|\,t_0\,|}{h}-1$. As an example, fig. 13.4 below illustrates the shape of $g_1(t)\exp(-t^2)$ for the case $f(y\,|\,m,\mu,\tau)=f(1\,|\,6,-5,1.5)$. In fig. 13.5 the corresponding bound on $|\,\varepsilon_2(k,h)\,|$ (termed "error bound" in fig. 13.5) calculated according to (13.35) is shown. It emerges

from fig. 13.5 that $k \geq 38$ if $|\,\varepsilon_2(k,h)\,| \leq 10^{-15-m}$. This number can be somewhat reduced if k_1 and k_2 are treated separately. That is, a closer inspection of the contribution from each of the terms in (13.35) reveals that if $k_1 = 24$ and $k_2 = 38$ we still get $|\,\varepsilon_2(k,h)\,| \leq 10^{-15-m}$.

Fig. 13.4 (left): $g_1(t)\exp(-t^2)$ in the case $f(y \mid m,\mu,\tau) = f(1 \mid 6,-5,1.5)$. Fig. 13.5 (right): -$\text{Log}_{10}$ to the error bound given by (13.35) as a function of k. $t_0 = -0.24, h = 0.137$, $A_1 \approx 10^{-1}$.



The identification of the set (k_1,k_2) which reduces $k_1 + k_2$ and keeps the error bound (13.35) below the acceptable limit demands an optimization algorithm. To avoid the time-consuming optimization step we have in the present implementation simply set $k_1 = k_2 = k$, and the subsequent determination of k is straightforward.

Factorization 2

Concerning *fac 2*, the function $g(t)$ is defined as $\dfrac{\exp(\alpha t)}{(1+\exp(\beta t))^m}$ whose poles are the points $t_j = \dfrac{i\pi(2j-1)}{\beta}$, where $j = \pm 0,1,2,...$ Thus, the distribution of the poles follows the same pattern as in fig. 13.2 but are now restricted to be located along the imaginary axis. Consequently, if we once again restrict κ to the interval $0 < \kappa \leq \pi/2\beta$, the contribution from the residues in (13.05) disappears. It can furthermore be shown that $|\,g_2(t \pm i\kappa)\,| \leq 1$ for all t if $\alpha > 0$ and $\beta > \alpha/m$. It follows that the bound on $\varepsilon_1(\kappa,h)$ derived in relation to *fac 1* also holds for *fac 2.*

A bound on the truncation error $\varepsilon_2(k_1,k_2,h)$ can be derived along lines similar to those being followed in connection with *fac 1.* Omitting the details we will simply state that

$$
\begin{aligned}
& h \cdot \sum_{n=k_{1+1}}^{\infty} g_2(t_0 + nh)\, e^{-(t_0+nh)^2} \\
= h \cdot & \sum_{n=k_{1+1}}^{\infty} \frac{e^{\alpha(t_0+nh)}}{(1+e^{\beta(t_0+nh)})^m} e^{-(t_0+nh)^2} \\
\leq h \cdot & \frac{e^{mt_0(\frac{\alpha}{m}-\beta)} e^{-\gamma_1^2}}{(1-e^{-2h\gamma_1})},
\end{aligned}
\tag{13.36}
$$

where $\gamma_1 = t_0 + (k_1+1)h > 0$, and

$$
\begin{aligned}
& h \cdot \sum_{n=k_{1+1}}^{\infty} g_2(t_0 - nh)\, e^{-(t_0-nh)^2} \\
\leq h \cdot & \frac{e^{-\gamma_2^2+\alpha\gamma_2}}{(1-e^{-2h\gamma_2})},
\end{aligned}
\tag{13.37}
$$

where $\gamma_2 = t_0 - (k_2+1)h < 0$. Consequently

$$
| \varepsilon_2(k_1,k_2,h) | \leq h[\frac{e^{mt_0(\frac{\alpha}{m}-\beta)} e^{-\gamma_1^2}}{(1-e^{-2h\gamma_1})} + \frac{e^{-\gamma_2^2+\alpha\gamma_2}}{(1-e^{-2h\gamma_2})}]. \tag{13.38}
$$

Just as in the case of *fac 1* we have in the present implementation set $k_1 = k_2 = k$, and given the step size h the subsequent determination of k is straightforward.

As $g(t)$ under *fac 3* in structure is similar to $g(t)$ under *fac 1*, there is no need for further error bound derivations.

13.6 Proof of Factorization Property

In paragraph 13.4 it was claimed that for any $f(y \mid m,\mu,\tau)$, there exists among the factorizations listed in table 13.1 at least one which satisfies that $A \leq 1$. The proof is straightforward and will be given here.

In the case of *fac 2* it was stated in paragraph 13.5 that $| g_2(t \pm i\kappa) | \leq 1$ for all t if $\alpha > 0$ and $\beta > \alpha/m$. As $\alpha = \sqrt{2}\tau(y + \frac{\mu}{\tau^2})$ and $\beta = \sqrt{2}\tau$, the restrictions on α and β imply that

$$
0 < y + \frac{\mu}{\tau^2} < m \tag{13.39}
$$

Consider now an arbitrary set of parameters (y, m, μ, τ) from a given $f(y \mid m, \mu, \tau)$. If (y, m, μ, τ) satisfies (13.39) it follows that $A \leq 1$ in the case of *fac 2* as $A_2 = \exp(\frac{-\mu^2}{2\tau^2}) \Big/ \sqrt{\pi}$.

If (y, m, μ, τ) does not satisfy (13.39), this implies that

$$y + \frac{\mu}{\tau^2} \leq 0 \Leftrightarrow y \leq -\frac{\mu}{\tau^2} \tag{13.40}$$

or

$$y + \frac{\mu}{\tau^2} \geq m \Leftrightarrow m - y \leq \frac{\mu}{\tau^2} \tag{13.41}$$

If the violation of (13.39) is due to (13.40), note that $A = \exp(\mu y + \frac{(y\tau)^2}{2}) \Big/ \sqrt{\pi}$ in the case of *fac 1*. Using the right inequality from (13.40) in the expression for A it follows that

$$\mu y + \frac{(y\tau)^2}{2} \leq \mu(-\frac{\mu}{\tau^2}) + \frac{(-\frac{\mu}{\tau^2} \cdot \tau)^2}{2} = -\frac{\mu^2}{2\tau^2} \leq 0. \tag{13.42}$$

Consequently, $A \leq 1$ if *fac 1* is used.

If the violation of (13.39) is due to (13.41), note that $A = \exp((y-m)[\frac{(y-m)\tau^2}{2} + \mu]) \Big/ \sqrt{\pi}$ in the case of *fac 3*. From (13.41) we have that

$$\mu \geq (m - y)\tau^2 \geq 0 \tag{13.43}$$

as $(m - y) \geq 0$ always. From this it follows that

$$\tfrac{1}{2}(y-m)\tau^2 + \mu \geq \tfrac{1}{2}(y-m)\tau^2 + (m-y)\tau^2 = \tfrac{1}{2}(m-y)\tau^2 \geq 0 \tag{13.44}$$

Consequently, $A \leq 1$ if *fac 3* is used. □

Chapter 14

Reference Priors

14.1. Introduction

To recapitulate what has been achieved so far, it was concluded in chapter 2 that the number of casualties in a mine affected area under fairly general assumptions can be considered to be the outcome of a binomial process. A forecast of the number of casualties can therefore be made if the binomial parameters characterizing the state of the minefield under study are known. The true binomial parameters will rarely be known in advance, but beliefs about these based on whatever information is available can be rephrased in terms of probability distributions. A convenient way to do so which provides the way for Bayesian data analysis is to express our previous knowledge through the priors

$$\pi_t(\tilde{m}) = \{\pi_t(0), \pi_t(1), ...\} \qquad \textbf{(14.01)}$$

and

$$\pi_t(\theta \mid \tilde{m}) \text{ for } \tilde{m} \geq 1. \qquad \textbf{(14.02)}$$

For $\tilde{m} \geq 1$ we may write $\pi_t(\tilde{m})$ and $\pi_t(\theta \mid \tilde{m})$ collectively as the prior joint distribution

$$\pi_t(\tilde{m}, \theta) = \pi_t(\theta \mid \tilde{m})\, \pi_t(\tilde{m}). \qquad \textbf{(14.03)}$$

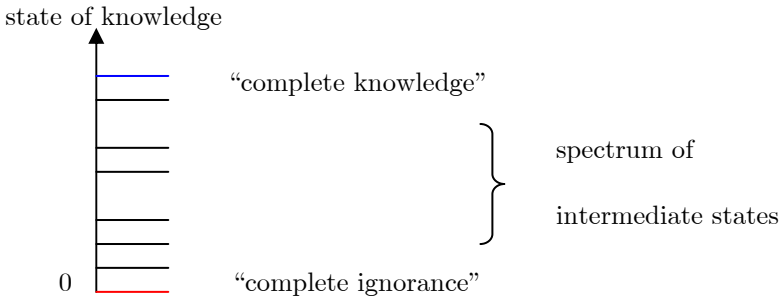
Up to now we have devised two ways of extracting information about the binomial parameter θ. Thus in chapter 4 it was demonstrated how a probability distribution $p(\theta)$ could be generated by combining accident statistics and clearance data from mine clearance operations. Similar information could in principle be provided through the analysis of accident statistics alone by the application of finite mixture models, as shown in chapter 5 to chapter 12.

Irrespective of how $p(\theta)$ has been provided, one has to extract the essential information contained in $p(\theta)$ and transfer it to (14.03) for every single minefield under consideration.

The most simple approach is to set $p(\theta) = \pi_t(\theta \mid \tilde{m})$ for all $\tilde{m}$. However, there might be cases where only essential features such as $E[\theta]$ and possibly $Var[\theta]$ should be extracted from $p(\theta)$ and incorporated with confidence into (14.03). The problem then arises how to piece together a prior $\pi_t(\tilde{m},\theta)$ from $E[\theta]$ and $Var[\theta]$ and similar fragments of information of relevance as to the distribution of $\tilde{m}$. More generally, we are looking for a method which can point out the particular prior that reflects our state of knowledge optimally.

The number of possible states of knowledge is infinite, but the spectrum of states can be considered to be bounded from above and below by two extreme states as sketched in figure 14.01 below. The state "complete knowledge" refers to the unique state where we a priori are absolutely certain about the true values of $(\tilde{m},\theta)$. The opposite to "complete knowledge" might be termed "complete ignorance". Consequently, the state "complete ignorance" denotes a kind of zero point as indicated in figure 14.01.

Fig. 14.01. Spectrum of states of previous knowledge about the binomial parameters $(\tilde{m},\theta)$.



To derive a prior which corresponds to the zero point in fig. 14.01, we need a clarification of the phrase "complete ignorance". In the present context we will use the phrase "complete ignorance" when our previous information about the parameters of interest is negligible *relative* to the information an experiment or observation can provide [Box and Tiao, 1973]. Thus in our search for a prior which reflects "complete ignorance" we will be looking for a probability distribution whose influence on the posterior distribution is marginal, that is, the posterior distribution should be dominated by the likelihood function as this is the factor through which observations modify our prior knowledge. Prior distributions guaranteed to play a minimal role in the posterior distribution are generally termed *noninformative priors*, a term which has already been used several times in the present report. Various approaches to generate noninformative priors are available as much

work has been done in this area [see for example Bernardo, 1979, Robert, 1994, Yang and Berger, 1998].

The derivation of a noninformative prior is of central importance but not sufficient in the present context as we want priors matching the intermediate states in fig. 14.01 as well. This together with the fact that the parameter space of $(\tilde{m},\theta)$ is $\mathbb{N}\times]0;1[$ makes the identification of suitable priors a challenging task. An intuitively appealing approach introducing so-called *reference priors* is due to Bernardo [Bernardo, 1979, Bernardo et al., 1994]. Bernardo's reference priors refers to a class of priors which in a certain sense maximize the information gained from observations. The derivation of a reference prior is in the general case technically demanding. However, the reference prior approach is adaptable to a variety of situations, and we will therefore base our derivation of prior distributions on Bernardo's concept of reference priors.

The introduction and derivation of reference priors in the remainder of the present chapter will cover the following topics: In paragraph 14.2, Bernardo's definition of reference priors for the general one-dimensional case is introduced. In paragraph 14.3, we set up the constrained functional which determines the two-dimensional reference prior $\pi_t(\tilde{m},\theta) = \pi_t(\theta \mid \tilde{m})\,\pi_t(\tilde{m})$. In paragraph 14.4, the reference prior corresponding to the "zero-point" state from fig. 14.01 is presented without proof. In paragraph 14.5, the joint posterior $\pi_t(\tilde{m},\theta \mid z)$ based on the reference prior from paragraph 14.4 is shown and compared with the likelihood function $p(z \mid m,\theta)$. In paragraph 14.6 we discuss how to set up reference priors when partial information is available. Paragraph 14.7 closes with concluding remarks. Appendix B contains technical details as to the derivation of the reference prior.

14.2. The Reference Prior Concept

To introduce the approach suggested by Bernardo, let X be some random variable taking values in some sample space where $p(X = x)$ depends on the value of a scalar parameter θ, that is, $p(X = x) = p(x \mid \theta)$. Let furthermore $\pi(\theta)$ denote the prior distribution of θ.

Assume now that an experiment e provides a single observation x. Let $\pi(\theta \mid x)$ denote the corresponding posterior distribution of θ. To quantify the information gained from the

observation x about θ, Bernardo makes use of the *Kullback-Leibler entropy distance* $K[\pi(\theta \mid x), \pi(\theta)]$ defined as

$$K[\pi(\theta \mid x), \pi(\theta)] = \int \log\left[\frac{\pi(\theta \mid x)}{\pi(\theta)}\right] \pi(\theta \mid x)\, d\theta. \qquad \textbf{(14.04)}$$

In general the Kullback-Leibler entropy distance for two normalized density functions $f(x)$ and $g(x)$ is defined as [Kullback, 1959]:

$$K[f(x), g(x)] = \int \log\left[\frac{f(x)}{g(x)}\right] f(x)\, dx. \qquad \textbf{(14.05)}$$

The use of the Kullback-Leibler entropy distance as a measure of information makes intuitively sense. That is, if a decision maker's previous knowledge about the true value of θ is accurate, the information gained from performing an experiment will be relatively low. Put in another way: If the accurate previous knowledge is reflected in the prior $\pi(\theta)$, the posterior $\pi(\theta \mid x)$ will almost certainly resemble the prior distribution. This induces a low value of $K[\pi(\theta \mid x), \pi(\theta)]$. If the decision maker on the other hand is ignorant about the true value of θ, the information gained from performing an experiment will be high. In this case the posterior distribution will be dominated by the likelihood function. This usually implies that the posterior distribution and the prior distribution are far apart in space which in turn generates a large value of $K[\pi(\theta \mid x), \pi(\theta)]$.

It can be shown that the Kullback-Leibler entropy distance is always non-negative and equals zero if and only if $f(x) = g(x)$ [Lehmann et al., 1998, p. 47]. In (14.05) the variable x is for convenience assumed to be a continuous variable but might as well be discrete in which case the integration is replaced by a summation.

The entropy distance $K[\pi(\theta \mid x), \pi(\theta)]$ depends on the particular observation x. The *expected information* $I(e, \pi(\theta))$ provided by a single observation is obtained by averaging (14.04) over the marginal distribution of x:

$$I(e, \pi(\theta)) = \int K[\pi(\theta \mid x), \pi(\theta)]\, p(x)\, dx, \qquad \textbf{(14.06)}$$

where $p(x)$ is given as

$$p(x) = \int p(x \mid \theta)\pi(\theta)\, d\theta. \tag{14.07}$$

Consider now a hypothetical experiment $e(k)$ yielding k independent observations. The expected information $I(e(k), \pi(\theta))$ can be calculated as

$$I(e(k), \pi(\theta)) = \int K[\pi(\theta \mid c_k), \pi(\theta)]\, p(c_k)\, dc_k, \tag{14.08}$$

where $c_k = (x_1, x_2, ..., x_k)$, $dc_k = dx_1 dx_2 ... dx_k$, and

$$\begin{aligned} p(c_k) &= \int p(c_k \mid \theta)\pi(\theta)\, d\theta \\ &= \int \prod_{i=1}^{k} p(x_i \mid \theta)\, \pi(\theta)\, d\theta \end{aligned} \tag{14.09}$$

In the limit $k \to \infty$ we will eventually obtain perfect information about the true value of θ. The corresponding quantity $I(e(\infty), \pi(\theta))$ defined as

$$I(e(\infty), \pi(\theta)) = \lim_{k \to \infty} I(e(k), \pi(\theta)) \tag{14.10}$$

measures, if it exists, our *missing information* about θ. The missing information depends on the function $\pi(\theta)$ and is therefore referred to as a missing information functional. If we search for a prior distribution containing negligible information about θ relative to what an observation can provide, the particular prior which *maximizes* the missing information functional appears to be the optimal choice. Bernardo terms the maximizing prior the *reference prior* [Bernardo et al., 1994]. Thus the determination of a non-informative prior has been transformed into a maximization problem of an information functional.

Even though the approach outlined above appears straightforward, the actual derivation of reference priors might get involved in specific cases. If the parameter of interest, say θ, can take only a finite number of values, the quantity $I(e(\infty), \pi(\theta))$ is always finite. As a consequence, the reference prior for θ can be derived directly from (14.08). For a continuous θ, however, $I(e(\infty), \pi(\theta))$ is typically infinite. To circumvent this problem, an asymptotic expansion of the information $I(e(k), \pi(\theta))$ might be derived from which the maximizing prior can be identified.

14.3. Information Functional in the Two-Dimensional Case

We will now set up the missing information functional from which a 2-dimensional reference prior $\pi(\tilde{m},\theta)$ can be derived. To avoid a cluttered notation, the symbol $\tilde{m}$ will in what follows be replaced by m.

Let us once again consider a hypothetical experiment $e(k)$ yielding k independent observations $(z(1),z(2),...,z(k)) = c_k$, where $z(i) \sim Bi(m,\theta)$. The expected information provided by k observations can in analogy to (14.08) be written as

$$I(e(k),\pi(m,\theta)) = \sum_{c_k} p(c_k)\, K[\pi(m,\theta \mid c_k),\pi(m,\theta)] \qquad \textbf{(14.11)}$$

where $p(c_k)$ is given as

$$\begin{aligned} p(c_k) &= \sum_m \pi(m) \int p(c_k \mid m,\theta)\, \pi(\theta \mid m)\, d\theta \\ &= \sum_m \pi(m) \int \prod_{i=1}^{k} p(z(i) \mid m,\theta)\, \pi(\theta \mid m)\, d\theta \end{aligned} \qquad \textbf{(14.12)}$$

In (14.11) we have extended the Kullback-Leibler entropy distance to include density functions of two variables. To simplify (14.11) the following identity will be useful [Kullback, 1959, p. 13]:

$$\begin{aligned} &K[f_1(x,y), f_2(x,y)] \\ &= \int\int \log[\frac{f_1(x,y)}{f_2(x,y)}]\, f_1(x,y)\, dx\, dy \\ &= \int \log[\frac{g_1(x)}{g_2(x)}]\, g_1(x)\, dx + \int g_1(x)\, [\int \log[\frac{h_1(y \mid x)}{h_2(y \mid x)}]\, h_1(y)\, dy]dx \\ &= K[g_1(x), g_2(x)] + \int g_1(x)\, K[h_1(y \mid x), h_2(y \mid x)]\, dx, \end{aligned} \qquad \textbf{(14.13)}$$

where $g_i(x)$ and $h_i(y \mid x)$ for $i = 1,2$ are defined as

$$g_i(x) = \int f_i(x,y)\, dy \qquad \textbf{(14.14)}$$

$$h_i(y \mid x) = \frac{f_i(x,y)}{g_i(x)}. \qquad \textbf{(14.15)}$$

By means of (14.13) the entropy distance $K[\pi(m,\theta \mid c_k),\pi(m,\theta)]$ can be written as

$$
\begin{aligned}
&K[\pi(m,\theta \mid c_k), \pi(m,\theta)] \\
&= K[\pi(m \mid c_k), \pi(m)] + \sum_m \pi(m \mid c_k)\, K[\pi(\theta \mid m, c_k), \pi(\theta \mid m)].
\end{aligned}
\tag{14.16}
$$

(14.16) allows us to express $I(e(k), \pi(m,\theta))$ from (14.11) as

$$
\begin{aligned}
&I(e(k), \pi(m,\theta)) \\
&= \sum_{c_k} p(c_k)\, K[\pi(m \mid c_k), \pi(m)] \\
&+ \sum_{c_k} \sum_m p(c_k) \pi(m \mid c_k) K[\pi(\theta \mid m, c_k), \pi(\theta \mid m)] \\
&= I(e(k), \pi(m)) \\
&+ \sum_m \pi(m) [\sum_{c_k} p(c_k \mid m) K[\pi(\theta \mid m, c_k), \pi(\theta \mid m)]] \\
&= I(e(k), \pi(m)) + \sum_m \pi(m)\, I(e(k), \pi(\theta \mid m)),
\end{aligned}
\tag{14.17}
$$

where we in (14.17) have used the identity $p(c_k)\pi(m \mid c_k) = p(c_k \mid m)\pi(m)$. Consequently, the prior $\pi(m,\theta)$ which in the limit $k \to \infty$ maximizes the functional

$$
I(e(k), \pi(m,\theta)) = I(e(k), \pi(m)) + \sum_m \pi(m)\, I(e(k), \pi(\theta \mid m)) \tag{14.18}
$$

is our two-dimensional reference prior.

14.4 Derivation of Two-Dimensional Reference Prior

To identify the prior $\pi(m,\theta)$ which in the limit $k \to \infty$ maximizes (14.18) we need asymptotic expansions of $I(e(k), \pi(m))$ and $I(e(k), \pi(\theta \mid m))$.

Regarding the term $I(e(k), \pi(\theta \mid m))$ it can be shown due to a theorem of Clarke and Barron [Clarke et al., 1990] that as $k \to \infty$,

$$
I(e(k), \pi(\theta \mid m)) = \frac{1}{2} \log \frac{k}{2\pi e} + \int \pi(\theta \mid m) \log \left[\frac{\left| I(\theta \mid m) \right|^{1/2}}{\pi(\theta \mid m)} \right] d\theta + R_k, \tag{14.19}
$$

where $R_k \to 0$ for $k \to \infty$. The term $\left| I(\theta \mid m) \right|$ denotes the Fisher information, i.e.

$$I(\theta \mid m) = E[-\frac{\partial^2}{\partial \theta^2} \log p(z \mid m, \theta)] = \frac{m}{\theta(1-\theta)}, \tag{14.20}$$

where we in (14.20) have used that $Z \sim Bi(m, \theta)$. From (14.20) it follows that (14.19) can be rewritten as

$$\begin{aligned} I(e(k), \pi(\theta \mid m)) &= \frac{1}{2} \log \frac{km\pi}{2e} + \int \pi(\theta \mid m) \log \left[\frac{Be(\theta \mid \frac{1}{2}, \frac{1}{2})}{\pi(\theta \mid m)} \right] d\theta + R_k \\ &= \frac{1}{2} \log \frac{km\pi}{2e} - K[\pi(\theta \mid m), Be(\theta \mid \tfrac{1}{2}, \tfrac{1}{2})] + R_k. \end{aligned} \tag{14.21}$$

As $\pi(\theta \mid m)$ only enters into the integral in (14.21), it is evident that $I(e(k), \pi(\theta \mid m))$ is maximized if

$$\pi(\theta \mid m) = Be(\theta \mid \tfrac{1}{2}, \tfrac{1}{2}) \; \forall m. \tag{14.22}$$

Regarding the term $I(e(k), \pi(m))$ from (14.18), let us rewrite the expression for $I(e(k), \pi(m))$ as

$$\begin{aligned} I(e(k), \pi(m)) &= \sum_{c_k} p(c_k) [\sum_m \pi(m \mid c_k) \log [\frac{\pi(m \mid c_k)}{\pi(m)}]] \\ &= \sum_{c_k} \sum_m p(c_k) \pi(m \mid c_k) \log \pi(m \mid c_k) \\ &- \sum_{c_k} \sum_m p(c_k) \pi(m \mid c_k) \log \pi(m) \\ &= \sum_{c_k} \sum_m p(c_k) \pi(m \mid c_k) \log \pi(m \mid c_k) - \sum_m \pi(m) \log \pi(m). \end{aligned} \tag{14.23}$$

In what follows we will assume that m takes only a *finite* number of different values, that is, $m \in M = \{m_1, m_2, ..., m_{MAX}\}$. It can then be shown (see appendix B) that

$$\sum_{c_k} \sum_{m \in M} p(c_k) \pi(m \mid c_k) \log \pi(m \mid c_k) \rightarrow 0 \tag{14.24}$$

in the limit $k \rightarrow \infty$. This allows us to write $I(e(k), \pi(m))$ from (14.23) as

$$I(e(k), \pi(m)) = -\sum_{m \in M} \pi(m) \log \pi(m) + R_k. \tag{14.25}$$

Now let $\sum_{m\in M}\pi(m)=1$ and define $h=\sum_{m\in M}\sqrt{m}$. From the derived expansions of $I(e(k),\pi(m))$ and $I(e(k),\pi(\theta\mid m))$ it follows from (14.18) that $I(e(k),\pi(m,\theta))$ for large values of k can be written as

$$\begin{aligned}
&I(e(k),\pi(m,\theta))=\\
&-\sum_{m\in M}\pi(m)\log\pi(m)\\
&+\sum_{m\in M}\pi(m)\Big[\frac{1}{2}\log\frac{km\pi}{2e}-K[\pi(\theta\mid m),Be(\theta\mid\tfrac{1}{2},\tfrac{1}{2})]\Big]+R_k\\
&=\frac{1}{2}\log\frac{k\pi h^2}{2e}-K[\pi(m),\frac{\sqrt{m}}{h}]-\sum_{m\in M}\pi(m)\,K[\pi(\theta\mid m),Be(\theta\mid\tfrac{1}{2},\tfrac{1}{2})]+R_k.
\end{aligned}\tag{14.26}$$

It is evident from (14.26) that $I(e(k),\pi(m,\theta))$ is maximized if

$$\pi(m)=\frac{\sqrt{m}}{h}\tag{14.27}$$

$$\pi(\theta\mid m)=Be(\theta\mid\tfrac{1}{2},\tfrac{1}{2})\ \forall m\in M\tag{14.28}$$

Consequently, the reference prior can be identified as

$$\boxed{\pi(m,\theta)=\frac{\sqrt{m}}{h}\,Be(\theta\mid\tfrac{1}{2},\tfrac{1}{2})\ \forall m\in M}\tag{14.29}$$

Fig. 14.02 below sketches the reference prior $\pi(m,\theta)$ when $m\in\{1,2,\ldots,20\}$.

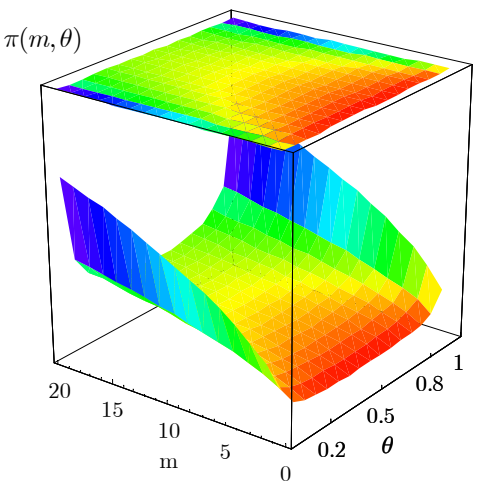


Fig. 14.02. $\pi(m,\theta)\propto\sqrt{m}\,Be(\theta\mid\frac{1}{2},\frac{1}{2})$ for $m\in\{1,2,\ldots,20\}$.

For the sake of clarity, the integer m is treated as a continuous variable in the 3D-plot.

The flat surface on top of the 3D-graph represents a density plot of $\pi(m,\theta)$.

14.5. Joint and Marginal Posterior Distributions Based on Reference Prior

Before we proceed with further analysis of (14.29), let us illustrate by a few examples how the derived reference prior influences the joint posterior distribution $\pi(m,\theta \mid z)$ and the derived marginal distributions for various observations.

To generate an analytical expression for $\pi(m,\theta \mid z)$ it is convenient to factorize the posterior joint distribution as $\pi(m,\theta \mid z) = \pi(\theta \mid m,z)\,\pi(m \mid z)$. The two one-dimensional posterior distributions are by definition given as

$$\pi(\theta \mid m,z) = \frac{p(z \mid m,\theta)\pi(\theta \mid m)}{p(z \mid m)}, \tag{14.30}$$

and

$$\begin{aligned} \pi(m \mid z) &= \frac{\pi(m)p(z \mid m)}{p(z)}. \\ &= \frac{\pi(m)\int p(z \mid m,\theta)\pi(\theta \mid m)\,d\theta}{p(z)}. \end{aligned} \tag{14.31}$$

Using (14.29) as our prior distribution, the conditioned posterior $\pi(\theta \mid m,z)$ becomes

$$\pi(\theta \mid m,z) = Be(\theta \mid \tfrac{1}{2}+z, \tfrac{1}{2}+m-z), \tag{14.32}$$

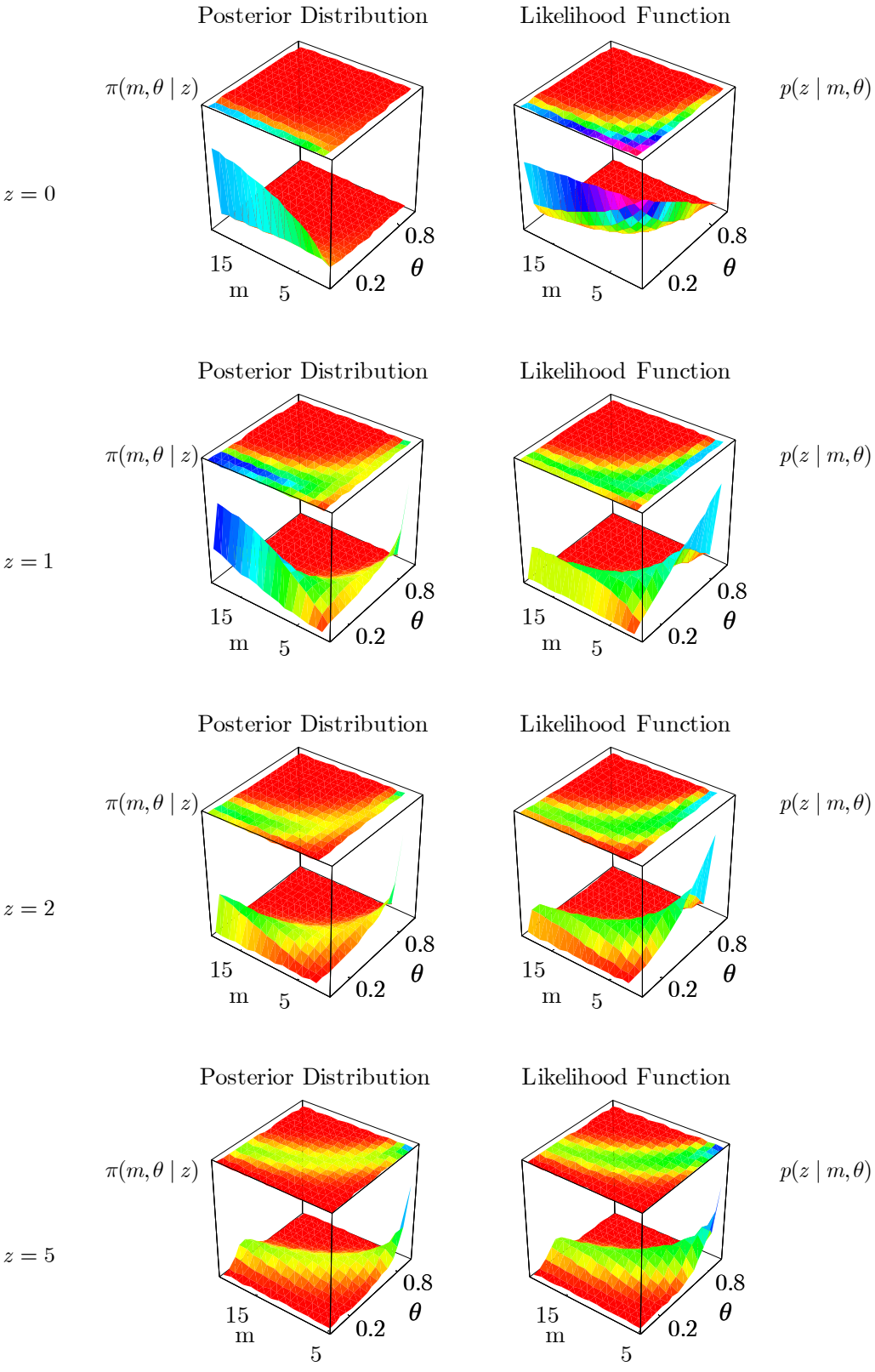
whereas the marginal posterior $\pi(m \mid z)$ takes the form

$$\begin{aligned} \pi(m \mid z) &\propto \pi(m)p(z \mid m) \\ &\propto \sqrt{m}\int_0^1 p(z \mid m,\theta)Be(\theta \mid \tfrac{1}{2},\tfrac{1}{2})\,d\theta \\ &\propto \sqrt{m}\binom{m}{z}\mathrm{B}(\tfrac{1}{2}+z, \tfrac{1}{2}+m-z). \end{aligned} \tag{14.33}$$

Combining (14.32) and (14.33) we get the joint posterior distribution

$$\pi(m,\theta \mid z) \propto \sqrt{m}\binom{m}{z}\mathrm{B}(\tfrac{1}{2}+z, \tfrac{1}{2}+m-z)\,\mathrm{B}e(\theta \mid \tfrac{1}{2}+z, \tfrac{1}{2}+m-z). \tag{14.34}$$

Fig. 14.03. Posterior distributions based on reference prior (14.29) for different values of z.



In fig. 14.03 (on the previous page) the joint posterior $\pi(m,\theta \mid z)$ from (14.34) is shown for different values of z. For the sake of clarity, the integer m is treated as a continuous variable in all 3D-plots. Included in fig. 14.03 is for comparison the likelihood function $p(z \mid m,\theta)$. As it emerges from fig. 14.03, the shape of the posterior distribution is clearly dominated by the likelihood function.

For completeness, plots of the marginal posterior distributions $\pi(m \mid z)$ and $\pi(\theta \mid z)$ for different values of z are included in fig. 14.04 and fig. 14.05 below.

Fig. 14.04. Marginal posterior of m.

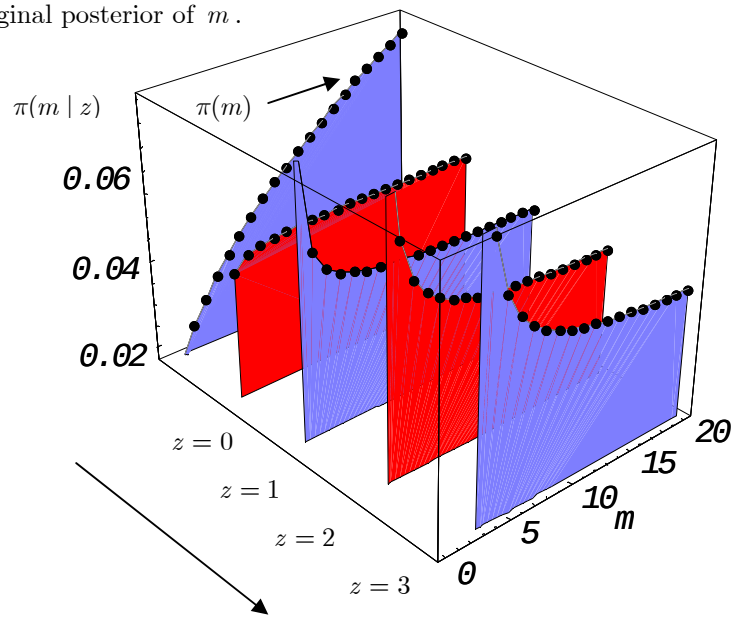
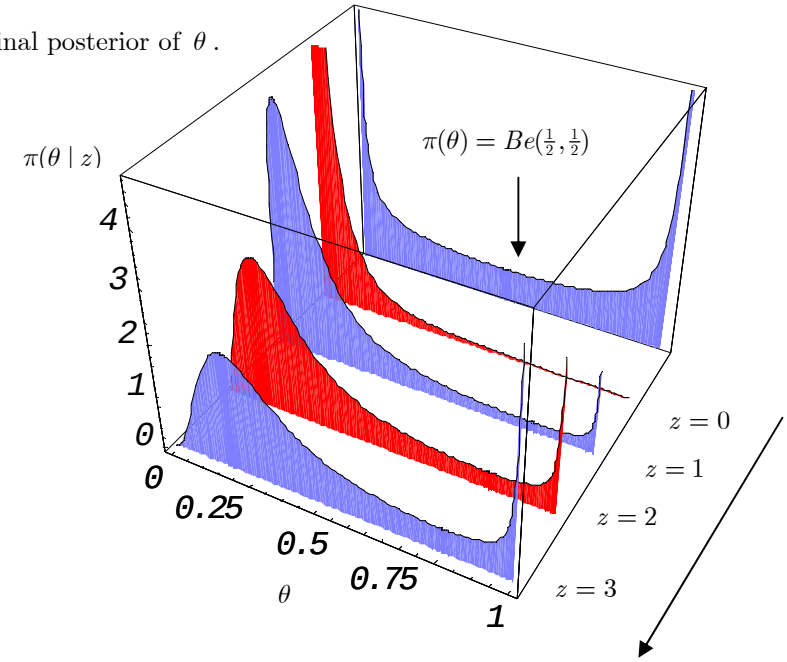


Fig. 14.05. Marginal posterior of θ.



14.6 Derivation of Reference Priors when Partial Information is Available

Situations where no information is available about neither m nor θ will rarely occur. The content of the present paragraph is an elaboration of how reference priors can be derived from the missing information functional in (14.26) when partial information of some kind is available. Thus we are aiming at reference priors corresponding to the intermediate states in fig. 14.01.

In what follows we will assume that the partial information can be rephrased as a set of constraints on the priors $\pi(m)$ and $\pi(\theta \mid m)$, that is,

$$\sum_{m \in M} \pi(m)\, g_j(m) = \mu_j \text{ for } j \in \{1,2,...,k\}, \tag{14.35}$$

$$\int \pi(\theta \mid m)\, g_{j,m}(\theta)\, d\theta = \mu_{j,m} \text{ for } j \in \{1,2,...,l\} \text{ and } m \in M. \tag{14.36}$$

Major simplifications in the derivation of reference priors can be obtained if one or more of the conditions listed below are met:

1) independence between m and θ.
2) $\pi(\theta \mid m)$ available for all $m \in M$.
3) no restrictions on $\pi(m)$.

We will examine each of the above conditions in turn.

14.6.1. Independence Between m and θ

The most simple case arises if we make the restriction

$$\pi(\theta \mid m) = \pi(\theta)\ \forall m \in M, \tag{14.37}$$

that is, we do not a priori assume any dependence between the number of mines in the minefield under study and the distribution of θ. In that case (14.26) can be written

$$I(e(k), \pi(m,\theta)) = \frac{1}{2}\log\frac{k\pi h^2}{2e} - K[\pi(m), \frac{\sqrt{m}}{h}] - K[\pi(\theta), Be(\theta \mid \tfrac{1}{2}, \tfrac{1}{2})] + R_k. \tag{14.38}$$

In (14.38) the contributions from $\pi(m)$ and $\pi(\theta)$ are separated into two independent terms which can be maximized separately. Given that restrictions like (14.35) have been enforced on $\pi(m)$ due to partial information, the prior $\pi(m)$ which maximizes (14.38) can be found as the solution to the following constrained maximization problem:

$$\begin{aligned} Max & -\sum_{m\in M}\pi(m)\log\left[\frac{\pi(m)}{\sqrt{m}}\right] \\ s.t. & \sum_{m\in M}\pi(m)=1, \\ & \sum_{m\in M}\pi(m)\,g_j(m)=\mu_j \text{ for } j\in\{1,2,...,k\}, \\ & \pi(m)>0\ \forall m\in M. \end{aligned} \tag{14.39}$$

With respect to $\pi(\theta)$ we are left with the maximization problem

$$\begin{aligned} Max & \int \pi(\theta)\log\left[\frac{Be(\theta\mid\frac{1}{2},\frac{1}{2})}{\pi(\theta)}\right]d\theta \\ s.t & \int \pi(\theta)\,g_j(\theta)\,d\theta=\mu_j \ for\ j\in\{1,2,...,l\}, \\ & \int \pi(\theta)\,d\theta=1. \end{aligned} \tag{14.40}$$

According to Bernardo [Bernardo, 1994, p. 319], the solution to (14.40) can (given it exists) be written as

$$\pi(\theta)=Be(\theta\mid\tfrac{1}{2},\tfrac{1}{2})\exp(\sum_{j=1}^{l}\lambda_j\,g_j(\theta)), \tag{14.41}$$

where the λ_j's are constants to be determined from the constraints in (14.40). To give an example where (14.40) might be brought into play, consider the case where estimates of $E[\theta]$ and $Var[\theta]$ makes up the partial information about θ (the moments $E[\theta]$ and $Var[\theta]$ can for example be obtained from a finite mixture model calculation). It follows that $g_1(\theta)=\theta$ and $g_2(\theta)=(\theta-E[\theta])^2$ in (14.40).

14.6.2. Conditioned Priors Available

A different situation arises if the conditioned priors $\pi(\theta\mid m)$ are available for all m due to partial information. In the most simple case we have that $\pi(\theta\mid m)=\pi(\theta)$ for all m where

$\pi(\theta)$ might be based on accident statistics and clearance data from mine clearance operations as discussed in chapter 4. In this case we simply ignore (14.40) and determine $\pi(m)$ from (14.39). On the other hand, if the available priors $\pi(\theta \mid m)$ depend on m, the corresponding reference prior $\pi(m)$ is to be found as the solution to the maximization problem

$$\begin{aligned} Max & -\sum_{m\in M}\pi(m)\log\left[\frac{\pi(m)}{\sqrt{m}}\right]-\sum_{m\in M}\pi(m)\,K[\pi(\theta\mid m),Be(\theta\mid \tfrac{1}{2},\tfrac{1}{2})] \\ s.t. \quad & \sum_{m\in M}\pi(m)=1, \\ & \sum_{m\in M}\pi(m)\,g_j(m)=\mu_j \text{ for } j\in\{1,2,...,k\}, \\ & \pi(m)>0\ \forall m\in M. \end{aligned} \tag{14.42}$$

14.6.3. No Restrictions on $\pi(m)$.

In the case that no restrictions are imposed on $\pi(m)$, we will rewrite the expansion of $I(e(k),\pi(m,\theta))$ from (14.26) as

$$\begin{aligned} & I(e(k),\pi(m,\theta))= \\ & \frac{1}{2}\log\frac{k\pi}{2e}+\log(s)-K[\pi(m),\frac{\sqrt{m}\exp(-K[\pi(\theta\mid m),Be(\theta\mid \tfrac{1}{2},\tfrac{1}{2})])}{s}]+R_k, \end{aligned} \tag{14.43}$$

where $s=\sum_{m\in M}\sqrt{m}\exp(-K[\pi(\theta\mid m),Be(\theta\mid \tfrac{1}{2},\tfrac{1}{2})])$.

As no restrictions are imposed on $\pi(m)$, it follows that (14.43) is maximized if

$$\pi(m)=\frac{\sqrt{m}\exp(-K[\pi(\theta\mid m),Be(\theta\mid \tfrac{1}{2},\tfrac{1}{2})])}{s}, \tag{14.44}$$

for any choice of $\pi(\theta\mid m)$.

To determine $\pi(\theta\mid m)$ we simply maximize $\log(s)$ from (14.43), that is, the set of reference priors $\{\pi(\theta\mid m)\}$ can be identified as solutions to the maximization problem

$$
\begin{aligned}
&Max\ \log\left[\sum_{m\in M}\sqrt{m}\,\exp(-K[\pi(\theta\mid m), Be(\theta\mid \tfrac{1}{2},\tfrac{1}{2})])\right]\\
&s.t.\int \pi(\theta\mid m)\,d\theta = 1\ \forall m\in M,\\
&\int \pi(\theta\mid m)\,g_{j,m}(\theta)\,d\theta = \mu_{j,m}\ \text{for all}\ (j,m),
\end{aligned}
\tag{14.45}
$$

or alternatively

$$
\begin{aligned}
&Max\int \pi(\theta\mid m)\log\left[\frac{Be(\theta\mid \tfrac{1}{2},\tfrac{1}{2})}{\pi(\theta\mid m)}\right]d\theta\ \forall m\in M\\
&s.t\int \pi(\theta\mid m)\,g_{j,m}(\theta)\,d\theta = \mu_{j,m}\ \text{for all}\ (j,m)\\
&\int \pi(\theta\mid m)\,d\theta = 1\ \forall m\in M.
\end{aligned}
\tag{14.46}
$$

From (14.46) it follows that the maximizing priors $\pi(\theta\mid m)$ can be written as

$$
\pi(\theta\mid m) = Be(\theta\mid \tfrac{1}{2},\tfrac{1}{2})\exp(\sum_{j=1}^{l}\lambda_{j,m}\,g_{j,m}(\theta))\ \forall m\in M, \tag{14.47}
$$

where the $\lambda_{j,m}$'s are constants to be determined from the constraints in (14.46).

14.7. Summary and Conclusions

A prerequisite for the employment of the Bayesian risk model derived in chapter 2 is the provision of a prior distribution $\pi_t(m,\theta) = \pi_t(\theta\mid m)\,\pi_t(m)$. As a decision maker's previous knowledge about the minefield under study can be anything from "complete ignorance" at the one end to a state of "complete knowledge" at the other end, it is a delicate matter how to embed an arbitrary level of knowledge into the two-dimensional probability distribution $\pi_t(m,\theta)$. Of particular importance is not to impose features on $\pi_t(m,\theta)$ which are without foundation in the available information.

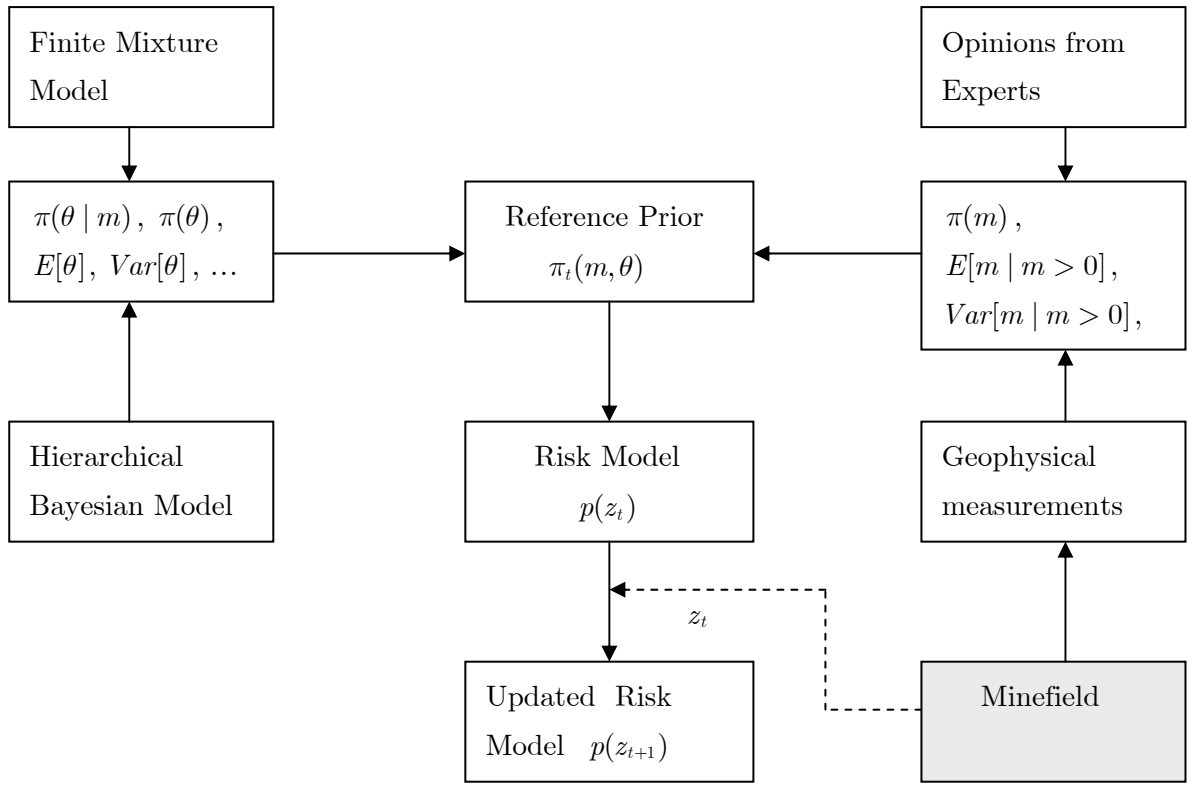
The aim of the present chapter has been to set up a general procedure for the construction of prior distributions which overcome the above difficulties. The concept of *reference priors* as defined by Bernardo has in this context turned out to be of great value. Thus through the application of the Kullback-Leibler entropy distance as a measure of information it has been possible to identify the distribution which maximizes the

information gained from an infinite number of observations. The distribution in question is termed a reference prior. The reference prior is to be considered as a noninformative prior which in a certain sense reflects a state of "complete ignorance".

Given the decision maker has previous knowledge which can be rephrased as a set of constraints on $\pi_t(m,\theta)$, a constrained reference prior can be identified as the solution to a constrained maximization problem. In certain special cases, as discussed in paragraph 14.6, the maximization problem takes on a simple form from which the identification of the constrained reference prior is straightforward. It is notable that the marginal reference prior $\pi(m)$ found by the above method deviates from a uniform distribution even when m and θ are assumed to be independent.

With the above method in place it is at last possible to associate the various pieces of models derived in the present report. In fig. 14.06 below, a hypothetical Bayesian risk assessment module is sketched illustrating the interrelationships between the various model components.

Fig. 14.06. Information flow in Bayesian risk assessment model. See text for further details.



By going through fig. 14.06 we take the opportunity to recapitulate what has been achieved so far and what is still left to be developed before a risk assessment system based on Bayesian data analysis is operational.

In fig. 14.06, the minefield under study is represented by the grey box positioned at the right lower corner. To make a risk assessment of the minefield for the coming observation period $\Delta(t)$, information of relevance as to the possible values of m and θ are to be collected. Regarding θ, a probability distribution can be provided through the hierarchical Bayesian model derived in chapter 4 or the finite mixture model discussed in chapter 5-12. This is indicated in the left upper part of fig. 14.06. If necessary it may be decided only to extract certain pieces of information such as $E[\theta]$ and $Var[\theta]$.

Regarding m, a probability distribution can be provided through the synthesis of individual estimates from various local or regional experts. In the future it may be possible to complement these estimates by actual geophysical measurements (or some other kind of measurement) from the minefield. We have not in the present report discussed how to provide and synthesize the above type of information.

Having somehow provided information about m and θ, all pieces of information are put together and embedded into a reference prior distribution. The reference prior thus constitutes the core component in the risk model. Subsequently, a probability distribution $p(z_t)$ can be set up founded on the reference prior.

---------- 0 ----------

Chapter 15

Summary, Conclusions and Suggestions for Further Work

Humanitarian Mine Action has undergone an impressive development since its advent in the late eighties. This development can be registered specifically at the organizational level among the practitioners in the HMA sector and more generally as an improved understanding of the complexities of the mine contamination problem and its impact on mine affected countries. The present lack of a fast and reliable mine detection technology means nonetheless that the worldwide mine contamination problem cannot be eliminated in the foreseeable future but has to be managed in several years to come. To sustain the sectors capacity for development, we request decision makers involved in HMA to be aware of disciplines such as of operations research and statistics which might offer powerful analytical tools enabling the HMA sector to optimize ongoing procedures with existing technologies.

The present thesis represents a first attempt to develop a minefield risk assessment model based on principles from operations research and statistics which might support decision makers in their attempt to classify and prioritize potential minefields according to risk. It should be emphasized however that only the first step in this direction has been taken. In what follows the major findings in the present research project will therefore be summarized in two steps: Firstly, with reference to the objectives stated in chapter 1, the main features of the derived risk model will be summarized. Secondly, various directions along which the presented model can be improved and adjusted to real-life applications are suggested.

15.1 Main Features of Derived Risk Model

The main objective set up in chapter 1 was to derive a mathematical model by which a decision maker can rank an arbitrary number of minefields according to risk. This objective has been met through the formulation of the stochastic binomial model derived in chapter 2. By incorporating the binomial model into a Bayesian framework it has

furthermore been possible to make the risk model dynamic in the sense that the risk assessment of a given minefield can be updated over time by incoming accident statistics through the application of Bayes' rule.

Apart from making the risk model dynamic, the application of Bayesian data analysis has given the risk model a very flexible structure which allows it to accommodate to the varied circumstances found in HMA with respect to accessible information. That is, due to the approach followed in Bayesian data analysis where prior beliefs about all entering variables are expressed in terms of probability distributions, it is possible to impart information from a variety of different sources into the risk model. Such information may be of a quantitative nature (e.g. accident statistics) or it may be of a more subjective or qualitative nature such as expert opinions concerning the degree of mine contamination in a given area. An overall prescription for the synthesis of different pieces of information and its transfer to the risk model is formulated in chapter 14 dealing with reference priors.

The derived risk model seems to overcome many of the shortcomings identified in the landmine impact score model referred to in chapter 1. Thus unlike the mine impact score model the risk model makes a balanced weighing of the decision makers previous knowledge about the minefield under study and later incoming accident statistics. The risk model is well suited for long term planning purposes due to its ability to make very graduated risk assessments. This contrasts with the mine impact score model which classifies all minefields with no records of recent victims as "Low". Finally, as risk in the present context is defined in probabilistic terms, it is possible to compare minefield risk assessments with other sources of risk in the society.

15.2 Suggestions for Further Work

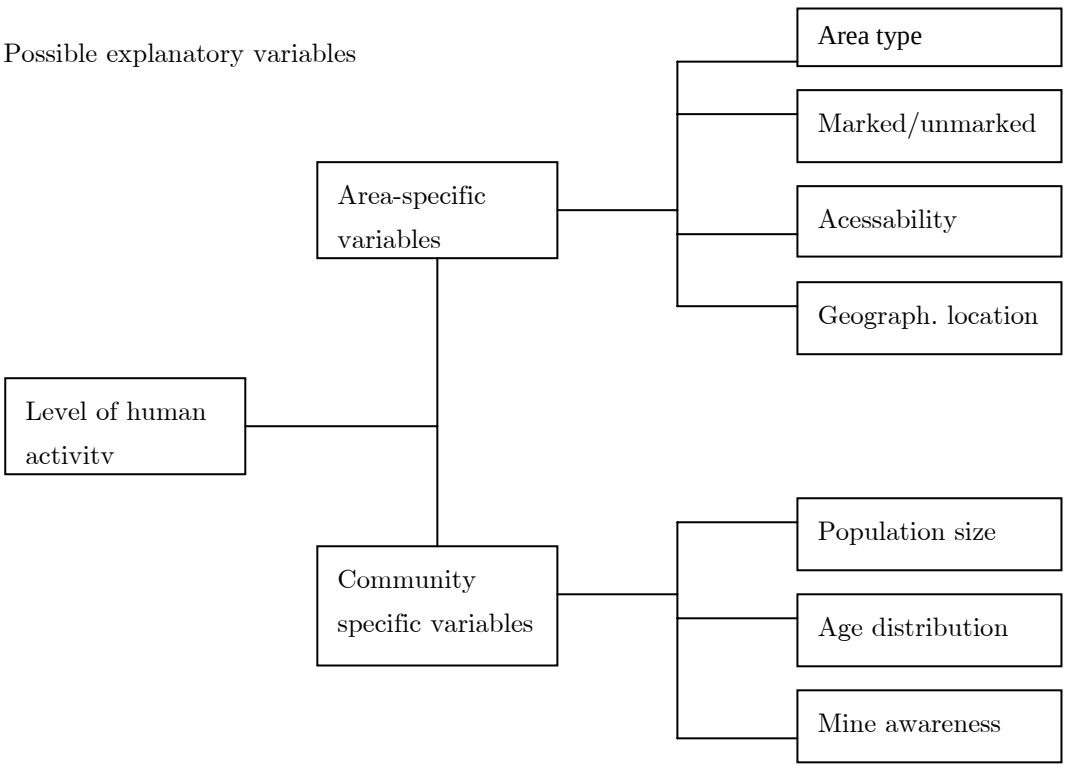
An appealing feature of the derived risk model is that only two parameters are needed to characterize the state of a given minefield. The analytical challenge is then to estimate these parameters by the collection and synthesis of various types of information of relevance. Concerning this estimation process, we have in the present thesis presented just two different approaches as to the estimation of the probability parameter θ, and none of these methods have involved explanatory variables. Consequently, some work remains to

be done before the risk model can be operational in real-life applications. In what follows we will give suggestions to areas where improvements are needed.

To take the finite mixture model calculations as our first example, the omission of explanatory variables has had the practical consequence that the posterior $p(\theta \mid y)$ generated from a given mixture model has been amenable to control against simpler models. There is no doubt, however, that to exploit the full potential of the mixture model concept in a real life application, one should aim at the more generalized versions of mixture models in which explanatory variables enter into the expression of both the mixture components and the mixture parameters. Such advanced models can be considered as a special case of so-called *Mixture-of-Experts Models* [see for example Jacobs et al., 1991] which may be generalized one step further to *Hierarchical Mixture-of-Experts Models* [Jordan & Jacobs, 1992]. Consequently, there exist several options for extensions of the application of the finite mixture model concept.

Given that explanatory variables are decided to be included in future work, the lack of relevant statistical material leaves us to speculate on what might candidate as explanatory variables. In fig. 15.1 below, likely variables are suggested which might correlate with θ. All variables may possibly influence what might loosely be termed the level of human activity which again determines the probability parameter θ.

Fig. 15.1. Possible explanatory variables



As shown in fig. 15.1, the explanatory variables are for clarity split up into two groups: area-specific variables and community-specific variables. By area-specific variables we mean variables which describe characteristic features of the mine affected area of interest such as area type (residential area, agricultural area, pasture, forest, road, etc.), whether the area is marked or unmarked, public access to area (level of accessibility), geographical location of area relative to important community facilities such as water facilities, and the number of recent victims known to the public. Similarly, by community specific variables we mean variables which describe characteristic features of the community. It goes without saying that to determine the statistically most significant explanatory variables, the data collection process in HMA has to be broadened considerably to include a much wider spectrum of data.

A second problem which has not been touched on in the present thesis is the estimation of the binomial parameter $\tilde{m}$. As noted in the introduction to chapter 4, various sources which might provide information of a more subjective or qualitative nature about the possibility of mine contamination in a given area include military staff and other ex-combatants with local knowledge, and local or regional authoritatives. Information of a quantitative nature might be available in terms of military mine maps and related archives. In the future it may furthermore become technically possible to undertake geophysical measurements or some other kind of measurements from outside the borders of a minefield. In any way, it is an open question how to combine these various pieces of information into a probability distribution.

In the publication "A Study of Soci-Economic Approaches to Mine Action" [GICHD, 2001] it is stated that humanitarian mine action is just as much about *data processing* as it is about mines. It seems fair to conclude that the Bayesian framework set up in the present thesis represents substantial new thinking concerning data processing in Humanitarian Mine Action, and the outlined approach to risk assessment and ranking of minefields can, if properly used, in the future support decision makers considerably in their aim at improving the impact of national mine action programmes. There exist however several options for extensions and improvements of the work presented in the present thesis which may gradually turn the theoretical model considerations into a decision tool of practical value to the Humanitarian Mine Action sector.

Appendix A. Sampling from Conditioned Distributions

To apply the single-component Metropolis-Hastings algorithm introduced in chapter 6 sampling has to be carried out from the four conditioned distributions $p(\zeta \mid y, \mu, \tau, \lambda)$, $p(\lambda \mid y, \zeta, \mu, \tau)$, $p(\mu \mid y, \zeta, \tau, \lambda)$ and $p(\tau \mid y, \zeta, \mu, \lambda)$. We here describe in brief how samples can be obtained from each of the conditioned distributions.

A1. Sampling from $p(\zeta \mid y, \mu, \tau, \lambda)$

Let

$$\begin{aligned} p(\zeta_{jm} &= 1 \mid y_j = k, \mu, \tau, \lambda) \\ &= \frac{\lambda_m f(k \mid m, \mu, \tau)}{\sum_{m' \geq k} \lambda_{m'} f(k \mid m', \mu, \tau)} \\ &\equiv Z_{km}, \end{aligned} \qquad \textbf{(A.1)}$$

where $m \in \{m_1, m_2, ..., m_g\}$ and

$$\sum_m Z_{km} = 1. \qquad \textbf{(A.2)}$$

Let $Z_{km}^{acc} = \sum_{l=0}^{m} Z_{kl}$ denote accumulated probabilities. Sampling from the conditioned distribution $p(\zeta \mid y, \mu, \tau, \lambda)$ can now be carried out as follows: For a given (μ, τ, λ) and a given observation $y_j = k$, calculate the accumulated probabilities Z_{km}^{acc}, $m \in \{m_1, m_2, ..., m_g\}$. Sample then a real number t where $t \sim U(0,1)$. Let $\zeta_{jm} = 1$ for the smallest m for which $t \leq Z_{km}^{acc}$. All other components in ζ_j are fixed to 0. Repeat the procedure for all observations y_j.

A2. Sampling from $p(\lambda \mid y, \mu, \tau, \zeta)$

Define X_K as

$$X_k = \{\# y_j \in y \mid y_j \text{ is associated with component } k \textit{ via } \zeta\}, \qquad \textbf{(A.3)}$$

i.e., X_K denotes the number of observations originating from component k according to the indicator variable ζ. Let furthermore λ follow a Dirichlet distribution, i.e. $\lambda \sim \textit{Dirichlet}(\alpha_1, \alpha_2, ..., \alpha_g)$. It follows that

$$\lambda \mid y, \mu, \tau, \zeta \sim Dirichlet(\alpha_1 + X_1, \alpha_2 + X_2, ..., \alpha_g + X_g). \tag{A.4}$$

A formal proof of (A.4) can be given sketched as follows: Note that

$$\begin{aligned} & p(\lambda \mid y, \mu, \tau, \zeta) \\ & \quad = \frac{p(\lambda, y, \mu, \tau, \zeta)}{p(y, \mu, \tau, \zeta)} \\ & \quad = \frac{p(y \mid \mu, \tau, \lambda, \zeta) p(\zeta \mid \lambda) p(\lambda)}{\int p(y \mid \mu, \tau, \lambda, \zeta) p(\zeta \mid \lambda) p(\lambda)\, d\lambda}, \end{aligned} \tag{A.5}$$

which can be written as

$$\begin{aligned} & = \frac{\prod_j \prod_{m \geq y_j} f(y_j \mid m, \mu, \tau)^{\zeta_{jm}} {\lambda_m}^{\zeta_{jm}} p(\lambda)}{\int \prod_j \prod_{m \geq y_j} f(y_j \mid m, \mu, \tau)^{\zeta_{jm}} {\lambda_m}^{\zeta_{jm}} p(\lambda)\, d\lambda} \\ & = \frac{\prod_{k=1}^{g} {\lambda_k}^{X_k} p(\lambda)}{\int \prod_{k=1}^{g} {\lambda_k}^{X_k} p(\lambda)\, d\lambda}. \end{aligned} \tag{A.6}$$

By definition

$$p(\lambda) \propto \prod_{k=1}^{g} \lambda_k^{\alpha_k - 1}, \tag{A.7}$$

and (A.4) follows immediately from (A.6) and (A.7). Sampling from the Dirichlet dist. can be implemented as described in Gelman (2003, p. 582).

A3. Sampling from $p(\mu \mid y, \tau, \lambda, \zeta)$ and $p(\tau \mid y, \mu, \lambda, \zeta)$.
Concerning $p(\mu \mid y, \tau, \lambda, \zeta)$ we have that

$$\begin{aligned} & p(\mu \mid y, \tau, \lambda, \zeta) \\ & = \frac{p(\mu, y, \tau, \lambda, \zeta)}{p(y, \tau, \lambda, \zeta)} \\ & = \frac{p(y \mid \mu, \tau, \lambda, \zeta) p(\zeta \mid \lambda) p(\mu, \tau, \lambda)}{p(y, \tau, \lambda, \zeta)} \\ & = \frac{\prod_j \prod_{m \geq y_j} f(y_j \mid m, \mu, \tau)^{\zeta_{jm}} {\lambda_m}^{\zeta_{jm}} p(\mu, \tau, \lambda)}{p(y, \tau, \lambda, \zeta)} \\ & \propto \prod_j \prod_{m \geq y_j} f(y_j \mid m, \mu, \tau)^{\zeta_{jm}} p(\mu) \end{aligned} \tag{A.8}$$

where the constant of proportionality in the last line of (A.8) is constant for fixed values of y, τ, λ and ζ.

Sampling from $p(\mu \mid y, \tau, \lambda, \zeta)$ can be carried out by use of the Metropolis-Hastings algorithm discussed in chapter 6. If μ^0 denotes the value of μ from the preceding iteration, a new μ^* is sampled from a jumping distribution $J(\mu^* \mid \mu^0) = N(\mu^* \mid \mu^0, d)$ where d is a constant. The draw μ^* is subsequently evaluated by the calculation of the quotient

$$r = \frac{\prod_j \prod_{m \geq y_j} f(y_j \mid m, \mu^*, \tau)^{\zeta_{jm}} p(\mu^*) / N(\mu^* \mid \mu^0, d)}{\prod_j \prod_{m \geq y_j} f(y_j \mid m, \mu^0, \tau)^{\zeta_{jm}} p(\mu^0) / N(\mu^0 \mid \mu^*, d)}. \qquad \textbf{(A.9)}$$

If μ^* is accepted $\mu^1 = \mu^*$. Otherwise $\mu^1 = \mu^0$. If the proportion of accepted draws is too low or too high, the acceptance rate can be adjusted by adjusting d. Sampling from the conditioned distribution $p(\tau \mid y, \mu, \lambda, \zeta)$ can be carried out in a similar way. We have that

$$\begin{aligned} & p(\tau \mid y, \mu, \lambda, \zeta) \\ &= \frac{p(\tau, y, \mu, \lambda, \zeta)}{p(y, \mu, \lambda, \zeta)} \\ &= \frac{p(y \mid \mu, \tau, \lambda, \zeta) p(\zeta \mid \lambda) p(\tau) p(\mu, \lambda)}{p(y, \mu, \lambda, \zeta)} \\ &\propto \prod_j \prod_{m \geq y_j} f(y_j \mid m, \mu, \tau)^{\zeta_{jm}} p(\tau), \end{aligned} \qquad \textbf{(A.10)}$$

where the constant of proportionality in the last line of (A.10) is constant for fixed values of y, μ, λ and ζ.

As $\tau \geq 0$ we cannot in connection with the Metropolis-Hastings algorithm apply a normal distribution as jumping distribution. Instead we apply a *scaled* $inv\chi^2$-distribution. If τ^0 denotes the value of τ from the preceding iteration, a new τ^* is sampled from the distribution $inv\chi^2(\tau^* \mid \tau^0, s)$, and the draw is subsequently evaluated by calculating the quotient

$$r = \frac{\prod_j \prod_{m \geq y_j} f(y_j \mid m, \mu, \tau)^{\zeta_{jm}} p(\tau^*) / inv\chi^2(\tau^* \mid \tau^0, s)}{\prod_j \prod_{m \geq y_j} f(y_j \mid m, \mu, \tau)^{\zeta_{jm}} p(\tau^0) / inv\chi^2(\tau^0 \mid \tau^*, s)}. \qquad \textbf{(A.11)}$$

If τ^* is accepted $\tau^1 = \tau^*$. Otherwise $\tau^1 = \tau^0$. The acceptance rate can be controlled by adjustment of the parameter s.

Appendix B. Reference Prior Derivation

In paragraph 14.4 it was claimed that

$$\sum_{c_k}\sum_{m\in M} p(c_k)\pi(m \mid c_k)\log\pi(m \mid c_k) \to 0 \qquad \textbf{(B.01)}$$

in the limit $k \to \infty$. In the present appendix we will set out to prove (B.01).

Consider an experiment yielding k observations $(z(1), z(2),..., z(k-1), z(k)) = c_k$ where $z(i) \sim Bi(m^*, \theta^*)$, $i \in \{1,2,...,k-1,k\}$. Note that $m^* \in M = \{m_1, m_2,..., m_{MAX}\}$ and $\theta^* \in]0;1[$. According to Bayes' rule, $\pi(m \mid c_k)$ can be written as

$$\begin{aligned}
\pi(m \mid c_k) &= \frac{\pi(m)\int p(c_k \mid m,\theta)\,\pi(\theta \mid m)\,d\theta}{p(c_k)} \\
&= \frac{\pi(m)\int p(c_k \mid m,\theta)\,\pi(\theta \mid m)\,d\theta}{\sum_{m\in M}\pi(m)\int p(c_k \mid m,\theta)\,\pi(\theta \mid m)\,d\theta} \\
&= \frac{\pi(m)\int \exp\log[\frac{p(c_k \mid m,\theta)}{p(c_k \mid m^*,\theta^*)}]\,\pi(\theta \mid m)\,d\theta}{\sum_{m\in M}\pi(m)\int \exp\log[\frac{p(c_k \mid m,\theta)}{p(c_k \mid m^*,\theta^*)}]\,\pi(\theta \mid m)\,d\theta}.
\end{aligned} \qquad \textbf{(B.02)}$$

Consider now the log-term from (B.02) which can be written as

$$\begin{aligned}
\log\frac{p(c_k \mid m,\theta)}{p(c_k \mid m^*,\theta^*)} &= \log\prod_{i=1}^{k}\frac{p(z(i) \mid m,\theta)}{p(z(i) \mid m^*,\theta^*)} \\
&= \log\prod_{j=0}^{m^*}\left[\frac{p(j \mid m,\theta)}{p(j \mid m^*,\theta^*)}\right]^{s_k(j)} \\
&= \sum_{j=0}^{m^*} s_k(j)\log\left[\frac{p(j \mid m,\theta)}{p(j \mid m^*,\theta^*)}\right],
\end{aligned} \qquad \textbf{(B.03)}$$

where $s_k(j)$ denotes the number of times the outcome $j \in \{0,1,2,...,m^*\}$ occurs in the vector c_k. As the corresponding stochastic variable $S_k(j) \sim Bi(k, p(j \mid m^*,\theta^*))$, it follows that

$$E\left[\frac{S_k(j)}{k}\right] = p(j \mid m^*, \theta^*), \tag{B.04}$$

and

$$Var\left[\frac{S_k(j)}{k}\right] = \frac{p(j \mid m^*, \theta^*)(1 - p(j \mid m^*, \theta^*))}{k}. \tag{B.05}$$

According to Chebychevs' Inequality we consequently have

$$p\left(\left|\frac{S_k(j)}{k} - p(j \mid m^*, \theta^*)\right| > \varepsilon\right) \leq \frac{p(j \mid m^*, \theta^*)(1 - p(j \mid m^*, \theta^*))}{k\varepsilon^2} \tag{B.06}$$

which implies that

$$p\left(\left|\frac{S_k(j)}{k} - p(j \mid m^*, \theta^*)\right| > \varepsilon\right) \to 0 \text{ for } k \to \infty. \tag{B.07}$$

Consequently, as $k \to \infty$,

$$\begin{aligned} \sum_{j=0}^{m^*} S_k(j) \log\left[\frac{p(j \mid m, \theta)}{p(j \mid m^*, \theta^*)}\right] = \quad & k \sum_{j=0}^{m^*} \frac{S_k(j)}{k} \log\left[\frac{p(j \mid m, \theta)}{p(j \mid m^*, \theta^*)}\right] \\ & \to -k \sum_{j=0}^{m^*} p(j \mid m^*, \theta^*) \log\left[\frac{p(j \mid m^*, \theta^*)}{p(j \mid m, \theta)}\right] \\ & = -k\, K[p(j \mid m^*, \theta^*), p(j \mid m, \theta)], \end{aligned} \tag{B.08}$$

implying that

$$\pi(m \mid c_k) \to \frac{\pi(m) \int \exp(-k\, K[p(j \mid m^*, \theta^*), p(j \mid m, \theta)])\, \pi(\theta \mid m)\, d\theta}{\sum_{m' \in M} \pi(m') \int \exp(-k\, K[p(j \mid m^*, \theta^*), p(j \mid m', \theta)])\, \pi(\theta \mid m')\, d\theta}. \tag{B.09}$$

By assumption $\pi(m^*) > 0$ and $\pi(\theta \mid m^*) > 0$ for $\theta \in]0;1[$. Therefore $\sum_{m \in M} \pi(m \mid c_k) = 1$. Defining $g(m,k)$ as

$$g(m,k) = \int \exp(-k\, K[p(j \mid m^*, \theta^*), p(j \mid m, \theta)])\, \pi(\theta \mid m)\, d\theta, \tag{B.10}$$

let us consider the following three cases:

$$1)\ m < m^*$$
$$2)\ m > m^*$$
$$3)\ m = m^*.$$

Based on the above three cases we want to show that in the limit $k \to \infty$,

$$g(m,k) \to 0 \text{ if } m \neq m^* \tag{B.11}$$

from which it follows

$$\pi(m \mid c_k) \to 0 \text{ if } m \neq m^*. \tag{B.12}$$

$$\pi(m \mid c_k) \to 1 \text{ if } m = m^* \tag{B.13}$$

Case 1: $m < m^*$ **.**

In this case the entropy distance $K[p(j \mid m^*,\theta^*), p(j \mid m,\theta)]$ amounts to

$$\sum_{j=0}^{m^*} p(j \mid m^*,\theta^*) \log\left[\frac{p(j \mid m^*,\theta^*)}{p(j \mid m,\theta)}\right] = \infty \tag{B.14}$$

as $p(j \mid m,\theta) = 0$ for $j > m$. Consequently, $g(m,k)$ is equal to zero if $m < m^*$. It follows from (B.09) that $\pi(m \mid c_k) \to 0$ if $m < m^*$

□

Case 2: $m > m^*$ **.**

As to the entropy distance $K[p(j \mid m^*,\theta^*), p(j \mid m,\theta)]$ we have that

$$\begin{aligned}
&\sum_{j=0}^{m^*} p(j \mid m^*,\theta^*) \log\left[\frac{p(j \mid m^*,\theta^*)}{p(j \mid m,\theta)}\right] \\
= &\sum_{j=0}^{m^*} p(j \mid m^*,\theta^*)(-) \log\left[\frac{p(j \mid m,\theta)}{p(j \mid m^*,\theta^*)}\right] \\
> &-\log\left[\sum_{j=0}^{m^*} p(j \mid m^*,\theta^*)\left[\frac{p(j \mid m,\theta)}{p(j \mid m^*,\theta^*)}\right]\right] \\
= &-\log\left[\sum_{j=0}^{m^*} p(j \mid m,\theta)\right] > -\log\left[\sum_{j=0}^{m} p(j \mid m,\theta)\right] = -\log(1) = 0,
\end{aligned} \tag{B.15}$$

where we from the second to the third line have applied Jensen's Inequality for a strictly convex function. Consequently, $K[p(j \mid m^*,\theta^*), p(j \mid m,\theta)]$ is strictly positive for *all* values of θ. To illuminate the behaviour of $K[p(j \mid m^*,\theta^*), p(j \mid m,\theta)]$ as a function of θ note that

$$\frac{\partial K[p(j \mid m^*,\theta^*), p(j \mid m,\theta)]}{\partial \theta} = -\frac{m^*\theta^*}{\theta} + \frac{m - m^*\theta^*}{\theta} \qquad \textbf{(B.16)}$$

which reveals that $K[p(j \mid m^*,\theta^*), p(j \mid m,\theta)]$ has one extremum located at

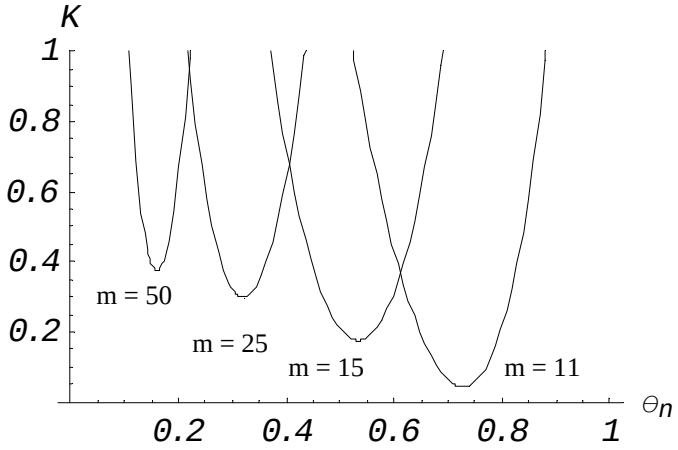
$$\theta = \frac{m^*\theta^*}{m}. \qquad \textbf{(B.17)}$$

The second derivative of $K[p(j \mid m^*,\theta^*), p(j \mid m,\theta)]$ evaluated at the extremum point takes the value

$$\left.\frac{\partial^2 K[p(j \mid m^*,\theta^*), p(j \mid m,\theta)]}{\partial \theta^2}\right|_{\theta=\frac{m^*\theta^*}{m}} = \frac{m^3}{m^*\theta^*(m - m^*\theta^*)} \qquad \textbf{(B.18)}$$

which is strictly positive for $m > m^*$. Hence $K[p(j \mid m^*,\theta^*), p(j \mid m,\theta)]$ has a well defined minimum as illustrated in fig. 14.06 below.

Fig. 14.06. $K = K[p(j \mid 10, 0.8), p(j \mid m,\theta)]$ as a function of θ for different values of m.



Let $K_{\min}$ denote the minimum of $K[p(j \mid m^*,\theta^*), p(j \mid m,\theta)]$. As $K_{\min} > 0$ we have for $k \geq 0$,

$$
\begin{aligned}
&\int \exp(-k\,K[p(j \mid m^*,\theta^*), p(j \mid m,\theta)])\,\pi(\theta \mid m)\,d\theta \\
&\leq \int \exp(-k\,K_{\min})\,\pi(\theta \mid m)\,d\theta \\
&= \exp(-k\,K_{\min})
\end{aligned}
\tag{B.19}
$$

from which it follows that $\lim_{k\to\infty} g(m,k) = 0$ when $m > m^*$. It follows from (B.09) that $\lim_{k\to\infty} \pi(m \mid c_k) = 0$ if $m > m^*$. □

Case 3: $m = m^*$.

The two previous cases showed that $\lim_{k\to\infty} \pi(m \mid c_k) = 0$ if $m \neq m^*$. As $\sum_{m\in M} \pi(m \mid c_k) = 1$, it follows that $\lim_{k\to\infty} \pi(m \mid c_k) = 1$ when $m = m^*$.

Note that $g(m^*,k)$ is different from zero in the limit $k \to \infty$ because the entropy distance $K[p(j \mid m^*,\theta^*), p(j \mid m^*,\theta)] = 0$ when $\theta = \theta^*$. That is, the entropy distance is not strictly positive.

□

We now return to the sum from (B.01):

$$
\begin{aligned}
&\sum_{c_k} p(c_k)[\sum_{m\in M} \pi(m \mid c_k) \log \pi(m \mid c_k)] \\
&= \sum_{c_k} p(c_k)[\pi(m^* \mid c_k) \log \pi(m^* \mid c_k) + \sum_{m\neq m^*} \pi(m \mid c_k) \log \pi(m \mid c_k)].
\end{aligned}
\tag{B.20}
$$

As $\pi(m \mid c_k) \to 0$ if $m \neq m^*$ and $\pi(m \mid c_k) \to 1$ if $m = m^*$, it follows that

$$
\pi(m^* \mid c_k) \log \pi(m^* \mid c_k) \to 0, \tag{B.21}
$$

$$
\sum_{m\neq m^*} \pi(m \mid c_k) \log \pi(m \mid c_k) \to 0. \tag{B.22}
$$

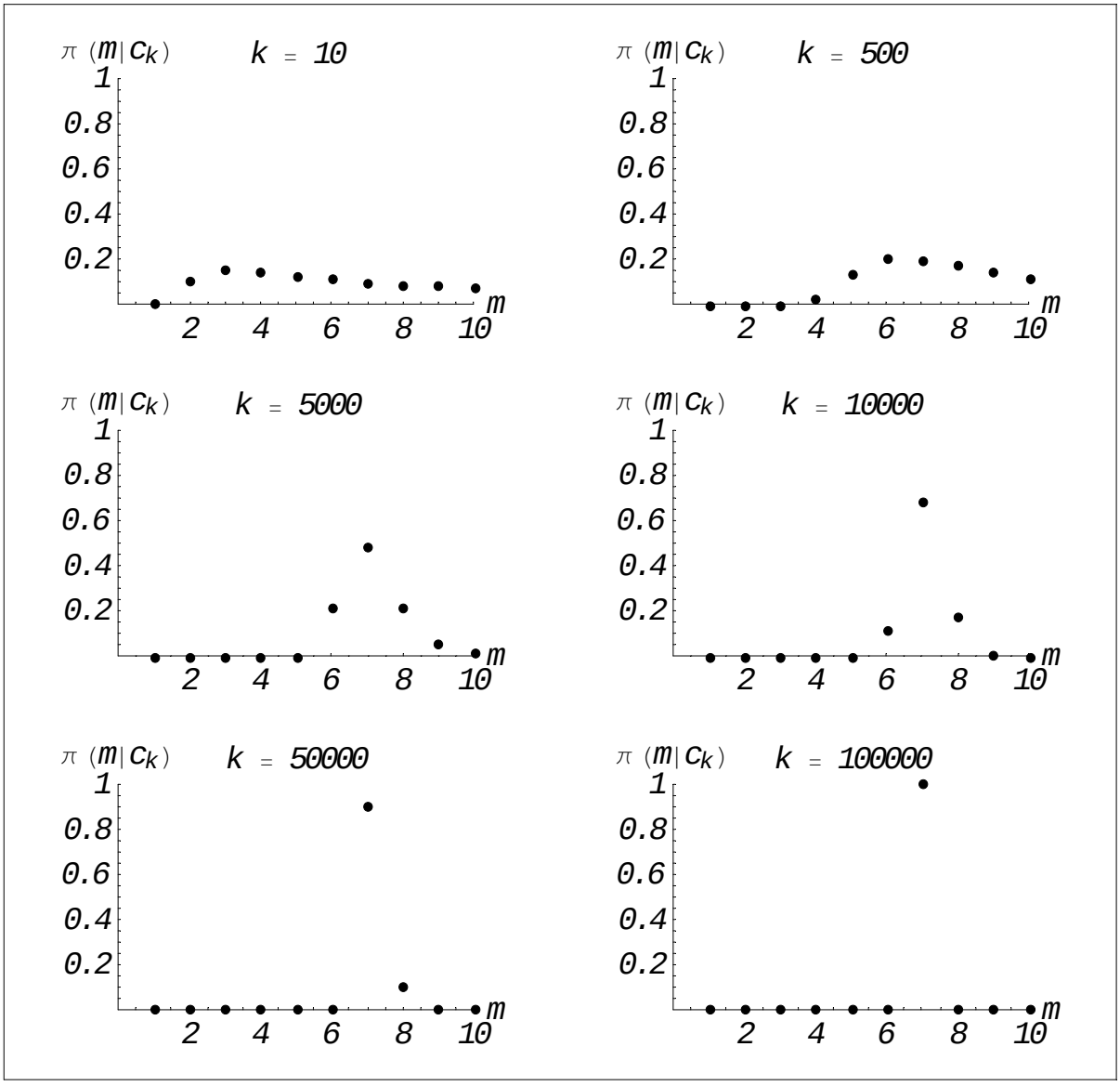
Consequently (B.20) goes to zero in the limit $k \to 0$, as we set out to prove.

□

Numerical study

To test the validity of the proof stated above, a simple numerical study was conducted in the following way: By use of a random number generator a sequence of vectors c_k of increasing length were build from sampled observations $Z \sim Bi(m^*,\theta^*)$. The binomial parameters (m^*,θ^*) were randomly chosen and set to the values (7, 0.18). Fig. B.1 below illustrates the posterior $\pi(m \mid c_k)$ for different values of m and for increasing values of k. As expected, $\pi(7 \mid c_k) \to 1$ for increasing values of k.

Fig. B.1 Simulation study: $\pi(m \mid c_k)$ for increasing values of k. $\pi(m) = \frac{1}{10}$, $\pi(\theta \mid m) = Be(1,1)$ for all m.



Note: The posterior distribution $\pi(m \mid c_k)$ in the numerical study was calculated from the formula →

$$\pi(m \mid c_k) = \frac{\pi(m)\int_0^1 p(c_k \mid m,\theta)\,\pi(\theta \mid m)\,d\theta}{p(c_k)} \propto \pi(m)\prod_{i=1}^{k}\binom{m}{z(i)} Beta(1+\sum_{i=1}^{k} z(i), 1+km-\sum_{i=1}^{k} z(i)) \tag{B.23}$$

References

1) Bajic, M. & Sambunjak, R. (2003): "Empirical statistical model of the density distribution of landmines and UXO", Int. Conf. Requirements and Technologies for the Detection, Removal and Neutralization of Landmines and UXO, 15-18 sept. 2003, Brussels, Belgium
2) Berger, James O. (1980): "Statistical Decision Theory", Springer-Verlag.
3) Bernardo, J.M. (1979): "Reference posterior distributions for Bayesian inference", J. Roy. Statist. Soc. B **41**, pp. 113-147.
4) Bernardo, J.M. and Smith, Adrian F. M. (1994): "Bayesian Theory", John Wiley & Sons
5) Box, George E. P. and Tiao, George C. (1973): "Bayesian Inference in Statistical Analysis", Addison-Wesley.
6) Canadian International Demining Corps & Paul F. Wilkinson & Associates Inc. (2001): "Landmine Impact Survey – Republic of Mozambique"
7) Clarke, B.S. and Barron, A. R. (1990): "Information-theoretic asymptotics of Bayes methods", IEEE Trans. Inform. Theory, vol. **36**, no.3, pp. 453-471
8) Crouch, A.C & Spiegelman, Donna (1990): "The Evaluation of Integrals of the Form $\int_{-\infty}^{+\infty} f(t)\exp(-t^2)\,dt$: Application to Logistic-Normal Models", Journal of the American Statistical Association, Vol. 85, No. 410
9) Gelman A. , and Rubin, D (1992):"Inference from Iterative Simulation Using Mulitple Sequences", Statistical Sciences, Vol. 7, no. 4, pp. 457-511
10) Gelman, A., Carlin, John B., Stern, Hal S. and Rubin, Donald B. (2003): "Bayesian Data Analysis", Chapman & Hall/CRC, 2nd edition
11) Geweke, John (1989): "Bayesian Inference in Econometric Models Using Monte Carlo Integration", Econometrica, Vol. 57, no. 6
12) GICHD (Geneva International Centre for Humanitarian Demining, 2001): "A Study of Soci-Economic Approaches to Mine Action", www.gichd.org.
13) GICHD (Geneva International Centre for Humanitarian Demining, 2004): "A Guide to Mine Action", 2nd edition, www.gichd.org.
14) GICHD (Geneva International Centre for Humanitarian Demining,2002): "Socio-Economic Approaches to Mine Action – An Operational Handbook", www.gichd.org.

15) Gilks, W.R, Richardson, S., and Spiegelhalter, D.J. (1996): "Markov Chain Monte Carlo in Practice", Chapman & Hall

16) Handicap International: "The Use of Mechanical Means for Humanitarian Demining Operations", edit. Phil Paterson, 2000.

17) Hastings, W.K. (1970): "Monte Carlo sampling methods using Markov chains and their applications", Biometrika, 57, 1, p. 97

18) Jacobs, R.A., Jordan, M. I., Nowlan, S.J., and Hinton, G. E. (1991): "Adaptive Mixutures of Local Experts", Neural Computation 3, pp. 79-87

19) Jeffreys, H. (1946): "An invariant form for the prior probability in estimation problems." Proc. R. Soc. London A, **186**, 453-461.

20) Jordan, M. I., and Jacobs, R. A. (1992): "Hierarchies of Adaptive Experts", in Advances in Neural Information Processing Systems 4, J. Moody, S. Hanson, and R. Lippmann (Eds.), San Mateo, California, Morgan Kaufmann, pp. 985-993.

21) Kullback, Solomon (1959): "Information Theory and Statistics", John Wiley & Sons, New York.

22) Laird, Nan M. & Louis, Thomas A. (1989): "Empirical Bayes Ranking Methods", Journal of Educational Statistics, Vol. 14, no. 1, pp. 29-46

23) Lehmann, E.L. and Casella, George (1998): "Theory of Point Estimation", Springer.

24) Lowrance, William W. (1976): "Of Acceptable Risk", Harvard University, William Kaufmann, Inc., California

25) MacDonald, J. & Lockwood, J.R. (2003): "Alternatives For Landmine Detection", RAND, Science and Technology Policy Institute, Santa Monica.

26) Millard, A. & Harpviken, K. (2001): "Community Studies in Practice", PRIO, International Peace Research Institute, Oslo

27) Millard, A. & Harpviken, K. (2000): "Reassessing the Impact of Humanitarian Mine Action", PRIO, International Peace Research Institute, Oslo

28) Nelder, J.A., and Wedderburn, R.W.M. (1972): "Generalized linear models", Journal of the Royal Statistical Society, A 135, 370-384

29) Richardson, S. & Green, P. J. (1997): "On Bayesian Analysis of Mixtures with an Unknown Number of Components", Journal of the Royal Statistical Society, B 59, no. 4, pp. 731-792

30) Robert, Christian P. (1994): "The Bayesian Choice – A Decision-Theoretic Motivation", Springer.
31) Spiegelhater, D. J., Best, N. G., Carlin, B., P., and van der Linde, A. (2002): "Bayesian measures of model complexity and fit", Journal of the Royal Statistical Society, B 64, part 4, pp. 583-639
32) Stephens, M. (2000): "Bayesian Analysis of Mixture Models with an Unknown Number of Components – an Alternative to Reversible Jump Methods", The Annals of Statistics, Vol. 28, no. 1, pp. 40-74
33) Survey Action Center & Mine Clearance Planning Agency (2000): "Landmine Impact Survey – Republic of Yemen"
34) Tanner, Martin A. (1993): "Tools for Statistical Inference", 2nd edition, Springer Verlag
35) Trevelyan, J. (1997): "Statistical Analysis of Minefield Clearance Data", Technical Report, Department of Mechanical and Materials Engineering, The University of Western Australia
36) Unitied Nations (1994): "Assistance in Mine Clearance", General Asssembly, doc. A/49/357, september 6
37) Yang, R. and Berger, J.O. (1998): "A Catalog of Noninformative Priors".